The Pennsylvania State University
The Graduate School

STATISTICAL INFERENCE WITH CORRUPTED DATA

A Dissertation in
Statistics
by
Mengyan Li

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2020

The dissertation of Mengyan Li was reviewed and approved by the following:

Yanyuan Ma
Professor of Statistics
Dissertation Co-Advisor, Committee Co-Chair

Runze Li
Eberly Family Chair in Statistics; Associate Department Head
Dissertation Co-Advisor, Committee Co-Chair

Michael Akritas
Professor of Statistics

Bing Li
Verne M. Willaman Professor of Statistics

Lan Kong
Professor of Biostatistics and Bioinformatics

Ephraim Hanks
Associate Professor of Statistics
Chair of Graduate Program

Abstract

Corrupted data are ubiquitous in many applications where measurement errors or missing data cannot be ignored. For example, measurement errors arise frequently in nutriology, biomedical science, etc. Missing data are common in research involving human subjects, such as health-related studies and sample surveys. Statistical inference with corrupted data is believed to be challenging, and improper treatments can lead to biased estimation and erroneous inference. Assumptions are imposed to guarantee the model identifiability and to facilitate the establishment of theoretical results. Those assumptions can be too restrictive in some applications. Another issue on analyzing corrupted data is impaired estimation efficiency. Extra noises or partial observations often lead to a lack of power. In this dissertation, we focus on developing robust or efficient methodologies for analyzing corrupted data and establishing asymptotic properties of the newly proposed estimators to make the statistical inference feasible.

In Chapter 3, we consider a given parametric regression model with a covariate measured with heteroscedastic error and the error distribution can have an arbitrary form. Both the variance function of the measurement error and the distribution of the error-prone covariate are left completely unspecified. We avoid performing deconvolution, the standard treatment in the prior literature, by using a novel spline-assisted semiparametric approach. Its most distinctive feature is to embed the B-splines approximation of the variance function in a semiparametric treatment; this achieves robust estimation that allows misspecification of the covariate distribution. By combining the knowledge of B-splines technique, integral equations, and semiparametric analysis, we establish our estimator's theoretical properties.

In Chapter 4, we study statistical inference on parameters associated with a finite number of error-prone covariates in high-dimensional linear measurement error models. In the high-dimensional settings, the main challenges posed by measurement errors are nonconvexity and lack of closed-form solutions which significantly complicate the analysis of standard regularization methods such as Lasso and Dantzig selector. To counteract the effect of high-dimensional nuisance parameters and correct the biases introduced by measurement errors, we propose a new corrected decorrelated score test and a corresponding one-step estimator. By adapting the bias-correction and the decorrelation operations to our model, we show that our test statistic is asymptotically normal and retains power under the local alternatives around zero. Further, our one-step estimator has significantly better convergence performance than other existing estimators, and it is semiparametrically efficient.

In Chapter 5, we consider the data where all the covariates are fully observed and the

scalar response is subject to nonignorable missingness, i.e., the missingness mechanism depends on the missing values themselves. In such cases, model identifiability and model misspecification can be two critical problems. We assume a flexible semiparametric exponential tilting propensity where the relationship between the missing indicator and the response is totally unspecified and estimated nonparametrically, while the relationship between the missingness indicator and the covariates are modeled parametrically. To guarantee that the model is identifiable, we model the fully observed part of the data parametrically. We devise two estimators for the parameter of interest in the parametric parts using a semiparametric treatment. The first one is robust against the misspecification of the distribution of the covariates, while the second estimator is semiparametrically efficient.

Table of Contents

List of Figures

List of Tables

Acknowledgments

The work presented here gives undeniable testimony to a community of contributors who have enriched this period of my life through their support and friendship, and enabled this research through their expertise. These are gratefully acknowledged.

I have had the privilege of working with two advisers, Professor Yanyuan Ma and Professor Runze Li, who have shaped the way I approach statistical problems. Their contributions to this work through our discussions and collaborations over the last four years are substantial. Professor Yanyuan Ma is not only my dissertation advisor but also a lifelong mentor and friend. I deeply value the high standard that she demands, which helps me form a rigorous scientific attitude. She also taught me the virtues of respect, discipline and hardworking. She has been always there when I was lost or felt frustrated, providing me with large encouragement and support. Professor Runze Li opened the door of Ph.D. study in statistics for me five years ago. He also provided me with great opportunities to learn more about the beauty of statistics and how to become a successful researcher. I appreciate, in particular, his invaluable advice on both research and life that helps me make better decisions in this important stage of my life.

I also would like to express my sincere gratitude to the rest of my dissertation committee members, Professor Michael Akritas, Professor Bing Li and Professor Lan Kong. I am grateful to Professor Akritas for his helpful suggestions on my dissertation as well as his substantial support for my career development. I took four courses taught by Professor Li and have learned a lot from him. I appreciate all his help and insightful comments. I also want to thank Professor Kong for her great comments from an applied point of view, which expands my vision beyond the statistics.

Last but not the least, my support system is founded on the unconditional love from my mother Jie Zhao, father Hongxing Li and my boyfriend Yunxing Lu. They are my heroes. I cannot express my gratitude enough for all that they have done for me.

Chapter 1 Introduction

1.1 Measurement Error Model with Heteroscedastic Error

Covariate measurement errors are common in many applications. Instead of a precise measurement, only an error-prone surrogate of the unobserved covariate is available. Existing methods often make restrictive assumptions on measurement errors such as normality and heteroscedasticity. In practice, these assumptions can be violated resulting in biased estimation and erroneous inference. Measurement error heteroscedasticity is an issue requiring attention but a challenging problem to handle. We consider a given parametric regression model allowing for one covariate measured with heteroscedastic error in low-dimensional settings. We allow both the variance function of the measurement errors and the distribution of the unobservable covariate to be completely unspecified. We approximate the variance function using B-spines, and the distribution of the unobservable covariate can be replaced with a working model due to a semiparametric treatment. The resulting semiparametric estimator is consistent and enjoys good inference properties. Its finite sample performance is demonstrated through simulation studies and a real data example.

1.2 High-dimensional Lineal Measurement Error Model

High-dimensional data become more and more common these days. Extensive research on both the estimation and the inference of high-dimensional models with clean data has been conducted. While high-dimensional model with corrupted data is an interesting but difficult topic with less research on it. Here we focus on a high-dimensional linear

model with a finite number of covariates measured with additive and homoscedastic error. In this case, estimation itself is already a challenging problem. To correct the bias introduced by measurement errors, the objective function becomes non-convex. Several methods have been proposed to solve the problems caused by the non-convexity (Belloni, Rosenbaum & Tsybakov 2017, Datta et al. 2017, Loh & Wainwright 2012). We move one-step further by considering the statistical inference in high-dimensional linear models with measurement errors. This work was motivated by an empirical analysis of a real data, where both finite-dimensional phenotypic covariates and high-dimensional single nucleotide polymorphisms (SNPs) are available and one of the phenotypic covariates is of clinical interest but measured with error. We can decompose a high-dimensional model into two parts, the low-dimensional component which is of interest and a high-dimensional component which is nuisance. From a semiparametric perspective, to conduct hypothesis testing on low-dimensional sub-parameters in a high dimensional sparse model, we need to control the effect of the estimation of the high-dimensional nuisance parameters. In analogy to the efficient scores in semiparametrics, we construct corrected decorrelated scores and propose a new corrected decorrelated score test. Instead of solving the estimating equation based on the corrected decorrelated scores, we propose a one-step estimator which has closed form and is semiparametrically efficient. The finite-sample performance of the proposed inference procedure is examined through simulation studies. We further illustrate the proposed procedure via an empirical analysis of a real data example.

1.3 Nonignorable Missing Data

The reason for missingness is crucial in statistical analysis with missing data. We consider a scalar nonignorable missing response, where the missingness depends on the missing value themselves. One intrinsic difficulty for analyzing nonignorable missing data is on modeling the missing mechanism. There have been many papers modeling it parametrically (Chang & Kott 2008, Ibrahim & Lipsitz 1996, Morikawa & Kim 2016, Qin et al. 2002, Rotnitzky & Robins 1997, Wang et al. 2014). However, the parametric assumption is generally uncheckable and can be too restrictive. Due to the uncheckableness, flexible models are desired. Tang et al. (2003) and Zhao & Ma (2019) treated the missingness mechanism as nuisance and allowed it to be unspecified. Tang et al. (2003) assumed the mechanism only depends on the partially observed response variable, which is very restrictive; while Zhao & Ma (2019) allowed the mechanism to

depend on some covariates, but the efficiency in estimation is impaired.

We consider a parametric regression model with a semiparametric exponential tilting propensity where the relationship between the missingness indicator and the response is totally unspecified and estimated nonparametrically, and the relationship between the missingness indicator and the covariates are modeled parametrically. Since the response is partially observed, it makes more sense to model the relationship between the missingness indicator and the response nonparametrically in the missingness mechanism. We propose two estimators for the parameter of interest in the parametric parts using a semiparametric treatment. The first one is robust against the misspecification of the covariates distribution, and the second estimator is semiparametrically efficient.

1.4 Organization of the Dissertation

Literature review is given in Chapter 2. In Chapter 3, the project on measurement error model with heteroscedastic error is discussed in detail. We specify the model for the data with mismeasured covariate and heteroscedastic measurement error in Section 3.2, and develop the estimation procedure in Section 3.3. Regularity conditions and asymptotic properties of the estimator are described in Section 3.4. We further explain how to implement our method in Section 3.5. To assess the performance of our method, we conduct simulation studies and perform an empirical data analysis in Section 3.6. The conclusion and discussion are given in Section 3.7. The technical proofs can be found in Appendix A.

In Chapter 4, the project on inference in high-dimensional linear measurement error models is discussed in depth. Model setup is given in Section 4.2. The theory for the corrected decorrelated score test and the one-step estimator is given in Section 4.3. Results of empirical studies are given in Section 4.4. Discussion and future research directions are given in Section 4.5. The corresponding technical proofs can be found in Appendix B.

In Chapter 5, we study the regression model with nonignorable and unspecified missingness carefully. The model specification is given in Section 5.2. Since model identifiability is a crucial issue for nonignorable missing data, we show the identifiability of our model in Section 5.3. Detailed semiparametric methodology is in Section 5.4. We summarize our method and explain the implementation in Section 5.5. Theoretical properties of our semiparametric estimator are given in Section 5.6, and the proofs are given in Appendix C. Simulation results are given in Section 5.7.

Chapter 2
Literature Review

2.1 Measurement Error Model

Measurement error model commonly consists of two parts. The first part is underlying main model for the response Y in terms of the covariates, which can be linear or nonlinear. We distinguish between two kinds of covariates: $\mathbf{X}$ represents the covariates that cannot be observed exactly, and $\mathbf{Z}$ represents those that are measured without error. The second part is measurement error process. We can observe a variable $\mathbf{W}$ which is the surrogate of $\mathbf{X}$. In general, there are two approaches to modeling the measurement error process. The first approach is error models, where conditional distribution of $\mathbf{W}$ given $(\mathbf{X}, \mathbf{Z})$ is modeled; the second approach is regression calibration models, where the conditional distribution of $\mathbf{X}$ given $(\mathbf{W}, \mathbf{Z})$ is modeled. Here we focus on error models.

Obviously, the parameters in the main model cannot be estimated directly by fitting Y to $(\mathbf{X}, \mathbf{Z})$. The goal of measurement error model is to obtain good estimates of these parameters indirectly using the observed data $(Y, \mathbf{W}, \mathbf{Z})$. The simplest idea might be fitting the model substituting $\mathbf{W}$ for $\mathbf{X}$ without any adjustment. However, it can lead to seriously biased estimates. Here we take the simple linear model with additive measurement error as an example to show the effects of measurement error in model fitting. Consider the model

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \text{and} \quad W = X + U,$$

where X has mean zero and variance σ_X^2, regression error ϵ is independent of X with mean zero and variance σ_ϵ^2, and measurement error U is also independent of X with mean zero and variance σ_U^2. In this model, W is unbiased measure of X. It has been proved that the ordinary least squares regression of Y on W is a consistent estimate of

$\lambda\beta_1$, where

$$\lambda = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2} \leqslant 1.$$

Besides biasedness, Carroll et al. (2006) shows that the existence of measurement errors also make the data more noisy. The variance of Y given W is

$$\begin{aligned} \text{var}(Y|W) &= \sigma_\epsilon^2 + \frac{\beta_1^2\sigma_U^2\sigma_X^2}{\sigma_X^2 + \sigma_U^2} \\ &= \text{var}(Y|X) + \frac{\beta_1^2\sigma_U^2\sigma_X^2}{\sigma_X^2 + \sigma_U^2}. \end{aligned}$$

Therefore, careful analysis is required in measurement error models for both estimation and inference.

Fuller (1987, 2009) summarized an enormous literature on linear measurement error models and systematically studied the effects of measurement error in simple and multiple regression with one or multiple covariates measured with error. As discussed above, the key is to correct the bias caused by measurement errors. Two commonly used methods are method of moments and orthogonal regression. Nonlinear measurement error modeling began in earnest in the early 1980s. A series of papers were published on diverse topics: survival analysis Prentice (1982), generalized linear models Carroll et al. (1984), Stefanski & Carroll (1985) and estimating equations Stefanski (1985). Carroll et al. (2006) summarized different approaches for nonlinear measurement error models including simulation extrapolation Carroll et al. (1996), Stefanski & Cook (1995), likelihood and quasi-likelihood Carroll et al. (1984), Wang et al. (1996), Whittemore & Gong (1991), Bayesian methods Mallick & Gelfand (1996), Müller & Roeder (1997), semiparametric methods Carroll & Wand (1991), Robins et al. (1995), etc. Yi (2016) summarized the literature from a different perspective. It discussed the applications of measurement error models in survival data, recurrent event data, longitudinal data, multi-state models and case-control studies.

One challenge in nonlinear measurement error models is approximating the unknown distribution of the unovservable covariates $\mathbf{X}$. Based on the assumption of the distribution of $\mathbf{X}$, the above methods can be divided into two categories. The first one is structural modeling, in which the distribution of X is given a parametric form. The second category is functional modeling, which makes no assumption about the distribution of $\mathbf{X}$. Tsiatis & Ma (2004) proposed a class of semiparametric estimators in general settings of functional

measurement error models following from estimating equations based on semiparametric efficient score, where the distributional assumption for $\mathbf{X}$ is allowed to be misspecified. The idea has been utilized in many other papers on measurement error models Ma & Carroll (2006), Ma & Li (2010*a*), and is also crucial to our work.

Note that normality and homoscedasticity of the measurement error are commonly assumed in measurement error models. However, these two assumptions can be too restrictive and unrealistic in certain cases. In practice, the measurement error distribution may have heavy tails and the variance may not be constant. The violation of the two assumptions can lead to erroneous estimation and inference. Staudenmayer et al. (2008) carefully studied the bias issue in density estimation caused by incorrectly assuming homoscedasticity of the measurement error. It considered the situation in which there are replicate measurements $W_1, \ldots, W_m$ with assumption that $W_j|X = x \sim \mathcal{N}\{x, v(x, \boldsymbol{\theta})\}$. In its framework, the approximate bias at x is

$$\left\{\frac{v(x) - \sigma_U^2}{m}\right\}\frac{f''(x)}{2},$$

where $v(x)$ is the true variance function and σ_U^2 is the expected value of $v(X)$ with respect to the density of X, $f(X)$. Similarly, ignoring the heteroscedasticity of the measurement error in regression problems can also cause biases in estimation and affect the quality of inference.

Actually, little literature is available on heteroscedasticity in measurement error. Devanarayan & Stefanski (2002) proposed empirical SIMEX method to accommodate heteroscedastic measurement error. However, they assumed normality of the measurement error and only provided an approximate solution. Cheng & Riu (2006) studied linear relationships when both the response variable and the covariates are subject to heteroscedastic error using maximum likelihood method, method-of-moments and generalized least squares method, under a critical but restrictive assumption that the variance of the normally distributed measurement and regression errors for each observation are both known. Guo & Little (2011) extended the regression calibration and multiple imputation methods to allow heterocsedastic measurement error, while assuming the normality of the conditional density of the measurement errors given the unobservable covariate and assuming the variance function is a power function of the unobservable covariate. Sarkar et al. (2014) studied the regression model with heteroscedastic errors in covariates in a Bayesian hierarchical framework and avoided the assumptions about normality and homoscedasticity of the measurement as well as regression errors. However, due to the

complexity of using both B-splines and Dirichlet processes, the theoretical properties of the estimator is not established.

2.2 Semiparametric Regular Asymptotic Linear (RAL) Estimator and Efficiency

Statistical problems are described using probability models, where data are envisioned as realizations of a vector of random variables $\mathbf{Z}$. The densities in a model are often identified through a set of parameters, and the value of the parameters or some subset of the parameters is believed to have vital importance. The class of densities can be described as

$$\mathcal{P} = \{p(\mathbf{z}, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Omega\}.$$

For finite-dimensional parametric models, $\Omega \subset \mathbb{R}^p$, where the dimension p is some finite positive integer. In some cases, we may want to consider models where the class of densities is too large that the parameter $\boldsymbol{\theta}$ is infinite-dimensional. For some problems, it is natural to partition the infinite-dimensional parameter $\boldsymbol{\theta}$ as $(\boldsymbol{\beta}, \boldsymbol{\eta})$, where $\boldsymbol{\beta}$ is the finite-dimensional parameter of interest and $\boldsymbol{\eta}$ is the infinite-dimensional nuisance parameter. These models are referred to as semiparametric models in the literature, because they are described using both the parametric component $\boldsymbol{\beta}$ and the nonparametric component $\boldsymbol{\eta}$. Compared with parametric models, by allowing the parameters to be infinite-dimensional, we put less restrictions on the semiparametric model, which leads to robustness and greater applicability of the solutions. Tsiatis (2007) summarized the theory for semiparametrics, including different kinds of semiparametric estimators of different types of semiparametric models, semiparametric efficiency, etc.

Here we focus on the construction of semiparametric regular asymptotically linear (RAL) estimators. An asymptotically linear estimator for $\boldsymbol{\beta}$ is defined as

$$n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = n^{1/2} \sum_{i=1}^{n} \phi(\mathbf{Z}_i) + o_p(1),$$

where the p-dimensional measurable random function $\phi(\mathbf{Z}_i)$ is called the ith influence function of the estimator $\widehat{\boldsymbol{\beta}}_n$. It has mean-zero and $\mathrm{E}(\phi\phi^{\mathrm{T}})$ is finite and nonsingular. Further, to exclude the unnatural super-efficient estimator, we restrict ourselves to regular estimators, satisfying that $n^{1/2}(\widehat{\boldsymbol{\beta}}_n - \beta_0)$ has a limiting distribution which does not depend

on local data generating process. From the asymptotic perspective, an RAL estimator is identified through its influence function. To be more specific, the asymptotic variance of an RAL estimator is the variance of its influence function. Efficient influence function and efficient RAL estimator for parametric model with finite-dimensional nuisance parameters are discussed in detail in Chapter 3 in Tsiatis (2007).

A common idea to deal with infinite-dimensional problem is to first work with a finite-dimensional problem as an approximation and then take limits to infinity. Similarly, to construct semiparametric RAL estimators, we first consider a simpler finite-dimensional parametric model contained within the semiparametric model. Recall that in a semiparametric model, the data $\mathbf{Z}_1, \ldots \mathbf{Z}_n$ are iid random vectors generated from a density that belongs to the class

$$\mathcal{P} = [p(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\eta}), \text{where } \boldsymbol{\beta} \text{ is p-dimensional and } \boldsymbol{\eta} \text{ is infinite-dimensional}]$$

with respect to some dominating measure $\boldsymbol{\nu}_{\mathbf{Z}}$. Denote the true density by $p_0\{\mathbf{z}, \boldsymbol{\beta}_0, \eta_0\}$. Define the parametric submodel, denoted by $\mathcal{P}_{\boldsymbol{\beta},\boldsymbol{\gamma}} = \{p(\mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma})\}$, as a class of densities characterized by the finite-dimensional parameter $(\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ such that $\mathcal{P}_{\boldsymbol{\beta},\boldsymbol{\gamma}} \subset \mathcal{P}$ and $p_0(\mathbf{z}, \boldsymbol{\beta}_0, \boldsymbol{\eta}_0) \in \mathcal{P}_{\boldsymbol{\beta},\boldsymbol{\gamma}}$. Denote the dimension of nuisance parameter $\boldsymbol{\gamma}$ by a finite integer r. Then the theory and methods developed for parametric RAL estimators can be applied to the parametric submodels. Note that different from parametric model, parametric submodel is only a conceptual idea that is used to develop the theory for semiparametric models and cannot be used in data analysis.

The nuisance tangent space for parametric submodel is a subspace of the Hilbert space $\mathcal{H}$ of p-dimensional mean-zero finite-variance measurable functions equipped with the covariance inner product. Denote it by

$$\Lambda_{\boldsymbol{\gamma}} = \left\{\mathbf{B}\mathbf{S}_{\boldsymbol{\gamma}}(\mathbf{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0), \text{for all matrix } \mathbf{B} \in \mathbb{R}^{p\times r}\right\},$$

where $\mathbf{S}_{\boldsymbol{\gamma}}(\mathbf{z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0) = \{\partial \log p(\mathbf{z}, \beta_0, \boldsymbol{\gamma})/\partial \boldsymbol{\gamma}\}|_{\gamma=\gamma_0}$. The influence functions of the RAL estimators for $\boldsymbol{\beta}$ of a parametric submodel belong to the subspace of $\mathcal{H}$ that is orthogonal to the parametric submodel nuisance tangent space $\Lambda_{\boldsymbol{\gamma}}$. The efficient influence function for the parametric submodel is given by

$$\phi_{\text{eff},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{Z}) = [\mathrm{E}\{\mathbf{S}_{\text{eff},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\mathbf{S}^{\mathrm{T}}_{\text{eff},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)\}]^{-1}\mathbf{S}_{\text{eff},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0),$$

where $\mathbf{S}_{\text{eff},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{Z}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0)$, the parametric submodel efficient score, is defined as the residual

of the score for $\boldsymbol{\beta}$ after projecting onto the parametric submodel nuisance tangent space $\Lambda_{\boldsymbol{\gamma}}$. That is,

$$\mathbf{S}_{\text{eff},\boldsymbol{\beta},\boldsymbol{\gamma}}(\mathbf{Z},\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0) = \mathbf{S}_{\boldsymbol{\beta}}(\mathbf{Z},\boldsymbol{\beta}_0,\boldsymbol{\eta}_0) - \Pi\{\mathbf{S}_{\boldsymbol{\beta}}(\mathbf{Z},\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)|\Lambda_{\boldsymbol{\gamma}}\},$$

where $\mathbf{S}_{\boldsymbol{\beta}}(\mathbf{Z},\boldsymbol{\beta}_0,\boldsymbol{\eta}_0) = \{\partial \log p(\mathbf{Z},\boldsymbol{\beta},\boldsymbol{\eta}_0)/\partial\boldsymbol{\beta}\} \mid_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$.

Back to the semiparametric model, the semiparametric nuisance tangent space is defined as the mean-square closure of parametric submodel nuisance tangent spaces. That is,

$$\begin{aligned}\Lambda = \quad & [\mathbf{h}(\mathbf{Z}) \in \mathcal{H} : \|\mathbf{h}(\mathbf{Z})\|^2 < \infty, \exists \text{ a sequence of parametric submodels } \mathbf{B}_j\mathbf{S}_{\boldsymbol{\gamma}j}(\mathbf{Z}) \\ & \text{such that} \|\mathbf{h}(\mathbf{Z}) - \mathbf{B}_j\mathbf{S}_{\boldsymbol{\gamma}j}(\mathbf{Z})\|^2 \to 0, \text{as } j \to \infty],\end{aligned}$$

where $\|\mathbf{h}(\mathbf{Z})\|^2 = \mathrm{E}\{\mathbf{h}(\mathbf{Z})^{\mathrm{T}}\mathbf{h}(\mathbf{Z})\}$. In order for the projection Theorem to be guaranteed to apply, from now on, we always assume that Λ is a closed linear space. The semiparametric efficent score for $\boldsymbol{\beta}$ is defined as

$$\mathbf{S}_{\text{eff}}(\mathbf{Z},\boldsymbol{\beta}_0,\boldsymbol{\eta}_0) = \mathbf{S}_{\boldsymbol{\beta}}(\mathbf{Z},\boldsymbol{\beta}_0,\boldsymbol{\eta}_0) - \Pi\{\mathbf{S}_{\boldsymbol{\beta}}(\mathbf{Z},\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)|\Lambda\}.$$

It has been proved that the semiparametric efficiency bound is equal to the inverse of the variance matrix of $\mathbf{S}_{\text{eff}}(\mathbf{Z},\boldsymbol{\beta}_0,\boldsymbol{\gamma}_0)$, i.e., $\{\mathrm{E}(\mathbf{S}_{\text{eff}}\mathbf{S}_{\text{eff}}^{\mathrm{T}})\}^{-1}$. Hence, the semiparametric RAL estimator constructed with efficient influence function $\phi_{\text{eff}}(\mathbf{Z}) = \{\mathrm{E}(\mathbf{S}_{\text{eff}}\mathbf{S}_{\text{eff}}^{\mathrm{T}})\}^{-1}\mathbf{S}_{\text{eff}}(\mathbf{Z},\boldsymbol{\beta}_0,\boldsymbol{\eta}_0)$ is semiparametrically efficient.

2.3 B-Splines Approximation

There can be little doubt that smoothing has an crucial place in statistics. A variety of smoothing methods have been developed, such as kernel smoothers Silverman (2018), local polynomialsCleveland (1979), Stone (1977) and wavelets Daubechies (1992). Spline is also an important smoothing method, which comes in several varieties: smoothing splines, regression splines Eubank (1999), B-splinesDe Boor (2001), Dierckx (1995), P-splines Eilers & Marx (1996), etc.

A B-spline, short for basis spline, consists of polynomial pieces that are connected in a special way. De Boor (2001) is the basic reference. Start with a nondecreasing sequence

$\mathbf{t} := (t_j)$, the B-splines of order 1 are defined as

$$B_{j1}(t) := \begin{cases} 1, & t_j \leqslant t < t_{j+1}; \\ 0, & \text{otherwise}, \end{cases}$$

which are the characteristic functions of the partition (t_j). Based on the first-order B-splines, the higher-order B-splines are obtained by recurrence:

$$B_{jk} := \omega_{jk} B_{j,k-1} + (1 - \omega_{j+1,k}) B_{j+1,k-1},$$

where

$$\omega_{jk}(t) := \begin{cases} \frac{t - t_j}{t_{j+k-1} - t_j}, & t_j \neq t_{j+k-1}; \\ 0, & \text{otherwise}. \end{cases}$$

The kth order B-spline has the form

$$B_{jk} = \sum_{l=j}^{j+k-1} b_{lk} B_{l1},$$

where b_{lk} is a polynomial of degree $k-1$. Thus, B_{jk} is a piecewise polynomial of degree $k-1$ which vanishes outside the interval $[t_j, t_{j+k})$. For example, the second-order B-splines is given by $B_{j2}(t) = \omega_{j2} B_{j1}(t) + (1 - \omega_{j+1,2}) B_{j+1,1}(t)$, which consists of two linear pieces and vanishes outside the interval $[t_j, t_{j+2})$. The third order B-splines is given as $B_{j3}(t) = \omega_{j3}\omega_{j2} B_{j1}(t) + \{\omega_{j3}(1 - \omega_{j+1,2}) + (1 - \omega_{j+1,3})\omega_{j+1,2}\} B_{j+1,1}(t) + (1 - \omega_{j+1,3})(1 - \omega_{j+2,2}) B_{j+2,1}(t)$, which consists of three quadratic pieces and vanishes outside the interval $[t_j, t_{j+3})$.

Note that the kth order B-spline B_{jk} is completely determined by the $k+1$ knots $t_j, \ldots, t_{j+k}$. Many important properties of B-splines are derived by considering the linear span of all B-splines of a given order k and partition $\mathbf{t} = (t_j)$, which is denoted by $S_{k,t} := \left\{ \sum_j B_{jk} \boldsymbol{\gamma}_j \right\}$. De Boor (2001) proved that $S_{k,\mathbf{t}}$ contains the collection of all polynomials of degree $< k$. It also coincide with the space of all piecewise polynomials of degree $< k$ with break points t_j that are $k - 1 - n_j$ times continuously differentiable at t_j, where n_j is the multiplicity of knot t_j.

Consider the case where the support is an interval. To give full support to the first and last B-splines, besides N interior knots, we need to extend the partition by adding k endpoints on each side. The values of the endpoints do not matter, and usually the two boundary points are just repeated. Consider a regression of n data points (x_i, y_i) on

a set of $N + k$ B-splines B_{jk} for a given partition (t_j). Without loss of generality, we assume the support is $[0, 1]$ and define the knots $t_{-k+1} = \cdots = t_0 = 0 < t_2 < \cdots < t_N < 1 = t_{N+1} = \cdots t_{N+k}$. A fitted curve $\widehat{y}_i = \sum_{j=1}^{N+k} \boldsymbol{\gamma}_j B_{jk}(x_i)$ is a linear combination of B_{jk}. Parameter $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{N+k})$ is called spline coefficient. The simplest way to obtain a fair estimator of $\boldsymbol{\gamma}$ can be minimizing the least squares objective function

$$\sum_{i=1}^{n} \left\{ y_i - \sum_{j=1}^{N+k} \boldsymbol{\gamma}_j B_{jk}(x_i) \right\}^2 .$$

De Boor (2001) also discussed the evaluation of B-splines approximation. We denote the true function by $y = f_0(x)$, satisfying $f_0(x) \in C^q([0, 1])$, $1 \leqslant q \leqslant k$. Assume that the number of interior knots N satisfies $N \to \infty$, $N^{-1}n(\log n)^{-1} \to \infty$ and $Nn^{-1/(2q)} \to \infty$ when $n \to \infty$. Let h_j be the distance between the jth and $(j-1)$th interior knots, and let $h_b = \max_{2 \leqslant j \leqslant N} h_j$, $h_s = \min_{2 \leqslant j \leqslant N} h_j$. Assume $h_b/h_s < c_h$, where c_h is a positive finite constant. Then De Boor (2001) showed that there exists $\boldsymbol{\gamma}_0$ such that $\sup_{x \in [0,1]} \left| \sum_{j=1}^{N+k} \boldsymbol{\gamma}_{0j} B_{jk}(x) - f_0(x) \right| = O_p(h_b^q)$. This property is crucial to establishing the convergence rates of certain estimators whose construction involves B-splines approximation.

Different from kernel smoothers, when we approximate a unknown function with B-splines, we operationally would deal with a parametric model. In measurement error models, the idea of using spline representation is not new, especially in Bayesian framework Sarkar et al. (2014), Staudenmayer et al. (2008). One advantage is that with B-splines approximation, we avoid deconvolution which converges very slow. However, it becomes challenging to derive convergence rates of the resulting estimators. By assuming that the distribution of the unobservable covariate is compactly supported, Jiang & Ma (2018) established the asymptotic properties of the proposed spline-assisted semiparametric estimator in nonparametric measurement error models. The assumption about compact support is crucial, because it restricts the impact of the tails of the measurement error distribution. The success of this spline-assisted approach in achieving better convergence rates is encouraging and inspiring. The establishment of the asymptotic properties of our estimators proposed in Chapter 3 refers to its proofs.

2.4 Inference for High-Dimensional Data

In general, inference for high-dimensional data is a challenging problem with relatively small literature available. The major challenge lies in the difficulty of obtaining an accurate estimate of standard error for the resulting estimate in high-dimensional settings. Consider the linear regression model

$$Y_i = \boldsymbol{\theta}_0^{\mathrm{T}} \mathbf{X}_i + \epsilon_i, \quad \text{and} \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{2.1}$$

where $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ are i.i.d., and $\boldsymbol{\theta} \in \mathbb{R}^p$. In matrix form, let $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^{\mathrm{T}} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, and $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$. Then we have

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta}_0 + \boldsymbol{\epsilon}, \quad \text{and} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{n\times n}).$$

In low-dimensional settings, where n is greater than p, the ordinary least squares estimator, $\widehat{\boldsymbol{\theta}}_{\mathrm{OLS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}\mathbf{Y}$, is Gaussian with mean $\boldsymbol{\theta}_0$ and covariance $\sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. Confidence intervals can be directly constructed. However, in high-dimensional settings, matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is no longer invertible due to the deficient rank. Hence, one has to resort to biased estimators. The Lasso is one of the most popular approaches, which promotes sparse estimators via an l_1 penalty:

$$\widetilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 \right\}.$$

With the sparsity assumption, Fan et al. (2012) discussed the issue about spurious correlation that is inherent in high-dimensional problems in detail. Actually, the residual variance can be seriously underestimated when irrelevant covariates which have large sample correlations with the realized noises are selected in the model.

One effective way to attenuate the influence of spurious correlation is data splitting. The general idea is that we first randomly split the data into two halves. We use the first half to do model selection, and then use the other half to estimate parameters and construct confidence intervals based on the model selected using the first half of the data. Since the regression coefficients in the first stage are discarded and refitted using the second half of the data, the spurious correlations in the first stage are significantly reduced at the second stage. Dezeure et al. (2015), Wasserman & Roeder (2009) considered p-values based on sample splitting technique. However, loss of estimation accuracy and power is inevitable because only part of the data is used for parameter estimation and

inference. Fithian et al. (2014) argued that inference obtained by data splitting are only valid conditional on the model that was selected on the first half of the data. With this idea of 'conditional' inference, Lee et al. (2016) proposed post-model-selection-inference for model (2.1), which targets on regression coefficients conditional on the model selected by the Lasso. Let θ_j^M denote the jth parameter in model M, where $M \subset \{1, \ldots, p\}$. To construct the post-selection interval C_j^M for θ_j^M which satisfies

$$\Pr\left(\theta_j^M \in C_j^M \mid \widehat{M} = M\right) \geqslant 1 - \alpha,$$

the conditional distribution $\boldsymbol{\theta}_M^{\mathrm{T}}\mathbf{y} \mid \{\widehat{M} = M\}$ was studied carefully, where the model $\widehat{M}$ is selected from the Lasso. Although the optimal and exact confidence intervals for θ_j^M conditional on $\{\widehat{M} = M\}$ can be constructed, the validity of this procedure relies heavily on the performance of the models selection at the first step.

Javanmard & Montanari (2014), Zhang & Zhang (2014) constructed confidence intervals for coefficients of model (2.1) in high-dimensional setting from a totally different perspective. Also based on the Lasso, they targeted on the coefficients of the true model, rather than the coefficients of the selected model. In high-dimensional settings, the Lasso estimator $\widetilde{\boldsymbol{\theta}}$ is unavoidably biased. Thus, a natural idea is to de-bias. Javanmard & Montanari (2014) proposed a de-biased estimator with the form

$$\widehat{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}} + \frac{1}{n}\mathbf{M}\mathbf{X}^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\widetilde{\boldsymbol{\theta}}).$$

The basic intuition is that the bias caused by l_1 penalty in the Lasso will be compensated by adding a term proportional to the subgradient of the objective function at the Lasso solution. Let $\widehat{\boldsymbol{\Sigma}} = \mathbf{X}^{\mathrm{T}}\mathbf{X}/n$. By simple calculations,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}}\mathbf{M}\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon} + \sqrt{n}(\mathbf{M}\widehat{\boldsymbol{\Sigma}} - \mathbf{I}_{p\times p})(\boldsymbol{\theta}_0 - \widetilde{\boldsymbol{\theta}}),$$

where $\mathbf{M}\mathbf{X}^{\mathrm{T}}\boldsymbol{\epsilon}/\sqrt{n} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{M}\widehat{\boldsymbol{\Sigma}}\mathbf{M}^{\mathrm{T}})$. To minimize the bias and maximize the efficiency, the jth row of the $p \times p$ matrix $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_p)^{\mathrm{T}}$ is the solution of the following convex program:

$$\text{minimize} \;\; \mathbf{m}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\mathbf{m} \;\; \text{subject to} \;\; \|\widehat{\boldsymbol{\Sigma}}\mathbf{m} - \mathbf{e}_j\|_\infty \leqslant \lambda',$$

where $\mathbf{e}_j \in \mathbb{R}^p$ is the vector with one at the jth position and zero everywhere else. The asymptotic properties of the proposed unbiased estimator $\widehat{\boldsymbol{\theta}}$ was established carefully.

Zhang & Zhang (2014) focused on inference for individual coefficients in high-dimensional linear models (2.1) and corrected the bias in linear estimators using low dimensional projections. In low-dimensional settings, the ordinary least squares estimator of one regression coefficient θ_j is defined as

$$\widehat{\theta}_{j,\mathrm{OLS}} = \frac{(\mathbf{x}_j^{\perp})^{\mathrm{T}}\mathbf{Y}}{(\mathbf{x}_j^{\perp})^{\mathrm{T}}\mathbf{x}_j} = \theta_j + \frac{(\mathbf{x}_j^{\perp})^{\mathrm{T}}\boldsymbol{\epsilon}}{(\mathbf{x}_j^{\perp})^{\mathrm{T}}\mathbf{x}_j} + \sum_{k \neq j} \frac{(\mathbf{x}_j^{\perp})^{\mathrm{T}}\mathbf{x}_k\theta_k}{(\mathbf{x}_j^{\perp})^{\mathrm{T}}\mathbf{x}_j},$$

where $\mathbf{x}_j^{\perp}$ is the projection of $\mathbf{x}_j$ to the orthogonal complement of the column space of $X_{-j} = (\mathbf{x}_k, k \neq j)$. However, in high-dimensional settings, when $\mathbf{X}$ is in general position, $\mathrm{rank}(\mathbf{X}_{-j}) = n$ and $\mathbf{x}_j^{\perp} = 0$. Further, it is impossible to have $\mathbf{x}_j^{\perp} \neq 0$ and $(\mathbf{x}_j^{\perp})^{\mathrm{T}}\mathbf{x}_k = 0$ for all $k \neq j$. In order to make the estimator well-defined and correct the bias caused by $(\mathbf{x}_j^{\perp})^{\mathrm{T}}\mathbf{x}_k\theta_k$, the low dimensional projection estimator (LDPE) is defined as

$$\begin{aligned} \widehat{\theta}_j &= \frac{\mathbf{z}_j^{\mathrm{T}}\mathbf{Y}}{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j} - \sum_{k \neq j} \frac{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_k\widetilde{\theta}_{k,\mathrm{init}}}{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j} \\ &:= \widetilde{\theta}_{j,\mathrm{init}} + \frac{\mathbf{z}_j^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}_{\mathrm{init}})}{\mathbf{z}_j^{\mathrm{T}}\mathbf{x}_j}, \end{aligned}$$

where $\widetilde{\boldsymbol{\beta}}_{\mathrm{init}}$ is an initial estimator selected by the scaled Lasso, and

$$\mathbf{z}_j = \mathbf{x}_j - \mathbf{X}_{-j}\widehat{\boldsymbol{\zeta}}_j, \quad \widehat{\boldsymbol{\zeta}}_j = \arg\min_{\mathbf{b}} \left\{ \frac{\|\mathbf{x}_j - \mathbf{X}_{-j}\mathbf{b}\|_2^2}{2n} + \lambda_j'\|\mathbf{b}\|_1 \right\}.$$

Actually, $\mathbf{z}_j$ is a relaxed orthogonalization of $\mathbf{x}_j$ against other design vectors. Due to the intrinsic link between projection and regression, the unrelaxed $\mathbf{x}_j^{\perp}$ is just the residual of the least squares fit of $\mathbf{x}_j$ on $\mathbf{X}_{-j}$. One natural relaxation is penalized least squares method. In Zhang & Zhang (2014), l_1 penalty is chosen so that $\mathbf{z}_j$ is the residual of the Lasso.

Ning et al. (2017) generalized the idea of low dimensional projection and provided a general framework for high-dimensional inference by constructing decorrelated score functions. They considered general penalized M-estimator

$$\widetilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \Omega} l(\boldsymbol{\theta}) + P_{\lambda}(\boldsymbol{\theta}),$$

where $l(\boldsymbol{\theta})$ is a general loss function (e.g., the negative log-likelihood) and $P_{\lambda}(\boldsymbol{\theta})$ is a penalty function with tuning parameter λ. To assess the uncertainty for low-dimensional

components in high-dimensional models, parameter $\boldsymbol{\theta}$ is partitioned as $(\beta, \boldsymbol{\gamma})$, where β is the finite-dimensional parameter of interest and $\boldsymbol{\gamma}$ is the nuisance parameter. For simplicity, assume β is univariate. Denote the variance matrix of score $\nabla_\beta l(\boldsymbol{\theta})$ by matrix

$$\mathbf{I} = \begin{pmatrix} I_{\beta\beta} & \mathbf{I}_{\beta\gamma} \\ \mathbf{I}_{\gamma\beta} & \mathbf{I}_{\gamma\gamma} \end{pmatrix}.$$

The deccorelated score function for β is defined as

$$S(\beta, \boldsymbol{\gamma}) = \nabla_\beta l(\beta, \boldsymbol{\gamma}) - \boldsymbol{\omega}^{\mathrm{T}} \nabla_\gamma l(\beta, \boldsymbol{\gamma}), \quad \text{with } \boldsymbol{\omega}^{\mathrm{T}} = \mathbf{I}_{\beta\gamma} \mathbf{I}_{\gamma\gamma}^{-1}.$$

It is obvious that $S(\beta, \boldsymbol{\gamma})$ is uncorrelated with the nuisance score functions $\nabla_\gamma l(\beta, \boldsymbol{\gamma})$, which is crucial to control the variability of higher order terms in Taylor expansions and make the high-dimensional inference feasible. In low-dimensional settings, $S(\beta, \boldsymbol{\gamma})$ is essentially similar to efficient score function and the corresponding RAL estimator is semiparametrically efficient. However, in high-dimensional settings, the sample version of $\mathbf{I}_{\gamma\gamma}$ is not invertible, then $\widehat{\boldsymbol{\omega}}^{\mathrm{T}}$ is chosen to be the best sparse linear combination of the nuisance score functions $\nabla_\gamma l(\beta, \boldsymbol{\gamma})$ to approximate the score function $\nabla_\beta l(\beta, \boldsymbol{\gamma})$. Geometrically, roughly speaking, they assumed that the the projection of $\nabla_\beta l(\beta, \boldsymbol{\gamma})$ to nuisance tangent space spanned by the nuisance score functions is identical to the projection of $\nabla_\beta l(\beta, \boldsymbol{\gamma})$ to a low-dimensional subspace of the nuisance tangent space. Actually, decorrelated score function generates the idea of low dimensional projection in Zhang & Zhang (2014) from multivariate normal distribution to general score functions based on the asymptotic normality of score functions.

2.5 Missing Data

In statistical data analysis, missing data is ubiquitous. Special treatments are needed when analyzing data suffering from missingness. For example, one cannot simply set up a model for a single partially observed response y given a set of fully observed covariates $\mathbf{X}$. To decide how to handle missing data, it is helpful to know why they are missing. There are three types of "missingness mechanisms".

- Missing completely at random (MCAR). A variable is missing completely at random if the probability of missingness is the same for all units.
- Missing at random (MAR). In this case, missingness can be fully accounted for by

variables where there is complete information. We also call it ignorable missingness.

- Missing not at random (also known as nonignorable missingness). Finally, a particularly difficult situation arises when the probability of missingness depends on the missing values themselves.

Obviously, the first two kinds of missingness are easier to handle. When data are MCAR, complete-case analysis, which excludes all units for which the response or any of the covariates are missing, gives consistent estimators. Although it can be inefficient when there are a lot of missing values. When data are MAR, two commonly used approaches are inverse-probability weighting (IPW) (Höfler et al. 2005) and multiple imputation (MI) (Little & Rubin 2002). In IPW, only complete cases are included in the analysis, but weights are used to rebalance the set of complete cases so that it is representative of the whole sample. IPW requires only a model for the probability that a subject has complete data. In MI, missing data are replaced by data drawn from an imputation model, and it requires a model for the joint distribution of the missing data given the observed data. MI is generally more efficient than IPW but more complex. IPW and MI yield consistent estimators when the data are MAR and the imputation and weighting models, respectively, are correctly specified (Seaman et al. 2012). There also have been papers combining IPW and MI in different ways to construct estimators with preferred properties, such as robustness (Robins & Wang 2000, Seaman et al. 2012). MAR data has been thoroughly studied with extensive literature available, including many great books (Kim & Shao 2013, Little & Rubin 2002, Rubin 1987, Schafer 1997, Tsiatis 2006).

Unfortunately, we generally cannot tell whether data are MAR, or whether the missingness depends on the missing data themselves. The fundamental and intrinsic difficulty is that these potential "lurking variables" are partially observed, and so we can never rule them out. In other words, MAR in general is an assumption that is impossible to verify statistically. We must rely on its substantive reasonableness. The uncheckableness of MAR motivates research on nonignorable missingness, to some extent. There are also direct evidence supporting nonignorable missingness in many applications. For example, a man failed to fill in a depression survey because of his level of depression. Another good example can be that people with higher earnings are less likely to respond to the earnings question.

Nonignorable missing data is a challenging topic with less research on it. Model identifiability is a notorious issue for analyzing nonignorable missing data. Assumptions on the data-generating process or the missingness mechanism are required Robins &

Ritov (1997). d'Haultfoeuille (2010), Tang et al. (2003) assumed that the missingness only depends on the partially missing variable itself, which can be too restrictive in many applications. Shadow variable (Kott 2014), also known as nonresponse instrument, is prevalent in survey sampling designs. A variable is called shadow variable if it is conditionally independent of the missingness indicator given the response and the other covariates. Much more flexible missingness mechanism can be adopted if an appropriate shadow variable can be identified (Kott 2014, Miao et al. 2016, 2019, Shao & Wang 2016, Shao & Zhao 2013, Wang et al. 2014, Zhao & Ma 2018, Zhao & Shao 2015). With a suitable shadow variable, the exact model identifiability conditions still need to be investigated on a case-by-case fashion.

Chapter 3
Semiparametric Regression for Measurement Error Model with Heteroscedastic Error

3.1 Introduction

Covariate measurement errors arise frequently in many areas, such as chemistry, biological science, medicine and epidemiological studies. Instead of a precise measurement, we only have error-prone surrogates of the unobservable covariate. There has been extensive attention on the problem about measurement errors in covariates in the literature, see Fuller (1987, 2009) for measurement error in linear model, Carroll et al. (2006) for measurement error in nonlinear model and Yi (2016) for measurement error in a wider ranges of applications such as survival analysis, case-control studies, etc. Above work focuses on low-dimensional measurement error models, where the sample size n is smaller than the number of covarites. Existing methods often make restrictive and unrealistic assumptions about the measurement error distribution such as normality and homoscedasticity. However, in practice, the error density may violate these assumptions, which leads to erroneous estimation and inference Bertrand et al. (2017).

Measurement error heteroscedasticity is an issue requiring attention but is not an easy problem to handle. Staudenmayer et al. (2008) carefully studied the bias issues in density estimation if one incorrectly assumes homoscedasticity of the measurement errors. Similarly, in regression problems, ignoring the heteroscedasticity of the measurement errors can affect both the accuracy of the estimation and the quality of the inference. In an attempt to properly treat the heteroscedastic measurement error, Devanarayan &

Stefanski (2002) proposed empirical SIMEX method to accommodate heteroscedastic measurement error. However, they assumed normality of the measurement error and only provided an approximate solution. Cheng & Riu (2006) studied linear relationships when both the response variable and the covariates are subject to heteroscedastic errors using maximum likelihood method, method-of-moments and generalized least squares method, under a critical but restrictive assumption that the variances of the normally distributed measurement and regression errors for each observation are both known. Guo & Little (2011) extended the regression calibration and multiple imputation methods to allow heterocsedastic measurement error, while assuming the normality of the conditional density of the measurement errors given the unobservable covariate and assuming the variance function is a power function of the unobservable covariate. Sarkar et al. (2014) studied the regression model with heteroscedastic errors in covariates in a Bayesian hierarchical framework and avoided the assumptions about normality and homoscedasticity of the measurement as well as regression errors. However, due to the complexity of using both B-splines and Dirichlet processes, the theoretical properties of the estimator is not established.

In this Chapter, we consider a general low-dimensional parametric regression model with one of its covariates measured with heteroscedastic error when both the conditional density of the unobservable covariate given the error-free covariates and the variance function of the measurement errors are unknown. We allow the distributions of the regression error and the measurement error to have any form, hence not limited to the normal distribution family. This problem is difficult in the sense that we need to estimate two nuisance functions. A key observation is that if the unknown variance function had been parametric, we would have a simpler problem. Hence, we approximate the variance function with B-splines to convert the model to a simpler setting operationally. Through a semiparametric approach, we also avoid performing deconvolution. In fact, the idea of using B-spline approximation in the measurement error model has been used in the Bayesian framework Berry et al. (2002), Sarkar et al. (2014), although the theoretical impact of the B-spline approximation is unclear. Recently, in the nonparametric model framework, Jiang & Ma (2018) proposed a spline-assisted semiparametric approach to measurements error models where the asymptotic properties of the nonparametric estimator was established in homoscedastic measurement error case. Although the B-splines approximation makes the estimation of the variance function feasible, we still face the challenge of handling the conditional density function of the unobservable covariate. We avoid having to estimate it. Instead, we propose an estimator that allows

for misspecification of the conditional density function based on efficient score functions. The final estimator is obtained by solving estimating equations based on estimating functions that are approximated and without a closed form, hence it is very challenging to establish the asymptotic properties of the estimator. The method of analysis is also very different from that of typical analysis about splines. Our method enjoys good asymptotic properties in terms of convergence rate. Generally, estimating nuisance parameters will alter, often inflate the estimation variance for the parameter of interest. However, our method does not suffer this issue. Further, if density function of the unobservable covariate is correctly specified, the estimator of the parameter of interest is efficient.

3.2 Notation and Model Setup

3.2.1 Model Specification

In this Chapter, we use bold fonts for vectors and matrices and regular fonts for scalars. Let Y be the response variable and $\mathbf{Z}$ be the vector of observed error-free covariates. Let X be an unobservable latent covariate that is measured in an error-prone way. Let W be the observed surrogate of X. We assume the support of X to be finite; without loss of generality, let the support be $[0, 1]$. We are interested in the relationship between the response variable Y and the true covariates $(X, \mathbf{Z})$. In particular, we link the response to the covariates using a parametric model

$$f_{Y|X,\mathbf{Z}}(y, x, \mathbf{z}, \boldsymbol{\beta}), \tag{3.1}$$

where $f_{Y|X,\mathbf{Z}}$ is the specified conditional probability density function or probability mass function of Y given X and $\mathbf{Z}$, and $\boldsymbol{\beta}$ is a p-dimensional vector of unknown parameters.

We further assume that W, the observed surrogate of X, is linked to X through a heteroscedastic measurement error model

$$W = X + \sigma(X)U, \tag{3.2}$$

where U is a random variable with a known density function $f_U(u)$, which is not restricted to the normal family. Without loss of generality, we assume that U has mean zero and variance 1. We also assume that U is independent of X and $\mathbf{Z}$. Let $\sigma(X)$ be an unknown positive nuisance function that describes the heteroscedasticity of the measurement error. Our goal is to estimate the regression parameter $\boldsymbol{\beta}$ based on a sample

$(Y_1, W_1, \mathbf{Z}_1), \ldots, (Y_n, W_n, \mathbf{Z}_n)$.

3.2.2 Identifiability Considerations

To ensure the identifiability of the model, we assume that two repeated measurements of the error-prone covariate X are available. Let W_1 and W_2 be two measurements of X, which are independent with each other conditional on X. To prove identifiability, we aim to show that if the density or mass function of the observed data conditional on $\mathbf{Z}$ satisfies $f_{Y,W_1,W_2|\mathbf{Z}}(y, w_1, w_2, \mathbf{z}, \boldsymbol{\beta}, \sigma, f_{X|\mathbf{Z}}) = f_{Y,W_1,W_2|\mathbf{Z}}(y, w_1, w_2, \mathbf{z}, \widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}, \widetilde{f}_{X|\mathbf{Z}})$, then $(\boldsymbol{\beta}, \sigma, \mathbf{f}_{X|\mathbf{Z}}) = (\widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}, \widetilde{f}_{X|\mathbf{Z}})$. Here $f_{X|\mathbf{Z}}(x|\mathbf{z})$ is the conditional density of X given $\mathbf{Z}$. Note that

$$
\begin{aligned}
&f_{Y,W_1,W_2|\mathbf{Z}}(y, w_1, w_2, \mathbf{z}, \boldsymbol{\beta}, \sigma, f_{X|\mathbf{Z}}) \\
&\qquad = \int f_{Y|X,\mathbf{Z}}(y|x, \mathbf{z}, \boldsymbol{\beta}) f_U \left\{ \frac{w_1 - x}{\sigma(x)} \right\} f_U \left\{ \frac{w_2 - x}{\sigma(x)} \right\} f_{X|\mathbf{Z}}(x|\mathbf{z}) \, \frac{1}{\sigma^2(x)} \, dx
\end{aligned}
$$

and, likewise,

$$
\begin{aligned}
&f_{Y,W_1,W_2|\mathbf{Z}}\{y, w_1, w_2, \mathbf{z}, \widetilde{\boldsymbol{\beta}}, \widetilde{\sigma}, \widetilde{f}_{X|\mathbf{Z}}\} \\
&\qquad = \int f_{Y|X,\mathbf{Z}}(y|x, \mathbf{z}, \widetilde{\boldsymbol{\beta}}) f_U \left\{ \frac{w_1 - x}{\widetilde{\sigma}(x)} \right\} f_U \left\{ \frac{w_2 - x}{\widetilde{\sigma}(x)} \right\} \widetilde{f}_{X|\mathbf{Z}}(x|\mathbf{z}) \, \frac{1}{\widetilde{\sigma}^2(x)} \, dx.
\end{aligned}
$$

When we leave the model of Y given X and $\mathbf{Z}$ as an arbitrary known parametric model, it is very difficult to prove identifiability without adding many conditions that are difficult to check. Thus, we believe it is a better strategy to establish identifiability in a case-by-case fashion. Here, as an example, we consider a specific situation where the main model is linear with heteroscedastic normal or Laplace measurement errors, i.e., $Y = \beta_0 + X\beta_1 + \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_2 + \epsilon$ and $W_j = X + \sigma(X)U_j$ for $j \in \{1, 2\}$, where ϵ has a mean zero normal distribution with variance σ_ϵ^2 and U_1, U_2 are standard normal random variables or Laplace random variables with mean 0 and variance 1. The identifiability of this specific model can be established by computing the first and the second moments of Y, W_1 and W_2 given $\mathbf{Z}$ and using Fourier transform. The detailed proof is given in Subsection A.1.

3.3 Methodology

3.3.1 Estimator of Original Model

The score function of the parametric model given in (3.1) is $\partial \ln\{f_{Y|X,\mathbf{Z}}(y|x,\mathbf{z},\boldsymbol{\beta})\}/\partial\boldsymbol{\beta}$. If the covariate X were observed precisely, a consistent estimate of the parameter $\boldsymbol{\beta}$ could be obtained by solving the sample version of

$$\mathrm{E}[\partial \ln\{f_{Y|X,\mathbf{Z}}(Y|X,\mathbf{Z},\boldsymbol{\beta})\}/\partial\boldsymbol{\beta}] = \mathbf{0}.$$

However, since X is unobservable and only W_1 and W_2 are available, we have to rely on the conditional density of (Y, W_1, W_2) given $\mathbf{Z}$,

$$f_{Y,W_1,W_2|\mathbf{Z}}(y,w_1,w_2,\mathbf{z},\boldsymbol{\beta},\sigma,f_{X|\mathbf{Z}}) = \int \frac{1}{\sigma^2(x)}\, f_{Y|X,\mathbf{Z}}(y|x,\mathbf{z},\boldsymbol{\beta}) f_U(u_1) f_U(u_2) f_{X|\mathbf{Z}}(x|\mathbf{z}) dx,$$

where $u_j = (w_j - x)/\sigma(x)$, for $j \in \{1,2\}$. Here the p-dimensional parameter $\boldsymbol{\beta}$ is of interest, and infinite-dimensional parameters $f_{X|\mathbf{Z}}$ and σ are nuisance. The "observed" score function with respect to $\boldsymbol{\beta}$ is given as

$$\mathbf{S}_\beta(y,w_1,w_2,\mathbf{z},\boldsymbol{\beta},\sigma,f_{X|\mathbf{Z}}) = \frac{\int\{\partial f_{Y|X,\mathbf{Z}}(y|x,\mathbf{z},\boldsymbol{\beta})/\partial\boldsymbol{\beta}\} f_U(u_1) f_U(u_2) f_{X|\mathbf{Z}}(x|\mathbf{z})/\sigma^2(x) dx}{\int f_{Y|X,\mathbf{Z}}(y|x,\mathbf{z},\boldsymbol{\beta}) f_U(u_1) f_U(u_2) f_{X|\mathbf{Z}}(x|\mathbf{z})/\sigma^2(x) dx},$$

which is the partial derivative of the logarithm of the likelihood $f_{Y,W_1,W_2|\mathbf{Z}}$ with respect to the parameter $\boldsymbol{\beta}$. Although $\mathrm{E}\{\mathbf{S}_\beta(Y,W_1,W_2,\mathbf{z},\boldsymbol{\beta},\sigma,f_{X|\mathbf{Z}})\} = \mathbf{0}$ at the true parameter values, it is impossible to estimate $\boldsymbol{\beta}$ by solving

$$\sum_{i=1}^{n} \mathbf{S}_\beta(y_i,w_{i1},w_{i2},\mathbf{z},\boldsymbol{\beta},\sigma,f_{X|\mathbf{Z}}) = \mathbf{0}$$

directly due to the presence of the nuisance parameters.

We thus take a different approach and try to construct estimators $\widehat{\boldsymbol{\beta}}_n$ by directly identifying the influence functions. A regular asymptotic linear estimator $\widehat{\boldsymbol{\beta}}_n$ can be written as

$$\sqrt{n}\,(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = n^{-1/2} \sum_{i=1}^{n} \phi(Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}) + o_p(1),$$

where $\phi(Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta})$ is a p-dimensional zero-mean random vector referred to as

the ith influence function of the estimator $\widehat{\boldsymbol{\beta}}_n$. From a geometric point of view, in the original model described by (3.1) and (3.2), an influence function of a single observation lies in the Hilbert space $\mathcal{H}$ of all p-dimensional zero-mean measurable functions of the observed data with finite variance, equipped with the inner product $< h_1, h_2 >= \mathrm{E}\{h_1^{\mathrm{T}}(Y, W_1, W_2, \mathbf{Z})h_2(Y, W_1, W_2, \mathbf{Z})|\mathbf{Z}\}$, where $h_1,\ h_2 \in \mathcal{H}$. Further, influence functions belong to the linear space orthogonal to the nuisance tangent space, which is defined as the mean squared closure of the nuisance tangent spaces of parametric sub-models spanned by the nuisance score vectors.

