**Leverage Sampling for Single-Index Models**

by

Basmah Almutairi

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
May 1, 2021

Keywords: Generalize Leverage, Single-Index Models(SIM), Sampling, Sub-sampling, robust estimation



Approved by

Asheber Abebe, Chair, Professor of Mathematics and Statistics
Guanqun Cao, Associate Professor of Mathematics and Statistics
Philippe Gaillard, Associate Professor of Mathematics and Statistics

Abstract

In this thesis, a generalized leverage-based sub-sampling method for single-index models is proposed. The approach gives more efficient estimators than random sub-samples of the same size. Also, robust rank-based estimators of single-index models using leverage sub-samples provide estimators that are robust to outliers and heavy tails. A common bottleneck for rank-based estimators is the lack of computational efficiency, which is overcome using sub-samples. A simulation study was performed and, as expected the rank-based index direction estimator was comparable to the least squares index direction estimator when the errors follow a normal distribution. However, the rank-based index direction estimator was more efficient when the data followed a heavy-tailed error distribution. Finally, the results from a real data example are presented to highlight the performance of the proposed estimators.

Acknowledgments

I will begin by thanking God for strength and health. Then, I am grateful for the financial support from Al-Majmaah University and the Saudi Government.

To my advisor, Dr. Asheber Abebe, I am thankful for always being available with guidance, support, help, and encouragement, without your advice this dissertation would not have been possible.

I would also like to thank Dr. Guanqun Cao and Dr. Philippe Gaillard for their contribution and for taking the time to serve as committee members.

Special thanks to my parents and family whom always stand up for me with unlimited help.

To my husband, Abdulrahman, who represents the key to my success by providing all necessary advice and encouragement. Finally, I would love to thank my daughters, Ghadah, Shatha, and Lubna, for their love and for sharing this journey with me.

Table of Contents

List of Figures

List of Tables

# Chapter 1

# Introduction

## 1.1 Background

Data science has become a prominent tool for decision making and tackling problems faced by society. The ubiquity of big data, including streaming data, has made finding algorithms that efficiently use a small part of the data to provide useful inference extremely important. One approach is based on subsampling including leverage sampling. Such methods have been studied only for linear models. In this study, we considered a more complex semi-parametric regression setup to evaluate the usefulness of leverage sampling approaches in estimation.

In this thesis, we propose a method that employs a generalized leverage weighted random sampling approach for single-index models. This model is especially amicable for such an approach due to its dependence on an inner product of the regressor $x$ and the parameter $\theta$ which is then transformed non-parametrically. We will show how this leads to a reduced mean squared error (MSE) of the estimates in comparison with random sub-samples of the same size. We also propose a rank-based version of this same leverage sampling approach for single-index models to minimize the impact of outliers without significant loss in efficiency.

The main results of this study are as follows:

(a) We propose a generalized leverage based subsampling method for single-index models that gives more efficient estimators than random sub-samples of the same size.

(b) We extend these results to robust rank-based estimators of single-index models to provide estimators that are robust to outliers and heavy tails.

(c) We demonstrate that the approach has high computational efficiency.

This is important, especially for rank-based estimators, since a huge drawback of the methods is the computational inefficiency which increases as a function of the sample size.

### 1.2 Literature Review

Increase in data collection and the use of technology has led to availability of massive amounts of data. For society to reap the benefits from these data through the improvement of the quality of both government and industry, it has increased investment in big data and many universities and researchers have embarked on the analysis of big data and its translation to aspects of science and technology. Previous studies have investigated leverage sub-sampling methods in the linear model for streaming data. However, this has not been studied for more complex structures. We argue that the idea of generalized leverage can provide a means for such cases and that this can be extended to robust models.

In this work we consider the generalized leverage described in Wei and Hu (1998) and defined as

$$GL = \frac{\partial \hat{Y}}{\partial Y^T} = \left( \frac{\partial \hat{Y}_i}{\partial Y_j} \right)_{n \times n} \tag{1.1}$$

Researchers have evaluated leverage points and local influence in the context of the general M-estimation problem. Leverage points play a significant role on model fitting on linear and non-linear regression and points with large generalized leverage have a great influence on the fitted model (Wei and Hu, 1998). The study, however, is limited in its application as they chose to focus on the exponential family generalized linear model. It is of interest to expand

the scope of the study to other distributions and study the relationship between generalized leverage and local influence on the fitted model.

Subsampling methods have garnered attention in recent years for the analysis of large streaming data. More recent attention has focused on the provision of leveraging methods and how much influence on model fitting especially in big data (Ma and Sun, 2015). Ma et al. (2015) provide a review of the use of weighted leverage methods and unweighted leverage methods for subsampling. The weighted leverage method is a sampling from the full sample using small sub-sample when points with high leverage value have a high chance to be selected. It is one of the more successful technique for estimating and produce a model Ma et al. (2015) for linear models. The process involves using the weighted leverage algorithm to produce a sampling probability to be included as weights in a regression problem Ma et al. (2015). This approach was shown to provide efficient estimates while at the same tim reducing the computational expense for big data analytics (Ma et al.,2015).

McWilliam et al. (2014) is an important study on how the weighted sub-sampling algorithm can lead to fast and robust estimation in the linear model. Generally, these studies highlight the beneficial effects of leverage sampling methods on model fitting to help solve the big data problems.

Rank estimation has been used for estimation in regression problems. A general approach based on the rank pseudo norm was proposed by Jaeckel (1972) and Jureckova (1971). This method results in robust and efficient estimates of regression parameters (Hettemanperger and McKean, 2011). This method has been used for general nonlinear models (Bindele and Abebe, 2012), generalized linear models (Miakonkana and Abebe, 2014), partial linear models (Bindele and Abebe, 2015), among others. For single-index models, Bindele at al. (2018) established the theory of rank estimators and this has been utilized for cases with

missing data (Abebe et al., 2020) as well as variable selection (Bindele et al., 2019). They found that the rank regression estimation approach can improve the model fit while providing a robust and efficient alternative compared to the least squares regression approach that is more sensitive to outliers and high leverage points. This thesis will provide a study of the use of subsampling for single index models, including for rank estimation.

### 1.3 Organization

The thesis is organized as follows. In Chapter 2, we state the review the main concept of regression analysis and discuss in detail the methods which are used in this study. We also provide a computation algorithm for obtaining the generalized leverage sampling for the single-index model. In Chapter 3, we start the chapter by giving a brief summary of the sampling methods and then we compare the two ways of sampling methods (random and leverage) and explain why sampling based on leverage weight is more accurate method to use. In Chapter 4, we present a simulation study and a real world application. This is followed by the conclusion and recommendation in Chapter 5.

# Chapter 2

## Regression Analysis

Regression analysis is one of the most well-known tools for analysis, estimation, modeling and prediction in statistics.

Suppose we have a linear regression model,

$$y = X\beta + \varepsilon \ , \tag{2.1}$$

where $y = (y_1, \ldots, y_n)^T$ is the vector of response variable $(n \times 1)$, $X = (x_1, \ldots, x_p)^T$ is the design matrix of the predictor variables $(n \times p)$, $\beta = (\beta_1, \ldots, \beta_p)^T$ is the vector of regression parameters, which is an unknown constant $(1 \times p)$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is a random errors vector. It is often assumed that the errors follow a normal distribution with zero mean and finite variance. We will not need that assumption in this study.

### 2.1 Least Squares (LS)

We will start by reviewing the least squares method estimates the regression parameters by minimizing the Euclidean norm of the errors. So, the estimator of $\beta$, $\widehat{\beta}$, is defined as:

$$\widehat{\beta} = \text{argmin}_\beta \|Y - X\beta\|_2 \tag{2.2}$$

which gives

$$\widehat{\beta}_{LS} = (X^T X)^{-1} X^T Y \ . \tag{2.3}$$

The least squares method is optimal when the errors follow the normal distribution. However, a limitation of this approach is that it is not robust. The least squares method is easily influenced by outliers and high-leverage points. Robust regression methods may be used to fit the linear regression model in a manner where the negative consequences of outliers and high-leverage points are avoided.

The leverage values for the LS estimation are represented as the diagonal of the $n \times n$ matrix $\partial \hat{Y} / \partial Y$. Since $\hat{Y} = X\hat{\beta}$, this is given by

$$H = X(X^T X)^{-1} X^T \ .$$

For weighted regression with weight matrix $W$, the estimate of $\beta$ is $\widehat{\beta}_{LS} = (X^T W X)^{-1} X^T W Y$ and the corresponding leverage matrix is

$$H_W = X(X^T W X)^{-1} X^T W \ .$$

## 2.2 Robust regression methods

In practice, the response variable $Y$ rarely follows the normal distribution, where the least squares method is optimal. The distribution may have outliers which affect the least squares fit. Since this is a common occurrence in real-life data, robust regression methods are often needed. Another instance is when the responses are derived from a heavy-tailed distribution. In heavy-tailed distributions, outliers tend to be generated with a large probability thus affecting the least squares fit. Sensitivity to outliers could also be an issue for streaming data and finding ways such as robust regression to deal with those is highly desirable.

