2016

Model-adaptive tests for regressions

Xuehu Zhu
Hong Kong Baptist University

Model-adaptive Tests for Regressions

ZHU Xuehu

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 2015

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 or dissertation submitted to this or any other institution for a degree, diploma or other qualifications.

Signature: _______________

Date: August 2015

ABSTRACT

In this thesis, we firstly develop a model-adaptive checking method for partially parametric single-index models, which combines the advantages of both dimension reduction technique and global smoothing tests. Besides, we propose a dimension reduction-based model adaptive test of heteroscedasticity checks for nonparametric and semi-parametric regression models. Finally, to extend our testing approaches to nonparametric regressions with some restrictions, we consider significance testing under a nonparametric framework.

In Chapter 2, **"Model Checking for Partially Parametric Single-index Models: A Model-adaptive Approach"**, we consider the model checking problems for more general parametric models which include generalized linear models and generalized nonlinear models. We develop a model-adaptive dimension reduction test procedure by extending an existing directional test. Compared with traditional smoothing model checking methodologies, the procedure of this test not only avoids the curse of dimensionality but also is an omnibus test. The resulting test is omnibus adapting the null and alternative models to fully utilize the dimension-reduction structure under the null hypothesis and can detect fully nonparametric global alternatives, and local alternatives distinct from the null model at a convergence rate as close to square root of the sample size as possible. Finally, both Monte Carlo simulation studies and real data analysis are conducted to compare with existing tests and illustrate the finite sample performance of the new test.

In Chapter 3, **"Heteroscedasticity Checks for Nonparametric and Semi-parametric Regression Model: A Dimension Reduction Approach"**, we consider heteroscedasticity checks for nonparametric and semi-parametric regression models. Existing local smoothing tests suffer severely from the curse of dimensionality even when the number of covariates is moderate because of use of nonparametric estimation. In this chapter, we propose a dimension reduction-based model adaptive test that behaves like a local smoothing test as if the number of covariates is equal to the number of their linear combinations in the mean regression function, in particular,

equal to 1 when the mean function contains a single index. The test statistic is asymptotically normal under the null hypothesis such that critical values are easily determined. The finite sample performances of the test are examined by simulations and a real data analysis.

In Chapter 4, **"Dimension Reduction-based Significance Testing in Nonparametric Regression"**, as nonparametric techniques need much less restrictive conditions than those required for parametric approaches, we consider to check nonparametric regressions with some restrictions under sufficient dimension reduction structure. A dimension-reduction-based model-adaptive test is proposed for significance of a subset of covariates in the context of a nonparametric regression model. Unlike existing local smoothing significance tests, the new test behaves like a local smoothing test as if the number of covariates is just that under the null hypothesis and it can detect local alternative hypotheses distinct from the null hypothesis at the rate that is only related to the number of covariates under the null hypothesis. Thus, the curse of dimensionality is largely alleviated when nonparametric estimation is inevitably required. In the cases where there are many insignificant covariates, the improvement of the new test is very significant over existing local smoothing tests on the significance level maintenance and power enhancement. Simulation studies and a real data analysis are conducted to examine the finite sample performance of the proposed test.

Finally, we conclude the main results and discuss future research directions in Chapter 5.

Keywords: Model checking; Partially parametric single-index models; Central mean subspace; Central subspace; Partial central subspace; Dimension reduction; Ridge-type eigenvalue ratio estimate; Model-adaption; Heteroscedasticity checks; Significance testing.

ACKNOWLEDGEMENTS

First and foremost, I would like to take this opportunity to express my greatest appreciation to my supervisor Prof. ZHU Lixing, who patiently taught me numerous methods in the fields of statistics, mathematics and research. Besides, his diverse excellent qualities really affect me a lot. Without his support and guidance, this work would not have been completed. Besides, I would like to thank my co-supervisor Dr. PENG Heng, for his invaluable and irreplaceable advices and helpful comments.

Secondly, I am really appreciated for the continuous guidance and encouragement of Prof. LIN Lu in Institute for Financial Studies, Shandong University.

