University of Cincinnati

Date: 6/13/2018

I, Shaobo Li, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Business Administration.

It is entitled:

Two Essays on High-Dimensional Robust Variable Selection and an Application to Corporate Bankruptcy Prediction

Student's name: **Shaobo Li**

This work and its defense approved by:

Committee chair: Yichen Qin

Committee chair: Yan Yu, Ph.D.

Committee member: Alexander Borisov

30488

Two Essays on High-Dimensional Robust Variable Selection and an Application to Corporate Bankruptcy Prediction

A dissertation submitted to the
Graduate School of the University of Cincinnati
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in the Department of Operations, Business Analytics,
and Information Systems
of the Carl H. Lindner College of Business
by

Shaobo Li
M.S. Statistics, University of Cincinnati

June 2018

Committee Chair: Yan Yu, Ph.D. & Yichen Qin, Ph.D.

Abstract

High-dimensional statistical problems have been encountered in numerous modern application fields including finance, biology, and engineering. The key of high-dimensional statistics is to identify important variables among many features with sparse representation. This dissertation consists of two essays.

In Essay I, we introduce a new class of mean regression estimators — penalized maximum tangent likelihood estimation — for high-dimensional regression estimation and variable selection. We first explain the motivations for the key ingredient, a novel robust method called maximum tangent likelihood estimation (MTE), and establish its asymptotic properties. The proposed MTE is highly efficient, and we numerically demonstrate this under various simulation settings. Unlike traditional robust methods, the proposed MTE protects against violation of any particularly assumed parametric model, hence can be easily extended to various statistical models in practice, while we focus on linear regression in this article. To robustly select important variables under high dimensional feature space, we further propose the penalized MTE. The optimal rate of convergence in the order of $\sqrt{\ln(d)/n}$ has been established for ultra-high dimensional regression where the number of variables d grows exponentially with the sample size n, while $\sqrt{n}$-consistency and oracle property have been shown for fixed dimensional linear regression. The proposed penalized MTE has a broad spectrum that consists of penalized ℓ_2 distance, penalized exponential squared loss, penalized least trimmed square and penalized least square as special cases, and can be regarded

as a mixture of minimum Kullback-Leibler distance estimation and minimum ℓ_2 distance estimation. We conduct extensive simulation studies and real data analysis to demonstrate the advantages of the penalized MTE.

In Essay II, we introduce a flexible yet easy-interpretable index hazard model for corporate bankruptcy prediction under a semiparametric modeling framework. Motivated by the long debate between accounting and finance researchers, we propose a penalized double-index hazard model with automatic variable selection. The two indices are naturally constructed by separate market and accounting based bankruptcy predictors. The unknown functions are estimated by polynomial splines. In order to identify important predictors, a nonconcave penalty function, SCAD, is adopted due to its attractive statistical properties. We develop a comprehensive database of the publicly traded firms in North America manufacturing sector and focus our empirical studies on this largest sector among all industries. We show that the proposed index hazard model reveals a novel nonlinear relationship. The proposed double-index hazard model is superior to the state-of-the-art Shumway's linear discrete hazard model using Altman's Z-score variables. The two newly constructed composite indices: market and accounting index may be of great potential interest in practice. In addition, we find that the accounting index would consist of more accounting based predictors as the prediction horizon increases, while market index would include fewer market based variables.

Acknowledgments

I would like to express my greatest gratitude to my advisors Dr. Yan Yu and Dr. Yichen Qin for their continuous support and inspiring guidance during the past five years. Their immense knowledge and attitude toward research have deeply influenced me to always "stay hungry and stay foolish". This work would not be possible without their constant encouragement and tremendous patience. I would also like to thank my committee member Dr. Alexander Borisov, who has brought insightful comments to this work.

I am very grateful to department head Dr. Michael Fry, and doctoral program director Dr. Suzanne Masterson and coordinator Dr. Craig Frohle, who have been working hard to help me with financial support for tuitions and conference travels especially in my last year at the college. They always strive to help doctoral students from all aspects, and I feel very fortunate to study in such a great doctoral program at College of Business.

I would also like to extend my appreciation to all OBAIS faculty members, staffs, and fellow PhD students and friends for their generous help and support during the past five years. They together have created an extremely friendly working and learning environment.

Finally, my special and sincere gratitude goes to my family. I thank my parents for their never-ending support throughout my life. I thank my beloved wife Jin for her everlasting love, encouragement and understanding. I also dedicate this Ph.D thesis to my lovely little boy Ryan, who has brought me infinite joy and happiness since he was born.

Table of Contents

Page

List of Tables

List of Figures

Chapter 1:

Penalized Maximum Tangent Likelihood Estimation and Robust Variable Selection

1.1 Introduction

Selecting explanatory variables has become one of the most important tasks in statistics. Tremendous progresses have been accomplished with incredible amount of efforts in studying regularized methods. These accomplishments include but not limited to LASSO [53], SCAD [18], adaptive-Lasso [72], and MCP [67] in both theoretical and practical viewpoints (see e.g., [18, 69, 41] for theoretical and [20, 61] for practical contributions). For regression problems, these regularized estimators usually can be expressed as penalized likelihood estimation,

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} \ln f(\mathbf{z}_i; \boldsymbol{\beta}) - n \sum_{j=1}^{d} p_\lambda(\beta_j) \right\}, \tag{1.1}$$

where $\{\mathbf{z}_i\}_{i=1}^n = \{y_i, \mathbf{x}_i^T\}_{i=1}^n$ represents the response variable and covariates, f is likelihood function that is commonly assumed to be normal density with zero mean for linear regression (note we use $f(\mathbf{z}_i; \boldsymbol{\beta})$ and $f(y_i - \mathbf{x}_i^T\boldsymbol{\beta})$ interchangeably) and other discrete distribution functions for generalized linear models (GLMs), and $p_\lambda(\beta_j)$ is a penalty function for jth regression coefficient, $j = 1, \ldots, d$.

However, the performance of such an estimator usually degrades drastically once the data disagree with the assumed distribution f. Such effect can be severe even if a small proportion of data is contaminated while majority consist with the assumed distribution. A natural question is then how to protect violation of f. To address this issue, we first propose the maximum tangent likelihood estimation (MTE) as

$$\tilde{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^d} \sum_{i=1}^{n} \ln_t(f(\mathbf{z}_i;\boldsymbol{\beta})), \tag{1.2}$$

and then its penalized version as

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \sum_{i=1}^{n} \ln_t(f(\mathbf{z}_i;\boldsymbol{\beta})) - n\sum_{j=1}^{d} p_{\lambda_{nj}}(|\beta_j|) \right\} \tag{1.3}$$

for robust variable selection, where $\ln_t(\cdot)$ is a piecewise function defined as

$$\ln_t(u) = \begin{cases} \ln(u) & \text{if } u > t, \\ \ln(t) + \sum_{k=1}^{p} \frac{\partial^k \ln(v)}{\partial v^k}\big|_{v=t} \frac{(u-t)^k}{k!} & \text{if } 0 \le u \le t. \end{cases} \tag{1.4}$$

Here $t \ge 0$ is a tuning parameter that balances robustness and efficiency tradeoff. $\ln_t(u)$ is essentially a p-th order Taylor expansion of $\ln(u)$ for $0 \le u < t$. Figure 1.1 compares the shape of $\ln_t(\cdot)$ (bold black curve) with $\ln(\cdot)$ (blue dashed curve) for various p and t. When $t = 0$, $\ln_t(u) = \ln(u)$ (1st figure from left), hence MTE contains the maximum likelihood estimation (MLE) as a special case. Although p also determines the shape of $\ln_t(\cdot)$ (2nd-4th figures), we found out through simulation that its effect is much less significant than that of t. We stress that $p = 1$ unless indicated otherwise throughout this article, hence the name "tangent". However, our results are expected to hold for a general p.

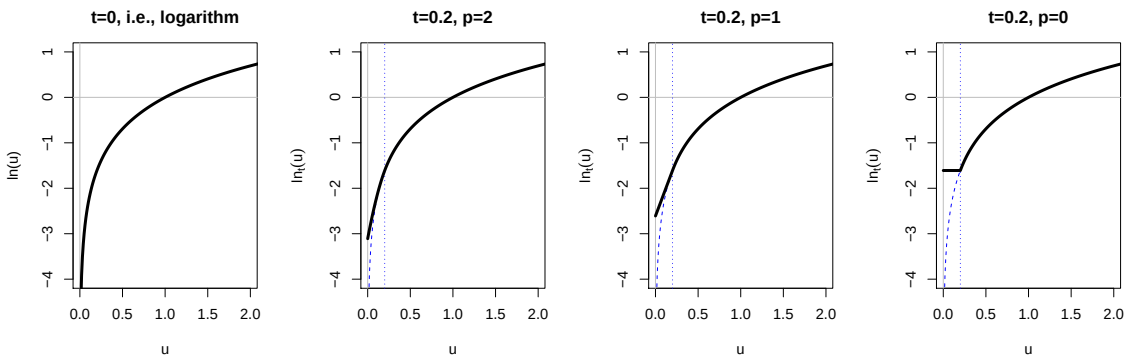

Figure 1.1: Illustration of $\ln_t(u)$ (in bold black) with different p and t. Blue dashed curve represents nature logarithm function for comparison. Blue dotted line indicates the value of tuning parameter t.

The mechanism of protecting violation of assumed distribution f can be easily seen from above formulations. In contrast to MLE, the second piece in the formula (1.4) offers lower bound by assuming $t > 0$, which would otherwise tend to negative infinity for observations that deviate far from f. A small value of t is generally preferred. The intuition is that it keeps the portion of MLE, the first piece of formula (1.4), as much as possible, so that MTE would maintain high efficiency. Following such intuition, a data-driven approach is proposed in section 1.5 to select optimal t in order to achieve high efficiency. With such formulation and the intuitions behind, the proposed MTE is an ideal robust statistical procedure, which performs nearly optimally when model assumptions are valid ($t \to 0$) and still maintains high performance when the assumptions are violated ($t > 0$ is selected to maximize efficiency). We will elaborate the proposed MTE, and discuss its statistical properties in section 1.2.

As a by-product, MTE can be considered as a mixture of minimum Kullback-Leibler (KL) distance estimation and minimum ℓ_2 distance (MD) estimation [37] or equivalently, exponential squared loss (ESL) estimation [60] when estimating the linear regression coefficients.

It allows MTE to combine the merits of both, so that it obtains remarkable robustness and still performs well for clean data. We will further show this nice hybrid in section 1.2.2.

Equipping MTE with penalty function $\sum_{j=1}^{d} p_{\lambda_{nj}}(|\beta_j|)$, the penalized MTE (1.3), called MTE-Lasso, is able to robustly select variables and estimate their coefficients simultaneously. As an attraction paralleling to MTE, with a single tuning parameter t, MTE-Lasso bridges many existing variable selection methods such as penalized least square, penalized exponential squared loss (ELS-Lasso) [60], penalized ℓ_2 distance estimation (MD-Lasso) [37], and penalized least trimmed square estimation (Sparse LTS) [1]. In addition, Wang et al. [56] has proposed to incorporate Lasso penalty to least absolute deviation (LAD-Lasso) for robust linear regression, and Wang [59] studies theoretical properties of LAD-Lasso under high dimensional regime. Zou and Yuan [73] has proposed composite quantile regression (CQR-Lasso) for the case where the error variance is infinite. However, both LAD and CQR are quantile based estimation, which may produce unreliable results for mean regression if error term is asymmetric. Fan et al. [19] tackles this issue by proposing penalized Huber's loss (RA-Lasso) for asymmetric errors. However, with Huber's loss, RA-Lasso may still be sensitive to extreme values, while our proposed MTE-Lasso is able to completely degrade their effect due to the natural of redescending influence function. Furthermore, unlike LAD and Huber's loss, our proposed estimator is statistically more efficient under linear regression with normality assumption, and enjoys the highest finite sample breakdown point of 0.5 [16, 64]. In section 1.6, we numerically demonstrate that MTE produces highest efficiency for normal mean estimation under various settings.

Another advantage of the proposed MTE is that practically it can be readily extended to other statistical models such as widely used logistic regression, Poisson regression, Gaussian graphical models, and mixture models as long as appropriate forms of distribution function

f in (1.2) and (1.3) are specified. However, it is not trivial for existing methods such as Huber's loss and LAD to be applied to these statistical models. Despite the immediate applicability of MTE to different models, we leave their theoretical properties and details in future works, yet in this article we focus on MTE and MTE-Lasso under linear regression models.

The rest of this paper is organized as follows. In Section 1.2, we formally introduce MTE, study its properties and discuss its links to other estimators. In Section 1.3, we further introduce the penalized MTE for variable selection, and demonstrate its asymptotic properties through an analysis of consistency, oracle property. We show the robustness properties in Section 1.4. We discuss the implementation aspect of the method such as selection of tuning parameters in Section 1.5 and present numerical results in Section 1.6. Finally, we conclude with a discussion in Section 1.7 and relegate the proofs to Section 1.8.

1.2 Maximum Tangent Likelihood Estimation

1.2.1 General framework

Let $\{\mathbf{z}_i\}_{i=1}^n$ be an i.i.d. random sample from a probability model $f(\mathbf{z};\boldsymbol{\beta})$ with parameter $\boldsymbol{\beta} \in \mathbb{R}^d$. We define the maximum tangent likelihood estimator (MTE) of $\boldsymbol{\beta}$ as in (1.2). Unlike traditional log-likelihood, the tangent likelihood function $\ln_t(\cdot)$ is a piecewise continuous function with breakpoint $t > 0$. Therefore, a weighted score function is derived, so that solving optimization problem (1.2) translates to solving a weighted likelihood equation (assuming the regularities conditions in the appendix),

$$0 = \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\beta}} \ln_t(f(\mathbf{z}_i;\boldsymbol{\beta})) = \sum_{i=1}^{n} w_i \frac{\partial}{\partial \boldsymbol{\beta}} \ln(f(\mathbf{z}_i;\boldsymbol{\beta})), \tag{1.5}$$

where $w_i = [1 - (1 - f(\mathbf{z}_i; \boldsymbol{\beta})/t)^p]^{\mathbb{1}\{f(\mathbf{z}_i;\boldsymbol{\beta})<t\}}$ and $\mathbb{1}\{\cdot\}$ is an indicator function. Clearly we see that $w_i \to 1$ if $t \to 0$, and $w_i = 1$ if set $\{i : f(\mathbf{z}_i; \boldsymbol{\beta}) < t\} = \varnothing$, which happens with probability 1 when $t = 0$, hence MLE. On the other hand, $w_i \to 0$ if $f(\mathbf{z}_i; \boldsymbol{\beta})/t \to 0$, that is when observation i is deviated far from the assumed model.

Equation (1.5) can be efficiently solved with an iterative reweighting algorithm as the weight w_i depends on the updated parameter estimates $\tilde{\boldsymbol{\beta}}$. Specifically, we iterate the procedures of solving the parameter $\tilde{\boldsymbol{\beta}}^{(k)}$ given the weights $\boldsymbol{w}_t(\tilde{\boldsymbol{\beta}}^{(k-1)})$ and updating the weights $\boldsymbol{w}_t(\tilde{\boldsymbol{\beta}}^{(k)})$ with new parameter estimates $\tilde{\boldsymbol{\beta}}^{(k)}$, where k is iteration step. The tuning parameter t can be optimally chosen at every iteration with certain data-driven approaches, one of which has been discussed in section 1.5.

To show the broad spectrum of MTE, following we shall briefly discuss a few special cases for $p = 1$ and $p = 0$, and later on we show the hybrid of MTE given $p = 1$ for estimating linear regression model. When $p = 1$, the weight simplifies to $w_i = \min\{1, f(\mathbf{z}_i, \boldsymbol{\beta})/t\}$, hence the tangent likelihood equation becomes

$$0 = \sum_{i=1}^{n} \left[\frac{\partial}{\partial \boldsymbol{\beta}} \ln(f(\mathbf{z}_i; \boldsymbol{\beta})) \right] \min \left\{ 1, \frac{f(\mathbf{z}_i, \boldsymbol{\beta})}{t} \right\}. \tag{1.6}$$

So if the observation has a likelihood below t, it is assigned partial weight, $f(\mathbf{z}_i, \boldsymbol{\beta})/t$. Otherwise, the observation is assigned full weight. When estimating the mean of a normal distribution, we have $\tilde{\mu} = (\sum_{i=1}^{n} w_i \mathbf{z}_i)/\sum_{i=1}^{n} w_i$ where $w_i = \min(1, \varphi(\mathbf{z}_i; \tilde{\mu}, \tilde{\sigma^2})/t)$ and $\varphi(\cdot)$ is the Gaussian density function. $\tilde{\mu}$ is essentially a weighted mean.

When $p = 0$, we have $w_i = \mathbb{1}\{f(\mathbf{z}_i; \boldsymbol{\beta}) \geq t\}$ and the tangent likelihood equation becomes

$$0 = \sum_{i=1}^{n} \left[\frac{\partial}{\partial \boldsymbol{\beta}} \ln(f(\mathbf{z}_i; \boldsymbol{\beta})) \right] \mathbb{1}\{f(\mathbf{z}_i; \boldsymbol{\beta}) \geq t\} = \sum_{i \in \mathcal{A}} \frac{\partial}{\partial \boldsymbol{\beta}} \ln(f(\mathbf{z}_i; \boldsymbol{\beta})),$$

where $\mathcal{A} = \{i : f(\mathbf{z}_i; \boldsymbol{\beta}) \geq t\}$. That is, we completely discard the data points whose likelihoods are below t. This follows similar spirit as in the trimmed likelihood/least square estimation proposed by Hadi and Luceno [23] and Alfons et al. [1]. When estimating the mean of a normal distribution, we have $\tilde{\mu} = (\sum_{i \in \mathcal{A}} \mathbf{z}_i)/|\mathcal{A}|$ where $\mathcal{A} = \{i : \varphi(\mathbf{z}_i; \tilde{\mu}, \tilde{\sigma^2}) \geq t\}$, i.e., a trimmed mean with data points whose likelihoods below t are removed. We again stress that $p = 1$ throughout this article unless otherwise indicated.

The maximum tangent likelihood estimator is essentially a redescending M-estimator [33, 49], which usually rejects data points with extreme values while take moderate outliers partially into account. This is because their score function, also called ψ-function, the first-order derivative of loss, redescends after certain threshold. Figure 1.2 shows the shape of ψ-function of our proposed MTE (i.e., right-hand side of (1.6)) by assuming $f(\cdot)$ is standard normal density. As the ψ-function descends to 0 as $|z| \to \infty$, the effect of large outlier is negligible. In contrast, for robust estimators whose ψ-function becomes a constant rather than redescending to 0, such as LAD and Huber's loss, extreme outliers contribute the same as moderate ones, hence their efficiency may be affected under certain model assumptions, for instance, linear regression with normality assumption. Furthermore, redescending M-estimators usually possess highest breakdown point of 0.5, i.e., the maximum fraction of contaminations that is allowed without destroying the estimate [16], while the estimators with monotone ψ-function have breakdown point of 0 [38]. We demonstrate the advantage of our proposed MTE in Section 1.6.

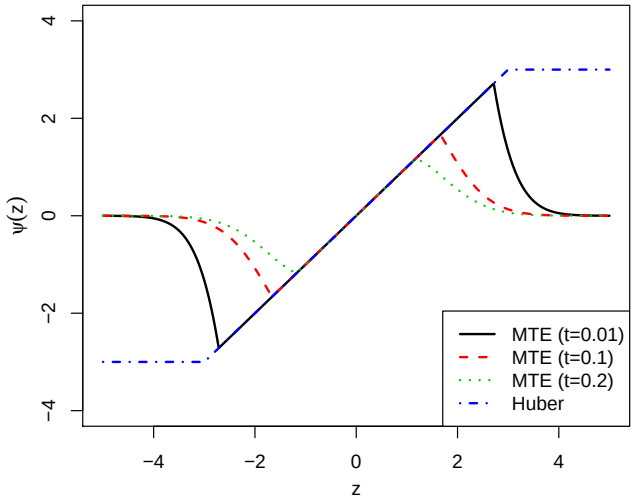

Figure 1.2: ψ-function of MTE with different values of t. ψ-function of Huber loss is also drawn.

1.2.2 MTE for linear regression

Consider a linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \qquad i = 1, \ldots, n, \tag{1.7}$$

where $\mathbf{z}_i = (y_i, \mathbf{x}_i^T)$ is the ith observation, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$ is an unknown regression coefficient vector, and ϵ_i is the random error that is independently and identically distributed with a parametric distribution.

The most widely used distribution assumption for ϵ_i is normal distribution with zero mean and constant variance despite that the actual residual could have much heavier tail than the assumed Gaussian density. Due to its optimality and convenience with the statistics pillar — maximum likelihood estimation, such normality assumption is the most fundamental and very first assumption for linear regression in almost every statistics textbook. *Nonetheless,*

the model estimation is optimal only if the actual data consist with that assumption. Even a very small deviation could destroy such optimality, and lead the estimation problematic. As discussed in section 1.1, our proposed MTE is designed to protect violation of any presumed parametric distribution. Thus, without changing or replacing the convenient and widely used Gaussian assumption, the MTE offers nearly optimal performance as if the normality assumption is valid. Although we assume $f(\cdot)$ to be a Gaussian probability density function with zero mean and constant variance σ^2, it is expected that our methodology holds for a wide range of densities well beyond the Gaussian density.

As mentioned earlier, the MTE is related to the minimum KL and ℓ_2 distance estimation for linear regression. To show this, we start by rewriting (1.2) as

$$\tilde{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \sum_{i\in\mathcal{A}} \ln(f(\mathbf{z}_i;\boldsymbol{\beta})) + \frac{1}{t}\sum_{i\in\mathcal{A}^c} f(\mathbf{z}_i;\boldsymbol{\beta}) \right\}, \tag{1.8}$$

where $\mathcal{A} = \{i : f(\mathbf{z}_i;\boldsymbol{\beta}) \geq t\}$. First, note that the minimum KL distance estimate $\tilde{\boldsymbol{\beta}}_{\mathrm{KL}}$ is essentially the MLE, that is

$$\tilde{\boldsymbol{\beta}}_{\mathrm{KL}} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \sum_{i\in\mathcal{A}} \ln(f(\mathbf{z}_i;\boldsymbol{\beta})) + \sum_{i\in\mathcal{A}^c} \ln(f(\mathbf{z}_i;\boldsymbol{\beta})) \right\}. \tag{1.9}$$

Second, the minimum ℓ_2 distance estimate $\tilde{\boldsymbol{\beta}}_{\ell_2}$ for linear regression [47, 37] is equivalent to

$$\tilde{\boldsymbol{\beta}}_{\ell_2} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \sum_{i\in\mathcal{A}} f(\mathbf{z}_i;\boldsymbol{\beta}) + \sum_{i\in\mathcal{A}^c} f(\mathbf{z}_i;\boldsymbol{\beta}) \right\}. \tag{1.10}$$

Remark 1. *To understand* (1.10)*, consider the* ℓ_2 *distance between the parametric distribution of* y *given* $\mathbf{x}$*,* $p(y|\mathbf{x},\boldsymbol{\beta})$*, and the true distribution of* y *given* $\mathbf{x}$*,* $p(y|\mathbf{x})$*,*

$$\begin{aligned}\int (p(y|\mathbf{x},\boldsymbol{\beta}) - p(y|\mathbf{x}))^2 dy &= \int p(y|\mathbf{x},\boldsymbol{\beta})^2 dy + \int p(y|\mathbf{x})^2 dy \\ &\quad - 2\int p(y|\mathbf{x},\boldsymbol{\beta}) p(y|\mathbf{x}) dy \\ &= \int p(y|\mathbf{x},\boldsymbol{\beta})^2 dy + \int p(y|\mathbf{x})^2 dy - 2\mathbb{E} f(y - \mathbf{x}^T\boldsymbol{\beta}).\end{aligned}$$

For linear regressions, $\int p(y|\mathbf{x},\boldsymbol{\beta})^2 dy = \int f(y - \mathbf{x}^T\boldsymbol{\beta})^2 dy$ *does not depend on* $\boldsymbol{\beta}$*. Hence, minimizing the* ℓ_2 *distance with respect to* $\boldsymbol{\beta}$ *is equivalent to maximizing* $\mathbb{E} f(y - \mathbf{x}^T\boldsymbol{\beta})$*. When observing a sample, we replace* $\mathbb{E} f(y - \mathbf{x}^T\boldsymbol{\beta})$ *with its empirical mean* $\sum_{i=1}^n f(\mathbf{z}_i;\boldsymbol{\beta})/n$*, and obtain* $\tilde{\boldsymbol{\beta}}_{\ell_2} = \arg\max_{\boldsymbol{\beta}\in\mathbb{R}^d} \sum_{i=1}^n f(\mathbf{z}_i;\boldsymbol{\beta})$*.*

Comparing (1.8) with (1.9) and (1.10), we understand that MTE can be considered as minimizing a mixture of KL and ℓ_2 distances. When $t = 0$, all the observations fall into the set $\mathcal{A}$, and MTE becomes the minimum KL distance estimation. As t gradually increases away from 0, some observations with relatively low likelihoods gradually move from $\mathcal{A}$ to $\mathcal{A}^c$. When t is sufficiently large, all observations have moved from $\mathcal{A}$ to $\mathcal{A}^c$, and MTE becomes the minimum ℓ_2 distance estimation.

With an appropriately selected t, we have observations in both $\mathcal{A}$ and $\mathcal{A}^c$. The observations in $\mathcal{A}^c$ are the potential outliers. If they were to be used in the pure minimum KL distance estimation, we would have an unstable estimate. Meanwhile, the observations in $\mathcal{A}$ are the good observations. If they were to be used in the pure minimum ℓ_2 distance estimation, we would have an inefficient estimate. Therefore, MTE minimizes the KL distance for the observations in $\mathcal{A}$ and minimizes the ℓ_2 distance for the observations in $\mathcal{A}^c$ to preserve efficiency and gain robustness.

Finally, we summarize the links between MTE and other estimators for linear regression as special cases. Suppose T is a sufficiently large number. When $0 < t < T$ and $p = 0$, MTE is asymptotically equivalent to LTS [23, 1]. When $0 < t < T$ and $p = 1$, MTE can be considered as a mixture of minimum KL distance and minimum ℓ_2 distance. When $t \geq T$ and $p = 1$, MTE is equivalent to L2D or ESL. Lastly, when $t = 0$ or when $p = +\infty$, MTE is essentially MLE or minimum KL distance estimation.

1.2.3 Asymptotic properties of MTE

We present asymptotic properties of MTE. First define $\boldsymbol{\beta}_t^* = \arg\max_{\boldsymbol{\beta}\in\mathcal{B}} \mathbb{E}_{\boldsymbol{\beta}_0} \ln_t(f(\mathbf{z};\boldsymbol{\beta}))$ where $\boldsymbol{\beta}_0$ is the true parameter and $t \geq 0$.

Theorem 1.2.1. *Under the regularity conditions specified in the supplementary materials, with probability going to 1, there exists a unique solution $\tilde{\boldsymbol{\beta}}$ for equation* (1.2)*. Furthermore, we have $\tilde{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_t^*$ as $n \to \infty$.*

Theorem 1.2.2. *Under the regularity conditions specified in the supplementary materials,*

$$\sqrt{n}\boldsymbol{\Omega}^{-1/2}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t^*) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}) \quad as \quad n \to \infty,$$

where $\mathbf{I}$ is a $d \times d$ identity matrix, $\boldsymbol{\Omega} = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$, $\mathbf{A} = \partial^2 \mathbb{E}_{\boldsymbol{\beta}_0}\big[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))\big]/\partial\boldsymbol{\beta}^2$, and $\mathbf{B} = \mathbb{E}_{\boldsymbol{\beta}_0}\big[(\partial \ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))/\partial\boldsymbol{\beta})(\partial \ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))/\partial\boldsymbol{\beta})^T\big]$. When $t \to 0^+$, we have $\boldsymbol{\beta}_t^* \to \boldsymbol{\beta}_0$ and $\boldsymbol{\Omega}$ becomes the inverse of Fisher information matrix.

In general, $\boldsymbol{\beta}_t^*$ is not necessarily the same as $\boldsymbol{\beta}_0$ for $t > 0$. However, when $\boldsymbol{\beta}_0$ represents the location parameter of a symmetric distribution such as linear regression coefficients, then we have $\boldsymbol{\beta}_t^* = \boldsymbol{\beta}_0$, which means MTE is indeed a consistent estimator and has asymptotic normality for such a case.

Theorem 1.2.3 (Consistency and asymptotic normality)**.** *Under the regularity conditions specified in the supplementary materials, for linear regression* $y_i = \mathbf{x}_i^T \boldsymbol{\beta}_0 + \epsilon_i$*, suppose the error* ϵ_i *follows a symmetric distribution with zero mean. Then we have* $\boldsymbol{\beta}_t^* = \boldsymbol{\beta}_0$ *for any* $t > 0$*. That is, MTE of the regression coefficient* $\tilde{\boldsymbol{\beta}}$ *defined in equation* (1.2) *is consistent and asymptotically normal for any* $t > 0$*.*

With a consistent MTE, we can further apply it into variable selection problem for linear regression and study its properties.

1.3 Penalized MTE for Variable Selection

Selecting explanatory variables accurately and robustly is one of the most important tasks in modern statistical problems. Due to computational attraction for large scale dataset, penalized methods have been well studied and widely used for linear regression problems. However, it remains challenging to consistently select and estimate coefficients in the presence of contaminations as any model violation could easily cause instability in both selection and estimation. We propose penalized MTE, defined in (1.3), for robust variable selection for linear regression models. Without loss of generality, we assume $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0S}^T, \boldsymbol{\beta}_{0S^c}^T)^T \in \mathbb{R}^d$ to be the true regression coefficients, where $S = \{j : \beta_{0j} \neq 0, j = 1, \ldots, d\} = \{1, \ldots, s\}$ and the cardinality $|S| = s$. In the rest of this section, we develop consistency and asymptotic normality for MTE-Lasso when the dimension of feature space is fixed. We also establish the optimal convergence rate of $\hat{\boldsymbol{\beta}}$ under the modern high dimensional regime, where the dimensionality d is allowed to grow exponentially.

1.3.1 Asymptotic properties with fixed dimensionality

When the number of covariates d is fixed and the sample size $n \to \infty$, the penalized MTE is $\sqrt{n}$-consistent under mild regularity conditions. In addition, it enjoys the oracle property when the penalty function satisfies certain conditions. Let $a_n = \max\{p'_{\lambda_{nj}}(|\beta_{0j}|) : \beta_{0j} \neq 0\}$ and $b_n = \max\{p''_{\lambda_{nj}}(|\beta_{0j}|) : \beta_{0j} \neq 0\}$. We provide following theorems.

