5-2010

Shrinkage Estimation in Partially Linear Models with Measurement Error

Yifang Li
Clemson University, yifangl@clemson.edu

Shrinkage Estimation in Partially Linear Models with Measurement Error

---

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Mathematical Sciences

---

by
Yifang Li
May 2010

Accepted by:
Dr. K. B. Kulasekera, Dr. Colin M. Gallagher, Committee Chair
Dr. Xiaoqian Sun

Abstract

In practice, measurement error in the covariates is often encountered. Measurement error has several effects when using ordinary least squares for the regression problems. In this thesis, we introduce the basic idea of correcting the bias caused by different types of measurement error. We then focus on the variable selection for partially linear models when some of the covariates are measured with additive errors. The bias caused by the measurement error is corrected by subtracting a bias correction term in the squared loss function. Adaptive LASSO is used for the variable selection procedure. The rate of convergence and the asymptotic normality of the estimators resulted by the proposed procedure are established. We also proved that, with the correct choice of the rate of the regularization parameter, the proposed procedure asymptotically performs as well as when the true model is known in advance. This is the so-called oracle properties.

Dedication

I dedicate this work to my beloved parents.

Acknowledgments

I would like to express my heartiest gratitude to my advisors, Dr. Kulasekera and Dr. Gallagher, for their insightful guidance and constant support. It is their unending encouragement and help that makes this work a success. I would also like to thank Dr. Sun for being on my committee and working with me on this thesis. A very special thank to my beloved parents. Without their love and support, I would have never made through the life and study here, and finally made this thesis a complete one.

Table of Contents

Title Page . . . . . **i**

Abstract . . . . . **ii**

Dedication . . . . . **iii**

Acknowledgments . . . . . **iv**

List of Tables . . . . . **vi**

1 Introduction **1**
1.1 Measurement Error 1
1.2 Variable Selection Procedures and the Oracle Properties 3
1.3 Variable Selection for Measurement Error Data 8

2 Measurement Error **10**
2.1 Taxonomy of Measurement Error Models 10
2.2 Nondifferential Error and Differential Error 11
2.3 Estimation for Different Types of Measurement Error Models in Linear Regression . 12

3 Variable Selection for Partially Linear Models with Adaptive LASSO **24**
3.1 The Model, Assumptions and Notations 24
3.2 The Oracle Properties of Adaptive LASSO 27

Appendices **45**
A Regularity Conditions 46

Bibliography **47**

List of Tables

2.1 Summary of Simple Linear Regression 23

Chapter 1

Introduction

1.1 Measurement Error

In regression problems, measurement error in the covariates is often encountered in many fields, such as epidemiology, economics and biology. Measurement errors are sometimes hard to avoid due to the accuracy of instrument and the way that the data are collected. For example, the National Cancer Institute's OPEN study by Subar, et al. [14], concerns the relation between cancer and dietary protein intake. However, the long-term intake, denoted by $\boldsymbol{X}$, cannot be observed in practice. Instead, another variable, urinary nitrogen, denoted by $\boldsymbol{W}$ is measured and used in the analysis. As a surrogate of the true value $\boldsymbol{X}$, $\boldsymbol{W}$ has a random error against $\boldsymbol{X}$. If it is used directly in the regression analysis, it may cause problems in the estimation. Generally speaking, measurement error in covariates has three effects:

- It causes bias in parameter estimation for statistical models.
- It leads to a loss of power, sometimes profound, for detecting interesting relationships among variables.
- It masks the features of the data, making graphical model analysis difficult.

The first two are called the *double whammy* of measurement error, i.e., not only is the slope attenuated if ordinary estimation methods are used, but the data are also more noisy, with an increased error about the regression line. In this article, the following notations are used to describe this measurement error problem.

- $\boldsymbol{Y}$ denotes the response in the underlying regression model in terms of predictors $\boldsymbol{X}$ and $\boldsymbol{Z}$, which are observed without error.
- $\boldsymbol{X}$ denotes the unobservable true covariate of concern. It cannot be measured exactly and so is subject to error. It is often called the *error-prone predictor* or the *latent predictor*.
- $\boldsymbol{W}$ denotes the observed value of the mismeasured variable $\boldsymbol{X}$. It may or may not be unbiased for $\boldsymbol{X}$.
- $\boldsymbol{Z}$ represents those predictors that for all practical purposes are measured without error. We will treat them as constant throughout our context, and the analysis will be conditioned on their values.
- $\boldsymbol{U}$ is the measurement error on $\boldsymbol{X}$.

We are interested in relating the response $\boldsymbol{Y}$ to the true predictors $(\boldsymbol{X}, \boldsymbol{Z})$. With the existence of measurement error, we need to make careful corrections according to different types and nature of the error, and the sources of data.

Specification of a model is a prerequisite for analyzing a measurement error problem. Here we give the examples of two fundamental types of measurement error. In the example mentioned at the beginning of this section, in trying to measure urinary nitrogen repeatedly, there are various sources of error including simple machine recording error, administration error, time of day and season of the year. It seems that $\boldsymbol{W}$ has more variability than $\boldsymbol{X}$ and thus it is reasonable to hypothesize that the structure is additive error model, which we write as

$$\boldsymbol{W} = \boldsymbol{X} + \boldsymbol{U}, \tag{1.1}$$

where we assume the error $\boldsymbol{U}$ is independent of $\boldsymbol{X}$ or at least $\mathrm{E}[\boldsymbol{U}|\boldsymbol{X}] = 0$. This model is also said to be *classical error model*. The other basic type is called *Berkson model*. We take the herbicide study by Rudemo, et al. [12] as an example. In that study, a certain measured amount $\boldsymbol{W}$ of herbicide was applied to a plant. However, the actual amount $\boldsymbol{X}$ absorbed differed from plant to plant. In this case, it makes sense that we believe that the true $\boldsymbol{X}$ varies more than $\boldsymbol{W}$ since it is fixed by design. We write the error model as

$$\boldsymbol{X} = \boldsymbol{W} + \boldsymbol{U}, \tag{1.2}$$

in which the error $\boldsymbol{U}$ is assumed to be independent of the observed value $\boldsymbol{W}$ or at least $\mathrm{E}[\boldsymbol{U}|\boldsymbol{W}] = 0$. Determining what error model to use depends on the circumstances and the design of the experiments. We will discuss this more in Chapter 2.

Further overview and systematic survey on this research area can be seen in Fuller [10] and Carroll, et al. [2]. In this thesis, we will talk about some details of the taxonomy of measurement error models and the correcting bias methods for linear regression with different types of measurement error.

1.2 Variable Selection Procedures and the Oracle Properties

Variable selection is an important topic in linear regression analysis. In practice, many predictors are generally collected and considered in the initial modeling. However, the underlying model may only have a sparse representation. In this situation, selecting the true variables that explain the response gives a more accurate model to the problem. Moreover, a large number of predictors in the regression model causes difficulty in interpretation of the fitted model. Hence identifying significant predictors is of fundamental interest in statistical learning and analysis.

Ordinary least squares is a widely used method to estimate the parameters in a linear model. It gives unbiased estimators. However, it never sets coefficients to be exactly zero. It does not accomplish variable selection. Traditionally, two standard techniques are used for variable selection: subset selection and stepwise selection. Several criteria are developed and frequently used such as AIC [1], BIC [13] and RIC [9]. These techniques give interpretable models, but also suffer from several drawbacks. First, these procedures are discrete methods in the sense that predictors are either retained or dropped from the model, and hence the resulting estimators can be extremely variable. As a consequence, a small change in data can produce a very different fitted model. They also ignore the stochastic errors inherited in the stages of variable selections. To avoid the instability caused by discontinuity, continuous shrinkage methods are considered.

Several shrinkage estimation procedures, which can simultaneously estimate the parameters and select variables, have been developed. Fan and Li [7] proposed that a good variable selection procedure should possess three good properties:

- *Unbiasedness*: the resulting estimator is nearly unbiased, especially when the true unknown parameter is large.

- *Sparsity*: the resulting estimator is a thresholding rule, which sets some estimated parameters to be exactly zero to accomplish variable selection requirement.
- *Continuity*: the resulting estimator should be continuous to avoid the high variability in model prediction.

Fan and Li [7] also argued that a variable selection procedure is said to have *oracle properties* if it asymptotically identifies the correct model and estimates the parameters as if the true model is known in advance. In other words, the estimator should have some components correctly set to be zero, and for the other nonzero estimated coefficients, they have consistency and asymptotic normality.

We assume that the model is

$$\mathrm{E}[\boldsymbol{Y}|\boldsymbol{X}] = \boldsymbol{X}\beta, \tag{1.3}$$

where $\boldsymbol{Y}$ is the response vector, and $\boldsymbol{X}$ is the $n \times d$ design matrix. Without loss of generality, we assume that the data are centered so that the intercept is not included in the model. Let $p_\lambda(|\beta|)$ be the penaly function. A penalized least squares (PLS) is defined as

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})^2 + n\sum_{j=1}^{d} p_\lambda(|\beta_j|), \tag{1.4}$$

where $\boldsymbol{X}_i$ is the transpose of the row vector in $\boldsymbol{X}$, so it is a $d \times 1$ vector. The three properties can be interpreted in terms of the penaly function $p_\lambda(|\beta|)$ as in Fan and Li [8]:

- *Unbiasedness*: $p'_\lambda(|\beta|) = 0$ for large $|\beta|$.
- *Sparsity*: $\min_\beta\{|\beta| + p'_\lambda(|\beta|)\} > 0$.
- *Continuity*: if and only if $\arg\min_\beta\{|\beta| + p'_\lambda(|\beta|)\} = 0$.

Tibshirani [15] developed the least absolute shrinkage and selection operator (LASSO) method. The PLS with LASSO is defined as

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{d}|\beta_j|, \tag{1.5}$$

where λ is a nonnegative regularization parameter. The LASSO is a $\mathcal{L}_1$ penalty and it is singular at the origin. The LASSO continuously shrinks the coefficients toward 0 as λ increases. It automatically sets some parameters to be exactly 0 when λ is sufficiently large. It has been shown that under certain conditions, the $\mathcal{L}_1$ penalization procedure can identify the right sparse model (Donoho et al. [6]; Donoho et al. [5]; Donoho [4]). However, it has been shown that the optimal λ for prediction gives inconsistent variable selection results. Zou [17] proved that the LASSO is consistent only when the underlying model satisfies some nontrivial conditions. Zou [17] also gave an example showing that the LASSO is inconsistent in variable selection.

Fan and Li [7] proposed a smoothly clipped absolute deviation (SCAD) penalty and proved that it has the oracle properties in the nonconcave penalized likelihood. Because the derivative of the LASSO penalty is a constant, they conjectured that $\mathcal{L}_1$ does not have the oracle properties. Set up the linear regression model as

$$\boldsymbol{Y} = \boldsymbol{X}\beta + \epsilon, \tag{1.6}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector and $\boldsymbol{X}$ is an $n \times d$ matrix. In Fan and Li [7], they further assumed that the columns of $\boldsymbol{X}$ is orthonormal. The SCAD is defined as

$$p_\lambda(|\beta|) = \begin{cases} \lambda|\beta|, & \text{if } 0 \le |\beta| < \lambda; \\ -\frac{|\beta|^2 - 2a\lambda|\beta| + \lambda^2}{2(a-1)}, & \text{if } \lambda \le |\beta| < a\lambda; \\ \frac{(a+1)\lambda^2}{2}, & \text{if } |\beta| \ge a\lambda. \end{cases} \tag{1.7}$$

They further suggested using $a = 3.7$ in (1.7). In Fan and Li [7], set $\boldsymbol{V} = (\boldsymbol{X}_i, Y_i)$, $i = 1, \ldots, n$. $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n$ be independent and identically distributed, and suppose each with a density $f(\boldsymbol{V}, \boldsymbol{\beta})$. The penalized likelihood function with SCAD was considered:

$$Q(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) - n\sum_{j=1}^{d} p_{\lambda_n}(|\beta_j|), \tag{1.8}$$

where $L(\boldsymbol{\beta})$ is the log-likelihood function of the observations $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n$. Without loss of generality, they denoted the true parameters as

$$\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0s}, \beta_{0(s+1)}, \ldots, \beta_{0d})^T = \begin{pmatrix} \boldsymbol{\beta}_{01} \\ \boldsymbol{\beta}_{02} \end{pmatrix}, \tag{1.9}$$

where $\boldsymbol{\beta}_{02} = \mathbf{0}$, and $\boldsymbol{\beta}_{01}$ contains all the nonzero components in $\boldsymbol{\beta}_0$ and its size is $s \times 1$. Under some regulating conditions, they proved the oracle properties for the nonconcave penalized likelihood satisfying certain conditions. We state the oracle properties in the following theorem, which is a combination of Theorem 1 and 2 in Fan and Li [7] specifically for SCAD:

Theorem 1.2.1. *If $\lambda_n \to 0$ and $\sqrt{n}\lambda_n \longrightarrow \infty$ as $n \to \infty$, then with probability tending to 1, the root n consistent local maximizers $\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix}$ must satisfy:*

1. *Unbiasedness: There exists a local maximizer $\widehat{\boldsymbol{\beta}}$ of $Q(\boldsymbol{\beta})$ such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-\frac{1}{2}})$.*

2. *Sparsity: $\widehat{\boldsymbol{\beta}}_2 = \mathbf{0}$.*

3. *Asymptotic normality:*

$$\sqrt{n}(I_1(\boldsymbol{\beta}_{01}) + \Sigma)^{-1}\{\widehat{\boldsymbol{\beta}_1} - \boldsymbol{\beta}_{01} + (I_1(\boldsymbol{\beta}_{01}) + \Sigma)^{-1}\boldsymbol{b}\} \xrightarrow{\mathcal{D}} N\{\mathbf{0}, I_1(\boldsymbol{\beta}_{01})\},$$

where $I_1(\boldsymbol{\beta}_{01}) = I_1(\boldsymbol{\beta}_{01}, \mathbf{0})$ is the Fisher information knowing $\boldsymbol{\beta}_2 = \mathbf{0}$, and

$$\Sigma = \mathrm{diag}\{p''_{\lambda_n}(|\beta_{01}|), \ldots, p''_{\lambda_n}(|\beta_{0s}|)\},$$

$$\boldsymbol{b} = (p'_{\lambda_n}(|\beta_{01}|)\mathrm{sgn}(\beta_{01}), \ldots, p'_{\lambda_n}(|\beta_{0s}|)\mathrm{sgn}(\beta_{0s})).$$

As a consequence, the asymptotic covariance matrix of $\widehat{\boldsymbol{\beta}}_1$ is

$$\frac{1}{n}\{I_1(\boldsymbol{\beta}_{01}) + \Sigma\}^{-1} I_1(\boldsymbol{\beta}_{01})\{I_1(\boldsymbol{\beta}_{01}) + \Sigma\}^{-1},$$

which approximately equals $\frac{1}{n}[I_1(\boldsymbol{\beta}_{01})]^{-1}$ for the SCAD thresholding penalties if λ_n tends to 0.

Based on the conditions required in this theorem, Fan and Li [7] also pointed out that, since the derivative of $\mathcal{L}_1$ penalty is always λ_n, the root-n consistency requires that $\lambda_n = O_p(n^{-\frac{1}{2}})$. However, it is also required in Theorem 1.2.1 that $\sqrt{n}\lambda_n \to \infty$. These two conditions cannot be satisfied for LASSO simultaneously. They conjectured that the oracle properties do not hold for $\mathcal{L}_1$ penalty.