In the original models from (3.1) and (3.2), the nuisance space is given as $\Lambda = \Lambda_{f_{X|\mathbf{Z}}} + \Lambda_\sigma$, where

$$\Lambda_{f_{X|\mathbf{Z}}} = \{\mathrm{E}\{\mathbf{a}(X, \mathbf{Z})|Y, W_1, W_2, \mathbf{Z}\} : \mathrm{E}\{\mathbf{a}(X, \mathbf{Z})|\mathbf{Z}\} = \mathbf{0}\},$$
$$\Lambda_\sigma = \{\mathrm{E}\{V(U_1, U_2)\mathbf{b}(X)|Y, W_1, W_2, \mathbf{Z}\} : \forall \mathbf{b}(X)\}.$$

Here $V(U_1, U_2) = f'_U(U_1)U_1/f_U(U_1) + f'_U(U_2)U_2/f_U(U_2) + 2$, where $U_j = (W_j - X)/\sigma(X)$, for $j \in \{1, 2\}$. The detailed proof of the result concerning Λ is in Appendix A.2. Note that the two subspaces $\Lambda_{f_{X|\mathbf{Z}}}$ and Λ_σ are not orthogonal to each other, hence we write the orthogonal complement of Λ as $\Lambda^\perp = \Lambda^\perp_{f_{X|\mathbf{Z}}} \cap \Lambda^\perp_\sigma$. Here, $\Lambda^\perp_{f_{X|\mathbf{Z}}}$ is the orthogonal complement of $\Lambda_{f_{X|\mathbf{Z}}}$ and has the form

$$\Lambda^\perp_{f_{X|\mathbf{Z}}} = \{\mathbf{h}(Y, W_1, W_2, \mathbf{Z}) : \mathrm{E}\{\mathbf{h}(Y, W_1, W_2, \mathbf{Z})|X, \mathbf{Z}\} = \mathbf{0} \text{ almost everywhere}\},$$

while $\Lambda^\perp_\sigma$ is the orthogonal complement of Λ_σ and is given by

$$\Lambda^\perp_\sigma = \{\mathbf{h}(Y, W_1, W_2, \mathbf{Z}) : \mathrm{E}\{\mathbf{h}(Y, W_1, W_2, \mathbf{Z})V(U_1, U_2)|X, \mathbf{Z}\} = \mathbf{0} \text{ almost everywhere}\}.$$

Even without an explicit form of $\Lambda^\perp$, we can still derive the orthogonal projection of $\mathbf{S}_\beta(y, w_1, w_2, \mathbf{z}, \boldsymbol{\beta}, \sigma, f_{X|\mathbf{Z}})$ onto $\Lambda^\perp$. We write the orthogonal projection as $\mathbf{S}_{\text{eff}}(y, w_1, w_2, \mathbf{z}, \boldsymbol{\beta}, \sigma, f_{X|\mathbf{Z}})$ and call it the efficient score. It is obvious that the asymptotic variance of a regular asymptotically linear estimator equals the variance of its influence function. Consequently, the optimal estimator among a class of regular asymptotically linear estimators is the one whose influence function has the smallest variance, which we call the efficient influence function. The efficient score directly leads to the efficient influence function through the form

$$\boldsymbol{\phi}_{\text{eff}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta})$$

$$= [\mathrm{E}\{\mathbf{S}_{\text{eff}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \sigma, f_{X|\mathbf{Z}})^{\otimes 2}\}]^{-1}\mathbf{S}_{\text{eff}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \sigma, f_{X|\mathbf{Z}}),$$

where $\mathbf{a}^{\otimes 2} \equiv \mathbf{a}\mathbf{a}^{\mathrm{T}}$ for any vector or matrix $\mathbf{a}$, and this convention is used throughout this chapter. Thus, for the purpose of constructing estimating equations, we only need to identify $\mathbf{S}_{\text{eff}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \sigma, f_{X|\mathbf{Z}})$. It is easy to verify that

$$\begin{aligned}\mathbf{S}_{\text{eff}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \sigma, f_{X|\mathbf{Z}}) &= \mathbf{S}_{\boldsymbol{\beta}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \sigma, f_{X|\mathbf{Z}}) \\ &\quad - \mathrm{E}\{\mathbf{a}(X, \mathbf{Z})|Y, W_1, W_2, \mathbf{Z}\} - \mathrm{E}\left\{V(U_1, U_2)\mathbf{b}(X)|Y, W_1, W_2, \mathbf{Z}\right\},\end{aligned}$$

where $\mathbf{a}(X, \mathbf{Z})$ and $\mathbf{b}(X)$ are functions that satisfy

$$\begin{aligned}&\mathrm{E}\{\mathbf{S}_{\boldsymbol{\beta}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \sigma, f_{X|\mathbf{Z}})|X, \mathbf{Z}\} \\ &\quad = \mathrm{E}[\mathrm{E}\{\mathbf{a}(X, \mathbf{Z})|Y, W_1, W_2, \mathbf{Z}\}|X, \mathbf{Z}] + \mathrm{E}[\mathrm{E}\{V(U_1, U_2)\mathbf{b}(X)|Y, W_1, W_2, \mathbf{Z}\}|X, \mathbf{Z}],\end{aligned}$$

and

$$\begin{aligned}&\mathrm{E}\{\mathbf{S}_{\boldsymbol{\beta}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \sigma, f_{X|\mathbf{Z}})V(U_1, U_2)|X, \mathbf{Z}\} \\ &\qquad = \mathrm{E}[\mathrm{E}\{\mathbf{a}(X, \mathbf{Z})|Y, W_1, W_2, \mathbf{Z}\}V(U_1, U_2)|X, \mathbf{Z}] \\ &\qquad\quad + \mathrm{E}[\mathrm{E}\{V(U_1, U_2)\mathbf{b}(X)|Y, W_1, W_2, \mathbf{Z}\}V(U_1, U_2)|X, \mathbf{Z}].\end{aligned}$$

By the definition of $\mathbf{a}(X, \mathbf{Z})$ and $\mathbf{b}(X)$, it is obvious that

$$\mathrm{E}\{\mathbf{S}_{\text{eff}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \sigma, f_{X|\mathbf{Z}})|X, \mathbf{Z}\} = \mathbf{0}.$$

Hence, an efficient estimator of $\boldsymbol{\beta}$ can be obtained by solving the estimating equation $\sum_{i=1}^{n} \mathbf{S}_{\text{eff}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \sigma, f_{X|\mathbf{Z}}) = \mathbf{0}$ if we know the true σ and $f_{X|\mathbf{Z}}$.

Remark 3.3.1. *One way to ensure the identifiability of a problem is to ensure that the efficient score is not identically equal to zero. This is true as long as $\Lambda^{\perp}$ is not a zero space. In Subsection A.3, we show this to be the case for the logistic regression model with normal measurement errors. In other models, when a rigorous proof for identifiability is hard to obtain, numerical calculation of $\mathbf{S}^{*}_{\text{eff}}$ can provide some insights to the identifiability issue. For example, in all the simulation studies and real data example conducted in Section 3.6, $\mathbf{S}^{*}_{\text{eff}}$ is not close to zero at any arbitrary parameter values, indicating the identifiability of the corresponding models.*

An interesting discovery here is that even without knowing the true σ and $f_{X|\mathbf{Z}}$, we

can still construct the estimating equation in the same fashion after adopting a working model $f^*_{X|\mathbf{Z}}$. A Similar observation is also made in Tsiatis & Ma (2004). Specifically, we find that

$$\mathrm{E}\{\mathbf{S}^*_{\text{eff}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \sigma, f^*_{X|\mathbf{Z}})|X, \mathbf{Z}\} = \mathrm{E}^*\{\mathbf{S}^*_{\text{eff}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \sigma, f^*_{X|\mathbf{Z}})|X, \mathbf{Z}\} = \mathbf{0},$$

where the superscript * denotes the corresponding quantities, such as the efficient score $\mathbf{S}_{\text{eff}}$ and the expectations, calculated with the unknown $f_{X|\mathbf{Z}}$ replaced by the possibly misspecified working model $f^*_{X|\mathbf{Z}}$ everywhere it appears in the construction. Thus,

$$\sum_{i=1}^{n} \mathbf{S}^*_{\text{eff}}(Y_i, W_{1i}, W_{2i}, \mathbf{Z}_i, \boldsymbol{\beta}, \sigma, f^*_{X|\mathbf{Z}}) = \mathbf{0}$$

is a consistent estimating equation set.

3.3.2 Estimator of Approximate Model

We have seen that although we have obtained $\mathbf{S}_{\text{eff}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \sigma, f_{X|\mathbf{Z}})$, it is not realistic to use it. One reason is that it relies on the unknown conditional density function $f_{X|\mathbf{Z}}$. We have circumvented this difficulty by adopting a working model $f^*_{X|\mathbf{Z}}$ as in Section 3.3.1. The other obstacle we encounter in implementing $\mathbf{S}_{\text{eff}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \sigma, f_{X|\mathbf{Z}})$ lies in σ, and it also holds in implementing $\mathbf{S}^*_{\text{eff}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \sigma, f^*_{X|\mathbf{Z}})$. To overcome this obstacle, we propose to estimate $\sigma(X)$ using spline approximation $\mathbf{B}(X)^{\mathrm{T}}\boldsymbol{\gamma}$. If a suitable estimator $\widehat{\boldsymbol{\gamma}}$ can be obtained, then we can use $\widehat{\sigma}(X) = \mathbf{B}(X)^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}$ in place of $\sigma(X)$ to facilitate the construction of the estimating equations using $\mathbf{S}^*_{\text{eff}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \widehat{\sigma}, f^*_{X|\mathbf{Z}})$.

To estimate $\boldsymbol{\gamma}$, we consider the approximate model

$$f_{Y|X,\mathbf{Z}}(y|x, \mathbf{z}, \boldsymbol{\beta}) \quad \text{and} \quad W = X + \mathbf{B}(X)^{\mathrm{T}}\boldsymbol{\gamma}U, \tag{3.3}$$

which allows us to estimate $\boldsymbol{\gamma}$ at any $\boldsymbol{\beta}$. In approximate model (3.3), the density of the observed data conditional on $\mathbf{Z}$ is given by

$$\begin{aligned} &f_{a,Y,W_1,W_2|\mathbf{Z}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f_{X|\mathbf{Z}}) \\ &\qquad = \int f_{Y|X,\mathbf{Z}}(y|x, \mathbf{z}, \boldsymbol{\beta}) f_U\left\{\frac{w_1 - x}{\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}}\right\} f_U\left\{\frac{w_2 - x}{\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}}\right\} f_{X|\mathbf{Z}}(x|\mathbf{z}) \frac{1}{\{\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}\}^2}\, dx, \end{aligned}$$

where the p-dimensional parameter $\boldsymbol{\beta}$ is of interest, while the d_γ-dimensional parameter $\boldsymbol{\gamma}$ and function $f_{X|\mathbf{Z}}$ are nuisance parameters. Here and throughout the text, we use the

subscript $_a$ to denote quantities pertaining to approximate model (3.3). In this model, the influence functions of a single observation for regular asymptotically linear estimators of $\boldsymbol{\beta}$ lie in the Hilbert space $\mathcal{H}_a$ of all p-dimensional zero-mean measurable functions of observed data with finite variance, equipped with the inner product $< h_1, h_2 >= \mathrm{E}_a\{h_1^{\mathrm{T}}(Y, W_1, W_2, \mathbf{Z})h_2(Y, W_1, W_2, \mathbf{Z})|\mathbf{Z}\}$, where $h_1,\ h_2 \in \mathcal{H}_a$.

As in Section 3.3.1, the nuisance tangent space is $\Lambda_a = \Lambda_{a,f_{X|\mathbf{Z}}} + \Lambda_{a,\boldsymbol{\gamma}}$, where

$$\begin{aligned}\Lambda_{a,f_{X|\mathbf{Z}}} &= \{\mathrm{E}_a\{\mathbf{c}(X, \mathbf{Z})|Y, W_1, W_2, \mathbf{Z}\} : \mathrm{E}_a\{\mathbf{c}(X, \mathbf{Z})|\mathbf{Z}\} = \mathbf{0}\},\\ \Lambda_{a,\boldsymbol{\gamma}} &= \{\mathrm{E}_a\{V(U_{a,1}, U_{a,2})\mathbf{KB}(X)|Y, W_1, W_2, \mathbf{Z}\} : \mathbf{K} \text{ is a } p \times d_{\boldsymbol{\gamma}} \text{ constant matrix}\},\end{aligned}$$

and $U_{a,j} = (W_j - X)/\{\mathbf{B}(X)^{\mathrm{T}}\boldsymbol{\gamma}\}$, for $j \in \{1, 2\}$. If we treat $\boldsymbol{\gamma}$ as part of the parameters of interest, the efficient score for $\boldsymbol{\gamma}$ is the residual of the score vector for $\boldsymbol{\gamma}$ after projecting it on to $\Lambda_{a,f_{X|\mathbf{Z}}}$. Following a derivation similar to that for $\mathbf{S}_{\mathrm{eff}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \sigma, f_{X|\mathbf{Z}})$, the efficient score for $\boldsymbol{\gamma}$ is given as

$$\begin{aligned}&\mathbf{S}_{a,\mathrm{eff},\boldsymbol{\gamma}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f_{X|\mathbf{Z}})\\ &\qquad = \mathbf{S}_{a,\boldsymbol{\gamma}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f_{X|\mathbf{Z}}) - \mathrm{E}_a\{\mathbf{c}(X, \mathbf{Z})|Y, W_1, W_2, \mathbf{Z}\}, \quad (3.4)\end{aligned}$$

where $\mathbf{S}_{a,\boldsymbol{\gamma}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f_{X|\mathbf{Z}}) \equiv \partial \ln\{f_{a,Y,W_1,W_2|\mathbf{Z}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f_{X|\mathbf{Z}})\}/\partial\boldsymbol{\gamma}$ is the score vector for $\boldsymbol{\gamma}$, and $\mathbf{c}(X, \mathbf{Z})$ satisfies

$$\mathrm{E}_a\{\mathbf{S}_{a,\boldsymbol{\gamma}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f_{X|\mathbf{Z}})|X, \mathbf{Z}\} = \mathrm{E}_a[\mathrm{E}_a\{\mathbf{c}(X, \mathbf{Z})|Y, W_1, W_2, \mathbf{Z}\}|X, \mathbf{Z}]. \quad (3.5)$$

The detailed derivation is in Subsection A.4. We can estimate $\boldsymbol{\gamma}$ as a function of $\boldsymbol{\beta}$ through solving the estimating equations

$$\sum_{i=1}^{n} \mathbf{S}^*_{a,\mathrm{eff},\boldsymbol{\gamma}}(Y_i, W_{1i}, W_{2i}, \mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}) = \mathbf{0},$$

where $f^*_{X|\mathbf{Z}}$ is the working model.

Summarizing the above methods, we describe the detailed estimation procedure for $\boldsymbol{\beta}$ in approximate model (3.3) in the following algorithm.

Algorithm

1: Adopt a working model for $f_{X|\mathbf{Z}}(x|\mathbf{z})$ and denote it as $f^*_{X|\mathbf{Z}}(x|\mathbf{z})$.

2: Select a B-spline representation $\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}$ for $\sigma(x)$ with spline order r. Define the knots $t_{-r+1} = \cdots = t_0 = 0 < t_1 < \cdots < t_N < 1 = t_{N+1} = \cdots = t_{N+r}$, where N is the number of interior knots.

3: Solve the estimating equation for $\boldsymbol{\gamma}$

$$\sum_{i=1}^{n} \mathbf{S}^*_{a,\mathrm{eff},\boldsymbol{\gamma}}(Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}})$$
$$= \sum_{i=1}^{n} [\mathbf{S}^*_{a,\boldsymbol{\gamma}}(Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}) - \mathrm{E}^*_a\{\mathbf{c}(X, \mathbf{Z}_i)|Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i\}] = \mathbf{0}$$

to obtain $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta})$, where $\mathbf{c}(X, \mathbf{Z})$ satisfies

$$\mathrm{E}_a\{\mathbf{S}^*_{a,\boldsymbol{\gamma}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}})|X, \mathbf{Z}\} = \mathrm{E}_a[\mathrm{E}^*_a\{\mathbf{c}(X, \mathbf{Z})|Y, W_1, W_2, \mathbf{Z}\}|X, \mathbf{Z}]. \quad (3.6)$$

4: Solve the estimating equation for $\boldsymbol{\beta}$

$$\sum_{i=1}^{n} \mathbf{S}^*_{\mathrm{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta}), f^*_{X|\mathbf{Z}}\} = \sum_{i=1}^{n} [\mathbf{S}^*_{\boldsymbol{\beta}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta}), f^*_{X|\mathbf{Z}}\}$$
$$- \mathrm{E}^*_a\{\mathbf{a}(X, \mathbf{Z}_i)|Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i\} - \mathrm{E}^*_a\{V(U_{a,i1}, U_{a,i2})\mathbf{b}(X)|Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i\}] = \mathbf{0} \quad (3.7)$$

to obtain $\widehat{\boldsymbol{\beta}}$, where $\mathbf{a}(X, \mathbf{Z})$ and matrix $\mathbf{b}(X)$ satisfy

$$\mathrm{E}_a[\mathbf{S}^*_{\boldsymbol{\beta}}\{Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta}), f^*_{X|\mathbf{Z}}\}|X, \mathbf{Z}] = \mathrm{E}_a[\mathrm{E}^*_a\{\mathbf{a}(X, \mathbf{Z})|Y, W_1, W_2, \mathbf{Z}\}|X, \mathbf{Z}]$$
$$+ \mathrm{E}_a[\mathrm{E}^*_a\{V(U_{a,1}, U_{a,2})\mathbf{b}(X)|Y, W_1, W_2, \mathbf{Z}\}|X, \mathbf{Z}], \quad (3.8)$$

and

$$\mathrm{E}_a[\mathbf{S}^*_{\boldsymbol{\beta}}\{Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta}), f^*_{X|\mathbf{Z}}\}V(U_{a,1}, U_{a,2})|X, \mathbf{Z}]$$
$$= \mathrm{E}_a[\mathrm{E}^*_a\{\mathbf{a}(X, \mathbf{Z})|Y, W_1, W_2, \mathbf{Z}\}V(U_{a,1}, U_{a,2})|X, \mathbf{Z}]$$
$$+ \mathrm{E}_a[\mathrm{E}^*_a\{V(U_{a,1}, U_{a,2})\mathbf{b}(X)|Y, W_1, W_2, \mathbf{Z}\}V(U_{a,1}, U_{a,2})|X, \mathbf{Z}].$$

$\mathbf{S}^*_{\mathrm{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta}), f^*_{X|\mathbf{Z}}\}$ is $\mathbf{S}^*_{\mathrm{eff}}(Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \sigma, f^*_{X|\mathbf{Z}})$ with $\sigma(X_i)$ replaced by $\mathbf{B}(X_i)^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta})$.

Note that all the calculations with * should be conducted under the posited model $f^*_{X|\mathbf{Z}}$. We note that since the functions $\mathbf{a}(X, \mathbf{Z})$ and $\mathbf{c}(X, \mathbf{Z})$ satisfy (3.8) and (3.6), they automatically satisfy $\mathrm{E}^*_a\{\mathbf{a}(X, \mathbf{Z})|\mathbf{Z}\} = \mathbf{0}$, and $\mathrm{E}^*_a\{\mathbf{c}(X, \mathbf{Z})|\mathbf{Z}\} = \mathbf{0}$. Hence, $\mathrm{E}^*_a\{\mathbf{c}(X, \mathbf{Z})|Y, W_1, W_2, \mathbf{Z}\}$ and $\mathrm{E}^*_a\{\mathbf{a}(X, \mathbf{Z})|Y, W_1, W_2, \mathbf{Z}\}$ are indeed in $\Lambda^*_{a,f_{X|\mathbf{Z}}}$.

Remark 3.3.2. So far, we have assumed that there are at least two observations, W_1 and W_2, for each subject. Our method can also be applied to the cases with partial replication. For the subjects with replication, we construct estimating equations following the method given in Section 3.3 exactly. For the subjects that have only one W available, we simply modify our model using one W, then construct corresponding estimating equations. A consistent estimator for θ can be obtained by solving the two sets of estimating equations together. More details about the model with one W are given in Subsection A.7.

3.4 Asymptotic Properties

To facilitate the proof of the asymptotic results, we first provide a list of regularity conditions.

(C1) The true density $f_{X|\mathbf{Z}}(x|\mathbf{z})$ at any $\mathbf{z}$ is a bounded function of x with compact support.

(C2) The function $\sigma(x) \in C^q([0, 1])$, $q > 1$, is bounded with compact support.

(C3) The spline order $r \geqslant q$.

(C4) In B-splines approximation, let the number of interior knots N satisfy $N \to \infty$, $N^{-1}n\{\ln(n)\}^{-1} \to \infty$ and $Nn^{-1/(2q)} \to \infty$ as $n \to \infty$. Let $d_{\boldsymbol{\gamma}}$ denote the number of spline bases and $d_{\boldsymbol{\gamma}} = N + r$.

(C5) Let h_j be the distance between the jth and $(j-1)$th interior knots. Let $h_b = \max_{1\leqslant j\leqslant N} h_j$ and $h_s = \min_{1\leqslant j\leqslant N} h_j$. There exists a constant $c_h \in (0, \infty)$ such that $h_b/h_s < c_h$. Hence, $h_b = O(N^{-1})$ and $h_s = O(N^{-1})$.

(C6) The equation set

$$\mathrm{E}\{\mathbf{S}^*_{\mathrm{eff}}(Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}})\} = \mathbf{0}, \quad \mathrm{E}\{\mathbf{S}^*_{a,\mathrm{eff},\boldsymbol{\gamma}}(Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}})\} = \mathbf{0}$$

has a unique root for $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$ in the neighborhood of the true parameters $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^{\mathrm{T}}, \boldsymbol{\gamma}_0^{\mathrm{T}})^{\mathrm{T}}$. The derivatives with respect to $\boldsymbol{\theta}$ on the left-hand side are smooth

functions of $\boldsymbol{\theta}$, with singular values bounded above and bounded away from $\mathbf{0}$ in this neighborhood. Let the unique root be $\boldsymbol{\theta}^*$. Note that $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}^*$ are functions of N, that is, for any sufficiently large N; there is a unique root $\boldsymbol{\theta}^*$ in the neighborhood of $\boldsymbol{\theta}_0$.

(C7) For a matrix $\mathbf{A} = (a_{ij})$, denote $\|\mathbf{A}\|_\infty = \max_i \sum_j |a_{ij}|$ and $\|A\|_2 = \lambda_{\max}(\mathbf{A})$, where $\lambda_{\max}(\mathbf{A})$ represents the largest singular value of matrix $\mathbf{A}$. The following terms are integrable:

$$\|\partial \mathbf{S}^*_{\text{eff}}(y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, f^*_{X|\mathbf{Z}})/\partial \boldsymbol{\gamma}_0^{\mathrm{T}}\|_\infty, \quad \|\mathbf{S}^*_{\text{eff}}(y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \sigma, f^*_{X|\mathbf{Z}})\|_\infty,$$

$$\|\mathbf{S}^*_{\text{eff}}(y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \sigma, f^*_{X|\mathbf{Z}})\{\mathbf{G}^{\mathrm{T}}\mathbf{S}_{a,\gamma}(y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, f^*_{X|\mathbf{Z}})\}^{\mathrm{T}}\|_\infty$$

and

$$\|\{\mathbf{G}^{\mathrm{T}}\mathbf{S}_{a,\gamma}(y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, f_{X|\mathbf{Z}})\}^{\mathrm{T}} f_{Y,W_1,W_2|\mathbf{Z}}(y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \sigma, f_{X|\mathbf{Z}}) f_{\mathbf{Z}}(\mathbf{z}_i)\|_\infty,$$

where $\mathbf{S}^*_{a,\text{eff},\gamma}(Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \sigma, f^*_{X|\mathbf{Z}})$ is defined as $\mathbf{S}^*_{a,\text{eff},\gamma}(Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}})$ with $\mathbf{B}(X_i)^{\mathrm{T}}\boldsymbol{\gamma}$ replaced by $\sigma(X_i)$, and $\mathbf{G}$ is arbitrary $d_\gamma \times p$ matrix with $\|\mathbf{G}\|_2 = 1$.

Remark 3.4.1. From (C2) and (C5), there exists a d_γ-dimensional spline coefficient vector $\boldsymbol{\gamma}_0$ such that $\sup_{x\in[0,1]} |\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}_0 - \sigma(x)| = O(h_b^q)$ (De Boor 2001). Note that the dimension of $\boldsymbol{\gamma}_0$ goes to infinity, as $n \to \infty$.

We now establish the consistency of $\widehat{\boldsymbol{\beta}}_n$ and $\widehat{\boldsymbol{\gamma}}_n$, as well as the asymptotic distribution property of $\widehat{\boldsymbol{\beta}}_n$. The proofs of the following results are in Appendices A.5 and A.6.

Theorem 3.4.1. *Assume that Conditions* (C1)–(C6) *hold. Let* $\widehat{\boldsymbol{\theta}}_n$ *satisfy*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{S}^*_{\text{eff}}(Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\gamma}}_n, f^*_{X|\mathbf{Z}}) = \mathbf{0},$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{S}^*_{a,\text{eff},\gamma}(Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\gamma}}_n, f^*_{X|\mathbf{Z}}) = \mathbf{0}.$$

Then $\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 = o_p(1)$ *element-wise.*

The result in Theorem 3.4.1 is used to further establish the asymptotic properties of the estimator of the parameters of interest $\widehat{\boldsymbol{\beta}}_n$.

Theorem 3.4.2. *Assume that Conditions* (C1)–(C7) *hold and let*

$$\mathbf{Q} \equiv \mathrm{E}\left[\frac{\partial \mathbf{S}^*_{\mathrm{eff}}(Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}})}{\partial \boldsymbol{\beta}_0^{\mathrm{T}}} \Big|_{\mathbf{B}(X)^{\mathrm{T}}\boldsymbol{\gamma}=\sigma(X)} \right].$$

Here the subscript $\mathbf{B}(X)^{\mathrm{T}}\boldsymbol{\gamma} = \sigma(X)$ *means replacing* $\mathbf{B}(X)^{\mathrm{T}}\boldsymbol{\gamma}$ *with the true function* $\sigma(X)$*. Then*

$$\sqrt{n}\,(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = -\mathbf{Q}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{S}^*_{\mathrm{eff}}(Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \sigma, f^*_{X|\mathbf{Z}}) + o_p(1).$$

Consequently, $\sqrt{n}\,(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \rightsquigarrow \mathcal{N}(\mathbf{0}, \mathbf{V})$ *in distribution when* $n \to \infty$*, where*

$$\mathbf{V} = \mathbf{Q}^{-1} var\{\mathbf{S}^*_{\mathrm{eff}}(Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \sigma, f^*_{X|\mathbf{Z}})\}(\mathbf{Q}^{-1})^{\mathrm{T}}.$$

In addition, if the working model $f^*_{X|\mathbf{Z}}$ *is correctly specified, i.e., if* $f^*_{X|\mathbf{Z}}(x|\mathbf{z}) = f_{X|\mathbf{Z}}(x|\mathbf{z})$*, then the estimator* $\widehat{\boldsymbol{\beta}}$ *is semiparametric efficient, where* $\widehat{\boldsymbol{\beta}}$ *achieves the optimal estimation variance bound* $[\mathrm{E}\{\mathbf{S}_{\mathrm{eff}}(Y, W_1, W_2, \mathbf{Z}, \sigma, f_{X|\mathbf{Z}})^{\otimes 2}\}]^{-1}$*.*

Remark 3.4.2. Typically, estimating the nuisance parameters will alter, often inflate, the estimation variance for the parameter of interest. However, in our construction, if we knew the true function $\sigma(x)$ and used it in the estimating equation derived in Section 3.3.1, the variance of the $\boldsymbol{\beta}$ estimation would not change, as shown in Theorem 3.4.2. In other words, the estimation of $\sigma(x)$ does not inflate the variance of $\widehat{\boldsymbol{\beta}}_n$ asymptotically.

Remark 3.4.3. The semiparametric efficiency of our estimator $\widehat{\boldsymbol{\beta}}$ relies on that the working model $f^*_{X|\mathbf{Z}}$ is correctly specified. In practice, the conditional density function $f_{X|\mathbf{Z}}$ can be estimated consistently using B-splines approximation $\mathbf{B}(X|\mathbf{Z})^{\mathrm{T}}\boldsymbol{\zeta}$. The density function of W_1, W_2 given $\mathbf{Z}$ is

$$f_{W_1,W_2|\mathbf{Z}}(w_1, w_2|\mathbf{z}) = \int f_{W_1|X,\mathbf{Z}}(w_1|x, \mathbf{z}) f_{W_2|X,\mathbf{Z}}(w_2|x, \mathbf{z}) f_{X|\mathbf{Z}}(x|\mathbf{z}) dx.$$

Since W_1, W_2 and $\mathbf{Z}$ are observable, a consistent estimator of $\boldsymbol{\zeta}$ can be obtained by maximizing the log-likelihood of W_1 and W_2 given $\mathbf{Z}$ with respect to $\boldsymbol{\zeta}$. Specifically, the likelihood is given as

$$\prod_{i=1}^{n}\int f_U\left\{\frac{w_{i1} - x}{\widehat{\sigma}(x)}\right\} f_U\left\{\frac{w_{i2} - x}{\widehat{\sigma}(x)}\right\} \widehat{\sigma}(x)^{-2}\mathbf{B}(x|\mathbf{z}_i)^{\mathrm{T}}\boldsymbol{\zeta} dx,$$

where the estimated function $\widehat{\sigma}(x) = \mathbf{B}(x)^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}$. One can then iteratively update the $\boldsymbol{\gamma}$ and $\boldsymbol{\zeta}$ estimates. However, it is not clear yet whether consistent estimation of $f_{X|\mathbf{Z}}$ is sufficient to achieve efficiency in estimating $\boldsymbol{\beta}$ or whether a certain convergence rate is needed. Therefore, we did not pursue this approach further.

The asymptotic covariance matrix $\mathbf{V}$ can be estimated using the sample version of the matrix $\mathbf{Q}$ and the variance of $\mathbf{S}^*_{\mathrm{eff}}(Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \sigma, f^*_{X|\mathbf{Z}})$. The detailed formulation is given in Section 3.6.1.

3.5 Implementation

To simplify the implementation, instead of profiling $\boldsymbol{\gamma}$ as a function of $\boldsymbol{\beta}$, we estimate $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ together. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\gamma}^{\mathrm{T}})^{\mathrm{T}}$. To estimate $\boldsymbol{\theta}$, we need to compute the efficient score for $\boldsymbol{\theta}$, denoted by $\mathbf{S}^*_{a,\mathrm{eff},\boldsymbol{\theta}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\theta}, f^*_{X|\mathbf{Z}})$. This entails solving the integral equations

$$\mathrm{E}_a\{\mathbf{S}^*_{a,\boldsymbol{\theta}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\theta}, f^*_{X|\mathbf{Z}})|X, \mathbf{Z}\} = \mathrm{E}_a[\mathrm{E}^*_a\{\mathbf{a}(X, \mathbf{Z})|Y, W_1, W_2, \mathbf{Z}\}|X, \mathbf{Z}]$$

to obtain $\mathbf{a}(X, \mathbf{Z})$, where

$$\mathbf{S}^*_{a,\boldsymbol{\theta}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\theta}, f^*_{X|\mathbf{Z}}) \equiv \partial \ln\{f_{a,Y,W_1,W_2|\mathbf{Z}}(Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\theta}, f^*_{X|\mathbf{Z}})\}/\partial\boldsymbol{\theta}$$

is the score vector for $\boldsymbol{\theta}$. A simple approach to solving integral equations is discretization, which is mathematically equivalent to approximating $f_{X|\mathbf{Z}}(x|\mathbf{z})$ with a discrete distribution with mass at L points $0 < x_1 < \cdots < x_L < 1$ with the corresponding weights $d_1, \ldots, d_L$. We write

$$f^*_{X|\mathbf{Z}}(x|\mathbf{z}) = \sum_{j=1}^{L} d_j \mathbf{1}(x = x_j),$$

where $d_j \geqslant 0$ and $d_1 + \cdots + d_L = 1$.

The joint density of Y, W_1, W_2 conditional on $X = x_j, \mathbf{Z} = \mathbf{z}$ is

$$\begin{aligned} &f_{a,Y,W_1,W_2|X=x_j,\mathbf{Z}}(y, w_1, w_2, x_j, \mathbf{z}, \boldsymbol{\theta}) \\ &\qquad = f_{Y|X,\mathbf{Z}}(y|x_j, \mathbf{z}, \boldsymbol{\beta}) f_U\left\{\frac{w_1 - x_j}{\mathbf{B}(x_j)^{\mathrm{T}}\boldsymbol{\gamma}}\right\} f_U\left\{\frac{w_2 - x_j}{\mathbf{B}(x_j)^{\mathrm{T}}\boldsymbol{\gamma}}\right\} \frac{1}{\{\mathbf{B}(x_j)^{\mathrm{T}}\boldsymbol{\gamma}\}^2}. \end{aligned}$$

Thus we obtain

$$\mathbf{S}^*_{a,\boldsymbol{\beta}}(y, w_1, w_2, \mathbf{z}, \boldsymbol{\theta}, f^*_{X|\mathbf{Z}}) =$$
$$\frac{\sum_{j=1}^{L} \left[\{\partial f_{Y|X,\mathbf{Z}}(y|x_j, \mathbf{z}, \boldsymbol{\beta})/\partial\boldsymbol{\beta}\}/f_{Y|X,\mathbf{Z}}(y|x_j, \mathbf{z}, \boldsymbol{\beta})\right] f_{a,Y,W_1,W_2|X=x_j,\mathbf{Z}}(y, w_1, w_2, x_j, \mathbf{z}, \boldsymbol{\theta}) d_j}{\sum_{j=1}^{L} f_{a,Y,W_1,W_2|X=x_j,\mathbf{Z}}(y, w_1, w_2, x_j, \mathbf{z}, \boldsymbol{\theta}) d_j},$$

$$\mathbf{S}^*_{a,\boldsymbol{\gamma}}(y, w_1, w_2, \mathbf{z}, \boldsymbol{\theta}, f^*_{X|\mathbf{Z}})$$
$$= \frac{\sum_{j=1}^{L} \left[-V(u_{j1}, u_{j2})\mathbf{B}(x_j)/\{\mathbf{B}(x_j)^{\mathrm{T}}\boldsymbol{\gamma}\}\right] f_{a,Y,W_1,W_2|X=x_j,\mathbf{Z}}(y, w_1, w_2, x_j, \mathbf{z}, \boldsymbol{\theta}) d_j}{\sum_{j=1}^{L} f_{a,Y,W_1,W_2|X=x_j,\mathbf{Z}}(y, w_1, w_2, x_j, \mathbf{z}, \boldsymbol{\theta}) d_j},$$

and

$$\mathrm{E}^*_a\{\mathbf{a}(X, \mathbf{Z})|Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\theta}\}$$
$$= \frac{\sum_{j=1}^{L} \mathbf{a}(x_j, \mathbf{z}) f_{a,Y,W_1,W_2|X=x_j,\mathbf{z}}(y, w_1, w_2, x_j, \mathbf{z}, \boldsymbol{\theta}) d_j}{\sum_{j=1}^{L} f_{a,Y,W_1,W_2|X=x_j,\mathbf{z}}(y, w_1, w_2, x_j, \mathbf{z}, \boldsymbol{\theta}) d_j}, \quad (3.9)$$

where $u_{jk} = (w_k - x_j)/\{\mathbf{B}(x_j)^{\mathrm{T}}\boldsymbol{\gamma}\}$, for $j \in \{1, \ldots, L\}$ and $k \in \{1, 2\}$. Note that

$$\mathbf{S}^*_{a,\boldsymbol{\theta}}(y, w_1, w_2, \mathbf{z}, \boldsymbol{\theta}, f^*_{X|\mathbf{Z}}) = \{\mathbf{S}^*_{a,\boldsymbol{\beta}}(y, w_1, w_2, \mathbf{z}, \boldsymbol{\theta}, f^*_{X|\mathbf{Z}})^{\mathrm{T}}, \mathbf{S}^*_{a,\boldsymbol{\gamma}}(y, w_1, w_2, \mathbf{z}, \boldsymbol{\theta}, f^*_{X|\mathbf{Z}})^{\mathrm{T}}\}^{\mathrm{T}}.$$

At any $\mathbf{Z}$, let $\mathbf{A}^{(\mathbf{Z})}(\boldsymbol{\theta})$ be a $L \times L$ matrix with its (i, j) entry

$$A^{(\mathbf{Z})}_{ij}(\boldsymbol{\theta}) = \int \frac{f_{a,Y,W_1,W_2|X=x_j,\mathbf{Z}}(y, w_1, w_2, x_j, \mathbf{z}, \boldsymbol{\theta}) d_j}{\sum_{j=1}^{L} f_{a,Y,W_1,W_2|X=x_j,\mathbf{Z}}(y, w_1, w_2, x_j, \mathbf{z}, \boldsymbol{\theta}) d_j}$$
$$\times f_{a,Y,W_1,W_2|X=x_i,\mathbf{Z}}(y, w_1, w_2, x_i, \mathbf{z}, \boldsymbol{\theta}) dy dw_1 dw_2.$$

Define $\mathbf{H}^{(\mathbf{Z})}(\boldsymbol{\theta})$ as a $(p + d_{\boldsymbol{\gamma}}) \times L$ matrix whose ith column is given by

$$\mathbf{H}^{(\mathbf{Z})}_i(\boldsymbol{\theta}) = \int \mathbf{S}^*_{a,\boldsymbol{\theta}}(y, w_1, w_2, \mathbf{z}, \boldsymbol{\theta}, f^*_{X|\mathbf{Z}}) f_{Y,W_1,W_2|X=x_i,\mathbf{Z}}(y, w_1, w_2, x_i, \mathbf{z}, \boldsymbol{\theta}) dy dw_1 dw_2.$$

Let $\mathbf{a}^{(\mathbf{Z})} = (\mathbf{a}(x_1, \mathbf{z}), \ldots, \mathbf{a}(x_L, \mathbf{z}))$. Then we obtain $\mathbf{H}^{(\mathbf{Z})}(\boldsymbol{\theta}) = \mathbf{a}^{(\mathbf{Z})}\{\mathbf{A}^{(\mathbf{Z})}(\boldsymbol{\theta})\}^{\mathrm{T}}$ and $\mathbf{a}^{(\mathbf{Z})} = \mathbf{H}^{(\mathbf{Z})}(\boldsymbol{\theta})[\{\mathbf{A}^{(\mathbf{Z})}(\boldsymbol{\theta})\}^{\mathrm{T}}]^{-1}$, as long as $\mathbf{A}^{(\mathbf{Z})}(\boldsymbol{\theta})$ is nonsingular. To emphasize the dependence of the resulting $\mathbf{a}(X, \mathbf{Z})$ on $\boldsymbol{\theta}$, we write the solution as $\mathbf{a}(X, \mathbf{Z}, \boldsymbol{\theta})$. This allows us to form $\mathrm{E}^*_a\{\mathbf{a}(X, \mathbf{Z}, \boldsymbol{\theta})|Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\theta}\}$ using (3.9). The resulting estimating

equation for $\boldsymbol{\theta}$ is thus

$$\frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{S}^{*}_{a,\boldsymbol{\theta}}(y_i, w_{i1}, w_{12}, \mathbf{z}_i, \boldsymbol{\theta}, f^{*}_{X|\mathbf{Z}}) - \mathrm{E}^{*}_{a}\{\mathbf{a}(X, \mathbf{Z}_i, \boldsymbol{\theta})|Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\theta}\}\right] = \mathbf{0}. \quad (3.10)$$

By solving (3.10), we can obtain semiparametric estimator $\widehat{\boldsymbol{\theta}}_n = (\widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}, \widehat{\boldsymbol{\gamma}}_n^{\mathrm{T}})^{\mathrm{T}}$.

3.6 Empirical Studies

3.6.1 Simulation Studies

We conducted two simulation studies to investigate the finite-sample performance of the proposed method. We also compared the results of our method with those of the method that ignores the measurement error and the method that assumes the variance of the measurement error is constant. We set the sample size $n = 1000$ and generated 1000 samples in each simulation study. The error-prone covariate X_i is generated from the uniform distribution from -2.7 to 0.7, and the error-free covariate Z_i is a Bernoulli random variable independent of X_i with success probability 0.5. Moreover, U_{i1} and U_{i2} are generated from independent standard normal distributions. We then formed $W_{ik} = X_i + \sigma(X_i)U_{ik}$, for each $k \in \{1, 2\}$, and $i \in \{1, \ldots, n\}$.

We set discretization points of X to be $x_j = 3.4j/L - 2.7$ for $j \in \{1, \ldots, L\}$ and set the working model

$$f^{*}_{X|\mathbf{Z}}(x|\mathbf{z}) = \sum_{j=1}^{L} d_j \mathbf{1}(x = x_j).$$

Two different working models are considered. In the first model, $d_j = 1/L$, corresponding to a uniform working model. In the second model,

$$d_j = \frac{\phi\{(x_j + 1)/3.4\}}{\sum_{j=1}^{L}\phi\{(x_j + 1)/3.4\}},$$

where ϕ is the density of standard normal distribution, corresponding to a normal working model.

In the first simulation study, the response Y_i is generated from a logistic regression model

$$\mathrm{logit}\{\Pr(Y_i = 1|X_i, Z_i)\} = \beta_0 + \beta_1 X_i + \beta_2 Z_i,$$

with true values $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}} = (0.5, 0.2, -0.2)^{\mathrm{T}}$ and the true $\sigma(X_i) = (X_i^2 + 3)/13.5$. We used quadratic splines with six knots to approximate $\sigma(x)$ and set L to be 20. We computed the sample mean and standard deviation of the estimates $\widehat{\boldsymbol{\beta}}_n$ over 1000 data sets and estimated the asymptotic covariance matrix using sandwich formula $\widehat{\mathbf{V}} = \widehat{\mathbf{Q}}^{-1}\widehat{\boldsymbol{\Sigma}}(\widehat{\mathbf{Q}}^{-1})^{\mathrm{T}}$, where

$$\widehat{\mathbf{Q}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial \widehat{\boldsymbol{\beta}}_n^{\mathrm{T}}} \mathbf{S}^*_{\mathrm{eff}}(Y_i, W_{i1}, W_{i2}, Z_i, \widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\gamma}}_n, f^*_{X|\mathbf{Z}}),$$

and

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{S}^*_{\mathrm{eff}}(Y_i, W_{i1}, W_{i2}, Z_i, \widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\gamma}}_n, f^*_{X|\mathbf{Z}})\mathbf{S}^*_{\mathrm{eff}}(Y_i, W_{i1}, W_{i2}, Z_i, \widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\gamma}}_n, f^*_{X|\mathbf{Z}})^{\mathrm{T}}.$$

Further, 95% confidence intervals for β_0, β_1, β_2 are constructed in each simulated data set based on the asymptotic normal distribution of $\widehat{\boldsymbol{\beta}}_n$ to compute the empirical percentage covering the true values. We compared the results of our method with those of the naive logistic regression method, which ignores the measurement error and existing estimating equation method, which incorrectly assumes a constant error variance, i.e., $W = X + \sigma U$, $\sigma > 0$. The estimating equations are constructed based on efficient scores that are derived by treating σ as a nuisance parameter and adopting working model $f^*_{X|\mathbf{Z}}(x|\mathbf{z})$. The results of the first simulation study are summarized in Table 3.1, where median of the estimated standard deviation was reported as "est sd."

In the second method assuming homoscedastic measurement error, the estimated standard deviation of measurement error is $\widehat{\sigma} = 0.4066$ under both uniform and normal working models. Figure 3.1 shows the performance of the B-spline approximation of the nuisance function $\sigma(x)$ of our method under different working models of $f_{X|\mathbf{Z}}(x|\mathbf{z})$. The solid line represents the true function $\sigma(x)$, while the three dashed lines represent the 1/4, 1/2 and 3/4 sample quantiles of the estimated function $\mathbf{B}(x)^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_n$. Note that the median curve is almost overlapping with the true $\sigma(x)$.

We also experimented with heteroscedastic Laplace measurement errors. The results are given in Table 3.2 and Figure 3.2. The performance of our method is still satisfactory.

In the second simulation study, we consider a relatively more complex model, where $\sigma(X_i) = 0.4\exp\{-0.15(X_i + 1)^2\}$. We generate the response Y_i from a quadratic logistic regression model, viz.

$$\mathrm{logit}\{\Pr(Y_i = 1|X_i, Z_i)\} = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 Z_i,$$

Table 3.1: Results of the first simulation study

	β_0	β_1	β_2			
Truth	0.5	0.2	-0.2			
	No measurement error					
Mean	0.4858	0.1869	-0.1970			
Emp sd	0.1193	0.0629	0.1341			
Est sd	0.1115	0.0630	0.1281			
Emp cov	93.2%	94.0%	94.7%			
	β_0	β_1	β_2	β_0	β_1	β_2
Truth	0.5	0.2	-0.2	0.5	0.2	-0.2
	Homoscedastic measurement error					
	Uniform working model			Normal working model		
Mean	0.5049	0.2058	-0.1970	0.5049	0.2058	-0.1970
Emp sd	0.1220	0.0678	0.1355	0.1220	0.0678	0.1355
Est sd	0.6240	0.1844	0.8492	0.6484	0.1914	0.8875
Emp cov	96.2%	90.3%	96.6%	96.4%	90.6%	96.6%
	Heteroscedastic measurement error					
	Uniform working model			Normal working model		
Mean	0.5024	0.2038	-0.1948	0.5032	0.2043	-0.1954
Emp sd	0.1183	0.0665	0.1317	0.1180	0.0668	0.1318
Est sd	0.1149	0.0680	0.1283	0.1149	0.0680	0.1283
Emp cov	95.3%	95.3%	94.8%	94.7%	95.0%	94.9%

In Table 3.1, "emp sd" denotes the empirical standard deviation of the estimates, "est sd" denotes the estimated asymptotic standard deviation and "emp cov" denotes the empirical coverage of the estimated 95% confidence intervals.

with true values for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^{\mathrm{T}} = (-0.8, 0.5, 0.2, 2.0)^{\mathrm{T}}$. We used similar working distribution models $f^*_{X|Z}(x|\mathbf{z})$ with $L = 20$. A quadratic spline with seven knots was used to approximate the nuisance function $\sigma(x)$. The results of the three different methods in the second simulation study are summarized in Table 3.3.

In the second method, the estimated standard deviation of the measurement error is $\widehat{\sigma} = 0.3514$ under both uniform and normal working models, while Figure 3.3 shows the performance of the B-splines approximation of the standard deviation function $\sigma(x)$ of the measurement error under different working models in our method.

In the simulation studies, our method performs well in both the simple and the complex model settings. The estimates have very small biases, the medians of estimated

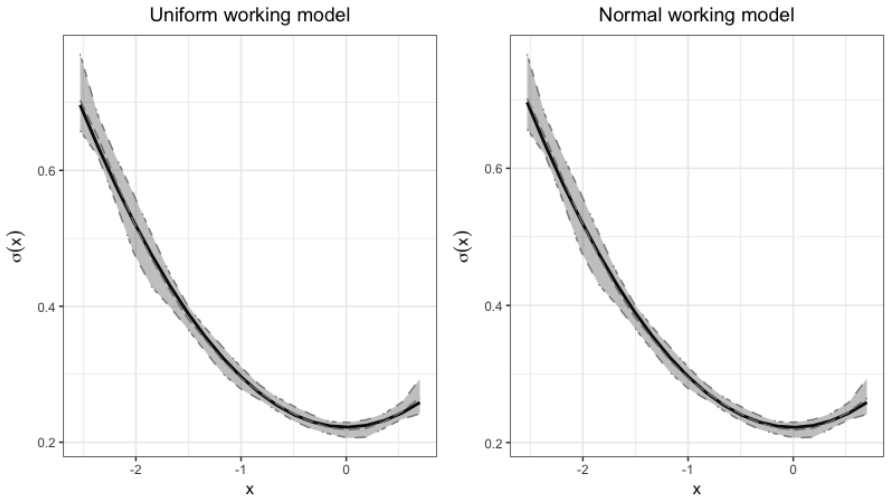

Figure 3.1: Performance of the B-splines approximation of $\sigma(x)$ under two working models of $f_{X|\mathbf{Z}}$ in the first simulation study

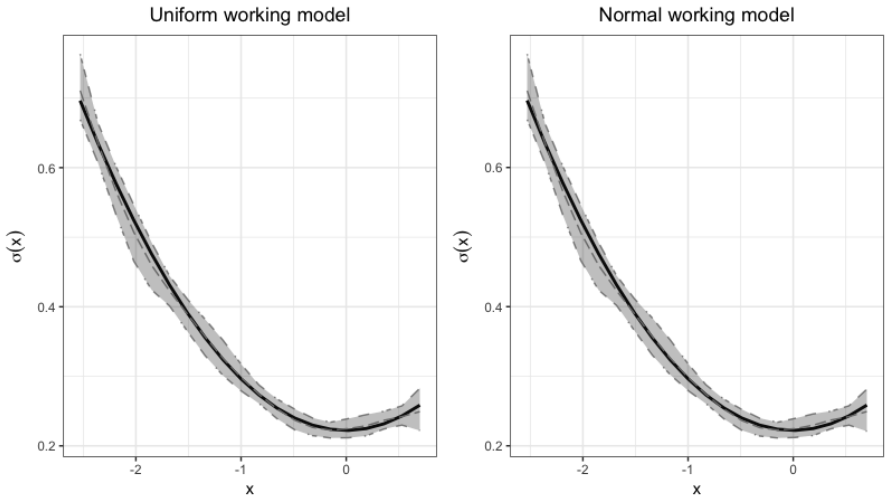

Figure 3.2: Performance of the B-splines approximation of $\sigma(x)$ under two working models of $f_{X|\mathbf{Z}}$ in the first simulation study with heteroscedastic Laplace measurement errors

standard deviations closely approximate the empirical standard deviations, and the empirical coverages of the estimated 95% confidence interval are close to the nominal level. Further, the results of using different working models are similar in both simulations, which suggests the insensitivity of our method to the misspecification of $f_{X|\mathbf{Z}}(x|\mathbf{z})$. We also see that ignoring the measurement error or incorrectly assuming homoscedastic measurement error can cause bias issues in estimation, especially the estimation of the coefficients associated with the unobservable covariate, and affect the quality of influence.

Table 3.2: Results of the first simulation study with heteroscedastic Laplace measurement errors

	β_0	β_1	β_2			
Truth	0.5	0.2	−0.2			
	No measurement error					
Mean	0.4906	0.1852	−0.2046			
Emp sd	0.1130	0.0619	0.1252			
Est sd	0.1117	0.0631	0.1281			
Emp cov	95.2%	94.6%	95.4%			
	β_0	β_1	β_2	β_0	β_1	β_2
Truth	0.5	0.2	−0.2	0.5	0.2	−0.2
	Homoscedastic measurement errors					
	Uniform working model			Normal working model		
Mean	0.5054	0.2088	−0.2056	0.5053	0.2088	−0.2054
Emp sd	0.1168	0.0703	0.1298	0.1172	0.0703	0.1302
Est sd	0.1521	0.0553	0.1675	0.1518	0.0557	0.1705
Emp cov	93.8%	80.6%	93.9%	93.3%	80.1%	93.9%
	Heteroscedastic measurement error					
	Uniform working model			Normal working model		
Mean	0.5051	0.2037	−0.2046	0.5042	0.2035	−0.2041
Emp sd	0.1168	0.0677	0.1303	0.1162	0.0675	0.1298
Est sd	0.1147	0.0675	0.1282	0.1147	0.0675	0.1282
Emp cov	94.7%	94.9%	94.2%	95.0%	95.0%	94.3%

In Table 3.2, "emp sd" denotes the empirical standard deviation of the estimates, "est sd" denotes the estimated asymptotic standard deviation and "emp cov" denotes the empirical coverage of the estimated 95% confidence intervals.

3.6.2 Data Analysis

In this section, we illustrate our method by analyzing the data from the Childhood Asthma Management Program (CAMP) CAMP is a longitudinal study designed to explore the long-term impact of several daily treatments for mild to moderate asthma in children.

We formed the outcome variable Y_i based on the average asthma symptoms (`amsym`) recorded in daily record of eight months. `amsym` is a binary variable indicating the severity of the asthma symptoms for a child (indexed by i). If the average `amsym` is greater than 0.5, then the outcome variable Y_i equals 1, implying moderate asthma

Table 3.3: Results of the second simulation study

	β_0	β_1	β_2	β_3	β_0	β_1	β_2	β_3
Truth	−0.8	0.5	0.2	2.0	−0.8	0.5	0.2	2.0
	No measurement error							
Mean	−0.8201		0.4172		0.1611		2.0107	
Emp sd	0.1236		0.1648		0.0749		0.1436	
Est sd	0.1206		0.1675		0.0756		0.1441	
Emp cov	94.3%		92.8%		91.8%		94.1%	
	Homoscedastic measurement error							
	Uniform working model				Normal working model			
Mean	−0.8042	0.5294	0.2143	2.0151	−0.8042	0.5294	0.2144	2.0151
Emp sd	0.1263	0.2133	0.0998	0.1447	0.1263	0.2133	0.0998	0.1447
Est sd	0.9700	0.7301	0.1419	1.1875	1.0717	0.7767	0.1555	1.2770
Emp cov	95.6%	92.1%	83.2%	99.8%	95.7%	91.9%	83.6%	99.8%
	Heteroscedastic measurement error							
	Uniform working model				Normal working model			
Mean	−0.8047	0.5097	0.2044	2.0132	−0.8045	0.5109	0.2050	2.0136
Emp sd	0.1220	0.1990	0.0915	0.1436	0.1223	0.2004	0.0930	0.1420
Est sd	0.1249	0.2105	0.0970	0.1453	0.1249	0.2112	0.0973	0.1452
Emp cov	96.4%	96.3%	96.5%	94.8%	96.1%	96.4%	96.6%	94.9%

In Table 3.3, "emp sd" denotes the empirical standard deviation of the estimates, "est sd" denotes the estimated asymptotic standard deviation and "emp cov" denotes the empirical coverage of the estimated 95% confidence intervals.

symptoms. Otherwise, the outcome variable Y_i equals 0, implying mild asthma. The FEV1/FVC ratio (`preff`) is an important index used in diagnosis of asthma, which represents the proportion of a person's vital capability to expire in the first second of forced expiration to the full vital capacity. Four measurements of `preff` were recorded during the 8-month study for each child.

We let X_i be the unobserved `preff` and W_i be the average of four measurements with heteroscedastic measurement error. Other error-free variables $\mathbf{Z}_i$ are gender, age at baseline, and treatment group. Three treatment groups were included in the study, and we coded them using two dummy variables `trt1` and `trt2`, where `trt1` = 1 if the treatment is budesonide, `trt2` = 1 if the treatment is nedocromil, and `trt1` = `trt2` = 0

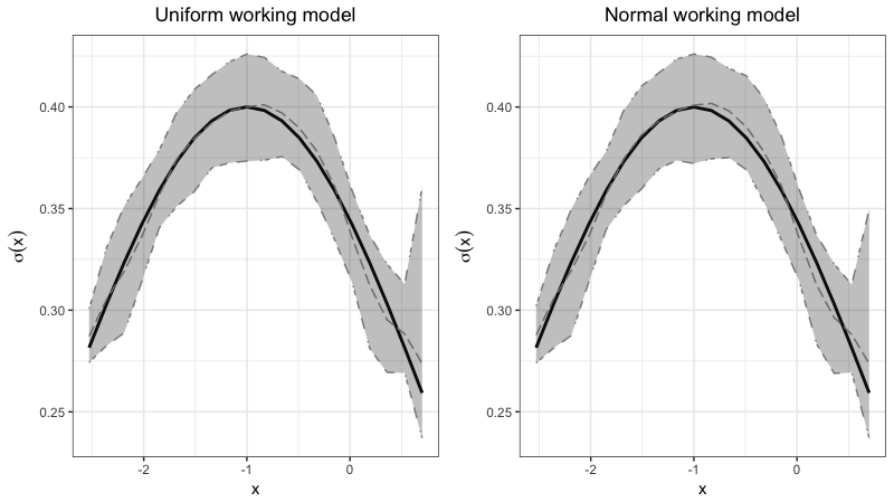

Figure 3.3: Performance of the B-splines approximation of $\sigma(x)$ under two working models of $f_{X|\mathbf{Z}}$ in the second simulation study

if the treatment is placebo.

The data set consisted of 737 children for whom $(Y_i, W_i, \mathbf{Z}_i)$ were measured. We considered the linear logistic regression model with heteroscedastic measurement error on X_i, viz.

$$\text{logit}\{\Pr(Y_i = 1|X_i, Y_i)\} = \beta_0 + \beta_x X_i + \boldsymbol{\beta}_z \mathbf{Z}_i \quad \text{and} \quad W_i = X_i + \sigma(X_i)U_i,$$

where $\boldsymbol{\beta}_z = (\beta_{z1}, \beta_{z2}, \beta_{z3}, \beta_{z4})^{\mathrm{T}}$, $\mathbf{Z}_i$ is a vector of gender, age, `trt1`, `trt2` for the ith child and $U_i \sim \mathcal{N}(0, 1)$. We used a uniform working model for the distribution of $\mathbf{X}$ and adopted $L = 20$ discretization points in the implementation. We used quadratic splines with 6 knots to approximate $\sigma(X)$.

We further compared the results from our analysis with those of the "naive" logistic regression method, which ignores the existence of measurement error, and the estimating equations approach, which treats the measurement error variance as constant. The results from all three methods are summarized in Table 3.4. In the second method which assumes constant measurement error variance, the estimated standard deviation is $\widehat{\sigma} = 0.0753$. The estimated heteroscedastic error standard deviation function $\widehat{\sigma}(x)$ in the third method is given in Figure 3.4.

In the naive logistic regression model that ignores the measurement error, besides the error-prone variable `preff`, we also detect variables gender and `trt1` to significantly influence the severity of asthma at the significance level $\alpha = 0.05$. In the second and third methods considering measurement errors, only the error-prone variable `preff` is

Table 3.4: Analysis of the CAMP data under uniform working model.

	β_0 intercept	β_x preff	β_{z1} gender	β_{z2} age	β_{z3} trt1	β_{z4} trt2
	No measurement error					
Est	0.3282	−2.8035	0.3774	0.0537	−0.7794	−0.3251
Est sd	0.5222	0.5297	0.1737	0.0411	0.2099	0.1978
p-value	0.5297	1.2e − 07	0.0298	0.1920	0.0002	0.1001
	Homosecdastic measurement error					
Est	0.4424	−2.9674	0.3859	0.0501	−0.7725	−0.3248
Est sd	0.4247	0.3518	3.8812	0.3296	0.5025	5.2613
p-value	0.2975	0.0000	0.9208	0.8791	0.1242	0.9508
	Heteroscedastic measurement error					
Est	0.4224	−2.9397	0.3846	0.0507	−0.7741	−0.3248
Est sd	0.1776	0.2655	3.2765	0.2263	0.4010	3.8781
p-value	0.0174	0.0000	0.9066	0.8226	0.0861	0.9332

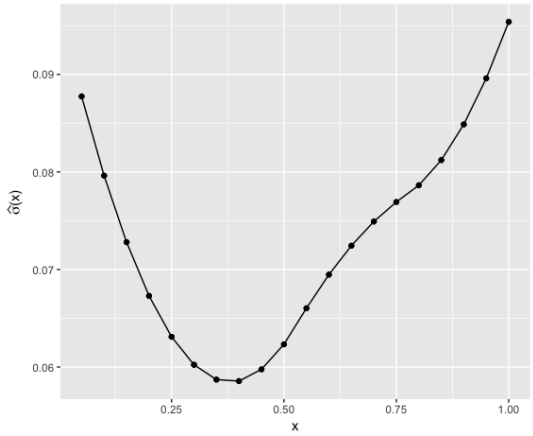

Figure 3.4: B-spline approximation of $\sigma(x)$ under uniform working model

significant, with the same sign and slightly greater absolute value as the estimate. The variable `trt1` is almost significant with p-value < 0.1 in the heteroscedastic measurement error model, but gender is not significant at all. This difference indicates that ignoring the measurement error could lead to misleading results. Further, the B-spline approximation of $\sigma(x)$ fluctuates with respect to x and is not always close to $\widehat{\sigma}$. Hence the assumption about constant measurement error variance may be too restrictive for this data set. From Figure 3.4, we can see that when x is close to 0 or 1, the measurement error variance tends to be larger.

We also tried using a normal working model for the distribution of X in the second and third methods. The estimates $\widehat{\boldsymbol{\beta}}$ are almost the same as those under the uniform working model. Specifically, with the normal working model, in the second method $\widehat{\boldsymbol{\beta}} = (0.4435, -2.9689, 0.3860, 0.0501, -0.7725, -0.3248)$, while in the third method $\widehat{\boldsymbol{\beta}} = (0.4244, -2.9416, 0.3847, 0.0507, 0.7741, -0.3248)$. However, due to the near singularity of the matrix $\widehat{\mathbf{Q}}$ in the sandwich formula, the inference results are inaccurate, so we did not pursue the normal working model further.

