2014

Checking the adequacy of regression models with complex data structure

Xu Guo
Hong Kong Baptist University

Checking the Adequacy of Regression Models with Complex Data Structure

GUO Xu

A thesis submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Principal Supervisor: Prof. ZHU Lixing

Hong Kong Baptist University

August 2014

Declaration

I hereby declare that this thesis represents my own work which has been done after registration for the degree of PhD at Hong Kong Baptist University, and has not been previously included in a thesis, dissertation submitted to this or other institution for a degree, diploma or other qualification.

Signature: _____

Date: August 2014

Abstract

In this thesis, we investigate the model checking problem for parametric regression model with missing response at random and nonignorable missing response. Besides, we also propose a hypothesis-adaptive procedure which is based on the dimension reduction theory. Finally, to extend our methods to missing response situation, we consider the dimension reduction problem with missing response at random.

The first part of the thesis introduces the model checking for parametric models with response missing at random which is a more general missing mechanism than missing completely at random. Different from existing approaches, two tests have normal distributions as the limiting null distributions no matter whether the inverse probability weight is estimated parametrically or nonparametrically. Thus, p-values can be easily determined. This observation shows that slow convergence rate of nonparametric estimation does not have significant effect on the asymptotic behaviours of the tests although it may have impact in finite sample scenarios. The tests can detect the alternatives distinct from the null hypothesis at a nonparametric rate which is an optimal rate for locally smoothing-based methods in this area. Simulation study is carried out to examine the performance of the tests. The tests are also applied to analyze a data set on monozygotic twins for illustration.

In the second part of the thesis, we consider model checking for general linear regression model with non-ignorable missing response. Based on an exponential tilting model, we first propose three estimators for the unknown parameter in the general linear regression model. Three empirical process-based tests are constructed. We discuss the asymptotic properties of the proposed tests under null and local alternative hypothesis with different scenarios. We find that these three tests perform the same in the asymptotic sense. Simulation studies are also carried out to assess the performance of our proposed test procedures.

In the third part, we revisit traditional local smoothing model checking procedures. Noticing that the general nonparametric regression model can be considered as a special multi-index model, we propose an adaptive testing procedure based on

the dimension reduction theory. To our surprise, our method can detect local alternative at faster rate than the traditional optimal rate. The theory indicates that in model checking problem, dimensionality may not have strong impact. Simulations are carried out to examine the performance of our methodology. A real data analysis is conducted for illustration.

In the last part, we study the dimension reduction problem with missing response at random. Based on the work in this part, we can extend the adaptive testing procedure introduced in the third part to the missing response situation. When there are many predictors, how to efficiently impute responses missing at random is an important problem to deal with for regression analysis because this missing mechanism, unlike missing completely at random, is highly related to high-dimensional predictor vector. In sufficient dimension reduction framework, the fusion-refinement (FR) method in the literature is a promising approach. To make estimation more accurate and efficient, two methods are suggested in this paper. Among them, one method uses the observed data to help on missing data generation, and the other one is an ad hoc approach that mainly reduces the dimension in the nonparametric smoothing in data generation. A data-adaptive synthesization of these two methods is also developed. Simulations are conducted to examine their performance and a HIV clinical trial dataset is analysed for illustration.

Keywords: Model checking; Inverse probability weight; Non-ignorable missing response; Adaptive; Central subspace; Dimension reduction; Data-adaptive Synthesization; Missing recovery; Missing response at random; Multiple imputation.

Acknowledgements

First and foremost, I would like to take this opportunity to express my great appreciation to my supervisor Prof. ZHU Lixing. Without his support and guidance, this work would not have been completed. Learning many different things from him in the fields of statistics, mathematics and research, and experiencing his diverse excellent qualities during the last years are what I will be ever thankful to him. I would also like to thank my co-supervisor Dr. TONG Tiejun, for his invaluable advice and helpful comments.

Also I would like to thank Prof. WONG Wing-Keung in the economics department of Hong Kong Baptist University. Without his continuous guidance and encouragement, I will not enter the economic area and enjoy much theoretical and empirical knowledge in economics and finance.

I wish to thank other faculty members in the mathematics department, especially CHUI Claudia, YUM Rainbow, LAM Tammy, YEUNG C. W., HUI Vicky and LI Candy for their excellent help. Also I would like to thank LO Kamfai of Graduate School for his great help.

Many people have helped and guided me–both professionally and personally–in the last few years. Prof. LIN Lu, Dr. XU Wangli, LI Qiuyue, YANG Yiping, PENG Heng and LI Gaorong have always been supportive and instructive. I am very grateful to Dr ZHU Liping, WU Jianhong, LI Zaixing, WU Ping, FANG Yun, YU Zhou, FENG Zhenghui, ZHANG Jun, FAN Yan, WANG Tao, XU Peirong, WANG Cheng, XIA Qiang, ZHOU Jingke, ZHU Xuehu and all the people of the Lixing research team for various combinations of help, support and inspiration. Special thanks go to TIAN Qiushi for her tremendous assistance and encouragement throughout my graduate career.

Finally, I wish to express my gratitude to my wife, NIU Cuizhen. I'm very happy and lucky to be married to this wonderful woman. Without her great support and

love, I would not complete this work smoothly. I also wish to thank my parents for their constant support and encouragement, both emotionally and intellectually.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Suppose that $Y \in \mathbb{R}$ is a univariate response and $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$ is a vector of predictors. In practice, we are always interested in how to characterize the relationship between the response Y and X. Due to easy interpretations, parametric regression models, especially linear regression models are wildly applied. If we are not sure about the parametric structure, to deal with the dimensionality, we are also interested in whether the relationship between Y and X is additive, whether the relationship is monotone or not. In other words, finding interesting data structures is an important topic. These parametric or non-parametric structures can be detected by developing suitable model checking procedures. On the other hand, we are also interested in finding some linear combinations of original predictors to reflect the information of X over Y, which can be achieved by dimension reduction. This thesis is concerned with the model checking problems with incomplete data. We also provide an interesting link between dimension reduction theory and model checking, which gives more insightful observations.

As two leitmotifs of statistics, model checking and dimension reduction are becoming progressively more prominent, both in theory and in practice.

1.1 Model Checking for Regression

If we have some priori information or knowledge about the form of the regression function, we may rely on linear regression, generalized linear regression or nonlinear regression models. On the other hand, if we have no idea about its form, we may adopt to the nonparametric regression models. As we know, when the dimension of X is high, the nonparametric estimators can not be accurate. It is also difficult to interpret the statistical results. Bandwidth selection in this case is also a difficult problem. When the parametric regression models hold, we can easily obtain efficient estimators. We can have better interpretability. However, statistical inferences are based on correct model building. If the parametric regression assumptions are in doubt, the inferences based on them can be very unreliable. Also we are interested in detecting whether there is some nonlinearity relationship between the response and the predictors. In sum, we should carry out some checking procedure to see if some simple parametric regression models are adequate to fit the data.

1.1.1 Smoothing-Based Tests

Considering a regression model with random design $Y = m(X)+\epsilon$, with $\{(X_i, Y_i)\}_{i=1}^n$ being a random sample of (X, Y). For a known parametric function $g(\cdot, \cdot)$, the goal is to test:

$$H_0 : m(X) = g(X, \theta_0),$$

for some θ_0 against alternative hypothesis

$$H_1 : m(X) \neq g(X, \theta),$$

for any θ.

Although there exist a large variety of smoothing methods for regression models, in the following, we will mainly focus on kernel type estimators given by $\hat{m}(x) =$

$\sum_{i=1}^{n} W_{ni}(x)Y_i$ with

$$W_{ni}(x) = \frac{K_h(x - X_i)}{\sum_{j=1}^{n} K_h(x - X_j)},$$

and $K_h(\cdot) = K(\cdot/h)/h^p$ with $K(\cdot)$ being a kernel function and h being the bandwidth.

Härdle and Mammen (1993) constructed a test statistic that is based on the L_2 distance between parametric and nonparametric estimators. To be precise, they proposed the following test statistic:

$$T_{HM} = \int \left(\sum_{i=1}^{n} W_{ni}(x)[Y_i - g(X_i, \hat{\theta})] \right)^2 \omega(x)dx,$$

here $\hat{\theta}$ is a $\sqrt{n}$-consistent estimator of θ_0 under H_0, such as the one obtained by least squares or maximum likelihood for Gaussian errors and $\omega(x)$ is some positive weight function.

Denote $\hat{m}_0(x) = \sum_{i=1}^{n} W_{ni}(x)g(X_i, \hat{\theta})$, the data smoother under H_0. This smoothed parametric estimator has the advantage of having the same bias as its nonparametric counterpart, which leads, in comparison with the nonsmoothed estimator $g(X_i, \hat{\theta})$, to considerable power improvements for small samples.

It can be shown that the limit distribution result of T_{HM} under the null hypothesis can be written as

$$\begin{aligned} & nh^{p/2} \left(T_{HM} - (nh^p)^{-1} \int K^2(x)dx \int \frac{\sigma^2(x)\omega(x)}{f(x)}dx \right) \\ & \Rightarrow N\left(0, 2\int (K * K)^2 dx \int \frac{\sigma^4(x)\omega^2(x)}{f^2(x)}dx\right), \end{aligned}$$

here $f(x)$ is the density of the explanatory variable X, $\sigma^2(x) = Var(Y|X = x)$ is the conditional variance, the symbol $*$ denotes the convolution operator and $K * K(x) = \int K(t)K(x - t)dt$.

González-Manteiga and Cao (1993) developed a discretized version of this test statistic as $T_{HM}^D = n^{-1} \sum_{i=1}^{n} W_{ni}(X_i)[Y_i - g(X_i, \hat{\theta})]^2\omega(X_i)$. This statistic consistently estimates the term $E(E^2(\epsilon_0|X)\omega(X))$ which is zero under H_0 with $\epsilon_0 = Y - g(X, \theta_0)$.

Zheng (1996) developed a quadratic form conditional moment test which was also independently proposed by Fan and Li (1996). Specially, Zheng proposed test

statistics which are consistent estimators of $E(\epsilon_0 E(\epsilon_0|X)f(X)\omega(X))$, another characteristics of the null hypothesis. Except for a negligible bias term, a natural estimate for this quantity is given by

$$T_{ZH} = \frac{1}{n(n-1)} \sum_{i \neq j} K_h(X_i - X_j)(Y_i - g(X_i, \hat{\theta}))(Y_j - g(X_j, \hat{\theta}))\omega(X_i).$$

Zheng (1996) presented the asymptotic distribution of this test statistic as follows:

$$nh^{p/2} T_{ZH} \Rightarrow N\left(0, 2\int K^2(x)dx \int \sigma^4(x)\omega^2(x)f^2(x)dx\right).$$

Thus, for T_{ZH}, there is no asymptotic bias and no bias-correction procedure is needed which is different from T_{HM}.

By noticing that $E\Big(\big[\epsilon_0^2 - (\epsilon_0 - E(\epsilon_0|X))^2\big]\omega(x)\Big)$ is zero under H_0, Dette (1999) introduced a test statistic based on the differences of the error variance estimates in the regression model. An estimate for this quantity is

$$T_{DE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i, \hat{\theta}))^2\omega(X_i) - \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{m}(X_i))^2\omega(X_i).$$

Denote $K^{2*} = 2K - K * K$, the asymptotic distribution of the variance difference statistic T_{DE} is

$$\begin{aligned} &nh^{p/2}\left(T_{DE} - (nh^p)^{-1}K^{2*}(0)\int \sigma^2(x)\omega(x)dx\right) \\ &\Rightarrow N\left(0, 2\int K^{2*}(x)dx \int \sigma^4(x)\omega^2(x)dx\right). \end{aligned}$$

For the comparison of these above three methods, kindly see Zhang and Dette (2004).

Inspired by the classical likelihood ratio test, Fan et al. (2001) considered a generalized likelihood ratio test which resembles the F-test construction for regression models. A significant property of this test statistic is that the asymptotic distribution does not depend on nuisance functions, exhibiting what is known as Wilks phenomenon. For more details, see also Fan and Jiang (2005, 2007).

For a general and updated review, see González-Manteiga and Crujeiras (2013). It should also be mentioned the book by Hart (1997), which collects a survey on the use of nonparametric smoothing methods for testing the fit of a parametric model.

1.1.2 Tests Based on Empirical Regression Processes

Another group of methodology for model checking is based on the empirical estimator of the integrated regression function $\mathcal{I}(x) = \int_{-\infty}^{x} m(t)dF(t) = E(YI(X \leq x))$, where $I(\cdot)$ is the indicator function. We can estimate $\mathcal{I}(x)$ by $\mathcal{I}_n(x) = n^{-1}\sum_{i=1}^{n} Y_i I(X_i \leq x)$. Then we can obtain an empirical process

$$\sqrt{n}(\mathcal{I}_n(x) - E_{\hat{\theta}}(\mathcal{I}_n(x))) = \sqrt{n}\sum_{i=1}^{n}(Y_i - g(X_i, \hat{\theta}))I(X_i \leq x).$$

Based on this empirical process, we can construct Cramér-von Mises or Kolmogorov-Smirnov type tests. Early studies for this kind of test statistic are due to Bierens (1982), Su and Wei (1991) and Stute (1997). However, for composite null hypothesis H_0, the above defined test statistics depend on the parametric form of $g(X, \theta)$ and also $\hat{\theta}$, thus, are not distribution free. Inspired by the Khmaladze transformation used in goodness of fit for distributions, Stute et al. (1998b) developed innovation martingale approach to obtain some distribution free tests for one dimension predictor. Khmaladze and Koul (2004) further studied the goodness-of-fit problem for errors in nonparametric regression.

Van Keilegom et al. (2008) considered continuous functionals of the distance between the empirical distribution of the residuals under the null and the alternative hypotheses. From this methodology, we can also construct test statistic for the goodness of fit for the error distribution. However, we should note that this test statistic needs the independence between the error term and the predictors, while other reviewed methods basically only assume that $E(\epsilon|X) = 0$. Huskova and Meintanis (2009) considered an alternative route based on the characteristic function requiring weaker conditions than their analogues using empirical distributions. For these two methods, see also Dette et al. (2007) and Huskova and Meintanis (2010).

Compared with smoothing-based tests, this kind of tests can avoid the selection of bandwidth, or at least do not depend too much on the bandwidth. Another advantage of this methodology is that the tests in this group can detect local alternatives

converging to the null hypothesis with the rate of $n^{-1/2}$, whereas for smoothing-based tests, the optimal rate is $(nh^p)^{-1/2}$. However, we should also mention that the higher detecting rate of empirical process based tests do not mean they can generally have larger power compared with the smoothing-based tests. It only means that they can have some power against closer alternatives. While these powers can be very low. In fact, empirical process based tests generally yard low powers against high frequency alternatives, see Fan and Li (2000). Furthermore these empirical process based tests are generally computationally intensive, especially when the dimension of X is high due to the sparsity of the data. Thus empirical process based tests and smoothing-based tests should be viewed as complements to each other.

1.2 Sufficient Dimension Reduction in Regression

To deal with the dimensionality problem, dimension reduction is necessary for us to efficiently work on regression analysis. Sufficient dimension reduction (SDR) has generated considerable interest in high-dimensional regressions. This general methodology aims at dealing with data sparseness in high-dimensional scenarios without parametric model structure and without loss of information on the regression of Y on X. The reduction is achieved by projecting raw predictors on to a lower-dimensional subspace.

In general, the central subspace (CS, Cook 1998), denoted by $\mathcal{S}_{Y|X}$, is defined as the intersection of all subspaces $\mathcal{S}$ of minimal dimension such that $Y \perp\!\!\!\perp X | P_{\mathcal{S}} X$, where $\perp\!\!\!\perp$ indicates statistical independence and $P_{(\cdot)}$ is a projection operator with respect to the usual inner product. When only the mean response $E(Y|X)$ is of interest, sufficient dimension reduction can be defined in a similar fashion. The space, denoted $\mathcal{S}_{E(Y|X)}$, is called the central mean subspace (CMS, Cook and Li 2002). It is, in effect, the intersection of all subspaces $\mathcal{S}$ such that $Y \perp\!\!\!\perp E(Y|X) | P_{\mathcal{S}} X$. In either case, sufficient dimension reduction permits us to restrict attention to a

number $d \leq p$ of new predictors, expressed as linear combinations of the original ones: $\beta_1^\top X, \ldots, \beta_d^\top X$, where $\{\beta_1, \ldots, \beta_d\}$ is a basis of $\mathcal{S}_{Y|X}$ or $\mathcal{S}_{E(Y|X)}$.

Since the pioneer research of Li (1991), many SDR methods have been developed. These estimation methods in the literature can be generally classified into three categories: inverse regression methods (e.g., Li 1991, Cook and Weisberg 1991, Cook and Ni 2005, Li and Wang 2007, Cook and Forzani 2009, Li and Dong 2009 and Zhu, Wang, Zhu and Ferré 2010), forward regression methods (e.g., Härdle et al. 1993, Hristache et al. 2001, Xia et al. 2002 and Xia 2006) and correlation approaches such as Fourier method (Zhu and Zeng, 2006), KL-distance (Yin and Cook, 2005, Yin, Li and Cook, 2008). Inverse regression methods are computationally simple and widely used. However, they require strong assumptions on the predictors such as the linearity condition (Li 1991) and often fail to estimate the central subspace exhaustively (Cook 1998). In contrast, direct regression methods need no strong requirements on the design distribution and have much better performance in finite samples.

1.2.1 Sliced Inverse Regression

Sliced inverse regression (SIR) method is a promising method for obtaining $\mathcal{S}_{Y|X}$. To implement SIR, a mild linearity condition is often assumed, that is, $E(X|B^\top X)$ is linear in $B^\top X$, here B denotes a basis of $\mathcal{S}_{Y|X}$. Given this condition, we can have that $\mathrm{Span}(\Sigma_X^{-1}\Sigma_{E(X|Y)}) \subseteq \mathcal{S}_{Y|X}$, where Σ_X is the non-singular covariance matrix of X and $\Sigma_{E(X|Y)} = \mathrm{Cov}\{\mathrm{E(X|Y)}\} \in \mathbb{R}^{\mathrm{p}\times\mathrm{p}}$. For simplicity, Li (1991) proposed a slicing estimation of $\mathrm{Cov}\{\mathrm{E(X|Y)}\}$. The range of Y are divided into M slices, $I_1, \cdots, I_M$, and $\mathrm{Cov}\{\mathrm{E(X|Y)}\}$ is approximated by

$$\Lambda = \sum_{s=1}^{M} p_s (m_s - \mu)(m_s - \mu)^\top, \tag{1.1}$$

where $\mu = E(X), p_s = P(Y \in I_s)$ and $m_s = E(X|Y \in I_s), s = 1, \cdots, M$. Let $\bar{X}$ and $\hat{\Sigma}_X$ denote the sample mean and sample covariance matrix of X, respectively. Let

$\hat{p}_s = n_s/n$ with n_s being the number of observations falling in slice s. Compute the sample mean of X in each slice, $\bar{X}_s, s = 1, \ldots, m$. Then we can obtain that

$$\widehat{\mathrm{Cov}}\{E(X|Y)\} = \sum_{s=1}^{M} \hat{p}_s(\bar{X}_s - \bar{X})(\bar{X}_s - \bar{X})^\top,$$

Conduct the spectral decomposition of $\widehat{\mathrm{Cov}}\{E(X|Y)\}$ with respect to $\hat{\Sigma}_X$ and let $\hat{\beta}_1, \ldots, \hat{\beta}_d$ be the eigenvectors corresponding to the d largest eigenvalues. Then, the associated SIR predictors are constructed as $\hat{\beta}_1^\top X, \ldots, \hat{\beta}_d^\top X$.

Under some mild conditions, Li (1991), Hsing and Carroll (1992) and Zhu and Ng (1995) showed the root n consistency of the SIR estimate, and Zhu and Ng (1995) particularly showed that SIR is not sensitive to the choice of the number of slices in theory, which echoes the empirical studies of Li (1991).

1.2.2 Minimum Average Variance Estimation

The minimum average variance estimation (MAVE) method of Xia et al. (2002) was originally proposed for dimension reduction for the conditional mean. Consider the following multiple-index model

$$Y = g(B^\top X) + \epsilon,$$

where g is an unknown smooth link function, B is a $p \times d$ orthogonal matrix with $d \leq p$, that is, $B^\top B = I_d$, and $E(\epsilon|X) = 0$ almost surely.

From the population, MAVE minimizes the objective function

$$E\{Y - E(Y|B^\top X)\}^2$$

over all $B \in \mathbb{R}^{p\times d}$. It's equivalent to minimize the following problem

$$\min_{B\in\mathbb{R}^{p\times d}} E\{\sigma_B^2(B^\top X)\} \quad \text{subject to } B^\top B = I_d,$$

where $\sigma_B^2(B^\top X) = E[\{Y - E(Y|B^\top X)\}^2|B^\top X]$ is the conditional variance of Y given $B^\top X$.

Now a random sample from (X, Y), $\{(X_i, y_i), i = 1, \ldots, n\}$, is available. For any given $X_0 \in \mathbb{R}^p$, $\sigma_B^2(B^\top X_0)$ can be approximated by local linear smoothing as

$$\sigma_B^2(B^\top X_0) \approx \sum_{i=1}^{n} \{y_i - E(y_i|B^\top X_i)\}^2 w_{i0} \approx \sum_{i=1}^{n} [y_i - \{a_0 + b_0^\top B^\top (X_i - X_0)\}]^2 w_{i0},$$

where w_{i0}'s are some positive weights satisfying $\sum_{i=1}^n w_{i0} = 1$, and $a_0 + b_0^\top B^\top (X_i - X_0)$ is the local linear expansion of $E(y_i|B^\top X_i)$ at X_0. Let $X_{ij} = X_i - X_j$. The MAVE procedure is to minimize

$$\frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} (y_i - a_j - b_j^\top B^\top X^{ij})^2 w_{ij} \tag{1.2}$$

over $a_j \in \mathbb{R}, b_j \in \mathbb{R}^d, j = 1, \ldots, n$, and $B \in \mathbb{R}^{p \times d}$ such that $B^\top B = I_d$.

The minimization problem in (1.2) can be solved by fixing $(a_j, b_j), j = 1, \ldots, n$, and B, alternatively. Given B we minimize, for $j = 1, \ldots, n$,

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - a_j - b_j^\top B^\top X_{ij})^2 w_{ij} \tag{1.3}$$

with respect to $a_j \in \mathbb{R}$ and $b_j \in \mathbb{R}^d$. Given $(a_j, b_j), j = 1, \ldots, n$, we minimize

$$\frac{1}{n} \sum_{j=1}^{n} \sum_{i=1}^{n} (y_i - a_j - b_j^\top B^\top X_{ij})^2 w_{ij} \tag{1.4}$$

over $B \in \mathbb{R}^{p \times d}$ subject to $B^\top B = I_d$.

The iterations between (1.3) and (1.4) can finally lead to the solutions. To reduce the effect of high dimension and improve the accuracy of estimation, Xia et al. (2002) proposed to use the lower-dimensional kernel weights

$$w_{ij} = \frac{K_h(B^\top X_{ij})}{\sum_{i=1}^n K_h(B^\top X_{ij})},$$

where B is the estimate from the previous iteration, and $K_h(\cdot)$ is a d-dimensional kernel function with bandwidth h.

MAVE has been found very useful in dimension reduction and semi-parametric modeling. Compared with other direct regression methods, the calculation for MAVE is much easier. Another important feature is that under-smoothing is unnecessary for

parameter estimators to achieve root-n consistency. Recently, the MAVE procedure was generalized to estimate the central subspace (Xia 2007; Wang and Xia 2008 and Yin and Li 2011).

1.3 Outline of the Thesis

In Chapter 2 we consider the model checking problem for parametric models with response missing at random which is a more general missing mechanism than missing completely at random. Different from existing approaches, two tests have normal distributions as the limiting null distributions no matter whether the inverse probability weight is estimated parametrically or nonparametrically. Thus, p-values can be easily determined. This observation shows that slow convergence rate of nonparametric estimation does not have significant effect on the asymptotic behaviours of the tests although it may have impact in finite sample scenarios. The tests can detect the alternatives distinct from the null hypothesis at a nonparametric rate which is an optimal rate for locally smoothing-based methods in this area. Simulation study is carried out to examine the performance of the tests. The tests are also applied to analyze a data set on monozygotic twins for illustration.

In survey about income or other sensitive quantities, non-ignorable missing response are encountered commonly. However, there has been no research about model checking with non-ignorable missing response. In Chapter 3, we consider model checking for general linear regression model with non-ignorable missing response. Based on an exponential tilting model, we first propose three estimators for the unknown parameter in the general linear regression model. Three empirical process-based tests are constructed. We discuss the asymptotic properties of the proposed tests under null and local alternative hypothesis with different scenarios. We find that these three tests perform the same in the asymptotic sense. Simulation studies are also carried out to assess the performance of our proposed test procedures.

Though the classical model checking procedures are useful and powerful when the dimension of X is small, their performances are greatly impacted by the dimension of X due to the nonparametric estimation of alternatives and data sparsity. In Chapter 4 we revisit traditional local smoothing model checking procedures. Noticing that the general nonparametric regression model can be considered as a special multi-index model, we propose an adaptive testing procedure based on the dimension reduction theory. To our surprise, our method can detect local alternative at faster rate than the traditional optimal rate. The theory indicates that in model checking problem, dimensionality may not have strong impact. Simulations are carried out to examine the performance of our methodology. A real data analysis is conducted for illustration.

To extend the adaptive testing procedure introduced in Chapter 4 to the missing response situation, we need to study the dimension reduction problem with missing response at random. Chapter 5 focuses on this topic. When there are many predictors, how to efficiently impute responses missing at random is an important problem to deal with for regression analysis because this missing mechanism, unlike missing completely at random, is highly related to high-dimensional predictor vector. In sufficient dimension reduction framework, the fusion-refinement (FR) method in the literature is a promising approach. To make estimation more accurate and efficient, two methods are suggested in this chapter. Among them, one method uses the observed data to help on missing data generation, and the other one is an ad hoc approach that mainly reduces the dimension in the nonparametric smoothing in data generation. A data-adaptive synthesization of these two methods is also developed. Simulations are conducted to examine their performance and a HIV clinical trial dataset is analysed for illustration.

Chapter 2

Model Checking for Parametric Regressions with Response Missing at Random

2.1 Introduction

The parametric regression model has received considerable attention, and relationship between the scalar response Y and the covariates X of dimension p is described as

$$Y = g(X, \theta_0) + \epsilon, \tag{2.1}$$

where $g(\cdot, \theta_0)$ is a known parametric function, θ_0 is an unknown parameter vector of m-dimension. It is assumed that the conditional expectation of ϵ given X is zero.

To prevent wrong conclusion and improve estimation efficiency, it is important to develop testing methods to ascertain whether the hypothetical parametric model is satisfied. When the response measurements are all available, there are a number of proposals available in the literature. For example, Härdle and Mammen (1993) constructed a test statistic that is based on the L_2 distance between parametric and nonparametric estimators with the assistance from the wild bootstrap for critical value determination. Zheng (1996) suggested a consistent test of functional form

of nonlinear regression models. Stute et al. (1998b) and Stute and Zhu (2002) considered to check the parametric regression models by replacing the residual cusum processes by their innovation martingale, and the resulting tests are asymptotically distribution-free. Aerts et al. (1999) constructed tests that are based on orthogonal series that involved selecting a nested model sequence in the bivariate regression. Fan and Huang (2001) introduced a method called Neyman threshold test by using the fact that the Fourier transform of the residuals compresses useful signals into low frequencies so that the power of the adaptive Neyman test can be enhanced. Stute et al. (2008) constructed a test that is based on the residual empirical process marked by proper functions of the regressors to deal with large dimension of the regressor vector. Eubank et al. (2005) proposed data-driven lack-of-fit tests based on nonparametric linear smoothers. All these tests can usually be classified into two categories: using local smoothing methods (nonparametric function fitting) and using global smoothing methods (empirical process). It is well known that the two types of methodologies have their own pros and cons. The former can be more sensitive to high-frequency alternative models than the tests based on the latter methodology, whereas the latter is more sensitive to smooth alternatives, and can detect the alternatives distinct from the null at faster convergence rate. Thus, both methodologies have been the main methodologies popularly used in practice. In this chapter, we construct tests that can be classified into the first category in our setting. We will see that the limiting null distributions are normal and thus determining p values is easily implemented. We will also make a limited comparison with a test in the latter category to see their advantages and disadvantages in the simulation study.

In practice, it is often the case that not all response measurements are observable. For example, due to limited budget, only the responses for a part of subjects among the fully cohort are measured. Individuals may refuse to answer certain questions, or investigators forget to write down the related information. Thus, it is of interest for us to investigate model checking with missing response. Among others,

González-Manteiga and Pérez-González (2006) constructed a test that is based on the L_2 distance between the nonparametric and parametric fits. Xu et al. (2012) defined a residual-marked empirical process to construct a test. Based on two completed samples, which are constructed by imputation and inverse probability weighting methods, Sun and Wang (2009) introduced two score type tests and two empirical process-based tests for the general linear models with missing response. Recently, Li (2012) proposed a test that is based on minimum integrated square distances between the nonparametric and parametric fits, which can be viewed as an extension of the minimum distance test proposed by Koul and Ni (2004) to handle missing responses.

In this Chapter, we propose two tests for model (2.1). For a known parameter function $g(\cdot, \cdot)$, almost surely, the null hypothesis is

$$H_0 : E(Y|X) = g(X, \theta_0), \tag{2.2}$$

for some θ_0 against alternative hypothesis

$$H_1 : E(Y|X) \neq g(X, \theta), \tag{2.3}$$

for any θ. The interesting feature of the newly proposed tests is that although the tests are also dependent on nonparametric smoothing, belonging to the category of local smoothing methodologies, the limiting null distributions are tractable for p-value determination. This advantage makes the tests easy to be implemented compared with existing ones. The tests can be regarded as an extension of Zheng (1996)'s test. As discussed above, there are many other possible approaches which can be used to handle the problem. We focus on the Zheng (1996)'s test in this chapter due to its technical tractability and easy computation. Dette and Von Lieres Und Wilkau (2001) compared several tests for additivity by kernel-based methods. They pointed out that, for realistic sample sizes, the bias has to be taken into account. For Zheng (1996)'s method, its standardized version has no bias converging to infinity and thus no bias-correction is needed. Gao et al. (2011) argued that a major advantage of Zheng (1996)'s method over its competitors is that an indirect estimator

of the unknown nonparametric $\sigma^2(X) = E(\epsilon^2|X)$ is used to replace $\sigma^2(X)$. They believed such a feature is attractive when the conditional variance function $\sigma^2(X)$ is a generally smooth function. Further, no matter whether the inverse probability is estimated parametrically or non-parametrically, the tests interestingly have the same asymptotic properties. Although in finite sample scenarios, they should have different performances. Compared with the test developed by Li (2012), both higher-order kernel functions and trimming on the boundary of a density function that are used in many applications of nonparametric regressions, are not needed. Also compared with Sun and Wang (2009), we do not need to construct the completed data set first for test statistic construction.

The rest of this Chapter is organized as follows. In Section 2.2, we construct the test statistics and derive their asymptotic properties under the null hypothesis and local alternatives. In Section 2.3, simulation results are reported to examine the performance of the tests and a real data analysis is carried out for illustration. We give some discussion in Section 2.4. The proofs are presented in the Appendix 2.5.

2.2 Test Procedures

2.2.1 Construction of Test Statistics

For model (2.1), it is assumed that the response Y is missing at random (MAR), while the observations for the covariate X are available. Let δ be the missing indicator for the individual whether Y is observed ($\delta = 1$) or not ($\delta = 0$). Then MAR implies

$$P(\delta = 1|Y, X) = P(\delta = 1|X) = \pi(X).$$

MAR is an usual missing mechanism in practice, which is more general than missing completely at random (MCAR), see Little and Rubin (1987).

Denote $\epsilon = Y - g(X, \theta_0)$, under the MAR assumption, we have

$$E(\frac{\delta}{\pi(X)}\epsilon|X) = E[\epsilon E(\frac{\delta}{\pi(X)}|X, Y)|X] = E[\epsilon|X].$$

Or equivalently

$$E(\delta\epsilon|X) = E[\epsilon E(\delta|X,Y)|X] = \pi(X)E(\epsilon|X).$$

Consequently, under H_0 of (2.2), we have

$$\begin{aligned} &E(\frac{\delta}{\pi(X)}\epsilon E(\frac{\delta}{\pi(X)}\epsilon|X)W(X)) = E(E^2(\frac{\delta}{\pi(X)}\epsilon|X)W(X)) = 0, \\ &E(\delta\epsilon E(\delta\epsilon|X)W(X)) = E(E^2(\delta\epsilon|X)W(X)) = 0, \end{aligned} \tag{2.4}$$

where $W(X)$ is some positive weight function which will be discussed below. Under the alternative hypothesis H_1, $E(\epsilon|X) \neq 0$, we have

$$\begin{aligned} &E(\frac{\delta}{\pi(X)}\epsilon E(\frac{\delta}{\pi(X)}\epsilon|X)W(X)) = E(E^2(\frac{\delta}{\pi(X)}\epsilon|X)W(X)) > 0, \\ &E(\delta\epsilon E(\delta\epsilon|X)W(X)) = E(E^2(\delta\epsilon|X)W(X)) > 0. \end{aligned} \tag{2.5}$$

Thus the null hypothesis H_0 holds if and only if the equations in (2.4) are zero. In other words, both can be used to be the bases for constructing test statistics. The empirical version of the left hand side in (2.4) can then be used to define test statistics. We will discuss their pros and cons later.

Let $(x_1, y_1, \delta_1), \cdots, (x_n, y_n, \delta_n)$ be an i.i.d. sample from (X, Y, δ). Estimate the terms $E(\delta\epsilon/\pi(X)|X = x)$ and $E(\delta\epsilon|X = x)$ by, respectively,

$$\begin{aligned} \hat{E}(\frac{\delta}{\pi(X)}\epsilon|x_i) &= \frac{1}{n-1}\sum_{j\neq i}^{n}\frac{\delta_j}{\hat{\pi}(x_j)}K_h(x_i - x_j)\hat{\epsilon}_j/\hat{f}(x_i), \\ \hat{E}(\delta\epsilon|x_i) &= \frac{1}{n-1}\sum_{j\neq i}^{n}\delta_j K_h(x_i - x_j)\hat{\epsilon}_j/\hat{f}(x_i). \end{aligned}$$

where $\hat{\epsilon}_j = y_j - g(x_j, \hat{\theta}_N)$ with $\hat{\theta}_N$ being an estimator of θ_0, which will be specified later, $\hat{\pi}(x)$ is a nonparametric estimator of $\pi(x)$, $K_h(\cdot) = K(\cdot/h)/h^p$ with $K(\cdot)$ being a kernel function and h being the bandwidth, and $\hat{f}(x)$ is the estimator of the density of X $f(x)$ defined as

$$\hat{f}(x_i) = \frac{1}{n-1}\sum_{j\neq i}^{n}\frac{\delta_j}{\hat{\pi}(x_j)}K_h(x_i - x_j).$$

Since our aim is to construct some efficient and simple tests, a natural selection of the weight function will be the density function $f(x)$ because it can eliminate the boundary effect of the kernel estimation. Two test statistics are defined as follows,

$$\begin{aligned} T_n^N &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}(x_i)} \hat{\epsilon}_i \sum_{j \neq i}^{n} \frac{\delta_j}{\hat{\pi}(x_j)} K_h(x_i - x_j) \hat{\epsilon}_j \\ &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\delta_i}{\hat{\pi}(x_i)} \frac{\delta_j}{\hat{\pi}(x_j)} K_h(x_i - x_j) \hat{\epsilon}_i \hat{\epsilon}_j; \\ R_n^N &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \delta_i \hat{\epsilon}_i \sum_{j \neq i}^{n} \delta_j K_h(x_i - x_j) \hat{\epsilon}_j \\ &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \delta_i \delta_j K_h(x_i - x_j) \hat{\epsilon}_i \hat{\epsilon}_j. \end{aligned} \tag{2.6}$$

In general, the function $\pi(X)$ is unknown and we can estimate it by a kernel estimator:

$$\hat{\pi}(x_i) = \frac{\sum_{j=1}^{n} \delta_j K_h(x_i - x_j)}{\sum_{j=1}^{n} K_h(x_i - x_j)}. \tag{2.7}$$

When $\pi(X)$ follows a parametric structure, that is, $\pi(X) = \pi(X, \alpha)$, we then only need to estimate the parameter α. As an example, for the logistic regression expressed as $\pi(x_i, \alpha) = (1 + \exp(-\alpha_0 - \alpha_1^\top x_i))^{-1}$, where $\alpha = (\alpha_0, \alpha_1)^\top$ is an unknown vector parameter, we can obtain consistent estimators of the regression coefficients $\hat{\alpha}$ by the maximum likelihood estimation. Then the corresponding estimator of $\pi(x, \alpha)$ follows as

$$\pi(x_i, \hat{\alpha}) = (1 + \exp(-\hat{\alpha}_0 - \hat{\alpha}_1 x_i))^{-1}. \tag{2.8}$$

Below we analyze the estimation of the regression parameter θ_0 by using the inverse probability weight least squares method:

$$\hat{\theta}_N = \arg\min \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}(x_i)} \{y_i - g(x_i, \theta)\}^2,$$

when $\pi(X)$ is estimated nonparametrically; and

$$\hat{\theta}_P = \arg\min \sum_{i=1}^{n} \frac{\delta_i}{\pi(x_i, \hat{\alpha})} \{y_i - g(x_i, \theta)\}^2,$$

when $\pi(X,\alpha)$ is estimated parametrically. Correspondingly, when the parameter function $\pi(X,\alpha)$ is estimated by $\pi(X,\hat{\alpha})$, we denote the statistics in (2.6) as, respectively,

$$\begin{aligned} T_n^P &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\delta_i}{\pi(x_i,\hat{\alpha})}\frac{\delta_j}{\pi(x_j,\hat{\alpha})}K_h(x_i-x_j)\hat{\epsilon}_i\hat{\epsilon}_j, \\ R_n^P &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\delta_i\delta_j K_h(x_i-x_j)\hat{\epsilon}_i\hat{\epsilon}_j, \end{aligned} \tag{2.9}$$

where $\hat{\epsilon}_j = y_j - g(x_j,\hat{\theta}_P)$.

Remark 2.1. *The proposed tests R_n^N and R_n^P are modifications of the complete case-based tests. Though we only use the completely observed units in the test constructions, an inverse probability weight method is adopted as seen on page 6 to estimate the parameter θ_0 and get $\hat{\epsilon}_i$. Use of the inverse probability weight method requires estimating $\pi(\cdot)$ or $\pi(\cdot,\alpha)$ by all the available data. Thus the asymptotic properties may not be directly derived from the transfer principle which was recently developed by Koul et al. (2012). In that paper, they proved the efficiency of complete case statistics in the situation of missing response at random situation, see also Müller and Van Keilegom (2012) and Chown and Müller (2013) for more discussions about the transfer principle. On the other hand, we note that the asymptotic properties of R_n^N and R_n^P are easier to develop compared with those for T_n^N and T_n^P. Thus in the appendix, we focus on the asymptotic properties of T_n^N and T_n^P.*

2.2.2 Asymptotic Behavior of the Test Statistics

Interestingly, we find that both the test statistics T_n^N in (2.6) and T_n^P in (2.9) have the same asymptotic properties. Also the asymptotic properties of R_n^N in (2.6) and R_n^P in (2.9) are equivalent. To state the theorems, we introduce some notations that

are related to the limiting variances of the test statistics. Let

$$\hat{\Sigma}^{TN} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{1}{h^m} \frac{\delta_i \delta_j}{\hat{\pi}^2(x_i)\hat{\pi}^2(x_j)} K^2(\frac{x_i - x_j}{h}) \hat{\epsilon}_i^2 \hat{\epsilon}_j^2,$$

$$\hat{\Sigma}^{RN} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{1}{h^m} \delta_i \delta_j K^2(\frac{x_i - x_j}{h}) \hat{\epsilon}_i^2 \hat{\epsilon}_j^2, \tag{2.10}$$

$\hat{\Sigma}^{TP}$ and $\hat{\Sigma}^{RP}$ are similarly defined as $\hat{\Sigma}^{TN}$ and $\hat{\Sigma}^{RN}$ respectively except for using $\pi(x_i, \hat{\alpha})$ and $\hat{\epsilon}_i = y_i - g(x_i, \hat{\theta}_P)$ instead of $\hat{\pi}(x_i)$ and $\hat{\epsilon}_i = y_i - g(x_i, \hat{\theta}_N)$, respectively. The asymptotic normalities for T_n^N, T_n^P and R_n^N, R_n^P under H_0 are stated below.

Theorem 2.1. *Under H_0 and the conditions in Appendix, we have*

$$nh^{p/2} T_n^N \to N(0, \Sigma^T), \text{ and } nh^{p/2} T_n^P \to N(0, \Sigma^T),$$

$$nh^{p/2} R_n^N \to N(0, \Sigma^R), \text{ and } nh^{p/2} R_n^P \to N(0, \Sigma^R),$$

where

$$\Sigma^T = 2 \int K^2(u)du \cdot \int \frac{(\sigma^2(x))^2 f^2(x)}{\pi^2(x)} dx.$$
$$\Sigma^R = 2 \int K^2(u)du \cdot \int (\sigma^2(x))^2 f^2(x) \pi^2(x) dx.$$

Moreover, Σ^T can be consistently estimated by $\hat{\Sigma}^{TN}$ or $\hat{\Sigma}^{TP}$, which depends on whether $\pi(x_i)$ is estimated parametrically or nonparametrically, and Σ^R can be consistently estimated by $\hat{\Sigma}^{RN}$ or $\hat{\Sigma}^{RP}$, that is: in probability

$$\hat{\Sigma}^{TN} \to \Sigma^T, \text{ and } \hat{\Sigma}^{TP} \to \Sigma^T;$$
$$\hat{\Sigma}^{RN} \to \Sigma^R, \text{ and } \hat{\Sigma}^{RP} \to \Sigma^R.$$

When there is no missing data, that is, $\pi(x) \equiv 1$, Σ^T and Σ^R are identical to Σ in Zheng (1996). Theorem 2.1 then is the same as Lemma 3.3 in Zheng (1996). Furthermore comparison with the results in the models without missing data, we can see clearly that the test statistics T_n^N and T_n^P induce larger asymptotic variances whereas R_n^N and R_n^P can have smaller asymptotic variances. However, this does not

mean that R_n^N and R_n^P generally are more powerful comparison with T_n^N and T_n^P since the powers of the tests also depend on the non-random drifts. We will discuss this point later. According to Theorem 2.1, the standardized versions of the test statistics $V_n^{TN}, V_n^{TP}, V_n^{RN}$ and V_n^{RP} can be defined as follows

$$V_n^{TN} = nh^{p/2}T_n^N/\sqrt{\hat{\Sigma}^{TN}}, \text{ and } V_n^{TP} = nh^{p/2}T_n^P/\sqrt{\hat{\Sigma}^{TP}}$$
$$V_n^{RN} = nh^{p/2}R_n^N/\sqrt{\hat{\Sigma}^{RN}}, \text{ and } V_n^{RP} = nh^{p/2}R_n^P/\sqrt{\hat{\Sigma}^{RP}}.$$

By Slusky Theorem, we have the following corollary.

Corollary 2.1. *Under H_0 and the conditions in Appendix, we have*

$$V_n^{TN} \to N(0,1), \text{ and } V_n^{TP} \to N(0,1);$$
$$V_n^{RN} \to N(0,1), \text{ and } V_n^{RP} \to N(0,1).$$

Thus, different from existing ones, it is easy to determine p values when our tests are applied. We now investigate the power behaviors of the tests under alternatives. We consider the following local alternatives:

$$H_{1n} : Y = g(X, \theta_0) + C_n G(X) + \eta, \tag{2.11}$$

where $E(\eta|X) = 0$ and the function $G(\cdot)$ satisfies $E(G^2(X)) < \infty$ and $\{C_n\}$ is a constant sequence. We have the following theorem.

Theorem 2.2. *Assume the same conditions as Theorem 1. Under the local alternatives H_{1n} we have, when $C_n = n^{-1/2}h^{-p/4}$,*

$$nh^{p/2}T_n^N \to N(\mu^T, \Sigma^T), \text{ and } nh^{p/2}T_n^P \to N(\mu^T, \Sigma^T)$$
$$nh^{p/2}R_n^N \to N(\mu^R, \Sigma^R), \text{ and } nh^{p/2}R_n^P \to N(\mu^R, \Sigma^R),$$

where

$$\mu^T = E\left[l^2(X)f(X)\right], \text{ and } \mu^R = E\left[l^2(X)\pi^2(X)f(X)\right],$$

with $\Sigma_1 = E(g'(X,\theta_0)g'^{\top}(X,\theta_0))$ and $l(X) = G(X) - g'^{\top}(X,\theta_0)\Sigma_1^{-1}E[G(X)g'(X,\theta_0)]$. The definitions of Σ^T and Σ^R are the same as those in Theorem 2.1.

