ROBUST A-OPTIMAL SUBSAMPLING FOR MASSIVE DATA ROBUST LINEAR REGRESSION

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Ziting Tang

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Fei Tan, Co-Chair

Department of Mathematical Sciences, IUPUI

Dr. Hanxiang Peng, Co-Chair

Department of Mathematical Sciences, IUPUI

Dr. Jyoti Sarkar

Department of Mathematical Sciences, IUPUI

Dr. Honglang Wang

Department of Mathematical Sciences, IUPUI

Dr. Guang Lin

Department of Mathematics, Purdue University

Approved by:

Dr. Evgeny Mukhin

Head of the Graduate Program, IUPUI

Dr. Gregery Buzzard

Head of the Graduate Program, Purdue University

This is dedicated to my parents and my husband Kang Lu.

ACKNOWLEDGMENTS

I would like to express my deepest appreciation to my advisors Prof. Hanxiang Peng and Prof. Fei Tan. The completion of my dissertation would not have been possible without their guidance, support and persistent help. More specifically, Prof. Hanxiang Peng taught me how to do rigorous research in the big data analysis. and how to find a solution to the problems in robust linear regression. I very much appreciate his invaluable instruction and help. I would like to thank Prof. Fei Tan for providing me a great opportunity to work on applied projects in addition to her advisement on dissertation. Prof. Fei Tan taught me how to write a better report and communicate with people who are not statisticians. This is a great help not only for my future work but also for my life.

I would like to extend my deepest gratitude to Prof. Jyoti Sarkar. I had great pleasure working with him and learned a lot in reliability. I sincerely appreciate his guidance and help. Besides, I would like to thank Prof. Honglang Wang for giving me invaluable insight into the advanced statistical inference. I would also like to thank Prof. Guang Lin for his effort and time in serving in my dissertation committee.

I would also like to extend my sincere thanks to Prof. Benzion Boukai, who encouraged me to speak up in class and recommended me to teach courses. His help has lasting effect on my life. I am also extremely grateful to Prof. Zhongmin Shen who encouraged me to pursue a PhD here. I very much appreciate his valuable advice on my career. I would also like to extend my gratitude to Prof. Evgeny Mukhin for his great suggestions and assistance during my PhD study. In addition, many thanks to the staff in the department for their assistance.

Thank you to all my friends for all the love and support. And special thanks to my parents and my husband Kang Lu for their endless patience and encouragement.

TABLE OF CONTENTS

Page

LIST OF TABLES vi

LIST OF FIGURES xi

ABSTRACT xiv

1 INTRODUCTION 1

1.1 Linear Regression and Classical Estimation 4

1.2 Robust Linear Regression and M-estimation 4

1.2.1 Two criteria for robustness 6

1.2.2 Examples of loss function 8

2 A-OPTIMAL SUBSAMPLING METHOD 12

3 ASYMPTOTIC BEHAVIOR OF M-ESTIMATORS WHEN ψ IS NOT DIFFERENTIABLE 16

3.1 Consistency 16

3.2 Asymptotic Normality 25

4 SIMULATION STUDY 42

4.1 Outlier Inclusion in Subsamples 42

4.2 Subsampling of Moderate Data 46

4.3 Subsampling of Big Data 54

4.4 Breakdown point 77

5 REAL DATA ANALYSIS 96

5.1 Beijing Multi-Site Air-Quality Data 96

5.2 Gas Sensor Array Data Set 105

REFERENCES 110

VITA 112

LIST OF TABLES

Table Page

4.1 $n = 1000,000$ and $p = 50$. Comparison of outlier ratio in subsamples selected using A-optimal sampling distributions from robust linear regressions (Bisquare, Huber, and Hampel), linear regression, and uniform sampling distribution. $\epsilon \sim N(0,1)$ when 5% outliers included. 43

4.2 $n = 1000,000$ and $p = 50$. Comparison of ratios of outliers selected using A-optimal sampling distributions from robust linear regressions (Bisquare, Huber, and Hampel), linear regression, and uniform sampling distribution. $\epsilon \sim N(0,1)$ when 10% outliers included. 45

4.3 $n = 1000,000$ and $p = 50$. Comparison of ratios of outliers selected using A-optimal sampling distributions from robust linear regressions (Bisquare, Huber, and Hampel), linear regression, and uniform sampling distribution. $\epsilon \sim N(0,1)$ when 20% outliers included. 46

4.4 $n = 10,000$ and $p = 50$. Comparison of Bias^2 and MSE using A-optimal and Uniform subsampling methods with different subsample size r for $\epsilon \sim t_3$ distribution. 48

4.5 $n = 10,000$ and $p = 50$. Comparison of MAD0 and MSE0 using A-optimal and Uniform subsampling methods with different subsample size r and different $\psi(x)$ function for $\epsilon \sim t_3$ distribution. 48

4.6 $n = 10,000$ and $p = 50$. Comparison of Bias^2 and MSE using A-optimal and Uniform subsampling methods with different subsample size r for $\epsilon \sim t_1$ distribution. 50

4.7 $n = 10,000$ and $p = 50$. Comparison of MAD0 and MSE0 using A-optimal and Uniform subsampling methods with different subsample size r and different $\psi(x)$ function for $\epsilon \sim t_1$ distribution. 50

4.8 $n = 10,000$ and $p = 50$. Comparison of Bias^2 and MSE using A-optimal and Uniform subsampling methods with different subsample size r for $\epsilon \sim N(0,1)$ when outliers included. 52

4.9 $n = 10,000$ and $p = 50$. Comparison of MAD0 and MSE0 using A-optimal and Uniform subsampling methods with different subsample size r and different $\psi(x)$ function for $\epsilon \sim N(0,1)$ when outliers included. 52

Table Page

4.10 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$. MSE comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers. 56

4.11 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$. MSE comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers. 57

4.12 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$. MSE comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers. 58

4.13 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$. MSE comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated. 59

4.14 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$. MSE comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated. 61

4.15 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$. MSE comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated. 62

4.16 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$. MSE0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers. 63

4.17 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$. MSE0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers. 64

4.18 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$. MSE0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers. 65

Table Page

4.19 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$. MSE comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated. 66

4.20 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$. MSE0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated. 68

4.21 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$. MSE0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated. 69

4.22 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$. MAD0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers. 70

4.23 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$. MAD0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers. 71

4.24 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$. MAD0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers. 72

4.25 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$. MAD0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated. 73

4.26 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$. MAD0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated. 75

4.27 $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$. MAD0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated. 76

Table Page

4.28 $n = 100,000$ and $p = 50$. MSE comparison of robust linear regression with bisquare function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. 79

4.29 $n = 100,000$ and $p = 50$. MSE0 comparison of robust linear regression with Bisquare function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. 81

4.30 $n = 100,000$ and $p = 50$. MAD0 comparison of robust linear regression with Bisquare function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. 83

4.31 $n = 100,000$ and $p = 50$. MSE comparison of robust linear regression with Huber function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. 85

4.32 $n = 100,000$ and $p = 50$. MSE0 comparison of robust linear regression with Huber function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. 87

4.33 $n = 100,000$ and $p = 50$. MAD0 comparison of robust linear regression with Huber function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. 89

4.34 $n = 100,000$ and $p = 50$. MSE comparison of robust linear regression with Hampel function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. 91

4.35 $n = 100,000$ and $p = 50$. MSE0 comparison of robust linear regression with Hampel function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. 93

4.36 $n = 100,000$ and $p = 50$. MAD0 comparison of robust linear regression with Hampel function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. 95

5.1 Comparison of MSEs for linear regression and robust linear regression with Bisquare function using A-optimal and Uniform subsampling distributions for different subsample size r. 98

5.2 Comparison of MADs for linear regression and robust linear regression with Bisquare function using A-optimal and Uniform subsampling distributions for different subsample size r. 99

5.3 Comparison of MSEs for linear regression and robust linear regression with Huber function using A-optimal and Uniform subsampling distributions for different subsample size r. 101

Table Page

5.4 Comparison of MADs for linear regression and robust linear regression with Huber function using A-optimal and Uniform subsampling distributions for different subsample size r. 102

5.5 Comparison of MSEs for linear regression and robust linear regression with Hampel function using A-optimal and Uniform subsampling distributions for different subsample size r. 103

5.6 Comparison of MADs for linear regression and robust linear regression with Hampel function using A-optimal and Uniform subsampling distributions for different subsample size r. 105

5.7 Comparison of MSEs for linear regression and robust linear regression with Bisquare function using A-optimal and Uniform subsampling distributions for different subsample size r. 107

5.8 Comparison of MADs for linear regression and robust linear regression with Bisquare function using A-optimal and Uniform subsampling distributions for different subsample size r. 108

LIST OF FIGURES

Figure Page

1.1 Huber ρ- and ψ-functions . . . 9

1.2 Tukey bisquare ρ- and ψ-functions . . . 10

1.3 Hampel ρ- and ψ-functions . . . 11

4.1 Comparison of outlier ratio in subsamples. $\epsilon \sim N(0,1)$ when 5% outliers included. . . . 44

4.2 Comparison of outlier ratio in subsamples. $\epsilon \sim N(0,1)$ when 10% outliers included. . . . 44

4.3 Comparison of outlier ratio in subsamples. $\epsilon \sim N(0,1)$ when 20% outliers included. . . . 45

4.4 Comparison of Bias^2 and MSE for $\epsilon \sim t_3$ distribution. . . . 49

4.5 Comparison of log(MAD0) for $\epsilon \sim t_3$ distribution. . . . 49

4.6 Comparison of log(MSE0) for $\epsilon \sim t_3$ distribution. . . . 49

4.7 Comparison of Bias^2 and MSE for $\epsilon \sim t_1$ distribution. . . . 51

4.8 Comparison of log(MAD0) for $\epsilon \sim t_1$ distribution. . . . 51

4.9 Comparison of log(MSE0) for $\epsilon \sim t_1$ distribution. . . . 51

4.10 Comparison of Bias^2 and MSE for $\epsilon \sim N(0,1)$ when outliers included. . . . 53

4.11 Comparison of log(MAD0) for $\epsilon \sim N(0,1)$ when outliers included. . . . 53

4.12 Comparison of log(MSE0) for $\epsilon \sim N(0,1)$ when outliers included. . . . 53

4.13 log(MSE) for $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$. . . . 56

4.14 log(MSE) for $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$. . . . 57

4.15 log(MSE) for $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$. . . . 58

4.16 log(MSE) for $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$ when truncated. . . . 60

4.17 log(MSE) for $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$ when truncated. . . . 60

4.18 log(MSE) for $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$ when truncated. . . . 61

4.19 log(MSE0) for $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$ . . . 63

Figure Page

4.20 log(MSE0) for $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$ . . . 64

4.21 log(MSE0) for $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$. . . . 65

4.22 log(MSE0) for $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$ when truncated . . . 67

4.23 log(MSE0) for $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$ when truncated . . . 67

4.24 log(MSE0) for $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$ when truncated. . . . 68

4.25 log(MAD0) for $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$ . . . 70

4.26 log(MAD0) for $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$ . . . 71

4.27 log(MAD0) for $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$. . . . 72

4.28 log(MAD0) for $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$ when truncated . . . 74

4.29 log(MAD0) for $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$ when truncated . . . 74

4.30 log(MAD0) for $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$ when truncated. . . . 75

4.31 log(MSE) for weighted robust linear regression with bisquare function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. . . . 78

4.32 log(MSE0) for weighted robust linear regression with Bisquare function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. . . . 80

4.33 log(MAD0) for weighted robust linear regression with Bisquare function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. . . . 82

4.34 log(MSE) for weighted robust linear regression with Huber function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. . . . 84

4.35 log(MSE0) for weighted robust linear regression with Huber function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. . . . 86

4.36 log(MAD0) for weighted robust linear regression with Huber function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. . . . 88

4.37 log(MSE) for weighted robust linear regression with Hampel function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. . . . 90

Figure Page

4.38 log(MSE0) for weighted robust linear regression with Hampel function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. 92

4.39 log(MAD0) for weighted robust linear regression with Hampel function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction. 94

5.1 Studentized residual plot for linear regression 97

5.2 log(MSE) for linear regression and robust linear regression with Bisquare function using A-optimal and Uniform subsampling distributions for different subsample size r. 98

5.3 log(MAD) for linear regression and robust linear regression with Bisquare function using A-optimal and Uniform subsampling distributions for different subsample size r. 99

5.4 log(MSE) for linear regression and robust linear regression with Huber function using A-optimal and Uniform subsampling distributions for different subsample size r. 100

5.5 log(MAD) for linear regression and robust linear regression with Huber function using A-optimal and Uniform subsampling distributions for different subsample size r. 102

5.6 log(MSE) for linear regression and robust linear regression with Hampel function using A-optimal and Uniform subsampling distributions for different subsample size r. 103

5.7 log(MAD) for linear regression and robust linear regression with Hampel function using A-optimal and Uniform subsampling distributions for different subsample size r. 104

5.8 Studentized residual plot for linear regression 106

5.9 log(MSE) for linear regression and robust linear regression with Bisquare function using A-optimal and Uniform subsampling distributions for different subsample size r. 107

5.10 log(MAD) for linear regression and robust linear regression with Bisquare function using A-optimal and Uniform subsampling distributions for different subsample size r. 108

ABSTRACT

Tang, Ziting Ph.D., Purdue University, December 2019. Robust A-optimal Subsampling for Massive Data Robust Linear Regression. Major Professors: Fei Tan, Hanxiang Peng.

This thesis is concerned with massive data analysis via robust A-optimally efficient non-uniform subsampling. Motivated by the fact that massive data often contain outliers and that uniform sampling is not efficient, we give numerous sampling distributions by minimizing the sum of the component variances of the subsampling estimate. And these sampling distributions are robust against outliers. Massive data pose two computational bottlenecks. Namely, data exceed a computer's storage space, and computation requires too long waiting time. The two bottle necks can be simultaneously addressed by selecting a subsample as a surrogate for the full sample and completing the data analysis. We develop our theory in a typical setting for robust linear regression in which the estimating functions are not differentiable. For an arbitrary sampling distribution, we establish consistency for the subsampling estimate for both fixed and growing dimension(as high dimensionality is common in massive data). We prove asymptotic normality for fixed dimension. We discuss the A-optimal scoring method for fast computing. We conduct large simulations to evaluate the numerical performance of our proposed A-optimal sampling distribution. Real data applications are also performed.

1. INTRODUCTION

In the past years, there is a huge growth in sample size and dimensions of the data which is called big data. Big data bring not only opportunities but also challenges to statisticians and data scientists. Specifically, big data has more information to be used to discover population patterns compared to small or medium-size data. But big data can have storage bottleneck, high computational cost and statistical challenges because of the exceptionally large sample size and very high dimensionality. Fan, et al. (2014) described about the impact of big data on statistical methods and computing architectures in details. The challenges caused by two features of big data can result in misleading statistical inferences and conclusions.

In big data analysis, the inaccuracy of an estimator arises mainly from random error, computing error and rounding error. Problems also will arise for large data set when we use bootstrap to estimate asymptotic distribution of the estimators. Due to the large sample size, the computation cost is large. It will be difficult and needs longer time to get the full sample estimates, especially to obtain thousands of estimates when the bootstrap method is used. It becomes more difficult for high dimensional data set. Subsampling method is one solution to solve the problems in big data analysis. It reduces substantially the sample sizes, improves errors and speeds up computation.Among the subsampling methods we have known, the uniform sampling is commonly used due to its simplicity and fast computation. Peng and Tan (2018) studied about uniform subsampling in a linear model. However, uniform sampling is not efficient in extracting information in data and sampling important observations. There are plenty of non-uniform subsampling methods can be found in the literature. Ma, et al. (2015) derived a non-uniform subsampling distribution by minimizing the trace of central part of a specific variance-covariance matrix. Peng and Tan (2019) derived the A-optimal probability distribution in linear regression and discussed the asymptotic expansion and normality of the subsampling estimator.

Another problem in big data is outliers. This is because outliers are common in massive data. For uniform sampling, all the data points including outliers in the sample have the same chance to be selected. This will have a great effect on the estimates in the linear regression. Many existing optimal sampling distributions are extremely not robust. Even a very small amount of outliers in data will ruin the subsampling estimates based on A-optimal on non-roubst sampling distributions. This is because A-optimality seeks "outliers" in the spirity of the Design of Experiements. For example, Ma, et al. (2015) derived a non-uniform subsampling distribution which sample influential data points with high probabilities. Moreover, there is problem about the outlier proportions in the bootstrap samples. They may have the same or even higher percentage of outliers than that in the original dataset. As a result, the estimates calculated based on the bootstrap samples will be affected by the outliers.Singh (1998) discussed bootstrap quantiles and calculated its breakdown point which is very low for some robust estimates. Salibian-Barrera and Zamar (2002) introduced how to approximate the distribution of weighted estimates in robust linear regression by applying a reweighted representation of the estimates. The decreasing functions are used as weights to deal with outliers in the data. They obtained higher breakdown points than those calculated from the bootstrap.

In linear regression, the ordinary least squares estimator is the optimal regression estimator under a set of assumptions. But in regression analysis, OLS estimates have bad performance when the error distribution is not normally distributed or if there exist outliers. That means, the least squares method is not appropriate on data sets containing outliers. As a result, we can get very misleading results. So, when the linear regression can't perform well on the data with outliers, we will use robust linear regression to deal with this problem. Three methods (M-estimation, S-estimation and MM-estimation) are the most often used in robust linear regression. In this dissertation, we study M estimation which was extended from the maximum likelihood estimation and introduced by Huber (1964) with subsampling methods. The optimal non-uniform subsampling in robust linear regression is derived from the criterion of A-optimality. That is, we seek the sampling distribution that minimizes the trace of certain dispersion matrix. We use approximated A-optimal

subsampling distribution to calculate the subsampling estimator in robust linear regression when the sample size of the data is extremely large. We also study the robustness of the subsampling estimators through their breakdown points.

In the robust linear regression, bisquare, Huber and Hampel functions are often used as the objective function. But Huber and Hampel functions are not differentiable at some points. Portnoy (1984) discussed the consistency of full sample estimator $\hat{\boldsymbol{\beta}}$ for all objective functions ψ including non-differentiable ones. They established the condition that p is allowed to be increased under a weaker condition $p \log p/n \to 0$ compared to $p^2/n \to 0$. After that, Portnoy (1985) discussed the asymptotic normality of the full sample estimator for all ψ functions. Our theory is established in a typical framework for robust linear regression. Specifically, we have proved the consistency and asymptotic normality when the estimating functions are not differentiable. Since massive data usually has high dimension, when we study the consistency of the subsampling estimate in this dissertation, we consider the situation in which the dimension grows with the increasing sample size. When we discuss the asymptotic normality of the subsampling estimate, we consider random covariates and fixed dimension p.

For future work, we will study the bias of the subsampling estimates and investigate the asymptotic distribution of the subsampling estimate for growing dimension p. We will establish the asymptotic normality when covariates are non-random. Besides that, we will investigate more robustness properties. Finally, we will apply our method in real data applications such as in astromical data.

The dissertation includes five chapters. Chapter 1 contains the introduction of the regression estimates of linear regression and robust linear regression. Chapter 2 presents the asymptotic distribution of subsample estimator when ψ function is differential. Chapter 3 illustrates consistency and asymptotic normality of subsampling estimator when ψ function is non-differential. Chapter 4 presents some simulation results. In chapter 5, we study about the real data sets on Beijing Multi-Site Air-Quality and which include outliers and compare the performance between different methods.

1.1 Linear Regression and Classical Estimation

Consider linear regression below which is used to model the linear relationship between dependent variables y_i and explanatory variables $\boldsymbol{x}_i (i = 1, \cdots, n)$:

$$y_i = \boldsymbol{x}_i' \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \cdots, n$$

where $\epsilon_i \sim N(0, \sigma^2)$ are independently and identically distributed.

The regression coefficients parameters are estimated from the data by ordinary least square method. Then the least squares estimator (LSE) $\hat{\boldsymbol{\beta}}$ can be obtained by minimizing the residual sum of squares (RSS), i.e.,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i' \boldsymbol{\beta})^2. \tag{1.1}$$

Taking the partial derivatives of the expression in (1.1) with respect to the regression coefficients $\boldsymbol{\beta}$ and letting them equal to 0, we get

$$\sum_{i=1}^{n} (y_i - \boldsymbol{x}_i' \boldsymbol{\beta}) \boldsymbol{x}_i = 0. \tag{1.2}$$

Then solve the normal equations in (1.2), we can get LSE $\hat{\boldsymbol{\beta}}$.

To check the fit of the estimated regression model to the data, we can look at the size of the residuals $r_i(\hat{\boldsymbol{\beta}}) = y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}}$ in residual plot. A point which has a large residual and is far from the horizontal line in the residual plot is called an outlier. Such points may be due to data entry errors. We can use studentized residuals to detect outliers.

1.2 Robust Linear Regression and M-estimation

Ordinary least squares estimates can perform badly, i.e. not robust when the normality of error distribution is violated or there are outliers. We can remove influential observations before fitting the linear regression model. Another way to solve this problem is using robust linear regression. That is, we use a different criterion which is less affected by unusual data than a quadratic function. A common method of robust linear regression was introduced

by Huber (1973). This method is generalized from maximum-likelihood estimation. So, it is called M-estimation.

When we replace the square function of residules used in OLS estimation by another function, we will get the M-estimators. That is, the estimates $\hat{\boldsymbol{\beta}}$ can be obtained by minimizing a objective function ρ over all $\boldsymbol{\beta}$ in M-estimation.

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho(r_i(\boldsymbol{\beta})), \tag{1.3}$$

where $r_i(\boldsymbol{\beta}) = y_i - \boldsymbol{x}_i'\boldsymbol{\beta}$. Usually, the objective function ρ will satisfy:

1. The function ρ is always nonnegative, i.e., $\rho(r) \geq 0$ for all residuals r.
2. The function ρ is zero if the residual is zero, i.e., $\rho(0) = 0$.
3. The function ρ is symmetric, i.e., $\rho(r) = \rho(-r)$.
4. The function ρ is monotone in $|r_i|$, i.e., $\rho(r_i) \geq \rho(r_j)$ for $|r_i| > |r_j|$.

If ρ is differentiable, then we take the partial derivatives of expression (1.4) with respect to the regression coefficients $\boldsymbol{\beta}$ and letting them equal to 0. Then we get

$$\sum_{i=1}^{n} \psi(r_i(\boldsymbol{\beta}))\boldsymbol{x}_i = 0, \tag{1.4}$$

where $\psi = \rho'$. Solving this estimating equation, we can obtain the M-estimator.

Although M-estimators are regression equivariant, they are not scale equivariant. We have to standardize the M-estimators by a robust scale estimate $\hat{\sigma}$:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho\Big(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\Big),$$

or solve

$$\sum_{i=1}^{n} \psi\Big(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\Big)\boldsymbol{x}_i = 0.$$

One common choice for the scale estimate $\hat{\sigma}$ is median absolute deviation (MAD):

$$\hat{\sigma} = C * MAD = C\ median_i\Big(|r_i - median_j(r_j)|\Big),$$

where C is a correction factor which is determined by the distribution. For data following normal distribution $N(\mu, \sigma)$, we have $median(|r - \mu|) \approx 0.6745\sigma$. So, $C = \frac{1}{0.6745} = 1.4826$.

There are many examples for M-estimators. For example, all MLEs including OLS estimator are M-estimators. The linear regression loss function for OLS estimator is $\rho(t) = t^2$ ($\psi(t) = t$) which increases dramatically as the size of the residual is increasing. Another example is $L1$ estimator which uses the absolute value as a loss function: $\rho(t) = |t|$ ($\psi(t) = sgn(t)$). This cab achieve robustness.

For many functions ρ and ψ, equation (1.4) has no closed form solution. In such cases, we can perform an iteratively reweighted least squares method to solve for M-estimators. Let $w_i = w\Big(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\Big) = \psi\Big(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\Big)/\Big(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\Big)$. Then we can rewrite the estimating equation:

$$\sum_{i=1}^{n} \psi\Big(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\Big)\boldsymbol{x}_i = \sum_{i=1}^{n} w_i\Big(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}}\Big)\boldsymbol{x}_i = 0. \tag{1.5}$$

Solving the estimating equation (1.4) becomes a weighted least-squares problem. By definition, the weights w_i are functions of residuals r_i which are calculated from $r_i(\hat{\boldsymbol{\beta}}) = y_i - \boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ and depend on $\hat{\boldsymbol{\beta}}$. But $\hat{\boldsymbol{\beta}}$ are calculated from (1.5) and depend on the weights w_i. Then the iteratively reweighted least squares method is performed to get an iterative solution of (1.5).

1.2.1 Two criteria for robustness

Breakdown point is used as a criterion for global robustness. Donoho and Huber (1983) introduced the breakdown point for finite sample.

Let $\boldsymbol{z} = (\boldsymbol{z}_1, ..., \boldsymbol{z}_n)'$ be an sample, where $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$. Denote $T(\boldsymbol{z})$ as the estimate of a parameter $\boldsymbol{\beta}$ based on the sample $\boldsymbol{z}$. If we replace any m data points in the sample $\boldsymbol{z}$ by arbitrary outliers, then we get the estimate $T(\boldsymbol{z}'_m)$ based on a new sample $\boldsymbol{z}'_m$. The maximum difference between $T(\boldsymbol{z})$ and $T(\boldsymbol{z}'_m)$ for such replacement is defined as bias.

$$bias(m; T, \boldsymbol{z}) = \sup_{\boldsymbol{z}'_m} \|T(\boldsymbol{z}'_m) - T(\boldsymbol{z})\|,$$

where $\|\cdot\|$ is Euclidean norm.

If there exist some m outliers such that the difference $\|T(\boldsymbol{z}'_m) - T(\boldsymbol{z})\|$ is arbitrary large, then $bias(m; T, \boldsymbol{z}) = \infty$. That means, the estimator breaks down when m outliers have very large influence. Then, the breakdown point (BP) ε^* of the estimator $\hat{\boldsymbol{\beta}}$ for finite sample is defined as

$$\varepsilon^*(m; T, \boldsymbol{z}) = \min_{1 \leq m \leq n} \left\{ \frac{m}{n} : bias(m; T, \boldsymbol{z}) = \infty \right\}$$

A higher BP value indicates higher robustness of the estimator to outliers. And the highest breakdown point is 0.5. Below are some examples:

1. The finite sample BP of sample mean is $\frac{1}{n}$. This is because one unusual observation can result in arbitrarily large sample mean. Since $\frac{1}{n} \to 0$ as $n \to \infty$, the asymptotic BP of sample mean is 0.

2. The finite sample BP of median is 0.5. This is because median can tolerate 50% outliers.

3. The finite sample BP of α-trimmed mean is α.

4. In OLS regression, The finite sample BP of coefficient estimates $\hat{\boldsymbol{\beta}}$ is $\frac{1}{n}$. This is because one outlier is enough to affect $\hat{\boldsymbol{\beta}}$. The asymptotic BP of coefficient estimates $\hat{\boldsymbol{\beta}}$ is 0.

5. Redescending M-estimators in robust linear regression have very high breakdown points.

To introduce the measurement of local robustness influence function, the sensitivity curve which is used to estimate the effect of one outlier on the estimator is defined as below. If a new observation $\boldsymbol{z}_0$ is added to the sample, then we get a new sample. The sensitivity curve (SC) of the estimate for this sample is defined as the difference between two estimates.

$$SC_n(\boldsymbol{z}_0, T) = T_{n+1}(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n, \boldsymbol{z}_0) - T_n(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)$$

Since there are $\frac{1}{n+1}$ outliers in the new sample, the standardized sensitivity curve (SC) is defined as

$$\begin{aligned} SC_n(\boldsymbol{z}_0, T) &= \frac{T_{n+1}(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n, \boldsymbol{z}_0) - T_n(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)}{1/(n+1)} \\ &= (n+1)(T_{n+1}(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n, \boldsymbol{z}_0) - T_n(\boldsymbol{z}_1, \cdots, \boldsymbol{z}_n)). \end{aligned}$$

The influence function (IF) of an estimate T is an asymptotic form of SC. It is a criterion for the local robustness. Suppose F is a distribution of sample in which identical outliers are included. Let ϵ denote the fraction of outliers and δ be the point mass of outlier. Then the IF is defined as

$$\begin{aligned} IF_T(\boldsymbol{z}_0, F) &= \lim_{\epsilon \to 0^+} \frac{T((1-\epsilon)F + \epsilon\delta_{\boldsymbol{z}_0}) - T(F)}{\epsilon} \\ &= \frac{\partial}{\partial \epsilon} T((1-\epsilon)F + \epsilon\delta_0)|_{\epsilon \to 0^+}. \end{aligned}$$

If σ is known, then the IF of an M-estimator is

$$IF_T(\boldsymbol{x}, y; F) = \frac{\sigma\psi(\frac{y - \boldsymbol{x}'\boldsymbol{\beta}}{\sigma})\boldsymbol{x}}{E(\psi'(\frac{y - \boldsymbol{x}'\boldsymbol{\beta}}{\sigma})\boldsymbol{x}\boldsymbol{x}')}.$$

1.2.2 Examples of loss function

In robust linear regression, Huber function, Tukey's bisquare function and Hampel function are often used as the objective function.

Huber function

One popular choice was proposed by Huber in 1964 and known as Huber functions. It is constructed by loss functions of OLS and L1. It is given by:

$$\rho(x) = \begin{cases} \frac{x^2}{2} & \text{if } |x| \le k \\ k|x| - \frac{k^2}{2} & \text{if } |x| > k \end{cases}$$

with derivative $\psi(x)$, where

$$\psi(x) = \begin{cases} x & \text{if } |x| \le k \\ sgn(x)k & \text{if } |x| > k \end{cases}$$

$k = 1.345\sigma$ is a good choice, where σ is the SD of the error distribution. If the errors have a normal distribution, then it is 95% as efficient as least squares asymptotically. In many other cases, it is also much more efficient.

We can see that ρ_k is quadratic in the central part $[-k, k]$ and increases linearly to infinity. The M-estimators in robust linear regression are the OLS and L1 estimators when $k \to \infty$ and $k \to 0$, respectively. ψ is constant outside $[-k, k]$.

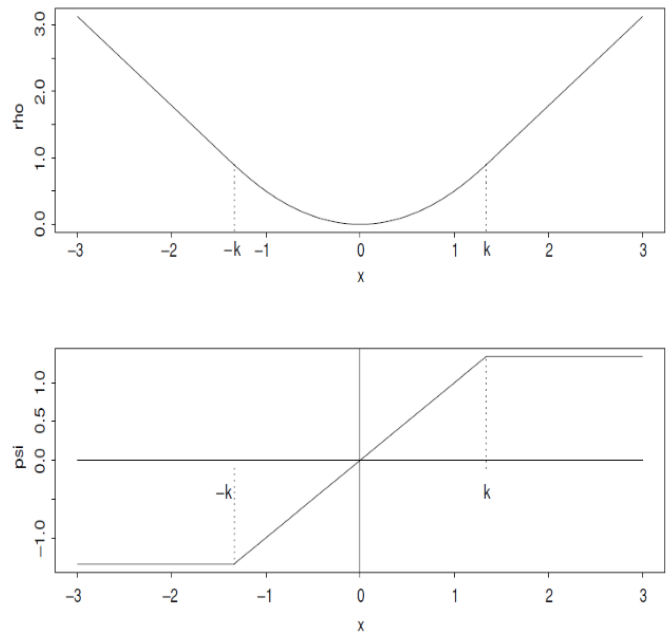

Fig. 1.1.: Huber ρ- and ψ-functions

Tukey's bisquare function

Another popular objective function was introduced by Tukey and known as Tukey's bisquare or Tukey's biweight. It is given by:

$$\rho(x) = \begin{cases} \frac{k^2}{6}\{1 - [1 - (\frac{x}{k})^2]^3\} & \text{if } |x| \leq k \\ \frac{k^2}{6} & \text{if } |x| > k \end{cases}$$

$$= \begin{cases} \frac{x^2}{2} - \frac{x^4}{2k^2} + \frac{x^6}{6k^4} & \text{if } |x| \leq k \\ \frac{k^2}{6} & \text{if } |x| > k \end{cases}$$

with derivative $\psi(x)$, where

$$\psi(x) = x\Big[1 - \Big(\frac{x}{k}\Big)^2\Big]^2 \mathrm{I}(|x| \leq k).$$

Similarly, the value $k = 4.685\sigma$ is usually used for Tukey's bisquare function, where σ is the SD of the error distribution.