Theorem 1.3.1 ($\sqrt{n}$-consistency)**.** *Under the regularity conditions specified in the supplementary materials, suppose $a_n = O_p(n^{-1/2})$, $b_n = o_p(1)$ and $t > 0$, then there exists a local maximizer $\hat{\boldsymbol{\beta}}$, such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 = O_p(n^{-1/2})$.*

Theorem 1.3.2 (Oracle property)**.** *Assume that the penalty function satisfies*

$$\liminf_{n\to\infty} \liminf_{\theta\to 0+} \left\{ \min_{s+1\le j\le d} p'_{\lambda_{nj}}(\beta)/\lambda_{nj} \right\} > 0, \tag{1.11}$$

and the regularization parameter λ_{nj} satisfies

$$\max_{1\le j\le s}(\sqrt{n}\lambda_{nj}) = o_p(1) \quad \textit{and} \quad 1/\min_{s+1\le j\le d}(\sqrt{n}\lambda_{nj}) = o_p(1). \tag{1.12}$$

Suppose $t > 0$, then $\hat{\boldsymbol{\beta}}$ satisfies:

(a) Sparsity: $\hat{\boldsymbol{\beta}}_{S^c} = \mathbf{0}$ with probability 1;

(b) Asymptotic normality for $\hat{\boldsymbol{\beta}}_S$:

$$\sqrt{n}(\mathbf{J}_S + \boldsymbol{\Sigma}_1)\left\{\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S} + (\mathbf{J}_S + \boldsymbol{\Sigma}_1)^{-1}\mathbf{b}\right\} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}_2),$$

where $\mathbf{b} = (p'_{\lambda_{n1}}(|\beta_{01}|)\mathrm{sgn}(\beta_{01}), \ldots, p'_{\lambda_{ns}}(|\beta_{0s}|)\mathrm{sgn}(\beta_{0s}))^T$, $\mathbf{J}_S = \mathbb{E}[\partial^2 \ln_t(f(\mathbf{z};\boldsymbol{\beta}_0))/\partial\boldsymbol{\beta}_S\partial\boldsymbol{\beta}_S^T]$, $\boldsymbol{\Sigma}_1 = \mathrm{diag}(p''_{\lambda_{n1}}(|\beta_{01}|), \ldots, p''_{\lambda_{ns}}(|\beta_{0s}|))$, *and* $\boldsymbol{\Sigma}_2 = \mathrm{cov}[\partial \ln_t(f(\mathbf{z};\boldsymbol{\beta}_0))/\partial\boldsymbol{\beta}_S]$.

Remark 2. *Conditions* (1.11) *and* (1.12) *are necessary for the oracle property, stating that the estimator is able to set apart zero and nonzero regression coefficients with probability 1 as if the true sets are known in advance. It is known that penalty functions such as adaptive-Lasso [72], where the regularization parameters λ_{nj} essentially varies across each individual β_j, satisfy conditions* (1.11) *and* (1.12)*, hence it enjoys oracle property. This property does not hold for estimators with the traditional Lasso penalty function due to the fixed value of regularization parameter λ_n for all regression coefficients. However, the penalized MTE with the traditional Lasso penalty is still a consistent estimator under certain conditions. We show this for high-dimensional regression in section 1.3.2.*

By Theorem 1.3.2, it is straightforward to derive the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_S$,

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_S) = \frac{1}{n}\{\mathbf{J}_S + \boldsymbol{\Sigma}_1\}^{-1}\boldsymbol{\Sigma}_2\{\mathbf{J}_S + \boldsymbol{\Sigma}_1\}^{-1}. \tag{1.13}$$

This asymptotic form of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_S$ enables us to select tuning parameter t that minimizes its determinant. The details are discussed in Section 1.5.2.

1.3.2 Consistency under high dimensional regression

We further consider the penalized MTE under modern high-dimensional linear regression setting, where the number of covariates d is allowed to approach infinity as well as the sample size n in model (1.7). In particular, we consider $\ln(d)/n \to 0$ as $n \to \infty$ and $d \to \infty$. Under this setting, the true coefficient vector $\boldsymbol{\beta}_0$ is usually assumed to be sparse. Regularization method with ℓ_1 penalty is among the popular methods to produce sparse solution for $\hat{\boldsymbol{\beta}}$. Combining MTE with the traditional Lasso penalty function, $p_{\lambda_n}(\boldsymbol{\beta}) = \lambda_n \sum_{j=1}^{d} |\beta_j|$, we are able to establish the statistical consistency of MTE-Lasso by establishing the ℓ_2-norm bound

$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$. Specifically, we rewrite (1.3) as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^d} \left\{ \mathcal{L}(\boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{d} |\beta_j| \right\}, \tag{1.14}$$

where $\mathcal{L}(\boldsymbol{\beta}) = -(1/n)\sum_{i=1}^{n} \ln_t(f(\mathbf{z}_i;\boldsymbol{\beta}))$ is MTE loss function, and λ_n is the regularization parameter of ℓ_1 penalty. Let $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ and define $\mathbb{C}(S) = \{\boldsymbol{\Delta} \in \mathbb{R}^d : 3\|\boldsymbol{\Delta}_S\|_1 \geq \|\boldsymbol{\Delta}_{S^c}\|_1\}$ where $\boldsymbol{\Delta}_S$ and $\boldsymbol{\Delta}_{S^c}$ are the projections of $\boldsymbol{\Delta}$ onto the coordinate sets S and S^c respectively. We further have the following assumptions.

A1 The regressors are bounded, i.e., $\|\mathbf{x}_i\|_\infty = M < +\infty$ for all $i = 1, \ldots, n$.

A2 The design matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^T$ satisfies the restricted eigenvalue condition, $\|\mathbf{X}\boldsymbol{\Delta}\|_2^2/n \geq \kappa_{\mathrm{RE}}\|\boldsymbol{\Delta}\|_2^2$, for all $\boldsymbol{\Delta} \in \mathbb{C}(S)$ where $\kappa_{\mathrm{RE}} > 0$.

A3 For any $\boldsymbol{\nu} \in \mathbb{R}^d$, and $\mathbf{x}_i, i = 1, \ldots, n$, the random variable $\mathbf{x}_i^T\boldsymbol{\nu}$ is sub-Gaussian with parameter at most $\kappa_s^2\|\boldsymbol{\nu}\|_2^2$.

A4 The error term ϵ_i is independently and identically distributed with symmetric distribution with mean 0.

In order to establish the bound for $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2$ in high-dimensional regressions, we need to verify two critical conditions: (1) the boundedness of the gradient of the loss function $\mathcal{L}$ at the true parameter $\boldsymbol{\beta}_0$ and (2) the restricted strong convexity (RSC) condition of the loss function $\mathcal{L}$ in the neighborhood of the true parameter $\boldsymbol{\beta}_0$.

We show that the first condition holds with high probability in the following Lemma.

Lemma 1. *Under Assumption 1, for $t > 0$, we have*

$$P\left(\left\| \frac{\partial \mathcal{L}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}} \right\|_\infty \leq \xi \sqrt{\frac{\ln(d)}{n}} \right) \geq 1 - 2\exp(-\alpha_1 \ln(d)),$$

where $\alpha_1 > 0$ *is a constant,* $\xi = C_t\sqrt{2(\alpha_1+1)}$ *and* $C_t = Mf(\sigma_R)/(t\sigma_R)$.

Lemma 1 shows that $\partial\mathcal{L}(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}$ is bounded with high probability and also provides the form of the bound. This bound plays an important role in deciding the convergence rate of $\hat{\boldsymbol{\beta}}$ as shown in Theorem 1.3.3. Since f represents the normal density function, when t increases, C_t decreases, hence the bound also decreases. It implies that the surface of the loss function around the true parameter $\boldsymbol{\beta}_0$ becomes flatter as t becomes larger. Lemma 1 corresponds to the sub-Gaussian tail condition, which ensures the boundedness of gradient of least square loss [41]. The proof is given in the supplementary materials. In the proof, we particularly discuss the normal density case and give the form of C_t.

It is understood that the estimation error $\hat{\boldsymbol{\Delta}}$ belongs to $\mathbb{C}(S)$ when the regularization parameter $\lambda_n \geq 2\|\partial\mathcal{L}(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}\|_\infty$ [41, Lemma 1, p.543-544]. Therefore, our Lemma 1 suggests that we could choose the regularization parameter $\lambda_n = 2\xi\sqrt{\ln(d)/n}$ in the penalized MTE to force $\hat{\boldsymbol{\Delta}} \in \mathbb{C}(S)$. Such a choice of λ_n is valid with probability at least $1 - 2\exp(-\alpha_2 n\lambda_n^2)$ where $\alpha_2 = \alpha_1/(4\xi^2)$.

Given that $\hat{\boldsymbol{\Delta}} \in \mathbb{C}(S)$, we next verify the RSC condition of the loss function $\mathcal{L}$ to establish the estimation error bound. Before showing the result, we provide the definition of RSC.

Definition 1 (Restricted strong convexity)**.** *The loss function* $\mathcal{L}$ *satisfies restricted strong convexity (RSC) with curvature* $\kappa_1 > 0$ *and tolerance* τ *over the set* $\mathbb{C}(S)$ *if* $\mathcal{L}(\boldsymbol{\beta}_0 + \boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\beta}_0) - [\partial\mathcal{L}(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}]^T\boldsymbol{\Delta} \geq \kappa_1\|\boldsymbol{\Delta}\|_2^2 + \tau^2$ *for all* $\boldsymbol{\Delta} \in \mathbb{C}(S)$.

Lemma 2. *Assume that the random error* ϵ *satisfies the tail condition*

$$P\left(|\epsilon| > \sqrt{c_0 R} - 2\kappa_s\sqrt{\ln n}\right) = \kappa_u \leq \left(1 + \frac{c_0}{c_1}2e^{-3/2}\right)^{-1},$$

where $c_0 = \sigma_R^2$, $c_1 = c_0^{3/2} t\sqrt{2\pi}$, *and* $R = -2\ln(t\sqrt{2\pi c_0})$ *with tuning parameter* t. *Under Assumptions A2–A4, consider the set* $\mathbb{H}(S,u) = \{\boldsymbol{\Delta} \in \mathbb{C}(S) : \|\boldsymbol{\Delta}\|_2 = u\}$, *for any* $u < \sqrt{c_0 R}/(2\kappa_s\sqrt{\ln n})$, *and* $\boldsymbol{\Delta} \in \mathbb{H}(S,u)$, *it holds that*

$$\mathcal{L}(\boldsymbol{\beta}_0 + \boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\beta}_0) - \left(\frac{\partial \mathcal{L}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{\Delta} \geq \kappa_1 \|\boldsymbol{\Delta}\|_2 (\|\boldsymbol{\Delta}\|_2 - \kappa_2 \sqrt{\frac{\ln(d)}{n}} \|\boldsymbol{\Delta}\|_1)$$

with probability at least $1 - \alpha_3 \exp(-\alpha_4 n)$ *for some positive constants* α_3 *and* α_4, *where* $\kappa_1 = (1/c_0 - c_2\kappa_u)\,\kappa_{\mathrm{RE}}/2$, $\kappa_2 = 49\kappa_s^2 c_2\sqrt{\ln n}/\kappa_1$, *and* $c_2 = 1/c_0 + 2e^{-3/2}/c_1$.

As we can see, the curvature of the loss function within the neighborhood of $\boldsymbol{\beta}_0$ in the direction of $\mathbb{C}(S)$ is measured by κ_1. It can be shown that this curvature increases as t decreases to 0. In particular, when t decreases to 0, R increases to $+\infty$, and κ_u decreases to 0. Furthermore, for most of the distributions of ϵ, it is straightforward to show that when t decreases to 0, κ_1 increases to $\kappa_{\mathrm{RE}}/(4c_0)$. It implies that as t decreases, the surface of the loss function become more convex which leads to a better convergence rate.

Note that since $\boldsymbol{\Delta} \in \mathbb{C}(S)$, we have $\|\boldsymbol{\Delta}\|_1 \leq 4\|\boldsymbol{\Delta}_S\|_1 \leq 4\sqrt{s}\|\boldsymbol{\Delta}\|_2$, therefore, the results of Lemma 2 becomes

$$\mathcal{L}(\boldsymbol{\beta}_0 + \boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\beta}_0) - \left(\frac{\partial \mathcal{L}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}}\right)^T \boldsymbol{\Delta} \geq \frac{\kappa_1}{2}\|\boldsymbol{\Delta}\|_2^2,$$

when $n > 64\kappa_2^2 s\ln(d)$. With the results provided by Lemmas 1 and 2, we are able to establish the bound for ℓ_2 norm of the estimation error.

Theorem 1.3.3. *Under the assumptions specified in Lemmas 1 and 2, with regularization parameter* $\lambda_n = 2\xi\sqrt{\ln(d)/n}$, *any of the solutions of equation* (1.14) *in the set* $\mathbb{K}_{\boldsymbol{\beta}_0} = \{\boldsymbol{\beta} + \boldsymbol{\Delta} :$

$\|\boldsymbol{\Delta}\|_2 \leq \sqrt{c_0 R}/(12M\sqrt{s})\}$, *$\hat{\boldsymbol{\beta}}$ satisfies*

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 \leq \frac{8\xi}{\kappa_1}\sqrt{\frac{s \ln d}{n}}$$

with probability at least $1-\alpha_5 \exp(-\alpha_6 n\lambda_n^2)$ *for* $n > \max\{64\kappa_2^2 s \ln d, 16^2\kappa_s^2\xi^2 s \ln n \ln d/(\kappa_1^2 c_0 R)\}$, *where* α_5 *and* α_6 *are positive constants.*

The theorem implies that the convergence rate of $\hat{\boldsymbol{\beta}}$ depends on two critical quantities, the bound of the gradient of the loss function at the true parameter and the curvature of the loss function around the true parameter. In particular, when the loss function becomes flatter at the true parameter and hence has a smaller bound of the gradient, the penalized MTE converges faster. Similarly, when the loss function becomes more convex (i.e. larger curvature) in the restricted direction within the neighborhood of the true parameter (i.e., $\mathbb{C}(S)$), the penalized MTE also converges faster.

However, as illustrated by Lemmas 1 and 2, the effects of t on these two quantities are often in the opposite directions. For example, as t increases, the entire loss function generally becomes flatter which leads to a smaller bound of the gradient at the true parameter. But an increasing t also leads to a smaller curvature. Therefore, selecting t involves controlling both the bound and the curvature. To gain a faster convergence rate, we need t to be large to control the bound of the gradient, but also need t to be small to increase the curvature of the loss function. Therefore, a trade-off has to be made when selecting t. Note that when $t > f(0)$, the penalized MTE becomes penalized minimum ℓ_2 distance estimation. Therefore, penalized MTE offers a more refined trade-off between efficiency and robustness.

1.4 Robustness Properties

In addition to the properties presented in the manuscript, we also derive the finite sample breakdown point and influence function that characterize the robustness properties of our proposed penalized MTE.

The finite sample breakdown point for an estimator is the maximum proportion of contaminated data points in a sample that the estimator can tolerate before it yields an arbitrarily bad result (i.e., breakdown). The finite sample breakdown point was introduced by Donoho and Huber [16]. Earlier versions of breakdown point can be also found in Hodges Jr [30], Hampel [26] and Hampel [24]. The finite sample breakdown point quantifies the estimator's overall resistance to outliers. Under regression settings, estimators such as MM-estimator [64] and S-estimator [45] can achieve the highest breakdown point of 1/2.

We denote the entire sample as $\mathbf{Z}_n = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$. Among these n observations, there are m bad data points $\mathbf{Z}_m = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$ and $n-m$ good data points $\mathbf{Z}_{n-m} = \{\mathbf{z}_{m+1}, \ldots, \mathbf{z}_n\}$. Following Donoho and Huber [16], the finite sample breakdown point is defined as

$$\mathrm{BP}(\hat{\boldsymbol{\beta}}; \mathbf{Z}_{n-m}) = \min \left\{ \frac{m}{n} : \sup_{\mathbf{Z}_m} \|\hat{\boldsymbol{\beta}}(\mathbf{Z}_n) - \hat{\boldsymbol{\beta}}(\mathbf{Z}_{n-m})\|_2 = \infty \right\}.$$

The following theorem shows the finite sample breakdown point of the penalized MTE.

Theorem 1.4.1 (Finite sample breakdown point)**.** *Let*

$$a_{nm} = \frac{1}{n-m} \max_{\boldsymbol{\beta} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \# \left\{ i : m+1 \le i \le n, \mathbf{x}_i^T \boldsymbol{\beta} = 0 \right\}.$$

For strictly increasing penalty function $p_{\lambda_{nj}}(|\beta_j|) = \lambda_{nj} g(|\beta_j|)$ *defined on* $[0, \infty]$ *with regularization parameter* $\lambda_{nj} > 0$*, if* $m/n \le \varepsilon < (1 - 2a_{nm})/(2 - 2a_{nm})$*,* $a_{nm} < 0.5$*, and*

$\zeta(t) < (1-\varepsilon)(1-a_{nm})$ *hold, then, for any initial estimator* $\tilde{\tilde{\boldsymbol{\beta}}}$*, we have*

$$\mathrm{BP}(\hat{\boldsymbol{\beta}}; \mathbf{Z}_{n-m}, t) \geq \min\left\{\mathrm{BP}(\tilde{\tilde{\boldsymbol{\beta}}}; \mathbf{Z}_{n-m}), \frac{1-2a_{nm}}{2-2a_{nm}}, 1-\frac{\zeta(t)}{2-2a_{nm}}\right\},$$

where $\zeta(t) = \frac{2m}{n} + \frac{2}{n}\sum_{i=m+1}^{n}\phi_t(y_i - \mathbf{x}_i^T\tilde{\tilde{\boldsymbol{\beta}}})$*, and* $\phi_t(r) = \frac{\ln_t(f(0))-\ln_t(f(r))}{\ln_t(f(0))-\ln_t(0)}$.

By Theorem 1.4.1, we can obtain a lower bound of the breakdown point for our penalized MTE. This lower bound depends on the breakdown point of the initial estimate $\tilde{\tilde{\boldsymbol{\beta}}}$. In addition, by the definition of a_{nm}, when $n \to \infty$, $(1-2a_{nm})/(2-2a_{nm}) \to 1/2$ and $1 - \zeta(t)/(2-2a_{nm}) \to 1/2$ if t is chosen such that $\zeta(t) \in (0, 1]$. Therefore, with an appropriate t and initial estimate, we can have the breakdown point of the penalized MTE as high as $1/2$. Note that the theorem holds for strictly increasing and unbounded penalty functions such as Lasso type penalty that has been adopted in this paper. For other types of bounded penalty functions, the theorem remains a challenge.

The influence function [25] is another important measure of robustness. It measures the effect of an infinitesimal contamination to an estimator. Let $T(F)$ be the estimator of interest (i.e., penalized MTE) at population level under the assumed distribution F, i.e., $T(F) = \arg\max_{\boldsymbol{\beta}}\{\int \ln_t(f(y-\mathbf{x}^T\boldsymbol{\beta}))dF - \sum_{j=1}^{d} p_{\lambda_{0j}}(|\beta_j|)\}$ where $p_{\lambda_{nj}}(\cdot)$ is supposed to have a limit denoted as $p_{\lambda_{0j}}(\cdot)$. Suppose we have a mixture distribution $F_\varepsilon = (1-\varepsilon)F + \varepsilon\delta_{\mathbf{z}}$ where $\delta_{\mathbf{z}}$ represents a point mass distribution function at fixed point $\mathbf{z} = (y^*, \mathbf{x}^{*T})^T$. Then the influence function of $T(F)$ at the point $\mathbf{z}$ is defined as $\mathrm{IF}(\mathbf{z}, T) = \lim_{\varepsilon\to 0}(T(F_\varepsilon) - T(F_0))/\varepsilon$. The following theorem shows that the influence function of the penalized MTE is bounded in the response variable domain.

Theorem 1.4.2 (Influence function)**.** *For the penalized MTE, the j-th element of its influence function* $\mathrm{IF}_j(\mathbf{z}, T)$ *takes the following form*

$$\mathrm{IF}_j(\mathbf{z}, T) = \begin{cases} 0 & \text{if } \beta_{0j} = 0 \\ \left[(-\int \frac{\partial^2}{\partial r^2} \ln_t(f(r))\mathbf{x}\mathbf{x}^T dF + v^*)^{-1}\right]_j & \\ \quad \times(\frac{\partial}{\partial r} \ln_t(f(r^*))(-\mathbf{x}^*) - v) & \text{if } \beta_{0j} \neq 0, \end{cases}$$

where $[\cdot]_j$ *denotes the j-th row of a matrix,* $r^* = y^* - \mathbf{x}^{*T}\boldsymbol{\beta}_0$*,* $v^* = \mathrm{diag}(p''_{\lambda_{01}}(|\beta_{01}|), \ldots, p''_{\lambda_{0d}}(|\beta_{0d}|))$*, and* $v = (p'_{\lambda_{01}}(|\beta_{01}|)\mathrm{sgn}(\beta_{01}), \ldots, p'_{\lambda_{0d}}(|\beta_{0d}|)\mathrm{sgn}(\beta_{0d}))^T$*.*

Note that $\boldsymbol{\beta}_0$ is the true parameter and $\boldsymbol{\beta}_0 = T(F_0)$ because of the consistency. Following condition (1.12) in Theorem 1.3.2, $\lambda_{0j} = 0$ if $\beta_{0j} \neq 0$ and $\lambda_{0j} = +\infty$ if $\beta_{0j} = 0$. Therefore, for the zero coefficients, the influence function is equal to zero, while for the nonzero coefficients, the influence function can be written as

$$\mathrm{IF}_j(\mathbf{z}, T) = \left[(-\int \frac{\partial^2}{\partial r^2} \ln_t(f(r))\mathbf{x}\mathbf{x}^T dF + v^*)^{-1}\right]_j (\frac{\partial}{\partial r} \ln_t(f(r^*))(-\mathbf{x}^*)).$$

1.5 Tuning Parameters and Algorithm

1.5.1 Choice of regularization parameter λ

The performance of penalized estimator heavily relies on the choice of the regularization parameter λ. For fixed dimensional regression, in order to achieve oracle property, we apply adaptive-Lasso penalty, where the λ_{nj} is chosen to satisfy condition (1.12). A simple BIC-type criterion [56, 60] is adopted for choosing the optimal $\hat{\lambda}_{nj}$. To be specific, we minimize

the following objective function

$$-\sum_{i=1}^{n} \ln_t \left(f(\mathbf{z}_i;\boldsymbol{\beta})\right) + n\sum_{j=1}^{d} \lambda_{nj}|\beta_j| - \ln(0.5n\lambda_{nj})\ln(n),$$

which leads to the solution of regularization parameter

$$\hat{\lambda}_{nj} = \frac{\ln(n)}{n|\tilde{\tilde{\beta}}_j|}, \tag{1.15}$$

where $\tilde{\tilde{\beta}}_j$ is an initial estimate of β_j. It is easy to see that the necessary condition for the oracle property (1.12) is satisfactory with above choice of regularization parameter. Note that for high-dimensional regression, our theoretical properties are established based on Lasso penalty, where the regularization parameter λ_n does not depend on jth regression coefficient β_j. Therefore the choice of regularization parameter as shown in (1.15) is not applicable due to lack of theoretical justification. Instead, the optimal λ_n is determined by minimizing median absolute prediction error through cross-validation over a grid.

1.5.2 Choice of tuning parameter t

As discussed in Sections 1.2 and 1.3, the tuning parameter t controls the trade-off between robustness and efficiency, hence the choice of t cannot be neglected. We use a simple data-driven method to grid search the optimal value of t such that it minimizes the determinant of asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}_S$ as in (1.13). The idea of this approach is that t is selected such that the proposed estimator has minimum variance in order to achieve high efficiency. Similar approach has been adopted by Wang et al. [60] to select the tuning parameter in the exponential squared loss function. As an illustration, Figure 1.3 shows one example of the

value of the determinant of (1.13) denoted as $\hat{H}(t)$ against different values of t under fixed dimensional setting.

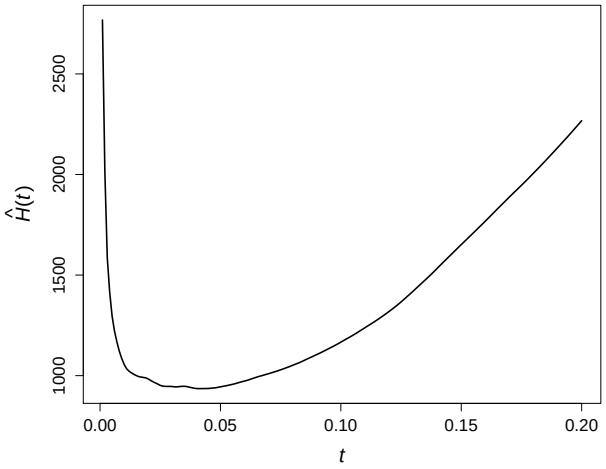

Figure 1.3: Determinant of covariance matrix $\hat{H}(t)$ against t

It remains challenging to select t under high dimensional settings due to instability of large variance-covariance matrix estimation. Certain regularized approaches can be adopted, but most of them require significant computing effort. To be computationally efficient, in practice we fix the value of t at first, and then the proposed data-driven approach is applied once the number of nonzero estimates are significantly dropped. This is often achieved after the first one or two iterations.

To see the appropriateness of this approach, we demonstrate that t has a relatively large forgiven region, in which model estimates differ slightly. With extensive simulations of both clean and contaminated data, Figure 1.4 shows the change of model error [18] along with its

1-standard error (vertical bars) across different values of t. The model error is defined as

$$\text{ME} = \frac{1}{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \mathbf{X}^T \mathbf{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0). \tag{1.16}$$

The clean data (left figure) are generated as follows: $\epsilon \overset{i.i.d.}{\sim} N(0,1)$, $\mathbf{x} \in \mathbb{R}^{500}$ are independently drawn from $N(\mathbf{0}, \mathbf{I}_{500})$, while the contaminated data (right figure) are generated by replacing 20% of the random error ϵ with heavy tailed data from $N(0, 20^2)$. Clearly, only if $t = 0$ under contaminated data is the model error extremely large, the model errors are very similar for the rest of nonzero t. However, we notice that the standard error increases when $t > 0.4$ roughly. This is due to the penalty function in the penalized MTE, where only the first component (tangent likelihood) involves t, so that optimizing the penalized MTE is more sensitive for $t > \max(f(\mathbf{z}, \boldsymbol{\beta}))$.

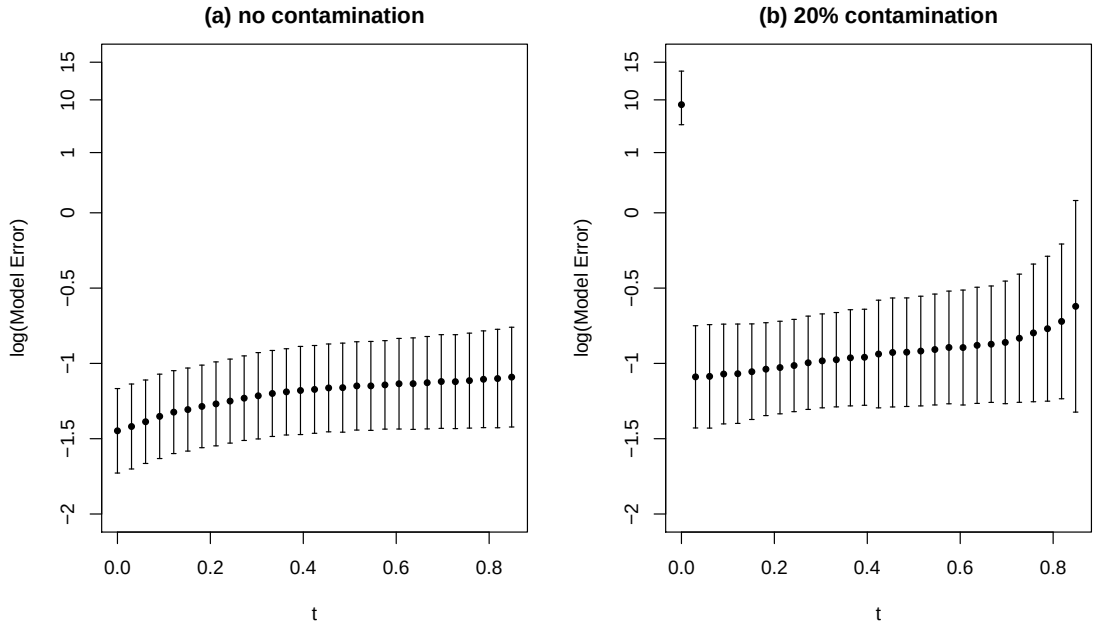

Figure 1.4: Model error against different initial value of t. The dot is mean of model error and vertical bar indicates 1-standard error over 100 random samples.

1.5.3 Choice of initial values

When solving the optimization problem (1.3) and (1.14), MTE could potentially lead to local maximums as the tangent likelihood loss function is nonconvex. Therefore, assigning suitable initial values for the optimization is critical. For our proposed method, we need to assign initial values for $\boldsymbol{\beta}$ as well as the preliminary scale estimate σ_R^2. For $\boldsymbol{\beta}$, we can use unpenalized LAD estimates as a candidate initial value because LAD is a monotone regression M-estimate whose objective function is always convex. For σ_R^2, we have adopted one of the well known robust scale parameter estimates, $\sigma_R = 1.4826 \times \text{MAD}$, where MAD can be the median absolute deviance of residuals from LAD estimates, as the initial estimate. Other types of robust scale parameter estimation are also well developed and available [46] to serve as potential initial values.

1.5.4 Computational algorithm

Coordinate descent (CD) algorithm has recently been well recognized and appreciated for its surprisingly fast and efficient capability in solving ℓ_1-regularization problem . It updates a single parameter one at a time while the rest are fixed. We choose the coordinate descent algorithm for its simplicity, speed and stability [61, 20, 21, 6], and apply it for both fixed and high-dimensional regression settings. We propose following 2-step iterative algorithm.

Step 1. Update tuning parameter t and λ_{nj}: Given current estimates $\hat{\boldsymbol{\beta}}^{(k-1)}$, find optimal value $t^{(k)}$ such that $t^{(k)}$ minimizes the determinant of (1.13) by grid search. Meanwhile, the optimal regularization parameter $\hat{\lambda}_{nj}^{(k)}$ can be calculated by (1.15). For high-dimensional problem, choice of t and λ follows approaches we discussed earlier to reduce computational effort.

Step 2. Update parameter estimates: Based on $t^{(k)}$ and $\hat{\lambda}_{nj}^{(k)}$ that are obtained from Step 1, we use the coordinate descent algorithm to solve the optimization problem (1.3), and obtain updated coefficients $\hat{\boldsymbol{\beta}}^{(j)}$. Then the scale parameter $\hat{\sigma}_R$ can be estimated using robust scale estimator such as $1.4826 \times \text{MAD}$. Repeat Steps 1 and 2 until $\hat{\boldsymbol{\beta}}$ converges.

This algorithm is directly applicable to both the fixed and high-dimensional regression settings with little modification (the optimal regularization parameter λ_n is chosen by cross-validation, and need not to be updated between two steps). In practice, the range of t in the grid-search procedure can be set from 0 to 0.2 in order to maintain high efficiency. From our limited numerical studies, the algorithm is computationally efficient with fast convergence.

1.6 Numerical Studies

1.6.1 Location parameter estimation

We first demonstrate the advantage of our proposed MTE for a single location parameter estimation $\hat{\mu}$. As discussed in Section 1.2.1, due to the redescending influence function, MTE offers more efficiency than traditional M-estimators such as Huber's and least absolute deviance (LAD) loss, while it maintains high robustness. This has been numerically evidenced in Figure 1.5, which compares MTE (black solid line) with different methods including MLE (red dashed line), LAD (green dotted line), Huber's loss (blue dash-dotted line), and a well-known redescending M-estimator, Tukey's Bisquare loss (light-green dashed line).

Figure 1.5 shows the mean squared error of normal mean estimates across different contamination ratios. We generate 1000 random samples with each sample size being 500, where the clean data are from standard normal distribution, while contaminations are from

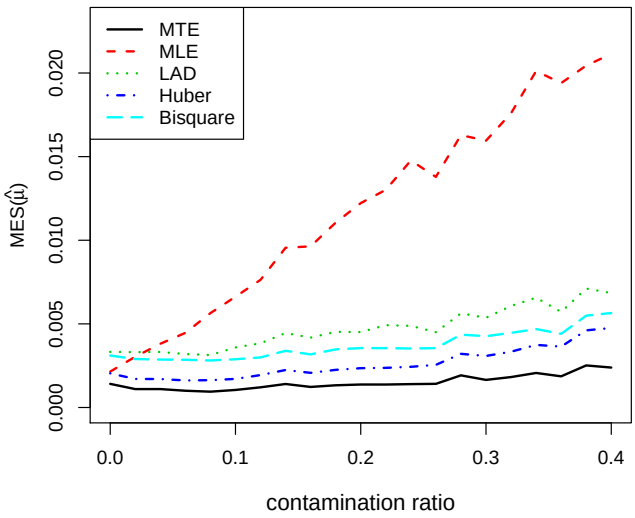

Figure 1.5: Mean squared error of normal mean estimation by different estimators. 1000 random samples are generated for different settings of contamination ratio from 0 to 40%.