Zou [17] proposed a new version of LASSO, which is called adaptive LASSO. Using the same

model as (1.6), the penalized least squares with adaptive LASSO is defined as

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})^2 + \lambda_n \sum_{j=1}^{d} \widehat{w}_j|\beta_j|. \tag{1.10}$$

Instead of simply using the absolute value of true parameters as the penalization, adaptive weights are added for penalizing different coefficients. For example,

$$\widehat{w}_j = \frac{1}{|\widehat{\beta}_j|^\gamma}$$

can be chosen as the weight, where $\widehat{\beta}_j$ comes from minimizing the least squares without penalty. By the definition, the weight goes to zero as the corresponding β_j is sufficiently large. This fixes the problem that the original LASSO contradicts the requirement for unbiasedness of a variable selection procedure. Adaptive LASSO is a convex optimization problem with an $\mathcal{L}_1$ constraint, and it can be solved by the same efficient algorithm used for LASSO. Moreover, Zou [17] also proved that, by letting λ_n vary in n and correctly setting the rate of λ_n, the oracle properties of adaptive LASSO hold.

Denote the true model by $\mathcal{A}$, in which there are totally s covariates, and the selected model from (1.10) by $\mathcal{A}^*$. Suppose $\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{C}$, where $\boldsymbol{C}$ is a positive definite matrix. Let

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{pmatrix},$$

where $\boldsymbol{C}_{11}$ is a $s \times s$ corresponding to the true nonzero components in the true $\boldsymbol{\beta}$. We state the oracle properties in the following theorem using Zou [17]'s notations. This theorem is the same as Theorem 2 in Zou [17].

Theorem 1.2.2. *Suppose that* $\frac{\lambda_n}{\sqrt{n}} \to 0$, *and* $\lambda_n\, n^{\frac{\gamma-1}{2}} \to \infty$ *as* $n \to \infty$. *Then the adaptive LASSO estimates satisfy the following:*

1. *Consistency in variable selection:*

$$\lim_{n\to\infty} P(\mathcal{A}_n^* = \mathcal{A}) = 1.$$

2. Asymptotic normality:

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \sigma^2 \boldsymbol{C}_{11}^{-1}),$$

where $\boldsymbol{\beta}_{01}$ is defined as the same as in (1.9) and $\widehat{\boldsymbol{\beta}}_1$ is its corresponding estimator.

This theorem shows that the adaptive LASSO is at least as competitive as other concave oracle procedures.

1.3 Variable Selection for Measurement Error Data

Most of the regular variable selection procedures are designed for models with observed predictors. However, with the existence of the measurement error on a covariate, they may not work properly. Under such circumstances, if the true predictor $\boldsymbol{X}$ is directly replaced by the observed data $\boldsymbol{W}$ in the ordinary least squares, the resulting estimator will be inconsistent for the true parameter because the loss function contains the error on the covariates and the expected value of the loss function is not zero. Hence, the powerful properties of the ordinary least squares may not hold in the presence of measurement error. Since the regular variable selection procedures are accomplished by adding a penalty function to the ordinary least squares, they may not work properly either when ignoring the measurement error. Thus seeking a variable selection method for models with measurement error is necessary for statisticians interested in models with measurement error. Liang et al. [11] proposed two variable selection procedures for the additive measurement error on one covariate in a partially linear regression model. One technique is accomplished by minimizing a penalized least squares with SCAD. The other one is penalized quantile regression. By choosing the correct penalty function and the tuning parameter, the oracle properties were also proved in that paper, which means that asymptotically the procedures work as well as when the true underlying models is already known. Correction of bias is one of the keys to the success of the two methods. For the penalized quantile regression, they used orthogonal regression proposed by Cheng et al. [3]. For the penalized least squares technique, a bias correcting term was subtracted from the ordinary least squares and so the expectation of the loss function without penalty was corrected to zero.

As mentioned in Section 1.2, Zou [17] proposed a new version of LASSO, adaptive LASSO, and proved its oracle properties. In this thesis, we consider the partially linear model with additive measurement error on the parametric predictors. We are interested in whether adaptive LASSO

under these circumstances has the oracle properties. In Chapter 3, we prove that, by correctly selecting the rate of λ_n, the oracle properties hold for this procedure.

Chapter 2

Measurement Error

In this chapter, we will define several different types of measurement error models. Details of bias correction methods for each type will also be discussed. Chapter 3 of Carroll, et al. [2] gives a thorough description of the methods by using the best linear predictor to correct the bias caused by measurement error. In our study, we used a different angle to understand the bias correction for measurement error models, and our way also leads to the same results as in Carroll, et al. [2]. We will state this approach in detail in this chapter.

2.1 Taxonomy of Measurement Error Models

There are two major ways to classify the measurement error models. One is based on the properties and the assumptions that can be made upon the unobserved true values $\boldsymbol{X}$s. The other one is determined by the structure and the relation between the observed value $\boldsymbol{W}$ and $\boldsymbol{X}$. These two defining characteristics are known as data structure and error structure respectively.

In the first category, if we regard $\boldsymbol{X}$s as a sequence of unknown fixed constants or parameters, then we say the models are *classical functional models*. If on the other hand, $\boldsymbol{X}$s are treated as random variables, the models are called *classical structural models*. In Carroll et al. [2], they pointed out that using maximum likelihood in classical functional models only works in linear regression, but it fails in any other case. In this thesis, we assume $\boldsymbol{X}$s to be random, then there are further classification for the models. If no or only minimal assumptions are made about the distribution of $\boldsymbol{X}$s, we call this modeling as *functional modeling*. In this case, the estimation procedures are robust

even if the distribution of the $\boldsymbol{X}$s are miss-specificated. The other type is *structural modeling*, which is a parametric approach since we place an assumption on the distribution of the random variables $\boldsymbol{X}$s.

The error structure is fundamental in analyzing a measurement error problem and correcting bias for it. It exposes how the observed $\boldsymbol{W}$s are related to the true $\boldsymbol{X}$s. There are generally two types:

- *Error models*, to which classical error models belong. In analyzing this type of measurement error, the conditional distribution of $\boldsymbol{W}$ given $(\boldsymbol{X}, \boldsymbol{Z})$ is modeled.
- *Regression calibration models*, including Berkson error models. In the study of this type of models, the conditional distribution of $\boldsymbol{X}$ given $(\boldsymbol{W}, \boldsymbol{Z})$ is of interest.

A natural question that arises is how one can choose an error model in a practical study. Take classical error models and Berkson error models as the representatives for each type, i.e., the choice is only between the two. If the error-prone covariate $\boldsymbol{X}$ is unique for each individual, and the observed value $\boldsymbol{W}$ can be measured repeatedly, we choose classical error models. If a same observed value is given to a group, whereas true value $\boldsymbol{X}$ of each individual in the group varies, Berkson error models will be the choice. For example, in a children's lung disease study, the NO_2 amount in a family was measured and treated as the amount of intake for every child in the family. However, the true absorbed value differed from child to child. Then Berkson error model is better in modeling such circumstances. As pointed out in Chapter 1, the assumption for the classical error model is that the error $\boldsymbol{U}$ is independent of $\boldsymbol{X}$ or at least $\mathrm{E}[\boldsymbol{U}|\boldsymbol{X}] = 0$, in which case, $\mathrm{Var}(\boldsymbol{W}) > \mathrm{Var}(\boldsymbol{X})$. On the other hand, Berkson error model assumes that the error $\boldsymbol{U}$ is independent of $\boldsymbol{W}$ or at least $\mathrm{E}[\boldsymbol{U}|\boldsymbol{W}] = 0$, so $\mathrm{Var}(\boldsymbol{X}) > \mathrm{Var}(\boldsymbol{W})$.

In this thesis, we consider the variable selection procedure especially for the classical error model, i.e., the measurement error is additive.

2.2 Nondifferential Error and Differential Error

It is important to make a distinction between *differential* and *nondifferential* measurement error.

Definition 2.2.1. The measurement error is *nondifferential* if the distribution of $\boldsymbol{Y}$ given $(\boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{W})$

depends only on $(\boldsymbol{X}, \boldsymbol{Z})$. In other words, $\boldsymbol{W}$ is conditionally independent of $\boldsymbol{Y}$ given $(\boldsymbol{X}, \boldsymbol{Z})$. In this case, $\boldsymbol{W}$ is said to be a *surrogate* of $\boldsymbol{X}$.

Under the assumption of nondifferential error, $\boldsymbol{W}$ contains no information about the response $\boldsymbol{Y}$ other than what is available in $\boldsymbol{X}$ and $\boldsymbol{Z}$. One can estimate the parameters in the underlying model for $\boldsymbol{Y}$ even though the true value of the predictors are not observed. On the other hand, if there is additional information available in $\boldsymbol{W}$ other than that contained in $\boldsymbol{X}$ and $\boldsymbol{Z}$, the error is *differential*. If this occurs, true values of the covariates must be observed for estimating the parameters in the target underlying model.

2.3 Estimation for Different Types of Measurement Error Models in Linear Regression

In this section, we focus on the estimation when the underlying regression models are linear, especially simple linear regression models. Measurement error has effects on many factors: it causes the attenuation in the estimation of the parameters; it hides the real effects of the true values $\boldsymbol{X}$; it sometimes exhibits relationships which are not present when the covariates are measured without error; the sign of the estimator may even be reversed. These effects caused by measurement error are determined by the properties of the underlying models: whether the model is simple or multiple, and whether the error-prone predictors are univariate or multivariate, and by the presence of the bias in the measurement error. Fuller [10] gives a comprehensive discussion of linear measurement error models, and some known results are also summarized in Carroll. et al [2].

In this section, we will show the danger of using the naive estimator for classical error models. We will also briefly discuss the approach in Carroll et al. [2] to adjust the attenuation caused by measurement error. We will also state another approach in this section.

2.3.1 Naive Estimator and the Idea of Correcting Bias

2.3.1.1 Naive Estimator of Classical Error Models and Berkson Error Models

Suppose the underlying model is a simple linear regression model, written as

$$Y = \beta_0 + \beta_x X + \epsilon, \tag{2.1}$$

where ϵ is independent of X with $\mathrm{E}[\epsilon] = 0$ and variance σ_ϵ^2. The measurement error model is additive, i.e.,

$$W = X + U, \tag{2.2}$$

where the error U is assumed to be independent of X with $\mathrm{E}[U] = 0$ and variance σ_u^2. Rewrite the additive measurement error structure (2.2) as

$$X = W - U. \tag{2.3}$$

The *naive estimator* is the estimator obtained by simply plugging (2.3) into the underlying model (2.1) and performing the ordinary estimation procedure. After plugging (2.3) into (2.1), we get

$$\begin{aligned} Y &= \beta_0 + \beta_x(W - U) + \epsilon \\ &= \beta_0 + \beta_x W + \epsilon - \beta_x U \end{aligned} \tag{2.4}$$

Now the error is $\epsilon - \beta_x U$, which has zero mean. Further assume that U is uncorrelated with ϵ, then the error has variance of $\sigma_\epsilon^2 + \beta_x^2\sigma_u^2$, which is larger than σ_ϵ^2. This means that (Y, W) data have more variability around a line than (Y, X) data. Another more important fact is that, the new covariate W and the new error $\epsilon - \beta_x U$ are correlated since $\mathrm{Cov}(W, U) = \mathrm{Cov}(X + U, U) = \sigma_u^2 \neq 0$. This is the main source of the bias in the estimator. Hence, the ordinary least squares cannot be applied directly to (2.4). Later we will show how to conduct estimation for these models and it can be seen that there is an *attenuation to the null* in the estimator.

However, under the Berkson model, the measurement error model is

$$X = W + U, \tag{2.5}$$

where U is independent of W with mean zero. If (2.5) is plugged in (2.1), then

$$\begin{aligned} Y &= \beta_0 + \beta_x(W + U) + \epsilon \\ &= \beta_0 + \beta_x W + \epsilon + \beta_x U. \end{aligned} \tag{2.6}$$

Although the new error is changed to $\epsilon + \beta_x U$, we can show that under the nondifferential assumption,

$\mathrm{Cov}(\epsilon, W) = 0$, then the error in (2.6) is still uncorrelated with the regressor W. Hence the estimator is unbiased for β_0 and β_x.

Thus special treatment has to be done in the classical error models. Actually, in classical error models, by regressing X on W, and replacing X in (2.1) with the best linear predictor, the classical error models can be transformed to Berkson error models. This is the basic idea of how we estimate the parameters with the existence of measurement error. Details will be given in Section 2.3.2.

2.3.1.2 Classical Error as Berkson Error

As mentioned before, the basic idea of correcting the bias is to change the classical error into Berkson error. Suppose the underlying model is still (2.1), and the error structure is

$$W = X + U_c, \tag{2.7}$$

where W is assumed to be a surrogate of X, i.e., W is conditionally independent of Y given X (we don't have Z in the underlying model). Also assume that $\mathrm{E}[U_c|X] = 0$, then we know that X and U_c are uncorrelated, and the variance of U_c is $\sigma^2_{u_c}$. Hence,

$$\sigma^2_w = \sigma^2_x + \sigma^2_{u_c}, \tag{2.8}$$

$$\rho^2_{wx} = \frac{\sigma^2_{xw}}{\sigma^2_x \sigma^2_w} = \frac{\sigma^4_x}{\sigma^2_x(\sigma^2_x + \sigma^2_{u_c})} = \frac{\sigma^2_x}{\sigma^2_x + \sigma^2_{u_c}}, \tag{2.9}$$

where ρ denotes the correlation between two random variables. We further assume that (Y, X, W) is multivariate normal. Define W_{blp} to be the best linear predictor of X based on W. Assume that

$$W_{blp} = \gamma_0 + \gamma_w W, \tag{2.10}$$

where γ_0 and γ_x minimizes $\mathrm{E}[\{X - (\gamma_0 + \gamma_w W)\}^2]$. Then

$$\begin{cases} \gamma_0 = \mu_x - \gamma_w \mu_w \\ \gamma_w = \dfrac{\sigma_{wx}}{\sigma^2_w} \end{cases} \tag{2.11}$$

So the best linear predictor is

$$W_{blp} = \mu_x + \frac{\sigma_{wx}}{\sigma_w^2}(W - \mu_w), \tag{2.12}$$

and the prediction error is

$$U^* = X - W_{blp} = X - \mu_x - \frac{\sigma_{wx}}{\sigma_w^2}(W - \mu_w). \tag{2.13}$$

Note that following from (2.9), $\frac{\sigma_{wx}}{\sigma_w^2}$ is just ρ_{wx}^2, and denote $\lambda = \rho_{wx}^2$, then since $\mu_x = \mu_w$, X can be written as

$$X = (1 - \lambda)\mu_x + \lambda W + U^*, \tag{2.14}$$

where the error U^* has mean zero and variance

$$\begin{aligned} \sigma_{u^*}^2 &= \mathrm{Var}(X - W_{blp}) = \sigma_x^2(1 - \rho_{xw}^2) \\ &= \sigma_x^2(1 - \frac{\sigma_x^2}{\sigma_x^2 + \sigma_{u_c}^2}) = \sigma_x^2 \frac{\sigma_{u_c}^2}{\sigma_x^2 + \sigma_{u_c}^2} = \sigma_{u_c}^2 \frac{\sigma_x^2}{\sigma_x^2 + \sigma_{u_c}^2} = \lambda \sigma_{u_c}^2. \end{aligned} \tag{2.15}$$

The regression error U^* is also uncorrelated with W because

$$\begin{aligned} \mathrm{Cov}(W, U^*) &= \mathrm{Cov}(W, X - \mu_x - \frac{\sigma_{wx}}{\sigma_w^2}(W - \mu_w)) = \mathrm{Cov}(W, X - \frac{\sigma_{wx}}{\sigma_w^2} W) \\ &= \mathrm{Cov}(W, X) - \frac{\sigma_{wx}}{\sigma_w^2}\mathrm{Cov}(W, W) = \sigma_{wx} - \sigma_{wx} = 0. \end{aligned} \tag{2.16}$$

Then we plug (2.14) into the underlying model (2.1),

$$\begin{aligned} Y &= \beta_0 + \beta_x((1 - \lambda)\mu_x + \lambda W + \beta_x U^*) + \epsilon \\ &= \beta_0 + \beta_x(1 - \lambda)\mu_x + \lambda \beta_x W + \epsilon + \beta_x U^*. \end{aligned} \tag{2.17}$$

What is important is that, after regressing X on W, the new error in (2.17), $\epsilon + \beta_x U^*$ is uncorrelated with W. Hence ordinary least squares estimation can be used to estimate the intercept and the slope and that gives unbiased estimators. Here the new intercept is $\beta_0 + \beta_x(1 - \lambda)\mu_x$, and the new slope is $\lambda\beta_x$. Note that $\lambda = \rho_{wx}^2 = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_{u_c}^2}$, so it is always less than 1. Thus ordinary least squares regression of Y on W produces an estimator of the slope that is attenuated to zero, which is referred to as the *attenuation (to the null)*. Another effect can be seen in the residual variance of this

regression of Y on W, which is

$$\begin{aligned}\mathrm{Var}(Y|W) &= \mathrm{Var}(\epsilon + \beta_x U^*|W) \\ &= \mathrm{Var}(\epsilon + \beta_x U^*) \\ &= \mathrm{Var}(\epsilon) + \beta_x^2 \mathrm{Var}(U^*) \\ &= \sigma_\epsilon^2 + \beta_x^2 \sigma_x^2 (1-\lambda). \end{aligned} \tag{2.18}$$

Thus measurement error causes a double whammy: not only is the slope attenuated, but the data are more noisy about the line.