We can see that ψ is differentiable everywhere and becomes 0 outside $[-k, k]$.

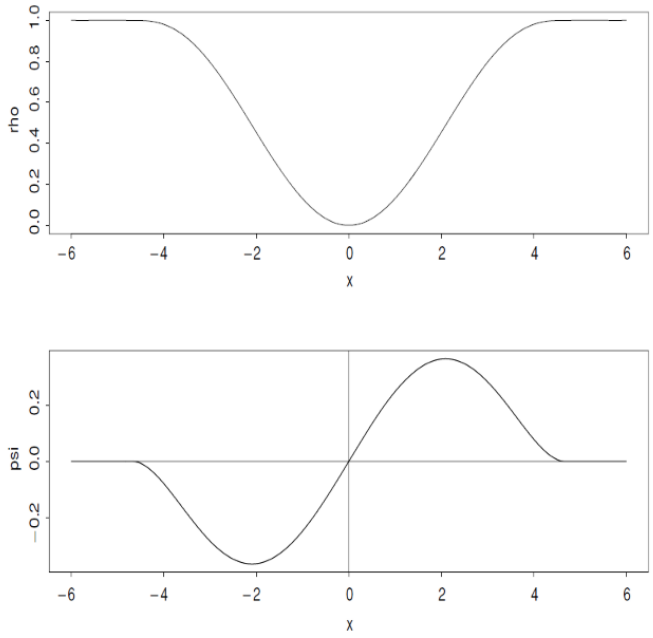

Fig. 1.2.: Tukey bisquare ρ- and ψ-functions

Hampel function

Another redescending function is Hampel function. It is very similar to bisquare function except that it is not differentible at some points. It is given by

$$\rho(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \leq a \\ a|x| - \frac{1}{2}a^2 & \text{if } a < |x| \leq b \\ a\frac{c|x| - \frac{1}{2}x^2}{c-b} - \frac{7a^2}{6} & \text{if } b < |x| \leq c \\ a(b + c - a) & |x| > c \end{cases}$$

with derivative $\psi(x)$, where

$$\psi(x) = \begin{cases} x & \text{if } |x| \leq a \\ a * sgn(x) & \text{if } a < |x| \leq b \\ a\frac{c-|x|}{c-b} * sgn(x) & \text{if } b < |x| \leq c \\ 0 & |x| > c \end{cases}$$

Similarly, the values $a = 2\sigma$, $b = 4\sigma$ and $c = 8\sigma$ are usually used for Hampel function, where σ is the SD of the error distribution.

We can see that ψ has a non-zero constant value inside $[-b, -a]$ or $[a, b]$ and becomes 0 outside $[-c, c]$.

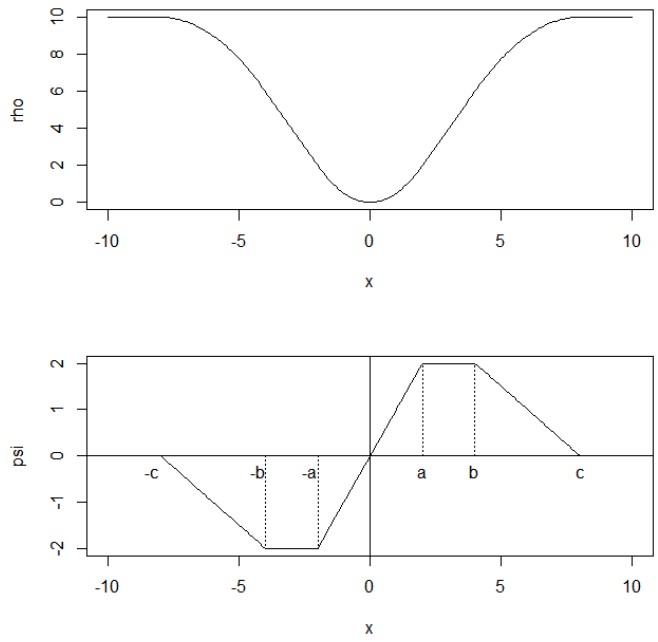

Fig. 1.3.: Hampel ρ- and ψ-functions

2. A-OPTIMAL SUBSAMPLING METHOD

In this section, we draw a subsample and construct a subsampling estimator to approximate the robust regression estimator. Cheung(2019) has derived the asyptotic property for the general estimating equations. Assume the ψ function in robust linear regression is differentiable, then it becomes a specific situation in Cheung(2019). Hence, we will omit the proof in this section.

Consider

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + \epsilon_i, \quad i = 1, \cdots, n$$

where ϵ_i follows a symmetric distribution. Suppose random sample $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)$ are i.i.d. Let $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_n)$ be a sampling distribution for the i.i.d sample points. Draw a subsample of size $r \ll n$ randomly from original sample with replacement according to corresponding probabilities $\boldsymbol{\pi}^* = (\pi_1^*, \cdots, \pi_r^*)$. Then we get a subsample $(\boldsymbol{x}_1^*, y_1^*), \cdots, (\boldsymbol{x}_r^*, y_r^*)$.

Let $\hat{\boldsymbol{\beta}}$ be the M-estimator obtained by the full sample, which is unknown and to be estimated by the subsampling method. It is the solution of equation

$$\sum_{i=1}^{n} \psi\Big(\frac{y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}}{\sigma}\Big)\boldsymbol{x}_i = 0.$$

Let $\boldsymbol{w} = (w_1, \cdots, w_n)$ have the scaled multinomial distribution with parameter vector $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_n)$ and number of trials r, that is

$$P\Big(w_1 = \frac{k_1}{r\pi_1}, \cdots, w_n = \frac{k_n}{r\pi_n}\Big) = \frac{r!}{\prod_{i=1}^n k_i!}\prod_{i=1}^{n}\pi^{k_i}, k_i \geq 0, \sum_{i=1}^{n} k_i = r.$$

It is easy to prove that

$$\mathbb{E}^*(w_i) = 1, \tag{2.1}$$

and

$$\mathbb{E}^*(w_i w_j) = 1 - \frac{1}{r} + \frac{1}{r\pi_i}I(i=j) = \begin{cases} \frac{1}{r}\left(\frac{1}{\pi_i} - 1\right) + 1, & \text{for } i = j \\ 1 - \frac{1}{r}, & \text{for } i \neq j \end{cases} \tag{2.2}$$

Let $\hat{\boldsymbol{\beta}}^*$ be weighted subsample estimator. We minimize weighted objective function on the subsample to get the weighted subsample estimator $\hat{\boldsymbol{\beta}}^*$.

$$\hat{\boldsymbol{\beta}}^* = \arg\min_{\beta} \sum_{i=1}^{r} w_i^* \rho\Big(\frac{y_i^* - \boldsymbol{x}_i^{*T}\boldsymbol{\beta}}{\sigma}\Big).$$

If ρ is differentiable, let $\rho' = \psi$, then we get $\hat{\boldsymbol{\beta}}^*$ by solving

$$\sum_{i=1}^{r} w_i^* \psi\Big(\frac{y_i^* - \boldsymbol{x}_i^{*T}\boldsymbol{\beta}}{\sigma}\Big)\boldsymbol{x}_i^* = 0.$$

If $\rho(z) = \frac{1}{2}z^2$, then we can get the ordinary least square estimator.

Denote

$$\boldsymbol{\psi}(\hat{\boldsymbol{\beta}}_n) = \Big(\psi\Big(\frac{y_1 - \boldsymbol{x}_1^T\hat{\boldsymbol{\beta}}}{\sigma}\Big), \cdots, \psi\Big(\frac{y_n - \boldsymbol{x}_n^T\hat{\boldsymbol{\beta}}}{\sigma}\Big)\Big)^T,$$

$$\dot{\boldsymbol{\Psi}} = diag\Big\{\frac{1}{\sigma}\psi'\Big(\frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\sigma}\Big)\Big\}_{i=1}^{n},$$

$$\boldsymbol{W} = diag\Big\{\frac{1}{r\pi_i}\Big\}, \quad \tilde{\boldsymbol{H}} = \boldsymbol{X}(\boldsymbol{X}^T\dot{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}\boldsymbol{X}^T.$$

Let

$$\boldsymbol{\Psi}_r^*(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{r} w_i^* \psi\Big(\frac{y_i^* - \boldsymbol{x}_i^{*T}\boldsymbol{\beta}}{\sigma}\Big)\boldsymbol{x}_i^*.$$

Then

$$\boldsymbol{\Psi}_r^*(\hat{\boldsymbol{\beta}}^*) = \frac{1}{n}\sum_{i=1}^{r} w_i^* \psi\Big(\frac{y_i^* - \boldsymbol{x}_i^{*T}\hat{\boldsymbol{\beta}}^*}{\sigma}\Big)\boldsymbol{x}_i^* = 0.$$

Assume ψ is a differentiable function. Then expand $\boldsymbol{\Psi}_r^*(\boldsymbol{\beta})$ into a Taylor series about $\hat{\boldsymbol{\beta}}$ and evaluating it at $\hat{\boldsymbol{\beta}}^*$, we have

$$0 = \boldsymbol{\Psi}_r^*(\hat{\boldsymbol{\beta}}^*) = \boldsymbol{\Psi}_r^*(\hat{\boldsymbol{\beta}}) + \dot{\boldsymbol{\Psi}}_r^*(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) + \boldsymbol{R}_r,$$

where $\dot{\boldsymbol{\Psi}}_r^*$ is the derivative of $\boldsymbol{\Psi}_r^*$, $\boldsymbol{R}_r = (\dot{\boldsymbol{\Psi}}_r^*(\boldsymbol{\beta}_*) - \dot{\boldsymbol{\Psi}}_r^*(\hat{\boldsymbol{\beta}}))(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})$ and $\boldsymbol{\beta}_*$ lies between $\hat{\boldsymbol{\beta}}^*$ and $\hat{\boldsymbol{\beta}}$. Solve above equation, we can get

$$\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}} - \dot{\boldsymbol{\Psi}}_r^*(\hat{\boldsymbol{\beta}})^{-1}(\boldsymbol{\Psi}_r^*(\hat{\boldsymbol{\beta}}) + \boldsymbol{R}_r).$$

For fixed p, choose function ψ_{ni} in Cheung(2019) as $\psi_{ni}(\hat{\boldsymbol{\beta}}) = \boldsymbol{x}_i\psi\Big(\frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\sigma}\Big)$, then we get specific conditions for below theorem from Cheung(2019).

Theorem 2.1. *With certain conditions, the following expansion holds,*

$$\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^T\dot{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{\psi}(\hat{\boldsymbol{\beta}})) + o_P(r^{-1}),$$

The bias of $\hat{\boldsymbol{\beta}}^*$ about $\hat{\boldsymbol{\beta}}$ can be expanded as

$$\mathbb{E}^*(\hat{\boldsymbol{\beta}}^*) - \hat{\boldsymbol{\beta}} = -\frac{1}{r}(\boldsymbol{X}^T\dot{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}\boldsymbol{X}^T\dot{\boldsymbol{\Psi}}Diag\Big(\frac{\tilde{\boldsymbol{h}}}{\boldsymbol{\pi}}\Big)\boldsymbol{\psi}(\hat{\boldsymbol{\beta}}) + o_P(r^{-1}),$$

where $\tilde{\boldsymbol{h}}$ is the vector composed of the diagonal elements of $\tilde{\boldsymbol{H}}$.

Moreover, the variance-covariance matrix of $\hat{\boldsymbol{\beta}}^*$ can be expressed by

$$\mathbf{Var}^*(\hat{\boldsymbol{\beta}}^*) = \frac{1}{r}(\boldsymbol{X}^T\dot{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}\boldsymbol{X}^TDiag\Big(\frac{\boldsymbol{\psi}^2(\hat{\boldsymbol{\beta}})}{\boldsymbol{\pi}}\Big)\boldsymbol{X}(\boldsymbol{X}^T\dot{\boldsymbol{\Psi}}\boldsymbol{X})^{-1} + o_P(r^{-1}),$$

Theorem 2.2. *Under same conditions of Theorem 2.1, then these is a sequence of subsample estimates $\hat{\boldsymbol{\beta}}^*$ such that as $r \to \infty$,*

$$V^{-\frac{1}{2}}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \to N(0, \boldsymbol{I})$$

in probability, where

$$V = \frac{1}{r}(\boldsymbol{X}^T\dot{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}\boldsymbol{X}^TDiag\Big(\frac{\boldsymbol{\psi}^2(\hat{\boldsymbol{\beta}})}{\boldsymbol{\pi}}\Big)\boldsymbol{X}(\boldsymbol{X}^T\dot{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}.$$

Minimize the variance-covariance matrix $\mathrm{Var}^*(\hat{\boldsymbol{\beta}}^*)$ of subsampling estimator $\hat{\boldsymbol{\beta}}^*$ in the sense of minimizing trace of the main term of covariance matrix

$$\begin{aligned}
\tau(\boldsymbol{\pi}) =& Tr\Big(\mathbf{Var}^*\Big[\Big(\frac{1}{\sigma}\sum_{i=1}^n \boldsymbol{x}_i\psi'\Big(\frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\sigma}\Big)\boldsymbol{x}_i^T\Big)^{-1}\Big(\sum_{i=1}^r w_i^*\psi\Big(\frac{y_i^* - \boldsymbol{x}_i^{*T}\hat{\boldsymbol{\beta}}}{\sigma}\Big)\boldsymbol{x}_i^*\Big)\Big]\Big) \\
=& \sum_{i=1}^n \frac{1}{r\pi_i}\big((\boldsymbol{X}^T\dot{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}\boldsymbol{x}_i\big)^T\big((\boldsymbol{X}^T\dot{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}\boldsymbol{x}_i\big)\psi^2\Big(\frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\sigma}\Big) \\
=& \sum_{i=1}^n \frac{1}{r\pi_i}\Big\|\Big(\sum_{j=1}^n \boldsymbol{x}_j\psi'\Big(\frac{y_j - \boldsymbol{x}_j^T\hat{\boldsymbol{\beta}}}{\sigma}\Big)\boldsymbol{x}_j^T\Big)^{-1}\boldsymbol{x}_i\Big\|^2\psi^2\Big(\frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\sigma}\Big). \qquad (2.3)
\end{aligned}$$

By the Lagrange multiplier method, we get the A-optimal subsampling probability

$$\begin{aligned}
\hat{\pi}_i =& \frac{\Big\|\big(\sum_{j=1}^n \boldsymbol{x}_j\psi'\big(\frac{y_j - \boldsymbol{x}_j^T\hat{\boldsymbol{\beta}}}{\sigma}\big)\boldsymbol{x}_j^T\big)^{-1}\boldsymbol{x}_i\Big\| \cdot \Big|\psi\big(\frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\sigma}\big)\Big|}{\sum_{i=1}^n \Big\|\big(\sum_{j=1}^n \boldsymbol{x}_j\psi'\big(\frac{y_j - \boldsymbol{x}_j^T\hat{\boldsymbol{\beta}}}{\sigma}\big)\boldsymbol{x}_j^T\big)^{-1}\boldsymbol{x}_i\Big\| \cdot \Big|\psi\big(\frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\sigma}\big)\Big|} \\
=& \frac{\big\|(\boldsymbol{X}^T\dot{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}\boldsymbol{x}_i\big\| \cdot \big|\psi\big(\frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\sigma}\big)\big|}{\sum_{i=1}^n \big\|(\boldsymbol{X}^T\dot{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}\boldsymbol{x}_i\big\| \cdot \big|\psi\big(\frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\sigma}\big)\big|},
\end{aligned}$$

where $\dot{\boldsymbol{\Psi}} = diag\{\psi'\big(\frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\sigma}\big)\}_{i=1}^n$. When $\psi(x) = x$, this becomes the situation for linear regression discussed by Peng and Tan (2018).

Theorem 2.3. *Assume that $(\boldsymbol{X}^T\dot{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}$ is invertible such that $a_i = (\boldsymbol{X}^T\dot{\boldsymbol{\Psi}}\boldsymbol{X})^{-1}x_i \neq 0$. Then there exists a unique A-optimal distribution $\hat{\boldsymbol{\pi}}$ for $\hat{\boldsymbol{\beta}}^*$ to approximate $\hat{\boldsymbol{\beta}}$, which is given by*

$$\hat{\pi}_i = \frac{\|a_i\| \cdot \big|\psi\big(\frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\sigma}\big)\big|}{\sum_{i=1}^n \|a_i\| \cdot \big|\psi\big(\frac{y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{\sigma}\big)\big|}, \quad i = 1, \cdots, n.$$

In big data analysis, the sample size of the data is usually large and calculation is difficult. Below is the weighted estimation algorithm by A-optimal subsampling method in Peng and Tan(2019).

1. Subsample(size r_0) with replacement from the data by uniform distribution. Use the robust scale estimate $\hat{\sigma}$ to construct the approximate A-optimal subsampling probability $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^n$.

2. Draw a subsample of size $r \ll n$ randomly according to the sampling distribution of $\boldsymbol{\pi}^*$ with replacement. Then we get a subsample $(\boldsymbol{x}_1^*, y_1^*), \cdots, (\boldsymbol{x}_r^*, y_r^*)$.

3. Solve weighted estimating equation below using the subsample to get the weighted subsample estimator $\hat{\boldsymbol{\beta}}^*$.

$$\sum_{i=1}^r \frac{1}{\pi_i^*}\psi\Big(\frac{y_i^* - \boldsymbol{x}_i^{*T}\hat{\boldsymbol{\beta}}^*}{\sigma}\Big)\boldsymbol{x}_i^* = 0$$

3. ASYMPTOTIC BEHAVIOR OF M-ESTIMATORS WHEN ψ IS NOT DIFFERENTIABLE

3.1 Consistency

Consider a linear regression model in which the response y_i and covariate $\boldsymbol{x}_i \in \mathbb{R}^p$ satisfy

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + R_i, \quad i = 1, \cdots, n$$

where $\boldsymbol{\beta}$ is an unknown parameter and $R_i (i = 1, 2, \cdots, n)$ are i.i.d. random errors. In this section, we consider non-random $\boldsymbol{X}$.

Consider the case that n is extremely large and the full-data estimator $\hat{\boldsymbol{\beta}}$ is not available either due to the physical limitation of computer's memory or too long waiting time.

Let $\pi_1, \cdots, \pi_n$ be a sampling distribution on the data points $(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_n, y_n)$. Using $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_n)$, we draw a subsample $(\boldsymbol{x}_1^*, y_1^*), \cdots, (\boldsymbol{x}_r^*, y_r^*)$ with $r << n$ from the full sample so that $y_j^* = \boldsymbol{x}_j^{*T} \boldsymbol{\beta}_0 + R_j^*$, where $\boldsymbol{\beta}_0$ is the true parameter. Let $\boldsymbol{\pi}^* = (\pi_1^*, \cdots, \pi_r^*)$ be the corresponding sampling probabilities. We now approximate $\hat{\boldsymbol{\beta}}$ by the sampling estimator $\hat{\boldsymbol{\beta}}^*$ which solves the equation

$$\sum_{j=1}^{r} \frac{1}{\pi_j^*} \boldsymbol{x}_j^* \psi \left(\frac{y_j^* - \boldsymbol{x}_j^{*T} \boldsymbol{\beta}}{\sigma} \right) = 0$$

where σ is some nuisance parameter, which can be estimated before/after the estimate of $\boldsymbol{\beta}$ is obtained. For convenience, we assume $\sigma = 1$. We are interested with robust estimation of $\boldsymbol{\beta}_0$. Let ψ be a function on reals. Our choice of ψ includes the commonly used Huber function. So it is not differentiable in general.

Portnoy(1984) established the consistency results for growing dimension p of parameter $\boldsymbol{\beta}$ when ψ is not differentiable. He obtained the results under the growth condition that $p \to \infty$ but $p \log p / n \to 0$, a weaker condition than $p^2/n \to \infty$ in literature. We shall

develop our consistency result using his framework. Below we quote his conditions P1-P2 and X1-X4. To this end, following Portnoy's assumption. We assume $\boldsymbol{\beta_0} = \mathbf{0}$.Then $y_i = R_i (i = 1, 2, \cdots, n)$.

For conditions X1 and X2, let

$$I(\boldsymbol{\theta}, c) = \left\{ i = 1, 2, \cdots, n : |\boldsymbol{x}_i^T \boldsymbol{\theta}| \leq c \right\}$$

and let $\mathcal{N}(\delta) = \{\boldsymbol{x} \in R^p : \|\boldsymbol{x}\| \leq \delta\}$ and $\mathcal{S} = \{\boldsymbol{x} \in R^p : \|\boldsymbol{x}\| = 1\}$.

P1: ψ is an absolutely continuous function with ψ' bounded satisfying $\mathbb{E}\psi(R) = 0, \mathbb{E}\psi'(R) > 0$, and $\mathbb{E}\psi^2(R) \leq B < +\infty$. Let c be a constant and define for r real

$$H(c; r) = \inf \left\{ \psi'(r - v) : |v| \leq c \right\} \tag{3.1}$$

P2: There exist positive constants b and c such that $H(c; \cdot)$ is measurable (hence, $H_i(c) \equiv H(c; R_i)$ is a random variable) and $\mathbb{E}H_i(c) \geq b$.

X1: For any constant $c > 0$, there are positive constants a, δ, and C such that for all $\boldsymbol{\beta} \in \mathcal{N}, \boldsymbol{\theta} \in \mathcal{S}$, and $n = 1, 2, \cdots$

$$\sum_{i \in J} \left(\boldsymbol{x}_i^T \boldsymbol{\theta} \right)^2 \geq an$$

where $J = I(\boldsymbol{\beta}, c) \cap I(\boldsymbol{\theta}, C)$.

X2: For any $c > 0$ and $\varepsilon > 0$ there are constants $\delta' > 0$, and $C > 0$ such that for all $\boldsymbol{\beta} \in \mathcal{N}, \boldsymbol{\theta} \in \mathcal{S}$, and $n = 1, 2, \cdots$

$$\sum_{i \notin J} \left(\boldsymbol{x}_i^T \boldsymbol{\theta} \right)^2 \leq \varepsilon n$$

where $J = I(\boldsymbol{\beta}, c) \cap I(\boldsymbol{\theta}, C)$.

X3: There exists a constant B such that for $n = 1, 2, \cdots$

$$\max \left\{ ||\boldsymbol{x}_i||^2, \quad i = 1, 2, \cdots, n \right\} \leq B.$$

X4: There exists a constant B such that for $i = 1, 2, \cdots$

$$\sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2 \leq Bpn.$$

We need the following A1-A3.

A1: For sampling distribution $\{\boldsymbol{\pi}\}$, there exists a constant ν_0 such that $\nu_n = \min_{1\le i\le n}(n\pi_i) \ge \nu_0 > 0$ uniformly in $n = 1, 2, \cdots$

A2: Let

$$b_n = \frac{1}{rn}\sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2 \left(\int_0^{\boldsymbol{x}_i^T\boldsymbol{\beta}} \psi'\left(R_i - v\right) dv\right)^2 = O_P(1).$$

Remark 3.1. A sufficient condition for A2 is that ψ' is bounded and that

$$\frac{1}{rn}\sum_{i=1}^{n} \|\boldsymbol{x}_i\|^4 = O_P(1).$$

A3: For $\delta > 0$,

$$\sup_{\|\boldsymbol{\beta}\|\le\delta}\sup_{\|\boldsymbol{y}\|=1} \frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{\pi_i}(\boldsymbol{x}_i^T\boldsymbol{y})^4\psi'^2\left(R_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right) = O_P(1).$$

Remark 3.2. A sufficient condition for A3 is that ψ' is bounded and that

$$\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|\boldsymbol{x}_i\|^4}{\pi_i} = O_P(1).$$

Definition 3.3. A sequence of events $\{A_r^*, r = 1, 2, \cdots\}$ satisfies

$$P^*(A_r^*) = 1 + o_P(1)$$

as $r \to \infty$ if for $\forall\, \epsilon > 0$, $\forall\, \eta > 0$, $\exists\, r \ge r_0$ s.t.

$$P(P^*(A_r^*) > 1 - \epsilon) > 1 - \eta.$$

Below is the Theorem 3.2 from Portnoy's paper (1984).

Theorem. *Assume conditions* P1, P2, X1, X2, X3, X4, *and that* $(p\log n)/n \to 0$. *Let* $\Psi_n : R^p \to R^p$ *be defined by*

$$\Psi_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\psi\left(R_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right). \tag{3.2}$$

Then there is a root $\hat{\boldsymbol{\beta}}$ *of the equation* $\Psi_n(\boldsymbol{\beta}) = 0$ *satisfying*

$$\|\hat{\boldsymbol{\beta}}\|^2 = \mathscr{O}_p(p/n).$$

We have similar consistency result for subsample estimator $\hat{\boldsymbol{\beta}}^*$ below. In this chapter, the same B may denote different constants.

Theorem 3.4. *Assume conditions* $\mathrm{P1}, \mathrm{P2}, \mathrm{X1}, \mathrm{X2}, \mathrm{X3}$*,* $\mathrm{X4}$*,* $\mathrm{A1}$ *and* $\mathrm{A2}$*, and that* $(p \log n)/n \to 0$*. Let* $\Psi_r^* : R^p \to R^p$ *be defined by*

$$\Psi_r^*(\boldsymbol{\beta}) = \frac{1}{r} \sum_{j=1}^{r} \frac{1}{n\pi_j^*} \boldsymbol{x}_j^* \psi \left(R_j^* - \boldsymbol{x}_j^{*T} \boldsymbol{\beta} \right). \tag{3.3}$$

If $p/r \to 0$*, then there exists a root* $\hat{\boldsymbol{\beta}}^*$ *of the equation* $\Psi_r^*(\boldsymbol{\beta}) = 0$ *satisfying*

$$\|\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\|^2 = O_P(p/r).$$

Proof. By result 6.3.4 of Ortega and Rheinholdt (1970, page 163), it suffices to show that $r\boldsymbol{\beta}^T \Psi_r^*(\boldsymbol{\beta}) < 0$ for $\|\boldsymbol{\beta}\|^2 = Bp/r$ in probability where B is a constant. Noting

$$\begin{aligned} r\boldsymbol{\beta}^T \Psi_r^*(\boldsymbol{\beta}) &= \sum_{j=1}^{r} \left(\boldsymbol{x}_j^{*T} \boldsymbol{\beta} \right) \frac{1}{n\pi_j^*} \psi \left(R_j^* - \boldsymbol{x}_j^{*T} \boldsymbol{\beta} \right) \\ &= \sum_{j=1}^{r} \left(\boldsymbol{x}_j^{*T} \boldsymbol{\beta} \right) \frac{1}{n\pi_j^*} \psi \left(R_j^* \right) - \sum_{j=1}^{r} \left(\boldsymbol{x}_j^{*T} \boldsymbol{\beta} \right) \frac{1}{n\pi_j^*} \int_0^{\boldsymbol{x}_j^{*T} \boldsymbol{\beta}} \psi' \left(R_j^* - v \right) dv. \\ &=: A_1^* - A_2^*. \end{aligned}$$

We have

$$|A_1^*| \leq \|\beta\| \left\| \sum_{j=1}^{r} \frac{1}{n\pi_j^*} \boldsymbol{x}_j^{*T} \psi \left(R_j^* \right) \right\|,$$

and

$$
\begin{aligned}
&\mathbb{E}^* \left\| \sum_{j=1}^{r} \frac{1}{n\pi_j^*} \boldsymbol{x}_j^{*T} \psi\left(R_j^*\right) \right\|^2 \\
=&\sum_{i=1}^{r} \sum_{j=1}^{r} \mathbb{E}^* \frac{\boldsymbol{x}_i^{*T} \boldsymbol{x}_j^*}{n^2 \pi_i^* \pi_j^*} \psi\left(R_i^*\right) \psi\left(R_j^*\right) \\
=&\sum_{i=1}^{r} \mathbb{E}^* \frac{\left\|\boldsymbol{x}_i^*\right\|^2}{\left(n\pi_i^*\right)^2} \psi^2\left(R_i^*\right) + \sum_{i \neq j} \mathbb{E}^* \frac{\boldsymbol{x}_i^{*T} \boldsymbol{x}_j^*}{n^2 \pi_i^* \pi_j^*} \psi\left(R_i^*\right) \psi\left(R_j^*\right) \\
=&\frac{r}{n^2} \sum_{i=1}^{n} \frac{1}{\pi_i} \|\boldsymbol{x}_i\|^2 \psi^2\left(R_i\right) + \frac{r(r-1)}{n^2} \mathbb{E}^* \frac{\boldsymbol{x}_1^{*T} \boldsymbol{x}_2^*}{\pi_1^* \pi_2^*} \psi\left(R_1^*\right) \psi\left(R_2^*\right) \\
=&\frac{r}{n^2} \sum_{i=1}^{n} \frac{1}{\pi_i} \|\boldsymbol{x}_i\|^2 \psi^2\left(R_i\right) + \frac{r(r-1)}{n^2} \left[\mathbb{E}^* \frac{\boldsymbol{x}_1^*}{\pi_1^*} \psi\left(R_1^*\right)\right]^T \left[\mathbb{E}^* \frac{\boldsymbol{x}_2^*}{\pi_2^*} \psi\left(R_2^*\right)\right] \\
=&\frac{r}{n^2} \sum_{i=1}^{n} \frac{1}{\pi_i} \|\boldsymbol{x}_i\|^2 \psi^2\left(R_i\right) + \frac{r(r-1)}{n^2} \left\| \sum_{i=1}^{n} \boldsymbol{x}_i \psi(R_i) \right\|^2 \\
=:&A_{11} + A_{12}.
\end{aligned}
$$

It follows that

$$\mathbb{E}^* \left|A_1^*\right|^2 \leq \|\beta\|^2 (A_{11} + A_{12}).$$

Since $\nu_n = \min(n\pi_i) \geq \nu_0 > 0$ by assumption A1, then

$$A_{11} \leq \frac{r}{n} \nu_n^{-1} \sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2 \psi^2\left(R_i\right).$$

By assumptions P1 and X4, we know

$$\sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2 \psi^2\left(R_i\right) = O_P(pn).$$

Therefore,

$$A_{11} = \frac{r}{n} \nu_n^{-1} O_P(pn) = O_P\left(\frac{rp}{\nu_n}\right).$$

By X4,

$$\mathbb{E} \left\| \sum_{i=1}^{n} \boldsymbol{x}_i \psi(R_i) \right\|^2 = \sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2 \mathbb{E}\psi^2(R_i) \leq B^2 np.$$

