Topics on Regression Percentiles

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

Sen Yuan

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

Snigdhansu Chatterjee, Adviser

May 2014

Acknowledgements

There are many people that have earned my gratitude for their contribution to my time in School of Statistics. I would first thank my adviser, Professor Snigdhansu Chatterjee who initialed my research interest on regression quantiles and irregular problems in M-estimation. Doing research under his supervision was great pleasure. Also I would like to thank NSF Expeditions in Computing: Understanding Climate Change (A Data Driven Approach) who offered my research assistant stipends from Jan. 2012 to Jan. 2014. Last but not least, I am grateful for the professors, staffs, students in School of Statistics who helped me in every aspect of my life here.

Dedication

To my parents: Hua Yuan and Leyuan Han

Abstract

The task of statistical regression is to learning conditional distribution of response given predictors. To learn the conditional distribution directly is hard so people instead focus on the functionals of the conditional distribution. Traditionally people model conditional mean which is an intuitive and useful functional of the conditional distribution. Nevertheless the conditional mean can hardly capture a full picture of the conditional distribution, for instance, distributional tail behaviors. Quantile Regression (QR) [40] and Expectile Regression (ER) [19] are introduced by Koenker and Efron respectively to learn the regression percentiles which provide broader views than conditional mean regarding gaining insights about the conditional distribution. In this thesis, we propose a new boosting algorithm to learn regression percentiles under the context of QR. Also, we provide LARS-like variable selection strategies for ER and provide the solution path. Finally, irregular problems of M-estimation is studied in this thesis. We discuss its connections to extreme values and some recent algorithms for solving QR.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Regression is about learning the conditional distribution of response variables Y given predictors X. This task is tough because estimating the whole distribution requires too much information. Imaging if we collect 100 samples and predictor values are all unique in these 100 samples. Therefore given a particular predictor value (multidimensional) we only have one response value. This unique value attached to a single predictor sample can be barely sufficient to reconstruct a sensible conditional distribution. Thus people came up with a simplified idea which is modeling the conditional mean instead of the whole distribution. Each response value corresponds to the unique predictor value can be collectively used for estimating the conditional mean. Some people realize although modeling mean is a brilliant idea for location learning but can be potentially problematic for what is beyond the mean. There are lots of situations where people need broader information than mean. To model the risk of extreme, rare events for instance is one such occasion. Specifically, consider if we model the annual highest sea level in Shanghai, it is not reasonable if the model is to predict the mean of the next few years. In extreme value theory, we know extreme values does not behave the same as mean asymptotically. So if we use the model of conditional mean, it can lead to very large deviation. Therefore, studying extreme of the distribution sounds like equally important as studying the mean. Indeed, there

are huge amount of literatures discussing modeling the extreme values. In biostatistics, motivating examples include the analysis of survival at extreme durations [41]. In finance, people use extreme value theory as a tool for measuring financial risk [26]. In climate science, extreme events like hurricanes, flood, draught are things of particular interest [36]. What goes beyond the mean is not limited to the extreme of the distribution but in every percentile of one's interest. Suppose one needs a more robust way to learn about the "center" part of the conditional distribution, one can use "median regression" to prevent against the outliers. So the interest for the percentile of the conditional distribution instead of mean produces the idea of quantile regression [40]. "In ecology, quantile regression has been proposed and used as a way to discover more useful predictive relationships between variables in cases where there is no relationship or only a weak relationship between the means of such variables. The need for and success of quantile regression in ecology has been attributed to the complexity of interactions between different factors leading to data with unequal variation of one variable for different ranges of another variable" [11]. There are huge amount of papers about QR and its application. For example, Koenker [38] gives comprehensive introduction to QR. Recently, Meinshausen [43] brought in random forest for predicting conditional quantiles. Additive models are also involved in quantile regression [39]. As for high dimensional settings, Belloni A. and Chernozhukov V. [3] developed results of L1 penalized QR in high dimensional sparse models. Wang, L. [55] has proposed QR for analyzing heterogeneity in ultra-high dimension.

Parallel with QR, Expectile Regression (ER) is another tool for studying regression percentile introduced by Efron [19]. Newey and Powell [45] pointed out that estimating conditional expectile is also effective for studying regression percentile. The quantile loss function is $\phi(x, \tau) = |xI(x \leq 0) - \tau x|$ while expectile loss is $\psi(x, \omega) = |I(x \leq 0)x^2 - \omega x^2|$. Compared to quantile q_τ which specifies the position below which $100\tau\%$ of the probability mass of X lies, expectile e_τ determines the

position $100\tau\%$ of the mean distance between it and X comes from the mass below it. Therefore expectile relies on the distance of observations at the price of increasing the outlier sensitivity. For this reason, it has been claimed that expectile uses the data more efficiently than quantile. The connections between QR and ER has been discussed in many literatures. The QR and ER have their own advantages over each other. For example, QR is more robust than ER. Newey and Powell [45] stated that expectile regression has two major advantages over quantile regression: 1) Fast Computation: the ALS loss is differentiable everywhere while the check loss is singular at zero. 2) the calculation of the asymptotic covariance matrix of the multiple linear expectile regression estimator does not involve calculating the values of the density function of the errors. As for the application of ER, Taylor [53] has already used ER for estimating Value at Risk and Expected Shortfall in Finance. He argued that using expectile has the appeal of avoiding distributional assumptions.

In this thesis, we will discuss both QR and ER. In chapter 2, we propose a new boosting algorithm called "Random Gradient Boosting" (RGB) for studying QR. Gradient Boosting was introduced to address both classification and regression problems with great power. The statistics community studied the boosting with L2 loss intensively both in theory and practice. However, the L2 loss is not proper if the problem is about learning quantities other than conditional mean such as conditional quantiles. There are huge amount of literatures studying quantile regression with various methods including machine learning techniques like random forest and boosting.The success of both random forest and gradient boosting brings in a natural question: Why not combine them together ? The RGB embraces the merits of both random forest and gradient boosting. Then, we compare the empirical results between this new method and some competitive ones. Finally, we provide some reasonings to support the fact that Gradient boosting is a rational method when it comes to predicting conditional quantiles.

In chapter 3, we comprehensively study how data mining algorithms like random forest and boosting brings statistical power in ER. Besides this, the variable selection strategies will be proposed by using LARS-like algorithms. Li and Zhu (2007) have developed a regularization path algorithm for QR based on solving KKT equations. However this algorithm cost time of complexity order $O(n^3p)$ when $n > p$. This is not as fast as conventional LARS algorithm for OLS which only requires $O(n^2p)$. What is more crucial is that Lasso though has its strong power to introduce sparsity, but may not work well when strong colinearity presents between predictors, see Zou (2005) [59], Hastie (2006) [29]. The alternative approach to take account the full picture of the conditional distribution is ER. We will show that LARS-like algorithms can be carried out with worst time of order $O(n^2p)$ if $n > p$. Plus with small modification for LARS algorithm we can obtain both Lasso regularization path and forward stage-wise regression. The later one usually outperforms lasso if collinearity exists. The LARS-like algorithms can also be developed for a fused lasso penalty for ER.

In chapter 4, we will present the irregular problem of M-estimation. With this abstract framework, we can somehow better understand the asymptotic behavior of extreme regression percentiles. M-estimators represent a broad class of estimators by minimizing the sum of functions of data point. That is, $\hat{\theta} = \arg\min_\theta \sum_{i=1}^n \rho(X_i, \theta)$, It was first introduced by Peter Huber [32] to study robust estimators and their relevant asymptotic properties and has been studied extensively for its theoretical properties. See, for example, Huber(1964, 1972) [32] [33], Portnoy(1977) [48], Collins(1976) [17], Freedman and Diaconis(1982) [21], Niemiro(1992) [46], He and Shao(1996) [31]. One can use quantile loss function as objective function for M-estimation. In this chapter, we consider a type of M-estimator defined by minimizing a sequence of convex functions with a control parameter that varies according to sample size. This problem is a natural extension of Niemiro's studies(1992) [46] on asymptotics for M-estimators. The additional control parameter has its practical meaning in

various application problems. For example, Colin Chen (2007) [15] proposed a finite smoothing algorithm for quantile regression which matches our formulation where the control parameter plays a role of controlling the distance between their objective function and the quantile objective function. Another motivation is from studying the irregular problems in M-estimators where asymptotic Hessian matrix is singular. This problem is important for understanding multivariate extreme quantile studies. Our approach to tackle this problem is to "slow down" the multivariate extremes convergence by limiting the converging rate of control parameters. We established the strong and weak convergence results for both regular and irregular problems. Moreover, we provided the upper bound for the converging speed of control parameter to force the asymptotic normality.

In chapter 5, we will discuss the variable selection power of regression percentile based gradient boosting. We compare it with other competitive variable selection methods for the high dimensional heteroskedasitic model. More importantly, we will show its power working with real data. Most variable selection methods are based on linear model, when it comes to real data it may fail due to the complex nature of real world data. Boosting with tree learner is a step function approximation to learning curve. For this reason, it may have the strength to capture the structure that can be hardly captured by linear models. In the last section, we propose a Lars-like algorithm for expectile regression called ERLars.

Chapter 2

Gradient Boosting for Quantile Regression

2.1 Overview

he AdaBoost Algorithm proposed [22] by Freund and Schapire is one of the most attractive methods for classification in the machine learning community, due to its good performance practically with a variety of datasets. The AdaBoost essentially ensembles a bunch of weak learners to form a strong one by assigning the weights to data and each base learner according to its classification error in that round. Later, It is noticed by Breiman and Friedman that AdaBoost algorithm, from another perspective, can be viewed as a gradient based line search optimization procedure with respect to the underlying functional. Friedman proposed Functional Gradient Boosting [23] [24] which targets at regression problems of different kinds. For example, under the L2 loss, the gradient boosting is to construct a model to study conditional mean. This so called "L2 Boosting" has been thoroughly studied by Buhlmann and Yu [9] [6] [10] . Even though there is no satisfactory answer as for why gradient boosting does well empirically in general, however, people have already did extensive research when the loss is L2 type. Zhang and Yu [57] even extend the L2 loss to a class of loss functions (not include quantile loss) to study the consistency of boosting

algorithm with early stopping.

In summary, Gradient boosting demonstrates its power in the regime of predicting the mean. If one can study conditional mean, one can naturally ask can we apply gradient boosting to conditional quantile ? There are literatures about applying machine learning techniques to study quantile regression. For example, Meinshausen (2007) [43] studied quantile regression and found its good performance by applying a random forest based method. That method is named "Quantile Regression Forests"(QRF). The intuition is that for every fixed value y and ith sample (Y_i, X_i), one can use $I(Y_i < y)$ as response input for random forest algorithm. After this procedure, we obtain an estimate of conditional distribution $F(y|x)$, let's say, $\hat{F}(y|x)$. Then we take the quantiles of $\hat{F}(y|x)$ as our estimate for conditional quantile. Mathematically, the conditional distribution function of Y given X can be expressed as:

$$\mathbf{F}(y|X = x) = \mathbf{P}(Y \leq y|X = x) = \mathbf{E}(I_{(Y \leq y)}|X = x)$$

So one can see the connection between random forest for mean and for distribution. One the other hand, Gradient boosting has been used in practice to deal with quantile modeling problems. Hothorn [20] used gradient boosting to study quantile regression empirically. They claimed that boosting is an appropriate tool for estimation in linear and additive quantile regression models. The reason for their statement is: (i) flexibility in estimating nonlinear effects; (ii) The variable selection and model selection are implicitly supported when using boosting for model estimation. They did not provide any theoretical analysis. In this paper, we will make up this blank by providing a theoretical explanation as for why boosting algorithm is reasonable tool for estimating quantile. This chapter compromises two major contributions: (i) a new algorithm called Random Gradient Boosting (RGB) which performs well empirically; (ii) theoretical results for quantile gradient boosting. In the next section,

we will present the overview of functional boosting algorithms and the key results from Bulhmann and Yu. In section 2.3, we explain why quantile regression is important. In section 2.4, RGB algorithm is presented along with its numerical comparison to three competitive quantile regression methods including: Quantile Regression (QR), Quantile Regression Forest (QRF), Gradient Boosting (GB). In section 2.5, we offer theoretical justification of using boosting as a tool for quantile modeling.

2.2 Functional Gradient Boosting

Breiman [5] is the first person who realize boosting can be viewed as an optimization algorithm functionally. This observation brings the topic of boosting to statistics community. Here we briefly outline the algorithm of functional gradient boosting. Suppose $L(.,.) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ is the convex loss function. Our goal is to estimate the model function $F : \mathbb{R}^p \to R$ by minimizing the expected loss $\mathbf{E}[L(Y, F(X))]$ based on data $(Y_i, X_i)(i = 1, .., n)$. So the estimator $\hat{F}$ is obtained by minimizing sample version : $n^{-1} \sum_{i=1}^{n} L(Y_i, F(X_i))$. However, without any constraints, the minimizer will be as simple as $\hat{F}(X_i) = Y_i$. In order to learn the model structure rather than "fitting", one assume the true function has additive structure in each small component: $F(x) = \sum_{i=0}^{\infty} h(x, \theta_i)$, where θ is the parameter for each base learner. For example, θ is regression coefficients when the base learner is OLS or if the base learner is tree then θ describes the variable to be split, splitting point and fitted value for each leaf. Now the task is to estimate each component $h(x, \theta_i)$. Because if F minimize $\mathbf{E}[L(Y, F(X))]$, then each $h(x, \theta_i)$ plays a role very similar to negative gradient of $\mathbf{E}[L(Y, F(X))]$ times the optimal step length. If one just map functional gradient of $\mathbf{E}[L(Y, F(X))]$ with its sample version to each component $h(x, \theta_i)$, then θ_i can be estimated according this mapping criterion, for instance, least square. The generic functional gradient boosting works as follows, see [23] [6].

Generic Functional Gradient Boosting

Step 1 (Initialization) Initialize $\hat{F}_0(x)$ from a rough estimation of the function: $\hat{F}_0(x) = h(x, \hat{\theta}_0)$ where $\hat{\theta}_0 = \arg\min \sum_{i=1}^{n} (Y_i - h(X_i, \theta))^2/n$, set iteration number $m = 0$.

Step 2 (Projecting negative gradient to base learner) Calculate the negative gradient:

$$d_i = -\frac{\partial L(Y_i, F)}{\partial F}|_{F=\hat{F}_m(X_i)}, i = 1, ..., n \ .$$

Then project the vector $(d_1, ..., d_n)$ to the base learner using least square:

$\hat{\theta}_{m+1} = \arg\min n^{-1} \sum_{i=1}^{n} (d_i - h(X_i, \theta))^2$. Then $\hat{h}_{m+1}(x) = h(x, \hat{\theta}_{m+1})$.

Step 3 $\hat{F}_{m+1} = \hat{F}_m + \nu_{m+1}\hat{h}_{m+1}$ where ν_{m+1} is learning rate. One possible option is optimal step size.

Step 4 m= m+1 and repeat steps 2) and 3).

Notice that for the above algorithm, the mapping is defined by the rule of least square. However, the mapping rule is not limited to minimizing L2 norm. People can choose different loss function $L(.,.)$ to serve different purposes. We list some of the most popular lost functions:

(1) $L(y, f) = \exp(yf)$ with $y \in \{-1, 1\}$: loss function for AdaBoost;

(2) $L(y, f) = \log_2(1 + \exp(-2yf))$ with $y \in -1, 1$: loss function for LogitBoost;

(3) $L(y, f) = (y - f)^2/2$ with $y \in \mathbb{R}$: loss function for L2Boost;

(4) $L(y, f) = |y - f|/2 + (\tau - 1/2)(y - f)$ with $y \in \mathbb{R}$: loss function for QuantileBoost.

When the Loss is $L(y, f) = (y - f)^2$, the above boosting algorithm is called L2Boost by Buhlmann and Yu [10]. In their paper, they contributed the convergence rate of the estimator and found a bias-variance trade off. In the theoretical parts of their work, they assume the eigenvalues of the fitting operator S which maps y to $\hat{y}$ should be all positive which constrain the scope of their theorem. For instance, the hat matrix $S = X(X^TX)^{-1}X^T$ have some zero eigenvalues if $n > p$. Zhang and Yu [57] did very generic analysis of gradient boosting under a wide range of loss functions. However type (4) is not included in their paper. In this article, we focus on the loss function (4) which generate quantile estimation. In part 3, we proposed a novel algorithm called random gradient boosting which empirically dominates some other competitive methods for quantile prediction. In section 4, we develop theories under loss type (4).

2.3 Quantile Regression

Quantile Regression (QR) is to model the conditional quantile of distribution Y given X. Let us denote the τth conditional quantile as $Q_\tau(x) = \inf\{y : F(y|X = x) \geq \tau\}$.

So the conditional quantiles give more comprehensive information than the conditional mean alone. We will list three occasions where the QR is superior:

1) **Interaction with predictors** There are cases that the changes in the means of response cannot be well connected with the variability of predictors which limit the discovery of many factors if the conditional mean model is used. Many ecological applications would prefer QR for this reason. Authors like Terrell (1996) [54] , Cade (1999) [12] and Huston (2002) [35] suggested that if ecological limiting factors act as constraints on organisms, then the estimated effects for the measured factors were not well represented by changes in the means of response variable distributions. Yet there may exist stronger predictive association with other parts of the response variable distribution. Obviously, QR is nice tool model other parts of distribution instead of mean.

2)**Prediction Intervals** In statistical inference, we do not only satisfy with a single estimation. It is desirable to know a range with high credibility that a predicting value will fall into. This range is called prediction interval. QR friendly provides information about the prediction interval. Imagine if you build models on 0.025th and 0.975th conditional quantile $Q_{0.025}(x), Q_{0.975}(x)$, you can obtain the 95% prediction intervals for the new coming sample x_{new} by combining two predicted quantiles: $[Q_{0.025}(x_{new}), Q_{0.975}(x_{new})]$.

3)**Outlier Detection** The outlier in statistics basically means the sample deviates extremely from the predicted results of the model. Nevertheless, there is no universal standard for detecting outliers. Traditionally people would compare the predicted value with the median of the conditional distribution to check how large the gap is. The magnitude of gap can be measured by some robust distance metric

such as median absolute deviation or interquartile range. Luckily, QR offers sufficient numbers for both metrics.

Modeling Conditional Quantiles

Like linear regression can be viewed as an optimization procedure of minimizing ordinary least square. The Solution for QR can be cast into optimization as well. The quantile loss function $\phi(x, q, \tau) = |(x - q)I(x \leq q) - \tau(x - q)|$. It is not hard to check the conditional quantile $Q_\tau(x)$ minimizes the quantile loss function. One can parameterize $Q_\tau(x) = \beta_0 + \beta_1 X_1 + ..\beta_p X_p$ for instance. By doing so, the estimation for β vector is worked out through optimizing the quantile loss function. And this problem is essentially a linear programming problem. One can also model the conditional quantiles by using some non-parametric approaches such as smoothing splines (He 1998 [30]) , additive models or tree-based models (Chaudhuri 2002 [13]).

2.4 Random Gradient Boosting

The method we propose in this paper is called "Random Gradient Boosting" (RGB). This boosting procedure assumes tree base learner. The difference between RGB and ordinary Gradient Boosting lies in how we grow the tree. First, we do not implement tree as recursively growing tree like CART. The reason is because Gradient Boosting works best if the tree size is reasonably small. People usually limit the number of nodes under 15 to achieve best results. If the tree size is limited, then growing all leaves together would increase the size too fast. Alternatively, we seek to find the "winner" leaf which shows its optimality to be split. The optimality is defined as information gain, such as variance reduction, cross entropy gain etc..

Second, we treat each predictor unequally when growing the tree. Specifically,

we have "weights" assigned to each predictor as an indicator of their importance to response variable. The importance can be evaluated by different methods. For example, one can use any ensemble learning methods with tree base learner in the first stage. Then the variable importance can be calculated by taking account of the magnitude of variance reduction one particular variable contributes. There is, however, a simpler method which is calculating the marginal correlation between each predictor and the response. Then we scale correlation vector to make it probability distribution. Finally, each time when we need candidate set for splitting tree, we randomly to m out of p predictor under this probability distribution. This idea is borrowed from Random Forest. Nevertheless the unequal weights is our own ingredient. The motivation of "randomness" is to help reduce correlation between each tree in random forest context. In RGB, however, we are not intending to reduce the correlation. In fact, we use it as a way to accomplish variable screening. We increase the chance of involving highly correlated variables. We outline the random gradient boosting with quantile loss:

Random Gradient Boosting

Step0(Preprocessing) Calculating correlations $\rho_i = \text{corr}(Y, X^{(j)}), j = 1, .., p$.Then generate the probability distribution $w_i = |\rho_i|/(\sum_{i=1}^{n} |\rho_i|), i = 1, ..., n$.

Step1(Initialization) Use the τth quantile of $(Y_1, ..., Y_n)$ as initial estimation $\hat{F}_0(x)$.

Step2(Projecting negative gradient to base learner) Get negative gradient:

$d_i = -\frac{\partial L(Y_i,F)}{\partial F}|_{F=\hat{F}_m(X_i)} = \tau I(y_i \geq \hat{F}_m(X_i)) + (\tau - 1) I(y_i < \hat{F}_m(X_i)), i = 1, ..., n$

Fitting $d = (d_1, ..d_n)$ to a decision tree according to the growing rule:

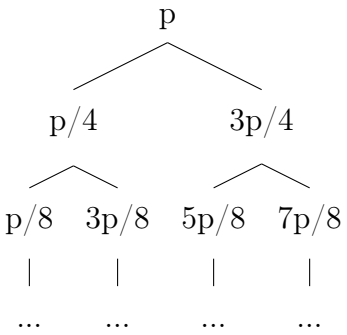
- Randomly select m out p predictors $(X^{(j_1)}, ..X^{(j_m)})$ as splitting candidates according to distribution $w = (w_1, ..w_p)$ in *Step0.*
- Split the tree by finding the best predictor and splitting points pair (k, l) in that:

  $(k, l) = \arg\min_{j,s}[\sum_{X_i \in R_1(j,s)}(d_i - c_i) + \sum_{X_i \in R_2(j,s)}(d_i - c_i)]$, where $R_1(j, s) = \{X_i^{(j)}; i = 1, .., n | X_i^{(j)} \leq X_s^{(j)}\}$, $R_2(j, s) = \{X_i^{(j)}; i = 1, .., n | X_i^{(j)} > X_s^{(j)}\}$, c1 and c2 are τth quantile in $R_1(j, s)$ and $R_2(j, s)$ respectively.
- Do (i) first and fix the candidate set. Running through all leaves, do splitting procedure (ii) for each leaf. Find the optimal leaf which generates maximum information gain: either maximum increase of Cross-entropy or reduction of variance.
- Repeat iii) until maximum number of nodes reached or no subsample flowing down in optimal leaf. Denote the fitted tree is $\hat{h}_{m+1}$

Step 3 $\hat{F}_{m+1} = \hat{F}_m + \nu_{m+1}\hat{h}_{m+1}$ where ν_{m+1} is learning rate. One possible option is optimal step size.