There are several $M$-estimation-based robust estimation techniques such as GM-estimation, S-estimation, M-estimation and MM-estimation. Our focus, however will be on the least absolute deviations (LAD) and rank-based regression ($R$-estimation) techniques. The LAD method is a special case of both $M$- and $R$- estimation techniques.

### 2.2.1 Least Absolute Deviations (LAD)

Historically, research investigating the LAD procedure has focused on both theory, applications, and how by using LAD we can overcome the disadvantages of LS. For example, LAD is a robust method and it is not sensitive to outliers compared to the LS method which is a strongly affected by outliers. A review of the academic literature on LAD is found in Dasgupta et al.(2004).

The major difference between the LS and LAD methods is that the LS minimizes the mean of squared errors (MSE) while the LAD minimizes the mean absolute values of errors (MAE). These are described mathematically as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \tag{2.4}$$

and

$$MAE = \frac{1}{n}\sum_{i=1}^{n} \mid y_i - \widehat{y}_i \mid \tag{2.5}$$

More details on the LAD method including asymptotic theory can be found in Bassett and Koenker (1978), Chen et al. (1990) and Pollard (1991).

### 2.2.2 Rank-Based estimation(R)

Rank-based (R) estimation was developed as a non-parametric and robust approach for regression analysis. It was proposed first by Jurecková (1971) and Jaeckel (1972). This was then extended by several researchers in various directions. The robustness and adaptability of the rank-based approach can provide a key tool for solving important problems in machine learning and algorithms. Here, in this study we will give a short review of this method. More details can be found in the books Hettmansperger and McKean (2011) and Kloke and McKean (2014).

$R$-estimators are found by minimizing the quantity defined for $\mathbf{u} = (u_1, \ldots, u_n)^T \in \mathbb{R}^n$ as

$$\|\mathbf{u}\|_\varphi = \frac{1}{n} \sum_{i=1}^{n} \varphi\left(\frac{R(u_i)}{n+1}\right) u_i$$

where $\varphi : (0,1) \to \mathbb{R}$ is a continuous non-decreasing function normalized so that $\int \varphi = 0$ and $\int \varphi^2 = 1$ and the function $R(\cdot)$ represents the rank function given by $R(u_t) = \sum_{i=1}^{n} \mathbb{1}\{u_i \leqslant u_t\}$. We can show that $\|\mathbf{u}\|_\varphi$ is a pseudo-norm on $\mathbb{R}^n$. It satisfies all the properties of the norm except the point separating property $\|\mathbf{u}\| = 0 \Leftrightarrow \mathbf{u} = \mathbf{0}$. In this case

$$\|\mathbf{u}\|_\varphi = 0 \;\Leftrightarrow\; u_1 = u_2 = \cdots = u_n \,.$$

It is easy to show that

$$\|\mathbf{u}\|_\varphi = \frac{1}{n} \sum_{i=1}^{n} \varphi\left(\frac{i}{n+1}\right) u_{(i)}$$

where $u(i)$ is the $i$th order statistics of $(u_1, \ldots, u_n)$. Jaeckel (1972) used this to show that for a regression problem

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$$

the quantity

$$\|Y_i - \mathbf{x}_i^T\boldsymbol{\beta}\|_\varphi = \frac{1}{n}\sum_{i=1}^{n}\varphi\left(\frac{R(Y_i - \mathbf{x}_i^T\boldsymbol{\beta})}{n+1}\right)(Y_i - \mathbf{x}_i^T\boldsymbol{\beta})$$

is a non-negative, continuous, and convex function of $\boldsymbol{\beta}$. Thus this can be minimized to find an estimate of $\boldsymbol{\beta}$. However, since $\|\cdot\|_\varphi$ does not have the point separating property and it is invariant to constant translations $\|\mathbf{u}\|_\varphi = \|\mathbf{u} \pm \mathbf{1}c\|_\varphi$, $c \in \mathbb{R}$, it cannot be used to estimate the intercept term.

To show that $R$-estimators can give LAD estimators, we note that $sgn(u)u = |u|$. Thus taking $\varphi(u) = sgn(u-1/2)$ in $R$-estimation will give the LAD estimator. So, in the following, we will treat only the $R$-estimators since the LAD is a byproduct.

Using $\int\varphi = 0$ Sievers and Abebe (2004) showed that the rank estimator can be computed as a sequence of weighted LS problems that minimize

$$\sum_{i=1}^{n} w_i(\boldsymbol{\beta})(Y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2$$

where

$$w_i(\boldsymbol{\beta}) = \begin{cases} \frac{\varphi\left(\frac{R(Y_i - \mathbf{x}_i^T\boldsymbol{\beta})}{n+1}\right)}{Y_i - \mathbf{x}_i^T\boldsymbol{\beta} - m} & \text{if } |Y_i - \mathbf{x}_i^T\boldsymbol{\beta} - m| > \delta \\ 0 & \text{otherwise} \end{cases}$$

and $m = \text{median}(Y_i - \mathbf{x}_i^T\boldsymbol{\beta})$ and $\delta$ a small tolerance. Given an initial estimate $\widehat{\boldsymbol{\beta}}^{(0)}$, the $R$-estimate of $\boldsymbol{\beta}$ can be iteratively computed as

$$\widehat{\boldsymbol{\beta}}^{(k)} = [X^T W(\widehat{\boldsymbol{\beta}}^{(k-1)})X]^{-1}X^T W(\widehat{\boldsymbol{\beta}}^{(k-1)})Y$$

with the corresponding leverage iteratively computed as

$$H^{(k)} = X[X^T W(\widehat{\boldsymbol{\beta}}^{(k)})X]^{-1}X^T W(\widehat{\boldsymbol{\beta}}^{(k)})\,.$$

## 2.3 Single-Index Models (SIM)

In recent years, there has been a increasing interest on the use of SIM in many disciplines including medical science, economics, and finance among others. The main aim of this research is to investigates a new approach for computing the generalized leverage of single-index models and how they can be used for intelligent sampling across non-parametric curves. The SIM is a semiparametric model where a nonparametric function acts upon a linear space. Thus it is amenable for the calculation of generalized leverage. Therefore, in this section we will give the definition of single-index models, their estimation via LS, LAD, and $R$ methods, and their generalized leverage computation. This is followed by the computational algorithm.

### 2.3.1 Definition of SIM

The SIM is a special case of semiparametric regression models and it avoids the curse of dimensionality by considering linear combinations of regressors and nonparametrically transforming the linear combination to "best" fit a response.

The SIM is defined as

$$Y_i = g(\mathbf{x}_i^T \boldsymbol{\beta}) + \varepsilon_i, \qquad 1 \leqslant i \leqslant n \,, \tag{2.6}$$

where $\boldsymbol{\beta} \in \mathcal{B} \subset \mathbb{R}^p$, with $\mathcal{B}$ closed and bounded, is a vector of parameters, $\mathbf{x}_i$ are vectors of regressors in some vector space $\mathcal{X}$, and $g$ is a real-valued function on the space defined by the inner-product of $\mathcal{B}$ and $\mathcal{X}$. It is assumed that $g(\cdot)$ is a three times continuously and boundedly differentiable function and the $\varepsilon_i$ are assumed to be iid. For model identifiability, it is assumed that $\|\boldsymbol{\beta}\| = 1$ and $\beta_p > 0$.

A number of authors have considered the important of using the single-index models (inWu et al, 2010; Bindele et al, 2018, Ash Abebe et al, 2020). Thus far, previous studies highlighted

factors that are associated with applying this approach for more details, it can be found in Wang et al. (1997), Wang et al. (2004), Wu et al. (2010).

### 2.3.2 LS estimation for the SIM

Least squares based estimates for SIM have been studied extensively beginning with Ichimura (1993) who used a Nadaraya-Watson type estimator of the SIM with an addition of a trimming function to avoid boundary effects. Other works include Horowitz and Härdle (1996), Carroll et al. (1997), Delecroix et al. (2006), Xia (2006), Wu et al. (2010a), Kong and Xia (2012), among others. The most common approach for estimating SIM involves the minimum average variance estimation (MAVE) method. For this we note that for $\mathbf{x}_i$ close to $\mathbf{x}$, we have a local linear approximation

$$g(\mathbf{x}_i^T\boldsymbol{\beta}) \approx g(\mathbf{x}^T\boldsymbol{\beta}) + g'(\mathbf{x}^T\boldsymbol{\beta})\mathbf{x}_i^{*T}\boldsymbol{\beta} = a + b\mathbf{x}_i^{*T}\boldsymbol{\beta}$$

where $\mathbf{x}_i^* = \mathbf{x}_i - \mathbf{x}$. MAVE estimates the SIM parameter $\boldsymbol{\beta}$ by minimizing

$$\min_{a,b} \sum_{i=1}^{n} \left\{Y_i - a - b\mathbf{x}_i^{*T}\boldsymbol{\beta}\right\}^2 w_i$$

where $w_i$ are weights such that $\sum w_i = 1$. In practice, we let $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$ and minimize

$$\frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n} \left\{Y_i - a_j - b_j\mathbf{x}_{ij}^T\boldsymbol{\beta}\right\}^2 w_{ij}$$

with respect to $(a_j, b_j)$ and $\boldsymbol{\beta}$. The weight function $w_{ij}$ is defined by

$$w_{ij} = K_h(\boldsymbol{\beta}^T\mathbf{x}_{ij}) / \sum_{j=1}^{n} K_h(\boldsymbol{\beta}^T\mathbf{x}_{ij}) \ ,$$

with $K_h(\cdot) = K(\cdot/h)$, $K$ a kernel function defined on the real line and $h$ the corresponding bandwidth.