So,

$$A_{12} = \frac{r(r-1)}{n^2} O_P(pn) = O_P\left(\frac{r^2 p}{n}\right).$$

Fix $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\| \neq 0$. Using Chebychev's inequality,

$$\begin{aligned}
& P^*\left\{A_1^* \geq B\sqrt{rp}\|\boldsymbol{\beta}\|\right\} \\
\leq & \frac{\mathbb{E}^*\left|A_1^*\right|^2}{(B\sqrt{rp}\|\boldsymbol{\beta}\|)^2} \leq \frac{1}{B^2 rp}(A_{11} + A_{12}) \\
= & \frac{1}{B^2 rp}\left(O_P\left(\frac{rp}{\nu_n}\right) + O_P\left(\frac{r^2 p}{n}\right)\right) \\
= & \frac{1}{B^2}\left(O_P\left(\frac{1}{\nu_n}\right) + O_P\left(\frac{r}{n}\right)\right) \to 0 \quad \text{as} \quad B \to \infty.
\end{aligned}$$

Then we have that for any $\varepsilon > 0$, there exists a constant B such that for all r

$$P^*\left\{A_1^* \leq B\sqrt{rp}\|\boldsymbol{\beta}\|\right\} = 1 - o_P(1). \tag{3.4}$$

By the definition of A_2^*,

$$\begin{aligned}
\mathbb{E}^* A_2^* = & \mathbb{E}^* \sum_{j=1}^{r} \left(\boldsymbol{x}_j^{*T}\boldsymbol{\beta}\right) \frac{1}{n\pi_j^*} \int_0^{\boldsymbol{x}_j^{*T}\boldsymbol{\beta}} \psi'\left(R_j^* - v\right) dv \\
= & \frac{r}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right) \int_0^{\boldsymbol{x}_i^T\boldsymbol{\beta}} \psi'\left(R_i - v\right) dv \\
=: & \frac{r}{n} A_2.
\end{aligned}$$

By (3.8) in Portnoy (page 1303), there is an event E_n with $P(E_n) \to 1$ such that on E_n,

$$A_2 \geq a_0 n\|\boldsymbol{\beta}\|^2,$$

where a_0 is a positive constant. Hence,

$$\mathbb{E}^* A_2^* = \frac{r}{n} A_2 \geq a_0 r\|\boldsymbol{\beta}\|^2. \tag{3.5}$$

Suppress $\boldsymbol{\beta}$ and let

$$a_j^* = \left(\boldsymbol{x}_j^{*T}\boldsymbol{\beta}\right) \frac{1}{n\pi_j^*} \int_0^{\boldsymbol{x}_j^{*T}\boldsymbol{\beta}} \psi'\left(R_j^* - v\right) dv.$$

Then $A_2^* = \sum_{j=1}^r a_j^*$. $a_1^*, \cdots, a_j^*$ are i.i.d and

$$\begin{aligned}\mathbb{E}^* a_1^{*2} =& \frac{1}{n^2}\sum_{i=1}^n \frac{\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2}{\pi_i}\left(\int_0^{\boldsymbol{x}_i^T\boldsymbol{\beta}} \psi'\left(R_i - v\right) dv\right)^2 \\ \leq& \|\boldsymbol{\beta}\|^2 \nu_n^{-1} \frac{1}{n}\sum_{i=1}^n \|\boldsymbol{x}_i\|^2 \left(\int_0^{\boldsymbol{x}_i^T\boldsymbol{\beta}} \psi'\left(R_i - v\right) dv\right)^2 \\ =& \|\boldsymbol{\beta}\|^2 \nu_n^{-1} r b_n,\end{aligned}$$

where b_n is given in A2. Hence,

$$\mathbb{E}^*|A_2^* - \mathbb{E}^* A_2^*|^2 = \mathbb{E}^*\left|\sum_{j=1}^r (a_j^* - \mathbb{E}^* a_j^*)\right|^2 = r\mathbb{E}^*(a_1^* - \mathbb{E}^* a_1^*)^2 \leq r\mathbb{E}^* a_1^{*2} = r^2\|\boldsymbol{\beta}\|^2\nu_n^{-1} b_n,$$

By A1 and A2,

$$P^*\left(\frac{|A_2^* - \mathbb{E}^* A_2^*|}{r} \geq M\right) \leq \frac{\mathbb{E}^*|A_2^* - \mathbb{E}^* A_2^*|^2}{r^2 M^2} \leq \frac{\|\boldsymbol{\beta}\|^2\nu_n^{-1} b_n}{M^2} = O_P\left(\frac{1}{M^2}\right) \to 0, \; as \, M \to \infty.$$

Since (3.5) and

$$\frac{A_2^*}{r} \geq \frac{\mathbb{E}^* A_2^*}{r} - \frac{|A_2^* - \mathbb{E}^* A_2^*|}{r}$$

on E_n, it follows that there exists some constant a and event E_n' with $P(E_n') \to 1$ such that on E_n'

$$\frac{A_2^*}{r} \geq a\|\boldsymbol{\beta}\|^2$$

for all $\boldsymbol{\beta}$ with $\|\boldsymbol{\beta}\| \leq \delta$. Thus, there is N such that for $r \geq N$ on E_n'

$$P^*\left\{A_1^* - A_2^* \leq B\sqrt{rp}\|\boldsymbol{\beta}\| - ar\|\boldsymbol{\beta}\|^2 \text{ for all } \boldsymbol{\beta} \text{ with } \|\boldsymbol{\beta}\| \leq \delta\right\} \geq 1 - 2\varepsilon$$

Choose $N' > N$ so that $Bp/r \leq \delta^2$ when $r \geq N'$. Let $B_0 = a\sqrt{B}/2$.Then

$$B_0\sqrt{rp}\|\boldsymbol{\beta}\| - ar\|\boldsymbol{\beta}\|^2 = \frac{aBp}{2} - aBp = -\frac{aBp}{2} < 0.$$

It follows that for $r \geq N'$

$$\begin{aligned}&P^*\left\{r\boldsymbol{\beta}^T\Psi_r^*(\boldsymbol{\beta}) < 0 \text{ for all } \boldsymbol{\beta} \text{ with } \|\boldsymbol{\beta}\|^2 = Bp/r\right\} \\ \geq& P^*\left\{A_1^* - A_2^* \leq -1/2Bap \text{ for all } \beta \text{ with } \|\boldsymbol{\beta}\|^2 = Bp/r\right\} \geq 1 - 2\varepsilon.\end{aligned}$$

Hence, according to the result in Ortega and Rheinboldt, we have

$$\|\hat{\boldsymbol{\beta}}^*\|^2 = O_P(p/r)$$

By Theorem 3.2 of Portnoy's result, there exists event E_n'' with $P(E_n'') \to 1$ such that on E_n''

$$\|\hat{\boldsymbol{\beta}}\|^2 = O_P(p/n).$$

Then on $E_n' \cap E_n''$,

$$\|\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\| \leq \|\hat{\boldsymbol{\beta}}^*\| + \|\hat{\boldsymbol{\beta}}\| = O_P(\sqrt{p/r}).$$

$$\|\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\|^2 = O_P(p/r).$$

□

We cite Portnoy's result below (Corollary 3.3 in Portnoy's paper(1984)).

Corollary. *Under the hypotheses of Theorem 3.2(in Portnoy's paper), $\hat{\boldsymbol{\beta}}$ is unique on $\{\|\boldsymbol{\beta}\| : \|\boldsymbol{\beta}\| \leq \delta\}$ in probability. If in addition ψ' is nonnegative (everywhere), then $\hat{\boldsymbol{\beta}}$ is unique on R^p in probability.*

We have similar result for subampling method below.

Corollary 3.5. *Under the hypotheses of Theorem 3.2, $\hat{\boldsymbol{\beta}}^*$ is unique on $\{\|\boldsymbol{\beta}\| : \|\boldsymbol{\beta}\| \leq \delta\}$ in probability. If in addition ψ' is nonnegative (everywhere), then $\hat{\boldsymbol{\beta}}^*$ is unique on R^p inprobability.*

Proof. Let

$$\boldsymbol{F}^*(\boldsymbol{\beta}) = n\Psi_r^*(\boldsymbol{\beta}),$$

then

$$\boldsymbol{F}^*(\boldsymbol{\beta}) = \sum_{j=1}^{r} \frac{1}{r\pi_j^*} \boldsymbol{x}_j^* \psi\left(R_j^* - \boldsymbol{x}_j^{*T}\boldsymbol{\beta}\right).$$

Then the $p \times p$ derivative matrix $\dot{\boldsymbol{F}}^*(\boldsymbol{\beta})$ satisfies

$$\dot{\boldsymbol{F}}^*(\boldsymbol{\beta}) = -\sum_{j=1}^{r} \frac{1}{r\pi_j^*} \boldsymbol{x}_j^* \boldsymbol{x}_j^{*T} \psi'\left(R_j^* - \boldsymbol{x}_j^{*T}\boldsymbol{\beta}\right),$$

Thus, for any $\boldsymbol{y} \in R^p$ with $\boldsymbol{y} \neq 0$,

$$\boldsymbol{y}^T \dot{\boldsymbol{F}}^*(\boldsymbol{\beta})\boldsymbol{y} = -\sum_{j=1}^{r} \frac{1}{r\pi_j^*}(\boldsymbol{x}_j^{*T}\boldsymbol{y})^2 \psi'\left(R_j^* - \boldsymbol{x}_j^{*T}\boldsymbol{\beta}\right) := -\sum_{j=1}^{r} \frac{1}{r} z_j^*,$$

where

$$z_j^* = \frac{1}{\pi_j^*}(\boldsymbol{x}_j^{*T}\boldsymbol{y})^2 \psi'\left(R_j^* - \boldsymbol{x}_j^{*T}\boldsymbol{\beta}\right).$$

Accordingly,

$$z_i = \frac{1}{\pi_i}(\boldsymbol{x}_i^{T}\boldsymbol{y})^2 \psi'\left(R_i - \boldsymbol{x}_i^{T}\boldsymbol{\beta}\right).$$

$$\mathbb{E}^* z_j^* = \mathbb{E}^* z_1^* = -\sum_{i=1}^{n} (\boldsymbol{x}_i^T\boldsymbol{y})^2 \psi'\left(R_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right) = \boldsymbol{y}^T \dot{\boldsymbol{F}}(\boldsymbol{\beta})\boldsymbol{y}.$$

Now $\dot{\boldsymbol{F}}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T \psi'\left(R_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right)$. Then by assumption A3, we have

$$\begin{aligned}
&\frac{1}{n^2}\mathbb{E}^*\left(\boldsymbol{y}^T \dot{\boldsymbol{F}}^*(\boldsymbol{\beta})\boldsymbol{y} - \boldsymbol{y}^T \dot{\boldsymbol{F}}(\boldsymbol{\beta})\boldsymbol{y}\right)^2 \\
=&\frac{1}{n^2}\mathbb{E}^*\left\{\frac{1}{r}\sum_{j=1}^{r}[z_j^* - \mathbb{E}^* z_j^*]\right\}^2 \\
=&\frac{1}{rn^2}\sum_{i=1}^{n}(z_i - \boldsymbol{y}^T \dot{\boldsymbol{F}}(\boldsymbol{\beta})\boldsymbol{y})^2 \pi_i \\
\leq&\frac{1}{rn^2}\sum_{i=1}^{n}\pi_i z_i^2 = \frac{1}{rn^2}\sum_{i=1}^{n}\frac{1}{\pi_i}(\boldsymbol{x}_i^T\boldsymbol{y})^4 \psi'^2\left(R_i - \boldsymbol{x}_i^T\boldsymbol{\beta}\right) \\
=&o_P(1).
\end{aligned}$$

It follows that

$$\boldsymbol{y}^T \dot{\Psi}_r^*(\boldsymbol{\beta})\boldsymbol{y} - \boldsymbol{y}^T \dot{\Psi}_n(\boldsymbol{\beta})\boldsymbol{y} = \frac{1}{n}\boldsymbol{y}^T \dot{\boldsymbol{F}}^*(\boldsymbol{\beta})\boldsymbol{y} - \frac{1}{n}\boldsymbol{y}^T \dot{\boldsymbol{F}}(\boldsymbol{\beta})\boldsymbol{y} = o_P(1).$$

From the proof of Corollary 3.3 of Portnoy (1984), $\dot{\boldsymbol{F}}(\boldsymbol{\beta})$ is strictly negative definite on $\{\|\boldsymbol{\beta}\| : \|\boldsymbol{\beta}\| \leq \delta\}$. For any $\boldsymbol{y} \neq 0$, we have $\boldsymbol{y}^T \dot{\Psi}_n(\boldsymbol{\beta})\boldsymbol{y} = \frac{1}{n}\boldsymbol{y}^T \dot{\boldsymbol{F}}(\boldsymbol{\beta})\boldsymbol{y} < 0$. Since

$$\boldsymbol{y}^T \dot{\Psi}_r^*(\boldsymbol{\beta})\boldsymbol{y} \to \boldsymbol{y}^T \dot{\Psi}_n(\boldsymbol{\beta})\boldsymbol{y} \quad \text{as} \quad r \to \infty.$$

It follows that

$$\boldsymbol{y}^T \dot{\Psi}_r^*(\boldsymbol{\beta})\boldsymbol{y} < 0.$$

Hence, $\dot{\Psi}_r^*(\boldsymbol{\beta})$ is strictly negative definite on some $\{\|\boldsymbol{\beta}\| : \|\boldsymbol{\beta}\| \leq \delta\}$. Thus, $\hat{\boldsymbol{\beta}}_r^*$ is unique on this set. If ψ' is nonnegative, then $\dot{\Psi}_r^*(\boldsymbol{\beta})$ is nonpositive definite everywhere. Since it is negative definite on a neighborhood of $\hat{\boldsymbol{\beta}}_r^*$) $\hat{\boldsymbol{\beta}}_r^*$ is unique on R^p.

□

Remark 3.6. When $\{\boldsymbol{x}_i\}$ are i.i.d., Portnoy(1984) showed that conditions X1, X2, X3, and X4 hold in probability when below conditions are satisfied.

(1) Assume $(p \log r)/r \to 0$.

(2) $\mathbb{E}_X x_{ij}^2 \leq B_0 < \infty$ (for all $i = 1, \cdots, n$ and $j = 1, \cdots, p$).

(3) For conditions X1 and X2 define

$$U_i(c, C) = \mathbf{1}\left(|\boldsymbol{x}_i^T\boldsymbol{\beta}| \leq c, |\boldsymbol{x}_i^T\boldsymbol{\theta}| \leq C\right)\left(\boldsymbol{x}_i^T\boldsymbol{\theta}\right)^2$$

For any positive constants c and ε, there exist positive constants δ and C such that for all $\boldsymbol{\beta} \in \mathscr{S}$ and $\boldsymbol{\theta} \in \mathscr{S}^*$,

$$\mathbb{E}U_i(c, C) \geq 1 - \varepsilon.$$

3.2 Asymptotic Normality

In this section, we will consider the situation in which dimension p is fixed. We further assume $X_1, \cdots, X_n$ are i.i.d random vectors with $\mathbb{E}\|X_i\|^4 < \infty$, although the result will typically hold for non-random vectors. We shall use capital $X_i (i = 1, \cdots, n)$ for $\boldsymbol{x}_i (i = 1, \cdots, n)$ to remind that X_i are i.i.d random vectors.

Let

$$\phi_{\boldsymbol{\beta}}(\boldsymbol{x}, y) = \psi(y - \boldsymbol{x}^T\boldsymbol{\beta})\boldsymbol{x} \quad \text{for} \quad \boldsymbol{x} \in \mathbb{R}^p, y \in \mathbb{R}.$$

Let $\mathbb{P}_n$ be the empirical measure with probability mass $\frac{1}{n}$ at (X_i, Y_i) for $i = 1, \cdots, n$. Then

$$\Psi_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \psi(Y_i - X_i^T\boldsymbol{\beta})X_i = \mathbb{P}_n\phi_{\boldsymbol{\beta}}.$$

Also denote $\Psi(\boldsymbol{\beta}) = \mathbb{E}\psi(Y_i - X_i^T\boldsymbol{\beta})X_i = \mathbb{P}\phi_{\boldsymbol{\beta}}$.

Let $\hat{\mathbb{P}}_n$ be the empirical measure with probability mass $\frac{1}{rn\pi_j^*}$ at (X_j^*, Y_j^*) for $j = 1, \cdots, r$. Let

$$\Psi_r^*(\boldsymbol{\beta}) = \frac{1}{r}\sum_{j=1}^{r}\frac{\psi(Y_j^* - X_j^{*T}\boldsymbol{\beta})X_j^*}{n\pi_j^*}.$$

Then $\Psi_r^*(\boldsymbol{\beta}) = \hat{\mathbb{P}}_n\phi_{\boldsymbol{\beta}}$ and $\mathbb{E}^*\Psi_r^*(\boldsymbol{\beta}) = \Psi_n(\boldsymbol{\beta}) = \mathbb{P}_n\phi_{\boldsymbol{\beta}}$.

Below A4 is part of P3 in Portnoy(1985).

A4: For u and v real, define

$$Q(u,v) = \frac{\psi(u) - \psi(u-v)}{v} - d. \tag{3.6}$$

where $d = \mathbb{E}\psi'(R)$. $Q(u,v) = 0$ when $v = 0$. Then $Q(u,v)$ is uniformly bounded.

Remark 3.7. If A4 holds, then there is a constant C such that for any $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^p$

$$|\psi(y - \boldsymbol{x}^T\boldsymbol{\beta}_1) - \psi(y - \boldsymbol{x}^T\boldsymbol{\beta}_2)| \le C\|\boldsymbol{x}\|\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|, \quad \boldsymbol{x} \in \mathbb{R}^p, y \in \mathbb{R}$$

A5: Assume $\mathbb{E}(\|\psi(Y - X^T\boldsymbol{\beta}_0)X\|^2) < \infty$ and that the map $\boldsymbol{\beta} \mapsto \mathbb{E}[\psi(Y - X^T\boldsymbol{\beta})X]$ is continuously differentiable at a zero $\boldsymbol{\beta}_0$, with nonsingular derivative matrix $D_{\boldsymbol{\beta}_0}$.

A6:

$$\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|X_i\|^4}{\pi_i} = O_P(1).$$

Remark 3.8. A6 is a sufficient condition for A3. Under assumption A1, if

$$\frac{1}{n}\sum_{i=1}^{n}\|X_i\|^4 = O_P(1),$$

then A6 exists.

A7: There exist constants b and B such that

$$\Sigma_n(\boldsymbol{\beta}) = \mathbb{E}^*\left(\frac{\psi^2(Y_i^* - X_i^{*T}\boldsymbol{\beta})X_i^*X_i^{*T}}{n^2\pi_j^{*2}}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\frac{\psi^2(Y_i - X_i^T\boldsymbol{\beta})}{\pi_i}X_iX_i^T$$

satisfies

$$0 \le b \le \lambda_{\min}(\Sigma_n(\boldsymbol{\beta}_0)) \le \lambda_{\max}(\Sigma_n(\boldsymbol{\beta}_0)) \le B < \infty, \quad \forall n.$$

A8:

$$\frac{\log r}{r\nu_n^2}\max_i\|X_i\|^4 = O_P(1).$$

By Theorem 5.21 in A.W. van der Vaart(1998), we have the following result.

Theorem 3.9. *Consider a neighborhood of $\boldsymbol{\beta}$, let $\psi(x)$ be a measurable vector-valued function. Assume A4 and A5. If an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ satisfy $\Psi_n(\hat{\boldsymbol{\beta}}) = o_P(n^{-\frac{1}{2}})$ and $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}_0 + o_P(1)$, then we have the expansion*

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = -D_{\boldsymbol{\beta}_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(Y_i - X_i^T \boldsymbol{\beta}_0) X_i + o_P(1).$$

As a result, $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ has a asymptotic normal distribution with mean zero and covariance matrix $D_{\boldsymbol{\beta}_0}^{-1} \Sigma D_{\boldsymbol{\beta}_0}^{-T}$, i.e.

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \Longrightarrow N(0, D_{\boldsymbol{\beta}_0}^{-1} \Sigma D_{\boldsymbol{\beta}_0}^{-T}),$$

where $\Sigma = \mathbb{E}(\psi^2(Y - X^T \boldsymbol{\beta}_0) X X^T)$.

Below is the corollary from Yurinskii on page 491. This is an extension of the vector Beinstein inequality.

Corollary (Yurinskii)**.** *Let $\xi_1, \cdots, \xi_n$ be independent random vectors with $\mathbb{E}\xi_i = 0$ for each i. Suppose there exist constants $b_1, \cdots, b_n$ and H such that*

$$\mathbb{E} \left\| \xi_i \right\|^m \leqslant \left(\frac{m!}{2} \right) b_i^2 H^{m-2}, \quad m = 2, 3, \cdots$$

Let $B_n^2 = b_1^2 + \cdots + b_n^2$. Then for $x \geq 0$,

$$\mathbf{P} \left\{ \left\| \xi_1 + \cdots + \xi_n \right\| \geqslant x B_n \right\} \leqslant 2 \exp \left\{ - \left(\frac{x^2}{2} \right) \left(1 + 1.62 \left(\frac{xH}{B_n} \right) \right)^{-1} \right\}.$$

Remark 3.10. Suppose $\mathbb{E}\xi_i = 0$ and $\|\xi_i\| \leq M, i = 1, \cdots, n$ for some constant $M > 0$. Then the conditions in Yurinskiio's corollary are satisfied with $b_i = H = M (i = 1, \cdots, n)$. In this case, for any $t > 0$,

$$\mathbf{P} \left\{ \left\| \sum_1^n \xi_i \right\| \geqslant t \right\} \leqslant 2 \exp \left\{ - \frac{\frac{1}{2} t^2}{nM^2 + 1.62 t M} \right\}. \tag{3.7}$$

Let $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbb{P})$ and $\hat{\mathbb{G}}_r = \sqrt{r}(\hat{\mathbb{P}}_n - \mathbb{P}_n)$. The following result constitutes the main part of our proof for asymptotic normality.

Lemma 3.11. *Assume $\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}} = O_P(r^{-\frac{1}{2}})$ and $\hat{\boldsymbol{\beta}} = O_P(n^{-\frac{1}{2}})$. Under conditions for the theorem, we have the following result $\hat{\mathbb{G}}_r \phi_{\hat{\boldsymbol{\beta}}^*} - \hat{\mathbb{G}}_r \phi_{\hat{\boldsymbol{\beta}}} = o_P(1)$.*

Proof. Let $\Delta^*(\boldsymbol{b}) = \hat{\mathbb{G}}_r(\phi_{\boldsymbol{b}} - \phi_{\hat{\boldsymbol{\beta}}})$.

Since $\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}} = O_P(r^{-\frac{1}{2}})$, it then suffices to show that for any $B > 0$, and $\epsilon > 0$,

$$P^*\left(\sup_{\|\boldsymbol{b}\| \leq \frac{B}{\sqrt{r}}} \|\Delta^*(\boldsymbol{b})\| \geq \epsilon\right) = o_P(1).$$

By definition of $\hat{\mathbb{G}}_r$,

$$\begin{aligned}
\hat{\mathbb{G}}_r \phi_{\hat{\boldsymbol{\beta}}^*} =& \sqrt{r}(\hat{\mathbb{P}}_n - \mathbb{P}_n)\phi_{\hat{\boldsymbol{\beta}}^*} \\
=& \sqrt{r}\left(\frac{1}{r}\sum_{j=1}^{r} \frac{\psi(Y_j^* - X_j^{*T}\hat{\boldsymbol{\beta}}^*)X_j^*}{n\pi_j^*} - \frac{1}{n}\sum_{i=1}^{n}\psi(Y_i - X_i^T\hat{\boldsymbol{\beta}}^*)X_i\right) \\
=& \sum_{j=1}^{r}\left(\frac{1}{\sqrt{r}}\frac{\psi(Y_j^* - X_j^{*T}\hat{\boldsymbol{\beta}}^*)X_j^*}{n\pi_j^*} - \frac{1}{\sqrt{r}}\frac{1}{n}\sum_{i=1}^{n}\psi(Y_i - X_i^T\hat{\boldsymbol{\beta}}^*)X_i\right).
\end{aligned}$$

Since $\hat{\boldsymbol{\beta}}^*$ is random due to sampling method, we consider function

$$\hat{\mathbb{G}}_r \phi_{\boldsymbol{b}} = \sum_{j=1}^{r}\left(\frac{1}{\sqrt{r}}\frac{\psi(Y_j^* - X_j^{*T}\boldsymbol{b})X_j^*}{n\pi_j^*} - \frac{1}{\sqrt{r}}\frac{1}{n}\sum_{i=1}^{n}\psi(Y_i - X_i^T\boldsymbol{b})X_i\right).$$

Let

$$\xi_j^*(\boldsymbol{b}) = \frac{(\psi(Y_j^* - X_j^{*T}\boldsymbol{b}) - \psi(Y_j^* - X_j^{*T}\hat{\boldsymbol{\beta}}))X_j^*}{\sqrt{r}n\pi_j^*}. \tag{3.8}$$

Then

$$\xi_i(\boldsymbol{b}) = \frac{(\psi(Y_i - X_i^T\boldsymbol{b}) - \psi(Y_i - X_i^T\hat{\boldsymbol{\beta}}))X_i}{\sqrt{r}n\pi_i}. \tag{3.9}$$

From (3.8), we get

$$\mathbb{E}^*\xi_j^*(\boldsymbol{b}) = \frac{1}{\sqrt{r}n}\sum_{i=1}^{n}(\psi(Y_i - X_i^T\boldsymbol{b}) - \psi(Y_i - X_i^T\hat{\boldsymbol{\beta}}))X_i.$$

Let $\tilde{\xi}_j^*(\boldsymbol{b}) = \xi_j^*(\boldsymbol{b}) - \mathbb{E}^*\xi_j^*(\boldsymbol{b})$. Then $\Delta^*(\boldsymbol{b})$ is a sum of conditionally i.i.d zero mean random vectors given the data, that is

$$\Delta^*(\boldsymbol{b}) = \sum_{j=1}^{r}\tilde{\xi}_j^*(\boldsymbol{b}).$$

Let

$$Q_i = \frac{(\psi(Y_i - X_i^T\boldsymbol{b}) - \psi(Y_i - X_i^T\hat{\boldsymbol{\beta}}))X_i}{X_i^T(\boldsymbol{b} - \hat{\boldsymbol{\beta}})} - d.$$

By (3.9), we get

$$\sqrt{r}n\pi_i\xi_i(\boldsymbol{b}) = (Q_i + d)X_iX_i^T(\boldsymbol{b} - \hat{\boldsymbol{\beta}})$$

From assumptions A1 and A4, Q_i is uniformly bounded by a constant, say q_0. Then we have

$$\|\xi_j^*(\boldsymbol{b})\| \leq \max_i \|\xi_i(\boldsymbol{b})\| \leq \max_i \frac{1}{\sqrt{r}n\pi_i}(q_0+d)\|X_i\|^2\|\boldsymbol{b}-\hat{\boldsymbol{\beta}}\| \leq \frac{1}{\sqrt{r}\nu_n}(q_0+d)\|\boldsymbol{b}-\hat{\boldsymbol{\beta}}\| \max_i \|X_i\|^2.$$

Then

$$\|\tilde{\xi}_j^*(\boldsymbol{b})\| \leq \frac{2}{\sqrt{r}\nu_n}(q_0 + d)\|\boldsymbol{b} - \hat{\boldsymbol{\beta}}\| \max_i \|X_i\|^2.$$

Since $\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}} = O_P(\frac{1}{\sqrt{r}})$, then for any $\epsilon > 0$, there exists a constant B such that $P^*(\|\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\| \leq \frac{B}{\sqrt{r}}) > 1 - \epsilon$. Hence,

$$P^*\left(\|\tilde{\xi}_j^*(\hat{\boldsymbol{\beta}}^*)\| \leq \frac{2B}{r\nu_n}(q_0 + d)\max_i \|X_i\|^2\right) > 1 - \epsilon.$$

Let B be an arbitrary fixed constant. Partition the ball $B_n = \{\boldsymbol{b} \in \mathbb{R}^p : \|\boldsymbol{b}\| \leq \frac{B}{\sqrt{r}}\}$ into cubes with equal sides of length $\frac{2B}{\sqrt{r}K}$. Then the ball can be covered by K^p cubes $C_k (k = 1, \cdots, K^p)$. Let $\boldsymbol{b}_k$ be a point in C_k. Then $\{\boldsymbol{b}_k\}$ consists a grid on B_n, and for $\boldsymbol{b} \in C_k$, $\|\boldsymbol{b} - \boldsymbol{b}_k\| \leq \sqrt{p}\frac{2B}{\sqrt{r}K}$. Then

$$\begin{aligned} P^*\left(\sup_{\|\boldsymbol{b}\| \leq \frac{B}{\sqrt{r}}} \|\Delta^*(\boldsymbol{b})\| \geq \epsilon\right) \leq & K^p P^*\left(\|\Delta^*(\boldsymbol{b}_k)\| \geq \frac{1}{2}\epsilon\right) \\ & + P^*\left(\max_k \sup_{C_k} \|\Delta^*(\boldsymbol{b}) - \Delta^*(\boldsymbol{b}_k)\| \geq \frac{1}{2}\epsilon\right) \end{aligned} \tag{3.10}$$

By Beinstein's inequality in (3.7), for any ϵ we have

$$\begin{aligned} P^*\left(\|\Delta^*(\boldsymbol{b}_k)\| \geq \frac{1}{2}\epsilon\right) = & P^*\left(\left\|\sum_{j=1}^r \tilde{\xi}_j^*(\boldsymbol{b}_k)\right\| \geq \frac{1}{2}\epsilon\right) \\ \leq & 2\exp\left\{-\frac{\frac{1}{8}\epsilon^2}{rM^2 + 0.81\epsilon M}\right\}, \end{aligned}$$

where $M = \frac{2B}{r\nu_n}(q_0 + d)\max_i \|X_i\|^2$.