Step 4 m= m+1 and repeat steps 2) and 3).

In *Step*0, we use correlation as weights for each variable in a hope to screen out noisy variables. The choice of weights can extend to any sensible variable relative influence measurements such as variable influence from gradient boosting model. However, bias selection may present if one use the default variable importance measurements in CART. The information gain criterion such as variance reduction, cross entropy and gini index may be not fair in favor of predictor variables with more categories. The permutation importance is a generally acknowledged heuristics for correcting the bias caused by information gain criterion, see Altmann [1]. The choice of m is flexible but not as robust as in random forest. In random forest, even $m = 1$ can produce amazing result. In our method, empirical experiments suggest m should be slightly greater than $p/3$. Nevertheless, we will show later the choice of m is in fact rather insensitive to the prediction. There are two suggested strategy for choosing m: 1)cross validation: giving a set of values for m, find the value minimizes the estimated prediction error according to k-folds cross validation; 2) binary search:



As one can see in this tree diagram, one can first compare $m = p/4$ with $3p/4$ to check which is better. Then flow down to the better branch and repeat the thing procedure for that subtree. Stop this process when reaching the bottom or the maximum pre-specified number of rounds. In the rest of this section, we will demonstrate the ef-

fectiveness of this algorithm regarding on learning additive and interaction structures.

We will list two numerical simulation examples to empirically illustrate the good performance of RGB. We will compare RGB to Quantile Regression (QR) from quantreg package: `http://cran.r-project.org/web/packages/quantreg/index.html`, Quantile Regression Forests (QRF) from quantregForest package: `http://cran.r-project.org/web/packages/quantregForest/` and Gradient Boosting (GB) from gbm package: `http://cran.r-project.org/web/packages/gbm/index.html`. We first generate from a simulation model. Then we fit the generated data to different quantile models. Finally, a test data set is generated from the same simulation model for estimating the predicting error. The predictor error is measured the same way as in [43] through a quantile loss function.

Model I: $X = (X_1, ..., X_p)$ $\sim$Normal(0,I)

$$Y = 2X_1 + X_2^2 + 0.25X_3^3 + 0.5 * rnorm(0, 1)$$

Model II: $X = (X_1, ..., X_p)$ $\sim$Unif($[0, 1]^p$)

$$Y = 10\sin(\pi X_1 X_2) + 20(X_3 - 0.5)^2 + 10X_4 + 5X_5 + rnorm(0, 1)$$

The first model involves additive structures as well as quadratic and cubic. The second one is very famous Friedman #1 model used in lots of literatures. It is first used in Friedman's MARS paper in 1991 to test the ability of MARS on identifying interaction effect. For both model, we specify $p = 20$ and training sample size $n = 100$. The testing sample size is also 100. The number of simulation round is 100. The simulation results are presented in terms of table and box plots (Figure 1).

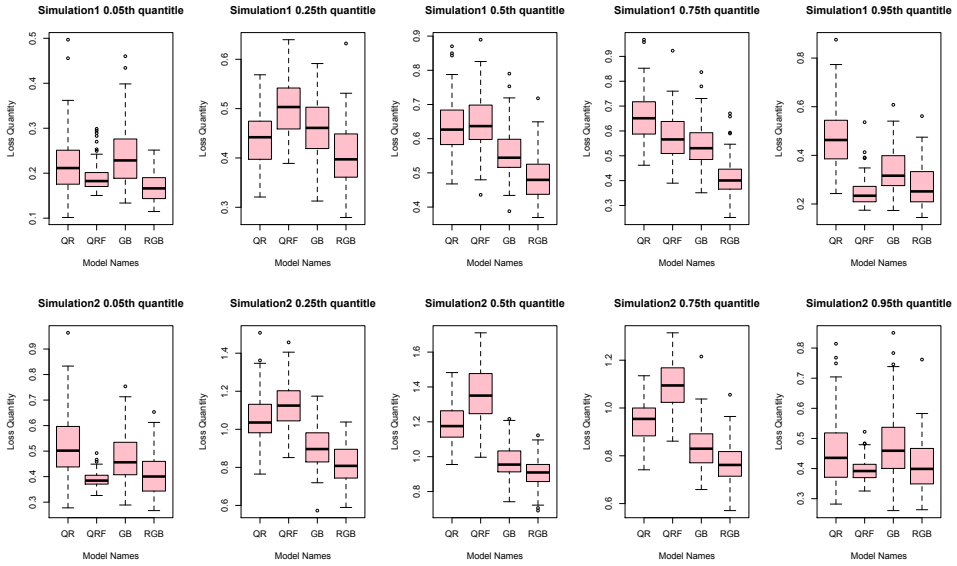


Figure 2.1: n = 100, p = 20

Models	τ	QR	QRF	GB	RGB
I	0.05	0.220(0.067)	0.191(0.032)	0.238(0.069)	0.169(0.032)
	0.25	0.436(0.052)	0.499(0.059)	0.461(0.059)	0.409(0.063)
	0.5	0.635(0.077)	0.649(0.078)	0.556(0.069)	0.484(0.063)
	0.75	0.658(0.097)	0.577(0.088)	0.544(0.089)	0.411(0.073)
	0.95	0.471(0.126)	0.247(0.057)	0.337(0.082)	0.270(0.083)
II	0.05	0.523(0.126)	0.389(0.031)	0.472(0.095)	0.408(0.081)
	0.25	1.053(0.131)	1.120(0.112)	0.909(0.116)	0.822(0.100)
	0.5	1.180(0.112)	1.359(0.149)	0.970(0.093)	0.908(0.081)
	0.75	0.945(0.087)	1.103(0.107)	0.836(0.093)	0.768(0.087)
	0.95	0.455(0.112)	0.395(0.035)	0.477(0.114)	0.411(0.083)

Table 2.1: Prediction Errors

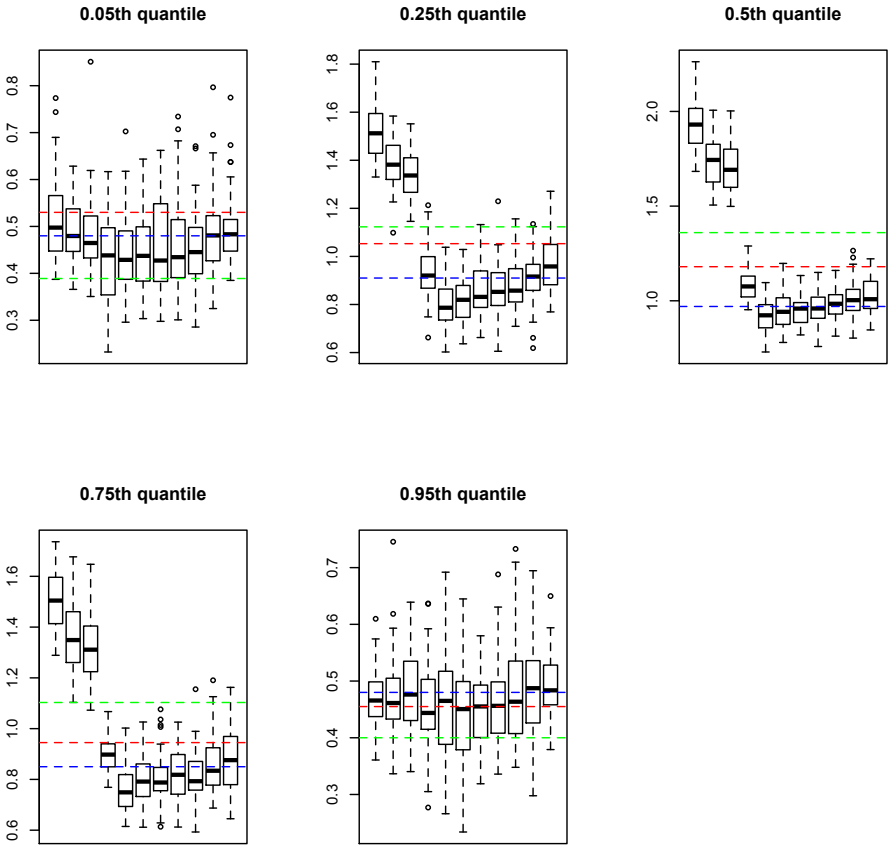


Figure 2.2: Sensitivity for Model II

In the first simulation model, we used default setting for QR and QRF. QRF is claimed to be very robust so using default setting is as good as we can do. As for GB, we used 1000 trees with learning rate 0.1. RGB, we set 1000 trees with only three nodes, accordingly the learning rate is 0.5. We choose $m = 5$ as the number of candidate set. RGB almost dominates all other methods except for quantile at 0.95 where QRF seems best. QRF did very job in extreme value estimation but not that good in quartiles and median. In additive model, GB and QR has similar performance in the first two plots but GB is better in last three. Overall, it is clear RGB is the best.

In the second simulation model, we have the same smoothing parameter setting as in model I. we again observe the competitiveness of RGB in estimating extreme value and its downside in estimating quartiles and median. QR did not perform as well as it did in model I. Maybe it is because the interaction effect that shrinks its power. GB did better job than methods besides RGB. RGB is on top in this model again.

In order to study the role of smoothing parameter m in RGB, we let m ranges from 1 to 10 in simulation model II. We plot the result to check how sensitive it depends on m. As baseline comparison, we also add three dashed lines in each plot. Red, Green and Blue represents the median of simulation results from QR, QRF and GB respectively. Apparently, for upper and lower extremes estimation, the RGB is not sensitive to the change of m. Again, the power of QRF in predicting extreme values is shown. For 0.25th, 0.75th and 0.5th quantiles, besides the first three values of m, the result is basically stable. The QRF is worse among all when it comes to these three quantiles.

2.5 Childhood Malnutrition in India

The malnutrition dataset is downloaded from Demographic and Health Surveys (DHS) website. This organization conduct surveys across 75 countries on fertility, child health, HIV, AIDS, nutrition etc.. There are two important previous research papers that have analyzed this dataset for learning risk factors for malnutrition of children in India. They both focused on modeling lower extreme quantiles of child's height. Fenske, Kenib, and Hothorn (FKH) [20] have adjusted the height for the age effect while Koenker's approach is to keep original height as response variable. Koenker argued that adjusting the effect of age to height seems to presuppose that none of the other predictors matter. In this paper, we adopt Koenker's proposal of keeping height unchanged because we can estimate the age effects altogether with other predictor effects in the model. As for the statistical tool to model the data, FKH used gradient boosting and Koencker used additive quantile regression models. The plots of factor effects from the two different models present most consistency with slight difference on some factors. We will use RGB to analyze the data with three nodes for each tree. Three quantiles $0.05, 0.1, 0.5$ are of our interest. Particularly, the median is selected as a measurement of middle part of distribution. We are curious about different those factors can affect medium than extreme. IAKR52SD is the data folder with data stored in SAS dataset. We choose predictors similar to FKH and Knoenker. All missing values and answers with uncertainty has been cleared out before the data analysis. The variable names and their meanings are listed in the table. And the univariate summary is presented in the table, the first variable X is just the index vector that indicates the indices for observations from the raw data.

Malnutrition
23 Variables **12361 Observations**

X

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
12361	0	12361	25089	1879	5280	13197	24805	35835	44027	48180

```
lowest :     9    11    12    14    15
highest: 51542 51546 51547 51549 51552
```

---

electricity

n	missing	unique	Sum	Mean
12361	0	2	10844	0.8773

---

radio

n	missing	unique	Sum	Mean
12361	0	2	4830	0.3907

---

tv

n	missing	unique	Sum	Mean
12361	0	2	8214	0.6645

---

fridge

n	missing	unique	Sum	Mean
12361	0	2	3653	0.2955

---

bike

n	missing	unique	Sum	Mean
12361	0	2	5562	0.45

---

motorcycle

n	missing	unique	Sum	Mean
12361	0	2	3824	0.3094

---

car

n	missing	unique	Sum	Mean
12361	0	2	840	0.06796

---

religion

n	missing	unique	Mean
12361	0	9	1.626

	1	2	3	4	5	6	7	9	10
Frequency	8171	1734	1912	291	144	61	2	3	43
%	66	14	15	2	1	0	0	0	0

---

telephone

n	missing	unique	Sum	Mean
12361	0	2	2900	0.2346

---

wealth

n	missing	unique	Mean
12361	0	5	3.811

	1	2	3	4	5
Frequency	636	1334	2274	3604	4513
%	5	11	18	29	37

---

no..of.living.children

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
12361	0	12	2.414	1	1	2	2	3	4	5

	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	2431	5558	2581	1025	418	192	84	45	16	3	6	2
%	20	45	21	8	3	2	1	0	0	0	0	0

---

mother.s.age

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
12361	0	34	27.22	21	22	24	27	30	34	36

lowest : 15 16 17 18 19, highest: 44 45 46 47 48

---

gender

n	missing	unique	Mean
12361	0	2	1.486

1 (6349, 51%), 2 (6012, 49%)

---

age.in.years

n	missing	unique	Mean
12361	0	5	2.831

	0	1	2	3	4
Frequency	210	1271	3004	3789	4087
%	2	10	24	31	33

---

lives.with.whom

n	missing	unique	Mean
12361	0	2	0.001942

0 (12355, 100%), 4 (6, 0%)

---

breastfeeding.in.months

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
12361	0	53	16.65	3	6	10	15	24	30	36

lowest : 0 1 2 3 4, highest: 48 51 55 58 60

---

size.of.child.at.birth

n	missing	unique	Mean
12361	0	7	2.995

	1	2	3	4	5	8	9
Frequency	474	2494	7159	1535	551	139	9
%	4	20	58	12	4	1	0

---

height

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
12361	0	641	908.8	735	778	845	915	981	1031	1061

```
lowest :  232  254  350  374  382, highest: 1302 1312 1320 1350 1477
```

mother.s.bmi

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
12361	0	1797	2128	1620	1695	1849	2063	2336	2659	2877

```
lowest : 1291 1300 1327 1334 1338, highest: 4120 4143 4147 5867 5962
```

child.s.age.in.months

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
12361	0	59	39.68	17	22	30	41	50	56	58

```
lowest :  1  2  3  4  5, highest: 55 56 57 58 59
```

mother.s.ed

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
12361	0	21	8.724	3	4	6	9	11	15	15

```
lowest :  0  1  2  3  4, highest: 16 17 18 19 20
```

father.s.ed

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
12361	0	21	9.232	0	3	7	10	12	15	16

```
lowest :  0  1  2  3  4, highest: 16 17 18 19 20
```

The five-folds cross validation is used to determine the optimal smoothing parameters. The optimality is achieved when $m = 11$, iteration round is 1000 and learning rate is 0.1 for all three quantiles. The relative influence of each variable can be calculated through its contribution to variance reduction. This means if variable X was chosen as splitting variable in tree indices $i_1, .., i_M$ (each tree only has two leaves in our case). Then the summation of variance reduction in those M trees is the influence measurement. The important influence factors except child's age which of course is extremely large are shown in the figure.

We do care about how those factors influence the estimated curve. The partial dependence plots are shown with three curves in each plot. The solid, dashed and dotted lines corresponds to 0.05th, 0.1th and 0.5th quantiles respectively.

We see that the trend of Mother's BMI first climbs up to a peak around 3300

Variable	Explanation
Height(target)	Child's Height in millimeters.
electricity	Household has electricity. (1 yes, 0 no)
radio	Household has radio. (1 yes, 0 no)
tv	Household has TV. (1 yes, 0 no)
fridge	Household has refrigerator. (1 yes, 0 no)
bike	Household has bicycle. (1 yes, 0 no)
motorcycle	Household has motorcycle. (1 yes, 0 no)
car	Household has car. (1 yes, 0 no)
telephone	Household has telephone. (1 yes, 0 no)
religion	Religion of the mother.
wealth	Wealth index (0-5, poorest-richest)
living	Number of living children in a family.
Mage	Mother's age in years.
gender	Gender of child.
age in years	Age of child in years.
breastfeeding	Duration of breastfeeding (in months)
Size of Child at birth	Self explained.
MBMI	Mother's BMI.
Child's age (in months)	Self explained.
Med	Mother's education years
Fed	Father's education years

Table 2.2: Variable Explanation

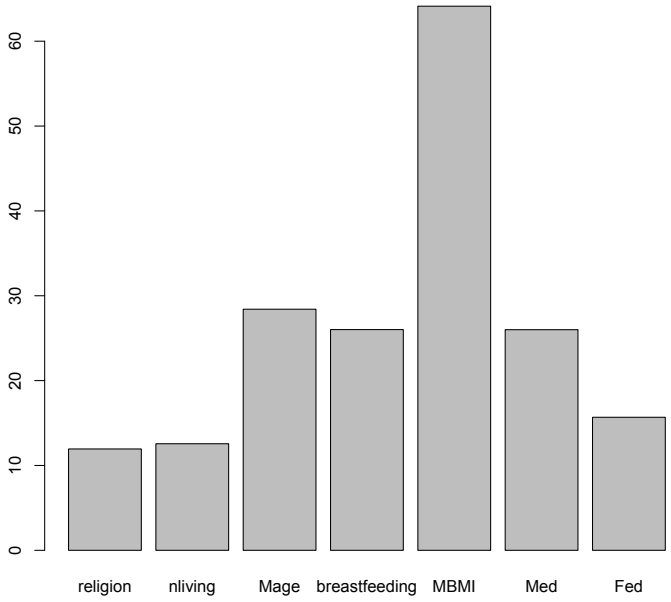


Figure 2.3: Important Factors

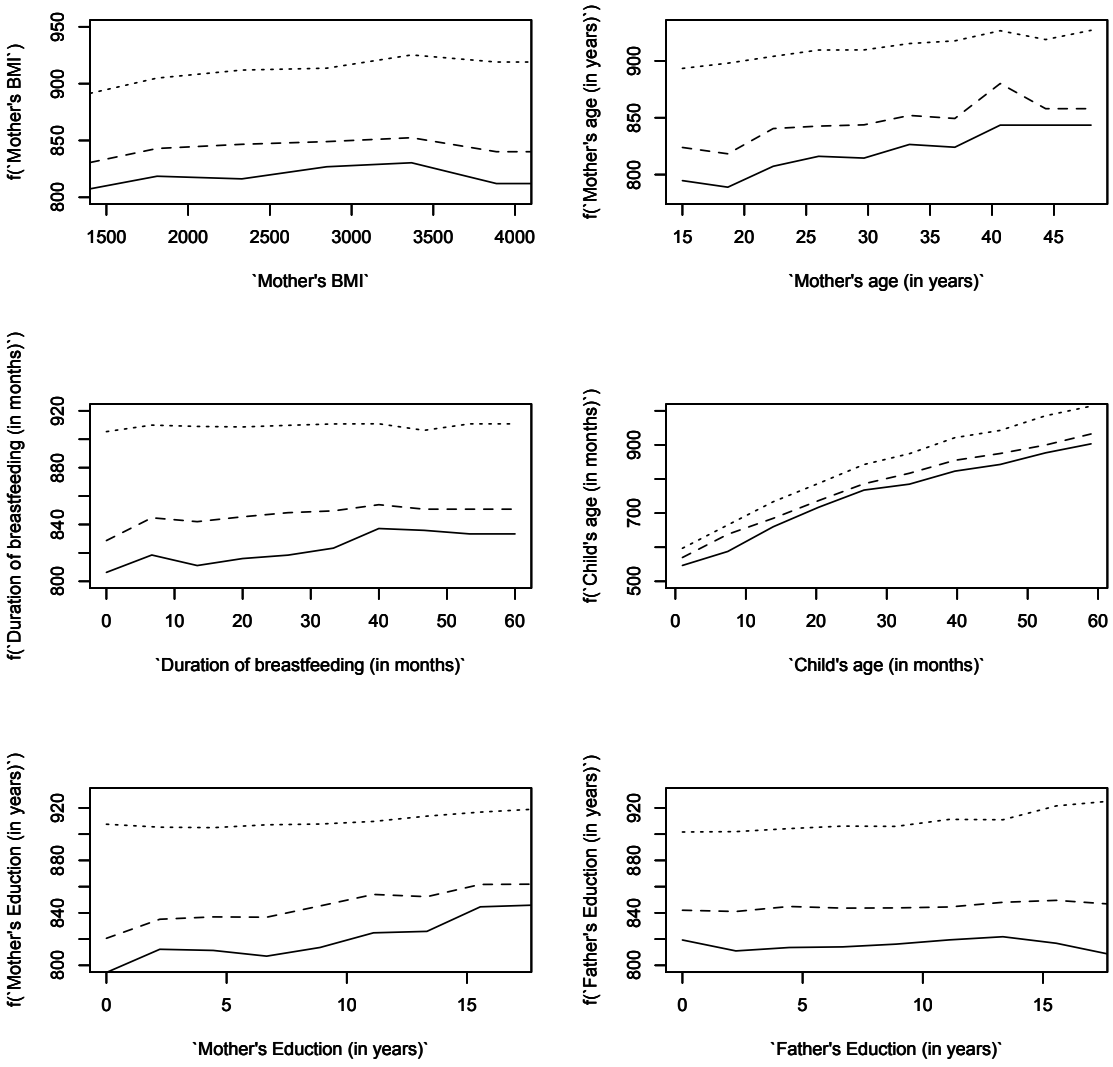


Figure 2.4: Partial Dependence

then goes down for both 0.05th and 0.1th quantiles. The median however do not reflect this scene very strongly. This peak is also seen from Koenker but not in FKH when the quantile is 0.05. The effect of Mother's age has a strictly increasing effect on the estimation curve. This phenomenon matches what described in Koenker but again deviate from FKH in that they found quadratic trend in mother's age. The duration of breastfeeding again has monotone growing trends when quantile is 0.05. For median, this trend does not appear evidently. However both Koenker and FKH has found a decline after 30. Child's age is obviously an increasing predictor for height target. We have found Mother's Education level seems to be positively associated with estimated child's height at lower quantiles while indifferent to the median. This result partially agrees with FKH in lower quantiles. Their plots also shows increasing trend for mother's education at median. Lastly, the education year of father seems not related to our fitting curve. This agrees with both Koenker and FKH.

2.6 Theoretical Results

.