This is solved iteratively as

$$\begin{pmatrix} \hat{a}_j \\ \hat{b}_j \end{pmatrix} = \left\{ \sum_{i=1}^{n} w_{ij} \begin{pmatrix} 1 \\ \mathbf{x}_{ij}\boldsymbol{\beta} \end{pmatrix} \begin{pmatrix} 1 \\ \mathbf{x}_{ij}\boldsymbol{\beta} \end{pmatrix}^T \right\}^{-1} \sum_{i=1}^{n} w_{ij} \begin{pmatrix} 1 \\ \mathbf{x}_{ij}\boldsymbol{\beta} \end{pmatrix} Y_i$$

and

$$\widehat{\boldsymbol{\beta}} = \left\{ \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(b_j\mathbf{x}_{ij})(b_j\mathbf{x}_{ij})^T \right\}^{-1} \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(b_j\mathbf{x}_{ij})(Y_i - a_j) \ .$$

These can be represented in matrix notation as $\hat{a}_j = \mathbf{c}_j^T\mathbf{Y}$, $\hat{b}_j = \mathbf{d}_j^T\mathbf{Y}$, and $\widehat{\boldsymbol{\beta}} = T\mathbf{Y}$, where $\mathbf{c}_j$ and $\mathbf{d}_j$ are $n$-vectors and $T$ is a $p \times n$ matrix. At each iteration step, we calculate $\hat{Y}_j \approx \hat{a}_j = \mathbf{c}_j^T\mathbf{Y}$. Letting $\mathbf{C}$ be the $n \times n$ matrix

$$\mathbf{C} = \begin{bmatrix} \mathbf{c}_1^T \\ \vdots \\ \mathbf{c}_n^T \end{bmatrix}$$

we can write $\widehat{\mathbf{Y}} = \mathbf{C}\mathbf{Y}$. Thus the leverage matrix is

$$GL = \partial\widehat{\mathbf{Y}}/\partial\mathbf{Y} = \mathbf{C}$$

which is iteratively updated. Since $\mathbf{C}$ measures the rate of change in $\widehat{\mathbf{Y}}$ as $\mathbf{Y}$ changes, its diagonals tend to be large around the points where the estimated function changes substantially.

### 2.3.3 Rank-Based estimation for the SIM

The R-estimator for the SIM is defined similarly to the MAVE approach given above. To this end, define the residual

$$\xi_i(a_j, b_j, \boldsymbol{\beta}) = Y_i - a_j - b_j \mathbf{x}_{ij}^T \boldsymbol{\beta} \ .$$

The $R$-estimator minimizes

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \varphi \left( \frac{R(\xi_i(a_j, b_j, \boldsymbol{\beta}))}{n+1} \right) \xi_i(a_j, b_j, \boldsymbol{\beta}) w_{ij} \ .$$

Estimation is performed using the iteratively reweighted LS approach given by Sievers and Abebe (2004) where the weight function is now a product of $w_{ij}$ and

$$v_{ij}(\boldsymbol{\beta}) = \begin{cases} \frac{\varphi\left(\frac{R(\xi_i(a_j, b_j, \boldsymbol{\beta}))}{n+1}\right)}{\xi_i(a_j, b_j, \boldsymbol{\beta}) - m} & \text{if } |\xi_i(a_j, b_j, \boldsymbol{\beta}) - m| > \delta \\ 0 & \text{otherwise} \end{cases}$$

and $m = \text{median}\{\xi_i(a_j, b_j, \boldsymbol{\beta})\}$.

Again, generalized leverage is calculated iteratively.

# Chapter 3

# Sub-Sampling

In this chapter, we provide a brief description of the sub-sampling methods used in the study. All computation was performed using the R package.

## 3.1 Introduction

During the last few years, the link between leverage points and sampling efficiency has garnered significant attention. This tracks the advent of big and streaming data where sampling efficiency is desirable. Many researchers have investigated various aspects of this subject. The importance of using the big data in statistics and data science has increased significantly (see, e.g., Ma and Sun (2015); Wei et al. (1998). The data generating mechanisms has also gotten more complex. However, many of the subsampling methods only apply to linear regression models.

Taking a small representative sample to use in computation of the SIM is especially important for the $R$-estimators which are computationally burdensome as they require multiple fits of the least squares procedure. Therefore, in this thesis we investigate two sub-sampling approaches: random sampling and leverage sampling. The latter uses the generalized leverage of SIMs to assign sampling weights to data points.

### 3.2 Random Sub-sampling

In random sub-sampling, every point in the original data has the same probability of being included in the sub-sample.

### 3.3 Leverage Sub-sampling

Let $H$ be the leverage matrix. In leverage sub-sampling the data point $(\mathbf{x}_i, Y_i)$ is included in the sub-sample with probability $H_{ii}/\sum_j H_{jj}$. Thus, for the SIM, higher probability is assigned for data points with high $\mathbf{C}_{ii}$. These are data points for which the estimated function shows a large amount of change.

### 3.4 Error Calculation

For evaluating the efficiency of sub-sampling, we use the MSE and the MAE of the SIM function estimates. We will illustrate this using the MSE. Suppose we have data $(x_1, y_1), \ldots, (x_n, y_n)$ with the true model as

$$y_i = g(x_i\beta) + \varepsilon \tag{3.1}$$

and the estimated model using the full-data as

$$\widehat{y}_i = \widehat{g}(x_i\widehat{\beta}) \tag{3.2}$$

Using a sub-sample $(x_1^*, y_1^*), \ldots, (x_m^*, y_m^*)$ we can estimate the SIM as

$$\widehat{y}_i^* = \widehat{g}^*(x_i^*\widehat{\beta}^*)\ . \tag{3.3}$$

The sub-sample prediction on the full data is

$$\widetilde{y}_i = \widehat{g}^*(x_i\widehat{\beta}^*) \tag{3.4}$$

and the full-data prediction on the sub-sample is

$$\widetilde{y}_i^* = \widehat{g}(x_i^*\widehat{\beta}) \ . \tag{3.5}$$

Small values of

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - \widetilde{y}_i)^2$$

and

$$MSE^* = \frac{1}{m}\sum_{i=1}^{n}(\widehat{y}_i^* - \widetilde{y}_i^*)^2$$

indicate the sub-sampling is efficient. If we know the true model $g$ (eg. simulation studies), then we can also calculate

$$MSE_0^* = \frac{1}{m}\sum(\widehat{g}^*(x_i^*\widehat{\beta}^*) - g(x_i^*\widehat{\beta}^*))^2 \tag{3.6}$$

$$MSE_0 = \frac{1}{n}\sum(\widehat{g}(x_i^*\widehat{\beta}^*) - g(x_i^*\widehat{\beta}^*))^2 \tag{3.7}$$

to evaluate the efficiency of sub-sampling.

Again, the major reason for performing sub-sampling is reducing the computation time of our algorithms. We do that by using *proc.time*() function in R. We expect the computation time to decrease as the sub-sample size decreases but at a loss of efficiency.

### 3.5 Cross-validation

Cross-validation of the most common tool for assessing the predictive ability of models. This involves splitting the data into two parts: the training set which used to build the model and the testing which used to validate the model by estimating the prediction error. However, in this case, we did not perform cross-validation. By evaluating the sub-sample models using the full data repeatedly, we are performing a kind of cross-validation since the performance is evaluated on data not used in building the model.

Chapter 4

Simulation Studies and Real Data Application

## 4.1 Simulation

To demonstrate the performance of the proposed methods, we performed an extensive simulation study for the estimation of the index parameters $\boldsymbol{\beta}$ of SIMs as well as the nonparametric function $g$. The simulation study was done by using the R 4.0.2 (R Development Core Team, 2020). We considered the LS, LAD, and $R$ estimators. All were computed using the algorithm described in Sun et al. (2019) as the intercept term of the single-index varying coefficient model (SIVCM). The initial value of the single index direction $\boldsymbol{\beta}$ was found using sliced inverse regression (Xia, 2006).