$N(0, 5^2)$. For each individual sample, the optimal tuning parameters of MTE, Huber's, and Bisquare loss are chosen over a grid such that the squared error of $\hat{\mu}$ is minimum [1]. Clearly, our proposed MTE dominates the alternatives at all contamination ratio. For clean data, where contamination ratio is zero, although theoretically MTE and MLE should have exactly the same (optimal) performance, the little difference is due to bias brought by sample size. The form of Huber loss is defined as

$$\rho_h(z) = \begin{cases} z^2 & \text{if } |z| \leq \alpha^{-1}; \\ 2\alpha^{-1}|z| - \alpha^{-2} & \text{if } |z| > \alpha^{-1}, \end{cases}$$

[1] We minimize squared error given the true parameter in order to have all candidate methods to produce their best possible estimation results hence fair enough to be compared, although this selection criteria is not possible for real data due to unknown true parameters.

and Tukey's Bisquare loss is defines as

$$\rho_b(z) = \begin{cases} \frac{k^2}{6}\left\{1 - \left[1 - \left(\frac{z}{k}\right)^2\right]^3\right\} & \text{if } |z| \leq k; \\ \frac{k^2}{6} & \text{if } |z| > k. \end{cases}$$

We further breakdown above settings and show how the statistical efficiency changes across different tuning parameter t. Under three different settings: (1) clean data ($z \overset{i.i.d.}{\sim} N(0,1)$); (2) 10% contamination; and (3) 20% contamination, where the contaminated data are from $N(0, 5^2)$, Figure 1.6 compares MTE with Huber's, LAD and MLE in terms of empirical efficiency, which is calculated as $\mathcal{I}^{-1}(\mu)/Var(\hat{\mu})$, where $\mathcal{I}(\mu) = 1/N$ is the fisher information given true parameter. With clean data (left), MLE is the most efficient estimator, while LAD has the lowest efficiency. MTE and Huber's loss achieve the highest efficiency when they are in the special forms that are equivalent to MLE, i.e., $t = 0$, $\alpha = 0$, while their efficiency decrease as the tuning parameters increase. In both contaminated data (middle and right), MLE loses efficiency significantly due to the model violation, while LAD maintains relatively good efficiency as it resists to outliers in certain degree. MTE again has the highest efficiency when the tuning parameter t is optimally chosen.[2]

Last, we also numerically demonstrate the advantage of MTE when the data are asymmetrically distributed. Similar to the settings as in Figure 1.5, we change the distribution of contaminations to $N(5, 5^2)$ so that the sample as a whole is asymmetric. Figure 1.7 shows both MSE in log scale (left) and variance (right) of the mean estimation. Again, for all different ratios of contamination, MTE produces the dominating results, while Tukey's bisquare loss performs similar to MTE only when the data is heavily contaminated.

[2]Note that the tuning parameters of MTE and Huber's loss are not associated vertically in Figure 1.6. We plot their efficiency against relatively larger range of the tuning parameters as we are rather interested in the optimal efficiency for both estimators and how it changes across the tuning parameter.

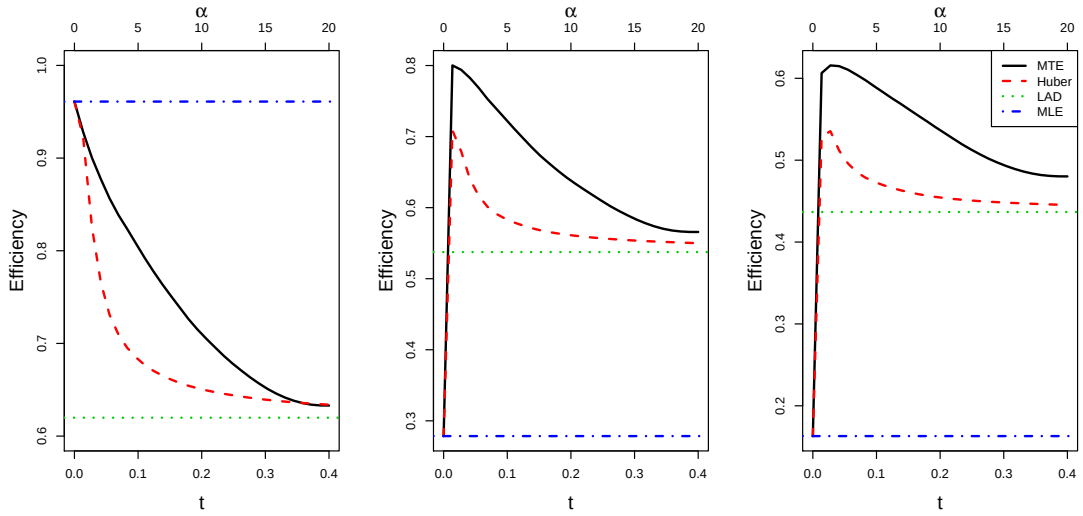

Figure 1.6: Empirical efficiency of mean estimation under clean (left), 10% (middle) and 20% contaminated data (right). Clean data are from standard normal distribution, and contaminations are from $N(0, 5^2)$.

1.6.2 Fixed dimensional regressions

For fixed dimensional regression, in order to achieve oracle estimates, we adopt the adaptive-Lasso penalty for MTE as well as its competitors, LAD [56], ESL [60], CQR [73] and MLE [3] [72]. The criteria used for comparison are median and median absolute deviation (MAD) of model error as defined in (1.16). Model selection errors are measured by false negative rate (FNR) and false positive rate (FPR). Specifically, FNR is defined as the proportion of zero coefficient estimates whose corresponding true coefficients are nonzero, i.e., $\#\{j : \hat{\beta}_j = 0, \beta_{0j} \neq 0\}/\#\{j : \beta_{0j} \neq 0\}$. FPR is defined as the proportion of nonzero coefficient estimates whose corresponding true coefficients are zero, i.e., $\#\{j : \hat{\beta}_j \neq 0, \beta_{0j} = 0\}/\#\{j : \beta_{0j} = 0\}$.

We set the true regression coefficient $\boldsymbol{\beta}_0 = (1, 1.5, 2, 1, 0, 0, 0, 0, -2.5, -1, 0, 0)^T \in \mathbb{R}^{12}$, and consider following simulation designs: (1) $\epsilon_i \overset{\text{iid}}{\sim} 0.7N(0, 1) + 0.3\text{Unif}(-10, 50)$

[3]For CQR and MLE with adaptive-Lasso penalty, we directly employ the existing `R` packages `cqrReg` and `parcor`, respectively.

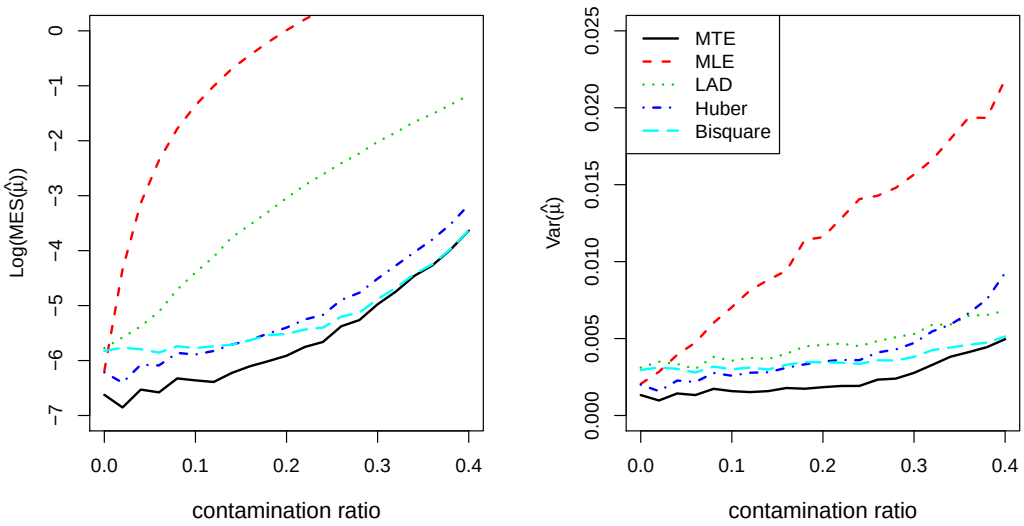

Figure 1.7: Mean squared error (in log scale) and variance of the mean estimation by different estimators. 1000 random samples are generated for different settings of contamination ratio from 0 to 40%.

and $\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, \mathbf{\Omega})$; (2) $\epsilon_i \overset{\text{iid}}{\sim} 0.7N(0,1) + 0.3N(10, 10^2)$ and $\mathbf{x}_i \overset{\text{iid}}{\sim} 0.8N(\mathbf{0}, \mathbf{I}) + 0.2N(\mathbf{3}, \mathbf{\Omega})$, where $\mathbf{I}$ is a 12×12 identity matrix, and $\mathbf{\Omega} = \{\Sigma_{ij}\}_{12\times 12}$ is a 12×12 covariance matrix with $\Sigma_{ij} = 0.5^{|i-j|}$. Under each setting, we simulate 1000 Monte Carlo samples for different sample sizes, $n = 100, 200, 400, 800$. The results are reported in Tables 1.1 and 1.2.

As Tables 1.1 and 1.2 illustrate, MTE outperforms all other methods in terms of model errors and variable selection accuracy. As the sample size n increases, the performance of all methods improve, but MTE dominates all other methods uniformly.

1.6.3 High dimensional regressions

We further demonstrate the performance of MTE under high-dimensional regression settings with $d = 500$ through a Monte Carlo simulation. We set the true coefficient $\boldsymbol{\beta}_0 = (3, 1.5, 2, -2.5, -2, 3, 1.5, 2, -2.5, -2, 0, \ldots, 0)^T \in \mathbb{R}^{500}$, a 500-dimensional coefficient

Table 1.1: Monte Carlo Simulation for regression models with error following mixture distribution: $\epsilon_i \overset{\text{iid}}{\sim} 0.7N(0,1) + 0.3\text{Unif}(-10,50)$ and covariates following distribution: $\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, \mathbf{\Omega})$.

				Model Error	
n	Method	FNR	FPR	Median	MAD
100	MTE	0.010	0.000	0.126	0.054
	LAD	0.019	0.006	0.237	0.113
	ESL	0.557	0.000	3.198	2.960
	CQR	0.343	0.234	10.951	1.367
	MLE	0.649	0.136	16.884	5.112
200	MTE	0.000	0.000	0.056	0.022
	LAD	0.001	0.002	0.097	0.040
	ESL	0.387	0.000	2.208	2.111
	CQR	0.334	0.204	10.202	0.923
	MLE	0.460	0.191	10.054	3.880
400	MTE	0.000	0.000	0.025	0.010
	LAD	0.000	0.000	0.046	0.019
	ESL	0.014	0.000	0.111	0.066
	CQR	0.333	0.169	9.932	0.561
	MLE	0.286	0.220	4.877	1.746
800	MTE	0.000	0.000	0.011	0.005
	LAD	0.000	0.000	0.021	0.009
	ESL	0.000	0.000	0.030	0.012
	CQR	0.333	0.141	9.818	0.346
	MLE	0.175	0.225	2.627	0.843

vector with 3 non-zeros. We conduct 100 Monte Carlo simulations from model (1.7) with sample size $n = 200$. We consider three types of covariates: (1) $\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, \mathbf{I})$; (2) $\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, \mathbf{\Omega})$; and (3) $\mathbf{x}_i \overset{\text{iid}}{\sim} 0.8N(\mathbf{0}, \mathbf{I}) + 0.2N(\mathbf{3}, \mathbf{\Omega})$, where $\mathbf{I}$ is a $d \times d$ identity matrix, and $\mathbf{\Omega} = \{\Sigma_{ij}\}_{d\times d}$ with $\Sigma_{ij} = 0.5^{|i-j|}$. We also consider six types of random errors:

(1) $\epsilon_i \overset{\text{iid}}{\sim} N(0,1)$;

(2) $\epsilon_i \overset{\text{iid}}{\sim} 0.8N(0,1) + 0.2N(0,20^2)$;

(3) $\epsilon_i \overset{\text{iid}}{\sim} 0.8N(0,1) + 0.2N(50,10^2)$;

(4) $\epsilon_i \overset{\text{iid}}{\sim} 0.6N(0,1) + 0.2N(20,10^2) + 0.2N(-50,10^2)$;

Table 1.2: Monte Carlo Simulation for regression models with random error following mixture distribution: $\epsilon_i \overset{\text{iid}}{\sim} 0.7N(0,1) + 0.3N(10, 10^2)$ and covariates following mixture distribution: $\mathbf{x}_i \overset{\text{iid}}{\sim} 0.8N(\mathbf{0}, \mathbf{I}) + 0.2N(\mathbf{3}, \mathbf{\Omega})$.

				Model Error	
n	Method	FNR	FPR	Median	MAD
100	MTE	0.009	0.001	0.126	0.058
	LAD	0.011	0.007	0.263	0.122
	ESL	0.654	0.000	9.336	9.073
	CQR	0.336	0.206	35.432	4.658
	MLE	0.306	0.255	7.584	2.743
200	MTE	0.000	0.000	0.057	0.023
	LAD	0.000	0.002	0.125	0.051
	ESL	0.278	0.000	2.269	2.144
	CQR	0.333	0.172	32.780	3.070
	MLE	0.137	0.295	4.639	1.372
400	MTE	0.000	0.000	0.025	0.010
	LAD	0.000	0.001	0.066	0.026
	ESL	0.000	0.000	0.085	0.033
	CQR	0.333	0.153	31.479	1.721
	MLE	0.051	0.294	3.042	0.781
800	MTE	0.000	0.000	0.012	0.005
	LAD	0.000	0.001	0.043	0.015
	ESL	0.000	0.000	0.027	0.017
	CQR	0.333	0.129	30.924	1.241
	MLE	0.008	0.267	2.257	0.467

(5) $\epsilon_i \overset{\text{iid}}{\sim}$ Cauchy;

(6) $\epsilon_i \overset{\text{iid}}{\sim} t(2)$.

We compare our methods to famous robust estimators, Huber (RA-Lasso) [19] and LAD-Lasso [59]. All methods are equipped with Lasso penalty function. We also add traditional LASSO (implemented using `R` package `parcor`) in the comparison. The optimal tuning parameter λ is chosen by minimizing median absolute prediction error through cross-validation. Figure 1.8 shows the box plots of model errors. The range of vertical axis is truncated from above for better comparison. As we can see, traditional LASSO estimator fails when the

data is contaminated. For the rest three robust estimator, MTE performs the best in most scenarios. We exclude CQR in the comparison because the `R` package `cqrReg` yields poor performance using the default algorithm and may not be appropriate for high-dimensional settings. We do not include ESL because to our best knowledge, there is no published work that studies ESL in high-dimensional regression.

We also report mean, median and MAD of model errors in Table 1.3. In addition, we further investigate the variable selection accuracy, and report the averaged counts of true positive covariates (TP) and false positive covariates (FP), i.e., $\text{TP} = \#\{j : \hat{\beta}_j \neq 0, \beta_{0j} \neq 0\}$ and $\text{FP} = \#\{j : \hat{\beta}_j \neq 0, \beta_{0j} = 0\}$.

1.6.4 Real data examples

We demonstrate the performance of the proposed penalized MTE using some real data examples. We first apply it to Boston housing price dataset (https://archive.ics.uci.edu/ml/datasets/Housing), which is commonly used as an example for regressions. It is particularly of interest for robust regression analysis as the dataset contains outliers and skewed variables. There are 14 variables in total: *medv*, *rm*, *tax*, *ptratio*, *lstat*, *nox*, *dis*, *crim*, *zn*, *indus*, *age*, *black*, *chas*, *rad*. Detailed explanations of these variables can be found in the supplementary materials. We use *medv* (median house price) as the response variable. Following Wu et al. [62] and references therein, we take logarithm of variables *crim*, *lstat* and *tax*, and standardize all variables before fitting the model. Table 1.4 gives the variable selection results. Standard errors are obtained based on 500 bootstrapping samples. We find that the traditional adaptive-Lasso (MLE) selects many (10 out of 13) variables. MTE and CQR select 5 variables *rm*, *ln(tax)*, *ptratio*, *ln(stat)*, and *dis*. This finding is largely consistent with variables commonly used in the literature. For example, four variables *rm*,

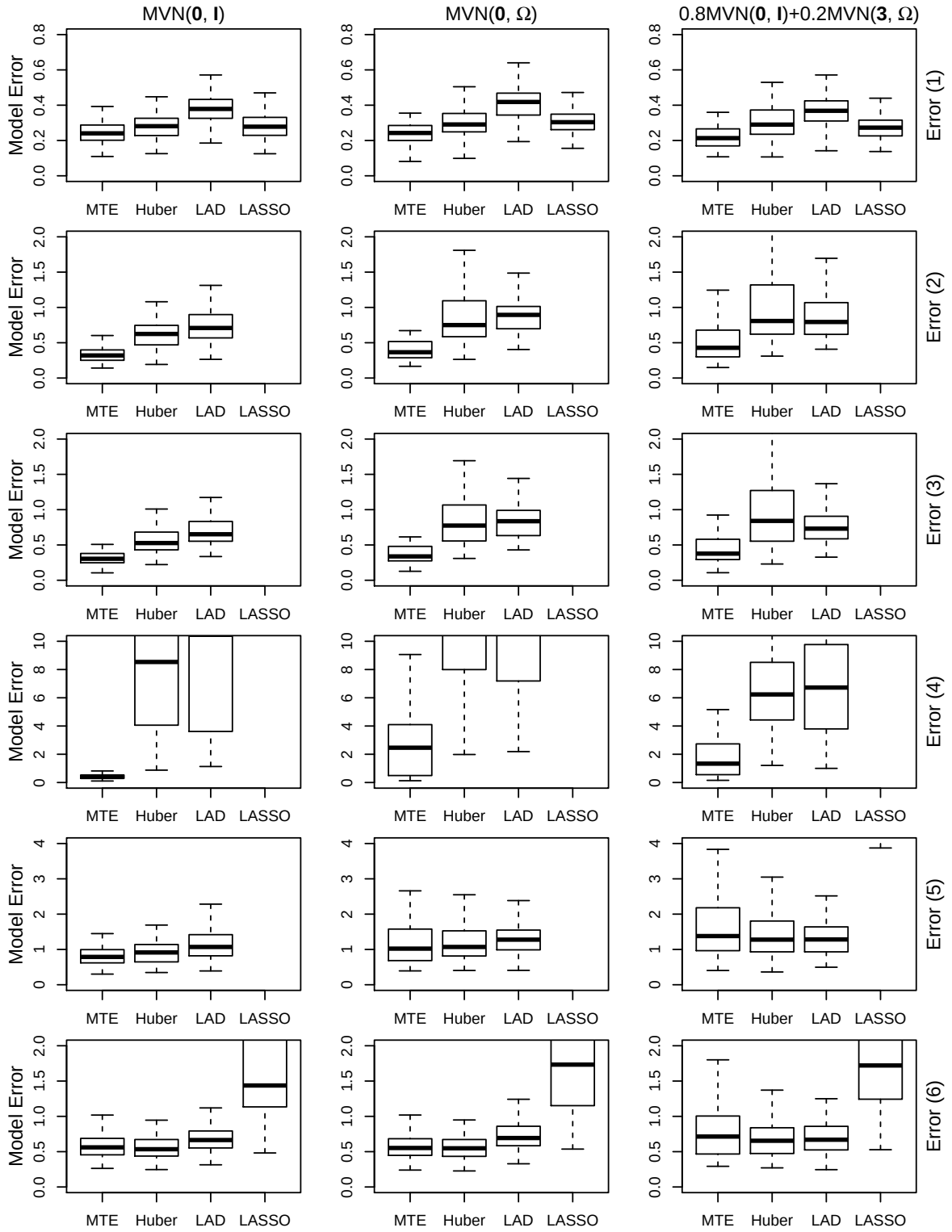

Figure 1.8: Box plots of model errors for different methods. Six types of errors are in row direction and three types of covariates are in column direction.

Table 1.3: Comparison of MTE, Huber, LAD and LASSO on model error and variable selection accuracy under high-dimensional regression setting with $n = 200, d = 500$. TP is the average count of correctly estimated nonzero coefficients; and FP is the average count of nonzero estimates whose corresponding true coefficients are zero. Note that there are 10 nonzero and 490 zero true coefficients in total. The average is based on 100 Monte Carlo simulations.

ϵ		$\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, \mathbf{I})$					$\mathbf{x}_i \overset{\text{iid}}{\sim} N(\mathbf{0}, \mathbf{\Omega})$					$\mathbf{x}_i \overset{\text{iid}}{\sim} 0.8N(\mathbf{0}, \mathbf{I}) + 0.2N(\mathbf{3}, \mathbf{\Omega})$				
		Mean	Med.	MAD	TP	FP	Mean	Med.	MAD	TP	FP	Mean	Med.	MAD	TP	FP
$\epsilon(1)$	MTE	0.24	0.24	0.04	10.0	28.4	0.29	0.24	0.04	9.9	26.9	0.25	0.21	0.05	10.0	18.6
	Huber	0.28	0.28	0.05	10.0	29.5	0.31	0.29	0.05	10.0	29.9	0.33	0.29	0.06	9.9	26.3
	LAD	0.37	0.38	0.05	10.0	48.2	0.41	0.42	0.06	10.0	55.0	0.37	0.37	0.06	10.0	51.5
	Lasso	0.28	0.28	0.05	10.0	41.6	0.30	0.30	0.04	10.0	43.9	0.28	0.27	0.04	10.0	43.8
$\epsilon(2)$	MTE	0.33	0.32	0.07	10.0	21.9	0.68	0.37	0.10	9.9	26.3	0.64	0.43	0.14	9.9	20.7
	Huber	0.64	0.62	0.14	10.0	25.4	1.12	0.75	0.22	9.9	28.2	1.05	0.81	0.26	9.9	25.3
	LAD	0.77	0.71	0.16	10.0	47.2	0.93	0.89	0.16	10.0	50.4	0.88	0.79	0.20	10.0	48.0
	Lasso	21.20	19.97	4.63	8.3	33.5	21.45	20.96	3.56	6.3	27.1	16.49	16.36	1.91	3.9	19.1
$\epsilon(3)$	MTE	0.31	0.30	0.06	10.0	23.4	0.78	0.34	0.08	9.8	31.8	0.58	0.38	0.12	9.8	24.6
	Huber	0.57	0.53	0.11	10.0	26.6	1.16	0.77	0.22	9.9	33.4	1.00	0.84	0.35	9.8	35.0
	LAD	0.71	0.65	0.13	10.0	51.4	0.83	0.84	0.18	10.0	58.4	0.76	0.73	0.15	10.0	56.3
	Lasso	48.21	48.55	3.57	0.4	1.0	45.78	46.79	2.61	0.5	1.2	24.89	24.62	3.88	0.4	2.8
$\epsilon(4)$	MTE	1.01	0.39	0.13	9.8	16.5	2.91	2.46	1.95	9.2	22.3	1.78	1.34	0.88	9.4	29.1
	Huber	11.12	8.54	5.50	9.0	23.5	13.19	12.42	4.50	7.8	22.7	6.51	6.24	1.89	8.2	29.5
	LAD	12.34	10.46	7.27	8.8	37.4	12.88	11.66	5.64	8.2	38.4	7.08	6.72	2.93	8.3	35.7
	Lasso	50.70	50.16	4.26	0.7	3.9	47.81	47.50	3.77	0.6	3.7	27.35	27.24	4.20	0.3	5.0
$\epsilon(5)$	MTE	0.86	0.79	0.19	10.0	22.4	1.38	1.02	0.42	9.8	25.0	1.66	1.38	0.57	9.8	34.1
	Huber	0.97	0.91	0.25	10.0	28.2	1.29	1.07	0.28	9.9	30.1	1.42	1.28	0.39	9.8	31.7
	LAD	1.15	1.07	0.27	10.0	47.1	1.37	1.28	0.28	10.0	52.2	1.32	1.28	0.35	10.0	46.7
	Lasso	35.90	40.87	12.12	4.0	14.1	35.00	40.36	10.78	3.1	13.1	21.09	20.67	6.97	2.2	9.6
$\epsilon(6)$	MTE	0.59	0.56	0.12	10.0	26.7	0.71	0.55	0.12	9.9	26.8	0.88	0.71	0.26	9.9	23.7
	Huber	0.56	0.53	0.11	10.0	29.2	0.60	0.55	0.12	10.0	27.5	0.72	0.65	0.18	9.9	29.0
	LAD	0.69	0.66	0.12	10.0	50.1	0.72	0.69	0.14	10.0	52.3	0.70	0.67	0.15	10.0	50.0
	Lasso	2.96	1.44	0.44	9.9	38.7	3.42	1.73	0.71	9.8	43.1	2.64	1.72	0.62	9.7	41.8

ln(tax), *ptratio*, and *ln(stat)* are considered in Opsomer and Ruppert [43], Yu and Lu [65] and Wu et al. [62], whereas three variables *rm*, *ln(stat)*, *dis* are used in Chaudhuri et al. [11].

Next, we apply the proposed method to an expression quantitative trait loci (eQTL) dataset under a high-dimensional regression. The dataset can be accessed at NCBI Gene Expression Omnibus data repository (http://www.ncbi.nlm.nih.gov/geo) with access number GSE3330. The dataset contains a sample of $n = 60$ individuals of F2-ob/ob(B) mice with 22,575 different Affymetrix probe sets. The expression value for each prob set is microarray-derived gene expression measurements (mRNA abundance traits), and they are obtained

Table 1.4: Coefficients estimates of Boston housing price data using different methods. The standard errors of coefficient estimates are in parenthesis and they are based on 500 bootstrap samples. "0" indicates that the corresponding variable is not selected.

Variable	MTE	LAD	ESL	CQR	MLE
rm	0.379 (*0.108*)	0.323 (*0.134*)	0.308 (*0.209*)	0.448 (*0.146*)	0.200 (*0.063*)
ln(tax)	-0.131 (*0.070*)	0	0	-0.019 (*0.034*)	-0.134 (*0.044*)
ptratio	-0.161 (*0.031*)	-0.156 (*0.060*)	-0.130 (*0.071*)	-0.083 (*0.036*)	-0.201 (*0.026*)
ln(lstat)	-0.436 (*0.078*)	-0.436 (*0.125*)	-0.453 (*0.177*)	-0.453 (*0.119*)	-0.609 (*0.077*)
nox	0	0	0	0	-0.152 (*0.045*)
dis	-0.069 (*0.068*)	0	0	-0.025 (*0.038*)	-0.233 (*0.043*)
ln(crim)	0	0	0	0	0
zn	0	0	0	0	0
indus	0	0	0	0	0
age	0	0	0	0	0.037 (*0.052*)
black	0	0	0	0	0.078 (*0.029*)
chas	0	0	0	0	0.054 (*0.036*)
rad	0	0	0	0	0.140 (*0.060*)

using the Affymetrix MOE430B microarrays (Array B of GeneChip Mouse Expression Set 430). Lan et al. [34] developed and studied this sample to identify regulatory networks. We investigate the linear relationship of gene expressions and PEPCK, the numbers of phosphoenopyruvate carboxykinase (NM_011044) measured by quatitative real-time RT-PCR. Similar study has been done by [51]. First, we pre-screened all 22,575 probes variables by calculating the correlation coefficients with the response variable PEPCK. We use 1000 gene expression variables who have the highest marginal correlation to repsonse variable as covariates. We compare our method with LAD-Lasso, Huber-Lasso, and LASSO.

MTE selects four probe sets: "1438937_x_at", "1437871_at", "1439163_at", and "1439617_s_at". Among them, "1438937_x_at" is the common one that has been selected by all methods, and "1437871_at" has been selected by three methods. More importantly, the four selected probe sets by MTE are all covered by LASSO, which has selected five probe sets. The selection

results from LAD and Huber, however, are very different from MTE and LASSO. By exploratory analysis, we found that the response variable in this dataset is little contaminated. In this case, as we expected, MTE and LASSO should produce similar estimates.

We further evaluate the out-of-sample prediction performance of these methods. The dataset is randomly split to training set (54 observations) and testing set (6 observations). Table 1.5 reports the average mean squared prediction error (MSPE) and average model size, i.e. number of significant genes, over 100 random splits. From Table 1.5, we can see that the out-of-sample prediction performance of MTE is uniformly better than the other methods. We notice that the standard deviation of model size (number of selected variables) of MTE is also the smallest among all methods.

Table 1.5: Mean squared prediction errors (MSPE) and model sizes obtained from different methods using the eQTL dataset. The average MSPE and model size based on 100 random splits are reported. Numbers in the parenthesis are standard errors.

Methods	MSPE	Model Size
MTE	0.565 (*0.034*)	5.58 (*1.210*)
LAD	0.683 (*0.038*)	5.02 (*1.461*)
Huber	0.574 (*0.034*)	6.16 (*1.436*)
LASSO	0.712 (*0.039*)	5.80 (*3.296*)

1.7 Conclusion

We have proposed a new class of robust mean regression estimators that can produce robust and efficient estimates. Our proposed maximum tangent likelihood estimate (MTE) covers a number of existing estimators, such as MLE, minimum distance estimator, Mallows type estimator, and trimmed likelihood estimator as special cases. More interestingly, we

show that solving the proposed MTE is equivalent to minimizing a combination of Kullback-Leibler (KL) and ℓ_2 distance, where the weights depend on the choice of tuning parameter t. Our proposed penalized maximum tangent likelihood estimator performs well in robust estimation and variable selection under both fixed and high-dimensional regression. In addition to various numerical studies that demonstrate superior performance in practice, we have shown that the unpenalized MTE enjoys nice theoretical properties such as consistency and asymptotic normality, and the oracle property holds for the penalized MTE under fixed dimensional regression. Further, we show that under an ultra-high-dimensional regression setting when d can grow exponentially with n, for any positive t, the penalized MTE is consistent in the optimal order of $\sqrt{\ln(d)/n}$.