When $n^{-1/2}h^{-p/4} = o(C_n)$, the test statistics converge in probability to infinity.

Theorem 2.2 indicates that the proposed tests have asymptotic power 1 for the local alternatives which are distinct from the null hypothesis at the rate slower than $n^{-1/2}h^{-p/4}$. Also the tests can still detect the alternatives converging to the null hypothesis at the rate $n^{-1/2}h^{-p/4}$, which is the same rate as that in Li (2012). However, since the asymptotic variances in Li (2012) and our are very different, it is not easy to tell which one can outperform the other in theory. Thus, a comparison will be made through simulation studies. Denote $D_1 = \mu^T/\sqrt{\Sigma^T}$ and $D_2 = \mu^R/\sqrt{\Sigma^R}$, from the above theorems, it can be also shown that, the asymptotic powers of T_n^N (or T_n^P) and R_n^N (or R_n^P) are $2-\Phi(z_{\alpha/2}-D_1)-\Phi(z_{\alpha/2}+D_1)$ and $2-\Phi(z_{\alpha/2}-D_2)-\Phi(z_{\alpha/2}+D_2)$, respectively for the alternatives, which are distinct from the null ones at rate $n^{-1/2}h^{-p/4}$. Here $\Phi(\cdot)$ is the standard normal distribution function, and $z_{\alpha/2}$ is the $\alpha/2$-th quantile. When the response is missing completely at random, that is, $0 < \pi(X) = c \leq 1$, after some simple calculations, we have $D_1 = D_2$. As a result, the test statistics T_n^N (or T_n^P) and R_n^N (or R_n^P) have the same asymptotic power in this special case.

When there is no missing data, we get similar results as Theorem 3 in Zheng (1996). We should note that we model the local alternatives slightly different from Zheng (1996). Zheng (1996) set the alternative as $m(X) = g(X, \tilde{\theta}_0) + C_n G(X)$. Here $\tilde{\theta}_0$ is the value of θ that minimizes $\tilde{S}_{0n}(\theta) = E[(m(X) - g(X,\theta))^2]$ with $m(X) = E(Y|X)$. Under H_0, $\tilde{\theta}_0 = \theta_0$. If the alternative hypothesis holds, $\tilde{\theta}_0$ will typically depend on $f(X)$. Note that under the local alternatives we design, $m(X) = g(X, \tilde{\theta}_0) + g(X, \theta_0) - g(X, \tilde{\theta}_0) + C_n G(X)$ and $\tilde{\theta}_0 - \theta_0 = C_n \Sigma_1^{-1} E[G(X)g'(X, \theta_0)]$. Thus $m(X) = g(X, \tilde{\theta}_0) + C_n l(X)$. Here $l(X) = G(X) - g'^{\top}(X, \theta_0)\Sigma_1^{-1} E[G(X)g'(X, \theta_0)]$.

Recall that $\tilde{\theta}_0$ is the value of θ that minimizes $\tilde{S}_{0n}(\theta) = E[(m(X) - g(X,\theta))^2]$. Thus, in Zheng's setting, we can have $E[G(X)g'(X, \tilde{\theta}_0)] = 0$. While, in our setting, there is also an orthogonality condition, that is, $E[l(X)g'(X, \theta_0)] = 0$. If we adopt the setting used in Zheng (1996), μ^T and μ^R will be equal to $E[G^2(X)f(X)]$ and $E[G^2(X)\pi^2(X)f(X)]$, respectively. Compared with the situations with no missing data, though the asymptotic variances of T_n^N and T_n^P are larger, the drift of them is

the same. That is, though the asymptotic variances of R_n^N and R_n^P are smaller, the drift of them is also smaller.

Consider the fixed alternative, $H_1 : m(X) = g(X, \theta_0) + G(X) = g(X, \tilde{\theta}_0) + \Delta(X)$, here $\Delta(X) = g(X, \theta_0) - g(X, \tilde{\theta}_0) + G(X)$. Note that even under the fixed alternative, according to White (1981), $\hat{\theta}_N$ is still a root-n consistent estimator of $\tilde{\theta}_0$. It is easy to see that $T_n^N = E[\Delta^2(X)f(X)] + o_p(1)$ and $R_n^N = E[\pi^2(X)\Delta^2(X)f(X)] + o_p(1)$. Similar results can be obtained for T_n^P and R_n^P. Thus the consistencies of the proposed tests are proved. Dette (1999) found an interesting phenomenon, that is, generally for the local smoothing based test procedures, the rate of convergence is different under the null hypothesis and the fixed alternatives. To be precise, while the rate is $(n^2h^p)^{-1}$ under H_0, it is of order n^{-1} under the fixed alternative. Dette and his coauthors showed that this is generally true in many different testing problems, see also Dette (2002), Dette and Spreckelsen (2003, 2004) and Dette and Hildebrandt (2012). We can prove that this is still true even when there are some missing data. Specially, we can have $\sqrt{n}(T_n^N - E[\Delta^2(X)f(X)]) \to N(0, \sigma_T^2)$ and $\sqrt{n}(R_n^N - E[\pi^2(X)\Delta^2(X)f(X)] \to N(0, \sigma_R^2)$. This can be proven by using Lemma 1 in appendix and the fact that $\sqrt{n}(\hat{\theta}_N - \tilde{\theta}_0) = O_p(1)$. We omit the details for saving space.

2.3 Numerical Analysis

2.3.1 Simulation Study

We now carry out simulations to examine the performance of the proposed test statistics and to compare the proposed statistics with the tests proposed by Li (2012) and Sun and Wang (2009), respectively. There are two statistics with the notations T_{n1}^S and T_{n2}^S in Sun and Wang (2009) for checking the adequacy of general linear models with missing response. Note that the behaviors of the T_{n1}^S and T_{n2}^S are very similar according to the simulation results in Sun and Wang (2009), we then only consider

T_{n2}^S. It's also interesting to compare with the tests based on empirical process. We note that Sun and Wang (2009) also developed two empirical process based tests with notations T_{n1}^E and T_{n2}^E, respectively. For the tests based on empirical process, see also Sun et al. (2009), where one procedure is given for testing the general partial linear model. Following their simulation studies, we know that T_{n1}^E and T_{n2}^E perform similarly, and thus we only consider T_{n2}^E in the following.

Study 1. To make the simulations comparable, we consider the same setting as that in Li (2012). The hypothetical model is linear as

$$Y = \theta^\top l(X) + \epsilon, \tag{2.12}$$

where $\theta = (0.5, 0.8)^\top$ and $l(X) = X = (X_1, X_2)$. The covariates $X_i = (X_{1i}, X_{2i}), i = 1, 2, \cdots, n$, are i.i.d. from bivariate normal distribution $N(0, \Sigma_j), j = 1, 2$ with

$$\Sigma_1 = \begin{pmatrix} 0.36 & 0.00 \\ 0.00 & 1.00 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} 1.00 & 0.64 \\ 0.64 & 1.00 \end{pmatrix},$$

, respectively. As for the random error term, we consider two distributions as Li (2012) did: $N(0, 0.3^2)$ and the double exponential distribution $DE(0, 3/10\sqrt{2})$ with density $f(x) = 5\sqrt{2}/3 \exp(-10\sqrt{2}/3|x|)$. Two missing probability mechanisms are considered for model (2.12), that is,

Case 1. $\pi_1(x) = P(\delta = 1|X = x) = 1/(1 + \exp(-(0.8 + 0.5x_1 + 0.5x_2))$.

Case 2. $\pi_2(x) = P(\delta = 1|X = x) = 1/(1 + \exp(-(0.2 + 0.3x_1 + 0.3x_2))$.

For the above two cases, the missing rates are 0.320 and 0.450, respectively. Here we consider the power performance of the tests under certain alternatives as follows:

$$H_{11} : Y = \theta^\top l(X) + 0.5(X_1 - 0.2)(X_2 - 0.4) + \epsilon;$$
$$H_{12} : Y = \theta^\top l(X) + 0.5(X_1 X_2 - 1) + \epsilon;$$
$$H_{13} : Y = \theta^\top l(X) + 2(\exp -0.4X_1^2 - \exp 0.6X_2^2) + \epsilon;$$
$$H_{14} : Y = \theta^\top l(X) + X_1 I_{X_2 > 0.2} + \epsilon.$$

where $I_{X_2>0.2}$ is an indicator, which equals to one if $X_2 > 0.2$ and otherwise zero. The kernel function takes the form $K(u,v) = K^1(u)K^1(v)$ with $K^1(u) = 0.75(1-u^2)I_{|u|\leq 1}$. The sample sizes are $n = 50$, 100 and 200. As for the bandwidth, we set it to be $n^{-1/4.5}$ which was used in Li (2012). Obviously, this bandwidth satisfies the conditions in Appendix. All the simulations are based on 1000 replications. The nominal level is set to be $\alpha = 0.05$. We denote the test proposed by Li (2012) as LI, the test T_{n2}^S and T_{n2}^E by Sun and Wang (2009) as SW^S and SW^E, and for our tests T_n^N as GXZ_{TN}, T_n^P as GXZ_{TP}, R_n^N as GXZ_{RN} and R_n^P as GXZ_{RP}, respectively.

Table 2.1 gives the empirical sizes and powers for testing H_0 against $H_{1i}, i = 1, \cdots, 4$ with design $X \sim N(0, \Sigma_1)$, $\epsilon \sim N(0, 0.3^2)$ when the data are randomly missing under either of the two missing mechanisms. The empirical sizes of these tests are all very close to the nominal level $\alpha = 0.05$. It is reasonable that the larger the sample size is, the closer the empirical sizes is to the nominal level. Among these tests, SW^S and SW^E have the best control on empirical size. About the power performance, it can be seen from Table 2.1 that all of our proposed tests are more powerful than LI and SW^S under all the designed alternative hypotheses except H_{14}. Under H_{14}, the most powerful one is the test LI. However, our tests are still competitive, that is, the gains of LI over our methods are limited. The power performance of SW^S is the worst under H_{11} and H_{14}. However, we also note that SW^E improves SW^S greatly under H_{11} and H_{14}. Specifically, under H_{11}, SW^E has similar power performance as those of our proposed tests. Under H_{14}, though there is some improvement of SW^E over SW^S, SW^E is still inferior to LI and our proposed tests. Further, under H_{13}, LI is the worst and our gain over LI, SW^S and SW^E is obvious. The impact of missing mechanisms on the empirical powers is evident. Generally, with the first missing mechanism, all these tests have greater powers. This is reasonable since there are less missing data with the first mechanism. Comparing the test GXZ_{TN} with GXZ_{TP}, GXZ_{TP} works better under H_{11} and H_{14} with larger power, where as GXZ_{TN} is better under H_{13}. It seems using a parametric estimation

of π gains not much compared with that using a nonparametric estimation. When we compare T_n with R_n, we can have the following results: under H_{11} and H_{14}, generally, R_n has larger power, while under H_{13}, T_n performs slightly better. In other words, overall, R_n works better. This seems to suggest that T_n may be affected by boundary effect more seriously to deteriorate its performance although in theory, it does not have such an issue. However, these comparisons cannot firmly say that among our proposed test statistics, which one is the best. But it seems that the tests with parametric estimation of $\pi(\cdot)$ is recommendable although it may have a misspecification issue.

Table 2.2 summaries the empirical sizes and powers for testing H_0 against $H_{1i}, i = 1, \cdots, 4$ when $X \sim N(0, \Sigma_2)$, $\epsilon \sim N(0, 0.3^2)$. In this case, X_1 is dependent of X_2. The only difference from the previous example is that X is from $N(0, \Sigma_2)$: the components are correlated. However, we can see, by a comparison with Table 2.1, that the performance on maintaining the nominal level is similar to that in the independent case. We can also see that the powers of all our proposed tests increase under H_{11} and H_{14}, while decrease under the other alternatives. Under H_{11}, SW^E is the most powerful and SW^S is also more powerful than our proposed tests and LI which is much different from that in the previous example. However, the opposite phenomenon happens under H_{12} for SW^S and SW^E. That is, the independence between X_1 and X_2 has a great influence on the behavior of SW^S and SW^E. LI is slightly more powerful than our tests under H_{12} whereas the proposed statistics are the most powerful under H_{13} and H_{14} in the most cases. Overall speaking, the proposed test is powerful under all the scenarios, SW^S and SW^E is not robust to the distribution of the covariates and further SW^E generally performs better than SW^S.

We now report the results about the testing for H_0 against $H_{1i}, i = 1, \cdots, 4$ with $\epsilon \sim DE(0, 3/10\sqrt{2})$ and $X \sim N(0, \Sigma_1)$ in Table 2.3, and $X \sim N(0, \Sigma_2)$ in Table 2.4 respectively. The comparisons between those in Table 2.1 and Table 2.3, or those in Table 2.2 and Table 2.4, are made to see the impact from the distributions.

The powers of the proposed tests increase greatly under H_{11} and H_{14} in Table 2.3 when comparing with those in Table 2.1, while the performance of SW^S and SW^E is slightly affected by the error distribution. Compared with LI, SW^S and SW^E, our tests have highest power under any alternatives in Table 2.3, and under H_{12}, and H_{13} in Table 2.4. From all these four tables, we can know that the proposed tests perform well and are the most powerful under many scenarios. For T_n and R_n, it seems that the latter performs better overall.

In study 1, normal distribution is used to determine critical values such that the empirical sizes and powers of the proposed tests can be conveniently computed. However, it is also well known that the rate of convergence to the normal limit is slow, see also Härdle and Mammen (1993), Stute et al. (1998a) and González-Manteiga and Crujeiras (2013). Thus the use of the asymptotic normality may be inappropriate for small sample sizes. This can also be seen from Tables 2.1-2.4 in study 1. To be precise, though the simulated sizes are very close to the nominal level, they generally underestimate it. As an alternative for calibrating critical values, we consider the residual based bootstrap in the following. Let the bootstrap errors $\epsilon_1^*, \cdots, \epsilon_n^*$ be an independent sample from the empirical distribution function of the centered residuals $\tilde{\epsilon}_i = \hat{\epsilon}_i - n^{-1}\sum_{l=1}^n \hat{\epsilon}_l$. Then we generate the bootstrap observations:

$$y_i^* = g(x_i, \hat{\theta}_N) + \epsilon_i^*.$$

Let T_n^{N*} be defined similarly as T_n^N, basing on the bootstrap sample $(x_1, y_1^*), \cdots, (x_n, y_n^*)$. The null hypothesis is rejected if T_n^N is bigger than the corresponding quantile of the bootstrap distribution of T_n^{N*}. We denote the bootstrap version of T_n^N as GXZ_{TN}^*. The bootstrap version of other proposed tests can be similarly developed and are denoted as GXZ_{TP}^*, GXZ_{RN}^* and GXZ_{RP}^* respectively. The study of the asymptotic validity of this procedure in the presence of missing response will be undertaken elsewhere. Here, we investigate the empirical properties of this bootstrap procedure when the same settings with $X \sim N(0, \Sigma_1)$ and $\epsilon \sim N(0, 0.3^2)$ as in study 1 are considered. For comparison, we use the same bandwidth $n^{-1/4.5}$ first. The number of

replications was 1000 and for each replication 500 bootstrap samples were generated. The results are presented in Table 2.5. From this table, we can see clearly that even with the sample size $n = 25$, the resampling method can control the type I error very well. As for the power performance, compare Table 2.1 with Table 2.5, we can conclude that the bootstrap performs better than the normal approximation under the alternative H_{11}, H_{12} and H_{14}. While under H_{13}, the normal approximation is the winner. Overall, the resampling method is recommendable for calibration especially when the sample size is small. Due to the computational intensiveness, when we have sufficient observations, the normal approximation is still recommendable.

Notice in the above study, we don't investigate the impact of the bandwidth on the performance of our tests and the considered alternative hypothesises are limited to the fixed alternatives. In the following study, we aim to study the bandwidth selection problem and the performances of our proposed tests under some local alternative.

Study 2. The local alternative is taken to be:

$$H_{1n} : Y = \theta^\top l(X) + C_n(X_1 - 0.2)(X_2 - 0.4) + \epsilon, \tag{2.13}$$

here we adopt the same setting as that in study 1 except we consider the local alternative indexed by C_n instead the fixed alternative H_{11} in study 1. It is evident that the null hypothesis $H_0 : Y = \theta^\top l(X) + \epsilon$ is valid if and only if $C_n = 0$. In this study, we set $X \sim N(0, \Sigma_1)$ and $\epsilon \sim N(0, 0.3^2)$ for space considerations. We only consider the first missing probability mechanism in study 1 to save space.

Zhu and Ng (2003) pointed out it's still an open problem about how to select optimal bandwidth in the testing problems. Though in nonparametric estimation literature the bandwidth selection has been discussed extensively, the selected optimal bandwidth for estimation may not yard the optimal power and size performance under different alternatives. To investigate the impact of bandwidth selection on our proposed tests, we take the bandwidth h to be $j/100$ for $j = 11, 15, 19 \cdots, 99$. Based on the 1000 simulations, we plot the estimated size and power curve against the above bandwidth sequences with the sample size 50, missing mechanism $\pi_1(x)$

and $C_n = Cn^{-1/2}$ with $C = 0, 2, 4$, which is shown in Figure 2.1. This strategy is also conducted by many authors, such as Sun and Wang (2009) and Lopez and Patilea (2009). From Figure 2.1, we have the following observations: (1) with the sample size $n = 50$, the proposed tests with different bandwidths can control the size reasonably. To be precise, the empirical sizes are all contained in the range of $(0.04, 0.06)$. In one word, the bandwidth selection for the size control for our proposed tests are not too critical. (2) generally, we can get larger powers if we use a relatively larger bandwidth. However, when we take the bandwidth not too small, the gain by employing a larger bandwidth is marginal. As discussed by Sperlich (2013), most, if not all, of the known methods are computationally expensive and somewhat very complex. Based on the above observations and suggestions by Lavergne and Vuong (2000), in the following simulations, we choose $h = 1.25 \times n^{-1/6}$ which is at the same rate as the optimal bandwidth derived in nonparametric estimation. Also note in study 1, we set $h = n^{-1/4.5}$ which is smaller than the selected bandwidth $h = 1.25 \times n^{-1/6}$. Further from Figure 1, we know the powers of our proposed tests with $h = n^{-1/4.5}$ are smaller than that with $h = 1.25 \times n^{-1/6}$. In other words, the power performance of the proposed tests in study 1 can be improved by using the selected bandwidth $h = 1.25 \times n^{-1/6}$.

We evaluate the performance of our proposed tests under the above defined local alternative (2.13) through varying the values of C_n, different sample size $n = 25, 50$ and $n = 100$ and missing mechanism $\pi_1(x)$. The simulation results are shown in Table 2.6. From the table, we can have similar findings except the following observations. When the local alternative hypothesis holds, that's, $C_n \neq 0$, the powers of our tests increases quickly as C_n in (2.13) increases. To be precise, the power of R_n^P is 0.903 under the sample size $n = 100$, missing mechanism $\pi_1(x)$ and $C_n = 6n^{-1/2} = 0.6$. In other wards, the tests are very sensitive to the alternatives. Moreover from these two tables, we can also conclude R_n^P performs best among these four proposed tests.

2.3.2 Real Data Analysis

Consider the data set about monozygotic twins with the sample size 50. In this data set, the response Y stands for birth-weight of a baby and two corresponding covariates X_{AC}(=AC) and X_{BDP} (=BDP) respectively for abdominal circumference and biparietal diameter. The data set has been used to test whether the nonparametric part in a partial linear model with missing response is of parametric form by Xu et al. (2012). They found that a parametric model is plausible. However, when we get the residual plots and found that a linear model would be further plausible. Thus, we make a further check to see whether a linear model is adequate.

Consider the null hypothesis

$$H_0 : E(Y|X) = X_{BDP}\beta_{BDP} + X_{AC}\beta_{AC} \tag{2.14}$$

for some β_{BDP} and β_{AC}. First all the variables are centered and the same notations are given without confusion. We illustrate our methods by missing 20% of the response randomly. Then we try to obtain the results by 2000 simulation runs in which each time we use the kernel function in subsection 2.3.1 for computation. We present the scatter plot for the covariates and the outcome shown in Figure 2.2. From our simulations in the study 2 in the subsection 2.3.1, we set the bandwidth to $\text{std}(x_{AC}) \times n^{-1/6}$. Under these settings, the p-values for T_n^N, R_n^N, T_n^P and R_n^P are 0.900, 0.865, 0.802 and 0.875 respectively. As a result, the null hypothesis in (2.14) cannot be rejected.

2.4 Discussion

In this Chapter, we extend Zheng (1996)'s method to adopt to missing response at random due to its technical tractability and easy computation. The asymptotic properties are developed for our proposed tests under null and local alternatives. The intensive simulation studies suggest that our proposed tests can perform well. To

better control the empirical size, we also propose to use the residual based bootstrap. Through our simulations, we also find that the use of different bandwidth has almost no effect on the empirical sizes of our proposed tests. On the other hand, the powers of the tests can be improved if we use a slightly larger bandwidth. Based on these observations, we suggest to use a rule of thumb which may be not optimal in all directions. The problem of choosing the bandwidth to optimize the power remains an open problem faced by all smoothing-based tests due to the obstacle that there are infinitely many alternatives. We prefer to use the suggested simple method instead of using other computation intensive and complex methodologies to find the optimal bandwidth.

Another direction which needs more attention is the model checking with missing covariates at random. We discuss this issue here briefly. Denote $X = (U, V)$ here U and V are p_1- and p_2- dimensional random vectors with $p_1 + p_2 = p$. Consider the situation that U is missing at random, whereas other variables Y and V are observed completely. Let δ be the missing indicator for the individual whether U is observed ($\delta = 1$) or not ($\delta = 0$). Assume that U is missing at random which implies

$$P(\delta = 1|Y, X) = P(\delta = 1|Y, V) = \pi(Z),$$

here $Z = (Y, V)$. Note that in this situation, $E(\delta\epsilon|X) = E[E(\delta\epsilon|X, Y)|X] = E(\epsilon\pi(Z)|X)$ may not be equal to zero. This further implies that

$$E(\delta\epsilon E(\delta\epsilon|X)W(X)) = E(E^2(\delta\epsilon|X)W(X))$$

may not be zero even under the null hypothesis. Thus we can not construct test statistics similarly as R_n^N and R_n^P. However, we note that $E(\delta/\pi(Z)\epsilon|X) = E[E(\delta/\pi(Z)\epsilon|X, Y)|X] = E(\epsilon|X) = 0$ under the null hypothesis. This suggests that test statistics could be constructed similarly as T_n^N and T_n^P. For the formal development of the related results, we leave them to further studies.

2.5 Appendix. Proofs of Theorems

The following conditions are required for the theorems in Section 2.2.

1) $g(x,\theta)$ is a Borel measurable function on R^p for each θ and a twice continuously differentiable real function on a compact subset of $\mathbb{R}^m$ Θ for each $x \in \mathbb{R}^p$; $\tilde{\theta}_0$, the value of θ that minimizes $\tilde{S}_{0n}(\theta) = E[(E(Y|X) - g(X,\theta))^2]$, is an interior point of Θ and is the unique minimizer of the function $\tilde{S}_{0n}$; $\Sigma_1 = E(g'(X,\theta_0)g'^{\top}(X,\theta_0))$ is nonsingular.

2) $\pi(x)$ has bounded partial derivatives up to order 2 almost surely and $\inf_x \pi(x) > 0$;

3) $\sup E(\varepsilon^4|X=x) < \infty$, $E|X|^4 < \infty$ and $E|Y|^4 < \infty$;

4) $nh^{3p/2} \to \infty$ and $h \to 0$;

5) The density of X, say $f(x)$ on support $\mathcal{C}$, exists and has bounded derivatives up to order 2 and satisfies

$$0 < \inf_{x\in\mathcal{C}} f(x) \leq \sup_{x\in\mathcal{C}} f(x) < \infty;$$

6) The continuous kernel function $K(\cdot)$ satisfies: i) the support of $K(\cdot)$ is the interval $[-1,1]$; ii) $K(\cdot)$ is symmetric about 0; iii) $\int_{-1}^{1} K(u)du = 1$ and $\int_{-1}^{1} |u|K(u)du \neq 0$.

Remark 2.2. *Conditions 4) and 6) are typical for obtaining convergence rates when non-parametric estimation is applied. Condition 2) is a common assumption in missing data study, for example, Sun and Wang (2009). The conditions 1) and 3) are necessary for the asymptotic normality of the least squares estimator. Condition 5) is aimed for avoiding tedious proofs of the theorems, see, e.g. Xue (2009). Without this condition, we have to resort to some truncation technique to control small values in the denominators.*

Lemma 2.1. *Under the null hypothesis and conditions above, we have*

$$W_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\delta_i \delta_j}{\pi(x_i)\pi(x_j)} K_h(x_i - x_j)\epsilon_i M(x_j) = O_p(1/\sqrt{n}). \tag{2.15}$$

where $M(\cdot)$ is continuously differentiable and $|M(x)| \leq b(x)$ for $x \in R^p$ and some $b(x)$ satisfying $E[b^2(X)] < \infty$.

This can be obtained following the same argument as Zheng (1996), so we omit the details.

Lemma 2.2. *Under the conditions in Appendix and the alternative H_{1n}, the asymptotic properties of $\sqrt{n}(\hat{\theta}_N - \theta_0)$ is as follows*

$$\begin{aligned} \sqrt{n}(\hat{\theta}_N - \theta_0) &= C_n \sqrt{n} \Sigma_1^{-1} E(g'(X, \theta_0) G(X)) \\ &\quad + \frac{\Sigma_1^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i g'(x_i, \tilde{\theta}_0)(y_i - g(x_i, \tilde{\theta}_0))}{\pi(x_i)} + o_p(1). \end{aligned} \tag{2.16}$$

where $\Sigma_1 = E(g'(X, \theta_0) g'^{\top}(X, \theta_0))$.

Lemma 2.3. *Under conditions in Appendix and the alternative H_{1n}, the asymptotic properties of $\sqrt{n}(\hat{\theta}_P - \theta_0)$ is*

$$\begin{aligned} \sqrt{n}(\hat{\theta}_P - \theta_0) &= C_n \sqrt{n} \Sigma_1^{-1} E(g'(X, \theta_0) G(X)) \\ &\quad + \frac{\Sigma_1^{-1}}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i g'(x_i, \tilde{\theta}_0)(y_i - g(x_i, \tilde{\theta}_0))}{\pi(x_i, \alpha)} + o_p(1). \end{aligned} \tag{2.17}$$

Lemma 2.2 and Lemma 2.3 can be similarly obtained from the Lemma 4.2 in Van Keilegom et al. (2008) and the Lemmas in Guo and Xu (2012) respectively, and we omit the detailed proof here.

The proof for R_n^N is similar to that for T_n^N, so we omit the detail for R_n^N in Theorems 1 and 2. Below we give the proof for T_n^N in Theorems 1 and 2.

Proof of Theorem 2.1. For T_n^N in (2.6), it can be decomposed as follows

$$\begin{aligned} T_n^N &= \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\hat{\epsilon}_i\hat{\epsilon}_j \\ &\quad -\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \frac{\delta_i\delta_j(\hat{\pi}(x_j)-\pi(x_j))}{\hat{\pi}(x_i)\hat{\pi}(x_j)\pi(x_j)}K_h(x_i-x_j)\hat{\epsilon}_i\hat{\epsilon}_j \\ &\quad -\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \frac{\delta_i\delta_j(\hat{\pi}(x_i)-\pi(x_i))}{\hat{\pi}(x_j)\hat{\pi}(x_i)\pi(x_i)}K_h(x_i-x_j)\hat{\epsilon}_i\hat{\epsilon}_j \\ &\quad -\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \frac{\delta_i\delta_j(\hat{\pi}(x_i)-\pi(x_i))(\hat{\pi}(x_j)-\pi(x_j))}{\hat{\pi}(x_j)\hat{\pi}(x_i)\pi(x_j)\pi(x_i)}K_h(x_i-x_j)\hat{\epsilon}_i\hat{\epsilon}_j \\ &=: T_{n1}-T_{n2}-T_{n3}-T_{n4}. \end{aligned} \tag{2.18}$$

Below we analyse the term T_{n1} in (2.18) first. It can be further divided as

$$\begin{aligned} T_{n1} &= \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\epsilon_i\epsilon_j \\ &\quad -\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\epsilon_i(g(x_j,\hat{\theta}_N)-g(x_j,\theta_0)) \\ &\quad -\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\epsilon_j(g(x_i,\hat{\theta}_N)-g(x_i,\theta_0)) \\ &\quad +\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n \Big(\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)(g(x_i,\hat{\theta}_N)-g(x_i,\theta_0)) \\ &\quad \times(g(x_j,\hat{\theta}_N)-g(x_j,\theta_0))\Big) \\ &=: T_{n1,1}-T_{n1,2}-T_{n1,3}+T_{n1,4}. \end{aligned} \tag{2.19}$$

Notice that $T_{n1,1}$ is a second order degenerate U-statistic. By some tedious calculations and according to Theorem 1 of Hall (1984), we can have

$$nh^{p/2}T_{n1,1}\to N(0,\Sigma^T). \tag{2.20}$$

where $\Sigma^T=\int K^2(u)du\cdot\int(\sigma^2(x))^2f^2(x)\pi^{-2}(x)dx$.

Below we prove that $nh^{p/2}T_{n1,2} = o_p(1)$. It can be divided as

$$\begin{aligned} T_{n1,2} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\epsilon_i\frac{\partial g(x_j,\theta_0)}{\partial\theta^\top}(\hat\theta_N-\theta_0) \\ &\quad +(\hat\theta_N-\theta_0)^\top\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\epsilon_i\frac{\partial^2 g(x_j,\tilde\theta)}{\partial\theta\partial\theta^\top}(\hat\theta_N-\theta_0) \\ &= R_{n1,1}(\hat\theta_N-\theta_0)+(\hat\theta_N-\theta)^\top R_{n1,2}(\hat\theta_N-\theta_0), \end{aligned}$$

where $\tilde\theta$ lies between $\hat\theta_N$ and θ_0.

Recalling Lemma 2.1, we have $R_{n1,1} = O_p(1/\sqrt{n})$. Let $\tilde A_{j,st}$ and $A_{j,st}$ denote the (s,t) element of $\partial^2 g(x_j,\tilde\theta)/\partial\theta\partial\theta^\top$ and $\partial^2 g(x_j,\theta_0)/\partial\theta\partial\theta^\top$ respectively. Due to the fact that $\hat\theta_N-\theta_0 = o_p(1)$ and the continuity of $\partial^2 g(x_j,\theta)/\partial\theta\partial\theta^\top$ as a function of θ, we can assert that

$$\begin{aligned} E\left|\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\epsilon_i\tilde A_{j,st}\right| &= E\left(\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)|\epsilon_i||A_{j,st}|\right)+o_p(1) \\ &= E\big(K_h(x_i-x_j)|A_{j,st}|E(|\epsilon_i||x_i)\big)=O(1). \end{aligned}$$

Thus we can have $R_{n1,2} = O_p(1)$. According to Lemma 2.2, we can have $\sqrt{n}(\hat\theta_N-\theta_0) = O_p(1)$. Then we can conclude that

$$T_{n1,2} = O_p(n^{-1/2})\cdot O_p(n^{-1/2}) + O_p(n^{-1/2})\cdot O_p(n^{-1/2}) = O_p(n^{-1}).$$

Thus

$$nh^{p/2}T_{n1,2} = O_p(h^{p/2}) = o_p(1). \tag{2.21}$$

Similarly as the derivation for $T_{n1,2}$, we have

$$nh^{p/2}T_{n1,3} = O_p(h^{p/2}) = o_p(1). \tag{2.22}$$

For $T_{n1,4}$ in (2.19), we have

$$\begin{aligned} T_{n1,4} &= (\hat\theta_N-\theta_0)^\top\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\Big(\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\frac{\partial g(x_i,\tilde\theta_1)}{\partial\theta} \\ &\quad \times\frac{\partial g(x_j,\tilde\theta_2)}{\partial\theta^\top}\Big)(\hat\theta_N-\theta_0) \\ &=: (\hat\theta_N-\theta_0)^\top R_{n1,3}(\hat\theta_N-\theta_0). \end{aligned}$$

Similar to the argument for $R_{n1,2}$, we can conclude that $R_{n1,3} = O(1)$. We have

$$T_{n1,4} = O_p(n^{-1/2}) \cdot O_p(1) \cdot O_p(n^{-1/2}) = O_p(n^{-1}).$$

Thus,

$$nh^{p/2}T_{n1,4} = O_p(h^{p/2}) = o_p(1). \tag{2.23}$$

Based on (2.19), (2.20), (2.21), (2.22) and (2.23), we have

$$nh^{p/2}T_{n1} = nh^{p/2}T_{n1,1} + o_p(1) \to N(0, \Sigma^T). \tag{2.24}$$

Following the argument for T_{n1}, $nh^{p/2}T_{n2} = o_p(1)$ can be proved by proving $nh^{p/2}T_{n2,1} = o_p(1)$, here

$$T_{n2,1} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\delta_i \delta_j (\hat{\pi}(x_j) - \pi(x_j))}{\hat{\pi}(x_i)\hat{\pi}(x_j)\pi(x_j)} K_h(x_i - x_j)\epsilon_i \epsilon_j.$$

Note that by computing the variance of $T_{n2,1}$, we can have

$$\begin{aligned} \mathrm{Var}(T_{n2,1}) &= \frac{1}{n^2(n-1)^2h^{2p}} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \sum_{k=1}^{n} \sum_{l \neq k}^{n} E\Big[\frac{\delta_i \delta_j \delta_k \delta_l (\hat{\pi}(x_j) - \pi(x_j))(\hat{\pi}(x_l) - \pi(x_l))}{\hat{\pi}(x_i)\hat{\pi}(x_j)\pi(x_j)\hat{\pi}(x_k)\hat{\pi}(x_l)\pi(x_l)} \\ &\quad \times K(\frac{x_i - x_j}{h}) K(\frac{x_k - x_l}{h}) \epsilon_i \epsilon_j \epsilon_k \epsilon_l \Big]. \end{aligned}$$

For the above summands, only the terms with $i = k, j = l$ and $i = l, j = k$ are non-zero. When $i = k, j = l$, we can have

$$\mathrm{Var}(T_{n2,1}) = \frac{1}{n^2(n-1)^2h^{2p}} \sum_{i=1}^{n} \sum_{j \neq i}^{n} E\Big[\frac{\delta_i \delta_j (\hat{\pi}(x_j) - \pi(x_j))^2}{\hat{\pi}^2(x_i)\hat{\pi}^2(x_j)\pi^2(x_j)} K^2(\frac{x_i - x_j}{h}) \epsilon_i^2 \epsilon_j^2 \Big].$$

Further note that

$$\begin{aligned} & E\Big[\frac{\delta_i \delta_j (\hat{\pi}(x_j) - \pi(x_j))^2}{\hat{\pi}^2(x_i)\hat{\pi}^2(x_j)\pi^2(x_j)} K^2(\frac{x_i - x_j}{h}) \epsilon_i^2 \epsilon_j^2 \Big] \\ = \; & E\Big[\frac{(\hat{\pi}(x_j) - \pi(x_j))^2}{\pi(x_i)\pi^3(x_j)} K^2(\frac{x_i - x_j}{h}) \sigma^2(x_i)\sigma^2(x_j) \Big] + o(1) \\ \leq \; & E\Big[\frac{\sigma^2(x_i)\sigma^2(x_j)}{\pi(x_i)\pi^3(x_j)} K^2(\frac{x_i - x_j}{h}) \times \sup_x (\hat{\pi}(x) - \pi(x))^2 \Big] \\ \leq \; & E^{1/2}\Big[\frac{\sigma^4(x_i)\sigma^4(x_j)}{\pi^2(x_i)\pi^6(x_j)} K^4(\frac{x_i - x_j}{h}) \Big] \times E^{1/2}[\sup_x (\hat{\pi}(x) - \pi(x))^4] \\ = \; & \left[h^m \int K^4(u)du \cdot \int (\sigma^4(x))^2 f^2(x) \pi^{-8}(x) dx \right]^{1/2} \times O\left(\sqrt{\frac{\ln(n)}{nh^m}}^2 \right) = o(h^p). \end{aligned}$$

For the last equation to hold, we need the condition that $nh^{3p/2} \to \infty$. Thus we can have $\mathrm{Var}(T_{n2,1}) = o(n^{-2}h^{-p})$. Similarly, for the term with $i = l, j = k$, we can also obtain that $\mathrm{Var}(T_{n2,1}) = o(n^{-2}h^{-p})$. Thus, $T_{n2,1} = o_p(n^{-1}h^{-p/2})$ which is also true for T_{n2}. Consequently,

$$nh^{p/2}T_{n2} = O_p(1) \cdot o_p(1) = o_p(1). \tag{2.25}$$

Similarly, we can obtain that

$$nh^{p/2}T_{n3} = o_p(1) \text{ and } nh^{p/2}T_{n4} = o_p(1). \tag{2.26}$$

Combining the equations (2.18), (2.24), (2.25) and (2.26) together, we have

$$nh^{p/2}T_n = nh^{p/2}T_{n1,1} + o_p(1) \to N(0, \Sigma^T). \tag{2.27}$$

Note that

$$\sup_{a \le x \le b} |\pi(x, \hat{\alpha}) - \pi(x, \alpha)| = O_p(\frac{1}{\sqrt{n}}) = o_p(1),$$

by a similar derivation as T_n^N, we have

$$nh^{p/2}T_n^P \to N(0, \Sigma^T). \tag{2.28}$$

Below we prove the consistency of $\hat{\Sigma}^{TN}$ based on U-statistics theory. Similarly as the derivation for T_n^N, we can verify

$$\hat{\Sigma}^{TN} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{1}{h^p} \frac{\delta_i \delta_j}{\pi^2(x_i)\pi^2(x_j)} K^2(\frac{x_i - x_j}{h}) \epsilon_i^2 \epsilon_j^2 + o_p(1).$$

Invoking the U-statistics theory, if

$$E\Big[\frac{1}{h^p} \frac{\delta_i \delta_j}{\pi^2(x_i)\pi^2(x_j)} K^2(\frac{x_i - x_j}{h}) \epsilon_i^2 \epsilon_j^2\Big]^2 = o(n), \tag{2.29}$$

we can have

$$\begin{aligned} \hat{\Sigma}^{TN} &= 2E(\frac{1}{h^p} \frac{\delta_i \delta_j}{\pi^2(x_i)\pi^2(x_j)} K^2(\frac{x_i - x_j}{h}) \epsilon_i^2 \epsilon_j^2) + o_p(1) \\ &= 2 \int K^2(u)du \cdot \int \frac{(\sigma^2(x))^2 f^2(x)}{\pi^2(x)} dx + o_p(1). \end{aligned}$$

In fact, the equation (2.29) can be proved based on the following fact

$$
\begin{aligned}
& E\Big[\frac{1}{h^{2p}}\frac{\delta_i\delta_j}{\pi^4(x_i)\pi^4(x_j)}K^4(\frac{x_i-x_j}{h})\epsilon_i^4\epsilon_j^4\Big] \\
= & \frac{1}{h^{2p}}\int\int\frac{K^4((x_1-x_2)/h)\tilde{\sigma}^4(x_1)\tilde{\sigma}^4(x_2)}{\pi^3(x_1)\pi^3(x_2)}f(x_1)f(x_2)dx_1dx_2 \\
= & \frac{1}{h^{p}}\int\int\frac{K^4(u)\tilde{\sigma}^4(x)\tilde{\sigma}^4(x-hu)}{\pi^3(x)\pi^3(x-hu)}f(x)f(x-hu)dxdu \\
= & \frac{1}{h^{p}}\int K^4(u)du\cdot\int\frac{(\tilde{\sigma}^4(x))^2f^2(x)}{\pi^6(x)}dx+o(1/h^p) \\
= & O(1/h^p)=o(n),
\end{aligned}
$$

here $\tilde{\sigma}^4(x)=E(\epsilon_1^4|x_1)$. The consistent of $\hat{\Sigma}^{TP}$ can be similarly derived when $\pi(x,\alpha)$ is a parameter function. We finish the proof for Theorem 2.1 $\square$.

Proof of Theorem 2.2. Denote

$$
\bar{T}_n^N=\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\hat{\epsilon}_i\hat{\epsilon}_j.
$$

Under the local alternatives (2.11) and recalling that $\epsilon_i=y_i-g(x_i,\theta_0)$, the term T_n^N can be decomposed as $T_n^N=\bar{T}_n^N+o_p(\bar{T}_n^N)$. As for $\bar{T}_n^N$, we have the expansion

$$
\begin{aligned}
\bar{T}_n^N = & \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\epsilon_i\epsilon_j \\
& -\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\epsilon_i(g(x_j,\hat{\theta}_N)-g(x_j,\theta_0)) \\
& -\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\epsilon_j(g(x_i,\hat{\theta}_N)-g(x_i,\theta_0)) \\
& +\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n\Big(\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)(g(x_i,\hat{\theta}_N)-g(x_i,\theta_0)) \\
& \times(g(x_j,\hat{\theta}_N)-g(x_j,\theta_0))\Big) \\
= & \bar{T}_{n1}-\bar{T}_{n2}-\bar{T}_{n3}+\bar{T}_{n4}+o_p(1). \qquad (2.30)
\end{aligned}
$$

For the term $\bar{T}_{n2}$ in (2.30), it follows that

$$\begin{aligned}\bar{T}_{n2} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\eta_i g'(x_j,\tilde{\theta})(\hat{\theta}_N-\theta_0)\\ &\quad +C_n\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)G(x_i)g'(x_j,\tilde{\theta})(\hat{\theta}_N-\theta_0)\\ &= \bar{T}_{n2,1}(\hat{\theta}_N-\theta_0)+C_n\bar{T}_{n2,2}(\hat{\theta}_N-\theta_0),\end{aligned}$$

where $\tilde{\theta}$ lies between $\hat{\theta}_N$ and θ_0.

Based on the conclusion from Lemma 2.1, we have $\bar{T}_{n2,1}=O_p(n^{-1/2})$. It can also be proved that

$$\begin{aligned}\bar{T}_{n2,2} &= E(G(X_1)g'(X_2,\theta_0)K_h(X_1-X_2))+o_p(1)\\ &= E(G(X)g'(X,\theta_0)f(X))+o_p(1).\end{aligned}$$

When $C_n=n^{-1/2}h^{-p/4}$, Lemma 2 implies that

$$\begin{aligned}nh^{p/2}\bar{T}_{n2} &= nh^{p/2}\Big[O_p(n^{-1/2})O_p(C_n)\\ &\quad +C_n^2E(G(X)g'(X,\theta_0)f(X))\Sigma_1^{-1}E(G(X)g'(X,\theta_0))\Big]\\ &= E(G(X)g'(X,\theta_0)f(X))\Sigma_1^{-1}E(G(X)g'(X,\theta_0))+o_p(1).\end{aligned} \tag{2.31}$$

For $\bar{T}_{n3}$ in (2.30), we can similarly derive that

$$nh^{p/2}\bar{T}_{n3}=E(G(X)g'(X,\theta_0)f(X))\Sigma_1^{-1}E(G(X)g'(X,\theta_0))+o_p(1). \tag{2.32}$$

Using a similar argument as that for proving Theorem 2.1 and Lemma 2.2, we have the expansion for $\bar{T}_{n4}$:

$$\begin{aligned}\bar{T}_{n4} &= (\hat{\theta}_N-\theta_0)^\top E[g'(X,\theta_0)g'^\top(X,\theta_0)f(X)](\hat{\theta}_N-\theta_0)+o_p(C_n^2)\\ &= C_n^2E^\top[G(X)g'(X,\theta_0)]\Sigma_1^{-1}E[g'(X,\theta_0)g'^\top(X,\theta_0)f(X)]\\ &\quad \times\Sigma_1^{-1}E[G(X)g'(X,\theta_0)]+o_p(C_n^2).\end{aligned}$$

As a result, when $C_n=n^{-1/2}h^{-p/4}$, we can obtain

$$\begin{aligned}nh^{p/2}\bar{T}_{n4} &= E^\top[G(X)g'(X,\theta_0)]\Sigma_1^{-1}E[g'(X,\theta_0)g'^\top(X,\theta_0)f(X)]\\ &\quad \times\Sigma_1^{-1}E[G(X)g'(X,\theta_0)]+o_p(1).\end{aligned} \tag{2.33}$$

Now we turn to investigate the term $\bar{T}_{n1}$ in (2.30), it can be decomposed as follows

$$\begin{aligned}
\bar{T}_{n1} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\eta_i\eta_j \\
&\quad +C_n\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\eta_i G(x_j) \\
&\quad +C_n\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)\eta_j G(x_i) \\
&\quad +C_n^2\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{\delta_i\delta_j}{\pi(x_i)\pi(x_j)}K_h(x_i-x_j)G(x_i)G(x_j) \\
&= \bar{T}_{n1,1}+C_n\bar{T}_{n1,2}+C_n\bar{T}_{n1,3}+C_n^2\bar{T}_{n1,4}.
\end{aligned}$$

From the proof for Theorem 2.1 and the conclusion of Lemma 2.1, we know that

$$\begin{aligned}
&nh^{p/2}\bar{T}_{n1,1}\to N(0,\Sigma^T); \\
&\bar{T}_{n1,2}=O_p(n^{-1/2}); \\
&\bar{T}_{n1,3}=O_p(n^{-1/2}); \\
&\bar{T}_{n1,4}=E(G^2(X)f(X))+o_p(1).
\end{aligned}$$

Consequently, when $C_n=n^{-1/2}h^{-p/4}$, we can obtain that:

$$nh^{p/2}T_n^N\to N(\mu^T,\Sigma^T) \tag{2.34}$$

where

$$\mu^T = E\left[\Big(G(X)-g'^{\top}(X,\theta_0)\Sigma_1^{-1}E[G(X)g'(X,\theta_0)]\Big)^2 f(X)\right].$$

Combining equations (2.31),(2.33) and (2.34), we can have

$$nh^{p/2}T_n^N\to N(\mu^T,\Sigma^T).$$

It can be similarly verified for $\pi(X,\alpha)$ be estimated as $\pi(X,\hat{\alpha})$. When C_n has a slower convergence rate than $n^{-1/2}h^{-p/4}$, the above proof can show that the test statistic goes to infinity in probability. We omit the details. Theorem 2.2 is proved. □

Table 2.1: Study 1: Empirical sizes and powers for H_0 vs $H_{1i}, i = 1, \cdots, 4$ with $X \sim N(0, \Sigma_1)$ and $\epsilon \sim N(0, 0.3^2)$.