In our simulation study, we used the underlying model

$$Y = 1.5 \sin((\boldsymbol{\beta}_0^T \mathbf{x})\pi) + \varepsilon, \tag{4.1}$$

where the true model parameter $\boldsymbol{\beta}_0 \in \mathbb{R}^4$ was taken to be $\boldsymbol{\beta}_0^T = (1, 2, 0, 2)/3$. We took $\mathbf{x} \sim (U[-1, 1])^{\otimes 4}$ and we considered two possible distributions for the model error $\varepsilon$. The first is the standard normal $N(0, 1)$ where we expect the LS method to give superior performance. The second is the t-distribution with three degrees of freedom $t_3$ where we expect the robust methods LAD and $R$ to perform better. We generated $n = 200$ samples. We considered subsample sizes of $m = 50$ and $m = 75$ for both random and leverage sub-sampling methods.

We calculated the estimates of $\boldsymbol{\beta}$ and $g$ using LS, LAD, and $R$ estimates for the full data and the sub-samples. The entire process was repeated 100 times.

In addition to the estimated curves $\widehat{g}$, we also estimated the MAE of the estimated functions $\widehat{g}(t)$ from $g_0(t) = 1.5\sin(\pi t)$ as well as the mean and SD of the angle between the estimated $\widehat{\boldsymbol{\beta}}$ and the true $\boldsymbol{\beta}_0$, which defined as:

$$A(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}_0) = (180/\pi)\arccos(|\widehat{\boldsymbol{\beta}}^T \boldsymbol{\beta}_0|). \tag{4.2}$$

where the angle $A$ is defined between $[0, 90]$. Smaller values of the MAE indicate a good fit of the nonparametric function and values of $A$ close to zero indicate a good approximation of the single-index direction.

The results of the simulation study are presented in Table 4.1 and Table 4.2. Moreover, the estimated functions from each run of the simulation are overlaid on the true $g_0$ for a a comparison of the estimated and true functions in Figures 4.1, 4.2, 4.3, 4.4, 4.5, 4.6 and 4.7.

Considering the MAE of $\widehat{g}$ (Table 4.1) the following results are noted:

- As expected, the MAE of the model using full data is always the smallest compared to the sub-samples for all the three methods. For the fits using the full data, the rank method gives the best performance in comparison to LS and LAD. This is a phenomenon also observed in Sun et al. (2019) and Bindele et al. (2018).

- For the $N(0, 1))$ error, the rank method gives the smallest MAE followed by the LS for random sub-sampling. However, as expected, the LS gives the smallest MAE followed by the rank method for leverage sub-sampling. The LAD gave the worst performance of all the methods compared.

- For the $t_3$ error, the rank method gives consistently the lowest MAE except for the case with the leverage sub-sample of 50 where the LS method performs the best. Again, the LAD method gave the worst performance.

- Leverage sub-sampling gave lower MAEs than random sub-sampling with the exception of the LAD method (all cases) and the rank method ($N(0,1), m = 75$).

Table 4.1: The mean absolute errors of the functional coefficients and standard deviation (SD) of the overall mean absolute deviation (MAE).

| Dist | Sampling Methods | $n$ | LS | | LAD | | Rank | |
|---|---|---|---|---|---|---|---|---|
| | | | MAE | SD | MAE | SD | MAE | SD |
| $N(0,1)$ | Random | 50 | 0.4156 | 0.3340 | 0.4894 | 0.3426 | 0.3852 | 0.4012 |
| | | 75 | 0.2978 | 0.0954 | 0.3936 | 0.2441 | 0.2548 | 0.1903 |
| | Leverage weight | 50 | 0.3543 | 0.3132 | 0.6029 | 0.6157 | 0.3743 | 0.3082 |
| | | 75 | 0.2727 | 0.1902 | 0.4250 | 0.3356 | 0.3365 | 0.3606 |
| | Full | 200 | 0.2271 | 0.0487 | 0.2719 | 0.0659 | 0.1601 | 0.0579 |
| $t_3$ | Random | 50 | 0.5790 | 0.4204 | 0.6643 | 0.4913 | 0.5713 | 0.4924 |
| | | 75 | 0.4882 | 0.3974 | 0.5239 | 0.4502 | 0.4009 | 0.4134 |
| | Leverage weight | 50 | 0.4768 | 0.4685 | 0.6692 | 0.5883 | 0.5082 | 0.4427 |
| | | 75 | 0.3995 | 0.3401 | 0.5960 | 0.5584 | 0.3761 | 0.3455 |
| | Full | 200 | 0.2372 | 0.0901 | 0.2891 | 0.0866 | 0.1771 | 0.0726 |

Considering the angle between the estimated $\widehat{\boldsymbol{\beta}}$ and the true $\boldsymbol{\beta}_0$ (Table 4.2), we noted

- For the model fit with full data, the LS did the best for $N(0,1)$ errors and the rank method did the best for $t_3$ errors.

- For random sub-samples under the $N(0,1)$ errors, the LS estimate gave the smallest mean angles followed by the rank estimate for both $m = 50$ and $m = 75$. The LAD was vastly inferior.

- For leverage sub-samples under the $N(0,1)$ errors, the LS estimate gave the smallest mean angles followed by the rank estimate for both $m = 75$ but the two methods were comparable for $m = 50$. The LAD was vastly inferior.

- For $t_3$ error distribution, the rank method provided the least angles for both random and leverage sub-sampling scenarios as well as both small $m = 50$ and moderate $m = 75$ sub-sample sizes. Leverage sub-sampling gave smaller mean angles for the small sub-sample sizes ($m = 50$). However, leverage sub-sampling did not improve the angles for $m = 75$.

Table 4.2: The Mean and SD of the Angle between the estimated $\widehat{\beta}$ and the true $\beta_0$; $A(\widehat{\beta}, \beta_0)$

| Dist | Sampling Methods | $n$ | LS | | LAD | | Rank | |
|---|---|---|---|---|---|---|---|---|
| | | | Mean | SD | Mean | SD | Mean | SD |
| | Random | 50 | 12.1611 | 10.3222 | 18.2154 | 12.0374 | 15.0347 | 11.1278 |
| | | 75 | 6.8763 | 3.0025 | 13.8438 | 11.1600 | 9.6370 | 5.1471 |
| $N(0,1)$ | Leverage weight | 50 | 15.0383 | 6.8162 | 20.5134 | 8.8440 | 15.0136 | 7.4074 |
| | | 75 | 10.8718 | 6.6991 | 17.1982 | 8.8501 | 13.4142 | 8.0427 |
| | Full | 200 | 3.8368 | 1.5713 | 5.5422 | 2.2311 | 4.5952 | 1.8868 |
| | Random | 50 | 27.0592 | 22.8669 | 27.7052 | 22.3592 | 24.1248 | 21.8432 |
| | | 75 | 17.2231 | 16.2578 | 19.2005 | 16.5982 | 15.4956 | 12.1393 |
| $t_3$ | Leverage weight | 50 | 21.01897 | 12.37403 | 27.10005 | 15.3207 | 20.0822 | 11.8546 |
| | | 75 | 17.2881 | 9.6200 | 21.7514 | 11.6795 | 15.72766 | 10.6102 |
| | Full | 200 | 6.1660 | 2.5849 | 6.3389 | 2.8905 | 5.5018 | 2.2433 |

One of the aims of this study was to investigates a new approach for computing the generalized leverage of single-index models and how they can be used for intelligent sampling across non-parametric curves. As a result, we found that the leverage-based sub-sampling generally captures the curves better than random sub-sampling. This is depicted in Figures 4.1, 4.2, 4.3 and 4.4.

The results in this chapter demonstrate that the rank estimator is a serious competitor to the LS for the estimation of the SIM under the $N(0,1)$ error distribution and it is superior to the LS for the heavier tailed $t_3$ distribution. The LAD is inferior to both the rank and the LS for both function estimation as well as single-index direction estimation. The rank method gave the best performance overall for the estimation of the nonparametric function.