Suppose the model is $y_i = f(x_{1i}, ..., x_{pi}) + \epsilon_i,\ i = 1, ..., n$. The i.i.d. error ϵ_i satisfies $P(\epsilon < 0) = \alpha$ such that $Q_{Y|X}(\alpha) = f(x_1, ..x_p)$. The standard deviation of each ϵ_i is δ and mean is 0 (Without loss of generality). Let $y = (y_1, ..., y_n)^T$. Let $x^{(i)} = (x_{1i}, ..., x_{pi})^T$ be ith observation and $x_j = (x_{j1}, ..., x_{jn})^T$ be the jth variable. The quantile loss function is: $L(y, f) = |y - f| + (\alpha - 1/2)(y - f)$. Suppose the current functional estimate at observed data is: $f_{m-1} = (f_{m-1}(x^{(1)}), ..., f_{m-1}(x^{(n)}))^T$. Then the negative functional gradient of the quantile loss evaluated at f_{m-1} is:

$$-\frac{dL(y,f)}{df}_{f=f_{m-1}} = d_{m-1} = \alpha I(y \geq f_{m-1}) + (\alpha - 1)I(y < f_{m-1}).$$

Where the $\leq$ and $>$ are applied to each coordinate of y and f_{m-1}. If we define S_{m-1} as the operator that maps d_{m-1} to fitted values under the base learner. Then f_m is updated following the relationship below:

$$f_m = f_{m-1} + \nu_{m-1} S_{m-1} d_{m-1},$$

where $\nu_{m-1} > 0$ is the learning rate. In this article, we assume each variable $x_j = (x_{j1}, ..., x_{jn})^T$ has been standardized such that $x_j^T e = 0$ and $||x_j|| = 1$. The norm $||.||$ used in this article is L2 norm. Additionaly, we restrict our attentions on three types of base leaners: 1, OLS; 2, Coordinate-wise OLS; 3, regression trees. Let $r_m = y - f_m$, then $d_m = (\alpha - \frac{1}{2})e + \frac{1}{2}\text{diag}(1/|r_m^{(i)}|)r_m$ where $e = (1, .., 1)^T$. we define $0/0 = 1$ in our case if for some $r_m^{(i)} = 0$. Therefore, the updating equation is:

$$r_m = (I - \tfrac{1}{2} S_{m-1}\Lambda_{m-1}) r_{m-1},$$

where $\Lambda_{m-1} = \text{diag}(\nu_{m-1}/|r_{m-1}^{(i)}|)$. From now on we assume that $v_{m-1} = \min_i |r_{m-1}^{(i)}|$. In particular if $v_{m-1} = 0$, then $v_{m-1}/|r_{m-1}^{(i)}| = 1$ if $r_{m-1}^{(i)} = 0$. Therefore Λ_{m-1} will never be zero such that the residual will keep updating. Further $\Lambda_{m-1} \leq I$. And if we define the condition number of Λ_{m-1} as $cn_{m-1} = \frac{\max_i |r_{m-1}^{(i)}|}{\min_i |r_{m-1}^{(i)}|}$. For the purpose of theory development, we assume that $|1 - \frac{1}{2cn_{m-1}}| < c < 1$ for some $0 < c < 1$. This is equivalent to $1 \leq cn_{m-1} < \frac{1}{2(1-c)}$ for some c. In practice such c always exists because you can choose c close to 1 if conditional number is large. We will see later that larger c corresponds slower convergence rate. Therefore the smaller the conditional number is, the smaller the constant c is hence ensure a faster convergence rate for residual. This results coincide with the results in numerical linear algebra where we also state that small conditional number ensure fast convergence.

We initialize estimate $f_0 = 0$ such that $Sf_0 = f_0$. Therefore $r_0 = y - f_0 = y$. We assume fixed design matrix X. Similar to Buhlmann and Yu (2002), we introduce MSE$(m) = 1/n||f_m - f||^2$. Again $f_m = (f_m(x^{(1)}), ..f_m(x^{(n)}))^T$ and $f = (f(x^{(1)}), ..., f(x^{(n)})^T$. Therefore the randomness is only from y.

Lemma I Suppose the operator (matrix) S satisfies $S^2 = S$ and $S^T = S$. Assume $S^c = I - S$ is the complement of S. Then MSE(m) $= \frac{1}{n}\mathbf{E}||f_m - f||^2 \leq 2\mathbf{E}||Sr_m||^2/n + f^T S^c f/n + 2\text{tr}(S)\delta^2/n$.

Proof:

Since $Sf_0 = f_0$. let us assume $Sf_{m-1} = f_{m-1}$, then $Sf_m = Sf_{m-1} + \nu_{m-1} SSd_{m-1} = f_{m-1} + \nu_{m-1} Sd_{m-1} = f_m$. According to Induction, we have $Sf_m = f_m$ for all m.

$$||f_m - f||^2 = ||Sf_m - f||^2 = ||S(f_m - f) - S^c f||^2 = ||Sf_m - Sf||^2 + f^T S^c f.$$

$$||Sf_m - Sf||^2 = ||Sr_m - S\epsilon||^2 \leq 2||Sr_m||^2 + 2||S\epsilon||^2$$

Notice that $\mathbf{E}\epsilon^T S\epsilon = \mathbf{E}\text{tr}(S\epsilon\epsilon^T) = \text{tr}\ (S\mathbf{E}(\epsilon\epsilon^T)) = \delta^2\text{tr}(S)$. Therefore we conclude: $\frac{1}{n}\mathbf{E}||f_m - f||^2 \leq 2\mathbf{E}||Sr_m||^2/n + f^T S^c f/n + 2\text{tr}(S)\delta^2/n$

Remark I Then MSE is bounded by three parts, the first part $\mathbf{E}||Sr_m||^2/n$ is related how boosting algorithm fit the data. The second part $f^T S^c f/n$ is about how well our chosen operator match the true model. So that means how you choose base learner reflects the image of true model in your head. The last term shows the MSE is related to model complexity or degree of freedom in general.

Let us look at the first term. If we consider the true function linear: $f = X\beta$, once we use hat matrix as S then $f^T S^c f/n = \beta^T X^T (I - X(X^T X)^{-1} X^T) X\beta/n = 0$. If f is nonlinear, we will assume f has polynomial expansion, for instance, cubics splines. Then Assume $f(x^{(k)}) = B^{(k)T} C$, where $B^{(k)}$ is the basis for kth sample and C is the coefficient vector. To things more concrete, $x^{(k)} = (x_1^{(k)}, ...x_p^{(k)})^T$, then $B^{(k)}$ is polynomials with p variables with certain degree, and C is the polynomial coefficients. Denote $B = (B^{(1)T}, ..., B^{(nT)})^T$, then $f = BC$. Once we choose $S = B(B^T B)^{-1} B$ as our operator, this term will disappear. In summary, S can be appropriate constructed to force the first term goes to 0.

For the second term, we know if f is linear, $\text{tr}(S)/n \to 0$ is equivalent to $p/n \to 0$. For nonlinear case, $\text{tr}(S)$ is still determined by p and degree of polynomial. If we make these two parameters fixed, then the second term will go to zero. So the conclusion of this remark is MSE will approach 0 as $n \to \infty$ if proper S is used (the construction of S has also shown in this remark).

Theorem I If the base learner is Ordinary Least Square, and if we have $1 \leq cn_{m-1} < \frac{1}{2(1-c)}$ for some c. Then we have the following conclusions:

1) if $p \geq n$ and $\text{rank}(X) = n$, then $||r_m|| \to 0$ at rate $o(c^m)$ almost surely, $\text{MSE}(m) \leq \delta^2 + c^{2m}/n + 2\delta c^m/n$.

2) if $p < n$ and fix n, p, $\text{MSE} = \lim_m \text{MSE(m)} \leq (\frac{2c^2}{n(1-c^2)} + \frac{1}{n}) f^T S^c f + (\frac{2c^2}{1-c^2}\frac{n-p}{n} + \frac{2p}{n})\delta^2$.

Proof:

For OLS, $S_{m-1} = S = X(X^TX)^-X^T$. (if X^TX is singular then it is general inversion otherwise normal inversion). Since $S^2 = S$, the eigenvalues are either 1 or 0. Let S^c be the orthonormal complement of S.

First we will discuss the MSE for both $p \geq n$ and $p < n$ and then draw conclusion respectively.

1) if $p \geq n$, and if span(X) is n dimensional full space, then S is just identity matrix.Therefore:

$||r_m||^2 = r_{m-1}^T(I - \frac{1}{2}\Lambda_{m-1})^2 r_{m-1} \leq c^2||r_{m-1}||^2$. The last inequality is because we have $|1 - \frac{1}{2cn_{m-1}}| < c$. From here we know $||r_m|| \to 0$ at rate $o(c^m)$.

As for MSE, because $S = I$, tr$(S) = n$ if rank$(X) = n$. According to lemma I, The conclusion holds.

2) If $p < n$,

$Sr_m = (S - \frac{1}{2}S\Lambda_{m-1})r_{m-1}$. Let $D_m = I - \frac{1}{2}\Lambda_m$, then $Sr_m = SD_{m-1}r_{m-1}$. $||Sr_m||^2 = ||SD_{m-1}r_{m-1}||^2 \leq ||S||^2||D_{m-1}||^2||r_{m-1}||^2$. It is clear the maximum eigenvalue of S is 1, and maximum eigenvalue of D_{m-1} is bounded by c. So $||Sr_m||^2 \leq c^2||r_{m-1}||^2 = c^2||Sr_{m-1}||^2 + c^2||S^c r_{m-1}||^2$. Because $r_m = (I - 1/2S\Lambda_{m-1})r_{m-1}$, thus $S^c r_m = S^c r_{m-1} = ...S^c r_0 = (I - S)y$. So the recursive relation we have is:

$||Sr_m||^2 \leq c^2||Sr_{m-1}||^2 + c^2||y - Sy||^2$

Applying this recursion for $m = 1, ..m$, we get $||Sr_m||^2 \leq c^{2m}||Sy||^2 + (1 - c^{2m})c^2/(1-c^2)||y - Sy||^2$. $\mathbf{E}||Sr_m||^2 \leq c^{2m}\mathbf{E}||Sy||^2 + (1-c^{2m})c^2/(1-c^2)E||y - Sy||^2$.

$\mathbf{E}||Sy||^2 == \mathbf{E}\text{tr}(Syy^T) = \text{tr}(S\mathbf{E}yy^T) = f^TSf + p\delta^2$. Similarly, $\mathbf{E}||(I-S)y||^2 = f^TS^cf + (n-p)\delta^2$. So after arranging terms according to Lemma I, MSE(m)$\leq$ $2c^{2m}/n(f^TSf+p\delta^2)+2(1-c^{2m})c^2/n(1-c^2)(f^TS^cf+(n-p)\delta^2)+2p/n\delta^2+f^TS^cf/n$ Let $m\to\infty$, MSE $=\lim_m$ MSE(m) $=(\frac{2c^2}{n(1-c^2)}+\frac{1}{n})f^TS^cf+(\frac{2c^2}{1-c^2}\frac{n-p}{n}+\frac{2p}{n})\delta^2$.

Remark II We are concerned with the MSE bound as $m\to\infty$. Normally this means overfitting. However, we can still show a bound exists for this "over-fitting". Notice that our learning rate v_m is shrinking as m becomes large, therefore this factor serves similar as early stopping. We can also consider large sample behavior. For fixed p, if n is large and as described in remark I we can always choose S such that $1/nf^TS^cf$ goes to zero as $n\to\infty$. So the MSE is bounded by $2c^2/(1-c^2)\delta^2$.

Theorem II If the base learner is coordinate-wise Ordinary Least Square. under the same conditions of theorem I, the following conclusions hold:

$$\text{MSE}(m)\leq\frac{p}{n}\delta^2+f^TS^cf/n+o(n^2p(\min((1-\lambda/p)^m,c^{2m})).$$

Proof:

If the procedure is coordinate-wise OLS, each iteration round the predictor with maximum correlation with the residual will be chosen. Suppose $\{i_1,...,i_m\}$ are chosen for algorithm in order. The projection operator at mth round is $S_m = x_{i_m}x_{i_m}^T$. Suppose the set $\{x_{i_1},...x_{i_m}\}=\{x_1,...,x_q\}(q\leq p)$. Let $X=(x_1,...,x_q)$, then $S = X(X^TX)^{-1}X$. Therefore S is projection matrix onto column space of X. Therefore $SS_m=S_m$, $S_mS=S_m$.

$||Sr_m||^2 = r_{m-1}(S-\frac{1}{2}SS_{m-1}\Lambda_{m-1})^T(S-\frac{1}{2}SS_{m-1}\Lambda_{m-1})r_{m-1}$

$$||Sr_m||^2 = ||Sr_{m-1}||^2 + r_{m-1}^T(\tfrac{1}{4}\Lambda_{m-1}S_{m-1}\Lambda_{m-1} - \tfrac{1}{2}\Lambda_{m-1}S_{m-1} - \tfrac{1}{2}S_{m-1}\Lambda_{m-1})r_{m-1}.$$

Let $D_m = \frac{1}{2}\Lambda_m - I$. It is true that $\frac{1}{4}\Lambda_{m-1}S_{m-1}\Lambda_{m-1} - \frac{1}{2}\Lambda_{m-1}S_{m-1} - \frac{1}{2}S_{m-1}\Lambda_{m-1} = D_{m-1}S_{m-1}D_{m-1} - S_{m-1}$. Thus,

$$||Sr_m||^2 = ||Sr_{m-1}||^2 - ||S_{m-1}r_{m-1}||^2 + r_{m-1}^T D_{m-1}S_{m-1}D_{m-1}r_{m-1}$$

Now we will prove $||S_{m-1}r_{m-1}||^2 \geq \lambda_{\min}(X^TX)/q||Sr_{m-1}||^2$.

Denote $v = (v_1, .., v_q)^T = (x_1^T r_{m-1}, ..., x_q^T r_{m-1})^T$.

$||Sr_{m-1}||^2 = ||X(X^TX)^{-1}X^T r_{m-1}||^2 = ||X(X^TX)^{-1}v||^2 = V^T(X^TX)^{-1}V \leq \lambda_{\max}((XTX)^{-1})(v_1^2 + ...v_q^2) = 1/\lambda_{\min}(X^TX)(v_1^2 + ...v_q^2)$.

$||S_{m-1}r_{m-1}||^2 = ||x_{i_{m-1}}x_{i_{m-1}}^T r_{m-1}||^2 = ||x_{i_{m-1}}v_{i_{m-1}}||^2 = v_{i_{m-1}}^2$. Since coordinate-wise OLS choose maximum correlation with current residual. So $v_1^2 + ...v_q^2 \leq qv_{i_{m-1}}^2 = q||S_{m-1}r_{m-1}||^2$. Thus $||Sr_{m-1}||^2 \leq q/\lambda_{\min}(X^TX)||S_{m-1}r_{m-1}||^2$.From now on, we replace $\lambda_{\min}(X^TX)$ with λ.

Then, $||Sr_m||^2 \leq (1 - \lambda/q)||Sr_{m-1}||^2 + r_{m-1}^T D_{m-1}S_{m-1}D_{m-1}r_{m-1}$.

Following the same proof as in theorem I, $r_{m-1}^T D_{m-1}S_{m-1}D_{m-1}r_{m-1} \leq np||y||^2c^{2m}$. Therefore,

$$||Sr_m||^2 \leq (1 - \lambda/p)||Sr_{m-1}||^2 + np||y||^2c^{2m};$$

Divide both sides by factor $(1 - \lambda/p)^m$ and use the telescope sum, it is easy to see:

$||Sr_m||^2/(1-\lambda/p)^m - ||Sr_0||^2 \leq np||y||^2 \sum_{k=1}^{m}(c^2/(1-\lambda/p))^k$

$||Sr_o||^2 = ||Sy||^2 \leq \lambda_{\max}(SS)||y||^2 = ||y||^2$, Thus

$||Sr_m||^2 \leq (1-\lambda/p)^m||y||^2 + \frac{(||y||^2/n)n^2pc^2((1-\lambda/p)^m - c^{2m})}{(1-\lambda/p-c^2)}$

First, $||y||^2/nc^2/(1-\lambda/p-c^2)$ is constant asymptotically. In the second term, because the order of $(1-\lambda/p)^m - c^2m$ is $o(\min((1-\lambda/p)^m, c^{2m}))$, hence the first term is faster so we need to ignore. Therefore the conclusion proved.

As for the residual projected onto completement of S, $||S^c r_m||^2 = ||S^c r_{m-1}||^2$.

Since $SS_m = S_m$, so if one starts with initialization that satisfies $Sf_0 = f_0$, by induction we have $Sf_m = f_m$ holds.

Therefore by following the same reasoning of theorem I's proof, we obtain the conclusion.

Chapter 3

Asymptotics for M-estimator with additional control parameter

3.1 Introduction

.

M-estimators represent a broad class of estimators by minimizing the sum of functions of data point. That is, $\hat{\theta} = \arg\min_\theta \sum_{i=1}^n \rho(X_i, \theta)$, It was first introduced by Peter Huber [32] to study robust estimators and their relevant asymptotic properties and has been studied extensively for its theoretical properties. See, for example, Huber(1964, 1972) [32] [33], Portnoy(1977) [48], Collins(1976) [17], Freedman and Diaconis(1982) [21], Niemiro(1992) [46], He and Shao(1996) [31]. One can choose different objective functions ρ to obtain estimators with desirable properties. For instance, one can choose $\log f(.)$ to get Maximum Likelihood Estimators and $(x-\theta)^2/2I(|x-\theta| \leq C) + C^2/2I(|x-\theta| > C)$ to get Huber estimators. For the reason that nonconvex ρ function can possibly produce inconsistent M-estimators, see Freedman and Diaconis (1982) [21]. Therefore people would often put convex constraint on ρ function. In particular, Haberman(1989) [28], Niemiro(1992) [46] provided strong consistency and asymptotic normality of M-estimators when ρ is convex. Moreover, Niemiro generalized a Bahadur(1966) [2] type representation of

sample quantiles to M-estimators. It sheds light on the precise description of asymptotics of regression quantiles and spatial quantiles based on some "regular" conditions. The word "regular" in this paper basically refers to the asymptotic Hessian matrix is non-singular at true parameter point. A natural question arose is that how would the asymptotic behaviour change if one considers the "non-regular" cases such as extreme quantiles or singular Hessian matrix.

The non-regular problem has been studied quite a bit for regression quantiles. Regression quantiles serve to estimate conditional quantile of a response given predictors, see Koenker and Bassett(1978) [40], Chaudhuri(1991) [13], Portnoy, S. and Jureckova, J.(1999) [49]. Generally, suppose $Y \in \mathbb{R}$ is the response, and $X \in R^{p+1}$ with the first column all 1s. The population conditional quantile is $F^{-1}_{Y|X}(\tau|X=x) = \inf\{y : F_{Y|X}(y|X=x) > \tau\}$, where $\tau \in [0,1], F_{Y|X}$ is conditional c.d.f. . Then we model conditional quantile as a linear combination of predictors : $F^{-1}_{Y|X}(\tau|X=x) = x^T\beta(\tau)$. Assuming that $(Y_i, X_i); i = 1, ...n$ is one set of data. Then the inference about $\beta\tau$ is made through regression quantile statistics $\hat{\beta}(\tau)$ defined by the least asymmetric absolute deviation problem:

$$\hat{\beta}(\tau) = \arg\min_{\beta} \sum_{i=1}^{n} \rho_\tau(Y_i - X_i^T\beta) \quad \text{where} \quad \rho_\tau(u) = (\tau - I(u \leq 0)u \tag{3.1}$$

The non-regular problem for regression quantile corresponds to the case where $\tau = 1$. Smith (1994) [52] did careful studies for a class of nonregular models which are essentially equivalent to extremal regression quantiles. Later, Chernozhukov(2005) [16] developed a theory of quantile regression in tails. The author defined extremal (extreme order if $n\tau_n \to 0$ and intermediate order $n\tau_n \to \infty$) quantile regression

for the class of linear quantile regression models. Asymptotically, extreme order regression quantiles converge weakly to minimizers of functionals of stochastic integrals of Poisson processes that is determined by predictors, while intermediate regression quantiles and their functionals converge to normal vectors with variance matrices depend on the tail parameters and predictors. However all the above is constrained to quantile regression. What if we consider similar problems in M-estimators ? Then the extreme order and intermediate order M-estimators are specified according to the convergent speed of our control parameter α_n. If this control parameter converges fast to α^* then it matches extreme order. However if it only has intermediate speed convergence, then it is called intermediate order. Under both cases, strong convergence is guaranteed provided some mild conditions are satisfied. As for weak convergence result, This paper includes the results for intermediate order so far. We figured out the fastest converging rate for control parameters beyond which no asymptotic normality can be ensured for every ϕ within our framework. That means, for some least favorable objective function ϕ there is no asymptotic normality beyond that point.

This paper is organized as follows. In section 2, we will introduce some key notations and conditions for the later use. Particularly we will define what we mean by "regular" and "non-regular" in our problem setup. In section 3, we will develop both strong convergence and weak convergence. Specifically, we will show that the in probability convergence rate is exponential fast for any ϵ. For regular situation, the estimator is asymptotic normal if $\alpha_n - \alpha^* = O(n^{-1/2})$. However for non-regular situation, we do not have root n consistency. The scalar is determined by the Hessian matrix of expected objective function and the convergence rate of α_n. In section 4, we provides two application examples to check how our theory can be used. The first example is Chaudhuri's spatial quantile. Our results give implications for extreme spatial quantile with intermediate order. Particularly, we listed simulation examples

for single variable case to illustrate the effect of sharpness of control parameter converging rate. Another example is a novel algorithm called finite smoothing quantile which is proposed by Chen(2007) [15] to estimate quantile regression coefficients. We will show that according the theory we developed, the asymptotic efficiency can be ensured for that algorithm.

3.2 Definitions and assumptions.

The notation introduced in this section will be used throughout the paper. Let $X \sim F(.)$ be a random variable with its value from $\mathbb{R}^p$. Let $\phi : \mathbb{R}^p \times \mathbb{R}^q \times \mathbb{R}^d \to \mathbb{R}^1$ be a real function. We define closed unit ball $\mathbb{B}_p = \{x : ||x|| \leq 1, x \in \mathbb{R}^p\}$. We introduce α_n and α^* as controlling parameter for the finite sample and infinite sample situations respectively. Moreover, $\alpha_n, \alpha^* \in \mathbb{B}_d$ and $||\alpha_n - \alpha^*|| \to 0$. The norm $||.||$ used in this paper is Euclidean for vector and Spectral for a positive semidefinite matrix which is largest eigenvalue if there is no special notification . Let $x_1, ..., x_n \in \mathbb{R}^p$ are i.i.d. samples from the c.d.f. $F(x)$ of random variable X. Let $q \in \mathbb{R}^q, \alpha \in \mathbb{R}^d$ be two real variables. Our basic problem setting is as following:

$$\hat{\phi}_n(q, \alpha) = \sum_{i=1}^{n} \phi(x_i, q, \alpha)/n \tag{3.2}$$

$$\tilde{\phi}(q, \alpha) = \mathbf{E}_X \phi(X, q, \alpha) \tag{3.3}$$

$$\tilde{q} = \arg\min_{q} \tilde{\phi}(q, \alpha^*) \tag{3.4}$$

$$\tilde{q}_n = \arg\min_{q} \tilde{\phi}(q, \alpha_n) \tag{3.5}$$

$$\hat{\phi}_n(\hat{q}_n, \alpha_n) = \min_{q} \hat{\phi}_n(q, \alpha_n) \tag{3.6}$$

We assume throughout this paper three assumptions:

i) $\tilde{q}$ and $\tilde{q}_n$ are unique, however uniqueness not necessarily holds for $\hat{q}_n$.

ii) $\mathbf{E}|\phi(X, q, \alpha^*)| < \infty$ for fixed q.

iii) $\phi(x, q, \alpha)$ is convex with respect to q for fixed x and α.