	$n = 50$		$n = 100$		$n = 200$	
	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$
H_0 LI	0.020	0.027	0.029	0.029	0.033	0.034
SW^S	0.049	0.052	0.058	0.052	0.053	0.062
SW^E	0.045	0.046	0.054	0.047	0.052	0.047
GXZ_{TN}	0.029	0.036	0.043	0.041	0.042	0.042
GXZ_{TP}	0.031	0.028	0.036	0.034	0.046	0.041
GXZ_{RN}	0.035	0.041	0.045	0.039	0.050	0.042
GXZ_{RP}	0.031	0.036	0.041	0.044	0.042	0.041
H_{11} LI	0.102	0.079	0.278	0.176	0.633	0.513
SW^S	0.069	0.080	0.098	0.102	0.137	0.142
SW^E	0.183	0.136	0.454	0.333	0.857	0.725
GXZ_{TN}	0.151	0.096	0.351	0.228	0.745	0.508
GXZ_{TP}	0.159	0.114	0.311	0.313	0.811	0.632
GXZ_{RN}	0.139	0.115	0.396	0.291	0.786	0.673
GXZ_{RP}	0.185	0.119	0.459	0.323	0.842	0.714
H_{12} LI	0.993	0.941	1.000	1.000	1.000	1.000
SW^S	1.000	1.000	1.000	1.000	1.000	1.000
SW^E	1.000	1.000	1.000	1.000	1.000	1.000
GXZ_{TN}	0.989	0.937	1.000	1.000	1.000	1.000
GXZ_{TP}	0.988	0.939	1.000	1.000	1.000	1.000
GXZ_{RN}	0.992	0.943	1.000	1.000	1.000	1.000
GXZ_{TP}	0.997	0.950	1.000	1.000	1.000	1.000
H_{13} LI	0.315	0.203	0.351	0.271	0.375	0.338
SW^S	0.555	0.498	0.686	0.681	0.747	0.695
SW^E	0.465	0.453	0.635	0.639	0.701	0.724
GXZ_{TN}	0.811	0.717	0.911	0.878	0.929	0.931
GXZ_{TP}	0.786	0.673	0.870	0.844	0.891	0.903
GXZ_{RN}	0.799	0.661	0.861	0.817	0.884	0.853
GXZ_{TP}	0.774	0.704	0.869	0.848	0.905	0.891
H_{14} LI	0.241	0.159	0.671	0.497	0.983	0.921
SW^S	0.055	0.043	0.045	0.040	0.046	0.054
SW^E	0.091	0.068	0.166	0.129	0.467	0.357
GXZ_{TN}	0.208	0.132	0.588	0.386	0.950	0.828
GXZ_{TP}	0.252	0.154	0.600	0.439	0.957	0.874
GXZ_{RN}	0.241	0.153	0.634	0.501	0.968	0.894
GXZ_{TP}	0.261	0.189	0.653	0.472	0.975	0.893

Table 2.2: Study 1: Empirical sizes and powers for H_0 vs $H_{1i}, i = 1, \cdots, 4$ with $X \sim N(0, \Sigma_2)$ and $\epsilon \sim N(0, 0.3^2)$.

	$n = 50$		$n = 100$		$n = 200$	
	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$
H_0 LI	0.031	0.023	0.041	0.036	0.045	0.043
SW^S	0.047	0.048	0.038	0.047	0.055	0.051
SW^E	0.043	0.045	0.046	0.043	0.044	0.044
GXZ_{TN}	0.045	0.033	0.042	0.037	0.041	0.056
GXZ_{TP}	0.045	0.043	0.043	0.041	0.045	0.041
GXZ_{RN}	0.032	0.031	0.046	0.035	0.049	0.047
GXZ_{RP}	0.041	0.036	0.031	0.042	0.039	0.045
H_{11} LI	0.115	0.103	0.199	0.164	0.479	0.373
SW^S	0.627	0.559	0.919	0.917	0.999	0.999
SW^E	0.733	0.646	0.985	0.975	1.000	1.000
GXZ_{TN}	0.334	0.248	0.762	0.676	0.993	0.968
GXZ_{TP}	0.422	0.277	0.812	0.679	0.960	0.969
GXZ_{RN}	0.381	0.269	0.783	0.703	0.961	0.985
GXZ_{RP}	0.441	0.361	0.883	0.773	1.000	0.995
H_{12} LI	0.965	0.831	0.999	0.991	1.000	1.000
SW^S	0.693	0.581	0.902	0.846	0.989	0.981
SW^E	0.714	0.610	0.927	0.882	0.999	0.996
GXZ_{TN}	0.923	0.783	1.000	1.000	1.000	1.000
GXZ_{TP}	0.911	0.769	0.998	0.992	1.000	0.999
GXZ_{TN}	0.956	0.809	1.000	0.995	1.000	1.000
GXZ_{TP}	0.934	0.803	1.000	0.993	1.000	1.000
H_{13} LI	0.237	0.187	0.272	0.209	0.274	0.227
SW^S	0.495	0.502	0.628	0.634	0.683	0.682
SW^E	0.538	0.526	0.668	0.680	0.751	0.767
GXZ_{TN}	0.770	0.654	0.897	0.862	0.936	0.910
GXZ_{TP}	0.728	0.622	0.848	0.811	0.893	0.862
GXZ_{RN}	0.765	0.624	0.868	0.804	0.887	0.870
GXZ_{RP}	0.767	0.656	0.833	0.822	0.888	0.878
H_{14} LI	0.203	0.144	0.596	0.471	0.957	0.892
SW^S	0.491	0.448	0.806	0.776	0.987	0.981
SW^E	0.636	0.545	0.955	0.923	1.000	0.999
GXZ_{TN}	0.448	0.315	0.890	0.785	1.000	0.993
GXZ_{TP}	0.499	0.345	0.907	0.837	1.000	0.996
GXZ_{RN}	0.465	0.337	0.908	0.839	1.000	0.997
GXZ_{RP}	0.552	0.380	0.954	0.874	1.000	0.996

Table 2.3: Study 1: Empirical sizes and powers for H_0 vs $H_{1i}, i = 1, \cdots, 4$ with $X \sim N(0, \Sigma_1)$ and $\epsilon \sim DE(0, 3/10\sqrt{2})$.

	$n = 50$		$n = 100$		$n = 200$	
	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$
H_0 LI	0.026	0.030	0.041	0.035	0.049	0.047
SW^S	0.045	0.041	0.044	0.058	0.045	0.050
SW^E	0.038	0.032	0.043	0.047	0.054	0.046
GXZ_{TN}	0.032	0.029	0.039	0.025	0.043	0.041
GXZ_{TP}	0.031	0.036	0.029	0.034	0.038	0.034
GXZ_{RN}	0.032	0.034	0.038	0.039	0.051	0.042
GXZ_{RP}	0.031	0.034	0.038	0.044	0.036	0.036
H_{11} LI	0.083	0.076	0.222	0.227	0.682	0.549
SW^S	0.122	0.101	0.148	0.136	0.262	0.271
SW^E	0.432	0.329	0.861	0.775	0.998	0.993
GXZ_{TN}	0.745	0.544	0.991	0.951	1.000	1.000
GXZ_{TP}	0.746	0.575	0.972	0.952	1.000	1.000
GXZ_{RN}	0.752	0.562	0.996	0.954	1.000	1.000
GXZ_{RP}	0.770	0.598	0.991	0.972	1.000	1.000
H_{12} LI	0.986	0.931	1.000	1.000	1.000	1.000
SW^S	1.000	1.000	1.000	1.000	1.000	1.000
SW^E	1.000	1.000	1.000	1.000	1.000	1.000
GXZ_{TN}	1.000	0.986	1.000	1.000	1.000	1.000
GXZ_{TP}	1.000	0.988	1.000	1.000	1.000	1.000
GXZ_{RN}	1.000	0.994	1.000	1.000	1.000	1.000
GXZ_{RP}	1.000	0.993	1.000	1.000	1.000	1.000
H_{13} LI	0.281	0.208	0.335	0.274	0.417	0.329
SW^S	0.548	0.529	0.681	0.683	0.738	0.757
SW^E	0.484	0.459	0.641	0.636	0.708	0.734
GXZ_{TN}	0.814	0.683	0.896	0.875	0.942	0.922
GXZ_{TP}	0.752	0.728	0.864	0.854	0.907	0.892
GXZ_{RN}	0.777	0.712	0.852	0.814	0.984	0.880
GXZ_{RP}	0.780	0.703	0.863	0.847	0.897	0.901
H_{14} LI	0.192	0.176	0.688	0.524	0.979	0.905
SW^S	0.052	0.053	0.041	0.035	0.045	0.033
SW^E	0.131	0.115	0.418	0.292	0.927	0.802
GXZ_{TN}	0.741	0.581	0.994	0.956	1.000	1.000
GXZ_{TP}	0.748	0.624	0.992	0.959	1.000	1.000
GXZ_{RN}	0.776	0.602	0.994	0.969	1.000	1.000
GXZ_{RP}	0.754	0.577	0.986	0.968	1.000	1.000

Table 2.4: Study 1: Empirical sizes and powers for H_0 vs $H_{1i}, i = 1, \cdots, 4$ with $X \sim N(0, \Sigma_2)$ and $\epsilon \sim DE(0, 3/10\sqrt{2})$.

	$n = 50$		$n = 100$		$n = 200$	
	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$
H_0 LI	0.028	0.032	0.033	0.042	0.035	0.046
SW^S	0.043	0.039	0.041	0.045	0.042	0.047
SW^E	0.036	0.044	0.043	0.045	0.044	0.047
GXZ_{TN}	0.026	0.027	0.053	0.042	0.067	0.048
GXZ_{TP}	0.031	0.032	0.035	0.039	0.037	0.038
GXZ_{RN}	0.035	0.027	0.046	0.043	0.049	0.040
GXZ_{RP}	0.033	0.032	0.040	0.049	0.032	0.054
H_{11} LI	0.117	0.095	0.237	0.194	0.523	0.433
SW^S	0.762	0.704	0.981	0.968	1.000	1.000
SW^E	0.861	0.800	0.997	0.989	1.000	1.000
GXZ_{TN}	0.714	0.561	0.957	0.901	1.000	0.998
GXZ_{TP}	0.692	0.528	0.914	0.871	0.966	0.979
GXZ_{RN}	0.717	0.565	0.969	0.916	1.000	0.992
GXZ_{RP}	0.756	0.602	0.970	0.921	1.000	0.998
H_{12} LI	0.949	0.826	1.000	0.99	1.000	1.000
SW^S	0.776	0.686	0.941	0.905	1.000	0.991
SW^E	0.801	0.726	0.970	0.944	0.999	1.000
GXZ_{TN}	0.993	0.958	1.000	1.000	1.000	1.000
GXZ_{TP}	0.988	0.949	0.998	1.000	1.000	1.000
GXZ_{RN}	0.993	0.955	1.000	1.000	1.000	1.000
GXZ_{RP}	0.993	0.950	1.000	1.000	1.000	1.000
H_{13} LI	0.244	0.184	0.265	0.225	0.272	0.242
SW^S	0.488	0.497	0.581	0.621	0.672	0.718
SW^E	0.538	0.525	0.687	0.696	0.752	0.773
GXZ_{TN}	0.799	0.661	0.891	0.861	0.914	0.919
GXZ_{TP}	0.712	0.627	0.838	0.809	0.887	0.873
GXZ_{RN}	0.768	0.615	0.852	0.828	0.888	0.872
GXZ_{RP}	0.759	0.614	0.869	0.809	0.879	0.868
H_{14} LI	0.259	0.207	0.626	0.483	0.944	0.881
SW^S	0.613	0.589	0.932	0.908	1.000	1.000
SW^E	0.801	0.730	0.990	0.985	1.000	1.000
GXZ_{TN}	0.782	0.599	0.998	0.988	1.000	1.000
GXZ_{TP}	0.806	0.625	0.982	0.979	1.000	1.000
GXZ_{RN}	0.793	0.654	0.994	0.975	1.000	1.000
GXZ_{RP}	0.807	0.659	0.994	0.984	1.000	1.000

Table 2.5: Simulated size and power under different sample sizes $n = 25, 50$ and $n = 100$, missing mechanisms $\pi_1(x)$ and $\pi_2(x)$ for bootstrap calibration.

	$n = 25$		$n = 50$		$n = 100$	
	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$
H_0 GXZ^*_{TN}	0.057	0.049	0.047	0.048	0.047	0.049
GXZ^*_{TP}	0.046	0.059	0.056	0.051	0.056	0.051
GXZ^*_{RN}	0.059	0.050	0.044	0.052	0.048	0.050
GXZ^*_{RP}	0.044	0.058	0.053	0.055	0.055	0.048
H_{11} GXZ^*_{TN}	0.129	0.124	0.269	0.175	0.481	0.366
GXZ^*_{TP}	0.144	0.115	0.251	0.216	0.526	0.455
GXZ^*_{RN}	0.137	0.131	0.293	0.192	0.531	0.427
GXZ^*_{RP}	0.142	0.118	0.271	0.228	0.559	0.455
H_{12} GXZ^*_{TN}	0.844	0.706	0.999	0.978	1.000	1.000
GXZ^*_{TP}	0.844	0.681	0.999	0.983	1.000	1.000
GXZ^*_{RN}	0.844	0.709	0.999	0.983	1.000	1.000
GXZ^*_{RP}	0.854	0.708	0.999	0.984	1.000	1.000
H_{13} GXZ^*_{TN}	0.509	0.382	0.598	0.585	0.682	0.598
GXZ^*_{TP}	0.445	0.393	0.580	0.542	0.632	0.574
GXZ^*_{RN}	0.491	0.358	0.576	0.557	0.648	0.565
GXZ^*_{RP}	0.454	0.382	0.588	0.538	0.654	0.565
H_{14} GXZ^*_{TN}	0.195	0.118	0.357	0.262	0.673	0.513
GXZ^*_{TP}	0.168	0.154	0.372	0.261	0.739	0.604
GXZ^*_{RN}	0.188	0.119	0.381	0.294	0.741	0.604
GXZ^*_{RP}	0.163	0.156	0.375	0.267	0.764	0.628

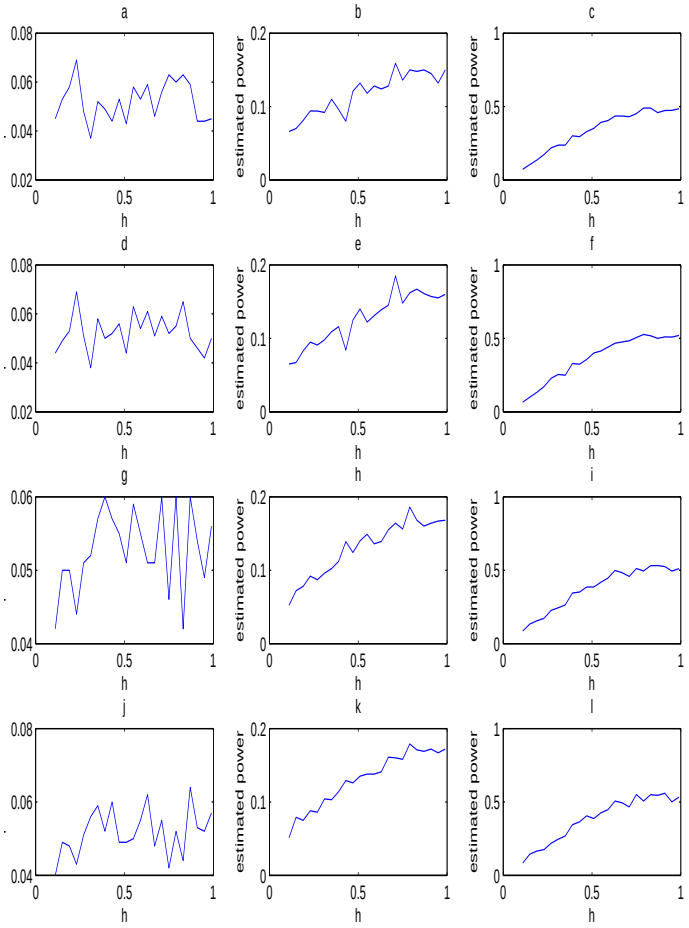

Figure 2.1: The estimated size and power curves of the tests GXZ^*_{TN}, GXZ^*_{RN}, GXZ^*_{TP}, GXZ^*_{RP} against the bandwidth h with missing mechanisms $\pi_1(x)$ and sample size 50 under different choices of C_n for testing problem (2.13). (a) $GXZ^*_{TN}, C_n = 0$; (b) $GXZ^*_{TN}, C_n = 2n^{-1/2}$; (c) $GXZ^*_{TN}, C_n = 4n^{-1/2}$. (d) $GXZ^*_{RN}, C_n = 0$; (e) $GXZ^*_{RN}, C_n = 2n^{-1/2}$; (f) $GXZ^*_{RN}, C_n = 4n^{-1/2}$. (g) $GXZ^*_{TP}, C_n = 0$; (h) $GXZ^*_{TP}, C_n = 2n^{-1/2}$; (i) $GXZ^*_{TP}, C_n = 4n^{-1/2}$. (j) $GXZ^*_{RP}, C_n = 0$; (k) $GXZ^*_{RP}, C_n = 2n^{-1/2}$; (l) $GXZ^*_{RP}, C_n = 4n^{-1/2}$.

Table 2.6: Simulated size and power under different sample sizes $n = 25, 50$ and $n = 100$, missing mechanisms $\pi_1(x)$, and different C_n for Study 2.

$n = 25$				
C_n	GXZ^*_{TP}	GXZ^*_{RP}	GXZ^*_{TN}	GXZ^*_{RN}
0.0	0.045	0.049	0.050	0.051
0.2	0.076	0.071	0.074	0.075
0.4	0.136	0.140	0.126	0.132
0.6	0.232	0.231	0.205	0.215
0.8	0.314	0.321	0.305	0.322
1.0	0.386	0.394	0.396	0.410
$n = 50$				
C_n	GXZ^*_{TP}	GXZ^*_{RP}	GXZ^*_{TN}	GXZ^*_{RN}
0.0	0.055	0.056	0.048	0.047
0.2	0.089	0.093	0.086	0.096
0.4	0.267	0.274	0.233	0.248
0.6	0.520	0.536	0.470	0.515
0.8	0.683	0.693	0.659	0.701
1.0	0.844	0.855	0.814	0.833
$n = 100$				
C_n	GXZ^*_{TP}	GXZ^*_{RP}	GXZ^*_{TN}	GXZ^*_{RN}
0.0	0.056	0.055	0.052	0.057
0.2	0.149	0.157	0.126	0.138
0.4	0.537	0.543	0.481	0.541
0.6	0.893	0.903	0.852	0.879
0.8	0.983	0.987	0.972	0.978
1.0	0.994	0.997	0.997	0.997

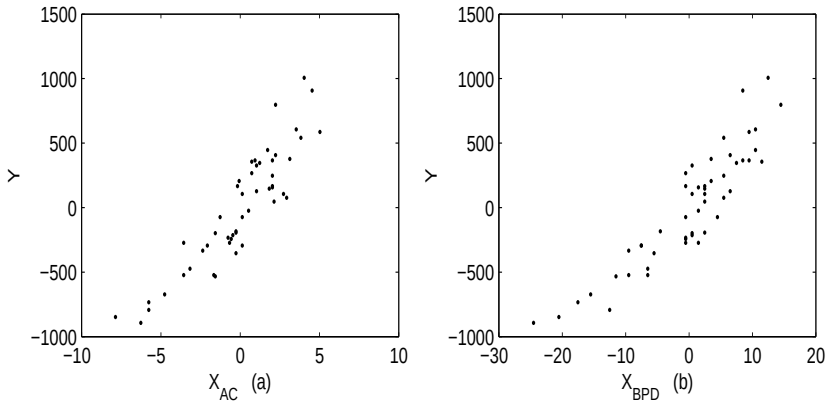

Figure 2.2: The plot for real data set: (a) for X_{AC} and Y; (b) for X_{BPD} and Y.

Chapter 3

Model Checking for General Linear Regression with Nonignorable Missing Response

3.1 Introduction

Due to its easy interpretation and well developed theories, the linear regression is wildly used to describe the relationship between scalar response Y and covariates X of dimension p. Consider the following general linear regression model with the form

$$Y = g^{\top}(X)\beta + \epsilon, \tag{3.1}$$

here $g(\cdot)$ is a known smooth function of dimension m, and β is an unknown parameter to be estimated. Further, ϵ is the error term satisfying $E(\epsilon|X) = 0$ and $E(\epsilon^2|X) = \sigma^2(X) < \infty$. The superscript $\top$ in (3.1) denotes the transpose. When $g(X) \equiv X$, model (3.1) becomes the classical linear model. Compared with the classical linear model, the model (3.1) is more flexible and applicable because interaction and high order terms of the covariates can be included in this model.

On the other hand, missing response is often encountered in practice. For instance, the response Y's may be very expensive to measure and due to limited budget, only

the responses for a part of subjects are available. The sampled individuals may refuse to supply the desired information to some survey questions, or investigators fail to gather correct information. Simply excluding the units with missing response and carrying out statistical analysis based only on the completely observed samples can often lead to biased and inefficient parameter estimates when the data are not missing completely at random (see Little and Rubin 1987). To develop more accurate and useful methodology in the presence of missing response, we often need to make some assumptions on the missing mechanism.

Define δ_i as the missing indicator for the ith individual whether Y_i is observed ($\delta_i = 1$) or not ($\delta_i = 0$). If the response Y is independent of the missing indicator δ, given the covariates X, then the response is missing at random(MAR) or ignorable. On the other hand, if whether the response is missing or not depends on the value of the response, we call this kind missing response is non-ignorable or not missing at random(NMAR). The NMAR is common in social survey, especially when the questions are sensitive. Consider the survey of personal income, low socioeconomic groups are more likely to refuse to provide the desired information.

To prevent wrong conclusions and improve the interpretations, the statistical analysis within the model (3.1), should be accompanied by a check of whether the hypothetical parametric model is satisfied at all. When the response variable Y is missing at random, there are some literatures investigated the model checking. Among others, Manteiga and González (2006) extended Härdle and Mammen's (1993) method to test the goodness of fit of a linear regression model with missing response data, and builded test statistics based on the L_2 distance between the nonparametric and parametric fits. For the general linear model (3.1) with missing response at random, Sun and Wang (2009) imputed the incomplete observations by imputation and inverse probability weighting methods and then proceed to construct two score type tests and two empirical process tests with the completed samples. Recently, Li (2012) proposed a test that is based on minimum integrated square distances between the

nonparametric and parametric fits.

However, these above mentioned works are limited to missing response at random. It's unclear and difficult to extend these works to the non-ignorable missing response situation because the missingness is also related to the unobserved responses. Recently, Kim and Yu (2011) proposed an exponential tilting model to handle the non-ignorable missing response data. They studied the estimation of mean functions when the tilting parameter is estimated as well as known. For the interval confidence construction of the mean functions, Zhao et al. (2013) applied the empirical likelihood method introduced by Owen(1988) to account for the non-ignorable missing response.

In this Chapter, we aim to obtain a model checking procedure for the model (3.1) with non-ignorable missing response and completely observed covariates, that is, to check the null hypothesis

$$H_0 : E(Y|X) = g^{\top}(X)\beta, \tag{3.2}$$

for some β and known $g(\cdot)$ with nonignorable missing response.

We first discuss how to estimate the unknown parameters in the null hypothesis. We propose three estimators for the unknown β. The first two estimators are based on imputed completed data sets and the third one is based on inverse probability weight method. We then construct three residual based empirical process test statistics. Based on the exponential tilting model, we study the asymptotic properties of our proposed tests under different situations in details.

The rest of this Chapter is organized as follows. In Section 3.2, we construct the test statistics. In Section 3.3, we derive their asymptotic properties under null hypothesis and local alternative hypothesis in three different situations. In Section 3.4, some simulation are reported to illustrate the proposed tests. The proofs of the asymptotic results are presented in the Appendix 3.5.

3.2 Construction of Test Statistics

Denote $P(\delta = 1|Y, X) = \pi(X, Y)$ as the selection probability function which is the probability given X and Y, a unit is observed. In this Chapter, we assume that the response is non-ignorable missing, thus $\pi(X, Y)$ is generally a function of X and Y. If $\pi(X, Y)$ only depends on X, then the response is missing at random.

Intuitively, we may ignore the missing information and construct statistics based on the empirical version of the expectation $E\Big(\delta(Y - g^\top(X)\beta)|X\Big)$. However note that

$$E\Big(\delta(Y - g^\top(X)\beta)|X\Big) \;=\; E\big(\pi(X, Y)(Y - g^\top(X)\beta)|X\big).$$

Under H_0, the expression above may not be zero. As a result, it may not be reasonable to construct statistic based on $E\Big(\delta(Y - g^\top(X)\beta)|X\Big)$. If the response is missing at random(MAR), that's, $\pi(X, Y) = P(\delta = 1|Y, X) = \pi(X)$, the expression above will be zero. Thus, in the MAR setting, the above expression can be used to construct test statistics. However, in the NMAR setting, this idea fails.

In the following, we first impute the missing response and construct two completed data sets as follows:

$$\begin{aligned}
\tilde{y}_{i1} &= \delta_i y_i + (1 - \delta_i)\hat{m}_0(x_i), \\
\tilde{y}_{i2} &= \frac{\delta_i}{\hat{\pi}(x_i, y_i)} y_i + (1 - \frac{\delta_i}{\hat{\pi}(x_i, y_i)})\hat{m}_0(x_i), i = 1, \cdots, n,
\end{aligned}$$

here $m_0(x_i) = E(Y|x_i, \delta = 0)$, $\hat{m}_0(x_i)$ and $\hat{\pi}(x_i, y_i)$ are the estimators of $m_0(x_i)$ and $\pi(x_i, y_i)$ respectively which will be specified later.

Then the residual-based empirical process test statistics can be constructed as follows:

$$R_{nk}(x) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Big(\tilde{y}_{ik} - g^\top(x_i)\hat{\beta}_k\Big) I(x_i \le x), k = 1, 2, \tag{3.3}$$

where $\hat{\beta}_k$ are the estimators of β defined as

$$\hat{\beta}_k = \Big(\sum_{i=1}^{n} g(x_i)g^\top(x_i)\Big)^{-1} \sum_{i=1}^{n} g(x_i)\tilde{y}_{ik}, k = 1, 2.$$

Then the test statistics can be defined by

$$T_{nk} = \int (R_{nk}(x))^2 dF_n(x), k = 1, 2, \tag{3.4}$$

where $F_n(x)$ is the empirical distribution based on $x_1, x_2, \cdots, x_n$.

Inspired by the inverse probability weight method, we consider the third approach. Note that under H_0, the following equation holds,

$$E\Big(\frac{\delta}{\pi(X,Y)}(Y - g^\top(X)\beta)|X\Big) = E\Big(Y - g^\top(X)\beta|X\Big) \equiv 0.$$

It is worth mentioning that our idea for constructing tests is to weight the observed residuals by inverse probability function. The residual-based empirical process test statistic is constructed as follows

$$R_{n3}(x) \;=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}(x_i, y_i)} \Big(y_i - g^\top(x_i)\hat{\beta}_3\Big) I(x_i \le x), \tag{3.5}$$

where $\hat{\beta}_3$ is the inverse probability weighted estimator of β with the following form

$$\hat{\beta}_3 = \Big(\sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}(x_i, y_i)} g(x_i) g^\top(x_i)\Big)^{-1} \sum_{i=1}^{n} \frac{\delta_i}{\hat{\pi}(x_i, y_i)} g(x_i) y_i.$$

Then the test statistics can be defined by

$$T_{n3} = \int (R_{n3}(x))^2 dF_n(x). \tag{3.6}$$

Different from the MAR setting, it's difficult to estimate the $m_0(x_i)$ and $\pi(x_i, y_i)$. To deal with this issue, we adopt the methodology developed in Kim and Yu (2011).

Assume that the $\pi(x_i, y_i)$ follows a semi-parametric model

$$\pi(x_i, y_i) = \frac{\exp(\phi(x_i) - \gamma y_i)}{1 + \exp(\phi(x_i) - \gamma y_i)}. \tag{3.7}$$

Here $\phi(\cdot)$ is a generally unknown smooth function and γ is called the tilting parameter which controls the degree of nonignorable missing mechanism. When $\gamma \equiv 0$, we will turn back to the MAR situation. Define $O(x_i, y_i) = \Pr(\delta_i = 0|x_i, y_i)/\Pr(\delta_i = 1|x_i, y_i) = \pi^{-1}(x_i, y_i) - 1$ as the conditional odds of non-response. Under the model

(3.7), we can rewrite the odd function as $O(x_i, y_i) = \exp(-\phi(x_i) + \gamma y_i)$. Define $\alpha(X;\gamma) := \exp(-\phi(X)) = O(X,Y)/\exp(\gamma Y)$, we then have

$$\alpha(X;\gamma)E[\delta \exp(\gamma Y)|X] = E[\delta O(X,Y)|X] = E(1-\delta|X).$$

Thus under the semiparametric logistic regression model (3.7), a nonparametric estimator of $\pi(x_i, y_i)$ can be obtained by $\hat{\pi}(x_i, y_i;\gamma) = \hat{\pi}(x_i, y_i)$, where

$$\hat{\pi}(x_i, y_i) = \{1 + \hat{\alpha}(x_i;\gamma)\exp(\gamma y_i)\}^{-1},$$

and

$$\hat{\alpha}(x_i;\gamma) = \frac{\sum_{j=1}^{n}(1-\delta_j)K_h(x_i, x_j)}{\sum_{j=1}^{n}\delta_j \exp(\gamma y_j)K_h(x_i, x_j)},$$

here $K_h(u,x) = K\{(u-x)/h\}/h^p$ with $K(\cdot)$ being a kernel function and h being a bandwidth. Note that for $m_0(X)$, we can have

$$m_0(X) = \frac{E[(1-\delta)Y|X]}{E(1-\delta|X)} = \frac{E[\delta O(X,Y)Y|X]}{E(\delta O(X,Y)|X)} = \frac{E[\delta \exp(\gamma Y)Y|X]}{E(\delta \exp(\gamma Y)|X)}.$$

Thus we can estimate $m_0(x_i)$ by

$$\hat{m}_0(x_i) := \hat{m}_0(x_i;\gamma) = \sum_{j=1}^{n}\omega_{i0}(x_i;\gamma)y_j, \tag{3.8}$$

where the weight

$$\omega_{i0}(x_i;\gamma) = \frac{\delta_j \exp(\gamma y_j)K_h(x_i, x_j)}{\sum_{j=1}^{n}\delta_j \exp(\gamma y_j)K_h(x_i, x_j)}. \tag{3.9}$$

We should note that $E\Big(\delta g(X)(Y - g^\top(X)\beta)\Big) = E\Big(\pi(X,Y)g(X)(Y - g^\top(X)\beta)\Big)$ may not be zero, thus the complete case estimator of β in the following form

$$\hat{\beta}_{CC} = \Big(\sum_{i=1}^{n}\delta_i g(x_i)g^\top(x_i)\Big)^{-1}\sum_{i=1}^{n}\delta_i g(x_i)y_i$$

can be biased and can not be used in general.

We denote γ^* as the true value of γ in model (3.7). When γ^* is unknown, it can be estimated from either an independent survey or a validation sample, which is a subsample of the non-respondents. Let $\hat{\gamma}$ be the corresponding estimator of γ^*. Then semi-parametric estimators of $m_0(x_i)$ and $\pi(x_i, y_i)$ can be given by $\hat{m}_0(x_i;\hat{\gamma})$ and $\hat{\pi}(x_i, y_i;\hat{\gamma})$ respectively.

3.3 Asymptotic Behavior of the Test Statistics

In this section, we investigate the asymptotic behaviors of our proposed test statistics under the null hypothesis as well as some local alternatives. We first discuss the case with known γ^* in advance.

3.3.1 Asymptotic Properties with Known γ^*

To state the theorems, we introduce some notations. Denote $Z = (X, Y)$, $\Sigma = E(g(X)g^\top(X))$, $L(X;x) = I(X \le x) - E\big(g^\top(X)I(X \le x)\big)\Sigma^{-1}g(X)$ and

$$J(\delta_i, x_i, y_i; x) \;=\; L(x_i;x)\frac{\delta_i\varepsilon_i + (\pi(z_i) - \delta_i)E(\varepsilon|x_i, \delta = 0)}{\pi(z_i)}.$$

We have the following theorem for the asymptotic behavior of our proposed tests under the null hypothesis (3.2).

Theorem 3.1. *Under H_0 and the conditions in Appendix, we have that*

$$R_{nk}(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} J(\delta_i, x_i, y_i; x) + o_p(1), \quad k = 1, 2, 3$$

will converge in distribution to $R(x)$ in the Skorokhod space $D[-\infty, +\infty]$, where $R(x)$ is a centered continuous Gaussian process with the covariance function,

$$Cov(R(x_1), R(x_2)) = E(J(\delta, X, Y; x_1)J(\delta, X, Y; x_2))$$

for any x_1 and x_2. Hence, T_{nk} converge in distribution to $T := \int R^2(x)dF(x)$ for $k = 1, 2, 3$, where $F(\cdot)$ is the distribution function of X.

From the above Theorem, we can know that the three proposed tests have the same asymptotic properties under the null hypothesis. Interestingly, this phenomenon continues to hold under the following local alternatives:

$$H_{1n} : Y = g^\top(X)\beta + C_nG(X) + \eta,$$

where $E(\eta|X) = 0$ and the function $G(\cdot)$ satisfies $E(G^2(X)) < \infty$. Actually, we should note that the asymptotic behaviors of the three proposed estimators of β also

are the same under the null as well as the above local alternative hypothesis. Denote $S(t) = E(G(X)L(X;x))$, we have the following theorem,

Theorem 3.2. *Under H_{1n} and conditions in Appendix, if $C_n\sqrt{n} \to 1$, R_{nk} converge in distribution to $R(t)+S(t)$ for $k = 1, 2, 3$, where $S(t)$ is a non-random shift function and T_{nk} converge in distribution to $\int (R(t) + S(t))^2 dF(t)$. If $n^r C_n \to a, 0 < r < 1/2$, then T_{nk} converge to ∞.*

From the expression of $S(t)$, we realize that it cannot be null unless $G(X) = C_0 g^\top(X)\theta$ with C_0 being any constant and θ being any parametric vector. Further, the larger $S(t)$ is, T_{nk} can yard the more powers. From Theorem 2, we can also know that: (1). when the local alternatives are distinct from the null hypothesis at the rate n^{-r} with $0 < r < 1/2$, the asymptotic powers of the proposed tests are all 1 in asymptotic sense; (2). when the alternatives converge to the null hypothesis at the rate $n^{-1/2}$, the proposed test can also detect them.

3.3.2 Asymptotic Properties with Estimated $\hat{\gamma}$ from Independent Survey

In many cases, γ^* is unknown and has to be estimated. In general, we can get the estimator of γ^* from an independent survey or a validation sample (Kim and Yu, 2011). In this subsection, we first consider the case with estimated $\hat{\gamma}$ from an independent survey with sample size n. We assume that $\sqrt{n}(\hat{\gamma} - \gamma^*) \Rightarrow N(0, V_\gamma)$. Denote $H(x) = E\Big((1-\delta)(Y-m_0(X))^2 L(X;x)\Big)$, we then have the following Theorem:

Theorem 3.3. *Under H_0 and the conditions in Appendix, we have that*

$$R_{nk}(x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} J(\delta_i, x_i, y_i; x) + H(x)\sqrt{n}(\hat{\gamma} - \gamma^*) + o_p(1), \quad k = 1, 2, 3$$

will converge in distribution to $R^*(x)$ in the Skorokhod space $D[-\infty, +\infty]$, where $R^*(x)$ is a centered continuous with the covariance function,

$$Cov(R^*(x_1), R^*(x_2)) = E(J(\delta, X, Y; x_1)J(\delta, X, Y; x_2)) + H(x_1)H(x_2)V_\gamma$$

for any x_1 and x_2. Hence, T_{nk} converge in distribution to $T^* := \int (R^*(x))^2 dF(x)$ for $k = 1, 2, 3$, where $F(\cdot)$ is the distribution function of X.

When $\hat{\gamma}$ is exactly estimated, that's, $V_\gamma = 0$, the above Theorem is just Theorem 1. Compared with Theorem 1, we can conclude that estimating unknown parameter γ can induce larger variance and thus reduce the power performance of our proposed tests. Below we turn to study the sensitivity of our proposed tests under local alternative H_{1n}. We obtain the following results:

Theorem 3.4. *Under H_{1n} and conditions in Appendix, if $C_n\sqrt{n} \to 1$, R_{nk} converges in distribution to $R^*(t) + S(t)$ for $k = 1, 2, 3$ and T_{nk} converges in distribution to $\int (R^*(t) + S(t))^2 dF(t)$. If $n^r C_n \to a, 0 < r < 1/2$, then T_{nk} converges to ∞.*

From this Theorem, we can know clearly that the shift function is the same as that obtained with known γ^*. The effects of estimating parameter γ is still to induce larger variance.

3.3.3 Asymptotic Properties with Estimated $\hat{\gamma}$ from Validation Sample

We now turn to the case when a validation sample is randomly selected from the set of nonrespondents and the responses are obtained for all the units in the validation sample. We also call the validation sample as follow-up sample since it's generally obtained after the first stage of sample. We obtain the estimator $\hat{\gamma}$ of γ^* by solving the following equation

$$\sum_{i=1}^{n} (1 - \delta_i) r_i \{y_i - \hat{m}_0(x_i; \gamma)\} = 0.$$

here r_i is an indicator function, which takes 1 if unit i belongs to the follow-up sample and takes 0 otherwise.

Denote

$$M = E\Big[(1-\delta)r\big(E(Y^2|X,\delta=0) - m_0^2(X,\gamma^*)\big)\Big]$$
$$\Delta_i = \frac{1}{M}(\eta_i - E(\eta|x_i,\delta=0))\Big[(1-\delta_i)r_i - \delta_i\nu(\frac{1}{\pi(z_i)}-1)\Big]$$
$$\tilde{J}(\delta_i, r_i, x_i, y_i; x) = J(\delta_i, x_i, y_i; x) + H(x)\Delta_i.$$

here $\nu = E(r|\delta=0)$.

Theorem 3.5. *Under H_0 and the conditions in Appendix, we have that*

$$R_{nk}(x) \;=\; \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\tilde{J}(\delta_i; r_i; x_i; y_i; x) + o_p(1), \quad k = 1,2,3$$

will converge in distribution to $\tilde{R}(x)$ in the Skorokhod space $D[-\infty,+\infty]$, where $\tilde{R}(x)$ is a centered continuous Gaussian process with the covariance function,

$$Cov(\tilde{R}(x_1), \tilde{R}(x_2)) = E(\tilde{J}(\delta, r, X, Y; x_1)\tilde{J}(\delta, r, X, Y; x_2))$$

for any x_1 and x_2. Hence, T_{nk} converges in distribution to $\tilde{T} := \int \tilde{R}^2(x)dF(x)$ for $k = 1,2,3$, where $F(\cdot)$ is the distribution function of X.

For the power performances of our proposed tests under local alternative hypothesis H_{1n}, we can have the following Theorem:

Theorem 3.6. *Under H_{1n} and conditions in Appendix, if $C_n\sqrt{n} \to 1$, R_{nk} converges in distribution to $\tilde{R}(t) + S(t)$ for $k = 1,2,3$, T_{nk} converge in distribution to $\int(\tilde{R}(t) + S(t))^2 dF(t), k = 1,2,3$. If $n^r C_n \to a, 0 < r < 1/2$, then $T_{ni}, i = 1,2,3$ converge to ∞.*

3.3.4 Monte Carlo Approximation

From Theorems 3.1, 3.3 and 3.5, we can obtain the asymptotic covariances of $R_{nk}(x)$ under different situations. However, it seems that the variances of the test statistic T_{nk} are very complicated for practical use. In below, to determine the critical value, we adopt the Nonparametric Monte Carlo approach, developed by Zhu (2005), and Zhu

and Neuhaus (2000) to approximate the asymptotic distribution of the test statistics under null hypothesis. This procedure has some desirable features, for instance, the test procedure is self-scale invariant, and we can determine the p-values without any additional standardization. We illustrate this methodology for T_{n1} under the situation that γ^* is estimated from a validation sample.

The Monte Carlo test procedure for determining p-values is as follows.

Step 1. Generate random variables $e_i (i = 1, 2, \cdots, n)$ independently with mean zero and variance one. Let $E_n := (e_1, \cdots, e_n)$ and define the conditional counterpart of R_n as

$$\tilde{R}_{n1}(E_n, x) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} e_i \hat{\tilde{J}}(\delta_i; r_i; x_i; y_i; x),$$

where $\hat{\tilde{J}}(\delta_i; r_i; x_i; y_i; x)$ is the estimator of $\tilde{J}(\delta_i; r_i; x_i; y_i; x)$, which is defined as

$$\hat{\tilde{J}}(\delta_i; r_i; x_i; y_i; x) \quad = \quad \hat{J}(\delta_i, x_i, y_i; x) + \hat{H}(x)\hat{\Delta}_i.$$

First, we get the estimator $\hat{J}(\delta_i, x_i, y_i; x)$ as follows:

$$J(\delta_i, x_i, y_i; x) = \hat{L}(x_i; x) \frac{\delta_i \hat{\varepsilon}_i + (\hat{\pi}(z_i, \hat{\gamma}) - \delta_i)\hat{E}(\varepsilon | x_i, \delta = 0)}{\hat{\pi}(z_i, \hat{\gamma})},$$

with

$$\hat{L}(x_i; x) = I(x_i \le x) - \sum_{i=1}^{n} g^\top(x_i) I(x_i \le x) \left(\sum_{i=1}^{n} g(x_i) g^\top(x_i) \right)^{-1} g(x_i);$$

$$\hat{\varepsilon}_i = y_i - g^\top(x_i)\hat{\beta}_1; \quad \hat{E}(\varepsilon | x_i, \delta = 0) = \sum_{j=1}^{n} \omega_{i0}(x_i, \hat{\gamma})\hat{\varepsilon}_j.$$

Here $\hat{\gamma}$ is an estimator of γ^* by solving this equation $\sum_{i=1}^{n}(1 - \delta_i) r_i \{y_i - \hat{m}_0(x_i; \gamma)\} = 0$ and $\hat{\pi}(z_i, \hat{\gamma}), \hat{m}_0(x_i; \gamma)$ and $\omega_{i0}(x_i, \gamma)$ has been defined in section 3.2.