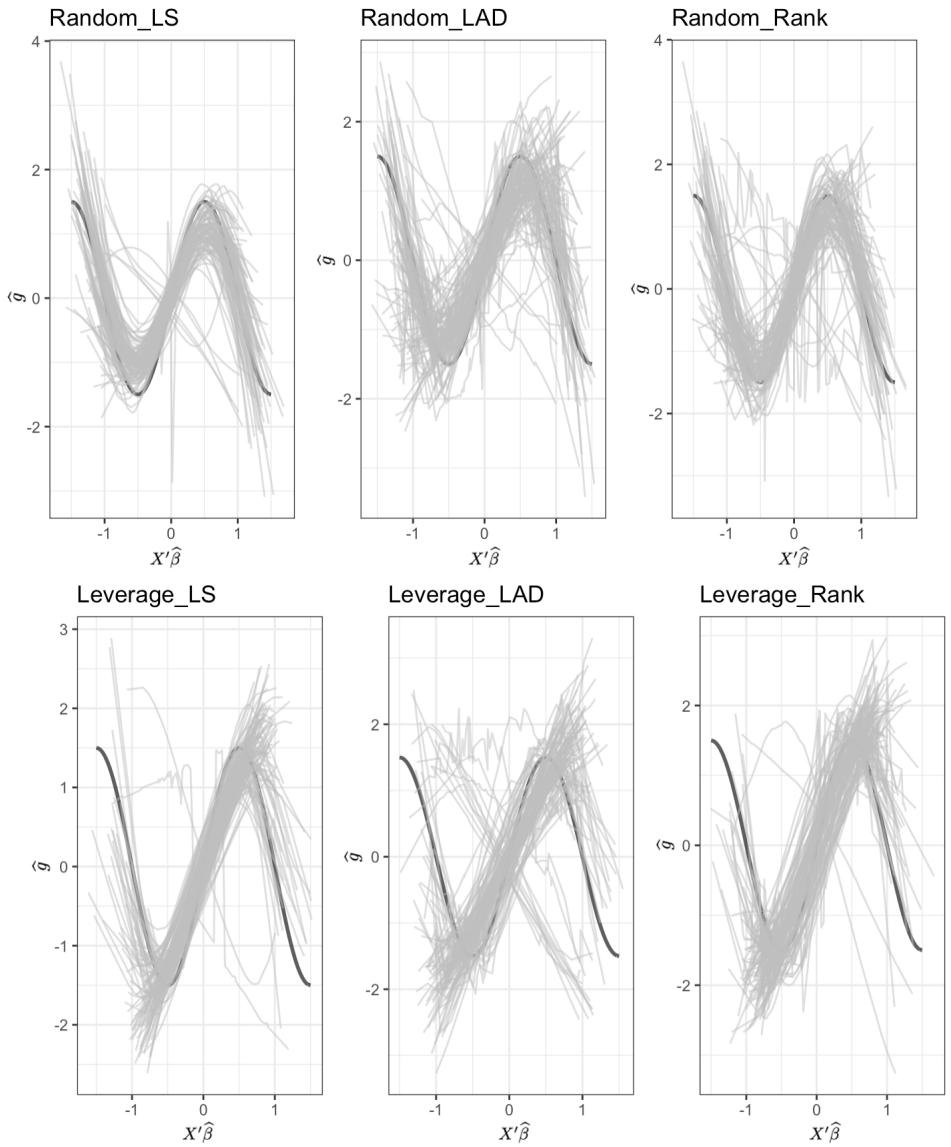


Figure 4.1: Estimated coefficient functions under the standard normal error distribution$N(0,1)$ with 50 points selected. *Top*: Randomly Sampling. *Bottom* : Leverage weight Sampling.

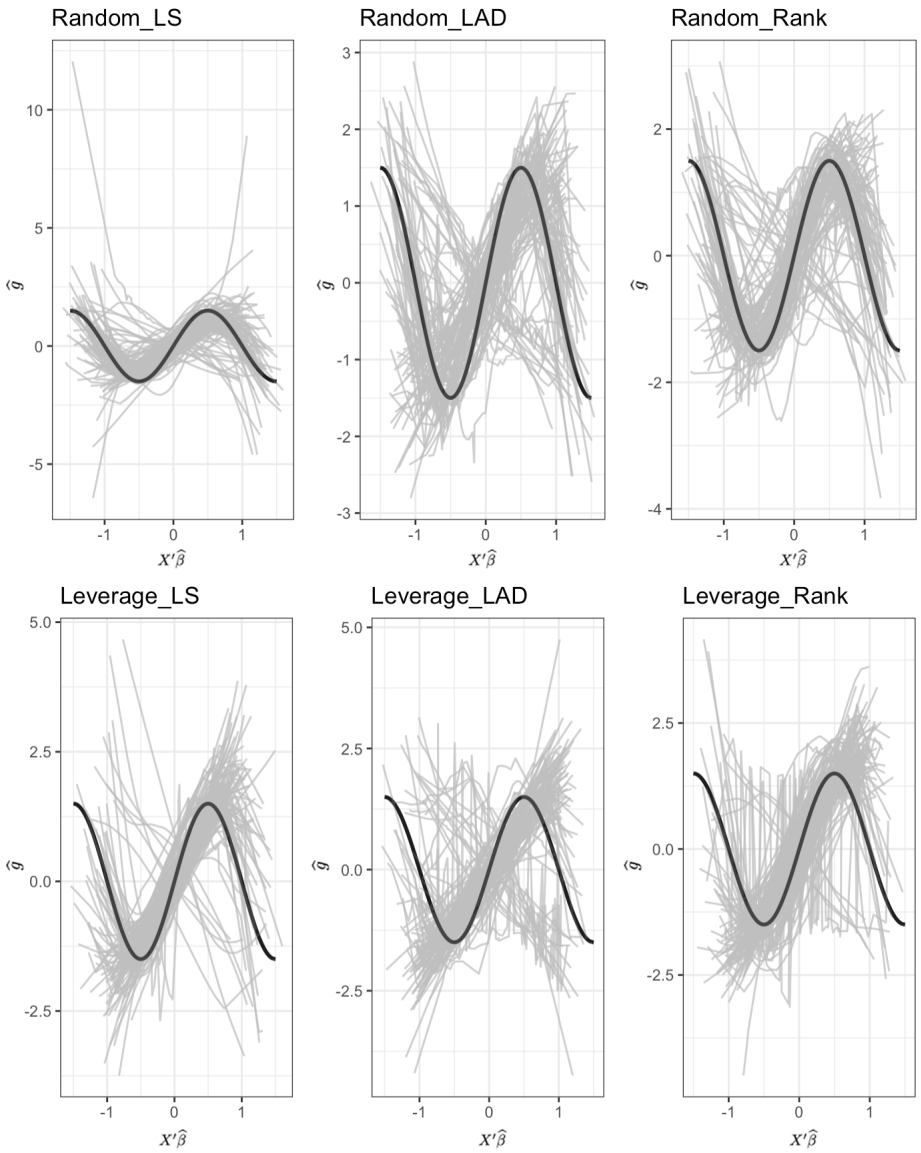


Figure 4.2: Estimated coefficient functions under the heavy tailed distribution $t_3$ with 50 points selected. *Top*: Randomly Sampling. *Bottom* : Leverage weight Sampling.

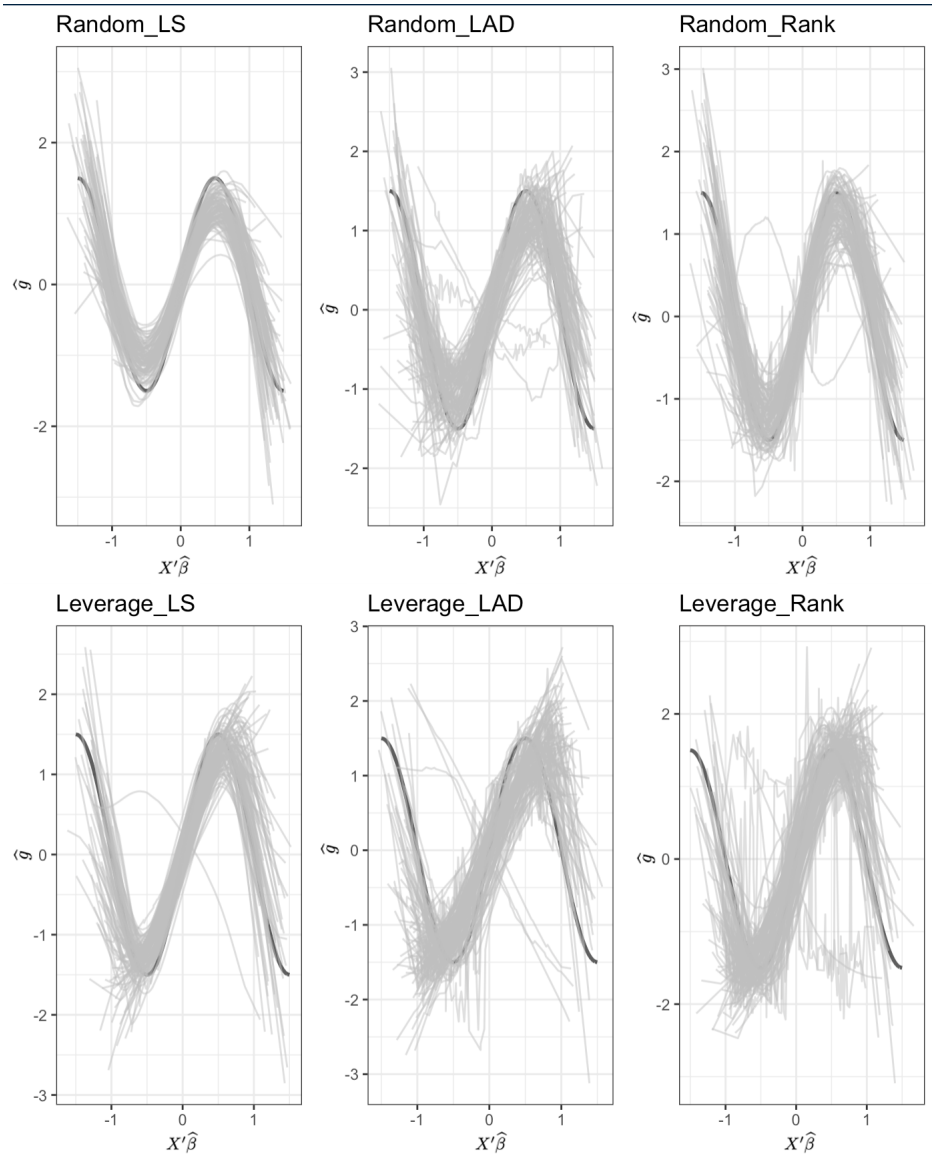


Figure 4.3: Estimated coefficient functions under the standard normal error distribution$N(0,1)$ with 75 points selected. *Top*: Randomly Sampling. *Bottom* : Leverage weight Sampling.