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

Fourthly, I would like to give my thanks to so many researchers who have helped and guided me both professionally and personally in the last few years. Prof. TONG Xingwei, Dr. TONG Tiejun and Dr. LI Gaorong were supportive and instructive in my research career. I am also very grateful to Prof. ZHU Liping, Dr. XU Wangli, Dr. WU Jianhong, Dr. LI Zaixing, Dr. WU Ping, Dr. FANG Yun, Dr. YU Zhou, Dr. FENG Zhenghui, Dr. ZHANG Jun, Dr. FAN Yan, Dr. WANG Tao, Dr. XU Peirong, Dr. GUO Xu, Dr. WANG Cheng, Dr. XIA Qiang, Dr. ZHOU Jingke, Mr. XIE Chuanlong and all the people of the Lixing's research team. Especially, I would give my special thanks to Mrs. TIAN Qiushi for her tremendous assistance and encouragement throughout my graduate career.

Last but not least, I would like to owe my loving thanks to my family and my girlfriend for their constant support and encouragement, both emotionally and intellectually.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Goodness-of-fit proposed by Pearson at the beginning of the twentieth century is one of the most important areas in economics and statistics. MathScinet database reported that there have been nearly 3000 references about goodness-of-fit until October 2012 (González-Manteiga and Crujeiras, 2013). There is no doubt that goodness-of-fit for regression is an important research topic in economics and statistics. This thesis focuses primarily on three kinds of models: the parametric models, the semi-parametric models and the nonparametric models with some restrictions. Study of these models includes model checks and heteroscedasticity checks.

We always like to explore the relationship between a response Y and a vector of predictors $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$. Due to a better interpretability and avoiding the curse of dimensionality usually occurred in nonparametric regression models, parametric regression models, especially generalized linear models, are commonly adopted. We are also found of whether a subset of explanatory variable X is the redundant variables to the response variable Y or not. Broadly speaking, studying interesting data structures is an important problem in economics and statistics. However, whether those structures hold or not are not all so clear-cut in practice. Thus, it is essential to develop some suitable testing procedures to check these structures.

Additionally, when we detect these structures with a large number of the predictors, the curse of dimensionality appears naturally in checking procedures. Therefore, when the dimension of X is large, overwhelming testing methodologies fail to work. To overcome this serious problem caused by dimensionality, it is necessary to adopt some linear combinations of the predictors X to summarize all information of X over Y as well as possible, which can be completed by dimension reduction technique. This thesis focuses on checking these structures by combining strongpoints of both dimension reduction technique and traditional testing methodologies.

As two leitmotifs of statistics, goodness-of-fit and dimension reduction are becoming progressively more and more prominent, both in theory and in practice.

1.1 Goodness-of-Fit for Regression

The purpose of goodness-of-fit for regression is to check whether the mean function for a set of data belongs to a certain parametric form, such as linear models and generalized linear models. If the certain parametric form for the mean function is not clear-cut, nonparametric regression models will be adopted. When the number of predictors X is large, nonparametric estimators, such as Nadaraya-Watson estimator (Nadaraya 1964, Watson 1964), are unsuitable, even they are not consistent for the mean function. Additionally, it is difficult to get the interpretability for nonparametric estimators. However, the parametric estimators have several advantages as follows: (1) efficient estimators, (2) better interpretability for the parametric regression. Furthermore, statistical inferences are built on right regression models. If there is no any priori information or without statistical evidence to support the assumption, the inferences based on them maybe very skeptical. Therefore, it is essential to implement some testing procedure to check whether some parametric form is adequate to fit the data or not.

1.1.1 Smoothing-Based Tests

When a sample $\{x_i, y_i\}_{i=1}^n$ of (X, Y) is available, we consider a regression model as

$$y_i = m(x_i) + \epsilon_i,$$

where $m(\cdot) = E(Y|X = \cdot)$ is the mean function of the response Y over the predictors X and ϵ_i is the residual term such that $E(\epsilon_i|x_i) = 0$. We are also found of exploring the structure of the mean regression $m(\cdot)$. Thus, it is our purpose to check

$$H_0: \ m(X) \in \mathcal{M}_\theta = \{m(X, \theta), \theta \in \Theta \subset R^q\} \ \text{ versus } \ H_1: \ m(X) \notin \mathcal{M}_\theta.$$

Although there are many specific smoothing methodologies for mean function in the literature, kernel type estimators will be considered, for example, Nadaraya-Watson estimator (Nadaraya 1964, Watson 1964) defined as

$$\hat{m}(x) = \sum_{i=1}^n \omega_{ni}(x) y_i \ \text{ with } \ \omega_{ni}(x) = \frac{K_h(x - x_i)}{\sum_{j=1}^n K_h(x - x_j)},$$

where $K_h(\cdot) = K(\cdot/h)/h^p$, $K(\cdot)$ is a kernel function and h is a bandwidth.