By making these three assumptions, $\tilde{\phi}(q, \alpha)$ is convex in q and finite value on space $\mathbb{R}^q \times \mathbb{R}^d$. In order to avoid ambiguity we will set the value to be ∞ if no minimum can be achieved for certain functions. We denote $g(x, q, \alpha)$ as a subgradient function for $\phi(x, q, \alpha)$, in other words, $(q_2 - q_1)^T g(x, q_1, \alpha) \leq \phi(x, q_2, \alpha) - \phi(x, q_1, \alpha)$ for all $q_1, q_2 \in \mathbb{R}^q$, $x \in \mathbb{R}^p$, $\alpha \in R^d$. According to Niemiro(1992) , one can always select a subgradient function $g(x, q, \alpha)$ which is measurable in x. The operator $\mathbf{D}\tilde{\phi}(q, \alpha)$ is gradient with respect to q, $\mathbf{H}\tilde{\phi}(q, \alpha)$ is Hessian with respect to q. We use $\mathbf{D}$, $\mathbf{H}$, $\mathbf{D}_n$ and $\mathbf{H}_n$ as simplified the notations for $\mathbf{D}\tilde{\phi}(\tilde{q}, \alpha^*)$, $\mathbf{H}\tilde{\phi}(\tilde{q}, \alpha^*)$, $\mathbf{D}\tilde{\phi}(\tilde{q}_n, \alpha_n)$ and

$\mathbf{H}\tilde{\phi}(\tilde{q}_n, \alpha_n)$ respectively. The notation $S_n = \sum g(x_i, \tilde{q}_n, \alpha_n)$ will be constantly used unless stated otherwise. We also need to state other assumptions which might be used in each specific theorem.

iv) There exists a function $M : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^+$, $||\phi(X, q, \alpha_1) - \phi(X, q, \alpha_2)|| \leq M(X, q)||\alpha_1 - \alpha_2||$. Moreover, $\mathbf{E}M(X, q) < \infty$ for fixed q.

v) The Hessian matrix $\mathbf{H}\tilde{\phi}(q, \alpha)$ exists and continuous in a neighbourhood of $(\tilde{q}, \alpha^*)$ and $\mathbf{H}\tilde{\phi}(\tilde{q}, \alpha^*)$ is positive definite.

vi) The subgradient $||g(X, q, \alpha)|| < M$ for all X and (q, α)in a neighbourhood of $(\tilde{q}, \alpha^*), \mathbf{E}g(X, \tilde{q}, \alpha^*)g(X, \tilde{q}, \alpha^*)^T$ is postive definite.

vii) $g(X, q, \alpha)$ is piecewise differentiable for all three arguments, the non-differentiable points in the first and second arguments are of measure zero and irrelevant to the value of the third argument. There exists a function $M : \mathbb{R}^p \times \mathbb{R}^d \to \mathbb{R}^+$ such that $||g(X, q_1, \alpha) - g(X, q_2, \alpha)|| \leq M(X, \alpha)||q_1 - q_2||$, $\mathbf{E}M(X, \alpha) < \infty$.

viii) (Irregular Condition: (a), (b) and (c) all hold)

(a) The Hessian of $\mathbf{H}\tilde{\phi}(q, \alpha)$ exists and continous in a neighborhood of $(\tilde{q}, \alpha^*)$ and $\mathbf{H}\tilde{\phi}(\tilde{q}, \alpha^*)$ is singular. $\mathbf{H}_n$ is positive definite for all n and there exists a $\kappa_n > 0$ such that $\kappa_n \mathbf{H}_n^{-1} \to \mathbf{H}$, where $\mathbf{H}$ is a positive definite matrix.

(b) There exists a sequence r_n such that $||r_n g(X, \tilde{q}_n, \alpha_n) - g|| \to 0$ where g satisfy condition vi).

(c) $n^{1/4} r_n \kappa_n \to \infty$.

ix) There exists a sequence $c_n \to 0, c_n r_n^2 \to 0$ and a function $L(q)$ such that $n^{1/2} c_n^{-1}(\tilde{\phi}(\tilde{q}_n + c_n q, \alpha_n) - \tilde{\phi}(\tilde{q}_n, \alpha_n)) \to L(q)$ which is finite on an open set and for almost every x, $q^T x + L(q)$ has a unique minimizer.

3.3 Main results and proofs

Theorem I If iv) holds , $\hat{\phi}_n(q, \alpha_n) \to \tilde{\phi}(q, \alpha^*)$ almost surely for fixed q.

The pointwise convergence for convex functions follows easily under assumption iv). We will in in section 4 iv) is a pretty mild condition to hold.

Proof :
$|\hat{\phi}_n(q, \alpha_n) - \tilde{\phi}(q, \alpha^*)| = |\sum_{i=1}^n (\phi(X_i, q, \alpha_n) - \phi(X_i, q, \alpha^*))/n + \sum_{i=1}^n \phi(X_i, q, \alpha^*)/n - \tilde{\phi}(q, \alpha^*)|$
$\leq 1/n \sum_{i=1}^n |\phi(X_i, q, \alpha_n) - \phi(X_i, q, \alpha^*)| + |\sum_{i=1}^n \phi(X_i, q, \alpha^*)/n - \tilde{\phi}(q, \alpha^*)|$ According to assumption ii), the last term goes to zero almost surely. $1/n \sum_{i=1}^n |\phi(X_i, q, \alpha_n) - \phi(X_i, q, \alpha^*)| \leq 1/n \sum_{i=1}^n |M(X_i, q)|||\alpha_n - \alpha^*||$. According to assumption iv), $1/n \sum_{i=1}^n M(X_i, q)$ $\to \mathbf{E}M(X, q) < \infty$. Hence , $|\hat{\phi}_n(q, \alpha_n) - \tilde{\phi}(q, \alpha^*)| \to 0$.

Theorem II If iv) holds,

$$\sup_{||q|| \leq M} |\hat{\phi}_n(q, \alpha_n) - \tilde{\phi}(q, \alpha^*)| \to 0$$

almost surely for any $M > 0$.

The proof of the theorem is a direct result from the following lemma cited from lemma 3 Niemiro [46].

Lemma I Let $f_n(q,\omega), n = 1, 2...$, be random functions on R^d, convex in q for each ω. Let $f(q,\omega)$ be a random function such that for each fixed q, $f_n(q) \to f(q)$ (a) almost surely; (b) in probability. Then for each $M > 0$, $\sup_{||q||\leq M} |f_n(q) - f(q)| \to 0$ (a) almost surely; (b) in probability. [See Niemiro(1992) [46], lemma 3; Rockafeller(1970) [51], Theorem 10.8].

This uniform convergence plays a central role in the proofs of rest results. Particularly, uniform convergence on a compact set ensure that the two minimizers are very close as n goes large. That is the key idea to prove weak convergence for the M-estimator. The proof is almost a direct result from lemma I.

Proof:

Except for a measure zero set, for every fixed ω, we have point-wise convergence according to theorem I, therefore according to lemma I, we have uniform convergence almost surely for that ω.

Theorem III If iv) holds, and if $\tilde{q} = \infty$, then $\hat{q}_n \to \infty$ almost surely. If $\tilde{q}$ is bounded, then $\hat{q}_n \to \tilde{q}$ almost surely.

Proof:

The following statements in this proof are based on fixed ω

1)When $\tilde{q}$=∞, but if $\hat{q}_n$ does not go to $+\infty$, then there exists a subsequence $||\hat{q}_{n_k}|| < M$. Thus there exists a further subsequence of $\{n_k\}$(for the sake of convenience I will still use $\{n_k\}$ to denote that further subsequence. Thus $\hat{q}_{n_k} \to q_0$ which is bounded. $\tilde{\phi}(\hat{q}_{n_k}, \alpha^*) = \hat{\phi}_{n_k}(\hat{q}_{n_k}, \alpha_{n_k}) + [\tilde{\phi}(\hat{q}_{n_k}, \alpha^*) - \hat{\phi}_{n_k}(\hat{q}_{n_k}, \alpha_{n_k})] \leq \hat{\phi}_{n_k}(q, \alpha_{n_k}) + [\tilde{\phi}(\hat{q}_{n_k}, \alpha^*) - \hat{\phi}_{n_k}(\hat{q}_{n_k}, \alpha_{n_k})]$ for any $q \in \mathbb{R}^p$. Let $k \to +\infty$, by Theorem II, $\tilde{\phi}(q_0, \alpha^*) \leq \tilde{\phi}(q, \alpha^*)$ for any $q \in \mathbb{R}^p$. Therefore q_0 is the minimizer which contradicts with the fact that the minimizer is ∞.

2) If $\tilde{q}$ is bounded, to prove $\hat{q}_n \to \tilde{q}$, it is equivalent to prove $\tilde{\phi}(\hat{q}_n, \alpha^*) \to \tilde{\phi}(\tilde{q}, \alpha^*)$. This is simple consequence of the fact that $\tilde{\phi}$ is convex and the minimizer is unique. Since $\hat{\phi}(\hat{q}_n, \alpha_n) \leq \hat{\phi}(\tilde{q}, \alpha_n)$, $\limsup \hat{\phi}(\hat{q}_n, \alpha_n) \leq \lim \hat{\phi}(\tilde{q}, \alpha_n) = \tilde{\phi}(\tilde{q}, \alpha^*)$.

On the other hand, because $\tilde{\phi}(\tilde{q}, \alpha^*) \leq \tilde{\phi}(\hat{q}_n, \alpha^*)$ for any n. Further more, for any given $\epsilon_0 > 0$, there exists a N_0, when $n \geq N_0$

$$\sup_q |\hat{\phi}(q, \alpha_n) - \tilde{\phi}(q, \alpha^*)| \leq \epsilon_0$$

,

therefore, $\tilde{\phi}(\tilde{q}, \alpha^*) \leq \tilde{\phi}(\hat{q}_n, \alpha^*) \leq \hat{\phi}(\hat{q}_n, \alpha_n) + \epsilon_0$. Because ϵ_0 is any given,

$\tilde{\phi}(\tilde{q}, \alpha^*) \leq \liminf \hat{\phi}(\hat{q}_n, \alpha_n)$, therefore $\lim \hat{\phi}(\hat{q}_n, \alpha_n) = \tilde{\phi}(\tilde{q}, \alpha^*)$

According to theorem II, $|\hat{\phi}(\hat{q}_n, \alpha_n) - \tilde{\phi}(\hat{q}_n, \alpha^*)| \to 0$.

$||\hat{\phi}(\hat{q}_n, \alpha_n) - \tilde{\phi}(\tilde{q}, \alpha^*)| - |\tilde{\phi}(\tilde{q}, \alpha^*) - \tilde{\phi}(\hat{q}_n, \alpha^*)|| \leq |\hat{\phi}(\hat{q}_n, \alpha_n) - \tilde{\phi}(\hat{q}_n, \alpha^*)| \to 0$

And the first term on left goes to 0 thus we have the conclusion proved.

We introduce a few lemmas below to provide a convenient way to develop the convergent rate.

lemma II Assume $X_1, X_2, ...$ are i.i.d. random variables such that $\mathbf{E}e^{t|X_n|} < \infty$ for some $t > 0$. Let $S_n = X_1 + ... + X_n$, $\mu = \mathbf{E}X_n$. Then for each $\epsilon > 0$ there exists $a > 0$ such that

$\mathrm{P}(|\frac{S_n}{n} - \mu| > \epsilon) = O(e^{-an})$, $n \to \infty$.

For the proof, see Durrett(1991), [18].

lemma III $f_n(x)$ and $f(x)$ are convex functions with their minimizer x_n^* and x^* respectively, and the minimizer for $f(x)$ is unique. Suppose there is a closed ball B such that $x^* \in B$. If $\sup_{x\in B} |f_n(x) - f(x)| \to 0$, then $x_n^* \to x^*$. That is, for any given $\epsilon > 0$, if $\sup_{x\in B} |f_n(x) - f(x)| < \epsilon$, then there must exist a $\epsilon' > 0$, $|x_n^* - x^*| < \epsilon'$.

Proof:

For any given $\epsilon > 0$, there is a N when $n > N$, for any $x \in B$, $|f_n(x) - f(x)| < \epsilon$. $f(x_n^*) \geq f(x^*) > f_n(x^*) - \epsilon \geq f_n(x_n^*) - \epsilon > f(x_n^*) - 2\epsilon$. $f(x^*) > f(x_n^*) - 2\epsilon$. Since the minimizer is unique, $x_n^* \to x^*$.And from the last inequality, it is easy to see x_n^* must be in a correpsonding neiborghood of x^*. Moreover, as we choose $\epsilon \to 0$, we can obtain corresponding $\epsilon' \to 0$ as well.

lemma IV If there is a closed ball B such that $\sup_{x\in B} |f_n(x) - f(x)| \to 0$, then for any given $\epsilon > 0$, there exists a $\delta > 0$, whenever $x_1, x_2 \in B$ satisfy $|x_1 - x_2| < \delta$,

then $|f_n(x_1) - f_n(x_2)| < \epsilon$ for all n.

Proof:

$|f_n(x_1) - f_n(x_2)| = |f_n(x_1) - f(x_1) + f(x_1) - f(x_2) + f(x_2) - f_n(x_2)| \leq |f_n(x_1) - f(x_1)| + |f_(x_1) - f(x_2)| + |f(x_2) - f_n(x_2)|$. For any given $\epsilon > 0$, there exists a N, when $n > N$ $\sup_x |f_n(x) - f(x)| < \frac{\epsilon}{3}$. We can choose a $\delta > 0$ such that $|f_n(x_1) - f_n(x_2)| < \epsilon$ for $n \leq N$ and $|f(x_1) - f(x_2)| < \frac{\epsilon}{3}$. Then it is easy to see such δ leads to our conclusion.

Theorem IV Suppose conditions iv) holds, then for every $\epsilon > 0$, there exists $a > 0$ such that $\mathrm{P}(||\hat{q}_n - \tilde{q}|| > \epsilon) = O(e^{-an})$, $n \to \infty$.

Proof:

Let $Y_{ni} = \phi(x_i, \tilde{q}, \alpha^*) - \tilde{\phi}(q, \alpha^*), S_n = \sum_{i=1}^n Y_{ni}$. For any given ϵ, there exists N, if $n > N$, $|\hat{\phi}_n(\tilde{q}, \alpha^*) - \hat{\phi}_n(\tilde{q}, \alpha_n))| < \epsilon$.

$\mathrm{P}(|\hat{\phi}_n(\tilde{q}, \alpha_n) - \tilde{\phi}(\tilde{q}, \alpha^*)| > \epsilon) = \mathrm{P}(\frac{S_n}{n} > \epsilon + \hat{\phi}_n(\tilde{q}, \alpha^*) - \hat{\phi}_n(\tilde{q}, \alpha_n)) + \mathrm{P}(\frac{S_n}{n} < -\epsilon + \hat{\phi}_n(\tilde{q}, \alpha^*) - \hat{\phi}_n(\tilde{q}, \alpha_n)) \leq \mathrm{P}(\frac{S_n}{n} > 2\epsilon) + \mathrm{P}(\frac{S_n}{n} < -2\epsilon) = \mathrm{P}(|S_n| > 2\epsilon) = O(e^{-an})$

This is due to the facts that $\hat{\phi}_n(q, \alpha_n) - \hat{\phi}_n(q, \alpha^*) \to 0$ almost surely, and S_n satisfies conditions of lemma II. There exists a small enough ball $B(\epsilon)$ with $\tilde{q} \in B(\epsilon)$. If $|\hat{\phi}_n(\tilde{q}, \alpha_n) - \tilde{\phi}(\tilde{q}, \alpha^*)| < \epsilon$, and $q^* \in B(\epsilon)$ is the maximizer within this ball.

$$\begin{aligned}
&\sup_{q\in B(\epsilon)} |\hat{\phi}_n(q,\alpha_n) - \tilde{\phi}(q,\alpha^*)| = |\hat{\phi}_n(q^*,\alpha_n) - \tilde{\phi}(q^*,\alpha^*)| \\
&= |\hat{\phi}_n(q^*,\alpha_n) - \hat{\phi}_n(\tilde{q},\alpha_n) + \hat{\phi}_n(\tilde{q},\alpha_n) - \tilde{\phi}(\tilde{q},\alpha^*) + \tilde{\phi}(\tilde{q},\alpha^*) - \tilde{\phi}(q^*,\alpha^*)| \\
&\quad \leq |\hat{\phi}_n(q^*,\alpha_n) - \hat{\phi}_n(\tilde{q},\alpha_n)| + |\hat{\phi}_n(\tilde{q},\alpha_n) - \tilde{\phi}(\tilde{q},\alpha^*)| + |\tilde{\phi}(\tilde{q},\alpha^*) - \tilde{\phi}(q^*,\alpha^*)|
\end{aligned} \tag{3.7}$$

According to lemma IV, $\hat{\phi}_n(q,\alpha_n)$ is equi-continuous, so we can choose a radius δ for $B(\epsilon)$ such that $|\hat{\phi}_n(q^*,\alpha_n) - \hat{\phi}_n(\tilde{q},\alpha_n)| < \epsilon$, $|\tilde{\phi}(\tilde{q},\alpha^*) - \tilde{\phi}(q^*,\alpha^*)| < \epsilon$. Therefore ,

$$\mathrm{P}(\sup_{q\in B(\epsilon)} |\hat{\phi}_n(q,\alpha_n) - \tilde{\phi}(q,\alpha^*)| < 3\epsilon) = 1 - e^{-an}. \tag{3.8}$$

The lemma III guarantees that,

$$\mathrm{P}(||\hat{q}_n - \tilde{q}|| < \epsilon') = 1 - e^{-an} \tag{3.9}$$

Since for every ϵ' we can back track the corresponding ϵ. In other words, for every ϵ' we have a according ϵ such that equation (3.2) leads to equation (3.3). Thus the statement holds.

We will need implicit function theorem to get the bias-correction term for $\tilde{q}_n$. Also multivariate Lindeberg-Feller is required to force asymptotic normality.

Lemma V (Implicit function theorem) Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be a continuously differentiable function, and let $\mathbb{R}^n \times \mathbb{R}^m$ have coordinates (x, y). Fix a point $(a, b) \in \mathbb{R}^n \times \mathbb{R}^m$ with $f(a, b) = c \in \mathbb{R}^m$.If the matrix $[\frac{\partial f_i}{\partial x_j}(a, b)]$ is invertible, then there exists an open set U containing a, an open sent V containing b, and a unique

continuously differentiable function $g : V \to U$ such that:

$$\{(g(y), y)|y \in V\} = \{(x, y) \in U \times V | f(x, y) = c\}$$

Lemma VI (Multivariate Lindeberg-Feller) Let $X_{n,i}, i = 1, 2, ..., n$ be independent real-valued random vectors with $\mathbf{E}X_{n,i} = 0$, and let $S_n = \sum_{i=1}^{n} X_{i,n}$. Suppose that:

1. $\lim_{n\to\infty} \mathbf{E}S_n S_n^T = \lim_{n\to\infty} \mathbf{E}X_{i,n}X_{i,n}^T = \Sigma$.

2. For all $\epsilon > 0 \lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{E}(||X_{i,n}||^2; ||X_{i,n}|| > \epsilon) = 0$.

Then S_n converges weakly to a normal random vector with mean zero and covariance Σ.

Theorem V If condition iv), v), vi), vii) hold, then $\sqrt{n}(\hat{q}_n - \tilde{q}_n) = -\frac{1}{\sqrt{n}}\mathbf{H}^{-1}S_n + o_p(1)$, and $-\frac{1}{\sqrt{n}}\mathbf{H}^{-1}S_n \to N(0, \mathbf{H}^{-1}\mathbf{Var}g\mathbf{H}^{-1})$ where $g = g(X, \tilde{q}, \alpha^*)$. Additionally, if $||\alpha_n - \alpha^*|| = o(n^{-1/2})$, then $\sqrt{n}(\hat{q}_n - \tilde{q}) \to N(0, \mathbf{H}^{-1}\mathbf{Var}g\mathbf{H}^{-1})$

Proof:

$$X_{ni} = \phi(x_i, \tilde{q}_n + \tfrac{q}{\sqrt{n}}, \alpha_n) - \phi(x_i, \tilde{q}_n, \alpha_n) - \tfrac{q^T}{\sqrt{n}} g(x_i, \tilde{q}_n, \alpha_n)$$

$$\sum_{i=1}^{n} X_{ni} = n\hat{\phi}(\tilde{q}_n + \tfrac{q}{\sqrt{n}}, \alpha_n) - n\hat{\phi}(\tilde{q}_n, \alpha_n) - \tfrac{q^T}{\sqrt{n}} S_n$$

$$\mathbf{E}g(x_i, \tilde{q}_n, \alpha_n) = \mathrm{D}\tilde{\phi}(\tilde{q}_n, \alpha_n) = 0; \ \mathbf{E}X_{ni} = \tilde{\phi}(\tilde{q}_n + \tfrac{q}{\sqrt{n}}, \alpha_n) - \tilde{\phi}(\tilde{q}_n, \alpha_n)$$

The change of integral is valid because let $\epsilon < 0$, $e \in R^q$

$$\frac{-1}{\epsilon}(\phi(X, q - \epsilon e, \alpha) - \phi(X, q, \alpha)) \leq e^T g(X, q, \alpha) \leq \frac{1}{\epsilon}(\phi(X, q + \epsilon e, \alpha) - \phi(X, q, \alpha))$$

Taking expectations and letting $\epsilon \to 0$ we get change of integral. Because ϕ is convex and g is its subgradient, thus by definition we know that $X_{ni} \geq 0$.