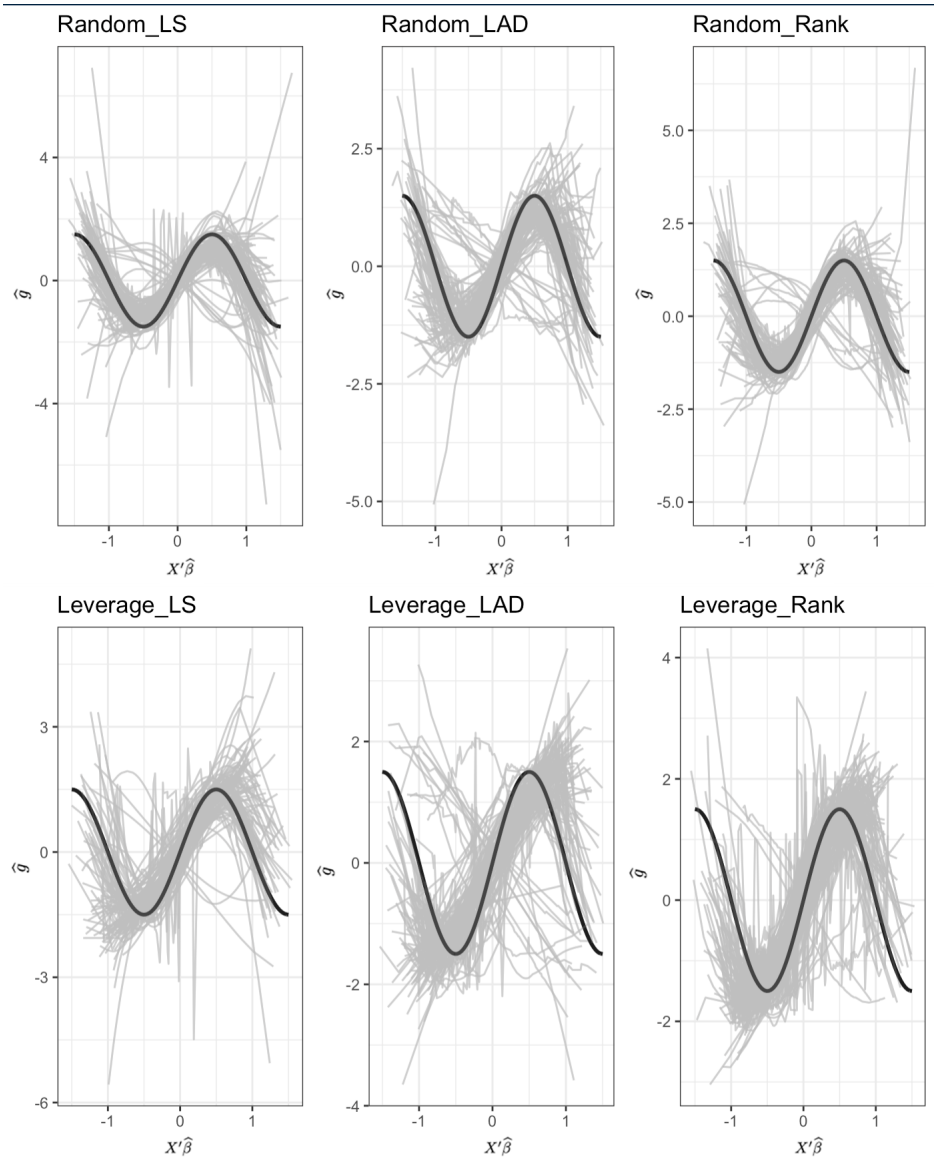


Figure 4.4: Estimated coefficient functions under the heavy tailed distribution $t_3$ with 75 points selected. *Top*: Randomly Sampling. *Bottom* : Leverage weight Sampling.

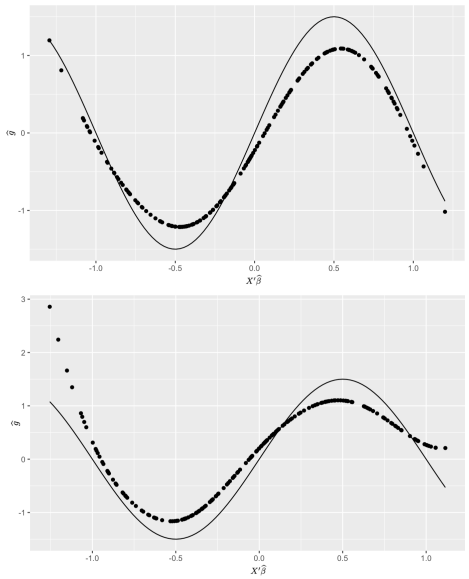


Figure 4.5: A single realization of true and estimated functions (LS). *Top*: under $N(0,1)$. *Bottom* : under $t_3$.

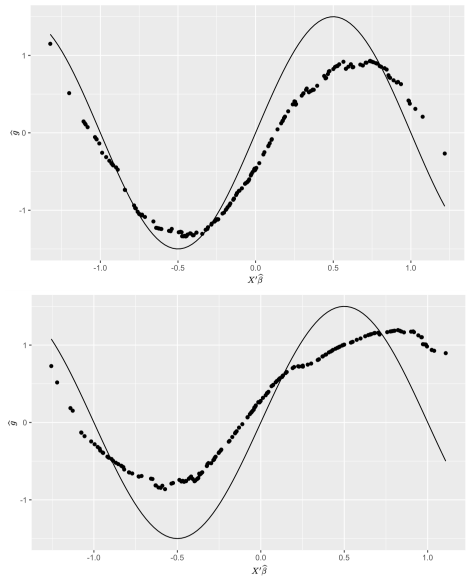


Figure 4.6: A single realization of true and estimated functions (LAD). *Top*: under $N(0,1)$. *Bottom* : under $t_3$.

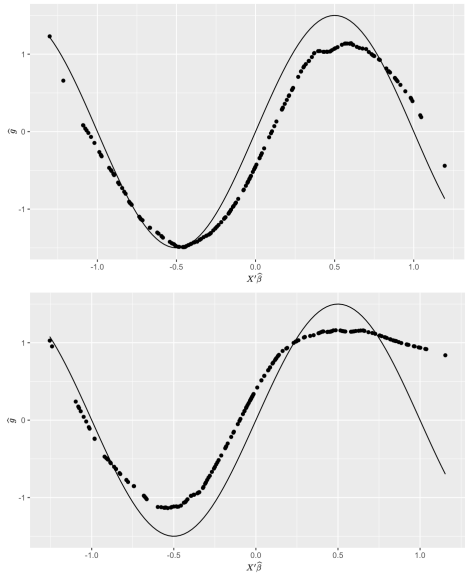


Figure 4.7: A single realization of true and estimated functions (Rank). *Top*: under $N(0,1)$. *Bottom* : under $t_3$.

## 4.2 Real Data Application

In this section, we analyze a real dataset applying the methodology developed in this thesis for prediction and estimation. The application was done by using the R software. One of the main reasons for using a robust regression approach is accurate prediction in the presence of outliers, which is a common feature of real data.

We use the popular Boston Housing dataset to illustrate the benefits of the proposed methodology. The dataset contains 506 observations with 14 variables on home prices and possible predictors. The Median value of owner-occupied homes in $\$1000's$ is used as a response variable $y$ and the remaining 13 variables were used as predictor variables $\mathbf{x} = (x_1, x_2, ...., x_{13})^T$. All the variables are described in Figure (4.8).

| CODE | DESCRIPTION |
|---|---|
| CRIM | per capita crime rate by town |
| ZN | proportion of residential land zoned for lots over 25,000 sq.ft. |
| INDUS | proportion of non-retail business acres per town |
| CHAS | Charles River dummy variable (= 1 if tract bounds river; 0 otherwise) |
| NOX | nitric oxides concentration (parts per 10 million) |
| RM | which means the square of average number of rooms per dwelling. |
| AGE | which means the proportion of owner-occupied units built prior to 1940. |
| DIS | Weighted distances to five Boston employment centers. |
| RAD | index of accessibility to radial highways |
| TAX | Full-value property-tax rate per USD 10,000. |
| PTRATIO | pupil-teacher ratio by town |
| B | 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town |
| LSTAT | % lower status of the population (percent) |
| MEDV | the median value of owner- occupied homes in $1000's. |

Figure 4.8: Variable description for the Boston Housing Data

To evaluate the performance of the proposed approaches on the single-index model for the Boston Housing data, we compared between the three methods: Least Squares (LS), Rank-Based estimation (R), and the Least Absolute Deviation (LAD) estimate. We used both leverage sub-sampling and random sub-sampling techniques.