1.8 Technical Proofs

Proof of Theorem 1.2.1. For some $\boldsymbol{\beta}$, let $B_l \downarrow \boldsymbol{\beta}$ be a decreasing sequence of open balls around $\boldsymbol{\beta}$ of diameter l converging to 0. Define $m_B(\mathbf{z}) = \sup_{\boldsymbol{\beta}\in B} \ln_t(f(\mathbf{z};\boldsymbol{\beta}))$, and the sequence $m_{B_l}(\mathbf{z})$ is decreasing and greater than $\ln_t(f(\mathbf{z};\boldsymbol{\beta}))$ for every l. Since $\ln_t(f(\mathbf{z};\boldsymbol{\beta}))$ is upper-semicontinuous in $\boldsymbol{\beta}$ for almost all $\mathbf{z}$, we have

$$\limsup_{\boldsymbol{\beta}_n\to\boldsymbol{\beta}} \ln_t(f(\mathbf{z};\boldsymbol{\beta}_n)) < \ln_t(f(\mathbf{z};\boldsymbol{\beta})). \quad \text{a.s.}$$

Furthermore, we know $m_{B_l}(\mathbf{z}) \downarrow \ln_t(f(\mathbf{z};\boldsymbol{\beta}))$ almost surely. Since $\mathbb{E}_{\boldsymbol{\beta}_0}[\sup_{\boldsymbol{\beta}\in B} \ln_t(f(x;\boldsymbol{\beta}))] < \infty$, by the monotone convergence theorem, we have $\mathbb{E}_{\boldsymbol{\beta}_0}[m_{B_l}(\mathbf{z})] \downarrow \mathbb{E}_{\boldsymbol{\beta}_0}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}))]$. For any $\boldsymbol{\beta} \neq \boldsymbol{\beta}_t^*$, $\mathbb{E}_{\boldsymbol{\beta}_0}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}))] < \mathbb{E}_{\boldsymbol{\beta}_0}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))]$. Therefore, there exists an open ball $B_{\boldsymbol{\beta}}$ around $\boldsymbol{\beta}$ with $\mathbb{E}_{\boldsymbol{\beta}_0}[m_{B_{\boldsymbol{\beta}}}(\mathbf{z})] < \mathbb{E}_{\boldsymbol{\beta}_0}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))]$. The set $D = \{\boldsymbol{\beta} \in \boldsymbol{\mathcal{B}} : ||\boldsymbol{\beta} - \boldsymbol{\beta}_t^*|| \geq \delta\}$ is compact and is covered by the balls $\{B_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in D\}$. Let $B_{\boldsymbol{\beta}_1}$, ...,$B_{\boldsymbol{\beta}_q}$ be a finite subcover.

Then, by the law of large numbers,

$$\sup_{\boldsymbol{\beta}\in D}\frac{1}{n}\sum_{i=1}^{n}\ln_t(f(\mathbf{z}_i;\boldsymbol{\beta}))\leq\sup_{j=1,\ldots,q}\frac{1}{n}\sum_{i=1}^{n}m_{B_{\boldsymbol{\beta}_j}}(\mathbf{z}_i)\overset{a.s.}{\to}\sup_{j=1,\ldots q}\mathbb{E}_{\boldsymbol{\beta}_0}[m_{B_{\boldsymbol{\beta}_j}}(\mathbf{z})]<\mathbb{E}_{\boldsymbol{\beta}_0}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))].$$

If $\tilde{\boldsymbol{\beta}}\in D$, then $\sup_{\boldsymbol{\beta}\in D}\frac{1}{n}\sum_{i=1}^{n}\ln_t(f(\mathbf{z}_i;\boldsymbol{\beta}))\geq\frac{1}{n}\sum_{i=1}^{n}\ln_t(f(\mathbf{z}_i;\tilde{\boldsymbol{\beta}}))$. Since $\tilde{\boldsymbol{\beta}}$ is the maximizer of $\frac{1}{n}\sum_{i=1}^{n}\ln_t(f(\mathbf{z}_i;\boldsymbol{\beta}))$, we also have $\frac{1}{n}\sum_{i=1}^{n}\ln_t(f(\mathbf{z}_i;\tilde{\boldsymbol{\beta}}))\geq\frac{1}{n}\sum_{i=1}^{n}\ln_t(f(\mathbf{z}_i;\boldsymbol{\beta}_t^*))$. By the law of large numbers, $\frac{1}{n}\sum_{i=1}^{n}\ln_t(f(\mathbf{z}_i;\boldsymbol{\beta}_t^*))=\mathbb{E}_{\boldsymbol{\beta}_0}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))]-o_p(1)$. Therefore,

$$\{\tilde{\boldsymbol{\beta}}\in D\}\subseteq\Big\{\sup_{\boldsymbol{\beta}\in D}\frac{1}{n}\sum_{i=1}^{n}\ln_t(f(\mathbf{z}_i;\boldsymbol{\beta}))\geq\mathbb{E}_{\boldsymbol{\beta}_0}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))]-o_p(1)\Big\}.$$

Since $P(\sup_{\boldsymbol{\beta}\in D}\frac{1}{n}\sum_{i=1}^{n}\ln_t(f(\mathbf{z}_i;\boldsymbol{\beta}))\geq\mathbb{E}_{\boldsymbol{\beta}_0}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))]-o_p(1))\to 0$, we have $\tilde{\boldsymbol{\beta}}\xrightarrow{p}\boldsymbol{\beta}_t^*$. □

Proof of Theorem 1.2.2. First we define $\mathbb{G}_n[g]=n^{-1/2}\sum_{i=1}^{n}\big(g(\mathbf{z}_i)-\mathbb{E}[g(\mathbf{z})]\big)$. Since we have the Lipschitz property and the differentiability of the function $\boldsymbol{\beta}\mapsto\ln_t(f(\mathbf{z};\boldsymbol{\beta}))$, by Lemma 19.31 in van der Vaart [54], for every random sequence h_n, a $d\times 1$ vector, that is bounded in probability, we have

$$\mathbb{G}_n\Big[\sqrt{n}\big(\ln_t(f(\mathbf{z}_i;\boldsymbol{\beta}_t^*+h_n/\sqrt{n}))-\ln_t(f(\mathbf{z}_i;\boldsymbol{\beta}_t^*+h_n/\sqrt{n}))\big)-h_n^T\frac{\partial}{\partial\boldsymbol{\beta}}\ln_t(f(\mathbf{z}_i;\boldsymbol{\beta}_t^*))\Big]\xrightarrow{p}0.$$

In addition, by Corollary 5.53 in van der Vaart [54], the Lipschitz condition and the twice differentiability of the function $\boldsymbol{\beta}\mapsto\ln_t(f(\mathbf{z};\boldsymbol{\beta}))$ also imply that the sequence $\sqrt{n}(\tilde{\boldsymbol{\beta}}-\boldsymbol{\beta}_t^*)$ is bounded in probability. By the twice differentiability, we also have

$$\sum_{i=1}^{n}\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*+h_n/\sqrt{n}))-\sum_{i=1}^{n}\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))=\frac{1}{2}h_n^T\frac{\partial^2}{\partial\boldsymbol{\beta}^2}\Big[\mathbb{E}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))]\Big]h_n^T+$$
$$h_n^T\mathbb{G}_n\left[\frac{\partial\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))}{\partial\boldsymbol{\beta}}\right]+o_p(1).$$

Since the sequence $\tilde{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent, it is valid for h_n if we let

$$\tilde{h}_n = \sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_t^*),$$

or let

$$\check{h}_n = -\left[\frac{\partial^2}{\partial \boldsymbol{\beta}^2}\Big[\mathbb{E}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))]\Big]\right]^{-1}\mathbb{G}_n\left[\frac{\partial \ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))}{\partial \boldsymbol{\beta}}\right].$$

Therefore, we have

$$\begin{aligned}\sum_{i=1}^{n}\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^* + \tilde{h}_n/\sqrt{n})) - \sum_{i=1}^{n}\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*)) =& \frac{1}{2}\tilde{h}_n^T\frac{\partial^2}{\partial \boldsymbol{\beta}^2}\Big[\mathbb{E}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))]\Big]\tilde{h}_n^T + \\ & \tilde{h}_n^T\mathbb{G}_n\left[\frac{\partial \ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))}{\partial \boldsymbol{\beta}}\right] + o_p(1) \qquad (1.17)\end{aligned}$$

and

$$\begin{aligned}&\sum_{i=1}^{n}\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^* + \check{h}_n/\sqrt{n})) - \sum_{i=1}^{n}\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*)) \\ =&\frac{1}{2}\check{h}_n^T\frac{\partial^2}{\partial \boldsymbol{\beta}^2}\Big[\mathbb{E}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))]\Big]\check{h}_n + \check{h}_n^T\mathbb{G}_n\left[\frac{\partial \ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))}{\partial \boldsymbol{\beta}}\right] + o_p(1) \\ =&-\frac{1}{2}\mathbb{G}_n\left[\frac{\partial \ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))}{\partial \boldsymbol{\beta}}\right]^T\left[\frac{\partial^2}{\partial \boldsymbol{\beta}^2}\Big[\mathbb{E}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))]\Big]\right]^{-1}\mathbb{G}_n\left[\frac{\partial \ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))}{\partial \boldsymbol{\beta}}\right] + o_p(1). \quad (1.18)\end{aligned}$$

Because $\tilde{\boldsymbol{\beta}}$ is the maximizer of $\sum_{i=1}^{n}\ln_t(f(\mathbf{z};\boldsymbol{\beta}))$, the left side of equation (1.17) is larger than the left side of equation (1.18) up to $o_p(1)$, hence the same relation is true for the right sides. Taking the difference of these two, complete the square, we have

$$\frac{1}{2}\left\{\tilde{h}_n + \left[\frac{\partial^2}{\partial \boldsymbol{\beta}^2}\Big[\mathbb{E}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))]\Big]\right]^{-1}\mathbb{G}_n\left[\frac{\partial \ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))}{\partial \boldsymbol{\beta}}\right]\right\}^T\frac{\partial^2}{\partial \boldsymbol{\beta}^2}\Big[\mathbb{E}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))]\Big]\left\{\tilde{h}_n +\right.$$

$$\left[\frac{\partial^2}{\partial\boldsymbol{\beta}^2}\left[\mathbb{E}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))]\right]\right]^{-1}\mathbb{G}_n\left[\frac{\partial\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))}{\partial\boldsymbol{\beta}}\right]\Bigg\}+o_p(1)\geq 0.$$

Note that $\partial^2\mathbb{E}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))]/\partial\boldsymbol{\beta}^2$ is negative definite, therefore, the quadratic form has to converge to 0 in probability, which implies

$$\left\|\tilde{h}_n+\left[\frac{\partial^2}{\partial\boldsymbol{\beta}^2}\left[\mathbb{E}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))]\right]\right]^{-1}\mathbb{G}_n\left[\frac{\partial\ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))}{\partial\boldsymbol{\beta}}\right]\right\|\xrightarrow{p}0.$$

The normality result follows by applying Slutsky's Lemma. □

Proof of Theorem 1.2.3. To prove the consistency of MTE for $\boldsymbol{\beta}_0$ in linear regressions, we only need to show $\boldsymbol{\beta}_0=\arg\max_{\boldsymbol{\beta}\in\mathcal{B}}\mathbb{E}_{\boldsymbol{\beta}_0}\ln_t(f(\mathbf{z};\boldsymbol{\beta}))$. By the first order condition, we have

$$\begin{aligned}
\frac{\partial}{\partial\boldsymbol{\beta}}\mathbb{E}_{\boldsymbol{\beta}_0}\ln_t(f(\mathbf{z};\boldsymbol{\beta}))&=\int\int\ln_t(f(\mathbf{z};\boldsymbol{\beta}))dF(\mathbf{x},\epsilon)\\
&=\int\int h(f(y-\mathbf{x}\boldsymbol{\beta}))f'(y-\mathbf{x}\boldsymbol{\beta})(-\mathbf{x})dF(\mathbf{x},\epsilon)\\
&=\int\int h(f(\mathbf{x}^T(\boldsymbol{\beta}_0-\boldsymbol{\beta})+\epsilon))f'(\mathbf{x}^T(\boldsymbol{\beta}_0-\boldsymbol{\beta})+\epsilon)(-\mathbf{x})dF(\mathbf{x},\epsilon)\\
&=\int\int h(f(\mathbf{x}^T(\boldsymbol{\beta}_0-\boldsymbol{\beta})+\epsilon))f'(\mathbf{x}^T(\boldsymbol{\beta}_0-\boldsymbol{\beta})+\epsilon)(-\mathbf{x})dF(\mathbf{x})dF(\epsilon)\\
&=\int\left(\int h(f(\mathbf{x}^T(\boldsymbol{\beta}_0-\boldsymbol{\beta})+\epsilon))f'(\mathbf{x}^T(\boldsymbol{\beta}_0-\boldsymbol{\beta})+\epsilon)f(\epsilon)d\epsilon\right)(-\mathbf{x})dF(\mathbf{x}),
\end{aligned}$$

where we have used the independence between $\mathbf{x}$ and ϵ. Note that $\int h(f(\epsilon))f'(\epsilon)f(\epsilon)d\epsilon=0$ which is due to the fact that f is an even function (i.e., a symmetric distribution) and f' is an odd function, therefore, the integral of an odd function equals to zero. In order to have $\frac{\partial}{\partial\boldsymbol{\beta}}\mathbb{E}_{\boldsymbol{\beta}_0}\ln_t(f(\mathbf{z};\boldsymbol{\beta}))=0$, we need $\mathbf{x}^T(\boldsymbol{\beta}_0-\boldsymbol{\beta})=0\ \forall\mathbf{x}$, by the regularity condition R8, it implies $\boldsymbol{\beta}_0=\boldsymbol{\beta}$. Hence, $\boldsymbol{\beta}_0=\arg\max_{\boldsymbol{\beta}\in\mathcal{B}}\mathbb{E}_{\boldsymbol{\beta}_0}\ln_t(f(\mathbf{z};\boldsymbol{\beta}))$. By Theorems 1.2.1 and 1.2.2, we obtain the consistency and asymptotic normality. □

Proof of Theorem 1.3.1. Given an appropriate choice of t, define

$$Q_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln_t(f(\mathbf{z}_i;\boldsymbol{\beta})) - n\sum_{j=1}^{d} p_{\lambda_{nj}}(|\beta_j|); \tag{1.19}$$

$$\mathcal{L}_n(\boldsymbol{\beta}) = \sum_{i=1}^{n} \ln_t(f(\mathbf{z}_i;\boldsymbol{\beta})). \tag{1.20}$$

Let $\alpha_n = n^{-1/2} + a_n$, where $a_n = \max\{p'_{\lambda_{nj}}(|\beta_{0j}|) : \beta_{0j} \neq 0\}$. To show the results, we need to show that for any given $\epsilon > 0$, there exists a large constant C such that

$$P\left\{\sup_{\|\mathbf{u}\|=C} Q_n(\boldsymbol{\beta}_0 + \alpha_n\mathbf{u}) < Q_n(\boldsymbol{\beta}_0)\right\} \geq 1-\epsilon, \tag{1.21}$$

where $\mathbf{u}$ is a d-dimensional vector such that $\|\mathbf{u}\| = C$. This implies with probability at least $1-\epsilon$ that there exists a local maximizer such that $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\| = O_p(\alpha_n)$.

Note that by SLLN, we have $\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \ln_t(f(\mathbf{z}_i,\boldsymbol{\beta})) = \mathbf{J}(\boldsymbol{\beta})\{1+o_p(1)\}$. By Taylor expansion, given an appropriate t fixed, we have

$$\begin{aligned}
D_n(\mathbf{u}) &\equiv Q_n(\boldsymbol{\beta}_0 + \alpha_n\mathbf{u}) - Q_n(\boldsymbol{\beta}_0) \\
&= \mathcal{L}_n(\boldsymbol{\beta}_0 + \alpha_n\mathbf{u}) - \mathcal{L}_n(\boldsymbol{\beta}_0) - n\sum_{j=1}^{s}\left\{p_{\lambda_{nj}}(|\beta_{0j} + \alpha_n u_j|) - p_{\lambda_{nj}}(|\beta_{0j}|)\right\} \\
&\leq \alpha_n\mathcal{L}'_n(\boldsymbol{\beta}_0)^T\mathbf{u} - \frac{1}{2}n\alpha_n^2\mathbf{u}^T[-\mathbf{J}(\boldsymbol{\beta}_0)]\mathbf{u}\{1+o_p(1)\} \\
&\quad - \sum_{j=1}^{s}\left[n\alpha_n p'_{\lambda_{nj}}(|\beta_{0j}|)\mathrm{sgn}(\beta_{0j})u_j + \frac{1}{2}n\alpha_n^2 p''_{\lambda_{nj}}(|\beta_{0j}|)u_j^2\{1+o(1)\}\right] \\
&= \alpha_n[\mathcal{L}'_n(\boldsymbol{\beta}_0) + o_p(\sqrt{n})]^T\mathbf{u} - \frac{1}{2}n\alpha_n^2\mathbf{u}^T[-\mathbf{J}(\boldsymbol{\beta}_0) + o_p(1)]\mathbf{u}\{1+o_p(1)\} \\
&\quad - \sum_{j=1}^{s}\left[n\alpha_n p'_{\lambda_{nj}}(|\beta_{0j}|)\mathrm{sgn}(\beta_{0j})u_j + \frac{1}{2}n\alpha_n^2 p''_{\lambda_{nj}}(|\beta_{0j}|)u_j^2\{1+o(1)\}\right],
\end{aligned} \tag{1.22}$$

where $\mathbf{J}(\boldsymbol{\beta}) = E_{\boldsymbol{\beta}_0}\left[\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\ln_t(f(\mathbf{z};\boldsymbol{\beta}))\right]$ is a negative definite matrix. By Theorem 1.2.3 and CLT, $n^{-1/2}\mathcal{L}_n'(\boldsymbol{\beta}_0) = O_p(1)$. Thus, the first term on the right-hand side of (1.22) is of the order $O_p(n^{1/2}\alpha_n)$. Furthermore, the condition $a_n = O_p(n^{-1/2})$ implies that $O_p(n^{1/2}\alpha_n) = O_p(n\alpha_n^2)$. By choosing a sufficiently large C, the second term dominates the first term uniformly in $\|\mathbf{u}\| = C$. In addition, the third term is bounded by $\sqrt{s}n\alpha_n a_n\|\mathbf{u}\| + \frac{1}{2}n\alpha_n^2 b_n\|\mathbf{u}\|^2$, where $b_n = \max\{p''_{\lambda_{nj}}(|\beta_{0j}|) : \beta_{0j} \neq 0\}$. Since $b_n = o_p(1)$, the third term is also dominated by the second term of (1.22). Hence, by choosing a sufficiently large C, (1.21) holds. This completes the proof of Theorem 1.3.1. □

To prove Theorem 1.3.2, we first prove the following lemma that presents the sparsity of the penalized tangent likelihood estimation.

Lemma 3. *Under the conditions in Theorem 1.3.1, for any given $\boldsymbol{\beta}$ satisfying $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2})$ and any constant C, we have*

$$Q_n((\boldsymbol{\beta}_S, \mathbf{0})) = \max_{\|\boldsymbol{\beta}_{S^c}\| \leq Cn^{-1/2}} Q_n((\boldsymbol{\beta}_S, \boldsymbol{\beta}_{S^c}))$$

with probability 1, where $Q_n(\boldsymbol{\beta})$ is defined in (1.19).

Proof of Lemma 3. First we show that with probability 1, for any $\boldsymbol{\beta}_S$ such that $\boldsymbol{\beta}_S = \boldsymbol{\beta}_{0S} + O_p(n^{-1/2})$, and for some small $\epsilon_n = Cn^{-1/2}$, and $j = s+1, \ldots, d$,

$$\begin{aligned} \frac{\partial Q_n(\boldsymbol{\beta})}{\partial\beta_j} &< 0 \qquad \text{for} \quad 0 < \beta_j < \epsilon_n \\ \frac{\partial Q_n(\boldsymbol{\beta})}{\partial\beta_j} &> 0 \qquad \text{for} \quad -\epsilon_n < \beta_j < 0, \end{aligned}$$

where $Q_n(\boldsymbol{\beta})$ is defined in (1.19). By Taylor expansion, we have

$$\begin{aligned}\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \beta_j} &= \frac{\partial \mathcal{L}_n(\boldsymbol{\beta})}{\partial \beta_j} - np'_{\lambda_{nj}}(|\beta_j|)\text{sgn}(\beta_j) \\ &= \frac{\partial \mathcal{L}_n(\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{d} \frac{\partial^2 \mathcal{L}_n(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l}(\beta_l - \beta_{0l}) \\ &\quad + \sum_{l=1}^{d}\sum_{k=1}^{d} \frac{\partial^3 \mathcal{L}_n(\boldsymbol{\beta}^*)}{\partial \beta_j \partial \beta_l \partial \beta_k}(\beta_l - \beta_{0l})(\beta_k - \beta_{0k}) - np'_{\lambda_{nj}}(|\beta_j|)\text{sgn}(\beta_j), \end{aligned} \tag{1.23}$$

where $\boldsymbol{\beta}^*$ is between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. Note that

$$\frac{1}{n}\frac{\partial \mathcal{L}_n(\boldsymbol{\beta}_0)}{\partial \beta_j} = O_p(n^{-1/2}),$$

and

$$\frac{1}{n}\frac{\partial^2 \mathcal{L}_n(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l} = \mathbb{E}_{\boldsymbol{\beta}_0}\left[\frac{\partial^2}{\partial \beta_j \partial \beta_l} \ln_t f(\mathbf{z}, \boldsymbol{\beta}_0)\right] + o_p(1).$$

Since $\boldsymbol{\beta} - \boldsymbol{\beta}_0 = O_p(n^{-1/2})$ by Theorem 1.3.1, the first three term of (1.23) is $O_p(n^{1/2})$. We can write (1.23) as

$$\begin{aligned}\frac{\partial Q_n(\boldsymbol{\beta})}{\partial \beta_j} &= -np'_{\lambda_{nj}}(|\beta_j|)\text{sgn}(\beta_j) + O_p(n^{1/2}) \\ &= n\lambda_{nj}\left\{-\text{sgn}(\beta_j)p'_{\lambda_{nj}}(|\beta_j|)/\lambda_{nj} + O_p(n^{-1/2}/\lambda_{nj})\right\}.\end{aligned}$$

By conditions (1.11) and (1.12), that is

$$\liminf_{n\to\infty}\liminf_{\theta\to 0+}\left\{\min_{s+1\le j\le d} p'_{\lambda_{nj}}(\theta)/\lambda_{nj}\right\} > 0 \quad \text{and} \quad 1/\min_{s+1\le j\le d}(\sqrt{n}\lambda_{nj}) = o_p(1),$$

the sign of the derivative $\partial Q_n(\boldsymbol{\beta})/\partial \beta_j$ is completely determined by the sign of β_j. This completes the proof of Lemma 3. □

Proof of Theorem 1.3.2. Part (a) holds due to Lemma 3. For part (b), we have shown that there exists a $\hat{\boldsymbol{\beta}}_S$ that is $\sqrt{n}$-consistent local maximizer of $Q_n((\boldsymbol{\beta}_S, \mathbf{0}))$ such that

$$\frac{\partial Q_n((\hat{\boldsymbol{\beta}}_S, \mathbf{0}))}{\partial \beta_j} = 0 \qquad \text{for } j = 1, \ldots, s.$$

Because $\hat{\boldsymbol{\beta}}_S$ is a consistent estimator, we have

$$\begin{aligned} 0 &\equiv \frac{\partial \mathcal{L}_n((\hat{\boldsymbol{\beta}}_S, \mathbf{0}))}{\partial \beta_j} - np'_{\lambda_{nj}}(|\hat{\beta}_j|)\mathrm{sgn}(\hat{\beta}_j) \\ &= \frac{\partial \mathcal{L}_n(\boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{s} \left\{ \frac{\partial^2 \mathcal{L}_n(\boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l} + o_p(1) \right\} (\hat{\beta}_l - \beta_{0l}) \\ &\quad - n \left[p'_{\lambda_{nj}}(|\beta_{0j}|)\mathrm{sgn}(\beta_{0j}) + \left\{ p''_{\lambda_{nj}}(|\beta_{0j}|) + o_p(1) \right\} (\hat{\beta}_j - \beta_{0j}) \right]. \end{aligned}$$

It follows by Slutsky's lemma and the central limit theorem that

$$\sqrt{n}(\mathbf{J}_S(\boldsymbol{\beta}_0) + \boldsymbol{\Sigma}_1) \left\{ \hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S} + (\mathbf{J}_S(\boldsymbol{\beta}_0) + \boldsymbol{\Sigma}_1)^{-1}\mathbf{b} \right\} \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}_2).$$

Note that for the results above, it is supposed that $\ln_t(f(\mathbf{z}, \boldsymbol{\beta}))$ admits a third-order Taylor expansion (i.e. $p \geq 3$) to allow continuous differentiability for ease of presentation. □

Lemma 4. *Let* $\mathbf{Z}_n = \{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ *be any sample of size* n, $\mathbf{Z}_m = \{\mathbf{z}_1, \ldots, \mathbf{z}_m\}$ *be a contamination sample of size* m, *and* $a_{nm} = \frac{1}{n-m} \max_{\boldsymbol{\beta} \in \mathbb{R}^d} \# \left\{ i : m+1 \leq i \leq n, \mathbf{x}_i^T \boldsymbol{\beta} = 0 \right\}$. *Assume* $a_{nm} < 0.5$, $\varepsilon < (1 - 2a_{nm})/(2 - 2a_{nm})$ *and* $\zeta(t) < (1 - \varepsilon)(2 - 2a_{nm})$. *For the weighted vector* $\lambda = (\lambda_{n1}, \ldots, \lambda_{nd})$, *if*

$$0 < \min_{\{j : 1 \leq j \leq d\}} \lambda_{nj} < +\infty,$$

there exists a b such that $m/n \leq \varepsilon$ implies

$$\sup_{\|\boldsymbol{\beta}\| \geq b} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ln_t(f(y_i - \mathbf{x}_i^T \boldsymbol{\beta})) - \sum_{j=1}^{d} \lambda_{nj} g(|\beta_j|) \right\} < \frac{1}{n} \sum_{i=1}^{n} \ln_t(f(y_i - \mathbf{x}_i^T \tilde{\tilde{\boldsymbol{\beta}}})) - \sum_{j=1}^{d} \lambda_{nj} g(|\tilde{\tilde{\beta}}_j|),$$

where $\tilde{\tilde{\boldsymbol{\beta}}}$ is an initial estimator of $\boldsymbol{\beta}$.

Proof of Lemma 4. First, we define

$$\phi(r) = \frac{\ln_t(f(0)) - \ln_t(f(r))}{\ln_t(f(0)) - \ln_t(0)}$$
$$\zeta(t) = \frac{2m}{n} + \frac{2}{n} \sum_{i=m+1}^{n} \phi(y_i - \mathbf{x}_i^T \tilde{\tilde{\boldsymbol{\beta}}}).$$

Since $a_{nm} = \frac{1}{n-m} \max_{\boldsymbol{\beta} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \# \left\{ i : m+1 \leq i \leq n, \mathbf{x}_i^T \boldsymbol{\beta} = 0 \right\}$, for all $\boldsymbol{\beta}$, we have,

$$\begin{aligned} 1 - a_{nm} &= \frac{n - m - \max_{\boldsymbol{\beta} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \# \left\{ i : m+1 \leq i \leq n, \mathbf{x}_i^T \boldsymbol{\beta} = 0 \right\}}{n-m} \\ &= \frac{1}{n-m} \inf_{\boldsymbol{\beta} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \# \left\{ i : m+1 \leq i \leq n, \mathbf{x}_i^T \boldsymbol{\beta} \neq 0 \right\}. \end{aligned}$$

Because $\varepsilon < (1 - 2a_{nm})/(2 - 2a_{nm})$ and $\zeta(t) < (1-\varepsilon)(2 - 2a_{nm})$ in the assumptions, we can find a $a_n^* > a_{nm}$ such that $\varepsilon < (1 - 2a_n^*)(2 - 2a_n^*)$ and $\zeta(t) < (1-\varepsilon)(2-2a_n^*)$. Using the compacity argument [64], there exists a $\delta > 0$ such that

$$1 - a_n^* \leq \frac{1}{n-m} \inf_{\|\boldsymbol{\beta}\| = 1} \# \left\{ i : m+1 \leq i \leq n, |\mathbf{x}_i^T \boldsymbol{\beta}| > \delta \right\}.$$

Since $\varepsilon < (1 - 2a_n^*)/(2 - 2a_n^*)$, we have $(1-\varepsilon)(1 - a_n^*) > 1/2$. Therefore, there exists a $\eta \in (1 - (1-\varepsilon)(1-a_n^*), 1/2)$ such that $(1-\varepsilon)(1-a_n^*)/(1-\eta) > 1$. In addition, since

$\zeta(t) < (1-\varepsilon)(2-2a_n^*)$, we have $(1-\varepsilon)(2-2a_n^*)/\zeta(t) > 1$. Therefore, we define

$$a_0 = \frac{(1-\eta)(1+\delta^*)\zeta(t)}{(1-\varepsilon)(2-2a_n^*)},$$

where $\delta^* > 0$ and

$$\delta^* < \min\left(\frac{(1-\varepsilon)(1-a_n^*)}{1-\eta}, \frac{(1-\varepsilon)(2-2a_n^*)}{\zeta(t)}\right) - 1$$

We can show that $a_0 < 1-\eta$ and $a_0 < \zeta(t)/2$. Then $m/n \leq \varepsilon$ implies $(n-m) \geq n(1-\varepsilon)$. By the definition of $\phi_t(r)$ (i.e., continuous, bounded and even function), there exists a $R \geq 0$ such that $\phi_t(R) = a_0/(1-\eta) < 1$. Let $b_1 \geq (R + \max_{m+1\leq i\leq n}|y_i|)/\delta$, given that $m/n \leq \varepsilon$, we have

$$\begin{aligned}
\inf_{\|\boldsymbol{\beta}\|\geq b_1}\left\{\sum_{i=1}^n \phi_t(y_i - \mathbf{x}_i^T\boldsymbol{\beta})\right\} &\geq \inf_{\|\boldsymbol{\beta}\|=1}\left\{\sum_{i\in A}\phi_t(|y_i| - b_1|\mathbf{x}_i^T\boldsymbol{\beta}|)\right\} \geq (n-m)(1-a_n^*)\phi_t(R) \\
&= (n-m)(1-a_n^*)\frac{a_0}{1-\eta} \geq n(1-\varepsilon)(1-a_n^*)\frac{a_0}{1-\eta} \\
&= 0.5n(1+\delta^*)\zeta(t) > 0.5n\zeta(t) \geq \sum_{i=1}^n \phi_t(y_i - \mathbf{x}_i^T\tilde{\tilde{\boldsymbol{\beta}}}),
\end{aligned}$$

where $A = \{i : m+1 \leq i \leq n, |\mathbf{x}_i\boldsymbol{\beta}| > \delta\}$. Therefore,

$$\sup_{\|\boldsymbol{\beta}\|\geq b_1}\left\{\sum_{i=1}^n \ln_t(f(y_i - \mathbf{x}_i^T\boldsymbol{\beta}))\right\} \leq \sum_{i=1}^n \ln_t(f(y_i - \mathbf{x}_i^T\tilde{\tilde{\boldsymbol{\beta}}})). \tag{1.24}$$