On the other hand, $X_{ni} \leq \frac{q^T}{\sqrt{n}}(g(x_i, \tilde{q}_n + \frac{q}{\sqrt{n}}, \alpha_n) - g(x_i, \tilde{q}_n, \alpha_n))$. Therefore,

$\sum_{i=1}^n \mathbf{E}X_{ni}^2 \leq \mathbf{E}[q^T(g(X, \tilde{q}_n + \frac{q}{\sqrt{n}}, \alpha_n) - g(X, \tilde{q}_n, \alpha_n))]^2$. Assumption vii) enables $q^T(g(X, \tilde{q}_n + \frac{q}{\sqrt{n}}, \alpha_n) - g(X, \tilde{q}_n, \alpha_n))]^2$ converges to 0 almost surely. By Dominated Convergent Theorem we know $\mathbf{E}[q^T(g(X, \tilde{q}_n + \frac{q}{\sqrt{n}}, \alpha_n) - g(X, \tilde{q}_n, \alpha_n))]^2 \to 0$, therefore $\mathbf{Var}(\sum_{i=1}^n X_{ni}) \to 0$. By applying Chebyshev inequality, we have

$$n\hat{\phi}(\tilde{q}_n + \frac{q}{\sqrt{n}}, \alpha_n) - n\hat{\phi}(\tilde{q}_n, \alpha_n) - \frac{q^T}{\sqrt{n}}S_n - n\tilde{\phi}(\tilde{q}_n + \frac{q}{\sqrt{n}}, \alpha_n) - n\tilde{\phi}(\tilde{q}_n, \alpha_n) \to 0 \quad (3.10)$$

for each fixed q in probability. $n\tilde{\phi}(\tilde{q}_n + \frac{q}{\sqrt{n}}, \alpha_n) - n\tilde{\phi}(\tilde{q}_n, \alpha_n) = \frac{1}{2}q^T\mathbf{H}\tilde{\phi}(\tilde{q}, \alpha^*)q + o_p(1)$. Because $n\hat{\phi}(\tilde{q}_n + \frac{q}{\sqrt{n}}, \alpha_n) - n\hat{\phi}(\tilde{q}_n, \alpha_n) - \frac{q^T}{\sqrt{n}}S_n$ is convex and converges to $\frac{1}{2}q^T\mathbf{H}\tilde{\phi}(\tilde{q}, \alpha^*)q$ in probability. Therefore according to lemma I, there exists a M and N as $n > N$,

$$\sup_{|q| \leq M} |n\hat{\phi}(\tilde{q}_n + \frac{q}{\sqrt{n}}, \alpha_n) - n\hat{\phi}(\tilde{q}_n, \alpha_n) - \frac{q^T}{\sqrt{n}}S_n - \frac{1}{2}q^T\mathbf{H}\tilde{\phi}(\tilde{q}, \alpha^*)q| < \epsilon \quad (3.11)$$

holds with probability at least $1 - \epsilon$. Since $\mathbf{H}\tilde{\phi}(\tilde{q}, \alpha^*)$ is bounded,

$$||\mathbf{H}^{-1}\tilde{\phi}(\tilde{q}, \alpha^*)\frac{S_n}{\sqrt{n}}|| < M - 1 \quad (3.12)$$

for such $n > N$ with probability at least $1 - \epsilon$. Let $g(q) = \frac{q^T}{\sqrt{n}}S_n + \frac{1}{2}q^T\mathbf{H}q$, $f(q) = n\hat{\phi}(\tilde{q}_n + \frac{q}{\sqrt{n}}, \alpha_n) - n\hat{\phi}(\tilde{q}_n, \alpha_n)$, $q_1 = \arg\min g(q) = -\frac{1}{\sqrt{n}}\mathbf{H}^{-1}S_n$, $q_2 = \arg\min f(q) =$

$\sqrt{n}(\hat{q}_n - \tilde{q}_n)$.

Let $C = 2(\inf_{|e|=1} e^T \mathbf{H} e)^{-1/2}$. Then consider the sphere $S = \{q : ||q - q_1|| = C\epsilon^{1/2}\}$.

From equation (3.5) we know that $f(q)|_{q\in S} > g(q)|_{q\in S} - \epsilon$. Because $g(q)$ can be written as: $g(q) = \frac{1}{2}(\mathbf{H}^{1/2}q - \mathbf{H}^{-1/2}\frac{S_n}{\sqrt{n}})^T(\mathbf{H}^{1/2}q - \mathbf{H}^{-1/2}\frac{S_n}{\sqrt{n}}) - S_n^T\mathbf{H}^{-1}S_n/2n$. Then for $q \in S$, suppose $q = q_1 + l$. $g(q_1 + l) = \frac{1}{2}(l/C\epsilon^{1/2})^T\mathbf{H}(l/C\epsilon^{1/2})C^2\epsilon + g(q_1) \geq 1/2 \times 4/C^2 \times C^2\epsilon + g(q_1) = 2\epsilon + g(q_1)$, Therefore

$$f(q)|_{q\in S} > g(q)|_{q\in S} - \epsilon > \epsilon + g(q_1) > f(q_1) \tag{3.13}$$

This equation indicates on the whole sphere S, $f(q)$ has its value larger than the center of sphere. Combined with convexity of $f(q)$, we have minimizer of $f(q)$ must be inside S otherwise there must exists a point on sphere with its function value less than center. Therefore we get $||\sqrt{n}\hat{q}_n + \mathbf{H}^{-1}\frac{S_n}{\sqrt{n}}|| < C\epsilon^{1/2}$ holds with probability at least $1 - 2\epsilon$.

According to assumption vi), let $Z_{i,n} = g(x_i, \tilde{q}_n, \alpha_n)$, $\lim_{n\to\infty} n\mathbf{E}\ Z_{1,n}Z_{1,n}^T = \mathbf{E}(\lim_{n\to\infty} g(x_1, \tilde{q}_n, \alpha_n)g(x_1, \tilde{q}_n, \alpha_n)^T) = \mathbf{E}(g(x_1, \tilde{q}, \alpha^*)g(x_1, \tilde{q}, \alpha^*)^T) = \Sigma$

The change of integral is because we assume uniform boundedness for subgradient.

$\lim_{n\to\infty} \sum_{i=1}^n \mathbf{E}(||Z_{i,n}||^2; ||Z_{i,n}|| > \epsilon) = \lim_{n\to\infty} \mathbf{E}(||g(x_1, \tilde{q}_n, \alpha_n)||^2; ||g(x_1, \tilde{q}_n, \alpha_n)|| > \sqrt{n}\epsilon) = 0$. Therefore $\sqrt{n}(\hat{q}_n - \tilde{q}_n)$ is asymptotic normal.

Because $\mathbf{D}\tilde{\phi}(\tilde{q}, \alpha^*) = 0$ and $\mathbf{H}\tilde{\phi}(\tilde{q}, \alpha^*) > 0$. Then according to implicit function theorem, there exists an open set U containing $\tilde{q}$, an open set V containing α^*, and a unique continuously differential function $g : V \to U$ such that:

$$\{(g(y), y)|y \in V\} = \{(x, y) \in U \times V|\mathbf{D}\tilde{\phi}(x, y) = 0\}$$

Because $\mathbf{D}\tilde{\phi}(\tilde{q}_n, \alpha_n) = 0$. Thus for very large n, since g is unique, $\tilde{q}_n = g(\alpha_n) = g(\alpha^*) + g'(*)^T(\alpha_n - \alpha^*) = \tilde{q} + g'(*)^T(\alpha_n - \alpha^*)$, where * is between α_n and α^*. Since * is in a closed set hence $g'(*)$ bounded. Up to this point, it is clear that if $||\alpha_n - \alpha^*|| = O_p(n^{-1/2})$, then $\sqrt{n}(\tilde{q}_n - \tilde{q})$ is asymptotic normal.

Theorem VI If iv), vii), viii) hold, then $\sqrt{n}r_n\kappa_n(\hat{q}_n - \tilde{q}_n) = -\frac{\kappa_n\mathbf{H}_n^{-1}}{\sqrt{n}}r_nS_n + o_p(1)$. $\sqrt{n}r_n\kappa_n(\hat{q}_n - \tilde{q}_n) \to N(0, \mathbf{H}\Sigma\mathbf{H})$, where $\kappa_n\mathbf{H}_n^{-1} \to \mathbf{H}$, $||r_ng(X, \tilde{q}_n, \alpha_n) - g|| \to 0$, $\Sigma = \mathbf{Var}g$.

Proof:

Let $K(n) = n^{-1/2}(r_n\kappa_n)^{-1}$. Consider $X_{ni} = \phi(x_i, \tilde{q}_n + K(n)q, \alpha_n) - \phi(x_i, \tilde{q}_n, \alpha_n) - q^TK(n)g(x_i, \tilde{q}_n, \alpha_n)$. By following the same reasoning of proof V we can get:

$\sum_{i=1}^n \mathbf{E}X_{ni}^2 \leq \mathbf{E}[n^{1/2}K(n)q^T(g(x_1, \tilde{q}_n + K(n)q, \alpha_n) - g(x_1, \tilde{q}_n, \alpha_n)]^2$
$\leq (r_n\kappa_n)^{-2}\mathbf{E}||q||^2E||g(x_1, \tilde{q}_n + K(n)q, \alpha_n) - g(x_1, \tilde{q}_n, \alpha_n)||^2 \leq n^{-1}(r_n\kappa_n)^{-4}||q||^4$.

This is by applying Cauchy-Schwarz inequality, Lipchitz condition for sub gradient and the definition of matrix norm. Recall from the irregular condition that $n^{1/4}(r_n\kappa_n) \to \infty$. According to assumptions vi), vii) and viii), we conclude that this quantity goes to 0 almost surely.

$$n\hat{\phi}(\tilde{q}_n + K(n)q, \alpha_n) - n\hat{\phi}(\tilde{q}_n, \alpha_n) - K(n)q^T S_n - \frac{nK^2(n)q^T \mathbf{H}_n q}{2} \to 0$$

Let $f(q) = n\hat{\phi}(\tilde{q}_n + K(n)q, \alpha_n) - n\hat{\phi}(\tilde{q}_n, \alpha_n)$, $g(q) = K(n)q^T S_n + \frac{nK^2(n)q^T \mathbf{H}_n q}{2}$,

Since it is obvious to see $g(q) \to 0$. Therefore $f(q) \to 0$ as well. And because they are both convex, by applying lemma I and we have:

$$\sup_{||q|| \leq M} |f(q) - g(q)| < \sup_{||q|| \leq M} |f(q)| + \sup_{||q|| \leq M} |g(q)| \leq \epsilon \tag{3.14}$$

holds with probability at least $1 - \epsilon$.

The minimizer of $g(q)$ is $q_1 = -(K^2(n)\mathbf{H}_n)^{-1}K(n)S_n/n = -r_n k_n S_n/n^{1/2}$ and we can select a M such that $||q_1|| \leq M - 1$ with probability at least $1 - \epsilon$. The minimizer of $f(q)$ is $q_2 = n^{1/2} r_n k_n (\hat{q}_n - \tilde{q}_n)$.

According to assumption viii), we know that $C_n = 2(\inf_{|e|=1} e^T \kappa_n H_n^{-1} e)^{-1/2} > 0$ and bounded. Similar to the previous proof, let us consider the sphere $S^{(n)} = \{q : |q - q_1| \leq C_n \epsilon^{1/2}\}$. One can easily prove that:

$$f(q)|_{q \in S^{(n)}} > g(q)|_{q \in S^{(n)}} - \epsilon > \epsilon + g(q_1) > f(q_1) \tag{3.15}$$

Therefore we can argue that the minimizer of $f(q)$ should be in $S^{(n)}$. Finally we have $||q1 - q2|| = ||n^{1/2} r_n k_n (\hat{q}_n - \tilde{q}_n) + r_n k_n S_n / n^{1/2}|| < C_n \epsilon^{1/2}$ holds with at least probability $1 - 2\epsilon$. The asymptotic normality follows for the same reasoning as in the proof of theorem V.

Theorem VI says if Hessian matrix $\mathbf{H}_n$ asymptotically singular, we can still have asymptotic normality by multiplying a scalar different than $\sqrt{n}$ to the difference between the M-estimator and its target under some conditions. The following theorem will establish a result by multiplying a scalar c_n to $\hat{q}_n - \tilde{q}_n$. Notice that the result will indicate the M-estimator may not necessarily converge to normal in distribution. The asymptotic distribution depends how fast $\alpha_n \to \alpha^*$ and the shape of $F(.)$.

Theorem VII If conditions iv) vi) vii) ix) hold, then $(c_n r_n)^{-1}(\hat{q}_n - \tilde{q}_n) \to \arg\min_q V(q)$ in distribution, where $V(q) = q^T W + L(q)$, $W \sim \mathrm{N}(0, \Sigma)$ with $\Sigma = \mathbf{E}g(X, \tilde{q}, \alpha^*)g(X, \tilde{q}, \alpha^*)^T$.

Lemma VII (Convexity Lemma)

Assuming Z_n and Z are random lower semicontinuous function and Z_n epi-converges to Z in distribution.
(A) U_n is an $\epsilon_n - \arg\min$ of Z_n where $\epsilon \to 0$ in probability.
(B) $U_n = O_p(1)$
(C) Z has an almost sure unique argmin U.
Then $U_n \to U$ in distribution. Particularly, If $Z_n \to Z$ in finite dimensional distribution and Z has probability 1 to be finite on an open set, then Z_n epi-converges to Z.

This Lemma is the combination of theorem 1 and theorem 5 from Knight(1999) [37].

Proof:

Consider $V_n(q) = n^{-1/2} c_n^{-1} \sum_{i=1}^n (\phi(x_i, \tilde{q}_n + c_n r_n q, \alpha_n) - \phi(x_i, \tilde{q}_n, \alpha_n))$. It can be

decomposed into three parts:

$$R_{1n}(q) = n^{-1/2}c_n^{-1}\sum_{i=1}^{n}(\phi(x_i,\tilde{q}_n+c_nr_nq,\alpha_n)-\phi(x_i,\tilde{q}_n,\alpha_n)-c_nr_nq^Tg(x_i,\tilde{q}_n,\alpha_n))-$$
$$n^{1/2}c_n^{-1}(\tilde{\phi}(\tilde{q}_n+c_nr_nq)-\tilde{\phi}(\tilde{q}_n,\alpha_n))$$

$$R_{2n}(q) = n^{1/2}c_n^{-1}(\tilde{\phi}(\tilde{q}_n+c_nr_nq)-\tilde{\phi}(\tilde{q}_n,\alpha_n))$$

$$R_{3n}(q) = n^{-1/2}r_n\sum_{i=1}^{n}q^Tg(x_i,\tilde{q}_n,\alpha_n)$$

(a) First of all, by assumption ix), we know that $R_{2n}(q)$ converges to $L(q)$ pointwisely. Again by Lemma I), we know that this convergence can happen uniformly on any compact set S.

(b) $R_{1n}(q) = \sum_{i=1}^{n}X_{ni}$ where $X_{ni} = n^{-1/2}c_n^{-1}[(\phi(x_i,\tilde{q}_n+c_nr_nq,\alpha_n)-\phi(x_i,\tilde{q}_n,\alpha_n)-$
$c_nr_nq^Tg(x_i,\tilde{q}_n,\alpha_n))-(\tilde{\phi}(\tilde{q}_n+c_nr_nq)-\tilde{\phi}(\tilde{q}_n,\alpha_n))]$

$$\mathbf{Var}(R_{1n}(q)) = n\mathbf{E}X_{n1}^2 = c_n^{-2}\mathbf{Var}(\phi(x_i,\tilde{q}_n+c_nr_nq,\alpha_n)-\phi(x_i,\tilde{q}_n,\alpha_n)-c_nr_nq^Tg(x_i,\tilde{q}_n,\alpha_n))$$

$$\leq c_n^{-2}\mathbf{E}(\phi(x_1,\tilde{q}_n+c_nr_nq,\alpha_n)-\phi(x_1,\tilde{q}_n,\alpha_n)-c_nr_nq^Tg(x_1,\tilde{q}_n,\alpha_n))^2$$

Because $0 \leq \phi(x_1,\tilde{q}_n+c_nr_nq,\alpha_n)-\phi(x_1,\tilde{q}_n,\alpha_n)-c_nr_nq^Tg(x_1,\tilde{q}_n,\alpha_n) \leq c_nr_nq^T(g(x_1,\tilde{q}_n+$
$c_nr_nq,\alpha_n)-g(x_1,\tilde{q}_n,\alpha_n))$; Therefore,

$$c_n^{-2}\mathbf{E}(\phi(x_1,\tilde{q}_n+c_nr_nq,\alpha_n)-\phi(x_1,\tilde{q}_n,\alpha_n)-c_nr_nq^Tg(x_1,\tilde{q}_n,\alpha_n))^2 \leq ||q||^2r_n^2\mathbf{E}||g(x_1,\tilde{q}_n+$$
$c_nr_nq,\alpha_n)-g(x_1,\tilde{q}_n,\alpha_n)||^2 \leq ||q||^4c_n^2r_n^4$ for fixed q. According to condition ix), this term goes to 0. Hence $\mathbf{Var}(R_{1n}(q)) \to 0$.

Therefore $R_{1n}(q) \to 0$ in probability for each fixed q. Noticed that $R_{1n}(q) =$

$n^{-1/2}c_n^{-1}\sum_{i=1}^n(\phi(x_i,\tilde{q}_n+c_nq,\alpha_n)-\phi(x_i,\tilde{q}_n,\alpha_n)-c_nq^Tg(x_i,\tilde{q}_n,\alpha_n))-R_{2n}(q)$, therefore $n^{-1/2}c_n^{-1}\sum_{i=1}^n(\phi(x_i,\tilde{q}_n+c_nq,\alpha_n)-\phi(x_i,\tilde{q}_n,\alpha_n)$ converges pointwisely. Again because this function is convex hence according to Lemma I it converges uniformly on any compact set S. Therefore $R_{1n}(q)$ converges to 0 in probability on any compact set S.

(c) Lastly, according to the same reasoning of Theorem V, we have $n^{-1/2}\sum_{i=1}^n q^Tg(x_i,\tilde{q}_n,\alpha_n)\to q^TW$ where $W\sim \mathrm{N}(0,\Sigma)$ with $\Sigma=\mathbf{E}g(x_1,\tilde{q},\alpha^*)g(x_1,\tilde{q},\alpha^*)^T$.

To synthesize (a), (b) and (c), we have $V_n(q)\to V(q)$ in finite dimensional distribution. This is because Slutsky's theorem and uniform convergence for $R_{1n}(q)$ and $R_{2n}(q)$ on any compact set. Finally by applying Lemma VII we claim that: $(c_nr_n)^{-1}(\hat{q}_n-\tilde{q}_n)\to\arg\min_q V(q)$ in distribution.

Remark To rule out the infinite values with positive probability for Z is crucial for Lemma VII. It forces us to require $L(q)$ to be finite on any open set. Without this finite value assumption, one can easily be mislead by the following counter example provided by Knight(1999) [37]:

Let $Z_n(u)=n(u-U_n)^2$ where $U_n\to U$ in distribution with U a continuous random variable. Then $Z_n\to\infty$ in finite distribution while Z_n epi-converges to $Z=\infty I(u\neq U)+0I(u=U)$.

Corollary I Consider the one dimensional case, if $n^{1/2}\ \phi''(\tilde{q}_n,\alpha_n)\to\infty$, then the necessary and sufficient condition for asymptotic normality is and $n^{1/4}r_n^{1/2}\tilde{\phi}''(\tilde{q}_n,\alpha_n)$ $|\tilde{\phi}'''(\tilde{q}_n,\alpha_n)|^{-1/2}\to\infty$. And $\sup\{\beta:n^\beta r_n^{1/2}\tilde{\phi}''(\tilde{q}_n,\alpha_n)\to\infty\}=1/4$ is required to guarantee asymptotic normality for all distributions F. In other words, any $\beta>1/4$

cannot guarantee asymptotic normality for some distributions F and ϕ.

Proof:

By using Taylor expansion,

$$\tilde{\phi}(\tilde{q}_n + c_n r_n q, \alpha_n) - \tilde{\phi}(\tilde{q}_n, \alpha_n) = \tilde{\phi}''(\tilde{q}_n, \alpha_n)(c_n r_n q)^2/2 + \tilde{\phi}'''(*, \alpha_n)(c_n r_n q)^3/6 \quad (3.16)$$

The $*$ is between $\tilde{q}_n$ and $\tilde{q}_n + c_n r_n q$. In order to force normality with variance Σ, we require $L(q) = n^{1/2} c_n^{-1}(\tilde{\phi}(\tilde{q}_n + c_n r_n q, \alpha_n) - \tilde{\phi}(\tilde{q}_n, \alpha_n)) = q^2/2$. From the first term of equation (3.11) we know that $c_n = [n^{1/2} r_n^2 \tilde{\phi}''(\tilde{q}_n, \alpha_n)]^{-1}$. The condition $n^{1/2} f(\tilde{q}_n) \to \infty$ leads to $c_n r_n^2 \to \infty$ which is what we need in condition ix). In addition, we need to ensure the last term goes to zero, therefore $n^{1/4} r_n^{1/2} \tilde{\phi}''(\tilde{q}_n, \alpha_n) |\tilde{\phi}'''(\tilde{q}_n, \alpha_n)|^{-1/2} \to \infty$.

We need to show that $\tilde{\phi}'''(\tilde{q}_n, \alpha_n)$ is not necessarily diminishing to 0 as $n \to \infty$ for some distributions. If this is true, then we must at least require $n^{1/4} r_n^{1/2} \tilde{\phi}''(\tilde{q}_n, \alpha_n) \to \infty$ inorder to force asymptotic normality. We can easily construct such an example. For example, if we choose ϕ as quantile objective function, and the underlying density function is $f(x) = \frac{6}{\pi} \frac{\sin(x^3)}{x} I(x > 0)$. It is easy to check $f(x)$ is density function by applying Dirichlet Integral. Under this case, the second derivative of $\tilde{\phi}$ is $2f(x)$ which goes to zero as $x \to \infty$. However the third derivative is $\frac{6}{\pi}(3x\cos(x^3) - \sin(x^3)/x^2)$ which is certainly not going to zero as $x \to \infty$. So far, we have proved that $\sup\{\beta : n^{\beta} r_n^{1/2} \tilde{\phi}''(\tilde{q}_n, \alpha_n) \to \infty\} \leq 1/4$ because we found a case that requires $n^{1/4} r_n^{1/2} \tilde{\phi}''(\tilde{q}_n, \alpha_n) \to \infty$. However according to Theorem VI, we have proved that if $n^{1/4} r_n^{1/2} \tilde{\phi}''(\tilde{q}_n, \alpha_n) \to \infty$ then the asymptotic normality is a sure thing, which means $\sup\{\beta : n^{\beta} r_n^{1/2} \tilde{\phi}''(\tilde{q}_n, \alpha_n) \to \infty\} \geq 1/4$. To sum up,

$\sup\{\beta : n^{\beta} r_n^{1/2} \tilde{\phi}''(\tilde{q}_n, \alpha_n) \to \infty\} = 1/4.$

Corollary II Suppose X has density function $f(x)$ with support $[x_m, \infty)$ where x_m can be $-\infty$. Suppose q_n and $\hat{q}_n$ are the α_nth population and sample quantile respectively and $n^{1/2} f(q_n) \to \infty$. We have the following conclusions:

(a) $n^{1/2}|f(q_n)/f'(q_n)|(1-\alpha_n)^{1/2} \to \infty$.
(b) $f'(x)/f(x) \leq -1$ for large x.