3.7 Discussion

In this chapter, we proposed a new method in the frame of general measurement error model with heteroscedastic error. The method can be applied to any parametric model with unspecified heteroscedastic measurement error variance structure. We have assumed the distribution of the error U to be known, but not restricted to the normal distribution. In practice, the density $f_U(\cdot)$ needs to be determined using external information such as validation data. If $f_U(\cdot)$ cannot be determined, one may be able to approximate it with $f_U(u) \approx \exp\{\mathbf{B}(u)^{\mathrm{T}}\boldsymbol{\alpha}\}/\int \exp\{\mathbf{B}(u)^{\mathrm{T}}\boldsymbol{\alpha}\}du$ and estimate $\boldsymbol{\alpha}$ together with other parameters, provided that the problem is still identifiable. The identifiability issue in this case is difficult and warrants further research. Overall, identifiability in measurement error model is a hard problem and is often established case by case. We have proved the identifiability of linear model with heteroscedastic normal and Laplace measurement errors rigorously. We have also shown the identifiability of $\boldsymbol{\beta}$ for logistic model with heteroscedastic normal measurement errors. In addition, $f_{X|\mathbf{Z}}(\cdot)$ is estimable (Staudenmayer et al. 2008). Then $\sigma(\cdot)$ is identifiable based on $W_1 - W_2 = \sigma(X)(U_1 - U_2)$, where the distribution of $W_1 - W_2$ is estimable and the distribution of $U_1 - U_2$ is known. Therefore, the whole problem for logistic model with heteroscedastic normal measurement errors is identifiable.

We have assumed that the unobservable covariate X is a scalar for simplicity of presentation. If there are more than one unobservable covariates $\mathbf{X} = (X_1, \ldots, X_m)$ in the model and the measurement W_j given X_j is independent of other unobservable covariates, conceptually we can use $\mathbf{B}(X_j)^{\mathrm{T}}\boldsymbol{\gamma}_j$ to approximate the jth unknown function $\sigma_j(X_j)$ then append the estimating equations with these additional estimating equations obtained from the corresponding score functions for $\boldsymbol{\gamma}_j$, $j = 1, \ldots, m$. The computation may be more challenging. Additionally, for simplicity, we used $\mathbf{B}(X)^{\mathrm{T}}\boldsymbol{\gamma}$ to approximate $\sigma(X)$ in our implementation. To ensure positivity of $\sigma(X)$, we could instead use $\exp\{\mathbf{B}(X)^{\mathrm{T}}\boldsymbol{\gamma}\}$ to approximate $\sigma(X)$. The theoretical properties of $\widehat{\boldsymbol{\beta}}_n$ would not change.

We also would like to point out that in general, multiple roots is a potential issue for estimating equations approaches. Choosing the correct estimator from multiple roots of the estimating equations may not be easy and straightforward. Heyde & Morton (1998) invented a criteria to discriminate the consistent estimator from multiple roots of estimating equation. More discussions can be found in Hanfelt & Liang (1995) and Section 13.3 of Heyde (1997, 2008). In practice, using empirical knowledge or using estimates from more primitive but simpler methods to form starting values of our method can be a sensible option.

Chapter 4
Inference in High-Dimensional Linear Measurement Error Models

4.1 Introduction

High dimensional data becomes more and more common in diverse fields such as computational biology, economics and climate science. Many statistical procedures have been developed for analysis of high dimensional data. However, most of them often assume that all covariates are measured accurately. In reality, measurement errors are ubiquitous in many high-dimensional problems, for example, measurements of gene expression with cDNA or oligonucleotide arrays (Rocke & Durbin 2001) and sensor network data (Slijepcevic et al. 2002). This work was motivated by an empirical analysis of a real data set in Section 4.4.2, where both finite-dimensional phenotypic covariates and high-dimensional SNPs are available and one of the phenotypic covariates is of clinical interest but measured with error.

The classical measurement error models, where the number of covariates p is fixed or is smaller than the sample size n, have been studied systematically, see Fuller (1987, 2009), Carroll et al. (2006), Yi (2016) and Ma & Li (2010*b*). Penalized methods have been developed for high-dimensional linear measurement error models with $p > n$. Consider the model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta}_0 + \boldsymbol{\epsilon}, \quad \text{and} \quad \mathbf{W} = \mathbf{X} + \mathbf{U}, \tag{4.1}$$

where random vectors $\mathbf{Y}, \boldsymbol{\epsilon} \in \mathbb{R}^n$, the $n \times p$ matrix $\mathbf{X}$ is unobservable, $\mathbf{W}$ is its observed surrogate, and the matrix $\mathbf{U}$ is random noise, i.e. measurement error. This is a difficult problem. In fact, even in the absence of measurement error, Zhao & Yu (2006) and

Meinshausen et al. (2006) showed that the Lasso or Dantzig selector often fails in identifying significant covariates in high-dimensional models. With measurement error, Rosenbaum et al. (2010) showed that the true selection is likely to be outside of the feasible set of the Dantzig selector. S**o**rensen et al. (2015) analyzed the impact of measurement error on the standard Lasso and showed that treating $\mathbf{W}$ as the true $\mathbf{X}$ leads to erroneous results.

To correct the bias caused by the measurement error $\mathbf{U}$, a corrected objective function is

$$\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \frac{1}{n}\mathbf{W}^{\mathrm{T}}\mathbf{y} + P_{\lambda}(\boldsymbol{\theta}),$$

where $P_{\lambda}(\boldsymbol{\theta})$ is a penalty with tuning parameter λ, $\widehat{\boldsymbol{\Sigma}} = \mathbf{W}^{\mathrm{T}}\mathbf{W}/n - \mathbf{D}$, and $\mathbf{D}$ is the $p \times p$ covariance matrix of $\mathbf{U}_i$. Since $\widehat{\boldsymbol{\Sigma}}$ can have negative eigenvalues when p is larger than n, the loss function $\boldsymbol{\theta}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta}/2 - \mathbf{X}^{\mathrm{T}}\mathbf{y}/n$ is no longer convex. To overcome the difficulties caused by the non-convexity, Loh & Wainwright (2012) proposed a projected gradient descent algorithm that finds a possible local optimum with strong performance guarantees. Chen & Caramanis (2013) developed a simple variant of orthogonal matching pursuit algorithm that performs at the minimax optimal rate. Later, Belloni, Rosenbaum & Tsybakov (2017) proposed the compensated matrix uncertainty (MU) selector, which can be written as a second-order cone programming minimization problem and the estimator attains the minimax efficiency bound. Loh et al. (2017) developed a primal-dual witness proof framework to establish the estimator error bounds in different norms in general sparse regression problems with non-convex loss function and penalty. This work does not require the typical incoherence condition, but need to impose the constraint $\|\boldsymbol{\theta}_0\|_1 < R$. Datta et al. (2017) proposed CoCoLasso estimator which forces the non-convex problem to be convex by applying a nearest positive semi-definite matrix projection operator to $\widehat{\boldsymbol{\Sigma}}$, which can be solved by the ADMM algorithm, and analyzed its error bounds with deterministic design matrix $\mathbf{X}$. Under a slightly stronger sparsity conditions, the asymptotic sign-consistency properties were established.

The aforementioned works focus on the theory and numerical algorithms of regularization methods rather than statistical inference. It is important to quantify the uncertainty of an estimator in high dimensional linear measurement error models. Recently, significant progress has been made regarding hypothesis testing on low dimensional sub-parameters in high dimensional sparse models. From a semiparametric perspective, the challenges in these problems lie in how to handle the effect of high-dimensional

nuisance parameters and correct the bias of the estimators for the low dimensional parameters of interest caused by the penalty. Zhang & Zhang (2014) proposed a low dimensional projection (LDP) approach to construct bias-corrected linear Lasso estimator and corresponding confidence intervals without assuming the uniform signal strength condition (Wainwright 2009*a*). Van de Geer et al. (2014) exploited the idea of inverting the Karush-Kuhn-Tucker characterization to desparsify Lasso, which essentially leads to the same results as in Zhang & Zhang (2014) for a linear model. Javanmard & Montanari (2014) proposed to debias the Lasso estimator by adding a term proportional to the subgradient of the ℓ_1 norm at the Lasso solution, and the confidence intervals constructed based on the debiased estimator have nearly optimal size. All these works assume either linear or generalized linear models. Ning et al. (2017) provided a general framework for high-dimensional inference by proposing a decorrelated score function. By applying a decorrelation operation on the high-dimensional score functions, the derived decorrelated score function is uncorrelated with the nuisance score function. In this case, the efficiency of the estimators for the parameters of interest will not be impaired provided that the estimators for the nuisance parameters are consistent at sufficient rate.

Inference for high dimensional measurement error models is believed to be a difficult topic due to the bias and lack of power introduced by measurement error as well as high dimensional nuisance parameters. Recently, Belloni, Chernozhukov & Kaul (2017) constructed simultaneous confidence regions for the parameters of interest in high-dimensional linear models with error-in-variables using multiplier bootstrap. Wang et al. (2019) employed a de-biasing approach and constructed component-wise confidence intervals in a sparse high-dimensional linear regression model when some covariates of the design matrix are missing completely at random. In this paper, we consider the setting where only a fixed number of covariates are measured with error and our goal is to develop statistical inference procedures for the coefficients of these covariates. In practice, it is common that not all covariates are corrupted. For example, in the real data example analyzed in Section 4.2, covariates such as gender and age are measured precisely. Moreover, it is in general very difficult to find a good estimate for the $p \times p$ covariance matrix $\mathbf{D}$ of measurement error without any strong and restrictive assumptions.

We extend the inference results of low dimensional linear measurement error models to high dimensional settings, which is important yet challenging, and requires vastly different treatments. In the spirit of semiparametrics, we employ decorrelation operation to control the impact of high-dimensional nuisance parameters, and construct a corrected decorrelated score function for the parameters of interest. The performance of the

corrected decorrelated score test relies on the convergence rate of the initial estimator. The asymptotic normality of the corrected decorrelated score test statistic holds provided that the initial estimator is statistically consistent at certain rate. Here, we take the CoCoLasso estimator (Datta et al. 2017) as an example. Indeed, any estimator with sufficient convergence rate can be served as the initial estimator in forming the decorrelated score function. Different from the settings in Datta et al. (2017), we assume that the design is random and sub-Gaussian, and only a fixed number of covariates, without loss of generality, one covariate, is measured with error. We rederive the theoretical properties of the CoCoLasso estimator in our new settings, which is one of the contributions of this work. Our corrected decorrelated score test statistics retain power under the local alternatives around 0, because we essentially do not impose any penalty on the parameter of interest in the construction. We further construct confidence intervals by proving the limiting distribution of the one-step estimator, which is semiparametrically efficient. Note that although we write our development for one variable with measurement error, the proposed method is directly applicable to a finite number of covariates with measurement error naturally.

Our work extends the key idea of semiparametrics to inference in high dimensional linear measurement error models. We handle the sparsity assumptions differently from Belloni, Rosenbaum & Tsybakov (2017) and Loh & Wainwright (2012), and extend the results in Datta et al. (2017) to random sub-Gaussian designs. Although a general framework of inference was provided in Ning et al. (2017), the existence of measurement errors imposes many special challenges in methodology and theoretical proofs, which requires innovative technical treatments, as illustrated in the main text of the paper. Compared to Belloni, Chernozhukov & Kaul (2017), we avoid solving estimating equations completely. Our one-step estimator has the same limiting distribution as that of the root of estimating equations but is much easier to compute.

Notations and Preliminaries: Before we pursue further, let us introduce some notation and some preliminaries used in this Chapter. For a vector $\mathbf{v} = (v_1, \ldots, v_p)^{\mathrm{T}} \in \mathbb{R}^p$, we define $\|\mathbf{v}\|_0 = |\mathrm{supp}(\mathbf{v})|$, where $\mathrm{supp}(\mathbf{v}) = \{\mathrm{j} : \mathrm{v_j} \neq 0\}$ and $|A|$ is the cardinality of a set A. Denote $\|\mathbf{v}\|_\infty = \max_{1 \leqslant j \leqslant p} |v_j|$ and $\mathbf{v}^{\otimes 2} = \mathbf{v}\mathbf{v}^{\mathrm{T}}$. For $S \subseteq \{1, \ldots, p\}$, let $\mathbf{v}_S = \{v_j : j \in S\}$ and S^C be the complement of S . For a matrix $\mathbf{M} = [\mathbf{M}_{jk}]$, let $\|\mathbf{M}\|_{\max} = \max_{j,k} |M_{jk}|$, $\|M\|_\infty = \max_j \sum_k |M_{jk}|$ and $\mathbf{M}^{\otimes 2} = \mathbf{M}\mathbf{M}^{\mathrm{T}}$. If $\mathbf{M}$ is symmetric, then $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ are the minimal and maximal eigenvalues of $\mathbf{M}$. For two positive sequences a_n and b_n, we use $a_n \lesssim b_n$ to denote $a_n \leqslant C b_n$ for some constant $C > 0$, and use $a_n \asymp b_n$ to denote $C \leqslant a_n / b_n \leqslant C'$ for some constants $C, C' > 0$. Denote

$\Phi(\cdot)$ to be the cumulative distribution function of the standard normal distribution. For simplicity, we use $E(\cdot)$ and $\Pr(\cdot)$ to denote the expectation and probability calculated under the true model, respectively.

The sub-exponential norm of a random variable X is defined as

$$\|X\|_{\psi_1} = \sup_{q \geqslant 1} q^{-1} \{E(|X|^q)\}^{1/q}.$$

Note that $\|X\|_{\psi_1} < C_1$ for some constant C_1, if X is sub-exponential. The sub-Gaussian norm of X is defined as $\|X\|_{\psi_2} = \sup_{q \geqslant 1} q^{-1/2} \{E(|X|^q)\}^{1/q}$. Note that $\|X\|_{\psi_2} < C_2$ for some constant C_2, if X is sub-Gaussian. More properties regarding sub-exponential and sub-Gaussian random variables are given in Appendix B.6.1.

4.2 Model Setup and Proposed Method

4.2.1 Model Specification

Suppose that $\{Y_i, W_i, \mathbf{Z}_i\}$, $i = 1, \ldots, n$, is an independent and identically distributed sample from a linear model with one of the covariates measured with additive error

$$Y_i = \beta_0 X_i + \boldsymbol{\gamma}_0^{\mathrm{T}} \mathbf{Z}_i + \epsilon_i \quad \text{and} \quad W_i = X_i + U_i. \tag{4.2}$$

Covariate $X_i \in \mathbb{R}$ is unobservable, and W_i is its error-prone surrogate. Covariate vector $\mathbf{Z}_i \in \mathbb{R}^{p-1}$ is measured precisely. Assume that $(X_i, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}}$ is sub-Gaussian element-wise with mean $\mathbf{0}$ and unit diagonal covariance matrix. To exclude the intercept term in the model, we let the response Y_i have mean 0 as well. The regression error ϵ_i is sub-Gaussian with mean 0, variance σ_ϵ^2, and sub-Gaussian norm K_ϵ. The measurement error U_i is also sub-Gaussian with mean 0, variance σ_U^2, and sub-Gaussian norm K_U. It is independent of ϵ_i, X_i and $\mathbf{Z}_i$. As in the literature, we assume that σ_U^2 and $E(U_i^4)$ are known.

Let $\mathbf{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\mathbf{X} = (X_1, \ldots, X_n)^{\mathrm{T}}$, $\mathbf{W} = (W_1, \ldots, W_n)^{\mathrm{T}}$ and $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_n)^{\mathrm{T}}$ denote the corresponding vector or matrix version of n samples. In practice, we only need to center all variables, and standardize the columns of the data matrix such that $\sum_{i=1}^n Z_{ij}^2/n = 1$ and $\sum_{i=1}^n W_i^2/n = 1 + \sigma_U^2$ for $j = 1, \ldots, p-1$ and $i = 1, \ldots, n$.

For the purpose of theoretical proofs, we have the following standard assumptions.

Assumption 4.2.1. *Assume that*

(i) $2\kappa \leqslant \lambda_{\min}[E\{(X_i, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}\otimes 2}\}] \leqslant \lambda_{\max}[E\{(X_i, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}\otimes 2}\}] \leqslant 2/\kappa$ *for some constant*

$\kappa > 0$;

(ii) $\|Z_{ij}\|_{\psi_2}$ *and* $\|X_i\|_{\psi_2}$ *are uniformly bounded by some constant* K *for* $j = 1, \ldots, p-1$;

(iii) *The true parameter* $\boldsymbol{\theta}_0 = (\beta_0, \boldsymbol{\gamma}_0^{\mathrm{T}})^{\mathrm{T}}$ *is sparse with support* S, *and* $|S| = s_0$; *Let* $\|\boldsymbol{\theta}_0\|_\infty \leqslant K_0$, *where* K_0 *is a positiveconstant;*

(iv) $E(X_i\mathbf{Z}_i^{\mathrm{T}})\{E(\mathbf{Z}_i^{\otimes 2})\}^{-1}$ *is sparse with support* S' *and* $|S'| = s'$. *Moreover,* $\|E(X_i\mathbf{Z}_i^{\mathrm{T}})\{E(\mathbf{Z}_i^{\otimes 2})\}^{-1}\|_1 \leqslant K_{\boldsymbol{\omega}}$ *for some constant* $K_{\boldsymbol{\omega}} > 0$.

In Assumption 4.2.1, (i) and (ii) are common assumptions for high dimensional random designs. Assumption (iii) is about the sparsity of the true model (4.2). Instead of assuming $\|\boldsymbol{\theta}_0\|_1$ is bounded, we only assume the l_∞ norm of $\boldsymbol{\theta}_0$ is bounded. Assumption (iv) is crucial in the inference framework of Ning et al. (2017). When conducting decorrelation operation, their key assumption is that the projection of the score function for β to the linear space spanned by the nuisance score functions for $\boldsymbol{\gamma}$, denoted as $\Lambda_{\boldsymbol{\gamma}}$, is identical to the projection of the score function for β to a low dimensional subspace of $\Lambda_{\boldsymbol{\gamma}}$. More details about the motivation of sparse projection and the formation of $E(X_i\mathbf{Z}_i^{\mathrm{T}})\{E(\mathbf{Z}_i^{\otimes 2})\}^{-1}$ will be discussed in Section 4.2.2.

Our goal is to test the hypothesis $H_0 : \beta_0 = \beta^*$ and construct valid confidence intervals for β_0 when the dimension of $\boldsymbol{\theta}_0 = (\beta_0, \boldsymbol{\gamma}_0^{\mathrm{T}})^{\mathrm{T}}$ is much larger than the sample size n, that is, $p \gg n$. Note that when $\beta^* = 0$, under the null hypothesis, the model degenerates to a linear model without measurement error, hence testing procedures for high dimensional sparse linear models can be applied. In this paper, we consider a general hypothesis test setting where $\beta^* \in \mathbb{R}$.

4.2.2 Decorrelated Score Function

If covariate X is observed with no measurement error, it is known that the loss function based on least squares is $\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\theta}/2 - \boldsymbol{\rho}^{\mathrm{T}}\boldsymbol{\theta}$, where $\boldsymbol{\Sigma} = (\mathbf{X}, \mathbf{Z})^{\mathrm{T}}(\mathbf{X}, \mathbf{Z})/n$ and $\boldsymbol{\rho} = (\mathbf{X}, \mathbf{Z})^{\mathrm{T}}\mathbf{Y}/n$. For our corrupted data $(\mathbf{Y}, \mathbf{W}, \mathbf{Z})$, as emphasized above, instead of treating $\mathbf{W}$ as $\mathbf{X}$ in the loss function directly, we define the corrected loss function as

$$l(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \widehat{\boldsymbol{\rho}}^{\mathrm{T}}\boldsymbol{\theta}, \tag{4.3}$$

where

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}(\mathbf{W}, \mathbf{Z})^{\mathrm{T}}(\mathbf{W}, \mathbf{Z}) - \begin{pmatrix} \sigma_U^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \text{and} \quad \widehat{\boldsymbol{\rho}} = \frac{1}{n}(\mathbf{W}, \mathbf{Z})^{\mathrm{T}}\mathbf{Y}.$$

By assumption, U_i is independent of X_i, $\mathbf{Z}_i$ and ϵ_i, it is easy to verify that $E(\widehat{\boldsymbol{\Sigma}}) = E(\boldsymbol{\Sigma})$ and $E(\widehat{\boldsymbol{\rho}}) = E(\boldsymbol{\rho})$.

The gradient of the loss function plays an important role in statistical analysis. Because our corrected loss function is no longer the log-likelihood, we name it the gradient corrected score function, which has the form $\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\mathbf{S}_{i\boldsymbol{\theta}}(\boldsymbol{\theta}) = \widehat{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \widehat{\boldsymbol{\rho}}$. Because we aim at conducting inference on the parameter β, we treat the $p-1$ dimensional parameter $\boldsymbol{\gamma}$ as nuisance. Then the corrected score function can be decomposed as

$$\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \begin{pmatrix} S_{\beta}(\beta,\boldsymbol{\gamma}) \\ \mathbf{S}_{\boldsymbol{\gamma}}(\beta,\boldsymbol{\gamma}) \end{pmatrix} = \begin{pmatrix} \widehat{\Sigma}_{11}\beta + \widehat{\boldsymbol{\Sigma}}_{12}\boldsymbol{\gamma} - \widehat{\rho}_1 \\ \widehat{\boldsymbol{\Sigma}}_{21}\beta + \widehat{\boldsymbol{\Sigma}}_{22}\boldsymbol{\gamma} - \widehat{\boldsymbol{\rho}}_2 \end{pmatrix},$$

where $\widehat{\Sigma}_{11} = \mathbf{W}^{\mathrm{T}}\mathbf{W}/n - \sigma_U^2$, $\widehat{\boldsymbol{\Sigma}}_{12} = \mathbf{W}^{\mathrm{T}}\mathbf{Z}/n$, $\widehat{\boldsymbol{\Sigma}}_{21} = \mathbf{Z}^{\mathrm{T}}\mathbf{W}/n$, $\widehat{\boldsymbol{\Sigma}}_{22} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}/n$, $\widehat{\rho}_1 = \mathbf{W}^{\mathrm{T}}\mathbf{Y}/n$ and $\widehat{\boldsymbol{\rho}}_2 = \mathbf{Z}^{\mathrm{T}}\mathbf{Y}/n$.

Similar to the standard score function, it can be easily verified that $E\{\mathbf{S}_{i\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\} = \mathbf{0}$. Define the $p \times p$ corrected score covariance matrix as

$$\mathbf{I}(\boldsymbol{\theta}) = E\{\mathbf{S}_{i\boldsymbol{\theta}}(\boldsymbol{\theta})\mathbf{S}_{i\boldsymbol{\theta}}(\boldsymbol{\theta})^{\mathrm{T}}\} = \begin{pmatrix} I_{\beta\beta} & \mathbf{I}_{\beta\boldsymbol{\gamma}} \\ \mathbf{I}_{\boldsymbol{\gamma}\beta} & \mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}} \end{pmatrix}.$$

Note that the covariance matrix $\mathbf{I}(\boldsymbol{\theta})$ is no longer equal to $E\{\partial\mathbf{S}_{i\boldsymbol{\theta}}(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^{\mathrm{T}}\}$ due to the bias correction procedure in constructing the loss function. In fact, the matrix $\mathbf{I}(\boldsymbol{\theta})$ has more complex form. With standardized data matrix $(\mathbf{X}, \mathbf{Z})$, by simple calculations we obtain that

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{pmatrix} (\sigma_\epsilon^2 + \beta^2\sigma_U^2) + \sigma_\epsilon^2\sigma_U^2 + \beta^2 E(U_i^4) - \beta^2\sigma_U^4 & (\sigma_\epsilon^2 + \beta^2\sigma_U^2)E(X_i\mathbf{Z}_i)^{\mathrm{T}} \\ (\sigma_\epsilon^2 + \beta^2\sigma_U^2)E(X_i\mathbf{Z}_i) & (\sigma_\epsilon^2 + \beta^2\sigma_U^2)E(\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}) \end{pmatrix}. \tag{4.4}$$

To control the impact of high-dimensional nuisance parameter $\boldsymbol{\gamma}$ on the inference of the parameter of interest β, we define the corrected decorrelated score function for β as

$$S(\beta,\boldsymbol{\gamma}) = S_{\beta}(\beta,\boldsymbol{\gamma}) - \boldsymbol{\omega}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{\gamma}}(\beta,\boldsymbol{\gamma}),$$

where $\boldsymbol{\omega}^{\mathrm{T}} = \mathbf{I}_{\beta\boldsymbol{\gamma}}\mathbf{I}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}^{-1} = E(X_i\mathbf{Z}_i^{\mathrm{T}})E(\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}})^{-1} = E(W_i\mathbf{Z}_i^{\mathrm{T}})E(\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}})^{-1}$. Under the assumption that the minimal eigenvalue of $E\{(X_i, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}\otimes 2}\}$ is bounded and bounded away from 0, it is easy to show that the $(p-1)\times(p-1)$ matrix $E(\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}})$ is invertible. Note that this construction ensures that $S(\beta,\boldsymbol{\gamma})$ is uncorrelated with the nuisance score function $\mathbf{S}_{\boldsymbol{\gamma}}(\beta,\boldsymbol{\gamma})$, i.e. $E\{S(\beta_0,\boldsymbol{\gamma}_0)\mathbf{S}_{\boldsymbol{\gamma}}(\beta_0,\boldsymbol{\gamma}_0)\} = \mathbf{0}$. The detailed verification is in Appendix B.1.1.

We denote the variance of $S(\beta, \boldsymbol{\gamma})$ as $\sigma^2_{\beta|\boldsymbol{\gamma}}$, and it is easy to show that

$$\sigma^2_{\beta|\boldsymbol{\gamma}} = I_{\beta\beta} - \mathbf{I}_{\beta\boldsymbol{\gamma}}\mathbf{I}^{-1}_{\boldsymbol{\gamma}\boldsymbol{\gamma}}\mathbf{I}_{\boldsymbol{\gamma}\beta}. \tag{4.5}$$

Under the null hypothesis $H_0 : \beta_0 = \beta^*$, to construct score test statistic, we need to find estimators for the nuisance parameter $\boldsymbol{\gamma}$ and the $p-1$ dimensional vector $\boldsymbol{\omega}$. For $\boldsymbol{\gamma}$, we can use any consistent estimator $\widetilde{\boldsymbol{\gamma}}$ with sufficient convergence rate due to the decorrelation operation. More details about $\widetilde{\boldsymbol{\gamma}}$ as well as the initial estimator $\widetilde{\beta}$ for β will be discussed in Section 4.2.3. For $\boldsymbol{\omega}$, an intuitive estimator is its sample version $\widehat{\boldsymbol{\Sigma}}_{12}\widehat{\boldsymbol{\Sigma}}^{-1}_{22}$. However, matrix $\widehat{\boldsymbol{\Sigma}}_{22}$ is not invertible when $p-1 > n$. Ning et al. (2017) imposed sparsity assumption on $\boldsymbol{\omega}$ to control the estimation error. Many different penalized methods can be applied to obtain a sparse estimator of $\boldsymbol{\omega}$. For example, the Dantzig type estimator $\widehat{\boldsymbol{\omega}}$ can be obtained as follows:

$$\widehat{\boldsymbol{\omega}} = \text{argmin}\|\boldsymbol{\omega}\|_1 \ \text{ s.t. } \ \|\widehat{\boldsymbol{\Sigma}}_{12} - \boldsymbol{\omega}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22}\|_\infty \leqslant \lambda', \tag{4.6}$$

where λ' is a tuning parameter. Note that in our model $\widehat{\boldsymbol{\Sigma}}_{12}$ and $\widehat{\boldsymbol{\Sigma}}_{22}$ do not depend on $\boldsymbol{\theta}$. Then the estimated corrected decorrelated score function is defined as $\widehat{S}(\beta, \widetilde{\boldsymbol{\gamma}}) = S_\beta(\beta, \widetilde{\boldsymbol{\gamma}}) - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{\gamma}}(\beta, \widetilde{\boldsymbol{\gamma}})$.

Under the null hypothesis, we construct the test statistic as

$$\widehat{T}_n = n^{1/2}\widehat{S}(\beta^*, \widetilde{\boldsymbol{\gamma}})(\widehat{\sigma}^2_{\beta|\boldsymbol{\gamma},H_0})^{-1/2},$$

where

$$\begin{aligned}\widehat{\sigma}^2_{\beta|\boldsymbol{\gamma},H_0} &= \{\widehat{I}_{\beta\beta} - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\mathbf{I}}_{\boldsymbol{\gamma}\beta}\}|_{\beta=\beta^*} \\ &= (\widehat{\sigma}^2_{\epsilon,H_0} + \beta^{*2}\sigma^2_U)(1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}) + \beta^{*2}E(U_i^4) + \widehat{\sigma}^2_{\epsilon,H_0}\sigma^2_U - \beta^{*2}\sigma^4_U, \end{aligned} \tag{4.7}$$

and $\widehat{\sigma}^2_{\epsilon,H_0} = n^{-1}\sum_{i=1}^n (Y_i - \beta^* W_i - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)^2 - \beta^{*2}\sigma^2_U$. The detailed derivation is given in Appendix B.1.2. Under some assumptions we will specify in Section 4.4, the test statistic $\widehat{T}_n$ is asymptotically standard normal, see Corollary 4.3.6.

For confidence interval construction, define the one-step estimator for β as the root of the first order approximation of the approximately unbiased estimating equation $\widehat{S}(\beta, \widetilde{\boldsymbol{\gamma}}) = 0$ around the initial estimator $\widetilde{\beta}$, i.e.,

$$\widehat{\beta} \ = \ \widetilde{\beta} - \widehat{S}(\widetilde{\boldsymbol{\theta}})/\{\partial\widehat{S}(\beta, \widetilde{\boldsymbol{\gamma}})/\partial\beta\}|_{\beta=\widetilde{\beta}}$$

$$= \widetilde{\beta} - \widehat{S}(\widetilde{\boldsymbol{\theta}})/(\widehat{\boldsymbol{\Sigma}}_{11} - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}).$$

Of course, we could use the true root of $\widehat{S}(\beta, \widetilde{\boldsymbol{\gamma}}) = 0$ as $\widehat{\boldsymbol{\beta}}$. Here, we choose to use the one-step update for its computational simplicity. In fact, we have proved that the asymptotic distribution of the one-step estimator is identical to that of the true root because we have a relatively good initial estimator $\widetilde{\boldsymbol{\beta}}$. We will show that the one-step estimator $\widehat{\beta}$ is consistent and asymptotically normal with asymptotic variance σ_β^2 under suitable assumptions in Theorem 4.3.8. Hence, the $(1-\alpha)100\%$ confidence interval for β_0 can be constructed as $\left(\widehat{\beta} - z_\alpha\sqrt{\widehat{\sigma}_\beta^2/n}, \widehat{\beta} + z_\alpha\sqrt{\widehat{\sigma}_\beta^2/n}\right)$, where $\Phi(z_\alpha) = 1 - \alpha/2$, and $\widehat{\sigma}_\beta^2$ is an estimate of σ_β^2 whose specific form is given in Theorem 4.3.8.

4.2.3 Initial Estimator

In the literature, estimation theories under different assumptions have been developed for model (4.1), where all covariates are measured with error, see Loh & Wainwright (2012), Chen & Caramanis (2013), Belloni, Rosenbaum & Tsybakov (2017), Loh et al. (2017) and Datta et al. (2017). With slight modifications, these methods can all be applied to our model to construct desired initial estimators. Here, we take CoCoLasso estimator proposed by Datta et al. (2017) as an example to show how the convergence performance of the initial estimator affects the inferential results of β.

The CoCoLasso estimator is defined as

$$\widetilde{\boldsymbol{\theta}} = \mathrm{argmin}_{\boldsymbol{\theta}} \frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\theta} - \widehat{\boldsymbol{\rho}}^{\mathrm{T}}\boldsymbol{\theta} + \lambda\|\boldsymbol{\theta}\|_1, \tag{4.8}$$

where $\widetilde{\boldsymbol{\Sigma}} = (\widehat{\boldsymbol{\Sigma}})_+$ and λ is a tuning parameter. The nearest positive semi-definite matrix projection operator $(\cdot)_+$ is defined as follows: for any matrix $\mathbf{K}$,

$$(\mathbf{K})_+ = \mathrm{argmin}_{\mathbf{K}_1 \succcurlyeq \mathbf{0}}\|\mathbf{K} - \mathbf{K}_1\|_{\max}.$$

The ADMM algorithm is used to find the nearest positive semi-definite matrix. For more details, see Fan et al. (2016) and Datta et al. (2017).

As mentioned in the introduction, since we consider sub-Gaussian design with fixed number of covariates measured with error, which is different from the settings in Datta et al. (2017), we modified their theoretical proofs under our settings and the error bounds are different in terms of certain constants. We give the l_1, l_2 and prediction error bounds of $\widetilde{\boldsymbol{\theta}}$ in the following Lemma.

Lemma 4.2.1. *Let* $\lambda = C_\lambda s_0\sqrt{n^{-1}logp} = o(1)$. *For* $C_\lambda > \max\left(8K_0K_2/C'', 8\sqrt{2}K_0K_3/\sqrt{C''}\right)$ *and* $\lambda \leqslant \min(8K_1, 16KK_\epsilon, 8K_0K_2, 8K_0K_3)$, *with probability at least* $1 - C_1\exp(-C_2 logp)$, *we have*

$$\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 \leqslant 16\lambda s_0/\kappa, \quad \|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2 \leqslant \sqrt{32s_0}\lambda/\kappa, \quad \textit{and} \quad \|(\mathbf{X}, \mathbf{Z})(\boldsymbol{\theta}_0 - \widetilde{\boldsymbol{\theta}})\|_2/\sqrt{n} \leqslant \lambda\sqrt{32s_0/\kappa},$$

where $\|\boldsymbol{\theta}_0\|_\infty \leqslant K_0$, C'' *is a universal constant,* C_1 *and* C_2 *are positive constants depending on* K, K_ϵ, K_U, K_0, κ *and* σ_U^2 *given in the proof,* $K_1 = 2K_U(K_0K + K_\epsilon)$, $K_2 = 4K(K + K_U) + 2K_U^2 + \sigma_U^2$ *and* $K_3 = 4(K + K_U)^2 + 2\sigma_U^2$.

The detailed proof is given in Appendix B.2.1. It is based on the closeness condition for $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\rho}}$, and the restricted eigenvalue (RE) condition for matrix $\boldsymbol{\Sigma}$. Different from deterministic design, Bernstein inequalities were used repeatedly and we have shown that under the assumption that $s_0\sqrt{n^{-1}\log p} = \mathrm{o}(1)$, the RE condition for sub-Gaussian matrix $\boldsymbol{\Sigma}$ holds with probability at least $1 - 2p^{-\zeta}$ in Lemma B.3.2.

For the l_∞ error bound, for simplicity, slightly different notations are used here. Specifically, we write $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{0,S}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$, $(\mathbf{X}, \mathbf{Z}) = (\mathbf{Q}_S, \mathbf{Q}_{S^C})$, and then partition the matrix $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \begin{pmatrix} n^{-1}\mathbf{Q}_S^{\mathrm{T}}\mathbf{Q}_S & n^{-1}\mathbf{Q}_S^{\mathrm{T}}\mathbf{Q}_{S^C} \\ n^{-1}\mathbf{Q}_{S^C}^{\mathrm{T}}\mathbf{Q}_S & n^{-1}\mathbf{Q}_{S^C}^{\mathrm{T}}\mathbf{Q}_{S^C} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{S,S} & \boldsymbol{\Sigma}_{S,S^C} \\ \boldsymbol{\Sigma}_{S^C,S} & \boldsymbol{\Sigma}_{S^C,S^C} \end{pmatrix}.$$

To clarify, the above partition is based on the true support of model (3.3), that is, whether $\mathbf{X}$ is a part of $\mathbf{Q}_S$ depends on the true value β_0. Actually, when deriving the l_∞ error bound for $\widetilde{\boldsymbol{\theta}}$, whether β_0 equals 0 would not affect the proof as well as the theoretical result. To derive the l_∞ error bound for $\widetilde{\boldsymbol{\theta}}$, we need to further assume that

$$\lambda_{\min}\{E(\boldsymbol{\Sigma}_{S,S})\} = \kappa_S > 0, \quad \text{and} \quad \|E(\boldsymbol{\Sigma}_{S^C,S})\{E(\boldsymbol{\Sigma}_{S,S})\}^{-1}\|_\infty \leqslant 1 - \boldsymbol{\gamma}, \tag{4.9}$$

for some $\boldsymbol{\gamma} \in (0, 1]$. Let $\|E(\boldsymbol{\Sigma}_{S,S})^{-1}\|_\infty = \phi$ and $\|E(\boldsymbol{\Sigma}_{S,S})^{-1}\|_\infty = \Phi$. The l_∞ error bound result is stated as follows, which are similar to those given in Theorem 2 in Datta et al. (2017) with minor modifications. The detailed proof is given in the Appendix B.2.2.

Lemma 4.2.2. *Let* $\lambda = C_\lambda s_0\sqrt{n^{-1}logp} = o(1)$. *Under the assumptions given in (4.9) and* $C_\lambda > 8K_4/(\gamma\sqrt{C''})$, *where* $K_4 = 2K^2K_0 + 2KK_\epsilon$

(a) With probability at least $1 - p_1(\delta)$, *there exists a unique solution* $\widetilde{\boldsymbol{\theta}}$ *minimizing* $\boldsymbol{\theta}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\theta}/2 - \widehat{\boldsymbol{\rho}}^{\mathrm{T}}\boldsymbol{\theta} + \lambda\|\boldsymbol{\theta}\|_1$ *whose support is a subset of the true support.*

(b) With probability at least $1 - p_2(\delta')$, $\|\widetilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0S}\|_\infty \leqslant C_\infty \lambda$, *where* $C_\infty = 8\phi$.

Probabilities $p_1(\delta)$ *and* $p_2(\delta')$ *go to zero as* n *goes to infinity and the detailed expressions are given in Appendix B.2.2.*

Note that Parts (a) and (b) of Lemma 4.2.2 imply that under the given conditions, $\|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_\infty = \|\widetilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0S}\|_\infty \leqslant C_\infty \lambda$ with probability at least $1 - p_1(\delta) - p_2(\delta')$.

Remark 4.2.1. *Note that we use* $\widetilde{\boldsymbol{\Sigma}}$ *in the loss function for CoCoLasso estimator to make the problem convex, but use* $\widehat{\boldsymbol{\Sigma}}$ *in the loss function to construct decorrelated score function. This discrepancy does not cause any problem when deriving the theoretical properties of our corrected score test statistic and one-step estimator.*

Remark 4.2.2. *For CoCoLasso estimator, the tuning parameter* λ *has the order* $s_0\sqrt{n^{-1}logp}$. *However, in Loh et al. (2017), the tuning parameter has the order* $\sqrt{n^{-1}logp}$ *under the assumption that* $\|\boldsymbol{\theta}_0\|_1$ *is bounded. In our proofs, we only assume that* $\|\boldsymbol{\theta}_0\|_\infty$ *is bounded. With the stronger assumption that* $\|\boldsymbol{\theta}_0\|_1$ *is bounded, the error bounds of CoCoLasso estimator would have the same order as those proposed in Loh et al. (2017) and Belloni, Rosenbaum & Tsybakov (2017).*

4.2.4 Algorithm

Now we summarize the proposed estimation procedure as the following algorithm.

1. Calculate the initial CoCoLasso estimator $\widetilde{\boldsymbol{\theta}} = (\widetilde{\beta}, \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$.

2. Estimate $\boldsymbol{\omega}$ by the Dantzig type estimator $\widehat{\boldsymbol{\omega}}$,

$$\widehat{\boldsymbol{\omega}} = \operatorname{argmin}\|\boldsymbol{\omega}\|_1 \quad \text{s.t.} \quad \|\widehat{\boldsymbol{\Sigma}}_{12} - \boldsymbol{\omega}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22}\|_\infty \leqslant \lambda',$$

   where $\lambda' = O(\sqrt{\log p/n})$. For the detailed algorithm, see Candes et al. (2007). Note that other penalized M-estimators can also be used to solve for $\widehat{\boldsymbol{\omega}}$, for example, the Lasso.

3. Calculate the estimated decorrelated score function

$$\widehat{S}(\beta, \widetilde{\boldsymbol{\gamma}}) = S_\beta(\beta, \widetilde{\boldsymbol{\gamma}}) - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{\gamma}}(\beta, \widetilde{\boldsymbol{\gamma}}),$$

   and the test statistic $\widehat{T}_n = n^{1/2}\widehat{S}(\beta^*, \widetilde{\boldsymbol{\gamma}})(\widehat{\sigma}^2_{\beta|\boldsymbol{\gamma},H_0})^{-1/2}$, where $\widehat{\sigma}^2_{\beta|\boldsymbol{\gamma},H_0}$ is given in (4.7). Under the conditions given in Theorem 4.3.5, the test statistic $\widehat{T}_n$ is asymptotically standard normal.

4. Calculate the one-step estimator

$$\begin{aligned}\widehat{\beta} &= \widetilde{\beta} - \widehat{S}(\widetilde{\boldsymbol{\theta}})/\{\partial\widehat{S}(\beta,\widetilde{\boldsymbol{\gamma}})/\partial\beta\}|_{\beta=\widetilde{\beta}} \\ &= \widetilde{\beta} - \widehat{S}(\widetilde{\boldsymbol{\theta}})/(\widehat{\boldsymbol{\Sigma}}_{11} - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}).\end{aligned}$$

Construct the $(1-\alpha)100\%$ confidence interval for β_0 as $[\widehat{\beta}-z_\alpha\sqrt{\widehat{\sigma}^2_\beta/n}, \widehat{\beta}+z_\alpha\sqrt{\widehat{\sigma}^2_\beta/n}]$, where $\Phi(z_\alpha) = 1-\alpha/2$, and $\widehat{\sigma}^2_\beta$ is given in Theorem 4.3.8.

4.3 Theory for Test and Confidence Intervals

4.3.1 Technical Lemmas

We first establish four technical lemmas 4.3.1, 4.3.2, 4.3.3 and 4.3.4 to ensure the asymptotic normality of the corrected score test statistic $\widehat{T}_n$ and the one-step estimator $\widehat{\beta}$.

Lemma 4.3.1. *Recall that $S' = \mathrm{supp}(\boldsymbol{\omega})$ and $|S'| = s'$. Let $\lambda' = C_{\lambda'}\sqrt{n^{-1}logp}$. The Dantzig type estimator $\widehat{\boldsymbol{\omega}}$ satisfies $\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}\|_1 = O_P(s'\sqrt{n^{-1}logp})$, when $C_{\lambda'} > \sqrt{2K_5/C''}$. Here C'' is a universal constant and $K_5 = 2K(K + K_U + KK_\omega)$.*

Lemma 4.3.2. *Let $\boldsymbol{\nu} = (1, -\boldsymbol{\omega}^{\mathrm{T}})^{\mathrm{T}}$. The gradient and Hessian of the corrected loss function (4.3) satisfy $\|\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\|_\infty = O_p(\sqrt{n^{-1}logp})$ and $\|\boldsymbol{\nu}^{\mathrm{T}}\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0) - E\{\boldsymbol{\nu}^{\mathrm{T}}\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\}\|_\infty = O_P(\sqrt{n^{-1}logp})$.*

Lemma 4.3.3. *Let $\widetilde{\boldsymbol{\theta}}_{H_0} = (\beta^*, \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$, $\widehat{\boldsymbol{\nu}} = (1, -\widehat{\boldsymbol{\omega}}^{\mathrm{T}})^{\mathrm{T}}$. Assume that*

$$\frac{s_0(s' \vee s_0)logp}{\sqrt{n}} = o(1).$$

Then $\boldsymbol{\nu}^{\mathrm{T}}\{\mathbf{S}_{\boldsymbol{\theta}}(\breve{\boldsymbol{\theta}}) - \mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0) - \nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)(\breve{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\} = 0$, and $(\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu})^{\mathrm{T}}\{\mathbf{S}_{\boldsymbol{\theta}}(\breve{\boldsymbol{\theta}}) - \mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\} = o_P(n^{-1/2})$, for both $\breve{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}_{H_0}$ and $\breve{\boldsymbol{\theta}} = \widetilde{\boldsymbol{\theta}}$.

Lemma 4.3.4. *When (4.2) does not degenerate, i.e., the corrected decorrelated score function $S(\boldsymbol{\theta}) \not\equiv 0$ a.s., then*

$$\sqrt{n}\boldsymbol{\nu}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)(\sigma^2_{\beta|\boldsymbol{\gamma},0})^{-1/2} \to \mathcal{N}(0,1)$$

in distribution. Here $\sigma^2_{\beta|\boldsymbol{\gamma},0} = (\sigma^2_\epsilon + \beta_0^2\sigma^2_U)\{1 - \boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)\} + \beta_0^2E(U_i^4) + \sigma^2_\epsilon\sigma^2_U - \beta_0^2\sigma^4_U$ by (4.4) and (4.5), and $\sigma^2_{\beta|\boldsymbol{\gamma},0} \geqslant C$ for some positive constant C.

Lemma 4.3.1, together with Lemma 4.2.1, states the consistency properties for initial estimators $\widetilde{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\omega}}$, which are crucial to the asymptotic performance of our corrected test statistic and one-step estimator. Lemma 4.3.2 and Lemma 4.3.3 describe the concentration properties of the gradient and Hessian of the corrected loss function (4.3), and its local smoothness properties, respectively. For high-dimensional random designs, it is important to quantify the distance between sample level statistic and its corresponding population level value, especially for critical statistics like the score function and the Hessian matrix. For local smoothness, Ning et al. (2017) require $(s_0 \vee s')n^{-1/2}\log p = o(1)$. However, using CoCoLasso estimator as the initial estimator, we need a stronger condition on dimensionality and sparsity to guarantee the $n^{-1/2}$ rate local smoothness of the corrected loss function. Lemma 4.3.4 is the central limit theorem for corrected decorrelated sore function $S(\boldsymbol{\theta}_0)$, which is a linear combination of $S_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)$. Because we define the score function as the gradient of the corrected loss function, which is different from negative log-likelihood, the variance $\sigma^2_{\beta|\boldsymbol{\gamma},0}$ of $S(\boldsymbol{\theta}_0)$ has relatively complex form. Detailed proofs of the four lemmas are given in Appendices B.3.1, B.3.2, B.3.3 and B.3.4, respectively.

4.3.2 Corrected Score Test

Theorem 4.3.5. *Under conditions of Lemmas 4.3.1 - 4.3.3 and under $H_0 : \beta_0 = \beta^*$, it follows that*

$$n^{1/2}\widehat{S}(\beta^*, \widetilde{\boldsymbol{\gamma}})(\sigma^2_{\beta|\boldsymbol{\gamma},0})^{-1/2} \to \mathcal{N}(0,1)$$

in distribution.

In Theorem 4.3.5, we state the asymptotic normality of the decorrelated score test statistic by assuming its true variance is known. The detailed proof is given in Appendix B.4.1. To show the asymptotic properties of the test statistic $\widehat{T}_n$ with estimated variance $\widehat{\sigma}^2_{\beta|\boldsymbol{\gamma},H_0}$, we need to further study the difference between $\widehat{\sigma}^2_{\beta|\boldsymbol{\gamma},H_0}$ and $\sigma^2_{\beta|\boldsymbol{\gamma},0}$, which is more complex than that of linear models without measurement error. We need to use l_∞ error bound of $\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0$ to facilitate the proof. Under a stronger assumption that $s_0^3\sqrt{n^{-1}\log p} = o(1)$, we show that $\widehat{T}_n$ is still asymptotically standard normal in the following corollary and detailed proof can be found in Appendix B.4.2.

Corollary 4.3.6. *Suppose that $s_0^3\sqrt{n^{-1}\log p} = o(1)$. Under conditions of Lemmas 4.3.1 - 4.3.3 and under H_0, it follows that*

$$n^{1/2}\widehat{S}(\beta^*, \widetilde{\boldsymbol{\gamma}})(\widehat{\sigma}^2_{\beta|\boldsymbol{\gamma},H_0})^{-1/2} \to \mathcal{N}(0,1)$$

in distribution.

Remark 4.3.1. *Assume that* $log(p) = O(n^{a_1})$, $s_0 = O(n^{a_2})$ *and* $s' = O(n^{a_3})$. *Then the conditions in Corollary 4.3.6 together with* $s_0(s_0 \vee s')n^{-1/2}logp = o(1)$, *imply that*

$$a_2 + (a_2 \vee a_3) + a_1 < 1/2 \quad and \quad 3a_2 + a_1/2 < 1/2.$$

The inference framework of Ning et al. (2017) requires $(a_2 \vee a_3) + a_1 < 1/2$, *while the consistency of CoCoLasso estimator of Datta et al. (2017) requires* $2a_2 + a_1/2 < 1/2$. *Our requirement on* (n, p, s_0, s') *here is stronger. This is because the CoCoLasso estimator converges more slowly than standard penalized M-estimators for high-dimensional linear models. On the other hand, the inference framework based on decorrelation operation needs stronger assumptions on dimensionality and sparsity compared with pure estimation theory.*

We further study the power of our test statistic $\widehat{T}_n$ at local alternatives in the following corollary, and its proof is given in Appendix B.4.3.

Corollary 4.3.7. *Consider the local alternative* $\beta_n = \beta^* + h/\sqrt{n}$, *where* h *is a constant. Under the assumptions given in Corollary 4.3.6, we have* $\widehat{T}_n + h(\sigma^2_{\beta_n})^{-1/2}$ *converges to* $\mathcal{N}(0,1)$ *in distribution under the local alternatives, where* $\sigma^2_{\beta_n} = [E\{\partial S(\beta, \boldsymbol{\gamma}_0)/\partial\beta \mid_{\beta=\beta_n}\}]^{-2}\sigma^2_{\beta_n|\boldsymbol{\gamma},0}$, *and* $\sigma^2_{\beta_n|\boldsymbol{\gamma},0}$ *is* $\sigma^2_{\beta|\boldsymbol{\gamma},0}$ *with* β_0 *replaced by* β_n.

4.3.3 Confidence Interval

In addition to hypothesis testing, we also construct asymptotic confidence intervals for the parameter of interest β based on the one-step estimator $\widehat{\beta}$. Its asymptotic normality is given in the following theorem and the detailed proof is given in Appendix B.5.1.

Theorem 4.3.8. *Suppose conditions of Lemmas 4.3.1 - 4.3.4 are valid, if*

$$E[\{\partial S(\beta, \boldsymbol{\gamma}_0)/\partial\beta\}|_{\beta=\beta_0}] \geqslant C$$

for some positive constant C, *then*

$$n^{1/2}(\widehat{\beta} - \beta_0) = -\left[E\left\{\frac{\partial S(\beta, \boldsymbol{\gamma}_0)}{\partial\beta}\Big|_{\beta=\beta_0}\right\}\right]^{-1} n^{1/2}S(\beta_0, \boldsymbol{\gamma}_0) + o_P(1) \to N(0, \sigma^2_\beta)$$

in distribution, where the asymptotic variance $\sigma_\beta^2 = \{E(X_i^2) - \boldsymbol{\omega}^{\mathrm{T}} E(X_i \mathbf{Z}_i)\}^{-2} \sigma_{\beta|\boldsymbol{\gamma},0}^2$. *The variance* σ_β^2 *can be estimated as*

$$\widehat{\sigma}_\beta^2 = \left(1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{21}\right)^{-2} \left\{ (\widehat{\sigma}_\epsilon^2 + \widehat{\beta}^2 \sigma_U^2)(1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_{21}) + \widehat{\beta}^2 E(U_i^4) + \widehat{\sigma}_\epsilon^2 \sigma_U^2 - \widehat{\beta}^2 \sigma_U^4 \right\}, \quad (4.10)$$

where $\widehat{\sigma}_\epsilon^2 = n^{-1} \sum_{i=1}^n (Y_i - \widehat{\beta} W_i - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}} \mathbf{Z}_i)^2 - \widehat{\beta}^2 \sigma_U^2$.

Remark 4.3.2. *Lemma 4.2.2 shows that the sign consistency property of the CoCoLasso estimator is ensured by the minimal signal condition* $\min_{j \in S} |\boldsymbol{\theta}_j| > C_\infty \lambda$. *That is, when* $|\beta_0| < C_\infty \lambda$, *then the CoCoLasso estimate* $\widetilde{\beta}$ *will be set to* 0 *with high probability. With the decorrelation operation, the convergence performance of our one-step estimator* $\widehat{\beta}$ *is improved significantly. Meanwhile, our test statistic* $\widehat{T}_n$ *retains power under the local alternatives around* 0.

Remark 4.3.3. *In low dimensional case, Nakamura (1990) provided inference results of generalized linear models with measurement error using corrected score functions. We have established inference results in high-dimensional settings. Since* $\sigma_{\beta|\boldsymbol{\gamma}}^2$ *is the variance of the corrected decorrelated score* $S(\beta, \boldsymbol{\gamma})$, *the form of our asymptotic variance* $\sigma_{\boldsymbol{\beta}}^2$ *is similar to theirs. Further, we show that our one-step estimator is semiparametrically efficient.*

4.4 Empirical Studies

4.4.1 Simulation Studies

We conducted simulation studies in Matlab/R2017b under different settings to investigate the performance of our proposed corrected decorrelated score test and the one-step estimator. The code is available for public use. To generate the data matrix $(\mathbf{X}, \mathbf{Z})$, we simulated $n = 100$ and $n = 200$ independent and identically distributed samples from a multivariate Gaussian distribution $N_p(\mathbf{0}, \boldsymbol{\Sigma})$, where $p = 250$ and $\boldsymbol{\Sigma}$ is the autoregressive matrix with its entry $\boldsymbol{\Sigma}_{jk} = \rho^{|j-k|}$. We considered two cases, where $\rho = 0.25$ and $\rho = 0.5$. To generate the responses $\mathbf{Y}$, we added the regression error $\boldsymbol{\epsilon}$ following the normal distribution $N(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_n)$, where $\sigma_\epsilon = 0.2$. The measurement error $\mathbf{U}$ was generated from $N(\mathbf{0}, \sigma_U^2 \mathbf{I}_n)$. Three different values of σ_U are considered, where $\sigma_U = 0.1, 0.15$ and 0.2 respectively. Both estimation and inference become progressively more difficult with larger measurement error variance. We considered two scenarios for the true parameter

$\boldsymbol{\theta}_0 = (\beta_0, \boldsymbol{\gamma}_0^{\mathrm{T}})^{\mathrm{T}}$. In the first scenario, $\boldsymbol{\theta}_0 = (1, 1, 0, \ldots 0)^{\mathrm{T}}$. In the second scenario, we set $\boldsymbol{\theta}_0 = (1, 0.8, 1.5, 0, \ldots 0)^{\mathrm{T}}$. Our goal is to test $H_0 : \beta_0 = 1$ versus $H_1 : \beta_0 \neq 1$.

For the initial CoCoLasso estimator $\widetilde{\boldsymbol{\theta}}$, we first perform variable selection using (4.8). Then refit the model using the selected covariates and set the coefficients of the rest of the covariates to zero. During the procedure, the tuning parameter λ' in (4.6) is selected by a K-fold cross-validation, where $K = 4$. Specifically, the optimal λ' is chosen in the sense of l_2 prediction for the test sample, see Bickel (2007).

In each setting, 1000 simulations are conducted. The averaged type I error rates at significance levels $\alpha = 1\%, 5\%$ and 10% of our test are summarized in Table 4.1. We can see that the type I error rates are very close to the nominal significance levels in all the simulation settings. To examine the power of our test, we regenerated data with $\beta_0 = 1.05, 1.10, 1.15$ and report the rejection rate at different significance levels ranging from 1% to 10%. The results, together with the rejection rates under H_0 when $\beta_0 = \beta^* = 1$, are shown in Figure 4.1, as well as Figures B.1, B.2 and B.3 in Appendix. Overall, the test has very good performance in terms of level under H_0, reflected in the close approximation of the observed rejection rates and the nominal levels. The power performance is also satisfactory in general, where the curves representing the rejection rates under all three alternatives are well separated from the null rejection curve, and the power increases when sample size increases, the correlation ρ decreases, the nonzero covariates number is smaller, or the measurement error variance decreases.

We also provide the performance of our one-step estimator $\widehat{\beta}$ in Table 2, where we report the mean and standard deviation of 1000 estimates of $\widehat{\beta}$, as well as the average of the estimated asymptotic standard deviation calculated based on (4.10). In addition, we constructed the 95% confidence intervals in each simulation using the asymptotic normality of $\widehat{\beta}$, and computed the empirical coverage of the true value β_0. We find that the one-step estimator performs well in different simulations settings. In each setting, the difference between the mean of the estimates and the true value is very small, the mean of estimated standard deviations closely approximates the empirical value, and the empirical coverage of the estimated 95% confidence intervals is reasonably close to the nominal level.

We have assumed σ_U^2 and $E(U_i^4)$ to be known. In this section, we further conducted simulation studies to examine the impact of $\widehat{\sigma}_U^2$ and $\widehat{E}(U_i^4)$. The simulation results are in the Appendix B.7.2.

Table 4.1: Type I error of the corrected decorrelated score test at different significance levels

		Scenario 1				Scenario 2			
		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.25$		$\rho = 0.5$	
σ_U	α	$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
0.1	1%	1.1%	0.9%	1.6%	1.1%	1.0%	0.8%	1.4%	1.1%
	5%	5.6%	4.4%	5.2%	5.4%	5.6%	4.4%	5.5%	5.5%
	10%	9.8%	10.6%	9.4%	12.0%	10.2%	10.7%	9.3%	12.0%
0.15	1%	1.4%	1.0%	1.3%	1.1%	1.1%	1.1%	1.4%	0.9%
	5%	5.4%	4.9%	5.4%	5.9%	4.6%	5.4%	5.9%	5.9%
	10%	11.3%	10.8%	9.9%	11.4%	9.2%	10.9%	10.5%	12.0%
0.2	1%	1.5%	1.0%	1.4%	0.8%	2.1%	1.2%	1.5%	0.7%
	5%	6.2%	5.9%	5.9%	5.7%	6.0%	6.3%	6.4%	5.7%
	10%	10.9%	10.9%	10.7%	11.7%	11.1%	10.5%	11.3%	11.3%

Table 4.2: Performance of the one-step estimator $\widehat{\beta}$

		Scenario 1				Scenario 2			
		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.25$		$\rho = 0.5$	
σ_U		$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$	$n = 100$	$n = 200$
0.1	Mean	1.0008	1.0000	1.0010	1.0001	1.0008	1.0000	1.0003	0.9996
	Est sd	0.0237	0.0163	0.0259	0.0177	0.0235	0.0162	0.0257	0.0177
	Emp sd	0.0246	0.0167	0.0259	0.0185	0.0246	0.0167	0.0260	0.0185
	Emp cvg	94.1%	94.6%	93.6%	93.8%	93.9%	94.7%	93.7%	93.6%
0.15	Mean	1.0024	1.0011	1.0031	1.0012	1.0024	1.0011	1.0028	1.0009
	Est sd	0.0268	0.0183	0.0294	0.0199	0.0267	0.0183	0.0292	0.0199
	Emp sd	0.0279	0.0183	0.0299	0.0204	0.0270	0.0184	0.0302	0.0204
	Emp cvg	94.2%	94.9%	93.3%	94.0%	94.6%	94.7%	93.4%	93.7%
0.2	Mean	1.0049	1.0015	1.0062	1.0024	1.0085	1.0016	1.0061	1.0021
	Est sd	0.0309	0.0211	0.0341	0.0230	0.0313	0.0211	0.0339	0.0229
	Emp var	0.0324	0.0217	0.0355	0.0234	0.0333	0.0218	0.0359	0.0234
	Emp cvg	94.1%	94.1%	93.0%	93.5%	92.9%	93.9%	92.2%	93.2%

In Table 4.2, "Est sd" denotes the mean of 1000 estimated asymptotic standard deviations; "Emp sd" denotes the empirical standard deviation of 1000 estimates; "Emp cvg" denotes the empirical coverage of the estimated 95% CI for $\boldsymbol{\beta}_0$.