Choose K such that $\log K / \log r \to 0$. Then by condition A8,

$$rM^2 \log K = 4B^2(q_0 + d)^2 \frac{\log K}{\log r}\frac{\log r}{r\nu_n^2}\max_i \|X_i\|^4 = o_P(1).$$

Hence,

$$\frac{\frac{1}{8}\epsilon^2}{rM^2+0.81\epsilon M} - p\log K = \log K\left(\frac{\epsilon^2}{8\log K(rM^2+0.81\epsilon M)} - p\right) \to +\infty.$$

Hence, as $K \to \infty$ but $\log K/\log r \to 0$,

$$P^*\left(\|\Delta^*(\boldsymbol{b}_k)\| \geq \frac{1}{2}\epsilon\right) = o_P(1).$$

Noting

$$\begin{aligned}
&\Delta^*(\boldsymbol{b}) - \Delta^*(\boldsymbol{b}_k) \\
=&\frac{1}{\sqrt{r}}\sum_{j=1}^{r}\left(\frac{(\psi(Y_j^* - X_j^{*T}\boldsymbol{b}) - \psi(Y_j^* - X_j^{*T}\boldsymbol{b}_k))X_j^*}{n\pi_j^*} - \frac{1}{n}\sum_{i=1}^{n}(\psi(Y_i - X_i^T\boldsymbol{b}) - \psi(Y_i - X_i^T\boldsymbol{b}_k))X_i\right),
\end{aligned}$$

we have

$$\begin{aligned}
\mathbb{E}^*\|\Delta^*(\boldsymbol{b}) - \Delta^*(\boldsymbol{b}_k)\|^2 \leq&\mathbb{E}^*\left\|\frac{(\psi(Y_1^* - X_1^{*T}\boldsymbol{b}) - \psi(Y_1^* - X_1^{*T}\boldsymbol{b}_k))X_1^*}{n\pi_1^*}\right\|^2 \\
\leq&\frac{1}{n^2}\sum_{i=1}^{n}\frac{\|(\psi(Y_i - X_i^T\boldsymbol{b}) - \psi(Y_i - X_i^T\boldsymbol{b}_k))X_i\|^2}{\pi_i} \\
\leq&\frac{\|\boldsymbol{b}-\boldsymbol{b}_k\|^2}{n^2}\sum_{i=1}^{n}\frac{C)\|X_i\|^4}{\pi_i} \\
\leq&\frac{4B^2p}{K^2rn^2}\sum_{i=1}^{n}\frac{C\|X_i\|^4}{\pi_i}.
\end{aligned}$$

By Markov's inequality, we have

$$\begin{aligned}
&P^*\Big(\max_k\sup_{C_k}\|\Delta^*(\boldsymbol{b}) - \Delta^*(\boldsymbol{b}_k)\| \geq \frac{1}{2}\epsilon\Big) \\
\leq&K^pP^*\Big(\sup_{C_k}\|\Delta^*(\boldsymbol{b}) - \Delta^*(\boldsymbol{b}_k)\| \geq \frac{1}{2}\epsilon\Big) \\
\leq&K^p\frac{4\mathbb{E}^*\|\Delta^*(\boldsymbol{b}_k) - \Delta^*(\boldsymbol{b}_k)\|^2}{\epsilon^2} \\
\leq&\frac{4K^{p-2}}{rn^2\epsilon^2}\sum_{i=1}^{n}\frac{C\|X_i\|^4}{\pi_i}.
\end{aligned}$$

From assumption A6, then we get

$$P^*\left(\max_k\sup_{C_k}\|\Delta^*(\boldsymbol{b}) - \Delta^*(\boldsymbol{b}_k)\| \geq \frac{1}{2}\epsilon\right) = o_P(1).$$

Hence, from (3.10), we proved

$$\Delta^*(\hat{\boldsymbol{\beta}}^*) = \hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}^*} - \hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}} = o_P(1).$$

□

Lemma 3.12. *Let* $\phi_{\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\beta}}} = [\psi(Y - X^T\hat{\boldsymbol{\beta}}^*) - \psi(Y - X^T\hat{\boldsymbol{\beta}})]X$. *Assume*

$$\|\hat{\boldsymbol{\beta}}\| = O_P(n^{-\frac{1}{2}}), \quad \|\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\| = O_P(r^{-\frac{1}{2}}). \tag{3.11}$$

Under conditions for the theorem, we have the following result

$$\Delta_n^* = \mathbb{P}_n\phi_{\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\beta}}} - \mathbb{E}\mathbb{P}_n\phi_{\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\beta}}} = o_P(r^{-\frac{1}{2}}).$$

Proof. Since (3.11), it suffices to show

$$P(\sup_{\|\boldsymbol{b}\le\frac{B}{\sqrt{r}}}\sup_{\|\boldsymbol{b}\le\frac{C}{\sqrt{r}}}\|\Delta_n(\boldsymbol{b},\boldsymbol{\beta}) \ge r^{-\frac{1}{2}}\epsilon\|) = o(1),$$

where

$$\Delta_n = \mathbb{P}_n\phi_{\boldsymbol{b},\boldsymbol{\beta}} - \mathbb{E}\mathbb{P}_n\phi_{\boldsymbol{b},\boldsymbol{\beta}}.$$

Let B, C be arbitrary fixed constants. Partition the ball $\mathbb{B}_r = \{\boldsymbol{b} \in \mathbb{R}^p : \|\boldsymbol{b}\| \le \frac{B}{\sqrt{r}}\}$ and $\mathbb{C}_n = \{\boldsymbol{\beta} \in \mathbb{R}^p : \|\boldsymbol{\beta}\| \le \frac{C}{\sqrt{r}}\}$ into cubes with equal sides of length $\frac{2B}{\sqrt{r}K}$ and $\frac{2C}{\sqrt{r}L}$, respectively. Then $\mathbb{B}_r$ and $\mathbb{C}_n$ can be covered by K^p and L^p cubes $B_k(k = 1, \cdots, K^p)$ and $C_l(l = 1, \cdots, L^p)$. Let $\boldsymbol{b}_k \in B_k$, $\boldsymbol{\beta}_l \in C_l$. Then $\{\boldsymbol{b}_k\}$ and $\{\boldsymbol{\beta}_l\}$ are grids on $\mathbb{B}_r$ and $\mathbb{C}_n$, and for $\boldsymbol{b} \in B_k$ and $\boldsymbol{\beta} \in C_l$,

$$\|\boldsymbol{b} - \boldsymbol{b}_k\| \le \sqrt{p}\frac{2B}{\sqrt{r}K}, \quad \|\boldsymbol{\beta} - \boldsymbol{\beta}_l\| \le \sqrt{p}\frac{2C}{\sqrt{n}L}.$$

Since

$$\|\Delta_n(\boldsymbol{b},\boldsymbol{\beta})\| \le \|\Delta_n(\boldsymbol{b}_k,\boldsymbol{\beta}_l)\| + \|\Delta_n(\boldsymbol{b},\boldsymbol{\beta}) - \Delta_n(\boldsymbol{b}_k,\boldsymbol{\beta}_l)\|,$$

for any $\epsilon > 0$, we have

$$
\begin{aligned}
& P(\sup_{\boldsymbol{b}\in B_k, \boldsymbol{\beta}\in C_l} \|\Delta_n(\boldsymbol{b},\boldsymbol{\beta})\| > \epsilon r^{-\frac{1}{2}}) \qquad (3.12)\\
\leq & P(\max_{k,l} \|\Delta_n(\boldsymbol{b}_k,\boldsymbol{\beta}_l)\| > \frac{1}{2}\epsilon r^{-\frac{1}{2}})\\
& + P(\max_{k,l} \sup_{\boldsymbol{b}_k\in B_k, \boldsymbol{\beta}_l\in C_l} \|\Delta_n(\boldsymbol{b},\boldsymbol{\beta}) - \Delta_n(\boldsymbol{b}_k,\boldsymbol{\beta}_l)\| > \frac{1}{2}\epsilon r^{-\frac{1}{2}})\\
\leq & \sum_k \sum_l P(\|\Delta_n(\boldsymbol{b}_k,\boldsymbol{\beta}_l)\| > \frac{1}{2}\epsilon r^{-\frac{1}{2}})\\
& + \sum_k \sum_l P(\sup_{\boldsymbol{b}_k\in B_k, \boldsymbol{\beta}_l\in C_l} \|\Delta_n(\boldsymbol{b},\boldsymbol{\beta}) - \Delta_n(\boldsymbol{b}_k,\boldsymbol{\beta}_l)\| > \frac{1}{2}\epsilon r^{-\frac{1}{2}}) \qquad (3.13)
\end{aligned}
$$

Suppress $\boldsymbol{b}_k$ and $\boldsymbol{\beta}_l$, let

$$\xi_i = [\psi(Y_i - X_i^T\boldsymbol{b}_k) - \psi(Y_i - X_i^T\boldsymbol{\beta}_l)]X_i.$$

Then

$$\Delta_n(\boldsymbol{b}_k,\boldsymbol{\beta}_l) = \frac{1}{n}\sum_{i=1}^n(\xi_i - \mathbb{E}\xi_i).$$

For any $\boldsymbol{x}$, we do truncation and define

$$X^{(m)} = X\mathbf{1}[\|X\| \leq m], \quad \bar{X}^{(m)} = X\mathbf{1}[\|X\| > m]$$

Then we have below decomposition $X = X^{(m)} + \bar{X}^{(m)}$ and $\xi_i = \xi_i^{(m)} + \bar{\xi}_i^{(m)}$. Then

$$\Delta_n(\boldsymbol{b}_k,\boldsymbol{\beta}_l,m) = \frac{1}{n}\sum_{i=1}^n(\xi_i^{(m)} - \mathbb{E}\xi_i^{(m)}).$$

Since

$$\|\xi_i^{(m)}\| \leq 2C\frac{B}{\sqrt{r}}\|X_i\|^2 \leq B\frac{m^2}{\sqrt{r}} =: M.$$

By Beinstein's inequality in (3.7), for any $\epsilon > 0$, choose $K = L$, we have

$$
\begin{aligned}
& \sum_k \sum_l P(\|\Delta_n(\boldsymbol{b}_k,\boldsymbol{\beta}_l,m)\| > \frac{1}{4}\epsilon r^{-\frac{1}{2}})\\
\leq & 2K^pL^p \exp\left\{-\frac{Bn\epsilon^2 r^{-1}}{(B_1m^2r^{-\frac{1}{2}})^2 + B_2m^2r^{-\frac{1}{2}}\epsilon r^{-\frac{1}{2}}}\right\}\\
= & 2\exp\left\{-\frac{Bn\epsilon^2}{B_1m^4 + B_2m^2\epsilon} + p\log K\right\}.
\end{aligned}
$$

If we choose $\frac{m^4}{n}\log K \to 0$, then

$$P(\|\Delta_n(\boldsymbol{b}_k, \boldsymbol{\beta}_l, m)\| > \frac{1}{4}\epsilon r^{-\frac{1}{2}}) \to 0. \tag{3.14}$$

By the definition, we have

$$\bar{\Delta}_n(\boldsymbol{b}_k, \boldsymbol{\beta}_l, m) = \frac{1}{n}\sum_{i=1}^{n}(\bar{\xi}_i^{(m)} - \mathbb{E}\bar{\xi}_i^{(m)}).$$

By assumption A4, then

$$\mathbb{E}\|\bar{\Delta}_n(\boldsymbol{b}_k, \boldsymbol{\beta}_l, m)\|^2 \leq \frac{1}{n}\mathbb{E}\|\bar{\xi}_i^{(m)} - \mathbb{E}\bar{\xi}_i^{(m)}\|^2 \leq \frac{B}{r^{\frac{1}{2}}n}\mathbb{E}\|\bar{X}^{(m)}\|^4.$$

By Markov's inequality,

$$P(\|\bar{\Delta}_n(\boldsymbol{b}_k, \boldsymbol{\beta}_l, m)\| > \frac{1}{4}\epsilon r^{-\frac{1}{2}}) \leq \frac{Br^{\frac{1}{2}}}{n\epsilon^2}\mathbb{E}\|X\|^4\mathbf{1}[\|X\| > m].$$

Since $\mathbb{E}\|X\|^4 < \infty$, we have

$$\mathbb{E}\|X\|^4\mathbf{1}[\|X\| > m] \to 0 \quad as \quad m \to \infty.$$

Then

$$P(\|\bar{\Delta}_n(\boldsymbol{b}_k, \boldsymbol{\beta}_l, m)\| > \frac{1}{4}\epsilon r^{-\frac{1}{2}}) \to 0 \quad as \quad m \to \infty. \tag{3.15}$$

Since we have below decomposition

$$\Delta_n(\boldsymbol{b}_k, \boldsymbol{\beta}_l) = \Delta_n(\boldsymbol{b}_k, \boldsymbol{\beta}_l, m) + \bar{\Delta}_n(\boldsymbol{b}_k, \boldsymbol{\beta}_l, m),$$

then

$$\begin{aligned} &P(\|\Delta_n(\boldsymbol{b}_k, \boldsymbol{\beta}_l)\| > \frac{1}{2}\epsilon r^{-\frac{1}{2}}) \\ \leq &P(\|\Delta_n(\boldsymbol{b}_k, \boldsymbol{\beta}_l, m)\| > \frac{1}{4}\epsilon r^{-\frac{1}{2}}) + P(\|\bar{\Delta}_n(\boldsymbol{b}_k, \boldsymbol{\beta}_l, m)\| > \frac{1}{4}\epsilon r^{-\frac{1}{2}}) \end{aligned} \tag{3.16}$$

If we choose m which satisfies $\frac{m^4}{n}\log K \to 0$ as $n \to \infty, m \to \infty, K \to \infty$, then we get

$$P(\|\Delta_n(\boldsymbol{b}_k, \boldsymbol{\beta}_l)\| > \frac{1}{2}\epsilon r^{-\frac{1}{2}}) \to 0. \tag{3.17}$$

We get

$$\|\Delta_n(\boldsymbol{b}, \boldsymbol{\beta})\| = o_P(r^{-\frac{1}{2}}) \quad as \quad \frac{m^4}{n}\log K \to 0, n \to \infty, m \to \infty. \tag{3.18}$$

Since $\boldsymbol{b} \in B_k, \boldsymbol{\beta} \in C_l$, if we choose $K = L$, then

$$\begin{aligned}\mathbb{E}\|\Delta_n(\boldsymbol{b},\boldsymbol{\beta}) - \Delta_n(\boldsymbol{b}_k,\boldsymbol{\beta}_l)\|^2 \leq & B\frac{1}{n}\left(\frac{B}{K\sqrt{r}} + \frac{C}{L\sqrt{n}}\right)^2 \mathbb{E}\|X_i\|^4 \\ \leq & \frac{B}{K^2}\frac{1}{rn}\mathbb{E}\|X_i\|^4\end{aligned}$$

By Markov's inequality, we have

$$\begin{aligned}&\sum_k \sum_l P(\sup_{\boldsymbol{b}_k \in B_k, \boldsymbol{\beta}_l \in C_l} \|\Delta_n(\boldsymbol{b},\boldsymbol{\beta}) - \Delta_n(\boldsymbol{b}_k,\boldsymbol{\beta}_l)\| > \frac{1}{2}\epsilon r^{-\frac{1}{2}}) \\ leq & \frac{\mathbb{E}\|\Delta_n(\boldsymbol{b},\boldsymbol{\beta}) - \Delta_n(\boldsymbol{b}_k,\boldsymbol{\beta}_l)\|^2}{\frac{1}{2}\epsilon r^{-\frac{1}{2}}} \\ \leq & \frac{K^{2p-2}B}{n\epsilon^2}\mathbb{E}\|X\|^4\end{aligned}$$

Choose k such that $\frac{K^{2p-2}}{n} \to 0$, then

$$\|\Delta_n(\boldsymbol{b},\boldsymbol{\beta}) - \Delta_n(\boldsymbol{b}_k,\boldsymbol{\beta}_l)\| \to 0 \quad as \quad n \to \infty.$$

Then we proved

$$P(\sup_{\boldsymbol{b} \in B_k, \boldsymbol{\beta} \in C_l} \|\Delta_n(\boldsymbol{b},\boldsymbol{\beta})\| > \epsilon r^{-\frac{1}{2}}) \to 0 \quad as \quad n \to \infty. \tag{3.19}$$

Hence,

$$\Delta_n^* = \mathbb{P}_n \phi_{\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\beta}}} - \mathbb{E}\mathbb{P}_n \phi_{\hat{\boldsymbol{\beta}}^*, \hat{\boldsymbol{\beta}}} = o_P(r^{-\frac{1}{2}}).$$

□

Lemma 3.13. *Under conditions for the theorem, we have*

$$\hat{\mathbb{G}}_r \phi_{\hat{\boldsymbol{\beta}}} = O_P(1).$$

Proof. By assumption A4,

$$|\psi(Y - X^T\hat{\boldsymbol{\beta}})| \leq |\psi(Y - X^T\boldsymbol{\beta}_0)| + C\|X\|\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|.$$

It follows that

$$|\psi(Y - X^T\hat{\boldsymbol{\beta}}) + \psi(Y - X^T\boldsymbol{\beta}_0)| \leq 2|\psi(Y - X^T\boldsymbol{\beta}_0)| + C\|X\|\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|. \tag{3.20}$$

By A4, (3.20) and Cauchy-Schwartz inequality, we get

$$
\begin{aligned}
&\|\Sigma_n(\hat{\boldsymbol{\beta}}) - \Sigma_n(\boldsymbol{\beta}_0)\| \\
\leq &\frac{1}{n^2}\sum_{i=1}^{n}\frac{|\psi^2(Y_i - X_i^T\hat{\boldsymbol{\beta}}) - \psi^2(Y_i - X_i^T\boldsymbol{\beta}_0)|}{\pi_i}\|X_i\|^2 \\
\leq &\frac{1}{n^2}\sum_{i=1}^{n}\frac{(2|\psi(Y_i - X_i^T\boldsymbol{\beta}_0)| + C\|X_i\|\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|)(C\|X_i\|\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|)}{\pi_i}\|X_i\|^2 \\
\leq &\frac{1}{n^2}\sum_{i=1}^{n}\frac{C\|X_i\|^2}{\sqrt{\pi_i}}\frac{(2|\psi(Y_i - X_i^T\boldsymbol{\beta}_0)| + C\|X_i\|\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|)\|X_i\|}{\sqrt{\pi_i}}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|) \\
\leq &\sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\frac{C\|X_i\|^4}{\pi_i}}\sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\frac{(2|\psi(Y_i - X_i^T\boldsymbol{\beta}_0)| + C\|X_i\|\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|)^2\|X_i\|^2}{\pi_i}}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|) \\
\leq &\sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\frac{C\|X_i\|^4}{\pi_i}}\sqrt{\frac{1}{n^2}\sum_{i=1}^{n}\frac{(4\psi^2(Y_i - X_i^T\boldsymbol{\beta}_0) + C\|X_i\|^2\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2)\|X_i\|^2}{\pi_i}}(\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|)
\end{aligned}
$$

From assumption A7,

$$
\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^{n}\frac{\psi^2(Y_i - X_i^T\boldsymbol{\beta})}{\pi_i}\|X_i\|^2 \\
=&tr\left(\frac{1}{n^2}\sum_{i=1}^{n}\frac{\psi^2(Y_i - X_i^T\boldsymbol{\beta})}{\pi_i}X_iX_i^T\right) \\
=&tr(\Sigma_n) \leq p\lambda_{\max}(\Sigma_n(\boldsymbol{\beta}_0)) \leq pB < \infty
\end{aligned} \tag{3.21}
$$

From assumption A6, (3.21) and the consistency of $\hat{\boldsymbol{\beta}}$, we have

$$
\begin{aligned}
&\frac{1}{n^2}\sum_{i=1}^{n}\frac{(4\psi^2(Y_i - X_i^T\boldsymbol{\beta}_0) + C\|X_i\|^2\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2)\|X_i\|^2}{\pi_i} \\
=&\frac{1}{n^2}\sum_{i=1}^{n}\frac{4\psi^2(Y_i - X_i^T\boldsymbol{\beta}_0)\|X_i\|^2}{\pi_i} + \frac{1}{n^2}\sum_{i=1}^{n}\frac{C\|X_i\|^4}{\pi_i}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2 \\
=&o_P(1).
\end{aligned}
$$

It follows that

$$\|\Sigma_n(\hat{\boldsymbol{\beta}}) - \Sigma_n(\boldsymbol{\beta}_0)\| = O_P(1).$$

Then there exist integer N and constant B such that if $n > N$ then

$$\lambda_{\max}(\Sigma_n(\hat{\boldsymbol{\beta}})) \leq B < \infty. \tag{3.22}$$

According to Markov's inequality and (3.22), we have

$$
\begin{aligned}
&P^*(\|\hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}}\| \geq M) \\
\leq &\frac{\mathbb{E}^*\|\hat{\mathbb{G}}_r(\psi(Y^* - X^{*T}\hat{\boldsymbol{\beta}})X^*)\|^2}{M^2} \\
= &\frac{tr\{\mathbb{E}^*(\hat{\mathbb{G}}_r(\psi(Y^* - X^{*T}\hat{\boldsymbol{\beta}})X^*))^T(\hat{\mathbb{G}}_r(\psi(Y^* - X^{*T}\hat{\boldsymbol{\beta}})X^*))\}}{M^2} \\
= &\frac{1}{M^2}tr\left\{\mathbb{E}^*\left(\frac{1}{r}\sum_{j=1}^{r}\frac{\psi(Y_j^* - X_j^{*T}\hat{\boldsymbol{\beta}}^*)X_j^*}{n\pi_j^*}\right)^T\left(\frac{1}{r}\sum_{j=1}^{r}\frac{\psi(Y_j^* - X_j^{*T}\hat{\boldsymbol{\beta}}^*)X_j^*}{n\pi_j^*}\right)\right\} \\
= &\frac{1}{M^2 r}\sum_{j=1}^{r}\mathbb{E}^*\left\|\frac{\psi(Y_j^* - X_j^{*T}\hat{\boldsymbol{\beta}}^*)X_j^*}{n\pi_j^*}\right\|^2 \\
= &\frac{1}{M^2 n^2}\sum_{i=1}^{n}\frac{1}{\pi_i}\|\psi(Y_i - X_i^T\hat{\boldsymbol{\beta}})X_i\|^2 \\
\leq &\frac{pB}{M^2}.
\end{aligned}
$$

Thus for fixed p,

$$\hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}} = O_P(1).$$

□

We have the following asymptotic normality result.

Theorem 3.14. *Consider a neighborhood of $\boldsymbol{\beta}$, let $\psi(x)$ be a measurable vector-valued function. Assume P1-P2, X1-X4 and A1-A7. Assume the map $\boldsymbol{\beta} \mapsto \mathbb{E}\left(\frac{\psi(Y_i^* - X_i^{*T}\boldsymbol{\beta})X_i^*}{n\pi_i^*}\right) = \mathbb{E}\psi(Y - X^T\boldsymbol{\beta})X =: \boldsymbol{\mu}(\boldsymbol{\beta}) \in \mathbb{R}^p$ is continuously differentiable at $\hat{\boldsymbol{\beta}}$ almost everywhere with respect to Lebesgue measure, with nonsingular derivative matrix $V_{\hat{\boldsymbol{\beta}}}$. If $\Psi_r^*(\hat{\boldsymbol{\beta}}^*) = o_P(r^{-\frac{1}{2}})$, $\hat{\boldsymbol{\beta}} = O_P(n^{-\frac{1}{2}})$ and $\hat{\boldsymbol{\beta}}^* = \hat{\boldsymbol{\beta}} + o_P(1)$, then*

$$V_{\hat{\boldsymbol{\beta}}}\sqrt{r}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) = -\frac{1}{\sqrt{r}}\sum_{j=1}^{r}\frac{\psi(Y_j^* - X_j^{*T}\hat{\boldsymbol{\beta}})X_j^*}{n\pi_j^*} + o_P(1).$$

As a consequence,

$$(V_{\hat{\boldsymbol{\beta}}}^{-1}\Sigma_n(\hat{\boldsymbol{\beta}})V_{\hat{\boldsymbol{\beta}}}^{-T})^{-\frac{1}{2}}\sqrt{r}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \Longrightarrow N(0, I), \quad \textit{in probability}.$$

Proof. For every $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ in a neighborhood of $\boldsymbol{\beta}_0$,

$$\left|\frac{\psi(Y - X^T\boldsymbol{\beta}_1)X}{n\pi} - \frac{\psi(Y - X^T\boldsymbol{\beta}_2)X}{n\pi}\right| \leq \frac{C\|X\|^2}{n\pi}\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|,$$

According to the assumption A6,

$$\mathbb{E}^*\left(\frac{\|X_i^*\|^2}{n\pi^*}\right)^2 = \frac{1}{n^2}\sum_{i=1}^{n}\frac{\|X_i\|^4}{\pi_i} = O_P(1).$$

By definition of $\hat{\mathbb{G}}_r$ and

$$\mathbb{P}_n\phi_{\hat{\boldsymbol{\beta}}} = \frac{1}{n}\sum_{i=1}^{n}\psi(Y_i - X_i^T\hat{\boldsymbol{\beta}})X_i = 0,$$

we can rewrite $\hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}^*}$ as follows

$$\begin{aligned}
\hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}^*} =& \sqrt{r}(\hat{\mathbb{P}}_n - \mathbb{P}_n)\phi_{\hat{\boldsymbol{\beta}}^*} \\
=& \sqrt{r}\left(\frac{1}{r}\sum_{j=1}^{r}\frac{\psi(Y_j^* - X_j^{*T}\hat{\boldsymbol{\beta}}^*)X_j^*}{n\pi_j^*} - \frac{1}{n}\sum_{i=1}^{n}\psi(Y_i - X_i^T\hat{\boldsymbol{\beta}}^*)X_i\right) \\
=& o_P(1) - \sqrt{r}\mathbb{P}_n\phi_{\hat{\boldsymbol{\beta}}^*} \\
=& \sqrt{r}\mathbb{P}_n(\phi_{\hat{\boldsymbol{\beta}}} - \phi_{\hat{\boldsymbol{\beta}}^*}) + o_P(1)
\end{aligned} \tag{3.23}$$

By Lemma 3.7, we have

$$\hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}^*} - \hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}} = o_P(1). \tag{3.24}$$

Combining (3.23) with (3.24), we obtain that

$$\hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}} + o_P(1) = \sqrt{r}\mathbb{P}_n(\phi_{\hat{\boldsymbol{\beta}}} - \phi_{\hat{\boldsymbol{\beta}}^*}) + o_P(1),$$

that is

$$\hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}} = \sqrt{r}\mathbb{P}_n(\phi_{\hat{\boldsymbol{\beta}}} - \phi_{\hat{\boldsymbol{\beta}}^*}) + o_P(1), \tag{3.25}$$

Denote

$$\mathbb{P}_n\phi_{\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\beta}}} = \mathbb{P}_n(\phi_{\hat{\boldsymbol{\beta}}^*} - \phi_{\hat{\boldsymbol{\beta}}}) = \frac{1}{n}\sum_{i=1}^{n}[\psi(Y_i - X_i^T\hat{\boldsymbol{\beta}}^*) - \psi(Y_i - X_i^T\hat{\boldsymbol{\beta}})]X_i.$$

Then equality (3.25) becomes

$$\hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}} = -\sqrt{r}\mathbb{P}_n\phi_{\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\beta}}} + o_P(1). \tag{3.26}$$

From Lemma 3.9,

$$\Delta_n^* = \mathbb{P}_n\phi_{\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\beta}}} - \mathbb{E}\mathbb{P}_n\phi_{\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\beta}}} = o_P(r^{-\frac{1}{2}}).$$

Then combine this with $\mathbb{P}_n\phi_{\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\beta}}} = (\mathbb{P}_n\phi_{\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\beta}}} - \mathbb{E}\mathbb{P}_n\phi_{\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\beta}}}) + \mathbb{E}\mathbb{P}_n\phi_{\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\beta}}}$. From (3.26), we get that

$$\hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}} = -\sqrt{r}\mathbb{E}\mathbb{P}_n\phi_{\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\beta}}} + o_P(1). \tag{3.27}$$

Since $\mathbb{E}\psi(Y - X^T\boldsymbol{\beta})X$ is differentiable at $\hat{\boldsymbol{\beta}}$, then let

$$V_{\boldsymbol{\beta}} = \frac{\partial}{\partial\boldsymbol{\beta}}(\mathbb{E}\psi(Y - X^T\boldsymbol{\beta})X).$$

By the continuity of the derivative, we have

$$\begin{aligned}
\mathbb{E}\mathbb{P}_n\phi_{\hat{\boldsymbol{\beta}}^*,\hat{\boldsymbol{\beta}}} =& \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}\psi(Y_i - X_i^T\hat{\boldsymbol{\beta}}^*)X_i - \mathbb{E}\psi(Y_i - X_i^T\hat{\boldsymbol{\beta}})X_i) \\
=& V_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) + (V_{\tilde{\boldsymbol{\beta}}^*} - V_{\hat{\boldsymbol{\beta}}})(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \\
=& V_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) + o_P(\|\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\|)
\end{aligned} \tag{3.28}$$

From (3.27) and (3.28), we find that

$$\hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}} = -\sqrt{r}(V_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) + o_P(\|\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\|)) + o_P(1). \tag{3.29}$$

From Lemma 3.10,

$$\hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}} = O_P(1).$$

It follows that

$$\sqrt{r}V_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) = o_P(\sqrt{r}\|\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\|) + O_P(1).$$

By the invertibility of $V_{\hat{\boldsymbol{\beta}}}$, we have

$$\sqrt{r}\|(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})\| \leq o_P(\sqrt{r}\|\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}\|) + O_P(1).$$

Then

$$\sqrt{r}\|(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})\|(1 - o_P(1)) = O_P(1).$$

$$\sqrt{r}\|(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}})\| = O_P(1). \tag{3.30}$$

This implies that $\hat{\boldsymbol{\beta}}^*$ is $\sqrt{r}$ consistent. Inserting (3.30) into (3.29), we obtain that

$$\sqrt{r}V_{\hat{\boldsymbol{\beta}}}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) = -\hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}} + o_P(1) = -\frac{1}{\sqrt{r}}\sum_{j=1}^{r}\frac{\psi(Y_j^* - X_j^{*T}\hat{\boldsymbol{\beta}})X_j^*}{n\pi_j^*} + o_P(1).$$

By simple calculation, we have

$$\mathbb{E}^*\hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}} = 0, \quad Var^*\hat{\mathbb{G}}_r\phi_{\hat{\boldsymbol{\beta}}} = \Sigma_n(\hat{\boldsymbol{\beta}}) = \frac{1}{n^2}\sum_{i=1}^{n}\frac{\psi^2(Y_i - X_i^T\hat{\boldsymbol{\beta}})}{\pi_i}X_iX_i^T.$$

Then

$$(V_{\hat{\boldsymbol{\beta}}}^{-1}\Sigma_n(\hat{\boldsymbol{\beta}})V_{\hat{\boldsymbol{\beta}}}^{-T})^{-\frac{1}{2}}\sqrt{r}(\hat{\boldsymbol{\beta}}^* - \hat{\boldsymbol{\beta}}) \Longrightarrow N(0, I) \quad \text{in probability,}$$

where

$$V_{\hat{\boldsymbol{\beta}}} = \frac{\partial}{\partial\boldsymbol{\beta}}\boldsymbol{\mu}(\boldsymbol{\beta}) = \frac{\partial}{\partial\boldsymbol{\beta}}\mathbb{E}(\psi(Y - X^T\boldsymbol{\beta})X).$$

□

Consider minimizing the trace of $\frac{1}{r}V_{\hat{\boldsymbol{\beta}}}^{-1}\Sigma_n(\hat{\boldsymbol{\beta}})V_{\hat{\boldsymbol{\beta}}}^{-T}$ which is the asymptotic covariance matrix of the subsampling estimator $\hat{\boldsymbol{\beta}}$, that is, we seek to minimize $\tau(\boldsymbol{\pi}) = tr\left(\frac{1}{r}V_{\hat{\boldsymbol{\beta}}}^{-1}\Sigma_n(\hat{\boldsymbol{\beta}})V_{\hat{\boldsymbol{\beta}}}^{-T}\right)$ over all samping distribution $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_n)$ on the data points. This is referred to as A-optimality in the literature. Using algebra, we get

$$\begin{aligned}
&\tau(\boldsymbol{\pi}) = Tr\left(\frac{1}{r}V_{\hat{\boldsymbol{\beta}}}^{-1}\Sigma_n(\hat{\boldsymbol{\beta}})V_{\hat{\boldsymbol{\beta}}}^{-T}\right)\\
=&Tr\left(\frac{1}{r}V_{\hat{\boldsymbol{\beta}}}^{-1}\left(\frac{1}{n^2}\sum_{i=1}^{n}\frac{\psi^2(Y_i - X_i^T\hat{\boldsymbol{\beta}})}{\pi_i}X_iX_i^T\right)V_{\hat{\boldsymbol{\beta}}}^{-T}\right)\\
=&\sum_{i=1}^{n}\frac{1}{r\pi_i n^2}\left(V_{\hat{\boldsymbol{\beta}}}^{-1}X_i\right)^T\left(V_{\hat{\boldsymbol{\beta}}}^{-1}X_i\right)\psi^2(Y_i - X_i^T\hat{\boldsymbol{\beta}})\\
=&\sum_{i=1}^{n}\frac{1}{r\pi_i n^2}\left\|V_{\hat{\boldsymbol{\beta}}}^{-1}X_i\right\|^2\psi^2(Y_i - X_i^T\hat{\boldsymbol{\beta}}).
\end{aligned} \tag{3.31}$$

Using the Lagrange multiplier method, we obtain the A-optimal probabilities below.

$$\hat{\pi}_i = \frac{\left\|V_{\hat{\boldsymbol{\beta}}}^{-1}X_i\right\|\left|\psi(Y_i - X_i^T\hat{\boldsymbol{\beta}})\right|}{\sum_{i=1}^{n}\left\|V_{\hat{\boldsymbol{\beta}}}^{-1}X_i\right\|\left|\psi(Y_i - X_i^T\hat{\boldsymbol{\beta}})\right|}, \quad i = 1, \cdots, n.$$

Theorem 3.15. *Assume that $V_{\hat{\boldsymbol{\beta}}}$ is invertible such that $a_i = V_{\hat{\boldsymbol{\beta}}}^{-1} X_i \neq 0$. Then there exists a unique A-optimal distribution $\hat{\boldsymbol{\pi}}$ for $\hat{\boldsymbol{\beta}}_r^*$ to approximate $\hat{\boldsymbol{\beta}}_n$, which is given by*

$$\hat{\pi}_i = \frac{\|a_i\| \cdot \left|\psi(Y_i - X_i^T \hat{\boldsymbol{\beta}})\right|}{\sum_{i=1}^n \|a_i\| \cdot \left|\psi(Y_i - X_i^T \hat{\boldsymbol{\beta}})\right|}, \quad i = 1, \cdots, n, \tag{3.32}$$

where $V_{\hat{\boldsymbol{\beta}}} = \frac{\partial}{\partial \boldsymbol{\beta}} \mathbb{E}(\psi(Y - X^T \boldsymbol{\beta}) X)$.