If both (a) and (b) hold, then $n^{1/2} f(q_n)(1-\alpha_n)^{-1/2}(\hat{q}_n - q_n)$ is asymptotically normal. Or,

If the underlying distribution function is von Mises function : $\lim_{x\to\infty} \frac{d}{dx}(\frac{1-F(x)}{f(x)}) = 0$, then the sufficient and necessary conditions for asymptotic normality for sample quantile is $n(1-\alpha_n) \to \infty$

Proof:

For part I, if both (a) and (b) hold, $\int_{q_n}^{\infty} f(x)dx = 1 - \alpha_n$;

$n^{1/2} f(q_n)/f'(q_n)(1-\alpha_n)^{1/2} = \int_{q_n}^{\infty} n^{1/2} f(q_n)/f'(q_n) f(x)(1-\alpha_n)^{-1/2} dx$

Let $G(t) = \int_t^{\infty} f(x)dx - f(t)$, then $G(\infty) = 0$ and $G'(t) = -f(t) - f'(t) \geq 0$ for large t. Therefore if n is lagre, $\int_{q_n}^{\infty} f(x)dx \leq f(q_n)$.
Hence $n^{1/2} f(q_n)/f'(q_n)(1-\alpha_n)^{1/2} \leq n^{1/2} f^2(q_n)/f'(q_n)(1-\alpha_n)^{-1/2}$.

Thus $n^{1/2} f^2(q_n)/f'(q_n)(1-\alpha_n)^{1/2} \to \infty$. According to Corollary, the asymptotic

normality is valid.

For part II, now instead of assuming (a) and (b), we consider the von Mises function.

$n^{1/2}(1-\alpha_n) = \frac{n^{1/2}f(q_n)^2}{f'(q_n)}\frac{f'(q_n)}{f(q_n)^2}(1-F(q_n))$. Let $A = \frac{f'(q_n)}{f(q_n)^2}(1-F(q_n))$

Simple calculation can show that $\frac{d}{dx}(\frac{1-F(x)}{f(x)}) = -1 - A$. Therefore $A \to -1$ as $x \to \infty$. So $n^{1/2}(1-\alpha_n)^{-1/2}$ has the same order as $\frac{n^{1/2}f(q_n)^2}{f'(q_n)}(1-\alpha_n)^{-1/2}$. Therefore according to corollary I, the sufficient and necessary condition for asymptotic normality is $n(1-\alpha_n) \to \infty$.

Remark

The condition (b) is valid for a wide range of distributions. For example, Normal distribution with $f(x) = 1/\sqrt{2\pi}\exp\{-x^2/2\}$, $f'(x) = -1/\sqrt{2\pi}2x\exp\{-x^2/2\}$ $\leq f(x)$ for large x. $\Gamma(\alpha,\beta)$distribution : $f(x|\alpha,\beta) = \beta^\alpha/\Gamma(\alpha)x^{\alpha-1}e^{-\beta x}I(x>0)$ and $f'(x|\alpha,\beta) = \beta^\alpha/\Gamma(\alpha)(\alpha-1)I(x>0)x^{\alpha-2}e^{-\beta x} + \beta^\alpha/\Gamma(\alpha)I(x>0)x^{\alpha-1}(-\beta)e^{-\beta x}$. Therefore as long as $\beta > 1$, condition (b) is valid.

There is a noticable fact that if condition (b) holds, then $1-\alpha_n$ can be at most as fast as $O(n^{-1/2})$ in order to achieve asymptotic normality. Therefore the order $O(n^{-1/2})$ is sharp under condition (b).

3.4 Application

The asymptotic theory in this paper can be applied to asymptotic quantiles under both regular and irregular conditions. Also it can be applied to some other problems

such as smoothing algorithm for quantile regression. By saying smoothing, it means that a sequence of smooth functions are used to approximate quantile objective function, see Chen(2007) [15], Hunter and Lange (2000) [34].

Example 1.

In p-dimensional space $\mathbb{R}^p$, for any random variable $X \in \mathbb{R}^p$ and every $u \in B_p$, the uth quantile $Q(u)$ is:

$$\tilde{\phi}(q,u) = \mathbf{E}[||X-q|| + u^T(X-q)] \tag{3.17}$$

The properties of Chaudhuri's spatial quantile has been well developed, see Chaudhuri(1996) [14]. Additionally, Mukhopadhyay and Chatterjee(2011) [44] obtained asymptotic distributions of generalized spatial quantiles under regular conditions. Sometimes instead of considering a fixed u, we are more interested in a changing u_n according to different sample size. Especially it makes great sense when trying to study the behavior of extreme quantile. In one dimensional case, for instance, maximum or minimum sample point could be outliers. Therefore we do not directly studies the extreme sample quantile and alternatively we choose α_nth quantile where $\alpha_n \to 1$. This sample quantile we obtain can still maintain asymptotic properties with its target population extreme quantile. Therefore the ϕ function we are considering is:

$$\phi(X,q,u_n) = ||X-q|| + u_n^T(X-q) \tag{3.18}$$

where $u_n \to u^* \in B_p$. It is easy to check this ϕ function satisfies our conditions i) iv). Therefore $\hat{q}_n$ is consistent. The subgradient is:

$$g(X, q, u) = (q - X)/||q - X|| - u \tag{3.19}$$

Again, it satisfies condition vi) and vii). The asymptotic normalilty will follow naturally if one assume the hessian of $\tilde{\phi}(\tilde{q}, u^*)$ is positive definite. However even if the hessian of $\tilde{\phi}(\tilde{q}, u^*)$ is not positive definite, once our irregular condition viii) holds, the sample minimizer still has weak convergence. It is particularly interesting if we look at one dimensional case where the second derivatives of $\phi(q, \alpha)$ is just $2f(q)$. If the upper bound of the support of $f(q)$ is infinity. Then for the extreme quantile the second derivative of $\phi(\infty, 1)$ is 0. To have asymptotic normal justified, we need $n^{1/2} f(\tilde{q}_n) \to \infty$. This means α_n should not go to 1 too rapidly.

Example 2

Let us investigate into one dimensional case (Corollary II of Theorem VII) in details. Let's look at the setting where X is sampled from exponential distribution with density function $f(x) = e^{-x} I(x > 0)$.

According to our corollary, the asymptotic distribution of $\sqrt{n(1-\alpha_n)}(\hat{q}_n + \log(1-\alpha_n))$ is normal if $n^{1/2}(1-\alpha_n) \to \infty$. In order to check this fact, we let $\alpha_n = 1 - \frac{1}{n^k}$ where $k = 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$ for each simulation round. We set simulation round at 2000 and for each round we have sample size 2000. The p value is calculated 100 times and then averaged out. What we can observe from the above plots panel is that the first three plots show a bell shape while the rest have different

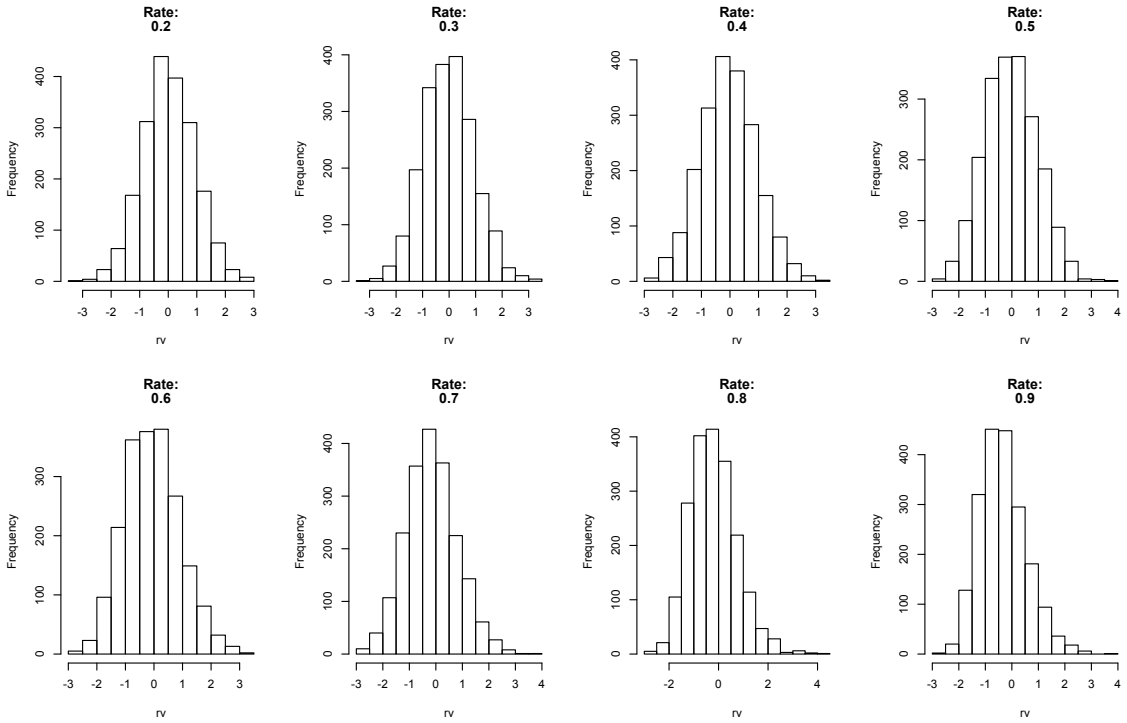


Figure 3.1: Histogram at different rate

degree of skewness. This observation supports our conclusion which is asymptotic normality follows if rate is less than 0.5. Numerically, the average p-value of Shapiro test results is: 0.367 0.324 0.231 0.085 0.010 2.604709e-05 8.102961e-07 1.056834e-11 for $k = 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$. This again numerically proved our conclusion.

Example 3.

Colin Chen(2007) used a Huber-like function which aims at avoiding nondifferentiable characteristics of quantile check function, see Figure 1. Numerical comparison shows that the finite smoothing algorithm outperforms the simplex algorithm in computing speed. Compared with the powerful interior point algorithm, it is competitive

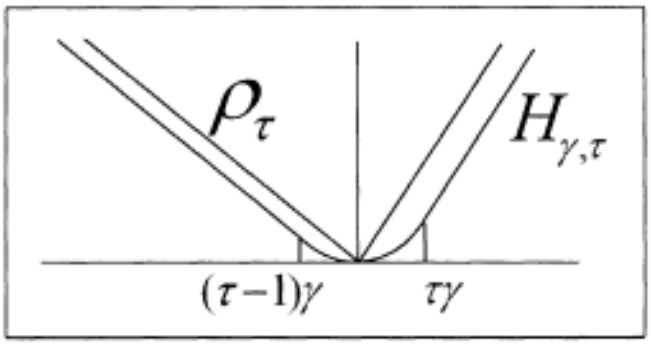


Figure 3.2: finite smoohting quantile function

overall;however, it is significantly faster than the interior point algorithm when the design matrix in quantile regression has a large number of covariates. Additionally, the new algorithm provides the same accuracy as the simplex algorithm. In contrast, the interior point algorithm gives only the approximate solutions in theory. Here we justify some asymptotic results which have not been studied yet. Instead of using ordinary quantile check function, Chen used the following function to model median:

$$\phi(X,\beta,\alpha_n) = (\frac{(y-x^T\beta)^2}{2\alpha_n}+\frac{\alpha_n}{2})I(|y-x^T\beta|\leq\alpha_n)+|y-x^T\beta|I(|y-x^T\beta|>\alpha_n) \quad (3.20)$$

Where $X=(y,x)^T$, $\alpha_n\to 0$ as $n\to\infty$. Our concern is the asymptotic properties of $\hat{\beta}_n$ which is the minimizer of $\sum_{i=1}^n \phi(X_i,\beta,\alpha_n)$. And suppose $\tilde{\beta}$ is the minimizer of $\mathbf{E}_{(y,x)}|y-x^T\beta|$. The first three conditions are easy to check. Since fourth condition is required to derive uniform convergence of $\hat{\phi}_n(q,\alpha_n)$. Therefore we only need to check equation (4.1) uniformly converges to $|y-x^T\beta|$ for fixed ω as $n\to\infty$ which is obvious. Then it follows from theorem III that $\hat{\beta}_n$ is a consistent estimator of $\tilde{\beta}$. Onc can show the gradient and Hessian are:

$$g(X,\beta,\alpha_n)=\frac{(x^T\beta-y)}{\alpha_n}xI(|x^T\beta-y|<\alpha_n)+xI(x^T\beta-y>\alpha_n)-xI(x^T\beta-y<\alpha_n) \tag{3.21}$$

$$\mathbf{H}_\beta\tilde{\phi}(y,\beta,\alpha_n)=\mathbf{E}_x\{(xx^T)\frac{1}{\alpha_n}(F_{y|x}(x^T\beta+\alpha_n)-F_{y|x}(x^T\beta-\alpha_n)\} \tag{3.22}$$

From the form of gradient and Hessian we can verify conditions v), vi) vii). Condition v) can be justified as long as F is continuous. Then we apply mean value theorem to get $H=\mathbf{H}_\beta\tilde{\phi}(y,\beta,0)=\mathbf{E}_X\{(xx^T)2f_{y|x}(x^T\beta)\}$ and so for finite β it corresponds to the regular condition. Therefore $\sqrt{n}(\hat{\beta}_n-\tilde{\beta})$ is asymptotically normal with asymptotic covariance matrix $H^{-1}\mathbf{Var}gH^{-1}$, where $g=xI(x^T\tilde{\beta}>y)-xI(x^T\tilde{\beta}<y)$.

3.5 Discussion

The paper presents the large sample properties of the M-estimators by minimizing a series of convex functions with a additional control parameter. Compared to the work of Niemiro (1992), we add a control parameter which extends model flexibility in multiple ways. In our examples stated in the previous section, the role of control parameter could be providing a robust inference for extreme quantile or obtaining asymptotics for finite smoothing quantile regression. Similarly, Chernozhukov, V. (2005) [16] considered a so-called "extremal quantile regression" where the control parameter τ_T also varies according to different sample size. The author proved asymptotics under extreme conditions $T\tau_T\to k>0$.The "extreme quantile regression"

is a sub-category of our problem set-up. We proposed intermedia order condition $||n^{1/4}r_n\mathbf{H}_n|| \to \infty$ which is essential for asymptotic normality. Both theoretical and numerical facts demonstrated this condition defines the converging rate beyond which the asymptotic distribution no longer falls into normal category. Nevertheless, what is beyond normal is still not revealed in this paper. Further research is required to learn more about the property of the arg min estimator in theorem VII. Similarly, Bose and Chatterjee(2001) [4] worked out the asymptotics of M-estimators under non-regular conditions which means F is not differentiable at some points. Keeping in mind that the irregular condition in this paper means asymptotic singular Hessian matrix which is different than theirs. Our future research interest is if $\alpha_n \to \alpha^*$ very fast, can we still work out the asymptotic distribution for the M-estimators ?

Another important aspect for further studies is Badahur representations of our M-estimators. Badahur representation describe a representation of sample quantile hence gains more insights into asymptotics of sample quantile. Niemiro(1992) extend this representation to M-estimator in theorem 5. It provides a precise order of weak convergence. Thus we may consider this as our future work to establish extended Bahadur representations under our model.

Chapter 4

Regression Percentiles based Variable Selection

4.1 Overview

In this chapter, we will devote the efforts to learning variable selection methods based on regression percentiles. The variable selection based on regression percentiles has the advantage of taking into account the full picture of the conditional distribution. Existing literatures have stressed the need for quantile regression or expectile regression to form a effective variable selection method. In Li, Y. and Zhu, J. (2007) [42], they pointed out that the fused quantile regression should be adopted to analyze array-CGH data. They believed considering quantile regression would gain more information to detect the regions of gains an losses. Wang, L. (2012) [55] has investigated in analyzing heterogeneity in ultra-high Dimension by using quantile regression. One of the take away messages from that paper is if heterogeneity is presented in the data, mean or median based variable based methods may not function well. In the simulation example in the paper, we saw extreme efficiency of quantile regression to detect the variables cause heterogeneity whereas other mean based methods can fail.

In chapter 2, the power of machine learning techniques (GB, QRF, RGB) for

learning quantile regression is demonstrated regarding prediction of quantiles. We have not yet studied the performance of these machine learning techniques for variable selection under the framework of quantile regression. Though the functionality of variable selection of random forests and gradient boosting is recognized by people for conditional mean model. Seldom did people notice the this success can still maintain even under quantile or expectile models. In this chapter, we will study the variable selection based on gradient boosting (GB or RGB) for regression percentiles. In section 4.2, we will go over gradient boosting based variable importance measurement. In section 4.3, we will propose heuristics variable selection methods based on gradient boosting for both quantile and expectile model. In section 4.4, the LARS-like solution path for expectile is provided to analyze the variable selection problem.

4.2 Gradient Boosting based Variable Selection

It has been pointed by Buhlmann [9] [6] that Boosting does variable selection and it assigns variable amount of degrees of freedom to the selected predictor variables. In Buhlmann"s 2006 paper [7], he proved that boosting with L2 loss is consistent for very high-dimensional models where the number of predictor variables is allowed to grow essentially as fast as exponential of sample size. The twin-boosting proposed by Buhlmann [8] is a parallel version of adaptive lasso [58] in boosting world. Essentially, the twin-boosting consists of two stages: 1) the first stage is to operate ordinary gradient boosting; 2) the second stage is enforced to resemble the first stage. Twin Boosting with componentwise linear least squares is proved to be equivalent to the adaptive Lasso for the case of an orthonormal linear model and it is empirically shown, in general and for various base procedures and models. It is demonstrated in that paper empirically that twin-boosting will choose much smaller number of total variables than boosting but get less incorrect number of variables at the same time.

One ting should be noticed that the variable selection is purely from selecting non-zero estimated value coefficients. This type of method may have risk in finding unrelated variables due to colinearity or different number levels for different variables. In this chapter, we adopt the variable important based heuristics for selecting variables for quantile and expectile regression rather than selecting all non-zero variables. Before we proceed with this heuristics, we review some background about the evaluation of variable importance in gradient boosting.

Variable Importance There is variable selection category which is based on variable importance ranking instead of shrinking inactive variables to zero. We will list some of them. Guyon et al. (2002) [27] and Rakotomamonjy (2003) [50], propose methods based on SVM scores and using descending elimination; Poggi et al. (2006) [47] propose a method based on CART scores and using stepwise ascending procedure with elimination step; Recently, Genuer et al. (2006) [25]suggest two different variable selection heuristics using random forests for two different variable scenarios: 1) to find important variables highly related to the response variable for interpretation purpose; 2) to find a small number of variables sufficient to a good prediction of the response variable. In this thesis, we will mainly focus the first type of variable selection. Let us consider more generally a tree-ensemble method (GB, RF...) such that the estimated function $\hat{f}(x) = \sum_{i=1}^{m} \hat{f}_i(x)$. Each $\hat{f}_i(x)$ is a CART. Let VI be variable importance. Then $\text{VI}(\hat{f})(x) = \sum_{i=1}^{m} \text{VI}\ (\hat{f}_i)(x)$. Let us assume the response is continuous, then the variable importance is measured by variable reduction (If the response is discrete, then the variable importance is measured by gini gain). Suppose f is a CART, for variable x, if x appears as the splitting variable at t different positions, assume v_{jT}, v_{jL}, v_{jR} are variance of subsamples at parent's node, left child and right child respectively. Then $\text{VI}(f)(x) = \sum_{j=1}^{t}(v_{jT} - v_{jL} - v_{jR})$. So it is obvious that the variables with importance zero is the ones that has never been

chosen as splitting variable.

Though we have the definition for variable importance, we have not yet drafted the blueprint for selection variables for the real data. Let us at this stage be consistent with the rule that only selecting the non-zero points according to variable importance. We will focus on the following numerical simulation example in this chapter:

Model I $y = x_1 + x_2 + x_3 + x_4 + x_5^2\epsilon$; $n = 500, p = 500, \epsilon \sim N(0, 1)$

For model I, $X \sim N(0, \Sigma)$, where the entry $\Sigma_{ij} = 0.5^{|i-j|}$. As one can notice the model is heteroscedastic models. We present three figures: 4.1, 4.2 and 4.3 with quantiles 0.1, 0.9, 0.5 respectively. Since there are as many as 500 variables, the empty circles most all cumulate at the bottom with y-axis value 0. So what you see as a solid bar is in fact a cumulation of empty circles. We use dashed blue line to locate the position of index 5. We are concerned with the value of points that are on the left side of this bar. The red dashed line is the horizontal at zero. For gradient boosting, we use the learning rate $\mu = 0.001$, interaction depth 1 which means the base learner is just stump. What is in common for different quantiles is that as the number of iterations go up to 1000, the power of detecting important variables will correspondingly increase. However, we may notice that for median, the gradient boosting performs poorest. It can never select the 5th variable which is the variable that defines the heteroscedastic effect. After all, the conditional mean or median cannot reflect the relative change of conditional variance in a model. That is why we need to get information from a more comprehensive picture instead of just peeking the center alone. So both 0.1th quantile and 0.9th quantile have successfully detecting all variables at 1000 iterations. Another boosting worth mentioning is called ER-boosting [56]which is gradient boosting method based on expectile instead

of quantile. In chapter 1, we have introduced expectile as an optimization solution of asymmetric least squared regression. It can also reflect the information from every part of the distribution. The advantage of this method is mentioned in chapter one. Out of curiosity, we used erboost package in R to conduct variable selection for model I. The result appears very appealing for this model. For 0.5th expectile, it is relatively weak. For illustration purpose without too much verbose, we will only provide figures for 0.1th expectile (0.9 will be similar, 0.5 is relatively poor). So from table 4.2 and figure 4.4 we can see that ERboosting have basically equal achievement as GB for quantile.

In the next example, we may exercise a comparison between gradient boosting with a variety of competitive variable selecting methods with regard to model I. The candidate models are: L-lasso, L-Alasso (adaptive lasso), L-scad, erboost, Q-lasso, Q-scad. Both Q-lasso and Q-scad are from Knoenker's package *quantreg*. L-lasso, L-Alasso and L-scad are from lqa package. All tunning parameters are determined by 10 folds cross validation. The simulation runs 100 times for each methods. The comparison is measured by two things: 1) the frequency of the true variables that have been chosen by the model (the closer to 1, the better); 2) the total frequency of choosing the incorrect variables with index from 6 to 500 (the closer to 0, the better). Particularly, one may investigate in the frequency that variable 5 is chosen which may reflect the power of selecting the heterogeneity variable. So the results are presented in the table. The prefix "L" means linear model based penalty, "Q" means quantile regression model based penalty. It is very clear that GB is the most powerful method to detect V5 which is the source of heterogeneity except for the median. This makes sense, since from model I, we know that the conditional median can not be influenced the variable V5. What surprises us is that Q-scad and Q-lasso are weak in finding V5. Although they both perform perfectly in finding V1-4, yet one may expect they find V5. Also, compared to GB, Q-scad and Q-lasso tend to include the

Methods	V1	V2	V3	V4	V5	Others
GB-0.1	1.00	1.00	1.00	1.00	0.99	8.71
GB-0.5	1.00	1.00	1.00	0.99	0.00	0.00
GB-0.9	1.00	1.00	1.00	1.00	0.99	9.70
Q-scad-0.1	1.00	1.00	1.00	1.00	0.11	17.04
Q-scad-0.5	1.00	1.00	1.00	1.00	0.05	8.26
Q-scad-0.9	1.00	1.00	1.00	1.00	0.06	17.05
Q-lasso-0.1	1.00	1.00	1.00	1.00	0.23	45.51
Q-lasso-0.5	1.00	1.00	1.00	1.00	0.15	35.42
Q-lasso-0.9	1.00	1.00	1.00	1.00	0.17	46.11
ERBoost-0.1	0.90	1.00	1.00	0.99	0.96	3.55
ERBoost-0.5	0.89	1.00	1.00	0.88	0.11	0.01
ERBoost-0.9	0.95	1.00	1.00	1.00	0.99	2.26
Lasso	1.00	1.00	1.00	1.00	0.34	124.51
ALasso	1.00	1.00	1.00	1.00	0.49	126.4
Scad	1.00	1.00	1.00	1.00	0.46	126.18

Table 4.1: Comparison among methods for variable selections

wrong variables. ERboost is the best in excluding the variables. It does a good job in detecting V5 but not as good as GB. It does poorly in finding V1 - V5 compared to other methods which all have 100% correctness. Finally, for linear model based methods: Lasso, Adaptive Lasso and Scad, they seems to work terribly on ruling out irrelevant variables. But what is surprising is that it has some power in find V5 even though the conditional mean should not be influenced by the heterogeneity variables.