Härdle and Mammen (1993) developed a test statistic relied on the L_2 distance between the underlying and the hypothetical mean function. To be exact, the test statistic can be written as follows:

$$T_{HM} = \int \left(\sum_{i=1}^n \omega_{ni}(x)[y_i - m(x_i, \theta_n)] \right)^2 W(x) dx,$$

where θ_n is, under the null hypothesis, a root-n consistent estimator of the true value θ_0, and it is gotten simply by the nonlinear least squares estimation, and $W(\cdot)$ is some positive weight function. Under some regularity conditions, the limiting null distribution of T_{HM} can be described as follows:

$$nh^{p/2} \left(T_{HM} - (nh^p)^{-1} \int K^2(x) dx \int \frac{\sigma^2(x) W(x)}{f(x)} dx \right)$$
$$\xrightarrow{\text{d}} N \left(0, 2 \int (K * K)^2 dx \int \frac{\sigma^4(x) W^2(x)}{f^2(x)} dx \right),$$

where the notation $\xrightarrow{\text{d}}$ stands for convergence in distribution, $f(\cdot)$ denotes the density function of the predictors X, $\sigma^2(x) = Var(Y|X = x)$ indicates the conditional variance and the notation $K * K$ stands for the self-convolution of the kernel function. González-Manteiga and Cao (1993) further extended Härdle and Mammen (1993)'s test to a discretized version as:

$$T_{GC} = n^{-1} \sum_{i=1}^{n} \omega_{ni}(x_i)[y_i - m(x_i, \theta_n)]^2 W(x_i).$$

Under the null hypothesis, this discrete version T_{GC} is a consistent estimator of $E[E^2(\epsilon_0|X)W(X)]$ where $\epsilon_0 = Y - m(X, \theta_0)$.

Zheng (1996) and Fan and Li (1996) independently developed nonparametric tests based on second order conditional moments. Particularly, Fan and Li (1996) constructed a test statistic depended on a consistent estimator of $E[\epsilon_0 f(X) E\{\epsilon_0 f(X)|X\} f(X)]$ under the null hypothesis. Naturally, this quantity can be estimated by

$$T_{FL} = \frac{1}{n(n-1)} \sum_{i \neq j} K_h(x_i - x_j)[\hat{\epsilon}_i \hat{f}(x_i)][\hat{\epsilon}_j \hat{f}(x_j)],$$

where $\hat{\epsilon}_i = y_i - m(x_i, \theta_n)$ and $\hat{f}(x_i)$ is a kernel estimator of $f(x_i)$. Fan and Li (1996) described its null limited distribution as follows:

$$n h^{p/2} T_{FL} \xrightarrow{\text{d}} N\left(0, \sigma^2\right),$$

with $\sigma^2 = 2 \int K^2(u)du \int \sigma^4(x) f^5(x)dx$ which can be consistently estimated as

$$\hat{\sigma}^2 = \frac{1}{n(n-1)} \sum_{i \neq j} K_h(x_i - x_j)[\hat{\epsilon}_i \hat{f}(x_i)]^2 [\hat{\epsilon}_j \hat{f}(x_j)]^2 \int K^2(x)dx.$$

Compared T_{FL} with T_{HM} and T_{GC}, it is preferable that there is no asymptotic bias for T_{FL} and then it is unnecessary for bias-correction procedure.

Additionally, since under the null hypothesis, $E\Big(\{\epsilon_0^2 - [\epsilon_0 - E(\epsilon_0|X)]^2\} W(x)\Big) = 0$ holds, Dette (1999) developed a test statistic based on the differences between the underlying and the hypothetical variance functions. A natural estimator of this quantity can be formulated as:

$$T_D = \frac{1}{n} \sum_{i=1}^{n} [y_i - m(x_i, \theta_n)]^2 W(x_i) - \frac{1}{n} \sum_{i=1}^{n} [y_i - \hat{m}(x_i)]^2 W(x_i).$$

Dette (1999) described the limiting distribution of T_D to be:

$$nh^{p/2}\left(T_D - (nh^p)^{-1}K^*(0)\int \sigma^2(x)W(x)dx\right)$$
$$\xrightarrow{\mathrm{d}} N\left(0, 2\int K^*(x)dx \int \sigma^4(x)W^2(x)dx\right).$$

where $K^* = 2K - K * K$. The more details about the comparison for above methods can be kindly referred to Zhang and Dette (2004).