We further define $b_2 = \sqrt{d}g^{-1}\left\{\sum_{j=1}^d \lambda_{nj}g(|\tilde{\tilde{\beta}}_j|)/(\min\{\lambda_{nj}\})\right\}$ such that $\lambda_{nj'}g(b_2/\sqrt{d}) \geq \sum_{j=1}^d \lambda_{nj}g(|\tilde{\tilde{\beta}}_j|)$ for $j' = 1, ..., d$. By defining $b = \max\{b_1, b_2\}$. We have

$$\sup_{\|\boldsymbol{\beta}\|\geq b}\left\{\frac{1}{n}\sum_{i=1}^n \ln_t(f(y_i - \mathbf{x}_i^T\boldsymbol{\beta})) - \sum_{j=1}^d \lambda_{nj}g(|\beta_j|)\right\}$$

$$
\begin{aligned}
&\leq \sup_{\|\boldsymbol{\beta}\|\geq b}\left\{\frac{1}{n}\sum_{i=1}^{n}\ln_t(f(y_i-\mathbf{x}_i^T\boldsymbol{\beta}))\right\}+\sup_{\|\boldsymbol{\beta}\|\geq b}\left\{-\sum_{j=1}^{d}\lambda_{nj}g(|\beta_j|)\right\}\\
&\leq \sup_{\|\boldsymbol{\beta}\|\geq b_1}\left\{\frac{1}{n}\sum_{i=1}^{n}\ln_t(f(y_i-\mathbf{x}_i^T\boldsymbol{\beta}))\right\}+\sup_{\|\boldsymbol{\beta}\|\geq b_2}\left\{-\sum_{j=1}^{d}\lambda_{nj}g(|\beta_j|)\right\}\\
&\leq \sum_{i=1}^{n}\ln_t(f(y_i-\mathbf{x}_i^T\tilde{\tilde{\boldsymbol{\beta}}}))+\sup_{\|\boldsymbol{\beta}\|\geq b_2}\left\{-\sum_{j=1}^{d}\lambda_{nj}g(|\beta_j|)\right\}. \qquad (1.25)
\end{aligned}
$$

where the last step uses (1.24). For the second term of (1.25), note that $\|\beta\| \geq b_2$ implies that there exists at least one element $\beta_{j'}$ (e.g., maximum) of $\boldsymbol{\beta}$ such that $|\beta_{j'}| \geq b_2/\sqrt{d}$ for some j'. Hence,

$$
\sup_{\|\boldsymbol{\beta}\|\geq b_2}\left\{-\sum_{j=1}^{d}\lambda_{nj}g(|\beta_j|)\right\} \leq -\lambda_{nj'}g(|\beta_j|) \leq -\lambda_{nj'}g(b_2/\sqrt{d}) \leq -\sum_{j=1}^{d}\lambda_{nj}g(|\tilde{\tilde{\beta}}_j|),
$$

where we have used the fact that g is strictly increasing. Therefore, substituting the result into (1.25), we have

$$
\begin{aligned}
&\sup_{\|\boldsymbol{\beta}\|\geq b}\left\{\frac{1}{n}\sum_{i=1}^{n}\ln_t(f(y_i-\mathbf{x}_i^T\boldsymbol{\beta}))-\sum_{j=1}^{d}\lambda_{nj}g(|\beta_j|)\right\}\\
&\qquad\leq \sum_{i=1}^{n}\ln_t(f(y_i-\mathbf{x}_i^T\tilde{\tilde{\boldsymbol{\beta}}}))-\sum_{j=1}^{d}\lambda_{nj}g(|\tilde{\tilde{\beta}}_j|).
\end{aligned}
$$

□

Proof of Theorem 1.4.1. Since $\tilde{\tilde{\boldsymbol{\beta}}}$ is the initial estimate and the penalized MTE $\hat{\boldsymbol{\beta}}$ maximizes the objective function, we have

$$
\frac{1}{n}\sum_{i=1}^{n}\ln_t\{f(y_i-\mathbf{x}_i^T\hat{\boldsymbol{\beta}})\}-\sum_{j=1}^{d}\lambda_{nj}g(|\hat{\beta}_{nj}|) \geq \frac{1}{n}\sum_{i=1}^{n}\ln_t\{f(y_i-\mathbf{x}_i^T\tilde{\tilde{\boldsymbol{\beta}}})\}-\sum_{j=1}^{d}\lambda_{nj}g(|\tilde{\tilde{\beta}}_{nj}|).
$$

However, for a contaminated sample $\mathbf{Z}_n$ with $m/n \leq \varepsilon$, by Lemma 4, we know that if $\|\hat{\boldsymbol{\beta}}\| \geq b$, we have

$$\frac{1}{n}\sum_{i=1}^{n} \ln_t\{f(y_i - \mathbf{x}_i^T\hat{\boldsymbol{\beta}})\} - \sum_{j=1}^{d} \lambda_{nj} g(|\hat{\beta}_{nj}|) < \frac{1}{n}\sum_{i=1}^{n} \ln_t\{f(y_i - \mathbf{x}_i^T\tilde{\tilde{\boldsymbol{\beta}}})\} - \sum_{j=1}^{d} \lambda_{nj} g(|\tilde{\tilde{\beta}}_{nj}|),$$

which is a contradiction. Therefore, we conclude

$$\mathrm{BP}(\hat{\boldsymbol{\beta}}; \mathbf{Z}_{n-m}, t) \geq \min\left\{\mathrm{BP}(\tilde{\tilde{\boldsymbol{\beta}}}; \mathbf{Z}_{n-m}), \frac{1-2a_{nm}}{2-2a_{nm}}, 1 - \frac{\zeta(t)}{2-2a_{nm}}\right\}.$$

□

Proof of Theorem 1.4.2. Under the contaminated distribution F_ε, the penalized MTE functional, $\boldsymbol{\beta}_\varepsilon$, is the solution of

$$0 = \frac{\partial}{\partial \boldsymbol{\beta}}\left\{\int \ln_t(f(y - \mathbf{x}^T\boldsymbol{\beta}))dF_\varepsilon - \sum_{j=1}^{d} p_{\lambda_{0j}}(|\beta_j|)\right\}.$$

In other words,

$$\begin{aligned}
0 &= (1-\varepsilon)\int \frac{\partial}{\partial \boldsymbol{\beta}} \ln_t(f(y - \mathbf{x}^T\boldsymbol{\beta}_\varepsilon))dF + \varepsilon \frac{\partial}{\partial \boldsymbol{\beta}} \ln_t(f(y^* - \mathbf{x}^{*T}\boldsymbol{\beta}_\varepsilon)) - v \\
0 &= (1-\varepsilon)\int \frac{\partial}{\partial r} \ln_t(f(r))\Big|_{r=y-\mathbf{x}^T\boldsymbol{\beta}_\varepsilon}(-\mathbf{x})dF + \varepsilon \frac{\partial}{\partial r} \ln_t(f(r))\Big|_{r=y^*-\mathbf{x}^{*T}\boldsymbol{\beta}_\varepsilon}(-\mathbf{x}^*) - v,
\end{aligned}$$

where $v(\boldsymbol{\beta}_\varepsilon) = (p'_{\lambda_{01}}(|\beta_{\varepsilon,1}|)\mathrm{sgn}(\beta_{\varepsilon,1}), ..., p'_{\lambda_{0d}}(|\beta_{\varepsilon,d}|)\mathrm{sgn}(\beta_{\varepsilon,d}))^T$. We further take derivative with respect to ε and let $\varepsilon \to 0$, we have

$$\begin{aligned}
0 = &-\int \frac{\partial}{\partial r} \ln_t(f(r))\Big|_{r=y-\mathbf{x}^T\boldsymbol{\beta}_0}(-\mathbf{x})dF + \int \frac{\partial^2}{\partial r^2} \ln_t(f(r))\Big|_{r=y-\mathbf{x}^T\boldsymbol{\beta}_0}(-\mathbf{x})(-\mathbf{x}^T)\frac{\partial}{\partial \varepsilon}\boldsymbol{\beta}_\varepsilon dF \\
&+ \frac{\partial}{\partial r} \ln_t(f(r))\Big|_{r=y^*-\mathbf{x}^{*T}\boldsymbol{\beta}_0}(-\mathbf{x}^*) - \frac{\partial}{\partial \varepsilon} v(\boldsymbol{\beta}_\varepsilon)\Big|_{\varepsilon=0}.
\end{aligned}$$

Note that when $\varepsilon \to 0$ we have $\boldsymbol{\beta}_\varepsilon \to \boldsymbol{\beta}_0$ because of the consistency. Furthermore, by the definition of $\boldsymbol{\beta}_\varepsilon$, we have $\int \frac{\partial}{\partial r} \ln_t(f(r))\Big|_{r=y-\mathbf{x}^T\boldsymbol{\beta}_0}(-\mathbf{x})dF - v(\boldsymbol{\beta}_0) = 0$.

$$-\int \frac{\partial^2}{\partial r^2} \ln_t(f(r))(-\mathbf{x})(-\mathbf{x}^T)\frac{\partial}{\partial \varepsilon}\boldsymbol{\beta}_\varepsilon dF + \frac{\partial}{\partial \varepsilon} v(\boldsymbol{\beta}_\varepsilon)\Big|_{\varepsilon=0} = \frac{\partial}{\partial r} \ln_t(f(r))\Big|_{r=y^*-\mathbf{x}^{*T}\boldsymbol{\beta}_0}(-\mathbf{x}^*) - v(\boldsymbol{\beta}_0)$$

Therefore, the influence function satisfies

$$\mathrm{IF}(\hat{\boldsymbol{\beta}})[-\int \frac{\partial^2}{\partial r^2} \ln_t(f(r))\mathbf{x}\mathbf{x}^T dF + v^*] = \frac{\partial}{\partial r} \ln_t(f(r^*))(-\mathbf{x}^*) - v(\boldsymbol{\beta}_0),$$

where $v^* = \mathrm{diag}(p''_{\lambda_{01}}(|\beta_{01}|) + p'_{\lambda_{01}}(|\beta_{01}|)\delta(\beta_{01}), ..., p''_{\lambda_{0d}}(|\beta_{0d}|) + p'_{\lambda_{0d}}(|\beta_{0d}|)\delta(\beta_{0d}))$ and

$$\delta(u) = \begin{cases} +\infty & \text{if } u = 0, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, we have

$$\mathrm{IF}(\hat{\boldsymbol{\beta}}) = \begin{cases} 0 & \text{if } \beta_{0j} = 0, \\ [[-\int \frac{\partial^2}{\partial r^2} \ln_t(f(r))\mathbf{x}\mathbf{x}^T dF + v^*]^{-1}]_j(\frac{\partial}{\partial r} \ln_t(f(r^*))(-\mathbf{x}^*) - v(\boldsymbol{\beta}_0)) & \text{if } \beta_{0j} \neq 0, \end{cases}$$

where $[\mathbf{A}]_j$ denotes the j-th row of the matrix $\mathbf{A}$. □

Proof of Lemma 1. Let $\epsilon_i = y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0$ be i.i.d. random error. The assumed density function is f. Suppose that the scale parameter is preliminarily estimated as σ_R^2 by a robust estimate. Define

$$d^j(\epsilon_1, ..., \epsilon_n) = \frac{\partial}{\partial \beta_j}\left[\frac{1}{n}\sum_{i=1}^n \ln_t(f(\mathbf{z}_i; \boldsymbol{\beta}))\right]\Big|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$$

$$
\begin{aligned}
&= \frac{\partial}{\partial \beta_j}\left[\frac{1}{n}\sum_{i=1}^{n}\ln_t(f(y_i-\mathbf{x}_i^T\boldsymbol{\beta}))\right]\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{f'(y_i-\mathbf{x}_i^T\boldsymbol{\beta})(-x_{ij})}{f(y_i-\mathbf{x}_i^T\boldsymbol{\beta})}w(f(y_i-\mathbf{x}_i^T\boldsymbol{\beta}))\Bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{f'(\epsilon_i)(-x_{ij})}{f(\epsilon_i)}w(f(\epsilon_i)).
\end{aligned}
$$

We first find bounds B_i such that for all ϵ_i and $\tilde{\epsilon}_i$, $i = 1, \ldots, n$,

$$
|d^j(\epsilon_1, ..., \epsilon_i, ..., \epsilon_n) - d^j(\epsilon_1, ..., \tilde{\epsilon}_i, ..., \epsilon_n)| \leq B_i. \tag{1.26}
$$

Because $\mathbb{E}[d^j(\epsilon_1, ..., \epsilon_n)] = 0$, by McDiarmid's inequality we have for all $v > 0$ and for all $j = 1, \ldots, d$,

$$
P(|d^j(\epsilon_1, ..., \epsilon_n)| > v) \leq 2\exp\left\{-\frac{2v^2}{\sum_{i=1}^{n}B_i^2}\right\}. \tag{1.27}
$$

Let $g^j(\epsilon_i) = \frac{f'(\epsilon_i)(-x_{ij})}{f(\epsilon_i)}w(f(\epsilon_i))$. We can show that $\max_{\epsilon_i}|g^j(\epsilon_i)| \leq C_t$, where C_t is a constant depending on the choice of t and σ_R. Given that the tangent likelihood order $p = 1$, we have

$$
\left|g^j(\epsilon_i)\right| = \left|\frac{\epsilon_i x_{ij}}{\sigma_R^2}\left[\frac{f(\epsilon_i)}{t}\right]^{\mathbb{1}(f(\epsilon_i)<t)}\right| = \begin{cases} \frac{1}{\sigma_R^2}|\epsilon_i x_{ij}| & \text{if } |\epsilon_i| < \epsilon_t \\ \frac{1}{\sigma_R^2}|\epsilon_i x_{ij} f(\epsilon_i)/t| & \text{if } |\epsilon_i| \geq \epsilon_t \end{cases}, \tag{1.28}
$$

where $\epsilon_t > 0$ such that $f(\epsilon_t) = t$. It is easy to see that the piecewise function (1.28) is continuous for all ϵ_i, and the first part is monotonically increasing while second part is bounded above. Therefore, the maximum value of $|g^j(\epsilon_i)|$ is taken at the second part of

(1.28). Hence,

$$\max_{\epsilon_i} \left| g^j(\epsilon_i) \right| \leq \frac{M}{\sigma_R^2} \max_{|\epsilon_i| \geq \epsilon_t} |\epsilon_i f(\epsilon_i)/t| \leq \frac{M}{\sigma_R^2} \max_{|\epsilon_i| \geq 0} |\epsilon_i f(\epsilon_i)/t| \,.$$

By assuming that $f(\cdot)$ is normal density with zero mean, $\arg\max_{\epsilon_i} |\epsilon_i f(\epsilon_i)/t| = \sigma_R$, so that $\frac{M}{\sigma_R^2} \max_{\epsilon_i} |\epsilon_i f(\epsilon_i)/t| = \frac{M}{t\sigma_R} f(\sigma_R)$.

It is worth noting that although $\max_{\epsilon_i} |g^j(\epsilon_i)|$ is globally bounded by $\frac{M}{t\sigma_R} f(\sigma_R)$, the maximum of $|g^j(\epsilon_i)|$ may have two different values depending on whether $\epsilon_t \geq \sigma_R$ or $\epsilon_t \leq \sigma_R$. If the threshold point $\epsilon_t > \sigma_R$, then $\max_{\epsilon_i} |g^j(\epsilon_i)| = \frac{M}{\sigma_R^2} \epsilon_t$, which is strictly smaller than the bound $\frac{M}{t\sigma_R} f(\sigma_R)$. On the other hand, if $\epsilon_t \leq \sigma_R$, $\max_{\epsilon_i} |g^j(\epsilon_i)| = \frac{M}{t\sigma_R} f(\sigma_R)$. Therefore, we have $C_t = \frac{M\check{\epsilon}}{t\sigma_R^2} f(\check{\epsilon})$, where $\check{\epsilon} = \max\{\sigma_R, \epsilon_t\}$.

In general, it can be shown that C_t is a finite constant that depends on t and σ_R^2. For fixed ϵ_1,... ϵ_{i-1},ϵ_{i+1},...ϵ_n, we have

$$\max_{\epsilon_i, \tilde{\epsilon}_i} \left| d^j(\epsilon_1, ..., \epsilon_i, ..., \epsilon_n) - d^j(\epsilon_1, ..., \tilde{\epsilon}_i, ..., \epsilon_n) \right| = \max_{\epsilon, \tilde{\epsilon}} \frac{1}{n} \left| g^j(\epsilon) - g^j(\tilde{\epsilon}) \right| \leq \frac{2}{n} C_t.$$

It implies that $\sum_{i=1}^n B_i^2 \leq n(2C_t/n)^2 = 4C_t^2/n$. By (1.27), we have

$$P(|d^j(\epsilon_1, ..., \epsilon_n)| > v) \leq 2 \exp\left\{ -\frac{nv^2}{2C_t^2} \right\}.$$

By union bound over the predictors we obtain

$$P\left\{ \max_{j=1,...,d} |d^j(\epsilon_1, \ldots, \epsilon_n)| > v \right\} \leq 2 \exp\left\{ -\frac{nv^2}{2C_t^2} + \ln(d) \right\}. \tag{1.29}$$

We can set $v = \lambda_n/2 = C_t \sqrt{2(\alpha_1 + 1)\ln(d)/n}$, where $\alpha_1 \geq 0$. This completes the proof.

□

Proof of Lemma 2. Denote $r_i = y_i - \mathbf{x}_i^T\boldsymbol{\beta}$, the tangent loss function $\mathcal{L}(\boldsymbol{\beta}) := -\frac{1}{n}\sum_{i=1}^{n} \ln_t(f(r_i))$. Consider the approximation error of the first order Taylor expansion of $\mathcal{L}(\boldsymbol{\beta})$ around $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_0 + \boldsymbol{\Delta}$, where $\boldsymbol{\Delta} \in \mathbb{R}^d$. For some $a \in [0, 1]$, we have

$$\begin{aligned}\delta\mathcal{L}(\boldsymbol{\beta}_0, \boldsymbol{\Delta}) &= \mathcal{L}(\boldsymbol{\beta}_0 + \boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\beta}_0) - \langle \nabla\mathcal{L}(\boldsymbol{\beta}_0), \boldsymbol{\Delta}\rangle \\ &= \boldsymbol{\Delta}^T \nabla^2\mathcal{L}(\boldsymbol{\beta}_0 + a\boldsymbol{\Delta})\boldsymbol{\Delta}. \end{aligned} \tag{1.30}$$

The following proof is based on the assumptions that f is normal density function and that the tangent likelihood order $p = 1$. Let $r_i = y_i - \mathbf{x}_i^T(\boldsymbol{\beta}_0 + a\boldsymbol{\Delta})$. By simple algebra,

$$\boldsymbol{\Delta}^T \nabla^2\mathcal{L}(\boldsymbol{\beta}_0 + a\boldsymbol{\Delta})\boldsymbol{\Delta} = \frac{1}{n}\sum_{i=1}^{n} h(z_i)(\mathbf{x}_i^T\boldsymbol{\Delta})^2,$$

with

$$h(z_i) = \begin{cases} 1/c_0 & \text{if } z_i \le R \\ \frac{1}{c_1}(1 - z_i)\exp(-z_i/2) & \text{if } z_i > R \end{cases},$$

where $c_0 = \sigma_R^2$, $z_i = r_i^2/c_0$, $c_1 = c_0^{3/2} t\sqrt{2\pi}$, and $R = -2\ln(t\sqrt{2\pi c_0})$. We require $0 < t < f(0)$, where $f(\cdot)$ is normal density function with mean equal to 0. Note that if $t > f(0)$, the MTE reduces to minimum ℓ_2 distance estimation and $h(z_i) = \frac{1}{c_1}(1 - z_i)\exp(-z_i/2)$ is always true. The corresponding proof can be found in Lozano et al. [37]. By requiring $0 < t < f(0)$, our proof is simplified without introducing extra constants, and needs weaker conditions than that of Lozano et al. [37].

It is easy to see that function $(1 - z_i)\exp(-z_i/2)$ attains its minimum $-2e^{-3/2}$ when $z_i = 3$. Then we have

$$\boldsymbol{\Delta}^T \nabla^2 \mathcal{L}(\boldsymbol{\beta}_0 + a\boldsymbol{\Delta})\boldsymbol{\Delta} \geq \frac{1}{n}\sum_{i=1}^{n} \phi(z_i)(\mathbf{x}_i^T \boldsymbol{\Delta})^2, \tag{1.31}$$

where

$$\phi(z_i) = \begin{cases} 1/c_0 & \text{if } z_i \leq R \\ -2e^{-3/2}/c_1 & \text{if } z_i > R \end{cases}.$$

To complete the proof, we will show that with high probability,

$$\frac{1}{n}\sum_{i=1}^{n} \phi(z_i)(\mathbf{x}_i^T \boldsymbol{\Delta})^2 \geq \kappa_1 \|\boldsymbol{\Delta}\|_2 \left(\|\boldsymbol{\Delta}\|_2 - \kappa_2 \sqrt{\frac{\ln d}{n}} \|\boldsymbol{\Delta}\|_1 \right) \tag{1.32}$$

for any $\boldsymbol{\Delta} \in \mathbb{H}(S, u; v) := \mathbb{C}(S) \cap \{\boldsymbol{\Delta} : \|\boldsymbol{\Delta}\|_1 = v, \|\boldsymbol{\Delta}\|_2 = u\}$. It is equivalent to show that the complement event of (1.32) holds with very small probability. In particular,

$$\frac{1}{n}\sum_{i=1}^{n} \phi(z_i)(\mathbf{x}_i^T \boldsymbol{\Delta})^2 < \kappa_1 u \left(u - \kappa_2 \sqrt{\frac{\ln d}{n}} v \right), \quad \text{for some } \boldsymbol{\Delta} \in \mathbb{H}(S, u, v). \tag{1.33}$$

Following is the outline of the proof:

1. Establish the lower bound for $\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} \phi(z_i)(\mathbf{x}_i^T \boldsymbol{\Delta})^2\right]$.

2. Show tail bound for $Q(v)$, where

$$Q(v) := \sup_{\boldsymbol{\Delta} \in \mathbb{H}(S,u,v)} \left| \frac{1}{n}\sum_{i=1}^{n} \left[\phi(z_i) - \mathrm{E}\phi(z_i)\right] (\mathbf{x}_i^T \boldsymbol{\Delta})^2 \right|. \tag{1.34}$$

To show the tail bound, we need to

2a. establish upper bound for $Q(v)$;

2b. establish upper bound for $\mathrm{E}\,[Q(v)]$.

3. Use the peeling argument in Negahban et al. [41] to show that v can be arbitrary.

First, we establish the lower bound for $\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\phi(z_i)(\mathbf{x}_i^T\boldsymbol{\Delta})^2\right]$. Note that

$$\mathrm{E}\phi(z_i) = \frac{1}{c_0}P(z_i \le R) - \frac{2e^{-3/2}}{c_1}P(z_i > R) = \frac{1}{c_0} - \left(\frac{1}{c_0} + \frac{2e^{-3/2}}{c_1}\right)P(z_i > R),$$

and

$$\begin{aligned} P(z_i > R) &= P(r_i^2 > c_0 R) \\ &= P\left(\left|(y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0) - a\mathbf{x}_i^T\boldsymbol{\Delta}\right| > \sqrt{c_0 R}\right) \\ &= P\left((y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0) > \sqrt{c_0 R} + a\mathbf{x}_i^T\boldsymbol{\Delta}\right) + P\left((y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0) < -\sqrt{c_0 R} + a\mathbf{x}_i^T\boldsymbol{\Delta}\right). \end{aligned}$$

By assumption [A3], that is, $\mathbf{x}_i^T\boldsymbol{\Delta}$ is sub-Gaussian with parameter at most $\kappa_s^2\|\boldsymbol{\Delta}\|_2^2$, we have

$$P(|\mathbf{x}_i^T\boldsymbol{\Delta}| \ge w) \le 2\exp\left(-\frac{w^2}{2\kappa_s^2}\|\boldsymbol{\Delta}\|_2^2\right) \quad \text{for all } w > 0. \tag{1.35}$$

It implies that $\max_i |\mathbf{x}_i^T\boldsymbol{\Delta}| \le 2\kappa_s\|\boldsymbol{\Delta}\|_2\sqrt{\ln n}$ with probability at least $1 - 1/n^2$. Then we have

$$\begin{aligned} &P\left((y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0) > \sqrt{c_0 R} + a\mathbf{x}_i^T\boldsymbol{\Delta}\right) \le P\left((y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0) > \sqrt{c_0 R} - 2a\kappa_s\|\boldsymbol{\Delta}\|_2\sqrt{\ln n}\right); \\ &P\left((y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0) < -\sqrt{c_0 R} + a\mathbf{x}_i^T\boldsymbol{\Delta}\right) \le P\left((y_i - \mathbf{x}_i^T\boldsymbol{\beta}_0) < -\sqrt{c_0 R} + 2a\kappa_s\|\boldsymbol{\Delta}\|_2\sqrt{\ln n}\right). \end{aligned}$$

Therefore, $P(z_i > R) \leq P\left(|y_i - \mathbf{x}_i^T \boldsymbol{\beta}_0| > \sqrt{c_0 R} - 2\kappa_s u \sqrt{\ln n}\right) = \kappa_u$. We need $u < \frac{\sqrt{c_0 R}}{2\kappa_s \sqrt{\ln n}}$ so that $\kappa_u < 1$. Further, we need $\kappa_u \leq (1 + \frac{c_0}{c_1} 2e^{-3/2})^{-1}$ in order to have

$$\mathrm{E}\phi(z_i) \geq \frac{1}{c_0} - \left(\frac{1}{c_0} + \frac{2e^{-3/2}}{c_1}\right)\kappa_u \geq 0. \tag{1.36}$$

By the restricted eigenvalue condition, we have

$$\mathrm{E}\left[\boldsymbol{\Delta}^T \nabla^2 \mathcal{L}(\boldsymbol{\beta}_0 + a\boldsymbol{\Delta})\boldsymbol{\Delta}\right] \geq \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\phi(z_i)(\mathbf{x}_i^T \boldsymbol{\Delta})^2 \geq \left(\frac{1}{c_0} - c_2\kappa_u\right)\kappa_{RE} u^2, \tag{1.37}$$

where $c_2 = \left(\frac{1}{c_0} + \frac{2e^{-3/2}}{c_1}\right)$.

Next, we show the tail bound of $Q(v)$. By Massart concentration inequality [39, 7],

$$P\left(Q(v) \geq \mathrm{E}Q(v) + \xi\right) \leq \exp\left(-\frac{\xi^2}{2L^2}\right), \tag{1.38}$$

where $L^2 = \sup_{\boldsymbol{\Delta} \in \mathbb{H}(S,u,v)} \sum_{i=1}^{n} \left[b_i(\boldsymbol{\Delta})\right]^2$, and $b_i(\boldsymbol{\Delta})$ is the upper bound of

$$\left|\frac{1}{n}\left[\phi(z_i) - \mathrm{E}\phi(z_i)\right](\mathbf{x}_i^T \boldsymbol{\Delta})^2\right|.$$

To show (1.38), we need to determine L^2, and the upper bound of $\mathrm{E}Q(v)$. We fisrt show

$$|\phi(z_i) - \mathrm{E}\phi(z_i)| \leq \left(\frac{1}{c_0} + \frac{2}{c_1}e^{-3/2}\right).$$

Note that $0 \leq \mathrm{E}\phi(z_i) \leq 1/c_0$, and $\phi(z_i) = 1/c_0$ if $z_i \leq R$, $\phi(z_i) = -2e^{-3/2}/c_1$ otherwise. Therefore, $|\phi(z_i) - \mathrm{E}\phi(z_i)| \leq \mathrm{E}\phi(z_i) - (-2e^{-3/2}/c_1) \leq \left(\frac{1}{c_0} + \frac{2}{c_1}e^{-3/2}\right)$. Hence,

$$\left|\frac{1}{n}\left[\phi(z_i) - \mathrm{E}\phi(z_i)\right](\mathbf{x}_i^T \boldsymbol{\Delta})^2\right| \leq \frac{1}{n}\left|\phi(z_i) - \mathrm{E}\phi(z_i)\right|(\mathbf{x}_i^T \boldsymbol{\Delta})^2 \leq \frac{1}{n} c_2 (\mathbf{x}_i^T \boldsymbol{\Delta})^2,$$

where $c_2 = \left(\frac{1}{c_0} + \frac{2}{c_1}e^{-3/2}\right)$, and $(\mathbf{x}_i^T\boldsymbol{\Delta})^2 \leq 4\kappa_s^2 u^2 \ln n$. Therefore, we have $L^2 = (4c_2\kappa_s^2 u^2 \ln n)^2/n$.

Next, we upper bound $\mathrm{E}Q(v)$. Let w_i be i.i.d. Rademacher variable. By symmetrization theorem (Theorem 14.3 in Bühlmann and Van De Geer [7]), we have

$$\begin{aligned}
\mathrm{E}Q(v) &\leq 2\mathrm{E}\sup_{\boldsymbol{\Delta}\in\mathbb{H}(S,u,v)}\left|\frac{1}{n}\sum_{i=1}^n w_i\phi(z_i)(\mathbf{x}_i^T\boldsymbol{\Delta})^2\right| \\
&\leq \frac{2}{c_0}\mathrm{E}\sup_{\boldsymbol{\Delta}\in\mathbb{H}(S,u,v)}\left|\frac{1}{n}\sum_{i=1}^n w_i\mathbb{1}\left\{z_i \leq R\right\}(\mathbf{x}_i^T\boldsymbol{\Delta})^2\right| \\
&\quad + \frac{4}{c_1}e^{-3/2}\mathrm{E}\sup_{\boldsymbol{\Delta}\in\mathbb{H}(S,u,v)}\left|\frac{1}{n}\sum_{i=1}^n w_i\mathbb{1}\left\{z_i > R\right\}(\mathbf{x}_i^T\boldsymbol{\Delta})^2\right| \\
&\leq 2c_2\mathrm{E}\sup_{\boldsymbol{\Delta}\in\mathbb{H}(S,u,v)}\left|\frac{1}{n}\sum_{i=1}^n w_i(\mathbf{x}_i^T\boldsymbol{\Delta})^2\right|.
\end{aligned}$$

Since $(\mathbf{x}_i^T\boldsymbol{\Delta})^2$ is Lipschitz continuous with parameter $K = 4\kappa_s u\sqrt{\ln n}$ for all $\boldsymbol{\Delta} \in \mathbb{H}(S,u,v)$, using Ledoux-Talagrand Contraction inequality [35], we have

$$\mathrm{E}Q(v) \leq 8\kappa_s u c_2\sqrt{\ln n}\mathrm{E}\sup_{\boldsymbol{\Delta}\in\mathbb{H}(S,u,v)}\left|\frac{1}{n}\sum_{i=1}^n w_i(\mathbf{x}_i^T\boldsymbol{\Delta})\right|.$$

Further, by Hölder's inequality,

$$\mathrm{E}Q(v) \leq 8\kappa_s uvc_2\sqrt{\ln n}\mathbb{E}\left\|\frac{1}{n}\sum_{i=1}^n w_i\mathbf{x}_i\right\|_\infty.$$

Since $\mathbf{x}_i^T\boldsymbol{\Delta}$ is sub-Gaussian with parameter $\kappa_s^2u^2$, $\frac{1}{n}\sum_{i=1}^n w_i\mathbf{x}_i$ is also sub-Gaussian with parameter κ_s^2/n. The existing bounds of expectation of sub-Gaussian maxima [35] yield

$$\mathrm{E}\left\|\frac{1}{n}\sum_{i=1}^n w_i\mathbf{x}_i\right\|_\infty \leq 6\kappa_s\sqrt{\frac{\ln d}{n}}.$$

Therefore,

$$\mathrm{E}Q(v) \leq 48\kappa_s^2 uvc_2\sqrt{\frac{\ln n \ln d}{n}}. \tag{1.39}$$

Now, combine (1.38) and (1.39), and let

$$\xi = \frac{1}{2}\left(\frac{1}{c_0} - c_2\kappa_u\right)u^2\kappa_{RE} + \kappa_s^2 uvc_2\sqrt{\frac{\ln n \ln d}{n}},$$

we have

$$\begin{aligned} P\left(Q(v) \geq \frac{1}{2}\left(\frac{1}{c_0} - c_2\kappa_u\right)u^2\kappa_{RE} + 49\kappa_s^2 uvc_2\sqrt{\frac{\ln n \ln d}{n}}\right) \\ \leq \exp\left(-\frac{n\left(\frac{1}{2}\left(\frac{1}{c_0} - c_2\kappa_u\right)u^2\kappa_{RE} + \kappa_s^2 uvc_2\sqrt{\frac{\ln n \ln d}{n}}\right)^2}{32\kappa_s^4 u^4 c_2^2 (\ln n)^2}\right). \end{aligned}$$

Note that $P(\sup|a-b| \geq c) \geq P(|a-b| \geq c) \geq P(a-b \leq -c) \geq P(a \leq -c+b^*)$ given $b \geq b^*$. Hence, for any $\mathbf{\Delta} \in \mathbb{H}(S,u,v)$, the event (1.33) holds with small probability. In particular,

$$\frac{1}{n}\sum_{i=1}^{n}\phi(z_i)\left(\mathbf{x}_i^T\mathbf{\Delta}\right)^2 \leq \frac{1}{2}\left(\frac{1}{c_0} - c_2\kappa_u\right)u^2\kappa_{RE} - 49\kappa_s^2 uvc_2\sqrt{\frac{\ln n \ln d}{n}}$$

holds with the probability at most

$$\begin{aligned} &\exp\left(-\frac{n\left(\frac{1}{2}\left(\frac{1}{c_0} - c_2\kappa_u\right)u^2\kappa_{RE} + \kappa_s^2 uvc_2\sqrt{\frac{\ln n \ln d}{n}}\right)^2}{32\kappa_s^4 u^4 c_2^2 (\ln n)^2}\right) \\ \leq &\exp\left(-\frac{n\left(\frac{1}{2}\left(\frac{1}{c_0} - c_2\kappa_u\right)\kappa_{RE}\right)^2}{32\kappa_s^4 c_2^2 (\ln n)^2} - \frac{\ln d}{32\ln n}\right). \end{aligned}$$

Therefore, by a peeling argument [44], we have the restrict strong convexity with probability at least $1 - \alpha_3 \exp(-\alpha_4 n)$. □

Proof of Theorem 1.3.3. We consider the set $\mathbb{H}(S, u) := \mathbb{C}(S) \cap \{\boldsymbol{\Delta} \in \mathbb{R}^d : \|\boldsymbol{\Delta}\|_2 = u\}$, where $u < \sqrt{c_0 R}/(2\kappa_s \sqrt{\ln n})$. Define function $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}$ as:

$$\mathcal{F}(\boldsymbol{\Delta}) := \mathcal{L}(\boldsymbol{\beta}_0 + \boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\beta}_0) + \lambda_n \left(\|\boldsymbol{\beta}_0 + \boldsymbol{\Delta}\|_1 - \|\boldsymbol{\beta}_0\|_1 \right).$$

We first give the following Lemma, and use this result to complete the theorem proof.