For the estimator of Δ_i, we can have:

$$\hat{\Delta}_i = \frac{1}{\hat{M}} (\hat{\varepsilon} - \hat{E}(\varepsilon | x_i, \delta = 0)) \Big[(1 - \delta_i) r_i - \delta_i \nu \big(\frac{1}{\hat{\pi}(z_i, \hat{\gamma})} - 1 \big) \Big],$$

with

$$\hat{M} = \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)r_i(\hat{E}(Y^2|x_i,\delta=0) - \hat{m}_0^2(x_i,\hat{\gamma}));$$
$$\hat{E}(Y^2|x_i,\delta=0) = \sum_{j=1}^{n}\omega_{i0}(x_i,\hat{\gamma})y_j^2.$$

Further note that

$$\begin{aligned} H(x) &= E\Big((1-\delta)(Y-m_0(X))^2L(X;x)\Big) \\ &= E\Big(E\big((1-\delta)(Y-m_0(X))^2L(X;x)|X\big)\Big) \\ &= E\Big[E\big((Y-m_0(X))^2L(X;x)|X,\delta=0\big)P(\delta=0|X)\Big]. \end{aligned}$$

Thus we can estimate $H(x)$ by the following equations:

$$\hat{H}(x) = \frac{1}{n}\sum_{i=1}^{n}\hat{E}\big((Y-m_0(X))^2L(X;x)|x_i,\delta=0\big)\hat{P}(\delta=0|x_i),$$

with

$$\hat{E}\big((Y-m_0(X))^2L(X;x)|x_i,\delta=0\big) = \sum_{j=1}^{n}\omega_{i0}(x_i,\hat{\gamma})(y_j-\hat{m}_0(x_j,\hat{\gamma}))^2\hat{L}(x_j;x);$$
$$\hat{P}(\delta=0|x_i) = 1 - \frac{\sum_{j=1}^{n}\delta_jK_h(x_i,x_j)}{\sum_{j=1}^{n}K_h(x_i,x_j)}.$$

The resultant conditional test statistic is

$$\tilde{T}_{n1}(E_n) = \int \tilde{R}_{n1}(E_n,x)^2dF_n(x).$$

Step 2. Generate t sets of E_n, say $E_n^{(i)}, i=1,\cdots,t$ and get t values of $\tilde{T}_{n1}(E_n)$, say $\tilde{T}_{n1}(E_n^{(i)}), i=1,\cdots,t$.

Step 3. The p -value is estimated by $\hat{p}_k = n_k/(t+1)$, where n_k is the number of $\tilde{T}_{n1}(E_n^{(i)})$ which is larger than or equal to T_{n1}. Reject H_0 when $\hat{p}_k \le \alpha$ for a designed level α.

For T_{n2} and T_{n3}, similar procedures can be carried out. We omit the details here. For the above algorithm, we expect that the determination of the critical values is not

affected no matter the null or the alternative hypothesis holds. Theorem 3.7 below shows that the conditional distribution based on the Monte Carlo approximation will always converge to the limit distribution under null hypothesis.

Theorem 3.7. *Either under null hypothesis or local alternatives with* $C_n = O(n^{-1/2})$ *and the conditions in Theorem 3.5, we have that for almost all sequences*

$$\{(y_1, \delta_1, r_1, x_1); \cdots ; (y_n, \delta_n, r_n, x_n), \cdots\}$$

the conditional distribution of $\tilde{T}_{n1}(E_n)$ *converges to the limit null distribution of* T_{n1}.

3.4 Simulation Study

In this section, we present the performance of our proposed test statistics through some simulation study.

Study 1. We generate the data from the model

$$Y = g^\top(X)\beta + aG(X) + \varepsilon, \tag{3.10}$$

where $g(X) = 1 + X^2$ with $X \sim U(0, 1)$, $\beta = 1$, $G(X) = X^3$ and $\varepsilon \sim N(0, 0.25)$. For model (3.10), the null hypothesis $H_0 : E(Y|X) = g^\top(X)\beta$ for some β is corresponding to $a = 0$. For this model with univariate covariate, we assume that Y is missing not at random. The follow-up rate is taken to be 20%.

Two missing probability mechanisms are considered for model (3.10)

Case 1. $\pi_1(x, y) = 1/(1 + \exp(-(1 + 0.3x + 0.3y))$;

Case 2. $\pi_2(x, y) = 1/(1 + \exp(-(y + 0.3y^2)))$.

For the above two cases, the mean response rates are $E\pi_1(x, y) \approx 0.82$ and $E\pi_2(x, y) \approx 0.86$ respectively. Case 1 follows our assumed model 3.7, while case 2 is used to evaluate the robustness of our proposed test statistics against some mis-specifications of the assumed missing mechanism.

The kernel function is taken to be the Gaussian kernel function $K(u) = \exp(-u^2/2)/(2\pi)^{1/2}$. In practice, we standardize our observations and a common bandwidth is used. To investigate the impact of bandwidth selection on our proposed tests, we take the bandwidth h to be $n^{-1/5}(0.25 + i/4)$ for $i = 0, \cdots, 8$. Each result is based on 1000 times running. For each sample, Monte Carlo approximation is based on 500 sets of reference data generated by Monte Carlo method. Let $\alpha = 0.05$. First, we plot the estimated power curve against the above bandwidth sequences with sample size 100, missing mechanisms $\pi_i(x, y), i = 1, 2$ and $a = 0, 0.5, 1$, which is shown in Fig 3.1 and Fig 3.2. From these two figures, we can conclude that the bandwidths have little effects on our proposed tests. According to Kim and Yu (2011)'s suggestion, we use $n^{-1/5}$ in the following studies. Further, from these two figures, we can observe that the sizes and powers of $T_{ni}, i = 1, 2, 3$ are almost the same. This is consistent with our theories developed in subsection 3.3. To save space, in the following, we only report the results for T_{n1}.

The performance of our proposed tests under the above model (3.10) are investigated through varying the values of a, different sample size $n = 100$ and $n = 200$ and different missing mechanism $\pi_i(x, y)(i = 1, 2)$. The results are presented in Table 3.1. From this table, we can observe that our proposed tests can control the size well and are very powerful to the alternatives. Further, it's in accordance with intuition that when the sample size is larger, our proposed tests can have larger powers. Moreover, from this table, we can also conclude that our proposed tests are robust against failure of the assumed missing mechanism model (3.7).

To study the performance of our test statistics against high frequency alternatives, we carry out the following simulation study.

Study 2. We generate the data from the model

$$Y = g^\top(X)\beta + aG(X) + \varepsilon, \tag{3.11}$$

We make the same settings as study 1 except the alternatives. In this study, we set $G(X) = \sin(2\pi X)$ to be a high frequency alternative. The results are presented in

Figure 3.3. From this figure, similar findings are obtained. We should note that in this study, the power performances of T_{n1} under $\pi_2(x, y)$ are comparable to those under $\pi_1(x, y)$. Again, this example can show that our proposed tests are robust against failure of the assumed missing mechanism model (3.7).

Finally, we give a further numerical example to show the performances of our proposed tests. The following simulations are conducted.

Study 3. We generate the data from the model

$$Y = 1 + 0.7X + aG(X) + \varepsilon, \tag{3.12}$$

here $G(X) = 0.5(X - 2.5)^2 - 0.7X$, $X \sim N(2, 1)$ and $\varepsilon \sim N(0, 1)$. For model (3.12), under the null hypothesis, that's, $a = 0$, it becomes the model A in Kim and Yu (2011), while when $a = 1$, it's the model B in that paper. The follow-up rate is still taken to be 20%.

Four missing probability mechanisms are considered for model (3.12)

Case 3. $\pi_3(x, y) = 1/(1 + \exp(-(-1.5 + x))$;

Case 4. $\pi_4(x, y) = 1/(1 + \exp(-(-0.85 + 0.3x + 0.3y))$;

Case 5. $\pi_5(x, y) = 1/(1 + \exp(-(-2 + 0.3x + 0.3x^2 + 0.3y))$;

Case 6. $\pi_6(x, y) = 1/(1 + \exp(-(-0.65 + 0.1x + 0.1y + 0.1y^2))$.

These four missing mechanisms are just the M1, M2, M3 and M5 in Kim and Yu (2011) respectively. The results are presented in Figure 3.4. From this figure, we can conclude again that our proposed test statistics can detect the alternatives efficiently.

3.5 Appendix. Proofs of Theorems

The following conditions are required for the theorems in Section 3.3.

1) $\pi(X, Y)$ has bounded partial derivatives up to order 2 , $\pi(X, Y) \geq c_0 > 0$ and $p(x) = E(\delta|x) \neq 1$ almost surely.

2) $\sup E(\varepsilon^4|X=x)<\infty$, $E|X|^4<\infty$ and $E|Y|^4<\infty$;

3) $nh^p\to\infty$ and $h\to 0$;

4) The density of X, say $f(x)$ on support $\mathcal{C}$, exists and has bounded derivatives up to order 2 and satisfies

$$0<\inf_{x\in\mathcal{C}} f(x)\le\sup_{x\in\mathcal{C}} f(x)<\infty;$$

5) $K(\cdot)$ is a spherical symmetric density function with a bounded derivative and support. All the moments of $K(\cdot)$ exist.

Remark 3.1. *Conditions 3), and 4) are typical for obtaining convergence rates when non-parametric estimation is applied. Condition 1) is a common assumption in missing data study, for example, Sun and Wang(2009) and Kim and Yu (2011). The conditions 2) are necessary for the asymptotic normality of the least squares estimator. Condition 5) is aimed for avoiding tedious proofs of the theorems, see, e.g. Xue(2009). Without this condition, we have to resort to some truncation technique to control small values in the denominators.*

Firstly, we give a lemma which is used for the proof of theorems in the situation that γ^* is known.

Lemma 3.1. *Under conditions 1-5 in the Appendix and the alternative H_{1n}, if the true value of γ, γ^* is known, the asymptotic properties of $\sqrt{n}(\hat{\beta}_k-\beta)$ is as follows*

$$\begin{aligned}&\sqrt{n}(\hat{\beta}_k-\beta)\\ =\;&\Big(\Big\{\frac{1}{n}\sum_{i=1}^n g(x_i)g(x_i)^\top\Big\}^{-1}-\Sigma^{-1}\Big)C_n\sqrt{n}E(g(X)G(X))+\Sigma^{-1}C_n\sqrt{n}E(g(X)G(X))\\ &+\frac{\Sigma^{-1}}{\sqrt{n}}\sum_{i=1}^n g(x_i)\frac{\delta_i\eta_i+(\pi(z_i)-\delta_i)E(\eta|x_i,\delta=0)}{\pi(z_i)}+o_p(1),k=1,2,3.\end{aligned}$$

here $\Sigma=E(g(X)g(X)^\top)$.

Proof of Lemma 3.1. Denote $Z = (X, Y)$ and note that

$$\begin{aligned}\sqrt{n}(\hat{\beta}_1 - \beta) &= \Big\{\frac{1}{n}\sum_{i=1}^{n} g(x_i)g(x_i)^\top\Big\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i)\left(\delta_i y_i + (1-\delta_i)\hat{m}_0(x_i) - g(x_i)^\top\beta\right)\\ &= A_1^{-1}A_2,\end{aligned}$$

we consider the properties of A_1 and A_2 below.

Similar to Kim and Yu(2011), it can be verified that

$$\begin{aligned}A_2 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i)\big[\delta_i y_i + (1-\delta_i)m_0(x_i) - g(x_i)^\top\beta + \delta_i O(z_i)(y_i - m_0(x_i))\big] + o_p(1)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i)\left[\frac{\delta_i}{\pi(z_i)}y_i + \Big(1 - \frac{\delta_i}{\pi(z_i)}\Big)m_0(x_i) - g(x_i)^\top\beta\right] + o_p(1).\end{aligned}$$

Under the local alternative $y_i = g(x_i)^\top\beta + C_n G(x_i) + \eta_i$, thus $m_0(x_i) = E(Y|x_i, \delta = 0) = g(x_i)^\top\beta + C_n G(x_i) + E(\eta|x_i, \delta = 0)$. As a result, we can have

$$A_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i)\left[\frac{\delta_i}{\pi(z_i)}\eta_i + \Big(1 - \frac{\delta_i}{\pi(z_i)}\Big)E(\eta|x_i, \delta = 0)\right] + C_n\sqrt{n}E(g(X)G(X)) + o_p(1).$$

Consequently,

$$\begin{aligned}&\sqrt{n}(\hat{\beta}_1 - \beta)\\ =\ &\Big(\Big\{\frac{1}{n}\sum_{i=1}^{n} g(x_i)g(x_i)^\top\Big\}^{-1} - \Sigma^{-1}\Big)C_n\sqrt{n}E(g(X)G(X)) + \Sigma^{-1}C_n\sqrt{n}E(g(X)G(X))\\ &+\frac{\Sigma^{-1}}{\sqrt{n}}\sum_{i=1}^{n} g(x_i)\frac{\delta_i\eta_i + (\pi(z_i) - \delta_i)E(\eta|x_i, \delta = 0)}{\pi(z_i)} + o_p(1). \qquad (3.13)\end{aligned}$$

Now we turn to consider the second estimator $\hat{\beta}_2$. Denote

$$A_3 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i)\Big(\delta_i y_i/\hat{\pi}(z_i) + (1 - \delta_i/\hat{\pi}(z_i))\hat{m}_0(x_i) - g(x_i)^\top\beta\Big).$$

Then it can be decomposed as follows:

$$\begin{aligned}A_3 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i)\left[\frac{\delta_i}{\pi(z_i)}y_i + \Big(1 - \frac{\delta_i}{\pi(z_i)}\Big)m_0(x_i) - g(x_i)^\top\beta\right]\\ &\quad+\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i)\left(\frac{\delta_i}{\hat{\pi}(z_i)} - \frac{\delta_i}{\pi(z_i)}\right)(y_i - m_0(x_i))\\ &\quad+\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i)\left(1 - \frac{\delta_i}{\hat{\pi}(z_i)}\right)(\hat{m}_0(x_i) - m_0(x_i))\\ &=: A_{31} + A_{32} + A_{33}.\end{aligned}$$

Now we turn to consider the term A_{32}. Note that

$$\hat{\pi}(z_i) = \{1 + \hat{\alpha}(x_i; \gamma^*) \exp(\gamma^* y_i)\}^{-1}; \qquad \pi(z_i) = \{1 + \alpha(x_i; \gamma^*) \exp(\gamma^* y_i)\}^{-1};$$

and

$$\hat{\alpha}(x_i; \gamma^*) = \frac{\sum_{j=1}^n (1 - \delta_j) K_h(x_i, x_j)}{\sum_{j=1}^n \delta_j \exp(\gamma^* y_j) K_h(x_i, x_j)}.$$

From Kim and Yu(2011), we can know that $n^{-1} \sum_{j=1}^n \delta_j \exp(\gamma^* y_j) K_h(x_i, x_j) = f(x_i)\{1 - p(x_i)\}\alpha^{-1}(x_i; \gamma^*) + o_p(1)$, here $p(x) = E(\delta|x)$. Thus, we can derive

$$\begin{aligned}
A_{32} &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \delta_i g(x_i)(y_i - m_0(x_i)) \exp(\gamma^* y_i) \\
&\quad \times \frac{\sum_{j=1}^n [(1 - \delta_j) - \delta_j \exp(\gamma^* y_j) \alpha(x_i; \gamma^*)] K_h(x_i, x_j)}{\sum_{j=1}^n \delta_j \exp(\gamma^* y_j) K_h(x_i, x_j)} + o_p(1) \\
&= \frac{1}{n^{3/2}} \sum_{i=1}^n \delta_i g(x_i)(y_i - m_0(x_i)) \exp(\gamma^* y_i) \\
&\quad \times \frac{\sum_{j=1}^n [(1 - \delta_j) - \delta_j \exp(\gamma^* y_j) \alpha(x_i; \gamma^*)] K_h(x_i, x_j)}{f(x_i)\{1 - p(x_i)\}\alpha^{-1}(x_i; \gamma^*)} + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{j=1}^n [(1 - \delta_j) - \delta_j O(z_j)] E\left(\frac{\delta g(X)(Y - m_0(X)) O(Z) K_h(X, x_j)}{f(X)\{1 - p(X)\}} | x_j \right) + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{j=1}^n [(1 - \delta_j) - \delta_j O(z_j)] g(x_j) \frac{E\big((1 - \delta)(Y - m_0(X))|x_j\big)}{E(1 - \delta|x_j)} + o_p(1) = o_p(1).
\end{aligned}$$

The last equation follows from the fact that $E((1-\delta)(Y - m_0(X))|x_j)/E(1-\delta|x_j) = E(Y - m_0(X)|x_j, \delta = 0) \equiv 0$. Similarly, we can prove $A_{33} = o_p(1)$. As a result, we can have

$$A_3 = \frac{1}{\sqrt{n}} \sum_{i=1}^n g(x_i) \left[\frac{\delta_i}{\pi(z_i)} y_i + \frac{\pi(z_i) - \delta_i}{\pi(z_i)} m_0(x_i) - g(x_i)^\top \beta \right] + o_p(1).$$

Recall that $\sqrt{n}(\hat{\beta}_2 - \beta) = A_1^{-1} A_3$, we finish the proof for $\hat{\beta}_2$ in Lemma 3.1.

Lastly, we investigate the estimator $\hat{\beta}_3$ in the following. Note that

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_3 - \beta) &= \left\{ \frac{1}{n} \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}(z_i)} g(x_i) g(x_i)^\top \right\}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\delta_i}{\hat{\pi}(z_i)} g(x_i)(y_i - g(x_i)^\top \beta) \\
&= \tilde{A}_1^{-1} A_4,
\end{aligned}$$

For A_4, it can be verified that

$$\begin{aligned} A_4 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{\pi(z_i)}g(x_i)(\eta_i + C_nG(x_i)) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\delta_i}{\hat{\pi}(z_i)} - \frac{\delta_i}{\pi(z_i)}\right)g(x_i)\eta_i \\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\delta_i}{\hat{\pi}(z_i)} - \frac{\delta_i}{\pi(z_i)}\right)g(x_i)C_nG(x_i) \\ &= C_1 + C_2 + C_3. \end{aligned} \tag{3.14}$$

For the term C_1, it is evident that

$$C_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{\pi(z_i)}g(x_i)\eta_i + C_n\sqrt{n}E(g(X)G(X)). \tag{3.15}$$

Recall the definitions of $\pi(z_i), \hat{\pi}(z_i)$ and $\hat{\alpha}(x_i;\gamma^*)$, similar to the deviation for the term A_{32}, we can derive

$$\begin{aligned} C_2 &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}[(1-\delta_j) - \delta_j O(z_j)]E\left(\frac{\delta g(X)\eta O(Z)}{f(X)\{1-p(X)\}}K_h(X,x_j)|x_j\right) + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}[1 - \frac{\delta_j}{\pi(z_j)}]g(x_j)E(\eta|x_j,\delta=0) + o_p(1). \end{aligned} \tag{3.16}$$

The last equation follows from the fact that $O(Z) = \pi^{-1}(Z) - 1$.

Similarly, we can derive that

$$C_3 = \frac{C_n}{\sqrt{n}}\sum_{j=1}^{n}\Big(1 - \frac{\delta_j}{\pi(z_j)}\Big)g(x_j)G(x_j) + o_p(1).$$

Based on the fact that

$$E\Big(\big(1 - \frac{\delta}{\pi(Z)}\big)g(X)G(X)\Big) \equiv 0,$$

we conclude that

$$C_3 = o_p(1). \tag{3.17}$$

Based on the equations (3.14), (3.15), (3.16) and (3.17), we have

$$A_4 = C_n\sqrt{n}E(g(X)G(X)) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}g(x_i)\frac{\delta_i\eta_i + (\pi(z_i)-\delta_i)E(\eta|x_i,\delta=0)}{\pi(z_i)} + o_p(1).$$

Now we consider the properties of $\tilde{A}_1$ below. For $\tilde{A}_1$, we have

$$\begin{aligned}\tilde{A}_1 &= \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi(z_i)}g(x_i)g(x_i)^\top + \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\delta_i}{\hat{\pi}(z_i)} - \frac{\delta_i}{\pi(z_i)}\right)g(x_i)g(x_i)^\top \\ &= \Sigma + \frac{1}{n}\sum_{j=1}^{n}[(1-\delta_j) - \delta_j O(z_j)]g(x_j)g(x_j)^\top + o_p(1) \\ &= \Sigma + o_p(1).\end{aligned}$$

Recall that $\sqrt{n}(\hat{\beta}_3 - \beta) = \tilde{A}_1^{-1}A_4$, we finish the proof for $\hat{\beta}_3$ in Lemma 3.1. □

Proof of Theorem 3.1. For $R_{n1}(x)$, it can be verified that

$$\begin{aligned}R_{n1}(x) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\delta_i y_i + (1-\delta_i)\hat{m}_0(x_i) - g(x_i)^\top\beta\right)I(x_i \le x) \\ &\quad -E(g(X)^\top I(X \le x))\sqrt{n}(\hat{\beta}_1 - \beta).\end{aligned}$$

Under the null hypothesis $H_0 : y_i = g(x_i)^\top\beta + \varepsilon_i$, we can have $m_0(x_i) = g(x_i)^\top\beta + E(\varepsilon|x_i, \delta = 0)$. Based on the argument and the conclusion for $\hat{\beta}_1$ in Lemma 3.1, it follows that

$$R_{n1}(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}L(x_i; x)\left(\frac{\delta_i\varepsilon_i + (\pi(z_i) - \delta_i)E(\varepsilon|x_i, \delta = 0)}{\pi(z_i)}\right) + o_p(1).$$

here $L(X; x) = I(X \le x) - E(g(X)^\top I(X \le x))\Sigma^{-1}g(X)$. Thus the asymptotic properties of T_{n1} follows by the continuous mapping theorem. From the argument for A_3 in Lemma 3.1, we can know that

$$R_{n2}(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\delta_i}{\pi(z_i)}y_i + (1 - \frac{\delta_i}{\pi(z_i)})m_0(x_i) - g(x_i)^\top\hat{\beta}_2\right)I(x_i \le x) + o_p(1),$$

and thus $R_{n2}(x)$ has the same asymptotic expansion as that for $R_{n1}(x)$.

Lastly, we consider the term $R_{n3}(x)$. It can be verified that

$$\begin{aligned} R_{n3}(x) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{\pi(z_i)}(y_i - g(x_i)^\top\beta)I(x_i \le x) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{\pi(z_i)}g(x_i)^\top I(x_i \le x)(\hat{\beta}_3 - \beta) \\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\delta_i}{\hat{\pi}(z_i)} - \frac{\delta_i}{\pi(z_i)}\right)(y_i - g(x_i)^\top\beta)I(x_i \le x) \\ &\quad - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\delta_i}{\hat{\pi}(z_i)} - \frac{\delta_i}{\pi(z_i)}\right)g(x_i)^\top I(x_i \le x)(\hat{\beta}_3 - \beta) + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i\varepsilon_i + (\pi(z_i) - \delta_i)E(\varepsilon|x_i, \delta = 0)}{\pi(z_i)}I(x_i \le x) \\ &\quad - E(g(X)^\top I(X \le x))\sqrt{n}(\hat{\beta}_3 - \beta) + o_p(1). \end{aligned}$$

By the asymptotic result for $\hat{\beta}_3$, we thus have

$$R_{n3}(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}L(x_i; x)\left(\frac{\delta_i\varepsilon_i + (\pi(z_i) - \delta_i)E(\varepsilon|x_i, \delta = 0)}{\pi(z_i)}\right) + o_p(1).$$

Thus the asymptotic properties of T_{n3} follows by the continuous mapping theorem. □

Proof of Theorem 3.2. Under the alternative H_{1n}, for the test $R_{n1}(x)$, we have

$$\begin{aligned} R_{n1}(x) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\delta_i}{\pi(z_i)}y_i + (1 - \frac{\delta_i}{\pi(z_i)})m_0(x_i) - g(x_i)^\top\beta\right)I(x_i \le x) \\ &\quad - E(g(X)^\top I(X \le x))\sqrt{n}(\hat{\beta}_1 - \beta). \end{aligned}$$

Under the local alternative $H_{1n} : y_i = g(x_i)^\top\beta + C_nG(x_i) + \eta_i$, we can have $m_0(x_i) = g(x_i)^\top\beta + C_nG(x_i) + E(\eta_i|x_i, \delta = 0)$. Consequently, we can have

$$\begin{aligned} R_{n1}(x) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\delta_i}{\pi(z_i)}\eta_i + (1 - \frac{\delta_i}{\pi(z_i)})E(\eta_i|x_i, \delta = 0)\right)I(x_i \le x) \\ &\quad + C_n\sqrt{n}E(G(X)I(X \le x)) - E(g(X)^\top I(X \le x))\sqrt{n}(\hat{\beta}_1 - \beta). \end{aligned}$$

If $n^{1/2}C_n \to 1$, by Lemma 3.1, we can get

$$R_{n1}(x) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}L(x_i; x)\left(\frac{\delta_i\eta_i + (\pi(z_i) - \delta_i)E(\eta|x_i, \delta = 0)}{\pi(z_i)}\right) + E(G(X)L(X; x)) + o_p(1).$$

If $n^rC_n \to a, 0 < r < 1/2$, then it yields $\sqrt{n}C_n \to \infty$, as $n \to \infty$. As a result, we have $R_{n1}(x) \to \infty$. The same conclusions for $R_{n2}(x)$ and $R_{n3}(x)$ can be obtained easily by using Lemma 3.1 and similar argument for $R_{n1}(x)$. □

Now we turn to consider the case that γ^* is estimated with $\hat{\gamma}$ from an independent survey. We first establish the following lemma which states the asymptotic properties of $\hat{\beta}_k, k = 1, 2, 3$ in this situation.

Lemma 3.2. *Under conditions 1-5 in the Appendix and the alternative H_{1n}, if $\hat{\gamma}$ is computed form an independent survey, the asymptotic properties of $\sqrt{n}(\hat{\beta}_k - \beta)$ is as follows*

$$\begin{aligned}
& \sqrt{n}(\hat{\beta}_k - \beta) \\
= \; & \Big(\Big\{\frac{1}{n}\sum_{i=1}^{n} g(x_i)g(x_i)^\top\Big\}^{-1} - \Sigma^{-1}\Big)C_n\sqrt{n}E(g(X)G(X)) + \Sigma^{-1}C_n\sqrt{n}E(g(X)G(X)) \\
& + \frac{\Sigma^{-1}}{\sqrt{n}}\sum_{i=1}^{n} g(x_i)\frac{\delta_i\eta_i + (\pi(z_i) - \delta_i)E(\eta|x_i, \delta = 0)}{\pi(z_i)} \\
& + \Sigma^{-1}E\Big[g(X)(1-\delta)(\eta - E(\eta|X, \delta = 0))^2\Big]\sqrt{n}(\hat{\gamma} - \gamma^*) + o_p(1), k = 1, 2, 3.
\end{aligned}$$

Proof of Lemma 3.2: Note that

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_1 - \beta) &= \Big\{\frac{1}{n}\sum_{i=1}^{n} g(x_i)g(x_i)^\top\Big\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i)(\delta_i y_i + (1-\delta_i)\hat{m}_0(x_i, \hat{\gamma}) - g(x_i)^\top\beta) \\
&= A_1^{-1}A_2^*,
\end{aligned}$$

For A_2^*, it can be verified that

$$\begin{aligned}
A_2^* &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(x_i)\Big[\big(\delta_i y_i + (1-\delta_i)\hat{m}_0(x_i, \gamma^*) - g(x_i)^\top\beta\big) \\
&\quad + (1-\delta_i)\big(\hat{m}_0(x_i, \hat{\gamma}) - \hat{m}_0(x_i, \gamma^*)\big)\Big] \\
&= A_2 + B_2^*. \qquad (3.18)
\end{aligned}$$

For B_2^*, we can have

$$B_2^* = \frac{1}{n}\sum_{i=1}^{n} g(x_i)(1-\delta_i)\frac{\partial\hat{m}_0(x_i, \gamma_1)}{\partial\gamma}\sqrt{n}(\hat{\gamma} - \gamma^*),$$

here γ_1 is in the line segment between $\hat{\gamma}$ and γ^*. Note that

$$\frac{\partial\hat{m}_0(x_i, \gamma)}{\partial\gamma} = \frac{\sum_{j=1}^{n}\delta_j\exp(\gamma y_j)y_j^2K_h(x_i, x_j)}{\sum_{j=1}^{n}\delta_j\exp(\gamma y_j)K_h(x_i, x_j)} - \hat{m}_0^2(x_i; \gamma) = \hat{E}(Y^2|x_i, \delta = 0) - \hat{m}_0^2(x_i; \gamma).$$

Thus we can have

$$\begin{aligned} B_2^* &= E\Big[g(X)(1-\delta)\big(E(Y^2|X,\delta=0)-m_0^2(X)\big)\Big]\sqrt{n}(\hat{\gamma}-\gamma^*)+o_p(1) \\ &= E\Big[g(X)(1-\delta)(Y-m_0(X))^2\Big]\sqrt{n}(\hat{\gamma}-\gamma^*)+o_p(1) \\ &= E\Big[g(X)(1-\delta)(\eta-E(\eta|X,\delta=0))^2\Big]\sqrt{n}(\hat{\gamma}-\gamma^*)+o_p(1). \end{aligned} \tag{3.19}$$

Based on the equations (3.18), (3.19) and Lemma 3.1, we can get the result for $\hat{\beta}_1$ in Lemma 3.2.

Now we turn to consider the second estimator $\hat{\beta}_2$. Denote

$$\begin{aligned} A_3^* &= \frac{1}{\sqrt{n}}\sum_{i=1}^n g(x_i)\Big(\delta_i y_i/\hat{\pi}(z_i,\hat{\gamma})+(1-\delta_i/\hat{\pi}(z_i,\hat{\gamma}))\hat{m}_0(x_i,\hat{\gamma})-g(x_i)^\top\beta\Big) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n g(x_i)A_3^*(z_i,\hat{\gamma}). \end{aligned}$$

Then it can be decomposed as follows:

$$\begin{aligned} A_3^* &= \frac{1}{\sqrt{n}}\sum_{i=1}^n g(x_i)(A_3^*(z_i,\gamma^*)+(A_3^*(z_i,\hat{\gamma})-A_3^*(z_i,\gamma^*)) \\ &=: A_3+\frac{1}{n}\sum_{i=1}^n g(x_i)\frac{\partial A_3^*(z_i,\gamma_1)}{\partial\gamma}\sqrt{n}(\hat{\gamma}-\gamma^*). \end{aligned} \tag{3.20}$$

Note that

$$\begin{aligned} \frac{\partial\hat{\pi}^{-1}(z_i,\gamma)}{\partial\gamma} &= \hat{\alpha}(x_i;\gamma)\exp(\gamma y_i)(y_i-\hat{m}_0(x_i;\gamma)); \\ \frac{\partial\hat{m}_0(x_i,\gamma)}{\partial\gamma} &= \hat{E}(Y^2|x_i,\delta=0)-\hat{m}_0^2(x_i;\gamma). \end{aligned}$$

Thus, we can derive

$$\begin{aligned} \frac{\partial A_3^*(z_i,\gamma)}{\partial\gamma} &= \delta_i(y_i-\hat{m}_0(x_i;\gamma))\frac{\partial\hat{\pi}^{-1}(x_i,\gamma)}{\partial\gamma}+\left(1-\frac{\delta_i}{\hat{\pi}(z_i;\gamma)}\right)\frac{\partial\hat{m}_0(x_i,\gamma)}{\partial\gamma} \\ &= \delta_i(y_i-\hat{m}_0(x_i;\gamma))^2\hat{\alpha}(x_i;\gamma)\exp(\gamma y_i)+\left(1-\frac{\delta_i}{\hat{\pi}(z_i;\gamma)}\right) \\ &\quad\times\big(\hat{E}(Y^2|x_i,\delta=0)-\hat{m}_0^2(x_i;\gamma)\big)+o_p(1). \end{aligned}$$

As a result, we can have

$$\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n} g(x_i)\frac{\partial A_3^*(z_i,\gamma_1)}{\partial\gamma}\sqrt{n}(\hat{\gamma}-\gamma^*) \\
= \; & \frac{1}{n}\sum_{i=1}^{n} g(x_i)\Big[\delta_i(y_i-m_0(x_i))^2 O(z_i)+\left(1-\frac{\delta_i}{\pi(z_i)}\right) \\
& \times\big(E(Y^2|x_i,\delta=0)-m_0^2(x_i)\big)\Big]\sqrt{n}(\hat{\gamma}-\gamma^*)+o_p(1) \\
= \; & E\Big[g(X)(1-\delta)(Y-m_0(X))^2\Big]\sqrt{n}(\hat{\gamma}-\gamma^*)+o_p(1) \\
= \; & E\Big[g(X)(1-\delta)(\eta-E(\eta|X,\delta=0))^2\Big]\sqrt{n}(\hat{\gamma}-\gamma^*)+o_p(1). \qquad (3.21)
\end{aligned}$$

Based on the equations (3.20), (3.21) and Lemma 3.1, we can get the result for $\hat{\beta}_2$ in Lemma 3.2.

Lastly, we turn to investigate the term $\hat{\beta}_3$.

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_3-\beta) \; = \; & \Big\{\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}(z_i;\hat{\gamma})}g(x_i)g(x_i)^\top\Big\}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}(z_i;\hat{\gamma})}g(x_i)(y_i-g(x_i)^\top\beta) \\
& = \tilde{A}_1^{*-1}A_4^*,
\end{aligned}$$

For the term A_4^*, we can have

$$\begin{aligned}
A_4^* \; = \; & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\delta_i}{\hat{\pi}(z_i;\gamma^*)}g(x_i)(y_i-g(x_i)^\top\beta)+\left(\frac{\delta_i}{\hat{\pi}(z_i;\hat{\gamma})}-\frac{\delta_i}{\hat{\pi}(z_i;\gamma^*)}\right)g(x_i)(y_i-g(x_i)^\top\beta) \\
= \; & A_4+\frac{1}{n}\sum_{i=1}^{n}\delta_i g(x_i)(y_i-g(x_i)^\top\beta)\frac{\partial\hat{\pi}^{-1}(z_i;\gamma_1)}{\partial\gamma}\sqrt{n}(\hat{\gamma}-\gamma^*) \\
= \; & A_4+\frac{1}{n}\sum_{i=1}^{n}\delta_i g(x_i)(y_i-g(x_i)^\top\beta)\hat{\alpha}(x_i;\gamma_1)\exp(\gamma_1 y_i)(y_i-\hat{m}_0(x_i;\gamma_1))\sqrt{n}(\hat{\gamma}-\gamma^*) \\
= \; & A_4+E\Big[\delta g(X)(Y-g(X)^\top\beta)O(Z;\gamma)(Y-m_0(X))\Big]\sqrt{n}(\hat{\gamma}-\gamma^*) \\
= \; & A_4+E\Big[\delta g(X)\eta O(Z)(\eta-E(\eta|X,\delta=0))\Big]\sqrt{n}(\hat{\gamma}-\gamma^*) \\
= \; & A_4+E\Big[g(X)(1-\delta)(\eta-E(\eta|X,\delta=0))^2\Big]\sqrt{n}(\hat{\gamma}-\gamma^*)+o_p(1).
\end{aligned}$$

The last second equation holds since

$$E\big[\delta g(X)G(X)O(Z)(Y-m_0(X))\big]=E\Big[g(X)G(X)E\big((1-\delta)(Y-m_0(X))|X\big)\Big]\equiv 0.$$

Based on the above equation and Lemma 3.1, we can get the result for $\hat{\beta}_3$ in Lemma 3.2. □

Proof of Theorem 3.3. For $R_{n1}(x)$, it can be verified that

$$\begin{aligned} R_{n1}(x) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big(\delta_i y_i + (1-\delta_i)\hat{m}_0(x_i;\gamma^*) - g(x_i)^\top\hat{\beta}_1\Big) I(x_i \le x) \\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)(\hat{m}_0(x_i;\hat{\gamma}) - \hat{m}_0(x_i;\gamma^*))I(x_i \le x) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\delta_i}{\pi(z_i)}y_i + (1-\frac{\delta_i}{\pi(z_i)})m_0(x_i) - g(x_i)^\top\beta\right) I(x_i \le x) \\ &\quad - E(g(X)^\top I(X \le x))\sqrt{n}(\hat{\beta}_1 - \beta) \\ &\quad + E\Big((1-\delta)I(X \le x)(Y - m_0(X))^2\Big)\sqrt{n}(\hat{\gamma} - \gamma^*). \end{aligned}$$

Under the null hypothesis $H_0 : y_i = g(x_i)^\top\beta + \varepsilon_i$ and based on Lemma 3.2, we can have

$$\begin{aligned} R_{n1}(x) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} L(x_i)\left(\frac{\delta_i\varepsilon_i + (\pi(z_i) - \delta_i)E(\varepsilon|x_i, \delta = 0)}{\pi(z_i)}\right) \\ &\quad + E\Big((1-\delta)(Y - m_0(X))^2 L(X;x)\Big)\sqrt{n}(\hat{\gamma} - \gamma^*) + o_p(1). \end{aligned}$$

Thus the asymptotic properties of T_{n1} follows by the continuous mapping theorem. From the arguments for $\hat{\beta}_2$ and $\hat{\beta}_3$ in Lemma 3.2 the results for $R_{n2}(x)$ and $R_{n3}(x)$ in Theorem 1, similarly, we can know that $R_{n2}(x)$ and $R_{n3}(x)$ have the same asymptotic expansion as that for $R_{n1}(x)$, thus we finish the proof here. □

Proof of Theorem 3.4. Under the alternative H_{1n}, for the test $R_{n1}(x)$, we have

$$\begin{aligned} R_{n1}(x) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{\delta_i}{\pi(z_i)}y_i + (1-\frac{\delta_i}{\pi(z_i)})m_0(x_i) - g(x_i)^\top\beta\right) I(x_i \le x) \\ &\quad - E(g(X)^\top I(X \le x))\sqrt{n}(\hat{\beta}_1 - \beta) \\ &\quad + E\Big((1-\delta)I(X \le x)(Y - m_0(X))^2\Big)\sqrt{n}(\hat{\gamma} - \gamma^*). \end{aligned}$$

Under the local alternative $H_{1n} : y_i = g(x_i)^\top\beta + C_n G(x_i) + \eta_i$ with $n^{1/2}C_n \to 1$, and based on Lemma 3.2, we can have

$$\begin{aligned} R_{n1}(x) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} L(x_i;x)\left(\frac{\delta_i\eta_i + (\pi(z_i) - \delta_i)E(\eta|x_i, \delta = 0)}{\pi(z_i)}\right) + E(G(X)L(X;x)) \\ &\quad + E\Big((1-\delta)(Y - m_0(X))^2 L(X;x)\Big)\sqrt{n}(\hat{\gamma} - \gamma) + o_p(1). \end{aligned}$$

If $n^r C_n \to a, 0 < r < 1/2$, then it yields $\sqrt{n}C_n \to \infty$, as $n \to \infty$. As a result, we have $R_{n1}(x) \to \infty$. The same conclusions for $R_{n2}(x)$ can be obtained easily.

Thus the asymptotic properties of T_{n1} follows by the continuous mapping theorem. From the arguments for $\hat{\beta}_2$ and $\hat{\beta}_3$ in Lemma 3.2 and the results for $R_{n2}(x)$ and $R_{n3}(x)$ in Theorem 3.2, similarly, we can know that $R_{n2}(x)$ and $R_{n3}(x)$ have the same asymptotic expansion as that for $R_{n1}(x)$, thus we finish the proof here. □

Lemma 3.3. *Under conditions 1-5 in the Appendix and the alternative H_{1n}, if $\hat{\gamma}$ is obtained from a validation sample, the asymptotic properties of $\sqrt{n}(\hat{\beta}_k - \beta)$ are as follows*

$$
\begin{aligned}
& \sqrt{n}(\hat{\beta}_k - \beta) \\
= \; & \Big(\Big\{\frac{1}{n}\sum_{i=1}^{n} g(x_i)g(x_i)^\top\Big\}^{-1} - \Sigma^{-1}\Big)C_n\sqrt{n}E(g(X)G(X)) + \Sigma^{-1}C_n\sqrt{n}E(g(X)G(X)) \\
& + \Sigma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{g(x_i)\Big[\frac{\delta_i}{\pi(z_i)}\eta_i + \Big(1 - \frac{\delta_i}{\pi(z_i)}\Big)E(\eta|x_i, \delta = 0)\Big] \\
& + HM^{-1}(\eta_i - E(\eta|x_i, \delta = 0))\Big[(1-\delta_i)r_i - \delta_i\nu(\frac{1}{\pi(z_i)} - 1)\Big]\Big\} + o_p(1), k = 1, 2, 3.
\end{aligned}
$$

here $H = E\Big[g(X)(1-\delta)(\eta - E(\eta|X, \delta = 0))^2\Big]$ *and* $M = E\Big[(1-\delta)r(E(Y^2|X, \delta = 0) - m_0^2(X, \gamma^*))\Big]$.

Proof of Lemma 3.3. Use the same notations, for $\hat{\beta}_1$, from the proof of Lemma 3.2, we can have $\sqrt{n}(\hat{\beta}_1 - \beta) = A_1^{-1}(A_2 + H\sqrt{n}(\hat{\gamma} - \gamma^*))$. The above formula is the same if $\hat{\gamma}$ is obtained from an independent survey or a validation sample. In the following, we first investigate the asymptotic property of $\hat{\gamma}$ in this situation.

Note that

$$
\begin{aligned}
0 &= \sum_{i=1}^{n}(1-\delta_i)r_i(y_i - \hat{m}_0(x_i, \hat{\gamma})) \\
&= \sum_{i=1}^{n}(1-\delta_i)r_i(y_i - \hat{m}_0(x_i, \gamma^*)) - \sum_{i=1}^{n}(1-\delta_i)r_i\frac{\partial\hat{m}_0(x_i, \gamma_1)}{\partial\gamma}(\hat{\gamma} - \gamma^*).
\end{aligned}
$$

Thus, we can have

$$\begin{aligned}\sqrt{n}(\hat{\gamma}-\gamma^*) &= \Big[\frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)r_i\frac{\partial\hat{m}_0(x_i,\gamma_1)}{\partial\gamma}\Big]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)r_i(y_i-\hat{m}_0(x_i,\gamma^*)) \\ &=: E\Big[(1-\delta)r(E(Y^2|X,\delta=0)-m_0^2(X,\gamma^*))\Big]^{-1}\times D_1+o_p(1).\end{aligned}$$

Notice that:

$$\begin{aligned}&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(1-\delta_i)r_i(m_0(x_i,\gamma^*)-\hat{m}_0(x_i,\gamma^*)) \\ =\ &\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\delta_j\exp(\gamma^* y_j)(m_0(x_j,\gamma^*)-y_j)E\left(\frac{(1-\delta)rK_h(X,x_j)}{f(X)(1-P(X)\alpha(X,\gamma^*))}|x_j\right) \\ =\ &\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\delta_j\left(\frac{1}{\pi(x_j,y_j)}-1\right)(m_0(x_j,\gamma^*)-y_j)\nu.\end{aligned}$$

Thus, we can have

$$\sqrt{n}(\hat{\gamma}-\gamma^*) = M^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(y_i-m_0(x_i,\gamma^*))\left[(1-\delta_i)r_i-\delta_i\nu\left(\frac{1}{\pi(x_j,y_j)}-1\right)\right].$$

Recall the results for A_2 from Lemma 3.1, we can have

$$\begin{aligned}&\sqrt{n}(\hat{\beta}_1-\beta) \\ =\ &\Big(\Big\{\frac{1}{n}\sum_{i=1}^{n}g(x_i)g(x_i)^\top\Big\}^{-1}-\Sigma^{-1}\Big)C_n\sqrt{n}E(g(X)G(X))+\Sigma^{-1}C_n\sqrt{n}E(g(X)G(X)) \\ &+\Sigma^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big\{g(x_i)\Big[\frac{\delta_i}{\pi(z_i)}\eta_i+\Big(1-\frac{\delta_i}{\pi(z_i)}\Big)E(\eta|x_i,\delta=0)\Big] \\ &+HM^{-1}(\eta_i-E(\eta|x_i,\delta=0))\Big[(1-\delta_i)r_i-\delta_i\nu\big(\frac{1}{\pi(z_i)}-1\big)\Big]\Big\}.\end{aligned}$$

We can similarly obtain the asymptotic properties for $\hat{\beta}_2$ and $\hat{\beta}_3$.