Proof. To find the minimizer of $\tau(\boldsymbol{\pi})$ in (3.31), we use the Lagrange multipliers,

$$L(\boldsymbol{\pi}, \lambda) = \tau(\boldsymbol{\pi}) + \lambda(\pi_1 + \cdots + \pi_n - 1).$$

Setting

$$\frac{\partial L}{\partial \pi_i} = -\frac{\left\| V_{\hat{\boldsymbol{\beta}}}^{-1} X_i \right\|^2 \psi^2(Y_i - X_i^T \hat{\boldsymbol{\beta}})}{r n^2 \pi_i^2} + \lambda = 0, \quad , i = 1, \cdots, n$$

we find

$$\pi_i = \frac{\left\| V_{\hat{\boldsymbol{\beta}}}^{-1} X_i \right\| |\psi(Y_i - X_i^T \hat{\boldsymbol{\beta}})|}{n\sqrt{r\lambda}},$$

As $\pi_1 + \cdots + \pi_n = 1$, we solve for $n\sqrt{r\lambda}$ and get the critical point

$$\hat{\pi}_i = \frac{\|a_i\| \cdot \left|\psi(Y_i - X_i^T \hat{\boldsymbol{\beta}})\right|}{\sum_{i=1}^n \|a_i\| \cdot \left|\psi(Y_i - X_i^T \hat{\boldsymbol{\beta}})\right|}, \quad i = 1, \cdots, n.$$

Since the second partial derivatives are given by

$$\frac{\partial^2 \tau(\boldsymbol{\pi})}{\partial \pi_i \partial \pi_j^T} = \frac{2\left\| V_{\hat{\boldsymbol{\beta}}}^{-1} X_i \right\|^2 \psi^2(Y_i - X_i^T \hat{\boldsymbol{\beta}})}{r n^2 \pi_i^3} \mathbf{1}[i = j], \quad i, j = 1, 2, \cdots, n,$$

where $\mathbf{1}[i = j]$ denotes the indicator of event $\{i = j\}$, the matrix of the second partial derivative of $\tau(\boldsymbol{\pi})$ is positive definite as it is diagonal with positive diagonal entries $\| \left(V_{\hat{\boldsymbol{\beta}}}^{-1} X_i \right) \|^2 \psi^2(Y_i - X_i^T \hat{\boldsymbol{\beta}})$ by assumption. Therefore, the critical point $\hat{\pi}_i$ is the unique minimizer of the trace $\tau(\boldsymbol{\pi})$. □

Remark 3.16. We approximate numerically $\boldsymbol{\mu}(\hat{\boldsymbol{\beta}})$ by

$$\hat{\mu}_k = \frac{\partial}{\partial \beta_k} \boldsymbol{\mu}(\hat{\boldsymbol{\beta}}) = \frac{1}{2c\sqrt{n}} \sum_{i=1}^n [\psi(Y_i - X_i^T(\hat{\boldsymbol{\beta}} + n^{-\frac{1}{2}} c \boldsymbol{e}_k)) - \psi(Y_i - X_i^T(\hat{\boldsymbol{\beta}} - n^{-\frac{1}{2}} c \boldsymbol{e}_k))]$$

where $\boldsymbol{e}_k \in \mathbb{R}^p$ with all components zero except the kth component equals to 1 and c is a constant.

Remark 3.17. From (3.32), we can see that the $\psi(Y - X^T\hat{\boldsymbol{\beta}})$ can be 0 for some data points. When we apply this A-optimal sampling method, we will truncate the probability distribution $\boldsymbol{\pi}$ by a fixed positive constant by assumption A1.

Remark 3.18. When we do simulations and real data analysis, we apply A-optimal scoring method in Peng and Tan(2019), Cheung(2019).

4. SIMULATION STUDY

4.1 Outlier Inclusion in Subsamples

We will compare outlier ratios in subsamples selected based on A-optimal probabilities calculated from robust linear regressions(using bisquare function, Huber function, Hampel function), A-optimal probabilities calculated from linear regression and Uniform probabilities.

Let $n = 1000,000$, $p = 50$, $r_0 = 1\%n$. We consider $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta} = (\mathbf{1}_{30}^T, 0.1 \cdot \mathbf{1}_{20}^T)^T$, $X_i \sim N(0,1)$ are i.i.d., $i = 1, \cdots, 50$. Consider the error density of ϵ:

$\epsilon \sim N(0,1)$ with outliers in y direction generated from $\epsilon \sim Unif(1000, 2000)$. The percentage of outliers varies among 5%, 10%, 20%.

To introduce outliers, we first generated errors from $N(0,1)$ distribution for n observations, then errors were replaced by realizations from Unif(1000, 2000) distribution

We apply A-optimal and Uniform subsampling methods to the data set with subsample size $r = 1000(0.1\%n), 3000(0.3\%n), 5000(0.5\%n), 10000(1\%n), 30000(3\%n), 50000(5\%n)$. The function ψ is chosen from bisquare function, Huber function, Hampel function and the identity function.

For every subsample size r and each sampling distribution (A-optimal sampling distributions from robust linear regressions, linear regression, and uniform sampling distribution) 1000 subsamples were generated. The average outlier ratio was calculated and compared.

We can see that for each subsample size, A-optimal sampling distribution calculated from robust linear regression using bisquare function and Hampel function selects the least outliers among five methods(A-optimal probabilities calculated from robust linear regression with bisquare function, Huber function, Hampel function, A-optimal probability calculated from linear regression, and Uniform probability). This is because both of them

are redescending functions. A-optimal sampling distribution for these two functions will give a 0 weight for extreme outliers. The proportion of outliers selected by A-optimal probabilities calculated from linear regression is highest.

Table 4.1.: $n = 1000,000$ and $p = 50$. Comparison of outlier ratio in subsamples selected using A-optimal sampling distributions from robust linear regressions (Bisquare, Huber, and Hampel), linear regression, and uniform sampling distribution. $\epsilon \sim N(0, 1)$ when 5% outliers included.

	r	Rlm Bisquare	Rlm Huber	Rlm Hampel	LM	Uniform
1	1000	0.000000	0.092850	0.000000	0.513529	0.050151
2	2000	0.000000	0.092456	0.000000	0.513020	0.050376
3	3000	0.000000	0.092761	0.000000	0.513241	0.049901
4	4000	0.000000	0.092682	0.000000	0.513704	0.050102
5	5000	0.000000	0.092527	0.000000	0.513775	0.050120
6	10000	0.000000	0.092718	0.000000	0.513377	0.049960
7	20000	0.000000	0.092655	0.000000	0.513369	0.049901
8	30000	0.000000	0.092598	0.000000	0.513518	0.049919
9	40000	0.000000	0.092699	0.000000	0.513375	0.050030
10	50000	0.000000	0.092719	0.000000	0.513423	0.050016

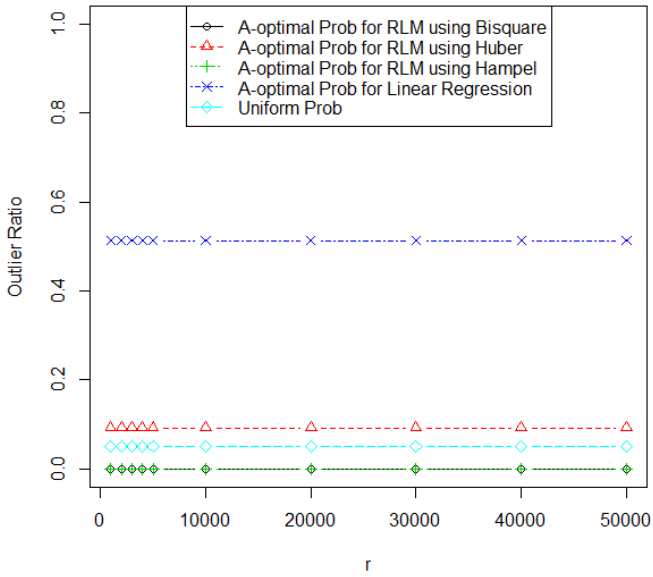

Fig. 4.1.: Comparison of outlier ratio in subsamples. $\epsilon \sim N(0,1)$ when 5% outliers included.

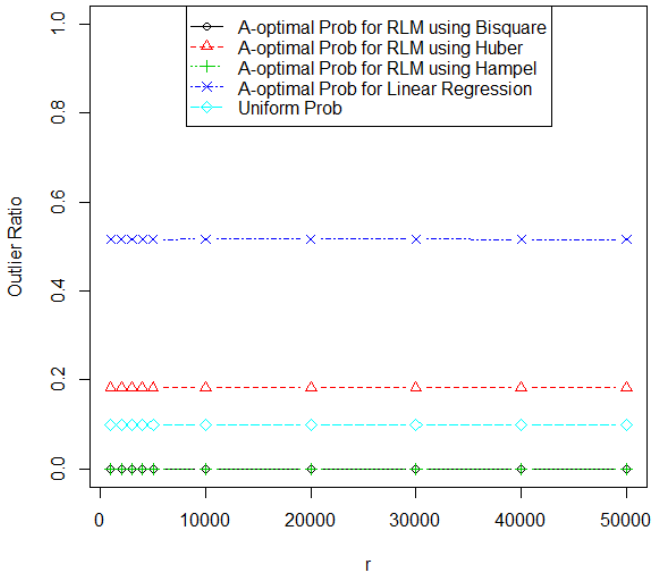

Fig. 4.2.: Comparison of outlier ratio in subsamples. $\epsilon \sim N(0,1)$ when 10% outliers included.

Table 4.2.: $n = 1000,000$ and $p = 50$. Comparison of ratios of outliers selected using A-optimal sampling distributions from robust linear regressions (Bisquare, Huber, and Hampel), linear regression, and uniform sampling distribution. $\epsilon \sim N(0,1)$ when 10% outliers included.

	r	Rlm Bisquare	Rlm Huber	Rlm Hampel	LM	Uniform
1	1000	0.000000	0.182923	0.000000	0.516029	0.100037
2	2000	0.000000	0.183008	0.000000	0.516196	0.100102
3	3000	0.000000	0.183114	0.000000	0.516465	0.099824
4	4000	0.000000	0.183094	0.000000	0.516172	0.099929
5	5000	0.000000	0.182589	0.000000	0.516115	0.100050
6	10000	0.000000	0.182832	0.000000	0.516385	0.099857
7	20000	0.000000	0.182836	0.000000	0.516383	0.100054
8	30000	0.000000	0.182879	0.000000	0.516359	0.100070
9	40000	0.000000	0.182946	0.000000	0.516275	0.100062
10	50000	0.000000	0.182846	0.000000	0.516231	0.099960

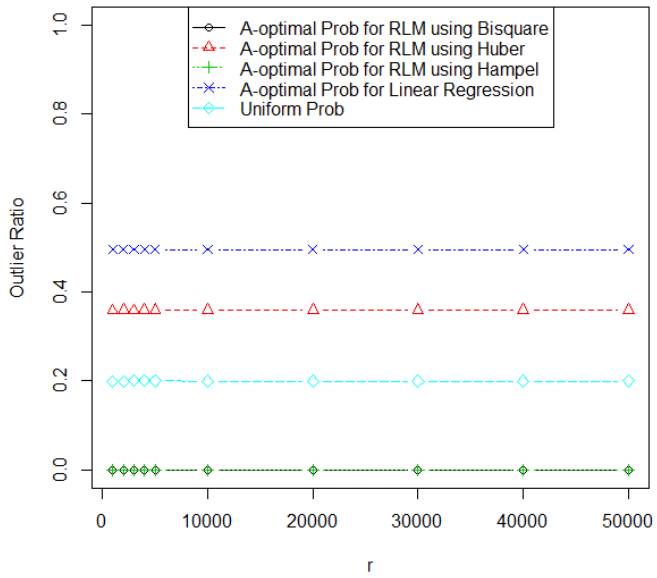

Fig. 4.3.: Comparison of outlier ratio in subsamples. $\epsilon \sim N(0,1)$ when 20% outliers included.

Table 4.3.: $n = 1000,000$ and $p = 50$. Comparison of ratios of outliers selected using A-optimal sampling distributions from robust linear regressions (Bisquare, Huber, and Hampel), linear regression, and uniform sampling distribution. $\epsilon \sim N(0,1)$ when 20% outliers included.

	r	Rlm Bisquare	Rlm Huber	Rlm Hampel	LM	Uniform
1	1000	0.000000	0.359059	0.000000	0.495605	0.199241
2	2000	0.000000	0.359244	0.000000	0.495684	0.199610
3	3000	0.000000	0.359180	0.000000	0.495395	0.200426
4	4000	0.000000	0.359510	0.000000	0.495500	0.200277
5	5000	0.000000	0.359455	0.000000	0.495322	0.200071
6	10000	0.000000	0.359348	0.000000	0.495426	0.199954
7	20000	0.000000	0.359421	0.000000	0.495248	0.199906
8	30000	0.000000	0.359484	0.000000	0.495094	0.199966
9	40000	0.000000	0.359378	0.000000	0.495270	0.199924
10	50000	0.000000	0.359521	0.000000	0.495176	0.200005

4.2 Subsampling of Moderate Data

Now we consider moderate data and compare A-optimal subsampling robust estimator, uniform subsampling robust estimators, A-optimal subsampling linear model estimator, and uniform subsampling linear model estimator. For robust estimators we used the Bisquare function. Let $n = 10,000$, $p = 50$, $r_0 = 3000$. Consider $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta} = (\mathbf{1}_{30}^T, 0.1 \cdot \mathbf{1}_{20}^T)^T$, $X_i \sim N(0,1)$ are i.i.d., $i = 1, \cdots, 50$. We consider the following three cases for the error distribution:

1. $\epsilon \sim t_3$ distribution,

2. $\epsilon \sim t_1$ distribution (Cauchy distribution),

3. $\epsilon \sim N(0,1)$ with 20% observations replaced by those generated from $\epsilon \sim N(0, 50^2)$.

We apply A-optimal and Uniform subsampling methods to the simulated data with subsample size $r = 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000$. For each r, we repeatedly

apply the following below weighted estimation algorithm for M = 1000 times to get the weighted subsample estimators $\boldsymbol{\beta}^*_{r,m}$, $m = 1, \cdots, M$.

The weighted estimation algorithm by A-optimal subsampling method:

1. Subsample(size r_0) with replacement from the data by uniform distribution. Use the robust scale estimate $\hat{\sigma}$ to construct the approximate A-optimal subsampling probability $\boldsymbol{\pi} = \{\pi_i\}_{i=1}^n$.

2. Draw a subsample of size $r \ll n$ randomly according to the sampling distribution of $\boldsymbol{\pi}^*$ with replacement. Then we get a subsample $(\boldsymbol{x}^*_1, y^*_1), \cdots, (\boldsymbol{x}^*_r, y^*_r)$.

3. Solve weighted estimating equation below using the subsample to get the weighted subsample estimator $\hat{\boldsymbol{\beta}}^*_r$.

$$\sum_{i=1}^{r} \frac{1}{\pi^*_i} \psi\Big(\frac{y^*_i - \boldsymbol{x}^{*T}_i \hat{\boldsymbol{\beta}}^*_r}{\sigma}\Big) \boldsymbol{x}^*_i = 0$$

First, for robust regression with psi = Bisquare in weighted estimating equation we compare the effect of A-optimal subsampling using the same ψ to that of uniform subsampling. Specifically, we calculate the mean squared error (MSE) and bias^2 as follows:

$$MSE(\boldsymbol{\beta}^*_r) = \frac{1}{M} \sum_{m=1}^{M} \|\boldsymbol{\beta}^*_{r,m} - \hat{\boldsymbol{\beta}}_n\|^2, bias^2(\boldsymbol{\beta}^*_r) = \|\frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\beta}^*_{r,m} - \hat{\boldsymbol{\beta}}_n\|^2$$

For comparison of subsampling methods under robust regression, we also calculate the MSE ratio and bias^2 ratio of A-optimal subsampling method to uniform.

To compare the effect of robust regression under subsampling to that of linear regression under subsampling we consider Bisquare and identify functions for ψ. We calculate the mean squared error (MSE0) and the median absolute deviation (MAD0 ; You, 1999) as follows:

$$MSE0(\boldsymbol{\beta}^*_r) = \frac{1}{M} \sum_{m=1}^{M} \|\boldsymbol{\beta}^*_{r,m} - \boldsymbol{\beta}_0\|^2, MAD0(\boldsymbol{\beta}^*_r) = median(\|\boldsymbol{\beta}^*_{r,m} - \boldsymbol{\beta}_0\|)$$

where $\boldsymbol{\beta}_0 = (0, \mathbf{1}^T_{30}, 0.1 \cdot \mathbf{1}^T_{20})^T$. We include MAD0 as it is a robust measure w.r.t. outliers.

Below are simulation results for these three error distributions.

Table 4.4.: $n = 10,000$ and $p = 50$. Comparison of Bias^2 and MSE using A-optimal and Uniform subsampling methods with different subsample size r for $\epsilon \sim t_3$ distribution.

r	A-opt Bias^2	Unif Bias^2	Bias^2 Ratio	A-opt MSE	Unif MSE	MSE Ratio
1000	0.00	0.00	9.75	0.07	0.09	0.84
3000	0.00	0.00	16.14	0.02	0.03	0.79
5000	0.00	0.00	30.25	0.01	0.02	0.77
7000	0.00	0.00	35.22	0.01	0.01	0.76
9000	0.00	0.00	55.22	0.01	0.01	0.78
10000	0.00	0.00	43.64	0.01	0.01	0.79

Table 4.5.: $n = 10,000$ and $p = 50$. Comparison of MAD0 and MSE0 using A-optimal and Uniform subsampling methods with different subsample size r and different $\psi(x)$ function for $\epsilon \sim t_3$ distribution.

r	A-opt RLM MAD0	A-opt RLM MSE0	Unif RLM MAD0	Unif RLM MSE0	A-opt LM MAD0	A-opt LM MSE0	Unif LM MAD0	Unif LM MSE0
1000	0.28	0.08	0.30	0.10	0.31	0.10	0.40	0.16
3000	0.17	0.03	0.19	0.04	0.20	0.04	0.25	0.06
5000	0.15	0.02	0.16	0.02	0.17	0.03	0.21	0.04
7000	0.13	0.02	0.14	0.02	0.16	0.02	0.18	0.04
9000	0.12	0.02	0.13	0.02	0.15	0.02	0.17	0.03
10000	0.12	0.01	0.13	0.02	0.15	0.02	0.17	0.03

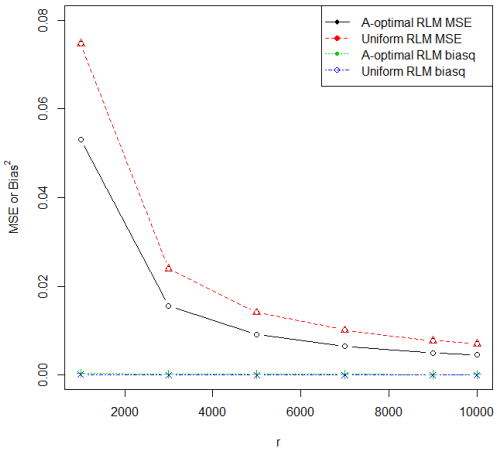

Fig. 4.4.: Comparison of Bias^2 and MSE for $\epsilon \sim t_3$ distribution.

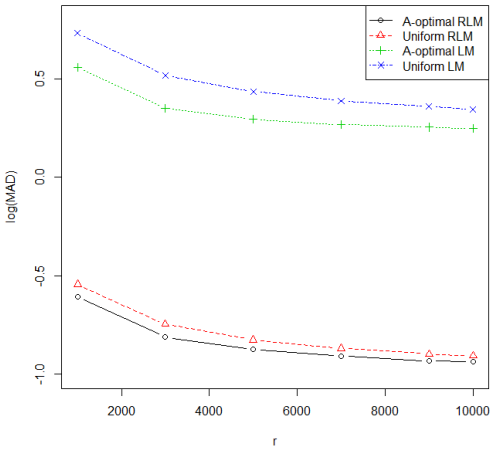

Fig. 4.5.: Comparison of log(MAD0) for $\epsilon \sim t_3$ distribution.

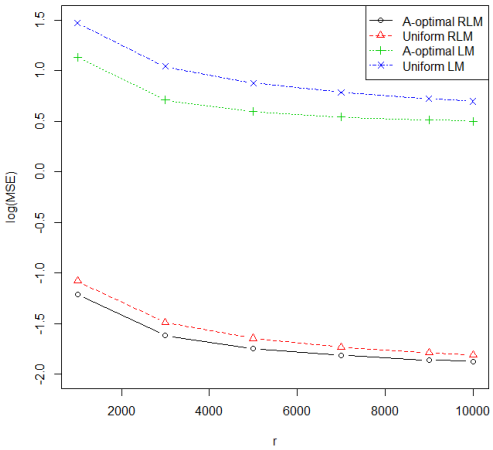

Fig. 4.6.: Comparison of log(MSE0) for $\epsilon \sim t_3$ distribution.

Table 4.6.: $n = 10,000$ and $p = 50$. Comparison of Bias^2 and MSE using A-optimal and Uniform subsampling methods with different subsample size r for $\epsilon \sim t_1$ distribution.

r	A-opt Bias^2	Unif Bias^2	Bias^2 Ratio	A-opt MSE	Unif MSE	MSE Ratio
1000	0.00	0.00	9.28	0.13	0.16	0.82
3000	0.00	0.00	22.47	0.04	0.05	0.77
5000	0.00	0.00	55.65	0.02	0.03	0.79
7000	0.00	0.00	48.10	0.02	0.02	0.80
9000	0.00	0.00	103.41	0.01	0.02	0.82
10000	0.00	0.00	94.45	0.01	0.01	0.83

Table 4.7.: $n = 10,000$ and $p = 50$. Comparison of MAD0 and MSE0 using A-optimal and Uniform subsampling methods with different subsample size r and different $\psi(x)$ function for $\epsilon \sim t_1$ distribution.

r	A-opt RLM MAD0	A-opt RLM MSE0	Unif RLM MAD0	Unif RLM MSE0	A-opt LM MAD0	A-opt LM MSE0	Unif LM MAD0	Unif LM MSE0
1000	0.37	0.14	0.41	0.17	10.23	108.47	12.83	909.30
3000	0.23	0.05	0.25	0.06	9.11	83.64	11.61	335.82
5000	0.20	0.04	0.21	0.04	8.88	79.45	12.21	226.46
7000	0.18	0.03	0.19	0.04	8.77	77.02	10.66	194.92
9000	0.17	0.03	0.17	0.03	8.70	75.94	10.09	156.82
10000	0.17	0.03	0.17	0.03	8.67	75.55	10.08	148.89

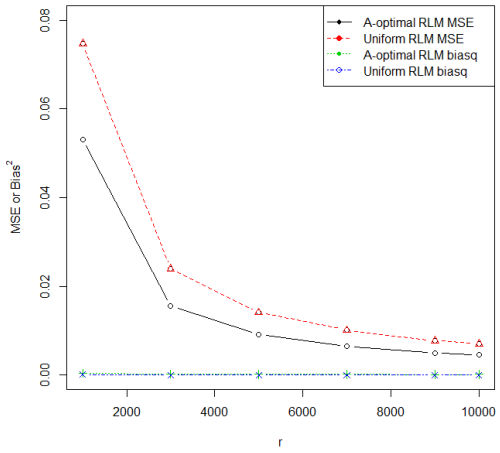

Fig. 4.7.: Comparison of Bias^2 and MSE for $\epsilon \sim t_1$ distribution.

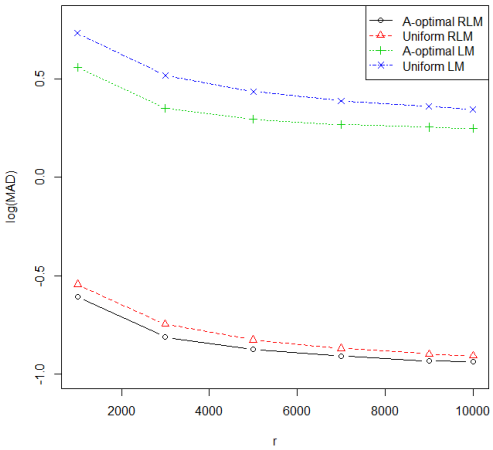

Fig. 4.8.: Comparison of log(MAD0) for $\epsilon \sim t_1$ distribution.

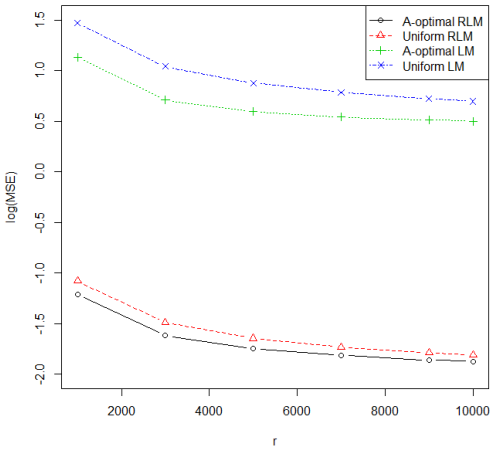

Fig. 4.9.: Comparison of log(MSE0) for $\epsilon \sim t_1$ distribution.

Table 4.8.: $n = 10,000$ and $p = 50$. Comparison of Bias^2 and MSE using A-optimal and Uniform subsampling methods with different subsample size r for $\epsilon \sim N(0,1)$ when outliers included.

r	A-opt Bias^2	Unif Bias^2	Bias^2 Ratio	A-opt MSE	Unif MSE	MSE Ratio
1000	0.00	0.00	6.84	0.05	0.07	0.71
3000	0.00	0.00	8.60	0.02	0.02	0.65
5000	0.00	0.00	8.73	0.01	0.01	0.64
7000	0.00	0.00	10.57	0.01	0.01	0.64
9000	0.00	0.00	17.47	0.00	0.01	0.64
10000	0.00	0.00	14.86	0.00	0.01	0.64

Table 4.9.: $n = 10,000$ and $p = 50$. Comparison of MAD0 and MSE0 using A-optimal and Uniform subsampling methods with different subsample size r and different $\psi(x)$ function for $\epsilon \sim N(0,1)$ when outliers included.

r	A-opt RLM MAD0	A-opt RLM MSE0	Unif RLM MAD0	Unif RLM MSE0	A-opt LM MAD0	A-opt LM MSE0	Unif LM MAD0	Unif LM MSE0
1000.00	0.25	0.06	0.29	0.08	3.62	13.44	5.40	29.62
3000.00	0.15	0.02	0.18	0.03	2.24	5.11	3.29	10.97
5000.00	0.13	0.02	0.15	0.02	1.97	3.92	2.73	7.52
7000.00	0.12	0.01	0.14	0.02	1.85	3.48	2.45	6.10
9000.00	0.12	0.01	0.13	0.02	1.80	3.26	2.29	5.30
10000.00	0.12	0.01	0.12	0.01	1.77	3.16	2.21	4.96

From the previous results, the Bias^2 of A-optimal subsampling estimate is significantly less than the corresponding MSE, which means that the variance plays the main role in MSE of A-optimal subsampling estimate.

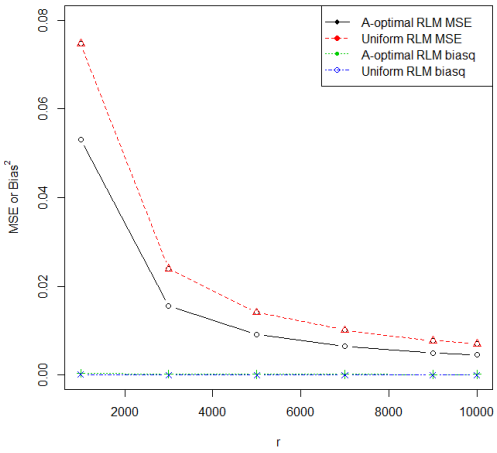

Fig. 4.10.: Comparison of Bias^2 and MSE for $\epsilon \sim N(0,1)$ when outliers included.

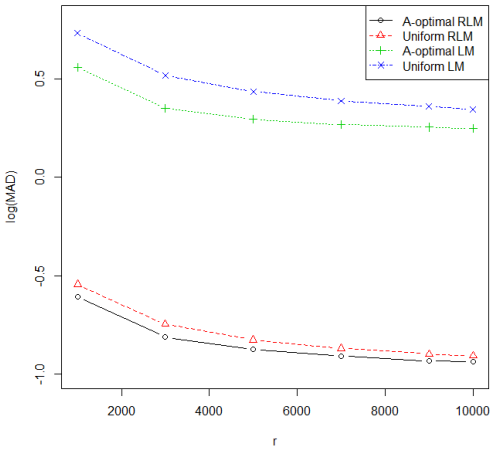

Fig. 4.11.: Comparison of log(MAD0) for $\epsilon \sim N(0,1)$ when outliers included.

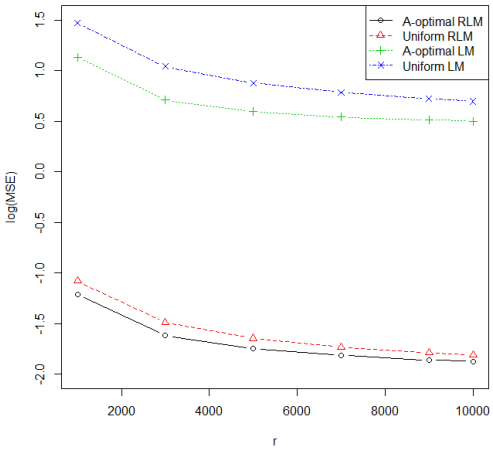

Fig. 4.12.: Comparison of log(MSE0) for $\epsilon \sim N(0,1)$ when outliers included.

The MSE ratio of A-optimal method to uniform method is less than one for all cases. This implies that A-optimal subsampling method outperforms Uniform subsampling method. And the benefit of A-optimal subsampling under robust regression is the largest for normal error with outliers compared to the other error distributions.

For each ψ function, MAD0 and MSE0 for A-optimal subsampling method are less than corresponding values for Uniform subsampling method. For each subsampling method, MAD0 and MSE0 for robust regression with $\psi(x) =$ bisquare are less than corresponding values for linear regression with $\psi(x) = x$. This means A-optimal subsampling estimate in robust regression outperforms that in linear regression and uniform subsampling estimates in robust or linear regression. The benefit of A-optimal subsampling in robust regression is especially evident in cases of Cauchy distribution and normal distribution with outliers.