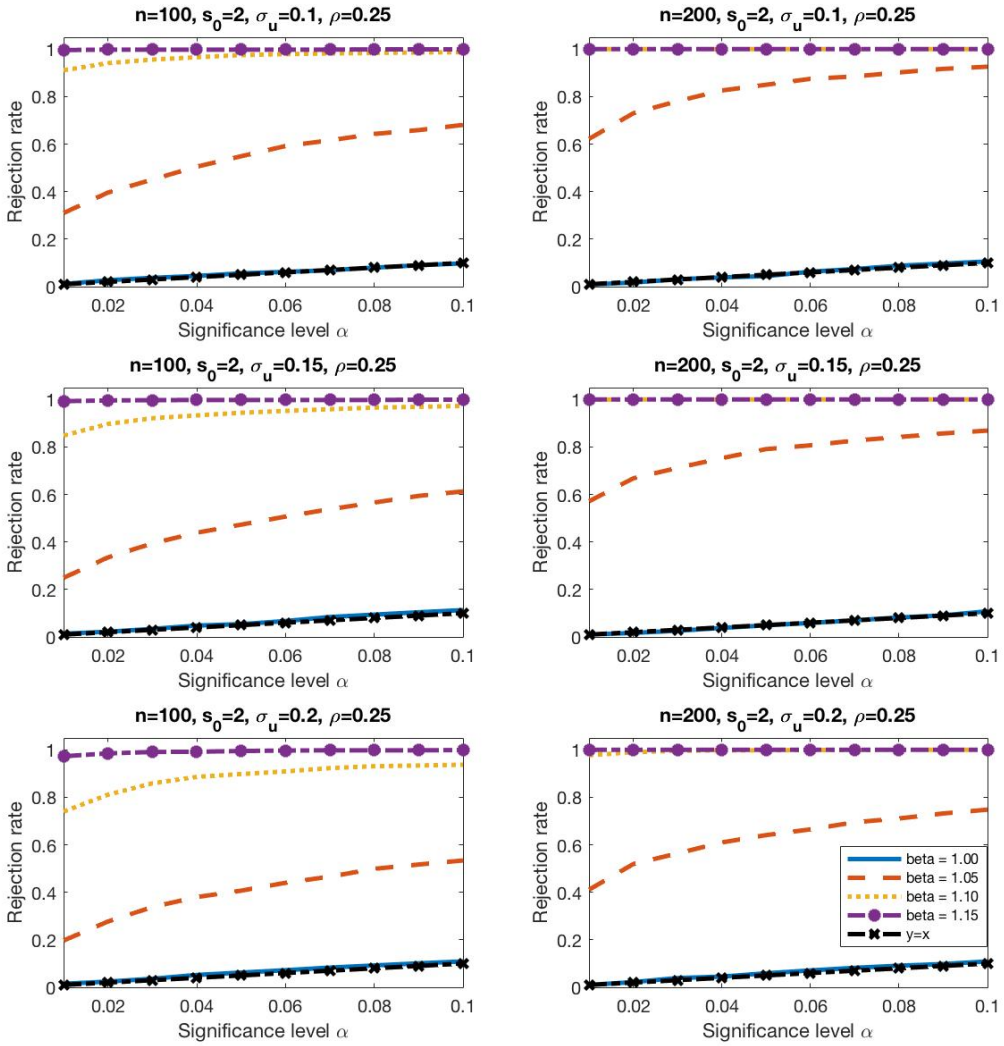

Figure 4.1: Power of the proposed corrected decorrelated score test at different significance levels in scenario 1 with $\rho = 0.25$

4.4.2 Real Data Analysis

We illustrate the proposed procedure via an empirical analysis of a data set analyzed in Chu et al. (2016). The data set was collected in a clinical trial designed to determine the long-term effects of different inhaled treatments for mild to moderate childhood asthma, where phenotypic information and genome-wide SNP data are accessible. The FEV1/FVC ratio is an important index used in diagnosis of obstructive and restrictive lung disease, which represents the proportion of a person's vital capability to expire in the first second

of forced expiration to the full vital capacity. We are interested in understanding how this ratio, often measured with errors, together with basic demographic variables and SNPs would affect the severity of asthma symptoms in children.

Here we focus on $n = 199$ subjects in the nedocromil treatment group, each had four clinical visits over 8 months. Exploratory data analysis was conducted on the four measurements of FEV1/FVC ratio, and no visible time trend was detected. The response variable Y_i is the average asthma symptoms (amsys). We let X_i be the unobserved FEV1/FVC ratio and W_i be the average of four measurements with homoscedastic measurement errors. Standard deviation and the fourth moment of measurement error U_i are estimated using the four measurements for each subject based on the fact that $W_{ik} - W_{ij} = U_{ik} - U_{ij}$, $\text{var}(U_{ik} - U_{ij}) = 2\sigma_U^2$, and $E(U_i^4) = [E\{(U_{ik} - U_{ij})^4\} - 6\sigma_U^4]/2$ for $i = 1, \ldots n$ and $j, k = 1, \ldots, 4$. Note that we do not need to assume the normality of measurement errors here. The estimated values are $\widehat{\sigma}_U = 0.4625$ and $\widehat{E}(U_i^4) = 0.1719$. The error-free variables $\mathbf{Z}_i$ are gender, age at baseline and 676 SNPs screened based on minor allele frequency (MAF). Here we treat SNPs as continuous variables by assuming that having two of the minor alleles has twice the effect on the phenotype as having one of the minor alleles, and zero means no effect.

Our goal is to first select significant variables among $p = 679$ variables in model (4.2), estimate the corresponding coefficients and then make inference for the error-prone variable FEV1/FVC ratio based on the proposed corrected decorrelated score test and the asymptotic properties of the one-step estimator. For the initial CoCoLasso estimator $\widetilde{\boldsymbol{\theta}}$, the tuning parameter λ is selected by cross validation with the criterion proposed in Datta et al. (2017). We find that besides FEV1/FVC ratio which is of interest, seven SNPs are selected. Detailed information about the selected SNPs is given in Table 4.3.

Under the null hypothesis $H_0 : \beta_0 = 0$, the corrected decorrelated score test statistic $\widehat{T}_n = 4.9806$. Hence, we reject the null hypothesis. The CoCoLasso estimate for β is -0.0654, while the one-step estimate is -0.1101 with confidence interval $(-0.1508, -0.0693)$. The negativeness of $\widehat{\beta}$ verifies the fact that the lower the FEV1/FVC ratio, the severer the obstruction of air escaping from the lungs.

Throughout the data analysis, we estimated the second and fourth moments of the measurement error using the four measurements of each subject. Because of the independent error assumption, $U_{ik} + U_{ij}$ is uncorrelated to $U_{ij} - U_{ik}$. Recall that the W_i relies on $U_{ik} + U_{ij}$, while the error moment estimates are based on $U_{ij} - U_{ik}$. Under normality assumption, the standard errors of the two moment estimates do not affect the performance of our proposed inference procedure.

Table 4.3: Information about the seven SNPs selected by CoCoLasso method

	SNP_1	SNP_2	SNP_3	SNP_4	SNP_5	SNP_6	SNP_7
SNP name	rs2830066	rs11798747	rs6961655	rs4432291	rs6860832	rs699770	rs4520841
Chromosome	21	X	7	17	5	1	16
Chr.position	26121885	18889776	136422490	72610903	8451644	119318352	26088794
Coefficient	0.0143	-0.0125	-0.0118	-0.0092	-0.0056	-0.0046	-0.0025

4.5 Discussion

In this Chapter, we have proposed an inference procedure for high-dimensional linear measurement error models based on corrected decorrelated score functions. With the decorrelation operation, our corrected score test statistic $\widehat{T}_n$ is asymptotically normal and retains power under the local alternatives around 0. Further, the convergence rate of the one-step estimator $\widehat{\boldsymbol{\beta}}$ has significantly improved compared to that of the initial estimator and achieves the semiparametric efficiency. Here we have assumed that the variance and the fourth moments of the measurement error are known. The framework in this paper still works if we treat σ_U^2 and $E(U_i^4)$ as nuisance parameters and then conduct decorrelation. Specifically, the new nuisance parameters are $(\boldsymbol{\gamma}^{\mathrm{T}}, \sigma_U^2, E(U_i^4))$. Note that we do not impose any penalty on σ_U^2 and $E(U_i^4)$.

One future research direction is to develop inference procedures when the number of covariates with measurement errors diverges with sample size n. Another possible consideration is to relax the sparsity assumption on $\boldsymbol{\omega}$. That is, extend the theory to cases where the ordered entries of $\boldsymbol{\omega}$ decay at a certain rate.

Chapter 5
Regression Models with Nonignorable and Partially Unspecified Missingness

5.1 Introduction

Missing data are ubiquitous in most areas of scientific inquiry, and especially in research involving human subjects, such as health related studies and sample surveys. Differentiating the different nature of missingness is crucial in statistical analysis with missing data. The missingness mechanism is called ignorable if it is independent of the missing values. Otherwise, it is called nonignorable. Extensive literature exists on ignorable missing data (Kim & Shao 2013, Little & Rubin 2002, Molenberghs et al. 2014, Robins et al. 1994, Rubin 1987, Schafer 1997, Tsiatis 2006), while nonignorable missing data is more challenging with less research on it.

The assumption on ignorable missingness can be violated in many applications. For example, when evaluating a new biomarker in analytical chemistry, scientists usually encounter a detection limit, defined as the lowest concentration of analyte distinguishable from the background noise (Carter et al. 2016). Concentration values below the detection limit are usually not released by laboratories. In this example, the missingness mechanism depends on the missing values themselves, thus it is nonignorable. Applying methods developed for ignorable missing data to nonignorable missing data can lead to biased estimation and erroneous inference.

One intrinsic difficulty for analyzing nonignorable missing data is on modeling the missingness mechanism. There have been many papers modeling the mechanism para-

metrically (Chang & Kott 2008, Ibrahim & Lipsitz 1996, Morikawa & Kim 2016, Qin et al. 2002, Rotnitzky & Robins 1997, Wang et al. 2014). However, the parametric model assumptions are usually uncheckable, because the mechanism depends on the missing values. Tang et al. (2003) considered a regression model with nonignorable missing response and proposed three pseudo-likelihood estimators treating the missingness mechanism as nuisance and allowing it to be unspecified. But they assumed that the mechanism only depends on the response variable, which is restrictive. Zhao & Ma (2019) proposed a versatile estimation procedure which allows the mechanism to depend on some covariates and does not require modeling or estimating the missingness mechanism. However, with a working model that can be different from the true mechanism, the efficiency in estimation is impaired. Efforts have also been made on semiparametric missingness mechanism. Kim & Yu (2011), Shao & Wang (2016) modeled the missingness mechanism with a semiparametric logistic regression model and proposed semiparametric estimators of unknown population parameters based on data with nonignorable missing responses. In the semiparametric mechanism, they modeled the relationship between the missingness indicator and the response parametrically and allowed the relationship between the missingness indicator and the covariates totally unspecified. The curse of dimensionality of multivariate nonparametric estimation puts a restriction on the number of covariates involved in the missingness mechanism.

In this chapter, we assume that all the covariates are fully observed and the scalar response is subject to nonignorable missingness. We consider a semiparametric exponential tilting propensity where the relationship between the missingness indicator and the response is totally unspecified and estimated nonparametrically, while the relationship between the missingness indicator and the covariates is modeled parametrically. Since the response is partially observed, it makes more sense to model the relationship between the missingness indicator and the response nonparametrically in the missingness mechanism. We also avoid multivariate kernel-type estimators. In other words, there is no restriction on the number of covariates that can be involved in the missingness mechanism practically.

In addition to robustness and flexibility, this new semiparametric missingness mechanism also brings new challenges methodologically and theoretically. The first challenge we face is the model identifiability. In fact, identifiability is a notorious issue for analyzing nonignorable missing data. Assumptions on the data-generating process or the missingness mechanism are needed to guarantee that the model is identifiable (Robins & Ritov 1997). Usually, the model identifiability conditions need to be investigated in a case-by-case fashion. Kott (2014), Shao & Zhao (2013), Wang et al. (2014), Zhao &

Shao (2015) introduced shadow variable/nonresponse instrument, which is conditionally independent of the missingness indicator given the response and the other covariates. Much more flexible missingness mechanism can be adopted if an appropriate shadow variable can be identified. For our missingness mechanism, even if we have identified the shadow variable properly, we still need to impose assumptions on the data generating process. Details are presented in Section 5.3. The partially observed response also causes difficulties in the estimation of nuisance functions and consequently in the establishment of asymptotic properties. More details can be found in Subsection 5.4.3.

5.2 Model Setup

Consider N subjects with $(\mathbf{x}_i, r_i y_i, r_i)$, $i = 1, \ldots, N$, which are independent and identically distributed realizations from $(\mathbf{X}, RY, R)$, where $\mathbf{X}$ is a p_0-dimensional fully observed covariate, R is a binary missingness indicator, and the scalar response Y is observed when the indicator $R = 1$. We decompose $\mathbf{X}$ as $(\mathbf{U}^{\mathrm{T}}, \mathbf{Z}^{\mathrm{T}})$, where $\mathbf{U}$ is q_0-dimensional and $\mathbf{Z}$ is $(p_0 - q_0)$-dimensional. We assume that Y is conditionally independent of $\mathbf{U}$ given $\mathbf{Z}$ and $R = 1$; R is conditionally independent of $\mathbf{Z}$ given Y and $\mathbf{U}$. That is,

$$f_{Y|\mathbf{X},R=1}(y, \mathbf{x}) = f_{Y|\mathbf{Z},R=1}(y, \mathbf{z}) \text{ and } \mathrm{pr}(R = 1 \mid Y, \mathbf{X}) = \mathrm{pr}(R = 1 \mid Y, \mathbf{U}),$$

where $f_{Y|\mathbf{X},R=1}(y, \mathbf{x})$ is the conditional probability density function (pdf) or probability mass function (pmf) of Y given $\mathbf{X}$ and $R = 1$, and $f_{Y|\mathbf{Z},R=1}(y, \mathbf{z})$ is the conditional pdf/pmf of Y given $\mathbf{Z}$ and $R = 1$.

Instead of modeling the data-generating process $f_{Y|\mathbf{X}}(y, \mathbf{x})$ that suffers from missing responses, we model $f_{Y|\mathbf{Z},R=1}(y, \mathbf{z})$ parametrically and assume semiparametric exponential tilting propensity. Specifically, we focus on the model

$$f_{Y|\mathbf{Z},R=1}(y, \mathbf{z}; \boldsymbol{\alpha}) = \exp\{y\eta(\mathbf{z}; \boldsymbol{\alpha}) + \rho(\mathbf{z}; \boldsymbol{\alpha}) + \tau(y; \boldsymbol{\alpha})\}, \tag{5.1}$$

with missingness mechanism

$$\mathrm{pr}(R = 1 \mid y, \mathbf{u}) = \pi(y, \mathbf{u}; \boldsymbol{\beta}, g) = \mathrm{expit}\{g(y) + h(\mathbf{u}, \boldsymbol{\beta})\}, \tag{5.2}$$

where $\boldsymbol{\alpha}$ is a $(p - q)$-dimensional unknown parameter, functions η, ρ and τ are known, $\boldsymbol{\beta}$ is a q-dimensional unknown parameter, function h is known, and function g is totally unspecified. For identifiability reason, we assume $f_{Y|\mathbf{Z},R=1}(y, \mathbf{z}; \boldsymbol{\alpha})$ belongs to the

exponential family, and assume $g(0) = 0$.

This model is also motivated by a real data example. The data set consists of around 1,300 ICU patients, and 30% have a missing serum albumin level (Y) while all other covariates (X) are fully observed. The covariates are mainly from two categories: demographics, such as age and gender, and lab tests, such as calcium level, red blood cell counts and white blood cell counts. The reason for missing Y is unknown, but one possible reason is the limit of detection issue; it may also because of the mistake of the nurses, or might be relevant to other clinical biomarkers in the same lab test. We want to argue that, the missingness should not depend on the demographics. Once the patient is in the ICU, the nurse should have the patient blood drawn and have all biomarkers tested no matter the patient's age, gender, etc. On the other hand, for the albumin-complete patients, we argue that the serum albumin level largely depends on the demographics supported by clinical literature, but not the other biomarkers.

Note that

$$f_{Y|\mathbf{X}}(y,\mathbf{x}) = \frac{f_{Y|\mathbf{X},R=1}(y,\mathbf{x})/\pi(y,\mathbf{u};\boldsymbol{\beta},g)}{\int f_{Y|\mathbf{X},R=1}(t,\mathbf{x})/\pi(t,\mathbf{u};\boldsymbol{\beta},g)dt} = \frac{f_{Y|\mathbf{Z},R=1}(y,\mathbf{z};\boldsymbol{\alpha})/\pi(y,\mathbf{u};\boldsymbol{\beta},g)}{\int f_{Y|\mathbf{Z},R=1}(t,\mathbf{z};\boldsymbol{\alpha})/\pi(t,\mathbf{u};\boldsymbol{\beta},g)dt}. \quad (5.3)$$

Then under model (5.1) and (5.2), the likelihood function of one observation, which is the joint pdf/pmf of (X, RY, R), is given as

$$\begin{aligned} f_{\mathbf{X},RY,R}(\mathbf{x},ry,r) &= f_{\mathbf{X}}(\mathbf{x})\{\pi(y,\mathbf{u};\boldsymbol{\beta},g)f_{Y|\mathbf{X}}(y,\mathbf{x})\}^r \left\{\int\{1-\pi(t,\mathbf{u};\boldsymbol{\beta},g)\}f_{Y|\mathbf{X}}(t,\mathbf{x})dt\right\}^{1-r} \\ &= f_{\mathbf{X}}(\mathbf{x})w(\mathbf{x};\boldsymbol{\beta},\boldsymbol{\alpha},g)f_{Y|\mathbf{Z},R=1}(y,\mathbf{z};\boldsymbol{\alpha})^r \left\{\frac{1-w(\mathbf{x};\boldsymbol{\alpha},\boldsymbol{\beta},g)}{w(\mathbf{x};\boldsymbol{\alpha},\boldsymbol{\beta},g)}\right\}^{1-r}, \end{aligned}$$

where $f_{\mathbf{X}}(\mathbf{x})$ is the pdf/pmf of $\mathbf{x}$ and

$$\begin{aligned} w(\mathbf{x};\boldsymbol{\alpha},\boldsymbol{\beta},g) &\equiv \mathrm{pr}(R=1 \mid \mathbf{X}=\mathbf{x}) \\ &= E\{\pi(Y,\mathbf{u};\boldsymbol{\beta},g) \mid \mathbf{X}=\mathbf{x}\} \\ &= \frac{1}{1+\exp\{-h(\mathbf{u};\boldsymbol{\beta})\}E[\exp\{-g(Y)\} \mid \mathbf{z},R=1]}. \end{aligned}$$

We treat $\boldsymbol{\theta} \equiv (\boldsymbol{\alpha}^{\mathrm{T}},\boldsymbol{\beta}^{\mathrm{T}})^{\mathrm{T}}$ as the parameter of interest, and treat $f_{\mathbf{X}}(\mathbf{x})$ and $g(y)$ as nuisance functions. For notation simplicity, in the following, we write $w(\mathbf{x};\boldsymbol{\theta},g)$ as $w(\mathbf{x})$ and $E(\cdot \mid \mathbf{z},R=1)$ as $E(\cdot \mid \mathbf{z},1)$.

5.3 Identifiability

Since $\mathbf{X}$ is fully observed, then $f_{\mathbf{X}}(\mathbf{x})$ is identifiable. Since when $R = 1$, both $\mathbf{X}$ and Y are fully observed, then $\boldsymbol{\alpha}$ is identifiable. We need to show that $\boldsymbol{\beta}$ and $g(y)$ are identifiable. Note that $w(\mathbf{x})$ is identifiable and

$$w(\mathbf{x})^{-1} = 1 + \exp\{-h(\mathbf{u};\boldsymbol{\beta})\}E[\exp\{-g(Y)\} \mid \mathbf{z}, 1].$$

We only need to show that for any $\mathbf{x}$, if

$$\exp\{-h(\mathbf{u};\boldsymbol{\beta})\}E[\exp\{-g(Y)\} \mid \mathbf{z}, 1] = \exp\{-h(\mathbf{u};\widetilde{\boldsymbol{\beta}})\}E[\exp\{-\widetilde{g}(Y)\} \mid \mathbf{z}, 1],$$

then $\boldsymbol{\beta} = \widetilde{\boldsymbol{\beta}}$ and $g(\cdot) = \widetilde{g}(\cdot)$. There exists a constant c such that

$$\frac{\exp\{-h(\mathbf{u};\boldsymbol{\beta})\}}{\exp\{-h(\mathbf{u};\widetilde{\boldsymbol{\beta}})\}} = \frac{E[\exp\{-\widetilde{g}(Y)\} \mid \mathbf{z}, 1]}{E[\exp\{-g(Y)\} \mid \mathbf{z}, 1]} = c, \quad \forall \mathbf{u}, \mathbf{z},$$

because the left hand side is a function of $\mathbf{u}$ only while the right hand side is a function of $\mathbf{z}$ only. Then $h(\mathbf{u};\boldsymbol{\beta}) = h(\mathbf{u};\widetilde{\boldsymbol{\beta}}) - \log(c)$ for all $\mathbf{u}$, and $\widetilde{g}(y) = g(y) - \log(c)$ for all y due to the invertibility of the Laplace transform. Taking into account the requirement that $g(0) = 0$, we obtain $c = 0$. Hence, $\boldsymbol{\beta}$ and g are also identifiable.

5.4 Methodology

In this section, we will control the effect of estimation of nuisance functions on the estimation of the parameters of interest using a semiparametric treatment, and construct regular asymptotically linear estimators for $\boldsymbol{\theta}$.

5.4.1 Nuisance Tangent Space and Its Orthogonal Complement

Consider the Hilbert space $\mathcal{H}$ of all p-dimensional zero-mean measurable functions of the observed data with finite variance, equipped with the inner product

$$\langle h_1, h_2 \rangle = E\{h_1^{\mathrm{T}}(\mathbf{X}, RY, R)h_2(\mathbf{X}, RY, R)\}$$

where h_1, $h_2 \in \mathcal{H}$. Nuisance tangent space is defined as the mean squared closure of the nuisance tangent spaces of parametric sub-models spanned by the nuisance score vectors.

By simple calculations, we can show that the nuisance tangent space for $f_{\mathbf{X}}(\mathbf{x})$ is

$$\Lambda_f = [\mathbf{a}(\mathbf{x}) \in \mathbb{R}^p : E\{\mathbf{a}(\mathbf{X})\} = \mathbf{0}].$$

The nuisance tangent space for $g(y)$ is

$$\begin{aligned}\Lambda_g &= \left(E[\mathbf{a}(Y)\{1-\pi(Y,\mathbf{u};\boldsymbol{\beta},g)\} \mid \mathbf{x}]\frac{r-w(\mathbf{x})}{1-w(\mathbf{x})} : \mathbf{a}(y) \in \mathbb{R}^p\right) \\ &= \left(E[\mathbf{a}(Y)\{\pi^{-1}(Y,\mathbf{u};\boldsymbol{\beta},g)-1\} \mid \mathbf{z},1]\frac{w(\mathbf{x})\{r-w(\mathbf{x})\}}{1-w(\mathbf{x})} : \mathbf{a}(y) \in \mathbb{R}^p\right).\end{aligned}$$

Since

$$E\left\{\frac{R-w(\mathbf{x})}{1-w(\mathbf{x})}\middle|\mathbf{x}\right\} = 0,$$

we can easily verify that $\Lambda_f \perp \Lambda_g$. Thus, the nuisance tangent space is $\Lambda = \Lambda_f \oplus \Lambda_g$.

Note that any function $\mathbf{d}(\mathbf{x}, ry, r)$ in $\mathcal{H}$ can be written as

$$\mathbf{d}(\mathbf{x}, ry, r) = r\mathbf{d}_1(y,\mathbf{x}) + (1-r)\mathbf{d}_0(\mathbf{x}).$$

It is easy to show that the orthogonal complement of Λ_f is

$$\begin{aligned}\Lambda_f^{\perp} &= [\mathbf{d}(\mathbf{x}, ry, r) \in \mathbb{R}^p : E\{\mathbf{d}(\mathbf{x}, RY, R) \mid \mathbf{x}\} = \mathbf{0}] \\ &= \left[\mathbf{d}(\mathbf{x}, ry, r) : \mathbf{d}_0(\mathbf{x}) = \frac{-w(\mathbf{x})}{1-w(\mathbf{x})}E\{\mathbf{d}_1(Y,\mathbf{x}) \mid \mathbf{x}, 1\}\right].\end{aligned}$$

Since any function $\mathbf{d}(\mathbf{x}, ry, r) \in \Lambda_g^{\perp}$ satisfies

$$\begin{aligned}\mathbf{0} &= E\left([E\{\mathbf{d}_1(Y,\mathbf{X}) \mid \mathbf{Z}, 1\} - \mathbf{d}_0(\mathbf{Z})]^{\mathrm{T}}E[\mathbf{a}(Y)\{1-\pi(Y,\mathbf{U};\boldsymbol{\beta},g)\} \mid \mathbf{X}]w(\mathbf{X})\right) \\ &= E\left\{E\left([E\{\mathbf{d}_1(Y,\mathbf{X}) \mid \mathbf{Z}, 1\} - \mathbf{d}_0(\mathbf{X})]^{\mathrm{T}}w(\mathbf{X})\{1-\pi(Y,\mathbf{U};\boldsymbol{\beta},g)\} \mid Y\right)\mathbf{a}(Y)\right\},\end{aligned}$$

for any $\mathbf{a}(y) \in \mathbb{R}^p$, then

$$\Lambda_g^{\perp} = \{\mathbf{d}(\mathbf{x}, ry, r) : E([E\{\mathbf{d}_1(Y,\mathbf{X}) \mid \mathbf{X}, 1\} - \mathbf{d}_0(\mathbf{X})]w(\mathbf{X})\{1-\pi(y,\mathbf{U};\boldsymbol{\beta},g)\} \mid y) = \mathbf{0}\}.$$

The orthogonal complement of the nuisance tangent space is

$$\Lambda^{\perp} = \Lambda_f^{\perp} \cap \Lambda_g^{\perp}$$

$$= \left(\mathbf{d}(\mathbf{x}, ry, r) : \mathbf{d}_0(\mathbf{x}) = \frac{-E\{\mathbf{d}_1(Y, \mathbf{x}) \mid \mathbf{z}, 1\}}{w^{-1}(\mathbf{x}) - 1}, \right.$$
$$\left. E\left[E\{\mathbf{d}_1(Y, \mathbf{X}) \mid \mathbf{Z}, 1\} \frac{w(\mathbf{X})}{1 - w(\mathbf{X})} \{1 - \pi(y, \mathbf{U}; \boldsymbol{\beta}, g)\} \mid y \right] = \mathbf{0} \right).$$

5.4.2 Efficient Score Functions

Let $\mathbf{S}(y, \mathbf{z}; \boldsymbol{\alpha}) \equiv \partial \log f_{Y|\mathbf{Z},R=1}(y, \mathbf{z}; \boldsymbol{\alpha}) / \partial \boldsymbol{\alpha}$. The score functions for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are

$$\mathbf{S}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}, g; \mathbf{x}, ry, r) = r\mathbf{S}(y, \mathbf{z}; \boldsymbol{\alpha}) - E\left\{\mathbf{S}(Y, \mathbf{z}; \boldsymbol{\alpha}) \mid \mathbf{x}\right\} \frac{r - w(\mathbf{x})}{1 - w(\mathbf{x})},$$

and

$$\mathbf{S}_{\beta}(\boldsymbol{\theta}, g; \mathbf{x}, ry, r) = \mathbf{h}'_{\beta}(\mathbf{u}; \boldsymbol{\beta})\{r - w(\mathbf{x})\},$$

respectively. Here $\mathbf{h}'_{\beta}(\mathbf{u}; \boldsymbol{\beta})$ is the derivative of $h(\mathbf{u}; \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. Detailed derivations for $\mathbf{S}_{\boldsymbol{\alpha}}$ and $\mathbf{S}_{\beta}$ are given in the Appendix C.1. Note that $E\{\mathbf{S}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}, g; \mathbf{x}, RY, R) \mid \mathbf{x}\} = \mathbf{0}$ and $E\{\mathbf{S}_{\beta}(\boldsymbol{\theta}, g; \mathbf{x}, RY, R) \mid \mathbf{x}\} = \mathbf{0}$. Thus, $\mathbf{G}_1\mathbf{S}_{\boldsymbol{\alpha}}(\boldsymbol{\theta}, g; \mathbf{x}, ry, r) \in \Lambda_g \oplus \Lambda^{\perp}$ for any $p \times (p - q)$ constant matrix $\mathbf{G}_1$, and $\mathbf{G}_2\mathbf{S}_{\beta}(\boldsymbol{\theta}, g; \mathbf{x}, ry, r) \in \Lambda_g \oplus \Lambda^{\perp}$ for any $p \times q$ constant matrix $\mathbf{G}_2$.

Define $\mathbf{S}_{\boldsymbol{\alpha},\text{eff}}(\boldsymbol{\theta}, g; \mathbf{x}, ry, r)$ and $\mathbf{S}_{\boldsymbol{\beta},\text{eff}}(\boldsymbol{\theta}, g; \mathbf{x}, ry, r)$ as the projections of $\mathbf{S}_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}, \boldsymbol{\beta}, g; \mathbf{x}, ry, r)$ and $\mathbf{S}_{\boldsymbol{\beta}}(\boldsymbol{\alpha}, \boldsymbol{\beta}, g; \mathbf{x}, ry, r)$ onto the space $\Lambda^{\perp}$, respectively. By simple calculations, we obtain

$$\begin{aligned}
& \mathbf{S}_{\boldsymbol{\alpha},\text{eff}}(\boldsymbol{\alpha}, \boldsymbol{\beta}, g; \mathbf{x}, ry, r) \\
= \; & r\mathbf{S}(y, \mathbf{z}; \boldsymbol{\alpha}) - \frac{r - w(\mathbf{x})}{1 - w(\mathbf{x})} \left(E\left\{\mathbf{S}(Y, \mathbf{z}; \boldsymbol{\alpha}) \mid \mathbf{x}\right\} + E[\mathbf{a}_0(Y)\{1 - \pi(Y, \mathbf{u}; \boldsymbol{\beta}, g)\} \mid \mathbf{x}] \right) \\
= \; & r\mathbf{S}(y, \mathbf{z}; \boldsymbol{\alpha}) \\
- \; & \frac{r - w(\mathbf{x})}{E[\exp\{-g(Y)\} \mid \mathbf{z}, 1]} \left(E[\mathbf{S}(Y, \mathbf{z}; \boldsymbol{\alpha}) \exp\{-g(Y)\} \mid \mathbf{z}, 1] + E[\mathbf{a}_0(y) \exp\{-g(Y)\} \mid \mathbf{z}, 1] \right),
\end{aligned}$$

where $\mathbf{a}_0(y)$ satisfies

$$\begin{aligned}
& E\left(E[\mathbf{a}_0(Y)\{1 - \pi(Y, \mathbf{U}; \boldsymbol{\beta}, g)\} \mid \mathbf{X}] \frac{w(\mathbf{X})\{1 - \pi(y, \mathbf{U}; \boldsymbol{\beta}, g)\}}{1 - w(\mathbf{X})} \mid y \right) \\
= \; & -E\left[E\left\{\mathbf{S}(Y, \mathbf{Z}; \boldsymbol{\alpha}) \mid \mathbf{X}\right\} \frac{w(\mathbf{X})\{1 - \pi(y, \mathbf{U}; \boldsymbol{\beta}, g)\}}{1 - w(\mathbf{X})} \mid y \right]. \qquad (5.4)
\end{aligned}$$

The efficient score for β is

$$\begin{aligned} & \mathbf{S}_{\beta,\text{eff}}(\boldsymbol{\alpha},\boldsymbol{\beta},g;\mathbf{x},ry,r) \\ = & \frac{r-w(\mathbf{x})}{1-w(\mathbf{x})}(\{1-w(\mathbf{x})\}\mathbf{h}'_{\boldsymbol{\beta}}(\mathbf{u};\boldsymbol{\beta}) - E[\mathbf{a}_1(Y)\{1-\pi(Y,\mathbf{u};\boldsymbol{\beta},g)\} \mid \mathbf{x}]) \\ = & \{r-w(\mathbf{x})\}\mathbf{h}'_{\boldsymbol{\beta}}(\mathbf{u};\boldsymbol{\beta}) - \frac{r-w(\mathbf{x})}{E[\exp\{-g(Y)\} \mid \mathbf{z},1]}E[\mathbf{a}_1(Y)\exp\{-g(Y)\} \mid \mathbf{z},1], \end{aligned}$$

where $\mathbf{a}_1(y)$ satisfies

$$\begin{aligned} & E\left(E[\mathbf{a}_1(Y)\{1-\pi(Y,\mathbf{U};\boldsymbol{\beta},g)\} \mid \mathbf{X}]\frac{w(\mathbf{X})\{1-\pi(y,\mathbf{U};\boldsymbol{\beta},g)\}}{1-w(\mathbf{X})} \mid y\right) \\ = & E[w(\mathbf{X})\{1-\pi(y,\mathbf{U};\boldsymbol{\beta},g)\}\mathbf{h}'_{\boldsymbol{\beta}}(\mathbf{U};\boldsymbol{\beta}) \mid y]. \end{aligned} \tag{5.5}$$

5.4.3 Estimation of Nuisance Functions

To construct estimating equations based on efficent score functions $\mathbf{S}_{\boldsymbol{\alpha},\text{eff}}$ and $\mathbf{S}_{\boldsymbol{\beta},\text{eff}}$, we need to deal with the two unknown nuisance functions $f_{\mathbf{X}}(\mathbf{x})$ and $g(y)$.

For $f_{\mathbf{X}}(\mathbf{x})$, due to the key observations

$$E\{R-w(\mathbf{x}) \mid \mathbf{x}\} = 0 \text{ and } E\{R\mathbf{S}(Y,\mathbf{z};\boldsymbol{\alpha}) \mid \mathbf{x}\} = \mathbf{0},$$

any working model $f^*_{\mathbf{X}}(\mathbf{x})$ can be used to construct estimating equations and the mean-zero property retains. One drawback of employing working model is the sacrifice of estimation efficiency. Alternatively, we can estimate it using $\mathbf{x}_i$'s. Note that all the covariates are fully observerd.

For $g(y)$, we propose to estimate it using local constant approximation. That is, at each y, we employ γ to replace $g(y)$. For any fixed y_0, we have

$$\pi(y_0,\mathbf{u};\boldsymbol{\beta},\gamma) \equiv \text{pr}(R=1 \mid \mathbf{X}=\mathbf{x}, Y=y_0) = \text{expit}\{\gamma + h(\mathbf{u};\boldsymbol{\beta})\},$$

and

$$w(\mathbf{x},y_0) \equiv \text{pr}(R=1 \mid \mathbf{X}=\mathbf{x}) = \text{expit}\{\gamma + h(\mathbf{u};\boldsymbol{\beta})\}.$$

The score function for γ is given as

$$S_\gamma(\boldsymbol{\beta},\boldsymbol{\gamma};\mathbf{x},r,ry_0) \;\; = \;\; r - w(\mathbf{x},y_0)$$

$$= \quad r - \pi(y_0, \mathbf{u}; \boldsymbol{\beta}, \gamma).$$

One intuitive way to estimate γ, i.e., $g(y_0)$ is to solve the following estimating equation

$$\frac{1}{N}\sum_{i=1}^{N}\{r_i - \pi(y_0, \mathbf{u}; \boldsymbol{\beta}, \gamma)\}K_h(y_i - y_0) = 0,$$

which is equivalent to

$$N^{-1}\sum_{i=1}^{n}\{1 - \pi(y_0, \mathbf{u}_i; \boldsymbol{\beta}, \gamma)\}K_h(y_i - y_0) = N^{-1}\sum_{i=n+1}^{N}\pi(y_0, \mathbf{u}_i; \boldsymbol{\beta}, \gamma)K_h(y_i - y_0). \quad (5.6)$$

However, y_i is missing when $i = n+1, \ldots, N$. We use $f_{\mathbf{X}|R}(\mathbf{x}, r)$ to denote the conditional pdf/pmf of $\mathbf{X}$ given R and the right-hand-side of (5.6) can be approximated by

$$\begin{aligned}
& \mathrm{pr}(R=0)E\left\{\pi(y_0, \mathbf{U}; \boldsymbol{\beta}, \gamma)K_h(Y - y_0) \mid R = 0\right\} \\
= \quad & \mathrm{pr}(R=0)E\left[\pi(y_0, \mathbf{U}; \boldsymbol{\beta}, \gamma)E\left\{K_h(Y - y_0) \mid \mathbf{X}, R = 0\right\} \mid R = 0\right] \\
= \quad & \mathrm{pr}(R=0)\int \pi(y_0, \mathbf{u}; \boldsymbol{\beta}, \gamma)\frac{E[K_h(Y - y_0)\exp\{-g(Y)\} \mid \mathbf{z}, 1]}{E[\exp\{-g(Y)\} \mid \mathbf{z}, 1]}f_{\mathbf{X}|R}(\mathbf{x}, 0)d\mathbf{x} \\
\approx \quad & N^{-1}\sum_{i=n+1}^{N}\frac{E[K_h(Y - y_0)\exp\{-g(Y)\} \mid \mathbf{z}_i, 1]}{E[\exp\{-g(Y)\} \mid \mathbf{z}_i, 1][1 + \exp\{-\gamma - h(\mathbf{u}_i, \boldsymbol{\beta})\}]}.
\end{aligned}$$

Hence, the approximate estimating equation for γ is

$$\begin{aligned}
& N^{-1}\sum_{i=1}^{n}\frac{1}{1 + \exp\{\gamma + h(\mathbf{u}_i; \boldsymbol{\beta})\}}K_h(y_i - y_0) \\
= \quad & N^{-1}\sum_{i=n+1}^{N}\frac{E[K_h(Y - y_0)\exp\{-g(Y)\} \mid \mathbf{z}_i, 1]}{E[\exp\{-g(Y)\} \mid \mathbf{z}_i, 1][1 + \exp\{-\gamma - h(\mathbf{u}_i, \boldsymbol{\beta})\}]}. \quad (5.7)
\end{aligned}$$

We propose to estimate $g(y)$ on L distinct points $(d_1, \ldots, d_L)$ approximately evenly distributed in the range of Y. Then estimate other $g(y)$ values by polynomial interpolation with degree $m-1$. Note that L goes to infinity when N goes to infinity, but $L \ll N$. Let $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_L)^{\mathrm{T}} = \{g(d_1), \ldots, g(d_L)\}^{\mathrm{T}}$, and we use $\widehat{g}(y; \boldsymbol{\gamma})$ to denote the approximate function. Employing the idea of profiling, at any $\boldsymbol{\theta}$, the L-dimensional vector $\boldsymbol{\gamma}(\boldsymbol{\theta})$ can be solved from the approximate estimating equation set that consists of L equations

$$N^{-1}\sum_{i=1}^{N}\mathbf{S}_g\{\boldsymbol{\theta}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta}); \mathbf{x}_i, r_i y_i, r_i\} = \mathbf{0}, \quad (5.8)$$

where the l-th component of $\mathbf{S}_g\{\boldsymbol{\theta}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta}); \mathbf{x}_i, r_i y_i, r_i\}$ is

$$\frac{r_i K_h(y_i - d_l)}{1 + \exp\{\gamma_l + h(\mathbf{u}_i; \boldsymbol{\beta})\}} + \frac{(r_i - 1)E[K_h(Y - y_0)\exp\{-g(Y)\} \mid \mathbf{z}_i, 1]}{E[\exp\{-\widehat{g}(Y; \boldsymbol{\gamma})\} \mid \mathbf{z}_i, 1][1 + \exp\{-\gamma_l - h(\mathbf{u}_i, \boldsymbol{\beta})\}]}.$$

5.5 Implementation and Algorithm

Every time we update $\boldsymbol{\theta}$, we first need to solve for functions $\mathbf{a}_0(y)$ and $\mathbf{a}_1(y)$ from integral equations (5.4) and (5.5), respectively. Note that $E[a(\mathbf{X}, Y)\{1 - \pi(Y, \mathbf{U}; \boldsymbol{\beta}, g)\} \mid Y] = 0$ is equivalent to $E[a(\mathbf{X}, Y)\exp\{-h(\mathbf{U}; \boldsymbol{\beta})\} \mid Y, 1] = 0$, for any function $a(\mathbf{x}, y)$. We also have

$$f_{\mathbf{X}|R=1}(\mathbf{x}, 1) = \frac{f_{\mathbf{X}}(\mathbf{x})w(\mathbf{x})}{\mathrm{pr}(R = 1)},$$

and the conditional pdf/pmf of $\mathbf{X}$ given Y and $R = 1$

$$\begin{aligned} f_{\mathbf{X}|Y,R=1}(\mathbf{x}, y) &= \frac{f_{Y|\mathbf{Z},R=1}(y, \mathbf{z}; \boldsymbol{\alpha}) f_{\mathbf{X}|R=1}(\mathbf{x})}{f_{Y|R=1}(y)} \\ &= \frac{f_{Y|\mathbf{Z},R=1}(y, \mathbf{z}; \boldsymbol{\alpha}) f_{\mathbf{X}}(\mathbf{x}) w(\mathbf{x})}{\int f_{Y|\mathbf{Z},R=1}(y, \mathbf{z}; \boldsymbol{\alpha}) f_{\mathbf{X}}(\mathbf{x}) w(\mathbf{x}) d\mathbf{x}}. \end{aligned}$$

Then (5.4) can be written as

$$\mathcal{L}(\mathbf{a}_0, f_{\mathbf{X}}; \boldsymbol{\theta}, g)(y) = \boldsymbol{\phi}_0(f_{\mathbf{X}}; \boldsymbol{\theta}, g)(y), \tag{5.9}$$

where $\mathcal{L}$ is a bilinear operator defined as

$$\mathcal{L}(\mathbf{a}_0, f_{\mathbf{X}}; \boldsymbol{\theta}, g)(y) = \iint \frac{w^2(\mathbf{x}) f_{Y|\mathbf{Z},R=1}(t, \mathbf{z}; \boldsymbol{\alpha}) f_{Y|\mathbf{Z},R=1}(y, \mathbf{z}; \boldsymbol{\alpha})}{\exp\{h(\mathbf{u}; \boldsymbol{\beta})\} E[\exp\{-g(Y)\} \mid \mathbf{z}, 1]} \exp\{-g(t)\} \mathbf{a}_0(t) f_{\mathbf{X}}(\mathbf{x}) dt d\mathbf{x},$$

and

$$\begin{aligned} & \boldsymbol{\phi}_0(f_{\mathbf{X}}; \boldsymbol{\theta}, g)(y) \\ = & -\int E[\mathbf{S}(Y, \mathbf{z}; \boldsymbol{\alpha}) \exp\{-g(Y)\} \mid \mathbf{z}, 1] \frac{w^2(\mathbf{x}) f_{Y|\mathbf{Z},R=1}(y, \mathbf{z}; \boldsymbol{\alpha})}{\exp\{h(\mathbf{u}; \boldsymbol{\beta})\} E[\exp\{-g(Y)\} \mid \mathbf{z}, 1]} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}. \end{aligned}$$

Similarly, (5.5) can be written as

$$\mathcal{L}(\mathbf{a}_1, f_{\mathbf{X}}; \boldsymbol{\theta}, g)(y) = \boldsymbol{\phi}_1(f_{\mathbf{X}}; \boldsymbol{\theta}, g)(y), \tag{5.10}$$

where

$$\boldsymbol{\phi}_1(f_{\mathbf{X}};\boldsymbol{\theta},g)(y)=\int\frac{\mathbf{h}'_{\boldsymbol{\beta}}(\mathbf{u};\boldsymbol{\beta})w^2(\mathbf{x})}{\exp\{h(\mathbf{u};\boldsymbol{\beta})\}}f_{Y|\mathbf{Z},R=1}(y,\mathbf{z};\boldsymbol{\alpha})f_{\mathbf{X}}(\mathbf{x})d\mathbf{x}.$$

Note that (5.9) and (5.10) are Fredholm integral equations of the first kind, then $\mathbf{a}_0(y)$ and $\mathbf{a}_1(y)$ can be obtained by solving two linear systems, respectively.

The procedure of estimating $\boldsymbol{\theta}$ is given as follows.

Algorithm

1: Posit a working model for $f_{\mathbf{X}}$, and denote it as $\widetilde{f}_{\mathbf{X}}$. It can be an arbitrary working model for $f_{\mathbf{X}}$, denoted as $f^*_{\mathbf{X}}$, or it can be the empirical $f_{\mathbf{X}}$, denoted as $\widehat{f}_{\mathbf{X}}$. Note that the support of the working model should cover the true support for $\mathbf{x}$.

2: At any $\boldsymbol{\theta}$, solve the approximate estimating equation set (5.8) to obtain $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})=\{\widehat{g}(d_1;\boldsymbol{\theta}),\ldots\widehat{g}(d_L;\boldsymbol{\theta})\}^{\mathrm{T}}$.

3: At the same $\boldsymbol{\theta}$, solve for the functions $\widetilde{\mathbf{a}}_0(y)$ and $\widetilde{\mathbf{a}}_1(y)$ from the following integral equations

$$\mathcal{L}(\widetilde{\mathbf{a}}_0,\widetilde{f}_{\mathbf{X}};\boldsymbol{\theta},\widehat{g})(y)=\boldsymbol{\phi}_0(\widetilde{f}_{\mathbf{X}};\boldsymbol{\theta},\widehat{g})(y),$$

and

$$\mathcal{L}(\widetilde{\mathbf{a}}_1,\widetilde{f}_{\mathbf{X}};\boldsymbol{\theta},\widehat{g})(y)=\boldsymbol{\phi}_1(\widetilde{f}_{\mathbf{X}};\boldsymbol{\theta},\widehat{g})(y),$$

respectively.

4: Let $\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}=(\mathbf{S}^{\mathrm{T}}_{\boldsymbol{\alpha},\mathrm{eff}},\mathbf{S}^{\mathrm{T}}_{\boldsymbol{\beta},\mathrm{eff}})^{\mathrm{T}}$. With $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})$, $\widetilde{\mathbf{a}}_0(y)$ and $\widetilde{\mathbf{a}}_1(y)$, solve the estimating equation set

$$\sum_{i=1}^{N}\widetilde{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}[\boldsymbol{\theta},\widehat{g}\{\cdot;\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})\};\mathbf{x}_i,r_iy_i,r_i]=\mathbf{0} \tag{5.11}$$

for $\widehat{\boldsymbol{\theta}}^*$ (using $f^*_{\mathbf{X}}$) or $\widehat{\boldsymbol{\theta}}$ (using $\widehat{f}_{\mathbf{X}}$).

Here $\widetilde{\mathbf{S}}_{\boldsymbol{\theta},\mathrm{eff}}[\boldsymbol{\theta},\widehat{g}\{\cdot;\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})\};\mathbf{x}_i,r_iy_i,r_i]$ is $\mathbf{S}_{\boldsymbol{\theta},\mathrm{eff}}[\boldsymbol{\theta},\widehat{g}\{\cdot;\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})\};\mathbf{x}_i,r_iy_i,r_i]$ with $\mathbf{a}_0(y)$ and $\mathbf{a}_1(y)$ replaced by $\widetilde{\mathbf{a}}_0(y)$ and $\widetilde{\mathbf{a}}_1(y)$, respectively.

5.6 Theoretical Properties

In this subsection, we establish the asymptotic properties of $\widehat{\boldsymbol{\theta}}$, which is constructed with $\widehat{f}_{\mathbf{X}}$. To facilitate the proof, we assume that $\widehat{f}_{\mathbf{X}}$ is independent of the data used to estimate $\boldsymbol{\theta}$ and $g(y)$. Sample splitting can be used to create the independence. Specifically, we randomly split the data into two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$, where $\mathcal{D}_k$ contains N_k observations and n_k out of N_k observations' responses are not missing. For simplicity, assume $n_k/N_K = n/N$, $k = 1, 2$. We use $\mathcal{I}_k$ to denote the indices of observations in the k-th subset. Subset $\mathcal{D}_1$ is used to estimate $f_{\mathbf{X}}(\mathbf{x})$, while $\mathcal{D}_2$ is used to estimate $\boldsymbol{\theta}$ and $g(y)$.

We further need the following regularity conditions.

(C1) Let $N_1 = CN^{\delta}$, where $\delta \in [0.5, 1)$ and C is a positive constant. Then $N_2 = N - CN^{\delta}$.

(C2) The degree of the polynomial interpolation is $m - 1$, where $m \geqslant 2$.

(C3) Let $L = ch^{-a}$, where c is a positive constant and $0 < a < 2$. The bandwidth $h = o(1)$ and satisfies $Nh^{4a+2} \to \infty$, $Nh^{8-4a} \to 0$, $Nh^{4am-4a} \to 0$, $Nh^{2am} \to 0$.

(C4) Let $A_0(\mathbf{z}) = E[\mathbf{a}_0(y)\exp\{-g(y)\} \mid \mathbf{z}, 1]$, and $A_1(\mathbf{z}) = E[\mathbf{a}_1(y)\exp\{-g(y)\} \mid \mathbf{z}, 1]$, which are functionals of $f_{\mathbf{X}}$. Assume that the Fréchet derivative of $A_0(\mathbf{z})$ with respect to $f_{\mathbf{X}}$ and the the Fréchet derivative of $A_1(\mathbf{z})$ with respect to $f_{\mathbf{X}}$ are bounded.

Remark 5.6.1. *To guarantee that* (C1)*,* (C2) *and* (C3) *hold simultaneously, by simple calculations, we need* $m \geqslant 4$*, and*

$$\frac{1}{m-2} < a < \frac{3}{4}.$$

We can set $m = 4$ *in implementation, which means that cubic interpolation is employed.*

The asymptotic normality of $\widehat{\boldsymbol{\theta}}$ is stated in Theorem 5.6.1, and the detailed proof can be found in the Appendix C.2.

Theorem 5.6.1. *Assume that Conditions* (C1) - (C4) *hold. Let*

$$\mathbf{Q} \equiv E\left\{\frac{\partial \mathbf{S}_{\boldsymbol{\theta},\text{eff}}(\boldsymbol{\theta}_0, g; \mathbf{X}, RY, R)}{\partial \boldsymbol{\theta}_0^{\mathrm{T}}}\right\} = -var\{\mathbf{S}_{\boldsymbol{\theta},\text{eff}}(\boldsymbol{\theta}_0, g; \mathbf{X}, RY, R)\}.$$

Then

$$N_2^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \;\; = \;\; -\mathbf{Q}^{-1} N_2^{-1/2} \sum_{i \in \mathcal{I}_2} \mathbf{S}_{\boldsymbol{\theta},\text{eff}}(\boldsymbol{\theta}_0, g; \mathbf{x}_i, r_i y_i, r_i) + o_p(1).$$

Consequently, $N_2^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \rightsquigarrow \mathcal{N}(\mathbf{0}, \mathbf{V})$ *in distribution when* $N_2 \to \infty$*, where*

$$\mathbf{V} = \mathbf{Q}^{-1} var\{\mathbf{S}_{\boldsymbol{\theta},\text{eff}}(\boldsymbol{\theta}_0, g; \mathbf{X}, RY, R)\}(\mathbf{Q}^{-1})^{\mathrm{T}} = [E\{\mathbf{S}_{\boldsymbol{\theta},\text{eff}}(\boldsymbol{\theta}_0, g; \mathbf{X}, RY, R)^{\otimes 2}\}]^{-1}.$$

Remark 5.6.2. *Typically, estimating nuisance parameters will alter, often inflate, the variance of the estimator for the parameter of interest. However, in our construction, this is not the case. In other words, even if we knew the true functions* $g(y)$ *and* $f_{\mathbf{X}}(\mathbf{x})$ *and used them in the estimating equation (5.11), the variance of* $\widehat{\boldsymbol{\theta}}$ *would not be further reduced. Indeed, the asymptotic variance of our* $\widehat{\boldsymbol{\theta}}$ *achieves the optimal estimation variance bound, and thus, it is semiparametrically efficient.*

5.7 Simulation Studies

In this subsection, we investigate the finite sample performance of our efficient method. Scalar covariates U and Z are generated from the uniform distribution on $(0, 1)$ and the normal distribution $\mathcal{N}(U, 0.5^2)$, respectively. The conditional distribution of Y given Z and $R = 1$ is $\mathcal{N}(Z\alpha, 1)$. Nuisance function $g(y)$ is set to 3expit(y) and we propose to estimate it on 13 distinct points, i.e., the dimension of $\boldsymbol{\gamma}$ is 13. We set true $\alpha = 2$, true $\beta = -1$, and sample size $N = 1000$ and 2000. The nuisance function $f_{\mathbf{X}}$ is estimated using 1000 samples of (U, Z) which are independent of the data used to estimate $\boldsymbol{\theta}$ and $g(y)$. Data $(\mathbf{x}_i, r_i y_i, r_i)$, $i = 1, \ldots, N$ are generated using rejection sampling method, and around 20% of the responses are missing. For each setting, we run 1000 simulation studies.

The estimation and inference results of $\boldsymbol{\theta}$ are summarized in Table 5.1, and the estimation results of $g(y)$ are given in Figure 5.1. We can see that when the sample size is 1000, the estimation of $\boldsymbol{\theta}$ is satisfactory, but the inference results are not very good. One reason is that the approximation of the right-hand-side of (5.6) introduces errors of the order $O_p\{(N-n)^{-1/2}\}$. To some extent, the effective sample size of estimation of $g(y)$ is only $N - n$. In our simulation setting with $N = 1000$, $N - n$ is around 200, which can be too small to make asymptotic confidence intervals valid. With larger sample size $N = 2000$, the biases of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\gamma}}$ become smaller, the 90% band of $\widehat{g}(y)$ becomes

Table 5.1: Estimation and inference results of the parameter of interest $\boldsymbol{\theta}$

	N=1000		N=2000	
	α	β	α	β
Truth	2	-1	2	-1
Mean	2.0008	-0.9874	2.0006	-0.9926
Emp sd	0.0629	0.2269	0.0336	0.2235
Est sd	0.0446	0.3083	0.0314	0.2182
Emp cvg	85.8%	97.8%	94.6%	95.6%

In Table 5.1, "Emp sd" denotes the empirical standard deviation of 1000 estimates; "Est sd" denotes the mean of 1000 estimated asymptotic standard deviations; "Emp cvg" denotes empirical coverage of the estimated 95% CI.

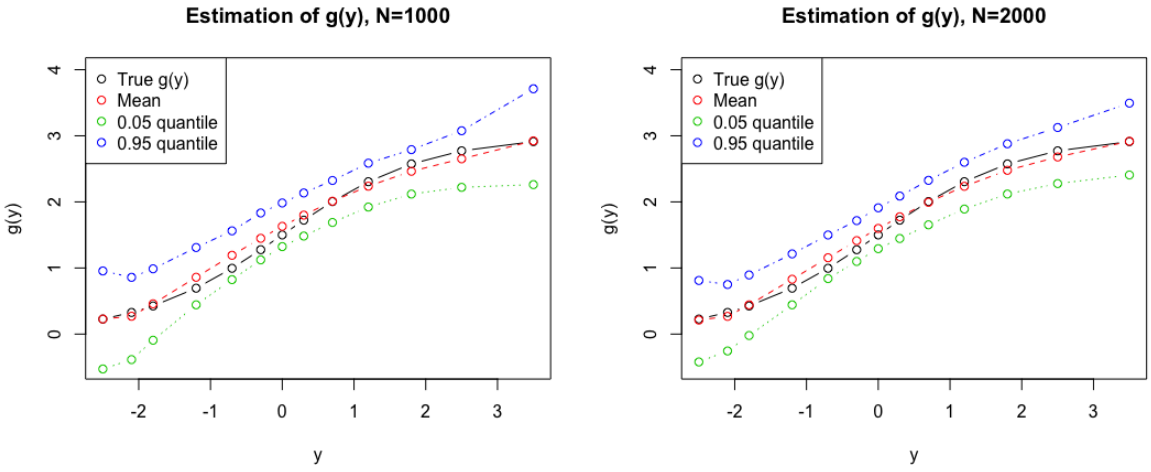

Figure 5.1: Performance of the local constant approximation of $g(y)$

narrower, and consequently the inference results become much better. Specifically, the mean of estimated standard deviations closely approximates the empirical one, and the empirical coverage of the estimated 95% confidence intervals is close to the nominal level.

Appendix A
Technical Proofs for Chapter 3

A.1 Proof of the Identifiability of the Linear Measurement Error Model

Computing the first and second order moments of Y, W_1 and W_2 given Z, we have

$$\begin{aligned}
\mathrm{E}(Y|\mathbf{Z}) &= \beta_0 + \beta_1 E(X|\mathbf{Z}) + \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_2 \\
\mathrm{E}(W_1|\mathbf{Z}) &= \mathrm{E}(X|\mathbf{Z}) \\
\mathrm{var}(Y|\mathbf{Z}) &= \beta_1^2 \mathrm{var}(X|\mathbf{Z}) + \sigma_\epsilon^2 \\
\mathrm{var}(W_1 + W_2|\mathbf{Z}) &= 4\mathrm{var}(X|\mathbf{Z}) + 2E\{\sigma^2(X)|\mathbf{Z}\} \\
\mathrm{var}(W_1 - W_2|\mathbf{Z}) &= 2\mathrm{E}\{\sigma^2(X)|\mathbf{Z}\} \\
\mathrm{cov}(W_1, Y|\mathbf{Z}) &= \beta_1 \mathrm{var}(X|\mathbf{Z}).
\end{aligned}$$

From the above equations, we get $\mathrm{var}(X|\mathbf{Z}) = \{\mathrm{var}(W_1 + W_2|\mathbf{Z}) - \mathrm{var}(W_1 - W_2|\mathbf{Z})\}/4$, hence $\mathrm{var}(X|\mathbf{Z})$ is identifiable. Subsequently, we have $\beta_1 = 4\mathrm{cov}(W_1, Y|\mathbf{Z})/\{\mathrm{var}(W_1 + W_2|\mathbf{Z}) - \mathrm{var}(W_1 - W_2|\mathbf{Z})\}$, hence β_1 is also identifiable. From

$$\begin{aligned}
\sigma_\epsilon^2 &= \mathrm{var}(Y|\mathbf{Z}) - \beta_1^2 \mathrm{var}(X|\mathbf{Z}) \\
&= \mathrm{var}(Y|\mathbf{Z}) - \frac{4\mathrm{cov}^2(W_1, Y|\mathbf{Z})}{\mathrm{var}(W_1 + W_2|\mathbf{Z}) - \mathrm{var}(W_1 - W_2|\mathbf{Z})},
\end{aligned}$$

we obtain the identifiability of σ_ϵ^2. Further, we have

$$\begin{aligned}
\beta_0 + \mathbf{Z}^{\mathrm{T}}\boldsymbol{\beta}_2 &= \mathrm{E}(Y|\mathbf{Z}) - \beta_1 \mathrm{E}(X|\mathbf{Z}) \\
&= \mathrm{E}(Y|\mathbf{Z}) - \frac{4\mathrm{cov}(W_1, Y|\mathbf{Z})\mathrm{E}(W_1|\mathbf{Z})}{\mathrm{var}(W_1 + W_2|\mathbf{Z}) - \mathrm{var}(W_1 - W_2|\mathbf{Z})},
\end{aligned}$$

then

$$\begin{aligned}
\beta_0 &= \mathrm{E}(Y|\mathbf{Z}=\mathbf{0}) - \frac{4\mathrm{cov}(W_1, Y|\mathbf{Z}=\mathbf{0})\mathrm{E}(W_1|\mathbf{Z}=\mathbf{0})}{\mathrm{var}(W_1+W_2|\mathbf{Z}=\mathbf{0}) - \mathrm{var}(W_1-W_2|\mathbf{Z}=\mathbf{0})}, \\
\boldsymbol{\beta}_2 &= \{\mathrm{E}(\mathbf{Z}\mathbf{Z}^{\mathrm{T}})\}^{-1}\mathrm{E}\left\{\mathbf{Z}\mathrm{E}(Y|\mathbf{Z}) - \frac{4\mathbf{Z}\mathrm{cov}(W_1, Y|\mathbf{Z})\mathrm{E}(W_1|\mathbf{Z})}{\mathrm{var}(W_1+W_2|\mathbf{Z}) - \mathrm{var}(W_1-W_2|\mathbf{Z})} - \beta_0\mathbf{Z}\right\}.
\end{aligned}$$

Therefore β_0 and $\boldsymbol{\beta}_2$ are also identifiable.