Lemma 5. *If $\mathcal{F}(\boldsymbol{\Delta}) > 0$ for all $\boldsymbol{\Delta} \in \mathbb{H}(S, u)$, then $\|\hat{\boldsymbol{\Delta}}\| \leq u$, where $\hat{\boldsymbol{\Delta}} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$.*

Proof of Lemma 5. We show the lemma by contradiction. Suppose that for some optimal $\hat{\boldsymbol{\beta}}$ (i.e. minimizer of $\mathcal{L}(\boldsymbol{\beta}) + \lambda_n \|\boldsymbol{\beta}\|_1$), $\|\hat{\boldsymbol{\Delta}}\|_2 > u$. Then the line joining $\hat{\boldsymbol{\Delta}}$ and 0 must intersect with the set $\mathbb{H}(S, u)$ at $a^* \hat{\boldsymbol{\Delta}}$ for some $a^* \in (0, 1)$. We know that $\mathcal{L}(\cdot)$ is locally convex, so is $\mathcal{F}(\cdot)$. Therefore, $\mathcal{F}(a^* \hat{\boldsymbol{\Delta}}) = \mathcal{F}(a^* \hat{\boldsymbol{\Delta}} + (1 - a^*)0) \leq a^* \mathcal{F}(\hat{\boldsymbol{\Delta}}) + (1 - a^*)\mathcal{F}(0) = a^* \mathcal{F}(\hat{\boldsymbol{\Delta}})$.

Since $\hat{\boldsymbol{\beta}}$ is optimal, $\mathcal{F}(\hat{\boldsymbol{\Delta}}) \leq 0$, hence $\mathcal{F}(a^* \hat{\boldsymbol{\Delta}}) \leq 0$. However, $a^* \hat{\boldsymbol{\Delta}} \in \mathbb{H}(S, \mu)$. This is a contradiction. □

We now complete the proof of Theorem 1.3.3. By (1.30), we know that $\mathcal{L}(\boldsymbol{\beta}_0 + \boldsymbol{\Delta}) - \mathcal{L}(\boldsymbol{\beta}_0) = \langle \nabla \mathcal{L}(\boldsymbol{\beta}_0), \boldsymbol{\Delta} \rangle + \delta \mathcal{L}(\boldsymbol{\beta}_0, \boldsymbol{\Delta})$. Then we have

$$\mathcal{F}(\boldsymbol{\Delta}) = \langle \nabla \mathcal{L}(\boldsymbol{\beta}_0), \boldsymbol{\Delta} \rangle + \delta \mathcal{L}(\boldsymbol{\beta}_0, \boldsymbol{\Delta}) + \lambda_n \left(\|\boldsymbol{\beta}_0 + \boldsymbol{\Delta}\|_1 - \|\boldsymbol{\beta}_0\|_1 \right).$$

By Lemma 2, we know that the restricted strong convexity, $\delta \mathcal{L}(\boldsymbol{\beta}_0, \boldsymbol{\Delta}) \geq \frac{\kappa_1}{2} \|\boldsymbol{\Delta}\|_2^2$, holds with probability at least $1 - \alpha_3 \exp(-\alpha_4 n)$. Therefore, with the same probability,

$$\mathcal{F}(\boldsymbol{\Delta}) \geq \langle \nabla \mathcal{L}(\boldsymbol{\beta}_0), \boldsymbol{\Delta} \rangle + \frac{\kappa_1}{2} \|\boldsymbol{\Delta}\|_2^2 + \lambda_n \left(\|\boldsymbol{\beta}_0 + \boldsymbol{\Delta}\|_1 - \|\boldsymbol{\beta}_0\|_1 \right)$$

$$\geq \langle \nabla \mathcal{L}(\boldsymbol{\beta}_0), \boldsymbol{\Delta} \rangle + \frac{\kappa_1}{2} \|\boldsymbol{\Delta}\|_2^2 + \lambda_n \left(\|\boldsymbol{\Delta}_{S^c}\|_1 - \|\boldsymbol{\Delta}_S\|_1 \right).$$

Note that $\boldsymbol{\beta}_0$ and $\boldsymbol{\Delta}_{S^c}$ are in complementary model subspaces. Hence, by definition of decomposability of ℓ_1 regularizer in Negahban et al. [41], $\|\boldsymbol{\beta}_0 + \boldsymbol{\Delta}_{S^c}\|_1 = \|\boldsymbol{\beta}_0\|_1 + \|\boldsymbol{\Delta}_{S^c}\|_1$.

By Hölder's inequality, $|\langle \nabla \mathcal{L}(\boldsymbol{\beta}_0), \boldsymbol{\Delta} \rangle| \leq \|\nabla \mathcal{L}(\boldsymbol{\beta}_0)\|_\infty \|\boldsymbol{\Delta}\|_1$. By Lemma 1, we know that $\lambda_n \geq 2\|\nabla \mathcal{L}(\boldsymbol{\beta}_0)\|_\infty$ holds with probability at least $1 - 2\exp(-\alpha_2 n \lambda_n^2)$. Then with the same probability, $|\langle \nabla \mathcal{L}(\boldsymbol{\beta}_0), \boldsymbol{\Delta} \rangle| \leq \frac{\lambda_n}{2} \|\boldsymbol{\Delta}\|_1$. Hence,

$$\begin{aligned}
\mathcal{F}(\boldsymbol{\Delta}) &\geq \frac{\kappa_1}{2} \|\boldsymbol{\Delta}\|_2^2 + \lambda_n \left(\|\boldsymbol{\Delta}_{S^c}\|_1 - \|\boldsymbol{\Delta}_S\|_1 \right) - \frac{\lambda_n}{2} \|\boldsymbol{\Delta}\|_1 \\
&= \frac{\kappa_1}{2} \|\boldsymbol{\Delta}\|_2^2 + \lambda_n \left(\frac{1}{2} \|\boldsymbol{\Delta}_{S^c}\|_1 - \frac{3}{2} \|\boldsymbol{\Delta}_S\|_1 \right) \\
&\geq \frac{\kappa_1}{2} \|\boldsymbol{\Delta}\|_2^2 - \frac{3\lambda_n}{2} \|\boldsymbol{\Delta}_S\|_1 \\
&\geq \frac{\kappa_1}{2} \|\boldsymbol{\Delta}\|_2^2 - \frac{3\lambda_n \sqrt{s}}{2} \|\boldsymbol{\Delta}\|_2.
\end{aligned}$$

Let $G(\|\boldsymbol{\Delta}\|_2) = \frac{\kappa_1}{2} \|\boldsymbol{\Delta}\|_2^2 - \frac{3\lambda_n \sqrt{s}}{2} \|\boldsymbol{\Delta}\|_2$ where $G(\|\boldsymbol{\Delta}\|_2)$. Then the root of $G(\cdot)$ is $\frac{3\lambda_n}{\kappa_1} \sqrt{s} > 0$. Therefore, $\mathcal{F}(\boldsymbol{\Delta}) > 0$ for all $\|\boldsymbol{\Delta}\|_2 > \frac{3\lambda_n}{\kappa_1} \sqrt{s}$ with probability at least $1 - 2\exp(-\alpha_2 n \lambda_n^2)$. Let $u = \frac{4\lambda_n}{\kappa_1} \sqrt{s}$ so that $\mathcal{F}(\boldsymbol{\Delta})$ is strictly positive. We then only need $\frac{4\lambda_n}{\kappa_1} \sqrt{s} \leq \sqrt{c_0 R}/(2\kappa_s \sqrt{\ln n})$, which holds as long as $n > \frac{16^2 \xi^2 \kappa_s^2 s \ln n \ln d}{\kappa_1^2 c_0 R}$. By results of Lemma 5, we have

$$\|\hat{\boldsymbol{\Delta}}\| \leq \frac{4\lambda_n}{\kappa_1} \sqrt{s}.$$

By choosing $\lambda_n = 2\xi \sqrt{\ln(d)/n}$, we have with probability at least $1 - 2\exp(-\alpha_2 n \lambda_n^2)$,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \leq \frac{8}{\kappa_1} \xi \sqrt{\frac{s \ln(d)}{n}}.$$

□

Chapter 2:

Corporate Bankruptcy Prediction:

A Penalized Semiparametric Index Hazard Model Approach

2.1 Introduction

Corporate bankruptcy prediction is of paramount interest in risk management. Default probabilities are necessary inputs to pricing credit derivatives [12]. Accurate default predictions are critical to financial institutions that are required by the international Basel Committee to reserve enough cash to cover risks incurred by operations. Consequently, corporate bankruptcy forecasting has attracted great attention in the past decades. Both reliable and easy-interpretable statistical models are desired to accurately predict firms bankruptcy risk in order to help government and financial institution to early detect default and minimize the potential losses.

One of the most cited bankruptcy prediction models, Altman's Z-score in a linear discriminant analysis, was introduced by Altman [2], where a static model with cross-sectional data was developed. Their empirical study is based on the largest sector–manufacturing sector on a small matched sample. Five financial ratios, *Working capital/total assets, Retained earnings/total assets, Earnings before interests and taxes/total assets, Market value equity/book value of total debt,* and *Sales/total assets* are used as predictors. Shumway [50] proposed a discrete hazard model, often termed as Shumway's model, arguing that static

model brings bias as it ignores the panel structure of data. Shumway's model is widely adopted by later researchers and becomes the state-of-the-art model in the current literature (see, e.g. [12], [8]). Ding et al. [15] note that the popular Shumway discrete hazard model is indeed the discrete logistic model [13] for time-varying covariates, and statistically equivalent to multi-period logit model.

The aforementioned popular bankruptcy prediction models are based on a linear model framework, whereas the linearity assumption may often not hold and the model may be subject to misspecification. In this article, we investigate potential nonlinearity between firm's default risk and financial variables by considering a semiparametric index hazard model [28, 27, 9, 32, 66] for bankruptcy prediction:

$$g\left(P\left[Y_{i,t+1}=1|Y_{i,t}=0,\mathbf{x}_{i,t}\right]\right)=\phi\left(\boldsymbol{\alpha}^T\mathbf{x}_{i,t}\right), \tag{2.1}$$

where $Y_{i,t}=1$ if firm i goes bankrupt at time t and 0 otherwise, $\mathbf{x}_{i,t}$ is a vector of predictors for firm i at time t, $\boldsymbol{\alpha}$ is a vector of index coefficient, $g(\cdot)$ is a prespecified canonic link function, i.e. the logit link function $g(\pi)=\log(\frac{\pi}{1-\pi})$ for $\pi\in(0,1)$ in our study, and $\phi(\cdot)$ is an unspecified link function that needs to be estimated. Model (2.1) is a natural combination of single-index models [28] and generalized linear model [40] for binary responses. The single-index model has been increasingly popular in many fields of quantitative research, exhibiting many appealing features. It circumvents the so-called "curse of dimensionality" as the "single-index" projects p-dimensional explanatory vector space to a one-dimensional vector so that it only demands a univariate nonparametric estimation. The unknown nonparametric function $\phi(\cdot)$ is flexible to handle nonlinearity. When $\phi(\cdot)$ is monotonic, the signs of single-index coefficients can be interpreted similarly as in the linear model. When

$\phi(\cdot)$ is identity, (2.1) reduces to Shumway's model. In our study, the "single-index" can be viewed as a composite financial index, which makes model (2.1) particularly appealing to the bankruptcy prediction. Many developed algorithms can be adopted to estimate model (2.1).

With numerous financial variables being available, it is crucial to determine important variables to be included in the predictive model. This investigation may also shed light on the debate between accounting and market researchers [52]. Accounting ratios are often adopted to predict default risk in early works by accounting researchers [2, 42, 71]. Shumway [50] is among the first to introduce market based variables as predictors and show noticeable gains. Campbell et al. [8] further modified accounting based variables with market information. Chava and Jarrow [12] show clear and even dominating advantages using market variables. Most previous work prespecifies somewhat different fixed set of explanatory variables. Many accounting and market based variables have been suggested, yet there are few studies to formally determine important predictors. Tian et al. [52] are among the first to introduce LASSO variable selection to the bankruptcy literature under the linear hazard model framework and find important roles of both accounting ratios and market variables. In fact, identifying important variables among a large number of predictors is challenging but crucial in many scientific research fields. Irrelevant variables may bring additional noise, and simpler model is often preferred for its interpretability. The traditional best subset variable selection method not only causes heavy computational duty but also brings stochastic errors (see, e.g., 18). Instead, the regularized methods, such as LASSO [53], Adaptive-Lasso [72], SCAD [18], and MCP [67], can select important variables and estimate the coefficients simultaneously. Furthermore, such regularized methods are computationally efficient and

scalable to large dataset with high-dimensional features whereas classical stepwise variable selection is infeasible.

In order to capture potential nonlinearity and select important predictors, in this paper, we propose a penalized index hazard model for bankruptcy prediction and variable selection. Motivated by the long debate between accounting and finance researchers, we further propose a novel penalized double-index hazard model. Market and accounting variables naturally form the candidate variable sets of the two indices. Automatic variable selection is achieved by the shrinkage estimation with a penalty, such as SCAD. We show that the proposed penalized double-index hazard model is specifically tailored for corporate bankruptcy prediction with the following advantages: (1) important market variables and accounting ratios are used to construct two indices naturally; (2) two nonparametric link functions allow nonlinear effect of the constructed two indices, market and accounting index, to the firm's default risk; (3) important variables can be automatically selected; and (4) more interestingly, interpretation of the index coefficients as well as two constructed composite indices: market and accounting index may be of great potential interest in practice.

We focus on the publicly traded manufacturing firms that form the largest industry sector in size with the highest bankruptcy rates. We construct the comprehensive database from 1980 to 2015 by merging monthly equity data from the Center for Research in Security Prices (CRSP) and quarterly financial data from the Standard & Poor's COMPUSTAT. The dataset is discrete in nature over the time. We construct a total of 23 financial variables, among which 10 are market variables and the rest are accounting ratios that have been considered in the previous bankruptcy literature (e.g. [2, 42, 71, 50, 12]). Default dummy is the response of interest, indicating a company's filing of bankruptcy protection under either Chapter 7 or Chapter 11. We describe further details in Section 2.4.

The rest of this paper is organized as following: Section 2.2 describes the database we have constructed and used for our empirical study. Section 2.3 introduces the proposed double-index hazard model along with the penalized estimation and algorithm. A Monte Carlo simulation study has been conducted in the end of Section 2.3 in order to demonstrate the effectiveness of the proposed model and efficient estimation algorithm. Our empirical results are shown in Section 2.4.

2.2 Data

In this section, we describe our bankruptcy database. Our bankruptcy database is consisting of a panel dataset of 65,220 firm-year observations for a total of 5,547 firms in the manufacturing sector from 1980 to 2015.

To estimate the default risk, we need a binary response variable that indicates a firms bankruptcy status and a set of explanatory variables. To construct firm-level explanatory predictor variables, we merge daily and monthly equity information from CRSP with the annually updated accounting information from COMPUSTAT database through the Wharton Research Data Services (WRDS). In this study, we consider both market-based and accounting-based predictor variables at each individual firm level. There are a total of 23 candidate explanatory variables. Such set of predictor variables is a tailored list of variables that have appeared in studies including [3, 2, 42, 71, 50, 12, 4, 29, 5], and many others. We further partition this predictor variable set into two groups: market and accounting variables. In specific, if the variable is formulated only by using the balance sheet or income statement data from COMPUSTAT, we classify the predictor variable as an accounting variable, for example, the leverage ratio of total liability over the total assets and the profitability ratio of net income over the total assets. If the variable is constructed using the market trading

information from CRSP, for example, the stock price or return, we classify the predictor variable as a market variable. As a result, our predictor variable set includes 10 market and 13 accounting based variables. The detailed description for each predictor variable is summarized in Table 2.1.

Table 2.1: Variable names and descriptions of bankruptcy predictors

Variable	Description
	Panel A: Market-based Variables
SIGMA	Stock volatility
EXRET	Excess return over S&P 500 index
NIMTA	Net income/(market equity+total liabilities)
LTMTA	Total liabilities/(total liabilities+market equity)
CASHMTA	Cash and short-term investment/(market equity + total liabilities)
SIZE	log(market capitalization)
LOG_PRICE	log(price)
MBE	Market-to-book ratio
LCTMTA	Current liabilities/(market equity+total liabilities)
MVEF	Market equity/total debt
	Panel B: Accounting-based Variables
LTAT	Total liabilities/total assets
LCTAT	Current liabilities/total asset
NIAT	Net income/total assets
EBITAT	Earnings before interest and tax/total asset
REAT	Retained earnings/total assets
RELCT	Retained earnings/current liabilities
LCTSALE	Current liabilities/sales
LOG_SALE	log(sales)
CHAT	Cash/total asset
ACTLCT	Current asset/current liabilities
WCAPAT	Working capital/total assets
LCTLT	Current liabilities/total liabilities
SALEAT	Sales/total assets

In this study, a bankruptcy is defined to occur only if the company filed under either Chapter 7 (liquidation) or Chapter 11 (reorganization) protection code. Even though it is

not necessary for firms that filed for Chapter 11 to end with Chapter 7 filing, such definition is quite common in most bankruptcy literatures in order to identify the firms with financial difficulties. To estimate a company's default risk, the bankruptcy indicator is set to unity in the year that the firm exits the database due to either Chapter 7 or Chapter 11 filing deletion. The bankruptcy indicator is set to zero for all other firms that are either healthy or deleted or delisted due to other reasons such as merge and acquisition. For bankrupted firms, bankruptcy indicator is set to zero until the time the company survives through prior to the deletion or delisting. As a result, we have a total of 543 bankrupted firms in manufacturing industry over the sampling period. Table 2.2 summarizes the firm distribution by year. The first column shows the number of bankruptcies reported and the second column reports the number of active firms each year. The last column summarizes the corresponding bankruptcy percentage. It is quite apparent that high bankruptcy frequency and the business recession coincide, for example, the early 1990s recessions, the internet bubble in early 2000s, the 2008 financial crisis and the recent subprime mortgage crisis. Figure 2.1 plots the bankruptcy frequency across years, where we can see that it changes along with the economy.

Consistent to prior literatures [50, 12, 8], we also carefully align the company's fiscal year to the calendar year and lag the temporally aligned annual records by four month to ensure the accounting information is available to the market at the time of prediction. Winsorization is common in bankruptcy literature to avoid potential outliers. In this work, we winsorize selected predictors at either or both top and bottom percentiles if their histograms suggest a heavy-tail feature. Specifically, we replace any value that is lower than the 1st percentile or higher than the 99th percentile with its 1st percentile value or its 99th percentile value for winsorization.

Table 2.2: Count of bankruptcy firms and total number of firms over year.

Year	Bankruptcies	Active Firms	(%)
1980	11	1637	0.67
1981	12	1691	0.71
1982	12	1781	0.67
1983	16	1796	0.89
1984	14	1937	0.72
1985	15	1977	0.76
1986	24	1992	1.20
1987	9	2027	0.44
1988	16	2084	0.77
1989	19	2013	0.94
1990	31	2001	1.55
1991	31	1935	1.60
1992	18	1944	0.93
1993	17	1984	0.86
1994	18	2127	0.85
1995	11	2217	0.50
1996	10	2307	0.43
1997	24	2433	0.99
1998	35	2444	1.43
1999	17	2276	0.75
2000	20	2149	0.93
2001	30	2087	1.44
2002	36	1955	1.84
2003	15	1767	0.85
2004	7	1664	0.42
2005	10	1594	0.63
2006	4	1530	0.26
2007	5	1502	0.33
2008	11	1454	0.76
2009	17	1386	1.23
2010	3	1324	0.23
2011	5	1293	0.39
2012	9	1267	0.71
2013	4	1221	0.33
2014	3	1201	0.25
2015	4	1223	0.33

Table 2.3 reports the summary statistics for the winsorized data set. In specific, the left (right) panel summarizes the distribution for the bankrupted (non-bankrupted) firms at the firm-year level. One apparent conclusion we observe from Table 2.3 is that the bankruptcy group demonstrates quite different financial behaviors from the non-bankruptcy firm group.

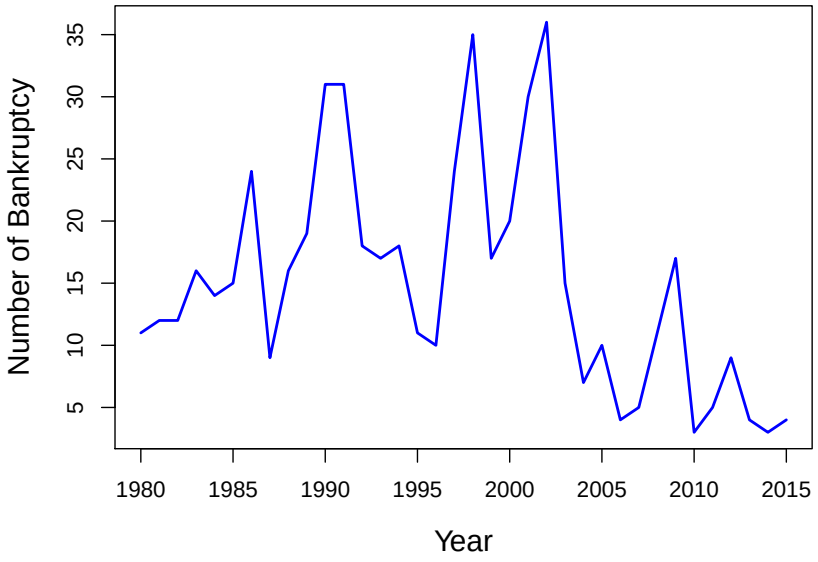

Figure 2.1: Number of bankrupted firms across years from 1980-2015.

The bankruptcy firms tend to have high debt and liabilities relative to their assets, smaller size in terms of their asset values and market capitalization, lesser profitability values and very negative reported earnings and returns. The table also shows that the bankruptcy firms are usually more volatile, where the average market return volatility for the bankruptcy group is 1.553 while it is only 0.612 for the non-bankruptcy group. The bankrupted firms also have a lower average trading price of -0.488, comparing to 1.938 at log scale for the non-bankruptcy firms.

Table 2.3: Summary statistics for bankruptcy predictors

	Bankrupt Firm (No. Firm-year = 543)					*Nonbankrupt Firm (No. Firm-year = 64677)*				
Variable	Mean	Std.	Min	Median	Max	Mean	Std.	Min	Median	Max
Panel A: Market Variable										
SIGMA	1.553	0.806	0.000	1.549	2.722	0.612	0.466	0.000	0.475	2.722
EXRET	-0.732	0.786	-1.638	-0.664	2.345	-0.116	0.535	-1.638	-0.064	3.376
NIMTA	-0.251	0.273	-0.787	-0.165	0.214	-0.013	0.138	-0.787	0.025	0.214
LTMTA	0.796	0.228	0.057	0.892	0.999	0.434	0.261	0.001	0.413	1.002
CASHMTA	0.108	0.152	-0.010	0.043	0.587	0.093	0.125	-0.009	0.045	0.587
SIZE	-13.573	1.667	-18.567	-13.759	-7.617	-10.461	2.193	-18.949	-10.505	-2.878
LOG_PRICE	-0.488	1.224	-1.856	-0.693	2.708	1.938	1.074	-1.856	2.526	2.708
MBE	0.665	2.066	-3.412	0.204	10.386	1.901	1.897	-3.412	1.457	10.386
LCTMTA	0.464	0.271	0.015	0.409	0.981	0.205	0.163	0.000	0.161	0.996
MVEF	0.031	0.183	0.000	0.001	1.776	0.081	0.290	0.000	0.005	1.776
Panel B: Accounting Variable										
LTAT	0.796	0.299	0.042	0.799	1.301	0.530	0.231	0.005	0.533	1.301
LCTAT	0.442	0.252	0.009	0.393	0.848	0.255	0.152	0.000	0.224	0.848
NIAT	-0.304	0.380	-1.259	-0.164	0.239	-0.026	0.218	-1.259	0.034	0.239
EBITAT	-0.193	0.305	-1.015	-0.091	0.754	0.025	0.195	-1.015	0.068	0.736
REAT	-1.035	1.655	-6.546	-0.383	0.830	-0.147	1.061	-6.546	0.123	3.186
RELCT	-2.418	4.439	-17.292	-0.904	6.704	-0.205	3.866	-17.292	0.563	6.704
LCTSALE	0.577	0.479	0.075	0.390	1.782	0.316	0.290	0.075	0.235	1.782
LOG_SALE	3.875	2.248	-4.605	3.926	11.912	5.168	2.382	-6.908	5.116	12.478
CHAT	0.075	0.114	-0.014	0.035	0.604	0.084	0.114	-0.069	0.039	0.604
ACTLCT	1.584	1.534	0.016	1.182	11.844	2.458	1.915	0.000	1.954	11.844
WCAPAT	0.065	0.286	-0.400	0.074	0.850	0.250	0.239	-0.400	0.243	0.971
LCTLT	0.603	0.295	0.016	0.623	1.000	0.538	0.266	0.001	0.516	1.000
SALEAT	1.092	0.763	0.000	1.031	3.017	1.024	0.590	-0.149	0.973	3.017

2.3 Semiparametric Index Model

2.3.1 Double-index hazard model

We propose the double-index hazard model for bankruptcy prediction, which is a simple extension of the single-index model (2.1). In particular, the proposed model is of form

$$g\left(P\left[Y_{i,t}=1|\mathbf{x}_{i,t-l},\mathbf{z}_{i,t-l}\right]\right)=\phi_1\left(\boldsymbol{\alpha}^T\mathbf{x}_{i,t-l}\right)+\phi_2\left(\boldsymbol{\beta}^T\mathbf{z}_{i,t-l}\right),\quad \|\boldsymbol{\alpha}\|=\|\boldsymbol{\beta}\|=1, \tag{2.2}$$

where $\mathbf{x}_{i,t}\in\mathbb{R}^s$ and $\mathbf{z}_{i,t}\in\mathbb{R}^d$ are two different sets of predictive information, i.e., market-based and accounting-based variables, that are observed from firm i at time t. Comparing

to model (2.1), the only difference is that the double-index model (2.2) involves a second unknown link function $\phi_2(\cdot)$ for another index term $\boldsymbol{\beta}^T\mathbf{z}$. A key restriction condition for the index coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is that $\|\boldsymbol{\alpha}\| = 1, \|\boldsymbol{\beta}\| = 1$. This is a common assumption in semi-parametric index model such that the index parameters can be uniquely identified through certain type of reparameterization. One popular way to reparameterize the index coefficient is called "delete-one" method. In particular and without loss of generality, for parameter vector $\boldsymbol{\beta}$, we first let $\boldsymbol{\beta}^* = \boldsymbol{\beta}/\beta_1 = (1, \beta_2^*, \ldots, \beta_d^*)$ for the sign identifiability, and then normalize the parameter vector as $\boldsymbol{\beta}^{**} = \boldsymbol{\beta}^*/\|\boldsymbol{\beta}^*\| = (1, \beta_2^*, \ldots, \beta_d^*)/(1 + \sum_{j=2}^d \beta_j^{*2})^{1/2}$, so that only $d-1$ parameters need to be estimated. The parameter vector $\boldsymbol{\alpha}$ can be reparameterized using the same way.

Model (2.2) covers many statistical models as special cases. If one or both of the unknown nonparametric functions are identity, (2.2) reduces to partially linear single-index models [9, 36] or traditional Shumway's hazard model. Note that the single-index hazard model (2.1) can be also viewed as a special case of (2.2). Specifically for our bankruptcy prediction, a composite index with both market and accounting based variables can be constructed under the single-index hazard model (2.1), while the double-index hazard model (2.2) can characterize group effect of the two types of financial variables separately through the two unknown link functions. Throughout this section, we replace the subscript $\{it\}$ with $\{i\}$ in order to simplify the notation, unless otherwise indicated.

2.3.2 Polynomial spline approximation

To estimate model (2.2), the most common approach is to use maximum likelihood estimation (MLE) by maximizing the quasi log-likelihood function [40]

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \left\{ y_i \eta_i - \log(1 + e^{\eta_i}) \right\}, \tag{2.3}$$

where y_i is the binary response, and $\eta_i = \phi_1\left(\mathbf{x}_i^T \boldsymbol{\alpha}\right) + \phi_2\left(\mathbf{z}_i^T \boldsymbol{\beta}\right)$, is a functional predictor term instead of the linear predictor term in traditional logit model. Clearly, the key to estimate the index hazard model (2.2) with MLE comprises two parts, (1) estimating the unknown univariate functions $\phi_1(\cdot)$ and $\phi_2(\cdot)$ nonparametrically, and (2) estimating the index coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ parametrically. These two parts of estimation can be achieved in two steps with an iterative algorithm, which will be discussed shortly in Section 2.3.3.

Typical methods for nonparametric regression include kernel smoothing [55] and spline approximations [22, 14]. Kernel smoothing is a local method that commonly applies Gaussian density as the local weight and obtains the estimates through weighted maximum likelihood estimation. As an alternative, spline regression essentially employs basis functions, the spline basis, and approximates the unknown function piece-wisely with a linear combination of the pre-specified basis functions. Higher polynomial order of the spline ensures smoothness at the piece-wise boundaries. In this paper, we adopt polynomial spline approximation to estimate the nonparametric component in model (2.2), due to its fast computation and many attractive statistical properties [22, 70, 66]. A computationally stable B-Spline basis functions [17, 10] are implemented with available statistical software packages. For illustration purpose, below we present the truncated power basis [66], which has little difference with B-Spline in terms of numerical performance. In particular, the unknown link function $\phi(\cdot)$

can be expressed as

$$\phi(u) \approx \gamma_0 + \gamma_1 u + \ldots + \gamma_p u^p + \sum_{k=1}^{K} \gamma_{p+k}(u - t_k)_+^p = \boldsymbol{\gamma}^T \mathbf{B}(u), \tag{2.4}$$

where $\mathbf{B}(u) = (1, u, \ldots, u^p, (u-t_1)_+^p, \ldots, (u-t_K)_+^p)^T$ is the truncated power basis of order p with K interior knots that take values at $t_1, \ldots, t_K$. $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \ldots, \gamma_{p+k})^T$ is the coefficient vector of spline basis. The truncation function $(u-t_k)_+^p = (u-t_k)^p$ if $u > t_k$ and 0 otherwise. A popular way to choose the knots is to place them at equally-spaced sample quantiles, such that each piece-wise interval has the same number of data points, which ensures the estimation stability. In practice, 2 to 4 knots are adequate in our application, while larger number of knots may be placed along with the roughness penalty. In our empirical study, the 2 equally-spaced quantile knots are determined through a simple grid search. For the spline order, we set $p = 3$, which is the commonly used cubic spline. The cubic spline has continuous second-order derivatives so that smoothness at boundaries is guaranteed.