We first fit the single-index model using the full Boston Housing data (see Figure (4.9)). Then we considered 10% and 15% sub-samples of the dataset for fitting the single-index model. These used random and leverage sub-sampling where each was repeated for 50 iterations (see Figure (4.10)). Predictions for the model with full data as well as the models with sub-sampled data were made on the same grid of indices. These were then used to compute the MAE between sub-sample estimated nonparametric function and the full data estimated nonparametric function. We also computed the mean and SD of the angle between the sub-sample estimated $\boldsymbol{\beta}$ and the full data estimated $\boldsymbol{\beta}$.

Figure (4.9) shows that the relationship between the response and the single index is nonlinear. Also, the plot shows that the range of the estimated index ($x$-axis) for LS is narrower than the range for LAD and rank estimates. This indicates that the estimated index direction for the LS is very different from the rank and LAD methods which have almost the same single index range. This indicates that the angles between sub-sample and full sample data may not be directly comparable for LS versus the remaining methods.

As can be seen from the Table (4.3) (below), the mean and SD of the angle between the sub-sample estimated $\widehat{\boldsymbol{\beta}}$ and the full data $\widehat{\boldsymbol{\beta}}_{full}$; $A(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\beta}}_{full})$ shows that leverage sub-sampling did not give an improvement over random sub-sampling. The rank and LAD methods were comparable in performance. Since the single-index ranges were different between the LS and the other methods, we were not able to make a comparison between LS and rank or LS and LAD.

Moreover, Figure (4.10) gives leverage sub-sampling estimated functions with only 10% of the data used for the three methods. Overall, by looking at the LS estimates, it is easy to see how it can be erroneous in some cases when there are outliers. However, the robust regression methods rank and LAD, were less sensitive to the same outliers.

Figure (4.11) compares the mean absolute error ($MAEs$) of the leverage sub-sampling methods versus full data between all the three methods. Overall, it is clear that the rank method does a slightly better job than the LAD method and a significantly better job than the LS.

Based on the previous results we noted that the $X$ matrix are variables that are thought to affect the median value of owner-occupied homes in $\$1000's$. Although these data were originally fit using linear models, it is clear that the relationship between $X$ and the response is nonlinear. Robust estimation methods such as the rank-based technique perform better than the LS since the data contain outliers (see Fig 4.10 and Fig 4.11).

Overall, a major benefit of the leverage sub-sampling methods is that they can actually capture more points near points at which the curve changes substantially to reduce the total expense of the data analysis.

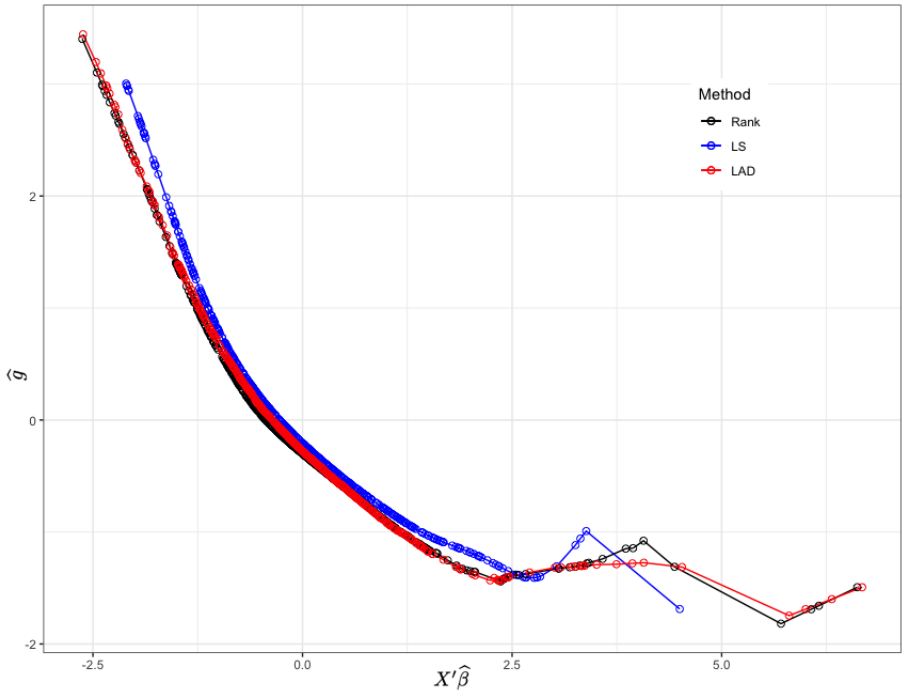


Figure 4.9: Comparing the Curve-Fitting Effectiveness of the data on Different Models.

Table 4.3: The Mean and SD of the Angle between the estimated $\widehat{\beta}$ and the true $\beta_0$; $A(\widehat{\beta}, \beta_0)$

| Sampling Methods | $n$ | LS | | LAD | | Rank | |
|---|---|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD | Mean | SD |
| Random | 50 | 74.67673 | 6.665399 | 67.43941 | 10.24543 | 68.53673 | 9.686003 |
| | 75 | 73.36251 | 5.010421 | 65.94731 | 9.450909 | 65.0805 | 9.487671 |
| Leverage weight | 50 | 76.3414 | 3.768954 | 85.11641 | 4.994919 | 85.42325 | 2.645459 |
| | 75 | 75.53126 | 3.082132 | 84.42777 | 3.85406 | 83.80424 | 5.743461 |

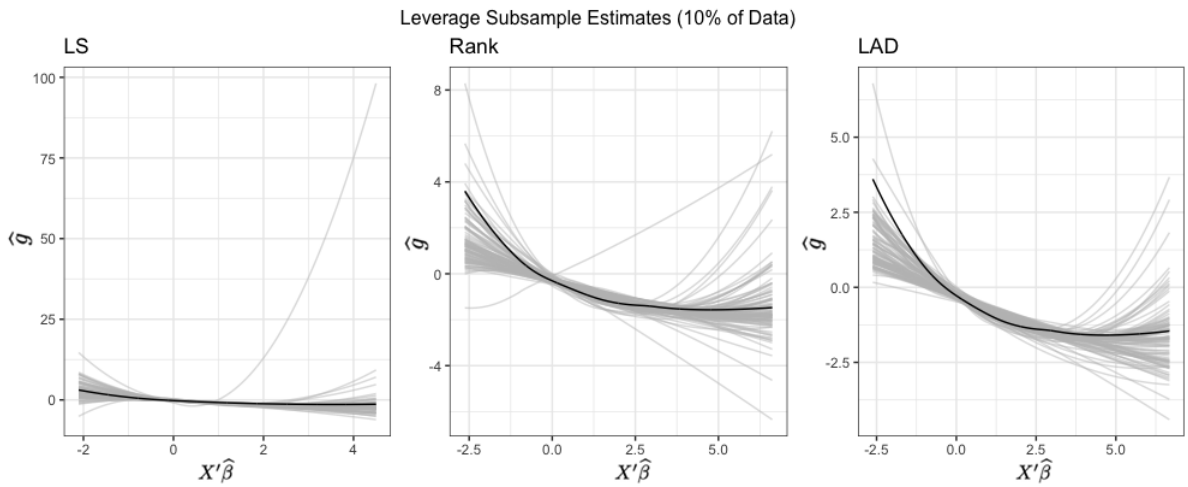


Figure 4.10: Leverage Sampling Estimates of the 10% of the Data

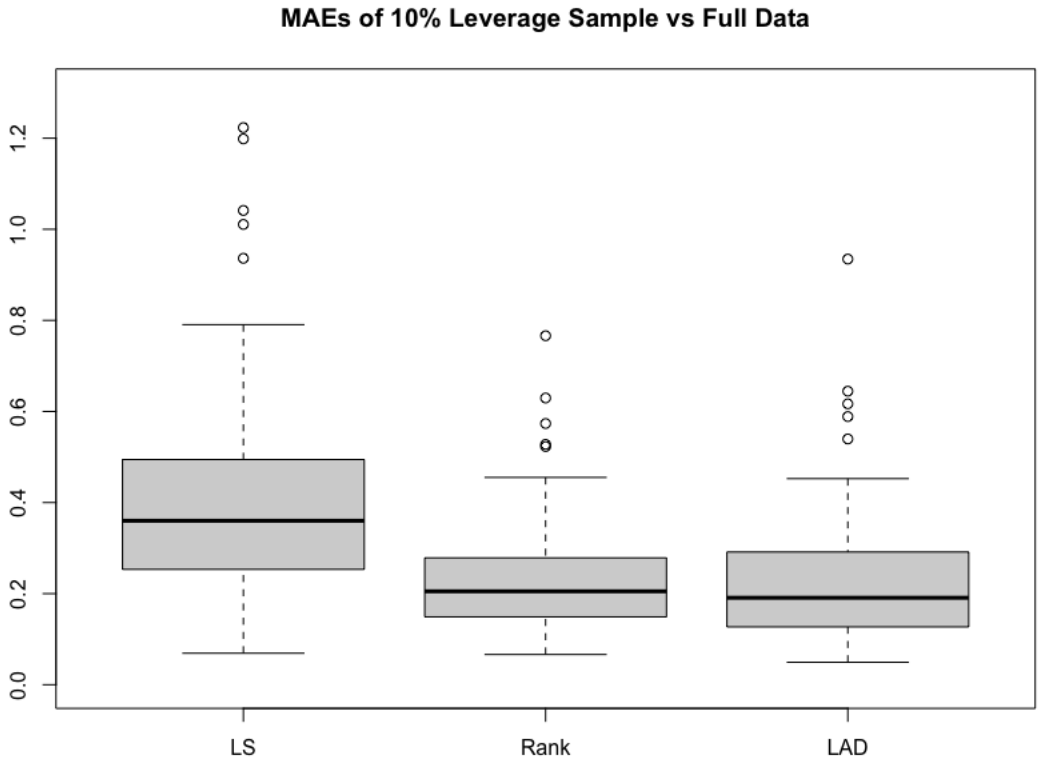


Figure 4.11: Comparison between the MAEs of 10% of Leverage Sampling vs Full data