When the error distribution is a heavy-tailed distribution or is contaminated with outliers, the MSE0 and MAD0 of uniform and A-optimal linear model estimators become much larger than those of A-optimal robust estimator. This means in these situations uniform and A-optimal linear model estimators are unstable, while the A-optimal robust estimator is still stable.

4.3 Subsampling of Big Data

Now we consider big data and let $n = 1000,000$, $p = 50$, $r_0 = 1\%n$. Consider $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta} = (\mathbf{1}_{30}^T, 0.1 \cdot \mathbf{1}_{20}^T)^T$. Consider different combinations of distribution of X and distribution of ϵ:

1. $X_i \sim N(0,1)$ are i.i.d., $i = 1, \cdots, 50$. $\epsilon \sim N(0,1)$ with 10% observations replaced by outliers in y direction generated from $\epsilon \sim Unif(1000, 2000)$.

2. $X_i \sim N(0,1)$ are i.i.d., $i = 1, \cdots, 50$. $\epsilon \sim Laplace(0, \frac{1}{\sqrt{2}})$ with 10% observations replaced by outliers in y direction generated from $\epsilon \sim Unif(1000, 2000)$.

3. $X_i \sim LN(0,1)$ are i.i.d., $i = 1, \cdots, 50$. $\epsilon \sim N(0,1)$ with 10% observations replaced by outliers in y direction generated from $\epsilon \sim Unif(1000, 2000)$.

For each of robust and regular linear regressions, we apply A-optimal and Uniform subsampling methods to the simulated data with subsample size $r = 100(0.01\%n)$, $300(0.03\%n)$, $500(0.05\%n)$, $1000(0.1\%n)$, $3000(0.3\%n)$, $5000(0.5\%n)$. Bisquare function was used as ψ for robust linear model, and identify function was used as ψ to obtain linear model results. For each r, we repeatedly apply weighted estimation algorithm for M = 1000 times to get the weighted subsample estimators $\boldsymbol{\beta}^*_{r,m}$, $m = 1, \cdots, M$. For each subsampling method and each subsample size, we calculate the mean squared errors $(MSE, MSE0)$ and the median absolute deviation (MAD0; You, 1999) as follows:

$$MSE(\boldsymbol{\beta}^*_r) = \frac{1}{M}\sum_{m=1}^{M} \|\boldsymbol{\beta}^*_{r,m} - \hat{\boldsymbol{\beta}}_n\|^2,$$

$$MSE0(\boldsymbol{\beta}^*_r) = \frac{1}{M}\sum_{m=1}^{M} \|\boldsymbol{\beta}^*_{r,m} - \boldsymbol{\beta}_0\|^2,$$

$$MAD0(\boldsymbol{\beta}^*_r) = median(\|\boldsymbol{\beta}^*_{r,m} - \boldsymbol{\beta}_0\|)$$

where $\boldsymbol{\beta}_0 = (0, \mathbf{1}^T_{30}, 0.1 \cdot \mathbf{1}^T_{20})^T$.

For above simulations, we consider two situations. We use the original sampling distribution or the truncated sampling distribution in these two types of simulation. When we do truncation, we use the number which is the smallest positive probability in the sampling distribution as the truncation level.

Below are the MSEs for different cases when sampling distribution is the original one:

Table 4.10.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$. MSE comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	3.4559	1.886	524307.6988	222963.4042
0.02%	200	0.4095	0.419	73301.1613	73010.6801
0.03%	300	0.2199	0.2382	30459.5966	43443.9554
0.04%	400	0.1452	0.1657	18426.1498	30905.6841
0.05%	500	0.1072	0.1264	13145.0717	24098.1943
0.1%	1000	0.0460	0.0586	4895.6266	11263.0165
0.2%	2000	0.0211	0.0290	2166.3241	5487.0712
0.3%	3000	0.0135	0.0191	1396.4656	3633.6931
0.4%	4000	0.0100	0.0142	1009.5619	2749.7793
0.5%	5000	0.0079	0.0115	806.0514	2206.1708

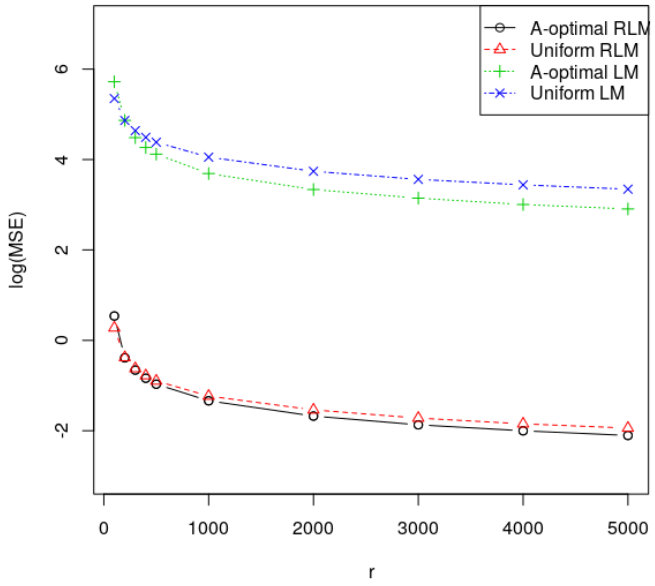

Fig. 4.13.: log(MSE) for $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$.

Table 4.11.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$. MSE comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	3.0309	1.4632	558711.6302	226143.1565
0.02%	200	0.3588	0.3306	77191.8446	73185.7533
0.03%	300	0.1823	0.1834	31901.1288	42881.2060
0.04%	400	0.1173	0.1277	18926.9967	30910.1512
0.05%	500	0.0873	0.0943	13232.7310	23801.3658
0.1%	1000	0.0358	0.0417	5118.0037	11258.7311
0.2%	2000	0.0157	0.0204	2194.1005	5566.0605
0.3%	3000	0.0097	0.0134	1389.4948	3601.6806
0.4%	4000	0.0073	0.0100	1018.0993	2758.0216
0.5%	5000	0.0057	0.0079	808.7604	2152.1429

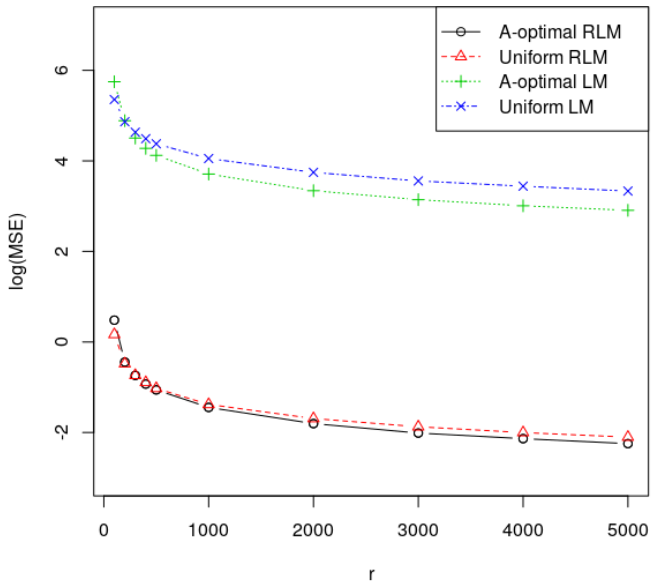

Fig. 4.14.: log(MSE) for $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$.

Table 4.12.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$. MSE comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	2.9086	1.9188	373130.2141	224315.2776
0.02%	200	0.2945	0.3817	61749.0496	71024.1589
0.03%	300	0.1458	0.2176	24317.4139	38977.8965
0.04%	400	0.0953	0.1444	13816.3780	29166.4494
0.05%	500	0.0692	0.1118	9834.1094	20453.9124
0.1%	1000	0.0301	0.0504	3263.1361	9319.8513
0.2%	2000	0.0128	0.0242	1480.4045	4382.3974
0.3%	3000	0.0082	0.0157	896.9277	2985.0254
0.4%	4000	0.0062	0.0117	659.8395	2213.4370
0.5%	5000	0.0048	0.0092	529.8405	1757.3076

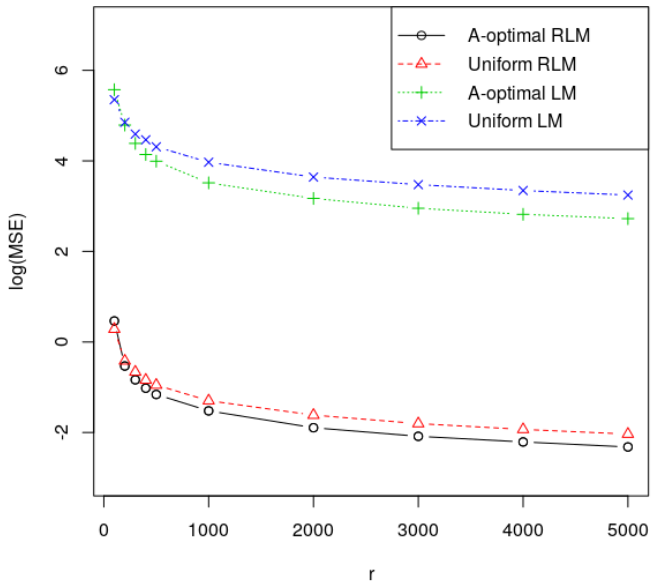

Fig. 4.15.: log(MSE) for $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$.

Below are the MSEs for different cases when sampling distribution is truncated:

Table 4.13.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$. MSE comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	3.4595	1.8671	556735.8074	225598.3397
0.02%	200	0.4143	0.4125	78248.2859	72720.6866
0.03%	300	0.2182	0.2385	32006.2891	43340.4611
0.04%	400	0.1443	0.1683	18906.525	30956.5055
0.05%	500	0.1077	0.1269	13176.5226	23921.2265
0.1%	1000	0.0465	0.0601	4996.4308	11344.4441
0.2%	2000	0.0208	0.0291	2114.717	5511.4154
0.3%	3000	0.0133	0.0193	1374.5657	3627.252
0.4%	4000	0.0099	0.0143	1007.814	2717.039
0.5%	5000	0.0079	0.0114	796.2953	2168.8922

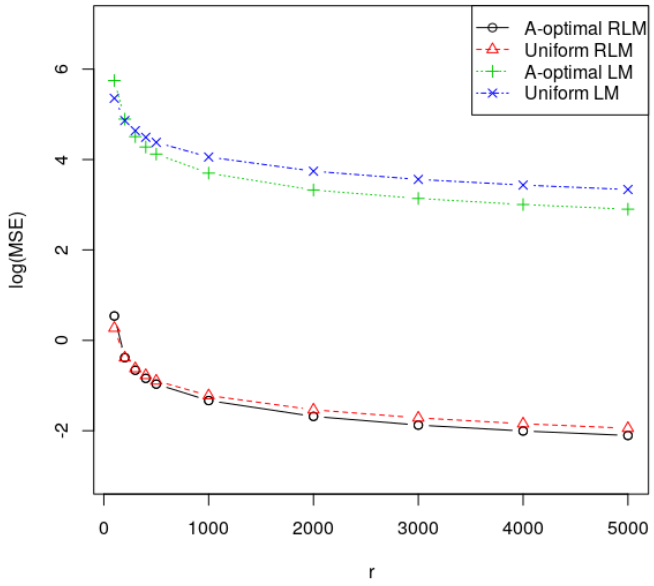

Fig. 4.16.: log(MSE) for $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$ when truncated.

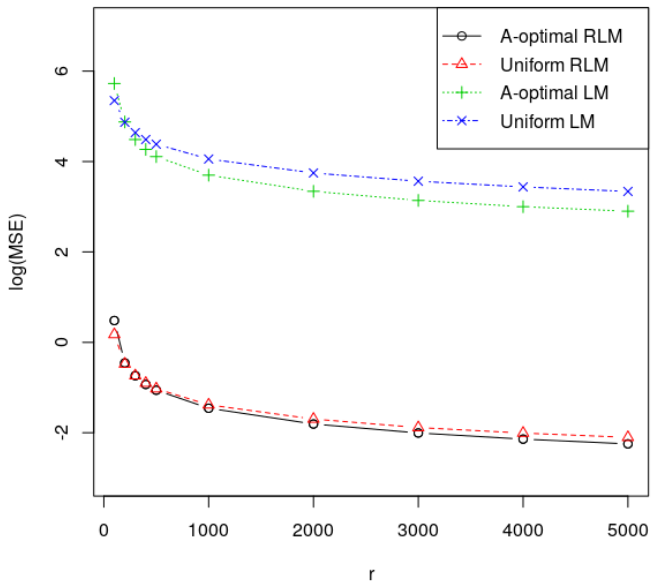

Fig. 4.17.: log(MSE) for $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$ when truncated.

Table 4.14.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$. MSE comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	3.0347	1.4913	531724.0245	222591.4705
0.02%	200	0.3493	0.3322	74933.0404	73813.1314
0.03%	300	0.1828	0.1833	30631.5534	43524.4599
0.04%	400	0.117	0.1253	18504.9972	30637.0618
0.05%	500	0.0873	0.0943	12835.5122	24118.2255
0.1%	1000	0.0349	0.0415	4981.6428	11347.1171
0.2%	2000	0.0156	0.0201	2188.953	5572.4994
0.3%	3000	0.0099	0.013	1377.5359	3662.9044
0.4%	4000	0.0072	0.0098	1000.5944	2742.2095
0.5%	5000	0.0057	0.0079	795.1459	2176.8212

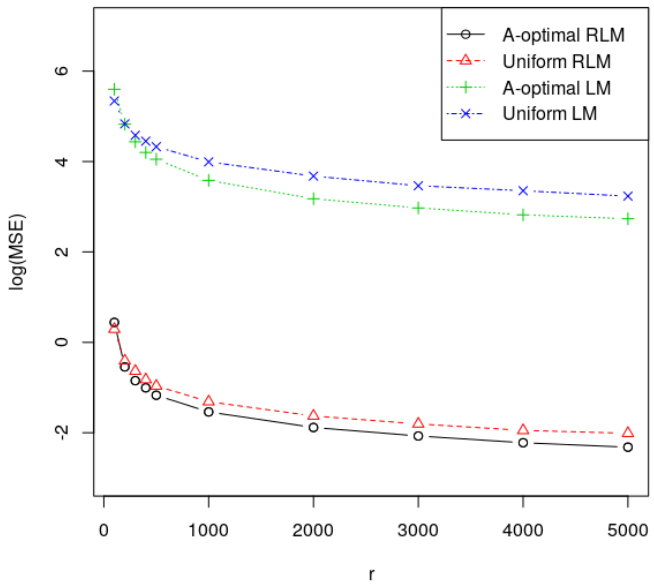

Fig. 4.18.: log(MSE) for $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$ when truncated.

Table 4.15.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$. MSE comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	2.7654	1.9563	396810.48	218181.6004
0.02%	200	0.2863	0.3915	66830.3188	68020.8447
0.03%	300	0.1421	0.2312	27335.3013	38027.5165
0.04%	400	0.0986	0.1502	15814.4264	28184.7247
0.05%	500	0.0676	0.1085	11231.7084	21243.1748
0.1%	1000	0.0289	0.0487	3825.4982	9810.8458
0.2%	2000	0.0131	0.0236	1500.7429	4756.3997
0.3%	3000	0.0085	0.0157	939.1146	2907.8552
0.4%	4000	0.006	0.0113	661.7802	2256.9189
0.5%	5000	0.0048	0.0098	540.8007	1713.0251

Below are the MSE0s for different cases when sampling distribution is the original one:

Table 4.16.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$. MSE0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	3.4561	1.8861	618364.8658	245005.9778
0.02%	200	0.4096	0.419	120314.7306	94812.1887
0.03%	300	0.22	0.2383	67067.7088	66169.5902
0.04%	400	0.1453	0.1658	50930.6015	53253.5515
0.05%	500	0.1073	0.1265	43769.7656	46493.3224
0.1%	1000	0.046	0.0586	30991.8895	33774.3911
0.2%	2000	0.0212	0.029	26508.1618	28071.1771
0.3%	3000	0.0136	0.0191	25082.6306	26166.3174
0.4%	4000	0.01	0.0143	24396.9124	25292.9472
0.5%	5000	0.0079	0.0115	24019.0855	24789.3175

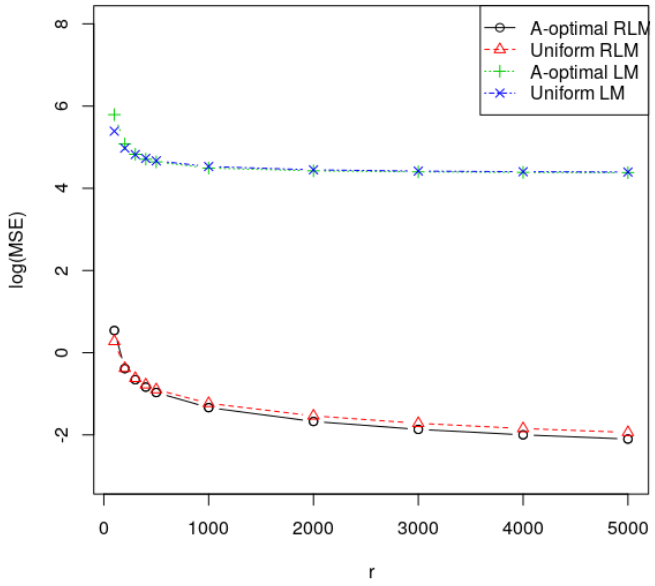

Fig. 4.19.: log(MSE0) for $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$

Table 4.17.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$. MSE0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	3.0311	1.4634	657535.4989	249880.7805
0.02%	200	0.3589	0.3307	125711.7982	95514.9263
0.03%	300	0.1824	0.1835	69211.4169	65095.5737
0.04%	400	0.1174	0.1278	51920.1796	53625.5132
0.05%	500	0.0874	0.0944	43893.3555	46342.3444
0.1%	1000	0.0358	0.0418	31512.8781	33803.7404
0.2%	2000	0.0158	0.0204	26546.4572	28048.9484
0.3%	3000	0.0098	0.0134	25048.6392	26133.188
0.4%	4000	0.0074	0.01	24375.747	25316.6427
0.5%	5000	0.0057	0.008	24023.3109	24575.6695

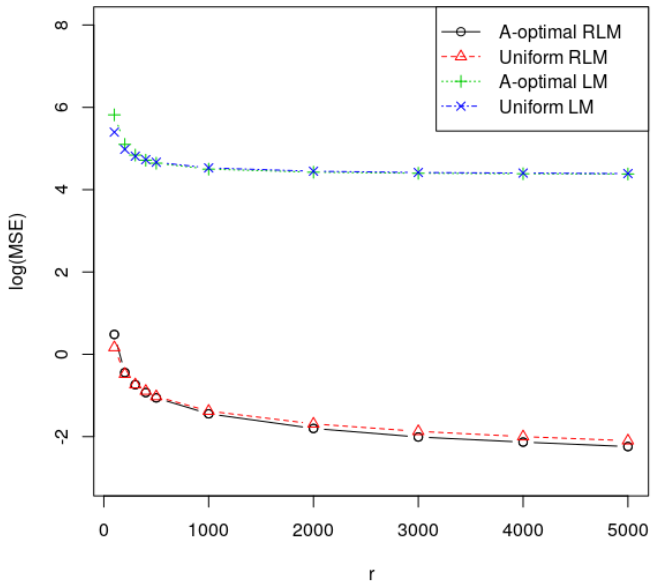

Fig. 4.20.: log(MSE0) for $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$

Table 4.18.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$. MSE0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	2.9088	1.919	347158.9787	246074.5653
0.02%	200	0.2945	0.3817	54211.2317	92855.1134
0.03%	300	0.1458	0.2177	29599.6497	61401.625
0.04%	400	0.0953	0.1444	23826.6701	51543.8083
0.05%	500	0.0692	0.1118	22017.3926	41899.1074
0.1%	1000	0.0301	0.0505	20845.5627	30621.328
0.2%	2000	0.0128	0.0243	20754.0602	26496.0392
0.3%	3000	0.0083	0.0158	20804.6139	24668.2576
0.4%	4000	0.0062	0.0118	21214.9839	23683.2772
0.5%	5000	0.0048	0.0092	21369.3344	23370.714

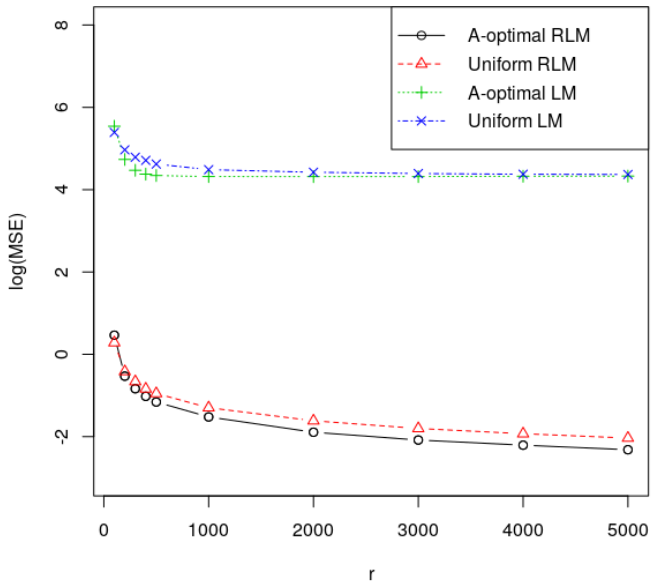

Fig. 4.21.: log(MSE0) for $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$.

Below are the MSE0s for different cases when sampling distribution is truncated:

Table 4.19.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$. MSE comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	3.4596	1.8671	653951.0926	248152.453
0.02%	200	0.4145	0.4126	127075.6281	95272.0855
0.03%	300	0.2183	0.2385	69715.9953	65809.1992
0.04%	400	0.1444	0.1683	52006.0305	53745.8345
0.05%	500	0.1078	0.127	43922.2971	46236.6371
0.1%	1000	0.0466	0.0602	31366.2393	34118.0196
0.2%	2000	0.0209	0.0291	26238.5678	28065.5655
0.3%	3000	0.0134	0.0193	25002.863	26154.0752
0.4%	4000	0.0099	0.0144	24386.5436	25176.338
0.5%	5000	0.0079	0.0114	24012.336	24669.0117

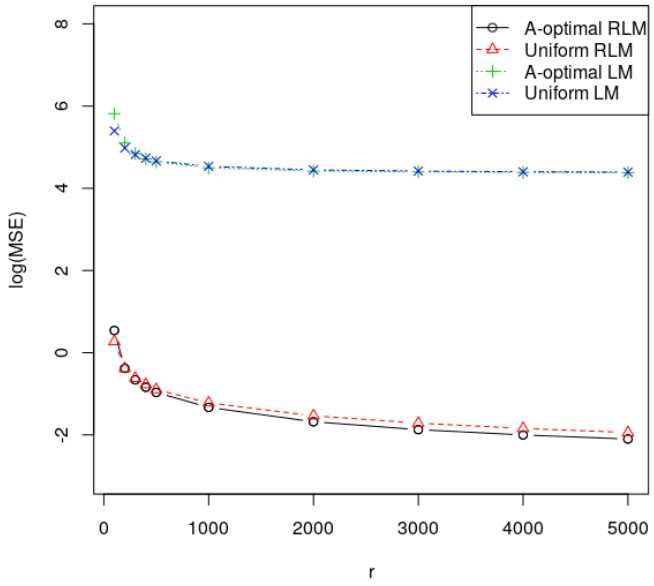

Fig. 4.22.: log(MSE0) for $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$ when truncated

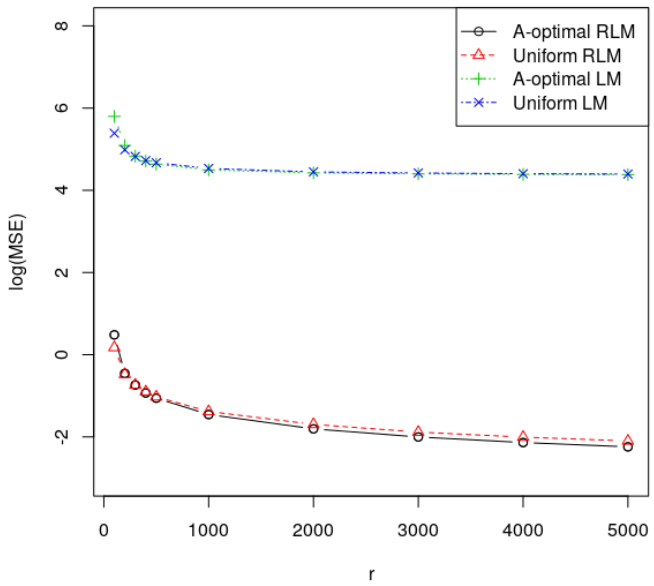

Fig. 4.23.: log(MSE0) for $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$ when truncated

Table 4.20.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$. MSE0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	3.0348	1.4914	625893.2047	244737.5811
0.02%	200	0.3494	0.3322	122831.7439	96597.0051
0.03%	300	0.1829	0.1834	67516.909	66564.3792
0.04%	400	0.1171	0.1253	51230.4463	52700.1743
0.05%	500	0.0873	0.0943	43183.624	46797.9115
0.1%	1000	0.0349	0.0415	31286.3997	33972.1779
0.2%	2000	0.0157	0.0201	26517.7118	28017.0515
0.3%	3000	0.0099	0.0131	25101.3981	26342.9539
0.4%	4000	0.0073	0.0099	24392.1154	25272.0501
0.5%	5000	0.0057	0.008	24074.5701	24796.5678

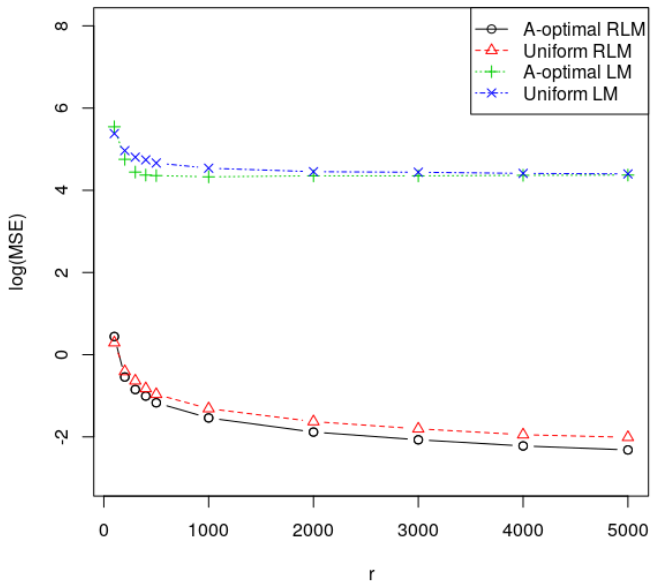

Fig. 4.24.: log(MSE0) for $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$ when truncated.

Table 4.21.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$. MSE0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	2.7656	1.9562	350483.4742	239614.7534
0.02%	200	0.2863	0.3915	56581.0183	92155.8212
0.03%	300	0.142	0.2312	27748.2745	63786.3042
0.04%	400	0.0986	0.1502	23420.9437	54931.9447
0.05%	500	0.0676	0.1085	22750.7615	46088.2339
0.1%	1000	0.0289	0.0487	21244.9044	34443.5018
0.2%	2000	0.0131	0.0236	22478.2676	28325.2434
0.3%	3000	0.0085	0.0157	22436.3124	27458.8191
0.4%	4000	0.006	0.0113	23098.0652	25738.4723
0.5%	5000	0.0048	0.0098	23447.0138	24976.7101

Below are the MAD0s for different cases when sampling distribution is the original one:

Table 4.22.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$. MAD0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	1.8442	1.3011	757.7214	478.6533
0.02%	200	0.6313	0.6394	333.3065	300.6756
0.03%	300	0.4621	0.4821	252.8117	252.8283
0.04%	400	0.377	0.4034	222.8302	229.5451
0.05%	500	0.3241	0.3528	207.0961	213.3701
0.1%	1000	0.2125	0.2417	175.1075	182.2686
0.2%	2000	0.1447	0.1697	162.7017	166.707
0.3%	3000	0.1155	0.1371	158.2276	161.6124
0.4%	4000	0.0993	0.1186	156.1016	158.6157
0.5%	5000	0.0883	0.1066	154.718	157.2894

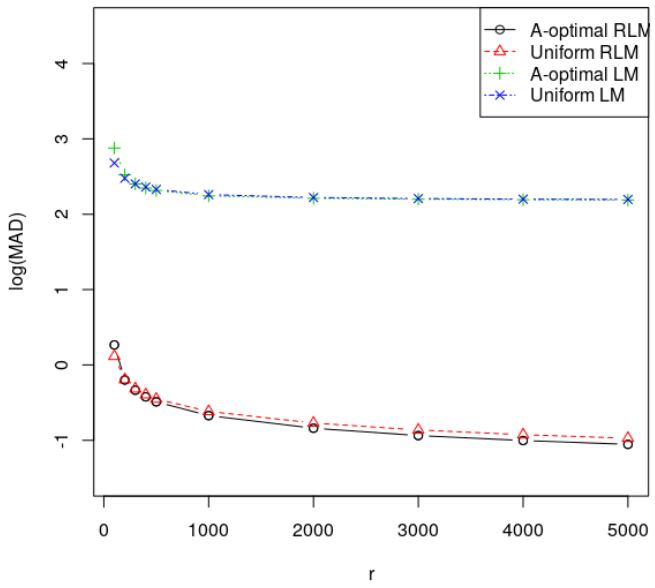

Fig. 4.25.: log(MAD0) for $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$

Table 4.23.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$. MAD0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	1.6987	1.1626	790.3447	490.9932
0.02%	200	0.5875	0.5653	341.2833	302.7485
0.03%	300	0.4216	0.4225	257.5121	252.0195
0.04%	400	0.3374	0.3549	224.7738	230.0359
0.05%	500	0.2919	0.3042	207.4624	212.4139
0.1%	1000	0.1875	0.2019	176.5912	182.5629
0.2%	2000	0.1245	0.1411	162.437	166.7376
0.3%	3000	0.0983	0.1154	157.918	161.48
0.4%	4000	0.0846	0.0991	156.0095	159.02
0.5%	5000	0.0753	0.0885	154.8397	156.5871

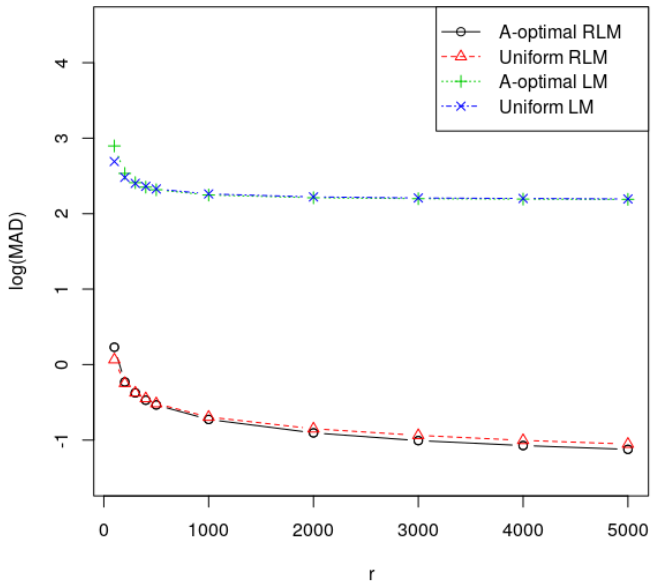

Fig. 4.26.: log(MAD0) for $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$

Table 4.24.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$. MAD0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	1.3558	1.0927	485.2207	381.8553
0.02%	200	0.4417	0.4766	193.6196	233.4021
0.03%	300	0.3183	0.3648	142.3177	187.5426
0.04%	400	0.2514	0.2974	130.2304	171.902
0.05%	500	0.2151	0.2621	130.3351	165.124
0.1%	1000	0.139	0.1823	137.5611	153.7263
0.2%	2000	0.0929	0.1208	139.6707	150.4054
0.3%	3000	0.0731	0.0978	141.9885	148.2635
0.4%	4000	0.0646	0.0837	144.2572	147.2651
0.5%	5000	0.0564	0.0736	144.2851	149.5901

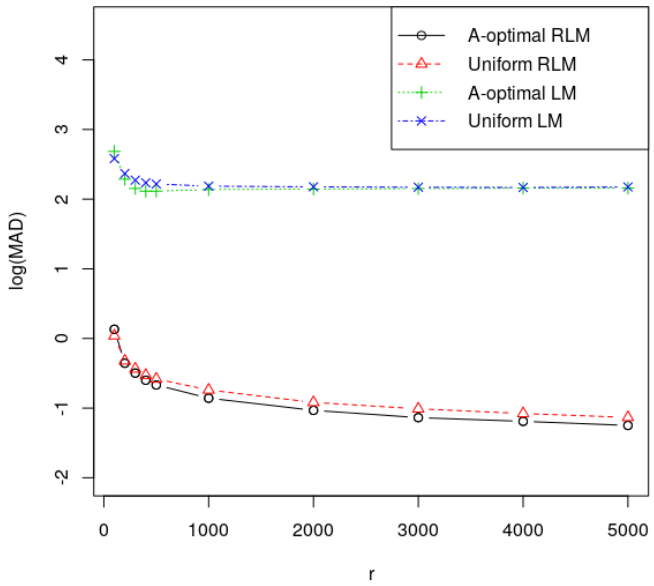

Fig. 4.27.: log(MAD0) for $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$.