Test for variable importance In this synthetic model, we see the extreme effectiveness of GB in terms of detecting important variables by strictly shrinking the irrelevant variables to 0. Nevertheless in reality it can be rare. In the real data example we will see later, we often face the situation that we have big magnitude for very important variables but relative small values for some variables which we do not know if we should include them. So it is pressing to work out a guideline to determine

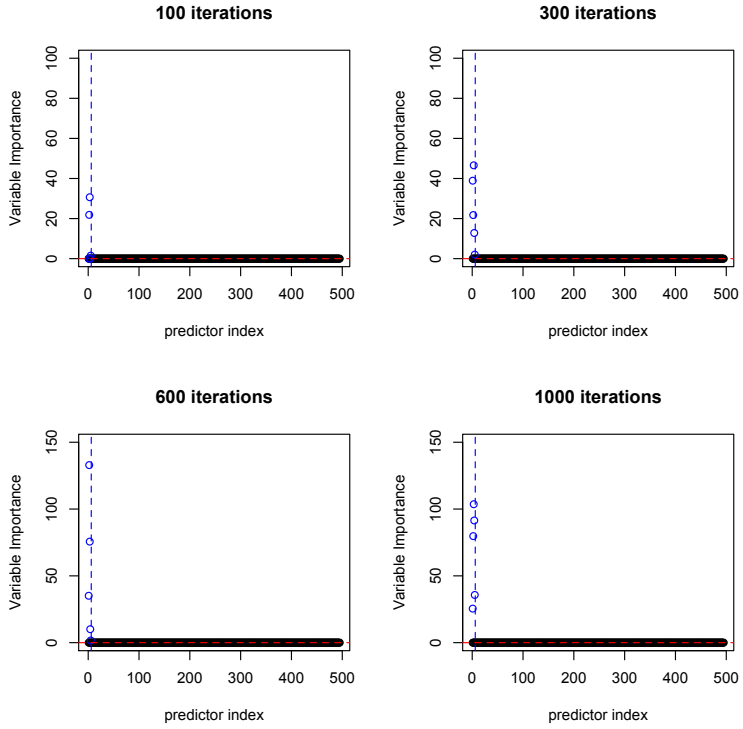


Figure 4.1: 0.1th Quantile

Quantiles	Iteration round	V1	V2	V3	V4	V5
0.1	100	42.95	587.41	319.88	65.10	0.00
	300	7.49	770.60	1079.08	761.21	17.43
	600	93.57	514.26	1382.72	694.42	644.11
	1000	783.25	1241.83	1907.44	1647.42	516.77
0.5	100	0.00	1958.18	1801.16	0.00	0.00
	300	23.70	7070.31	3057.38	0.00	0.00
	600	378.25	8125.11	8158.59	986.23	0.00
	1000	1525.91	11209.50	13831.15	1113.09	0.00
0.9	100	0.00	437.03	612.87	0.00	30.39
	300	778.05	434.97	931.54	256.14	40.66
	600	702.00	2657.87	1512.52	201.38	33.41
	1000	509.83	1594.98	2072.29	1828.81	714.06

Table 4.2: Variable Importance from GB for model I

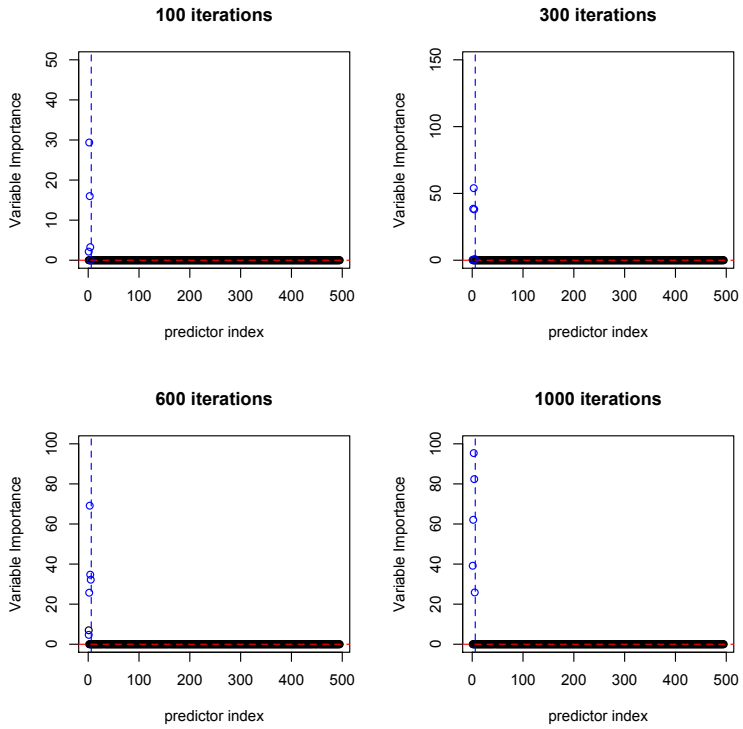


Figure 4.2: 0.9th Quanile

Iteration round	V1	V2	V3	V4	V5
100	0.00	10830.09	11775.58	0.00	988.40
300	464.99	26449.87	24287.18	194.80	5480.91
600	1097.84	35907.17	41923.44	2105.46	14934.00
1000	2902.11	43371.40	50681.99	5729.49	25942.70

Table 4.3: Variable Importance from ERBoosting at 0.1th expectile for model I

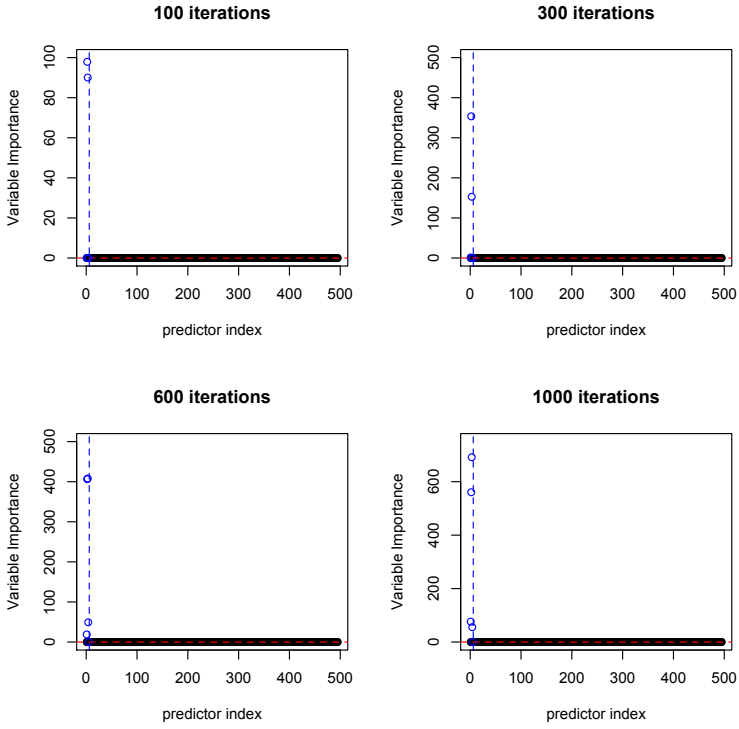


Figure 4.3: Median

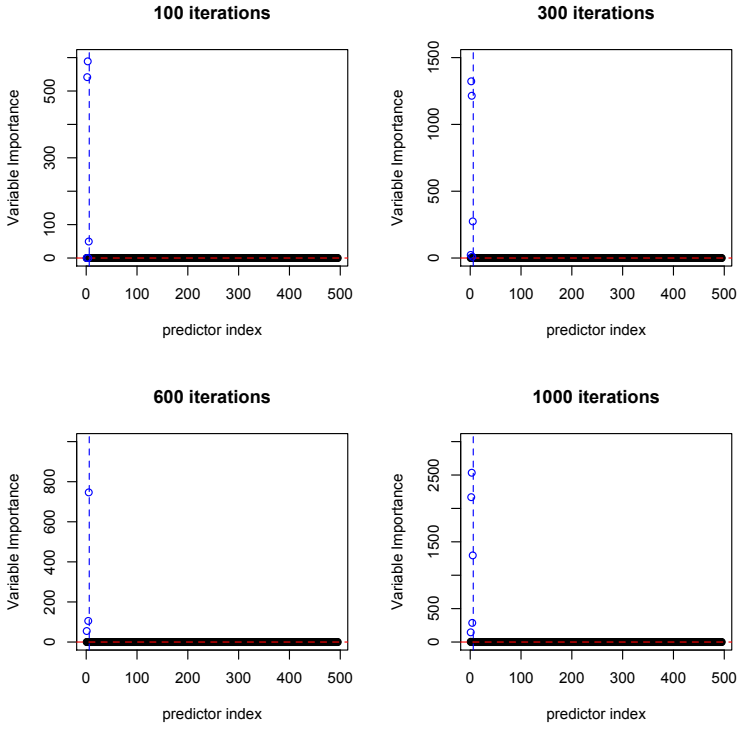


Figure 4.4: ERBooting variable selection at 0.1th expectile

the threshold under which those variables should not be selected. Since the GB is a tree-based model so it is not easy to carry out the statistical tests for the coefficients. Two ad-hoc methods can be used under this scenario. The first is bootstrap test, and the other is permeation test. The bootstrap test goes like this:

The intuition is based on permeation test. Imagine if you have a very important variable V1. Once you shuffle the sample points of V1 and leave other sample points unchanged. Then it is expected that the fitting error will be influenced to a large degree. However, if V1 is irrelevant to your target, what will happen if you shuffle the rows of V1 is that the fitting will basically remain the same. So this brings a natural question: what is the underlying hypothesis test ? And how do you carry out the test ? To carry out the test, we must be clear about the null hypothesis at least. There are two types of variable selection problems in statistics: one is selecting the subset of variables that optimize the prediction; the other is selecting the subset that gives best interpretation. For both task the null hypothesis are both the variables are not in the model which is reflected by the variable importance of value 0. However, since they have different aims so the test statistics for them are different. The test statistics for the first type variable selection is prediction error distribution. For the second type, the test statistics is just information gain. In our thesis, we only focus on the prediction error oriented variable selection. Both the bootstrap test and permutation can be carried out for this task. Here we will provide the hydride test of bootstrap and permeation to test the variable importance.

The flow chart gives us a outline as for how the hybrid test is conducted. Suppose we have a raw data then we run the gradient boosting to rank the variable importance. Once we have variables ranked, we can just ignore those with value 0. For the variables with nonzero values, we still need to determine whether they can should stay. To do this, we need first to bootstrap the sample b times so that we can use them to estimate the null distribution for fitting error, say, f_{null}. For a specific variable, we permute the

sample realization of this variable b times. For each time, we get a fitting error. So we have a distribution for fitting error for the permuted variable. The final decision is made according to the result of test of comparing these two distributions. If there is a significant different between these two distributions, then we take that variable as an active variable. Otherwise, we will consider it as irrelevant.

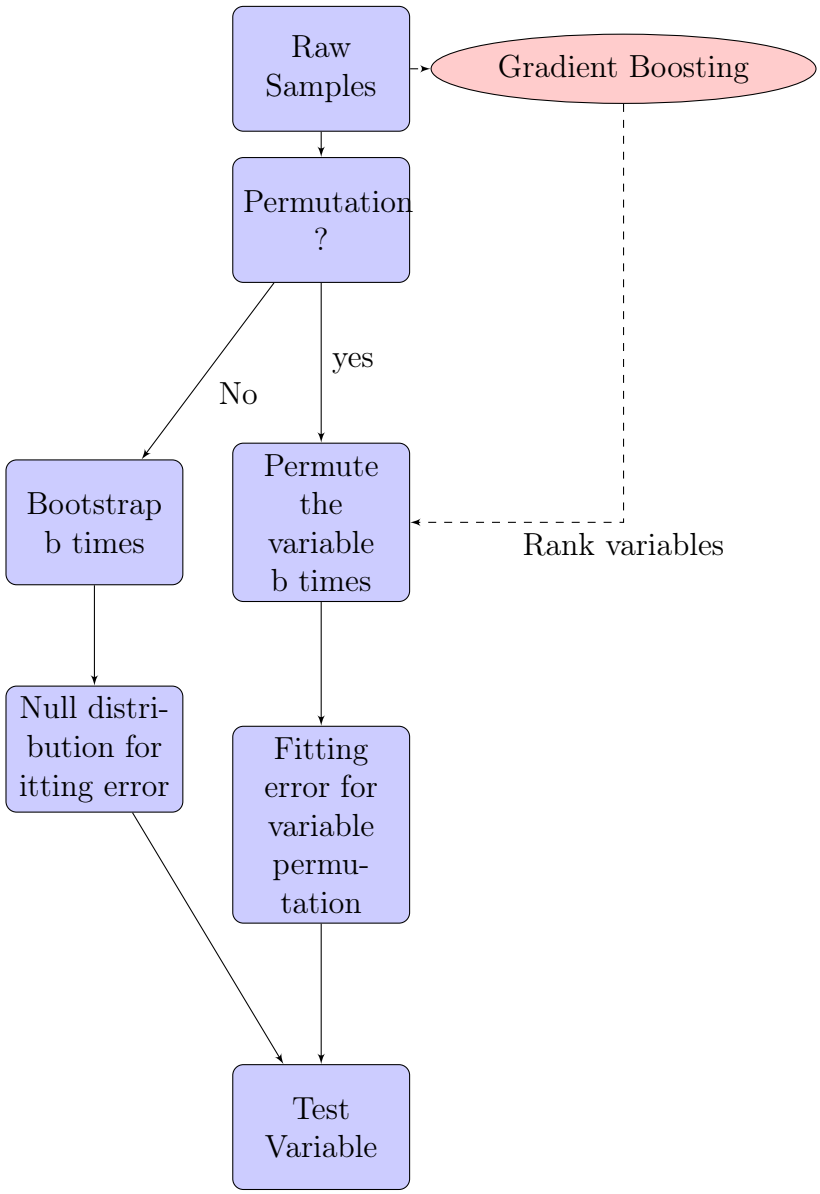


Figure 4.5: flow charts for permeation test

4.3 Birth Weight Data Analysis

The birth weight data is from R MASS package birthwt. The birthwt data frame has 189 rows and 10 columns. The data were collected at Baystate Medical Center, Springfield, Mass during 1986. One may be concerned with how many categories for each variable has if one is concerned with panel data analysis, say, the comparison between boys and girls. So we will also provide this information about the variable description in table 4.4. Also, the univariate variable summary can be accessible from table 4.5. This table tells us quantitative information for each single variable. Most analysis of this dataset is on top of conditional mean based models. However, the results from analyzing the mean weights may not reflect the relationship between low weight infants and associated factors. Hence it makes more sense if one can conduct analysis based on quantile regression which capture a comprehensive distributional traits than just mean alone.

Koencker(2001) did a thorough analysis of a very similar data by using quantile regression. Particularly, He did panel comparison for instance, how different are the corresponding weights of boys and girls, given a specification of the other conditioning variables. The disparity between birthweights of infants born to black and white mothers is also discussed in that paper. In this section, we focus on the application of permeation test for variable importance of gradient boosting algorithm for variable selection. To analyze this data, we shall first opt out the first variable which is low because this is useless for prediction. Including this variable will to a large extent deprive the importance of all other variables. The relative importance after running gbm with quantile 0.1 trees 1000 shrinkage 0.001 are shown in the figure. At a first glance, we may notice age, lwt and ui are most important factors. The rest factors has importance of certain magnitude. Now we will run permutation test for each

Variable	Description
low (factor)	indicator of birth weight less than 2.5 kg
age	Mother's age
lwt	mother's weight in pounds at last menstrual period
race	mother's race (1 = white, 2 = black, 3 = other)
smoke	smoking status during pregnancy
ptl	number of previous premature labours
ht	history of hypertension
ui	presence of uterine irritability
ftv	number of physician visits during the first trimester
bwt	birth weight in grams

Table 4.4: Variable description for birthwt data

variable to see if it should be kept or not.

The fitting error distributions are summarized in figure 4.7. The background distribution with light gray color is null distribution. The blue color distribution is fitting error distribution after permuting corresponding variable. According to KS test, we have p value greater than 0.05 only for ht (hypertension). But we also observe distributional similarities in smoke, ptl, ftv. Not surprisingly, those variables all have very small relative importance values. Notice that the null distribution has bell shape. So we conduct shapiro's test of normality for Null, smoke, ptl, ftv. It turns out Null, ptl and ftv are of insufficient evidence of non-normal. So we will still keep smoke according to KS test. Two sample normality test will further be applied to ptl and ftv. Again, they are proved to be significantly from null distribution. So the only variable should not be a factor is ht according to our analysis.

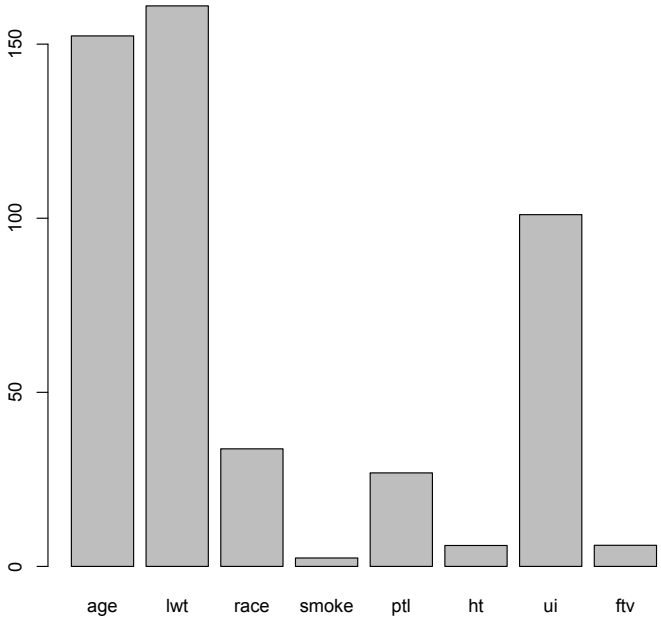


Figure 4.6: Relative Importance

birthwt
10 Variables 189 Observations

low

n	missing	unique	Sum	Mean
189	0	2	59	0.3122

age

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
189	0	24	23.24	16	17	19	23	26	31	32

```
lowest : 14 15 16 17 18, highest: 33 34 35 36 45
```

lwt

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
189	0	75	129.8	94.4	99.6	110.0	121.0	140.0	170.0	188.2

```
lowest :  80  85  89  90  91, highest: 215 229 235 241 250
```

race

n	missing	unique	Mean
189	0	3	1.847

```
1 (96, 51%), 2 (26, 14%), 3 (67, 35%)
```

smoke

n	missing	unique	Sum	Mean
189	0	2	74	0.3915

ptl

n	missing	unique	Mean
189	0	4	0.1958

```
0 (159, 84%), 1 (24, 13%), 2 (5, 3%), 3 (1, 1%)
```

ht

n	missing	unique	Sum	Mean
189	0	2	12	0.06349

ui

n	missing	unique	Sum	Mean
189	0	2	28	0.1481

ftv

n	missing	unique	Mean
189	0	6	0.7937

```
             0  1  2 3 4 6
Frequency  100 47 30 7 4 1
%           53 25 16 4 2 1
```

bwt

n	missing	unique	Mean	.05	.10	.25	.50	.75	.90	.95
189	0	131	2945	1801	2038	2414	2977	3487	3865	3997

```
lowest :  709 1021 1135 1330 1474, highest: 4167 4174 4238 4593 4990
```

Table 4.5: Birth Weight Data

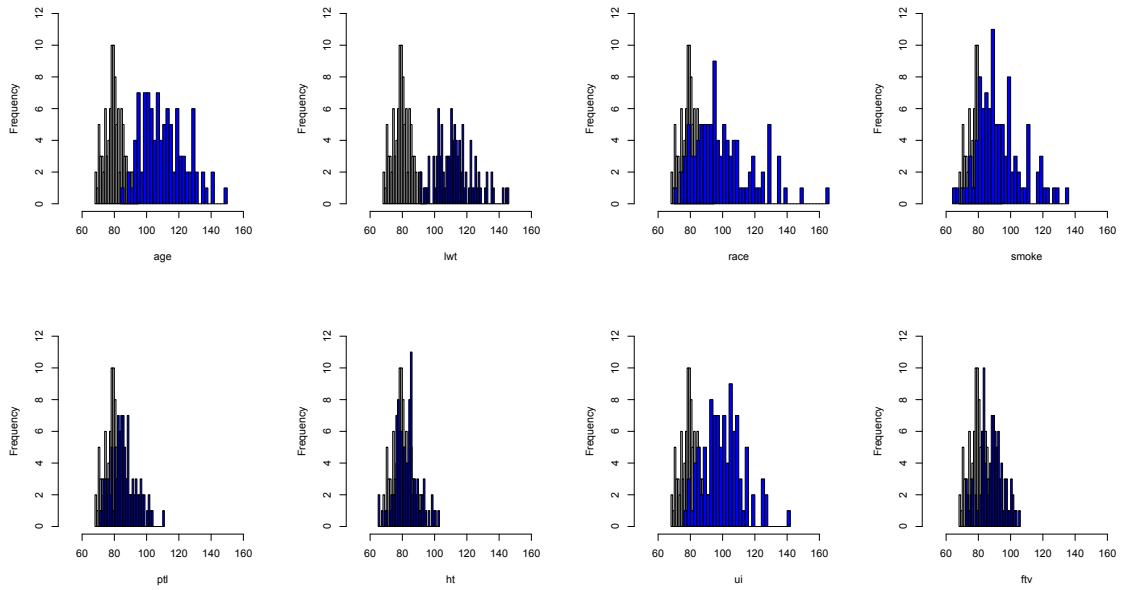


Figure 4.7: Fitting Error Distribution

4.4 Childhood Malnutrition in India

The detailed description of this data is in chapter 2. Here we use this dataset to evaluate the effectiveness of variable selection for different methods with focus on Gradient Boosting. The target is chidden's height. Notice in this data set, we have two very influential variables: children's age in years, children's age in months. Common sense can even tell children at early ages must have their height very closely tied to their ages. So these two variables should be chosen by any effective variable selection methods and should have very large weights. Tables 4.6 and Tables 4.7 respectively list up the variable importance of learning rate at 0.001 and 0.1. So if the shrinkage is large(0.001), only ages of child are selected as important variables. If the shrinkage is mild, almost all variables are included but still with largest weights attached to children's ages. In comparison, we have also conducted variable selections from Lasso, Scad, Adaptive-Lasso and Q-scad. Since Lasso, Adaptive-Lasso and Scad performs

very similarly. So we only shows the result of Lasso in table 4.8. Also the result of Q-scad is shown in table 4.9. The running parameter is chose by cross validation. (For lasso, 16, for Q-lasso, 30).

```
> m<-gbm(height~.,distribution =list(name = "quantile", alpha =0.1),
data = as.data.frame(mal),n.trees = 1000, shrinkage = 0.001)
```

The tables besides GB present results which are very much counter intuitive. They all conclude that electricity is the most factor that can influence the children's malnutrition (measured by height). And the Children's ages (in months or in years) are not even among the top three important factors in all tables. Moreover, we have observed a very implausible fact that the coefficients of child's ages in months and child's ages in years has opposite signs which is hard to explain. Possible reasons for this terrible performance for the traditional variable selections methods may stem from the following facts:

1) Many variables have collinearity, for instance, children's age in years and children's age in months. The methods besides gradients boosting we have here are not well known for handling the collinearity effect. However, gradient boosting even with coordinate-wise linear base learner have good performance as for tackling collinearity, see Hastie.