Proof of Theorem 3.5. For $R_{n1}(x)$, it can be verified that

$$\begin{aligned}R_{n1}(x) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\delta_i y_i+(1-\delta_i)\hat{m}_0(x_i;\hat{\gamma})-g(x_i)^\top\beta\right)I(x_i\le x) \\ &\quad -E(g(X)^\top I(X\le x))\sqrt{n}(\hat{\beta}_1-\beta).\end{aligned}$$

Under the null hypothesis $H_0 : y_i = g(x_i)^\top \beta + \varepsilon_i$ and based on Lemma 3.3, we can have

$$\begin{aligned} R_{n1}(x) &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} L(x_i; x) \left(\frac{\delta_i \varepsilon_i + (\pi(z_i) - \delta_i) E(\varepsilon | x_i, \delta = 0)}{\pi(z_i)} \right) \\ &\quad + E\Big((1-\delta)(Y - m_0(X))^2 L(X; x) \Big) \\ &\quad \times M^{-1} (\eta_i - E(\eta | x_i, \delta = 0)) \Big[(1-\delta_i) r_i - \delta_i \nu (\frac{1}{\pi(z_i)} - 1) \Big] + o_p(1). \end{aligned}$$

The results for R_{n2} and R_{n3} can be similarly obtained. We omit the details here.

Proof of Theorem 3.6. From the proof of Theorem 5 and the results in Lemma 3, we can prove this Theorem. We omit the details here.

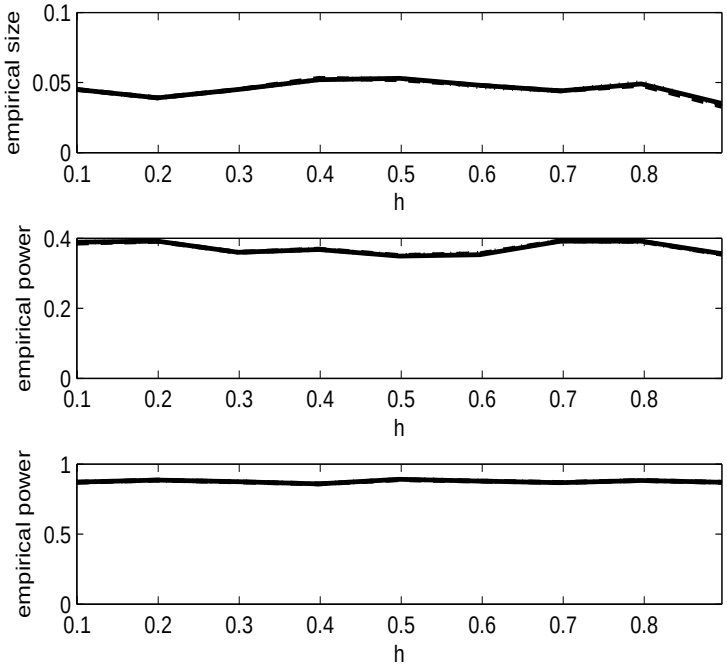

Figure 3.1: The estimated power curves of the tests against the bandwidth h with missing mechanisms $\pi_1(x, y)$ and sample size 100 under different choices of a for study 1 with $a = 0$(the above panel); $a = 0.5$(the central panel); and $a = 1$(the below panel). The solid line, dotted line and dashed line represent the results from T_{n1}, T_{n2} and T_{n3}.

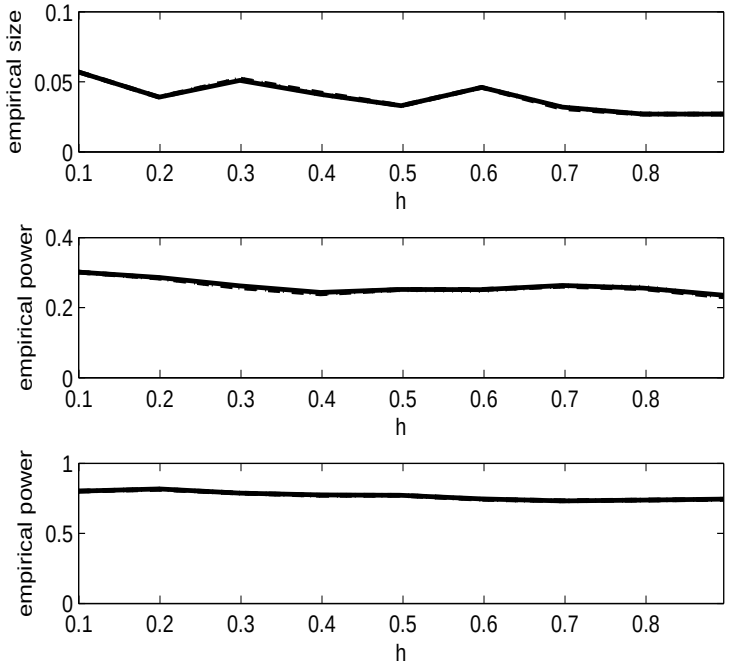

Figure 3.2: The estimated power curves of the tests against the bandwidth h with missing mechanisms $\pi_2(x, y)$ and sample size 100 under different choices of a for study 1 with $a = 0$(the above panel); $a = 0.5$(the central panel); and $a = 1$(the below panel). The solid line, dotted line and dashed line represent the results from T_{n1}, T_{n2} and T_{n3}.

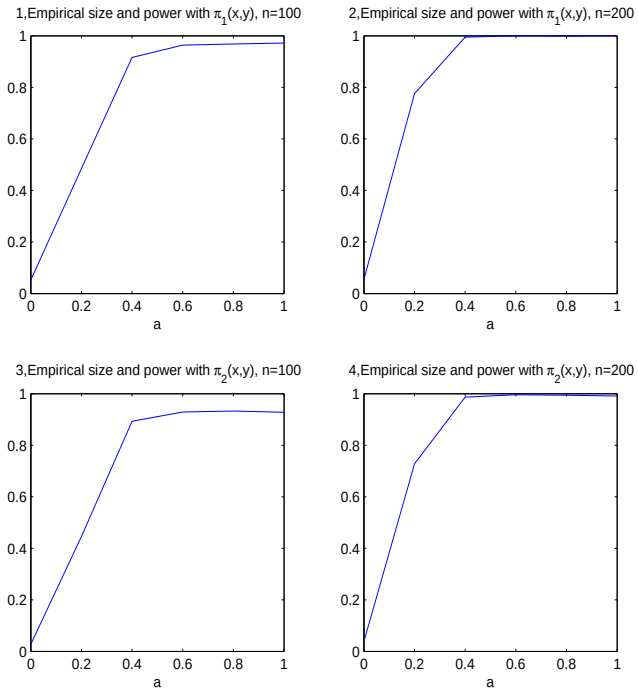

Figure 3.3: Empirical sizes and powers of T_{n1} for Study 2 with $n = 100$ and $n = 200$: (1) for $\pi_1(x, y)$ and $n = 100$; (2) for $\pi_1(x, y)$ and $n = 200$; (3) for $\pi_2(x, y)$ and $n = 100$ and (4) for $\pi_2(x, y)$ and $n = 200$.

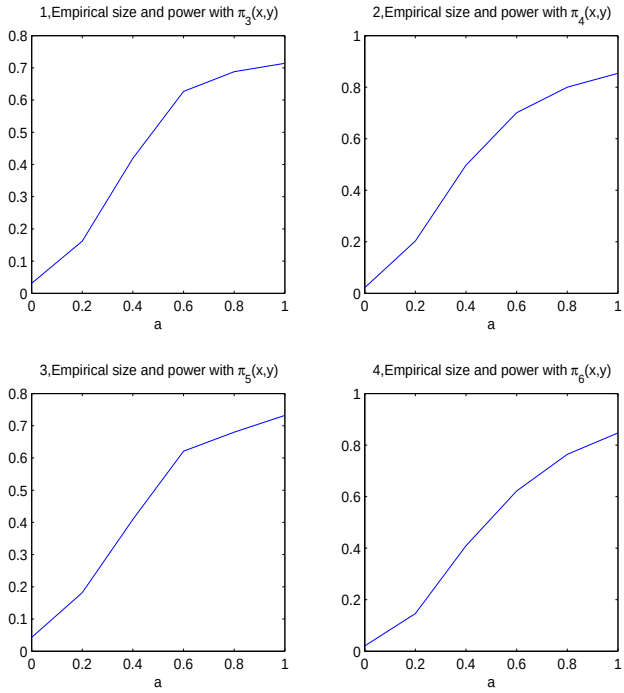

Figure 3.4: Empirical sizes and powers of T_{n1} for Study 3 with $n = 100$: (1) for $\pi_3(x, y)$; (2) for $\pi_4(x, y)$; (3) for $\pi_5(x, y)$ and (4) for $\pi_6(x, y)$.

Table 3.1: Empirical sizes and powers for study 1, with $n = 100, 200$ and missing mechanism $\pi_i(x, y), i = 1, 2$.

	a	$n = 100$	$n = 200$
$\pi_1(x, y)$	0.0	0.0480	0.0530
	0.3	0.1400	0.2960
	0.6	0.5110	0.8530
	0.9	0.8160	0.9920
	1.2	0.9450	0.9960
	1.5	0.9460	0.9990
	1.8	0.9570	0.9990
$\pi_2(x, y)$	0.0	0.0400	0.0430
	0.3	0.1030	0.1900
	0.6	0.3640	0.7290
	0.9	0.6940	0.9300
	1.2	0.8520	0.9580
	1.5	0.8650	0.9710
	1.8	0.8680	0.9790

Chapter 4

Model Checking for Generalized Linear Models: An Hypothesis-Adaptive Method

4.1 Introduction

Consider the following generalized linear regression model:

$$Y = g(\beta^\tau \mathbf{X}) + \epsilon,$$

here Y is the scalar response, $\mathbf{X}$ is a predictor vector of p dimension, $g(\cdot)$ is a known squared integrable continuous function, β is any p-dimensional unknown parameter vector, and $E(\epsilon|\mathbf{X}) = 0$.

To make the statistical inference based on regression models reliable, as demonstrated and reviewed in Chapter 1, we should carry out some suitable and efficient model checking procedures. However, traditional methods suffer from the curse of dimension. To be precise, for the smoothing-based, or local smoothing methods, they generally involve estimating some nonparametric function. When the dimension is high, we can not get accurate estimate for the nonparametric function. On the other hand, the empirical regression process based, or global smoothing methods depend

on some high dimensional process. Because of the sparsity, the power performance will drop greatly. Further it's generally believed that high dimensional process is computationally intensive.

Under the null hypothesis, we can have $E(\epsilon|\beta^\tau \mathbf{X}) = 0$. Now, let's turn to the general alternative hypothesis which takes the following form:

$$Y = G(\mathbf{X}) + \eta, \tag{4.1}$$

here $G(\cdot)$ is unknown smooth function and $E(\eta|\mathbf{X}) = 0$. Under the above alternative hypothesis, we can have $E(\epsilon|\mathbf{X}) = E(Y - g(\beta^\tau \mathbf{X})|\mathbf{X}) = E(G(\mathbf{X}) - g(\beta^\tau \mathbf{X})|\mathbf{X}) \neq 0$. This inequality can also be written as $E(\epsilon|B^\tau \mathbf{X}) \neq 0$ with B being any nonsingular $p \times p$ matrix. For identifiability, we assume that the matrix B satisfies $B^\tau B = I_p$ an identity matrix. Further the nonparametric regression model can also be written as $Y = G(B^\tau \mathbf{X}) + \eta$. In other words, the model can be considered as a special multi-index model with p indexes. On the other hand, if the alternative model is single index model or partial linear single index model, the number of the index in the above defined multi-index model will be one or two respectively. In practice, we generally have no idea about the number of index, instead we use d as the true number of index. The estimation of d will be specified later. These several observations can help us develop new hypothesis-adaptive testing procedures.

To this end, sufficient dimension reduction (SDR, Cook 1998) may help. The space, denoted $S_{E(Y|\mathbf{X})}$, is called the central mean subspace (CMS, Cook and Li 2002). It is, in effect, the intersection of all subspaces S_B such that $Y \perp\!\!\!\perp E(Y|\mathbf{X})|B^\tau \mathbf{X}$ for all $p \times d$ matrices B. The estimation methods for the d basis vectors of $S_{E(Y|\mathbf{X})}$ include the sliced inverse regression (SIR, Li 1991), sliced average variance estimation (SAVE, Cook and Weisberg 1991), contour regression (CR, Li, Zha and Chiaromonte 2005), directional regression (DR, Li and Wang 2007), likelihood acquired directions (LAD, Cook and Forzani 2009), discretization-expectation estimation (DEE, Zhu, Wang, Zhu and Ferré 2010) and average partial mean estimation (APME, Zhu, Zhu and Feng 2010). Zhu and Ng (1995) and Zhu and Fang (1996) proved the asymptotic

normality of the SIR directions, and Li and Zhu (2007) systematically investigated the asymptotic behaviors of the SAVE directions. All of these estimations can have root-n consistency. In addition, a minimum average variance estimation (MAVE, Xia et al, 2002; Xia 2006) can estimate relevant space with fewer regularity conditions on $\mathbf{X}$ and can exhaustively estimate the dimensions in the conditional mean function.

In this Chapter, we revisit the traditional local smoothing methods. Because of its technical tractability and easy computation, we focus on Zheng (1996)'s procedure. However, we should note that the basic idea can be extended to other traditional model checking procedures. For details, we leave this to the discussion part. The adaptive property of our methodology implies that under the null hypothesis, d is estimated to be one, while under the alternative hypothesis, we obtain a consistent estimator of d. In other words, our methodology can be adaptive to the true working model. In this way, we can improve the traditional local smoothing methods greatly in asymptotic sense. Our adaptive procedure can detect local alternative at faster rate than the traditional local smoothing methods. These points will be specified in the following sections.

The rest of this Chapter is organized as follows. In Section 4.2, we construct a test statistic. Then in Section 4.3, we derive its asymptotic properties under the null and alternative hypotheses. In Section 4.4, simulation results are reported and a real data analysis is carried out to illustrate the proposed test. Some discussions about further research are put forward in Section 4.5. The proofs of the theoretical results are postponed to the appendix 4.6.

4.2 Adaptive Test Procedure

Under the null hypothesis, $d = 1$ and then $\tilde{B} = \tilde{\beta} = c\beta$ for some scalar c. Thus, under the null hypothesis,

$$E(\epsilon|\mathbf{X}) = 0 \Longleftrightarrow E(\epsilon|\beta^\tau \mathbf{X}) = E(\epsilon|\tilde{B}^\tau \mathbf{X}) = 0.$$

With slightly notational abuse, we still use B in the place of $\tilde{B}$. Therefore, under H_0,

$$E(\epsilon E(\epsilon|B^\tau \mathbf{X})W(\mathbf{X})) = E(E^2(\epsilon|B^\tau \mathbf{X})W(\mathbf{X})) = 0, \tag{4.2}$$

where $W(X)$ is some positive weight function that is discussed below.

Under the alternative hypothesis H_1, $E(Y|B^\tau \mathbf{X}) \neq g(\beta^\tau \mathbf{X})$, we have

$$E(\epsilon E(\epsilon|B^\tau \mathbf{X})W(\mathbf{X})) = E(E^2(\epsilon|B^\tau \mathbf{X})W(\mathbf{X})) > 0, \tag{4.3}$$

The empirical version of the left hand side in (4.2) can then be used as a test statistic, and the null hypothesis can be rejected for large test statistic values. For simplicity of bandwidth selection and ease of exposition, after the standardization, a common bandwidth is used for all variables. In doing so, we estimate $E(\epsilon|B^\tau \mathbf{X})$ by

$$\hat{E}(\epsilon_i|\hat{B}^\tau \mathbf{x}_i) = \frac{1}{n-1}\sum_{j\neq i}^{n} \hat{\epsilon}_j K_h(\hat{B}^\tau \mathbf{x}_i - \hat{B}^\tau \mathbf{x}_j)/\hat{f}(\hat{B}^\tau \mathbf{x}_i).$$

In the above formula, $\hat{\epsilon}_j = y_j - g(\hat{\beta}^\tau \mathbf{x}_j)$, with $\hat{\beta}$ being the commonly used least squares estimate of β, $\hat{f}(\hat{B}^\tau \mathbf{X})$ is an estimate of the density function $f(\cdot)$ of $B^\tau \mathbf{X}$, $\hat{B}$ is a sufficient dimension reduction estimate, $K_h(\cdot) = K(\cdot/h)/h^{\hat{d}}$ with $K(\cdot)$ being a kernel function, $\hat{d}$ being a consistent estimate of d and h being a bandwidth. For the estimation of B and the selection of the number of index d, we will specify later. Let $\hat{f}(\hat{B}^\tau \mathbf{X})$ be a kernel estimate of $f(B^\tau \mathbf{X})$ of $B^\tau \mathbf{X}$ with the formula

$$\hat{f}(\hat{B}^\tau \mathbf{X}) = \frac{1}{n-1}\sum_{j\neq i}^{n} K_h(\hat{B}^\tau \mathbf{x}_i - \hat{B}^\tau \mathbf{x}_j).$$

When the weight $W(\cdot)$ is chosen to be $f(\cdot)$, a test statistic is defined by

$$V_n = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \hat{\epsilon}_i \hat{\epsilon}_j K_h(\hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j)). \tag{4.4}$$

Remark 4.1. *Zheng (1996) proposed the following test statistics*

$$V_n = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \hat{\epsilon}_i \hat{\epsilon}_j K_h(\mathbf{x}_i - \mathbf{x}_j), \tag{4.5}$$

here $K_h(\cdot) = K(\cdot/h)/h^p$. Compare the formulas (4.4) and (4.5), we can notice that the difference between our method and Zheng (1996)'s test is the use of $\hat{B}^\tau X$. As

shown before, under the null hypothesis, the working dimension of $\hat{B}^\tau X$ is one, while under the alternative, it will be $\hat{d}$, a consistent estimator of $d \leq p$. The working dimension is adaptive to the underlying model. On the other hand, for Zheng (1996)'s procedure, whether the null or the alternative hypothesis holds, the working dimension is always p, a fixed value.

4.2.1 Review of DEE

As described in above, our test procedure needs to estimate the matrix B. In this subsection, we first assume the the dimension d is known in ahead and then we discuss how to select the dimension d consistently. We first give a brief review of discretization-expectation estimation (DEE), see Zhu, Wang, Zhu and Ferré (2010) for details. In sufficient dimension reduction, SIR and SAVE are two popular methods which involve the partition of the range of Y into several slices and the choice of the number of slices. However, as documented by many authors, for instance, Li (1991), Zhu and Ng (1995) and Li and Zhu (2007), the choice of the number of slices may effect the efficiency and can even yard inconsistent estimators. To avoid the delicate choice of the number of slices, Zhu, Wang, Zhu and Ferré (2010) introduced the discretization-expectation estimation (DEE). The basic idea is simple. We first define the new response variable $Z(t) = I(Y \leq t)$, which takes the value 1 if $Y \leq t$ and 0 otherwise. Let $\mathcal{S}_{Z(t)|\mathbf{X}}$ be the central subspace and $\mathcal{M}(t)$ be a $p \times p$ positive semi-definite matrix such that $\text{span}\{\mathcal{M}(t)\} = \mathcal{S}_{Z(t)|\mathbf{X}}$. Define $\mathcal{M} = E\{\mathcal{M}(T)\}$. Under some mild conditions, we can have $\mathcal{M} = \mathcal{S}_{Y|\mathbf{X}}$.

In the discretization step, we construct sample $\{\mathbf{x}_i, z_i(y_j)\}$ with $z_i(y_j) = I(y_i \leq y_j)$. For each fixed y_j, we estimate $\mathcal{M}(y_j)$ by using SIR or SAVE. Let $\mathcal{M}_n(y_j)$ denote the candidate matrix obtained from a chosen basic method. In the expectation step, we can estimate $\mathcal{M}$ by $\mathcal{M}_{n,n} = n^{-1}\sum_{j=1}^{n}\mathcal{M}_n(y_j)$. The d eigenvectors of $\mathcal{M}_{n,n}$ corresponding to its d largest eigenvalues can be used to form an estimator of B. Denote the DEE procedure based on SIR and SAVE be DEE_{SIR} and DEE_{SAVE}

respectively. To save space, in this chapter, we only focus on these two basic methods.

4.2.2 Review of MAVE

As well known, the SIR and SAVE needs some extra conditions on the marginal distribution of the predictors to successfully estimate the matrix B. On the other hand, the Minimum Average conditional Variance Estimation (MAVE) is an efficient method with fewer regularity conditions on the predictors and can exhaustively estimate the dimensions in the conditional mean function. In the following, we apply MAVE to estimate B, see Xia et al. (2002) or Chapter 1 subsection 1.2.2 for details. The estimator for B is the minimizer of

$$\sum_{j=1}^{n}\sum_{i=1}^{n}(y_i - a_j - \mathbf{d}_j^\tau B^\tau \mathbf{x}_{ij})^2 K_h(B^\tau \mathbf{x}_{ij}),$$

over all B satisfying $B^\tau B = I_d$, a_j and $\mathbf{d}_j$, here $\mathbf{d}_j = G'(B^\tau \mathbf{x}_j)$ and $\mathbf{x}_{ij} = \mathbf{x}_i - \mathbf{x}_j$. The algorithm is as follows.

Step 1. For given B_0 being an initial estimate of B, calculate

$$\begin{aligned}(a_j^B, \mathbf{d}_j^B h)^\tau &= \Big\{\sum_{i=1}^{n} K_h(B_0^\tau \mathbf{x}_{ij})(1, \mathbf{x}_{ij}^\tau B_0/h)^\tau (1, \mathbf{x}_{ij}^\tau B_0/h)\Big\}^{-1} \\ &\quad \times \sum_{i=1}^{n} K_h(B_0^\tau \mathbf{x}_{ij})(1, \mathbf{x}_{ij}^\tau B_0/h)^\tau y_i.\end{aligned}$$

Step 2. Based on the estimator a_j^B and $\mathbf{d}_j^B$ from Step 1, calculate

$$\begin{aligned}\tilde{B} &= \Big\{\sum_{i,j=1}^{n} K_h(B_0^\tau x_{ij})\mathbf{d}_j^B x_{ij}^\tau \mathbf{x}_{ij} \mathbf{d}_j^{B\tau}/\hat{f}(B_0^\tau \mathbf{x}_j)\Big\}^{-1} \\ &\quad \times \sum_{i,j=1}^{n} K_h(B_0^\tau \mathbf{x}_{ij})\mathbf{d}_j^B \mathbf{x}_{ij}^\tau (y_i - a_j^B)/\hat{f}(B_0^\tau \mathbf{x}_j).\end{aligned}$$

Step 3. Repeat steps 1 and 2 with $\tilde{B}$ until convergence.

The final estimator is denoted by $\hat{B}$. In the above procedures, d is assumed to be given, when it is unknown, estimating it is involved.

4.2.3 Estimation of Dimension d

We first describe the BIC method used to estimate the dimension d for DEE. According to Zhu, Wang, Zhu and Ferré (2010), we determine $d = \dim(S_{Y|\mathbf{X}})$ by

$$\hat{d} = \arg\max_{l=1,\cdots,p} \left\{ \frac{n}{2} \times \frac{\sum_{i=1}^{l}\{\log(\hat{\lambda}_i + 1) - \hat{\lambda}_i\}}{\sum_{i=1}^{p}\{\log(\hat{\lambda}_i + 1) - \hat{\lambda}_i\}} - 2 \times D_n \times \frac{l(l+1)}{2p} \right\},$$

where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$ are the eigenvalues of $\mathcal{M}_{n,n}$. We should note that the first term in the bracket can be considered as likelihood ratio, and the second term is the penalty term with $l(l+1)/2$ free parameters when the dimension is l. Zhu, Wang, Zhu and Ferré (2010) explained the calculation of this number of free parameters in details. See also Zhu, Miao and Peng (2006) for more discussions of the BIC methodology. The major merit of this methodology (BIC) is that to show the consistency of the estimator of the dimension, we only need the convergence of the estimate of the relevant matrix. Zhu, Wang, Zhu and Ferré (2010) proved that under some suitable conditions, $\hat{d}$ is a consistent estimator of d. Following their suggestion, we choose D_n to be $n^{1/2}$.

For the MAVE, though the cross-validation method of Xia et al. (2002) can be adopted to determine these dimensions, it can be computationally intensive. We instead apply a BIC criterion proposed by Wang and Yin (2008) for MAVE. It has the following form:

$$BIC_k = \log(\frac{RSS_k}{n}) + \frac{\log(n)k}{nh^k},$$

where RRS_k is the residual sum of squares, and k is the estimate of the dimension. The form of RSS_k is as follows:

$$RSS_k = \sum_{j=1}^{n}\sum_{i=1}^{n}(y_i - \hat{a}_j - \hat{\mathbf{d}}_j^{\tau}\hat{B}_k^{\tau}\mathbf{x}_{ij})^2 K_h(\hat{B}_k^{\tau}\mathbf{x}_{ij}),$$

here we use B_k to denote the matrix B when the dimension is k.

The estimated dimension is then

$$\hat{d} = \min\{l : l = \arg\min_{1\leq k\leq p}\{BIC_k\}\}.$$

Wang and Yin (2008) showed that under some mild conditions, $\hat{d}$ is also consistent estimator of d.

To investigate the power performance of our proposed tests, it's important to study the asymptotic properties of the two estimated $\hat{d}$ under some local alternatives. We will discuss this point in details in the following.

4.3 Asymptotic Properties

We offer some notations before presenting the details of the following theoretical results. Let $\mathbf{Z} = B^\tau \mathbf{X}, \sigma^2(\mathbf{z}) = E(\epsilon^2|\mathbf{Z} = \mathbf{z})$,

$$
\begin{aligned}
var &= 2\int K^2(u)du \cdot \int (\sigma^2(\mathbf{z}))^2 f^2(\mathbf{z})d\mathbf{z}, \\
\widehat{var} &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i}^{n} \frac{1}{h^{\hat{d}}} K^2(\frac{\hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j)}{h}) \hat{\epsilon}_i^2 \hat{\epsilon}_j^2.
\end{aligned}
$$

We now state the asymptotic property of the test statistic under the null hypothesis.

Theorem 4.1. *Under H_0 and conditions in the Appendix, we have*

$$nh^{1/2}V_n \Rightarrow N(0, var).$$

Moreover, var can be consistently estimated by $\widehat{var}$.

We now standardize V_n to get a scale-invariant statistic. According to Theorem 4.1, the standardized V_n is

$$
\begin{aligned}
T_n &= \sqrt{\frac{n-1}{n}} \frac{nh^{1/2}V_n}{\sqrt{\widehat{var}}} \\
&= \frac{h^{(1-\hat{d})/2} \sum_{i=1}^{n} \sum_{j\neq i}^{n} \hat{\epsilon}_i \hat{\epsilon}_j K(\frac{\hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j)}{h})}{\{2\sum_{i=1}^{n} \sum_{j\neq i}^{n} K^2(\frac{\hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j)}{h}) \hat{\epsilon}_i^2 \hat{\epsilon}_j^2\}^{1/2}}.
\end{aligned}
$$

By the consistency of $\widehat{var}$, the application of the Slusky theorem yields the following corollary.

Corollary 4.1. *Assume that the same conditions as those in Theorem 1 hold. Under H_0, we have*

$$T_n^2 \Rightarrow \chi_1^2,$$

where χ_1^2 is the chi-square distribution with one degree of freedom.

From this corollary, we can then calculate the critical and p values easily by normal approximation for large sample size. However, it is also well known that the rate of convergence to the normal limit is slow. Thus the use of the asymptotic normality may be inappropriate for small sample sizes. As an alternative for calibrating critical values, we may need to consider some re-sampling methods. We will discuss these points in the simulation studies further.

4.3.1 Power Study

We now examine the power performance of the proposed test statistic under a sequence of local alternatives with the form

$$H_{1n} : Y = g(\beta^\tau \mathbf{X}) + C_n G(B^\tau \mathbf{X}) + \eta, \tag{4.6}$$

where $E(\eta|\mathbf{X}) = 0$ and the function $G(\cdot)$ satisfies $E(G^2(B^\tau \mathbf{X})) < \infty$. In this sequence of models, β is one of the columns in B.

Denote $g'(\beta^\tau \mathbf{X})\mathbf{X} = \text{grad}_\beta(g(\beta^\tau \mathbf{X}))^\tau$, $H(\mathbf{X}) = G(B^\tau \mathbf{X})g'(\beta^\tau \mathbf{X})\mathbf{X}$ and $\Sigma_x = E((g'(\beta^\tau \mathbf{X}))^2 \mathbf{X}\mathbf{X}^\tau)$.

Before we present the main result about the power performance under the local alternative, our Lemma 2 in Appendix states that under local alternative, d still can be selected to be one.

Now, we are ready to give the power performance of our proposed test procedure under the local alternative.

Theorem 4.2. *Under the conditions in Appendix, we have the following if $C_n = n^{-1/2}h^{-1/4}$, $nh^{1/2}V_n \Rightarrow N(\mu, var)$ and $T_n^2 \Rightarrow \chi_1^2(\mu^2/var)$, where*

$$\mu = E\left[\left(G(B^\tau \mathbf{X}) - g'(\beta^\tau \mathbf{X})\mathbf{X}^\tau \Sigma_x^{-1} E[H(\mathbf{X})]\right)^2 f(B^\tau \mathbf{X})\right],$$

here $\chi_1^2(\mu^2/var)$ is a noncentral chi-squared random variable with one degree of freedom and the noncentrality parameter μ^2/var.

We know that in the fixed p scenarios, the optimal rate for detectable alternative approaching the null is usually $C_n = n^{-1/2}h^{-p/4}$, see e.g. Härdle and Mammen (1993) and Zheng (1996). That is, when p is fixed and $C_n = n^{-1/2}h^{-p/4}$, T_n^2 may converge in probability to a noncentral chi-square variable. However, Theorem 2 tells a totally different story. By using adaptive procedure, our test can make great improvement on the optimal rate over the traditional local smoothing tests. The success of our method is due to the fact that under null and local alternative hypothesis, the dimension of B, d, is both selected to be one. Our method works on the one dimension index $\hat{B}^\tau \mathbf{X}$ instead the original p dimension covariates $\mathbf{X}$. In this way, the testing problem in some sense turns to be checking whether the single index model is generalized linear model or not. But we should note that our method is not limited to single index alternative. For general alternative (4.1), our method is still applicable.

4.4 Numerical Analysis

4.4.1 Simulations

We now carry out simulations to examine the performance of the proposed test. Because the situation is similar, we then consider linear models instead of generalized linear models as the hypothetical models in the following studies. Further to save space, we only consider the DEE procedure based on SIR and MAVE in the following.

Study 1. The hypothetical model is linear:

$$Y = \beta^\tau \mathbf{X} + \epsilon, \tag{4.7}$$

where $\beta = (1, 1, \cdots, 1)^\tau/\sqrt{p}$, $\mathbf{X} = (X_1, X_2, \cdots, X_p)^\tau$ and p is set to be 8.

The observations $\mathbf{x}_i, i = 1, 2, \cdots, n$, are i.i.d. respectively from multivariate nor-

mal distribution $N(0, \Sigma_j), j = 1, 2$, with

$$\Sigma_1 = I_{p\times p}; \quad \Sigma_2 = (0.5^{|j-l|})_{p\times p}.$$

As for ϵ, we consider the distribution $N(0, 1)$.

To examine the power performance, consider the alternatives as follows:

$$H_{11} : Y = \beta^\tau \mathbf{X} + a\cos(0.6\pi\beta^\tau \mathbf{X}) + \epsilon;$$
$$H_{12} : Y = \beta^\tau \mathbf{X} + a\exp\{-(\beta^\tau \mathbf{X})^2\} + \epsilon;$$
$$H_{13} : Y = \beta^\tau \mathbf{X} + a(\beta^\tau \mathbf{X})^2 + \epsilon.$$

The value $a = 0$ corresponds to the null hypothesis and $a \neq 0$ to the alternatives.

In a nonparametric estimation, the kernel function is taken to be $K(u) = 15/16(1-u^2)^2$, if $|u| \leq 1$; and 0 otherwise. In the simulations, the replication time is $2,000$. The bandwidth is taken to be $h = 1.5n^{-1/(4+\hat{d})}$ for simplicity. To investigate the impact of bandwidth selection on our proposed tests, we take the bandwidth h to be $n^{-1/(4+\hat{d})}(0.25 + i/4)$ for $i = 0, \cdots, 8$. We let the nominal level be $\alpha = 0.05$ in simulation study. First, we plot the estimated power curve against the above bandwidth sequences with $X \sim N(0, \Sigma_1)$, H_{11} and sample size 50 which is shown in Fig 4.1. We also consider other alternatives, similar patterns are found. To save space, we omit the figures for other situations. From this figure, we can conclude that the bandwidths have little effects on the empirical sizes of our proposed tests. Under different bandwidths, our proposed tests can control the size very well. On the other hand, the bandwidth does have impact on the empirical powers of our proposed test, especially when we set the bandwidth too small. Overall, from our empirical studies here, $h = 1.5n^{-1/(4+\hat{d})}$ can be suitably used.

We denote the test statistics T_n based on DEE and MAVE as T_n^{DEE} and T_n^{MAVE} respectively. We present the empirical sizes of our proposed test with different significance level $\alpha = 0.01, 0.05$ and 0.1 and different sample sizes $n = 50$ and 100 in Table 4.1. From this table, we can see clearly that the empirical sizes of our proposed test T_n^{DEE} for significance level $\alpha = 0.01$ is somewhat large and underestimate for

$\alpha = 0.10$. However, for significance level $\alpha = 0.05$, it controls the size very well even under small sample size $n = 50$. This suggests that at least for significance level $\alpha = 0.05$, we can rely on the use of the asymptotic normality for finite sample studies. For T_n^{MAVE}, our simulations which are not reported here show that the empirical sizes are larger than the nominal sizes. Thus we need some modifications for these two test statistics, especially for T_n^{MAVE}.

As an alternative for calibrating critical values, we consider the wild bootstrap introduced by Wu (1986) in the following. Let the bootstrap observations:

$$y_i^* = \hat{\beta}^\tau \mathbf{x}_i + \hat{\epsilon}_i \times V_i.$$

Here $\{V_i\}_{i=1}^n$ is a sequence of i.i.d. random variables with zero mean, unit variance and independent of the sequence $\{y_i, \mathbf{x}_i\}_{i=1}^n$. Usually, $\{V_i\}_{i=1}^n$ can be chosen to be i.i.d. Bernoulli variates with

$$P(V_i = \frac{1-\sqrt{5}}{2}) = \frac{1+\sqrt{5}}{2\sqrt{5}}, \qquad , \; P(V_i = \frac{1+\sqrt{5}}{2}) = 1 - \frac{1+\sqrt{5}}{2\sqrt{5}}.$$

Let T_n^* be defined similarly as T_n, basing on the bootstrap sample $(\mathbf{x}_1, y_1^*), \cdots, (\mathbf{x}_n, y_n^*)$. The null hypothesis is rejected if T_n is bigger than the corresponding quantile of the bootstrap distribution of T_n^*. We denote the bootstrap version of T_n based on DEE and MAVE as T_n^{DEE*} and T_n^{MAVE*} respectively. The study of the asymptotic validity of this procedure will be undertaken elsewhere. Here, we investigate the empirical properties of this bootstrap procedure.

For T_n^{MAVE}, we also consider another size-adjustment as follows:

$$\widetilde{T}_n^{MAVE} = \frac{T_n^{MAVE}}{1 + 4n^{-4/5}}.$$

After such an adjustment, the new tests $\widetilde{T}_n^{MAVE}$ can much better control type I errors as shown in table 1 than those without size-adjustment. The size-adjustment constant is selected via intensive simulation with many different values and this one is recommended.

From table 4.1, we can observe that the empirical sizes of T_n^{MAVE*} are slightly larger than the nominal sizes especially when $\mathbf{X} \sim N(0, \Sigma_1)$. On the other hand, $\widetilde{T}_n^{MAVE}$ can control the size acceptably under different situations. We can also see that T_n^{DEE*} can control the empirical sizes best, that's, under all the situations we consider, T_n^{DEE*} can maintain the type I error very well. In sum, for significance level $\alpha = 0.01, 0.05$ and 0.1, T_n^{DEE*} is the best option to control the size. If we only focus on the significance level $\alpha = 0.05$, T_n^{DEE} and $\widetilde{T}_n^{MAVE}$ would also be some suitable choice.

Now we turn to study the empirical powers of our proposed test against alternatives $H_{1i}, i = 1, 2, 3$ with nominal size $\alpha = 0.05$. The results are shown in tables 4.2-4.4. From the table 4.2 and 4.4, we can find that when $\mathbf{X} \sim N(0, \Sigma_1)$, $\widetilde{T}_n^{MAVE}$ generally yard larger powers compared with T_n^{DEE}. On the other hand, when we generate $\mathbf{X}$ from $N(0, \Sigma_2)$, T_n^{DEE} becomes the winner. Further, we can observe that compared with T_n^{MAVE*}, T_n^{DEE*} generally can control the empirical size better though it may lose some power. Moreover, for alternatives H_{11} and H_{12}, generally T_n^{MAVE*} has relatively larger power than $\widetilde{T}_n^{MAVE}$ while for alternative H_{13}, the size-adjustment procedure $\widetilde{T}_n^{MAVE}$ is more powerful than the bootstrapped version T_n^{MAVE*}. As for our test based on DEE, we can find that almost in all cases, T_n^{DEE} can obtain larger power than its bootstrapped version T_n^{DEE*}. These several tests are all very sensitive to the alternatives. To be precise, when a increases, the powers can increase very quickly. In sum, from this simulation study, we can conclude that T_n^{DEE} can not only control the size satisfactorily but also can yard reasonable powers.

To illustrate the performance of our test when there are more than one direction under the alternative hypothesis, we construct the following simulation:

Study 2. The data is generated from the following model:

$$Y = \beta_1^\tau \mathbf{X} + a(\beta_2^\tau \mathbf{X})^2 + \epsilon, \tag{4.8}$$

where $\beta_1 = (0.5, 0.5, 0.5, 0.5, 0, 0, 0, 0)^\tau$, $\beta_2 = (0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5)^\tau$. We also set

p to be 8. Thus, under the null, we have β_1 and under the alternatives, we have $B = (\beta_1, \beta_2)$.

We generate $\mathbf{X}$ and ϵ according to the same mechanism as in *Study 1*. That is, we generate the observations $\mathbf{x}_i, i = 1, 2, \cdots, n$ from multivariate normal distribution $N(0, \Sigma_j), j = 1, 2$ and ϵ from $N(0, 1)$ and the double exponential distribution $DE(0, \sqrt{2}/2)$ with density $f(x) = \sqrt{2}/2 \exp\{-\sqrt{2}|x|\}$ with mean zero and variance 1 respectively. To save space, we only consider T_n^{DEE} and T_n^{DEE*} in the following due to their well performance on size control and easy computation.

We present the empirical sizes of T_n^{DEE} and T_n^{DEE*} with different significance level $\alpha = 0.01, 0.05$ and 0.1 and different sample sizes $n = 50$ and 100 in Table 4.5. From this table, we can see clearly that the empirical sizes of T_n^{DEE*} can be very close to the nominal size in different situations. On the other hand, T_n^{DEE} generally overestimate the size for $\alpha = 0.01$ but underestimate for $\alpha = 0.1$. However, for significance level $\alpha = 0.05$, T_n^{DEE} can control the type I error very well. These findings are consistent with those found in study 1. From them, we can assert that for significance level $\alpha = 0.05$, T_n^{DEE} is a suitable choice.

We then show the empirical powers of T_n^{DEE} and T_n^{DEE*} in this study and compare with the bootstrapped version of Zheng (1996)'s test which is denoted as T_n^{ZH*}. Our preliminary simulations show that the original version of Zheng (1996)'s test generally underestimate the size. Thus we apply the bootstrapped version of Zheng (1996)'s test. The results are presented in Table 4.6. From this table, we can observe that when $X \sim N(0, \Sigma_1)$, T_n^{ZH*} can have very limited powers, on the other hand, our proposed tests T_n^{DEE} and T_n^{DEE*} performs very well. For instance, when $X \sim N(0, \Sigma_1)$ and $\epsilon \sim N(0, 1)$, the power T_n^{ZH*} is only 0.1330 with $n = 50$ and $a = 0.8$. However, at the same setting, the empirical powers of our test T_n^{DEE} and T_n^{DEE*} are 0.7510 and 0.7275 respectively. In sum, in this situation, Zheng (1996)'s test fails to detect the alternatives while our proposed test can still be very powerful. When $X \sim N(0, \Sigma_2)$, though the powers of T_n^{ZH*} are enlarged, they are still very limited.

From this numerical study, we can see clearly that our proposed tests can improve Zheng (1996)'s method greatly. From this simulation study, we can further conclude that T_n^{DEE} generally can have larger power compared with its bootstrapped version T_n^{DEE*}.

Recently, Stute and Zhu (2002) also considered dimension reduction tests for the correct specification of GLMs based on one-dimensional projected covariates. To be precise, they developed an innovation process transformation of the empirical process $n^{-1/2}\sum_{i=1}^{n}\Big(y_i - g(\hat{\beta}^\tau \mathbf{x}_i)\Big)I(\hat{\beta}^\tau \mathbf{x}_i \leq u)$. By introducing the transformation, they can yard asymptotically distribution-free test statistic, an attractive feature of their procedure. However, notice that their test only focuses on the direction β, the direction involved in the null hypothesis model and thus their approach is not ominous test and can fail or lose some power under certain alternatives. To compare with their test, we consider the following simulations.

Study 3. The data is generated from the following model:

$$Y = \beta_1^\tau \mathbf{X} + a(\beta_2^\tau \mathbf{X})^3 + \epsilon, \tag{4.9}$$

where $\beta_1 = (1,0,0)^\tau$, $\beta_2 = (0,1,0)^\tau$ for $p = 3$ and $\beta_1 = (1,1,0,0)^\tau/\sqrt{2}$, $\beta_2 = (0,0,1,1)^\tau/\sqrt{2}$ for $p = 4$. For $p = 3$, we consider sample size $n = 50, 100$ while for $p = 4$, we set $n = 100$. In both cases, we generate X and ϵ from multivariate standard normal distribution. Furthermore, consider $a = 0.0, 0.3, \cdots, 1.5$ in these several cases. Denote Stute and Zhu (2002)'s test as T_n^{SZ}. To save space, we only present the results of T_n^{DEE}. The results are presented in figure 4.2. From this figure, we can see clearly that our proposed test T_n^{DEE} performs better than T_n^{SZ} uniformly. Under certain alternatives, T_n^{SZ} can have very low powers. On the other hand, our proposed test can still detect the alternatives efficiently.

4.4.2 Real Data Analysis

This dataset is obtained from the Machine Learning Repository at the University of California-Irvine (http://archive.ics.uci.edu/ml/datasets/Auto+MPG). Recently, Xia (2007) analysed this data set by their method. The first analysis of this data set is duo to Quinlan (1993). There are 406 observations in the original data set. To illustrate our methods, we first clear the units with missing response and/or covariate and get 392 sample points. The response variable Y is miles per gallon (Y). There are other seven covariates: the number of cylinders (X_1), engine displacement (X_2), horsepower (X_3), vehicle weight (X_4), time to accelerate from 0 to 60 mph (X_5), model year (X_6) and origin of the car (1 = American, 2 = European, 3 = Japanese). Since the origin of the car contains more than two categories, we follow Xia (2007)'s suggestions and define two indictor variables. To be precise, let $X_7 = 1$ if a car is from America and 0 otherwise and $X_8 = 1$ if it is from Europe and 0 otherwise. For ease of explanation, all covariates are standardized separately. Quinlan (1993) aimed to predict the response in terms of the eight predictors $\mathbf{X} = (X_1, \cdots, X_8)^\top$. To achieve this goal, a simple linear regression model may be adopted. However, as argued in the introduction part, we need to check its adequacy to avoid model misspecification. We have $T_n^{DEE} = 86.5703$ and the p value is 0. Hence, we should not apply the linear regression model to predict the response. Moreover, the $\hat{d}$ is selected to be 1 by the BIC criterion for DEE. Thus a single index model may be applied. Our conclusion is also supported by the fig 4.3. Further $\widetilde{T}_n^{MAVE} = 98.2602$ and the p value is also 0. In sum, for this data sum, we can conclude that the linear regression model is not adequate.

4.5 Discussion

In this Chapter, we revisit Zheng (1996)'s model checking procedure and propose a new adaptive method to improve the power performance of Zheng (1996)'s method.

It's easy to extend our method to other local smoothing methods discussed in the introduction part and we expect that similar improvements and findings can be obtained.