This technique is thoroughly summarized in Chapter 3 in Carroll et al. [2] for all types of measurement error structures.

2.3.2 Bias Correction in Simple Linear Regression with Measurement Error

In this section, we will discuss how to correct the bias caused by measurement error in estimation for different types of measurement error. Our approach is slightly different from that in Carroll et al. [2], however, they are all based on the same idea.

In this section, the underlying regression model is

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{2.19}$$

where X and ϵ are independent and ϵ has mean zero and variance σ_ϵ^2. We further assume that (Y, X, W) is multivariate normal.

2.3.2.1 Nondifferential

Now we first look at the general case that the measurement error is nondifferential, i.e., W is conditionally independent of Y given X. However, the type of the error structure is not specified,

and W can be either biased or unbiased for X. It is worth noticing that, under such assumption,

$$\begin{aligned}\mathrm{Cov}(Y,W) &= \mathrm{Cov}(\beta_0+\beta_1 X+\epsilon, W)\\ &= \beta_1 \mathrm{Cov}(X,W)+\mathrm{Cov}(\epsilon,W),\end{aligned} \tag{2.20}$$

where

$$\begin{aligned}\mathrm{Cov}(\epsilon,W) &= \mathrm{E}[\epsilon\, W]\\ &= \mathrm{E}\big[\mathrm{E}[\epsilon\, W|X]\big]\\ &= \mathrm{E}\big[\mathrm{E}[\epsilon|X]\,\mathrm{E}[W|X]\big]\\ &= \mathrm{E}\big[\mathrm{E}[\epsilon]\,\mathrm{E}[W|X]\big]\\ &= \mathrm{E}\big[0\cdot \mathrm{E}[W|X]\big]=0.\end{aligned} \tag{2.21}$$

Hence,

$$\mathrm{Cov}(Y,W)=\beta_1\mathrm{Cov}(X,W)=\beta_1\sigma_{xw}. \tag{2.22}$$

Since we only have the data on W, we can look at the model

$$Y=\beta_0^*+\beta_1^* W+\eta, \tag{2.23}$$

where $\mathrm{E}[\eta]=0$. We want to estimate the parameters in this model and then study their relationship with the true parameters in the original model (2.19). The predictors minimizing $\mathrm{E}[\{Y-(\beta_0^*+\beta_1^*W)\}^2]$ are

$$\beta_1^*=\frac{\mathrm{Cov}(Y,W)}{\mathrm{Cov}(W,W)}=\frac{\sigma_{xw}}{\sigma_w^2}\beta_1, \tag{2.24}$$

$$\beta_0^*=\mu_y-\beta_1^*\mu_w=\beta_0+\beta_1\mu_x-\frac{\sigma_{xw}}{\sigma_w^2}\beta_1\mu_w. \tag{2.25}$$

Note that $\beta_0^*+\beta_1^*W$ is the best linear predictor of Y based on W. As a result, we also have $\mathrm{Cov}(W,\eta)=0$. The estimator of β_1^* is an unbiased estimator for the attenuated slope $\frac{\sigma_{xw}}{\sigma_w^2}\beta_1$. The

residual variance of this regression of Y on W is

$$\begin{aligned}
\mathrm{Var}(Y|W) &= \mathrm{Var}(\eta|W) = \mathrm{Var}(\eta) \qquad (2.26)\\
&= \sigma_y^2(1-\rho_{yw}^2)\\
&= \sigma_y^2\left(1-\frac{\mathrm{Cov}(Y,W)^2}{\sigma_y^2\sigma_w^2}\right)\\
&= \sigma_y^2\left(1-\frac{\beta_1^2\sigma_{xw}^2}{\sigma_y^2\sigma_w^2}\right)\\
&= \sigma_y^2-\beta_1^2\frac{\sigma_{xw}^2}{\sigma_w^2}\\
&= \sigma_\epsilon^2+\beta_1^2\sigma_x^2-\beta_1^2\frac{\sigma_{xw}^2}{\sigma_w^2\sigma_x^2}\sigma_x^2\\
&= \sigma_\epsilon^2+\beta_1^2\sigma_x^2(1-\rho_{xw}^2). \qquad (2.27)
\end{aligned}$$

(2.27) follows from the multivariate normality assumption. However, actually, as long as the variance of η is free of W, (2.27) also holds.

The above analysis shows that if W is a surrogate, then the original parameter β_1 can be recovered with knowledge of or estimability of σ_{xw}. The residual variance in the regression of Y based on W is always greater than the original σ_ϵ^2 since $\rho_{xw}^2 < 1$. In this sense, W is always less informative than X.

In the following three types of error models, we assume the measurement error is nondifferential. Each section deals with one specific error structure.

2.3.2.2 Classical Error Model

The measurement error structure is

$$W = X + U_c, \qquad (2.28)$$

where U_c is independent of X, and it has mean zero. Then

$$\mu_w = \mu_x, \qquad \sigma_w^2 = \sigma_x^2 + \sigma_{u_c}^2; \qquad (2.29)$$

and

$$\sigma_{xw} = \mathrm{Cov}(X,W) = \mathrm{Cov}(X, X+U_c) = \mathrm{Cov}(X,X) = \sigma_x^2; \qquad (2.30)$$

$$\rho_{xw}^2 = \frac{\sigma_{xw}^2}{\sigma_x^2 \sigma_w^2} = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_{u_c}^2}. \tag{2.31}$$

From (2.22), we get that

$$\mathrm{Cov}(Y, W) = \beta_1 \sigma_{xw} = \beta_1 \sigma_x^2. \tag{2.32}$$

Hence, from (2.24) and (2.25),

$$\beta_1^* = \beta_1 \frac{\sigma_x^2}{\sigma_x^2 + \sigma_{u_c}^2}; \tag{2.33}$$

$$\beta_0^* = \beta_0 + \beta_1 \mu_x - \beta_1 \frac{\sigma_x^2}{\sigma_x^2 + \sigma_{u_c}^2} \mu_x = \beta_0 + \beta_1 (1 - \rho_{xw}^2). \tag{2.34}$$

The variance of the residual has the same form as (2.27). If the error model is classical, with known or estimable λ, the bias in the parameters can be removed.

2.3.2.3 Berkson Error Model

The measurement error structure is

$$X = W + U_b, \tag{2.35}$$

where the error U_b is independent of W, and has mean zero. Then

$$\mu_w = \mu_x, \qquad \sigma_w^2 = \sigma_x^2 - \sigma_{u_b}^2; \tag{2.36}$$

and

$$\sigma_{xw} = \mathrm{Cov}(X, W) = \mathrm{Cov}(W + U_b, W) = \mathrm{Cov}(W, W) = \sigma_w^2; \tag{2.37}$$

$$\rho_{xw}^2 = \frac{\sigma_{xw}^2}{\sigma_x^2 \sigma_w^2} = \frac{\sigma_x^2 - \sigma_{u_b}^2}{\sigma_x^2}. \tag{2.38}$$

From (2.22), we get that

$$\mathrm{Cov}(Y, W) = \beta_1 \sigma_{xw} = \beta_1 \sigma_w^2 = \beta_1 (\sigma_x^2 - \sigma_{u_b}^2). \tag{2.39}$$

Hence, follows from (2.24) and (2.25),

$$\beta_1^* = \beta_1 \frac{\sigma_w^2}{\sigma_w^2} = \beta_1; \tag{2.40}$$

$$\beta_0^* = \beta_0 + \beta_1 \mu_x - \beta_1 \mu_w = \beta_0. \tag{2.41}$$

The variance of the residual has the same form as (2.27). The regression parameters are not biased for the original parameters. However, the residual variance is still greater than σ_ϵ^2.

2.3.2.4 Berkson/Classical Mixture Error Model

We now consider an error model that contains both classical and Berkson error components. Assume that

$$X = L + U_b, \tag{2.42}$$

$$W = L + U_c, \tag{2.43}$$

where U_b denotes the Berkson error, and U_c denotes the classical error. They are both independent of L and have mean zero. We also assume that U_b and U_c are independent. In this special case, when $U_b = 0$, then $X = L$, so the classical error model is obtained as a result. On the other hand, if $U_c = 0$, then the error model is Berkson. This error model has features of both the classical and Berkson error models. Under this error model,

$$\mu_x = \mu_w = \mu_l, \qquad \sigma_x^2 = \sigma_l^2 + \sigma_{u_b}^2, \qquad \sigma_w^2 = \sigma_l^2 + \sigma_{u_c}^2; \tag{2.44}$$

and

$$\sigma_{xw} = \mathrm{Cov}(X, W) = \mathrm{Cov}(L + U_b, L + U_c) = \mathrm{Cov}(L, L) = \sigma_l^2; \tag{2.45}$$

$$\rho_{xw}^2 = \frac{\sigma_{xw}^2}{\sigma_x^2 \sigma_w^2} = \frac{\sigma_l^4}{(\sigma_l^2 + \sigma_{u_b}^2)(\sigma_l^2 + \sigma_{u_c}^2)}. \tag{2.46}$$

From (2.22), we get that

$$\mathrm{Cov}(Y, W) = \beta_1 \sigma_{xw} = \beta_1 \sigma_l^2. \tag{2.47}$$

Hence, follows from (2.24) and (2.25)

$$\beta_1^* = \beta_1 \frac{\sigma_l^2}{\sigma_l^2 + \sigma_{u_c}^2}; \tag{2.48}$$

$$\beta_0^* = \beta_0 + \beta_1 \mu_x \left(1 - \frac{\sigma_l^2}{\sigma_l^2 + \sigma_{u_c}^2}\right). \tag{2.49}$$

Note that if $\sigma_{u_c}^2 > 0$, there is bias in the slope of the regression model, as in the classical error model. The form of the residual variance is still the same, but now ρ_{xw} has the effects of both U_b and U_c in it.

2.3.2.5 Differential

Despite the previous discussion, in this section the error is assumed to be differential. As defined in Section 2.2, this means that the observed data W contains some additional information which is not in the true X. In this case, other than the covariance between W and X, the covariance between W and ϵ, $\sigma_{w\epsilon}$, is also required to correct the bias in the parameters. The model we are trying to estimate is still (2.23), but the covariance between Y and W is

$$\text{Cov}(Y, W) = \text{Cov}(\beta_0 + \beta_1 X + \epsilon, W) = \beta_1 \sigma_{xw} + \sigma_{\epsilon w}. \tag{2.50}$$

Then, using (2.24) and (2.25),

$$\beta_1^* = \frac{\text{Cov}(Y, W)}{\text{Cov}(W, W)} = \frac{\beta_1 \sigma_{xw} + \sigma_{\epsilon w}}{\sigma_w^2} = \beta_1 \frac{\sigma_{xw}}{\sigma_w^2} + \frac{\sigma_{\epsilon w}}{\sigma_w^2}; \tag{2.51}$$

$$\beta_0^* = \beta_0 + \beta_1 \mu_x - \frac{\beta_1 \sigma_{xw} + \sigma_{\epsilon w}}{\sigma_w^2} \mu_w. \tag{2.52}$$

Since W is not conditionally independent of ϵ given X, in correcting the bias, $\sigma_{w\epsilon}$ is also needed. Notice that the residual variance is

$$\begin{aligned}
\mathrm{Var}(Y|W) &= \mathrm{Var}(\eta|W) = \mathrm{Var}(\eta) \\
&= \sigma_y^2 \left(1 - \rho_{yw}^2\right) \\
&= \sigma_y^2 \left(1 - \frac{(\beta_1 \sigma_{xw} + \sigma_{\epsilon w})^2}{\sigma_y^2 \sigma_w^2}\right) \\
&= \sigma_\epsilon^2 + \beta_1^2 \sigma_x^2 - \frac{(\beta_1 \sigma_{xw} + \sigma_{\epsilon w})^2}{\sigma_w^2}. \qquad (2.53)
\end{aligned}$$

The problem is that this variance can be smaller than the original residual σ_ϵ^2. This is dangerous because this means that the data (Y, W) can be more precise to the wrong model with bias.

2.3.3 Summary of Simple Linear Regression

We summarize all the previous results about different types of measurement error structures discussed in Section 2.3.2 in the table below. The same table can be found in Carroll et al. [2]. The order of the error models in Table 2.1 are arranged from most to least problematic in terms of the negative effects caused by measurement error. All the assumptions are stated in Section 2.3.2.

Table 2.1: Summary of Simple Linear Regression

Error Model	ρ^2_{xw}	Intercept (β_0^*)	Slope (β_1^*)	Residual Variance (Var($Y\|W$))
Differential	ρ^2_{xw}	$\beta_0 + \beta_1\mu_x - \frac{\beta_1\sigma_{xw}+\sigma_{\epsilon w}}{\sigma^2_w}\mu_w$	$\beta_1\frac{\sigma_{xw}}{\sigma^2_w} + \frac{\sigma_{\epsilon w}}{\sigma^2_w}$	$\sigma^2_\epsilon + \beta_1^2\sigma^2_x - \frac{(\beta_1\sigma_{xw}+\sigma_{\epsilon w})^2}{\sigma^2_w}$
Nondifferential	ρ^2_{xw}	$\beta_0 + \beta_1\mu_x - \frac{\sigma_{xw}}{\sigma^2_w}\beta_1\mu_w$	$\frac{\sigma_{xw}}{\sigma^2_w}\beta_1$	$\sigma^2_\epsilon + \beta_1^2\sigma^2_x(1-\rho^2_{xw})$
Classical	$\frac{\sigma^2_x}{\sigma^2_x+\sigma^2_{u_c}}$	$\beta_0 + \beta_1(1-\rho^2_{xw})$	$\beta_1\frac{\sigma^2_x}{\sigma^2_x+\sigma^2_{u_c}}$	$\sigma^2_\epsilon + \beta_1^2\sigma^2_x(1-\rho^2_{xw})$
B/C Mixture	$\frac{\sigma^4_l(\sigma^2_l+\sigma^2_{u_b})^{-1}}{(\sigma^2_l+\sigma^2_{u_c})}$	$\beta_0 + \beta_1\mu_x\left(1-\frac{\sigma^2_l}{\sigma^2_l+\sigma^2_{u_c}}\right)$	$\beta_1\frac{\sigma^2_l}{\sigma^2_l+\sigma^2_{u_c}}$	$\sigma^2_\epsilon + \beta_1^2\sigma^2_x(1-\rho^2_{xw})$
Berkson	$\frac{\sigma^2_x-\sigma^2_{u_b}}{\sigma^2_x}$	β_0	β_1	$\sigma^2_\epsilon + \beta_1^2\sigma^2_x(1-\rho^2_{xw})$
No Error	1	β_0	β_1	σ^2_ϵ

Chapter 3

Variable Selection for Partially Linear Models with Adaptive LASSO

In this chapter, we will consider the partially linear regression model with the additive error structure. We prove that unbiasedness, sparsity and asymptotic normality hold for the adaptive LASSO. We state the relevant results in the following theorems and include the proofs in this chapter.