2.3.3 Algorithm

By using (2.4) to approximate the unknown nonparametric functions $\phi_1(\cdot)$ and $\phi_2(\cdot)$, the systematic component η_i in likelihood function (2.3) can be written as

$$\eta_i = \boldsymbol{\gamma}_1^T \mathbf{B}_1\left(\mathbf{x}_i^T \boldsymbol{\alpha}\right) + \boldsymbol{\gamma}_2^T \mathbf{B}_2\left(\mathbf{z}_i^T \boldsymbol{\beta}\right). \tag{2.5}$$

Therefore, given the spline basis $\mathbf{B}_1(u)$ and $\mathbf{B}_2(u)$, i.e., given $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the optimization problem (2.3) is equivalent to the estimation of a traditional logit model with unknown

spline coefficients $\boldsymbol{\gamma}_1$ and $\boldsymbol{\gamma}_2$. For the case of single-index hazard model,

$$\hat{\boldsymbol{\gamma}} = \arg\max_{\gamma} \frac{1}{n} \sum_{i=1}^{n} \left[y_i \eta_i - \log(1 + e^{\eta_i}) \right], \tag{2.6}$$

while for the proposed double-index hazard model, $\hat{\boldsymbol{\gamma}}_1$ and $\hat{\boldsymbol{\gamma}}_2$ are obtained separately with back-fitting algorithm. Specifically, $\hat{\boldsymbol{\gamma}}_1$ is estimated by fixing $\hat{\phi}_2(\cdot)$ from last iteration, hence $\hat{\phi}_1(\cdot)$ is updated by plugging $\hat{\boldsymbol{\gamma}}_1$. Then we estimate $\hat{\boldsymbol{\gamma}}_2$ by fixing previously updated $\hat{\phi}_1(\cdot)$, hence $\hat{\phi}_2(\cdot)$ is estimated.

In next step, given the updated $\hat{\phi}_1(\cdot)$ and $\hat{\phi}_2(\cdot)$, the index coefficients $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ can be estimated with linear approximation on the link functions. We illustrate this approach by using single-index model for simplicity. That is

$$\phi(u) \approx \phi(u_0) + \phi'(u_0)(u - u_0), \tag{2.7}$$

where $\phi'(u_0)$ is first-order derivative of $\phi(u)$ evaluated at u_0, the estimated index term $\hat{\boldsymbol{\alpha}}^T \mathbf{x}$ from last iteration. Such linear approximation again reduces the optimization problem to traditional logit model estimation with unknown parameter $\boldsymbol{\alpha}$, where $\phi(u_0)$, $\phi'(u_0)$, and u_0 are all constant, while $u = \boldsymbol{\alpha}^T \mathbf{x}$ contains the unknown parameter $\boldsymbol{\alpha}$ as a linear function. Comparing to directly optimizing the objective function with respect to parameter $\boldsymbol{\alpha}$ using a nonlinear optimization algorithm, the linear approximation significantly reduces the computational cost, and standard software can be used. For double-index model, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are estimated separately as we shall specify different penalty levels for the two indices in the penalized estimation for the purpose of variable selection. More details will be discussed shortly in next section. Following we summarize our algorithm:

Step 0 Use Shumway's linear hazard model to obtain an initial estimates $\hat{\boldsymbol{\alpha}}^{(0)}, \hat{\boldsymbol{\beta}}^{(0)}$. One can also obtain the initial values by other estimators or use random starting values. Normalize $\hat{\boldsymbol{\alpha}}^{(0)}$ and $\hat{\boldsymbol{\beta}}^{(0)}$ separately, and multiply by sign of the first index element such that $\|\hat{\boldsymbol{\alpha}}^{(0)}\|_2 = \|\hat{\boldsymbol{\beta}}^{(0)}\|_2 = 1$, and the first element is positive.

Step 1 Given $\hat{\boldsymbol{\alpha}}^{(0)}$, fix $\phi_2(\mathbf{z}_i^T\boldsymbol{\beta}^{(0)})$ (set $\phi_2(\cdot) = 0$ at first step), and estimate the spline coefficient $\hat{\boldsymbol{\gamma}}_1$, where $\phi_1(\mathbf{x}_i^T\hat{\boldsymbol{\alpha}}^{(0)}) \approx \boldsymbol{\gamma}_1^T\mathbf{B}_1(\mathbf{x}_i^T\hat{\boldsymbol{\alpha}}^{(0)})$.

Step 2 Fix $\hat{\phi}_1(\mathbf{x}_i^T\boldsymbol{\alpha})$, and estimate the spline coefficients $\hat{\boldsymbol{\gamma}}_2$, where $\phi_2(\mathbf{z}_i^T\hat{\boldsymbol{\beta}}^{(0)}) \approx \boldsymbol{\gamma}_2^T\mathbf{B}_2(\mathbf{z}_i^T\hat{\boldsymbol{\beta}}^{(0)})$ similar as step 1 and hence $\hat{\phi}_2(\mathbf{z}_i^T\boldsymbol{\beta})$.

Step 3 Given $\hat{\phi}_1(\mathbf{x}_i^T\boldsymbol{\alpha})$ and $\hat{\phi}_2(\mathbf{z}_i^T\boldsymbol{\beta})$, we develop block-wise coordinate descent algorithm to first estimate $\hat{\boldsymbol{\alpha}}$ while $\hat{\phi}_2(\mathbf{z}_i^T\boldsymbol{\beta})$ is fixed. Then estimate $\hat{\boldsymbol{\beta}}$ by fixing $\hat{\phi}_1(\mathbf{x}_i^T\boldsymbol{\alpha})$.

Step 4 Repeat steps 1, 2, 3 until the parameters $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$ both converge.

This algorithm is computationally efficient and converges quickly, which is largely contributed by the coordinate descent (CD) algorithm in step 3 in which index coefficients are estimated. The idea of coordinate descent algorithm is to update a single parameter one at a time while the rest parameters are fixed. It has recently been well recognized and appreciated for its simplicity, speed and stability [61, 20, 21] in solving ℓ_1-regularization problem, i.e., LASSO [53], especially when predictor is in high dimensional space. Breheny and Huang [6] also shows that coordinate descent algorithm is significantly faster than other competing methods for regularized problem with nonconvex penalty functions such as SCAD [18] and MCP [67], which has been demonstrated to have attractive statistical properties.

Although above algorithm is proposed for estimating double-index hazard model (2.2), it can be simplified for the case of single-index hazard models (2.1). In particular, the

steps of estimating $\boldsymbol{\gamma}_2$ and $\boldsymbol{\beta}$ can be omitted. Our algorithm can be simply implemented with available statistical software such as `R` and Matlab. In our implementation, we use the function `bsplineS()` in `R` package "`fda`" to construct B-spline basis and its derivatives, and use standard function `glm()` to obtain the spline coefficients $\hat{\boldsymbol{\gamma}}$.

2.3.4 Penalized estimation for variable selection

A key research question in bankruptcy prediction is to identify important variables as bankruptcy predictors among many candidates. To address this question, we propose the penalized double-index hazard model, which is able to select important variables and estimate the model simultaneously.

Variable selection and dimension reduction are fundamental problems in statistical analysis. Traditional subset selection method through an exhaustive search suffers from computational infeasibility if there are many predictors. Regularized methods such as LASSO [53], Adaptive-Lasso [72], SCAD [18], and MCP [67], have been increasingly popular in modern data analytics due to its fast computational speed and statistical consistency. Such regularized method is able to select important variables and estimate the coefficient simultaneously by adding a certain type of penalty function for parameters, hence called penalized estimation. Statistical consistency and asymptotic normality has been proved for most penalized estimator under appropriate conditions. SCAD is among one of the most widely used penalty functions for penalized estimation. Its nonconcavity shape ensures the variable selection consistency or the so-called oracle property, i.e. with probability attending to 1,

true model can be identified. The SCAD penalty function is defined as

$$p_\lambda(\theta) = \begin{cases} \lambda|\theta|; & \text{if } |\theta| \leq \lambda \\ -\frac{|\theta|^2 - 2a\lambda|\theta| + \lambda^2}{2(a-1)}; & \text{if } \lambda < |\theta| \leq a\lambda \\ \frac{(a+1)\lambda^2}{2}; & \text{if } |\theta| > a\lambda, \end{cases} \tag{2.8}$$

where the tuning parameter $\lambda > 0$ can be chosen by cross-validation or a Bayesian information criteria (BIC) [36], and $a = 3.7$ has been suggested according to Fan and Li [18].

To select bankruptcy predictors and estimate their coefficients for constructing the composite indices of the proposed double-index hazard model (2.2), we attach the SCAD penalty function (2.8) to the quasi-likelihood function $L(\boldsymbol{\alpha}, \boldsymbol{\beta})$ that is defined in (2.3), and maximize the following penalized quasi-likelihood function,

$$Q_n(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} L(\boldsymbol{\alpha}, \boldsymbol{\beta}) - \sum_{j=1}^{s} p_{\lambda_\alpha}(\alpha_j) - \sum_{l=1}^{d} p_{\lambda_\beta}(\beta_l), \tag{2.9}$$

where $p_{\lambda_\alpha}(\cdot)$ and $p_{\lambda_\beta}(\cdot)$ are penalty functions defined in (2.8) with potentially different regularization parameters λ_α and λ_β, which controls the shrinkage level. A larger value of λ results in more sparse model. In other words, fewer predictors are selected. Specifying two penalty functions with different regularization parameters λ_α and λ_β for two indices provides following advantages to the proposed double-index model. First, it is flexible enough for practitioners to choose different penalization levels for market and accounting based variables with his/her expertise and preference. Second, it is convenient and straightforward to develop a block-wise coordinate descent algorithm enabling fast and stable computation.

As discussed above, appropriate choice of regularization parameter is essential for regularized method. Large value of λ may lead to high sparse model where only few variables are

selected, while small λ often yields complex model by including some unimportant variables. Different methods for tuning parameter selection have been proposed in previous studies (e.g. [57, 58, 68]). A Bayesian Information Criteria (BIC) based tuning parameter selector proposed by Wang et al. [57] has been shown to satisfy model selection consistency [48, 63]. In particular, the optimal value of λ_{BIC} is chosen by minimizing

$$\text{BIC}_\lambda = -2\log(\hat{L}) + \frac{\log(n)}{n}\text{DF}_\lambda, \tag{2.10}$$

where $\hat{L}$ the likelihood function defined in (2.3), DF_λ is the number of effective parameters defined in Fan and Li [18] with the form $\text{DF}_\lambda = \text{tr}\left\{X\left(X'X + n\Sigma_\lambda\right)^{-1}X'\right\}$, and $\Sigma_\lambda = \text{diag}\left\{p'_\lambda(|\hat{\alpha}_1|)/|\hat{\alpha}_1|, \ldots, p'_\lambda(|\hat{\alpha}_s|)/|\hat{\alpha}_s|\right\}$. For our proposed double-index hazard model, we choose λ_a and λ_b separately by minimizing (2.10) through a two-dimensional grid search.

2.3.5 Simulation study

In order to demonstrate the performance of our proposed penalized double-index model, we conduct Monte Carlo simulations. Consider model (2.2), we set the true index parameter vector $\boldsymbol{\alpha}_0 = (2, -1.5, 1, 0, 0, 0, 0, 0, 0)$ and $\boldsymbol{\beta}_0 = (1, -1, 1.5, 0, 0, 0, 0)$. The covariate $\mathbf{x}_i$ is generated from multivariate Gaussian distribution with mean zero and identity covariance matrix. For covariate $\mathbf{z}_i$, we let $z_{i1} \sim \text{N}(0, 1)$, $z_{i2} \sim \text{Bernoulli}(0.2)$, $z_{i3} \sim \text{U}(-2, 2)$, and $\mathbf{z}_{i,4-7} \sim \text{MN}(\mathbf{0}, \mathbf{I}_{4\times 4})$. Let $\phi_1(\mathbf{x}_i^T\boldsymbol{\alpha}) = 2\sin(\pi(\mathbf{x}_i^T\boldsymbol{\alpha} + 2)/3)$, and $\phi_2(\mathbf{z}_i^T\boldsymbol{\beta}) = -\sin(\pi(\mathbf{z}_i^T\boldsymbol{\beta} - 1)/1.5) + \mathbf{z}_i^T\boldsymbol{\beta}$. Then, the binary response variable Y_i can be simulated from Bernoulli distribution with probability of $1/(1+\exp(-\phi_1-\phi_2))$. We run 1000 replications with sample size n being 500, 1000, and 2000.

We report the average of nonzero coefficient estimates over 1000 Monte Carlo simulations in Table 2.4. Standard errors are also reported. To evaluate how the unknown link functions $\phi_1(\cdot)$ and $\phi_2(\cdot)$ are estimated, we report the averaged ℓ_2-norm across all simulated samples. In addition, the average false negative rate (FNR) and false positive rate (FPR) of both two index coefficients are reported in the Table 2.5 in order to assess the variable selection accuracy. FNR is defined as the proportion of zero coefficient estimates whose corresponding true coefficients are nonzero, i.e., $\#\{j : \hat{\beta}_j = 0, \beta_{0j} \neq 0\}/\#\{j : \beta_{0j} \neq 0\}$. FPR is defined as the proportion of nonzero coefficient estimates whose corresponding true coefficients are zero, i.e., $\#\{j : \hat{\beta}_j \neq 0, \beta_{0j} = 0\}/\#\{j : \beta_{0j} = 0\}$. We can see from Table 2.4 and Table 2.5, our proposed double-index model performs well in terms of estimation consistency and variable selection accuracy.

Table 2.4: Monte Carlo simulation for the proposed double-index hazard model. Avg.Est. is the averaged coefficient estimates across 1000 simulation samples. S.E. is the standard error of the coefficient estimates. $\|\hat{\phi} - \phi_0\|_2$ is ℓ_2 distance between estimated and the underlying true unknown nonparametric functions.

n		α_1	α_2	α_3	β_1	β_2	β_3
	True	0.743	-0.557	0.371	0.485	-0.485	0.728
	Avg.Est.	0.724	-0.557	0.375	0.467	-0.489	0.710
500	S.E.	0.078	0.078	0.085	0.072	0.157	0.071
	$\|\hat{\phi} - \phi_0\|_2$	0.394			0.362		
	Avg.Est.	0.740	-0.554	0.371	0.481	-0.484	0.723
1000	S.E.	0.041	0.051	0.055	0.044	0.086	0.044
	$\|\hat{\phi} - \phi_0\|_2$	0.235			0.250		
	Avg.Est.	0.742	-0.556	0.370	0.485	-0.482	0.725
2000	S.E.	0.027	0.033	0.038	0.031	0.062	0.031
	$\|\hat{\phi} - \phi_0\|_2$	0.199			0.192		

Table 2.5: False negative rate (FNR) and False positive rate (FPR) for variable selection for two indices. FNR is defined as the proportion of zero coefficient estimates whose corresponding true coefficients are nonzero, i.e., $\#\{j : \hat{\theta}_j = 0, \theta_{0j} \neq 0\}/\#\{j : \theta_{0j} \neq 0\}$. FPR is defined as the proportion of nonzero coefficient estimates whose corresponding true coefficients are zero, i.e., $\#\{j : \hat{\theta}_j \neq 0, \theta_{0j} = 0\}/\#\{j : \theta_{0j} = 0\}$.

n	$\boldsymbol{\alpha}$		$\boldsymbol{\beta}$	
	FNR	FPR	FNR	FPR
500	0.033	0.019	0.049	0.026
1000	0.003	0.000	0.010	0.005
2000	0.000	0.000	0.001	0.000

2.4 Empirical Results

We apply the proposed penalized double-index hazard model to the annual manufacturing database described in Section 2.2. In addition, the single-index model (2.1) is also implemented as a special case of model (2.2). We first build one-year ahead prediction model and assess the performance in Section 2.4.1. Two-year and three-year ahead prediction models are investigated in Section 2.4.2. For each bankruptcy prediction horizon, we align the bankruptcy indicator with predictors by 1, 2, and 3 lags for each firm, such that the predictive information is of the time prior to the bankruptcy event. As a result, the total number of bankruptcy in each sample is 490, 419, and 346 respectively. Such difference is due to that some companies filed bankruptcy within one, two, or three years after they went public, so that these firms would be removed after the lags. For comparison, in addition to our proposed index hazard models, we fit the popular Shumway's hazard models with different prespecified predictors as a benchmark: (1) variables constructed and used in Campbell et al. [8] (CHS), (2) financial ratios used in Altman's Z-score [2], and (3) automatic variable selection with LASSO for bankruptcy prediction [52] .

2.4.1 One-year ahead forecast

By applying our proposed semiparametric index hazard model, we find interesting nonlinearity between bankruptcy predictors and default risk through our empirical study. It is important to provide empirical evidence of nonlinearity between predictive variables and default risk before we show the model results. One way to explore the relationship between binary response and continuous covariate is to construct contingency table with the continuous variable being categorized based on quantiles. Graphically, a scatter plot between the bin average of the continuous variable against event (bankruptcy) occurrence frequency or proportion in each bin can serve as an exploratory analysis to visualize relationship between binary response and continuous covariate. Figure 2.2 shows the scatter plot of bankruptcy frequency against different predictors that have been frequently used in the literature of bankruptcy prediction. These predictors are also selected later in our empirical results (see Table 2.6) of one-year ahead forecast model. Smooth curve is fitted for each scatter plot for better view of the nonlinearity. Not surprisingly, nonlinearity is clearly shown, while the majority portion for each scatter plot appears to be monotone relationship. Intuitively, if individual variable has nonlinear relationship to the response variable, their linear combination is not guaranteed to be linearly related to the response. This coincides with our proposed index hazard model, which is able to capture the nonlinearity between the index, a linear combination of individual predictors, and default risk.

Table 2.6 shows the coefficient estimates and standard errors (reported in parenthesis) of one-year ahead forecast models based on the full sample period, i.e., 1980-2015. For the double-index hazard model, five market based variables: SIGMA, EXRET, NIMTA, LTMTA and LOG(PRICE), are selected to construct the market index, and three accounting based variables: LTAT, NIAT, EBITAT, are identified as the important variables to construct the

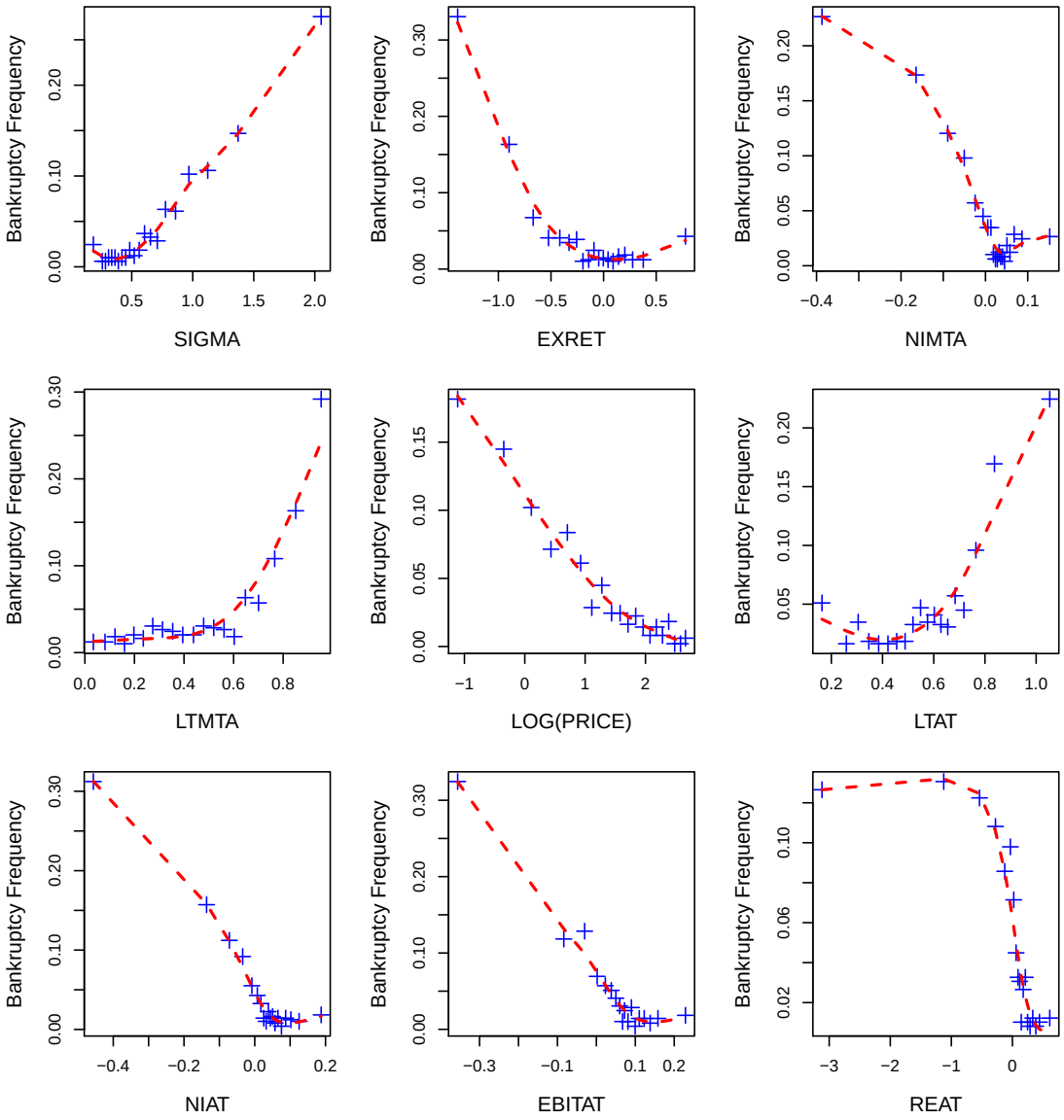

Figure 2.2: Bankruptcy frequency across quantile bins of individual predictors.

accounting index. For the single-index hazard model, selected bankruptcy predictors are consistent with the double-index model despite the model structural difference. LASSO selects the same market based variables while NIAT is excluded for accounting index. More importantly, we notice that the signs of coefficient estimates are consistent across different

modeling approaches. This implies that the interpretation of individual predictors are the same in a qualitative manner. In parenthesis we also report standard errors of coefficient estimates for which the index models and LASSO are obtained through bootstrapping with 500 resampling, while the standard formula is used for CHS and Altman. It is worth noting that, for double-index hazard model, one could adjust the tuning parameter for each index based on the analyst's intuition, so that the number of selected variables could be subjective yet flexible. The optimal values of λ's in our empirical results is based on the BIC criteria defined in (2.10) with a two-dimensional grid search method.

Figure 2.3 is the estimated nonparametric functions. From left to right, the first plot is the estimated link function of single-index hazard model, and the second and third plots correspond to the estimated functions of market and accounting indices in the double-index hazard model. Nonlinearity is obvious and majority portion of the functions are monotonic. The shape of the curve qualitatively agrees with the fitted curve of individual predictors shown in Figure 2.2 after applying the sign of estimated index coefficients. The single index function has a similar shape as the market index function estimated by the double-index model, while the accounting index function is nearly linear. This is because that for one-year ahead forecast model, more market than accounting based variables are selected as a result of data-driven automatic variable selection.

To evaluate the performance of our bankruptcy prediction models, we report three popular statistical measures: Pseudo-R^2, Area Under the Curve (AUC), and Hosmer-Lemeshow goodness-of-fit test statistic [31], which are widely used for assessing binary classification models. Pseudo-R^2 is defined as $1 - L_1/L_0$, where L_1 and L_0 are residual deviance from fitted model and null model which is estimated only with intercept. Larger value of pseudo-R^2 means better fitting. AUC is a commonly used prediction accuracy measure for binary

Table 2.6: Coefficient estimates for different models under one-year ahead forecasting horizon. The time period of the training dataset is 1980-2015. Panel A and B separates market and accounting based variables. The data used for penalized methods is scaled to (0, 1) range. Standard errors are reported in parenthesis with italic font. For index model, the standard error is obtained by 500 bootstrapped samples.

	Double-Index	Single-Index	Lasso	CHS	Altman
Intercept			-3.181 (*0.329*)	-6.381 (*0.484*)	-4.902 (*0.094*)
			Panel A: Market Variable		
SIGMA	0.137 (*0.042*)	0.060 (*0.022*)	0.471 (*0.236*)	0.168 (*0.086*)	
EXRET	-0.640 (*0.044*)	-0.568 (*0.039*)	-3.950 (*0.477*)	-0.768 (*0.096*)	
NIMTA	-0.618 (*0.058*)	-0.188 (*0.039*)	-0.085 (*0.319*)	-0.948 (*0.230*)	
LTMTA	0.199 (*0.034*)	0.175 (*0.023*)	2.054 (*0.278*)	2.564 (*0.228*)	
LOG(PRICE)	-0.387 (*0.032*)	-0.353 (*0.033*)	-2.452 (*0.277*)	-0.543 (*0.061*)	
CASHMTA				-1.025 (*0.363*)	
SIZE				-0.029 (*0.035*)	
MBE				0.072 (*0.026*)	
LCTMTA					
MVEF					-0.419 (*0.266*)
			Panel B: Accounting Variable		
LTAT	0.437 (*0.032*)	0.358 (*0.023*)	1.657 (*0.318*)		
NIAT	0.673 (*0.044*)	0.272 (*0.045*)			
EBITAT	-0.596 (*0.042*)	-0.501 (*0.052*)	-1.954 (*0.511*)		-2.725 (*0.226*)
REAT					0.069 (*0.046*)
WCAPAT					-1.669 (*0.193*)
SALEAT					0.423 (*0.066*)
LCTAT					
RELCT					
LCTSALE					
LOG(SALE)					
CHAT					
ACTLCT					
LCTLT					

classification problem. It evaluates the models discriminative power, where a value close to 1 indicates strong discriminative ability. The formal Hosmer-Lemeshow goodness-of-fit test is another popular statistical test for calibration performance. A p-value smaller than 0.05 indicates that the model is lack of fitting. Hosmer-Lemeshow test statistics is rarely reported in the bankruptcy literature because it is frequently rejected in practice [15]. In addition, we also report the cumulative Decile Ranking Tables that has been widely used

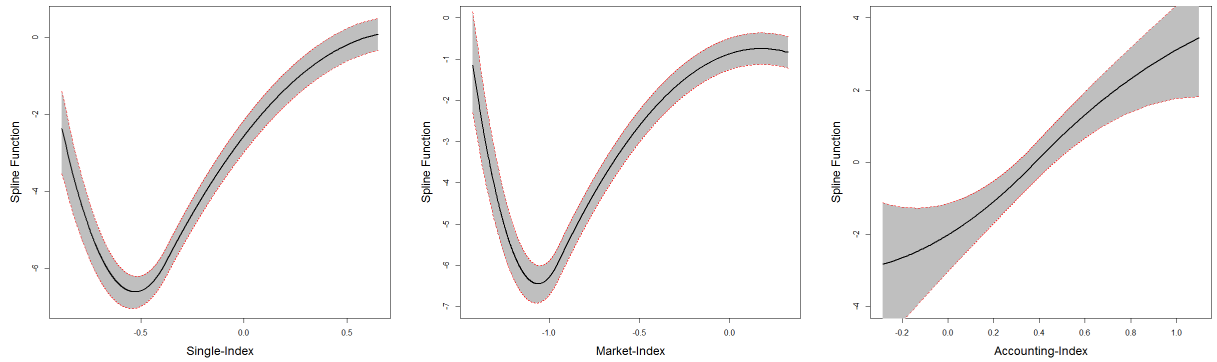

Figure 2.3: From left to right, the plots are estimated unknown link function of single-index (model (2.1)), market-index and accounting-index (model (2.2)) for one-year ahead forecasting horizon. The training dataset is based on full sample, i.e., the time period is 1980-2015.

in bankruptcy prediction literatures [50, 12] to evaluate the models discrimination and calibration power. In particular, the common decile ranking table is generated by ranking the predicted probability of default in deciles. The top decile contains the firms with highest predicted bankruptcy probability. Within each decile, we calculate the proportion of the bankruptcies that are captured in that decile over the total number of observed bankruptcies. Higher proportion in the top decile is more desirable, which implies a model with higher prediction accuracy. By cumulating, we obtain the cumulative decile ranking table. These measures are evaluated for both in and out-of-sample dimensions. Table 2.7 shows the in-sample results based on the full sample period 1980-2015, and Table 2.8 shows both training and testing performance with different periods (train: 1980-2007, 1980-2003, and 1980-1997; test: 2008-2015, 2004-2015, and 1998-2015) for the purpose of robustness check.

According to Table 2.7 and 2.8, our proposed double-index model uniformly dominates other approaches in terms of all assessments of both in-sample and out-of-sample performance. It is worth noting that both the double-index and single-index hazard model easily

Table 2.7: Decile ranking table, area under the curve (AUC), and the p-value of the Hosmer-Lemeshow goodness-of-fit test for different models under one-year ahead forecasting horizon. The time period of the training dataset is 1980-2015.

	Double-Index	Single-Index	Lasso	CHS	Altman
90-100%	0.706	0.702	0.692	0.686	0.408
80-100%	0.847	0.847	0.841	0.824	0.573
70-100%	0.908	0.902	0.898	0.900	0.657
60-100%	0.931	0.929	0.924	0.922	0.712
50-100%	0.957	0.939	0.941	0.941	0.765
0-100%	1.000	1.000	1.000	1.000	1.000
Pseudo-R^2	0.251	0.248	0.241	0.231	0.067
AUC	0.893	0.890	0.887	0.884	0.733
H-L pval	0.380	0.407	0.009	0.009	0.000

pass the Hosmer-Lemeshow goodness-of-fit test (H-L pval > 0.05), while the other models do not pass for most training and testing periods. For out-of-sample period 2008-2015, the Pseudo R^2 is 0.349 for the proposed double-index model comparing to 0.287 from the best linear model (LASSO), which delivers 22% improvements. The relatively underperformed model with Altman's variable again shows evidence that market based variables carry much more predictability than accounting based variables for one-year-ahead forecast horizon.

2.4.2 Different forecasting horizon

We further investigate how the predictors vary across different forecasting horizons. We report the selected variables of double and single-index hazard models in Table 2.9, and the model estimation in Appendix B. Meanwhile, the variables used in Campbell et al. [8] and Altman [2] are reported for reference. An interesting yet intuitive finding is that more market based variables play role in short forecasting horizon, while more accounting based predictors are more likely to be selected for longer forecasting horizons. This is intuitive because the change of market variables are very dynamic and can only reflect investors fear

Table 2.8: In-sample and out-of-sample performance for the one-year ahead forecast model. The criteria consist of decile ranking table, Pseudo-R^2, area under the curve (AUC), and the p-value of the Hosmer-Lemeshow goodness-of-fit test. The time period of training sample are 1980-2007, 1980-2003 and 1980-1997.