We can also extend the basic idea to global smoothing method. To be precise, as discussed in section 4.2, under the null hypothesis $Y = g(\beta^\tau \mathbf{X}) + \epsilon$ with $E(\epsilon|\mathbf{X}) = 0$, we can have $E(\epsilon|\beta^\tau \mathbf{X}) = E(\epsilon|B^\tau \mathbf{X}) = 0$. While under the alternative H_1, $E(\epsilon|B^\tau \mathbf{X}) = E(Y - g(\beta^\tau \mathbf{X})|B^\tau \mathbf{X}) = G(B^\tau \mathbf{X}) - g(\beta^\tau \mathbf{X}) \neq 0$. This motivates us to define the following test statistics:

$$R_n(\mathbf{z}) = n^{-1/2} \sum_{i=1}^{n} \Big(y_i - g(\hat{\beta}^\tau \mathbf{x}_i) \Big) I(\hat{B}^\tau \mathbf{x}_i \leq \mathbf{z}).$$

Stute (1997) proposed the following residual marked empirical process test

$$R_n^*(\mathbf{x}) = n^{-1/2} \sum_{i=1}^{n} \Big(y_i - g(\hat{\beta}^\tau \mathbf{x}_i) \Big) I(\mathbf{x}_i \leq \mathbf{x}).$$

Compare with these two empirical process, it's easy to see that the difference is the use of $\hat{B}^\tau \mathbf{x}_i$. Different from $R_n^*(\mathbf{x})$, under the null hypothesis, $\hat{B}^\tau \mathbf{x}_i$ is one dimension random variable, while under the alternative, it's $\hat{q}$ dimensional random vectors. In other words, $R_n(\mathbf{z})$ is adaptive to the underling true model. Since we avoid to use the high dimensional process indexed by the p-dimension vector $\mathbf{x}$, we expect $R_n(\mathbf{z})$ can be much easier to compute and have more power in finite sample. A similar approach was introduced by Stute and Zhu (2002) by setting $\hat{B} = \hat{\beta}$ in $R_n(\mathbf{z})$ for generalized linear model. However, they did not investigate the power performance under local alternative and may be inconsistent for some kinds of alternatives as verified by our study 3 in the subsection 4.4.1.

Extensions of our methodology to missing, censored data and dependent data set can also be considered. Take the missing response as an example. We give some notations first. Let δ_i be the missing indicator, that's, $\delta_i = 1$ if y_i is observed , otherwise it's equal to zero. We also assume that the response is missing at random. This means $P(\delta = 1|\mathbf{X}, Y) = P(\delta = 1|\mathbf{X}) := \pi(\mathbf{X})$. For more details, see Little and Rubin (1987). Again, we consider to test whether the following regression model holds

or not. That's, $H_0 : Y = g(\beta^\tau \mathbf{X}) + \epsilon$ with $E(\epsilon|\mathbf{X}) = 0$ and Y is missing at random. Notice that under the null hypothesis $E(\delta\epsilon/\pi(\mathbf{X})|\beta^\tau\mathbf{X}) = E(\delta\epsilon/\pi(\mathbf{X})|B^\tau\mathbf{X}) = 0$, while under the alternative $E(\delta\epsilon/\pi(\mathbf{X})|B^\tau\mathbf{X}) \neq 0$. Similarly, we can construct a consistent test statistic with the following form:

$$V_{n1} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{\delta_i}{\hat{\pi}(\mathbf{x}_i)} \frac{\delta_j}{\hat{\pi}(\mathbf{x}_j)} \hat{\epsilon}_i \hat{\epsilon}_j K_h(\hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j)),$$

here $\hat{\pi}(\mathbf{x}_i)$ is an estimator, say, the nonparametric or parametric estimator, of $\pi(\mathbf{x}_i)$, $\hat{\epsilon}_i = y_i - g(\hat{\beta}^\tau \mathbf{x}_i)$ and $\hat{\beta}$ and $\hat{B}$ is obtained from the complete observed units. Another possible test statistic takes the following form

$$V_{n2} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \delta_i \delta_j \hat{\epsilon}_i \hat{\epsilon}_j K_h(\hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j)).$$

This corresponds to the test statistics obtained from the complete case. The direct extension of Zheng (1996)'s method to this situation has been discussed by Guo, Xu and Zhu (2014).

Also we can consider to apply the procedure in this chapter to other testing problem such as testing the homoscedasticity, testing whether the underling quantile regression model is of some parametric form and testing whether the underling conditional density of Y given $\mathbf{X}$ belongs to some specific parametric families. Below we take the quantile regression testing problem as an illustration. Denote the conditional distribution function Y given $\mathbf{X}$ by $F(y|\mathbf{x})$. The θ-th$(0 < \theta < 1)$ conditional quantile $Q_\theta(\mathbf{x})$ of y_i given $\mathbf{x}_i = \mathbf{x}$ is defined as

$$Q_\theta(\mathbf{x}) = \inf\{y : F(y|\mathbf{x}) \geq \theta\}.$$

If, as is assumed throughout, the conditional distribution function $F(y|\mathbf{x})$ is absolutely continuous in y for almost all $\mathbf{x}$, then $F(Q_\theta(\mathbf{x})|\mathbf{x}) = \theta$. We want to test whether the θth$(0 < \theta < 1)$ conditional quantile $Q_\theta(\mathbf{x})$ is a specific parametric model or not. That's,

$$H_0 : F(g(\beta^\tau \mathbf{X})|\mathbf{X}) = \theta,$$

here $g(\cdot)$ is a known function. Note that $F(g(\beta^\tau \mathbf{X})|\mathbf{X}) = \theta$ is equivalent to $E(I(y_i \leq g(\beta^\tau \mathbf{x}_i)|\mathbf{x}_i) = \theta$. Define $\epsilon_i = I(y_i \leq g(\beta^\tau \mathbf{x}_i) - \theta$, we can have $E(\epsilon|\mathbf{X}) = 0$ and also $E(\epsilon|\beta^\tau \mathbf{X}) = E(\epsilon|B^\tau \mathbf{X}) = 0$. While under the alternative, $E(\epsilon|B^\tau \mathbf{X}) \neq 0$. Thus after we set $\hat{\epsilon}_i = I(y_i \leq g(\hat{\beta}^\tau \mathbf{x}_i) - \theta$, we can use the statistics defined in (4.4) to check the null hypothesis. Here $\hat{\beta}$ can be obtained from classical quantile estimation methods.

In sum, our methodology is an unified and powerful approach. It can be readily extended to many different situations. We leave these aspects to further studies.

4.6 Appendix. Proof of the Theorems

The following conditions are assumed for the theorems in Section 4.3.

1) $\sup E(X_l^2|B^\tau \mathbf{X}) < \infty, l = 1, \cdots, p$; $E(\eta^2|B^\tau \mathbf{X}) < \infty$, $\sup g'(\beta^\tau \mathbf{X}) < \infty$ and $\sup G^2(B^\tau \mathbf{X}) < \infty$.

2) $nh^2 \to \infty$;

3) The density $f(B^\tau \mathbf{X})$ of $B^\tau \mathbf{X}$ on support $\mathcal{C}$ exists and has 2 bounded derivatives and satisfies

$$0 < \inf_{B^\tau \mathbf{X} \in \mathcal{C}} f(B^\tau \mathbf{X}) \leq \sup_{B^\tau \mathbf{X} \in \mathcal{C}} f(B^\tau \mathbf{X}) < \infty;$$

4) $K(\cdot)$ is a spherical symmetric density function with a bounded derivative and support. All the moments of $K(\cdot)$ exist and $\int UU^\top K(U)dU = I$.

5) $\Sigma_x = E(g'^2(\beta^\tau \mathbf{X})\mathbf{X}\mathbf{X}^\tau)$ is positive definite.

Remark 4.2. *The conditions 1) are necessary for the convergence rate of the least squares estimate $\hat{\beta}$. Condition 2) is needed for the asymptotic normality of our statistic. In Condition 2), $nh^2 \to \infty$ is an usual assumption in nonparametric smoothing. To our surprise, Condition 3) is not necessary but aims to avoid tedious proofs.*

Without this condition, we have to resort to some truncation technique since the denominators in the statistic may be close to zero. Condition 4) is a typical condition for non-parametric estimation and is assumed by Xia et al. (2002).

Lemma 4.1. *Let $\mathcal{M}(t)$ be a $p\times p$ positive semi-definite matrix such that span$\{\mathcal{M}(t)\} = \mathcal{S}_{Z(t)|\mathbf{X}}$. We can have span$\{E\{\mathcal{M}(T)\}\}$=span$\{E\{\mathcal{M}(T)\rho(T)\}\}$, here $\rho(\cdot) > 0$ is some weight function.*

Proof of Lemma 4.1. Let $\mathbf{v}\perp$ span$\{E\{\mathcal{M}(T)\}\}$, we can have $0 = E\{\mathbf{v}^\tau\mathcal{M}(T)\mathbf{v}\}$. Due to the fact that $\mathcal{M}(t)$ is semi-definite matrix for any t, we can obtain that $\mathcal{M}(t)\rho(t)\mathbf{v} = 0$. In other words, $\mathbf{v}\perp$ span$\{\mathcal{M}(t)\rho(t)\}$ for any t. Further, we can get $E\{\mathcal{M}(T)\rho(T)\mathbf{v}\} = 0$. Thus $\mathbf{v}\perp$ span$\{E\{\mathcal{M}(T)\rho(T)\}\}$. This follows that span$\{E\{\mathcal{M}(T)\rho(T)\}\} \subseteq$ span$\{E\{\mathcal{M}(T)\}\}$. Another direction can be similarly shown. We conclude that span$\{E\{\mathcal{M}(T)\}\}$=span $\{E\{\mathcal{M}(T)\rho(T)\}\}$.

Lemma 4.2. *Under the assumptions in the Appendix and under the local alternative (4.6), we can have $\hat{d} \to 1$. Here $\hat{d}$ is selected by BIC criterion for DEE or MAVE.*

Proof of Lemma 4.2. We first investigate the $\hat{d}$ selected by BIC criterion for DEE. We also adopt the same conditions as those in Theorem 4 of Zhu, Wang, Zhu and Ferré (2010).

We turn to consider DEE_{SIR} first. From the argument used for proving theorem 3.2 of Li et al. (2008), we can know that to obtain that $\mathcal{M}_{n,n} - \mathcal{M} = O_p(C_n)$, we only need to show that $\mathcal{M}_n(t) - \mathcal{M}(t) = O_p(C_n)$ uniformly. Now we investigate the term $\mathcal{M}_n(t)$ for every t. Here $\mathcal{M}(t) = \Sigma^{-1}Var(E(\mathbf{X}|Z(t))) = \Sigma^{-1}(\mu_1-\mu_0)(\mu_1-\mu_0)^\tau p(1-p)$ here $\mu_j = E(\mathbf{X}|Z(t) = j)$ with $j = 0$ and 1 and $p = E(I(Y \le t))$. Further note that

$$\begin{aligned}\mu_1 - \mu_0 &= \frac{E(\mathbf{X}I(Y \le t))}{p} - \frac{E(\mathbf{X}I(Y > t))}{1-p} \\ &= \frac{E(\mathbf{X}I(Y \le t)) - E(\mathbf{X})E(I(Y \le t))}{p(1-p)}.\end{aligned}$$

Thus from our Lemma 4.1, the $\mathcal{M}(t)$ can also be taken to be

$$\mathcal{M}(t) = \Sigma^{-1}\Big[E\{(\mathbf{X} - E(\mathbf{X}))I(Y \le t)\}\Big]\Big[E\{(\mathbf{X} - E(\mathbf{X}))I(Y \le t)\}\Big]^\tau.$$

For ease of illustration, we denote $m(t) = E\{(\mathbf{X} - E(\mathbf{X}))I(Y \leq t)\}$, then $\mathcal{M}(t) = \Sigma^{-1}m(t)m(t)^{\tau} = \Sigma^{-1}L(t)$. We can estimate $m(t)$ by

$$m_n(t) =: n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})I(y_i \leq t).$$

Thus, $\mathcal{M}_n(t)$ can be taken to be

$$\mathcal{M}_n(t) = \hat{\Sigma}^{-1}m_n(t)m_n^{\tau}(t) = \hat{\Sigma}^{-1}L_n(t).$$

Denote the response under the null and local alternative as Y and Y_n respectively. Note that

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i I(y_{in} \leq t) - E(\mathbf{X}I(Y \leq t)) \\ = \; & \frac{1}{n}\sum_{i=1}^{n}[\mathbf{x}_i I(y_{in} \leq t) - E(\mathbf{X}I(Y_n \leq t))] + E(\mathbf{X}I(Y_n \leq t)) - E(\mathbf{X}I(Y \leq t)). \end{aligned}$$

By the Lindeberg-Levy central limit theorem, the first term has order $O_p(n^{-1/2})$. Now we consider the second term. Note that

$$E(\mathbf{X}I(Y_n \leq t)) - E(\mathbf{X}I(Y \leq t)) \;=\; E\Big(\mathbf{X}[P(Y_n \leq t|\mathbf{X}) - P(Y \leq t|\mathbf{X})]\Big).$$

Recall that $Y_n = Y + C_n G(B^{\tau}\mathbf{X})$ and denote the conditional density and conditional distribution function of Y given $\mathbf{X}$ as $f_{Y|\mathbf{X}}(\cdot)$ and $F_{Y|\mathbf{X}}(\cdot)$ respectively, thus we can have

$$\begin{aligned} P(Y_n \leq t|\mathbf{X}) - P(Y \leq t|\mathbf{X}) &= F_{Y|\mathbf{X}}(t - C_n G(B^{\tau}\mathbf{X})) - F_{Y|\mathbf{X}}(t) \\ &= -C_n G(B^{\tau}\mathbf{X}) f_{Y|\mathbf{X}}(t) + O_p(C_n). \end{aligned}$$

From this, we can conclude that $n^{-1}\sum_{i=1}^{n}\mathbf{x}_i I(y_{in} \leq t) - E(\mathbf{X}I(Y \leq t)) = O_p(C_n)$. Similarly, we can show that $m_n(t) - m(t) = O_p(C_n), L_n(t) - L(t) = O_p(C_n)$ and $\mathcal{M}_n(t) - \mathcal{M}(t) = O_p(C_n)$. Finally, similar to the argument used for proving theorem 3.2 of Li et al. (2008), we can obtain that $\mathcal{M}_{n,n} - \mathcal{M} = O_p(C_n)$. Now we turn to prove our Lemma for the BIC criterion for DEE_{SIR}. Note that for $l > 1$, we can have

$$G(1) - G(l) = D_n \frac{l(l+1)-2}{p} - \frac{n\sum_{i=2}^{l} \log(\hat{\lambda}_i + 1) - \hat{\lambda}_i}{2\sum_{i=1}^{p} \log(\hat{\lambda}_i + 1) - \hat{\lambda}_i}.$$

From $\mathcal{M}_{n,n} - \mathcal{M} = O_p(C_n)$, we can know that $\hat{\lambda}_i - \lambda_i = O_p(C_n)$. Note that $\log(\hat{\lambda}_i + 1) - \hat{\lambda}_i = -\hat{\lambda}_i^2 + o_p(\hat{\lambda}_i^2)$ and $\lambda_i = 0$ for any $i > 1$. We can obtain that $\sum_{i=2}^{l} \log(\hat{\lambda}_i + 1) - \hat{\lambda}_i = O_p(C_n^2)$ and $\sum_{i=1}^{p} \log(\hat{\lambda}_i + 1) - \hat{\lambda}_i \to b$ in probability for some $b < 0$.

Take $D_n = n^{1/2}$ and $C_n = n^{-1/2}h^{-1/4}$, we can have

$$\frac{nC_n^2}{D_n} = (nh)^{-1/2} \to 0.$$

Since $l(l+1) > 2$ for any $l > 1$, we can conclude that $P(G(1) > G(l)) \to 1$. In other words, we can have $P(\hat{d} = 1) \to 1$.

Now we investigate the properties of DEE_{SVAE}. Similar to DEE_{SIR}, we need to study the term $\mathcal{M}_n(t)$ for every t. For SAVE, the $\mathcal{M}(t)$ takes the following form: $\mathcal{M}(t) = \Sigma^{-1}E\{\Sigma - Var(\mathbf{X}|Z(t))\}^2$. Further note that

$$\Sigma = E[Var(\mathbf{X}|Z(t))] + Var(E(\mathbf{X}|Z(t))).$$

Thus the term $E\{\Sigma - Var(\mathbf{X}|Z(t))\}^2$ can be rewritten as

$$\Sigma Var(E(\mathbf{X}|Z(t))) - \Sigma^2 + Var(E(\mathbf{X}|Z(t)))\Sigma + E[Var^2(\mathbf{X}|Z(t))].$$

From the argument for DEE_{SIR}, we have studied the distance between the empirical version and the population version of $Var(E(\mathbf{X}|Z(t)))$, which is on the order of $O_p(C_n)$. Thus in the following, we only need to focus on the term $E[Var^2(\mathbf{X}|Z(t))]$. Notice that

$$\begin{aligned} E[Var^2(\mathbf{X}|Z(t))] &= pVar^2(\mathbf{X}|1) + (1-p)Var^2(\mathbf{X}|0) \\ Var^2(\mathbf{X}|1) &= E(\mathbf{X}\mathbf{X}^\tau|1) - E(\mathbf{X}|1)E(\mathbf{X}^\tau|1) \\ &= \{E(\mathbf{X}\mathbf{X}^\tau I(Y \le t))E(I(Y \le t)) - E(\mathbf{X}I(Y \le t))E(\mathbf{X}^\tau I(Y \le t))\}p^{-2}. \end{aligned}$$

From the proof for DEE_{SIR}, we conclude that $n^{-1}\sum_{i=1}^{n}\mathbf{x}_i I(y_{in}\le t)-E(\mathbf{X}I(Y\le t))=O_p(C_n)$. Here, same to the notations in the proof for DEE_{SIR}, denote the response under the null and local alternative as Y and Y_n respectively. Similarly, we can easily show that $n^{-1}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\tau}I(y_{in}\le t)-E(\mathbf{X}\mathbf{X}^{\tau}I(Y\le t))=O_p(C_n)$ and $n^{-1}\sum_{i=1}^{n}I(y_{in}\le t)-E(I(Y\le t))=O_p(C_n)$. These results can yard that the distance between the empirical version and the population version of $E[Var^2(\mathbf{X}|Z(t))]$ is also on the order of $O_p(C_n)$. Finally, we can obtain that $\mathcal{M}_n(t)-\mathcal{M}(t)=O_p(C_n)$ and $\mathcal{M}_{n,n}-\mathcal{M}=O_p(C_n)$. In the end, following the same line, we can get that $P(\hat{d}=1)\to 1$ by using BIC criterion for DEE to select d.

We now turn to consider the $\hat{d}$ being selected by BIC criterion for MAVE. We use the same assumptions as Wang and Yin (2008), and for simplicity we assume that $\mathbf{X}$ has a compact support over which its density is positive. Recall the definition of B_k, we can know that $Y=E(Y|B_k^{\tau}\mathbf{X})+\epsilon$, where $E(\epsilon|B_k^{\tau}\mathbf{X})=0$. Suppose that the orthogonal $p\times k$ matrix $\hat{B}_k$ is the sample estimator of B_k.

Follow Wang and Yin (2008), we can have

$$\frac{1}{n}RSS_k=E\{Y-E(Y|B_k^{\tau}\mathbf{X})\}^2+o_p(1).$$

Consequently, we can have the following results:

$$\begin{aligned}\frac{RSS_k-RSS_1}{n} &= E\{Y-E(Y|B_k^{\tau}\mathbf{X})\}^2-E\{Y-E(Y|\beta^{\tau}\mathbf{X})\}^2+o_p(1)\\ &= E\{-2YE(Y|B_k^{\tau}\mathbf{X})+E^2(Y|B_k^{\tau}\mathbf{X})\\ &\quad +2YE(Y|\beta^{\tau}\mathbf{X})-E^2(Y|\beta^{\tau}\mathbf{X})\}+o_p(1).\end{aligned}$$

Note that under local alternative (4.6), we can have

$$\begin{aligned}E(Y|\mathbf{X}) &= g(\beta^{\tau}\mathbf{X})+C_nG(B^{\tau}\mathbf{X})\\ &= E(Y|\beta^{\tau}\mathbf{X})+C_n\left(G(B^{\tau}\mathbf{X})-E\{G(B^{\tau}\mathbf{X})|\beta^{\tau}\mathbf{X}\}\right)\\ &= E(Y|\beta^{\tau}\mathbf{X})+o_p(1).\end{aligned}$$

Further note that

$$E\left(YE(Y|\beta^\tau \mathbf{X})\right) = E\left(E^2(Y|\beta^\tau \mathbf{X})\right);$$
$$E\left(YE(Y|B_k^\tau \mathbf{X})\right) = E\left(E(Y|B_k^\tau \mathbf{X})E(Y|\mathbf{X})\right) = E\left(E(Y|B_k^\tau \mathbf{X})E(Y|\beta^\tau \mathbf{X})\right) + o_p(1).$$

Thus, we obtain that

$$\frac{RSS_k - RSS_1}{n} = E\{E(Y|B_k^\tau \mathbf{X}) - E(Y|\beta^\tau \mathbf{X})\}^2 + o_p(1) \geq 0.$$

Recall that the BIC criterion is $BIC_k = \log(RSS_k/n) + \log(n)k/(nh^k)$. Because $\log(n)k/(nh^k) \to 0$, we thus have $BIC_k \geq BIC_1$. Hence, for large n, $\hat{d} \to 1$.

In the end, we remark that since $\hat{d}$ can only take discrete values, we can further conclude that $P(\hat{d} = 1) \to 1$.

Lemma 4.3. *Under the null hypothesis and conditions 1a)-4), we have*

$$W_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} K_h(\hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j))\epsilon_i M(x_{jl}) = O_P(1/\sqrt{n}), l = 1, \cdots, p.$$

where $M(\cdot)$ is continuously differentiable and $E(M^2(X_l)|B^\tau \mathbf{X}) \leq b(B^\tau \mathbf{X})$ for $X_l \in R$ and $E[b(B^\tau \mathbf{X})] < \infty$.

Proof of Lemma 4.3. For natational convenience, we denote $B_{ij} = B^\tau(\mathbf{x}_i - \mathbf{x}_j)$ and $\hat{B}_{ij} = \hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j)$. Define $\tilde{W}_n$ as follows:

$$\tilde{W}_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{1}{h} K(\frac{\hat{B}_{ij}}{h})\epsilon_i M(x_{jl}) = \frac{h^{\hat{d}}}{h} W_n.$$

Since $\hat{d} \to 1$ in probability, to prove Lemma 3, we only need to show $\tilde{W}_n = O_p(1/\sqrt{n})$. For $\tilde{W}_n$, we can have:

$$\begin{aligned}
\tilde{W}_n &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{1}{h} K(\frac{B_{ij}}{h})\epsilon_i M(x_{jl}) \\
&\quad + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{1}{h} \left(K(\frac{\hat{B}_{ij}}{h}) - K(\frac{B_{ij}}{h}) \right) \epsilon_i M(x_{jl}) \\
&= W_{n1} + W_{n2}.
\end{aligned}$$

Let $\mathbf{t}_i = (y_i, \mathbf{x}_i^\tau)^\tau$, then W_{n1} can be written in a U-statistic form with kernel,

$$H_n(\mathbf{t}_i, \mathbf{t}_j) = \frac{1}{2h} K(\frac{B_{ij}}{h})[\epsilon_i M(x_{jl}) + \epsilon_j M(x_{il})].$$

To apply the theory for non-degenerate U-statistic, we need to show $E[H_n^2(\mathbf{t}_i, \mathbf{t}_j)] = o(n)$. Let $\mathbf{Z} = B^\tau \mathbf{X}$, it can be verified that

$$\begin{aligned}
& E[H_n^2(\mathbf{t}_i, \mathbf{t}_j)] \\
\leq\ & 2E\left[\frac{1}{2h} K(\frac{B_{ij}}{h})\epsilon_i M(x_{jl})\right]^2 + 2E\left[\frac{1}{2h} K(\frac{B_{ij}}{h})\epsilon_j M(x_{il})\right]^2 \\
=\ & \int \frac{1}{h^2}\sigma^2(z_i) E(M^2(x_{jl})|z_j) K^2(\frac{z_i - z_j}{h}) f(z_i) f(z_j) dz_i dz_j \\
\leq\ & \int \frac{1}{h}\sigma^2(z_i) b(z_i - hu) K^2(u) f(z_i) f(z_i - hu) dz_i du \\
=\ & \int \frac{1}{h}\sigma^2(z) b(z) f^2(z) dz \cdot \int K^2(u) du + o(1/h) \\
=\ & O(1/h) = o(n).
\end{aligned}$$

For $H_n(\mathbf{t}_i, \mathbf{t}_j)$, since $E(\epsilon|\mathbf{X}) = 0$, it can be derived that $E(H_n(\mathbf{t}_i, \mathbf{t}_j)) = 0$. Now, we investigate the conditional expectation of $H_n(\mathbf{t}_i, \mathbf{t}_j)$. It can be proved that

$$\begin{aligned}
r_n(\mathbf{t}_i) &= E(H_n(\mathbf{t}_i, \mathbf{t}_j)|\mathbf{t}_i) = \frac{\epsilon_i}{2h} E\left(K(\frac{B^\tau(\mathbf{x}_i - \mathbf{X})}{h}) M(X_l)\right) \\
&= \frac{\epsilon_i}{2h} E\left(K(\frac{z_i - Z}{h}) E(M(X_l)|Z)\right) = \frac{\epsilon_i}{2} \int E(M(X_l)|z_i + hu) f(z_i + hu) K(u) du \\
&= \frac{\epsilon_i f(z_i) E(M(X_l)|z_i)}{2} + l_n(\mathbf{t}_i).
\end{aligned}$$

Denote $\hat{W}_n$ as the "projection" of the statistic W_{n1}. We have

$$\begin{aligned}
\sqrt{n}\hat{W}_n &= \frac{2}{\sqrt{n}} \sum_{i=1}^n r_n(\mathbf{t}_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \epsilon_i f(z_i) E(M(X_l)|z_i) + \frac{2}{\sqrt{n}} \sum_{i=1}^n l_n(\mathbf{t}_i) \\
&= O_P(1).
\end{aligned}$$

The last equation follows from the fact that $E(l_n^2(\mathbf{t}_i)) = O(h^2) \to 0$ due to the Lipschitz condition for the function $E(M(X_l)|\cdot)f(\cdot)$. As a result, we have $W_{n1} = O_P(\hat{W}_n) = O_P(1/\sqrt{n})$. Denote

$$W_{n2}^* = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i}^n \frac{1}{h} K'(\frac{B_{ij}}{h})(\mathbf{x}_i - \mathbf{x}_j)^\tau \epsilon_i M(X_{jl}) \times \frac{\hat{B} - B}{h}.$$

Then for the term W_{n2}, we have

$$W_{n2} \;=\; W_{n2}^* + o_P(W_{n2}^*).$$

As $||\hat{B} - B||_2 = O_P(1/\sqrt{n})$, and under the condition $1/nh^2 \to 0$ using a similar arguments for W_{n1}, we can obtain that $W_{n2} = o_P(1/\sqrt{n})$. Thus we can conclude that $W_n = O_P(1/\sqrt{n})$. The proof is completed. □

Before we establish the asymptotic theory of our statistic under the null and local alternatives, we develop the following lemma about the asymptotic property of $\hat{\beta}$. This is necessary because it is defined under the null hypothesis.

Lemma 4.4. *Under the local alternative and conditions 1), 5), we have*

$$\begin{aligned}\sqrt{n}(\hat{\beta} - \beta) \;&=\; \Sigma_x^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g'(\beta^\tau \mathbf{x}_i)\mathbf{x}_i\eta_i + \Sigma_x^{-1}C_n\sqrt{n}E(g'(\beta^\tau \mathbf{X})\mathbf{X}G(B^\tau \mathbf{X})) \\ &\quad +(C^{-1} - \Sigma_x^{-1})C_n\sqrt{n}E(g'(\beta^\tau \mathbf{X})\mathbf{X}G(B^\tau \mathbf{X})) + o_p(1).\end{aligned}$$

Here $C = \sum_{i=1}^{n} g'^2(\beta^\tau \mathbf{x}_i)\mathbf{x}_i\mathbf{x}_i^\tau/n$.

Proof of Lemma 4.4. Under the regularity conditions designed in Jennrich (1969), similar to the derivation of Theorems 6 and 7 in Jennrich (1969), $\hat{\beta}$ is a strongly consistent estimate of β. If we let $D_n = C^{-1}\sum_{i=1}^{n} g'(\beta^\tau \mathbf{x}_i)\mathbf{x}_i(y_i - g(\beta^\tau \mathbf{x}_i))/n$, we can further have:

$$\begin{aligned}\hat{\beta} - \beta \;&=\; C^{-1}\frac{1}{n}\sum_{i=1}^{n} g'(\beta^\tau \mathbf{x}_i)\mathbf{x}_i(y_i - g(\beta^\tau \mathbf{x}_i)) + o_P(D_n) \\ &=\; (C^{-1} - \Sigma_x^{-1})\frac{1}{n}\sum_{i=1}^{n} g'(\beta^\tau \mathbf{x}_i)\mathbf{x}_i\eta_i \\ &\quad +(C^{-1} - \Sigma_x^{-1})\frac{1}{n}\sum_{i=1}^{n} g'(\beta^\tau \mathbf{x}_i)\mathbf{x}_i C_n G(B^\tau \mathbf{x}_i) \\ &\quad +\Sigma_x^{-1}\frac{1}{n}\sum_{i=1}^{n} g'(\beta^\tau \mathbf{x}_i)\mathbf{x}_i\eta_i \\ &\quad +\Sigma_x^{-1}\frac{1}{n}\sum_{i=1}^{n} g'(\beta^\tau \mathbf{x}_i)\mathbf{x}_i C_n G(B^\tau \mathbf{x}_i) + o_P(D_n) \\ &=:\; \sum_{i=1}^{4} I_{ni} + o_P(D_n). \qquad (4.10)\end{aligned}$$

Due to the consistency of C for Σ_x, we can easily conclude that

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta) &= \Sigma_x^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g'(\beta^\tau \mathbf{x}_i)\mathbf{x}_i\eta_i + \Sigma_x^{-1}C_n\sqrt{n}E(g'(\beta^\tau\mathbf{X})\mathbf{X}G(B^\tau\mathbf{X})) \\ &\quad +(C^{-1}-\Sigma_x^{-1})C_n\sqrt{n}E(g'(\beta^\tau\mathbf{X})\mathbf{X}G(B^\tau\mathbf{X}))+o_p(1).\end{aligned}$$

Now we turn to prove Theorem 4.1 below.

Proof of Theorem 4.1. First, V_n can be decomposed as, noting the symmetry of $K_h(\cdot)$,

$$\begin{aligned}V_n &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K_h(\hat{B}_{ij})\epsilon_i\epsilon_j \\ &\quad -\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K_h(\hat{B}_{ij})\epsilon_i g'(\beta^\tau\mathbf{x}_j)\mathbf{x}_j^\tau(\hat{\beta}-\beta) \\ &\quad +(\hat{\beta}-\beta)^\tau\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K_h(\hat{B}_{ij})g'(\beta^\tau\mathbf{x}_j)g'(\beta^\tau\mathbf{x}_i)\mathbf{x}_j\mathbf{x}_i^\tau(\hat{\beta}-\beta) \\ &\quad +o_P(V_n^*) \\ &=: V_{n1}-V_{n2}+V_{n3}+o_P(V_n^*), \qquad (4.11)\end{aligned}$$

where V_n^* denotes the term $V_{n1}-V_{n2}+V_{n3}$.

Consider the term V_{n2}. Under the conditions designed for Theorem 4.1, and from Lemmas 4.3 and 4.4, we can get that $V_{n2}=O_P(1/n)$. This yields that $nh^{1/2}V_{n2}=o_P(1)$.

Now we deal with the term V_{n3}. Rewrite it as

$$V_{n3} = (\hat{\beta}-\beta)^\tau\cdot\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K_h(\hat{B}_{ij})g'(\beta^\tau\mathbf{x}_j)g'(\beta^\tau\mathbf{x}_i)\mathbf{x}_j\mathbf{x}_i^\tau\cdot(\hat{\beta}-\beta).$$

A similar argument for proving Lemma 4.3 can be used to derive that

$$\begin{aligned}&\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K_h(\hat{B}_{ij})g'(\beta^\tau\mathbf{x}_j)g'(\beta^\tau\mathbf{x}_i)\mathbf{x}_j\mathbf{x}_i^\tau \\ =\ & E(g'^2(\beta^\tau\mathbf{X})\mathbf{X}\mathbf{X}^\tau f(B^\tau\mathbf{X}))+o_P(1).\end{aligned}$$

By the rate of $\hat{\beta}-\beta$, $V_{n3}=O_P(1/n)$. Consequently, $nh^{1/2}V_{n3}=o_P(1)$.

Finally, deal with the term V_{n1}. Note that we have the following decomposition:

$$\begin{aligned} V_{n1} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K_h(B_{ij})\epsilon_i\epsilon_j \\ &\quad + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(K_h(\hat{B}_{ij}) - K_h(B_{ij}))\epsilon_i\epsilon_j \\ &=: V_{n1,1} + V_{n1,2}. \end{aligned}$$

For the term $V_{n1,1}$, since in this chapter, we always assume that the dimension of $B^\tau\mathbf{X}$ is fixed, it is an U-statistic. Note that under the null hypothesis, $d = 1$ and $\hat{d} \to 1$. It is easy to derive the asymptotic normality: $nh^{1/2}V_{n1,1} \to N(0, var)$. Here

$$var = 2\int K^2(u)du \cdot \int (\sigma^2(\mathbf{z}))^2 f^2(\mathbf{z})d\mathbf{z}$$

with $\mathbf{Z} = B^\tau\mathbf{X}, \sigma^2(\mathbf{z}) = E(\epsilon^2|\mathbf{Z} = \mathbf{z})$. See a similar argument as that for Lemma 3.3 a) in Zheng(1996). We then omit the details.

Denote

$$V^*_{n1,2} = \frac{h}{h^{\hat{d}}} \cdot \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{1}{h}K'(\frac{B_{ij}}{h})(\mathbf{x}_i - \mathbf{x}_j)^\tau\epsilon_i\epsilon_j \cdot \frac{\hat{B} - B}{h}.$$

An application of Taylor expansion yields

$$V_{n1,2} = V^*_{n1,2} + o_P(V^*_{n1,2}).$$

Using a similar argument for the term $V_{n1,1}$ above, together with $||\hat{B} - B||_2 = O_P(1/\sqrt{n})$ and $1/nh^2 \to 0$, it results in that $nh^{1/2}V^*_{n1,2} = o_P(1)$. Thus we can have $nh^{1/2}V_{n1} \to N(0, var)$.

Combining the above results for the terms $V_{ni}, i = 1, 2, 3$, we conclude that

$$nh^{1/2}V_n \Rightarrow N(0, var).$$

Since var is actually unknown, an estimate is defined as

$$\widehat{var} = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{1}{h^{\hat{d}}}K^2(\frac{\hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j)}{h})\hat{\epsilon}_i^2\hat{\epsilon}_j^2.$$

As the proof is rather straightforward, we then only give a very brief description. Since $\hat{\beta}$ is consistent under the null hypothesis, some elementary computations lead to an asymptotic presentation:

$$\widehat{var} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{1}{h^{\hat{d}}} K^2(\frac{\hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j)}{h}) \epsilon_i^2 \epsilon_j^2 + o_P(1).$$

Using a similar argument as that for Lemma 3, we get

$$\widehat{var} = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \frac{1}{h^{\hat{d}}} K^2(\frac{B^\tau(\mathbf{x}_i - \mathbf{x}_j)}{h}) \epsilon_i^2 \epsilon_j^2 + o_P(1).$$

The consistency will be derived by using U-statistic theory. The proof is finished. □

Proof of Theorem 4.2. Under the local alternatives, V_n follows the decomposition by Taylor expansion:

$$\begin{aligned} V_n &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} K_h(\hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j))(\eta_i + C_n G(B^\tau \mathbf{x}_i))(\eta_j + C_n G(B^\tau \mathbf{x}_j)) \\ &\quad - \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} K_h(\hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j))(\eta_i + C_n G(B^\tau \mathbf{x}_i)) g'(\beta^\tau \mathbf{x}_j) \mathbf{x}_j^\tau (\hat{\beta} - \beta) \\ &\quad + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} K_h(\hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j)) g'(\beta^\tau \mathbf{x}_j)(\hat{\beta} - \beta)^\tau g'(\beta^\tau \mathbf{x}_i) \mathbf{x}_j \mathbf{x}_i^\tau (\hat{\beta} - \beta) \\ &\quad + o_P(V_n^*) \\ &=: \overline{V}_{n1} - \overline{V}_{n2} + \overline{V}_{n3} + o_P(\overline{V}_n^*). \end{aligned} \tag{4.12}$$

where $\overline{V}_n^* = \overline{V}_{n1} - \overline{V}_{n2} + \overline{V}_{n3}$.

For the term $\overline{V}_{n2}$ in (4.12), it follows that

$$\begin{aligned} \overline{V}_{n2} &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} K_h(\hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j)) \eta_i g'(\beta^\tau \mathbf{x}_j) \mathbf{x}_j^\tau (\hat{\beta} - \beta) \\ &\quad + C_n \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} K_h(\hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j)) G(B^\tau \mathbf{x}_i) g'(\beta^\tau \mathbf{x}_j) \mathbf{x}_j^\tau (\hat{\beta} - \beta) \\ &= \overline{V}_{n2,1}(\hat{\beta} - \beta) + C_n \overline{V}_{n2,2}^\tau (\hat{\beta} - \beta). \end{aligned}$$

Based on the conclusion from Lemma 4.3, we have $\overline{V}_{n2,1} = O_p(n^{-1/2})$. It can also be proved that

$$\overline{V}_{n2,2} = E(G(B^\tau \mathbf{X}) g'(\beta^\tau \mathbf{X}) \mathbf{X} f(B^\tau \mathbf{X})) + o_p(1).$$

When $C_n = n^{-1/2}h^{-1/4}$, Lemma 4.4 implies that

$$\begin{aligned} nh^{1/2}\overline{V}_{n2} &= nh^{1/2}\Big[O_p(n^{-1/2})O_p(C_n) \\ &\quad +2C_n^2 E(G(B^\tau\mathbf{X})g'(\beta^\tau\mathbf{X})\mathbf{X}f(B^\tau\mathbf{X}))\Sigma_x^{-1}E(G(B^\tau\mathbf{X})g'(\beta^\tau\mathbf{X})\mathbf{X})\Big] \\ &= 2E(G(B^\tau\mathbf{X})g'(\beta^\tau\mathbf{X})\mathbf{X}f(B^\tau\mathbf{X}))\Sigma_x^{-1}E(G(B^\tau\mathbf{X})g'(\beta^\tau\mathbf{X})\mathbf{X}) + o_p(1). \end{aligned} \tag{4.13}$$

Now, we turn to consider the term $\overline{V}_{n3}$. We can have

$$\begin{aligned} \overline{V}_{n3} &= (\hat{\beta}-\beta)^\tau E[g'^2(\beta^\tau\mathbf{X})\mathbf{X}\mathbf{X}^\tau f(B^\tau\mathbf{X})](\hat{\beta}-\beta) + o_p(C_n^2) \\ &= C_n^2 E^\tau[G(B^\tau\mathbf{X})g'(\beta^\tau\mathbf{X})\mathbf{X}]\Sigma_x^{-1}E[g'^2(\beta^\tau\mathbf{X})\mathbf{X}\mathbf{X}^\tau f(B^\tau\mathbf{X})] \\ &\quad \times\Sigma_x^{-1}E[G(B^\tau\mathbf{X})g'(\beta^\tau\mathbf{X})\mathbf{X}] + o_p(C_n^2). \end{aligned}$$

As a result, when $C_n = n^{-1/2}h^{-1/4}$, we can obtain

$$\begin{aligned} nh^{1/2}\overline{V}_{n3} &= E^\tau[G(B^\tau\mathbf{X})g'(\beta^\tau\mathbf{X})\mathbf{X}]\Sigma_x^{-1}E[g'^2(\beta^\tau\mathbf{X})\mathbf{X}\mathbf{X}^\tau f(B^\tau\mathbf{X})] \\ &\quad \times\Sigma_x^{-1}E[G(B^\tau\mathbf{X})g'(\beta^\tau\mathbf{X})\mathbf{X}] + o_p(1). \end{aligned} \tag{4.14}$$

Now we investigate the term $\overline{V}_{n1}$ in (4.12), it can be decomposed as follows

$$\begin{aligned} \overline{V}_{n1} &= \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n K_h(\hat{B}^\tau(\mathbf{x}_i-\mathbf{x}_j))\eta_i\eta_j \\ &\quad +C_n\frac{2}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n K_h(\hat{B}^\tau(\mathbf{x}_i-\mathbf{x}_j))\eta_i G(B^\tau\mathbf{x}_j) \\ &\quad +C_n^2\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}^n K_h(\hat{B}^\tau(\mathbf{x}_i-\mathbf{x}_j))G(B^\tau\mathbf{x}_i)G(B^\tau\mathbf{x}_j) \\ &= \overline{V}_{n1,1} + C_n\overline{V}_{n1,2} + C_n^2\overline{V}_{n1,3}. \end{aligned}$$

From the proof for Theorem 4.1 and the conclusion of Lemma 4.3, we know that

$$\begin{aligned} \overline{V}_{n1,2} &= O_P(n^{-1/2}); \\ \overline{V}_{n1,3} &= E(G^2(B^\tau\mathbf{X})f(B^\tau\mathbf{X})) + o_P(1). \end{aligned}$$

Note that $nh^{1/2}\overline{V}_{n1,1} = nh^{\hat{d}/2}\overline{V}_{n1,1}\times h^{(1-\hat{d})/2}$. Since $nh^{\hat{d}/2}\overline{V}_{n1,1}\Rightarrow N(0,var)$, we can conclude that $nh^{1/2}\overline{V}_{n1,1}$. Consequently, if $C_n = n^{-1/2}h^{-1/4}$, we can obtain that

$$nh^{1/2}\overline{V}_{n1}\Rightarrow N(E(G^2(B^\tau\mathbf{X})f(B^\tau\mathbf{X})), var). \tag{4.15}$$

Combining equations (4.13),(4.14) and (4.15), we can have

$$nh^{1/2}V_n \Rightarrow N(\mu, var),$$

where

$$\begin{aligned}
\mu &= E[G^2(B^\tau \mathbf{X}) f(B^\tau \mathbf{X})] - 2E^\tau[H(\mathbf{X}) f(B^\tau \mathbf{X})]\Sigma_x^{-1} E[H(\mathbf{X})] \\
&\quad + E^\tau[H(\mathbf{X})]\Sigma_x^{-1} E[g'^2(\beta^\tau \mathbf{X})\mathbf{X}\mathbf{X}^\tau f(B^\tau \mathbf{X})]\Sigma_x^{-1} E[H(\mathbf{X})] \\
&= E\left[\Big(G(B^\tau \mathbf{X}) - g'(\beta^\tau \mathbf{X})\mathbf{X}^\tau \Sigma_x^{-1} E[H(\mathbf{X})]\Big)^2 f(B^\tau \mathbf{X})\right].
\end{aligned}$$

here $H(\mathbf{X}) = G(B^\tau \mathbf{X}) g'(\beta^\tau \mathbf{X})\mathbf{X}$.

Define T_n as the standardized version of V_n as follows:

$$T_n = \frac{h^{(1-\hat{d})/2} \sum_{i=1}^n \sum_{j\neq i}^n \hat{\epsilon}_i \hat{\epsilon}_j K(\frac{\hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j)}{h})}{\{2\sum_{i=1}^n \sum_{j\neq i}^n K^2(\frac{\hat{B}^\tau(\mathbf{x}_i - \mathbf{x}_j)}{h})\hat{\epsilon}_i^2 \hat{\epsilon}_j^2\}^{1/2}}.$$

Note that when $C_n = n^{-1/2}h^{-1/4}$, $\widehat{var}$ is still a consistent estimate of var. From Lemma 3, we can easily have that $\hat{\beta}$ is also consistent under the local alternative. Thus both $\hat{\beta}$ and $\widehat{var}$ are still consistent to β and var under the local alternatives. Thus $T_n^2 \Rightarrow \chi_1^2(\mu^2/var)$.