3.1 The Model, Assumptions and Notations

Suppose that $\{(\boldsymbol{W}_i, Z_i, Y_i), i = 1, \ldots, n\}$ is a random sample from the partially linear measurement error model:

$$Y = \boldsymbol{X}^T\boldsymbol{\beta} + v(Z) + \epsilon, \tag{3.1}$$

$$\boldsymbol{W} = \boldsymbol{X} + \boldsymbol{U}. \tag{3.2}$$

In the model, Z is a univariate covariate which is observed error-free. $\boldsymbol{X}$ is a $d \times 1$ vector representing the true unobserved predictor and it is measured subject to an additive error. $\boldsymbol{W}$ is the observed

surrogate of $\boldsymbol{X}$, and it is nondifferential, i.e., $\boldsymbol{W}$ is independent of Y given $(\boldsymbol{X}, Z)$. ϵ is the error in the underlying model, and it is independent of $(\boldsymbol{X}, Z)$ with a zero mean. $\boldsymbol{U}$ is the measurement error of $\boldsymbol{X}$. It is independent of $\boldsymbol{X}$ with a zero mean, and $\mathrm{E}[\boldsymbol{U}|Z] = 0$. In the following, we assume that $\boldsymbol{U}$ has a known covariance matrix Σ_{uu}. Σ_{uu} can be singular, so the covariate vector X may consist of some error-free variables. Under the assumption of nondifferential error, it can be shown that $\mathrm{Cov}(\epsilon, \boldsymbol{U}) = 0$.

It is worth noticing that, our assumptions on $\boldsymbol{X}$, $\boldsymbol{Y}$,$\boldsymbol{Z}$ and $\boldsymbol{W}$ are weaker than those in Liang and Li [11]. Liang and Li [11] required that $\boldsymbol{U}$ be independent of $(\boldsymbol{X}, Z, Y)$. However, we believe that this assumption is too strong because in the real world, it is rare that we have such independence conditions. Actually, the nondifferential error assumption is adequate for the oracle properties to hold in the penalized least squares. We note that $\boldsymbol{W}$ being a surrogate of $\boldsymbol{X}$ does not imply that $\boldsymbol{U}$ is independent of $(\boldsymbol{X}, Z, Y)$, and it is also natural in practice. Define the least squares function as

$$\sum_{i=1}^{n}\{Y_i - \boldsymbol{W}_i^T\boldsymbol{\beta} - v(Z_i)\}^2 - n\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{uu}\boldsymbol{\beta}. \tag{3.3}$$

The second term in (3.3) is needed to correct the bias in the squared loss function due to measurement error. It can be shown that with this term, the expectation of $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^2$ is zero. The nonparametric part in this model is an unknown function of Z, and needs to be estimated. Since $\mathrm{E}[\boldsymbol{U}|Z] = 0$, $\mathrm{E}[\boldsymbol{W}|Z] = \mathrm{E}[\boldsymbol{X}|Z]$. Taking conditional expectation given Z on both sides of (3.1), we have

$$v(Z) = \mathrm{E}[Y|Z] - \mathrm{E}[\boldsymbol{W}|Z]^T\boldsymbol{\beta}. \tag{3.4}$$

Plug (3.4) for $v(Z)$ in the least squares function (3.3) to obtain

$$\sum_{i=1}^{n}\{(Y_i - \mathrm{E}[Y_i|Z_i]) - (\boldsymbol{W}_i - \mathrm{E}[\boldsymbol{W}_i|Z_i])^T\boldsymbol{\beta}\}^2 - n\boldsymbol{\beta}^T\boldsymbol{\Sigma}_{uu}\boldsymbol{\beta}. \tag{3.5}$$

The conditional expectations $\mathrm{E}[Y_i|Z_i]$ and $\mathrm{E}[\boldsymbol{W}_i|Z_i]$ can be estimated by local linear regression. Denote the local linear regression estimators by $\widehat{\mathrm{E}}[Y_i|Z_i]$ and $\widehat{\mathrm{E}}[\boldsymbol{W}_i|Z_i]$ respectively, and denote

$$\widetilde{Y}_i = Y_i - \mathrm{E}[Y_i|Z_i]; \qquad \widehat{Y}_i = Y_i - \widehat{\mathrm{E}}[Y_i|Z_i]; \tag{3.6}$$

$$\widetilde{\boldsymbol{W}}_i = \boldsymbol{W}_i - \mathrm{E}[\boldsymbol{W}_i|Z_i]; \qquad \widehat{\boldsymbol{W}}_i = \boldsymbol{W}_i - \widehat{\mathrm{E}}[\boldsymbol{W}_i|Z_i]. \tag{3.7}$$

Using the notations above, (3.5) can be rewritten as

$$l(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \{\widehat{Y}_i - \widehat{\boldsymbol{W}}_i^T \boldsymbol{\beta}\}^2 - n\boldsymbol{\beta}^T \boldsymbol{\Sigma}_{uu} \boldsymbol{\beta}. \tag{3.8}$$

We define the penalized least squares function with adaptive LASSO as

$$\mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta}) = \sum_{i=1}^{n} (\widehat{Y}_i - \widehat{\boldsymbol{W}}_i^T \boldsymbol{\beta})^2 - n\boldsymbol{\beta}^T \boldsymbol{\Sigma}_{uu} \boldsymbol{\beta} + \lambda_n \sum_{j=1}^{d} \widehat{w}_j \, |\beta_j|, \tag{3.9}$$

where $\widehat{w}_j$ is the adaptive weight for the j-th parameter β_j. For $\gamma > 0$, the weight is defined as

$$\hat{w}_j = \frac{1}{|\hat{\beta}_j|^\gamma}. \tag{3.10}$$

In the original adaptive LASSO, $\hat{\beta}_j$ is the estimator obtained from the ordinary least squares. In this thesis, with the presence of the measurement error, we use the least squares function with the bias correction term, $l(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta})$, to obtain the $\hat{\beta}_j$, which is also an unbiased estimator. λ_n is the regularization parameter in adaptive LASSO penalty, and it varies with n. We use the λ_n with the same rate as Zou [17] when we prove the oracle properties of the adaptive LASSO estimator of $\boldsymbol{\beta}$. For $\gamma > 0$, we assume that as $n \to \infty$,

$$\frac{\lambda_n}{\sqrt{n}} \to 0 \quad \text{and} \quad \lambda_n \cdot n^{(\gamma-1)/2} \to \infty. \tag{3.11}$$

Without loss of generality, denote the true parameter as

$$\boldsymbol{\beta}_0 = (\beta_{01}, \ldots, \beta_{0s}, \beta_{0(s+1)}, \ldots, \beta_{0d})^T = \begin{pmatrix} \boldsymbol{\beta}_{01} \\ \boldsymbol{\beta}_{02} \end{pmatrix}, \tag{3.12}$$

where $\boldsymbol{\beta}_{02} = (\beta_{0(s+1)}, \ldots, \beta_{0d})^T = \mathbf{0}^T$, and $\boldsymbol{\beta}_{01}$ is a $s \times 1$ vector containing all the nonzero components in $\boldsymbol{\beta}_0$.

3.2 The Oracle Properties of Adaptive LASSO

In this section, the oracle properties of adaptive LASSO estimator for this partially linear regression model with additive error will be proved. We assume that the regularity conditions hold through out this section. The penalized least squares function, $\mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta})$ is defined as (3.9), with the weight and the regularization parameter defined as (3.10) and (3.11) respectively.

3.2.1 Consistency

We first claim the consistency property of the estimator from the penalized least squares.

Theorem 3.2.1. *As $n \to \infty$, with probability approaching one, there exists a local minimizer $\widehat{\boldsymbol{\beta}}$ of $\mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta})$ such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-\frac{1}{2}})$.*

It is clear from Theorem 3.2.1 that by choosing an appropriate regularization parameter λ_n, there exists a root-n consistent adaptive LASSO estimator. To prove Theorem 3.2.1, we need the following proposition and lemma. They are also stated in Liang et al [11].

Proposition 3.2.2. $\sup_z |\hat{m}_y(z) - m_y(z)| = o_p(n^{-\frac{1}{4}})$ *and* $\sup_z |\hat{m}_{w,j}(z) - m_{w,j}(z)| = o_p(n^{-\frac{1}{4}})$.

Lemma 3.2.3. *Assume that random variables $a_i(\boldsymbol{W_i}, Z_i, Y_i)$ and $c_i(\boldsymbol{W_i}, Z_i, Y_i)$, denoted by a_i and c_i, satisfy $a_i(\boldsymbol{W_i}, Z_i, Y_i) \equiv 1$ or $\mathrm{E}[a_i] = 0$ and $\max_{1 \le i \le n} |c_i| = o_p(n^{-\frac{1}{4}})$. Then*

$$\sum_{i=1}^{n} a_i c_i w_i = o_p(n^{\frac{1}{2}}),$$

where w_i are independent variables with mean zero and finite variance.

Now we prove Theorem 3.2.1.

Proof. Let $\alpha_n = n^{-\frac{1}{2}}$. To prove the theorem, we show that, for any given $\epsilon > 0$, there exists a large constant C such that

$$P\left\{ \inf_{||\boldsymbol{v}||=C} \mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta}_0 + \alpha_n \boldsymbol{v}) > \mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta}_0) \right\} \ge 1 - \epsilon. \tag{3.13}$$

This implies that with probability at least $1-\epsilon$, there exists a local minimum in the ball $\{\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{v} : \|\boldsymbol{v}\| \le C\}$. Hence, there exists a local minimizer $\widehat{\boldsymbol{\beta}}$ such that $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-\frac{1}{2}})$. The theorem

then follows. Define

$$D_n(\boldsymbol{v}) = \mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta}_0 + \alpha_n \boldsymbol{v}) - \mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta}_0). \tag{3.14}$$

Then

$$\begin{aligned} D_n(\boldsymbol{v}) = &\sum_{i=1}^{n} \left[\left(\widehat{Y}_i - \widehat{\boldsymbol{W}}_i^T (\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{v}) \right)^2 - \left(\widehat{Y}_i - \widehat{\boldsymbol{W}}_i^T \boldsymbol{\beta}_0 \right)^2 \right] \\ &- n \left\{ (\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{v})^T \boldsymbol{\Sigma}_{uu} (\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{v}) - \boldsymbol{\beta}_0^T \boldsymbol{\Sigma}_{uu} \boldsymbol{\beta}_0 \right\} + \lambda_n \sum_{j=1}^{d} \widehat{w}_j \left(|\beta_{0j} + \alpha v_j| - |\beta_{0j}| \right). \end{aligned} \tag{3.15}$$

Let $J_n(\boldsymbol{v})$ represent the first two terms in $D_n(\boldsymbol{v})$;

$$\begin{aligned} J_n(\boldsymbol{v}) = &\sum_{i=1}^{n} \left[\left(\widehat{Y}_i - \widehat{\boldsymbol{W}}_i^T (\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{v}) \right)^2 - \left(\widehat{Y}_i - \widehat{\boldsymbol{W}}_i^T \boldsymbol{\beta}_0 \right)^2 \right] \\ &- n \left\{ (\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{v})^T \boldsymbol{\Sigma}_{uu} (\boldsymbol{\beta}_0 + \alpha_n \boldsymbol{v}) - \boldsymbol{\beta}_0^T \boldsymbol{\Sigma}_{uu} \boldsymbol{\beta}_0 \right\}. \end{aligned} \tag{3.16}$$

Then $J_n(\boldsymbol{v})$ can be rewritten as

$$\begin{aligned} J_n(\boldsymbol{v}) = &-2\alpha_n \sum_{i=1}^{n} \left(\widehat{Y}_i \widehat{\boldsymbol{W}}_i^T - \widehat{\boldsymbol{W}}_i^T \boldsymbol{\beta}_0 \widehat{\boldsymbol{W}}_i^T + \boldsymbol{\beta}_0^T \boldsymbol{\Sigma}_{uu} \right) \boldsymbol{v} \\ &+ n\alpha_n^2 \boldsymbol{v}^T \left(n^{-1} \sum_{i=1}^{n} \widehat{\boldsymbol{W}}_i \widehat{\boldsymbol{W}}_i^T - \boldsymbol{\Sigma}_{uu} \right) \boldsymbol{v}. \end{aligned} \tag{3.17}$$

We then calculate the order of the first term in (3.17). Note that the summation in the first term can be written as

$$\begin{aligned} &\sum_{i=1}^{n} \left(\widehat{Y}_i \widehat{\boldsymbol{W}}_i^T - \widehat{\boldsymbol{W}}_i^T \boldsymbol{\beta}_0 \widehat{\boldsymbol{W}}_i^T + \boldsymbol{\beta}_0^T \boldsymbol{\Sigma}_{uu} \right) \\ &= \sum_{i=1}^{n} \left(\widehat{Y}_i - \widetilde{Y}_i \right) \left(\widehat{\boldsymbol{W}}_i - \widetilde{\boldsymbol{W}}_i \right)^T + \sum_{i=1}^{n} \left(\widehat{Y}_i - \widetilde{Y}_i \right) \widetilde{\boldsymbol{W}}_i^T \\ &+ \sum_{i=1}^{n} \widetilde{Y}_i \left(\widehat{\boldsymbol{W}}_i - \widetilde{\boldsymbol{W}}_i \right)^T + \sum_{i=1}^{n} \left(\widetilde{Y}_i \widetilde{\boldsymbol{W}}_i^T - \widetilde{\boldsymbol{W}}_i^T \boldsymbol{\beta}_0 \widetilde{\boldsymbol{W}}_i^T + \boldsymbol{\beta}_0^T \boldsymbol{\Sigma}_{uu} \right) \\ &- \sum_{i=1}^{n} \left(\widehat{\boldsymbol{W}}_i - \widetilde{\boldsymbol{W}}_i \right)^T \boldsymbol{\beta}_0 \left(\widehat{\boldsymbol{W}}_i - \widetilde{\boldsymbol{W}}_i \right)^T - \sum_{i=1}^{n} \widetilde{\boldsymbol{W}}_i^T \boldsymbol{\beta}_0 \left(\widehat{\boldsymbol{W}}_i - \widetilde{\boldsymbol{W}}_i \right)^T \\ &- \sum_{i=1}^{n} \left(\widehat{\boldsymbol{W}}_i - \widetilde{\boldsymbol{W}}_i \right)^T \boldsymbol{\beta}_0 \widetilde{\boldsymbol{W}}_i^T. \end{aligned} \tag{3.18}$$

We claim that the order of (3.18) is $O_p(n^{\frac{1}{2}})$. The orders of the first and the fifth terms are calculated by Proposition 3.2.2. Let $\boldsymbol{J} = (1, \ldots, 1)^T$. The first term

$$\begin{aligned}
&\sum_{i=1}^{n} \left(\widehat{Y}_i - \widetilde{Y}_i\right) \left(\widehat{\boldsymbol{W}}_i - \widetilde{\boldsymbol{W}}_i\right)^T \\
&\leq n \cdot \sup_z |\widehat{Y}_i - \widetilde{Y}_i| \cdot \sup_z |\widehat{W}_{ij} - \widetilde{W}_{ij}| \cdot \boldsymbol{J}^T \\
&= n \cdot \sup_z |\hat{m}_y(z) - m_y(z)| \cdot \sup_z |\hat{m}_{w,j}(z) - m_{w,j}(z)| \cdot \boldsymbol{J}^T \\
&\leq n \cdot o_p(n^{-\frac{1}{4}}) \cdot o_p(n^{-\frac{1}{4}}) \\
&= o_p(n^{\frac{1}{2}});
\end{aligned}$$

and the fifth term

$$\begin{aligned}
&\sum_{i=1}^{n} \left(\widehat{\boldsymbol{W}}_i - \widetilde{\boldsymbol{W}}_i\right)^T \boldsymbol{\beta}_0 \left(\widehat{\boldsymbol{W}}_i - \widetilde{\boldsymbol{W}}_i\right)^T \\
&\leq n \cdot \sup_z |\hat{m}_{w,j}(z) - m_{w,j}(z)| \cdot \boldsymbol{J}^T \cdot \boldsymbol{\beta}_0 \cdot \sup_z |\hat{m}_{w,j}(z) - m_{w,j}(z)| \cdot \boldsymbol{J}^T \\
&\leq n \cdot o_p(n^{-\frac{1}{4}}) \cdot \boldsymbol{\beta}_0 \cdot o_p(n^{-\frac{1}{4}}) \\
&= o_p(n^{\frac{1}{2}});
\end{aligned}$$