Below are the MAD0s for different cases when sampling distribution is truncated:

Table 4.25.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$. MAD0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	1.8197	1.2959	786.5195	480.6537
0.02%	200	0.6362	0.633	344.6008	303.1818
0.03%	300	0.4629	0.4838	257.5532	254.1804
0.04%	400	0.3755	0.4063	225	228.407
0.05%	500	0.3267	0.3543	206.786	213.1039
0.1%	1000	0.214	0.2437	176.4543	183.8921
0.2%	2000	0.1431	0.1697	161.9544	167.0787
0.3%	3000	0.1151	0.1383	157.8499	161.7436
0.4%	4000	0.0984	0.1193	155.773	158.2788
0.5%	5000	0.0881	0.106	154.7489	156.8343

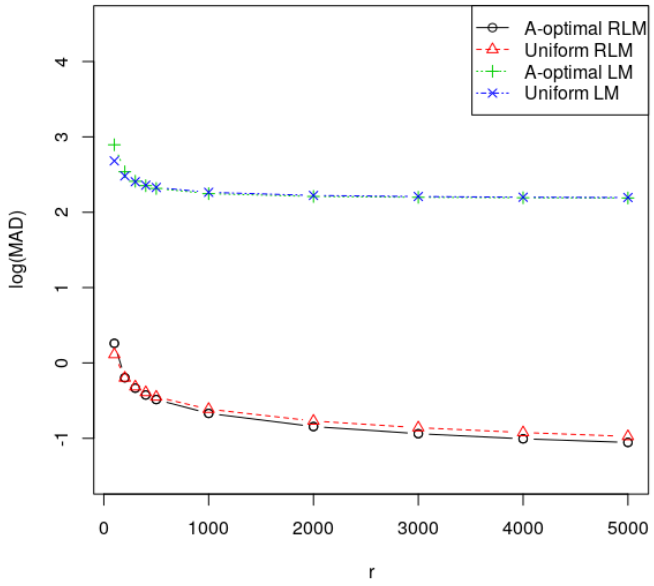

Fig. 4.28.: log(MAD0) for $\boldsymbol{X} \sim GA$, $\epsilon \sim GA$ when truncated

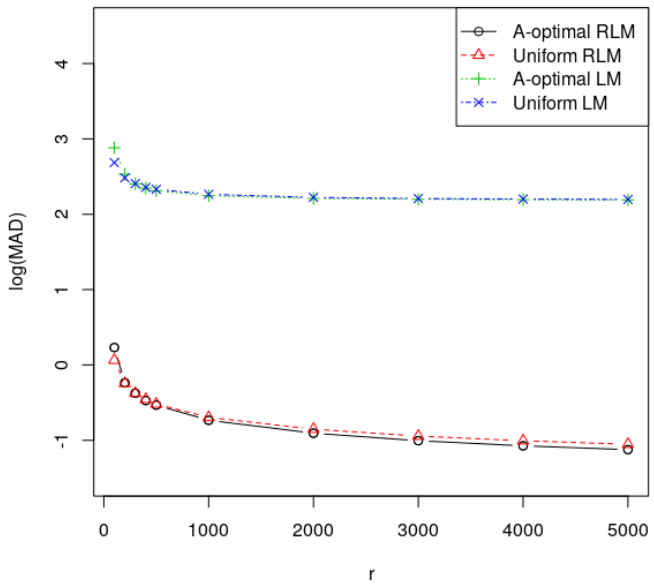

Fig. 4.29.: log(MAD0) for $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$ when truncated

Table 4.26.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim GA$, $\epsilon \sim Laplace$. MAD0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	1.7026	1.1591	762.4678	484.7652
0.02%	200	0.5803	0.5693	339.0357	304.3177
0.03%	300	0.4217	0.4206	253.3356	254.9008
0.04%	400	0.3369	0.3501	221.9994	227.2923
0.05%	500	0.2928	0.3031	205.2161	213.9226
0.1%	1000	0.1842	0.2013	175.7962	183.7532
0.2%	2000	0.1242	0.1405	162.5825	167.0774
0.3%	3000	0.0987	0.1137	158.0078	161.7805
0.4%	4000	0.0845	0.0987	156.0212	159.0116
0.5%	5000	0.0751	0.0886	154.8971	157.2947

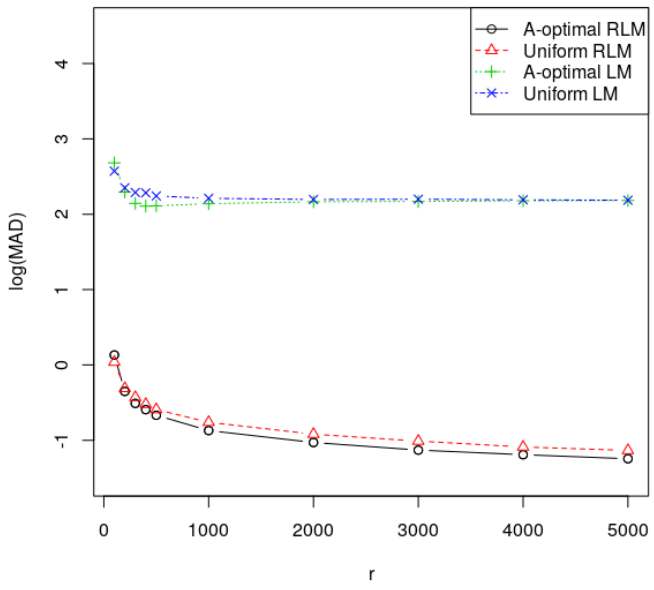

Fig. 4.30.: log(MAD0) for $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$ when truncated.

Table 4.27.: $n = 1000,000$ and $p = 50$. $\boldsymbol{X} \sim LN$, $\epsilon \sim GA$. MAD0 comparison of robust linear regression with bisquare function and linear regression using A-optimal probability and Uniform probability for different subsample sizes r with 10% outliers when truncated.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	100	1.3554	1.0929	480.6763	373.2731
0.02%	200	0.4451	0.4887	196.7637	224.5882
0.03%	300	0.3088	0.3716	139.4373	194.5681
0.04%	400	0.2552	0.3008	128.1133	192.3649
0.05%	500	0.2151	0.2562	129.4284	174.928
0.1%	1000	0.1345	0.1735	137.9613	162.8741
0.2%	2000	0.0935	0.1205	146.4575	157.0033
0.3%	3000	0.0741	0.0976	147.9493	158.3002
0.4%	4000	0.0644	0.0815	150.8639	155.451
0.5%	5000	0.0569	0.0734	151.9705	153.4914

We can see that, for MSEs, MSE0s and MAD0s, in almost all the situations, the values calculated from robust linear regression with bisquare function using A-optimal sampling distribution are the smallest among all four different methods. In most cases these values are smaller than those from robust linear regression with bisquare function using Uniform probability, linear regression using A-optimal probability, and linear regression using Uniform probability. Linear regression with Uniform probability performs worst. In the presence of outliers, the linear model subsampling estimators consistently generate huge MSE values in all simulation scenarios compared to robust regression subsampling estimators. Hence they are not as stable as the robust regression subsampling estimators when outliers are included. The A-optimal robust estimator outperforms the uniform robust estimator.

4.4 Breakdown point

Next we study how tolerant the methods are to outliers via breakdown point. Let $n = 100,000$, $p = 50$, $r_0 = 5\%n$. Consider $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\beta} = (\mathbf{1}_{30}^T, 0.1 \cdot \mathbf{1}_{20}^T)^T$, $X_i \sim N(0,1)$ are i.i.d., $i = 1, \cdots, 50$. We consider the following case for the error distribution:

$\epsilon \sim N(0,1)$ with a varying percentage of observations replaced by outliers in y direction generated from $\epsilon \sim Unif(1000, 2000)$.

Apply A-optimal and Uniform subsampling methods to simulated data with subsample size $r = 100(0.1\%n), 300(0.3\%n), 500(0.5\%n), 1000(1\%n), 3000(3\%n), 5000(5\%n)$. The percentage of outliers is chosen from 1%, 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, 49%, 50%. The number of repetition is $M = 500$. For A-optimal subsampling method, we calculate the mean squared errors $(MSE, MSE0)$ and the median absolute deviation (MAD0; You, 1999) for each subsample size and one of the ψ functions(bisquare, Huber, Hampel). In the plot, the logarithm with base 10 is taken for MSEs, MSE0s, MAD0s.

Below are the MSEs, MSE0s, MAD0s for bisquare function:

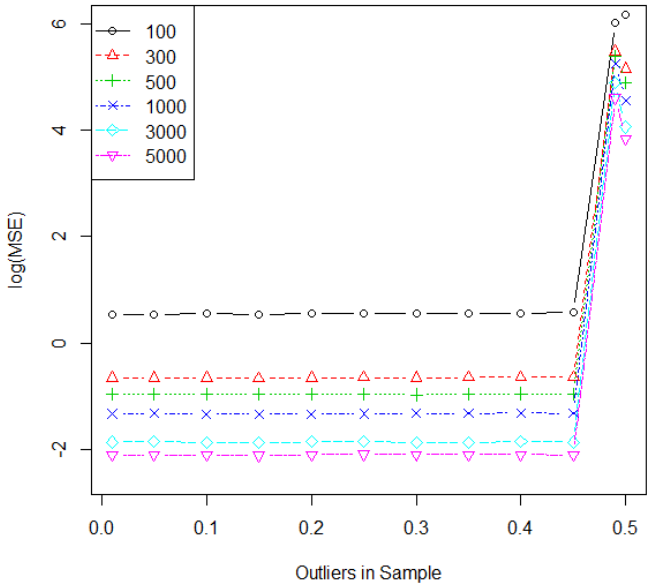

Fig. 4.31.: log(MSE) for weighted robust linear regression with bisquare function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

Table 4.28.: $n = 100,000$ and $p = 50$. MSE comparison of robust linear regression with bisquare function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

r	0.1%n	0.3%n	0.5%n	1%n	3%n	5%n
outlier	(100)	(300)	(500)	(1000)	(3000)	(5000)
1%	3.4732	0.2203	0.1095	0.0466	0.0137	0.0078
5%	3.3999	0.2205	0.1074	0.0476	0.0139	0.0079
10%	3.5964	0.2222	0.1073	0.046	0.0136	0.0079
15%	3.4317	0.2173	0.1089	0.0456	0.0133	0.0077
20%	3.5555	0.2191	0.1105	0.0463	0.0138	0.0079
25%	3.5219	0.2236	0.1103	0.0467	0.0139	0.0081
30%	3.5664	0.2183	0.106	0.047	0.0136	0.008
35%	3.6717	0.2247	0.1096	0.0472	0.0134	0.008
40%	3.5512	0.2265	0.1114	0.0481	0.0141	0.008
45%	3.7204	0.2246	0.1105	0.0471	0.0137	0.0079
49%	1054789.1515	302513.5272	247475.5828	179247.1255	79247.2208	40259.4036
50%	1517716.3912	142242.6194	77112.1688	36456.279	11444.6066	6943.8136

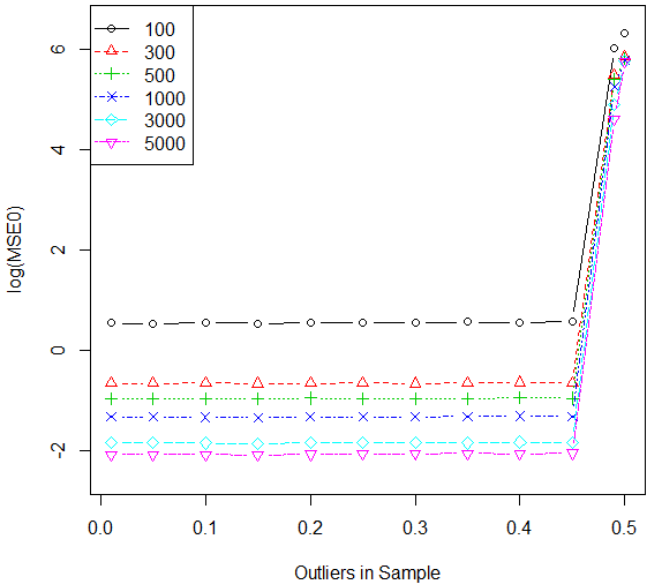

Fig. 4.32.: log(MSE0) for weighted robust linear regression with Bisquare function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

Table 4.29.: $n = 100,000$ and $p = 50$. MSE0 comparison of robust linear regression with Bisquare function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

r	0.1%n	0.3%n	0.5%n	1%n	3%n	5%n
outlier	(100)	(300)	(500)	(1000)	(3000)	(5000)
1%	3.473	0.2207	0.11	0.0472	0.0144	0.0084
5%	3.4011	0.2208	0.1078	0.0481	0.0144	0.0084
10%	3.5967	0.223	0.1078	0.0467	0.0142	0.0084
15%	3.4323	0.2179	0.1095	0.0462	0.0139	0.0083
20%	3.5568	0.2204	0.1116	0.0472	0.0146	0.0086
25%	3.5217	0.2237	0.1107	0.0471	0.0145	0.0087
30%	3.5669	0.2192	0.1066	0.0477	0.0143	0.0087
35%	3.6725	0.226	0.1104	0.048	0.0142	0.0089
40%	3.5521	0.2275	0.1123	0.049	0.0149	0.0087
45%	3.7212	0.2255	0.1115	0.0481	0.0148	0.009
49%	1054787.9033	302512.5575	247474.6676	179246.4925	79246.9194	40259.2576
50%	2088577.4437	691310.669	631654.0753	592929.2791	566778.8892	563540.0267

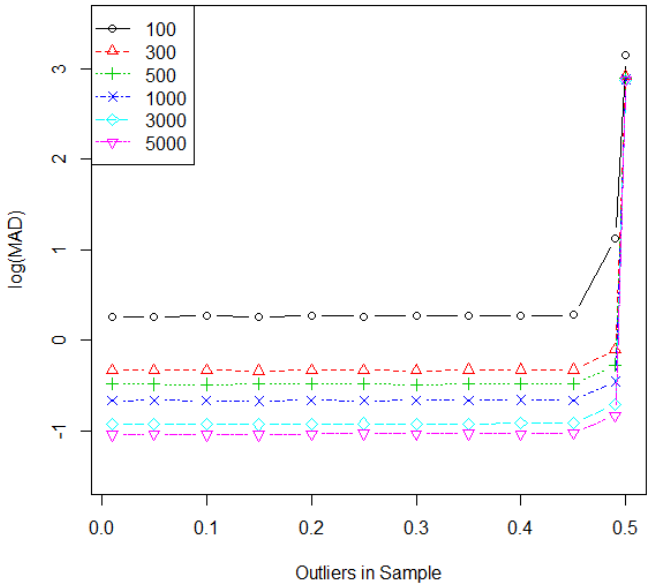

Fig. 4.33.: log(MAD0) for weighted robust linear regression with Bisquare function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

Table 4.30.: $n = 100,000$ and $p = 50$. MAD0 comparison of robust linear regression with Bisquare function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

r	0.1%n	0.3%n	0.5%n	1%n	3%n	5%n
outlier	(100)	(300)	(500)	(1000)	(3000)	(5000)
1%	1.83	0.4649	0.3306	0.2151	0.1186	0.0908
5%	1.8109	0.4668	0.3258	0.2177	0.1193	0.0913
10%	1.8542	0.4668	0.3249	0.2135	0.118	0.0908
15%	1.825	0.4588	0.3279	0.2127	0.1165	0.0901
20%	1.8494	0.4654	0.3319	0.2168	0.1194	0.0924
25%	1.8405	0.4662	0.3311	0.215	0.1197	0.0927
30%	1.8522	0.4616	0.3241	0.2174	0.1184	0.0923
35%	1.8577	0.4692	0.3294	0.2168	0.1192	0.0934
40%	1.8468	0.4702	0.3308	0.2197	0.1213	0.0925
45%	1.9019	0.4688	0.3285	0.2171	0.1211	0.094
49%	13.4581	0.7859	0.5354	0.351	0.1934	0.148
50%	1426.0666	830.6151	792.3451	770.9764	752.4168	750.1542

We can see that for MSEs, MSE0s and MAD0s, there is a huge increase near 49% or 50% outliers. The breakdown point of weighted subsample estimator for A-optimal robust linear regression with bisquare function for this case is close to 0.5 which is the largest possible value for breakdown point.

Below are the MSEs, MSE0s, MAD0s for Huber function:

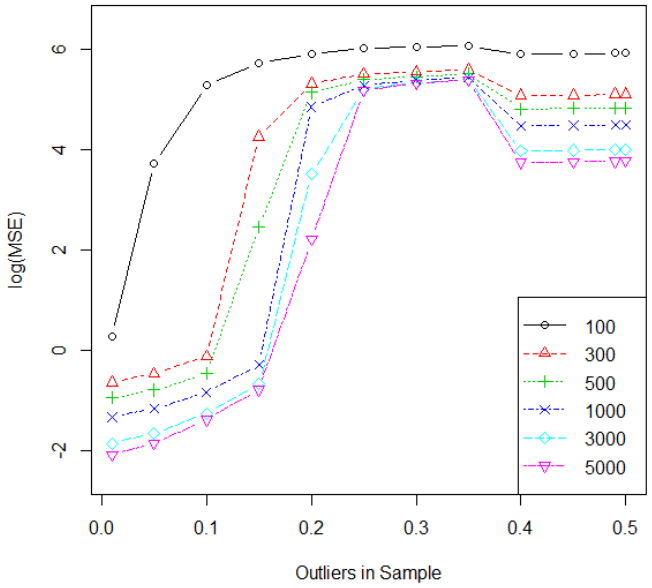

Fig. 4.34.: log(MSE) for weighted robust linear regression with Huber function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

Table 4.31.: $n = 100,000$ and $p = 50$. MSE comparison of robust linear regression with Huber function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

r	0.1%n	0.3%n	0.5%n	1%n	3%n	5%n
outlier	(100)	(300)	(500)	(1000)	(3000)	(5000)
1%	1.9356	0.2287	0.1122	0.0467	0.0141	0.0082
5%	5219.5495	0.3429	0.1615	0.0698	0.0219	0.014
10%	192822.5304	0.7598	0.3501	0.1464	0.0554	0.0414
15%	526773.3827	17726.9542	281.9221	0.5264	0.212	0.1641
20%	796879.5752	202212.6906	135793.9945	70098.3726	3221.47	162.597
25%	1022397.0375	312704.4221	243281.1998	189565.9232	157377.3816	149477.6672
30%	1101434.8495	351016.8777	279216.2068	237461.956	210749.1188	206138.8987
35%	1137698.3454	383592.111	319192.5862	272108.0203	246579.4154	242411.8658
40%	799065.0945	120339.4778	63648.2227	29378.0550	9249.4795	5464.2703
45%	809753.6731	120258.1583	65860.2054	30131.4024	9612.7692	5677.1151
49%	819467.587	124563.2162	65613.1096	30743.5225	9875.6850	5760.5773
50%	819515.213	126098.8006	66649.2207	30619.2287	9826.8386	5933.1124

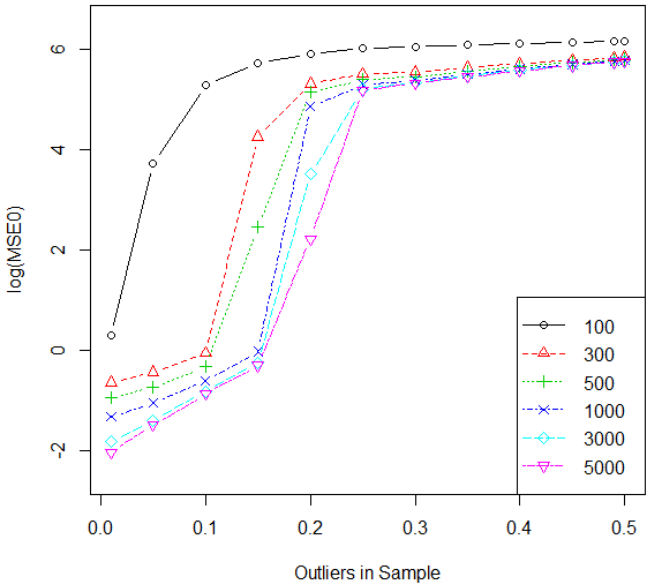

Fig. 4.35.: log(MSE0) for weighted robust linear regression with Huber function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

Table 4.32.: $n = 100,000$ and $p = 50$. MSE0 comparison of robust linear regression with Huber function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

r	0.1%n	0.3%n	0.5%n	1%n	3%n	5%n
outlier	(100)	(300)	(500)	(1000)	(3000)	(5000)
1%	1.9379	0.2299	0.1135	0.0479	0.0152	0.0092
5%	5220.2034	0.3655	0.1834	0.0894	0.0406	0.0324
10%	192865.7187	0.891	0.4684	0.2511	0.1502	0.1357
15%	526965.2684	17749.8085	282.965	0.947	0.5689	0.5054
20%	797324.4896	202560.6602	136099.2357	70302.4225	3235.4822	165.2646
25%	1023343.833	313481.8788	244025.0721	190263.4192	158045.3301	150136.208
30%	1104908.1907	353913.2181	281995.8789	240168.0598	213393.6764	208773.7209
35%	1188744.6607	426800.7546	361586.561	313408.6667	287383.3289	283152.2959
40%	1266853.8002	516817.6786	444767.2322	400268.6584	372364.0129	367894.3271
45%	1359313.8144	597812.1188	534880.7257	492666.3698	468371.2137	462510.3416
49%	1405995.4251	671023.0766	613091.8511	573927.0062	550898.2977	546916.4822
50%	1396076.704	690073.8688	628601.2639	596199.9729	573042.2759	568132.291

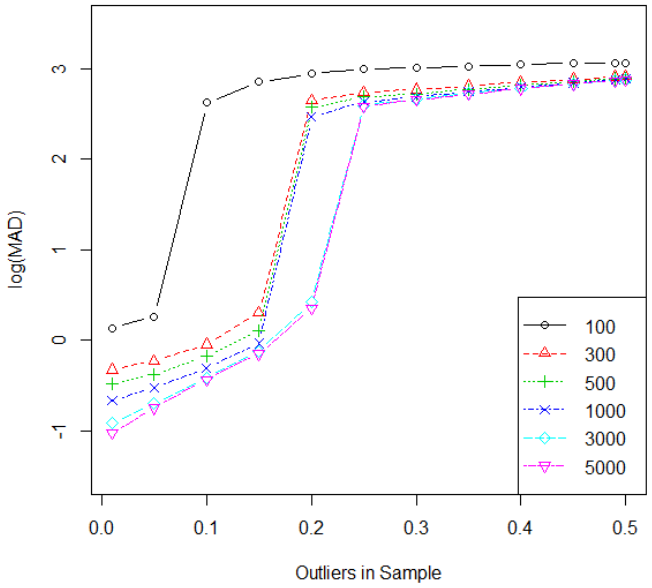

Fig. 4.36.: log(MAD0) for weighted robust linear regression with Huber function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

Table 4.33.: $n = 100,000$ and $p = 50$. MAD0 comparison of robust linear regression with Huber function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

r	0.1%n	0.3%n	0.5%n	1%n	3%n	5%n
outlier	(100)	(300)	(500)	(1000)	(3000)	(5000)
1%	1.3504	0.4692	0.3343	0.2159	0.1228	0.0951
5%	1.8349	0.5911	0.421	0.296	0.1998	0.1787
10%	425.2898	0.9042	0.6641	0.4944	0.3857	0.3673
15%	716.595	2.0289	1.2765	0.9235	0.751	0.709
20%	883.4018	450.7754	376.049	297.1991	2.6409	2.2458
25%	996.5681	554.0115	489.3197	434.1895	396.798	387.4063
30%	1039.0285	591.7994	531.1281	489.9046	461.2415	456.8651
35%	1078.7758	647.7022	597.6204	560.4974	535.6253	531.1443
40%	1114.4489	714.7719	666.5069	631.6665	610.1478	606.9513
45%	1155.0599	768.1869	728.0054	703.8719	684.3247	680.0259
49%	1170.5752	818.8342	780.6631	757.9019	742.5983	739.2662
50%	1172.2103	828.5515	791.9249	772.293	757.2967	753.3127

We can see that for MSEs, MSE0s and MAD0s, there is a huge increase between 10% or 25% outliers. The breakdown point of weighted subsample estimator for A-optimal robust linear regression with Huber function for this case is between 0.1 and 0.25 approximately.

Below are the MSEs, MSE0s, MAD0s for Hampel function:

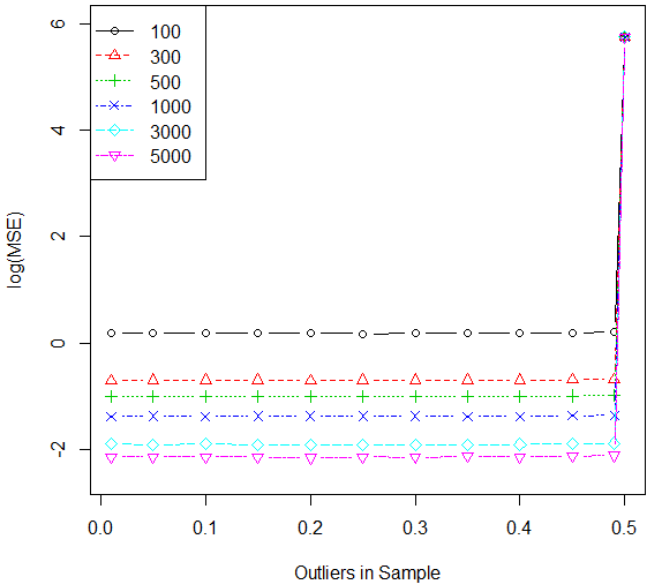

Fig. 4.37.: log(MSE) for weighted robust linear regression with Hampel function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

Table 4.34.: $n = 100,000$ and $p = 50$. MSE comparison of robust linear regression with Hampel function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

r	0.1%n	0.3%n	0.5%n	1%n	3%n	5%n
outlier	(100)	(300)	(500)	(1000)	(3000)	(5000)
1%	1.5097	0.1963	0.0978	0.0417	0.0125	0.0072
5%	1.5496	0.2008	0.0989	0.0424	0.0123	0.0072
10%	1.5191	0.1987	0.1000	0.0418	0.0125	0.0072
15%	1.5428	0.2005	0.0984	0.0420	0.0122	0.0072
20%	1.5558	0.1966	0.0973	0.0422	0.0122	0.0070
25%	1.4872	0.2009	0.0968	0.0422	0.0121	0.0072
30%	1.5410	0.2002	0.0976	0.0422	0.0123	0.0071
35%	1.5127	0.1999	0.1005	0.0419	0.0123	0.0074
40%	1.5388	0.2010	0.0995	0.0414	0.0124	0.0072
45%	1.5665	0.2037	0.1008	0.0431	0.0127	0.0074
49%	1.5824	0.2064	0.1031	0.0436	0.0130	0.0077
50%	563984.9634	563986.1619	563987.1733	563987.3112	563985.8453	563983.9114

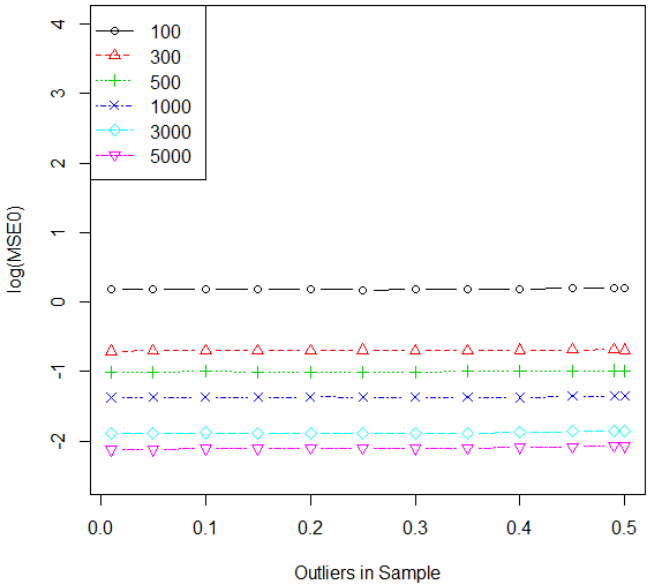

Fig. 4.38.: log(MSE0) for weighted robust linear regression with Hampel function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

Table 4.35.: $n = 100,000$ and $p = 50$. MSE0 comparison of robust linear regression with Hampel function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

r	0.1%n	0.3%n	0.5%n	1%n	3%n	5%n
outlier	(100)	(300)	(500)	(1000)	(3000)	(5000)
1%	1.5096	0.1966	0.0981	0.0421	0.0129	0.0076
5%	1.5496	0.2011	0.0992	0.0428	0.0128	0.0076
10%	1.5194	0.1994	0.1006	0.0425	0.0132	0.0079
15%	1.5428	0.2011	0.099	0.0425	0.0129	0.0079
20%	1.556	0.1978	0.0983	0.0432	0.0131	0.008
25%	1.4878	0.202	0.0977	0.043	0.0129	0.008
30%	1.5417	0.201	0.0984	0.0428	0.0131	0.0078
35%	1.5135	0.2008	0.1012	0.0426	0.013	0.008
40%	1.5392	0.2015	0.1002	0.0422	0.0133	0.0081
45%	1.5684	0.2051	0.1021	0.0443	0.0138	0.0084
49%	1.5813	0.2069	0.1034	0.0444	0.0139	0.0086
50%	1.5614	0.2033	0.1009	0.0446	0.0139	0.0085

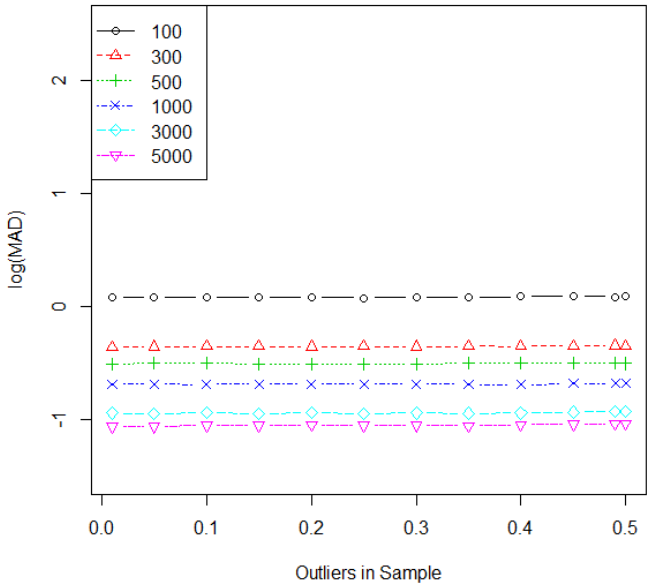

Fig. 4.39.: log(MAD0) for weighted robust linear regression with Hampel function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

Table 4.36.: $n = 100,000$ and $p = 50$. MAD0 comparison of robust linear regression with Hampel function using A-optimal probability for different subsample sizes r and different proportions of outliers in y direction.

r	0.1%n	0.3%n	0.5%n	1%n	3%n	5%n
outlier	(100)	(300)	(500)	(1000)	(3000)	(5000)
1%	1.1952	0.4369	0.3085	0.205	0.1136	0.0867
5%	1.2134	0.4405	0.3132	0.2044	0.1119	0.0864
10%	1.2085	0.4421	0.3146	0.204	0.1142	0.0888
15%	1.2076	0.4438	0.3102	0.2053	0.112	0.0886
20%	1.2172	0.4398	0.3108	0.2061	0.1143	0.0893
25%	1.1845	0.4444	0.311	0.2046	0.1131	0.0884
30%	1.2124	0.4398	0.3095	0.2045	0.1137	0.0881
35%	1.1967	0.4436	0.315	0.2044	0.1133	0.088
40%	1.2182	0.4435	0.3131	0.2028	0.1137	0.0896
45%	1.2232	0.4447	0.3146	0.2088	0.1161	0.0904
49%	1.2175	0.4481	0.317	0.2079	0.1179	0.0918
50%	1.2205	0.4449	0.3123	0.2082	0.1168	0.0917

Hampel function is similar to bisquare function. The outliers with extreme residules will receive a 0 weight. We can see that, for MSEs, there is a huge increase near 49% or 50% outliers. But for MSE0s or MAD0s, There're not so much change near 49% or 50% outliers. It doesn't perform stable when the percentage of outliers is close to 50%. This is because we can't distinguish between outliers and data points from original distribution. The breakdown point of weighted subsample estimator for A-optimal robust linear regression with Hampel function for this case is close to 0.5.