Having obtained the identifiability of β_0, β_1, $\boldsymbol{\beta}_2$ and σ_ϵ, we proceed to prove the identifiability of the unknown function $\sigma(\cdot)$ and the conditional density $f_{X|\mathbf{Z}}(\cdot)$. We first consider the normal measurement errors. Given the observed data $(Y, W_1, W_2, \mathbf{Z})$ and the identifiability of β_0, β_1, $\boldsymbol{\beta}_2$ and σ_ϵ, assume the model is still not identifiable. Then, there exist $\{\sigma(\cdot), f_{X|\mathbf{Z}}(\cdot)\}$ and $\{\widetilde{\sigma}(\cdot), \widetilde{f}_{X|\mathbf{Z}}(\cdot)\}$ that satisfy

$$\begin{aligned}
&\int_{-\infty}^{+\infty} f_{Y|X,\mathbf{Z}}(y|x,\mathbf{z},\beta_0,\beta_1,\boldsymbol{\beta}_2,\sigma_\epsilon) f_U\left\{\frac{w_1-x}{\sigma(x)}\right\} f_U\left\{\frac{w_2-x}{\sigma(x)}\right\} f_{X|\mathbf{Z}}(x|\mathbf{z})\frac{1}{\sigma^2(x)}dx \\
&= \int_{-\infty}^{+\infty} f_{Y|X,\mathbf{Z}}(y|x,\mathbf{z},\beta_0,\beta_1,\boldsymbol{\beta}_2,\sigma_\epsilon) f_U\left\{\frac{w_1-x}{\widetilde{\sigma}(x)}\right\} f_U\left\{\frac{w_2-x}{\widetilde{\sigma}(x)}\right\} \widetilde{f}_{X|\mathbf{Z}}(x|\mathbf{z})\frac{1}{\widetilde{\sigma}^2(x)}dx. \quad \text{(A.1)}
\end{aligned}$$

We rewrite (A.1) as convolution

$$\begin{aligned}
g \circ h &= \int_{-\infty}^{+\infty} g(y-t)h(t)dt \\
&= \int_{-\infty}^{+\infty} g(y-t)\widetilde{h}(t)dt \\
&= g \circ \widetilde{h},
\end{aligned}$$

where $t = \beta_0 + x\beta_1 + \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2$,

$$\begin{aligned}
g(y-t) &= \exp\left\{-\frac{(y-t)^2}{2\sigma_\epsilon^2}\right\}, \\
h(t) &= \exp\left[-\frac{\{w_1-(t-\beta_0-\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1\}^2 + \{w_2-(t-\beta_0-\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1\}^2}{2\sigma^2\{(t-\beta_0-\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1\}}\right] \\
&\quad \times \frac{f_{X|\mathbf{Z}}\{(t-\beta_0-\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1|\mathbf{z}\}}{\beta_1\sigma^2\{(t-\beta_0-\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1\}}, \\
\widetilde{h}(t) &= \exp\left[-\frac{\{w_1-(t-\beta_0-\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1\}^2 + \{w_2-(t-\beta_0-\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1\}^2}{2\widetilde{\sigma}^2\{(t-\beta_0-\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1\}}\right] \\
&\quad \times \frac{\widetilde{f}_{X|\mathbf{Z}}\{(t-\beta_0-\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1|\mathbf{z}\}}{\beta_1\widetilde{\sigma}^2\{(t-\beta_0-\mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1\}}.
\end{aligned}$$

By the convolution theorem of Fourier transform, we have $\mathcal{F}(g)\mathcal{F}(h) = \mathcal{F}(g)\mathcal{F}(\widetilde{h})$, then $\mathcal{F}(h) = \mathcal{F}(\widetilde{h})$. Hence $h(t) = \widetilde{h}(t)$ for any $t \in \mathbb{R}$ via the inverse Fourier transformation. Because w_1, w_2 can be any values, this directly leads to $\sigma(\cdot) = \widetilde{\sigma}(\cdot)$ and $f_{X|\mathbf{Z}}(\cdot) = \widetilde{f}_{X|\mathbf{Z}}(\cdot)$.

For Laplace measurement errors, we have

$$\begin{aligned} h(t) &= \exp\left[-\frac{|w_1 - (t - \beta_0 - \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1| + |w_2 - (t - \beta_0 - \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1|}{\sqrt{1/2}\sigma\{(t - \beta_0 - \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1\}}\right] \\ &\quad \times \frac{f_{X|\mathbf{Z}}\{(t - \beta_0 - \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1|\mathbf{z}\}}{\beta_1\sigma^2\{(t - \beta_0 - \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1\}}, \\ \widetilde{h}(t) &= \exp\left[-\frac{|w_1 - (t - \beta_0 - \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1| + |w_2 - (t - \beta_0 - \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1|}{\sqrt{1/2}\widetilde{\sigma}\{(t - \beta_0 - \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1\}}\right] \\ &\quad \times \frac{\widetilde{f}_{X|\mathbf{Z}}\{(t - \beta_0 - \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1|\mathbf{z}\}}{\beta_1\widetilde{\sigma}^2\{(t - \beta_0 - \mathbf{z}^{\mathrm{T}}\boldsymbol{\beta}_2)/\beta_1\}}. \end{aligned}$$

Note that $\sigma(X) > 0$ and $\widetilde{\sigma}(X) > 0$ for any X. Similar to the proof of normal measurement errors, we have $\sigma(\cdot) = \widetilde{\sigma}(\cdot)$ and $f_{X|\mathbf{Z}}(\cdot) = \widetilde{f}_{X|\mathbf{Z}}(\cdot)$.

This completes the proof of the identifiability of all components in the model. □

A.2 The Derivation of Nuisance Tangent Space

For a parametric model, the nuisance tangent space is the linear space in $\mathcal{H}$ spanned by the nuisance score vector. For semiparametric models, in which the nuisance parameter is infinite-dimensional, the nuisance tangent space is defined as the mean squared closure of all parametric submodel nuisance tangent spaces. The parametric submodel is a true parametric model contained in the semiparametric model. In our original model (3.1) and (3.2), the nuisance score vector for the parametric submodel $f_{X|\mathbf{Z}}(x|\mathbf{z}, \boldsymbol{\xi}_1)$ is

$$\begin{aligned} &\mathbf{S}_{f_{X|\mathbf{Z}}}\{Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \sigma(\cdot), \boldsymbol{\xi}_1\} \\ = &\frac{\int\left[\{\partial f_{X|\mathbf{Z}}(x|\mathbf{z}, \boldsymbol{\xi}_1)/\partial\boldsymbol{\xi}_1\}/f_{X|\mathbf{Z}}(x)\right] f_{Y,W_1,W_2,X|\mathbf{Z}}\{y, w_1, w_2, x, \mathbf{z}, \boldsymbol{\beta}, \sigma(\cdot), \boldsymbol{\xi}_1\}dx}{\int f_{Y,W_1,W_2,X|\mathbf{Z}}\{y, w_1, w_2, x, \mathbf{z}, \boldsymbol{\beta}, \sigma(\cdot), \boldsymbol{\xi}_1\}dx} \\ = &\mathrm{E}\left\{\frac{\partial f_{X|\mathbf{Z}}(X|\mathbf{Z}, \boldsymbol{\xi}_1)}{\partial\boldsymbol{\xi}_1}\frac{1}{f_{X|\mathbf{Z}}(X|\mathbf{Z}, \boldsymbol{\xi}_1)}\middle| Y, W_1, W_2, \mathbf{Z}\right\}, \end{aligned}$$

where $\mathrm{E}[\{\partial f_{X|\mathbf{Z}}(x|\mathbf{z}, \boldsymbol{\xi}_1)/\partial\boldsymbol{\xi}_1\}/f_{X|\mathbf{Z}}(x|\mathbf{z}, \boldsymbol{\xi}_1)|\mathbf{Z}] = 0$. Thus the nuisance tangent space with respect to $f_{X|\mathbf{Z}}$ is $\Lambda_{f_{X|\mathbf{Z}}} = [\mathrm{E}\{\mathbf{a}(X, \mathbf{Z})|Y, W_1, W_2, \mathbf{Z}\} : \mathrm{E}\{\mathbf{a}(X, \mathbf{Z})|\mathbf{Z}\} = \mathbf{0}]$. Similarly,

the nuisance score vector for the parametric submodel $\sigma(x, \boldsymbol{\xi}_2)$ is given as

$$\begin{aligned}
& \mathbf{S}_\sigma\{Y, W_1, W_2, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\xi}_2, f_{X|\mathbf{Z}}(\cdot)\} \\
= & \frac{\int -V(u_1,u_2)[\{\partial\sigma(x,\boldsymbol{\xi}_2)/\partial\boldsymbol{\xi}_2\}/\sigma(x,\boldsymbol{\xi}_2)] f_{Y,W_1,W_2,X|\mathbf{Z}}\{y,w_1,w_2,x,\mathbf{z},\boldsymbol{\beta},\boldsymbol{\xi}_2,f_{X|\mathbf{Z}}(\cdot)\}dx}{\int f_{Y,W_1,W_2,X|\mathbf{Z}}\{y,w_1,w_2,x,\mathbf{z},\boldsymbol{\beta},\boldsymbol{\xi}_2,f_{X|\mathbf{Z}}(\cdot)\}dx} \\
= & \mathrm{E}\left\{ -V(U_1,U_2)\frac{\partial\sigma(x,\boldsymbol{\xi}_2)}{\partial\boldsymbol{\xi}_2}\frac{1}{\sigma(x,\boldsymbol{\xi}_2)} \,\middle|\, Y, W_1, W_2, \mathbf{Z} \right\},
\end{aligned}$$

where $\partial \ln\{\sigma(x, \boldsymbol{\xi}_2)\}/\partial\boldsymbol{\xi}_2$ can be any function of x. Thus the nuisance tangent space with respect to σ is $\Lambda_\sigma = [E\{V(U_1, U_2)\mathbf{b}(X)|Y, W_1, W_2, \mathbf{Z}\} : \forall \mathbf{b}(X)]$. It is easy to verify that $\mathrm{E}\{V(U_1, U_2)\} = 0$. Since U is independent of X and $\mathbf{Z}$, $\mathrm{E}\{V(U_1, U_2)\mathbf{b}(X)|\mathbf{Z}\} = \mathrm{E}[\mathrm{E}\{V(U_1, U_2)|X, \mathbf{Z}\}\mathbf{b}(X)|\mathbf{Z}] = \mathbf{0}$ for arbitrary $\mathbf{b}(X)$. Thus

$$\mathrm{E}\{V(U_1, U_2)\mathbf{b}(X)|Y, W_1, W_2, \mathbf{Z}\} \in \mathcal{H}$$

indeed.

A.3 Nontrivial Orthogonal Complement of Nuisance Tangent Space for Logistic Model with Heteroscedastic Normal Measurement Errors

Following the proof in Tsiatis & Ma (2004), if we can find a nonzero function $h(Y, W_1, W_2, Z)$ such that $\mathrm{E}\{h(Y, W_1, W_2, Z)|X, Z\} = 0$ and $\mathrm{E}\{h(Y, W_1, W_2, Z)V(U_1, U_2)|X, Z\} = 0$, simultaneously, then the orthogonal complement of the nuisance tangent space is nontrivial. Hence, the root-n estimators exist, which means the parameters are estimable and then the problem is identifiable.

Consider logistic regression model $\mathrm{logit}\Pr(Y = 1|X, Z) = \beta_0 + \beta_1 X + \beta_2 Z$, and $W = X + \sigma(X)U$, where U is standard normal. Since Y is binary, any function of Y, W_1, W_2, Z can be written as $h(Y, W_1, W_2, Z) = Y h_1(W_1, W_2, Z) - h_2(W_1, W_2, Z)$. Since Y and W_j are conditionally independent given X and Z for $j = 1, 2$, then $\mathrm{E}\{h(Y, W_1, W_2, Z)|X, Z\} = \mathrm{E}(Y|X, Z)\mathrm{E}\{h_1(W_1, W_2, Z)|X, Z\} - \mathrm{E}\{h_2(W_1, W_2, Z)|X, Z\}$. The conditional expectation $\mathrm{E}\{h(Y, W_1, W_2, Z)|X, Z\} = 0$, if

$$\mathrm{E}\{h_1(W_1, W_2, Z)|X, Z\} = \{1 + \exp(-\beta_0 - \beta_1 X - \beta_2 Z)\}\mathrm{E}\{h_2(W_1, W_2, Z)|X, Z\}.$$

When the conditional distribution of W_j given $X = x$, $Z = z$ is $\mathcal{N}\{x, \sigma^2(x)\}$ for $j = 1, 2$, then the standard calculations for normal densities yield that

$$\mathrm{E}\{\exp(\beta_1 W)|X, Z\} = \exp(\beta_1 X)\exp\{\beta_1^2\sigma^2(X)/2\}.$$

Further, since U_1 and U_2 are independent, then $W_1 - W_2$ given $X = x$ and $Z = z$ is $\mathcal{N}\{0, 2\sigma^2(x)\}$. Similarly, we have

$$\mathrm{E}[\exp\{\beta_1(W_1 - W_2)/\sqrt{2}\}|X, Z] = \exp\{\beta_1^2\sigma^2(X)/2\}.$$

Let $h_2(W_1, W_2, Z) = \exp(\beta_0 + \beta_2 Z)\exp(\beta_1 W_1)$. Then

$$\mathrm{E}\{h_1(W_1, W_2, Z)|X, Z\} = \mathrm{E}\{h_2(W_1, W_2, Z)|X, Z\} + \exp\{\beta_1^2\sigma^2(X)/2\}$$

. Hence, a nontrivial solution exists by choosing $h_1(W_1, W_2, Z) = h_2(W_1, W_2, Z) + \exp\{\beta_1(W_1 - W_2)/\sqrt{2}\}$. It can be verified that this nontrivial solution $h(Y, W_1, W_2, Z)$ also satisfies

$$\mathrm{E}\{h(Y, W_1, W_2, Z)V(U_1, U_2)|X, Z\} = 0$$

based on standard calculations. Therefore, the orthogonal complement of the nuisance tangent space is nontrivial and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ is identifiable.

A.4 The Derivation of $\mathbf{S}_{a,\mathrm{eff},\boldsymbol{\gamma}}\{Y, W_1, W_2, \boldsymbol{\beta}, \boldsymbol{\gamma}, f_{X|\mathbf{Z}}(\cdot)\}$

To estimate $\boldsymbol{\gamma}$ at each $\boldsymbol{\beta}$ in the approximate model (3.3), we treat $\boldsymbol{\beta}, \boldsymbol{\gamma}$ as parameters of interest and $f_{X|\mathbf{Z}}(\cdot)$ as nuisance. The nuisance tangent space is

$$\Lambda_{a,f_{X|\mathbf{Z}}} = [\mathrm{E}_a\{\mathbf{c}(X, \mathbf{Z})|Y, W_1, W_2, \mathbf{Z}\} : \mathrm{E}_a\{\mathbf{c}(X, \mathbf{Z})|\mathbf{Z}\} = 0],$$

and its orthogonal complement is

$$[h(Y, W_1, W_2, \mathbf{Z}) : \mathrm{E}_a\{h(Y, W_1, W_2, \mathbf{Z})|X, \mathbf{Z}\} = \mathbf{0} \text{ almost surely}].$$

The score vector for $\boldsymbol{\gamma}$ is easily verified to be

$$\mathbf{S}_{a,\boldsymbol{\gamma}}\{y, w_1, w_2, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f_{X|\mathbf{X}}(\cdot)\}$$

$$= \frac{\int -V(u_{a,1}, u_{a,2})/\{\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}\} f_{a,Y,W_1,W_2,X|\mathbf{Z}}\{y, w_1, w_2, x, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f_{X|\mathbf{Z}}(\cdot)\}\mathbf{B}(x)dx}{\int f_{a,Y,W_1,W_2,X|\mathbf{Z}}\{y, w_1, w_2, x, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f_{X|\mathbf{Z}}(\cdot)\}dx}.$$

Therefore the efficient score $\mathbf{S}_{a,\text{eff},\boldsymbol{\gamma}}$ is the projection of $\mathbf{S}_{a,\boldsymbol{\gamma}}$ onto the orthogonal complement of nuisance tangent space and is easily verified to be as in (3.4), where $\mathbf{c}(X, \mathbf{Z})$ satisfies (3.5).

A.5 Proof of Theorem 3.4.1

By the definitions of $\mathbf{S}^*_{a,\text{eff},\boldsymbol{\gamma}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\}$ and $\mathbf{S}^*_{\text{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \sigma(\cdot), f^*_{X|\mathbf{Z}}(\cdot)\}$, we have

$$\begin{aligned}
\mathrm{E}[\mathbf{S}^*_{\text{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \sigma(\cdot), f^*_{X|\mathbf{Z}}(\cdot)\}|X, \mathbf{Z}_i] &= \mathbf{0}, \\
\mathrm{E}_a[\mathbf{S}^*_{a,\text{eff},\boldsymbol{\gamma}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, f^*_{X|\mathbf{Z}}(\cdot)\}|X, \mathbf{Z}_i] &= \mathbf{0}.
\end{aligned}$$

Then the un-conditional expectations also equal $\mathbf{0}$ almost everywhere, that is,

$$\begin{aligned}
\mathrm{E}[\mathbf{S}^*_{\text{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \sigma(\cdot), f^*_{X|Z}(\cdot)\}] &= \mathbf{0}, \\
\mathrm{E}_a[\mathbf{S}^*_{a,\text{eff},\boldsymbol{\gamma}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, f^*_{X|\mathbf{Z}}(\cdot)\}] &= \mathbf{0}.
\end{aligned}$$

This leads to

$$\begin{aligned}
\mathrm{E}[\mathbf{S}^*_{\text{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, f^*_{X|\mathbf{Z}}(\cdot)\}] &= o_p(1), \\
\mathrm{E}[\mathbf{S}^*_{a,\text{eff},\boldsymbol{\gamma}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, f^*_{X|\mathbf{Z}}(\cdot)\}] &= o_p(1),
\end{aligned}$$

element-wise by Remark 3.4.1. Condition (C6) ensures that as a vector function of $\boldsymbol{\theta}$,

$$(\mathrm{E}[\mathbf{S}^*_{\text{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\}]^{\mathrm{T}}, \mathrm{E}[\mathbf{S}^*_{a,\text{eff},\boldsymbol{\gamma}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\}]^{\mathrm{T}})^{\mathrm{T}}$$

is invertible near $\boldsymbol{\theta}^*$ and the first derivative of the inverse function is bounded in the neighborhood of its zero. Therefore, $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\|_2 = o_p(1)$.

On the other hand, since

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \mathbf{S}^*_{\text{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\gamma}}_n, f^*_{X|\mathbf{Z}}(\cdot)\} &= \mathbf{0}, \\
\frac{1}{n}\sum_{i=1}^{n} \mathbf{S}^*_{a,\text{eff},\boldsymbol{\gamma}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\gamma}}_n, f^*_{X|\mathbf{Z}}(\cdot)\} &= \mathbf{0},
\end{aligned}$$

we have

$$\begin{aligned} \mathrm{E}[\mathbf{S}^*_{\mathrm{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\gamma}}_n, f^*_{X|\mathbf{Z}}(\cdot)\}] &= o(1), \\ \mathrm{E}[\mathbf{S}^*_{a,\mathrm{eff},\boldsymbol{\gamma}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\gamma}}_n, f^*_{X|\mathbf{Z}}(\cdot)\}] &= o(1) \end{aligned}$$

element-wise. Using the same argument, we obtain $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*\|_2 = o_p(1)$. Hence $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|_2 = o_p(1)$. □

A.6 Proof of Theorem 3.4.2

To prove asymptotic normality, we first expand the estimating function (3.7) as a function of $\boldsymbol{\beta}$ about $\boldsymbol{\beta}_0$ keeping $\widehat{\boldsymbol{\gamma}}_n(\cdot)$ fixed, to obtain $\mathbf{T}_1 + \mathbf{T}_2(\widetilde{\boldsymbol{\beta}}_n)\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = \mathbf{0}$, where

$$\begin{aligned} \mathbf{T}_1 &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \mathbf{S}^*_{\mathrm{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\beta}_0), f^*_{X|\mathbf{Z}}(\cdot)\}, \\ \mathbf{T}_2(\boldsymbol{\beta}) &= \mathbf{T}_{21}(\boldsymbol{\beta}) + \mathbf{T}_{22}(\boldsymbol{\beta}) \frac{\partial \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}. \end{aligned}$$

Here

$$\begin{aligned} \mathbf{T}_{21}(\boldsymbol{\beta}) &= \frac{1}{n} \sum_{i=1}^n \frac{\partial \mathbf{S}^*_{\mathrm{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}_n, f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}, \\ \mathbf{T}_{22}(\boldsymbol{\beta}) &= \frac{1}{n} \sum_{i=1}^n \frac{\partial \mathbf{S}^*_{\mathrm{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\beta}), f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\beta})^{\mathrm{T}}}, \end{aligned}$$

and $\widetilde{\boldsymbol{\beta}}_n$ is on the line connecting $\boldsymbol{\beta}_0$ and $\widehat{\boldsymbol{\beta}}_n$. Since $\widehat{\boldsymbol{\gamma}}_n(\cdot)$ satisfies

$$n^{-1} \sum_{i=1}^n \mathbf{S}^*_{a,\mathrm{eff},\boldsymbol{\gamma}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\beta}), f^*_{X|\mathbf{Z}}(\cdot)\} = \mathbf{0}$$

for any $\boldsymbol{\beta}$,

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n \frac{\partial \mathbf{S}^*_{a,\mathrm{eff},\boldsymbol{\gamma}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}_n, f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}} \\ + & \frac{1}{n} \sum_{i=1}^n \frac{\partial \mathbf{S}^*_{a,\mathrm{eff},\boldsymbol{\gamma}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\beta}), f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\beta})^{\mathrm{T}}} \frac{\partial \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}} = \mathbf{0}. \end{aligned}$$

Then

$$\frac{\partial \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{\mathrm{T}}} = -\{\mathbf{T}_{23}(\boldsymbol{\beta})\}^{-1}\mathbf{T}_{24}(\boldsymbol{\beta}),$$

where

$$\mathbf{T}_{23}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbf{S}^*_{a,\mathrm{eff},\boldsymbol{\gamma}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\beta}), f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial \widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\beta})^{\mathrm{T}}},$$

$$\mathbf{T}_{24}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbf{S}^*_{a,\mathrm{eff},\boldsymbol{\gamma}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}_n, f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial \boldsymbol{\beta}^{\mathrm{T}}}.$$

Hence $\mathbf{T}_2(\widetilde{\boldsymbol{\beta}}_n) = \mathbf{T}_{21}(\widetilde{\boldsymbol{\beta}}_n) - \mathbf{T}_{22}(\widetilde{\boldsymbol{\beta}}_n)\{\mathbf{T}_{23}(\widetilde{\boldsymbol{\beta}}_n)\}^{-1}\mathbf{T}_{24}(\widetilde{\boldsymbol{\beta}}_n)$. We further expand $\mathbf{T}_1$ as a function of $\widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\beta}_0)$ about $\boldsymbol{\gamma}_0(\boldsymbol{\beta}_0)$ to obtain

$$\mathbf{T}_1 = \mathbf{T}_{11} + \mathbf{T}_{12}\{\widetilde{\boldsymbol{\gamma}}_n(\boldsymbol{\beta}_0)\}\sqrt{n}\{\widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0(\boldsymbol{\beta}_0)\},$$

where

$$\mathbf{T}_{11} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{S}^*_{\mathrm{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0(\boldsymbol{\beta}_0), f^*_{X|\mathbf{Z}}(\cdot)\},$$

$$\mathbf{T}_{12}\{\boldsymbol{\gamma}(\boldsymbol{\beta}_0)\} = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathbf{S}^*_{\mathrm{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}(\boldsymbol{\beta}_0), f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial \boldsymbol{\gamma}(\boldsymbol{\beta}_0)^{\mathrm{T}}},$$

and $\widetilde{\boldsymbol{\gamma}}_n(\boldsymbol{\beta}_0)$ is a value between $\widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\beta}_0)$ and $\boldsymbol{\gamma}_0(\boldsymbol{\beta}_0)$.

By the consistency of $\mathbf{B}(x)^{\mathrm{T}}\widetilde{\boldsymbol{\gamma}}_n$ to $\sigma(x)$, for arbitrary $d_{\boldsymbol{\gamma}} \times p$ matrix $\mathbf{G}$ with $\|\mathbf{G}\|_2 = 1$, we have

$$\mathbf{T}_{12}\{\widetilde{\boldsymbol{\gamma}}_n(\boldsymbol{\beta}_0)\}\mathbf{G} = \mathrm{E}\left[\frac{\partial \mathbf{S}^*_{\mathrm{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\mathbf{G}\,\Big|_{\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}=\sigma(\cdot)}\right]\{1 + o_p(1)\},$$

where

$$\mathrm{E}\left[\frac{\partial \mathbf{S}^*_{\mathrm{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial \boldsymbol{\gamma}^{\mathrm{T}}}\mathbf{G}\,\Big|_{\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}=\sigma(\cdot)}\right]$$

$$
\begin{aligned}
&= \int \left[\frac{\partial \mathbf{S}^*_{\text{eff}}\{y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial \boldsymbol{\gamma}^{\mathrm{T}}} \mathbf{G} \Big|_{\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}=\sigma(\cdot)} \right] \\
&\quad \times f_{Y|W_1,W_2,\mathbf{Z}}\{y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \sigma, f_{X|\mathbf{Z}}(\cdot)\} f_{\mathbf{Z}}(\mathbf{z}_i) dy_i dw_{i1} dw_{i2} d\mathbf{z}_i \\
&= \int \left[\frac{\partial \mathbf{S}^*_{\text{eff}}\{y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mathbf{G} + O_p(h_b^q) \right] \\
&\quad \times \left\{ f_{a,Y,W_1,W_2|\mathbf{Z}}\{y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, f_{X|\mathbf{Z}}(\cdot)\} f_{\mathbf{Z}}(\mathbf{z}_i) + O_p(h_b^q) \right\} dy_i dw_{i1} dw_{i2} d\mathbf{z}_i \\
&= \int \frac{\partial \mathbf{S}^*_{\text{eff}}\{y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mathbf{G} \\
&\quad \times f_{a,Y,W_1,W_2|\mathbf{Z}}\{y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, f_{X|\mathbf{Z}}(\cdot)\} f_{\mathbf{Z}}(\mathbf{z}_i) dy_i dw_{i1} dw_{i2} d\mathbf{z}_i + O_p(h_b^q) \\
&= \frac{\partial}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \int [\mathbf{S}^*_{\text{eff}}\{y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \sigma(\cdot), f^*_{X|\mathbf{Z}}(\cdot)\} + O_p(h_b^q)] \mathbf{G} \\
&\quad \times [f_{Y,W_1,W_2|\mathbf{Z}}\{y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \beta_0, \sigma, f_{X|\mathbf{Z}}(\cdot)\} f_{\mathbf{Z}}(\mathbf{z}_i) + O_p(h_b^q)] dy_i dw_{i1} dw_{i2} d\mathbf{z}_i \\
&\quad - \int [\mathbf{S}^*_{\text{eff}}\{y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \sigma(\cdot), f^*_{X|\mathbf{Z}}(\cdot)\} + O_p(h_b^q)] \mathbf{G} \\
&\quad \times \frac{\partial f_{a,Y,W_1,W_2|\mathbf{Z}}\{y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, f_{X|\mathbf{Z}}(\cdot)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} f_{\mathbf{Z}}(\mathbf{z}_i) dy_i dw_{i1} dw_{i2} d\mathbf{z}_i + O_p(h_b^q) \\
&= - \int \mathbf{S}^*_{\text{eff}}\{y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \sigma(\cdot), f^*_{X|\mathbf{Z}}(\cdot)\} [\mathbf{G}^{\mathrm{T}} \mathbf{S}_{a,\boldsymbol{\gamma}}\{y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, f_{X|\mathbf{Z}}(\cdot)\}]^{\mathrm{T}} \\
&\quad \times f_{Y,W_1,W_2|\mathbf{Z}}\{y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \sigma, f_{X|\mathbf{Z}}(\cdot)\} f_{\mathbf{Z}}(\mathbf{z}_i) dy_i dw_{i1} dw_{i2} d\mathbf{z}_i + O_p(h_b^p) \\
&= O_p(h_b^q).
\end{aligned}
$$

The second equality holds by Remark 3.4.1. The third equality holds because

$$
\|\partial \mathbf{S}^*_{\text{eff}}\{y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0), f^*_{X|\mathbf{Z}}(\cdot)\} / \partial \boldsymbol{\gamma}_0^{\mathrm{T}}\|_\infty
$$

is integrable by condition (C7) and $f_{a,Y,W_1,W_2|\mathbf{Z}}\{y_i, w_{i1}, w_{i2}, \mathbf{z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, f_{X|\mathbf{Z}}(\cdot)\} f_{\mathbf{Z}}(\mathbf{z}_i)$ is absolutely integrable. The fourth equality holds by Remark 3.4.1. The fifth equality holds because $\mathrm{E}[\mathbf{S}^*_{\text{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}, \sigma(\cdot), f^*_{X|\mathbf{Z}}(\cdot)\}] = \mathbf{0}$. For the last equality, first note that for any $p \times d_\gamma$ matrix $\mathbf{K}$, there exists a function $\mathbf{b}(X)$ such that $\mathbf{KB}(X) = \mathbf{b}(X)$. Then by Remark 3.4.1 and definitions of Λ_σ and $\Lambda_{a,\boldsymbol{\gamma}}$, for any $d_\gamma \times p$ matrix $\mathbf{G}$, there exists a function $\mathbf{g}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \sigma(\cdot)\} \in \Lambda_\sigma$ such that $\sup_X |\mathbf{G}^{\mathrm{T}} \mathbf{S}_{a,\boldsymbol{\gamma}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, f_{X|\mathbf{Z}}(\cdot)\} - \mathbf{g}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \sigma(\cdot)\}| = O_p(h_b^q)$. Further, $\mathbf{S}^*_{\text{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \sigma(\cdot), f^*_{X|\mathbf{Z}}(\cdot)\}$ is orthogonal to any function in Λ_σ, thus the last equality holds. Hence, we obtain $\|\mathbf{T}_{12}\{\widetilde{\boldsymbol{\gamma}}(\boldsymbol{\beta}_0)\}\|_2 = O_p(h_b^q)$.

Based on the asymptotic results of Proposition 4 in Jiang & Ma (2018), we have

$\|\widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0(\boldsymbol{\beta}_0)\|_2 = O_p\{(nh_b)^{-1/2}\}$. Then we have

$$\|\mathbf{T}_{12}\{\widetilde{\boldsymbol{\gamma}}_n(\boldsymbol{\beta}_0)\}\sqrt{n}\{\widehat{\boldsymbol{\gamma}}_n(\boldsymbol{\beta}_0) - \boldsymbol{\gamma}_0(\boldsymbol{\beta}_0)\}\|_2 = O_p(h_b^{q-1/2}).$$

Further, by Remark 3.4.1 we have

$$\mathbf{T}_{11} = n^{-1/2}\sum_{i=1}^{n}\mathbf{S}^*_{\text{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \sigma(\cdot), f^*_{X|\mathbf{Z}}(\cdot)\} + O_p(n^{1/2}h_b^q).$$

Since $h_b^{q-1/2} = o_p(n^{1/2}h_b^q)$, then

$$\mathbf{T}_{1} = n^{-1/2}\sum_{i=1}^{n}\mathbf{S}^*_{\text{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \sigma(\cdot), f^*_{X|\mathbf{Z}}(\cdot)\} + O_p(n^{1/2}h_b^q).$$

Note that $n^{1/2}h_b^q = o_p(1)$ by conditions (C4) and (C5).

Now consider $\mathbf{T}_2(\widetilde{\boldsymbol{\beta}}_n)$. By the consistency of $\widetilde{\boldsymbol{\beta}}_n$ to $\boldsymbol{\beta}_0$ and $\mathbf{B}(x)^{\mathrm{T}}\widehat{\boldsymbol{\gamma}}_n$ to $\sigma(x)$, we have

$$\mathbf{T}_{21}(\widetilde{\boldsymbol{\beta}}_n) = \mathrm{E}\left[\frac{\partial\mathbf{S}^*_{\text{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial\boldsymbol{\beta}_0^{\mathrm{T}}}\bigg|_{\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}=\sigma(\cdot)}\right]\{1 + o_p(1)\},$$

and

$$\mathbf{T}_{24}(\widetilde{\boldsymbol{\beta}}_n) = \mathrm{E}\left[\frac{\partial\mathbf{S}^*_{a,\text{eff},\boldsymbol{\gamma}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial\boldsymbol{\beta}_0^{\mathrm{T}}}\bigg|_{\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}=\sigma(\cdot)}\right]\{1 + o_p(1)\}.$$

We also have

$$\mathbf{T}_{22}(\widetilde{\boldsymbol{\beta}}_n) = \mathrm{E}\left[\frac{\partial\mathbf{S}^*_{\text{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial\boldsymbol{\gamma}^{\mathrm{T}}}\bigg|_{\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}=\sigma(\cdot)}\right]\{1 + o_p(1)\}.$$

We have already proved that

$$\mathrm{E}([\partial\mathbf{S}^*_{\text{eff}}\{Y_i, W_{i1}, W_{i2}, \boldsymbol{\beta}_0, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\}/\partial\boldsymbol{\gamma}^{\mathrm{T}}]\mathbf{G}|_{\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}=\sigma(\cdot)}) = O_p(h_b^q)$$

element-wise for any arbitrary $d_{\boldsymbol{\gamma}} \times p$ matrix $\mathbf{G}$ with $\|\mathbf{G}\|_2 = 1$ by showing that for any $d_{\boldsymbol{\gamma}} \times p$ matrix $\mathbf{G}$, there exists a function $\mathbf{g} \in \Lambda_\sigma$, which is orthogonal to $\mathbf{S}^*_{\text{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \sigma(\cdot), f^*_{X|\mathbf{Z}}(\cdot)\}$, such that

$$\sup_x |\mathbf{G}^{\mathrm{T}}\mathbf{S}_{a,\boldsymbol{\gamma}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}_0, f_{X|\mathbf{Z}}(\cdot)\} - \mathbf{g}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \sigma(\cdot)\}| = O_p(h_b^q).$$

For $\mathbf{T}_{23}(\widetilde{\boldsymbol{\beta}}_n)$, based on the proof of Proposition 4 in Jiang & Ma (2018), we have $\|\mathbf{T}_{23}(\widetilde{\boldsymbol{\beta}}_n)^{-1}\|_2 = O_p(h_b^{-1})$. Then we have $\mathbf{T}_{22}(\widetilde{\boldsymbol{\beta}}_n)\{\mathbf{T}_{23}(\widetilde{\boldsymbol{\beta}}_n)\}^{-1}\mathbf{T}_{24}(\widetilde{\boldsymbol{\beta}}_n) = O_p(h_b^{q-1})$, where $q > 1$ by condition (C2). Thus

$$\mathbf{T}_2(\widetilde{\boldsymbol{\beta}}_n) = \mathrm{E}\left[\frac{\partial \mathbf{S}^*_{\mathrm{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial \boldsymbol{\beta}_0^{\mathrm{T}}}\bigg|_{\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}=\sigma(\cdot)}\right]\{1 + o_p(1)\} + O_p(h_b^{q-1}).$$

Therefore,

$$\begin{aligned}
&\sqrt{n}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\
= \; & -\left(\mathrm{E}\left[\frac{\partial \mathbf{S}^*_{\mathrm{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\}}{\partial \boldsymbol{\beta}_0^{\mathrm{T}}}\bigg|_{\mathbf{B}(\cdot)^{\mathrm{T}}\boldsymbol{\gamma}=\sigma(\cdot)}\right]\right)^{-1} \\
& \times \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathbf{S}^*_{\mathrm{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \sigma(\cdot), f^*_{X|\mathbf{Z}}(\cdot)\} + o_p(1).
\end{aligned}$$

Since $n^{-1/2}\sum_{i=1}^{n} \mathbf{S}^*_{\mathrm{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \sigma(\cdot), f^*_{X|\mathbf{Z}}(\cdot)\}$ is the sum of independent zero-mean random vectors, this will converge in distribution to a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix

$$(\mathrm{E}[\mathbf{S}^*_{\mathrm{eff}}\{Y_i, W_{i1}, W_{i2}, \mathbf{Z}_i, \boldsymbol{\beta}_0, \sigma(\cdot), f^*_{X|\mathbf{Z}}(\cdot)\}^{\otimes 2}])^{-1}.$$

□

A.7 Estimating Equations for Subjects Without Replication of W

In the approximate model, the conditional density of (Y, W) given $\mathbf{Z}$ is

$$f^{(1)}_{a,Y,W|\mathbf{Z}}\{y, w, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f_{X|\mathbf{Z}}(\cdot)\} = \int \frac{f_{Y|X,|Z}(y|x, \mathbf{z}, \boldsymbol{\beta}) f_U\{(w-x)/\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}\} f_{X|\mathbf{Z}}(x|\mathbf{z})}{\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}} dx.$$

We use the superscript $^{(1)}$ to denote the corresponding quantities that are calculated when only one W is available. The corresponding scores for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are given as

$$\begin{aligned}
&\mathbf{S}^{(1)}_{a,\boldsymbol{\beta}}\{y, w, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f_{X|\mathbf{Z}}(\cdot)\} \\
= \; & \frac{\int \{\partial f_{Y|X,\mathbf{Z}}(y|x, \mathbf{z}, \boldsymbol{\beta})/\partial \boldsymbol{\beta}\} f_U\{(w-x)/\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}\} f_{X|\mathbf{Z}}(x|\mathbf{z})/\{\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}\} dx}{\int f_{Y|X,\mathbf{Z}}(y|x, \mathbf{z}, \boldsymbol{\beta}) f_U\{(w-x)/\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}\} f_{X|\mathbf{Z}}(x|\mathbf{z})/\{\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}\} dx},
\end{aligned}$$

and

$$\begin{aligned}
&\mathbf{S}^{(1)}_{a,\gamma}\{y, w, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f_{X|\mathbf{Z}}(\cdot)\} \\
= \quad & \frac{\int -V^{(1)}(u_a) f_{Y|X,\mathbf{Z}}(y|x, \mathbf{z}, \boldsymbol{\beta}) f_U\{(w-x)/\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}\} f_{X|\mathbf{Z}}(x|\mathbf{z})\mathbf{B}(x)/\{\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}\}^2 dx}{\int f_{Y|X,\mathbf{Z}}(y|x, \mathbf{z}, \boldsymbol{\beta}) f_U\{(w-x)/\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}\} f_{X|\mathbf{Z}}(x|\mathbf{z})/\{\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}\} dx},
\end{aligned}$$

where $V^{(1)}(u_a) = u_a f'_U(u_a)/f_U(u_a) + 1$ and $u_a = (w-x)/\mathbf{B}(x)^{\mathrm{T}}\boldsymbol{\gamma}$. Following the derivations in Section 3.3, the approximate efficient score for $\boldsymbol{\beta}$ with working model $f^*_{X|\mathbf{Z}}(\cdot)$ is

$$\begin{aligned}
&\mathbf{S}^{*(1)}_{a,\text{eff}}\{Y, W, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\} \\
= \quad & \mathbf{S}^{*(1)}_{a,\boldsymbol{\beta}}\{Y, W, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\} - \mathrm{E}^{*(1)}_a\{\mathbf{a}(X, \mathbf{Z})|Y, W, \mathbf{Z}\} \\
& -\mathrm{E}^{*(1)}_a\{V^{(1)}(U_a)\mathbf{b}(X)|Y, W, \mathbf{Z}\},
\end{aligned}$$

where $a(X, \mathbf{Z})$ and $\mathbf{b}(X)$ satisfy that

$$\begin{aligned}
&\mathrm{E}^{(1)}_a[\mathbf{S}^{*(1)}_{a,\boldsymbol{\beta}}\{Y, W, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\}|X, \mathbf{Z}] \\
= \quad & \mathrm{E}^{(1)}_a[\mathrm{E}^{*(1)}_a\{\mathbf{a}(X, \mathbf{Z})|Y, W, \mathbf{Z}\}|X, \mathbf{Z}] + \mathrm{E}^{(1)}_a[\mathrm{E}^{*(1)}_a\{V^{(1)}(U_a)\mathbf{b}(X)|Y, W, \mathbf{Z}\}|X, \mathbf{Z}]
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{E}^{(1)}_a[\mathbf{S}^{*(1)}_{a,\boldsymbol{\beta}}\{Y, W, \mathbf{Z}, \boldsymbol{\beta}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\beta}), f^*_{X|\mathbf{Z}}(\cdot)\}V(U_a)|X, \mathbf{Z}] \\
= \quad & \mathrm{E}^{(1)}_a[\mathrm{E}^{*(1)}_a\{\mathbf{a}(X, \mathbf{Z})|Y, W, \mathbf{Z}\}V(U_a)|X, \mathbf{Z}] \\
& +\mathrm{E}^{(1)}_a[\mathrm{E}^{*(1)}_a\{V(U_a)\mathbf{b}(X)|Y, W_1, W_2, \mathbf{Z}\}V(U_a)|X, \mathbf{Z}].
\end{aligned}$$

The efficient score for $\boldsymbol{\gamma}$ is given as

$$\mathbf{S}^{*(1)}_{a,\text{eff},\boldsymbol{\gamma}}\{Y, W, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\} = \mathbf{S}^{*(1)}_{a,\boldsymbol{\gamma}}\{Y, W, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\} - \mathrm{E}^{*(1)}_a\{\mathbf{c}(X, \mathbf{Z})|Y, W, \mathbf{Z}\},$$

where $\mathbf{c}(X, \mathbf{Z})$ satisfies

$$\mathrm{E}^{(1)}_a[\mathbf{S}^{*(1)}_{a,\boldsymbol{\gamma}}\{Y, W, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\}|X, \mathbf{Z}] = \mathrm{E}^{(1)}_a[\mathrm{E}^{*(1)}_a\{\mathbf{c}(X, \mathbf{Z})|Y, W, \mathbf{Z}\}|X, \mathbf{Z}].$$

Then the corresponding estimating equations can be constructed based on $\mathbf{S}^{*(1)}_{a,\text{eff}}\{Y, W, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\}$ and $\mathbf{S}^{*(1)}_{a,\text{eff},\boldsymbol{\gamma}}\{Y, W, \mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\gamma}, f^*_{X|\mathbf{Z}}(\cdot)\}$.

Appendix B
Technical Proofs for Chapter 4

B.1 Additional Details Regarding Decorrelated Score Function

B.1.1 Verification of the orthogonality between the decorrelated score function and nuisance score functions

Since $E\{S(\boldsymbol{\theta}_0)\} = 0$ and $E\{\mathbf{S}_\gamma(\boldsymbol{\theta}_0)\} = \mathbf{0}$, we have

$$\begin{aligned}
\text{cov}\{S(\boldsymbol{\theta}_0), \mathbf{S}_\gamma(\boldsymbol{\theta}_0)\} &= E\{S(\boldsymbol{\theta}_0)\mathbf{S}_\gamma(\boldsymbol{\theta}_0)^{\mathrm{T}}\} \\
&= E[\{S_\beta(\boldsymbol{\theta}_0) - \boldsymbol{\omega}^{\mathrm{T}}\mathbf{S}_\gamma(\boldsymbol{\theta}_0)\}\mathbf{S}_\gamma(\boldsymbol{\theta}_0)^{\mathrm{T}}] \\
&= \mathbf{I}_{\beta\gamma} - \boldsymbol{\omega}^{\mathrm{T}}\mathbf{I}_{\gamma\gamma} \\
&= \mathbf{0}.
\end{aligned}$$

B.1.2 Estimation of the variance of regression error

From (4.2), we have $\epsilon_i = Y_i - (W_i - U_i)\beta_0 - \boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{Z}_i$. Since $E(\epsilon_i) = 0$, then

$$\sigma_\epsilon^2 = E(\epsilon_i^2) = E\{(Y_i - \beta_0 W_i - \boldsymbol{\gamma}_0^{\mathrm{T}} Z_i)^2\} - \beta_0^2\sigma_U^2.$$

Under null hypothesis, the estimated variance is given as

$$\widehat{\sigma}_{\epsilon,H_0}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \beta^* W_i - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)^2 - \beta^{*2}\sigma_U^2.$$

Plugging the one-step estimator $\widehat{\beta}$, we have

$$\widehat{\sigma}_\epsilon^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \widehat{\beta}W_i - \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)^2 - \widehat{\beta}^2\sigma_U^2.$$

B.2 Proofs Regarding Properties of the Initial Estimator

B.2.1 Proof of Lemma 4.2.1

Proof. The general idea of the proof closely resembles the proof of Theorem 1 in Datta et al. (2017). They assumed deterministic data matrix but here we assume that the data matrix is random and sub-Gaussian.

From the definition of $\widetilde{\boldsymbol{\theta}}$ in (4.8), we have

$$\frac{1}{2}\widetilde{\boldsymbol{\theta}}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\theta}} - \widehat{\boldsymbol{\rho}}\widetilde{\boldsymbol{\theta}} + \lambda\|\widetilde{\boldsymbol{\theta}}\|_1 \leqslant \frac{1}{2}\boldsymbol{\theta}_0^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\theta}_0 - \widehat{\boldsymbol{\rho}}\boldsymbol{\theta}_0 + \lambda\|\boldsymbol{\theta}_0\|_1.$$

Let $\widetilde{\mathbf{v}} = \widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$, then we have

$$\begin{aligned}\frac{1}{2}\widetilde{\mathbf{v}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{v}} + \lambda\|\widetilde{\boldsymbol{\theta}}\|_1 &\leqslant \widetilde{\mathbf{v}}^{\mathrm{T}}(\widehat{\boldsymbol{\rho}} - \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\theta}_0) + \lambda\|\boldsymbol{\theta}_0\|_1 \\ &\leqslant \|\widetilde{\mathbf{v}}\|_1\|\widehat{\rho} - \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\theta}_0\|_\infty + \lambda\|\boldsymbol{\theta}_0\|_1. \end{aligned} \tag{B.1}$$

In order to obtain an upper bound for the left-hand side above, we first bound the quantity $\|\widehat{\boldsymbol{\rho}} - \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\theta}_0\|_\infty$. By triangle inequality, we have

$$\|\widehat{\boldsymbol{\rho}} - \widetilde{\boldsymbol{\Sigma}}\boldsymbol{\theta}_0\|_\infty \leqslant \|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_\infty + \|\boldsymbol{\rho} - \boldsymbol{\Sigma}\boldsymbol{\theta}_0\|_\infty + \|\{\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\}\boldsymbol{\theta}_0\|_\infty + \|\{\widehat{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Sigma}}\}\boldsymbol{\theta}_0\|_\infty.$$

Since $\widehat{\boldsymbol{\rho}} = (\mathbf{W}, \mathbf{Z})^{\mathrm{T}}\mathbf{Y}/n$ and $\boldsymbol{\rho} = (\mathbf{X}, \mathbf{Z})^{\mathrm{T}}\mathbf{Y}/n$, then

$$\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_\infty = \left|\frac{1}{n}\sum_{i=1}^{n} U_i Y_i\right|.$$

Recall that $Y_i = X_i\beta_0 + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_0 + \epsilon_i$. Since X_i, $\mathbf{Z}_i$ and ϵ_i are sub-Gaussian and $\|\boldsymbol{\theta}_0\|_1/s_0 \leqslant K_0$ which is finite, by triangle inequality we have that $\|Y_i/s_0\|_{\psi_2} \leqslant \|\boldsymbol{\theta}_0\|_1 K/s_0 + K_\epsilon/s_0 \leqslant K_0 K + K_\epsilon/s_0 \leqslant \infty$. Hence, by Lemma B.6.4 we have

$$\begin{aligned}\|U_i Y_i/s_0\|_{\psi_1} &\leqslant 2\|U_i\|_{\psi_2}\|Y_i/s_0\|_{\psi_2} \\ &\leqslant 2K_U(K_0 K + K_\epsilon/s_0)\end{aligned}$$

$$
\begin{aligned}
&\leqslant 2K_U(K_0K + K_\epsilon) \\
&\leqslant \infty.
\end{aligned}
$$

Let $K_1 = 2K_U(K_0K + K_\epsilon)$. Since $E(U_iY_i) = 0$, by Bernstein inequality, for any $t > 0$ we have

$$
\begin{aligned}
\Pr(\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_\infty / s_0 \geqslant t) &= \Pr\left(\frac{1}{n}\left|\sum_{i=1}^n \frac{U_iY_i}{s_0}\right| \geqslant t\right) \\
&\leqslant 2\exp\left\{-C'' \min\left(\frac{t^2}{K_1^2}, \frac{t}{K_1}\right) n\right\}.
\end{aligned}
$$

Let $t = \lambda/(8s_0) = o(1)$. When $\lambda \leqslant 8s_0K_1$, we have

$$
\Pr\left(\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_\infty \geqslant \frac{\lambda}{8}\right) \leqslant 2\exp\left(-C''\frac{\lambda^2 n}{64s_0^2K_1^2}\right).
$$

For the term $\|\boldsymbol{\rho} - \boldsymbol{\Sigma}\boldsymbol{\theta}_0\|_\infty$, we know that $\boldsymbol{\rho} - \boldsymbol{\Sigma}\boldsymbol{\theta}_0 = n^{-1}\sum_{i=1}^n (X_i, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}}\epsilon_i$. By Assumption 4.2.1 and Lemma B.6.4, we have $\|X_i\epsilon_i\|_{\psi_1} \leqslant 2\|X_i\|_{\psi_2}\|\epsilon_i\|_{\psi_2} \leqslant 2KK_\epsilon$, and $\|Z_{ij}\epsilon_i\|_{\psi_1} \leqslant 2\|Z_{ij}\|_{\psi_2}\|\epsilon_i\|_{\psi_2} \leqslant 2KK_\epsilon$, for $i = 1, \ldots, n$ and $j = 1, \ldots, p-1$. When $\lambda \leqslant 16KK_\epsilon$, by union bound inequality and Bernstein inequality, we have

$$
\begin{aligned}
\Pr\left(\|\boldsymbol{\rho} - \boldsymbol{\Sigma}\boldsymbol{\theta}_0\|_\infty \geqslant \frac{\lambda}{8}\right) &\leqslant 2p\exp\left\{-C''\min\left(\frac{\lambda^2}{256K^2K_\epsilon^2}, \frac{\lambda}{16KK_\epsilon}\right) n\right\} \\
&\leqslant 2p\exp\left(\frac{-C''\lambda^2 n}{256K^2K_\epsilon^2}\right).
\end{aligned}
$$

For the term $\|\{\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\}\boldsymbol{\theta}_0\|_\infty$, we first have $\|\{\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}})\}\boldsymbol{\theta}_0\|_\infty \leqslant \|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\|_{\max}\|\boldsymbol{\theta}_0\|_1 \leqslant \|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\|_{\max}s_0K_0$. By Assumption 4.2.1, we have $\|U_iZ_{ij}\|_{\psi_2} \leqslant 2KK_U$, and $\|W_i^2 - X_i^2 - \sigma_U^2\|_{\psi_2} \leqslant 4K(K + K_U) + 2K_U^2 + \sigma_U^2$, for $j = 1, \ldots, p-1$ and $i = 1, \ldots, n$. Let $K_2 \equiv 4K(K + K_U) + 2K_U^2 + \sigma_U^2$. Note that $E(\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}) = \mathbf{0}$. When $\lambda \leqslant 8s_0K_0K_2$, by union bound inequality and Bernstein inequality, we have

$$
\begin{aligned}
\Pr\left(\|\{\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\}\boldsymbol{\theta}_0\|_\infty \geqslant \frac{\lambda}{8}\right) &\leqslant \Pr\left(\|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\|_{\max} \geqslant \frac{\lambda}{8s_0K_0}\right) \\
&\leqslant 2(2p-1)\exp\left\{-C''\min\left(\frac{\lambda^2}{64s_0^2K_0^2K_2^2}, \frac{\lambda}{8s_0K_0K_2}\right) n\right\} \\
&\leqslant 2(2p-1)\exp\left(\frac{-C''\lambda^2 n}{64s_0^2K_0^2K_2^2}\right). \qquad \text{(B.2)}
\end{aligned}
$$

By the definition of $\widetilde{\mathbf{\Sigma}}$ which is the nearest positive semi-definite matrix of $\widehat{\mathbf{\Sigma}}$ with respect to $\|\cdot\|_{\max}$ and the assumption that $E(\mathbf{\Sigma})$ is positive semi-definite, we have

$$\|\widehat{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}}\|_{\max} \leqslant \|E(\mathbf{\Sigma}) - \widehat{\mathbf{\Sigma}}\|_{\max}. \tag{B.3}$$

By Assumption 4.2.1, we have $\|W_i^2 - \sigma_U^2\|_{\psi_1} \leqslant 2(K+K_U)^2 + \sigma_U^2$, $\|W_i Z_{ij}\|_{\psi_1} \leqslant 2K(K+K_U)$ and $\|Z_{ij}Z_{ik}\|_{\psi 1} \leqslant 2K^2$. Let $K_3 \equiv 2\max\{2(K+K_U)^2 + \sigma_U^2, 2K(K+K_U), 2K^2\} = 4(K+K_U)^2 + 2\sigma_U^2$. When $\lambda \leqslant 8s_0K_0K_3$, by union bound inequality and Bernstein inequality, we have

$$\begin{aligned}
\Pr\left(\|\{\widehat{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}}\}\boldsymbol{\theta}_0\|_\infty \geqslant \frac{\lambda}{8}\right) &\leqslant \Pr\left(\|\widehat{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}}\|_{\max}\|\boldsymbol{\theta}_0\|_1 \geqslant \frac{\lambda}{8}\right) \\
&\leqslant \Pr\left(\|\widehat{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}}\|_{\max} \geqslant \frac{\lambda}{8s_0K_0}\right) \\
&\leqslant \Pr\left(\|E(\mathbf{\Sigma}) - \widehat{\mathbf{\Sigma}}\|_{\max} \geqslant \frac{\lambda}{8s_0K_0}\right) \\
&\leqslant 2p^2\exp\left\{-C''\min\left(\frac{\lambda^2}{64s_0^2K_0^2K_3^2}, \frac{\lambda}{8s_0K_0K_3}\right)n\right\} \\
&\leqslant 2p^2\exp\left(\frac{-C''\lambda^2 n}{64s_0^2K_0^2K_3^2}\right).
\end{aligned} \tag{B.4}$$

Then we obtain

$$\begin{aligned}
\Pr\left(\|\widehat{\rho} - \widetilde{\mathbf{\Sigma}}\boldsymbol{\theta}_0\|_\infty < \frac{\lambda}{2}\right) &> 1 - 2\exp\left(\frac{-C''\lambda^2 n}{64s_0^2K_1^2}\right) - 2p\exp\left(\frac{-C''\lambda^2 n}{256K^2K_\epsilon^2}\right) \\
&\quad -2(2p-1)\exp\left(\frac{-C''\lambda^2 n}{64s_0^2K_0^2K_2^2}\right) - 2p^2\exp\left(\frac{-C''\lambda^2 n}{64s_0^2K_0^2K_3^2}\right) \\
&\equiv 1 - p_1
\end{aligned}$$

for $\lambda \leqslant \min(8s_0K_1, 16KK_\epsilon, 8s_0K_0K_2, 8s_0K_0K_3) \leqslant \min(8K_1, 16KK_\epsilon, 8K_0K_2, 8K_0K_3)$. Recall that $\lambda = C_\lambda s_0\sqrt{n^{-1}\log p}$. It is easy to show that p_1 goes to 0 as n goes to infinity, if $C_\lambda > \max\left(8K_0K_2/C'', 8\sqrt{2}K_0K_3/\sqrt{C''}\right)$.

Back to equation (B.1), we now have

$$\begin{aligned}
\frac{1}{2}\widetilde{\mathbf{v}}^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}\widetilde{\mathbf{v}} + \lambda\|\widetilde{\boldsymbol{\theta}}\|_1 &\leqslant \|\widetilde{\mathbf{v}}\|_1\|\widetilde{\rho} - \widetilde{\mathbf{\Sigma}}\boldsymbol{\theta}_0\|_\infty + \lambda\|\boldsymbol{\theta}_0\|_1 \\
&\leqslant \frac{\lambda}{2}\|\widetilde{\mathbf{v}}\|_1 + \lambda\|\boldsymbol{\theta}_0\|_1.
\end{aligned}$$

Since $\boldsymbol{\theta}_{0S^C} = \mathbf{0}$, then $\widetilde{\mathbf{v}}_{S^C} = \widetilde{\boldsymbol{\theta}}_{S^C}$ and $\|\boldsymbol{\theta}_0\|_1 = \|\boldsymbol{\theta}_{0S}\|_1$. We have

$$\frac{1}{2}\widetilde{\mathbf{v}}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{v}} + \lambda\|\widetilde{\boldsymbol{\theta}}_S\|_1 + \lambda\|\widetilde{\mathbf{v}}_{S^C}\|_1 \leqslant \frac{\lambda}{2}\|\widetilde{\mathbf{v}}_S\|_1 + \frac{\lambda}{2}\|\widetilde{\mathbf{v}}_{S^C}\|_1 + \lambda\|\boldsymbol{\theta}_{0S}\|_1.$$

By triangle inequality, $\|\widetilde{\boldsymbol{\theta}}_S\|_1 \geqslant \|\boldsymbol{\theta}_{0S}\|_1 - \|\widetilde{\mathbf{v}}_S\|_1$. Then we have

$$\widetilde{\mathbf{v}}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{v}} + \lambda\|\widetilde{\mathbf{v}}_{S^C}\|_1 \leqslant 3\lambda\|\widetilde{\mathbf{v}}_S\|_1. \tag{B.5}$$

Since $\widetilde{\mathbf{v}}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{v}} \geqslant 0$, then $\|\widetilde{\mathbf{v}}_{S^C}\|_1 \leqslant 3\|\widetilde{\mathbf{v}}_S\|_1$. Lemma B.3.2 implies that with probability at least $1 - 2p^{-\zeta}$, $\|\widetilde{\mathbf{v}}_S\|_1^2 \leqslant s_0\|\widetilde{\mathbf{v}}\|_2^2 \leqslant s_0\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{\Sigma}\widetilde{\mathbf{v}}/\kappa$. The remainder of the proof is derived with probability at least $1 - p_1 - 2p^{-\zeta}$. Hence, we obtain

$$\begin{aligned}
\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{\Sigma}\widetilde{\mathbf{v}} + \lambda\|\widetilde{\mathbf{v}}\|_1 &= \widetilde{\mathbf{v}}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\mathbf{v}} + \lambda\|\widetilde{\mathbf{v}}_S\|_1 + \lambda\|\widetilde{\mathbf{v}}_{S^C}\|_1 + \widetilde{\mathbf{v}}^{\mathrm{T}}(\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}})\widetilde{\mathbf{v}} \\
&\leqslant 4\lambda\|\widetilde{\mathbf{v}}_S\|_1 + \widetilde{\mathbf{v}}^{\mathrm{T}}(\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}})\widetilde{\mathbf{v}} \\
&\leqslant 4\lambda\sqrt{\frac{s_0\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{\Sigma}\widetilde{\mathbf{v}}}{\kappa}} + \widetilde{\mathbf{v}}^{\mathrm{T}}(\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}})\widetilde{\mathbf{v}} \\
&\leqslant \frac{\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{\Sigma}\widetilde{\mathbf{v}}}{4} + \frac{16\lambda^2 s_0}{\kappa} + |\widetilde{\mathbf{v}}^{\mathrm{T}}(\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}})\widetilde{\mathbf{v}}|.
\end{aligned}$$

The first inequality holds because of (B.5), the second inequality holds because Lemma B.3.2 and the last inequality holds because $4ab \leqslant a^2/4 + 16b^2$. Note that

$$\begin{aligned}
|\widetilde{\mathbf{v}}^{\mathrm{T}}(\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}})\widetilde{\mathbf{v}}| &\leqslant \|\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}}\|_{\max}\|\widetilde{\mathbf{v}}\|_1^2 \\
&= \|\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}}\|_{\max}(\|\widetilde{\mathbf{v}}_S\|_1 + \|\widetilde{\mathbf{v}}_{S^C}\|_1)^2 \\
&\leqslant 16\|\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}}\|_{\max}\|\widetilde{\mathbf{v}}_S\|_1^2 \\
&\leqslant 16\|\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}}\|_{\max}\frac{s_0\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{\Sigma}\widetilde{\mathbf{v}}}{\kappa}.
\end{aligned}$$

By triangle inequality and (B.3), we have

$$\begin{aligned}
\|\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}}\|_{\max} &\leqslant \|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\|_{\max} + \|\widehat{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Sigma}}\|_{\max} \\
&\leqslant \|\boldsymbol{\Sigma} - \widehat{\boldsymbol{\Sigma}}\|_{\max} + \|E(\boldsymbol{\Sigma}) - \widehat{\boldsymbol{\Sigma}}\|_{\max}.
\end{aligned}$$

By (B.2) and (B.4), we have

$$\Pr\left(16s_0\|\boldsymbol{\Sigma} - \widetilde{\boldsymbol{\Sigma}}\|_{\max} < \frac{\kappa}{4}\right)$$

$$
\begin{aligned}
&= \Pr\left(\|\boldsymbol{\Sigma}-\widetilde{\boldsymbol{\Sigma}}\|_{\max} < \frac{\kappa}{64s_0}\right) \\
&\geqslant \Pr\left(\|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}\|_{\max} + \|\widehat{\boldsymbol{\Sigma}}-E(\boldsymbol{\Sigma})\|_{\max} < \frac{\kappa}{64s_0}\right) \\
&\geqslant 1 - \Pr\left(\|\boldsymbol{\Sigma}-\widehat{\boldsymbol{\Sigma}}\|_{\max} \geqslant \frac{\kappa}{128s_0}\right) - \Pr\left(\|\widehat{\boldsymbol{\Sigma}}-E(\boldsymbol{\Sigma})\|_{\max} \geqslant \frac{\kappa}{128s_0}\right) \\
&\geqslant 1 - 2(2p-1)\exp\left\{-C''\min\left(\frac{\kappa^2}{128^2 s_0^2 K_2^2}, \frac{\kappa}{128 s_0 K_2}\right)n\right\} \\
&\quad -2p^2\exp\left\{-C''\min\left(\frac{\kappa^2}{128^2 s_0^2 K_2^3}, \frac{\kappa}{128 s_0 K_3}\right)n\right\} \\
&\equiv 1 - p_2.
\end{aligned}
$$

Since $s_0\sqrt{n^{-1}\log p} = o(1)$, then p_2 goes to zero as n goes to infinity. Hence, with probability at least $1-p_1-p_2-2p^{-\zeta}$, we have

$$
\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{\Sigma}\widetilde{\mathbf{v}} + \lambda\|\widetilde{\mathbf{v}}\|_1 \leqslant \frac{\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{\Sigma}\widetilde{\mathbf{v}}}{4} + \frac{16\lambda^2 s_0}{\kappa} + \frac{\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{\Sigma}\widetilde{\mathbf{v}}}{4}.
$$

Then we obtain

$$
\frac{\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{\Sigma}\widetilde{\mathbf{v}}}{2} + \lambda\|\widetilde{\mathbf{v}}\|_1 \leqslant \frac{16\lambda^2 s_0}{\kappa}. \tag{B.6}
$$

Since $\widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{\Sigma}\widetilde{\mathbf{v}}/2 > 0$ by Lemma B.3.2, then $\|\widetilde{\mathbf{v}}\|_1 = \|\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_1 \leqslant 16\lambda s_0/\kappa = 16C_\lambda s_0^2\sqrt{n^{-1}\log p}/\kappa$ with probability at least $1-p_1-p_2-2p^{-\zeta}$.