2) The real data may have complex structures can be hardly captured by linear models. No matter assuming the conditional mean or conditional quantile is linear

Variables	
electricity	0.0000000000
radio	0.0000000000
tv	0.0000000000
fridge	0.0000000000
bike	0.0000000000
motorcycle	0.0000000000
car	0.0000000000
religion	0.0000000000
telephone	0.0000000000
wealth	0.0000000000
no..of.living.children	0.0000000000
mother.s.age	0.0000000000
gender	0.0000000000
age.in.years	10379.0492918219
lives.with.whom	0.0000000000
breastfeeding.in.months	0.0000000000
size.of.child.at.birth	0.0000000000
mother.s.bmi	0.0000000000
child.s.age.in.months	76902.7970451835
mother.s.ed	0.0000000000
father.s.ed	0.0000000000

Table 4.6: Variable importance including children's Ages: GB,learning rate = 0.001

Variables	
electricity	7.69699665868986
radio	2.25877066898787
tv	1.89317967105749
fridge	2.56532333786323
bike	6.24400029507260
motorcycle	13.20581234343342
car	2.65837474383027
religion	22.14799290357281
telephone	4.50061983143209
wealth	34.16112278828038
no..of.living.children	36.34237475365907
mother.s.age	55.88550123702347
gender	11.82114839534978
age.in.years	215.62370581295548
lives.with.whom	0.00000000000000
breastfeeding.in.months	50.00259409553846
size.of.child.at.birth	19.34924034009987
mother.s.bmi	129.96547932762888
child.s.age.in.months	1270.02508938057690
mother.s.ed	66.71711576934936
father.s.ed	27.22893989448357

Table 4.7: Variable importance without children's ages: GB,learning rate = 0.1

Variables	
electricity	35.69776171
radio	−4.25290136
tv	−14.34947380
fridge	−28.35588087
bike	19.41620269
motorcycle	−1.93143517
car	−11.90755051
religion	5.14555084
telephone	−9.96948851
wealth	17.84615608
no..of.living.children	−1.07004168
mother.s.age	6.45780252
gender	29.82549716
age.in.years	−18.40306420
lives.with.whom	1.76304310
breastfeeding.in.months	1.07978322
size.of.child.at.birth	20.16878253
mother.s.bmi	0.08509025
child.s.age.in.months	8.69752770
mother.s.ed	3.04842073
father.s.ed	1.36915684

Table 4.8: Variable importance without children's ages: Lasso, $\lambda = 16$

Variables	
electricity	103.2040
radio	−3.937371
tv	−12.88814
fridge	−19.13837
bike	17.68419
motorcycle	0
car	−17.28178
religion	2.853051
telephone	−2.355277
wealth	14.42733
no..of.living.children	−0.7314436
mother.s.age	4.657492
gender	19.17150
age.in.years	−8.021516
lives.with.whom	0
breastfeeding.in.months	0.9625227
size.of.child.at.birth	11.94794
mother.s.bmi	0.05893068
child.s.age.in.months	7.712627
mother.s.ed	3.549903
father.s.ed	1.951347

Table 4.9: Variable importance without children's ages: Q-scad 0.1 quantile, $\lambda = 30$

combinations of variables. The underlying assumptions are both on top of linearity assumption. But in real data, it will be in question. When using gradient boosting, we use trees which are essentially step functions as approximation functions to complex structures.

4.5 ERLars

Since we have seen the power of variable selection based on reboots. We have reasons to believe expectile baed penalty method can have competitive performance. Li and Zhu (2007) have developed a regularization path algorithm based on solving KKT equations. However this algorithm cost time of complexity order $O(n^3p)$ when $n > p$. This is not as fast as conventional LARS algorithm for OLS which only requires $O(n^2p)$. What is more crucial is that Lasso though has its strong power to introduce sparsity, but may not work well when strong colinearity presents between predictors, see Zou (2005), Hastie (2006). There is alternative approach to consider the comprehensive picture of the conditional distribution which is called Asymmetric Least Square (ALS) or Expectile Regression (ER), see Efron (1992). We will show in this section LARS algorithms for ER can be carried out with worst time of order $O(n^2p)$ if $n > p$. This we will call ERLars/ Plus with small modification for LARS algorithm we can obtain both Lasso regularization path and foward stagewise regression. The later one outperforms lasso if colinearity exists. The LARS algorithms can also be developed for a fused lasso penalty for ER. Efron (1992) studied a global parallel version of quantile regression called expectile regression (ER). Compared to quantile q_τ which specifies the position below which $100\tau\%$ of the probability mass of X lies, expectile e_τ determines the position $100\tau\%$ of the mean distance between it and X comes from the mass below it. Therefore expectiles rely on the distance of observations at the price of increasing the outlier sensitivity. For this reason, it

has been claimed that expectiles use the data more efficiently than quantile (Newey and Powell, 1987). Further, expectiles are widely discussed in financial world. For example, the expected shortfall (ES) is a quantity to measure portfolio risk. The ES at the $100\tau\%$ level is the expected return on the portfolio in the worst $100\tau\%$ of the cases. It is related to expectiles, see Yee (2012).

Quantiles:

$$\rho_\tau(u) = \tau u_+ + (1-\tau)u_- \tag{4.1}$$

$$q_\tau = \arg\min_\theta \mathbf{E}\rho_\tau(X-\theta) \tag{4.2}$$

Expectiles:

$$\phi_\tau(u) = \tau {u_+}^2 + (1-\tau){u_-}^2 \tag{4.3}$$

$$e_\tau = \arg\min_\theta \mathbf{E}\phi_\tau(X-\theta) \tag{4.4}$$

Computationally, expectiles enjoy a superior benefits when compared to quantiles with the same goal to recover the comprehensive form of the conditional distribution. Because the expectile check function (1.3) is differentiable, one can derive a simple iteratively weighted least squares for estimating expectile even if in a nonlinear regression scenario. Our paper aims at deriving a LARS-like algorithm which is n times

faster than quantile regression $L-1$ regularization path. Despite Lasso can generate sparsity on covariate coefficients. However it is not a good choice if $p >> n$ or there are covariates heavily correlated. Alternatively, one can consider a method called Foward Stagewise Regression to recover the sparsity. This method has been proved to discourage frequent change in solution path which may be caused by lasso. Moreover, due its "boosting" nature, it is relatively hard to overfit empirically. Therefore LARs algorithm which can solve both fused lasso and foward stagewise regression solution paths by making a small modification to each.

For a general predictor matrix X, we denote $X_{(i)}^T$ as ith row and X_j as jth column. Further we assume this matrix is standardized so that $||X_{(i)}|| = 1$. We define left, right and equal regions as:

$$L = \{i : y_i - f_i(x) < 0\}, R = \{i : y_i - f_i(x) > 0\}, E = \{i : y_i = f_i(x)\}$$

For Expectile regression (ER), the sample version objective function is:

$$\phi_n(f) = \sum_L \tau(y_i - f_i(x))^2 + \sum_R (1-\tau)(y_i - f_i(x))^2 \tag{4.5}$$

where $f_i(x) = R_{(i)}^T\theta + \mu$. The lasso problem for ER is:

$$\hat{\theta}_{lasso} = \arg\min_\theta \sum_L \tau(y_i - \mu - X_{(i)}^T\theta)^2 + \sum_R (1-\tau)(y_i - \mu - X_{(i)}^T\theta)^2 + \lambda|\theta| \tag{4.6}$$

One can also study the fused lasso problem for ER:

$$\hat{f}_i = \arg\min_{f_i} \sum_L \tau(y_i - \mu - f_i)^2 + \sum_R (1-\tau)(y_i - \mu - f_i)^2 + \sum_{i=1}^n |f_i - f_{i-1}| \tag{4.7}$$

$f_0 = 0$, the fused lasso for QR has been proposed to formulate the detection of DNA copy number changes by Eilers and Menezes (2004). The f_i in the above equation is the smooth series that approximate y_i. Our guess is that fused lasso for ER can also parallelly achieve the same goal. And one can easily find that the fused lasso for ER is essentially the same as lasso if we do the following transformation: $\theta_i = f_i - f_{i-1}$, if using matrix notation: $\theta = Lf, f = R\theta, R = L^{-1}$, where L is the appropriate matrix.

Before we introduce the algorithm, we shall make notations clear in the first place. Let $r(t)$ be the residual vector at time t. $L(t), R(t), E(t)$ are left, right and equal regions respectively. $t_0 = 0$. Let A be the active set which is the set of variables chosen up to time t. And we will see later that the variables in this set have the same correlation with current residual.We will define two types of events as following:

event I: One of $r_i(t)$ hits 0.
event II: There is a $l \in A^c$ such that $X_l^T r(t) = \{X_j^T r(t) : j \in A\}$

If event I happens, it means that region L, R, E will change accordingly. if event II happens it means a new variable will join the active set A.

ERLars Algorithm

Step1(Initialization)$\theta = 0, \mu = \tau -$ expectile
$L(t_0) = \{i : y_i - \mu < 0\}, R(t_0) = \{i : y_i - \mu > 0\}, E(t_0) = \{i : y_i = \mu\}, A = \emptyset$
$y_i = \sqrt{\tau}(y_i - \mu)I(i \in L) + \sqrt{1-\tau}(y_i - \mu)I(i \in R)$
$X_{(i)} = \sqrt{\tau}X_{(i)}I(i \in L) + \sqrt{1-\tau}X_{(i)}I(i \in R)$

$r(t_0) = y = (y_1, ..., y_n)^T; t_0 = 0$

Step2(Moving in ERLars direction) For k in 1: m

Between $(k-1)$th and kth event, $t_{k-1} \leq t < t_k$,

If event is type I: Update $A = A \cup \{l\}$ then go to Moving Step.

If event is type II: if $i \in L(t_{k-1})$, $L(t) = L(t)/\{i\}$, $R(t) = R(t) \cup \{i\}$ vice versa.

Renew corresponding X_i and y_i by times $(\frac{\sqrt{\tau}}{\sqrt{1-\tau}})^{I(i\in L)-I(i\in R)}$, then standardize X such that $||X_{(i)}|| = 1$, then go to Moving Step.

Moving Step $\theta(t) = \theta(t_{k-1}) + td_A; d_A = (X_A^T X_A)^{-1} X_A^T r(t_{k-1})$

$r_i(t) = y_i - X_{(i)}^T \theta(t)$ for $i = 1, .., n$. Keep moving until the next event occurs.

Computational Details As we can see from above algorithm, the time for updating moving step is of order nk^2 assuming there are $k < n$ variables already in set A. However if we have a orthonormalizing procedure along, the time can be reduced to nk in direction updating. Suppose after initialization, the time when event II happened we will expand $X_A = (X_1, ..., X_l)$(WTLG, I use variable indices: 1,...,l). Additionally we keep another matrix $O_A = (O_1, ..., O_l)$, which is Gram-schmitt-orthonormalizing version of X_A. Moreover, we also record the coordinates matrix U_A which is upper triangular. $X_A = O_A U_A$ Here the direction we go is:

$$d_A = (U^T U)^{-1} U^T O^T r \tag{4.8}$$

The complexity of computing this direction is just nk. This is due to the fact that calculating $U^T O^T r$ takes time of order nk and solving this linear system $U^T(Ux) = U^T O^T r$ is of order k^2 if U is triangular. The only thing we sacrifice is space because

we need to keep track of O and U. Despite this little space expansion, we can achieve a efficient time reduced by a factor p. So instead of updating X_A, we need to store and update O_A and U_A. Now we also need to take account of the time for updating O_A and U_A.$O'_{l+1} = X_{l+1} - \sum_{i=1}^{l}(X_{l+1}^T O_i)O_i$.

For event I: A new variable comes, we will expand $O_A = (O_1, ...O_l, O_{l+1})$, where $O_{l+1} = O'_{l+1}/||O'_{l+1}||_2$. $U_{l+1} = (X_{l+1}^T O_1, ..., X_{l+1}^T O_l, ||O'_{l+1}||_2, 0, ..., 0)$.

For event II: Since the row $X_{(i)}$ for those i entered E is updated by times a factor. It is equivalent to write it as $X_{new} = \Gamma X$, where Γ is diagonal. To order to keep the column space O same as X (that is the key to keep LARS property !), $X_{new} = \Gamma X = \Gamma OU = \Gamma OSS^{-1}U$, $O_{new} = \Gamma OS$, $U_{new} = S^{-1}U$ where $S = diag(||O'_1||_2, ...||O'_l||_2)$. Therefore O_{new} is orthonormal and span$\{O_{new}\}$ = span $\{X\}$.

Again either event I or event II only costs nk. So the total cost is of order np^2.

Lars Properties Now we will check ERLars has Lars-property : the correlation between active variables and current residual are the same and decreases.Using induction:

1) From initialization before the first event is exactly the same as Lars.

2) Assuming the correlations the same and keep decreasing for variables in set A up to time t_{k-1}.

3) If $(k-1)$th event is type I, then L, R do not change. It is exactly the same as Lars. If $(k-1)$th event is type II, $r(t_{k-1}) = (X_{(i)}^T \theta(t_{k-1}) - y_{(i)})_{i=1,..n}^T$. Since we

only change those is for which $X_{(i)}^T\theta(t_{k-1}) - y_{(i)}$ are zero. Hence $r(t_{k-1})$ is in fact the same. Although X has been changed, but again suppose row i is changed, however $r(t_{k-1})_i$ is 0!. So sample covariance will not change. Remember we have standardized X every time after timing the factor, therefore the correlation is the same. And the correlation decreases between t_{k-1} and t_k. By continuity, it decreases between t_0 and t_k.

References

[1] A. Altmann, L. Tolosi, O. Sander, and T. Lengauer. Permutation importance: a corrected feature importanc measure. *Bioinformatics*, 26(10)(1340-7), 2010.

[2] R.R. Bahadur. A note on quantiles in large samples. *The Annals of Mathematical Statistics*, 37:577–580, 1966.

[3] A. Belloni and V. Chernozhukov. l_1 penalized quantile regression in high-dimenisonal sparse models. *The Annals of Statistics*, 39:82–130, 2011.

[4] A. Bose and Chatterjee S. Generalised bootstrap in non-regular m-estimation problems. *Statistics & Probability Letters*, 55:319–328, 2001.

[5] L. Breiman. Prediction games & arcing algorithms. *Neural Computation*, 11:1493 – 1517, 1999.

[6] P. Buhlmann. Boosting methods: Why they can be useful for high-dimensional data. *Proceedings of the 3rd International Workshop of Distributed Statistical Computing*, 2003.

[7] P. Buhlmann. Boosting for high-dimensional linear models. *The Annals of Statistics*, 34(2):559–583, 2006.

[8] P. Buhlmann. Twin boosting: Improved feature selection and prediction. *Technical Report, ETH, Zurich*, 2007.

[9] P. Buhlmann and T. Hothorn. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, pages 477–505, 2007.

[10] P. Buhlmann and B. Yu. Boosting with l2 loss: Regression and classification. *Journal of the American Statistical Association*, 98:324–339, 2003.

[11] B.S. Cade and B.R. Noon. A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 8:412–420, 2003.

[12] B.S. Cade, J.W. Terrell, and R.L. Schroeder. Estimating effects of limiting factors with regression quantiles. *Ecology*, 80(311-23), 1999.

[13] P. Chaudhuri. Nonparametric estimates of regression quantiles and their local bahadur representation. *The Annals of Statistics*, 19:760–777, 1991.

[14] P. Chauduri. On a geometric notion of quantiles for multivariate data. *Journal of American Statistical Association*, 91:862–872, 1996.

[15] C. Chen. A finite smoothing algorithm for quantile regression. *Journal of Computational and Graphical Statistics*, 16:136–164, 2007.

[16] V. Chernozhukov. Extremal quantile regression. *The Annals of Statistics*, 33:806–839, 2005.

[17] J.R. Collins. Robust estimation in the presence of symmetry. *The Annals of Statistics*, 4:68–85, 1976.

[18] R. Durret. *Probability: Theory and Examples.* Cambridge Univ. Press, 1991.

[19] B. Efron. Regression percentiles using asymmetric squared error loss. *Statistica Sinica*, 1:93–125, 1991.

[20] T. Fenke, N. Kneib and T. Hothorn. Identifying risk factors for severe childhood malnutrition by boosting additive quantitle regression. *Technical Report*, 2009.

[21] D.A. Freedman and P. Diaconis. On inconsistent m-estimators. *The Annals of Statistics*, 10:454–461, 1982.

[22] Y. Freud and R. Schapire. A decision-theoretic generalization of online learning and an application to boosting. *Computational Learning Theory*, 904:23–27, 1995.

[23] J. Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2001.

[24] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Annals of Statistics*, 28:337–655, 2000.

[25] R. Genuer, J.M. Poggi, and C. Tuleau-Malot. Variable selection using random forests. *Pattern Recognition Letters*, 14:2225–2236, 2010.

[26] M. Gilli and E. Kellezi. An application of extreme value theory for measuring financial risk. *Computational Ecomomics*, 27:1–23, 2006.

[27] I. Guyon, J. Weston, S. Barnhill, and V.N. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[28] S.J. Haberman. Concavity and estimation. *The Annals of Statistics*, 17:1631–1661, 1989.

[29] T. Hastie, J. Taylor, R. Tibshirani, and G. Walther. Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1:1–29, 2007.

[30] X. He, P. Ng, and S. Portnoy. Bivariate quantile smoothing splines. *J.R. Statist. Soc. B*, 3:537–550, 1998.

[31] X. He and Q. Shao. A general bahadur representation of m-estimators and its application to linear regression with nonstochastic designs. *The Annals of Statistics*, 24:2608–2630, 1996.

[32] P. J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35:73–101, 1964.

[33] P. J. Huber. Robust regression: Asymptotics, conjectures and monte carlo. *The Annals of Statistics*, 1:799–821, 1973.

[34] D. R. Hunter and K. Lange. Quantile regression via an mm algorithm. *Journal of Computational and Graphical Statistics*, 9:66–77, 2000.

[35] M.A. Huston. Introductory essay: critical issues for improving predictions. *Introductory essay: critical issues for improving predictions*, pages 7–21, 2002.

[36] R.W. Katz. Extreme value theory for precipitation. *Advanced in Water Resources*, 23:133–139, 1999.

[37] K. Knight. Epi-convergence in distribution and stochastic equisemicontinuity. *Technical Report, Dept. of Stat., Univ. of Toronto*, 1999.

[38] R. Koenker. Quantile regression. *New York: Cambridge University Press*, 2005.

[39] R. Koenker. Additive models for quantile regression: Model selection and confidence bandaids. *Brazilian Journal of Probability and Statistics*, 25:237–470, 2011.

[40] R. Koenker and G.S. Bassett. Regression quantiles. *Econometrica*, 46:33–50, 1978.

[41] R. Koenker and O.Geling. Reappraising medfly longevity: A quantile regression survial analysis. *Journal of American Statistical Association*, 96:458–468, 1996.

[42] Y. Li and J. Zhu. Analysis of array cgh data for cancer studies using fused quantile regression. *Bioinformatics*, 23:2470–2476, 2007.

[43] N. Meinshausen. Quantile regression forests. *The Journal of Machine Learning Research*, 7:983–999, 2006.

[44] N.D. Mukhopadhyay and S. Chatterjee. High dimensional data analysis using multivariate generalized spatial quantiles. *Journal of Multivariate Analysis*, 102:768–780, 2011.

[45] W. Newey and J. Powell. Asymmetric least squares estimation and testing. *Econometrica*, 55:819–847, 1987.

[46] W. Niemiro. Asymptotics for m-estimators defined by convex minimization. *The Annals of Statistics*, 20:1514–1533, 1992.

[47] C. Poggi, J.M. andTuleau. Classification supervisee en grande dimension. *Application a l'agrement de conduite automobile. Revue de Statistique Appliqu ee*, LIV(4)(39-58), 2006.

[48] S. Portnoy. Robust estimation in dependent situations. *The Annals of Statistics*, 5:22–43, 1977.

[49] S. Portnoy and J. Jureckova. On extreme regression quantiles. *Extremes*, 2:227–243, 1999.

[50] A Rakotomamonjy. Variable selection using svm-based criteria. *Journal of Machine Learning Research*, 3:1357–1370, 2003.

[51] R.T. Rockafellar. *Convex Analysis*. Princeton Univ. Press, 1970.

[52] R.L. Smith. Nonregular regression. *Biometrika*, 81:173–183, 1994.

[53] J.W. Taylor. Estimating value at risk and expected shortfall using expectiles. *Journal of Financial Econometrics*, 6:231–252, 2008.

[54] J.W. Terrell, B.S. Cade, J. Carpenter, and J.M. Thompson. Modeling stream fish habitat limitations from wedged-shaped patterns of variation in standing stock. *Trans Am Fish Soc*, 125(104-17), 1996.

[55] L. Wang. Quantile regression for analyzing heterogeneity in quantile regression for analyzing heterogeneity in ultra-high dimension. *Journal of American Statistical Association*, 107:214–222, 2012.

[56] Y. Yang and H. Zou. Nonparametric multiple expectile regression via er-boost. *Journal of Statistical Computation and Simulation*, 2014.

[57] T. Zhang and B. Yu. Boosting with early stopping: Convergence and consistency. *The Annals of Statistics*, 33:1538–1579, 2005.

[58] H. Zou. The adaptive lasso and its oracle properties. *Journal of American Statistical Association*, 101:1418–1429, 2006.

[59] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *J.R. Statist. Soc. B*, 67:301–320, 2005.