5. REAL DATA ANALYSIS

5.1 Beijing Multi-Site Air-Quality Data

We found this data on UCI Machine Learning Repository. This is a hourly air pollutants data set for 12 air-quality monitoring sites in Beijing. The data was collected from 03/01/2013 to 02/28/2017. 'NA' in the data set indicates missing values. There are 4 time variables(year, month, day, hour), 6 air pollutants(PM2.5: PM2.5 concentration, PM10: PM10 concentration, SO2: SO2 concentration, NO2: NO2 concentration, CO: CO concentration, O3: O3 conce) and 6 relevant meteorological variables, where units of concentration are $(\mu g/m^3)$. We use PM2.5 as the response variable and it measures the atmospheric PM, or fine particles, that are smaller than 2.5 μm in diameter. Due to their small size and light weight, fine particles in the air have a higher chance of being inhaled by humans and can enter deep into the lung or even the circulatory system. Fine particles are associated with many adverse health outcomes such as asthma, heart attack, bronchitis, lung cancer, premature death, birth defects, and ect.

There are 420768 observations in total. We removed 8739 observations due to missing value of the response variable PM2.5. The final sample size is 412029. For variables PM10, SO2, NO2, CO, O3, TEMP, PRES, DEWP, RAIN and WSPM, the missing values are imputed by the corresponding medians. For the categorical variable wind direction which has 16 levels, 7 levels including west(W) direction are combined to a new level "west" while the rest levels are combined to the other level "non-west". Variable station is the air-quality monitoring site with 12 levels, and dummy variables were created for it in analysis. Finally, the number of variables is $p = 26$.

After the application of the linear regression, the studentized residual plot is used to show outliers in y direction. Observations with standardized residuals beyond +/-4 are outliers.

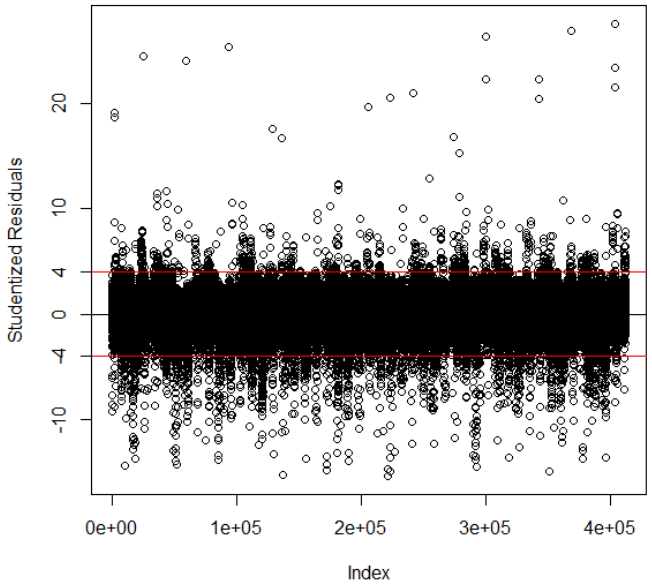

Fig. 5.1.: Studentized residual plot for linear regression

We apply A-optimal and Uniform subsampling methods to the data set with subsample size $r = 412(0.1\%n), 1236(0.3\%n), 2060(0.5\%n), 4120(1\%n), 12361(3\%n), 20601(5\%n)$. Number of repetitions is $M = 1000$. For each subsampling method and each subsample size, we calculate MSE and the approximated MAD as follows:

$$MSE(\boldsymbol{\beta}_r^*) = \frac{1}{M}\sum_{m=1}^{M} \|\boldsymbol{\beta}_{r,m}^* - \hat{\boldsymbol{\beta}}_n\|^2,$$

$$MAD(\boldsymbol{\beta}_r^*) = median(\|\boldsymbol{\beta}_{r,m}^* - \hat{\boldsymbol{\beta}}_n\|).$$

For robust linear regression with bisquare function, we have the following MSE results. We can see that the A-optimal subsampling method in robust linear regression has the smallest MSEs for each subsample size r for this dataset. It performs the best in reducing MSE in the presence of outliers among the four different methods. Uniform subsampling distribution in linear regression has the largest MSEs for each subsample size r for this dataset. It performs the worst, for this data, in reducing MSE in the presence of outliers among the four different methods.

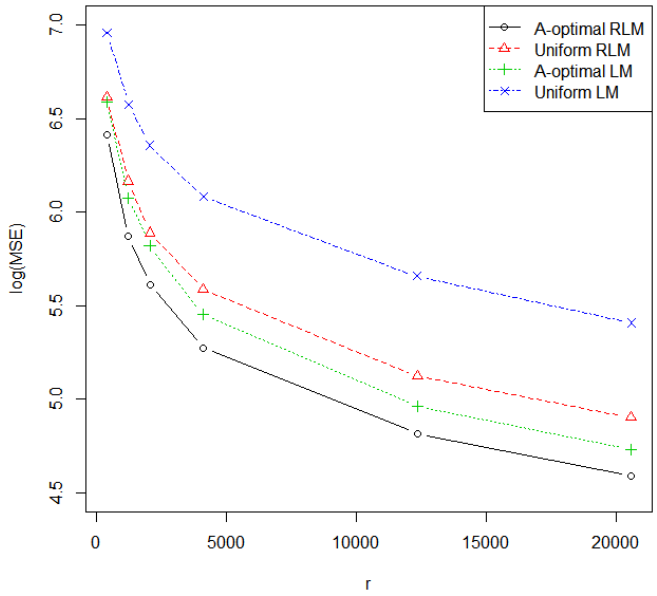

Fig. 5.2.: log(MSE) for linear regression and robust linear regression with Bisquare function using A-optimal and Uniform subsampling distributions for different subsample size r.

Table 5.1.: Comparison of MSEs for linear regression and robust linear regression with Bisquare function using A-optimal and Uniform subsampling distributions for different subsample size r.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.1%	412	2588817.1724	4114827.7885	3852783.6636	9063634.956
0.3%	1236	742696.2441	1463885.6896	1191795.308	3735844.5322
0.5%	2060	409862.4359	772042.6982	659470.0363	2272993.9797
1%	4120	188448.3827	386179.655	282806.6555	1207047.354
3%	12361	65302.6586	133343.6885	91751.8475	454005.487
5%	20601	38740.7249	79852.6902	53939.6756	255814.2036

For bisquare function, we have the following MAD results. We can see that the A-optimal subsampling method in robust linear regression has the smallest MAD for each

subsample size r for this dataset. It is better in the presence of outliers compared to the other 3 methods.

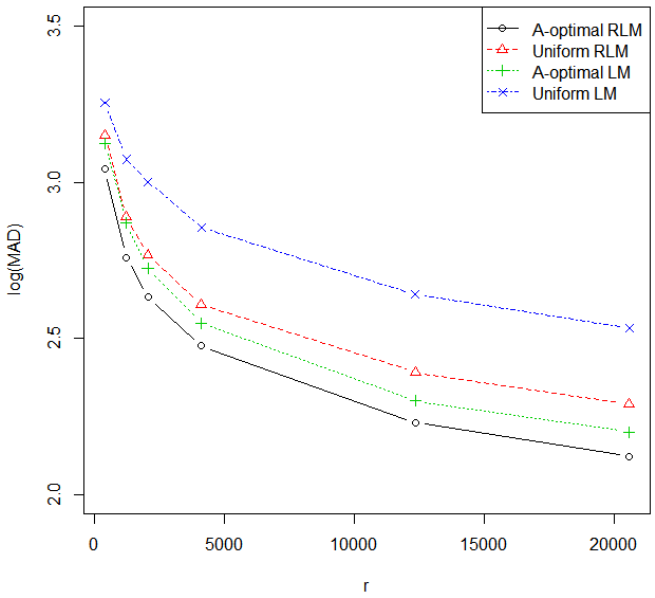

Fig. 5.3.: log(MAD) for linear regression and robust linear regression with Bisquare function using A-optimal and Uniform subsampling distributions for different subsample size r.

Table 5.2.: Comparison of MADs for linear regression and robust linear regression with Bisquare function using A-optimal and Uniform subsampling distributions for different subsample size r.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.1%	412	1100.4583	1413.255	1326.2759	1799.5802
0.3%	1236	574.3529	774.7698	742.1133	1183.6729
0.5%	2060	428.1704	584.8767	530.5962	1003.5927
1%	4120	298.9066	404.7682	354.1132	716.205
3%	12361	170.0925	246.2049	199.4815	437.2804
5%	20601	132.3733	194.9058	158.5105	342.1061

For robust linear regression with Huber function, we have the following MSE results. We can see that the A-optimal subsampling method in robust linear regression has the smallest MSE for each subsample size r for this dataset. It performs slightly better than A-optimal subsampling method in linear regression. This is because that Huber function has a constant value for large residuals.

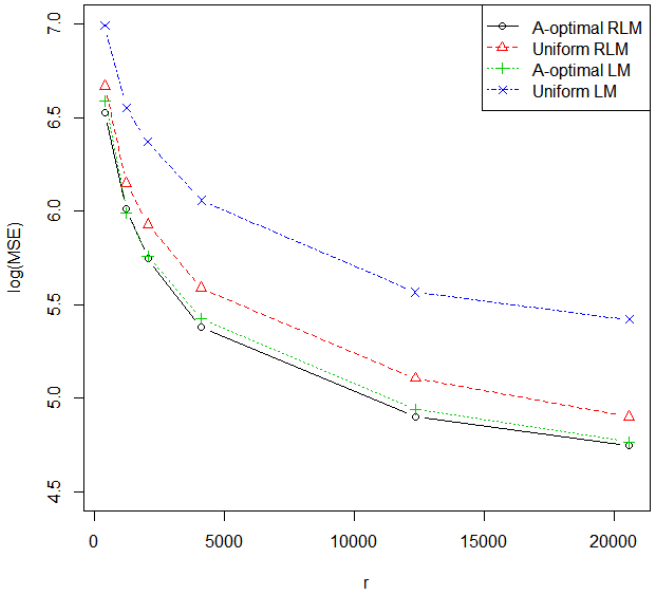

Fig. 5.4.: log(MSE) for linear regression and robust linear regression with Huber function using A-optimal and Uniform subsampling distributions for different subsample size r.

Table 5.3.: Comparison of MSEs for linear regression and robust linear regression with Huber function using A-optimal and Uniform subsampling distributions for different subsample size r.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.1%	412	3347061.5583	4658950.4226	3879368.2803	9816507.1163
0.3%	1236	1028945.1942	1408661.1826	980506.6688	3543168.0893
0.5%	2060	558384.7192	848506.0863	573216.2391	2343346.1903
1%	4120	239141.5048	388542.4508	265183.1937	1140838.7643
3%	12361	79349.4698	127805.9036	87159.7014	368142.7116
5%	20601	55773.9716	79441.4473	58223.5929	262954.3229

For Huber function, we have the following MAD results. We can see that the A-optimal subsampling method in robust linear regression has the smallest MAD for each subsample size r for this dataset. It is slightly better in the presence of outliers than A-optimal subsampling method in linear regression.

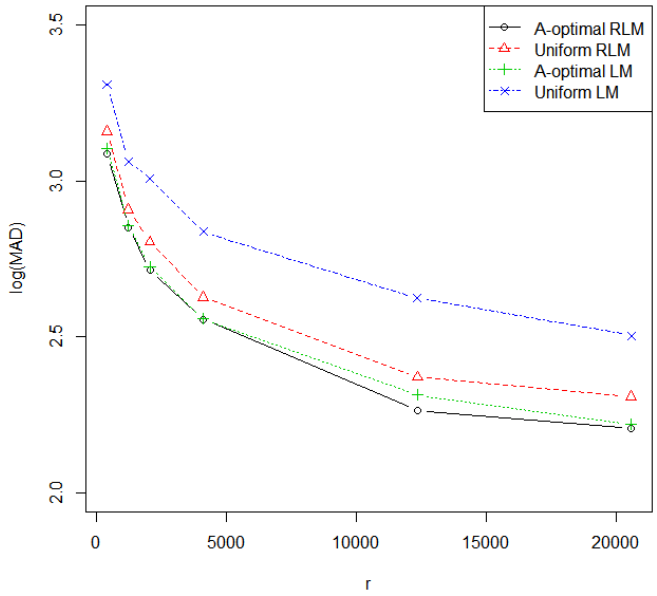

Fig. 5.5.: log(MAD) for linear regression and robust linear regression with Huber function using A-optimal and Uniform subsampling distributions for different subsample size r.

Table 5.4.: Comparison of MADs for linear regression and robust linear regression with Huber function using A-optimal and Uniform subsampling distributions for different subsample size r.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.1%	412	1216.0778	1439.3713	1268.9915	2036.7866
0.3%	1236	709.7398	808.8446	719.242	1150.7702
0.5%	2060	515.9131	636.9039	531.3388	1018.8622
1%	4120	360.1655	423.4923	360.552	687.4179
3%	12361	183.8277	234.5416	205.7784	420.2502
5%	20601	161.2337	203.1769	166.0725	318.9157

For Hampel function, we have the following MSE results. We can see that the A-optimal subsampling method in robust linear regression has the smallest MSE for each subsample size r for this dataset. Although Hampel function is also a redescending func-

tion, the outliers in this dataset are not so extreme. So the performance of Hampel function is similar to the performance of Huber function. It performs slightly better than A-optimal subsampling method in linear regression.

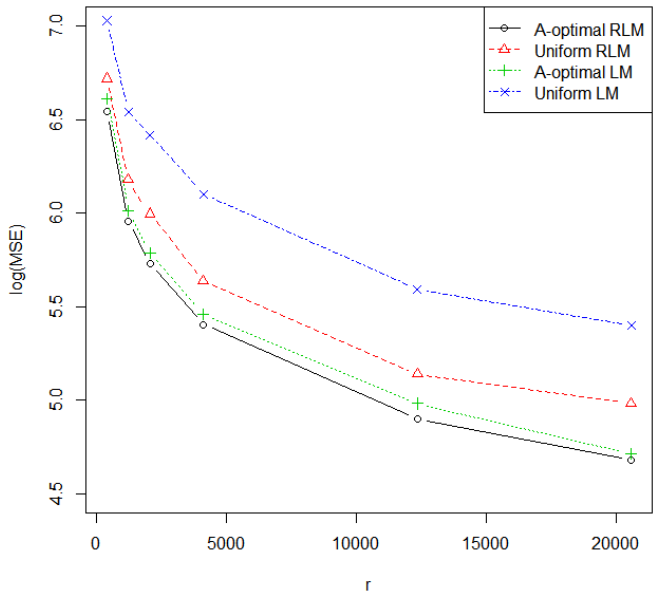

Fig. 5.6.: log(MSE) for linear regression and robust linear regression with Hampel function using A-optimal and Uniform subsampling distributions for different subsample size r.

Table 5.5.: Comparison of MSEs for linear regression and robust linear regression with Hampel function using A-optimal and Uniform subsampling distributions for different subsample size r.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.1%	412	3478805.0519	5217878.3421	4075391.6164	10666064.9054
0.3%	1236	907763.7299	1507597.2494	1022587.2296	3464293.4163
0.5%	2060	534597.3838	988622.8781	613061.0375	2616301.9548
1%	4120	253989.0098	432827.1309	286397.4529	1256585.9568
3%	12361	79203.8428	137700.8033	95914.5405	392234.6145
5%	20601	47652.4684	96064.9148	51456.0577	250922.7751

For Hampel function, we have the following MAD results. We can see that the A-optimal subsampling method in robust linear regression has the smallest MAD for each subsample size r for this dataset. It is slightly better in the presence of outliers than A-optimal subsampling method in linear regression.

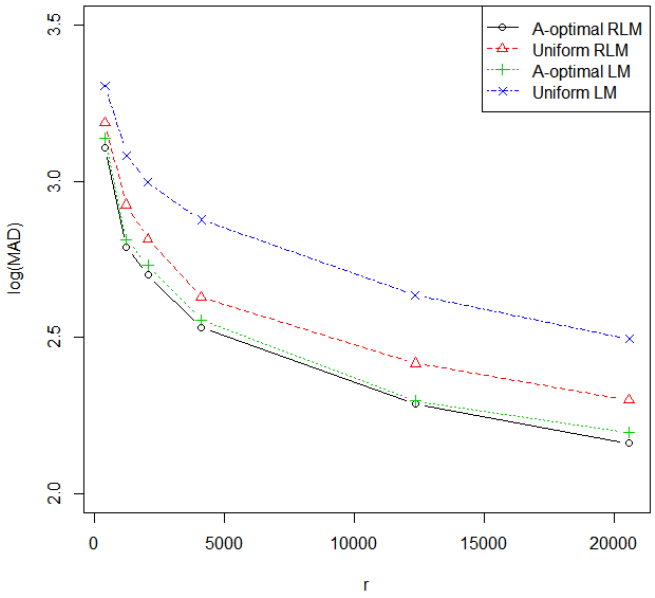

Fig. 5.7.: log(MAD) for linear regression and robust linear regression with Hampel function using A-optimal and Uniform subsampling distributions for different subsample size r.

Table 5.6.: Comparison of MADs for linear regression and robust linear regression with Hampel function using A-optimal and Uniform subsampling distributions for different subsample size r.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.1%	412	1280.1732	1541.1313	1371.0452	2023.5457
0.3%	1236	615.9271	840.268	648.0935	1209.7804
0.5%	2060	499.9069	652.1511	538.9018	993.3874
1%	4120	341.0216	424.5177	358.9226	752.3638
3%	12361	193.3214	261.0656	198.0421	432.3086
5%	20601	144.872	199.7202	156.8613	313.2319

The MSE and MAD of A-optimal subsampling estimate in robust linear regression are the smallest among the four methods. So, the A-optimal subsampling method in robust linear regression is a good candidate for analyzing big data with outliers in real world.

5.2 Gas Sensor Array Data Set

We found this data on UCI Machine Learning Repository. In this data set, values from 16 chemical sensors were measured when gas mixture Ethylene and Methane in air varies at different concentration levels randomly. The 16-sensor array signals were obtained continuously for 12 hours. This data set contains variables time (seconds), Methane or Ethylene concentration set point (ppm), and 16 readings of the chemical sensors. Here we use Methane concentration (ppm) as the response variable and recordings from 16 chemical sensors as independent variables. There are $n = 4,178,504$ observations in total in this data set. $p = 16$.

After the application of the linear regression, the studentized residual plot is used to show outliers in y direction. Observations with standardized residuals beyond +/-4 are outliers.

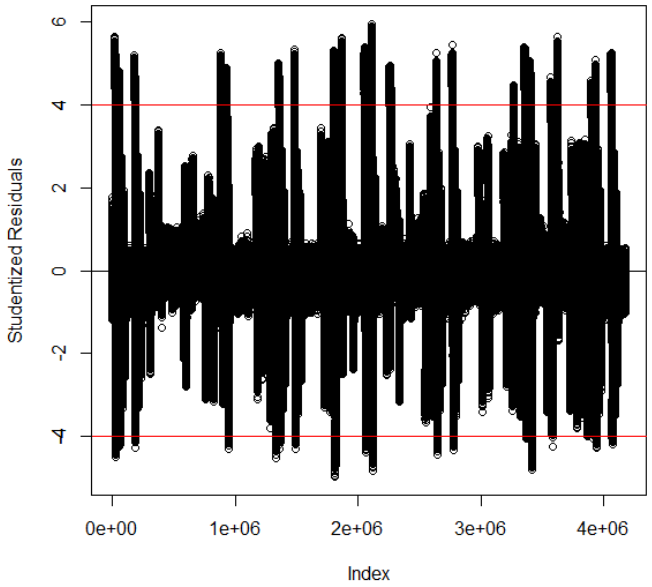

Fig. 5.8.: Studentized residual plot for linear regression

We apply A-optimal and Uniform subsampling methods to the data set with subsample size $r = 418(0.01\%n), 836(0.02\%n), 1254(0.03\%n), 1671(0.04\%n), 2089(0.05\%n)$. Number of repetitions is $M = 1000$. For each subsampling method and each subsample size, we calculate MSE and the approximated MAD as follows:

$$MSE(\boldsymbol{\beta}_r^*) = \frac{1}{M}\sum_{m=1}^{M} \|\boldsymbol{\beta}_{r,m}^* - \hat{\boldsymbol{\beta}}_n\|^2,$$

$$MAD(\boldsymbol{\beta}_r^*) = median(\|\boldsymbol{\beta}_{r,m}^* - \hat{\boldsymbol{\beta}}_n\|).$$

Here we use bisquare function as the objective function in robust linear regression. We have the following MSE results. We can see that the A-optimal subsampling method in robust linear regression has the smallest MSEs for each subsample size r for this dataset among the four different methods. It performs the best in reducing MSE when outliers are included in this data set. Uniform subsampling distribution in linear regression has the largest MSEs for each subsample size r for this dataset among the four different methods. It performs the worst in reducing MSE when outliers are included in this data set.

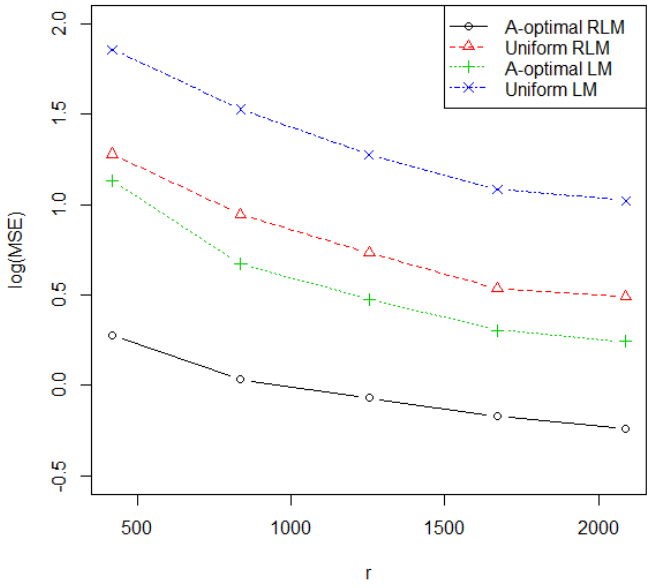

Fig. 5.9.: log(MSE) for linear regression and robust linear regression with Bisquare function using A-optimal and Uniform subsampling distributions for different subsample size r.

Table 5.7.: Comparison of MSEs for linear regression and robust linear regression with Bisquare function using A-optimal and Uniform subsampling distributions for different subsample size r.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	418	1.9044	18.996	13.6011	71.8033
0.02%	836	1.0786	8.8036	4.6873	33.715
0.03%	1254	0.8535	5.4304	2.9869	18.9454
0.04%	1671	0.6731	3.4345	2.0191	12.2017
0.05%	2089	0.58	3.1152	1.7514	10.541

For bisquare function, we have the following MAD results. We can see that the A-optimal subsampling method in robust linear regression has the smallest MAD for each

subsample size r for this dataset among the four different methods. It has the best performance in the presence of outliers for this dataset.

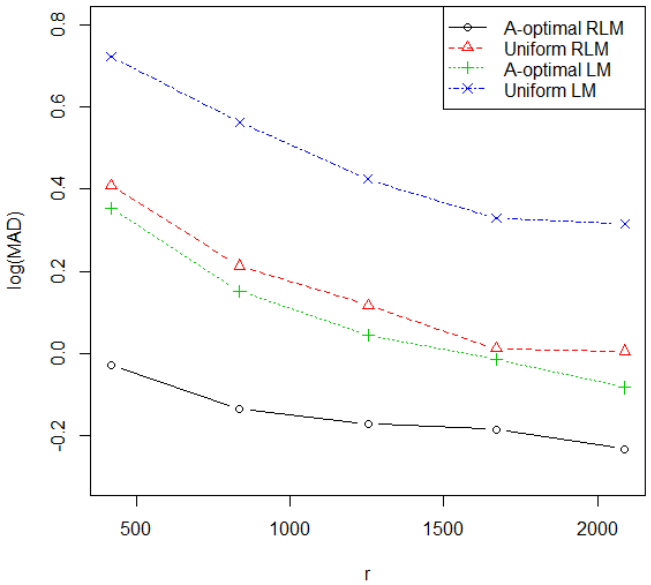

Fig. 5.10.: log(MAD) for linear regression and robust linear regression with Bisquare function using A-optimal and Uniform subsampling distributions for different subsample size r.

Table 5.8.: Comparison of MADs for linear regression and robust linear regression with Bisquare function using A-optimal and Uniform subsampling distributions for different subsample size r.

r/n	r	A-optimal RLM	Uniform RLM	A-optimal LM	Uniform LM
0.01%	418	0.9362	2.5665	2.2622	5.2854
0.02%	836	0.7338	1.6352	1.4171	3.6547
0.03%	1254	0.6742	1.3103	1.1096	2.6551
0.04%	1671	0.6545	1.0281	0.9694	2.1381
0.05%	2089	0.585	1.0133	0.8266	2.0703

The MSE and MAD of A-optimal subsampling estimate in robust linear regression are the smallest among the four methods. So, the A-optimal subsampling method in robust linear regression is a good candidate for analyzing big data with outliers in real world.

REFERENCES

REFERENCES

[1] Alma, O. G. (2011). Comparison of robust regression methods in linear regression. *International Journal of Contemporary Mathematical Sciences*, 6: 409– 421.

[2] Bickel, P., Gotze, F. and van Zwet, W. (1997). Resampling fewer than n observations: Gains, losses, and remedies for losses. *Statistica Sinica*, 7: 1-31.

[3] Chen C., Robust regression and outlier detection with the Robustreg procedure. *Statistics and Data Analysis*, 265-27.

[4] Chung K. L., (2001). A course in probability theory. *New York: Academic Press*

[5] Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown-point. *In A Festschrift for Erich L. Lehmann* 157–184.

[6] Draper, N. R. and Smith, H. Applied regression analysis, *Wiley Interscience Publication, United States, 1998*.

[7] Fan, J., Han, F. and Liu, H. (2014). Challenges of big data analysis. *National Science Review*, 1: 293-314.

[8] Gad, M. A. and Qura, E. M. (2016). Regression estimation in the presence of outliers: a comparative study. *International Journal of Probability and Statistics*, 5(3): 65-72.

[9] Huber, P. J. (1981). Robust statistics. *John Wiley and Sons, New York, 1981*.

[10] Hampel, F., Ronchetti, E., Rousseeuw, P., and Stahel, W. (1986). Robust statistics: The approach based on influence functions. *Wiley, New York*.

[11] Khalil, U., Khan, D.M., Khan, S.A., Ali, A., Alamgir and Qadir, F. (2016). Efficient Uk's re-descending M-estimator for robust regression. *Pakistan Journal of Statistics*, 32(2): 125-138.

[12] Ma, P. and Sun, X. (2014). Leveraging for big data regression. *Computational Statistics*, 7 (1): 70-76

[13] Mahoney, M. W. (2011). Randomized algorithms for matrices and data. Foundations and Trends in Machine Learning. NOW Publishers, Boston.

[14] Muthukrishnan, R. and Radha, M. (2010). M-Estimators in regression models. *Journal of Mathematics Research*, 2(4).

[15] Peng, H. and Tan, F. (2018). A big data linear regression via A-optimal subsampling. *Preprint*.

[16] Portnoy, S. (1984a). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I: Consistency. *Ann. Statist.*, 12(4) 1298-1309.

[17] Portnoy, S. (1985). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. II: Normality. *Ann. Statist.*, 13(4) 1403-1417.

[18] Pitselis, G. (2012). A review on robust estimators applied to regression credibility. *Journal of Computational and Applied Mathematics*, 231–249.

[19] Rousseeuw, P. J. and Leroy, A. M. (1987). Robust regression and outlier detection. *Wiley, New York.*

[20] Salibian-Barrera, M. and Zamar, R.H. (2002) Bootstrapping robust estimates of regression. *Ann. Statist.*, 30: 556–582

[21] Singh, K. (1998). Breakdown theory for bootstrap quantiles. *Ann. Statist.*, 26: 1719–1732.

[22] Susanti, Y., Pratiwi, H., Sulistijowati, H. S., and Liana, T.,(2014). M estimation, S estimation, and MM estimation in robust regression. *International Journal of Pure and Applied Mathematics*.

[23] Van der Vaart, A. W. (1998), Asymptotic Statistics, *Cambridge University Press, London.*

[24] Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.,* 15: 642–656

[25] You, J. (1999). A monte carlo comparison of several high breakdown and efficient estimators. *Computational Statistics and Data Analysis*, 30: 205–219.

[26] Yu, C. and Yao, W. (2017). Robust linear regression: A review and comparison. *Communications in Statistics - Simulation and Computation*, 46(8): 6261-6282.

[27] Zhu, R., Ma, P., Mahoney, M. W. and Yu, B. (2015). Optimal subsampling approaches for large sample linear regression. *arXiv:1509.0511.v1* [stat.ME].

VITA

VITA

My name is Ziting Tang. I received my bachelor's degree in mathematics from Shandong Normal University in 2011. I studied Finsler geometry and obtained my master's degree in mathematics from Zhejiang University in 2014. After that, I continued to pursue my PhD in IUPUI.