For the prediction error $(\mathbf{X},\mathbf{Z})\widetilde{\mathbf{v}}$, since $\lambda\|\widetilde{\mathbf{v}}\|_1 \geqslant 0$, using (B.6), we have

$$
\begin{aligned}
\frac{1}{n}\|(\mathbf{X},\mathbf{Z})\widetilde{\mathbf{v}}\|_2^2 &= \frac{1}{n}\sum_{i=1}^{n}\{(X_i,\mathbf{Z}_i^{\mathrm{T}})\widetilde{\mathbf{v}}\}^2 \\
&= \widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{\Sigma}\widetilde{\mathbf{v}} \\
&\leqslant \frac{32\lambda^2 s_0}{\kappa}.
\end{aligned}
$$

Therefore, $\|(\mathbf{X},\mathbf{Z})(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\|_2/\sqrt{n} \leqslant \lambda\sqrt{32s_0/\kappa}$ with probability at least $1-p_1-p_2-2p^{-\zeta}$. Further, since $\|\widetilde{\mathbf{v}}\|_2^2 \leqslant \widetilde{\mathbf{v}}^{\mathrm{T}}\boldsymbol{\Sigma}\widetilde{\mathbf{v}}/\kappa$, then

$$
\|\widetilde{\mathbf{v}}\|_2^2 \leqslant \frac{32\lambda^2 s_0}{\kappa^2},
$$

with probability at least $1-p_1-p_2-2p^{-\zeta}$.

□

B.2.2 Proof of Lemma 4.2.2

Analogous to Lemma 1 in Wainwright (2009*b*) and Lemma 5 in Datta et al. (2017), we have the following lemma:

Lemma B.2.1. *Let $\partial\|\mathbf{x}\|_1$ denotes the sub-gradient of $\|\mathbf{x}\|_1$ for any vector $\mathbf{x}$. Then we have the following results:*

(a) *$\widetilde{\boldsymbol{\theta}}$ is the optimal solution to $\widetilde{l}(\boldsymbol{\theta}) = \boldsymbol{\theta}^{\mathrm{T}}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{\theta}/2 - \widehat{\boldsymbol{\rho}}^{\mathrm{T}}\boldsymbol{\theta} + \lambda\|\boldsymbol{\theta}\|_1$ iff there exists a vector $\mathbf{b}$ in $\partial\|\widetilde{\boldsymbol{\theta}}\|_1$ such that*

$$\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{\theta}} - \widehat{\rho} + \lambda\mathbf{b} = \mathbf{0}.$$

(b) *If $|b_j| < 1$ for any $j \notin \mathrm{supp}(\widetilde{\boldsymbol{\theta}})$, then any other optimal solution $\widetilde{\boldsymbol{\theta}}'$ will have support $\mathrm{supp}(\widetilde{\boldsymbol{\theta}}') \subseteq \mathrm{supp}(\widetilde{\boldsymbol{\theta}})$.*

(c) *Under the conditions of part* (b)*, if matrix $\widetilde{\boldsymbol{\Sigma}}_{\mathrm{supp}(\widetilde{\boldsymbol{\theta}}),\mathrm{supp}(\widetilde{\boldsymbol{\theta}})}$ is invertible, then $\widetilde{\boldsymbol{\theta}}$ is the unique optimal solution.*

Proof. Since matrix $\widetilde{\boldsymbol{\Sigma}} \succcurlyeq \mathbf{0}$, the proof is analogous to the proof of Lemma 1 in Wainwright (2009*b*). □

Assume $\widetilde{\boldsymbol{\Sigma}}_{S,S}$ is invertible and assume Steps 1 through 3 of the Primal and Dual Witness (PDW) method succeed (Wainwright 2009*b*). By part (a) of Lemma 2 in Wainwright (2009*b*), we know that Lasso $\widetilde{l}(\boldsymbol{\theta})$ has a unique solution $\widetilde{\boldsymbol{\theta}}$ with $S(\widetilde{\boldsymbol{\theta}}) \subseteq S(\boldsymbol{\theta}_0)$.

Different from the derivations in Wainwright (2009*b*), we need to consider the properties of $\widetilde{\boldsymbol{\Sigma}}$ instead of $\boldsymbol{\Sigma}$. Define $\mathbf{G} = E(\boldsymbol{\Sigma}_{S^C,S})E(\boldsymbol{\Sigma}_{S,S})^{-1}$, $\widetilde{\mathbf{G}} = \widetilde{\boldsymbol{\Sigma}}_{S^C,S}\widetilde{\boldsymbol{\Sigma}}_{S,S}^{-1}$, $\mathbf{H} = \widetilde{\mathbf{G}} - \mathbf{G}$, $\mathbf{F} = \widetilde{\boldsymbol{\Sigma}}_{S,S}^{-1} - E(\boldsymbol{\Sigma}_{S,S})^{-1}$, $\mathbf{F}_1 = \boldsymbol{\Sigma}_{S,S}^{-1} - E(\boldsymbol{\Sigma}_{S,S})^{-1}$ and $\mathbf{D} = \widetilde{\boldsymbol{\Sigma}} - E(\boldsymbol{\Sigma})$. The following Lemmas discussed the invertibility of $\widetilde{\boldsymbol{\Sigma}}_{S,S}$, difference between $\widetilde{\mathbf{G}}$ and $\mathbf{G}$ and that between $\widetilde{\boldsymbol{\Sigma}}_{S,S}^{-1}$ and $E(\boldsymbol{\Sigma}_{S,S})^{-1}$.

Lemma B.2.2. $\Pr(\widetilde{\boldsymbol{\Sigma}}_{S,S} > 0) \geqslant 1 - 2p^2\exp\{-C''n\delta^2/(4s_0^2K_3^2)\}$*, where* $K_3 = 4(K + K_U)^2 + 2\sigma_U^2$ *and* $\delta = \min(2K_3, \kappa_S/2)$.

Proof. By inequalities among different types of matrix norm, we have

$$\begin{aligned}
\lambda_{\min}(\widetilde{\boldsymbol{\Sigma}}_{S,S}) &\geqslant \lambda_{\min}\{E(\boldsymbol{\Sigma}_{S,S})\} - |\lambda_{\max}(-\mathbf{D}_{S,S})| \\
&\geqslant \kappa_S - \|\mathbf{D}_{S,S}\|_2 \\
&\geqslant \kappa_S - s_0\|\mathbf{D}_{S,S}\|_{\max}
\end{aligned}$$

$$\geqslant \quad \kappa_S - s_0\|\mathbf{D}\|_{\max}.$$

Since $\delta = \min(2K_3, \kappa_S/2)$, by the definition of $\widetilde{\mathbf{\Sigma}}$ we have

$$\begin{aligned}
\Pr(s_0\|\mathbf{D}\|_{\max} \geqslant \kappa_S/2) &= \Pr(2s_0\|\widehat{\mathbf{\Sigma}} - E(\mathbf{\Sigma})\|_{\max} \geqslant \kappa_S/2) \\
&\leqslant \Pr(s_0\|\widehat{\mathbf{\Sigma}} - E(\mathbf{\Sigma})\|_{\max} \geqslant \delta/2) \\
&\leqslant 2p^2 \exp\left\{-C''n\min\left(\frac{\delta^2}{4s_0^2K_3^2}, \frac{\delta}{2s_0K_3}\right)\right\} \\
&\leqslant 2p^2 \exp\left(\frac{-C''n\delta^2}{4s_0^2K_3^2}\right), \qquad \text{(B.7)}
\end{aligned}$$

where $K_3 = 4(K + K_U)^2 + 2\sigma_U^2$. Therefore, $\lambda_{\min}(\widetilde{\mathbf{\Sigma}}_{S,S}) \geqslant \kappa_S/2$ with probability at least $1 - 2p^2\exp\{-C''n\delta^2/(4s_0^2K_3^2)\}$.

□

Lemma B.2.3. *If $\widehat{\mathbf{\Sigma}}$ and $\widehat{\boldsymbol{\rho}}$ satisfy closeness condition, then for every $\delta \leqslant \min(s_0K_3, 1/\phi)$,*

$$\begin{aligned}
\Pr(\|\mathbf{F}\|_\infty \geqslant \delta\phi^2(1-\phi\delta)^{-1}) &\leqslant 2p^2\exp\{-C''n\delta^2/(s_0^2K_3^2)\}, \\
\Pr(\|\mathbf{H}\|_\infty \geqslant \delta\phi(2-\gamma)(1-\phi\delta)^{-1}) &\leqslant 2p^2\exp\{-C''n\delta^2/(s_0^2K_3^2)\},
\end{aligned}$$

where $K_3 = 4(K + K_u)^2 + 2\sigma_u^2$.

Proof. This proof is analogous to that of Lemma 7 in Datta et al. (2017).

Let $\eta_1 = \|\mathbf{D}_{S,S}\|_\infty$, $\eta_2 = \|\mathbf{D}_{S^C,S}\|_\infty$ and $\eta_3 = \|\mathbf{F}\|_\infty$. Since $\sum_{j=1}^{s_0}|D_{ij}| \leqslant s_0\|\mathbf{D}\|_{\max}$ for $i = 1, \ldots p$, if $\|\mathbf{D}\|_{\max} \leqslant \delta/s_0$ then $\eta_1 \leqslant \delta$ and $\eta_2 \leqslant \delta$. Similar to (B.7), for $\delta \leqslant K_3$ we have

$$\begin{aligned}
\Pr(\eta_1 \leqslant \delta, \eta_2 \leqslant \delta) &\geqslant \Pr(\|\mathbf{D}\|_{\max} \leqslant \delta/s_0) \\
&= 1 - \Pr(\|\widehat{\mathbf{\Sigma}} - E(\mathbf{\Sigma})\|_{\max} \geqslant \delta/s_0) \\
&\geqslant 1 - 2p^2\exp\left\{-C''n\min\left(\frac{\delta^2}{s_0^2K_3^2}, \frac{\delta}{s_0K_3}\right)\right\} \\
&\geqslant 1 - 2p^2\exp\{-C''n\delta^2/(s_0^2K_3^2)\}. \qquad \text{(B.8)}
\end{aligned}$$

We have the decomposition

$$\begin{aligned}
& \|\widetilde{\mathbf{\Sigma}}_{S^C,S}\widetilde{\mathbf{\Sigma}}_{S,S}^{-1} - E(\mathbf{\Sigma}_{S^C,S})E(\mathbf{\Sigma}_{S,S})^{-1}\|_\infty \\
\leqslant\ & \|\widetilde{\mathbf{\Sigma}}_{S^C,S} - E(\mathbf{\Sigma}_{S^C,S})\|_\infty \times \|\widetilde{\mathbf{\Sigma}}_{S,S}^{-1} - E(\mathbf{\Sigma}_{S,S})^{-1}\|_\infty \\
+\ & \|\widetilde{\mathbf{\Sigma}}_{S^C,S} - E(\mathbf{\Sigma}_{S^C,S})\|_\infty \times \|E(\mathbf{\Sigma}_{S,S})^{-1}\|_\infty
\end{aligned}$$

$$
\begin{aligned}
&+ \quad \|E(\mathbf{\Sigma}_{S^C,S})E(\mathbf{\Sigma}_{S,S})^{-1}\|_\infty \times \|E(\mathbf{\Sigma}_{S,S}) - \widetilde{\mathbf{\Sigma}}_{S,S}\|_\infty \times \|E(\mathbf{\Sigma}_{S,S})^{-1}\|_\infty \\
&+ \quad \|E(\mathbf{\Sigma}_{S^C,S})E(\mathbf{\Sigma}_{S,S})^{-1}\|_\infty \times \|E(\mathbf{\Sigma}_{S,S}) - \widetilde{\mathbf{\Sigma}}_{S,S}\|_\infty \times \|\widetilde{\mathbf{\Sigma}}_{S,S}^{-1} - E(\mathbf{\Sigma}_{S,S})^{-1}\|_\infty \\
&\leqslant \quad \{(1-\gamma)\eta_1 + \eta_2\}(\phi + \eta_3).
\end{aligned}
$$

Moreover,

$$
\begin{aligned}
\eta_3 &\leqslant \|\widetilde{\mathbf{\Sigma}}_{S,S}^{-1}\|_\infty \times \|\widetilde{\mathbf{\Sigma}}_{S,S} - E(\mathbf{\Sigma}_{S,S})\|_\infty \times \|E(\mathbf{\Sigma}_{S,S})^{-1}\|_\infty \\
&\leqslant (\phi + \eta_3)\eta_1\phi.
\end{aligned}
$$

When $\eta_1\phi < 1$, we have $\eta_3 \leqslant \phi^2\eta_1(1-\eta_1\phi)^{-1}$ and hence

$$
\|\mathbf{H}\|_\infty = \|\widetilde{\mathbf{\Sigma}}_{S^C,S}\widetilde{\mathbf{\Sigma}}_{S,S}^{-1} - E(\mathbf{\Sigma}_{S^C,S})E(\mathbf{\Sigma}_{S,S})^{-1}\|_\infty \leqslant \{(1-\gamma)\eta_1 + \eta_2\}\phi(1-\phi\eta_1)^{-1}.
$$

By (B.8), we have

$$
\begin{aligned}
\Pr\{\|\mathbf{F}\|_\infty \leqslant \delta\phi^2(1-\phi\delta)^{-1}\} &\geqslant \Pr\{\|\mathbf{F}\|_\infty \leqslant \phi^2\eta_1(1-\eta_1\phi)^{-1}, \eta_1 \leqslant \delta, \eta_2 \leqslant \delta\} \\
&= \Pr(\eta_1 \leqslant \delta, \eta_2 \leqslant \delta) \\
&\geqslant 1 - 2p^2\exp\{-C''n\delta^2/(s_0^2K_3^2)\},
\end{aligned}
$$

and

$$
\begin{aligned}
\Pr(\|\mathbf{H}\|_\infty \leqslant \phi\delta(2-\gamma)(1-\phi\delta)^{-1}) &\geqslant \Pr(\|\mathbf{H}\|_\infty \leqslant \{(1-\gamma)\eta_1 + \eta_2\}\phi(1-\phi\eta_1)^{-1}, \eta_1 \leqslant \delta, \eta_2 \leqslant \delta) \\
&= \Pr(\eta_1 \leqslant \delta, \eta_2 \leqslant \delta) \\
&\geqslant 1 - 2p^2\exp\{-C''n\delta^2/(s_0^2K_3^2)\}.
\end{aligned}
$$

Therefore, when $\delta \leqslant \min(K_3, 1/\phi)$, we indeed have

$$
\begin{aligned}
\Pr(\|\mathbf{F}\|_\infty \geqslant \delta\phi^2(1-\phi\delta)^{-1}) &\leqslant 2p^2\exp\{-C''n\delta^2/(s_0^2K_3^2)\}, \\
\Pr(\|\mathbf{H}\|_\infty \geqslant \delta\phi(2-\gamma)(1-\phi\delta)^{-1}) &\leqslant 2p^2\exp\{-C''n\delta^2/(s_0^2K_3^2)\}.
\end{aligned}
$$

□

Corollary B.2.4. *For every $\delta \leqslant \min(K^2, 1/\phi)$, we have*

$$
\Pr(\|\mathbf{F}_1\|_\infty \geqslant \delta\phi^2(1-\phi\delta)^{-1}) \leqslant 2s_0^2\exp\{-C''n\delta^2/(16s_0^2K^4)\}.
$$

Proof. Let $\eta_4 = \|\mathbf{F}_1\|_\infty$ and $\eta_5 = \|\mathbf{\Sigma}_{S,S} - E(\mathbf{\Sigma}_{S,S})\|_\infty$. First we have

$$\begin{aligned} \eta_4 &\leqslant \|\mathbf{\Sigma}_{S,S}^{-1}\|_\infty \|\mathbf{\Sigma}_{S,S} - E(\mathbf{\Sigma}_{S,S})\|_\infty \|E(\mathbf{\Sigma}_{S,S})^{-1}\|_\infty \\ &\leqslant (\phi + \eta_4)\eta_5\phi. \end{aligned}$$

When $\eta_5\phi < 1$, we have $\eta_4 \leqslant \phi^2\eta_5(1-\eta_5\phi)^{-1}$. Since $\|Q_{ij}Q_{ik}\|_{\psi_1} \leqslant 2K^2$, $\|Q_{ij}Q_{ik} - [E(\mathbf{\Sigma}_{S,S})]_{jk}\|_{\psi_1} \leqslant 4K^2$ for $i = 1, \ldots, n$ and $j, k = 1, \ldots, s_0$. Then by union bound inequality and Bernstein inequality, for $\delta \leqslant 4K^2$ we have

$$\begin{aligned} \Pr(\eta_5 \leqslant \delta) &\geqslant \Pr(\|\mathbf{\Sigma}_{S,S} - E(\mathbf{\Sigma}_{S,S})\|_{\max} \leqslant \delta/s_0) \\ &\geqslant 1 - 2s_0^2 \exp\left\{-C'' \min\left(\frac{\delta^2}{16s_0^2K^4}, \frac{\delta}{4s_0K^2}\right) n\right\} \\ &\geqslant 1 - 2s_0^2 \exp\left\{-C''n\delta^2/(16s_0^2K^4)\right\}. \end{aligned}$$

Hence, for $\delta \leqslant \min(4K^2, 1/\phi)$, we obtain that

$$\begin{aligned} \Pr(\|\mathbf{F}_1\|_\infty \leqslant \delta\phi^2(1-\phi\delta)^{-1}) &\geqslant \Pr\{\|\mathbf{F}_1\|_\infty \leqslant \phi^2\eta_5(1-\eta_5\phi)^{-1}, \eta_5 \leqslant \delta\} \\ &= \Pr(\eta_5 \leqslant \delta) \\ &\geqslant 1 - 2s_0^2 \exp\left\{-C''n\delta^2/(16s_0^2K^4)\right\}. \end{aligned}$$

Then $\Pr(\|\mathbf{F}_1\|_\infty \geqslant \delta\phi^2(1-\phi\delta)^{-1}) \leqslant 2s_0^2 \exp\left\{-C''n\delta^2/(16s_0^2K^4)\right\}$. □

Proof of Lemma 4.2.2 part (a):

Similar to the proof of Theorem 2 in Datta et al. (2017), Primal and Dual Witness construction technique is used.

Let $\widetilde{\boldsymbol{\theta}}_S$ be the solution to the restricted modified Lasso problem, that is,

$$\widetilde{\boldsymbol{\theta}}_S = \text{argmin}_{\boldsymbol{\theta}_S \in \mathbb{R}^{s_0}} \left\{\frac{1}{2}\boldsymbol{\theta}_S^{\mathrm{T}}\widetilde{\mathbf{\Sigma}}_{S,S}\boldsymbol{\theta}_S - \widehat{\boldsymbol{\rho}}_S^{\mathrm{T}}\boldsymbol{\theta}_S + \lambda\|\boldsymbol{\theta}_S\|_1\right\}.$$

Let $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\theta}}_S^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{b} = (\mathbf{b}_S^{\mathrm{T}}, \mathbf{b}_{S^C}^{\mathrm{T}})^{\mathrm{T}}$ where $\mathbf{b}_S \in \partial\|\widetilde{\boldsymbol{\theta}}_S\|_1$ and $\mathbf{b}_{S^C}$ is some unspecified $(p - s_0) \times 1$ vector. From part (a) of Lemma B.2.1, $\widetilde{\boldsymbol{\theta}}$ is an optimal solution to (4.8) iff $\widetilde{\boldsymbol{\theta}}$ and $\mathbf{b}$ satisfy

$$\begin{aligned} \widetilde{\mathbf{\Sigma}}_{S,S}\widetilde{\boldsymbol{\theta}}_S - \widehat{\boldsymbol{\rho}}_S + \lambda\mathbf{b}_S &= \mathbf{0}, \\ \widetilde{\mathbf{\Sigma}}_{S^C,S}\widetilde{\boldsymbol{\theta}}_S - \widehat{\boldsymbol{\rho}}_{S^C} + \lambda\mathbf{b}_{S^C} &= \mathbf{0}. \end{aligned} \tag{B.9}$$

Solving for $\widetilde{\boldsymbol{\theta}}_S$ and $\mathbf{b}_{S^C}$ from equation (B.9), we have

$$\widetilde{\boldsymbol{\theta}}_S = \widetilde{\boldsymbol{\Sigma}}_{S,S}^{-1}(\widehat{\boldsymbol{\rho}}_S - \lambda \mathbf{b}_S), \quad \text{and} \quad \mathbf{b}_{S^C} = \widetilde{\mathbf{G}}\mathbf{b}_S + \frac{1}{\lambda}(\widehat{\boldsymbol{\rho}}_{S^C} - \widetilde{\mathbf{G}}\widehat{\boldsymbol{\rho}}_S). \tag{B.10}$$

From part (b) and (c) of Lemma B.2.1, we know that if $\widetilde{\boldsymbol{\Sigma}}_{S,S}$ is nonsingular and $|\mathbf{b}_j| < 1$ for $j \in S^C$, then $\widetilde{\boldsymbol{\theta}}$ is the unique solution to (4.8) and $\text{supp}(\widetilde{\boldsymbol{\theta}}) \subseteq \text{supp}(\boldsymbol{\theta}_0)$. Lemma B.2.2 provides lower bound for $\Pr(\widetilde{\boldsymbol{\Sigma}}_{S,S} > 0)$. We now derive the lower bound for $\Pr(\|\mathbf{b}_{S^C}\|_\infty < 1)$. We expand $\mathbf{b}_{S^C}$ as

$$\begin{aligned}
\mathbf{b}_{S^C} &= \mathbf{G}\mathbf{b}_S + \mathbf{H}\mathbf{b}_S + \frac{1}{\lambda}\{(\widehat{\boldsymbol{\rho}}_{S^C} - \boldsymbol{\rho}_{S^C}) + (\boldsymbol{\rho}_{S^C} - \mathbf{G}\boldsymbol{\rho}_S) + \mathbf{G}(\boldsymbol{\rho}_S - \widehat{\boldsymbol{\rho}}_S) - \mathbf{H}\widehat{\boldsymbol{\rho}}_S\} \\
&= \mathbf{G}\mathbf{b}_S + \mathbf{H}\left(\mathbf{b}_S + \frac{1}{\lambda}(\boldsymbol{\rho}_S - \widehat{\boldsymbol{\rho}}_S) - \frac{1}{\lambda}\boldsymbol{\rho}_S\right) + \frac{1}{\lambda}\left\{(\widehat{\boldsymbol{\rho}}_{S^C} - \boldsymbol{\rho}_{S^C})\right. \\
&\quad \left. +(\boldsymbol{\rho}_{S^C} - \mathbf{G}\boldsymbol{\rho}_S) + \mathbf{G}(\boldsymbol{\rho}_S - \widehat{\boldsymbol{\rho}}_S)\right\}.
\end{aligned}$$

Since $\|\mathbf{b}_S\|_\infty = 1$, by triangle inequality we have

$$\begin{aligned}
\|\mathbf{b}_{S^C}\|_\infty &\leqslant \|\mathbf{G}\mathbf{b}_S\|_\infty + \|\mathbf{H}\|_\infty\left(1 + \frac{1}{\lambda}\|\widehat{\boldsymbol{\rho}}_S - \boldsymbol{\rho}_S\|_\infty + \frac{1}{\lambda}\|\boldsymbol{\rho}_S\|_\infty\right) \\
&\quad + \frac{1}{\lambda}\|\boldsymbol{\rho}_{S^C} - \mathbf{G}\boldsymbol{\rho}_S\|_\infty + \frac{1}{\lambda}\left\{\|\widehat{\boldsymbol{\rho}}_{S^C} - \boldsymbol{\rho}_{S^C}\|_\infty + \|\mathbf{G}(\widehat{\boldsymbol{\rho}}_S - \boldsymbol{\rho}_S)\|_\infty\right\}.
\end{aligned}$$

Recall that we assume that $\|\mathbf{G}\|_\infty \leqslant 1-\gamma$, then $\|\mathbf{G}\mathbf{b}_S\|_\infty \leqslant 1-\gamma$. By Bernstein inequality, we have

$$\begin{aligned}
&\Pr\left[\frac{1}{\lambda}\{\|\widehat{\boldsymbol{\rho}}_{S^C} - \boldsymbol{\rho}_{S^C}\|_\infty + \|\mathbf{G}(\widehat{\boldsymbol{\rho}}_S - \boldsymbol{\rho}_S)\|_\infty\} < \frac{\gamma}{2}\right] \\
\geqslant\ &\Pr\left[\frac{1}{\lambda}\{\|\widehat{\boldsymbol{\rho}}_{S^C} - \boldsymbol{\rho}_{S^C}\|_\infty + (1-\gamma)\|(\widehat{\boldsymbol{\rho}}_S - \boldsymbol{\rho}_S)\|_\infty\} < \frac{\gamma}{2}\right] \\
\geqslant\ &\Pr\left[\frac{1}{\lambda}(2-\gamma)\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_\infty < \frac{\gamma}{2}\right] \\
\geqslant\ &\Pr\left(\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_\infty < \frac{\gamma\lambda}{4}\right) \\
\geqslant\ &1 - 2\exp\left\{-C''\min\left(\frac{\gamma^2\lambda^2}{16 s_0^2 K_1^2}, \frac{\gamma\lambda}{4 s_0 K_1}\right) n\right\},
\end{aligned}$$

where $K_1 = 2K_U(K_0 K + K_\epsilon)$.

For term $\|\boldsymbol{\rho}_{S^C} - \mathbf{G}\boldsymbol{\rho}_S\|_\infty$, first we know

$$E(\boldsymbol{\rho}_{S^C} - \mathbf{G}\boldsymbol{\rho}_S) = \frac{1}{n}E(\mathbf{Q}_{S^C}^{\mathrm{T}}\mathbf{Q}_S\boldsymbol{\theta}_{0S} + \mathbf{Q}_{S^C}^{\mathrm{T}}\boldsymbol{\epsilon} - \mathbf{G}\mathbf{Q}_S^{\mathrm{T}}\mathbf{Q}_S\boldsymbol{\theta}_{0S} - \mathbf{G}\mathbf{Q}_S^{\mathrm{T}}\boldsymbol{\epsilon})$$

$$
\begin{aligned}
&= \frac{1}{n}E(\mathbf{Q}_{S^C}^{\mathrm{T}}\mathbf{Q}_S\boldsymbol{\theta}_{0S}) - \frac{1}{n}E(\mathbf{Q}_{S^C}^{\mathrm{T}}\mathbf{Q}_S)E(\mathbf{Q}_S^{\mathrm{T}}\mathbf{Q}_S)^{-1}E(\mathbf{Q}_S^{\mathrm{T}}\mathbf{Q}_S)\boldsymbol{\theta}_{0S} \\
&\quad + \frac{1}{n}E(\mathbf{Q}_{S^C}^{\mathrm{T}}\boldsymbol{\epsilon} - \mathbf{G}\mathbf{Q}_S^{\mathrm{T}}\boldsymbol{\epsilon}) \\
&= \frac{1}{n}E(\mathbf{Q}_{S^C}^{\mathrm{T}}\boldsymbol{\epsilon} - \mathbf{G}\mathbf{Q}_S^{\mathrm{T}}\boldsymbol{\epsilon}) \\
&= \mathbf{0}.
\end{aligned}
$$

The last equality holds because ϵ_i is independent of X_i and $\mathbf{Z}_i$. By Lemma B.6.4, we know $\|Q_{ij}Y_i/s_0\|_{\psi_1} \leqslant 2K\|Y_i/s_0\|_{\psi_2} \leqslant 2K(\|\boldsymbol{\theta}_0\|_1 K/s_0 + K_\epsilon/s_0) \leqslant 2K^2K_0 + 2KK_\epsilon/s_0$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Let $K_4 \equiv 2K^2K_0 + 2KK_\epsilon$. Then

$$
\begin{aligned}
\|s_0^{-1}[\boldsymbol{\rho}_{S^C} - \mathbf{G}\boldsymbol{\rho}_S]_j\|_{\psi_1} &\leqslant \|s_0^{-1}[\boldsymbol{\rho}_{S^C}]_j\|_{\psi_1} + \|s_0^{-1}[\mathbf{G}\boldsymbol{\rho}_S]_j\|_{\psi_1} \\
&\leqslant K_4 + \|\mathbf{G}\|_\infty K_4 \\
&\leqslant 2K_4.
\end{aligned}
$$

By union bound inequality and Bernstein inequality, we have

$$
\begin{aligned}
\Pr\left(\frac{1}{\lambda}\|\boldsymbol{\rho}_{S^C} - \mathbf{G}\boldsymbol{\rho}_S\|_\infty \geqslant \frac{\gamma}{4}\right) &= \Pr\left(\frac{1}{s_0}\|\boldsymbol{\rho}_{S^C} - \mathbf{G}\boldsymbol{\rho}_S\|_\infty \geqslant \frac{\gamma\lambda}{4s_0}\right) \\
&\leqslant 2(p - s_0)\exp\left\{-C''\min\left(\frac{\gamma^2\lambda^2}{64s_0^2K_4^2}, \frac{\gamma\lambda}{8s_0K_4}\right)n\right\}.
\end{aligned}
$$

Recall that $\lambda = C_\lambda s_0\sqrt{n^{-1}\log p}$. Then $2(p-s_0)\exp[-C''\min\{\gamma^2\lambda^2/(64s_0^2K_4^2), \gamma\lambda/(8s_0K_4)\}n]$ goes to zero, when $C_\lambda > 8K_4/(\gamma\sqrt{C''})$.

For term $\|\mathbf{H}\|_\infty(1 + \lambda^{-1}\|\boldsymbol{\rho}_S - \widehat{\boldsymbol{\rho}}_S\|_\infty + \lambda^{-1}\|\boldsymbol{\rho}_S\|_\infty)$, first we have

$$
\begin{aligned}
\|\boldsymbol{\rho}_S\|_\infty &= \frac{1}{n}\|\mathbf{Q}_S^{\mathrm{T}}\boldsymbol{\epsilon}\|_\infty + \frac{1}{n}\|\mathbf{Q}_S^{\mathrm{T}}\mathbf{Q}_S\boldsymbol{\theta}_{0S}\|_\infty \\
&\leqslant \frac{1}{n}\|\mathbf{Q}_S^{\mathrm{T}}\boldsymbol{\epsilon}\|_\infty + \|\boldsymbol{\Sigma}_{S,S}\|_\infty\|\boldsymbol{\theta}_0\|_\infty \\
&\leqslant \frac{1}{n}\|\mathbf{Q}_S^{\mathrm{T}}\boldsymbol{\epsilon}\|_\infty + K_0\|\boldsymbol{\Sigma}_{S,S}\|_\infty.
\end{aligned}
$$

By the definition of $\widehat{\boldsymbol{\rho}}$, we have

$$
\begin{aligned}
\Pr(\|\widehat{\boldsymbol{\rho}}_S - \boldsymbol{\rho}_S\|_\infty \leqslant 1) &\geqslant \Pr(\|\widehat{\boldsymbol{\rho}} - \boldsymbol{\rho}\|_\infty \leqslant 1) \\
&\geqslant 1 - 2\exp\left\{-C''\min\left(\frac{1}{s_0^2K_1^2}, \frac{1}{s_0K_1}\right)n\right\}. \qquad \text{(B.11)}
\end{aligned}
$$

Recall that we assume $s_0\sqrt{n^{-1}\log p} = o(1)$ which implies $s_0^2/n = o(1)$, or equivalently

$n/s_0^2 \to \infty$. Then $2\exp\left[-C''\min\left\{(s_0^2K_1^2)^{-1},(s_0K_1)^{-1}\right\}n\right]$ goes to 0 as n goes to infinity. Since $\|Q_{ij}\epsilon_i\|_{\psi_1} \leqslant 2KK_\epsilon$, by union bound inequality and Bernstein inequality, we have

$$\Pr\left(\frac{1}{n}\|\mathbf{Q}_S^{\mathrm{T}}\boldsymbol{\epsilon}\|_\infty \leqslant 1\right) \geqslant 1 - 2s_0\exp\left\{-C''\min\left(\frac{1}{4K^2K_\epsilon^2},\frac{1}{2KK_\epsilon}\right)n\right\}. \tag{B.12}$$

Now we derive the upper bound for $\|\boldsymbol{\Sigma}_{S,S}\|_\infty$. First, recall that we assume $\|E(\boldsymbol{\Sigma}_{S,S})\|_\infty = \Phi$ in (4.9). By triangle inequality, we know $\|\boldsymbol{\Sigma}_{S,S}\|_\infty \leqslant \|\boldsymbol{\Sigma}_{S,S} - E(\boldsymbol{\Sigma}_{S,S})\|_\infty + \Phi$. Since $\|[\mathbf{Q}_{iS}\mathbf{Q}_{iS}^{\mathrm{T}} - E(\boldsymbol{\Sigma}_{S,S})]_{jk}\|_{\psi_1} \leqslant 4K^2$, for $i = 1,\ldots,n$ and $j,k = 1,\ldots,s_0$, then

$$\|s_0^{-1}\sum_{k=1}^{s_0}[\mathbf{Q}_{iS}\mathbf{Q}_{iS}^{\mathrm{T}} - E(\boldsymbol{\Sigma}_{S,S})]_{jk}\|_{\psi_1} \leqslant 4K^2,$$

for $j = 1,\ldots,s_0$. By union bound inequality and Bernstein inequality, we have

$$\Pr(\|\boldsymbol{\Sigma}_{S,S} - E(\boldsymbol{\Sigma}_{S,S})\|_\infty \leqslant 1) \geqslant 1 - 2s_0\exp\left\{-C''\min\left(\frac{1}{16s_0^2K^4},\frac{1}{4s_0K^2}\right)n\right\}.$$

Then

$$\Pr(\|\boldsymbol{\Sigma}_{S,S}\|_\infty \leqslant 1+\Phi) \geqslant 1 - 2s_0\exp\left\{-C''\min\left(\frac{1}{16s_0^2K^4},\frac{1}{4s_0K^2}\right)n\right\}. \tag{B.13}$$

Combining the above results with Lemma B.2.3, we have,

$$\|\mathbf{H}\|_\infty\left(1 + \frac{1}{\lambda}\|\widehat{\boldsymbol{\rho}}_S - \boldsymbol{\rho}_S\|_\infty + \frac{1}{\lambda}\|\boldsymbol{\rho}_S\|_\infty\right) \leqslant \frac{\delta\phi(2-\gamma)}{1-\phi\delta}\left[1 + \frac{1}{\lambda}\{2 + (1+\Phi)K_0\}\right] \leqslant \frac{\gamma}{8}$$

with probability at least

$$\begin{aligned}
&1 - 2\exp\left[-C''\min\left\{(s_0^2K_1^2)^{-1},(s_0K_1)^{-1}\right\}n\right]\\
&-2s_0\exp\left[-C''\min\left\{(4K^2K_\epsilon^2)^{-1},(2KK_\epsilon)^{-1}\right\}n\right]\\
&-2s_0\exp\left[-C''\min\left\{(16s_0^2K^4)^{-1},(4s_0K^2)^{-1}\right\}n\right] - 2p^2\exp\{-C''n\delta^2/(s_0^2K_3^2)\},
\end{aligned}$$

where $\delta = \min\{K_3, 1/\phi, \gamma(8\phi(2-\gamma)[1+\lambda^{-1}\{2+(1+\Phi)K_0\}] + \gamma\phi)^{-1}\}$.

In conclusion, we obtain that

$$\begin{aligned}
&\Pr\left(\|\mathbf{b}_{S^C}\|_\infty \geqslant 1 - \frac{\gamma}{8}\right)\\
\leqslant\ & 2\exp\left\{-C''\min\left(\frac{\gamma^2\lambda^2}{16s_0^2K_1^2},\frac{\gamma\lambda}{4s_0K_1}\right)n\right\}
\end{aligned}$$

$$
\begin{aligned}
&+2(p-s_0)\exp\left\{-C''\min\left(\frac{\gamma^2\lambda^2}{64s_0^2K_4^2},\frac{\gamma\lambda}{8s_0K_4}\right)n\right\}\\
&+2\exp\left\{-C''\min\left(\frac{1}{s_0^2K_1^2},\frac{1}{s_0K_1}\right)n\right\}\\
&+2s_0\exp\left\{-C''\min\left(\frac{1}{4K^2K_\epsilon^2},\frac{1}{2KK_\epsilon}\right)n\right\}\\
+\quad & 2s_0\exp\left\{-C''\min\left(\frac{1}{16s_0^2K^4},\frac{1}{4s_0K^2}\right)n\right\}+2p^2\exp\{-C''n\delta^2/(s_0^2K_3^2)\}\\
:=\quad & p_1(\delta),
\end{aligned}
$$

where $K_1 = 2K_U(K_0K + K_\epsilon)$, $K_3 = 4(K + K_u)^2 + 2\sigma_u^2$, $K_4 = 2K^2K_0 + 2KK_\epsilon$ and $\delta = \min\{K_3, 1/\phi, \gamma(8\phi(2-\gamma)[1+\lambda^{-1}\{2+(1+\Phi)K_0\}]+\gamma\phi)^{-1}\}$.

Proof of Lemma 4.2.2 part (b):

By (B.10), we expand $\widetilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0S}$ as

$$
\begin{aligned}
\widetilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0S} &= \widetilde{\boldsymbol{\Sigma}}_{S,S}^{-1}\left(\widehat{\boldsymbol{\rho}}_S - \boldsymbol{\rho}_S + \frac{1}{n}\mathbf{Q}_S^{\mathrm{T}}\mathbf{Q}_S\boldsymbol{\theta}_{0S} + \frac{1}{n}\mathbf{Q}_S^{\mathrm{T}}\boldsymbol{\epsilon} - \lambda\mathbf{b}_S\right) - \boldsymbol{\theta}_{0S}\\
&= \widetilde{\boldsymbol{\Sigma}}_{S,S}^{-1}\left(\widehat{\boldsymbol{\rho}}_S - \boldsymbol{\rho}_S + \frac{1}{n}\mathbf{Q}_S^{\mathrm{T}}\boldsymbol{\epsilon} - \lambda\mathbf{b}_S\right) + (\widetilde{\boldsymbol{\Sigma}}_{S,S}^{-1} - \boldsymbol{\Sigma}_{S,S}^{-1})\frac{1}{n}\mathbf{Q}_S^{\mathrm{T}}\mathbf{Q}_S\boldsymbol{\theta}_{0S}\\
&= \{\mathbf{F} + E(\boldsymbol{\Sigma}_{S,S})^{-1}\}\left(\widehat{\boldsymbol{\rho}}_S - \boldsymbol{\rho}_S + \frac{1}{n}\mathbf{Q}_S^{\mathrm{T}}\boldsymbol{\epsilon} - \lambda\mathbf{b}_S\right) + (\mathbf{F}-\mathbf{F}_1)\frac{1}{n}\mathbf{Q}_S^{\mathrm{T}}\mathbf{Q}_S\boldsymbol{\theta}_{0S}.
\end{aligned}
$$

By similar derivation as for (B.11), (B.12), we have

$$
\begin{aligned}
\Pr(\|\widehat{\boldsymbol{\rho}}_S - \boldsymbol{\rho}_S\|_\infty \leqslant \lambda) &\geqslant 1 - 2\exp\left\{-C''\min\left(\frac{\lambda^2}{s_0^2K_1^2},\frac{\lambda}{s_0K_1}\right)n\right\},\\
\Pr\left(\frac{1}{n}\|\mathbf{Q}_S^{\mathrm{T}}\boldsymbol{\epsilon}\|_\infty \leqslant \lambda\right) &\geqslant 1 - 2s_0\exp\left\{-C''\min\left(\frac{\lambda^2}{4K^2K_\epsilon^2},\frac{\lambda}{2KK_\epsilon}\right)n\right\},
\end{aligned}
$$

where $K_1 = 2K_U(K_0K + K_\epsilon)$. By (B.13), we have

$$
\begin{aligned}
\Pr\left\{\left\|\frac{1}{n}\mathbf{Q}_S^{\mathrm{T}}\mathbf{Q}_S\boldsymbol{\theta}_{0S}\right\|_\infty \leqslant (1+\Phi)K_0\right\} &\geqslant \Pr\left\{\left\|\frac{1}{n}\mathbf{Q}_S^{\mathrm{T}}\mathbf{Q}_S\boldsymbol{\theta}_{0S}\right\|_\infty \leqslant (1+\Phi)\|\boldsymbol{\theta}_{0S}\|_\infty\right\}\\
&\geqslant \Pr\left\{\left\|\frac{1}{n}\mathbf{Q}_S^{\mathrm{T}}\mathbf{Q}_S\right\|_\infty \leqslant 1+\Phi\right\}\\
&\geqslant 1 - 2s_0\exp\left\{-C''\min\left(\frac{1}{16s_0^2K^4},\frac{1}{4s_0K^2}\right)n\right\}.
\end{aligned}
$$

Combining with Lemma B.2.3, Corollary B.2.4, for $\delta' \leqslant \min[K_3, 4K^2, 1/(2\phi), \lambda\{\phi(1+$

$\Phi)K_0 + \phi\lambda\}^{-1}]$, we obtain

$$\begin{aligned}
\|\widetilde{\boldsymbol{\theta}}_S - \boldsymbol{\theta}_{0S}\|_\infty &\leqslant (\|\mathbf{F}\|_\infty + \|E(\boldsymbol{\Sigma}_{S,S})^{-1}\|_\infty)(\|\widehat{\boldsymbol{\rho}}_S - \boldsymbol{\rho}_S\|_\infty + \frac{1}{n}\|\mathbf{Q}_S^{\mathrm{T}}\boldsymbol{\epsilon}\|_\infty + \lambda\|\mathbf{b}_S\|_\infty) \\
&+ (\|\mathbf{F}\|_\infty + \|\mathbf{F}_1\|_\infty)\frac{1}{n}\|\mathbf{Q}_S^{\mathrm{T}}\mathbf{Q}_S\boldsymbol{\theta}_{0S}\|_\infty \\
&\leqslant \{\delta'\phi^2(1-\phi\delta')^{-1} + \phi\}3\lambda + 2\delta'\phi^2(1-\phi\delta')^{-1}(1+\Phi)K_0 \\
&\leqslant (\phi + \phi)3\lambda + 2\phi\lambda \\
&= 8\phi\lambda
\end{aligned}$$

with probability at least $1 - p_2(\delta')$, where

$$\begin{aligned}
p_2(\delta') &= 2s_0^2\exp\{-C''n\delta/(16s_0^2K^4)\} + 2p^2\exp\{-C''n\delta^2/(s_0^2K_3^2)\} \\
&+ 2\exp\left[-C''\min\left\{\lambda^2(s_0^2K_1^2)^{-1}, \lambda(s_0K_1)^{-1}\right\}n\right] \\
&+ 2s_0\exp\left[-C''\min\left\{\lambda^2(4K^2K_\epsilon^2)^{-1}, \lambda(2KK_\epsilon)^{-1}\right\}n\right] \\
&+ 2s_0\exp\left[-C''\min\left\{(16s_0^2K^4)^{-1}, (4s_0K^2)^{-1}\right\}n\right],
\end{aligned}$$

where $K_1 = 2K_U(K_0K + K_\epsilon)$ and $K_3 = 4(K + K_u)^2 + 2\sigma_u^2$.

B.3 Proofs Regarding Four Technical Conditions

B.3.1 Proof of Lemma 4.3.1

Lemma B.3.1. *Under the conditions of Lemma 4.3.1, with probability at least* $1 - 2p^{-1}$,

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(W_i\mathbf{Z}_i^{\mathrm{T}} - \boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}})\right\|_\infty \leqslant C_{\lambda'}\sqrt{\frac{logp}{n}},$$

for constant $C_{\lambda'} > \sqrt{2K_5^2/C''}$. *Here* C'' *is a universal constant and* $K_5 = 2K(K + K_U + KK_\omega)$, *and* K_ω *is the upper bound of* $\|\boldsymbol{\omega}\|_1$.

Proof. This proof follows the similar ideas of the proof of Lemma D2 in the supplementary materials to Ning et al. (2017) with some modifications.

Recall that W_i, Z_{ij} and $\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_i$ are all sub-Gaussian random variables. By Lemma B.6.4 and Assumption 4.2.1, we have $\|W_iZ_{ij}\|_{\psi_1} \leqslant 2\|W_i\|_{\psi_2}\|Z_{ij}\|_{\psi_2} \leqslant 2K(K + K_U)$ and $\|\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_iZ_{ij}\|_{\psi_2} \leqslant 2\|\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_i\|_{\psi_2}\|Z_{ij}\|_{\psi_2} \leqslant 2K^2K_\omega$, for $j = 1, \ldots, (p-1)$. By triangle inequality,

we have

$$\begin{aligned}\|W_iZ_{ij}-\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_iZ_{ij}\|_{\psi_1} &\leqslant \|W_iZ_{ij}\|_{\psi_1}+\|\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_iZ_{ij}\|_{\psi_1}\\ &\leqslant 2K(K+K_U+KK_\omega),\end{aligned}$$

for any $i=1,\ldots,n$ and $j=1,\ldots,p-1$. Let $K_5=2K(K+K_U+KK_\omega)$.

By Bernstein inequality in Lemma B.6.1 and the union bound inequality, we have that for any $t>0$,

$$\begin{aligned}\Pr\left(\left\|\frac{1}{n}\sum_{i=1}^{n}(W_i\mathbf{Z}_i^{\mathrm{T}}-\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}})\right\|_\infty\geqslant t\right) &\leqslant \sum_{j=1}^{p-1}\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n}(W_iZ_{ij}-\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_iZ_{ij})\right|\geqslant t\right)\\ &\leqslant 2(p-1)\exp\left\{-C''\min\left(\frac{t^2}{K_5^2},\frac{t}{K_5}\right)n\right\}\\ &\leqslant 2p\exp\left\{-C''\min\left(\frac{t^2}{K_5^2},\frac{t}{K_5}\right)n\right\},\end{aligned}$$

where $C''>0$ is a universal constant. Consider $t=C_{\lambda'}\sqrt{n^{-1}\log p}=o(1)$. Then

$$\begin{aligned}&\Pr\left(\left\|\frac{1}{n}\sum_{i=1}^{n}(W_i\mathbf{Z}_i^{\mathrm{T}}-\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}})\right\|_\infty\geqslant C_{\lambda'}\sqrt{n^{-1}\log p}\right)\\ \leqslant\ & 2p\exp\left\{-C''\min\left(\frac{C_{\lambda'}^2\log p}{K_5^2 n},\frac{C_{\lambda'}\sqrt{\log p}}{K_5\sqrt{n}}\right)n\right\}\\ \leqslant\ & 2p^{-\zeta},\end{aligned}$$

where $\zeta=C''C_{\lambda'}^2/K_5^2-1$. Thus we have proven that $\left\|\sum_{i=1}^{n}(W_i\mathbf{Z}_i^{\mathrm{T}}-\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}})/n\right\|_\infty\leqslant C_{\lambda'}\sqrt{n^{-1}\log p}$ holds with probability at least $1-2p^{-1}$, when $C''C_{\lambda'}^2/K_5^2>2$. □

Lemma B.3.2. *Under the assumption that $(s_0\vee s')\sqrt{n^{-1}\log p}=o(1)$, with probability at least $1-2p^{-\zeta}$, we have*

$$\begin{aligned}\kappa &\leqslant \mathrm{RE}=\min\left\{\frac{\mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v}}{\|\mathbf{v}\|_2^2}:\mathbf{v}\in\mathbb{R}^p\backslash\{0\},\|\mathbf{v}_{S^C}\|_1\leqslant 3\|\mathbf{v}_S\|_1\right\},\\ \kappa &\leqslant \mathrm{RE}'=\min\left\{\frac{\mathbf{v}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22}\mathbf{v}}{\|\mathbf{v}\|_2^2}:\mathbf{v}\in\mathbb{R}^{p-1}\backslash\{0\},\|\mathbf{v}_{S'^C}\|_1\leqslant 3\|\mathbf{v}_{S'}\|_1\right\},\end{aligned}$$

where $\boldsymbol{\Sigma}=n^{-1}\sum_{i=1}^{n}(X_i,\mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}\otimes 2}$, $\widehat{\boldsymbol{\Sigma}}_{22}=n^{-1}\sum_{i=1}^{n}\mathbf{Z}_i^{\otimes 2}$, ζ is a function of n and goes to ∞ as $n\to\infty$.

Proof. This proof follows the similar idea as the proof of Lemma J1 in the supplementary

materials to Ning et al. (2017) with some modifications.

For $\mathbf{v} \in \mathbb{R}^p \backslash \{\mathbf{0}\}$, we have

$$\begin{aligned}
\frac{\mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v}}{\|\mathbf{v}\|_2^2} &= \frac{\mathbf{v}^{\mathrm{T}}E(\boldsymbol{\Sigma})\mathbf{v}}{\|\mathbf{v}\|_2^2} + \frac{\mathbf{v}^{\mathrm{T}}\{\boldsymbol{\Sigma} - E(\boldsymbol{\Sigma})\}\mathbf{v}}{\|\mathbf{v}\|_2^2} \\
&\geqslant \lambda_{\min}\{E(\boldsymbol{\Sigma})\} - \left|\frac{\mathbf{v}^{\mathrm{T}}\{\boldsymbol{\Sigma} - E(\boldsymbol{\Sigma})\}\mathbf{v}}{\|\mathbf{v}\|_2^2}\right| \\
&= \lambda_{\min}[E\{(X_i, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}\otimes 2}\}] - \left|\frac{\mathbf{v}^{\mathrm{T}}\{\boldsymbol{\Sigma} - E(\boldsymbol{\Sigma})\}\mathbf{v}}{\|\mathbf{v}\|_2^2}\right| \\
&\geqslant 2\kappa - \frac{\|\mathbf{v}\|_1^2\|\boldsymbol{\Sigma} - E(\boldsymbol{\Sigma})\|_{\max}}{\|\mathbf{v}\|_2^2}.
\end{aligned}$$

The last inequality is because in Assumption 4.2.1, we assumed $\lambda_{\min}[E\{(X_i, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}\otimes 2}\}] \geqslant 2\kappa$. Since $\|\mathbf{v}_{S^C}\|_1 \leqslant 3\|\mathbf{v}_S\|_1$, then $\|\mathbf{v}\|_1^2 \leqslant 16\|\mathbf{v}_S\|_1^2 \leqslant 16 s_0\|\mathbf{v}\|_2^2$. We further have

$$\frac{\mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v}}{\|\mathbf{v}\|_2^2} \geqslant 2\kappa - 16 s_0\|\boldsymbol{\Sigma} - E(\boldsymbol{\Sigma})\|_{\max}.$$

By triangle inequality, Lemma B.6.2, Lemma B.6.4 and Assumption 4.2.1, we have $\|X_i^2 - E(X_i^2)\|_{\psi_1} \leqslant 4K^2$, $\|X_i Z_{ij} - E(X_i Z_{ij})\|_{\psi_1} \leqslant 4K^2$, and $\|Z_{ij}Z_{ik} - E(Z_{ij}Z_{ik})\|_{\psi_1} \leqslant 4K^2$ where $i = 1, \ldots, n$, $j = 1, \ldots, p-1$. By Bernstein inequality and the union bound inequality, we have

$$\begin{aligned}
\Pr\left(\frac{\mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v}}{\|\mathbf{v}\|_2^2} \geqslant \kappa\right) &\geqslant \Pr\left[2\kappa - 16 s_0\|\boldsymbol{\Sigma} - E(\boldsymbol{\Sigma})\|_{\max} \geqslant \kappa\right] \\
&= \Pr\left[\|\boldsymbol{\Sigma} - E(\boldsymbol{\Sigma})\|_{\max} \leqslant \kappa/(16 s_0)\right] \\
&\geqslant 1 - 2p^2 \exp\left\{-C'' \min\left(\frac{\kappa^2}{64^2 s_0^2 K^4}, \frac{\kappa}{64 s_0 K^2}\right) n\right\} \\
&= 1 - 2\exp\left\{2\log p - \min\left(\frac{C''\kappa^2 n}{64^2 s_0^2 K^4}, \frac{C''\kappa n}{64 s_0 K^2}\right)\right\},
\end{aligned}$$

where C'' is a universal constant. Define

$$\zeta \leqslant \min\left(\frac{C''\kappa n}{64 s_0 K^2 \log p} - 2, \frac{C''\kappa^2 n}{64^2 s_0^2 K^4 \log p} - 2\right).$$

Then it is easy to show that

$$\Pr\left(\frac{\mathbf{v}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{v}}{\|\mathbf{v}\|_2^2} \geqslant \kappa\right) = 1 - 2p^{-\zeta}.$$

By the assumption that $s_0\sqrt{n^{-1}\log p} = o(1)$, for n large enough, ζ is positive and goes to infinity as n goes to infinity. Hence, RE $\geqslant \kappa$ with probability at least $1 - 2p^{-\zeta}$, which goes to 1 when n goes to infinity.

Lemma B.6.5 implies that $\lambda_{\min}\{E(\mathbf{Z}_i^{\otimes 2})\} \geqslant 2\kappa$ as well. Similar to the above proofs, under the assumption that $s'\sqrt{\log p/n} = o(1)$, we can show that for any $\mathbf{v} \in \mathbb{R}^{p-1}\backslash\{0\}$ satisfying $\|\mathbf{v}_{S'^C}\|_1 \leqslant 3\|\mathbf{v}'_S\|_1$, $\mathbf{v}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22}\mathbf{v}/\|\mathbf{v}\|_2^2 \geqslant \kappa$ with probability at least $1 - 2p^{-\zeta}$, where

$$\zeta \leqslant \min\left(\frac{C''\kappa n}{64s'K^2\log p} - 2, \frac{C''\kappa^2 n}{64^2 s'^2 K^4 \log p} - 2\right).$$

□

Lemma B.3.3. *Under the conditions of Lemma 4.3.1, with probability at least* $1-2p^{-1}-2p^{-\zeta}$*, we have*

$$\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}\|_1 \leqslant 8C_{\lambda'}\kappa^{-1}s'\sqrt{\frac{log p}{n}}, \quad \textit{and} \quad (\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega})^{\mathrm{T}}\boldsymbol{\Sigma}_{22}(\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}) \leqslant \frac{16C_{\lambda'}^2}{\kappa}\frac{s' log p}{n},$$

where the constants $C_{\lambda'}$ *is given in Lemma B.3.1, and* ζ *is given in Lemma B.3.2.*

Proof. This proof follows the idea of Lemma D4 in the supplementary materials to Ning et al. (2017), with modifications.