	Double-index	Single-index	Lasso	CHS	Altman	Double-index	Index-index	Lasso	CHS	Altman
	Panel A					*Panel B*				
	In-sample (1980-2007)					*Out-of-sample (2008-2015)*				
90-100%	0.688	0.686	0.684	0.675	0.416	0.829	0.805	0.805	0.829	0.244
80-100%	0.835	0.831	0.826	0.813	0.568	0.976	0.976	0.951	0.951	0.341
70-100%	0.906	0.893	0.886	0.884	0.659	0.976	0.976	0.976	0.976	0.512
60-100%	0.924	0.915	0.920	0.915	0.713	0.976	0.976	0.976	0.976	0.610
50-100%	0.944	0.938	0.938	0.938	0.768	0.976	0.976	0.976	0.976	0.683
0-100%	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
AUC	0.888	0.884	0.881	0.877	0.735	0.947	0.942	0.939	0.938	0.651
H-L pval	0.322	0.340	0.077	0.077	0.002	0.548	0.344	0.014	0.009	0.174
Psuedo R^2	0.243	0.242	0.235	0.227	0.070	0.349	0.330	0.287	0.274	0.037
	In-sample (1980-2003)					*Out-of-sample (2004-2015)*				
90-100%	0.692	0.682	0.673	0.664	0.419	0.809	0.765	0.750	0.765	0.309
80-100%	0.829	0.834	0.818	0.801	0.566	0.926	0.926	0.956	0.956	0.397
70-100%	0.900	0.891	0.884	0.879	0.656	0.971	0.956	0.956	0.956	0.515
60-100%	0.922	0.910	0.912	0.912	0.718	0.985	0.985	0.985	0.971	0.618
50-100%	0.938	0.931	0.934	0.938	0.775	1.000	0.985	0.985	0.971	0.691
0-100%	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
AUC	0.885	0.880	0.877	0.873	0.734	0.936	0.934	0.932	0.929	0.672
H-L pval	0.257	0.315	0.075	0.065	0.011	0.527	0.281	0.003	0.004	0.003
Psuedo R^2	0.241	0.239	0.233	0.224	0.072	0.316	0.306	0.282	0.273	0.043
	In-sample (1980-1997)					*Out-of-sample (1998-2015)*				
90-100%	0.713	0.697	0.697	0.684	0.498	0.689	0.705	0.705	0.678	0.268
80-100%	0.850	0.837	0.827	0.827	0.632	0.869	0.858	0.863	0.842	0.432
70-100%	0.909	0.902	0.889	0.886	0.691	0.923	0.907	0.907	0.918	0.552
60-100%	0.932	0.912	0.912	0.925	0.749	0.951	0.945	0.945	0.940	0.645
50-100%	0.935	0.932	0.935	0.941	0.795	0.973	0.973	0.962	0.962	0.721
0-100%	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
AUC	0.890	0.885	0.881	0.881	0.765	0.904	0.903	0.898	0.894	0.664
H-L pval	0.457	0.523	0.061	0.108	0.008	0.390	0.163	0.037	0.010	0.059
Psuedo R^2	0.256	0.253	0.247	0.239	0.097	0.240	0.236	0.227	0.213	0.015

and favor in short term. In other words, the market variables are more informative for short-term investment, while investors often study firm's accounting ratios to make decision of long-term investment.

Similar to one-year ahead forecast model, we also report the same criteria in Table 2.10 for assessing the models' out-of-sample performance (in-sample performance is also available upon request). We notice that the out-of-sample Psuedo R^2's of the proposed double-index

Table 2.9: Variable selection of different models across different forecasting horizons.

	Double-Index			Single-Index			Lasso			CHS	Altman
	One	Two	Three	One	Two	Three	One	Two	Three		
Panel A: Market Variable											
SIGMA	X	X	X	X	X	X	X	X	X	X	
EXRET	X	X	X	X	X	X	X	X	X	X	
NIMTA	X	X	X	X	X	X	X			X	
LOG(PRICE)	X	X		X			X	X	X	X	
LTMTA	X			X			X	X		X	
CASHMTA										X	
SIZE								X	X	X	
MBE										X	
LCTMTA											
MVEF											X
Panel B: Accounting Variable											
LTAT	X	X	X	X	X	X	X	X	X		
NIAT	X	X	X	X	X	X					
EBITAT	X	X	X	X	X	X	X	X	X		X
LOG(SALE)		X	X		X	X		X	X		
REAT		X	X		X	X					X
RELCT			X			X			X		
WCAPAT											X
SALEAT											X
LCTAT											
LCTSALE								X			
CHAT											
ACTLCT											
LCTLT											

model are dominating across all different testing periods for two-year forecasting horizon. The Hosmer-Lemeshow goodness-of-fit test is again easily passed for the semiparametric models in most testing periods. These results provide strong evidences that our proposed double-index, as well as single-index hazard models dominate other approaches that are based on linear model framework for two-year ahead prediction. For three-year ahead forecast model, Shumway's hazard model performs better than semiparametric models. However, we note that even the best model for three-year ahead forecasting still performs poorly. This is true as the predictability of any information is weakened for longer horizon. Overall, short term prediction is more accurate than long term horizon.

Table 2.10: Out-of-sample performance for the two and three-year ahead forecast models. *Panel A (left)* is for two-year ahead and *Panel B (right)* is for three-year ahead forecast model. The criteria consist of decile ranking table, Pseudo-R^2, area under the curve (AUC), and the p-value of the Hosmer-Lemeshow goodness-of-fit test. The time period of training sample are 1980-2007, 1980-2003 and 1980-1997, respectively

	Double-index	Single-index	Lasso	CHS	Altman	Double-index	Index-index	Lasso	CHS	Altman
	Panel A: Two-year ahead forecast model					*Panel B: Three-year ahead forecast model*				
	Out-of-sample (2008-2015)					*Out-of-sample (2008-2015)*				
90-100%	0.435	0.391	0.391	0.435	0.217	0.389	0.444	0.278	0.222	0.167
80-100%	0.826	0.783	0.826	0.739	0.478	0.556	0.556	0.556	0.444	0.222
70-100%	0.870	0.870	0.870	0.826	0.522	0.667	0.667	0.778	0.722	0.222
60-100%	0.913	0.957	0.913	0.870	0.696	0.778	0.722	0.833	0.778	0.389
50-100%	1.000	0.957	0.957	0.913	0.739	0.833	0.889	0.944	0.944	0.500
0-100%	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
AUC	0.855	0.843	0.851	0.822	0.676	0.727	0.745	0.790	0.765	0.443
H-L pval	0.260	0.068	0.006	0.003	0.004	0.249	0.000	0.069	0.016	0.010
Psuedo R^2	0.179	0.147	0.122	0.083	0.060	0.082	0.078	0.077	0.042	0.003
	Out-of-sample (2004-2015)					*Out-of-sample (2004-2015)*				
90-100%	0.538	0.462	0.519	0.538	0.327	0.297	0.324	0.270	0.243	0.270
80-100%	0.750	0.788	0.827	0.750	0.481	0.541	0.541	0.595	0.486	0.351
70-100%	0.885	0.827	0.885	0.846	0.654	0.622	0.676	0.838	0.757	0.432
60-100%	0.942	0.904	0.923	0.904	0.731	0.838	0.757	0.892	0.892	0.459
50-100%	0.962	0.942	0.923	0.904	0.788	0.865	0.919	0.946	0.919	0.622
0-100%	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
AUC	0.869	0.841	0.852	0.835	0.721	0.740	0.759	0.800	0.785	0.577
H-L pval	0.422	0.001	0.009	0.016	0.000	0.002	0.000	0.002	0.001	0.003
Psuedo R^2	0.184	0.137	0.152	0.123	0.046	0.071	0.077	0.093	0.077	0.017
	Out-of-sample (1998-2015)					*Out-of-sample (1998-2015)*				
90-100%	0.490	0.443	0.443	0.409	0.208	0.351	0.315	0.270	0.279	0.225
80-100%	0.691	0.698	0.664	0.617	0.443	0.568	0.577	0.459	0.486	0.360
70-100%	0.805	0.819	0.805	0.752	0.591	0.757	0.712	0.766	0.640	0.514
60-100%	0.899	0.879	0.866	0.839	0.651	0.847	0.829	0.865	0.856	0.658
50-100%	0.933	0.919	0.913	0.866	0.711	0.910	0.937	0.946	0.910	0.730
0-100%	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
AUC	0.830	0.826	0.811	0.791	0.674	0.789	0.781	0.779	0.765	0.656
H-L pval	0.015	0.332	0.025	0.006	0.000	0.000	0.000	0.000	0.000	0.005
Psuedo R^2	0.123	0.127	0.101	0.075	0.018	0.071	0.066	0.058	0.045	0.011

Bibliography

[1] Andreas Alfons, Christophe Croux, and Sarah Gelper. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *Annals of Applied Statistics*, 7(1):226–248, 2013.

[2] Edward I Altman. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, 23(4):589–609, 1968.

[3] William H Beaver. Financial ratios as predictors of failure. *Journal of Accounting Research*, pages 71–111, 1966.

[4] William H Beaver, Maureen F McNichols, and Jung-Wu Rhie. Have financial statements become less informative? evidence from the ability of financial ratios to predict bankruptcy. *Review of Accounting studies*, 10(1):93–122, 2005.

[5] Sreedhar T Bharath and Tyler Shumway. Forecasting default with the merton distance to default model. *Review of Financial Studies*, 21(3):1339–1369, 2008.

[6] Patrick Breheny and Jian Huang. Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The annals of applied statistics*, 5(1):232, 2011.

[7] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[8] John Y Campbell, Jens Hilscher, and Jan Szilagyi. In search of distress risk. *The Journal of Finance*, 63(6):2899–2939, 2008.

[9] Raymond J Carroll, Jianqing Fan, Irene Gijbels, and Matt P Wand. Generalized partially linear single-index models. *Journal of the American Statistical Association*, 92 (438):477–489, 1997.

[10] Edwin Catmull and James Clark. Recursively generated b-spline surfaces on arbitrary topological meshes. *Computer-aided design*, 10(6):350–355, 1978.

[11] Probal Chaudhuri, Kjell Doksum, Alexander Samarov, et al. On average derivative quantile regression. *Annals of Statistics*, 25(2):715–744, 1997.

[12] Sudheer Chava and Robert A Jarrow. Bankruptcy prediction with industry effects. *Review of Finance*, 8(4):537–569, 2004.

[13] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[14] Carl De Boor, Carl De Boor, Etats-Unis Mathématicien, Carl De Boor, and Carl De Boor. *A practical guide to splines*, volume 27. Springer-Verlag New York, 1978.

[15] A Adam Ding, Shaonan Tian, Yan Yu, and Hui Guo. A class of discrete transformation survival models with application to default probability prediction. *Journal of the American Statistical Association*, 107(499):990–1003, 2012.

[16] David L Donoho and Peter J Huber. The notion of breakdown point. *A festschrift for Erich L. Lehmann*, pages 157–184, 1983.

[17] Paul HC Eilers and Brian D Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, pages 89–102, 1996.

[18] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

[19] Jianqing Fan, Quefeng Li, and Yuyan Wang. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):247–265, 2017.

[20] Jerome Friedman, Trevor Hastie, Holger Hofling, and Robert Tibshirani. Pathwise coordinate optimization. *Annals of Applied Statistics*, 1:302–332, 2007.

[21] Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[22] Peter J Green and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press, 1993.

[23] Ali S. Hadi and Alberto Luceno. Maximum trimmed likelihood estimators: a unified approach, examples, and algorithms. *Computational Statistics & Data Analysis*, 25(3): 251–272, 1997.

[24] Frank R. Hampel. A general qualitative definition of robustness. *Annals of Mathematical Statistics*, 42(6):1887–1896, 12 1971.

[25] Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. Wiley, first edition, 1986.

[26] Frank Rudolf Hampel. *Contributions to the theory of robust estimation*. University of California, 1968.

[27] Wolfgang Härdle and Enno Mammen. Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, pages 1926–1947, 1993.

[28] Wolfgang Härdle, Peter Hall, Hidehiko Ichimura, et al. Optimal smoothing in single-index models. *The Annals of Statistics*, 21(1):157–178, 1993.

[29] Wolfgang Härdle, Yuh-Jye Lee, Dorothea Schäfer, and Yi-Ren Yeh. Variable selection and oversampling in the use of smooth support vector machines for predicting the default risk of companies. *Journal of Forecasting*, 28(6):512–534, 2009.

[30] Joseph L Hodges Jr. Efficiency in normal samples and tolerance of extreme values for some estimates of location. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 163–186, 1967.

[31] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[32] Marian Hristache, Anatoli Juditsky, and Vladimir Spokoiny. Direct estimation of the index coefficient in a single-index model. *The Annals of Statistics*, pages 595–623, 2001.

[33] Peter J. Huber and Elvezio M. Ronchetti. *Robust Statistics*. Wiley, second edition, 2009.

[34] Hong Lan, Meng Chen, Jessica B Flowers, Brian S Yandell, Donnie S Stapleton, Christine M Mata, Eric Ton-Keen Mui, Matthew T Flowers, Kathryn L Schueler, Kenneth F Manly, et al. Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genet*, 2(1):e6, 2006.

[35] Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media, 2013.

[36] Hua Liang, Xiang Liu, Runze Li, and Chih-Ling Tsai. Estimation and testing for partially linear single-index models. *Ann. Statist.*, 38(6):3811–3836, 12 2010.

[37] Aurélie C Lozano, Nicolai Meinshausen, Eunho Yang, et al. Minimum distance lasso for robust high-dimensional regression. *Electronic Journal of Statistics*, 10(1):1296–1340, 2016.

[38] Ricardo Maronna, Doug Martin, and Victor Yohai. *Robust Statistics Theory and Methods*. Wiley, 2006.

[39] Pascal Massart. About the constants in talagrand's concentration inequalities for empirical processes. *Annals of Probability*, 28(2):863–884, 2000.

[40] Peter McCullagh and John A Nelder. *Generalized linear models*, volume 37. CRC press, 1989.

[41] Sahand N. Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science*, 27(4):538–557, 11 2012.

[42] James A Ohlson. Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, pages 109–131, 1980.

[43] Jean D Opsomer and David Ruppert. A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association*, 93(442): 605–619, 1998.

[44] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, 2010.

[45] Peter J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

[46] Peter J Rousseeuw and Christophe Croux. Alternatives to the median absolute deviation. *Journal of the American Statistical Association*, 88(424):1273–1283, 1993.

[47] David W Scott. Parametric statistical modeling by minimum integrated square error. *Technometrics*, 43(3):274–285, 2001.

[48] Jun Shao. An asymptotic theory for linear model selection. *Statistica Sinica*, pages 221–242, 1997.

[49] Georgy Shevlyakov, Stephan Morgenthaler, and Alexander Shurygin. Redescending m-estimators. *Journal of Statistical Planning and Inference*, 138(10):2906–2917, 2008.

[50] Tyler Shumway. Forecasting bankruptcy more accurately: A simple hazard model*. *The Journal of Business*, 74(1):101–124, 2001.

[51] Qifan Song and Faming Liang. High-dimensional variable selection with reciprocal l 1-regularization. *Journal of the American Statistical Association*, 110(512):1607–1620, 2015.

[52] Shaonan Tian, Yan Yu, and Hui Guo. Variable selection and corporate bankruptcy forecasts. *Journal of Banking & Finance*, 52:89–100, 2015.

[53] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[54] A. W. van der Vaart. *Asymptotics Statistics*. Cambridge University Press, 1998.

[55] Matt P Wand and M Chris Jones. *Kernel smoothing*. CRC Press, 1994.

[56] Hansheng Wang, Guodong Li, and Guohua Jiang. Robust regression shrinkage and consistent variable selection through the LAD-lasso. *Journal of Business & Economic Statistics*, 25:347–355, 2007.

[57] Hansheng Wang, Runze Li, and Chih-Ling Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568, 2007.

[58] Hansheng Wang, Bo Li, and Chenlei Leng. Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(3):671–683, 2009.

[59] Lie Wang. The l1 penalized LAD estimator for high dimensional linear regression. *Journal of Multivariate Analysis*, 120:135–151, 2013.

[60] Xueqin Wang, Yunlu Jiang, Mian Huang, and Heping Zhang. Robust variable selection with exponential squared loss. *Journal of the American Statistical Association*, 108 (502):632–643, 2013.

[61] Tong Tong Wu and Kenneth Lange. Coordinate descent algorithms for lasso penalized regressoon. *Annals of Applied Statistics*, 2:224–244, 2008.

[62] Tracy Z Wu, Keming Yu, and Yan Yu. Single-index quantile regression. *Journal of Multivariate Analysis*, 101(7):1607–1621, 2010.

[63] Yuhong Yang. Can the strengths of aic and bic be shared? a conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

[64] Victor J. Yohai. High breakdown-point and high efficiency robust estimates for regression. *Annals of Statistics*, 15(2):642–656, 06 1987.

[65] Keming Yu and Zudi Lu. Local linear additive quantile regression. *Scandinavian Journal of Statistics*, 31(3):333–346, 2004.

[66] Yan Yu and David Ruppert. Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97(460):1042–1054, 2002.

[67] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 04 2010.

[68] Yiyun Zhang, Runze Li, and Chih-Ling Tsai. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association*, 105 (489):312–323, 2010.

[69] Peng Zhao and Bin Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

[70] S. Zhou, X. Shen, and D.A. Wolfe. Local asymptotics for regression splines and confidence regions. *Ann. Statist.*, 26(5):1760–1782, 10 1998.

[71] Mark E Zmijewski. Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting research*, pages 59–82, 1984.

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

[73] Hui Zou and Ming Yuan. Composite quantile regression and the oracle model selection theory. *Annals of Statistics*, 36(3):1108–1126, 2008.

Appendix A: Supplementary Materials for "Penalized Maximum Tangent Likelihood Estimation and Robust Variable Selection"

A.1 Regularity Conditions

Regularity Conditions for Theorem 1.2.1

R1 The parameter space $\mathcal{B}$ is compact.

R2 The target parameter $\boldsymbol{\beta}_t^* = \arg\max_{\boldsymbol{\beta}\in\mathcal{B}} \mathbb{E}_{\boldsymbol{\beta}_0} \ln_t(f(\mathbf{z};\boldsymbol{\beta}))$ exists and is unique.

R3 $\boldsymbol{\beta}_t^*$ and $\boldsymbol{\beta}_0$ are interior points in $\mathcal{B}$.

R4 The function $\boldsymbol{\beta} \mapsto \ln_t(f(\mathbf{z};\boldsymbol{\beta}))$ is upper-semicontinuous for almost all $\mathbf{z}$.

R5 For every sufficiently small ball $B \subset \mathcal{B}$, the function $x \mapsto \sup_{\boldsymbol{\beta}\in B}\{\ln_t(f(\mathbf{z};\boldsymbol{\beta}))\}$ is measurable and satisfies

$$\mathbb{E}_{\boldsymbol{\beta}_0}[\sup_{\boldsymbol{\beta}\in B} \ln_t(f(\mathbf{z};\boldsymbol{\beta}))] < \infty.$$

Regularity Conditions for Theorem 1.2.2

R6 The function $\boldsymbol{\beta} \mapsto \mathbb{E}_{\boldsymbol{\beta}_0}[\ln_t(f(\mathbf{z};\boldsymbol{\beta}))]$ is twice continuously differentiable (admits a second-order Taylor expansion) in a neighborhood of $\boldsymbol{\beta}_t^*$ with a nonsingular symmetric second derivative matrix.

R7 For any $\boldsymbol{\beta}$ in an open subset of parameter space, let the function $\mathbf{z} \mapsto \ln_t(f(\mathbf{z};\boldsymbol{\beta}))$ be measurable such that $\boldsymbol{\beta} \mapsto \ln_t(f(\mathbf{z};\boldsymbol{\beta}))$ is differentiable at $\boldsymbol{\beta}_t^*$ for $\mathbf{z}$ almost everywhere, with derivative $\partial \ln_t(f(\mathbf{z};\boldsymbol{\beta}))/\partial\boldsymbol{\beta}$. For every $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ in a neighborhood of $\boldsymbol{\beta}_0$, we have

$$|\ln_t(f(\mathbf{z};\boldsymbol{\beta}_1)) - \ln_t(f(\mathbf{z};\boldsymbol{\beta}_2))| \leq ||\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2|| \frac{\partial \ln_t(f(\mathbf{z};\boldsymbol{\beta}_t^*))}{\partial\boldsymbol{\beta}}.$$

Regularity Conditions for Theorem 1.2.3

R8 All eigenvalues of the matrix $\mathbb{E}[\mathbf{x}\mathbf{x}^T]$ are finite and lower bounded by a positive value.

Regularity Conditions for Theorem 1.3.1

R9 The matrix

$$\mathbf{J}(\boldsymbol{\beta}) = \mathbb{E}_{\boldsymbol{\beta}_0}\left[\frac{\partial^2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}\ln_t(f(\mathbf{z};\boldsymbol{\beta}))\right]$$

is finite and negative definite at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

R10 Let $\mathcal{B}$ be the parameter space of $\boldsymbol{\beta}$. There exists an open subset $B \subset \mathcal{B}$ that contains the true parameter $\boldsymbol{\beta}_0$ such that for almost every $\mathbf{z} = (y, \mathbf{x}^T)$ the density $f(\mathbf{z};\boldsymbol{\beta})$ admits all third derivatives $(\partial^3 f(\mathbf{z};\boldsymbol{\beta}))/(\partial\beta_j\ \partial\beta_k\ \partial\beta_l)$ for all $\boldsymbol{\beta} \in B$. Furthermore, there exist functions M_{jkl} such that

$$\left|\frac{\partial^3}{\partial\beta_j\partial\beta_k\partial\beta_l}\ln_t(f(\mathbf{z};\boldsymbol{\beta}))\right| \leq M_{jkl}(\mathbf{z}) \quad \text{for all } \boldsymbol{\beta} \in B \text{ and for almost every } \mathbf{z}.$$

where $m_{jkl} = \mathbb{E}_{\boldsymbol{\beta}_0}[M_{jkl}(\mathbf{z})] < \infty$ for j, k, l.

A.2 List of Variables of Boston Housing Dataset

Response variable:

medv: median value of owner-occupied homes in thousand dollars.

Covariates:

rm: average number of rooms per dwelling.

tax: full-value property-tax rate per 10,000 dollars.

ptratio: pupil-teacher ratio by town.

lstat: pct. lower status of the population.

nox: nitric oxides concentration (parts per 10 million).

dis: weighted distances to five Boston employment centers.

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.

age: proportion of owner-occupied units built prior to 1940.

black: $1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town.

chas: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

rad: index of accessibility to radial highways.

Appendix B: Coefficient estimates of two and three-year ahead forecast models

Table B.1: Coefficient estimates for different models under two-year ahead forecasting horizon. The time period of the training dataset is 1980-2015. Panel A and B separates market and accounting based variables. The data used for penalized methods is scaled to (0, 1) range. Standard errors are reported in parenthesis with italic font. For index model, the standard error is obtained by 500 bootstrapped samples.

	Double-Index	Single-Index	Lasso	CHS	Altman
			Panel A: Market Variable		
Intercept			-2.625 (*0.389*)	-7.102 (*0.501*)	-5.075 (*0.106*)
SIGMA	0.365 (*0.142*)	0.203 (*0.022*)	0.986 (*0.280*)	0.330 (*0.102*)	
EXRET	-0.851 (*0.169*)	-0.396 (*0.037*)	-2.685 (*0.489*)	-0.463 (*0.097*)	
NIMTA	-0.264 (*0.107*)	-0.266 (*0.041*)		-0.688 (*0.300*)	
LTMTA			1.027 (*0.335*)	1.650 (*0.218*)	
LOG(PRICE)	-0.269 (*0.067*)		-0.703 (*0.313*)	-0.297 (*0.066*)	
CASHMTA				-0.900 (*0.415*)	
SIZE			-1.102 (*0.723*)	-0.130 (*0.036*)	
MBE				0.129 (*0.025*)	
LCTMTA					
MVEF					-0.784 (*0.360*)
			Panel B: Accounting Variable		
LTAT	0.528 (*0.041*)	0.451 (*0.028*)	2.475 (*0.309*)		
NIAT	0.420 (*0.053*)	0.376 (*0.043*)			
EBITAT	-0.577 (*0.057*)	-0.389 (*0.051*)	-0.872 (*0.443*)		-2.317 (*0.274*)
REAT	0.225 (*0.071*)	0.144 (*0.044*)			-0.012 (*0.053*)
WCAPAT					-0.824 (*0.215*)
SALEAT					0.460 (*0.075*)
LCTAT					
RELCT					
LCTSALE			-0.063 (*0.275*)		
LOG(SALE)	-0.402 (*0.064*)	-0.463 (*0.036*)	-2.942 (*0.831*)		
CHAT					
ACTLCT					
LCTLT					

Table B.2: Decile ranking table, area under the curve (AUC), and the p-value of the Hosmer-Lemeshow goodness-of-fit test for different models under two-year ahead forecasting horizon. The time period of the training dataset is 1980-2015.

	Double-Index	Single-Index	Lasso	CHS	Altman
90-100%	0.516	0.470	0.489	0.463	0.329
80-100%	0.692	0.709	0.687	0.663	0.518
70-100%	0.816	0.814	0.797	0.771	0.640
60-100%	0.890	0.869	0.862	0.852	0.706
50-100%	0.924	0.912	0.916	0.905	0.785
0-100%	1.000	1.000	1.000	1.000	1.000
Pseudo-R^2	0.137	0.131	0.122	0.106	0.039
AUC	0.834	0.824	0.822	0.808	0.713
H-L pval	0.598	0.787	0.010	0.005	0.000

Table B.3: Coefficient estimates for different models under three-year ahead forecasting horizon. The time period of the training dataset is 1980-2015. Panel A and B separates market and accounting based variables. The data used for penalized methods is scaled to (0, 1) range. Standard errors are reported in parenthesis with italic font. For index model, the standard error is obtained by 500 bootstrapped samples.

	Double-Index	Single-Index	Lasso	CHS	Altman
			Panel A: Market Variable		
Intercept			-2.342 (*0.401*)	-7.211 (*0.547*)	-5.157 (*0.120*)
SIGMA	0.534 (*0.096*)	0.156 (*0.021*)	0.862 (*0.335*)	0.280 (*0.123*)	
EXRET	-0.769 (*0.125*)	-0.286 (*0.028*)	-2.160 (*0.529*)	-0.325 (*0.107*)	
NIMTA	-0.351 (*0.137*)	-0.172 (*0.044*)		-0.773 (*0.367*)	
LTMTA				1.186 (*0.235*)	
LOG(PRICE)			-0.461 (*0.355*)	-0.237 (*0.076*)	
CASHMTA				-0.764 (*0.474*)	
SIZE			-1.834 (*0.668*)	-0.158 (*0.039*)	
MBE				0.149 (*0.026*)	
LCTMTA					
MVEF					0.018 (*0.236*)
			Panel B: Accounting Variable		
LTAT	0.417 (*0.029*)	0.398 (*0.024*)	2.618 (*0.282*)		
NIAT	0.438 (*0.047*)	0.388 (*0.042*)			
EBITAT	-0.466 (*0.038*)	-0.433 (*0.041*)	-0.373 (*0.490*)		-2.314 (*0.319*)
REAT	0.334 (*0.067*)	0.304 (*0.052*)			0.005 (*0.062*)
WCAPAT					-0.334 (*0.237*)
SALEAT					0.370 (*0.087*)
LCTAT					
RELCT	-0.207 (*0.037*)	-0.186 (*0.031*)	-0.477 (*0.379*)		
LCTSALE					
LOG(SALE)	-0.512 (*0.044*)	-0.490 (*0.039*)	-2.440 (*0.681*)		
CHAT					
ACTLCT					
LCTLT					

Table B.4: Decile ranking table, area under the curve (AUC), and the p-value of the Hosmer-Lemeshow goodness-of-fit test for different models under three-year ahead forecasting horizon. The time period of the training dataset is 1980-2015.

	Double-Index	Single-Index	Lasso	CHS	Altman
90-100%	0.393	0.387	0.364	0.344	0.315
80-100%	0.621	0.607	0.569	0.564	0.462
70-100%	0.746	0.743	0.760	0.697	0.581
60-100%	0.853	0.824	0.841	0.824	0.688
50-100%	0.916	0.884	0.902	0.902	0.769
0-100%	1.000	1.000	1.000	1.000	1.000
Pseudo-R^2	0.099	0.100	0.083	0.071	0.027
AUC	0.794	0.792	0.792	0.778	0.698
H-L pval	0.176	0.956	0.000	0.000	0.000

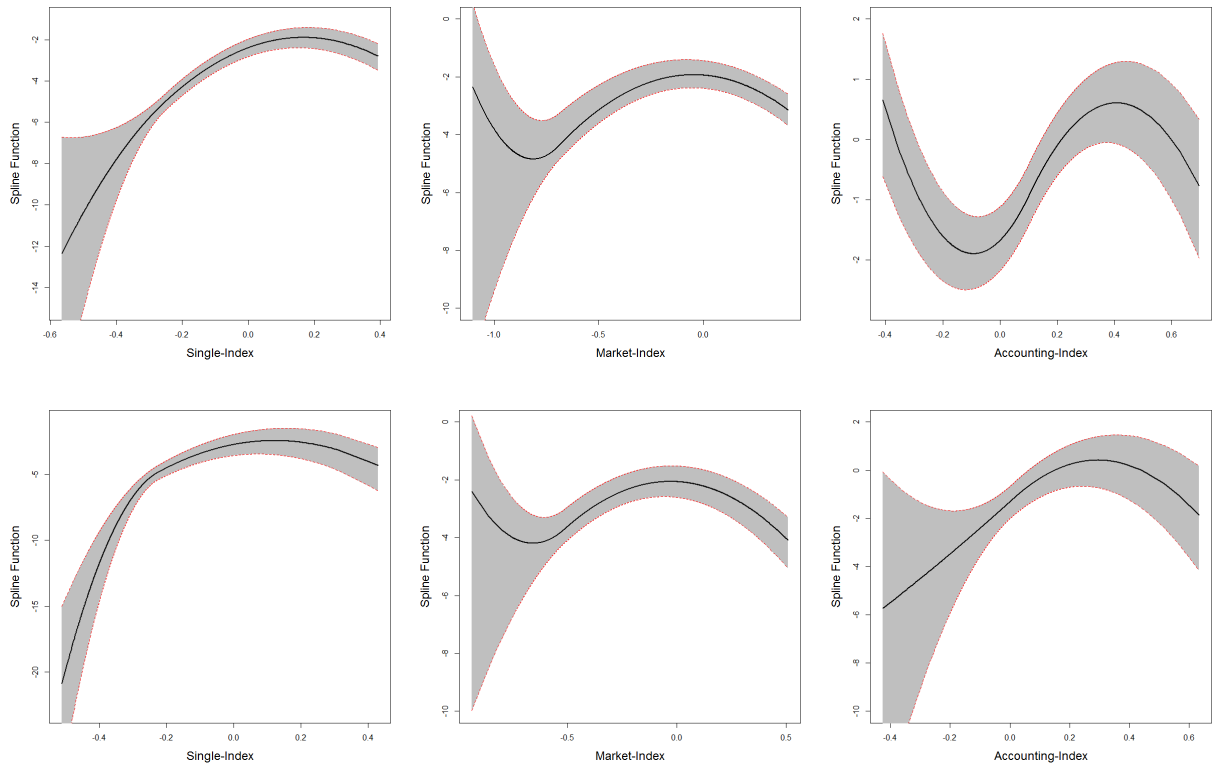

Figure B.1: From left to right, the plots are estimated unknown link function of single-index (mixture of market and accounting variables), market-index and accounting-index (double-index model). Top panel is for two-year ahead forecasting horizon, and bottom penal is for three-year ahead forecasting. The training dataset is based on full sample, i.e., the time period is 1980-2015.