The orders of the second, third, sixth and seventh follow from Lemma 3.2.3. For the second term, let $a_i = 1$, $c_i = \widehat{Y}_i - \widetilde{Y}_i = \widehat{\mathrm{E}}[Y_i|Z_i] - \mathrm{E}[Y_i|Z_i] = o_p(n^{-\frac{1}{4}})$, and $\widetilde{\boldsymbol{W}}_i$ are independent with a zero mean, so that

$$\sum_{i=1}^{n} (\widehat{Y}_i - \widetilde{Y}_i) \widetilde{\boldsymbol{W}}_i^T = o_p(n^{\frac{2}{1}}).$$

For the third term, let $a_i = 1$, $c_i = \widehat{\boldsymbol{W}}_i - \widetilde{\boldsymbol{W}}_i = \widehat{\mathrm{E}}[\boldsymbol{W}_i|Z_i] - \mathrm{E}[\boldsymbol{W}_i|Z_i] = o_p(n^{-\frac{1}{4}})$. Now, since $\widetilde{Y}_i$s are independent with zero mean, we have

$$\sum_{i=1}^{n} \widetilde{Y}_i (\widehat{\boldsymbol{W}}_i - \widetilde{\boldsymbol{W}}_i)^T = o_p(n^{\frac{2}{1}}).$$

Now for the sixth term, we let $a_i = 1$, $c_i = \widehat{\boldsymbol{W}}_i - \widetilde{\boldsymbol{W}}_i = \widehat{\mathrm{E}}[\boldsymbol{W}_i|Z_i] - \mathrm{E}[\boldsymbol{W}_i|Z_i] = o_p(n^{-\frac{1}{4}})$, and since

$\widetilde{\boldsymbol{W}}_i^T\boldsymbol{\beta}_0$ are independent with a zero mean, if follows that

$$\sum_{i=1}^{n}\widetilde{\boldsymbol{W}}_i^T\boldsymbol{\beta}_0(\widehat{\boldsymbol{W}}_i-\widetilde{\boldsymbol{W}}_i)^T=o_p(n^{\frac{2}{1}}).$$

In the seventh term, let $a_i=1$, $c_i=\widehat{\boldsymbol{W}}_i-\widetilde{\boldsymbol{W}}_i=\widehat{\mathrm{E}}[\boldsymbol{W}_i|Z_i]-\mathrm{E}[\boldsymbol{W}_i|Z_i]=o_p(n^{-\frac{1}{4}})$, and $\boldsymbol{\beta}_0\widetilde{\boldsymbol{W}}_i^T$ are independent with a zero mean, so that

$$\sum_{i=1}^{n}(\widehat{\boldsymbol{W}}_i-\widetilde{\boldsymbol{W}}_i)^T\boldsymbol{\beta}_0\widetilde{\boldsymbol{W}}_i^T=o_p(n^{\frac{2}{1}}).$$

Finally we calculate the order of the forth term in (3.18). Since $\left(\widetilde{Y}_i\widetilde{\boldsymbol{W}}_i^T-\widetilde{\boldsymbol{W}}_i^T\boldsymbol{\beta}_0\widetilde{\boldsymbol{W}}_i^T\right)$'s are iid, and $\mathrm{E}[\widetilde{Y}_i\widetilde{\boldsymbol{W}}_i^T-\widetilde{\boldsymbol{W}}_i^T\boldsymbol{\beta}_0\widetilde{\boldsymbol{W}}_i^T]<\infty$, by the weak law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\widetilde{Y}_i\widetilde{\boldsymbol{W}}_i^T-\widetilde{\boldsymbol{W}}_i^T\boldsymbol{\beta}_0\widetilde{\boldsymbol{W}}_i^T\right)\xrightarrow{p}\mathrm{E}\left[\widetilde{Y}_i\widetilde{\boldsymbol{W}}_i^T-\widetilde{\boldsymbol{W}}_i^T\boldsymbol{\beta}_0\widetilde{\boldsymbol{W}}_i^T\right].$$

Now, note that

$$\begin{aligned}
\mathrm{E}\left[\widetilde{Y}_i\widetilde{\boldsymbol{W}}_i^T-\widetilde{\boldsymbol{W}}_i^T\boldsymbol{\beta}_0\widetilde{\boldsymbol{W}}_i^T\right]&=\mathrm{E}\left[\left(\widetilde{Y}_i-\widetilde{\boldsymbol{W}}_i^T\boldsymbol{\beta}_0\right)\widetilde{\boldsymbol{W}}_i^T\right]\\
&=\mathrm{E}\left[\left(Y_i-\mathrm{E}[Y_i|Z_i]-(\boldsymbol{W}_i-\mathrm{E}[\boldsymbol{W}_i|Z_i])^T\boldsymbol{\beta}_0\right)\widetilde{\boldsymbol{W}}_i^T\right]\\
&=\mathrm{E}\left[\left(Y_i-\mathrm{E}[Y_i|Z_i]-(\boldsymbol{X}_i+\boldsymbol{U}_i-\mathrm{E}[\boldsymbol{X}_i|Z_i])^T\boldsymbol{\beta}_0\right)\widetilde{\boldsymbol{W}}_i^T\right]\\
&=\mathrm{E}\left[\left(Y_i-\boldsymbol{X}_i^T\boldsymbol{\beta}_0-\left(\mathrm{E}[Y_i|Z_i]-\mathrm{E}[\boldsymbol{X}_i|Z_i]^T\boldsymbol{\beta}_0\right)-\boldsymbol{U}_i^T\boldsymbol{\beta}_0\right)\widetilde{\boldsymbol{W}}_i^T\right]\\
&=\mathrm{E}\left[\left(v(Z_i)+\epsilon_i-\mathrm{E}\big[v(Z_i)+\epsilon_i\big|Z\big]-\boldsymbol{U}_i^T\boldsymbol{\beta}_0\right)\widetilde{\boldsymbol{W}}_i^T\right]\\
&=\mathrm{E}\left[\left(v(Z_i)+\epsilon_i-\left(v(Z_i)+\mathrm{E}[\epsilon_i|Z_i]\right)-\boldsymbol{U}_i^T\boldsymbol{\beta}_0\right)\widetilde{\boldsymbol{W}}_i^T\right]\\
&=\mathrm{E}\left[\left(\epsilon_i-0-\boldsymbol{U}_i^T\boldsymbol{\beta}_0\right)\widetilde{\boldsymbol{W}}_i^T\right]\\
&=\mathrm{E}\left[\epsilon_i\widetilde{\boldsymbol{W}}_i^T-\boldsymbol{U}_i^T\boldsymbol{\beta}_0\widetilde{\boldsymbol{W}}_i^T\right] \qquad (3.19)
\end{aligned}$$

We know that $\mathrm{Cov}(\boldsymbol{W},\epsilon)=0$, Z is independent of ϵ, and $\mathrm{E}[\epsilon]=0$. Hence the RHS of (3.19) further

reduces to

$$\begin{aligned}
&\mathrm{E}[\epsilon_i]\mathrm{E}\left[\widetilde{\boldsymbol{W}}_i^T\right] - \mathrm{E}\left[\boldsymbol{U}_i^T\boldsymbol{\beta}_0\widetilde{\boldsymbol{W}}_i^T\right] \\
&= 0 - \mathrm{E}\left[\boldsymbol{U}_i^T\boldsymbol{\beta}_0\left(\boldsymbol{X}_i + \boldsymbol{U}_i - \mathrm{E}[\boldsymbol{X}_i|Z_i]\right)^T\right] \\
&= -\mathrm{E}\left[\boldsymbol{U}_i^T\boldsymbol{\beta}_0\boldsymbol{X}_i^T + \boldsymbol{U}_i^T\boldsymbol{\beta}_0\boldsymbol{U}_i^T - \boldsymbol{U}_i^T\boldsymbol{\beta}_0\mathrm{E}[\boldsymbol{X}_i|Z_i]^T\right] \\
&= -\mathrm{E}[\boldsymbol{U}_i^T\boldsymbol{\beta}_0\boldsymbol{X}_i^T] + \mathrm{E}[\boldsymbol{U}_i^T\boldsymbol{\beta}_0\boldsymbol{U}_i^T] - \mathrm{E}[\boldsymbol{U}_i^T\boldsymbol{\beta}_0\mathrm{E}[\boldsymbol{X}_i|Z_i]^T]. \qquad (3.20)
\end{aligned}$$

Note that $\boldsymbol{U}$ is independent of $\boldsymbol{X}$, and $\mathrm{E}[\boldsymbol{U}] = 0$ so that the first term of the RHS of (3.20) is $-\mathrm{E}[\boldsymbol{U}_i^T]\boldsymbol{\beta}_0\mathrm{E}[\boldsymbol{X}_i^T] = 0$. We also have $\mathrm{E}[\boldsymbol{U}|Z] = 0$, and therefore $\mathrm{Cov}(\boldsymbol{U}, Z) = \mathrm{E}[\boldsymbol{U}Z] = \mathrm{E}\big[\mathrm{E}[\boldsymbol{U}Z|Z]\big] = \mathrm{E}\big[Z\,\mathrm{E}[\boldsymbol{U}|Z]\big] = 0$. The last term in the RHS of (3.20) is $\mathrm{E}\left[\mathrm{E}\left[\boldsymbol{U}_i^T\boldsymbol{\beta}_0\mathrm{E}[\boldsymbol{X}_i|Z_i]^T\right]|Z_i\right] = \mathrm{E}\left[\mathrm{E}[\boldsymbol{U}_i^T|Z_i]\boldsymbol{\beta}_0\mathrm{E}[\boldsymbol{X}_i|Z_i]^T\right] = 0$. Hence (3.20) becomes

$$0 + \mathrm{E}[\boldsymbol{U}_i^T\boldsymbol{\beta}_0\boldsymbol{U}_i^T] - 0 = \mathrm{E}[\boldsymbol{U}_i^T\boldsymbol{\beta}_0\boldsymbol{U}_i^T] = -\boldsymbol{\beta}_0^T\Sigma_{uu}.$$

Hence,

$$\mathrm{E}[\widetilde{Y}_i\widetilde{\boldsymbol{W}}_i^T - \widetilde{\boldsymbol{W}}_i^T\boldsymbol{\beta}_0\widetilde{\boldsymbol{W}}_i^T + \boldsymbol{\beta}_0^T\boldsymbol{\Sigma}_{uu}] = -\boldsymbol{\beta}_0^T\Sigma_{uu} + \boldsymbol{\beta}_0^T\Sigma_{uu} = 0 \qquad (3.21)$$

Hence, as n goes to infinity, the fourth term goes to 0 in probability, and from the central limit theorem, it is of the order $nO_p(n^{-\frac{1}{2}}) = O_p(n^{\frac{1}{2}})$. Thus, taking $||\boldsymbol{v}|| = C$, the order of the first term of $J_n(\boldsymbol{v})$ is

$$\alpha_n O_p(n^{\frac{1}{2}}) = O_p(\alpha_n n^{\frac{1}{2}}) = O_p(n^{-\frac{1}{2}} \cdot n^{\frac{1}{2}}) = O_p(1).$$

Now we consider the second term in (3.17). Note that

$$\widehat{\boldsymbol{W}}_i = \boldsymbol{W}_i - \mathrm{E}[\boldsymbol{W}_i|Z_i] + E[\boldsymbol{W}_i|Z_i] - \widehat{\mathrm{E}}[\boldsymbol{W}_i|Z_i];$$

so that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{W}}_i^T\widehat{\boldsymbol{W}}_i \\
&= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{W}_i - \mathrm{E}[\boldsymbol{W}_i|Z_i] + E[\boldsymbol{W}_i|Z_i] - \widehat{\mathrm{E}}[\boldsymbol{W}_i|Z_i])^T(\boldsymbol{W}_i - \mathrm{E}[\boldsymbol{W}_i|Z_i] + E[\boldsymbol{W}_i|Z_i] - \widehat{\mathrm{E}}[\boldsymbol{W}_i|Z_i]) \\
&= \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{W}_i - \mathrm{E}[\boldsymbol{W}_i|Z_i])^T(\boldsymbol{W}_i - \mathrm{E}[\boldsymbol{W}_i|Z_i]) \\
&\quad + \frac{2}{n}\sum_{i=1}^{n}(\boldsymbol{W}_i - \mathrm{E}[\boldsymbol{W}_i|Z_i])^T(\mathrm{E}[\boldsymbol{W}_i|Z_i] - \widehat{\mathrm{E}}[\boldsymbol{W}_i|Z_i]) \qquad (3.22) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(\mathrm{E}[\boldsymbol{W}_i|Z_i] - \widehat{\mathrm{E}}[\boldsymbol{W}_i|Z_i])^T(\mathrm{E}[\boldsymbol{W}_i|Z_i] - \widehat{\mathrm{E}}[\boldsymbol{W}_i|Z_i]).
\end{aligned}
$$

Recall that $\widetilde{\boldsymbol{W}}_i = \boldsymbol{W}_i - \mathrm{E}[\boldsymbol{W}_i|Z_i]$, so that the first term of (3.22) is $\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{W}}_i^T\widetilde{\boldsymbol{W}}_i$. Since $\widetilde{\boldsymbol{W}}_i$'s are iid, by the weak law of large numbers, the first term goes to $\mathrm{E}[\widetilde{\boldsymbol{W}}_i^T\widetilde{\boldsymbol{W}}_i]$ in probability as $n \to \infty$. Furthermore, we have the property that $\sup_z |\mathrm{E}[\boldsymbol{W}_i|Z_i] - \widehat{\mathrm{E}}[\boldsymbol{W}_i|Z_i]| = o_p(n^{-\frac{1}{4}})$. Hence the second term in (3.22) is bounded by $\frac{2}{n}\{\sup_z |\mathrm{E}[\boldsymbol{W}_i|Z_i] - \widehat{\mathrm{E}}[\boldsymbol{W}_i|Z_i]|\}|\sum_{i=1}^{n}\widetilde{\boldsymbol{W}}_i^T| = o_p(n^{-\frac{1}{4}})\,\frac{1}{n}|\sum_{i=1}^{n}\widetilde{\boldsymbol{W}}_i^T|$. Since $\widetilde{\boldsymbol{W}}_i$s are iid with a zero mean and finite variance, the weak law of large numbers gives $\frac{1}{n}\sum_{i=1}^{n}\widetilde{\boldsymbol{W}}_i^T \xrightarrow{P} 0$. This implies that the second term converges in probability to 0. The third term of (3.22) is $\frac{1}{n}\,n\,o_p((n^{-\frac{1}{4}})^2) = o_p(n^{-\frac{1}{2}})$, giving us

$$
\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{W}}_i\widehat{\boldsymbol{W}}_i^T \xrightarrow{p} \mathrm{E}[\widetilde{\boldsymbol{W}}_i\widetilde{\boldsymbol{W}}_i^T],
$$

as $n \to 0$. Furthermore,

$$
\begin{aligned}
&\mathrm{E}\left[\widetilde{\boldsymbol{W}}_i\widetilde{\boldsymbol{W}}_i^T\right] \\
&= \mathrm{E}\left[(\boldsymbol{W}_i - \mathrm{E}[\boldsymbol{W}_i|Z_i])(\boldsymbol{W}_i - \mathrm{E}[\boldsymbol{W}_i|Z_i])^T\right] \\
&= \mathrm{E}\left[\boldsymbol{W}_i\boldsymbol{W}_i^T - \boldsymbol{W}_i\mathrm{E}[\boldsymbol{W}_i|Z_i]^T - \mathrm{E}[\boldsymbol{W}_i|Z_i]\boldsymbol{W}_i + \mathrm{E}[\boldsymbol{W}_i|Z_i]\mathrm{E}[\boldsymbol{W}_i|Z_i]^T\right] \\
&= \mathrm{E}\left[(\boldsymbol{X}_i - \boldsymbol{U}_i)(\boldsymbol{X}_i - \boldsymbol{U}_i)^T - (\boldsymbol{X}_i - \boldsymbol{U}_i)\mathrm{E}[\boldsymbol{X}_i|Z_i]^T - \mathrm{E}[\boldsymbol{X}_i|Z_i](\boldsymbol{X}_i - \boldsymbol{U}_i)^T + \mathrm{E}[\boldsymbol{X}_i|Z_i]\mathrm{E}[\boldsymbol{X}_i|Z_i]^T\right] \\
&= \mathrm{E}\left[\boldsymbol{X}_i\boldsymbol{X}_i^T + \boldsymbol{U}_i\boldsymbol{U}_i - \boldsymbol{X}_i\mathrm{E}[\boldsymbol{X}_i|Z_i]^T - \mathrm{E}[\boldsymbol{X}_i|Z_i]\boldsymbol{X}_i + \mathrm{E}[\boldsymbol{X}_i|Z_i]\mathrm{E}[\boldsymbol{X}_i|Z_i]^T\right] \\
&= \mathrm{E}\left[\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^T\right] + \mathrm{E}\left[\boldsymbol{U}_i\boldsymbol{U}_i^T\right] \\
&= \mathrm{E}\left[\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^T\right] + \boldsymbol{\Sigma}_{uu}.
\end{aligned}
$$