Let S' denote the support set of $\boldsymbol{\omega}$. Recall the definition of Dantzig type estimator $\widehat{\boldsymbol{\omega}}$ in (4.6). $\widehat{\boldsymbol{\omega}}$ has the smallest L_1 norm among all $\boldsymbol{\omega}$ that satisfies $\|\widehat{\boldsymbol{\Sigma}}_{12} - \boldsymbol{\omega}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22}\|_\infty \leqslant \lambda'$. In Lemma B.3.1, we have proved that $\|\widehat{\boldsymbol{\Sigma}}_{12} - \boldsymbol{\omega}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22}\|_\infty \leqslant C_{\lambda'}\sqrt{n^{-1}\log p}$ with probability at least $1 - 2p^{-1}$. Let $\lambda' = C_{\lambda'}\sqrt{n^{-1}\log p}$. Then $\|\boldsymbol{\omega}\|_1 \geqslant \|\widehat{\boldsymbol{\omega}}\|_1$ with probability at least $1 - 2p^{-1}$. Further, since $\boldsymbol{\omega}_{S'^C} = \mathbf{0}$, we have $\|\boldsymbol{\omega}\|_1 = \|\boldsymbol{\omega}_{S'}\|_1 \geqslant \|\widehat{\boldsymbol{\omega}}_{S'}\|_1 + \|\widehat{\boldsymbol{\omega}}_{S'^C}\|_1$ with probability at least $1 - 2p^{-1}$. Denote $\widehat{\Delta} = \widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}$. By Lemma B.3.1 and definition of $\widehat{\boldsymbol{\omega}}$, we have

$$\|\widehat{\boldsymbol{\Sigma}}_{22}\widehat{\Delta}\|_\infty \leqslant \|\widehat{\boldsymbol{\Sigma}}_{21} - \widehat{\boldsymbol{\Sigma}}_{22}\widehat{\boldsymbol{\omega}}\|_\infty + \|\widehat{\boldsymbol{\Sigma}}_{21} - \widehat{\boldsymbol{\Sigma}}_{22}\boldsymbol{\omega}\|_\infty \leqslant 2C_{\lambda'}\sqrt{\frac{\log p}{n}},$$

with probability at least $1 - 2p^{-1}$. We also have $\|\widehat{\Delta}\|_1 \leqslant 2\|\widehat{\Delta}_{S'}\|_1 \leqslant 2s'^{1/2}\|\widehat{\Delta}_{S'}\|_2 \leqslant 2s'^{1/2}\|\widehat{\Delta}\|_2$. The first inequality holds because

$$\begin{aligned}
\|\widehat{\Delta}\|_1 &= \|\widehat{\Delta}_{S'} + \widehat{\Delta}_{S'^C}\|_1 \\
&\leqslant \|\widehat{\Delta}_{S'}\|_1 + \|\widehat{\Delta}_{S'^C}\|_1 \\
&\leqslant \|\widehat{\Delta}_{S'}\|_1 + \|\widehat{\boldsymbol{\omega}}_{S'^C}\|_1
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \|\widehat{\Delta}_{S'}\|_1 + \|\boldsymbol{\omega}_{S'}\|_1 - \|\widehat{\boldsymbol{\omega}}_{S'}\|_1 \\
&\leqslant \|\widehat{\Delta}_{S'}\|_1 + \|(\boldsymbol{\omega} - \widehat{\boldsymbol{\omega}})_{S'}\|_1 \\
&= 2\|\widehat{\Delta}_{S'}\|_1.
\end{aligned}
$$

This implies

$$
\widehat{\Delta}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22}\widehat{\Delta} \leqslant \|\widehat{\Delta}\|_1 \|\widehat{\boldsymbol{\Sigma}}_{22}\widehat{\Delta}\|_\infty \leqslant 2C_{\lambda'}\sqrt{\frac{\log p}{n}}\|\widehat{\Delta}\|_1 \leqslant 4C_{\lambda'}\sqrt{\frac{s'\log p}{n}}\|\widehat{\Delta}\|_2
$$

with probability at least $1-2p^{-1}$. Together with the RE$'$ condition that $\widehat{\Delta}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22}\widehat{\Delta} \geqslant \kappa\|\widehat{\Delta}\|_2^2$ with probability at least $1-2p^{-\zeta}$ in Lemma B.3.2, we have $\|\widehat{\Delta}\|_2 \leqslant 4\kappa^{-1}C_{\lambda'}\sqrt{s'\log p/n}$ with probability at least $1-2p^{-1}-2p^{-\zeta}$. Hence,

$$
\|\widehat{\Delta}\|_1 \leqslant 2s'^{1/2}\|\widehat{\Delta}\|_2 \leqslant 8s'\kappa^{-1}C_{\lambda'}\sqrt{\frac{\log p}{n}}
$$

with probability at least $1-2p^{-1}-2p^{-\zeta}$. Additionally,

$$
\widehat{\Delta}^{\mathrm{T}}\boldsymbol{\Sigma}_{22}\widehat{\Delta} \leqslant \frac{16C_{\lambda'}^2}{\kappa}\frac{s'\log p}{n}
$$

with probability at least $1-2p^{-1}-2p^{-\zeta}$. □

B.3.2 Proof of Lemma 4.3.2

Proof. First consider the score function $\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)$. Recall that

$$
\begin{aligned}
S_{i\beta}(\boldsymbol{\theta}_0) &= \beta_0(W_i^2 - \sigma_U^2) + W_i\mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma} - W_iY_i \\
&= \beta_0 W_iU_i - \beta_0\sigma_U^2 - W_i\epsilon_i.
\end{aligned}
$$

Then by triangle inequality, Lemma B.6.4 and Assumption 4.2.1, we have

$$
\begin{aligned}
\|S_{i\beta}(\boldsymbol{\theta}_0)\|_{\psi_1} &= |\beta_0|\|W_iU_i\|_{\psi_1} + |\beta_0|\sigma_U^2 + \|W_i\epsilon_i\|_{\psi_1} \\
&\leqslant 2|\beta_0|\|W_i\|_{\psi_2}\|U_i\|_{\psi_2} + |\beta_0|\sigma_U^2 + 2\|W_i\|_{\psi_2}\|\epsilon_i\|_{\psi_2} \\
&\leqslant 2|\beta_0|K_U(K+K_U) + |\beta_0|\sigma_U^2 + 2(K+K_U)K_\epsilon.
\end{aligned}
$$

Recall that for $j = 1, \ldots, (p-1)$

$$
S_{i\gamma_j}(\boldsymbol{\theta}_0) = \beta_0 W_iZ_{ij} + \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\gamma}_0 Z_{ij} - Y_iZ_{ij}
$$

$$= \quad \beta_0 U_i Z_{ij} - \epsilon_i Z_{ij}.$$

Then by triangle inequality, Lemma B.6.4 and Assumption 4.2.1, we have

$$\begin{aligned} \|S_{i\gamma_j}(\boldsymbol{\theta}_0)\|_{\psi_1} &\leqslant 2|\beta_0|\|U_i\|_{\psi_2}\|Z_{ij}\|_{\psi_2} + 2\|\epsilon_i\|_{\psi_2}\|Z_{ij}\|_{\psi_2} \\ &\leqslant 2|\beta_0|K_U K + 2K_\epsilon K. \end{aligned}$$

Let $K_6 = \max\{2|\beta_0|K_U(K+K_U) + |\beta_0|\sigma_U^2 + 2(K+K_U)K_\epsilon, 2|\beta_0|K_U K + 2K_\epsilon K\} = 2|\beta_0|K_U(K+K_U) + |\beta_0|\sigma_U^2 + 2(K+K_U)K_\epsilon$. By Bernstein inequality in Lemma B.6.1 and union bound inequality, we have

$$\Pr\left(\|\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\|_\infty \geqslant t\right) \leqslant 2p\exp\left\{-C''\min\left(\frac{t^2}{K_2^2}, \frac{t}{K_2}\right)n\right\},$$

where C'' is a universal constant. Let $t = C\sqrt{n^{-1}\log p}$. Since $t = o(1)$, we have

$$-C''\min\left(t^2/K_2^2, t/K_2\right)n = -C''t^2K_2^{-2}n = \log p^{-C''C^2K_2^{-2}}$$

for large enough n. For any $\varepsilon > 0$, choose C such that $C''C^2K_2^{-2} > 1 + \log 2/\log p$. Then there exists finite $N > 0$ such that

$$\begin{aligned} \Pr\left(\|\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\|_\infty \geqslant t\right) &\leqslant 2pp^{-C''C^2K_2^{-2}} \\ &= p^{1+\log 2/\log p - C''C^2K_2^{-2}} \\ &\leqslant \varepsilon, \end{aligned}$$

for any $n > N$. Therefore, $\|\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\|_\infty = O_P(\sqrt{n^{-1}\log p})$.

Second consider the term $\boldsymbol{\nu}^{\mathrm{T}}\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)$. Recall that

$$\begin{aligned} \boldsymbol{\nu}^{\mathrm{T}}\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0) &= \begin{pmatrix} \widehat{\Sigma}_{11} - \boldsymbol{\omega}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21} & \widehat{\boldsymbol{\Sigma}}_{12} - \boldsymbol{\omega}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22} \end{pmatrix} \\ &= \begin{pmatrix} \sum_{i=1}^n W_i^2/n - \sigma_U^2 - \sum_{i=1}^n W_i\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_i/n & \sum_{i=1}^n W_i\mathbf{Z}_i^{\mathrm{T}}/n - \sum_{i=1}^n \boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}/n \end{pmatrix}. \end{aligned}$$

By triangle inequality, Lemma B.6.4 and Assumption 4.2.1, we have

$$\begin{aligned} \|[\boldsymbol{\nu}^{\mathrm{T}}\nabla\mathbf{S}_{i,\boldsymbol{\theta}}(\boldsymbol{\theta}_0)]_1\|_{\psi_1} &= \|W_i^2 - \sigma_U^2 - W_i\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_i\|_{\psi_1} \\ &\leqslant \|W_i^2\|_{\psi_1} + \sigma_U^2 + \|W_i\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_i\|_{\psi_1} \\ &\leqslant 2\|W_i\|_{\psi_2}^2 + \sigma_U^2 + 2\|W_i\|_{\psi_2}\|\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_i\|_{\psi_2} \end{aligned}$$

$$\leqslant \quad 2(K+K_U)^2+\sigma_U^2+2(K+K_U)KK_\omega.$$

For $j=1,\ldots,(p-1)$, we have

$$\begin{aligned}\|[\boldsymbol{\nu}^{\mathrm{T}}\nabla\mathbf{S}_{i,\boldsymbol{\theta}}(\boldsymbol{\theta}_0)]_{j+1}\|_{\psi_1} &= \|W_iZ_{ij}-\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_iZ_{ij}\|_{\psi_1}\\ &\leqslant 2\|W_i\|_{\psi_2}\|Z_{ij}\|_{\psi_2}+2\|\boldsymbol{\omega}^{\mathrm{T}}\mathbf{Z}_i\|_{\psi_2}\|Z_{ij}\|_{\psi_2}\\ &\leqslant 2K(K+K_U)+2K^2K_{\boldsymbol{\omega}}.\end{aligned}$$

Let $K_7=2\max\{2(K+K_U)^2+\sigma_U^2+2(K+K_U)KK_\omega, 2K(K+K_U)+2K^2K_\omega\}=4(K+K_U)(K+K_U+KK_\omega)+2\sigma_U^2$. Then we have $\|[\boldsymbol{\nu}^{\mathrm{T}}\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)-E\{\boldsymbol{\nu}^{\mathrm{T}}\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\}]_j\|_{\psi_1}\leqslant K_7$ for $j=1,\ldots,p$. Similarly, we have that for any $\varepsilon>0$, choose C such that $C''C^2K_7^{-2}>1+\log 2/\log p$. Then there exists finite $N>0$ such that

$$\begin{aligned}\Pr\left(\|\boldsymbol{\nu}^{\mathrm{T}}\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)-E\{\boldsymbol{\nu}^{\mathrm{T}}\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\}\|_\infty\geqslant C\sqrt{n^{-1}\log p}\right) &\leqslant 2pp^{-C''C^2K_7^{-2}}\\ &= p^{1+\log 2/\log p-C''C^2K_7^{-2}}\\ &\leqslant \varepsilon,\end{aligned}$$

for any $n>N$. Therefore, $\|\boldsymbol{\nu}^{\mathrm{T}}\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)-E\{\boldsymbol{\nu}^{\mathrm{T}}\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\}\|_\infty=O_P(\sqrt{n^{-1}\log p})$.

□

B.3.3 Proof of Lemma 4.3.3

Proof. Recall that the corrected loss function (4.3) is a quadratic function of $\boldsymbol{\theta}$, and

$$\mathbf{S}_{\boldsymbol{\theta}}(\breve{\boldsymbol{\theta}})=\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)+\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)(\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}_0).$$

Hence, $\boldsymbol{\nu}^{\mathrm{T}}\{\mathbf{S}_{\boldsymbol{\theta}}(\breve{\boldsymbol{\theta}})-\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)-\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)(\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\}=0$. Further, we have

$$\begin{aligned}&(\widehat{\boldsymbol{\nu}}-\boldsymbol{\nu})^{\mathrm{T}}\{\mathbf{S}_{\boldsymbol{\theta}}(\breve{\boldsymbol{\theta}})-\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\}\\ =\;&(\widehat{\boldsymbol{\nu}}-\boldsymbol{\nu})^{\mathrm{T}}\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)(\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\\ =\;&(\widehat{\boldsymbol{\nu}}-\boldsymbol{\nu})^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}(\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\\ =\;&\begin{pmatrix}0 & (\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega})^{\mathrm{T}}\end{pmatrix}\left\{\begin{pmatrix}\mathbf{W}^{\mathrm{T}}\mathbf{W}/n & \mathbf{W}^{\mathrm{T}}\mathbf{Z}/n\\ \mathbf{Z}^{\mathrm{T}}\mathbf{W}/n & \mathbf{Z}^{\mathrm{T}}\mathbf{Z}/n\end{pmatrix}-\begin{pmatrix}\sigma_U^2 & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}\right\}(\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\\ =\;&(\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega})^{\mathrm{T}}\frac{1}{n}\sum_{i=1}^n\mathbf{Z}_i(W_i,\mathbf{Z}_i^{\mathrm{T}})(\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}_0).\end{aligned}$$

By Cauchy-Schwarz inequality, we have

$$\begin{aligned} & n^{1/2}|(\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega})^{\mathrm{T}}\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i(W_i,\mathbf{Z}_i^{\mathrm{T}})(\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)| \\ \leqslant\ & n^{1/2}\left|\frac{1}{n}\sum_{i=1}^{n}\{(\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega})^{\mathrm{T}}\mathbf{Z}_i\}^2\right|^{1/2}\left|\frac{1}{n}\sum_{i=1}^{n}\{(W_i,\mathbf{Z}_i^{\mathrm{T}})(\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\}^2\right|^{1/2}. \end{aligned}$$

In Lemma B.3.3, we have proved that with probability at least $1-2p^{-1}-2p^{-\zeta}$

$$\left|\frac{1}{n}\sum_{i=1}^{n}\{(\widehat{\boldsymbol{\omega}}-\boldsymbol{\omega})^{\mathrm{T}}\mathbf{Z}_i\}^2\right|^{1/2}\leqslant\frac{4C_{\lambda'}}{\sqrt{\kappa}}\sqrt{\frac{s'\log p}{n}},$$

where $C_{\lambda'}$ is given in Lemma B.3.1. Since $\lambda=C_\lambda s_0\sqrt{n^{-1}\log p}$. In Lemma 4.2.1, for $C_\lambda>\max\left(8K_0K_2/C'',8\sqrt{2}K_0K_3/\sqrt{C''}\right)$, we have proved that

$$\left|(\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^{\mathrm{T}}\boldsymbol{\Sigma}(\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\right|\leqslant\frac{32\lambda^2 s_0}{\kappa},\ \text{ and }\ \|\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_2^2\leqslant\frac{32\lambda^2 s_0}{\kappa^2},$$

with probability at least $1-C_1\exp(-C_2\log p)$. Since U_i is sub-Gaussian with $\|U_i\|_{\psi_2}\leqslant K_U$, then by Lemma B.6.2, U_i^2 is sub-exponential with $\|U_i^2\|_{\psi_1}\leqslant 2K_U^2$. Since $E(U_i^2)=\sigma_U^2$, then $U_i^2-\sigma_U^2$ has mean 0 and $\|U_i^2-\sigma_U^2\|_{\psi_1}\leqslant 4K_U^2$. By Bernstein inequality, we have

$$\begin{aligned} \Pr\left(\frac{1}{n}\left|\sum_{i=1}^{n}(U_i^2-\sigma_U^2)\right|\geqslant\sqrt{\frac{\log p}{n}}\right) &\leqslant 2\exp\left\{-C''\min\left(\frac{\log p}{16nK_U^4},\frac{\sqrt{\log p}}{4K_U^2\sqrt{n}}\right)n\right\} \\ &\leqslant 2\exp\left(-C''\frac{\log p}{16K_U^4}\right), \end{aligned}$$

for n large enough. Then we have

$$\begin{aligned} & \left|\frac{1}{n}\sum_{i=1}^{n}\{(W_i,\mathbf{Z}_i^{\mathrm{T}})(\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\}^2\right|^{1/2} \\ =\ & \left|\frac{1}{n}\sum_{i=1}^{n}\{(X_i,\mathbf{Z}_i^{\mathrm{T}})(\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)+U_i(\breve{\beta}-\beta_0)\}^2\right|^{1/2} \\ \leqslant\ & \left|\frac{2}{n}\sum_{i=1}^{n}\{(X_i,\mathbf{Z}_i^{\mathrm{T}})(\breve{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\}^2+\frac{2}{n}\sum_{i=1}^{n}U_i^2(\breve{\beta}-\beta_0)^2\right|^{1/2} \\ \leqslant\ & \left|\frac{64\lambda^2 s_0}{\kappa}+2(\breve{\beta}-\beta_0)^2\frac{1}{n}\sum_{i=1}^{n}(U_i^2-\sigma_U^2)+2(\breve{\beta}-\beta_0)^2\sigma_U^2\right|^{1/2} \end{aligned}$$

$$
\begin{aligned}
&\leqslant \left| \frac{64\lambda^2 s_0}{\kappa} + 2\|\breve{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2^2 \frac{1}{n} \left| \sum_{i=1}^{n} (U_i^2 - \sigma_U^2) \right| + 2\|\breve{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2^2 \sigma_U^2 \right|^{1/2} \\
&\leqslant \left| \frac{64\lambda^2 s_0}{\kappa} + \frac{64\lambda^2 s_0}{\kappa^2} \sqrt{\frac{\log p}{n}} + \frac{64\lambda^2 s_0}{\kappa^2} \sigma_U^2 \right|^{1/2} \\
&\leqslant \left| 64(\kappa + 1 + \sigma_U^2) \frac{\lambda^2 s_0}{\kappa^2} \right|^{1/2} \\
&\leqslant \frac{8 s_0}{\kappa} \sqrt{\frac{(\kappa + 1 + \sigma_U^2) s_0 \log p}{n}}
\end{aligned}
$$

with probability at least $1 - C_1 \exp\{-C_2 \log p\} - 2\exp\{-C'' \log p/(16K_U^4)\}$. Hence, we have

$$
n^{1/2}(\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu})^{\mathrm{T}} \{\mathbf{S}_{\boldsymbol{\theta}}(\breve{\boldsymbol{\theta}}) - \mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\} \leqslant \frac{32 C_{\lambda'} (\kappa + 1 + \sigma_U^2)^{1/2}}{\kappa^{3/2}} \frac{s_0 (s_0 \vee s') \log p}{\sqrt{n}}
$$

with probability at lest $1 - C_1' \exp(-C_2' \log p)$, where C_1' and C_2' are constants depending on K, K_U, K_ϵ, K_0, κ and σ_U^2. Under the assumption that $s_0(s_0 \vee s')\log p/\sqrt{n} = o(1)$, we have $(\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu})^{\mathrm{T}} \{\mathbf{S}_{\boldsymbol{\theta}}(\breve{\boldsymbol{\theta}}) - \mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\} = o_P(n^{-1/2})$. □

B.3.4 Proof of Lemma 4.3.4

Proof. Since $\boldsymbol{\nu}^{\mathrm{T}} \mathbf{S}_{i\boldsymbol{\theta}}(\boldsymbol{\theta}_0)$ is independent and identically distributed with mean zero and finite variance $\sigma_{\beta|\gamma,0}^2 > 0$, then by Central Limit Theorem, we have

$$
\sqrt{n} \boldsymbol{\nu}^{\mathrm{T}} \mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \boldsymbol{\nu}^{\mathrm{T}} \mathbf{S}_{i\boldsymbol{\theta}}(\boldsymbol{\theta}_0) \to N(0, \sigma_{\beta|\gamma,0}^2).
$$

□

B.4 Proofs Regarding Score Test Statistic

B.4.1 Proof of Theorem 4.3.5

Proof. Recall that $\widetilde{\boldsymbol{\theta}}_{H_0} = (\beta^*, \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}})^{\mathrm{T}}$, $\widehat{\boldsymbol{\nu}} = (1, -\widehat{\boldsymbol{\omega}}^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\nu} = (1, -\boldsymbol{\omega}^{\mathrm{T}})^{\mathrm{T}}$. We have

$$
\begin{aligned}
& \sqrt{n} |\widehat{S}(\widetilde{\boldsymbol{\theta}}_{H_0}) - S(\boldsymbol{\theta}_0)| \\
= \; & \sqrt{n} |\widehat{\boldsymbol{\nu}}^{\mathrm{T}} \mathbf{S}_{\boldsymbol{\theta}}(\widetilde{\boldsymbol{\theta}}_{H_0}) - \boldsymbol{\nu}^{\mathrm{T}} \mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)| \\
\leqslant \; & \sqrt{n} |\boldsymbol{\nu}^{\mathrm{T}} \{\mathbf{S}_{\boldsymbol{\theta}}(\widetilde{\boldsymbol{\theta}}_{H_0}) - S_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\}| + \sqrt{n} |(\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu})^{\mathrm{T}} \mathbf{S}_{\boldsymbol{\theta}}(\widetilde{\boldsymbol{\theta}}_{H_0})|
\end{aligned}
$$

$$= \quad D_1 + D_2,$$

where $D_1 \equiv \sqrt{n}|\boldsymbol{\nu}^{\mathrm{T}}\{\mathbf{S}_{\boldsymbol{\theta}}(\widetilde{\boldsymbol{\theta}}_{H_0}) - S_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\}|$ and $D_2 \equiv \sqrt{n}|(\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu})^{\mathrm{T}}\mathbf{S}_{\boldsymbol{\theta}}(\widetilde{\boldsymbol{\theta}}_{H_0})|$. Since the corrected loss function (4.3) is a quadratic function of $\boldsymbol{\theta}$, by Lemma 4.3.3, we have

$$\begin{aligned}
|D_1| &= \sqrt{n}|\boldsymbol{\nu}^{\mathrm{T}}\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)(\widetilde{\boldsymbol{\theta}}_{H_0} - \boldsymbol{\theta}_0)| \\
&= \sqrt{n}|\boldsymbol{\nu}^{\mathrm{T}}\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)(0, \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}} - \boldsymbol{\gamma}_0^{\mathrm{T}})^{\mathrm{T}}| \\
&\leqslant \sqrt{n}\|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1\|\widehat{\boldsymbol{\Sigma}}_{12} - \boldsymbol{\omega}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22}\|_\infty \\
&= \sqrt{n}\|\widetilde{\boldsymbol{\theta}}_{H_0} - \boldsymbol{\theta}_0\|_1\|\widehat{\boldsymbol{\Sigma}}_{12} - \boldsymbol{\omega}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22}\|_\infty \\
&= \sqrt{n}\|\widetilde{\boldsymbol{\theta}}_{H_0} - \boldsymbol{\theta}_0\|_1\|\widehat{\boldsymbol{\Sigma}}_{12} - \boldsymbol{\omega}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22} - E\{\widehat{\boldsymbol{\Sigma}}_{12} - \boldsymbol{\omega}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22}\}\|_\infty \\
&\quad + \sqrt{n}\|\widetilde{\boldsymbol{\theta}}_{H_0} - \boldsymbol{\theta}_0\|_1\|E\{\widehat{\boldsymbol{\Sigma}}_{12} - \boldsymbol{\omega}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22}\}\|_\infty \\
&= \sqrt{n}\|\widetilde{\boldsymbol{\theta}}_{H_0} - \boldsymbol{\theta}_0\|_1\|\widehat{\boldsymbol{\Sigma}}_{12} - \boldsymbol{\omega}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22} - E\{\widehat{\boldsymbol{\Sigma}}_{12} - \boldsymbol{\omega}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22}\}\|_\infty \\
&\leqslant \sqrt{n}\|\widetilde{\boldsymbol{\theta}}_{H_0} - \boldsymbol{\theta}_0\|_1\|\boldsymbol{\nu}^{\mathrm{T}}\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0) - E\{\boldsymbol{\nu}^{\mathrm{T}}\nabla\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\}\|_\infty. \qquad \text{(B.14)}
\end{aligned}$$

In the above derivation, we used the fact that $\nabla\mathbf{S}_{\boldsymbol{\theta}}(\theta_0) = \widehat{\boldsymbol{\Sigma}}$, and under H_0 the first element of $\widetilde{\boldsymbol{\theta}}_{H_0} - \boldsymbol{\theta}_0$ is 0. In addition, $\boldsymbol{\omega}^{\mathrm{T}} = E(W_i\mathbf{Z}_i^{\mathrm{T}})E(\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}})^{-1}$, and hence

$$E(\widehat{\boldsymbol{\Sigma}}_{12} - \boldsymbol{\omega}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{22}) = E(W_i\mathbf{Z}_i^{\mathrm{T}}) - \boldsymbol{\omega}^{\mathrm{T}}E(\mathbf{Z}_i\mathbf{Z}_i^{\mathrm{T}}) = \mathbf{0}.$$

By Lemmas 4.2.1 and 4.3.2, we have $D_1 \leqslant O_P\left\{s_0^2\sqrt{n^{-1}\log p}\cdot\sqrt{\log p}\right\} = o_P(1)$.

For D_2, Lemma 4.3.3 yields

$$\begin{aligned}
|D_2| &\leqslant \sqrt{n}|(\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu})^{\mathrm{T}}\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)| + \sqrt{n}|(\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu})^{\mathrm{T}}\{\mathbf{S}_{\boldsymbol{\theta}}(\widetilde{\boldsymbol{\theta}}_{H_0}) - \mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\}| \\
&\leqslant \sqrt{n}|(\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu})^{\mathrm{T}}\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)| + o_P(1) \\
&\leqslant \sqrt{n}\|\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu}\|_1\|\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)\|_\infty + o_P(1).
\end{aligned}$$

By Lemmas 4.3.1 and 4.3.2, we have $|D_2| \leqslant O_P\left\{s'\sqrt{n^{-1}\log p}\cdot\sqrt{\log p}\right\} + o_P(1) = o_P(1)$. Hence, we have $\sqrt{n}|\widehat{S}(\widetilde{\boldsymbol{\theta}}_{H_0}) - S(\boldsymbol{\theta}_0)| = o_P(1)$. Since $\sigma^2_{\beta|\boldsymbol{\gamma},0} > 0$, we obtain that

$$\sqrt{n}\mid \widehat{S}(\widetilde{\boldsymbol{\theta}}_{H_0})(\sigma^2_{\beta|\boldsymbol{\gamma},0})^{-1/2} - S(\boldsymbol{\theta}_0)(\sigma^2_{\beta|\boldsymbol{\gamma},0})^{-1/2} \mid = o_P(1).$$

By Lemma 4.3.4, $\sqrt{n}S(\boldsymbol{\theta}_0)(\sigma^2_{\beta|\boldsymbol{\gamma},0})^{-1/2} = \sqrt{n}\boldsymbol{\nu}^{\mathrm{T}}\mathbf{S}_{\boldsymbol{\theta}}(\boldsymbol{\theta}_0)(\sigma^2_{\beta|\boldsymbol{\gamma},0})^{-1/2} \to N(0,1)$ in distribution. Applying the Slutsky's theorem, we hence get $\sqrt{n}\widehat{S}(\beta^*, \widetilde{\boldsymbol{\gamma}})(\sigma^2_{\beta|\boldsymbol{\gamma},0})^{-1/2} \to N(0,1)$ in distribution under null hypothesis. This completes the proof. □

B.4.2 Proof of Corollary 4.3.6

Proof. Recall that $\widehat{T}_n = n^{1/2}\widehat{S}(\beta^*, \widetilde{\boldsymbol{\gamma}})(\widehat{\sigma}^2_{\beta|\boldsymbol{\gamma},H_0})^{-1/2}$. Let $T_n = n^{1/2}\widehat{S}(\beta^*, \widetilde{\boldsymbol{\gamma}})(\sigma^2_{\beta|\boldsymbol{\gamma},0})^{-1/2}$. Then

$$\widehat{T}_n - T_n = T_n\left(\frac{\sigma_{\beta|\boldsymbol{\gamma},0}}{\widehat{\sigma}_{\beta|\boldsymbol{\gamma},H_0}} - 1\right).$$

In Theorem 4.3.5, we have proved that $T_n \to N(0,1)$ in distribution, as $n \to \infty$. It remains to show that $\sigma_{\beta|\boldsymbol{\gamma},0}/\widehat{\sigma}_{\beta|\boldsymbol{\gamma},H_0} - 1 = o_p(1)$. We start from deriving the bound of $|\widehat{\sigma}^2_{\beta|\boldsymbol{\gamma},H_0} - \sigma^2_{\beta|\boldsymbol{\gamma},0}|$. Recall that

$$\begin{aligned}
\widehat{\sigma}^2_{\beta|\boldsymbol{\gamma},H_0} &= (\widetilde{\sigma}^2_{\epsilon,H_0} + \beta^{*2}\sigma^2_U)(1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}) + \beta^{*2}E(U_i^4) + \widetilde{\sigma}^2_{\epsilon,H_0}\sigma^2_U - \beta^{*2}\sigma^4_U,\\
\sigma^2_{\beta|\boldsymbol{\gamma},0} &= (\sigma^2_\epsilon + \beta_0^2\sigma^2_U)\{1 - \boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)\} + \beta_0^2E(U_i^4) + \sigma^2_\epsilon\sigma^2_U - \beta_0^2\sigma^4_U.
\end{aligned}$$

Since $\beta_0 = \beta^*$ under null hypothesis, then we have

$$\begin{aligned}
\widehat{\sigma}^2_{\beta|\boldsymbol{\gamma},H_0} - \sigma^2_{\beta|\boldsymbol{\gamma},0} &= (\widetilde{\sigma}^2_{\epsilon,H_0} - \sigma^2_\epsilon)\sigma^2_U + \beta_0^2\sigma^2_U\{\boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i) - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}\}\\
&+ \widetilde{\sigma}^2_{\epsilon,H_0}(1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}) - \sigma^2_\epsilon\{1 - \boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)\}.
\end{aligned}$$

Let $D_1 = (\widetilde{\sigma}^2_{\epsilon,H_0} - \sigma^2_\epsilon)\sigma^2_U$, $D_2 = \beta_0^2\sigma^2_U\{\boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i) - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}\}$ and $D_3 = \widetilde{\sigma}^2_{\epsilon,H_0}(1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}) - \sigma^2_\epsilon\{1 - \boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)\}$. For term D_1, recall that

$$\begin{aligned}
\widetilde{\sigma}^2_{\epsilon,H_0} &= n^{-1}\sum_{i=1}^n (Y_i - \beta^* W_i)^2 - n^{-1}\sum_{i=1}^n (\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)^2 - \beta^{*2}\sigma^2_U,\\
\sigma^2_\epsilon &= E\{(Y_i - \beta_0 W_i)^2\} - E\{(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{Z}_i)^2\} - \beta_0^2\sigma^2_U.
\end{aligned}$$

Then we have

$$\begin{aligned}
&\widetilde{\sigma}^2_{\epsilon,H_0} - \sigma^2_\epsilon\\
=\ & n^{-1}\sum_{i=1}^n (Y_i - \beta^* W_i)^2 - E\{(Y_i - \beta_0 W_i)^2\} - \left[n^{-1}\sum_{i=1}^n (\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)^2 - E\{(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{Z}_i)^2\}\right]\\
=\ & n^{-1}\sum_{i=1}^n (Y_i - \beta_0 W_i)^2 - E\{(Y_i - \beta_0 W_i)^2\} - \left[n^{-1}\sum_{i=1}^n (\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)^2 - E\{(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{Z}_i)^2\}\right].
\end{aligned}$$

First, by triangle inequality and Assumption 4.2.1, we know

$$\|\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{Z}_i\|_{\psi_2} \leqslant \sum_{j=1}^{p-1} |\gamma_{0j}| \|\mathbf{Z}_{ij}\|_{\psi_2} \leqslant K\|\boldsymbol{\gamma}_0\|_1 \leqslant s_0 K K_0.$$

Then we have

$$\begin{aligned}
\|s_0^{-2}(Y_i - \beta_0 W_i)^2\|_{\psi_1} &= \|s_0^{-2}(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{Z}_i + \epsilon_i - \beta_0 U_i)^2\|_{\psi_1} \\
&\leqslant 2\|s_0^{-1}(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{Z}_i + \epsilon_i - \beta_0 U_i)\|_{\psi_2}^2 \\
&\leqslant 2\left(\|s_0^{-1}\boldsymbol{\gamma}_0\mathbf{Z}_i\|_{\psi_2} + \|s_0^{-1}\epsilon_i\|_{\psi_2} + \|s_0^{-1}\beta_0 U_i\|_{\psi_2}\right)^2 \\
&\leqslant 2(K_0 K + s_0^{-1}K_\epsilon + s_0^{-1}|\beta_0|K_U)^2 \\
&\leqslant K_8,
\end{aligned}$$

where K_8 is a finite constant. Then by Bernstein inequality, for any $t > 0$ we have

$$\Pr\left(s_0^{-2}\left|n^{-1}\sum_{i=1}^n (Y_i - \beta_0 W_i)^2 - E\{(Y_i - \beta_0 W_i)^2\}\right| \geqslant t\right) \leqslant 2\exp\left\{-C''\min\left(\frac{t^2}{4K_8^2}, \frac{t}{2K_8}\right)n\right\}.$$

Let $t = \sqrt{n^{-1}\log p}$. Then for n large enough, we have

$$\Pr\left(\left|n^{-1}\sum_{i=1}^n (Y_i - \beta_0 W_i)^2 - E\{(Y_i - \beta_0 W_i)^2\}\right| \leqslant s_0^2\sqrt{n^{-1}\log p}\right) \geqslant 1 - 2\exp\left(\frac{-C''\log p}{4K_8^2}\right).$$

Thus,

$$\left|n^{-1}\sum_{i=1}^n (Y_i - \beta_0 W_i)^2 - E\{(Y_i - \beta_0 W_i)^2\}\right| \leqslant s_0^2\sqrt{n^{-1}\log p} \tag{B.15}$$

with probability tending to 1. Note that under the condition given in Lemma 4.3.1, $s_0^2\sqrt{n^{-1}\log p} = o(1)$. For term $n^{-1}\sum_{i=1}^n (\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)^2 - E\{(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{Z}_i)^2\}$, we first have

$$\begin{aligned}
&\left|n^{-1}\sum_{i=1}^n (\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)^2 - E\{(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{Z}_i)^2\}\right| \\
\leqslant\ &\left|n^{-1}\sum_{i=1}^n (\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}\mathbf{Z}_i)^2 - n^{-1}\sum_{i=1}^n (\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{Z}_i)^2\right| + \left|n^{-1}\sum_{i=1}^n (\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{Z}_i)^2 - E\{(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{Z}_i)^2\}\right| \\
\leqslant\ &\|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 \left\|n^{-1}\sum_{i=1}^n (\widetilde{\boldsymbol{\gamma}} + \boldsymbol{\gamma}_0)^{\mathrm{T}}\mathbf{Z}_i\mathbf{Z}_i\right\|_\infty + \left|n^{-1}\sum_{i=1}^n (\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{Z}_i)^2 - E\{(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{Z}_i)^2\}\right|.
\end{aligned}$$

By triangle inequality, Lemma B.6.4 and Lemma 4.2.1, we have

$$\begin{aligned}
\|(\widetilde{\boldsymbol{\gamma}}+\boldsymbol{\gamma}_0)^{\mathrm{T}}\mathbf{Z}_i Z_{ik}\|_{\psi_1} &= \left\|\sum_{j=1}^{p-1}(\widetilde{\gamma}_j+\gamma_{0j})Z_{ij}Z_{ik}\right\|_{\psi_1} \\
&\leqslant \sum_{j=1}^{p-1}|\widetilde{\gamma}_j+\gamma_{0j}|\|Z_{ij}Z_{ik}\|_{\psi_1} \\
&\leqslant 2K^2\|\widetilde{\boldsymbol{\gamma}}+\boldsymbol{\gamma}_0\|_1 \\
&\leqslant 2K^2 s_0(\|\widetilde{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_\infty+2\|\boldsymbol{\gamma}_0\|_\infty) \\
&\leqslant 2K^2 s_0(C_\infty C_\lambda s_0\sqrt{n^{-1}\log p}+2K_0) \\
&\leqslant 2K^2 s_0 K_0',
\end{aligned}$$

with probability tending to 1, where K_0' is a constant. The third last inequality holds because the support of CoCoLasso estimate $\widetilde{\boldsymbol{\theta}}$ is a subset of the true support with probability going to 1. The second last inequality used result (b) in Lemma 4.2.2 and that $\|\boldsymbol{\theta}\|_\infty$ is bounded. Then $\|s_0^{-1}(\widetilde{\boldsymbol{\gamma}}+\boldsymbol{\gamma}_0)^{\mathrm{T}}\mathbf{Z}_i Z_{ik}\|_{\psi_1}\leqslant 2K^2K_0'<\infty$, for $k=1,\ldots,p-1$. By the definition of sub-exponential norm, we know that $|E\{s_0^{-1}(\widetilde{\boldsymbol{\gamma}}+\boldsymbol{\gamma}_0)^{\mathrm{T}}\mathbf{Z}_i Z_{ik}\}|$ is also finite. By Bernstein inequality and union bound inequality, for any $t>0$ we have

$$\begin{aligned}
&\Pr\left(\left\|n^{-1}\sum_{i=1}^n s_0^{-1}(\widetilde{\boldsymbol{\gamma}}+\boldsymbol{\gamma}_0)^{\mathrm{T}}\mathbf{Z}_i\mathbf{Z}_i - E\{s_0^{-1}(\widetilde{\boldsymbol{\gamma}}+\boldsymbol{\gamma}_0)^{\mathrm{T}}\mathbf{Z}_i\mathbf{Z}_i\}\right\|_\infty \geqslant t\right) \\
\leqslant\ & 2p\exp\left\{-C''\min\left(\frac{t^2}{16K^4K_0'^2},\frac{t}{4K^2K_0'}\right)n\right\}.
\end{aligned}$$

Let $t=C\sqrt{n^{-1}\log p}$. Then for n large enough, we have

$$\begin{aligned}
&\Pr\left(\left\|n^{-1}\sum_{i=1}^n s_0^{-1}(\widetilde{\boldsymbol{\gamma}}+\boldsymbol{\gamma}_0)^{\mathrm{T}}\mathbf{Z}_i\mathbf{Z}_i - E\{s_0^{-1}(\widetilde{\boldsymbol{\gamma}}+\boldsymbol{\gamma}_0)^{\mathrm{T}}\mathbf{Z}_i\mathbf{Z}_i\}\right\|_\infty \leqslant C\sqrt{n^{-1}\log p}\right) \\
\geqslant\ & 1-2p\exp\left(\frac{-C''C^2\log p}{16K^4K_0'^2}\right).
\end{aligned}$$

When $C''C^2/(16K^4K_0')>1$, we have

$$\left\|n^{-1}\sum_{i=1}^n s_0^{-1}(\widetilde{\boldsymbol{\gamma}}+\boldsymbol{\gamma}_0)^{\mathrm{T}}\mathbf{Z}_i\mathbf{Z}_i\right\|_\infty \leqslant \|E\{s_0^{-1}(\widetilde{\boldsymbol{\gamma}}+\boldsymbol{\gamma}_0)^{\mathrm{T}}\mathbf{Z}_i\mathbf{Z}_i\}\|_\infty + C\sqrt{n^{-1}\log p}$$

with probability tending to 1. Hence, we obtain that

$$\begin{aligned}
& \|\widetilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 \left\| n^{-1} \sum_{i=1}^{n} (\widetilde{\boldsymbol{\gamma}} + \boldsymbol{\gamma}_0)^{\mathrm{T}} \mathbf{Z}_i \mathbf{Z}_i \right\|_\infty \\
\leqslant \;& s_0 \|\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 \left\{ \|E\{s_0^{-1} (\widetilde{\boldsymbol{\gamma}} + \boldsymbol{\gamma}_0)^{\mathrm{T}} \mathbf{Z}_i \mathbf{Z}_i\}\|_\infty + C\sqrt{n^{-1}\log p} \right\} \\
\leqslant \;& 16 s_0^2 \lambda \kappa^{-1} \{ \|E\{s_0^{-1} (\widetilde{\boldsymbol{\gamma}} + \boldsymbol{\gamma}_0)^{\mathrm{T}} \mathbf{Z}_i \mathbf{Z}_i\}\|_\infty + C\sqrt{n^{-1}\log p} \} \\
\leqslant \;& C_1 s_0^3 \sqrt{n^{-1}\log p},
\end{aligned}$$

for some constant C_1 with probability tending to 1.

For term $\left| n^{-1} \sum_{i=1}^{n} (\boldsymbol{\gamma}_0^{\mathrm{T}} \mathbf{Z}_i)^2 - E\{(\boldsymbol{\gamma}_0^{\mathrm{T}} \mathbf{Z}_i)^2\} \right|$, since $\|s_0^{-1} \boldsymbol{\gamma}_0^{\mathrm{T}} \mathbf{Z}_i\|_{\psi_2} \leqslant K K_0$, then

$$\|s_0^{-2} (\boldsymbol{\gamma}_0^{\mathrm{T}} \mathbf{Z}_i)^2\|_{\psi_1} \leqslant 2K^2 K_0^2 < \infty$$

by Lemma B.6.4. By Bernstein inequality, for any $t > 0$, we have

$$\begin{aligned}
& \Pr\left(\left| n^{-1} \sum_{i=1}^{n} s_0^{-2} (\boldsymbol{\gamma}_0^{\mathrm{T}} \mathbf{Z}_i)^2 - E\{s_0^{-2} (\boldsymbol{\gamma}_0^{\mathrm{T}} \mathbf{Z}_i)^2\} \right| \geqslant t \right) \\
& \leqslant 2 \exp\left\{ -C'' \min\left(\frac{t^2}{16K^4 K_0^4}, \frac{t}{4K^2 K_0^2} \right) n \right\}.
\end{aligned}$$

Let $t = \sqrt{n^{-1}\log p}$, then for n large enough we have

$$\Pr\left(\left| n^{-1} \sum_{i=1}^{n} (\boldsymbol{\gamma}_0^{\mathrm{T}} \mathbf{Z}_i)^2 - E\{(\boldsymbol{\gamma}_0^{\mathrm{T}} \mathbf{Z}_i)^2\} \right| \leqslant s_0^2 \sqrt{n^{-1}\log p} \right) \geqslant 1 - 2\exp\left(\frac{-C''\log p}{16K^4 K_0^4} \right).$$

Hence, we obtain that

$$\left| n^{-1} \sum_{i=1}^{n} (\widetilde{\boldsymbol{\gamma}}^{\mathrm{T}} \mathbf{Z}_i)^2 - E\{(\boldsymbol{\gamma}_0^{\mathrm{T}} \mathbf{Z}_i)^2\} \right| \leqslant C_1 s_0^3 \sqrt{n^{-1}\log p} + s_0^2 \sqrt{n^{-1}\log p} \tag{B.16}$$

with probability tending to 1. Therefore, from (B.15) and (B.16), we obtain

$$\begin{aligned}
|D_1| &= |\widetilde{\sigma}^2_{\epsilon, H_0} - \sigma^2_\epsilon| \sigma_U^2 \\
&\leqslant \left(2s_0^2 \sqrt{n^{-1}\log p} + C_1 s_0^3 \sqrt{n^{-1}\log p} \right) \sigma_U^2 \\
&\leqslant C_2 s_0^3 \sqrt{n^{-1}\log p} \sigma_U^2,
\end{aligned} \tag{B.17}$$

with probability tending to 1.

For term $D2$, by triangle inequality, we first have

$$\begin{aligned} |D_2| &\leqslant \beta_0^2\sigma_U^2\{|\boldsymbol{\omega}^{\mathrm{T}}\{E(X_i\mathbf{Z}_i) - \widehat{\boldsymbol{\Sigma}}_{21}\}| + |(\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega})^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}|\} \\ &\leqslant \beta_0^2\sigma_U^2\{\|\boldsymbol{\omega}\|_1\|\widehat{\boldsymbol{\Sigma}}_{21} - E(X_i\mathbf{Z}_i)\|_\infty + \|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}\|_1\|\widehat{\boldsymbol{\Sigma}}_{21}\|_\infty\}. \end{aligned}$$

In the proof of Lemma 4.2.1, we have showed that $\{\widehat{\boldsymbol{\Sigma}}_{21} - E(X_i\mathbf{Z}_i)\}_j$ is sub-exponential and $\|\{\widehat{\boldsymbol{\Sigma}}_{21} - E(X_i\mathbf{Z}_i)\}_j\|_{\psi_1} \leqslant 4K^2$ for $j = 1, \ldots, p-1$. Then by Bernstein inequality, for any $t > 0$ we have

$$\Pr\left(\|\widehat{\boldsymbol{\Sigma}}_{21} - E(X_i\mathbf{Z}_i)\|_\infty \geqslant t\right) \leqslant 2(p-1)\exp\left\{-C''\min\left(\frac{t^2}{16K^4}, \frac{t}{4K^2}\right)n\right\}.$$

Let $t = M\sqrt{n^{-1}\log p}$, where $M > 0$. Then for any $\varepsilon > 0$, there exists $M > \sqrt{16K^4/C''}$, such that

$$\Pr\left(\|\widehat{\boldsymbol{\Sigma}}_{21} - E(X_i\mathbf{Z}_i)\|_\infty \geqslant M\sqrt{n^{-1}\log p}\right) \leqslant \varepsilon, \tag{B.18}$$

for n large enough. Hence, $\|\widehat{\boldsymbol{\Sigma}}_{21} - E(X_i\mathbf{Z}_i)\|_\infty = O_P(\sqrt{n^{-1}\log p})$. By Assumption 4.2.1, then we have

$$\|\boldsymbol{\omega}\|_1\|\widehat{\boldsymbol{\Sigma}}_{21} - E(X_i\mathbf{Z}_i)\|_\infty \leqslant K_{\boldsymbol{\omega}}M\sqrt{n^{-1}\log p} \leqslant C_3\sqrt{n^{-1}\log p}$$

for some constant C_3, with probability tending to 1. Here, $K_{\boldsymbol{\omega}}$ is a positive constant satisfying $\|E(X_i\mathbf{Z}_i^{\mathrm{T}})\{E(\mathbf{Z}_i^{\otimes 2})\}^{-1}\|_1 \leqslant K_{\boldsymbol{\omega}}$. Since the data is standardized, $|E(X_iZ_{ij})| = |\mathrm{cor}(X_i, Z_{ij})| \leqslant 1$ for $j = 1, \ldots, p-1$. Then $\|E(X_i\mathbf{Z}_i)\|_\infty \leqslant 1$. By (B.18), $\|\widehat{\boldsymbol{\Sigma}}_{21}\|_\infty \leqslant \|E(X_i\mathbf{Z}_i)\|_\infty + M\sqrt{n^{-1}\log p} \leqslant 1 + M\sqrt{n^{-1}\log p}$ for some constant M and n large enough, with probability tending to 1. By Lemma 4.3.1, we have

$$\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}\|_1\|\widehat{\boldsymbol{\Sigma}}_{21}\|_\infty \leqslant \|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}\|_1 + \|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}\|_1 M\sqrt{n^{-1}\log p} \leqslant C_4 s'\sqrt{n^{-1}\log p} \tag{B.19}$$

for some constant C_4, with probability tending to 1. Then we obtain

$$|D_2| \leqslant \beta_0^2\sigma_U^2(C_3 + C_4)s'\sqrt{n^{-1}\log p}. \tag{B.20}$$

with probability tending to 1.

For term $D3$, by triangle inequality, (B.17) and (B.20), we have

$$|D_3| \leqslant \widetilde{\sigma}^2_{\epsilon,H_0}|\boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i) - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}| + |1 - \boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)| \cdot |\widetilde{\sigma}^2_{\epsilon,H_0} - \sigma^2_\epsilon|$$

$$
\begin{aligned}
&\leqslant \widetilde{\sigma}^2_{\epsilon,H_0}(C_3+C_4)s'\sqrt{n^{-1}\log p}+|1-\boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)|C_2s_0^3\sqrt{n^{-1}\log p}\\
&\leqslant (\sigma^2_\epsilon+C_2s_0^3\sqrt{n^{-1}\log p})(C_3+C_4)s'\sqrt{n^{-1}\log p}\\
&\quad +|1-\boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)|C_2s_0^3\sqrt{n^{-1}\log p}\\
&\leqslant C_5(s_0^3+s')\sqrt{n^{-1}\log p}
\end{aligned}
$$

with probability tending to 1, where C_5 is a constant. Note that $|1-\boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)|$ is bounded, because $\|\boldsymbol{\omega}\|_1 \leqslant K_{\boldsymbol{\omega}}$ and $\|E(X_i\mathbf{Z}_i)\|_\infty \leqslant 1$. Therefore,

$$
|\widehat{\sigma}^2_{\beta|\boldsymbol{\gamma},H_0}-\sigma^2_{\beta|\boldsymbol{\gamma},0}| \leqslant |D_1|+|D_2|+|D_3| = O_p\{(s_0^3+s')\sqrt{n^{-1}\log p}\}.
$$

Since we assume that the true parameter $\sigma^2_{\beta|\boldsymbol{\gamma},0}$ is bounded away from 0, $s_0^3\sqrt{n^{-1}\log p} = o(1)$ and $s'\sqrt{n^{-1}\log p} = o(1)$, then

$$
\begin{aligned}
\left|\frac{\sigma_{\beta|\boldsymbol{\gamma},0}}{\widehat{\sigma}_{\beta|\boldsymbol{\gamma},H_0}}-1\right| &= \frac{1}{\widehat{\sigma}_{\beta|\boldsymbol{\gamma},H_0}(\widehat{\sigma}_{\beta|\boldsymbol{\gamma},H_0}+\sigma_{\beta|\boldsymbol{\gamma},0})}|\sigma^2_{\beta|\boldsymbol{\gamma},0}-\widehat{\sigma}^2_{\beta|\boldsymbol{\gamma},H_0}|\\
&\leqslant \widehat{\sigma}^{-2}_{\beta|\boldsymbol{\gamma},H_0}|\sigma^2_{\beta|\boldsymbol{\gamma},0}-\widehat{\sigma}^2_{\beta|\boldsymbol{\gamma},H_0}|\\
&\leqslant C_6|\sigma^2_{\beta|\boldsymbol{\gamma},0}-\widehat{\sigma}^2_{\beta|\boldsymbol{\gamma},H_0}|,
\end{aligned}
$$

for some constant C_6 with probability tending to 1. Hence, $|\sigma_{\beta|\boldsymbol{\gamma},0}/\widehat{\sigma}_{\beta|\boldsymbol{\gamma},H_0}-1| = o_p(1)$ and $\widehat{T}_n \to N(0,1)$ in distribution as n goes to infinity. □

B.4.3 Proof of Corollary 4.3.7

Proof. Under local alternatives, we know $E\{S(\beta_n,\boldsymbol{\gamma}_0)\}=0$ and $\sqrt{n}\widehat{S}(\beta_n,\widetilde{\boldsymbol{\gamma}})(\widehat{\sigma}^2_{\beta_n|\boldsymbol{\gamma}})^{-1/2}$ converges to standard normal distribution by Corollary 4.3.6. Then by Taylor expansion, we have

$$
\begin{aligned}
\widehat{T}_n &= \frac{\sqrt{n}\widehat{S}(\beta_n,\widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\sigma}^2_{\beta_n|\boldsymbol{\gamma}}}}+\left\{\frac{\partial\widehat{S}(\beta_{0n},\widetilde{\boldsymbol{\gamma}})}{\partial\beta_{0n}}\frac{1}{\sqrt{\widehat{\sigma}^2_{\beta_{0n}|\boldsymbol{\gamma}}}}-\frac{\widehat{S}(\beta_{0n},\widetilde{\boldsymbol{\gamma}})}{2(\widehat{\sigma}^2_{\beta_{0n}|\boldsymbol{\gamma}})^{3/2}}\right\}\sqrt{n}(\beta^*-\beta_n)\\
&= \frac{\sqrt{n}\widehat{S}(\beta_n,\widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\sigma}^2_{\beta_n|\boldsymbol{\gamma}}}}+\left[E\left\{\frac{\partial S(\beta,\boldsymbol{\gamma}_0)}{\partial\beta}\bigg|_{\beta=\beta_n}\right\}\frac{1}{\sqrt{\sigma^2_{\beta_n|\boldsymbol{\gamma},0}}}-\frac{E\{S(\beta_n,\boldsymbol{\gamma}_0)\}}{2(\sigma^2_{\beta_n|\boldsymbol{\gamma},0})^{3/2}}\right]\sqrt{n}(\beta^*-\beta_n)\\
&\quad +o_p(1)\\
&= \frac{\sqrt{n}\widehat{S}(\beta_n,\widetilde{\boldsymbol{\gamma}})}{\sqrt{\widehat{\sigma}^2_{\beta_n|\boldsymbol{\gamma}}}}-hE\left\{\frac{\partial S(\beta,\boldsymbol{\gamma}_0)}{\partial\beta}\bigg|_{\beta=\beta_n}\right\}\frac{1}{\sqrt{\sigma^2_{\beta_n|\boldsymbol{\gamma},0}}}+o_p(1)\\
&\to N\{-h(\sigma^2_\beta)^{-1/2},1\}
\end{aligned}
$$

in distribution, where β_{0n} is between β^* and β_n, and $\sigma^2_{\beta_n|\boldsymbol{\gamma},0}$ is the variance of the decorrelated score $S(\beta_n, \boldsymbol{\gamma}_0)$ under local alternatives. Therefore, the power function converges to $\Pr\{|Z - h(\sigma_\beta^2)^{-1/2}| \geqslant Z_{\alpha/2}\}$, where Z is a standard normal random variable. □

B.5 Proofs Regarding Confidence Interval

B.5.1 Proof of Theorem 4.3.8

Proof. To prove the asymptotic normality of the one-step estimator $\widehat{\beta}$, we first show that $\{\partial \widehat{\mathbf{S}}(\beta, \widetilde{\boldsymbol{\gamma}})/\partial\beta\}|_{\beta=\widetilde{\beta}} = 1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}$ is consistent for $E[\{\partial S(\beta, \boldsymbol{\gamma}_0)/\partial\beta\}|_{\beta=\beta_0}] = 1 - \boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)$. By triangle inequality, we have the following decomposition

$$\begin{aligned}
|1 - \boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i) - (1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21})| &= |\widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21} - \boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)| \\
&\leqslant |\boldsymbol{\omega}^{\mathrm{T}}\{\widehat{\boldsymbol{\Sigma}}_{21} - E(X_i\mathbf{Z}_i)\}| + |(\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega})^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}| \\
&\leqslant \|\boldsymbol{\omega}\|_1\|\widehat{\boldsymbol{\Sigma}}_{21} - E(X_i\mathbf{Z}_i)\|_\infty + \|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}\|_1\|\widehat{\boldsymbol{\Sigma}}_{21}\|_\infty.
\end{aligned}$$

By (B.18), we know that $\|\widehat{\boldsymbol{\Sigma}}_{21} - E(X_i\mathbf{Z}_i)\|_\infty = O_P(\sqrt{n^{-1}\log p})$. Since $\|\boldsymbol{\omega}\|_1 \leqslant K_{\boldsymbol{\omega}}$, then $\|\boldsymbol{\omega}\|_1\|\widehat{\boldsymbol{\Sigma}}_{21} - E(X_i\mathbf{Z}_i)\|_\infty = O_P(\sqrt{n^{-1}\log p})$. By (B.19), we have $\|\widehat{\boldsymbol{\omega}} - \boldsymbol{\omega}\|_1\|\widehat{\boldsymbol{\Sigma}}_{21}\|_\infty = O_P(s'\sqrt{n^{-1}\log p})$. Hence, $|1 - \boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i) - (1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21})| \leqslant O_P(s'\sqrt{n^{-1}\log p}) = o_P(1)$.

Recall that $\widehat{\beta} = \widetilde{\beta} - \widehat{S}(\widetilde{\boldsymbol{\theta}})/(1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21})$. By plugging in the expression of $\widehat{\beta}$, we have

$$\begin{aligned}
n^{1/2}(\widehat{\beta} - \beta_0) &= n^{1/2}\left\{\widetilde{\beta} - \frac{\widehat{S}(\widetilde{\boldsymbol{\theta}})}{1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}} - \beta_0\right\} \\
&= n^{1/2}\left[\widetilde{\beta} - \beta_0 - \frac{1}{1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}}\left\{\widehat{S}(\beta_0, \widetilde{\boldsymbol{\gamma}}) + \frac{\partial\widehat{S}(\beta, \widetilde{\boldsymbol{\gamma}})}{\partial\beta}\bigg|_{\beta=\beta_0}(\widetilde{\beta} - \beta_0)\right\}\right] \\
&= n^{1/2}\left[\widetilde{\beta} - \beta_0 - \frac{1}{1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}}\left\{\widehat{S}(\beta_0, \widetilde{\boldsymbol{\gamma}}) + (1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21})(\widetilde{\beta} - \beta_0)\right\}\right] \\
&= -\frac{n^{1/2}\widehat{S}(\beta_0, \widetilde{\boldsymbol{\gamma}})}{1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}} \\
&= -\frac{n^{1/2}S(\beta_0, \boldsymbol{\gamma}_0) + o_P(1)}{1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}} \\
&= -\frac{n^{1/2}S(\beta_0, \boldsymbol{\gamma}_0)}{1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}} + o_P(1) \\
&= -\frac{n^{1/2}S(\beta_0, \boldsymbol{\gamma}_0)}{1 - \boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)}\frac{1 - \boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)}{1 - \widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}} + o_P(1)
\end{aligned}$$

$$
\begin{aligned}
&= -\frac{n^{1/2}S(\beta_0,\boldsymbol{\gamma}_0)}{1-\boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)}\{1+o_P(1)\}+o_P(1)\\
&= -\frac{n^{1/2}S(\beta_0,\boldsymbol{\gamma}_0)}{1-\boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)}+o_P(1).
\end{aligned}
$$

The second equality holds because the estimated decorrelated score $\widehat{S}(\beta,\boldsymbol{\gamma})$ is linear in β, then by expanding $\widehat{S}(\widetilde{\beta},\widetilde{\boldsymbol{\gamma}})$ around β_0, we obtain

$$
\widehat{S}(\widetilde{\beta},\widetilde{\boldsymbol{\gamma}})=\widehat{S}(\beta_0,\widetilde{\boldsymbol{\gamma}})+\left.\frac{\partial\widehat{S}(\beta,\widetilde{\boldsymbol{\gamma}})}{\partial\beta}\right|_{\beta=\beta_0}(\widetilde{\beta}-\beta_0).
$$

The fifth equality holds by Theorem 4.3.5 . The eighth equality holds because of the consistency of $1-\widehat{\boldsymbol{\omega}}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_{21}$ to $1-\boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)$.

By Lemma 4.3.4, we know $n^{1/2}S(\beta_0,\boldsymbol{\gamma}_0)\to N(0,\sigma^2_{\beta|\boldsymbol{\gamma},0})$ in distribution. Hence,

$$
n^{1/2}(\widehat{\beta}-\beta_0)=-\frac{n^{1/2}S(\beta_0,\boldsymbol{\gamma}_0)}{1-\boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)}+o_P(1)\to N(0,\sigma^2_\beta)
$$

in distribution, where $\sigma^2_\beta=\{1-\boldsymbol{\omega}^{\mathrm{T}}E(X_i\mathbf{Z}_i)\}^{-2}\sigma^2_{\beta|\boldsymbol{\gamma},0}$. □

B.6 Supplementary Lemmas

B.6.1 Inequalities about sub-exponential and sun-Gaussian random variables

Lemma B.6.1 (Bernstein Inequality)**.** *Let $X_1,\ldots,X_n$ be independent mean 0 sub-exponential random variables and $K=\max_i\|X_i\|_{\psi_1}$. Then for any $t>0$, we have*

$$
\Pr\left(\frac{1}{n}\left|\sum_{i=1}^n X_i\right|\geqslant t\right)\leqslant 2\exp\left\{-C''\min\left(\frac{t^2}{K^2},\frac{t}{K}\right)n\right\},
$$

where $C''>0$ is a universal constant.

For more details see Lemma K2 in the supplementary materials to Ning et al. (2017).