When C_n has a slower convergence rate than $n^{-1/2}h^{-1/4}$, the above proof can show that the test statistic goes to infinity in probability. We omit the details. Theorem 2 is proved. □

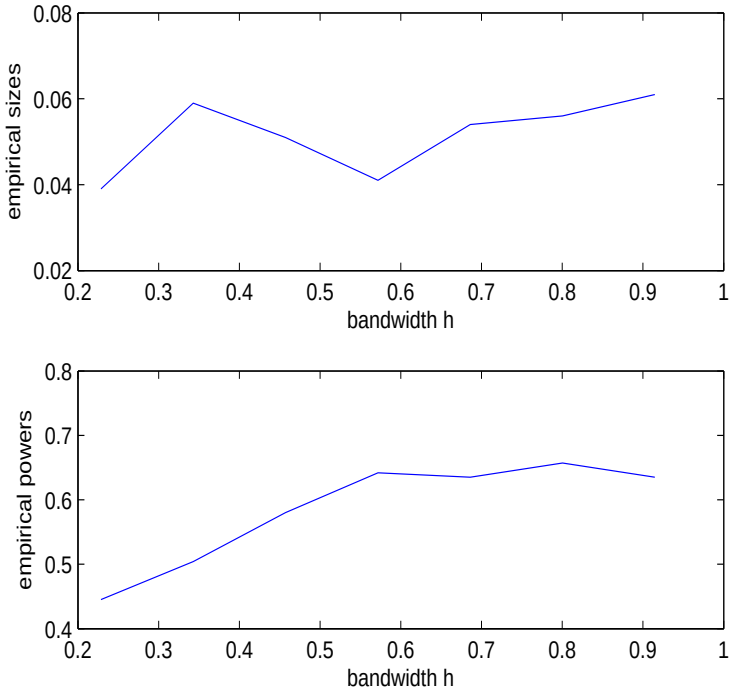

Figure 4.1: The empirical size and power curves of T_n^{DEE} against the bandwidth h with $X \sim N(0, \Sigma_1)$, $\epsilon \sim N(0, 1)$ and sample size 50 under different choices of a for study 1 with $a = 0$(the above panel) and $a = 1$(the below panel).

Table 4.1: Empirical sizes for H_0 with $\epsilon \sim N(0,1)$ and sample sizes $n = 50$ and 100.

	a	$\widetilde{T}_n^{MAVE}$		T_n^{DEE}	
		$n = 50$	$n = 100$	$n = 50$	$n = 100$
$X \sim N(0, \Sigma_1)$	0.01	0.0220	0.0203	0.0160	0.0160
	0.05	0.0517	0.0563	0.0480	0.0523
	0.10	0.0943	0.0970	0.0773	0.0777
$X \sim N(0, \Sigma_2)$	0.01	0.0170	0.0107	0.0170	0.0110
	0.05	0.0453	0.0503	0.0480	0.0510
	0.10	0.0760	0.0810	0.0800	0.0827
	a	T_n^{MAVE*}		T_n^{DEE*}	
		$n = 50$	$n = 100$	$n = 50$	$n = 100$
$X \sim N(0, \Sigma_1)$	0.01	0.0173	0.0173	0.0105	0.0090
	0.05	0.0753	0.0757	0.0450	0.0465
	0.10	0.1400	0.1297	0.0970	0.0965
$X \sim N(0, \Sigma_2)$	0.01	0.0117	0.0133	0.0135	0.0110
	0.05	0.0557	0.0587	0.0485	0.0510
	0.10	0.1110	0.1107	0.0940	0.1010

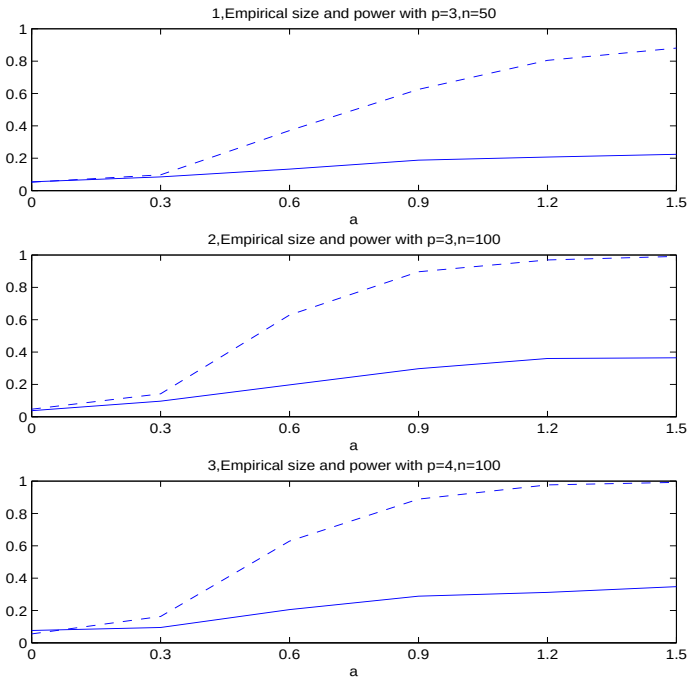

Figure 4.2: The empirical size and power curves of T_n^{SZ} and T_n^{DEE} in study 3. The solid and dash line represent the results from T_n^{SZ} and T_n^{DEE} respectively.

Table 4.2: Empirical sizes and powers of $\widetilde{T}_n^{MAVE}$ and T_n^{DEE} for H_0 vs. H_{11} and H_{12}, with $X \sim N(0, \Sigma_i)$, $i = 1, 2$ and $\epsilon \sim N(0, 1)$.

	a	$\widetilde{T}_n^{MAVE}$		T_n^{DEE}	
		$n = 50$	$n = 100$	$n = 50$	$n = 100$
$H_{11}, X \sim N(0, \Sigma_1)$	0	0.0630	0.0565	0.0470	0.0500
	0.2	0.0890	0.1535	0.0730	0.1263
	0.4	0.2175	0.4735	0.1623	0.3857
	0.6	0.4290	0.8125	0.3207	0.7227
	0.8	0.6470	0.9630	0.4910	0.9207
	1.0	0.8135	0.9990	0.6347	0.9803
$X \sim N(0, \Sigma_2)$	0	0.0460	0.0545	0.0480	0.0563
	0.2	0.0570	0.1050	0.0767	0.1173
	0.4	0.1165	0.3255	0.1647	0.3667
	0.6	0.2275	0.6530	0.3243	0.7203
	0.8	0.4100	0.8885	0.4953	0.9293
	1.0	0.5230	0.9690	0.6787	0.9887
$H_{12}, X \sim N(0, \Sigma_1)$	0	0.0535	0.0610	0.0490	0.0470
	0.2	0.1275	0.1845	0.0990	0.1687
	0.4	0.2950	0.5695	0.2657	0.5510
	0.6	0.5715	0.8895	0.5383	0.8980
	0.8	0.8100	0.9945	0.7763	0.9890
	1.0	0.9410	1.0000	0.9267	0.9993
$X \sim N(0, \Sigma_2)$	0	0.0395	0.0380	0.0460	0.0523
	0.2	0.0700	0.1160	0.0773	0.1140
	0.4	0.1545	0.3690	0.1820	0.3717
	0.6	0.3285	0.6920	0.3660	0.6953
	0.8	0.5490	0.9025	0.5723	0.9130
	1.0	0.7340	0.9805	0.7587	0.9857

Table 4.3: Empirical sizes and powers of T_n^{MAVE*} and T_n^{DEE*} for H_0 vs. H_{11} and H_{12}, with $X \sim N(0, \Sigma_i)$, $i = 1, 2$ and $\epsilon \sim N(0, 1)$.

	a	T_n^{MAVE*}		T_n^{DEE*}	
		$n = 50$	$n = 100$	$n = 50$	$n = 100$
$H_{11}, X \sim N(0, \Sigma_1)$	0	0.0697	0.0580	0.0470	0.0500
	0.2	0.1160	0.1890	0.0840	0.1425
	0.4	0.2740	0.5260	0.1635	0.3900
	0.6	0.4670	0.8510	0.3255	0.7235
	0.8	0.6750	0.9750	0.5115	0.9185
	1.0	0.8320	0.9960	0.6160	0.9780
$X \sim N(0, \Sigma_2)$	0	0.0490	0.0550	0.0500	0.0475
	0.2	0.0770	0.1160	0.0680	0.1285
	0.4	0.1600	0.3460	0.1515	0.3480
	0.6	0.2850	0.7170	0.3080	0.7155
	0.8	0.4520	0.8930	0.4835	0.9255
	1.0	0.5800	0.9740	0.6660	0.9820
$H_{12}, X \sim N(0, \Sigma_1)$	0	0.0770	0.0740	0.0435	0.0555
	0.2	0.1520	0.2070	0.1095	0.1680
	0.4	0.3320	0.5990	0.2545	0.5510
	0.6	0.6170	0.9200	0.5600	0.8975
	0.8	0.8410	0.9990	0.7690	0.9925
	1.0	0.9430	1.0000	0.9215	1.0000
$X \sim N(0, \Sigma_2)$	0	0.0570	0.058	0.0420	0.0540
	0.2	0.0990	0.1250	0.0705	0.1275
	0.4	0.2070	0.3560	0.1770	0.3495
	0.6	0.3710	0.7070	0.3380	0.6870
	0.8	0.5920	0.9150	0.5390	0.9190
	1.0	0.7600	0.9810	0.7445	0.9805

Table 4.4: Empirical sizes and powers for H_0 vs. H_{13}, with $X \sim N(0, \Sigma_i)$, $i = 1, 2$ and $\epsilon \sim N(0, 1)$.

	a	$\widetilde{T}_n^{MAVE}$		T_n^{DEE}	
		$n = 50$	$n = 100$	$n = 50$	$n = 100$
$X \sim N(0, \Sigma_1), \epsilon \sim N(0, 1)$	0	0.0515	0.0625	0.0507	0.0563
	0.2	0.1255	0.2245	0.1067	0.2000
	0.4	0.3465	0.7520	0.3330	0.7127
	0.6	0.6390	0.9790	0.6170	0.9580
	0.8	0.8335	0.9980	0.8023	0.9960
	1.0	0.9240	1.0000	0.8897	0.9993
$X \sim N(0, \Sigma_2), \epsilon \sim N(0, 1)$	0	0.0485	0.0505	0.0477	0.0480
	0.2	0.4160	0.8745	0.4163	0.8523
	0.4	0.8760	0.9995	0.8933	0.9993
	0.6	0.9515	1.0000	0.9680	0.9997
	0.8	0.9750	1.0000	0.9860	1.0000
	1.0	0.9765	1.0000	0.9907	1.0000
	a	T_n^{MAVE*}		T_n^{DEE*}	
		$n = 50$	$n = 100$	$n = 50$	$n = 100$
$X \sim N(0, \Sigma_1), \epsilon \sim N(0, 1)$	0	0.0767	0.0640	0.0490	0.0490
	0.2	0.1580	0.2570	0.1075	0.1835
	0.4	0.3860	0.7190	0.3120	0.6800
	0.6	0.6170	0.9560	0.5655	0.9410
	0.8	0.7940	0.9920	0.7400	0.9885
	1.0	0.8660	0.9930	0.8475	0.9930
$X \sim N(0, \Sigma_2), \epsilon \sim N(0, 1)$	0	0.0580	0.0435	0.0400	0.0475
	0.2	0.4170	0.8310	0.3925	0.7855
	0.4	0.7950	0.9880	0.7755	0.9870
	0.6	0.8770	0.9920	0.8985	0.9955
	0.8	0.8870	1.0000	0.9390	1.0000
	1.0	0.8880	1.0000	0.9445	1.0000

Table 4.5: Empirical sizes for T_n^{DEE} and T_n^{DEE*} in *Study 2* with sample sizes $n = 50$ and 100.

	a	T_n^{DEE}		T_n^{DEE*}	
		$n = 50$	$n = 100$	$n = 50$	$n = 100$
$X \sim N(0, \Sigma_1), \epsilon \sim N(0, 1)$	0.01	0.0177	0.0223	0.0100	0.0100
	0.05	0.0517	0.0557	0.0560	0.0510
	0.10	0.0797	0.0870	0.1095	0.1040
$X \sim N(0, \Sigma_1), \epsilon \sim DE(0, \sqrt{2}/2)$	0.01	0.0240	0.0240	0.0093	0.0097
	0.05	0.0553	0.0523	0.0500	0.0527
	0.10	0.0847	0.0827	0.1003	0.1073
$X \sim N(0, \Sigma_2), \epsilon \sim N(0, 1)$	0.01	0.0173	0.0223	0.0080	0.0103
	0.05	0.0497	0.0517	0.0503	0.0480
	0.10	0.0830	0.0860	0.1017	0.0940
$X \sim N(0, \Sigma_2), \epsilon \sim DE(0, \sqrt{2}/2)$	0.01	0.0193	0.0190	0.0090	0.0113
	0.05	0.0550	0.0500	0.0530	0.0523
	0.10	0.0860	0.0873	0.1043	0.0987

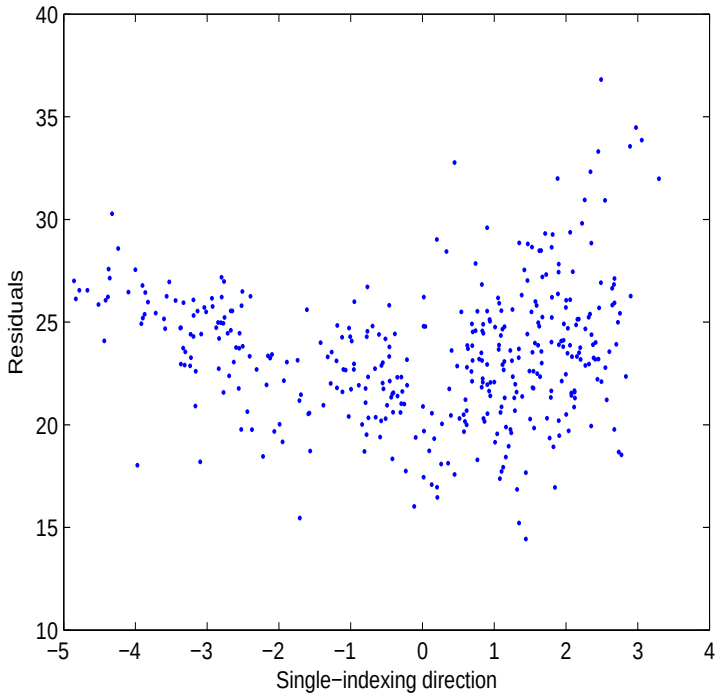

Figure 4.3: Plot of the residuals from the linear regression model against the single-indexing direction obtained from DEE.

Table 4.6: Empirical sizes and powers in *Study 2*, with $X \sim N(0, \Sigma_i)$, $i = 1, 2$ and $\epsilon \sim N(0, 1)$ or $DE(0, \sqrt{2}/2)$.

	a	T_n^{ZH*}		T_n^{DEE}		T_n^{DEE*}	
		$n = 50$	$n = 100$	$n = 50$	$n = 100$	$n = 50$	$n = 100$
$X \sim N(0, \Sigma_1), \epsilon \sim N(0, 1)$	0	0.0450	0.0415	0.0605	0.0460	0.0465	0.0495
	0.2	0.0500	0.0705	0.1400	0.2645	0.1345	0.2610
	0.4	0.0785	0.0930	0.3485	0.6990	0.3555	0.7145
	0.6	0.1035	0.1895	0.5905	0.9420	0.5555	0.9280
	0.8	0.1330	0.2770	0.7510	0.9855	0.7275	0.9890
	1.0	0.1630	0.3500	0.8475	0.9960	0.8170	0.9935
$X \sim N(0, \Sigma_1), \epsilon \sim DE(0, \sqrt{2}/2)$	0	0.0480	0.0430	0.0543	0.0517	0.0400	0.0540
	0.2	0.0655	0.0590	0.1463	0.2843	0.1395	0.2760
	0.4	0.0865	0.1240	0.3740	0.7240	0.3610	0.7365
	0.6	0.1295	0.2070	0.6073	0.9323	0.5990	0.9260
	0.8	0.1560	0.2925	0.7470	0.9873	0.7355	0.9860
	1.0	0.1880	0.3955	0.8510	0.9980	0.8170	0.9935
$X \sim N(0, \Sigma_2), \epsilon \sim N(0, 1)$	0	0.0480	0.0485	0.0463	0.0483	0.0450	0.0590
	0.2	0.0765	0.1660	0.3237	0.6443	0.2905	0.6595
	0.4	0.2245	0.4265	0.6970	0.9797	0.6780	0.9780
	0.6	0.3220	0.6570	0.8773	0.9993	0.8340	0.9980
	0.8	0.4290	0.7535	0.9237	0.9993	0.8985	0.9990
	1.0	0.4750	0.8020	0.9527	1.0000	0.9335	0.9990
$X \sim N(0, \Sigma_2), \epsilon \sim DE(0, \sqrt{2}/2)$	0	0.0460	0.0540	0.0507	0.0533	0.0475	0.0500
	0.2	0.1095	0.1915	0.3280	0.6583	0.3140	0.6600
	0.4	0.2370	0.4420	0.6977	0.9783	0.6975	0.9760
	0.6	0.3440	0.6495	0.8660	0.9993	0.8355	0.9985
	0.8	0.4285	0.7425	0.9250	0.9997	0.9110	1.0000
	1.0	0.4865	0.8095	0.9550	1.0000	0.9355	1.0000

Chapter 5

Dimension Reduction with Missing Response at Random

5.1 Introduction

In practice, as argued in Chapter 2 and Chapter 3, due to various reasons such as loss of information caused by uncontrollable factors, unwillingness of some sampled units to provide the desired information and so on, often not all responses are available. The most commonly used method to handle missing response problems simply resorts to the complete-case(CC) analysis by discarding all the incomplete measurements with missing values. However, this practice is undesirable since the resulting estimates are inconsistent unless the missing mechanism is missing completely at random (MCAR); that is, the missingness is independent of all the observed and unobserved variables (Wang and Chen 2009). A more general missing mechanism is missing at random (MAR) which will be investigated in this Chapter. Besides, inference built on the complete case analysis is generally inefficient as it throws away data with missing values. Many efforts have been devoted to address this issue. Generally speaking, there are two ways to handle missing data problems. The first is to impute a plausible value for each missing value and then analyze the data as if they were complete. See,

for example, linear regression imputation (Yates 1933), ratio imputation (Rao 1996), semiparametric regression imputation (Wang and Rao 2002), and kernel regression imputation (Cheng 1994), etc. Rubin (1987) proposed a popular and general multiple imputation (MI) procedure. The other is to use the inverse probability weighted (IPW) approach introduced by Robins, Rotnitzky and Zhao (1994); see Zhao, Lipsitz and Lew (1996), Wang, Wang, Zhao and Ou (1997), Robins, Rotnitzky and Zhao (1994), Wang, Linton and Härdle (2004), and Guo and Xu (2012). However, existing regression imputation and inverse probability weighted approaches involve high-dimensional smoothing for estimating the completely unknown regression function and selection probability function in nonparametric settings. This difficulty consequently hinders their applications due to the well known curse of dimensionality. One can refer to Little and Rubin (2002) and references therein for a comprehensive review of statistical methods dealing with missing data.

To deal with the dimensionality problem, dimension reduction is necessary for us to efficiently work on regression analysis. Another motivation is to extend our proposed hypothesis-adaptive testing procedure to incomplete response situation. As seen from Chapter 4, our introduced approaches rely on the dimension reduction. Thus it's necessary to develop the dimension reduction theory with missing response at random which is the aim of this Chapter. Sufficient dimension reduction (SDR) has generated considerable interest in high-dimensional regressions. This general methodology aims at dealing with data sparseness in high-dimensional scenarios without parametric model structure. A pioneer research is sliced inverse regression proposed by Li (SIR 1991). Let Y and X be respectively the response and predictor vector. In general, the central subspace (CS, Cook 1998), denoted by $\mathcal{S}_{Y|X}$ in our context, is defined as the subspace S of minimal dimension such that $Y \perp\!\!\!\perp X | P_{\mathcal{S}} X$, where $\perp\!\!\!\perp$ indicates statistical independence and $P_{(\cdot)}$ is a projection operator with respect to the usual inner product. Its dimension $d = \dim(\mathcal{S}_{Y|X})$ is often used to refer to structural dimension. We call the vectors forming a basis of $\mathcal{S}_{Y|X}$ dimension

reduction directions.

Recently, in the context of missing predictor, Li and Lu (2008) combined SIR and the augmented inverse probability weighted method for the dimension reduction problem. Zhu, Wang and Zhu (2012) introduced a nonparametric imputation procedure for semiparametric regressions with missing predictors. When the missingness depends on both the completely observed predictors and the response, they still needed a parametric model structure to impute missing values. Ding and Wang (2011) proposed a fusion-refinement (FR) procedure to target the dimension reduction problem with missing response. Let δ be the missingness indicator, which equals 1 if the response Y is observed and 0 otherwise. Following the literatures (Little and Rubin 2002) in the missing data area, we adopt the commonly used misssing mechanism missing at random (MAR) in this Chapter. To be precise, this means $P(\delta = 1|Y, X) = P(\delta = 1|X) = \pi(X)$ or in other words, $Y \perp\!\!\!\perp \delta|X$. Further $\pi(X)$ is called the selection probability. Assume that $\gamma \in \mathbb{R}^{p\times q}$ is a basis matrix of the central subspace $\mathcal{S}_{\delta|X}$, and $\beta \in \mathbb{R}^{p\times d}$ is a basis matrix of the central subspace $\mathcal{S}_{Y|X}$. In fusion stage, they first enlarged the target subspace to be the joint central subspace $\mathcal{S}_{(Y,\delta)|X}$. In the second stage of estimation, they used probability mass function (pmf) imputation method and their Theorem 2 in their paper to recover $\mathcal{S}_{Y|X}$ that is a subspace of $\mathcal{S}_{(\delta,Y)|X}$. The estimate is proved to be consistent. However, their numerical studies indicate that when the angle between the subspace $\mathcal{S}_{\delta|X}$ and $\mathcal{S}_{Y|X}$ is large, the estimation accuracy is not satisfactory in the sense that $\mathcal{S}_{Y|X}$ may not be extracted from $\mathcal{S}_{(\delta,Y)|X}$ straightforwardly. Ding and Wang (2011) then suggested an ad hoc way to determine a threshold value of the angle between these two estimated subspaces for the practical use. Their recommendation was based on the simulation results they conducted. It is necessary to have a data-adaptive approach to determine which method should be used for maximizing the estimation accuracy.

From the above observations, in this Chapter, two alternative approaches are proposed to promote the estimation accuracy. The first is to take care of the inefficiency

of nonparametric estimation when the dimension of $\mathcal{S}_{(\delta,Y)|X}$ is relatively large. For instance, it is well known that the kernel estimation can have an optimal rate of convergence $O_p(n^{-2/(4+d+q)})$ when the density function of $(\gamma, \beta)^\top X$ is two times differentiable. Thus we can see that a more efficient way may be to directly impute Y via the conditional distribution of Y given the projection of X onto $\mathcal{S}_{\delta|X}$ such that we can suffer less from the nonparametric estimation with data sparseness in high-dimensional space. This idea should also be consistent with the theme of dimension reduction investigated in this Chapter. Based on this motivation, A novel two-stage method is proposed. In the first stage, we obtain a basis estimate $\hat{\gamma}$ for $\mathcal{S}_{\delta|X}$, and impute Y through the conditional distribution of Y given $\hat{\gamma}^\top X$. This stage is called Selection Probability Assisted Recovery (SPAR). However, as $\mathcal{S}_{\delta|X}$ is not necessarily contained in $\mathcal{S}_{Y|X}$, we then use the CC method to assist. Since the estimate deduced from the CC method is consistent, we can develop a Complete Case Assisted Recovery (CCAR). First, obtain a basis estimate $\hat{\beta}$ for $\mathcal{S}_{Y|X}$ from the CC analysis and then impute missing responses through the conditional distributions of Y given $\hat{\beta}^\top X$.

From our comprehensive simulation studies we found that almost uniformly,

- When $S_{Y|X}$ is close to $S_{\delta|X}$, SPAR is better than CCAR, whereas when the angle between these two subspaces is large, CCAR is the winner.

Thus, a natural question is what angle is regarded as either small or large and then we should use either SPAR or CCAR. A straightforward idea is to choose the method which can more efficiently estimate the subspace $S_{Y|X}$. To this end, we propose a data-adaptive method to synthesize both SPAR and CCAR to produce a new estimate. The method can automatically choose one of them in a data-adaptive way to maximize the estimation accuracy. This adaptive approach is realized by the bootstrap method.

Thus, in this Chapter, we make a comparison between these two methods and the FR procedure through the numerical studies.

The rest of the Chapter is organized as follows. In Section 5.2 the methods that

are based on SPAR and CCAR are introduced. The methodologies are illustrated by adopting the commonly used SIR in Section 5.3. In Section 5.4 the data-adaptive approach for SPAR and CCAR is elaborated. Simulation studies are conducted to examine the performance of the methods with a comparison with the FR procedure. In Section 5.5 the proposed procedures is applied to analyze a real data. Conclusions are given in Section 5.6. The technical details are relegated to the Appendix 5.7.

5.2 Semiparametric Dimension Reduction Assisted Recovery

Since the conditional mean imputation, a commonly used imputation method, cannot be applied in our context because it focuses on only one characteristic of the conditional distribution, we then use the multiple imputation (Rubin 1987). Below we propose a general idea of semiparametric dimension reduction for imputing missing responses. Instead of estimating the conditional distribution $P_{Y|X}$, we estimate the conditional distribution $P_{Y|S(X)}$, where $S(X)$ is some unknown transformation of X. The function $S(\cdot)$ should satisfy the following two conditions:

(i) The dimension of $S(X)$ is smaller than that of X.

(ii) $P_{Y|S(X)} = P_{Y|S(X),\delta=1}$ or $P_{Y|X} = P_{Y|S(X),\delta=1}$.

Note that for the cases with missing values in Y, condition (ii) guarantees that we can use the observed data set (i.e. $\delta = 1$) to obtain consistent estimate of either $P_{Y|S(X)}$ or $P_{Y|X}$ and then use multiple imputation to generate new values for missing units. Further, since the dimension of $S(X)$ is smaller than that of X, the involved nonparametric estimate of the related conditional distribution can also have less estimation accuracy lose. As the result, the main issue is to get a suitable transformation of X for us to successfully estimate $P_{Y|S(X)}$ for imputation of Y. Below we propose two solutions.

5.2.1 Selection Probability Assisted Recovery

Suppose that $\gamma \in \mathbb{R}^{p\times q}$ is a basis matrix of the central subspace $S_{\delta|X}$; that is, the column space of γ equals $S_{\delta|X}$. We note that

$$\begin{aligned} P(\delta = 1|\gamma^\top X, Y) &= E\{E(\delta|X,Y)|\gamma^\top X, Y\} \\ &= E\{E(\delta|X)|\gamma^\top X, Y\} \\ &= E\{E(\delta|\gamma^\top X)|\gamma^\top X, Y\} \\ &= P(\delta = 1|\gamma^\top X). \end{aligned}$$

Thus, conditioning on $\gamma^\top X$, δ is independent of Y. Then for any fixed $y \in \mathbb{R}$, we have

$$\begin{aligned} P(Y \le y|\gamma^\top X, \delta = 1) &= \frac{P(\delta = 1|\gamma^\top X, Y \le y)P(Y \le y|\gamma^\top X)}{P(\delta = 1|\gamma^\top X)} \\ &= P(Y \le y|\gamma^\top X). \end{aligned}$$

It follows that $\gamma^\top X$ is a suitable transformation $S(X)$ of X satisfying the aforementioned conditions (i) and (ii). We now employ the pmf imputation to impute missing responses. Let $\hat{\gamma}$ be an estimate of γ. Following the method in Hall, Wolff and Yao (1999), when MAR assumption holds, we can estimate the conditional distribution of Y given $\gamma^\top X$ by

$$\hat{P}(Y < y|\hat{\gamma}^\top X = \hat{\gamma}^\top x) = \frac{\sum_{i=1}^n \delta_i I(y_i \le y) K_h(\hat{\gamma}^\top x_i - \hat{\gamma}^\top x)}{\sum_{i=1}^n \delta_i K_h(\hat{\gamma}^\top x_i - \hat{\gamma}^\top x)}, \tag{5.1}$$

where $K_h(u) = h^{-q}K(u/h)$, h is the bandwidth and $K(\cdot)$ is a kernel function. From (5.1), we can know that the jump at the points y_j corresponding to $\delta_j = 1$ of the estimated distribution of Y given $\hat{\gamma}^\top X$ is,

$$\begin{aligned} \hat{p}_{ij} &\equiv \hat{P}(Y = y_j|\hat{\gamma}^\top X = \hat{\gamma}^\top x_i) \\ &= \frac{\delta_j K_h(\hat{\gamma}^\top x_j - \hat{\gamma}^\top x_i)}{\sum_{j=1}^n \delta_j K_h(\hat{\gamma}^\top x_j - \hat{\gamma}^\top x_i)}. \end{aligned} \tag{5.2}$$

Generate m conditionally independent samples, $\{Y_{i1}^*, \cdots, Y_{im}^*\}$, from $\hat{P}_{\cdot|\gamma^\top X}$, and then obtain m imputed values for every Y_i as follows:

$$\tilde{Y}_{ij} = \delta_i Y_i + \Big(1 - \delta_i\Big) Y_{ij}^*, \quad j = 1, \cdots, m. \tag{5.3}$$

If m is large enough, $\hat{p}_{ij}$ is approximately the proportion of samples equal to y_j. Inspired by this observation, the pmf imputation realizes multiple imputation by drawing dummy samples $\{(\hat{\gamma}^\top x_i, y_j), i, j = 1, \cdots, n\}$ with weights $\{\delta_i I(i = j) + (1 - \delta_i)\hat{p}_{ij}\}$; see Ding and Wang (2011) for more discussions about the advantages of the pmf imputation.

5.2.2 Complete Case Assisted Recovery

Let $S_{Y|X}^{\{\delta=1\}}$ be the partial central subspace that is the minimal subspace $\mathcal{S}$ such that $Y \perp\!\!\!\perp X|(P_S X, \delta = 1)$ for a projection matrix P_S. When MAR assumption holds, we know that $P_{Y|X} = P_{Y|X,\delta=1}$ and $S_{Y|X}^{\{\delta=1\}} = S_{Y|X}$. Consequently, we can use the CC method to obtain an initial estimate of the basis of $S_{Y|X}$.

Let $A \in \mathbb{R}^{p\times d}$ be a basis matrix of $S_{Y|X}^{\{\delta=1\}}$, and let $\hat{A}$ be an estimate of A. Note that $Y \perp\!\!\!\perp X|(A^\top X, \delta = 1)$ and $P_{Y|X} = P_{Y|X,\delta=1} = P_{Y|A^\top X,\delta=1}$. Thus we find the second suitable transformation $S(X) = A^\top X$ of X satisfying conditions (i) and (ii), and $A = \beta$. As before, we estimate the conditional distribution of Y given $\hat{\beta}^\top X$ by

$$\hat{P}(Y < y|\hat{\beta}^\top X = \hat{\beta}^\top x) = \frac{\sum_{i=1}^n \delta_i I(y_i \le y) K_h(\hat{\beta}^\top x_i - \hat{\beta}^\top x)}{\sum_{i=1}^n \delta_i K_h(\hat{\beta}^\top x_i - \hat{\beta}^\top x)},$$

where $K_h(u) = h^{-d}K(u/h)$. Again, we use the pmf imputation method. Specifically, we draw dummy samples $\{(\hat{\beta}^\top x_i, y_j), i, j = 1, \cdots, n\}$ with weights $\{\delta_i I(i = j) + (1 - \delta_i)\hat{p}_{ij}^{cc}\}$, where

$$\hat{p}_{ij}^{cc} = \frac{\delta_j K_h(\hat{\beta}^\top x_j - \hat{\beta}^\top x_i)}{\sum_{j=1}^n \delta_j K_h(\hat{\beta}^\top x_j - \hat{\beta}^\top x_i)}. \tag{5.4}$$

We call this procedure Complete Case Assisted Recovery (CCAR).

Remark 5.1. *As we commented in the introduction, when we directly use the information contained in the missing indicator δ, we can suffer less estimation accuracy from nonparametric smoothing for the conditional distribution of Y either given $\gamma^T X$ in SPAR (an optimal rate of convergence may be $O_p(n^{-2/(4+q)})$) or given $\beta^T X$ in CCAR than that given $(\gamma, \beta)^T X$ (an optimal rate of convergence may be*

$O_p(n^{-2/(4+d+q)})$) in FR of Ding and Wang (2011). As a result, the proposed methods, SPAR and CCAR, may have good performance in finite sample scenarios.

5.3 SIR with Missing Response

In this section we show how to get the central subspace $S_{Y|X}$ with missing responses by SPAR and CCAR.

5.3.1 Application of SIR to SPAR and CCAR

As reviewed in Chapter 1, it is ready for us to apply SIR to SPAR. In the first stage, we get the SIR estimate $\hat{\gamma}$ of the basis vectors of $S_{\delta|X}$. For binary response, Cook and Lee (1999) discussed the dimension reduction problems in detail. Consequently, we can also use other methodologies in their paper to obtain $\hat{\gamma}$. As for binary response, SIR can find only one direction, we assume that $\dim(S_{\delta|X}) = 1$ as assumed in Ding and Wang (2011), otherwise, SIR can identify a subspace of $S_{\delta|X}$. We should note that this assumption depends on the SDR method we consider and can be abandoned if we use other SDR methods. Moreover, this assumption is mild compared with Li and Lu (2008) and Zhu, Wang and Zhu (2012) in which they posited parametric models for the selection probability $P(\delta = 1|X)$; it can be considered as a semi-parametric assumption which is more flexible and avoids suffering from the curse of dimensionality.

To obtain an estimate of the basis vectors of $S_{Y|X}$, $\hat{\beta}^{SPAR}$, through SPAR, the estimation procedure in following steps are formulated.

Step 1. We use SIR to obtain an estimator $\hat{\gamma}$. To be precise, $\hat{\gamma} = \hat{\Sigma}_x^{-1}(\bar{x}_{ob} - \bar{x}_{miss})'$, here $\hat{\Sigma}_x = n^{-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top$ is the sample covariance matrix, $\bar{x}$ is the sample mean of X, $\bar{x}_{ob}$ and $\bar{x}_{miss}$ are the sample means of X corresponding to $\delta = 1$ and to $\delta = 0$ respectively.

Step 2. We then can obtain $\hat{p}_{ij}$ by (5.2), that's,

$$\hat{p}_{ij} = \frac{\delta_j K_h(\hat{\gamma}^\top x_j - \hat{\gamma}^\top x_i)}{\sum_{j=1}^n \delta_j K_h(\hat{\gamma}^\top x_j - \hat{\gamma}^\top x_i)}.$$

Step 3. Estimate p_s and m_s respectively by

$$\hat{p}_s^{SPAR} = n^{-1} \sum_{i,j=1}^n \{[(1-\delta_i)\hat{p}_{ij} + \delta_i I(i=j)] I(y_j \in I_s)\}$$

and

$$\hat{m}_s^{SPAR} = n^{-1} \sum_{i,j=1}^n \{[(1-\delta_i)\hat{p}_{ij} + \delta_i I(i=j)] I(y_j \in I_s)\} \times x_i / \hat{p}_i^{SPAR}.$$

Then we have $\hat{\Lambda}_n = \sum_{s=1}^M \hat{p}_s^{SPAR}(\hat{m}_s^{SPAR} - \bar{x})(\hat{m}_s^{SPAR} - \bar{x})^\top$. Conduct the spectral decomposition of $\hat{\Lambda}_n$ with respect to $\hat{\Sigma}_x$. We thus obtain $\hat{\beta}^{SPAR}$, an estimate of the basis vectors of $S_{Y|X}$.

For CCAR, minor changes should be made. Instead of estimating p_{ij} by (5.2), it is estimated by (5.4). The resulting estimate is denoted to be $\hat{\beta}^{CCAR}$.

We are now in the position to study the asymptotic properties of $\hat{\beta}^{SPAR}$ and $\hat{\beta}^{CCAR}$. Under the conditions presented in the Appendix, we can show the root-n consistency of $\hat{\beta}^{CCAR}$ which is stated in the following.

Theorem 5.1. *Assume that* $K = \dim(S_{Y|X})$ *is given in advance and that the conditions in the Appendix are true. Then* $\mathrm{Span}(\hat{\beta}^{CCAR})$ *is root-n consistent estimate of* $S_{Y|X}$.

We realize that $\hat{\beta}^{SPAR}$ may not be consistent. This is because for SPAR, Y is imputed from the conditional distribution of Y given $\gamma^\top X$ which may be different from that given X. On the other hand, though we cannot prove that $\hat{\beta}^{SPAR}$ is also root-n consistency theoretically, the numerical performance of SPAR is preferred as shown in our simulation studies section. Further, in practice, there is no oracle to tell us the true value of $d = \dim(S_{Y|X})$ and it should be estimated from data. To this end, we apply the so-called "BIC-type" criterion.

5.3.2 Determination of the Structural Dimension

To determine the structural dimension, Li (1991) developed a sequential test method. For this method, many authors such as Schott (1994), Velilla (1998), Bura and Cook (2001) and Ferré (1998) made further extensions. Zhu, Miao and Peng (2006) first introduced a method of Bayesian information criterion (BIC) type to consistently determine d. This criterion was further modified by Zhu, Wang, Zhu and Ferré (2010). We adopt the modified BIC-type criterion. This criterion has been introduced in Chapter 4, subsection 4.2.3 and thus we omit the details. The only differences here are that $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$ are the eigenvalues of $\hat{\Lambda}_n$ with respect to $\hat{\Sigma}_x$.

Theorem 5.2. *Assume that $D_n \to \infty$ and $D_n/n \to 0$ as $n \to \infty$ and that the conditions in Appendix are true, $\hat{d}^{CCAR}$ is consistent estimate of the structural dimension.*

The detailed proof is omitted as it is similar to that of Zhu, Wang, Zhu and Ferré (2010). As for choosing D_n, Zhu, Wang, Zhu and Ferré (2010) recommended a value $D_n = n^{1/2}$ to avoid either overestimating or underestimating d. Following their recommendation, we also use this value in our numerical studies.

5.4 Simulation Studies

In this section, we examine the performance of SPAR and CCAR through carrying out extensive simulation studies and comparing them with the FR procedure and the CC method.

Consider the following four models for the purpose of comparison:

$$Y = \frac{\beta_1^\top X}{0.5 + (\beta_2^\top X + 1.5)^2} + \epsilon, \tag{5.5}$$

$$Y = \mathrm{sgn}(\beta_1^\top X) \log |\beta_2^\top X + 5| + \epsilon, \tag{5.6}$$

$$Y = \beta_1^\top X + (\beta_2^\top X + 3)\epsilon, \tag{5.7}$$

$$Y = \beta_1^\top X + \exp(|\beta_2^\top X|)\epsilon. \tag{5.8}$$

These models were used in Ding and Wang (2011). Specifically, both $X = (X_1, \cdots, X_p)^\top$ and $(X^\top, \epsilon)^\top$ follow standard normal distributions. In these models, $p = 10$, $\beta_1 = (0.5, 0.5, 0.5, 0.5, 0, \cdots, 0)^\top$ and $\beta_2 = (0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5, 0, 0)^\top$. We denote the basis matrix of $S_{Y|X}$ as $\beta = (\beta_1, \beta_2)$. Consider the following two types of missing mechanisms

$$\pi(x) = P(\delta = 1|X = x) = \frac{1}{1 + \exp\{-(\gamma^\top x + a_0)\}}; \tag{5.9}$$

$$\pi(x) = P(\delta = 1|X = x) = \frac{1}{1 + \exp[-\{(\gamma_1^\top X)^2 + \gamma_2^\top X + a_0\}]}, \tag{5.10}$$

where a_0 is introduced to determine missing proportion.

Let $\check{\beta}_1 = v_1\beta_1 + v_2\beta_2$ and $\check{\beta}_2 = v_2\beta_1 - v_1\beta_2$, where $v_1 \sim U(0, 1)$ and $v_2 \sim U(0, 1)$. For missing mechanism (5.9), define $\check{\gamma} = \check{\beta}_1/||\check{\beta}_1|| + (0, 0, \cdots, a_1)^\top$ and set $\gamma = \check{\gamma}/(1 + a_1^2)^{1/2}$. For (5.10), define $\check{\gamma}_1 = \check{\beta}_1/||\check{\beta}_1|| + (0, \cdots, 0, a_1, 0)^\top, \check{\gamma}_2 = \check{\beta}_2/||\check{\beta}_2|| + (0, \cdots, 0, a_1)$ and set $\gamma = (\gamma_1, \gamma_2)$, where $\gamma_1 = \check{\gamma}_1/(1 + a_1^2)^{1/2}$ and $\gamma_2 = \check{\gamma}_2/(1 + a_1^2)^{1/2}$.

To see whether a $p \times 1$ vector γ and a $p \times d$ matrix β is close to each other or not, we adopt the squared multiple correlation coefficient,

$$R^2(\gamma, \beta) = \max_{b \in \text{Span}(\beta)} \frac{(\gamma^\top \Sigma_x b)^2}{\gamma^\top \Sigma_x \gamma \cdot b^\top \Sigma_x b}.$$

Here a_1 is used to determine the closeness of γ and β. When $a_1 = 0$, $S_{\delta|X}$ is a subspace of $S_{Y|X}$; on the other hand, when a_1 is large enough, $S_{\delta|X}$ and $S_{Y|X}$ will be vertical to each other. Note that for the estimate of γ in model (5.9) or γ_1 and γ_2 in model (5.10), we adopt SIR for simplicity. Other methods are also available. Also, we note that γ is not the parameter of interest. As is well known, SIR can only select one direction when the response δ is binary. Thus, the use of SIR for model (5.10) can also be viewed as a robustness study. That is, even when we only estimate one direction of $S_{\delta|X}$, we still can get reasonable results by SPAR. For other discussions, refer to the beginning of Subsection 5.3.1.

To measure the estimation accuracy of $\hat{\beta}$, we adopt the average of the squared canonical correlation coefficients $R^2(\hat{\beta}, \beta)$ proposed by Li (1991). This coefficient

$R^2(\hat{\beta}, \beta)$ ranges between 0 and 1, the larger R^2, the closer $\hat{\beta}$ to β. Consequently, one estimate of β performs better than another estimate when it possesses larger mean and smaller standard deviation of R^2.

5.4.1 Estimation of the Central Subspace

The sample size is $n = 400$, the number of slices M is set to be 10, and we determine $K = \dim(S_{Y|X})$ by the BIC criterion described in Subsection 5.3.2. We use the kernel function $K(u) = \prod_i K_1(u_i)$ with $K_1(u_1) = 3/4(1-u_1^2)I(|u_1| \leq 1)$. For the bandwidth selection, we adopt the same setting as Ding and Wang (2011), that is, $h = n^{-1/(d+4)}$. We carry out 500 simulation runs. The choice of bandwidth h may be considered to be ad hoc. The first reason for this choice is to make a fair comparison with that in Ding and Wang(2011) as they used this bandwidth. Also note that the optimal bandwidth is of the order $n^{-1/(d+4)}$. Although data-adaptive procedures are often used, our empirical study shows the insensitiveness of the method to the bandwidth. Thus implementation convenience with much less computational burden than that data-driven procedures, we report the results with bandwidth $n^{-1/(d+4)}$.

The curves of mean and standard deviations of $R^2(\hat{\beta}, \beta)$ versus the angle between $S_{\delta|X}$ and $S_{Y|X}$ with 25%, 50%, and 75% missing proportions are presented in Figures 5.1- 5.5. We can see obviously that the proposed two-stage estimates $\hat{\beta}^{SPAR}$ and $\hat{\beta}^{CCAR}$ perform better than FR and CC respectively. Except in Figure 5.1 (e), we can see clearly SPAR improves FR substantially. Even in Figure 5.1 (e), we can also note that when these two subspaces are close to each other, SPAR still performs greatly better than FR.

From these figures, we can also know that when $S_{\delta|X}$ and $S_{Y|X}$ are close, generally SPAR performs best; while when $S_{\delta|X}$ is away from $S_{Y|X}$, CCAR is the best. As discussed by Ding and Wang(2011), when $S_{\delta|X}$ and $S_{Y|X}$ are close, $S_{\delta|X}$ generally carries some meaningful information about $S_{Y|X}$. Further, since δ is completely observed, we can estimate $S_{\delta|X}$ accurately. These facts can thus help us obtain

better estimate of $S_{Y|X}$ by SPAR.