Thus,

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{W}}_i\widehat{\boldsymbol{W}}_i^T \xrightarrow{p} \mathrm{E}\left[\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^T\right] + \boldsymbol{\Sigma}_{uu};$$

so that

$$\frac{1}{n}\sum_{i=1}^{n}\widehat{\boldsymbol{W}}_i\widehat{\boldsymbol{W}}_i^T - \boldsymbol{\Sigma}_{uu} \xrightarrow{p} \mathrm{E}\left[\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^T\right].$$

Taking $\|\boldsymbol{v}\| = C$, the order of the second term of $J_n(\boldsymbol{v})$ is

$$n\alpha_n^2 O_p(1) = O_p(n\alpha_n^2) = O_p(1).$$

Also, note that $\mathrm{E}\left[\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^T\right]$ is a positive definite matrix. We then consider the penalty term of $D_n(\boldsymbol{v})$,

$$\lambda_n\sum_{j=1}^{d}\hat{w}_j(|\beta_{0j}+\alpha_n v_j| - |\beta_{0j}|) \geq \lambda_n\sum_{j=1}^{s}\hat{w}_j(|\beta_{0j}+\alpha_j v_j| - |\beta_{0j}|)$$

By Taylor's expansion, for $\beta_{0j} \neq 0$ and n sufficiently large,

$$|\beta_{0j}+\alpha_n v_j| - |\beta_{0j}| = |\beta_{0j}| + \alpha_n v_j \operatorname{sgn}(\beta_{0j}) - |\beta_{0j}| = \alpha_n v_j \operatorname{sgn}(\beta_{0j}).$$

Thus the penalty portion of $D_n(\boldsymbol{v})$ reduces to

$$\begin{aligned}\lambda_n\sum_{j=1}^{s}\hat{w}_j(|\beta_{0j}+\alpha_j v_j| - |\beta_{0j}|) &= \lambda_n\sum_{j=1}^{s}\frac{1}{|\hat{\beta_{0j}}|^\gamma}\alpha_n\, v_j \operatorname{sgn}(\beta_{0j}) \\ &= \lambda_n\, n^{-\frac{1}{2}}\cdot\sum_{j=1}^{s}\frac{1}{|\hat{\beta_{0j}}|^\gamma}\, v_j \operatorname{sgn}(\beta_{0j}). \qquad (3.23)\end{aligned}$$

Denote $\widehat{\boldsymbol{\beta}}_0^*$, whose j-th component is

$$\widehat{\beta}_{0j}^* = \frac{\operatorname{sgn}(\beta_{0j})}{|\widehat{\beta}_{0j}|^\gamma}. \qquad (3.24)$$

By Cauchy-Schwartz inequality, and taking $\|\boldsymbol{v}\| = C$, the RHS of (3.23) is bounded by

$$\frac{\lambda_n}{n^{\frac{1}{2}}}\cdot\|\boldsymbol{v}\|\cdot\|\widehat{\boldsymbol{\beta}}_0^*\| = \frac{\lambda_n}{n^{\frac{1}{2}}}\cdot C\cdot\sum_{j=1}^{s}\frac{1}{|\hat{\beta}_{0j}|^{2\gamma}}. \qquad (3.25)$$

Since $\lambda = o_p(n^{\frac{1}{2}})$, the RHS of (3.25) is of order $o_p(1)$, and is of less order than that of the first term in (3.17).

In summarizing the above results, combining (3.15) and (3.17), we have

$$\begin{aligned} D_n(\boldsymbol{v}) &= J_n(\boldsymbol{v}) + \lambda_n \sum_{j=1}^{d} \widehat{w}_j \left(|\beta_{0j} + \alpha v_j| - |\beta_{0j}|\right) \\ &= -2\alpha_n \sum_{i=1}^{n} \left(\widehat{Y}_i \widehat{\boldsymbol{W}}_i^T - \widehat{\boldsymbol{W}}_i^T \boldsymbol{\beta}_0 \widehat{\boldsymbol{W}}_i^T + \boldsymbol{\beta}_0^T \boldsymbol{\Sigma}_{uu} \right) \boldsymbol{v} \\ &\quad + n\alpha_n^2 \boldsymbol{v}^T \left(n^{-1} \sum_{i=1}^{n} \widehat{\boldsymbol{W}}_i \widehat{\boldsymbol{W}}_i^T - \boldsymbol{\Sigma}_{uu} \right) \boldsymbol{v} \\ &\quad + \lambda_n \sum_{j=1}^{d} \widehat{w}_j \left(|\beta_{0j} + \alpha v_j| - |\beta_{0j}|\right). \end{aligned} \tag{3.26}$$

From the above analysis of the order of the three terms of $D_n(\boldsymbol{v})$, taking $\|\boldsymbol{v}\| = C$, we can rewrite $D_n(\boldsymbol{v})$ as

$$D_n(\boldsymbol{v}) = \boldsymbol{v}^T \boldsymbol{M} \boldsymbol{v} - \boldsymbol{A}^T \boldsymbol{v}, \tag{3.27}$$

where $\boldsymbol{v}^T \boldsymbol{M} \boldsymbol{v}$ is the second term in (3.26), and $\boldsymbol{M} \xrightarrow{\mathrm{P}} \mathrm{E}\left[\widetilde{\boldsymbol{X}}_i \widetilde{\boldsymbol{X}}_i^T\right]$, which is a positive definite matrix. Here $\boldsymbol{A}$ represents the combination of the first term and third term in (3.26), and thus $\boldsymbol{A} = O_p(1)$. Then, to prove (3.13), we only need to show that for any given ϵ, there exists a large constant C, such that for all $\|\boldsymbol{v}\| = C$,

$$\begin{aligned} P\left\{\mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta}_0 + \alpha_n \boldsymbol{v}) > \mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta}_0)\right\} &= P(D_n(\boldsymbol{v}) > 0) \\ &= P(\boldsymbol{v}^T \boldsymbol{M} \boldsymbol{v} - \boldsymbol{A}^T \boldsymbol{v} > 0) \\ &= P(\boldsymbol{v}^T \boldsymbol{M} \boldsymbol{v} > \boldsymbol{A}^T \boldsymbol{v}) \\ &\geq 1 - \epsilon. \end{aligned}$$

For the random matrix $\boldsymbol{M}$, if it is positive definite, let $\lambda_{\min}$ denote the smallest eigenvalue of $\boldsymbol{M}$,

then $\boldsymbol{v}^T\boldsymbol{M}\boldsymbol{v} \geq \lambda_{\min}\|\boldsymbol{v}\|^2$. Also, $|\boldsymbol{A}^T\boldsymbol{v}| \leq \|\boldsymbol{A}\|\|\boldsymbol{v}\|$. Hence

$$\begin{aligned}
P(\boldsymbol{v}^T\boldsymbol{M}\boldsymbol{v} > \boldsymbol{A}^T\boldsymbol{v}) &= P(\boldsymbol{v}^T\boldsymbol{M}\boldsymbol{v} > \boldsymbol{A}^T\boldsymbol{v},\ \boldsymbol{M} \text{ positive definite}) \\
&\quad + P(\boldsymbol{v}^T\boldsymbol{M}\boldsymbol{v} > \boldsymbol{A}^T\boldsymbol{v},\ \boldsymbol{M} \text{ not positive definite}) \\
&\geq P(\boldsymbol{v}^T\boldsymbol{M}\boldsymbol{v} > \boldsymbol{A}^T\boldsymbol{v},\ \boldsymbol{M} \text{ positive definite}) \\
&\geq P(\lambda_{\min}\|\boldsymbol{v}\|^2 > \|\boldsymbol{A}\|\|\boldsymbol{v}\|,\ \boldsymbol{M} \text{ positive definite}) \\
&= P(\lambda_{\min}\|\boldsymbol{v}\| > \|\boldsymbol{A}\|,\ \boldsymbol{M} \text{ positive definite})
\end{aligned}$$

Let $\lambda^*_{\min}$ denote the smallest eigenvalue of the nonrandom matrix $\mathrm{E}\left[\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^T\right]$. Then $\lambda^*_{\min} > 0$. Since $\boldsymbol{M} \xrightarrow{\mathrm{P}} \mathrm{E}\left[\widetilde{\boldsymbol{X}}_i\widetilde{\boldsymbol{X}}_i^T\right]$, we have $\lambda_{\min} \xrightarrow{\mathrm{P}} \lambda^*_{\min}$. Hence, as long as $\lambda_{\min} > \frac{\lambda^*_{\min}}{2} > 0$, $\boldsymbol{M}$ is positive definite. Then

$$P(\lambda_{\min}\|\boldsymbol{v}\| > \|\boldsymbol{A}\|,\ \boldsymbol{M} \text{ p.d.}) \geq P(\lambda_{\min}\|\boldsymbol{v}\| > \|\boldsymbol{A}\|,\ \lambda_{\min} > \frac{\lambda^*_{\min}}{2}).$$

Since $\lambda_{\min} \xrightarrow{\mathrm{P}} \lambda^*_{\min}$, for any $\epsilon > 0$, there exists a N_1, s.t. for any $n > N_1$

$$P\left(\lambda_{\min} > \frac{\lambda^*_{\min}}{2}\right) > 1 - \frac{\epsilon}{2}.$$

Also, since $\boldsymbol{A} = O_p(1)$, for this $\epsilon > 0$, there exists a M (a function of ϵ) and a N_2, s.t. for any $n > N_2$

$$P\left(\|\boldsymbol{A}\| > M\right) < \frac{\epsilon}{2}.$$

By taking $\|\boldsymbol{v}\| = C$, we have

$$\begin{aligned}
P\left(\lambda_{\min}\|\boldsymbol{v}\| > \|\boldsymbol{A}\|,\ \lambda_{\min} > \frac{\lambda^*_{\min}}{2}\right) &\geq P\left(\lambda_{\min} > \frac{\|\boldsymbol{A}\|}{\|\boldsymbol{v}\|},\ \lambda_{\min} > \frac{\lambda^*_{\min}}{2},\ \|\boldsymbol{A}\| < M\right) \\
&= P\left(\lambda_{\min} > \frac{\|\boldsymbol{A}\|}{C},\ \lambda_{\min} > \frac{\lambda^*_{\min}}{2},\ \|\boldsymbol{A}\| < M\right) \\
&\geq P\left(\frac{\lambda^*_{\min}}{2} > \frac{M}{C},\ \lambda_{\min} > \frac{\lambda^*_{\min}}{2},\ \|\boldsymbol{A}\| < M\right).
\end{aligned}$$

We can choose a large enough C, which is also a function of ϵ, such that for the ϵ and the M we

picked before, $\frac{\lambda^*_{\min}}{2} > \frac{M}{C}$ always holds. Then for any $n > N = \max(N_1, N_2)$,

$$\begin{aligned} P\left(\frac{\lambda^*_{\min}}{2} > \frac{M}{C}, \lambda_{\min} > \frac{\lambda^*_{\min}}{2}, \|\boldsymbol{A}\| < M\right) &= P\left(\lambda_{\min} > \frac{\lambda^*_{\min}}{2}, \|\boldsymbol{A}\| < M\right) \\ &= P\left(\lambda_{\min} > \frac{\lambda^*_{\min}}{2}\right) - P\left(\lambda_{\min} > \frac{\lambda^*_{\min}}{2}, \|\boldsymbol{A}\| > M\right) \\ &\geq P\left(\lambda_{\min} > \frac{\lambda^*_{\min}}{2}\right) - P\left(\|\boldsymbol{A}\| > M\right) \\ &> 1 - \frac{\epsilon}{2} - \frac{\epsilon}{2} = 1 - \epsilon. \end{aligned}$$

In conclusion, for any ϵ, there exists a large constant C, such that

$$P(D_n(\boldsymbol{v}) > 0) > 1 - \epsilon$$

holds uniformly for all $\|\boldsymbol{v}\| = C$. Thus, we prove the argument that for any ϵ, there exists a large constant C, such that

$$P\left\{\inf_{||\boldsymbol{v}||=C} \mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta}_0 + \alpha_n \boldsymbol{v}) > \mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta}_0)\right\} \geq 1 - \epsilon.$$

This implies that for the local minimizer $\widehat{\boldsymbol{\beta}}$,

$$P\left\{\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \leq C n^{-\frac{1}{2}}\right\} \geq 1 - \epsilon.$$

□

3.2.2 Sparsity

We now state the second property of the adaptive LASSO estimator, the sparsity.