Lemma B.6.2 (Sub-exponential is sub-Gaussian squared)**.** *A random variable X is sub-Gaussian if and only if X^2 is sub-exponential. Moreover,*

$$
\|X\|^2_{\psi_2}\leqslant\|X^2\|_{\psi_1}\leqslant 2\|X\|^2_{\psi_2}.
$$

Lemma B.6.3. *Consider any bounded random variable X, thus $|X| \leqslant M$ almost surely for some M. Then X is a sub-Gaussian random variable with $\|X\|_{\psi_2} \leqslant M$. That is, $\|X\|_{\psi_2} \leqslant \|X\|_\infty$.*

Lemma B.6.4. *Assume X and Y are sub-Gaussian random variables. Then $\|XY\|_{\psi_1} \leqslant 2\|X\|_{\psi_2}\|Y\|_{\psi_2}$.*

B.6.2 Other supplementary Lemmas

Lemma B.6.5. *Let $p \times p$ symmetric matrix $\mathbf{M} = \begin{bmatrix} 1 & \boldsymbol{\alpha}^{\mathrm{T}} \\ \boldsymbol{\alpha} & \mathbf{A} \end{bmatrix}$, and $\lambda_{\min}(\mathbf{M}) \geqslant 2\kappa$ for some positive constant κ. Then $\lambda_{\min}(\mathbf{A}) \geqslant 2\kappa$.*

Proof. For any $\mathbf{x} \in \mathbb{R}^{p-1}$, let $x_1 = C\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha}$, where

$$C = \frac{-1}{2} I\{\lambda_{\min}(\mathbf{M}) = 1\} + \frac{1}{\lambda_{\min}(\mathbf{M}) - 1} I\{\lambda_{\min}(\mathbf{M}) < 1\} + \frac{4}{\lambda_{\min}(\mathbf{M}) - 1} I\{\lambda_{\min}(\mathbf{M}) > 1\}.$$

By

$$\begin{aligned} x_1^2 + 2x_1\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha} + \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} &= \begin{bmatrix} x_1 & \mathbf{x}^{\mathrm{T}} \end{bmatrix} \mathbf{M} \begin{bmatrix} x_1 \\ \mathbf{x} \end{bmatrix} \\ &\geqslant \lambda_{\min}(\mathbf{M})(x_1^2 + \mathbf{x}^{\mathrm{T}}\mathbf{x}) \\ &= \lambda_{\min}(\mathbf{M})x_1^2 + \lambda_{\min}(\mathbf{M})\mathbf{x}^{\mathrm{T}}\mathbf{x}, \end{aligned}$$

using the form of C, we get

$$\begin{aligned} \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} &\geqslant \lambda_{\min}(\mathbf{M})\mathbf{x}^{\mathrm{T}}\mathbf{x} + (\lambda_{\min}(\mathbf{M})C^2 - C^2 - 2C)(\mathbf{x}^{\mathrm{T}}\boldsymbol{\alpha})^2 \\ &\geqslant \lambda_{\min}(\mathbf{M})\mathbf{x}^{\mathrm{T}}\mathbf{x}. \end{aligned}$$

Hence, $\lambda_{\min}(\mathbf{A}) \geqslant \lambda_{\min}(\mathbf{M}) \geqslant 2\kappa$. □

B.7 Figures and Tables in Simulation Studies

B.7.1 Power performance with known measurement error variance

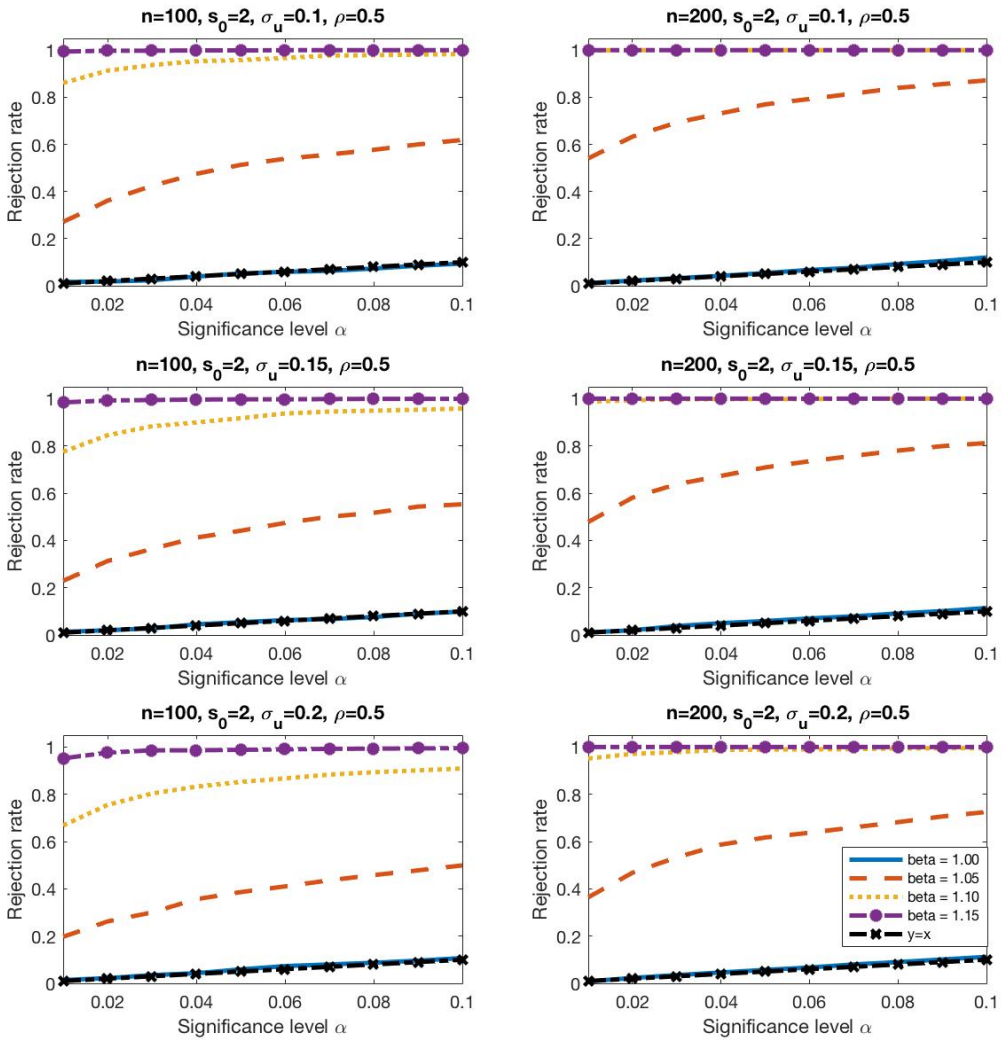

Figure B.1: Power of the proposed corrected decorrelated score test at different significance levels in scenario 1 with $\rho = 0.5$

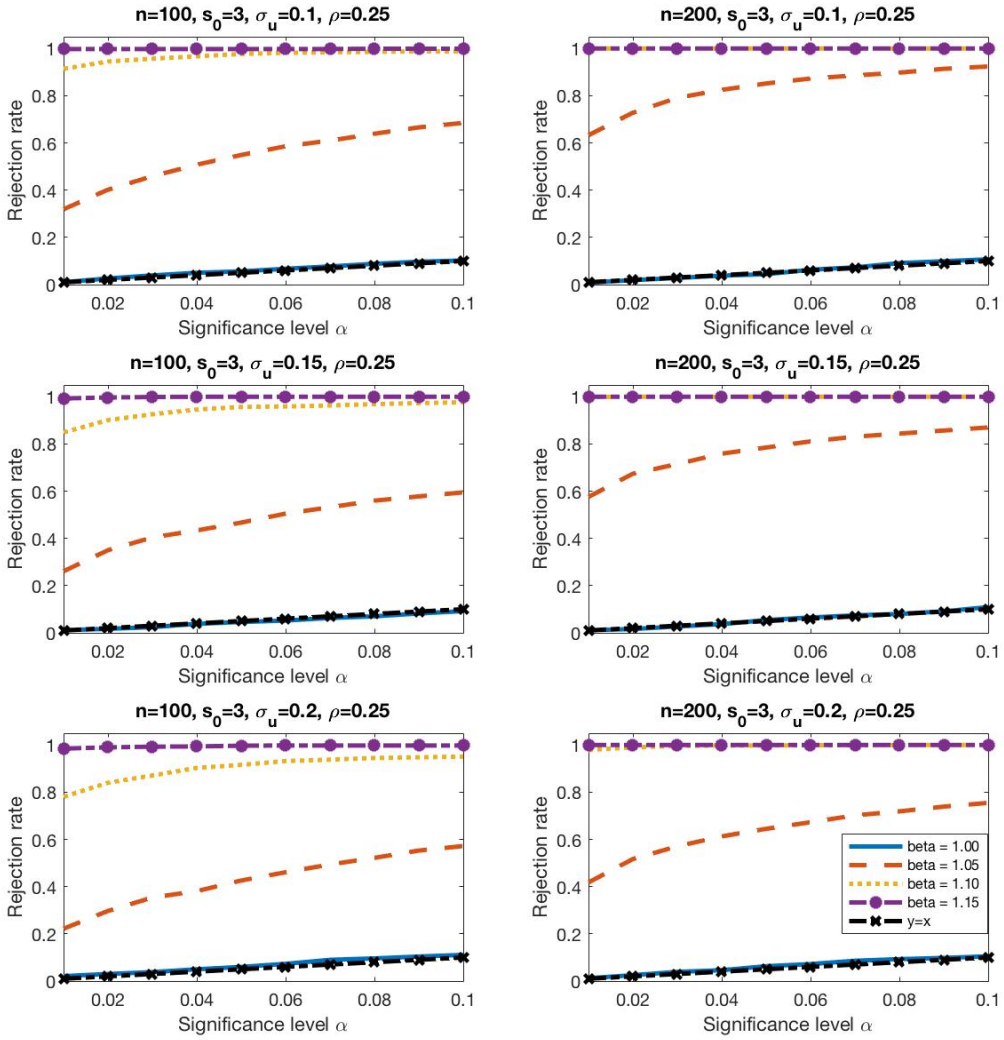

Figure B.2: Power of the proposed corrected decorrelated score test at different significance levels in scenario 2 with $\rho = 0.25$

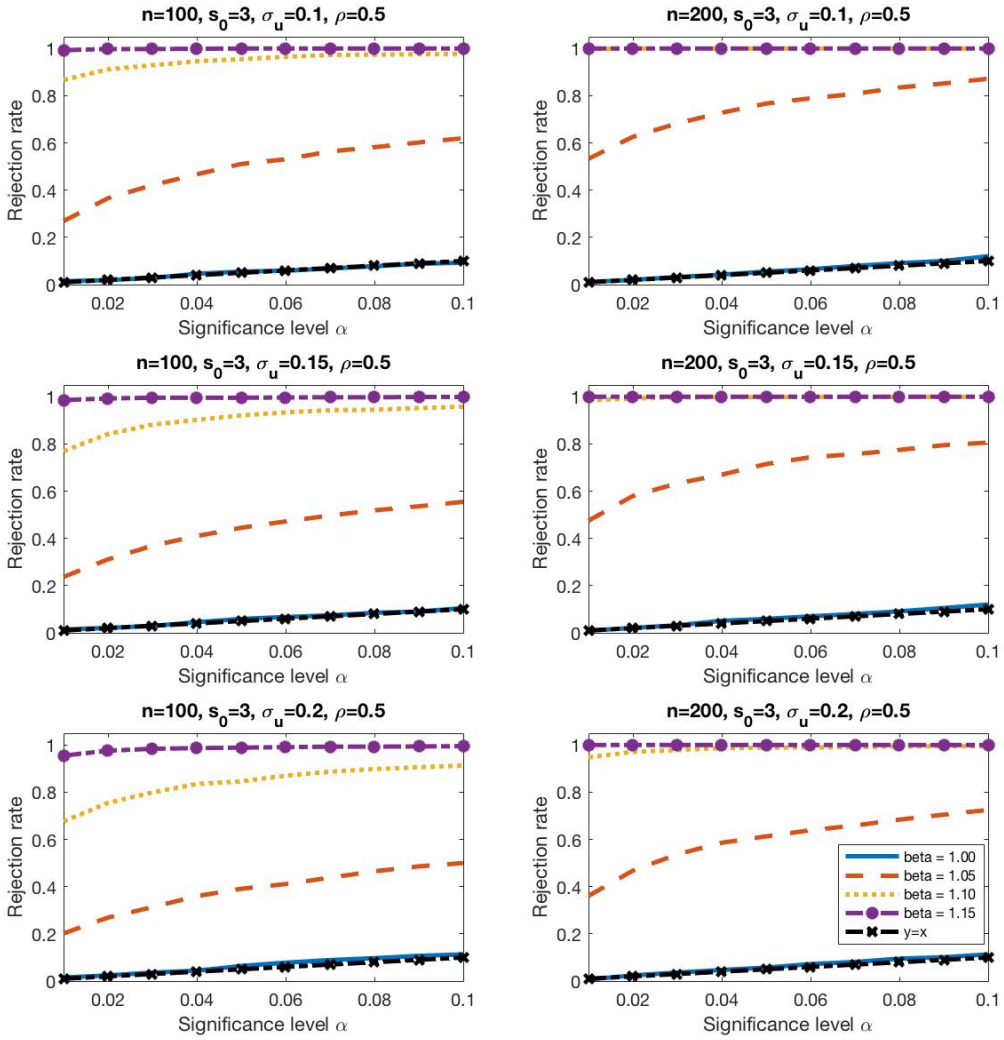

Figure B.3: Power of the proposed corrected decorrelated score test at different significance levels in scenario 2 with $\rho = 0.5$

B.7.2 Impact of estimated measurement error variance

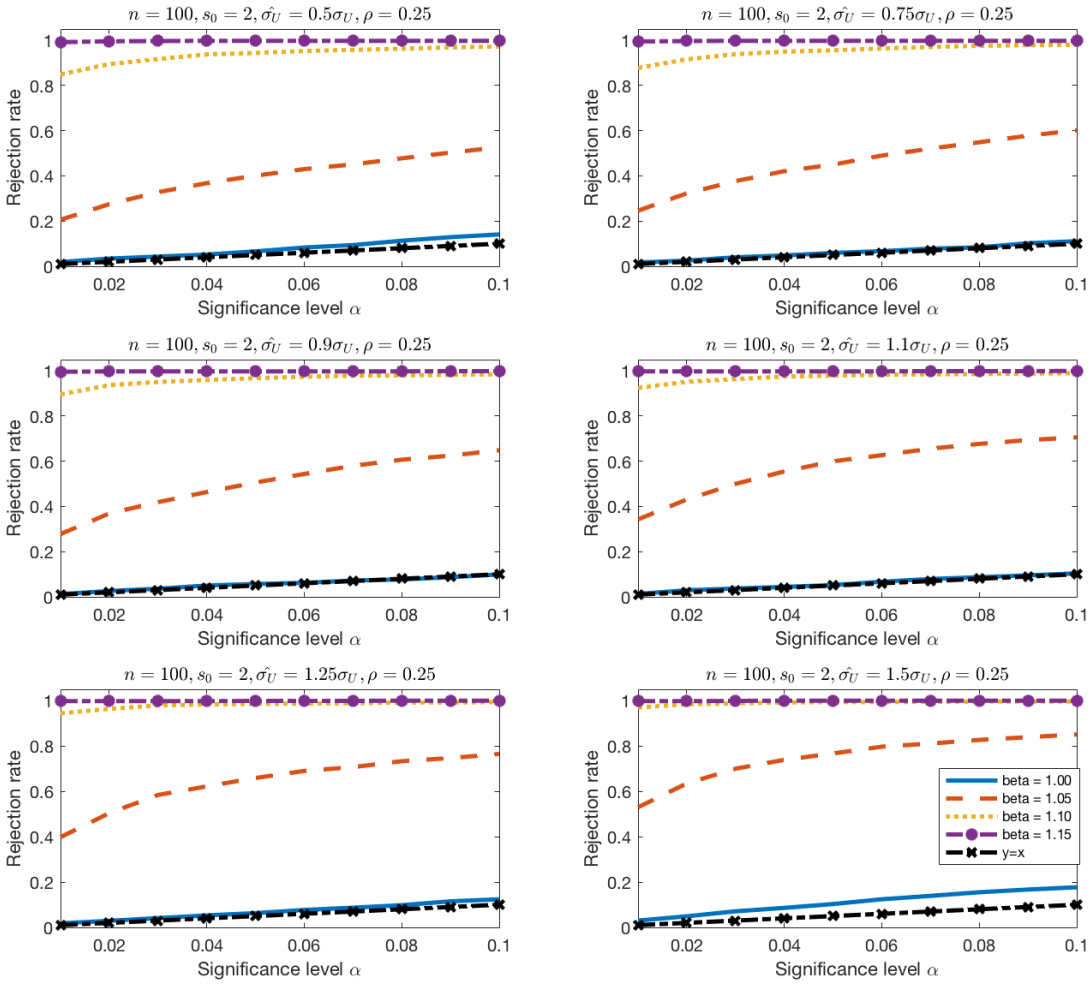

Figure B.4: The impact of $\widehat{\sigma}_U$ on the power performance of the proposed corrected decorrelated score test, where $n = 100$, $p = 250$, $\rho = 0.25$, $s_0 = 2$, $\sigma_u = 0.1$.

Table B.1: The impact of $\widehat{\sigma}_U$ on the one-step estimators, where $n = 100$, $p = 250$, $\rho = 0.25$, $s_0 = 2$, $\sigma_u = 0.1$

$\widehat{\sigma}_U$	$0.5\sigma_U$	$0.75\sigma_U$	$0.9\sigma_U$	$1.0\sigma_U$	$1.1\sigma_U$	$1.25\sigma_U$	$1.5\sigma_U$
Mean	0.9918	0.9955	0.9985	1.0008	1.0033	1.0076	1.0161
Est sd	0.0235	0.0236	0.0236	0.0237	0.0237	0.0238	0.0241
Emp sd	0.0243	0.0244	0.0245	0.0246	0.0247	0.0249	0.0254
Emp cvg	92.0%	94.2%	94.3%	94.1%	94.1%	92.6%	89.0%

"Est sd" denotes the mean of 1000 estimated asymptotic standard deviations; "Emp sd" denotes the empirical standard deviation of 1000 estimates; "Emp cvg" denotes the empirical coverage of the estimated 95% CI for $\boldsymbol{\beta}_0$.

Appendix C
Technical Proofs for Chapter 5

C.1 Derivations of score functions

The log-likelihood function is

$$\begin{aligned} & l(\boldsymbol{\alpha}, \boldsymbol{\beta}, g; \mathbf{x}, ry, r) \\ = \; & \log\{f_X(\mathbf{x})\} + r\log\{f_{Y|\mathbf{Z},R=1}(y, \mathbf{z}; \boldsymbol{\alpha})\} + r\log\{w(\mathbf{x})\} + (1-r)\log\{1-w(\mathbf{x})\}. \end{aligned}$$

Since

$$\begin{aligned} \frac{\partial w(\mathbf{x})}{\partial \boldsymbol{\beta}} &= w^2(\mathbf{x}) \int f_{Y|\mathbf{Z},R=1}(t, \mathbf{z}; \boldsymbol{\alpha}) \exp\{-g(t) - h(\mathbf{u}; \boldsymbol{\beta})\} \mathbf{h}'_{\beta}(\mathbf{u}; \boldsymbol{\beta}) dt \\ &= w^2(\mathbf{x}) \int f_{Y|\mathbf{Z},R=1}(t, \mathbf{z}; \boldsymbol{\alpha}) \{\pi^{-1}(t, \mathbf{u}; \boldsymbol{\beta}, g) - 1\} \mathbf{h}'_{\beta}(\mathbf{u}; \boldsymbol{\beta}) dt \\ &= w(\mathbf{x})\{1 - w(\mathbf{x})\} \mathbf{h}'_{\beta}(\mathbf{u}; \boldsymbol{\beta}), \\ \frac{\partial w(\mathbf{x})}{\partial \boldsymbol{\alpha}} &= -w^2(\mathbf{x}) \frac{\partial}{\partial \boldsymbol{\alpha}} \int f_{Y|\mathbf{Z},R=1}(t, \mathbf{z}; \boldsymbol{\alpha}) [1 + \exp\{-g(t) - h(\mathbf{u}; \boldsymbol{\beta})\}] dt \\ &= -w^2(\mathbf{x}) \exp\{-h(\mathbf{u}; \boldsymbol{\beta})\} E\left[\mathbf{S}(Y, \mathbf{z}; \boldsymbol{\alpha}) \exp\{-g(Y)\} \mid \mathbf{z}, 1\right], \end{aligned}$$

then

$$\begin{aligned} \mathbf{S}_{\beta}(\boldsymbol{\beta}, \boldsymbol{\alpha}, g; \mathbf{x}, ry, r) &= \frac{\partial w(\mathbf{x})}{\partial \boldsymbol{\beta}} \frac{r - w(\mathbf{x})}{w(\mathbf{x})\{1 - w(\mathbf{x})\}} \\ &= \{r - w(\mathbf{x})\} \mathbf{h}'_{\beta}(\mathbf{u}), \end{aligned}$$

and

$$\mathbf{S}_{\boldsymbol{\alpha}}(\boldsymbol{\beta}, \boldsymbol{\alpha}, g; \mathbf{x}, ry, r)$$

$$
\begin{aligned}
&= r\mathbf{S}(y,\mathbf{z},\boldsymbol{\alpha}) + \frac{\partial w(\mathbf{x})}{\partial \boldsymbol{\alpha}} \frac{r - w(\mathbf{x})}{w(\mathbf{x})\{1 - w(\mathbf{x})\}} \\
&= r\mathbf{S}(y,\mathbf{z};\boldsymbol{\alpha}) - E\left[\mathbf{S}(Y,\mathbf{z},\boldsymbol{\alpha}) \exp\{-g(Y)\} \mid \mathbf{z}, 1\right] \frac{w(\mathbf{x})\{r - w(\mathbf{x})\}}{1 - w(\mathbf{x})} \exp\{-h(\mathbf{u};\boldsymbol{\beta})\} \\
&= r\mathbf{S}(y,\mathbf{z};\boldsymbol{\alpha}) - E\left[\mathbf{S}(Y,\mathbf{z},\boldsymbol{\alpha})\{\pi^{-1}(Y,\mathbf{u};\boldsymbol{\beta},g) - 1\} \mid \mathbf{z}, 1\right] \frac{w(\mathbf{x})\{r - w(\mathbf{x})\}}{1 - w(\mathbf{x})} \\
&= r\mathbf{S}(y,\mathbf{z};\boldsymbol{\alpha}) - E\left[\mathbf{S}(Y,\mathbf{z},\boldsymbol{\alpha})\pi^{-1}(Y,\mathbf{u};\boldsymbol{\beta},g) \mid \mathbf{z}, 1\right] \frac{w(\mathbf{x})\{r - w(\mathbf{x})\}}{1 - w(\mathbf{x})} \\
&= r\mathbf{S}(y,\mathbf{z},\boldsymbol{\alpha}) - E\left\{\mathbf{S}(Y,\mathbf{z};\boldsymbol{\alpha}) \mid \mathbf{x}\right\} \frac{r - w(\mathbf{x})}{1 - w(\mathbf{x})}.
\end{aligned}
$$

C.2 Proof of Theorem 5.6.1

Proof. Let $\mathbf{o}_i = (\mathbf{x}_i, r_i y_i, r_i)$. Define $f_{\mathbf{X},RY,R}(\mathbf{o}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)$ as the likelihood of a parametric submodel with true parameters $\boldsymbol{\theta}_0$ and $\boldsymbol{\gamma}_0$. Note that any parametric submodel contains the true model, which implies that

$$
f_{\mathbf{X},RY,R}(\mathbf{o}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0) \;=\; f_{\mathbf{X},RY,R}(\mathbf{o}_i;\boldsymbol{\theta}_0,g).
$$

For term $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$, we expand the estimating equation (5.11) as a function for $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ about the truth $\boldsymbol{\theta}_0$ and $\boldsymbol{\gamma}_0$ to obtain

$$
\begin{aligned}
\mathbf{0} &= N_2^{-1/2} \sum_{i \in \mathcal{I}_2} \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}[\widehat{\boldsymbol{\theta}}, \widehat{g}\{\cdot;\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}})\};\mathbf{o}_i] \\
&= N_2^{-1/2} \sum_{i \in \mathcal{I}_2} \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, \widehat{g}(\cdot;\boldsymbol{\gamma}_0);\mathbf{o}_i\} + N_2^{-1} \sum_{i \in \mathcal{I}_2} \frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\widetilde{\boldsymbol{\theta}}, \widehat{g}(\cdot;\widetilde{\boldsymbol{\gamma}});\mathbf{o}_i\}}{\partial \widetilde{\boldsymbol{\theta}}^{\mathrm{T}}} N_2^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\
&\quad + N_2^{-1} \sum_{i \in \mathcal{I}_2} \frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\widetilde{\boldsymbol{\theta}}, \widehat{g}(\cdot;\widetilde{\boldsymbol{\gamma}});\mathbf{o}_i\}}{\partial \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}} N_2^{1/2}\{\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\gamma}_0\},
\end{aligned}
$$

where $\widetilde{\boldsymbol{\theta}}$ is on the line connecting $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. We will show that under regularity conditions (C1), (C2), (C3) and (C4)

$$
N_2^{-1} \sum_{i \in \mathcal{I}_2} \frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\widetilde{\boldsymbol{\theta}}, \widehat{g}(\cdot;\widetilde{\boldsymbol{\gamma}});\mathbf{o}_i\}}{\partial \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}} N_2^{1/2}\{\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\gamma}_0\} = o_p(1),
$$

element-wise.

First, for $j = 1, \ldots, p$ and $k = 1, \ldots, L$ we have

$$\begin{aligned} & N_2^{-1} \sum_{i \in \mathcal{I}_2} \left[\frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\widetilde{\boldsymbol{\theta}}, \widehat{g}(\cdot;\widetilde{\boldsymbol{\gamma}}); \mathbf{o}_i\}}{\partial \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}} \right]_{j,k} \\ = \; & N_2^{-1} \sum_{i \in \mathcal{I}_2} \left[\frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{o}_i\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \right]_{j,k} + O_p\{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 + \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 + h_L^m\} \\ = \; & E\left(\left[\frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{O}_i\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1 \right]_{j,k} \right) + O_p\{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 + \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 + h_L^m + N_2^{-1/2}\} \end{aligned} \tag{C.1}$$

Note that in $\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{o}_i\}$, $\widehat{\mathbf{a}}_0(y)$ and $\widehat{\mathbf{a}}_1(y)$ are estimated at the true $\boldsymbol{\theta}_0$ and $g(\cdot)$. Then $\widehat{\mathbf{a}}_0(\cdot)$ and $\widehat{\mathbf{a}}_1(\cdot)$ are deterministic conditional on $\mathcal{D}_1$. The first equality holds because

$$\| E\left[\frac{\partial}{\partial \boldsymbol{\theta}_0} \frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{O}_i\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1 \right] \|_{\max} \text{ and } \| E\left[\frac{\partial}{\partial \boldsymbol{\gamma}_0} \frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{O}_i\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1 \right] \|_{\max}$$

are bounded, and

$$\begin{aligned} & \sum_{l=1}^{L} N_2^{-1} \sum_{i \in \mathcal{I}_2} \frac{\partial}{\partial \gamma_{0,l}} \left[\frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{o}_i\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \right]_{j,k} (\widehat{\gamma}_l - \gamma_{0,l}) \\ = \; & \sum_{l=1}^{L} \left\{ E\left(\frac{\partial}{\partial \gamma_{0,l}} \left[\frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{O}_i\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1 \right]_{j,k} \right) + O_p(N_2^{-1/2}) \right\} (\widehat{\gamma}_l - \gamma_{0,l}) \\ = \; & O_p(\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1). \end{aligned} \tag{C.2}$$

Further, since the dimension of $\widehat{\boldsymbol{\theta}}$ is fixed, then any norm of $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ is of the same order.

We can also show that

$$\begin{aligned} & E\left[\frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{O}_i\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1 \right] \\ = \; & \int \frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{o}_i\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} f_{\mathbf{X},R,RY}(\mathbf{o}_i; \boldsymbol{\theta}_0, g) d\mathbf{o}_i \\ = \; & \int \frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{o}_i\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} f_{\mathbf{X},R,RY}(\mathbf{o}_i; \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0) d\mathbf{o}_i \\ = \; & \frac{\partial}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \int \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{o}_i\} f_{\mathbf{X},R,RY}(\mathbf{o}_i; \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0) d\mathbf{o}_i \\ & - \int \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{o}_i\} \frac{\partial f_{\mathbf{X},R,RY}(\mathbf{o}_i; \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} d\mathbf{o}_i \end{aligned}$$

$$
\begin{aligned}
&= -\int \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{o}_i\} \frac{\partial f_{\mathbf{X},R,RY}(\mathbf{o}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} d\mathbf{o}_i \\
&= -E\left[\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{O}_i\} \frac{\partial \log\{f_{\mathbf{X},R,RY}(\mathbf{O}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1\right] \\
&= -E\left[\mathbf{S}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{O}_i\} \frac{\partial \log\{f_{\mathbf{X},R,RY}(\mathbf{O}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}}\right] \\
&\quad -E\left(\left[\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{O}_i\} - \mathbf{S}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{O}_i\}\right] \frac{\partial \log\{f_{\mathbf{X},R,RY}(\mathbf{O}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1\right) \\
&= -E\left(\left[\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{O}_i\} - \mathbf{S}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0, g(\cdot;\boldsymbol{\gamma}_0); \mathbf{O}_i\}\right] \frac{\partial \log\{f_{\mathbf{X},R,RY}(\mathbf{O}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1\right) \\
&= O_p(N_1^{-1/2}),
\end{aligned}
$$

element-wise. Due to $f_{\mathbf{X}}$-robustness, we have

$$
\int \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}[\boldsymbol{\theta}, g\{\cdot;\boldsymbol{\gamma}(\boldsymbol{\theta})\}; \mathbf{o}_i] f_{\mathbf{X},R,RY}\{\mathbf{o}_i;\boldsymbol{\theta},\boldsymbol{\gamma}(\boldsymbol{\theta})\} d\mathbf{o}_i = \mathbf{0}
$$

for any parameters $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}(\boldsymbol{\theta})$. So the fourth equality holds. The seventh equality holds because $\mathbf{S}_{\boldsymbol{\theta},\text{eff}}$ is orthogonal to the nuisance tangent space Λ_g and the nuisance score for $\boldsymbol{\gamma}$ of any parametric submodel is in Λ_g. For the last equality, we have

$$
\begin{aligned}
&E\left[\{\widehat{\mathbf{S}}_{\boldsymbol{\beta},\text{eff}}(\boldsymbol{\theta}_0, g; \mathbf{O}_i) - \mathbf{S}_{\boldsymbol{\beta},\text{eff}}(\boldsymbol{\theta}_0, g; \mathbf{O}_i)\} \frac{\partial \log\{f_{\mathbf{X},R,RY}(\mathbf{O}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1\right] \\
&= E\left(\frac{w^3(\mathbf{X}_i)}{1-w(\mathbf{X}_i)} \exp\{-2h(\mathbf{U}_i;\boldsymbol{\beta}_0)\} E\left[\exp\{-g(Y;\boldsymbol{\gamma}_0)\} \frac{\partial g(Y;\boldsymbol{\gamma}_0)}{\partial \boldsymbol{\gamma}_0} \mid \mathbf{Z}_i, 1\right] \{\widehat{\mathbf{A}}_1(\mathbf{Z}_i) - \mathbf{A}_1(\mathbf{Z}_i)\} \mid \mathcal{D}_1\right)
\end{aligned}
$$

where

$$
\mathbf{A}_1(\mathbf{Z}_i) = E[\mathbf{a}_1(Y)\exp\{-g(Y)\} \mid \mathbf{Z}_i, 1], \text{ and } \widehat{\mathbf{A}}_1(\mathbf{Z}_i) = E[\widehat{\mathbf{a}}_1(Y;\boldsymbol{\theta}_0, g)\exp\{-g(Y)\} \mid \mathbf{Z}_i, 1].
$$

By (C4), we know that given $\mathcal{D}_1$, there exists a positive constant c such that

$$
\|[\mathbf{A}_1(\mathbf{z})]_k - [\widehat{\mathbf{A}}_1(\mathbf{z})]_k\|_2 \leqslant cN_1^{-1/2}, \quad k = 1, \ldots, p.
$$

Hence,

$$
E\left[\{\widehat{\mathbf{S}}_{\boldsymbol{\beta},\text{eff}}(\boldsymbol{\theta}_0, g; \mathbf{O}_i) - \mathbf{S}_{\boldsymbol{\beta},\text{eff}}(\boldsymbol{\theta}_0, g; \mathbf{O}_i)\} \frac{\partial \log\{f_{\mathbf{X},R,RY}(\mathbf{O}_i;\boldsymbol{\theta}_0,\boldsymbol{\gamma}_0)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1\right] = O_p(N_1^{-1/2}).
$$

Similarly, we can show that

$$E\left[\{\widehat{\mathbf{S}}_{\boldsymbol{\alpha},\text{eff}}(\boldsymbol{\theta}_0, g; \mathbf{O}_i) - \mathbf{S}_{\boldsymbol{\alpha},\text{eff}}(\boldsymbol{\theta}_0, g; \mathbf{O}_i)\}\frac{\partial \log\{f_{\mathbf{X},R,RY}(\mathbf{O}_i; \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0)\}}{\partial \boldsymbol{\gamma}_0^{\mathrm{T}}} \mid \mathcal{D}_1\right] = O_p(N_1^{-1/2}).$$

Therefore, by (C.1) we have

$$N_2^{-1}\sum_{i\in\mathcal{I}_2}\left[\frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\widetilde{\boldsymbol{\theta}}, \widehat{g}(\cdot;\widetilde{\boldsymbol{\gamma}}); \mathbf{o}_i\}}{\partial \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}}\right]_{j,k} = O_p\{\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_1 + \|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_1 + h_L^m + N_2^{-1/2} + N_1^{-1/2}\} \quad \text{(C.3)}$$

Second, we consider the term $\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\gamma}_0$. Since for any $\boldsymbol{\theta}$, we have

$$N_2^{-1}\sum_{i\in\mathcal{I}_2}\mathbf{S}_g\{\boldsymbol{\theta}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta}); \mathbf{o}_i\} = \mathbf{0},$$

then

$$\frac{\partial \widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^{\mathrm{T}}} = -\{\mathbf{T}_{11}(\boldsymbol{\theta})\}^{-1}\mathbf{T}_{12}(\boldsymbol{\theta}),$$

where

$$\mathbf{T}_{11}(\boldsymbol{\theta}) = N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial \mathbf{S}_g\{\boldsymbol{\theta}, \widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta}); \mathbf{o}_i\}}{\partial \widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta})^{\mathrm{T}}}, \quad \text{and} \quad \mathbf{T}_{12}(\boldsymbol{\theta}) = N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial \mathbf{S}_g(\boldsymbol{\theta}, \widehat{\boldsymbol{\gamma}}; \mathbf{o}_i)}{\partial \boldsymbol{\theta}^{\mathrm{T}}}.$$

By Taylor expansion we have

$$\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}}) - \widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta}_0) = \frac{\partial \widehat{\boldsymbol{\gamma}}(\widetilde{\boldsymbol{\theta}})}{\partial \widetilde{\boldsymbol{\theta}}^{\mathrm{T}}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = -\{\mathbf{T}_{11}(\widetilde{\boldsymbol{\theta}})\}^{-1}\mathbf{T}_{12}(\widetilde{\boldsymbol{\theta}})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

where $\widetilde{\boldsymbol{\theta}}$ is between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Further,

$$\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta}_0) - \boldsymbol{\gamma}_0 = -\{\mathbf{T}_{21}(\boldsymbol{\theta}_0)\}^{-1}\mathbf{T}_{22}(\boldsymbol{\theta}_0),$$

where

$$\mathbf{T}_{21}(\boldsymbol{\theta}_0) = N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial \mathbf{S}_g\{\boldsymbol{\theta}_0, \widetilde{\boldsymbol{\gamma}}(\boldsymbol{\theta}_0); \mathbf{o}_i\}}{\partial \widetilde{\boldsymbol{\gamma}}(\boldsymbol{\theta}_0)^{\mathrm{T}}}, \quad \text{and} \quad \mathbf{T}_{22}(\boldsymbol{\theta}_0) = N_2^{-1}\sum_{i\in\mathcal{I}_2}\mathbf{S}_g(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_0; \mathbf{o}_i),$$

and $\widetilde{\boldsymbol{\gamma}}(\boldsymbol{\theta}_0)$ is in the line connecting $\widehat{\boldsymbol{\gamma}}(\boldsymbol{\theta}_0)$ and $\boldsymbol{\gamma}_0$. We first consider the order of $\mathbf{T}_{22}(\boldsymbol{\theta}_0)$.

Since

$$\sup_y |\widehat{g}(y;\boldsymbol{\gamma}_0) - g(y)| = O_p(h_L^m),$$

then we have

$$E[\exp\{-\widehat{g}(Y;\boldsymbol{\gamma}_0)\} \mid \mathbf{z}_i, 1] = E[\exp\{-g(Y)\} \mid \mathbf{z}_i, 1] + O_p(h_L^m).$$

Due to the local linear approximation, we have

$$\|\mathbf{T}_{22}(\boldsymbol{\theta}_0)\|_\infty = O_p\{h_L^m + h^2 + (n_2 h)^{-1/2}\}.$$

Then we consider the order of off-diagonal elements in matrices $\mathbf{T}_{11}(\boldsymbol{\theta})$ and $\mathbf{T}_{21}(\boldsymbol{\theta}_0)$. For $j \neq l$, we have

$$\begin{aligned}
& N_2^{-1} \sum_{i \in \mathcal{I}_2} \frac{\partial S_{g,l}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\gamma}}; \mathbf{o}_i)}{\partial \widehat{\gamma}_j} \\
= \; & N_2^{-1} \sum_{i \in \mathcal{I}_2} \frac{(r_i - 1)\exp(-\widehat{\gamma}_l) f_{Y|\mathbf{Z},R=1}(d_l, \mathbf{z}_i; \widehat{\boldsymbol{\alpha}}) E[\exp\{-\widehat{g}(Y;\widehat{\boldsymbol{\gamma}})\}\{\partial \widehat{g}(Y;\widehat{\boldsymbol{\gamma}})/\partial \widehat{\gamma}_j\} \mid \mathbf{z}_i, 1]}{[1 + \exp\{-\widehat{\gamma}_l - h(\mathbf{u}_i; \widehat{\boldsymbol{\beta}})\}](E[\exp\{-\widehat{g}(Y;\widehat{\boldsymbol{\gamma}})\} \mid \mathbf{z}_i, 1])^2} \\
= \; & O_p(h_L).
\end{aligned}$$

The last equality holds because

$$\frac{\partial \widehat{g}(y;\widehat{\boldsymbol{\gamma}})}{\partial \widehat{\gamma}_j} \equiv 0, \quad \text{for } y \notin (d_{j-m+1}, d_{j+m-1}),$$

where $m-1$ is the degree of polynomial interpolation. Then we have

$$\|\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\gamma}_0\|_\infty = O_p\{h_L^m + h^2 + (n_2 h)^{-1/2}\}.$$

Under conditions C1, C2 and C3, by (C.3), we have

$$\begin{aligned}
& N_2^{-1} \sum_{i \in \mathcal{I}_2} \frac{\partial \widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\widetilde{\boldsymbol{\theta}}, \widehat{g}(\cdot;\widetilde{\boldsymbol{\gamma}}); \mathbf{o}_i\}}{\partial \widetilde{\boldsymbol{\gamma}}^{\mathrm{T}}} N_2^{1/2}\{\widehat{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\theta}}) - \boldsymbol{\gamma}_0\} \\
= \; & O_p(L N_2^{1/2}[L\{h_L^m + h^2 + (n_2 h)^{-1/2}\} + N_1^{-1/2} + N_2^{-1/2}]\{h_L^m + h^2 + (n_2 h)^{-1/2}\}) \\
= \; & O_p[L^2 N_2^{1/2}\{h_L^m + h^2 + (n_2 h)^{-1/2}\}^2 + L(N_2^{1/2} N_1^{-1/2} + 1)\{h_L^m + h^2 + (n_2 h)^{-1/2}\}] \\
= \; & o_p(1),
\end{aligned}$$

Hence, we have

$$N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0,\widehat{g}(\cdot;\boldsymbol{\gamma}_0);\mathbf{o}_i\}+N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\widetilde{\boldsymbol{\theta}},\widehat{g}(\cdot;\widetilde{\boldsymbol{\gamma}});\mathbf{o}_i\}}{\partial\widetilde{\boldsymbol{\theta}}^{\mathrm{T}}}N_2^{1/2}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=o_p(1).$$

Note that

$$\begin{aligned}
&N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\widetilde{\boldsymbol{\theta}},\widehat{g}(\cdot;\widetilde{\boldsymbol{\gamma}});\mathbf{o}_i\}}{\partial\widetilde{\boldsymbol{\theta}}^{\mathrm{T}}}\\
=\;&N_2^{-1}\sum_{i\in\mathcal{I}_2}\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0);\mathbf{o}_i\}}{\partial\boldsymbol{\theta}_0^{\mathrm{T}}}+O_p(\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_1+\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_1+h_L^m)\\
=\;&E\left[\frac{\partial\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0);\mathbf{O}_i\}}{\partial\boldsymbol{\theta}_0^{\mathrm{T}}}\mid\mathcal{D}_1\right]+O_p(\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_1+\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_1+h_L^m+N_2^{-1/2})\\
=\;&E\left\{\frac{\partial\mathbf{S}_{\boldsymbol{\theta},\text{eff}}(\boldsymbol{\theta}_0,g;\mathbf{O}_i)}{\partial\boldsymbol{\theta}_0^{\mathrm{T}}}\right\}+O_p(\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\|_1+\|\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_1+h_L^m+N_2^{-1/2}+N_1^{-1/2})\\
=\;&E\left\{\frac{\partial\mathbf{S}_{\boldsymbol{\theta},\text{eff}}(\boldsymbol{\theta}_0,g;\mathbf{O}_i)}{\partial\boldsymbol{\theta}_0^{\mathrm{T}}}\right\}+o_p(1).
\end{aligned}$$

Further, under conditions (C1), (C2) and (C3), we have

$$\begin{aligned}
N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0,\widehat{g}(\cdot;\boldsymbol{\gamma}_0);\mathbf{o}_i\}&=N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}\{\boldsymbol{\theta}_0,g(\cdot;\boldsymbol{\gamma}_0);\mathbf{o}_i\}+O_p(N_2^{1/2}h_L^m)\\
&=N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\mathbf{S}_{\boldsymbol{\theta},\text{eff}}(\boldsymbol{\theta}_0,g;\mathbf{o}_i)+O_p(N_1^{-1/2})+O_p(N_2^{1/2}h_L^m)\\
&=N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\mathbf{S}_{\boldsymbol{\theta},\text{eff}}(\boldsymbol{\theta}_0,g;\mathbf{o}_i)+o_p(1).
\end{aligned}$$

The second equality holds because $E\{R-w(\mathbf{X})\mid\mathbf{X}\}=0$ and

$$\begin{aligned}
&N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\{\widehat{\mathbf{S}}_{\boldsymbol{\theta},\text{eff}}(\boldsymbol{\theta}_0,g;\mathbf{o}_i)-\mathbf{S}_{\boldsymbol{\theta},\text{eff}}(\boldsymbol{\theta}_0,g;\mathbf{o}_i)\}\\
=\;&N_2^{-1/2}\sum_{i\in\mathcal{I}_2}\frac{r_i-w(\mathbf{x}_i)}{1-w(\mathbf{x}_i)}w(\mathbf{x}_i)\exp\{-h(\mathbf{u}_i;\boldsymbol{\beta}_0)\}\begin{pmatrix}\mathbf{A}_0(\mathbf{z}_i)-\widehat{\mathbf{A}}_0(\mathbf{z}_i)\\ \mathbf{A}_1(\mathbf{z}_i)-\widehat{\mathbf{A}}_1(\mathbf{z}_i)\end{pmatrix}\\
=\;&O_p(N_1^{-1/2}),
\end{aligned}$$

element-wise, where

$$\mathbf{A}_0(\mathbf{z}_i)=E[\mathbf{a}_0(Y)\exp\{-g(Y)\}\mid\mathbf{z}_i,1],\text{ and }\widehat{\mathbf{A}}_0(\mathbf{z}_i)=E[\widehat{\mathbf{a}}_0(Y;\boldsymbol{\theta}_0,g)\exp\{-g(Y)\}\mid\mathbf{z}_i,1].$$

Then we obtain that

$$N_2^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \;=\; -\left[E\left\{\frac{\partial \mathbf{S}_{\boldsymbol{\theta},\text{eff}}(\boldsymbol{\theta}_0, g; \mathbf{O}_i)}{\partial \boldsymbol{\theta}_0^{\mathrm{T}}}\right\}\right]^{-1} N_2^{-1/2} \sum_{i \in \mathcal{I}_2} \mathbf{S}_{\boldsymbol{\theta},\text{eff}}(\boldsymbol{\theta}_0, g; \mathbf{o}_i) + o_p(1).$$

□

Bibliography

Belloni, A., Chernozhukov, V. & Kaul, A. (2017), 'Confidence bands for coefficients in high dimensional linear models with error-in-variables', *arXiv preprint arXiv:1703.00469* .

Belloni, A., Rosenbaum, M. & Tsybakov, A. B. (2017), 'Linear and conic programming estimators in high dimensional errors-in-variables models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **79**(3), 939–956.

Berry, S. M., Carroll, R. J. & Ruppert, D. (2002), 'Bayesian smoothing and regression splines for measurement error problems', *Journal of the American Statistical Association* **97**(457), 160–169.

Bertrand, A., Legrand, C., Léonard, D. & Van Keilegom, I. (2017), 'Robustness of estimation methods in a survival cure model with mismeasured covariates', *Computational Statistics & Data Analysis* **113**, 3–18.

Bickel, P. J. (2007), 'Discussion: The dantzig selector: Statistical estimation when p is much larger than n', *Ann. Statist.* **35**(6), 2352–2357.

Candes, E., Tao, T. et al. (2007), 'The dantzig selector: Statistical estimation when p is much larger than n', *The Annals of Statistics* **35**(6), 2313–2351.

Carroll, R. J., Küchenhoff, H., Lombard, F. & Stefanski, L. A. (1996), 'Asymptotics for the simex estimator in nonlinear measurement error models', *Journal of the American Statistical Association* **91**(433), 242–250.

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. (2006), *Measurement error in nonlinear models: a modern perspective*, CRC press.

Carroll, R. J., Spiegelman, C. H., Lan, K. G., Bailey, K. T. & Abbott, R. D. (1984), 'On errors-in-variables for binary regression models', *Biometrika* **71**(1), 19–25.

Carroll, R. J. & Wand, M. P. (1991), 'Semiparametric estimation in logistic measurement error models', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 573–585.

Carter, R. L., Wrabetz, L., Jalal, K., Orsini, J. J., Barczykowski, A. L., Matern, D. & Langan, T. J. (2016), 'Can psychosine and galactocerebrosidase activity predict

early-infantile k rabbe's disease presymptomatically?', *Journal of neuroscience research* **94**(11), 1084–1093.

Chang, T. & Kott, P. S. (2008), 'Using calibration weighting to adjust for nonresponse under a plausible model', *Biometrika* **95**(3), 555–571.

Chen, Y. & Caramanis, C. (2013), Noisy and missing data regression: Distribution-oblivious support recovery, *in* 'International Conference on Machine Learning', pp. 383–391.

Cheng, C.-L. & Riu, J. (2006), 'On estimating linear relationships when both variables are subject to heteroscedastic measurement errors', *Technometrics* **48**(4), 511–519.

Chu, W., Li, R. & Reimherr, M. (2016), 'Feature screening for time-varying coefficient models with ultrahigh dimensional longitudinal data', *The Annals of Applied Statistics* **10**(2), 596.

Cleveland, W. S. (1979), 'Robust locally weighted regression and smoothing scatterplots', *Journal of the American statistical association* **74**(368), 829–836.

Datta, A., Zou, H. et al. (2017), 'Cocolasso for high-dimensional error-in-variables regression', *The Annals of Statistics* **45**(6), 2400–2426.

Daubechies, I. (1992), *Ten lectures on wavelets*, Vol. 61, Siam.

De Boor, C. (2001), *A practical guide to splines. Applied mathematical sciences.*, Vol. 27, Springer.

Devanarayan, V. & Stefanski, L. A. (2002), 'Empirical simulation extrapolation for measurement error models with replicate measurements', *Statistics & Probability Letters* **59**(3), 219–225.

Dezeure, R., Bühlmann, P., Meier, L. & Meinshausen, N. (2015), 'High-dimensional inference: Confidence intervals, p-values and r-software hdi', *Statistical Science* pp. 533–558.

d'Haultfoeuille, X. (2010), 'A new instrumental method for dealing with endogenous selection', *Journal of Econometrics* **154**(1), 1–15.

Dierckx, P. (1995), *Curve and surface fitting with splines*, Oxford University Press.

Eilers, P. H. & Marx, B. D. (1996), 'Flexible smoothing with b-splines and penalties', *Statistical science* pp. 89–102.

Eubank, R. L. (1999), *Nonparametric regression and spline smoothing*, CRC press.

Fan, J., Guo, S. & Hao, N. (2012), 'Variance estimation using refitted cross-validation in ultrahigh dimensional regression', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**(1), 37–65.

Fan, J., Xue, L. & Zou, H. (2016), 'Multitask quantile regression under the transnormal model', *Journal of the American Statistical Association* **111**(516), 1726–1735.

Fithian, W., Sun, D. & Taylor, J. (2014), 'Optimal inference after model selection', *arXiv preprint arXiv:1410.2597* .

Fuller, W. A. (1987, 2009), *Measurement error models*, Vol. 305, John Wiley & Sons.

Guo, Y. & Little, R. J. (2011), 'Regression analysis with covariates that have heteroscedastic measurement error', *Statistics in Medicine* **30**(18), 2278–2294.

Hanfelt, J. J. & Liang, K.-Y. (1995), 'Approximate likelihood ratios for general estimating functions', *Biometrika* **82**(3), 461–477.

Heyde, C. C. (1997, 2008), *Quasi-likelihood and its application: a general approach to optimal parameter estimation*, Springer Science & Business Media.

Heyde, C. & Morton, R. (1998), 'Multiple roots in general estimating equations', *Biometrika* **85**(4), 954–959.

Höfler, M., Pfister, H., Lieb, R. & Wittchen, H.-U. (2005), 'The use of weights to account for non-response and drop-out', *Social psychiatry and psychiatric epidemiology* **40**(4), 291–299.

Ibrahim, J. G. & Lipsitz, S. R. (1996), 'Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable', *Biometrics* pp. 1071–1078.

Javanmard, A. & Montanari, A. (2014), 'Confidence intervals and hypothesis testing for high-dimensional regression', *The Journal of Machine Learning Research* **15**(1), 2869–2909.

Jiang, F. & Ma, Y. (2018), 'A spline-assisted semiparametric approach to non-parametric measurement error models'.

Kim, J. K. & Shao, J. (2013), *Statistical Methods for Handling Incomplete Data*, Chapman & Hall/CRC.

Kim, J. K. & Yu, C. L. (2011), 'A semiparametric estimation of mean functionals with nonignorable missing data', *Journal of the American Statistical Association* **106**(493), 157–165.

Kott, P. (2014), Calibration weighting when model and calibration variables can differ, *in* F. F. Mecatti, L. P. Conti & G. M. Ranalli, eds, 'Contributions to Sampling Statistics', Springer International Publishing, Cambridge, pp. 1–18.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E. et al. (2016), 'Exact post-selection inference, with application to the lasso', *The Annals of Statistics* **44**(3), 907–927.

Little, R. J. & Rubin, D. B. (2002), *Statistical Analysis with Missing Data*, 2 edn, Wiley.

Loh, P.-L. & Wainwright, M. J. (2012), 'High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity', *Ann. Statist.* **40**(3), 1637–1664.

Loh, P.-L., Wainwright, M. J. et al. (2017), 'Support recovery without incoherence: A case for nonconvex regularization', *The Annals of Statistics* **45**(6), 2455–2482.

Ma, Y. & Carroll, R. J. (2006), 'Locally efficient estimators for semiparametric models with measurement error', *Journal of the American Statistical Association* **101**(476), 1465–1474.

Ma, Y. & Li, R. (2010*a*), 'Variable selection in measurement error models', *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability* **16**(1), 274.

Ma, Y. & Li, R. (2010*b*), 'Variable selection in measurement error models', *Bernoulli* **16**, 274–300.

Mallick, B. K. & Gelfand, A. E. (1996), 'Semiparametric errors-in-variables models a bayesian approach', *Journal of Statistical planning and Inference* **52**(3), 307–321.

Meinshausen, N., Bühlmann, P. et al. (2006), 'High-dimensional graphs and variable selection with the lasso', *The Annals of Statistics* **34**(3), 1436–1462.

Miao, W., Ding, P. & Geng, Z. (2016), 'Identifiability of normal and normal mixture models with nonignorable missing data', *Journal of the American Statistical Association* **111**(516), 1673–1683.

Miao, W., Liu, L., Tchetgen Tchetgen, E. & Geng, Z. (2019), 'Identification, doubly robust estimation, and semiparametric efficiency theory of nonignorable missing data with a shadow variable', *arXiv preprint arXiv:1509.02556* .

Molenberghs, G., Fitzmaurice, G., Kenward, M. G., Tsiatis, A. A. & Verbeke, G. (2014), *Handbook of Missing Data Methodology*, Boca Raton, Florida: Chapman & Hall/CRC Press.

Morikawa, K. & Kim, J. K. (2016), 'Semiparametric adaptive estimation with nonignorable nonresponse data', *arXiv preprint arXiv:1612.09207* .

Müller, P. & Roeder, K. (1997), 'A bayesian semiparametric model for case-control studies with errors in variables', *Biometrika* **84**(3), 523–537.

Nakamura, T. (1990), 'Corrected score function for errors-in-variables models: Methodology and application to generalized linear models', *Biometrika* **77**(1), 127–137.

Ning, Y., Liu, H. et al. (2017), 'A general theory of hypothesis tests and confidence regions for sparse high dimensional models', *The Annals of Statistics* **45**(1), 158–195.

Prentice, R. (1982), 'Covariate measurement errors and parameter estimation in a failure time regression model', *Biometrika* **69**(2), 331–342.

Qin, J., Leung, D. & Shao, J. (2002), 'Estimation with survey data under nonignorable nonresponse or informative sampling', *Journal of the American Statistical Association* **97**(457), 193–200.

Robins, J. M., Hsieh, F. & Newey, W. (1995), 'Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 409–424.

Robins, J. M. & Ritov, Y. (1997), 'Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models', *Statistics in Medicine* **16**(3), 285–319.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994), 'Estimation of regression coefficients when some regressors are not always observed', *Journal of the American statistical Association* **89**(427), 846–866.

Robins, J. M. & Wang, N. (2000), 'Inference for imputation estimators', *Biometrika* **87**(1), 113–124.

Rocke, D. M. & Durbin, B. (2001), 'A model for measurement error for gene expression arrays', *Journal of Computational Biology* **8**(6), 557–569.

Rosenbaum, M., Tsybakov, A. B. et al. (2010), 'Sparse recovery under matrix uncertainty', *The Annals of Statistics* **38**(5), 2620–2651.

Rotnitzky, A. & Robins, J. (1997), 'Analysis of semi-parametric regression models with non-ignorable non-response', *Statistics in Medicine* **16**, 81–102.

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, New York: John Wiley and Sons.

Sarkar, A., Mallick, B. K. & Carroll, R. J. (2014), 'Bayesian semiparametric regression in the presence of conditionally heteroscedastic measurement and regression errors', *Biometrics* **70**(4), 823–834.

Schafer, J. L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman & Hall.

Seaman, S. R., White, I. R., Copas, A. J. & Li, L. (2012), 'Combining multiple imputation and inverse-probability weighting', *Biometrics* **68**(1), 129–137.

Shao, J. & Wang, L. (2016), 'Semiparametric inverse propensity weighting for nonignorable missing data', *Biometrika* **103**(1), 175–187.

Shao, J. & Zhao, J. (2013), 'Estimation in longitudinal studies with nonignorable dropout', *Statistics and Its Interface* **6**, 303–313.

Silverman, B. W. (2018), *Density estimation for statistics and data analysis*, Routledge.

Slijepcevic, S., Megerian, S. & Potkonjak, M. (2002), 'Location errors in wireless embedded sensor networks: sources, models, and effects on applications', *ACM SIGMOBILE Mobile Computing and Communications Review* **6**(3), 67–78.

S**o**rensen, **O**., Frigessi, A. & Thoresen, M. (2015), 'Measurement error in lasso: Impact and likelihood bias correction', *Statistica Sinica* pp. 809–829.

Staudenmayer, J., Ruppert, D. & Buonaccorsi, J. P. (2008), 'Density estimation in the presence of heteroscedastic measurement error', *Journal of the American Statistical Association* **103**(482), 726–736.

Stefanski, L. A. (1985), 'The effects of measurement error on parameter estimation', *Biometrika* **72**(3), 583–592.

Stefanski, L. A. & Carroll, R. J. (1985), 'Covariate measurement error in logistic regression', *The Annals of Statistics* pp. 1335–1351.

Stefanski, L. A. & Cook, J. R. (1995), 'Simulation-extrapolation: the measurement error jackknife', *Journal of the American Statistical Association* **90**(432), 1247–1256.

Stone, C. J. (1977), 'Consistent nonparametric regression', *The annals of statistics* pp. 595–620.

Tang, G., Little, R. J. & Raghunathan, T. E. (2003), 'Analysis of multivariate missing data with nonignorable nonresponse', *Biometrika* **90**(4), 747–764.

Tsiatis, A. (2007), *Semiparametric theory and missing data*, Springer Science & Business Media.

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*, New York: Springer.

Tsiatis, A. A. & Ma, Y. (2004), 'Locally efficient semiparametric estimators for functional measurement error models', *Biometrika* **91**(4), 835–848.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R. et al. (2014), 'On asymptotically optimal confidence regions and tests for high-dimensional models', *The Annals of Statistics* **42**(3), 1166–1202.

Wainwright, M. J. (2009*a*), 'Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting', *IEEE Transactions on Information Theory* **55**(12), 5728–5741.

Wainwright, M. J. (2009*b*), 'Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1 -constrained quadratic programming (lasso)', *IEEE transactions on information theory* **55**(5), 2183–2202.

Wang, N., Carroll, R. & Liang, K.-Y. (1996), 'Quasilikelihood estimation in measurement error models with correlated replicates', *Biometrics* pp. 401–411.

Wang, S., Shao, J. & Kim, J. K. (2014), 'An instrumental variable approach for identification and estimation with nonignorable nonresponse', *Statistica Sinica* **24**, 1097–1116.

Wang, Y., Wang, J., Balakrishnan, S. & Singh, A. (2019), 'Rate optimal estimation and confidence intervals for high-dimensional regression with missing covariates', *Journal of Multivariate Analysis* **174**, 104526.

Wasserman, L. & Roeder, K. (2009), 'High dimensional variable selection', *Annals of statistics* **37**(5A), 2178.

Whittemore, A. S. & Gong, G. (1991), 'Poisson regression with misclassified counts: application to cervical cancer mortality rates', *Applied Statistics* pp. 81–93.

Yi, G. Y. (2016), *Statistical Analysis with Measurement Error Or Misclassification*, Springer.

Zhang, C.-H. & Zhang, S. S. (2014), 'Confidence intervals for low dimensional parameters in high dimensional linear models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76**(1), 217–242.

Zhao, J. & Ma, Y. (2018), 'Optimal pseudolikelihood estimation in the analysis of multivariate missing data with nonignorable nonresponse', *Biometrika* **105**, 479–486.

Zhao, J. & Ma, Y. (2019), 'A versatile estimation procedure without estimating the nonignorable missingness mechanism', *arXiv preprint arXiv:1907.03682* .

Zhao, J. & Shao, J. (2015), 'Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data', *Journal of the American Statistical Association* **110**(512), 1577–1590.

Zhao, P. & Yu, B. (2006), 'On model selection consistency of lasso', *Journal of Machine Learning Research* **7**(Nov), 2541–2563.

Vita

Mengyan Li

Mengyan Li was born in Taiyuan, China, in 1992. She received a B.S. degree in Mathematics and Applied Mathematics at Beijing Normal University in June 2015. She enrolled in the Ph.D. program in Statistics at the Pennsylvania State University in August 2015 and has been working under the supervision of Dr. Yanyuan Ma and Dr. Runze Li. Her research interests include semiparametric regression modeling, high-dimensional inference, measurement error models and nonignorable missing data.