For determining the structural dimension, we present the results in Table 5.1. To save space, we focus on the model (5.5). From Table 5.1, we can see that when $R^2(\gamma, \beta) = 1$, the SPAR estimate of d, d^{SPAR}, is comparable to the full data-based estimate, d^{FULL}, for the missing mechanism of (5.9) with missing proportion 25% and 50% and performs slightly better than d^{FULL} for the mechanism of (5.10) with missing proportion 50%. When $R^2(\gamma, \beta) = 0$, d^{CCAR} becomes comparable to d^{FULL} while d^{SPAR} performs not well. In contrast, d^{CC} obtained from the CC method does not perform well in most of cases expect in the mechanism (5.10) with missing proportion 50%. In this case, d^{CCAR} performs best. Therefore, when $R^2(\gamma, \beta)$ is large, SPAR can determine the dimension well and when $R^2(\gamma, \beta)$ is close to zero, CCAR is a better approach.

5.4.2 Data-Adaptive Synthesization

As suggested in the above observations, generally when the above mentioned two subspaces are close to each other, SPAR is recommendable, whereas when the angle is comparatively large, CCAR should be used. However, in practice, it is unknown about the closeness of these two subspaces and thus a threshold value of angle is difficult to determine. To handle this issue, we apply the bootstrap method to develop a data-adaptive synthesization to choose between SPAR and CCAR. To be precise, we first use the bootstrap procedure to generate a set of $\hat{\beta}_b^{SPAR}$ and $\hat{\beta}_b^{CCAR}, b = 1, \ldots, B$. We then calculate the squared canonical correlation coefficients $R^2(\hat{\beta}^{SPAR}, \hat{\beta}_b^{SPAR})$ and $R^2(\hat{\beta}^{CCAR}, \hat{\beta}_b^{CCAR})$ as a distance measure and use the mean of the nonnegative distances as a measure of the performance of $\hat{\beta}^{SPAR}$ and $\hat{\beta}^{CCAR}$. Among these two estimates, we will use the one with the larger mean squared canonical correlation coefficients as our preferred estimate.

We apply this data-adaptive approach to the models considered in the above subsection. We take 250 replications and 200 bootstrapping samples. The results

for model (5.7) are shown in Figure 5.6. We also compare them with Ding and Wang (2011)'s ad hoc selection with $\alpha_0 = 45°$: the angle between $S_{\delta|X}$ and $S_{Y|X}$ by their simulation results. That is, when the angle is smaller than $\alpha_0 = 45°$, FR is used, otherwise CC is applied. From Figure 5.6, the data-adaptive synthesization of SPAR and CCAR obviously overmatches the ad hoc synthesization of FR and CC. Further, we also find that for model (5.7), the data-adaptive procedure can have smaller standard deviation than the ad hoc method.

5.5 Application to A HIV Dataset

In this section we use our procedures to analyze a real world data collected from a HIV clinical trial. Patients in this study were randomly divided into four groups to receive (i) zidovudine (ZDV), (ii) didanosine (ddi), (iii) ZDV + ddi, and (iv) ZDV + zalcitabine respectively (Hammer et al. 1996; Hu, Follmann and Qin 2010). We are interested in comparison of the treatment effects of monotherapy, say (i), and combined therapies, say (ii)-(iv), for 746 male patients who had not received antiretroviral therapy before this trial. The response variable Y is the CD4 count at 96 ± 5 weeks post therapy. As for predictor vector X, we take six baseline variables: age, weight, CD4 cell counts at baseline and 20 ± 5 weeks, and CD8 cell counts at baseline and 20 ± 5 weeks. The indicator variable T for random assignment is 1 or 0 depending on whether a patient received combined therapy or not. Consequently, $E(Y|T = 1)$ is the mean response under the combined therapy and $E(Y|T = 0)$ is the mean response under the monotherapy. The treatment effect is defined to be $\Delta = E(Y|T = 1) - E(Y|T = 0)$. Among the 746 patients, 567 patients received combined therapy, while the others received monotherapy. Due to death and dropout, 199 patients with $T = 1$ and 74 patients with $T = 0$ have missing response values. The predictors for all patients are available.

Let γ and β denote the basis of $S_{\delta|X}$ and $S_{Y|X}$ respectively. Since there are many

missing observations in Y, we do not apply the CC method directly. Instead, we note that

$$E(Y|T=1) = E[E(Y|\gamma^\top X, T=1)|T=1] = E[E(Y|\beta^\top X, T=1)|T=1]. \quad (5.11)$$

Consequently, to get the estimate of the mean of Y with treatment $T=1$, we can first obtain the estimate of the conditional expectation $E(Y|\gamma^\top X, T=1)$ or $E(Y|\beta^\top X, T=1)$, and then average over all the unites with $T=1$. To do this, we first need to obtain the estimates of the basis of the subspace $S_{\delta|X}$ and $S_{Y|X}$ with $T=1$. For treatment $T=1$, the data-adaptive approach suggests that $\hat{\beta}_{T=1}^{CCAR} = (0.1031, -0.6678, -0.2086, -0.7037, -0.0312, 0.0610)^\top$ is the better choice between the proposed two estimates $\hat{\beta}_{T=1}^{CCAR}$ and $\hat{\beta}_{T=1}^{SPAR}$. On the other hand, the estimate for $S_{\delta|X}$ is $\hat{\gamma}_{T=1} = (0.9453, -0.3061, -0.0751, 0.0839, -0.0001,\ 0.0078)^\top$. The angle between $\hat{\beta}_{T=1}^{CCAR}$ and $\hat{\gamma}_{T=1}$ is about 72.146°, while that between $\hat{\beta}_{T=1}^{SPAR}$ and $\hat{\gamma}_{T=1}$ is about 62.438°. As for treatment $T=0$, the same strategy can be used to estimate $E(Y|T=0)$. The data-adaptive procedure selects $\hat{\beta}_{T=0}^{SPAR} = (-0.2014, 0.9412, -0.1486, -0.2260, -0.0071, 0.0216)^\top$ other than $\hat{\beta}_{T=0}^{CCAR}$ as the estimate of the basis of the subspace $S_{Y|X}$. Accordingly, the estimate for $S_{\delta|X}$ is $\hat{\gamma}_{T=0} = (0.4108, -0.8901, -0.0301, 0.1921, 0.0207, -0.0280)^\top$. The angle between $\hat{\beta}_{T=0}^{CCAR}$ and $\hat{\gamma}_{T=0}$ is about 33.535°, while that between $\hat{\beta}_{T=0}^{SPAR}$ and $\hat{\gamma}_{T=0}$ is about 23.169°. Thus, the data-adaptive method does work well to choose a better one between SPAR and CCAR.

To further study the relationship between the response variable and the dimension reduction predictors, we show scatter plots in Figure 5.7. From this figure, we can observe clearly that linear regression models are appropriate to describe the relationship. Simple linear regression yields

$$\begin{aligned} E(Y|\beta^\top X, T=1) &= -46.4069 - 1.0618(\hat{\beta}_{T=1}^{CCAR})^\top X, \\ E(Y|\beta^\top X, T=0) &= 154.0655 - 2.4833(\hat{\beta}_{T=0}^{SPAR})^\top X, \end{aligned} \quad (5.12)$$

and

$$\begin{aligned} E(Y|\gamma^\top X, T=1) &= 287.6551 + 3.4707(\hat{\gamma}_{T=1}^\top X), \\ E(Y|\gamma^\top X, T=0) &= 298.7726 + 3.6849(\hat{\gamma}_{T=0}^\top X). \end{aligned} \tag{5.13}$$

Under model (5.12), the estimated value of $\Delta = E(Y|T=1) - E(Y|T=0)$ is 83.9531, while under model (5.13) it is 81.4420.

5.6 Conclusion

In this article we introduce two novel two stage procedures SPAR and CCAR for dimension reduction with missing response at random. A popular SDR method, SIR, is used to explain our procedures. Through comprehensive simulation studies, the newly proposed procedures work better than existing ones and the data-adaptive algorithm is recommendable for a better synthesization of these two methods.

5.7 Appendix. Proof of the Theorems

The following conditions are required for the theorems in Section .

(C1) $E||X||^2 < \infty$.

(C2) Let $f(\beta^\top x)$ is the density of $\beta^\top x$, $\pi(\beta^\top x) = P(\delta = 1|\beta^\top X = \beta^\top x)$, which is assumed to be bounded away from zero and above. $m(\beta^\top x) = E(I(Y \in I_s)|\beta^\top X)$, $g(\beta^\top x) = \pi(\beta^\top x)f(\beta^\top x)$, $G(\beta^\top x) = m(\beta^\top x)g(\beta^\top x)$ $g(\beta^\top x)$ are defined on a compact support, and $\inf_{\beta^\top x} g(\beta^\top x) \geq c > 0$ for some constant c. $G(\beta^\top x), g(\beta^\top x), \pi(\beta^\top x)$ and $m(\beta^\top x)$ have bounded partial derivatives up to order 2.

(C3) The symmetric kernel function $K(\cdot)$ support on the interval $[-1,1]^d$ and satisfy: (a)$\int K(u)du = 1$. (b) There exists $m \geq 2$ such that $\int u_1^{l_1} \cdots u_p^{l_p} K(u)du = 0$, if $l_1 + \cdots + l_p < m$ and $\int u_i^m K(u)du \neq 0$ for $i = 1, \cdots p$.

(C4) As n tends to infinity, $nh^{2m} \to 0$ and $nh^{2d+2}/(\log(n))^2 \to \infty$.

Remark 5.2. *We discuss in brief the above regularity conditions. Condition 1) is often assumed to derive the asymptotic normality. Condition 2) is related to the smoothness of the response density function and regression curves, which is widely assumed in the literature. See, e.g., condition 1 in Zhu and Fang (1996) and Zhu and Zhu (2007). Condition 3) is chosen for simplicity of illustration, where the q-th order kernel is used. Condition 4) shows the range of bandwidth for the desired asymptotic.*

Proof for Theorem 5.1: Define the estimate of $g(\beta^\top x), G(\beta^\top x)$ and $m(\beta^\top x)$ as follows:

$$\begin{aligned}
\hat{g}(\hat{\beta}^\top x) &= n^{-1}\sum_{j=1}^{n}\delta_j K_h(\hat{\beta}^\top x_j - \hat{\beta}^\top x), \\
\hat{G}(\hat{\beta}^\top x) &= n^{-1}\sum_{j=1}^{n}\delta_j K_h(\hat{\beta}^\top x_j - \hat{\beta}^\top x_i) I(y_j \in I_s), \\
\hat{m}(\hat{\beta}^\top x) &= \frac{\sum_{j=1}^{n}\delta_j K_h(\hat{\beta}^\top x_j - \hat{\beta}^\top x) I(y_j \in I_s)}{\sum_{j=1}^{n}\delta_j K_h(\hat{\beta}^\top x_j - \hat{\beta}^\top x)}.
\end{aligned}$$

Recall β is the basis of $S_{Y|X}$, then the MAR assumption can imply that $E(I(Y \in I_s)|X) = m(\beta^\top x) = E(I(Y \in I_s)|\beta^\top X, \delta = 1)$ and $E(\delta I(Y \in I_s)|\beta^\top X) = m(\beta^\top x)E(\delta|\beta^\top X)$. Further we can have the following result:

$$\begin{aligned}
& E(\delta I(Y \in I_s) + (1-\delta)m(\beta^\top x)) \\
= \; & E\Big[E\Big(\delta I(Y \in I_s) + (1-\delta)m(\beta^\top x)\Big|\beta^\top X\Big)\Big] \\
= \; & E\Big[E(\delta I(Y \in I_s)|\beta^\top X) - m(\beta^\top x)E(\delta|\beta^\top X) + m(\beta^\top x)\Big] \\
= \; & P(Y \in I_s).
\end{aligned}$$

As a result, we can estimate $p_s = P(Y \in I_s)$ by

$$\begin{aligned}
\hat{p}_s^{CCAR} &= n^{-1}\sum_{i}^{n}\{\delta_i I(y_i \in I_s) + (1-\delta_i)\frac{\sum_{j=1}\delta_j K_h(\hat{\beta}^\top x_j - \hat{\beta}^\top x_i) I(y_j \in I_s)}{\sum_{j=1}^{n}\delta_j K_h(\hat{\beta}^\top x_j - \hat{\beta}^\top x_i)}\} \\
&= n^{-1}\sum_{i}^{n}\{\delta_i I(y_i \in I_s) + (1-\delta_i)\sum_{j=1}^{n}\hat{p}_{ij} I(y_j \in I_s)]\} \\
&= n^{-1}\sum_{i,j=1}^{n}\{[(1-\delta_i)\hat{p}_{ij} + \delta_i I(i=j)]I(y_j \in I_s)\},
\end{aligned}$$

where $\hat{p}_{ij}^{cc}$ is defined by

$$\hat{p}_{ij}^{cc} : = \frac{\delta_j K_h(\hat{\beta}^\top x_j - \hat{\beta}^\top x_i)}{\sum_{j=1}^n \delta_j K_h(\hat{\beta}^\top x_j - \hat{\beta}^\top x_i)}.$$

Similarly, we can have

$$\hat{m}_s^{CCAR} = n^{-1} \sum_{i,j=1}^n \{[(1-\delta_i)\hat{p}_{ij}^{cc} + \delta_i I(i=j)] I(y_j \in I_s)\} \times x_i / \hat{p}_s^{CCAR}.$$

Further note that

$$\begin{aligned} & E(\delta I(Y \in I_s)X + (1-\delta)m(\beta^\top x)X) \\ = \; & E(\delta I(Y \in I_s)X + (1-\delta)E(I(Y \in I_s)|X)X) = E(XI(Y \in I_s)). \end{aligned}$$

Consequently, the root n-consistency of $\hat{\beta}^{CCAR}$ follows from the fact that $\hat{\beta}$ is a root n-consistent estimate of basis of $S_{Y|X}$ and a similar argument in Ding and Wang (2011).

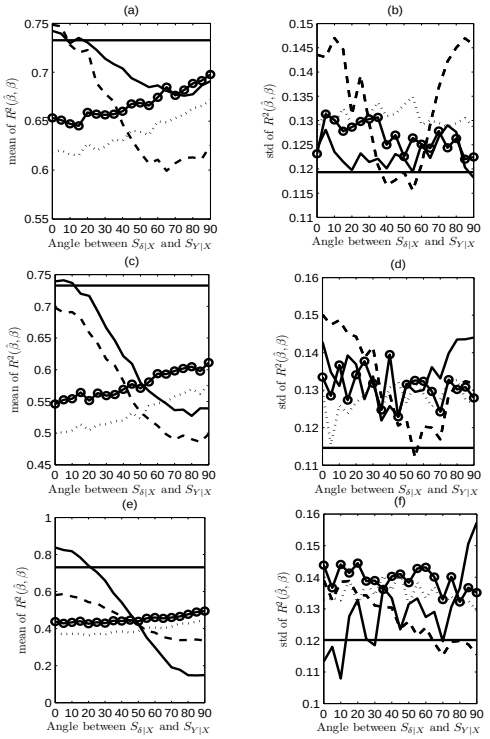

Figure 5.1: The x-axis is the angle between $S_{\delta|X}$ and $S_{Y|X}$ for model (5.5) with missingness (5.9); the y-axis is Mean (left) and SD (right) of $R^2(\hat{\beta}, \beta)$. The three rows are the results with 25%, 50% and 75% missing proportions respectively. The solid line, solid line marked by "o", dotted line, dashed line and horizontal line represent the results from SPAR, CCAR, complete-case, FR and the benchmark full-data estimates respectively.

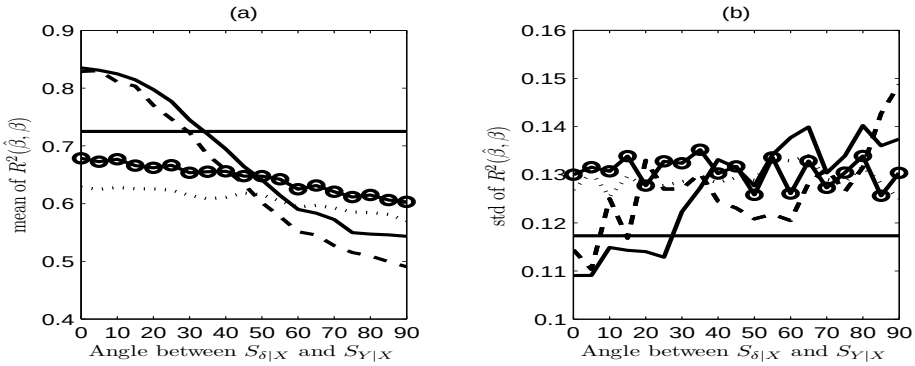

Figure 5.2: The x-axis is the angle between $S_{\delta|X}$ and $S_{Y|X}$ for model (5.5) with missingness (5.10); the y-axis is Mean (left) and SD (right) of $R^2(\hat{\beta}, \beta)$. The solid line, solid line marked by "o", dotted line, dashed line and horizontal line represent the results from SPAR, CCAR, complete-case, FR and the benchmark full-data estimates respectively.

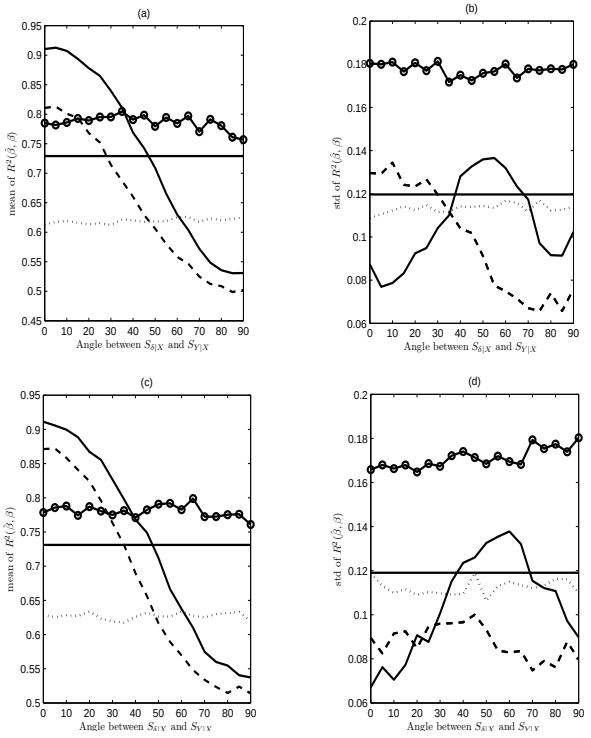

Figure 5.3: The x-axis is the angle between $S_{\delta|X}$ and $S_{Y|X}$ for model (5.6); the y-axis is Mean (left) and SD (right) of $R^2(\hat{\beta}, \beta)$. The two rows respectively correspond to missingness (5.9) with 50% missing proportion and missingness (5.10) with 50% missing proportion. The solid line, solid line marked by "o", dotted line, dashed line and horizontal line represent the results from SPAR, CCAR, complete-case, FR and the benchmark full-data estimates respectively.

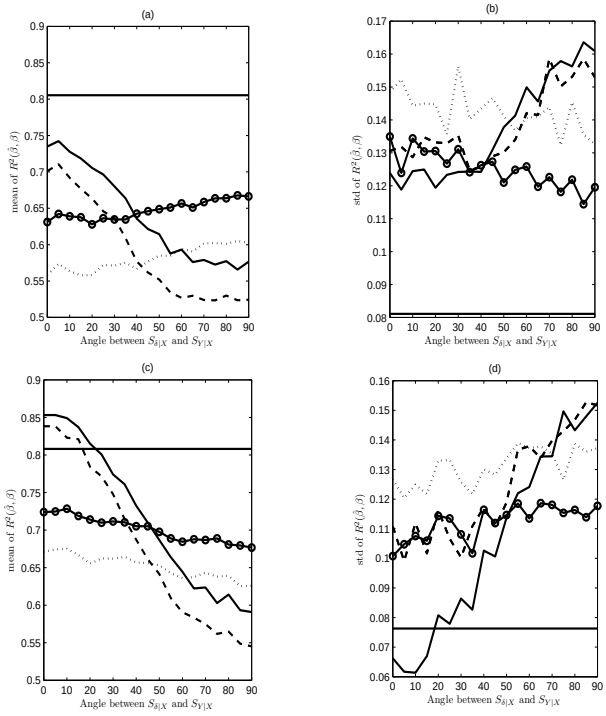

Figure 5.4: The x-axis is the angle between $S_{\delta|X}$ and $S_{Y|X}$ for model (5.7); the y-axis is Mean (left) and SD (right) of $R^2(\hat{\beta}, \beta)$. The two rows respectively correspond to missingness (5.9) with 50% missing proportion and missingness (5.10) with 50% missing proportion. The solid line, solid line marked by "o", dotted line, dashed line and horizontal line represent the results from SPAR, CCAR, complete-case, FR and the benchmark full-data estimates respectively.

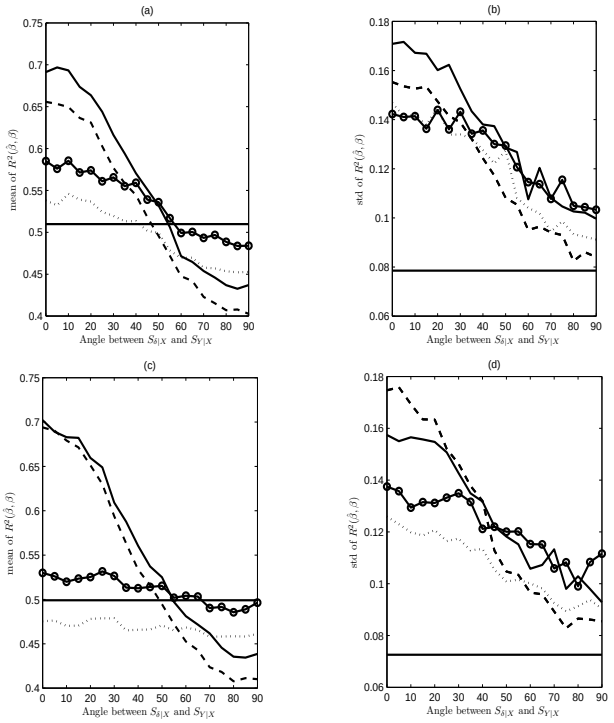

Figure 5.5: The x-axis is the angle between $S_{\delta|X}$ and $S_{Y|X}$ for model (5.8); the y-axis is Mean (left) and SD (right) of $R^2(\hat{\beta}, \beta)$. The two rows respectively correspond to missingness (5.9) with 50% missing proportion and missingness (5.10) with 50% missing proportion. The solid line, solid line marked by "o", dotted line, dashed line and horizontal line represent the results from SPAR, CCAR, complete-case, FR and the benchmark full-data estimates respectively.

Table 5.1: Distribution (in percentage) of the estimated structural dimension $d = \dim(S_{Y|X})$ for model 5.5 with missing mechanisms 5.9 and 5.10 and with $R^2(\gamma, \beta) = 1$ and $R^2(\gamma, \beta) = 0$ respectively.

$\hat{d}$	$R^2(\gamma,\beta)=1$			$R^2(\gamma,\beta)=0$		
	1	2	>2	1	2	>2
Case 1, missing mechanism 5.9 with 25% missing proportion						
$\hat{d}^{Full}$	0.0040	0.7300	0.2660	0.0040	0.7300	0.2660
$\hat{d}^{SPAR}$	0.0160	0.7020	0.2820	0.0020	0.4100	0.5880
$\hat{d}^{CCAR}$	0.0040	0.6240	0.3720	0.0020	0.6920	0.3060
$\hat{d}^{CC}$	0.0020	0.4700	0.5280	0.0020	0.6140	0.3840
Case 2, missing mechanism 5.9 with 50% missing proportion						
$\hat{d}^{Full}$	0.0020	0.7860	0.2120	0.0020	0.7860	0.2120
$\hat{d}^{SPAR}$	0.1720	0.6780	0.1500	0.0000	0.4260	0.5740
$\hat{d}^{CCAR}$	0.0080	0.4820	0.5100	0.0100	0.5580	0.4320
$\hat{d}^{CC}$	0.0000	0.3560	0.6440	0.0000	0.4460	0.5540
Case 3, missing mechanism 5.9 with 75% missing proportion						
$\hat{d}^{Full}$	0.0040	0.7400	0.2560	0.0040	0.7400	0.2560
$\hat{d}^{SPAR}$	0.8320	0.1600	0.0080	0.6020	0.3580	0.0400
$\hat{d}^{CCAR}$	0.0000	0.4680	0.5320	0.0040	0.5140	0.4820
$\hat{d}^{CC}$	0.0000	0.3520	0.6480	0.0020	0.4060	0.5920
Case 4, missing mechanism 5.10 with 50% missing proportion						
$\hat{d}^{Full}$	0.0040	0.7300	0.2660	0.0040	0.7300	0.2660
$\hat{d}^{SPAR}$	0.2060	0.7740	0.0200	0.0020	0.3340	0.6640
$\hat{d}^{CCAR}$	0.0400	0.7980	0.1620	0.0080	0.5740	0.4180
$\hat{d}^{CC}$	0.0060	0.7300	0.2640	0.0000	0.4400	0.5600

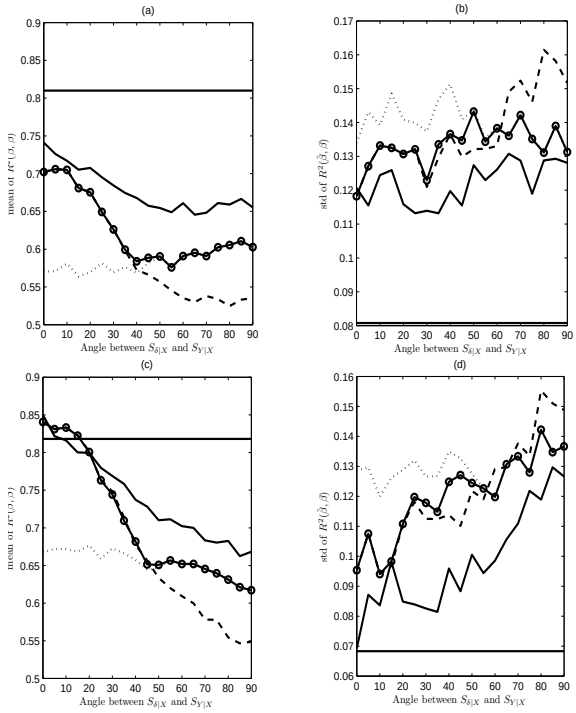

Figure 5.6: The x-axis is the angle between $S_{\delta|X}$ and $S_{Y|X}$ for model (5.7); the y-axis is Mean (left) and SD (right) of $R^2(\hat{\beta}, \beta)$. The two rows respectively correspond to the missingness of (5.9) with 50% missing proportion and the missingness of (5.10) with 50% missing proportion. The solid line, solid line marked by "o", dotted line, dashed line and horizontal line respectively represent the results from adaptive, Ding and Wang's ad hoc, complete-case, FR and the benchmark full data-based estimates.

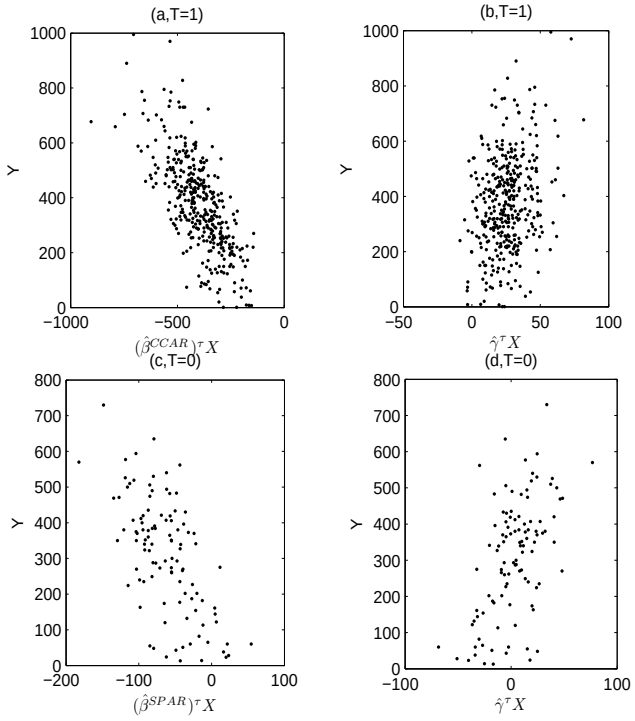

Figure 5.7: Scatter plots of CD4 counts at 96±5 weeks (Y) versus the estimated dimesion reduction predictors with the complete observations for treatment T. (a) $((\hat{\beta}^{CCAR}_{T=1})^\top X, Y, T = 1)$; (b) $(\hat{\gamma}^\top_{T=1} X, Y, T = 1)$; (c) $((\hat{\beta}^{SPAR}_{T=0})^\top X, Y, T = 0)$; and (d) $(\hat{\gamma}^\top_{T=0} X, Y, T = 0)$.

Bibliography

[1] Aerts, M., Claeskens, G., and Hart, J. D. (1999). Testing lack of fit in multiple regression, *Journal of the American Statistical Association*, **94**, 869-879.

[2] Bierens, H. J. (1982). Consistent model especification tests. *Journal of Econometrics*, **20**, 105-134.

[3] Bura, E., and Cook, R. D. (2001). Extending sliced inverse regression: the weighted chi-squared test. *Journal of the American Statistical Association*, **96**, 996-1003.

[4] Cheng, P. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, **89**, 81-87.

[5] Chown, J. and Müller, U. U. (2013). Efficiently estimating the error distribution in nonparametric regression with responses missing at random. *Journal of Nonparametric Statistics*, **25**, 665-677.

[6] Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*. New York: Wiley.

[7] Cook, R. D., and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, **104**, 197-208.

[8] Cook, R. D., and Lee, H. (1999). Dimension reduction in binary response regression. *Journal of the American Statistical Association*, **94**, 1187-1200.

[9] Cook, R. D., and Li, B. (2002). Dimension reduction for conditional mean in regression. *Annals of Statistics*, **30**, 455-474.

[10] Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, **100**, 410-428.

[11] Cook, R. D., and Weisberg, S. (1991). Discussion of "Sliced inverse regression for dimension reduction," by K. C. Li, *Journal of the American Statistical Association*, **86**, 316–342.

[12] Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimators. *Annals of Statistics*, **27**, 1012-1050.

[13] Dette, H. (2002). A consistent test for heteroscedasticity in nonparametric regression based on the kernel method. *Journal of Statistical Planning and Inference*, **103**, 311-329.

[14] Dette, H. and Hildebrandt, T. (2012). A note on testing hypotheses for stationary processes in the frequency domain. *Journal of Multivariate Analysis*, **104**, 101-114.

[15] Dette, H., Neumeyer, N., Van Keilegom, I. (2007). A new test for the parametric form of the variance function in nonparametric regression. *Journal of the Royal Statistical Society: Series B*, **69**, 903-971.

[16] Dette, H. and Spreckelsen, I. (2003). A note on a specification test for time series models based on spectral density estimation. *Scandinavian Journal of Statistics*, **30**, 481-491.

[17] Dette, H. and Spreckelsen, I. (2004). Some comments on specification tests in nonparametric absolutely regular processes. *Journal of Time Series Analysis*, **25**, 159-172.

[18] Dette, H. and von Lieres und Wilkau, C. (2001). Testing additivity by kernel-based methods-what is a reasonable test? *Bernoulli*, **7**, 669-697.

[19] Ding, X. B., and Wang, Q. H. (2011). Fusion-refinement procedure for dimension reduction with missing response at random. *Journal of the American Statistical Association*, **106**, 1193-1207.

[20] Eubank, R. L., Li, C. S. and Wang, S. J. (2005). Testing lack-of-fit of parametric regression models using nonparametric regression techniques, *Statistica Sinica*, **15**, 135-152.

[21] Fan, J. and Huang, L. (2001). Goodness-of-fit tests for parametric regression models, *Journal of the American Statistical Association*, **96**, 640-652.

[22] Fan, J. and Jiang, J. (2005). Nonparametric inference for additive models. *Journal of the American Statistical Association*, **100**, 890-907.

[23] Fan, J. and Jiang, J. (2007). Nonparametric inference with generalized likelihood ratio tests. *Test*, **16**, 409-444.

[24] Fan, J., Zhang, C. and Zhang, J. (2001) Generalized likelihood ratio statistics and Wilks phenomenon. *Annals of Statistics*, **29**, 153-193.

[25] Fan, Y., Li, Q., (1996). Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica*, **64**, 865-890.

[26] Fan, Y., Li, Q. (2000). Consistent Model Specification Tests: Kernel-Based Tests Versus Bierens' ICM Tests. *Econometric Theory*, **16**, 1016-1041.

[27] Ferré, L. (1998). Determining the dimension in sliced inverse regression and related methods. *Journal of the American Statistical Association*, **93** 132-140.

[28] Gao, J., Wang, Q., and Yin, J. (2011). Specification testing in nonlinear time series with long-range dependence. *Econometric Theory*, **27**, 260-284.

[29] González-Manteiga, W. and Cao, R. (1993) Testing the hypothesis of a general linear model using nonparametric regression estimation. *Test*, **2**, 161-188.

[30] González-Manteiga, W. and Crujeiras, R. M. (2013). An updated review of Goodness-of-Fit tests for regression models. *Test*, **22**, 361-411.

[31] González-Manteiga, W. and Pérez-González, A. (2006). Goodness-of-fit tests for linear regression models with missing response data. *Canadian Journal of Statistics*, **34**, 149-170.

[32] Guo, X. and Xu, W. L. (2012). Goodness-of-fit tests for general linear models with covariates missed at random. *Journal of Statistical Planning and Inference*, **142**, 2047-2058.

[33] Guo, X., Xu, W. L. and Zhu, L. X. (2014). Model checking for parametric regressions with response missing at random. *Annals of the Institute Statistical Mathematics*, accepted.

[34] Hall, P. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *Journal of Multivariate Analysis*, **14**, 1-16.

[35] Hall, P., Wolff, R. C. L., and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, **94**, 154-163.

[36] Hammer, S. M., et al. (1996). A Trial Comparing Nucleotide Monotherapy With Combined Therapy in HIV-Infected Adults With CD4 Cell Counts From 200 to 500 per Cubic Millimeter, *New England Journal of Medicine*, **335**, 1081-1090.

[37] Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, **21**, 157-178.

[38] Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Annals of Statistics*, **21**, 1926-1947.

[39] Hart, J. (1997). *Nonparametric smoothing and lack-of-fit tests*. Springer, Berlin.

[40] Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, **29**, 595-623.

[41] Hsing, T., and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics*, **20**, 1040-1061.

[42] Hu, Z. H., Follmann, D., Qin, J. (2010). Semiparametric dimension reduction estimation for mean response with missing. *Biometrika,* **97**, 301-319.

[43] Huskova, M. and Meintanis, S. (2009). Goodness-of-fit tests for parametric regression models based on empirical characteristic functions. *Kybernetika*, **45**, 960-971.

[44] Huskova, M. and Meintanis, S. (2010). Test for the error distribution in non-parametric possibly heterocedastic regression models. *Test*, **19**, 92-112.

[45] Khmadladze, E. V., Koul, H. L. (2004). Martingale transforms goodness-of-fit tests in regression models. *Annals of Statistics*, **37**, 995-1034.

[46] Kim, J. K. and Yu, C. L. (2011). A Semiparametric Estimation of Mean Functionals With Nonignorable Missing Data. *Journal of the American Statistical Association*, **106**, 157-165.

[47] Koul, H. L. and Ni, P. P. (2004). Minimum distance regression model checking. *Journal of Statistical Planning and Inference*, **119**,109-141.

[48] Koul, H. L., Müller, U. U. and Schick, A. (2012). The transfer principle: a tool for complete case analysis. *Annals of Statistics*, **40**, 3031-3049.

[49] Lavergne, P. and Vuong, Q. H. (2000). Nonparametric significance testing. *Econometric Theory*, **16**, 576-601.

[50] Li, B. and Dong, Y. (2009). Dimension reduction for non-elliptically distributed predictors. *The Annals of Statistics*, **37**, 1272-1298.

[51] Li, B., and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, **102**, 997-1008.

[52] Li, B., Wen, S. Q. and Zhu, L. X. (2008). On a Projective Resampling method for dimension reduction with multivariate responses. *Journal of the American Statistical Association*, **103**, 1177-1186.

[53] Li, B., Zha, H., and Chiaromonte, F. (2005). Contour Regression: A General Approach to Dimension Reduction, *Annals of Statistics*, **33**, 1580-1616.

[54] Li, K. C. (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86**, 86, 316–327.

[55] Li, L., and Lu, W. (2008). Sufficient dimension reduction with missing predictors. *Journal of the American Statistical Association*, **103**, 822-831.

[56] Li, X. Y. (2012). Lack-of-fit testing of regression model with response missing at random. *Journal of Statistical Planning and Inference*, **142**, 155-170.

[57] Li, Y. X. and Zhu, L. X. (2007). Asymptotics for sliced average variance estimation, *Annals of Statistics*, **35**, 41-69.

[58] Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.

[59] Lopez, O. and Patilea, V. (2009). Nonparametric lack-of-fit tests for parametric mean-regression models with censored data, *Journal of Multivariate Analysis*, **100**, 210-230.

[60] Manteiga, G. W. and González, P. A. (2006). Goodness-of-fit tests for linear regression models with missing response data. *Canadian Journal of Statistics*, **34**, 149-170.

[61] Müller, U. U. and Van Keilegom, I. (2012). Efficient parameter estimation in regression with missing responses. *Electronic Journal of Statistics*, **6**, 1200-1219.

[62] Pollard, D. (1984). *Convergence of Stochastic Processes*, Springer-Verlag, New York.

[63] Quinlan, R. (1993). Combining Instance-Based and Model-Based Learning. *In Proceedings on the Tenth International Conference of Machine Learning*, 236-243, University of Massachusetts, Amherst. Morgan Kaufmann.

[64] Rao, J. N. K. (1996). On variance estimation with imputed survey data (with discussion). *Journal of the American Statistical Association*, **91**, 499-520.

[65] Robins, J. M., Rotnitzky, A., and Zhao, L. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89**, 846-866.

[66] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

[67] Schott, J. R. (1994). Determining the dimensionality in sliced inverse regression. *Journal of the American Statistical Association*, **89**, 141-148.

[68] Sperlich, S. (2013). On the choice of regularization parameters in nonparametric specification testing. *Empirical Economics*, to appear.

[69] Stute, W. (1997). Nonparametric model checks for regression. *Annals of Statistics*, **25**, 613-641.

[70] Stute, W. and Zhu, L. X. (2002). Model checks for generalized linear models. *Scandinavian Journal of Statistics*, 29, 535-546.

[71] Stute, W., Gonzáles-Manteiga, W. and Presedo-Quindimil, M. (1998a). Bootstrap approximation in model checks for regression. *Journal of American Statistical Association*, 93, 141-149.

[72] Stute, W., Thies, S. and Zhu, L. X. (1998b). Model checks for regression: An innovation process approach. *Annals of Statistics*, **26**, 1916-1934.

[73] Stute, W., Xu, W. L. and Zhu, L. X. (2008). Model diagnosis for parametric regression in high-dimensional spaces. *Biometrika*, **95**, 451-467.

[74] Su, J. Q., Wei, L. J. (1991). A lack of fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association*, **86**, 420-426.

[75] Sun, Z. H. and Wang, Q. H., (2009). Checking the adequacy of a general linear model with responses missing at random. *Journal of Statistical Planning and Inference*, **139**, 3588-3604.

[76] Sun, Z., Wang, Q. and Dai, P. (2009). Model checking for partially linear models with missing responses at random. *Journal of Multivariate Analysis*, **100**, 636-651.

[77] Van Keilegom, I., Gonzáles-Manteiga, W. and Sánchez Sellero, C. (2008). Goodness-of-fit tests in parametric regression based on the estimation of the error distribution. *Test*, **17**, 401-415.

[78] Velilla, S. (1998). Assessing the number of linear components in a general regression problem. *Journal of the American Statistical Association*, **93**, 1088-1098.

[79] Wang, C. Y., Wang, S. J., Zhao, L. P., and Ou, S. T. (1997). Weighted semiparametric estimation in regression analysis regression with missing covariates data. *Journal of the American Statistical Association*, **92**, 512-525.

[80] Wang, D., and Chen S. X. (2009). Empirical likelihood for estimating equation with missing values. *The Annals of Statistics*, **37**, 490-517.

[81] Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, **103**, 811-821.

[82] Wang, Q. H., Linton, O., and Härdle, W. (2004). Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association*, **99**, 334-345

[83] Wang, Q. H., and Rao, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *The Annals of Statistics*, **30**, 896-924.

[84] Wang, Q. and Yin, X. R. (2008). A nonlinear multi-dimensional variable selection method for high dimensional data: Sparse MAVE. *Computational Statistics & Data Analysis*, **52**, 4512-4520.

[85] White, H. (1981). Consequences and detection of misspecified nonlinear regression models. *Journal of the American Statistical Association*, **76**, 419-433.

[86] Wu, C. F. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Annals of Statistics*, **14**, 1261-1295.

[87] Xia, Y. C. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, **22**, 1112-1137.

[88] Xia, Y. (2007). A constructive approach to the estimation of dimension reduction directions. *The Annals of Statistics*, **35**, 2654-2690.

[89] Xia, Y. C., Tong, H., Li, W. K. and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B*, **64**, 363-410.

[90] Xu, W. L., Guo, X. and Zhu, L. X. (2012). Goodness-of-fitting for partial linear model with missing response at random. *Journal of Nonparametric Statistics*, **24**, 103-118.

[91] Xue, L. G. (2009). Empirical likelihood for linear models with missing responses. *Journal of Multivariate Analysis*, **100**, 1353-1366.

[92] Yates, F. (1933). The analysis of replicated experiments when the field results are incomplete. *Empire Journal of Experimental Agriculture*, **1**, 129-142.

[93] Yin, X. and Cook, R. D. (2005). Direction estimation in single-index regressions. *Biometrika*, **92**, 371-384.

[94] Yin, X., Li, B. and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a Multiple-index regression. *Journal of Multivariate Analysis*, **99**, 1733-1757.

[95] Yin, X. R. and Li, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *The Annals of Statistics*, **39**, 3392-3416.

[96] Zhao, H., Zhao, P. Y., Tang, N. S. (2013). Empirical likelihood inference for mean functionals with nonignorably missing response data. *Computational Statistics and Data Analysis*, **66**, 101-116.

[97] Zhao, L. P., Lipsitz, S., and Lew, D. (1996). Regression analysis with missing covariate data using estimating equations. *Biometrics*, **52**, 1165-1182.

[98] Zhang, C., Dette, H. (2004). A power comparison between nonparametric regression tests. *Statistics and Probability Letters*, **66**, 289-301.

[99] Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, **75**, 263-289.

[100] Zhu, L. P., Wang, T., and Zhu, L. X. (2012). Sufficient dimension reduction in regression with missing predictors. *Statistica Sinica*, **22**, 1611-1637.

[101] Zhu, L. P., Wang, T., Zhu, L.X. and Ferré, L. (2010). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, **97**. 295-304.

[102] Zhu, L. P., Zhu, L. X. , and Feng, Z. H. (2010). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association* , **105**, 1455-1466.

[103] Zhu, L. P., and Zhu, L. X. (2007). On kernel method for sliced average variance estimation. *Journal of Multivariate Analysis*, **98**, 970-991.

[104] Zhu, L. P., and Zhu, L. X. (2009). On distribution-weighted partial least squares with diverging number of highly correlated predictors, *Journal of the Royal Statistical Society: Series B*, **71**, 525–548.

[105] Zhu, L. X. (2005). *Nonparametric Monte Carlo tests and their applications*. Springer, New York.

[106] Zhu, L. X. and Fang, K. T. (1996). Asymptotics for the kernel estimates of sliced inverse regression. *Annals of Statistics*, **24**, 1053-1067.

[107] Zhu, L. X., Miao, B. Q., and Peng, H. (2006). Sliced inverse regression with large dimensional covariates. *Journal of the American Statistical Association*, **101**, 630-643.

[108] Zhu, L. X. and Ng, K. (1995). Asymptotics for Sliced Inverse Regression. *Statistica Sinica*, **5**, 727- 736.

[109] Zhu, L. X. and Ng, K. W. (2003). Checking the adequacy of a partial linear model, *Statistica Sinica*, **13**, 763-781.

[110] Zhu, L. X. and Neuhaus, G. (2000). Nonparametric Monte Carlo tests for multivariate distributions. *Biometrika*, **87**, 919-928.

[111] Zhu, Y. and Zeng, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *Journal of the American Statistical Association*, **101**, 1638-1651.

Curriculum Vitae

GUO Xu

Academic qualifications of the thesis author, Mr. GUO Xu:

• Received the degree of Bachelor of Science (Mathematics) from China Agricultural University, Jul. 2009.

• Received the degree of Master of Science (Probability and Mathematical Statistics) from Renmin University of China, Jul. 2012.

August 2014