Theorem 3.2.4. *As $n \to \infty$, with probability approaching one, the root n consistent estimator $\widehat{\boldsymbol{\beta}}$ in Theorem 3.2.1 satisfies $\widehat{\boldsymbol{\beta}}_2 = 0$.*

Proof. To prove the sparsity, it suffices to show that

$$\mathcal{L}_p\left\{\boldsymbol{\Sigma}_{uu}, \begin{pmatrix} \boldsymbol{\beta}_1 \\ \mathbf{0} \end{pmatrix}\right\} = \min_{\|\boldsymbol{\beta}_2\| \leq \tilde{C} n^{-\frac{1}{2}}} \mathcal{L}_p\left\{\boldsymbol{\Sigma}_{uu}, \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}\right\}, \tag{3.28}$$

where $\boldsymbol{\beta}_1$ corresponds to the s vector $\boldsymbol{\beta}_{01}$ in (3.12), and $\boldsymbol{\beta}_2$ is the corresponding vector of $\boldsymbol{\beta}_{02}$ in (3.12). $\tilde{C}$ is a constant. It then suffices to show that for any given $\boldsymbol{\beta}_1$ satisfying $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| = O_p(n^{-\frac{1}{2}})$ and for some small $\epsilon_n = \widetilde{C} n^{-\frac{1}{2}}$ and $j = s+1, \ldots, d$,

$$\frac{\partial \mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta})}{\partial \beta_j} \begin{cases} > 0, & 0 < \beta_j < \epsilon_n; \\ \\ < 0, & -\epsilon_n < \beta_j < 0. \end{cases} \tag{3.29}$$

Rewrite $\mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta})$ as

$$\begin{aligned} \mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta}) &= \sum_{i=1}^{n} (\widehat{Y}_i - \widehat{\boldsymbol{W}}_i^T \boldsymbol{\beta})^2 - n\boldsymbol{\beta}^T \boldsymbol{\Sigma}_{uu} \boldsymbol{\beta} + \lambda_n \sum_{j=1}^{d} \widehat{w}_j \, |\beta_j| \\ &= \mathcal{M}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{d} \widehat{w}_j \, |\beta_j|. \end{aligned}$$

Then

$$\frac{\partial \mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta})}{\partial \beta_j} = \frac{\partial \mathcal{M}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta})}{\partial \beta_j} + \lambda_n \widehat{w}_j \operatorname{sgn}(\beta_j)$$

By Taylor's expansion, we have

$$\begin{aligned} & \frac{\partial \mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta})}{\partial \beta_j} \\ &= \frac{\partial \mathcal{M}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta})}{\partial \beta_j} + \lambda_n \widehat{w}_j \operatorname{sgn}(\beta_j) \\ &= \frac{\partial \mathcal{M}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta}_0)}{\partial \beta_j} + \sum_{l=1}^{d} \frac{\partial^2 \mathcal{M}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta}_0)}{\partial \beta_j \partial \beta_l} (\beta_l - \beta_{0l}) + \lambda_n \widehat{w}_j \operatorname{sgn}(\beta_j) \\ &= -2 \sum_{i=1}^{n} \left(\widehat{Y}_i - \widehat{\boldsymbol{W}}_i^T \boldsymbol{\beta}_0 \right) \widehat{W}_{ij} - 2n \sum_{l=1}^{d} \beta_l \Sigma_{lj} + \left(\sum_{i=1}^{n} \sum_{l=1}^{d} \left[2\widehat{W}_{il} \widehat{W}_{ij} + 2\Sigma_{lj} \right] \right) (\beta_l - \beta_{0l}) + \lambda_n \widehat{w}_j \operatorname{sgn}(\beta_j) \\ &= -2 \sum_{i=1}^{n} \left[\left(\widehat{Y}_i - \widehat{\boldsymbol{W}}_i^T \boldsymbol{\beta}_0 \right) \widehat{W}_{ij} + \sum_{l=1}^{d} \beta_l \Sigma_{lj} \right] + \left(2 \sum_{i=1}^{n} \sum_{l=1}^{d} \left[\widehat{W}_{il} \widehat{W}_{ij} + \Sigma_{lj} \right] \right) (\beta_l - \beta_{0l}) + \lambda_n \widehat{w}_j \operatorname{sgn}(\beta_j). \end{aligned} \tag{3.30}$$

By the argument in the proof of Theorem 3.2.1, the first term in (3.30) is of the order $O_p(n^{\frac{1}{2}})$. By the assumptions that $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01} = O_p(n^{-\frac{1}{2}})$, and $\boldsymbol{\beta}_2 = O(n^{-\frac{1}{2}})$, the second term is of the order

$nO_p(n^{-\frac{1}{2}}) = O_p(n^{\frac{1}{2}})$. Hence,

$$\begin{aligned}\frac{\partial \mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta})}{\partial \beta_j} &= O_p(n^{\frac{1}{2}}) + \lambda_n \widehat{w}_j \operatorname{sgn}(\beta_j) \\ &= \sqrt{n}\left(O_p(1) + \frac{\lambda_n}{\sqrt{n}}\widehat{w}_j \operatorname{sgn}(\beta_j)\right).\end{aligned}$$

For $j = s+1, \ldots, d$, $\beta_j = 0$, the un-penalized least squares estimator $\widehat{\beta}_j \xrightarrow{\mathrm{P}} \beta_{0j} = 0$ with an order of $O_p(n^{-\frac{1}{2}})$. Hence

$$\widehat{w}_j = \frac{1}{|\widehat{\beta}_j|^\gamma} \xrightarrow{\mathrm{P}} \infty$$

with an order of $O_p\left((n^{\frac{1}{2}})^\gamma\right) = O_p\left(n^{\frac{\gamma}{2}}\right)$ as $n \to \infty$. Then

$$\frac{\lambda_n}{\sqrt{n}} O_p\left(n^{\frac{\gamma}{2}}\right) = O_p\left(\frac{\lambda_n}{\sqrt{n}}\, n^{\frac{\gamma}{2}}\right) = O_p(\lambda_n\, n^{\frac{\gamma-1}{2}})$$

Hence

$$\frac{\partial \mathcal{L}_p(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta})}{\partial \beta_j} = \sqrt{n}\left\{O_p(1) + O_p(\lambda_n\, n^{\frac{\gamma-1}{2}})\operatorname{sgn}(\beta_j)\right\}$$

Whereas $\lambda_n\, n^{\frac{\gamma-1}{2}} \to \infty$ for any $\gamma > 0$, the sign of the derivative is completely determined by that of β_j. Hence equation (3.29) follows. □

3.2.3 Asymptotic Normality

In this section, for convenience, we define $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{A}^{\otimes 2}$.

Theorem 3.2.5. *Further assume that each component in $\boldsymbol{W}_i$ and Z_i have finite fourth moments. As $n \to \infty$, with probability approaching one, the nonzero part $\widehat{\boldsymbol{\beta}}_1$ of the root n consistent estimator $\widehat{\boldsymbol{\beta}}$ in Theorem 3.2.1, has an asymptotic normal distribution*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) \xrightarrow{\mathcal{D}} N\left(0, \boldsymbol{V}_{\boldsymbol{\beta}_{01}}^{-1} \mathrm{E}[\boldsymbol{\psi}_i(\boldsymbol{\beta}_{01})\boldsymbol{\psi}_i(\boldsymbol{\beta}_{01})^T](\boldsymbol{V}_{\boldsymbol{\beta}_{01}}^{-1})^T\right),$$

where

$$\boldsymbol{V}_{\boldsymbol{\beta}_{01}} = 2\mathrm{Cov}(\widetilde{\boldsymbol{X}}_{i1}),$$

and

$$\mathrm{E}[\boldsymbol{\psi}_i(\boldsymbol{\beta}_{01})\boldsymbol{\psi}_i(\boldsymbol{\beta}_{01})^T] = 4\mathrm{E}\left[\left(\widetilde{\boldsymbol{X}}_{i1}(\epsilon_i - \boldsymbol{U}_{i1}^T\boldsymbol{\beta}_{01}) + \epsilon_i\boldsymbol{U}_{i1} - (\boldsymbol{U}_{i1}\boldsymbol{U}_{i1}^T - \boldsymbol{\Sigma}_{uu1})\boldsymbol{\beta}_{01}\right)^{\otimes 2}\right].$$

Denote $\boldsymbol{S}_{*1}$ as the elements of $\boldsymbol{S}_*$ corresponding to $\boldsymbol{\beta}_{01}$ for any random variable or function vector $\boldsymbol{S}_*$, where $\boldsymbol{\beta}_{01}$ denotes the vector that consists of all the nonzero elements in $\boldsymbol{\beta}_0$. We use the following theorem (Theorem 5.21 in Van der Vaart [16]) to prove the asymptotical normality.

Theorem 3.2.6. *For each θ in an open subset of Euclidean space, let $x \to \psi_\theta(x)$ be a measurable vector-valued function such that, for every θ_1 and θ_2 in a neighborhood of θ_0 and a measurable function $\dot{\psi}$ with $\mathrm{E}[\dot{\psi}] < \infty$,*

$$\|\psi_{\theta_1}(x) - \psi_{\theta_2}(x)\| \le \dot{\psi}(x)\|\theta_1 - \theta_2\|.$$

Assume that $\mathrm{E}[\|\psi_{\theta_0}\|^2] < \infty$ and that the map $\theta \mapsto \mathrm{E}[\psi_\theta]$ is differentiable at a zero θ_0, with nonsingular derivative matrix V_{θ_0}. If $\frac{1}{n}\sum_{i=1}^n \psi_{\hat{\theta}_n} = o_p(n^{-\frac{1}{2}})$, and $\hat{\theta}_n \xrightarrow{P} \theta_0$, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \psi_{\theta_0}(X_i) + o_p(1).$$

In paticular, the sequence $\sqrt{n}(\hat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and covariance matrix $V_{\theta_0}^{-1}\mathrm{E}[\psi_{\theta_0}\psi_{\theta_0}^T](V_{\theta_0}^{-1})^T$.

Now we prove Theorem 3.2.5.

Proof. Define $\mathcal{L}_p^*(\boldsymbol{\Sigma}_{uu}, \boldsymbol{\beta}_1)$ as

$$\begin{aligned}
\mathcal{L}_p^*(\boldsymbol{\Sigma}_{uu1}, \boldsymbol{\beta}_1) &= \frac{1}{n}\mathcal{L}_p\left(\boldsymbol{\Sigma}_{uu}, \begin{pmatrix}\boldsymbol{\beta}_1 \\ \mathbf{0}\end{pmatrix}\right) \\
&= \frac{1}{n}\left[\sum_{i=1}^n\left(\widehat{Y}_i - \widehat{\boldsymbol{W}}_{i1}^T\boldsymbol{\beta}_1\right)^2 - \sum_{i=1}^n \boldsymbol{\beta}_1^T\boldsymbol{\Sigma}_{uu1}\boldsymbol{\beta}_1\right] + \frac{\lambda_n}{n}\sum_{i=1}^s \widehat{w}_j|\beta_j| \\
&= \frac{1}{n}\sum_{i=1}^n\left[\left(\widehat{Y}_i - \widehat{\boldsymbol{W}}^T\boldsymbol{\beta}_1\right)^2 - \boldsymbol{\beta}_1^T\boldsymbol{\Sigma}_{uu1}\boldsymbol{\beta}_1\right] + \frac{\lambda_n}{n}\sum_{i=1}^s \widehat{w}_j|\beta_j|. \qquad (3.31)
\end{aligned}$$

Take the derivative of $\mathcal{L}_p^*(\boldsymbol{\Sigma}_{uu1}, \boldsymbol{\beta}_1)$ to get,

$$\mathrm{D}\left[\mathcal{L}_p^*(\boldsymbol{\Sigma}_{uu1}, \boldsymbol{\beta}_1)\right] = \frac{1}{n}\sum_{i=1}^n\left[-2\widehat{\boldsymbol{W}}_{i1}(\widehat{Y}_i - \widehat{\boldsymbol{W}}_{i1}^T\boldsymbol{\beta}_1) - 2\boldsymbol{\Sigma}_{uu1}\boldsymbol{\beta}_1\right] + \frac{\lambda_n}{n}\widehat{\boldsymbol{\beta}}_1^*, \qquad (3.32)$$

where $\widehat{\boldsymbol{\beta}}_1^*$ is defined as (3.24). By equation (3.18), the first term of (3.32) is

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left[-2\widehat{\boldsymbol{W}}_{i1}(\widehat{Y}_i-\widehat{\boldsymbol{W}}_{i1}^T\boldsymbol{\beta}_1)-2\boldsymbol{\Sigma}_{uu1}\boldsymbol{\beta}_1\right]\\
=&\frac{1}{n}\sum_{i=1}^{n}\Big[-2\left(\widehat{Y}_i-\widetilde{Y}_i\right)\left(\widehat{\boldsymbol{W}}_i-\widetilde{\boldsymbol{W}}_i\right)-2\left(\widehat{Y}_i-\widetilde{Y}_i\right)\widetilde{\boldsymbol{W}}_{\boldsymbol{i}}\\
&\quad-2\widetilde{Y}_i\left(\widehat{\boldsymbol{W}}_i-\widetilde{\boldsymbol{W}}_i\right)-2\left(\widetilde{Y}_i\widetilde{\boldsymbol{W}}_i-\widetilde{\boldsymbol{W}}_i\boldsymbol{\beta}_1^T\widetilde{\boldsymbol{W}}_i+\boldsymbol{\Sigma}_{uu1}\boldsymbol{\beta}_1\right)\\
&\quad+2\left(\widehat{\boldsymbol{W}}_i-\widetilde{\boldsymbol{W}}_i\right)\boldsymbol{\beta}_1^T\left(\widehat{\boldsymbol{W}}_i-\widetilde{\boldsymbol{W}}_i\right)+2\widetilde{\boldsymbol{W}}_i\boldsymbol{\beta}_1^T\left(\widehat{\boldsymbol{W}}_i-\widetilde{\boldsymbol{W}}_i\right)\\
&\quad+2\left(\widehat{\boldsymbol{W}}_i-\widetilde{\boldsymbol{W}}_i\right)\boldsymbol{\beta}_1^T\widetilde{\boldsymbol{W}}_i\Big]
\end{aligned}\tag{3.33}$$

Denote the fourth term inside the brackets in (3.33) by $\boldsymbol{\psi}_i(\boldsymbol{\beta}_1)$, and all the other sums by H, then (3.33) can be rewritten as

$$\frac{1}{n}\sum_{i=1}^{n}\left[-2\widehat{\boldsymbol{W}}_{i1}(\widehat{Y}_i-\widehat{\boldsymbol{W}}_{i1}^T\boldsymbol{\beta}_1)-2\boldsymbol{\Sigma}_{uu1}\boldsymbol{\beta}_1\right]=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}_i(\boldsymbol{\beta}_1)+H \tag{3.34}$$

Define

$$\boldsymbol{\Psi}_n(\boldsymbol{\beta}_1)=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}_i(\boldsymbol{\beta}_1)$$

i.e.,

$$\boldsymbol{\Psi}_n(\boldsymbol{\beta}_1)=\frac{1}{n}\sum_{i=1}^{n}\left[-2\widetilde{\boldsymbol{W}}_{i1}(\widetilde{Y}_i-\widetilde{\boldsymbol{W}}_{i1}^T\boldsymbol{\beta}_1)-2\boldsymbol{\Sigma}_{uu1}\boldsymbol{\beta}_1\right].$$

$\boldsymbol{\psi}_i(\boldsymbol{\beta}_1)$ is a continuous and differentiable vector-valued function, and its j-th ($j=1,\ldots,d$) component is

$$[\boldsymbol{\psi}_i(\boldsymbol{\beta}_1)]_j=-2\widetilde{W}_{i1j}(\widetilde{Y}_i-\widetilde{\boldsymbol{W}}_{i1}^T\boldsymbol{\beta}_1)-2\boldsymbol{\Sigma}_{uu1j}^T\boldsymbol{\beta}_1,$$

where $\boldsymbol{\Sigma}_{uu1j}$ denotes the j-th column vector of Σ_{uu1}. Then its derivative is

$$\mathrm{D}[\boldsymbol{\psi}_i(\boldsymbol{\beta}_1)]_j=2\widetilde{W}_{i1j}\widetilde{\boldsymbol{W}}_{i1}-2\boldsymbol{\Sigma}_{uu1j}.$$

For any $\boldsymbol{\beta}_1^{(1)}$ and $\boldsymbol{\beta}_1^{(2)}$ in the neighborhood of $\boldsymbol{\beta}_0$ such that

$$\|\boldsymbol{\beta}_1^{(i)}-\boldsymbol{\beta}_{01}\|\le Cn^{-\frac{1}{2}},\qquad i=1,2;$$

we have the Lipschitz continuity condition for each component of $\boldsymbol{\psi}_i(\boldsymbol{\beta}_1)$,

$$|[\boldsymbol{\psi}_i(\boldsymbol{\beta}_1^{(1)})]_j - [\boldsymbol{\psi}_i(\boldsymbol{\beta}_1^{(2)})]_j| \leq \sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| \leq Cn^{-\frac{1}{2}}} \|\mathrm{D}[\boldsymbol{\psi}_i(\boldsymbol{\beta}_1)]_j\| \, \|\boldsymbol{\beta}_1^{(1)} - \boldsymbol{\beta}_1^{(2)}\|.$$

Define

$$\phi = \sqrt{d} \sup_{j=1,\dots d} \left\{ \sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| \leq Cn^{-\frac{1}{2}}} \|\mathrm{D}[\boldsymbol{\psi}_i(\boldsymbol{\beta}_1)]_j\| \right\} = \sqrt{d} \sup_{j=1,\dots d} \|2\widetilde{W}_{i1j} \widetilde{\boldsymbol{W}}_{i1} - 2\boldsymbol{\Sigma}_{uu1j}\|$$