# Chapter 5

## Conclusions and Recommendations

In conclusion, the purpose of the current study was to study the use of generalized leverage based sub-subsampling for single-index models. These give more efficient estimators than random sub-samples of the same size. Also, we extended the leverage based sub-subsampling approach to robust rank-based estimators of single-index models to provide estimators that are robust to outliers and heavy tails.

Moreover, we demonstrated that the approach has high computational efficiency and provides relatively accurate estimates. Thus, this approach will prove useful in expanding our understanding and practice of how the leverage sampling algorithm can facilitate inference for real data where the data are derived from complex processes.

Finally, it is worth pointing out that the performance of the proposed methods in this thesis can lead to future work. There are a couple of possible directions that need to be a further studied. First, in this study we considered a more complicated function (SIM) than the linear models considered in the literature. We would like to extend this to the analysis of streaming data from such structures. Second, according to our previous literature review we found that the variable selection can play a significant role in leverage sampling methods (Sun et al., 2019). The benefit of this approach is that by selecting the influential variables to reduce dimension and by applying the leverage sub-sampling technique on the reduced dimension should result in a substantial reduction of the computational burden involved in the analysis of big data.

References

A. Abebe and J. W. McKean. Weighted wilcoxon estimators in nonlinear regression. *Australian & New Zealand Journal of Statistics*, 55(4):401–420, 2013.

A. Abebe, H. F. Bindele, M. Otlaadisa, B. Makubate, et al. Robust estimation of single index models with responses missing at random. *Statistical Papers*, pages 1–31.

H. F. Bindele and A. Abebe. Bounded influence nonlinear signed-rank regression. *Canadian Journal of Statistics*, 40(1):172–189, 2012.

H. F. Bindele and A. Abebe. Semi-parametric rank regression with missing responses. *Journal of Multivariate Analysis*, 142:117–132, 2015.

H. F. Bindele, A. Abebe, and K. N. Meyer. General rank-based estimation for regression single index models. *Annals of the Institute of Statistical Mathematics*, 70(5):1115–1146, 2018.

P. Bloomfield and W. L. Steiger. *Least absolute deviations: theory, applications, and algorithms*. Springer, 1983.

R. J. Carroll, J. Fan, I. Gijbels, and M. P. Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92(438):477–489, 1997.

K. Chen, Z. Ying, H. Zhang, and L. Zhao. Analysis of least absolute deviation. *Biometrika*, 95(1):107–122, 2008.

M. Dasgupta and S. K. Mishra. Least absolute deviation estimation of linear econometric models: A literature review. *Available at SSRN 552502*, 2004.

M. Delecroix, W. Härdle, and M. Hristache. Efficient estimation in conditional single-index regression. *Journal of Multivariate Analysis*, 86(2):213–226, 2003.

M. Delecroix, M. Hristache, and V. Patilea. On semiparametric m-estimation in single-index regression. *Journal of Statistical Planning and Inference*, 136(3):730–769, 2006.

D. Draper. Rank-based robust analysis of linear models. i. exposition and review. *Statistical Science*, pages 239–257, 1988.

A. Giloni, J. S. Simonoff, and B. Sengupta. Robust weighted lad regression. *Computational Statistics & Data Analysis*, 50(11):3124–3140, 2006.

M. M. Guy-vanie and A. Abebe. Iterative rank estimation for generalized linear models. *Journal of Statistical Planning and Inference*, 151:60–72, 2014.

W. Hardle, P. Hall, and H. Ichimura. Optimal smoothing in single-index models. *The annals of Statistics*, pages 157–178, 1993.

F. E. Harrell. Regression modeling strategies. *BIOS*, 330:2018, 2017.

F. E. Harrell Jr. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

D. Harrison and D. Rubinfeld. Boston housing dataset, 2015.

T. Hettmansperger and J. McKean. Robust nonparametric statistical methods, volume 119 of. *Monographs on Statistics and Applied Probability*, 2011.

T. P. Hettmansperger and J. W. McKean. *Robust nonparametric statistical methods*. CRC Press, 2010.

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

O. Hössjer. Rank-based estimates in the linear model with high breakdown point. *Journal of the American Statistical Association*, 89(425):149–158, 1994.

P. J. Huber. *Robust statistics*, volume 523. John Wiley & Sons, 2004.

M. Hubert and P. J. Rousseeuw. Robust regression with both continuous and binary regressors. *Journal of Statistical Planning and Inference*, 57(1):153–163, 1997.

L. A. Jaeckel. Estimating regression coefficients by minimizing the dispersion of the residuals. *The Annals of Mathematical Statistics*, pages 1449–1458, 1972.

J. Jureckova. Nonparametric estimate of regression coefficients. *The Annals of Mathematical Statistics*, pages 1328–1338, 1971.

J. Kloke and J. W. McKean. *Nonparametric statistical methods using R*. CRC Press, 2014.

J. D. Kloke, J. W. McKean, and M. M. Rashid. Rank-based estimation and associated inferences for linear models with cluster correlated errors. *Journal of the American Statistical Association*, 104(485):384–390, 2009.

K. Knight. Limiting distributions for l1 regression estimators under general conditions. *Annals of statistics*, pages 755–770, 1998.

R. Koenker and G. Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

E. Kong and Y. Xia. A single-index quantile regression model and its estimation. *Econometric Theory*, pages 730–768, 2012.

J. Liu, R. Zhang, W. Zhao, and Y. Lv. A robust and efficient estimation method for single index models. *Journal of Multivariate Analysis*, 122:226–238, 2013.

P. Ma and X. Sun. Leveraging for big data regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(1):70–76, 2015.

X. Ma, Q. Zhang, Q. Zhu, W. Liu, Y. Chen, R. Qiu, B. Wang, Z. Yang, H. Li, Y. Lin, et al. A robust crispr/cas9 system for convenient, high-efficiency multiplex genome editing in monocot and dicot plants. *Molecular plant*, 8(8):1274–1284, 2015.

J. W. McKean. Robust analysis of linear models. *Statistical Science*, pages 562–570, 2004.

B. McWilliams, G. Krummenacher, M. Lucic, and J. M. Buhmann. Fast and robust least squares estimation in corrupted linear models. In *Advances in Neural Information Processing Systems*, pages 415–423, 2014.

D. Pollard. Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, 7(2):186–199, 1991.

G. L. Sievers and A. Abebe. Rank estimation of regression coefficients using iterated reweighted least squares. *Journal of Statistical Computation and Simulation*, 74(11):821–831, 2004.

W. Stute, L.-X. Zhu, et al. Nonparametric checks for single-index models. *The Annals of Statistics*, 33(3):1048–1083, 2005.

W. Sun, H. F. Bindele, A. Abebe, and H. Correia. General local rank estimation for single-index varying coefficient models. *Journal of Statistical Planning and Inference*, 202:57–79, 2019.

C. Wang, S. Wang, L.-P. Zhao, and S.-T. Ou. Weighted semiparametric estimation in regression analysis with missing covariate data. *Journal of the American Statistical Association*, 92(438):512–525, 1997.

Q. Wang, O. Linton, and W. Härdle. Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association*, 99(466):334–345, 2004.

B.-C. Wei, Y.-Q. Hu, and W.-K. Fung. Generalized leverage and its applications. *Scandinavian Journal of statistics*, 25(1):25–37, 1998.

T. Z. Wu, K. Yu, and Y. Yu. Single-index quantile regression. *Journal of Multivariate Analysis*, 101(7):1607–1621, 2010.

Y. Xia. Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, pages 1112–1137, 2006.

Y. Xia and W. K. Li. On single-index coefficient regression models. *Journal of the American Statistical Association*, 94(448):1275–1285, 1999.

X. Yin and R. D. Cook. Direction estimation in single-index regressions. *Biometrika*, 92(2): 371–384, 2005.