Then

$$\begin{aligned}
\|\boldsymbol{\psi}_i(\boldsymbol{\beta}_1^{(1)}) - \boldsymbol{\psi}_i(\boldsymbol{\beta}_1^{(2)})\|^2 &= \sum_{j=1}^{d} \left| [\boldsymbol{\psi}_i(\boldsymbol{\beta}_1^{(1)})]_j - [\boldsymbol{\psi}_i(\boldsymbol{\beta}_1^{(2)})]_j \right|^2 \\
&\leq \sum_{j=1}^{d} \left(\sup_{\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_{01}\| \leq Cn^{-\frac{1}{2}}} \|\mathrm{D}[\boldsymbol{\psi}_i(\boldsymbol{\beta}_1)]_j\| \right)^2 \|\boldsymbol{\beta}_1^{(1)} - \boldsymbol{\beta}_1^{(2)}\|^2 \\
&\leq \left(\frac{\phi}{\sqrt{d}} \right)^2 \sum_{j=1}^{d} \|\boldsymbol{\beta}_1^{(1)} - \boldsymbol{\beta}_1^{(2)}\|^2 \\
&= \phi^2 \|\boldsymbol{\beta}_1^{(1)} - \boldsymbol{\beta}_1^{(2)}\|^2.
\end{aligned}$$

Taking square root on both sides, we have

$$\|\boldsymbol{\psi}_i(\boldsymbol{\beta}_1^{(1)}) - \boldsymbol{\psi}_i(\boldsymbol{\beta}_1^{(2)})\| \leq \phi \, \|\boldsymbol{\beta}_1^{(1)} - \boldsymbol{\beta}_1^{(2)}\|$$

Assuming that each component in $\boldsymbol{W}_i$ and Z_i have finite fourth moments, for any $j = 1, \dots d$, we have

$$\begin{aligned}
\mathrm{E}\left[\|\widetilde{W}_{i1j} \widetilde{\boldsymbol{W}}_{i1} - \boldsymbol{\Sigma}_{uu1j}\| \right] &\leq \mathrm{E}\left[\|\widetilde{W}_{i1j} \widetilde{\boldsymbol{W}}_{i1}\| \right] + \|\boldsymbol{\Sigma}_{uu1j}\| \\
&= \mathrm{E}\left[|\widetilde{W}_{i1j}| \|\widetilde{\boldsymbol{W}}_{i1}\| \right] + \|\boldsymbol{\Sigma}_{uu1j}\| \\
&\leq \mathrm{E}\left[|\widetilde{W}_{i1j}|^2 \right]^{\frac{1}{2}} \mathrm{E}\left[\|\widetilde{\boldsymbol{W}}_{i1}\|^2 \right]^{\frac{1}{2}} + \|\boldsymbol{\Sigma}_{uu1j}\| \\
&= \mathrm{E}\left[|W_{i1j} - \mathrm{E}[W_{i1j}|Z_i]|^2 \right]^{\frac{1}{2}} \mathrm{E}\left[\|\boldsymbol{W}_{i1} - \mathrm{E}[\boldsymbol{W}_{i1}|Z_i]\|^2 \right]^{\frac{1}{2}} + \|\boldsymbol{\Sigma}_{uu1j}\| \\
&< \infty.
\end{aligned}$$

Since d is finite,

$$\mathrm{E}[\phi] < \infty.$$

We then consider the expectation of $\boldsymbol{\psi}_i(\boldsymbol{\beta}_1)$. By the calculation of the expectation of the forth term in equation (3.18) at $\boldsymbol{\beta}_0$, we showed that (see (3.21)),

$$\mathrm{E}[\boldsymbol{\psi}_i(\boldsymbol{\beta}_{01})] = 0.$$

Hence $\boldsymbol{\beta}_{01}$ is a zero of $\mathrm{E}[\boldsymbol{\psi}_i(\boldsymbol{\beta}_1)]$, and $\mathrm{E}[\boldsymbol{\psi}_i(\boldsymbol{\beta}_1)]$ is differentiable at $\boldsymbol{\beta}_{01}$. Moreover,

$$\begin{aligned}
\mathrm{E}\left[\|\boldsymbol{\psi}_i(\boldsymbol{\beta}_{01})\|^2\right] &= 4\mathrm{E}\left[\left\|-\widetilde{\boldsymbol{W}}_{i1}(\widetilde{Y}_i - \widetilde{\boldsymbol{W}}_{i1}^T\boldsymbol{\beta}_{01}) - \boldsymbol{\Sigma}_{uu1}\boldsymbol{\beta}_{01}\right\|^2\right] \\
&\leq 8\mathrm{E}\left[\left\|\widetilde{\boldsymbol{W}}_{i1}(\widetilde{Y}_i - \widetilde{\boldsymbol{W}}_{i1}^T\boldsymbol{\beta}_{01})\right\|^2\right] + 8\|\Sigma_{uu1}\boldsymbol{\beta}_{01}\| \\
&= 8\mathrm{E}\left[(\widetilde{Y}_i - \widetilde{\boldsymbol{W}}_{i1}^T\boldsymbol{\beta}_{01})^2\|\widetilde{\boldsymbol{W}}_{i1}\|^2\right] + 8\|\Sigma_{uu1}\boldsymbol{\beta}_{01}\| \\
&\leq 8\mathrm{E}\left[(\widetilde{Y}_i - \widetilde{\boldsymbol{W}}_{i1}^T\boldsymbol{\beta}_{01})^4\right]^{\frac{1}{2}} \cdot \mathrm{E}[\|\widetilde{\boldsymbol{W}}_{i1}\|^4]^{\frac{1}{2}} + 8\|\Sigma_{uu1}\boldsymbol{\beta}_{01}\| \\
&< \infty.
\end{aligned}$$

For the root n consistent local minimizer $\widehat{\boldsymbol{\beta}}$ we got from the previous results, with $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2 = 0$, such that

$$\mathrm{D}\left[\mathcal{L}_p^*(\boldsymbol{\Sigma}_{uu1}, \boldsymbol{\beta}_1)\right]\big|_{\boldsymbol{\beta}_1 = \widehat{\boldsymbol{\beta}}_1} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}_i(\widehat{\boldsymbol{\beta}}_1) + H + \frac{\lambda_n}{n}\widehat{\boldsymbol{\beta}}_1^* = 0$$

Hence

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}_i(\widehat{\boldsymbol{\beta}}_1) = -\frac{\lambda_n}{n}\widehat{\boldsymbol{\beta}}_1^* - H. \tag{3.35}$$

Since

$$\frac{\lambda_n}{n} = \frac{\lambda_n}{\sqrt{n}} \cdot \frac{1}{\sqrt{n}} = o(1)\,\frac{1}{\sqrt{n}} = o(n^{\frac{1}{2}}),$$

and

$$\widehat{\beta}_{1j}^* \xrightarrow{P} \frac{1}{|\beta_j|^\gamma},$$

we have

$$\frac{\lambda_n}{n}\widehat{\boldsymbol{\beta}}_1^* = o_p(n^{-\frac{1}{2}}). \tag{3.36}$$

Also, from the calculation of the order of the RHS of (3.18) we know that

$$H = o_p(n^{-\frac{1}{2}}). \tag{3.37}$$

Combine (3.35), (3.36), and (3.37) to get

$$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\psi}_i(\widehat{\boldsymbol{\beta}}_1) = o_p(n^{-\frac{1}{2}}).$$

Since $\widehat{\beta}_1$ is a root n consistent estimator of β_{01}, we have $\widehat{\beta}_1 \xrightarrow{P} \beta_{01}$. The derivative of $\boldsymbol{\psi}_i(\boldsymbol{\beta}_{01})$ is continuous at $\boldsymbol{\beta}_{01}$. Hence by Theorem 3.2.6,

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) = -\boldsymbol{V}_{\boldsymbol{\beta}_{01}}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\boldsymbol{\psi}_i(\boldsymbol{\beta}_{01}) + o_p(1),$$

where $\boldsymbol{V}_{\boldsymbol{\beta}_{01}} = \mathrm{E}[\mathrm{D}\boldsymbol{\psi}_i(\boldsymbol{\beta}_{01})]$, which is a $d \times d$ matrix. It then follows that $\widehat{\boldsymbol{\beta}}_1$ has an asymptotic normal distribution where

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{01}) \xrightarrow{\mathcal{D}} N\left(0, \boldsymbol{V}_{\boldsymbol{\beta}_{01}}^{-1}\mathrm{E}[\boldsymbol{\psi}_i(\boldsymbol{\beta}_{01})\boldsymbol{\psi}_i(\boldsymbol{\beta}_{01})^T](\boldsymbol{V}_{\boldsymbol{\beta}_{01}}^{-1})^T\right).$$

Then

$$\begin{aligned}
\mathrm{E}[\boldsymbol{\psi}_i(\boldsymbol{\beta}_{01})\boldsymbol{\psi}_i(\boldsymbol{\beta}_{01})^T] &= \mathrm{E}[\boldsymbol{\psi}(\boldsymbol{\beta}_{01})^{\otimes 2}] \\
&= 4\mathrm{E}[(\widetilde{\boldsymbol{W}}_{i1}(\widetilde{Y}_i - \widetilde{\boldsymbol{W}}_{i1}^T\boldsymbol{\beta}_1) + \boldsymbol{\Sigma}_{uu1}\boldsymbol{\beta}_1)^{\otimes 2}] \\
&= 4\mathrm{E}[(\epsilon_i\widetilde{\boldsymbol{W}}_{i1} - \boldsymbol{U}_{i1}^T\boldsymbol{\beta}_{01}\widetilde{\boldsymbol{W}}_{i1} + \boldsymbol{\Sigma}_{uu1}\boldsymbol{\beta}_{01})^{\otimes 2}] \\
&= 4\mathrm{E}[(\epsilon_i(\boldsymbol{W}_{i1} - \mathrm{E}[\boldsymbol{W}_{i1}|Z_i]) - \boldsymbol{U}_{i1}^T\boldsymbol{\beta}_{01}(\boldsymbol{W}_{i1} - \mathrm{E}[\boldsymbol{W}_{i1}|Z_i]) + \boldsymbol{\Sigma}_{uu1}\boldsymbol{\beta}_{01})^{\otimes 2}] \\
&= 4\mathrm{E}[(\epsilon_i(\widetilde{X}_{i1} + \boldsymbol{U}_{i1}) - \boldsymbol{U}_{i1}^T\boldsymbol{\beta}_{01}(\widetilde{X}_{i1} + \boldsymbol{U}_{i1}) + \boldsymbol{\Sigma}_{uu1}\boldsymbol{\beta}_{01})^{\otimes 2}] \\
&= 4\mathrm{E}\left[\left(\widetilde{\boldsymbol{X}}_{i1}(\epsilon_i - \boldsymbol{U}_{i1}^T\boldsymbol{\beta}_{01}) + \epsilon_i\boldsymbol{U}_{i1} - (\boldsymbol{U}_{i1}\boldsymbol{U}_{i1}^T - \boldsymbol{\Sigma}_{uu1})\boldsymbol{\beta}_{01}\right)^{\otimes 2}\right];
\end{aligned}$$

and

$$\begin{aligned}
\boldsymbol{V}_{\boldsymbol{\beta}_{01}} &= \mathrm{E}[\mathrm{D}\boldsymbol{\psi}_i(\boldsymbol{\beta}_{01})] \\
&= 2\mathrm{E}[\widetilde{\boldsymbol{W}}_{i1}\widetilde{\boldsymbol{W}}_{i1}^T] - 2\boldsymbol{\Sigma}_{uu1}^T \\
&= 2\mathrm{Cov}(\widetilde{\boldsymbol{X}}_{i1} + \boldsymbol{U}_{i1}) - 2\boldsymbol{\Sigma}_{uu1} \\
&= 2\mathrm{Cov}(\widetilde{\boldsymbol{X}}_{i1}) + 2\mathrm{Cov}(\boldsymbol{U}_{i1}) - 2\boldsymbol{\Sigma}_{uu1} \\
&= 2\mathrm{Cov}(\widetilde{\boldsymbol{X}}_{i1}).
\end{aligned}$$

□

This asymptotic normality result is very close to that of Liang et al. [11] using SCAD. However, in their result, there is a bias inherited from SCAD technique. Although the bias eventually goes to 0 as $n \to \infty$, it still exists for finite samples. Hence,their method only results in a nearly unbiased estimator. On the other hand, by adaptive LASSO, we have an unbiased estimator for all coefficients.

Appendices

Appendix A Regularity Conditions

The following regularity conditions are needed for the proof of the theorems. They may not be the weakest ones.

1. $\mathrm{Cov}(\widehat{\boldsymbol{X}}_{i1})$ is a positive definite matrix. $\mathrm{E}[|\epsilon|^3|\boldsymbol{X},\boldsymbol{Z}]<\infty$.

2. The bandwidths in estimating $\mathrm{E}[\boldsymbol{X}|Z]$ and $\mathrm{E}[Y|Z]$ are of order $n^{-\frac{1}{5}}$.

3. $K(\cdot)$ is a bounded symmetric density function with compact support and satisfies

$$\int K(u)\,du=1,\quad \int K(u)u\,du=0,\quad \int u^2K(u)\,du=1.$$

4. The density function of Z, $f_Z(z)$, and the density function of (Y,Z) are bounded away from 0 and have bounded continuous second derivative.

5. $\mathrm{E}[\boldsymbol{X}|Z]$ and $\mathrm{E}[Y|Z]$ have bounded and continuous second derivatives.

Bibliography

[1] Hirotugu Akaike. Maximum likelihood identification of Gaussian autoregressive moving average models. *Biometrika*, 60:255–265, 1973.

[2] Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement error in nonlinear models*, volume 105 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, Boca Raton, FL, second edition, 2006. A modern perspective.

[3] Chi-Lun Cheng and John W. Van Ness. *Statistical regression with measurement error*, volume 6 of *Kendall's Library of Statistics*. Arnold, London, 1999.

[4] David L. Donoho. For most large underdetermined systems of equations, the minimal l_1-norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.*, 59(7):907–934, 2006.

[5] David L. Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. *Proc. Natl. Acad. Sci. USA*, 100(5):2197–2202 (electronic), 2003.

[6] David L. Donoho and Xiaoming Huo. Beamlets and multiscale image analysis. In *Multiscale and multiresolution methods*, volume 20 of *Lect. Notes Comput. Sci. Eng.*, pages 149–196. Springer, Berlin, 2002.

[7] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, 96(456):1348–1360, 2001.

[8] Jianqing Fan and Runze Li. Statistical challenges with high dimensionality: feature selection in knowledge discovery. In *International Congress of Mathematicians. Vol. III*, pages 595–622. Eur. Math. Soc., Zürich, 2006.

[9] Dean P. Foster and Edward I. George. The risk inflation criterion for multiple regression. *Ann. Statist.*, 22(4):1947–1975, 1994.

[10] Wayne A. Fuller. *Measurement error models*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006. Reprint of the 1987 original, Wiley-Interscience Paperback Series.

[11] Hua Liang and Runze Li. Variable selection for partially linear models with measurement errors. *J. Amer. Statist. Assoc.*, 104(485):234–248, 2009.

[12] M. Rudemo, D. Ruppert, and J. C. Streibig. Random-effect models in nonlinear regression with applications to bioassay. *Biometrics*, 45(2):349–362, 1989.

[13] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464, 1978.

[14] A. F. Subar, F. E. Thompson, V. Kipnis, D. Midthune, P. Hurwitz, S. McNutt, A. McIntosh, and S. Rosenfeld. Comparative validation of the block, willett, and national cancer institute food frequency questionnaires : The eating at america's table study. *Am. J. Epidemiol.*, 154(12):1089–1099, 2001.

[15] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[16] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2007.

[17] Hui Zou. The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, 101(476):1418–1429, 2006.