Motivated by the conventional likelihood ratio test, Fan et al. (2001) and Fan and Jiang (2005, 2007) revised this test statistic to be generalized likelihood ratio tests for regression models. One of the major advantages for this type of tests is that their limiting distributions do not lie on nuisance parameters, see, Fan and Jiang (2005, 2007).

1.1.2 Tests Based on Empirical Processes

Another popular category of methodologies, motivated by the goodness-of-fit for distributions, is stemmed from the fact that under the null hypothesis, $E(Y - m(X,\theta)I(X \leq t) = 0$ for all $t \in R^p$. Thus, an empirical process can be defined as the following form:

$$V_n(x) = \sqrt{n}\{\mathcal{G}_n(x) - E_{\theta_n}(\mathcal{G}_n(x))\} = \sqrt{n}\sum_{i=1}^{n} \hat{\epsilon}_i I(x_i \leq x).$$

Here $E_{\theta_n}(\mathcal{G}_n(x))$ is a parametric estimator of the integrated mean function $\mathcal{G}(x) = \int_{-\infty}^{x} m(t)dF(t) = E(YI(X \leq x))$ with $I(\cdot)$ being the indicator function under the null hypothesis and $\hat{\epsilon}_i = y_i - m(x_i, \theta_n)$ with θ_n, under the null hypothesis, being a root-n consistent estimator of θ. The empirical process $V_n(x)$ can be viewed as the base to construct Cramér-von Mises or Kolmogorov-Smirnov type tests. The limiting distributions for this kind of tests rely on a Gaussian limit process that the empirical process $V_n(x)$ converges to. This type of tests was studies originally by Bierens (1982), Su and Wei (1991) and Stute (1997). On the other hand, Koul and Stute (1998) and Diebolt (1995) considered goodness-of-fit for regressions with non-random design relied on the empirical process. It is noteworthy to point out that the limiting null distributions of these tests depend on the parametric function of $m(X, \theta)$. Thus, these methods are not distribution free. Using ideas of Khmaladze (1981), Stute et al. (1998) proposed an innovation martingale approach for model checking to acquire a distribution free test for one-dimension predictor. Stute and Zhu (2002) developed a relevant asymptotically distribution-free test based on one-dimensional projected covariates to check generalized linear regression models. Khmaladze and Koul (2004) further studied the goodness-of-fit problem for errors when the Khmaladze martingale transformation of the empirical processes is applied to nonparametric regression.

Van Keilegom et al. (2008) developed a test statistic that is the distance between the empirical versions of the underlying and the hypothetical distribution functions. Huskova and Meintanis (2009) advised a test statistic based on the distance between the empirical characteristic function of the residuals under the underlying and the non-parametric models. It is essential to point out that both the tests are only suitable for both predictors and the error term to be independent. However, other relevant proposals principally suppose that $E(\epsilon|X) = 0$. Other relevant references include Dette et al. (2007) and Huskova and Meintanis (2010).

Compared with existing smoothing-based tests, this type of methodologies can detect local alternative hypotheses distinct from the null hypothesis at typical convergence rate $O(n^{-1/2})$, while smoothing-based tests can acquire the typical convergence rate $(nh^{p/2})^{-1/2}$. Another advantage for this type of these tests is not subjected to the selection of tuning parameters, such as bandwidth. Thus, these methodologies have theoretical advantages over existing local smoothing tests. It only implies that these methodologies may have some power against the closer alternative hypotheses. Actually, as Fan and Li (2000) pointed out, tests based on empirical processes commonly produce low powers against high frequency alternative hypotheses. Furthermore, practical evidence shows that the high-dimensional predictors X still deteriorate the power of global smoothing tests. Additionally, when the dimension of X is very large, the computational burden is also an issue because these tests concentrated on the empirical process are generally computationally intensive.

A comprehensive review of goodness-of-fit for regressions is González-Manteiga and Crujeiras (2013). It is also needed to point out the book of Hart (1997), which collects some classical testing process for the goodness-of-fit problem, for example, the likelihood ratio tests, the von Neumann's test, and some nonparametric smoothing methods used for the goodness-of-fit tests of parametric models.

1.2 Overview of Sufficient Dimension Reduction

Since the sparsity of data structure restricts the application of local smoothing testing methods with high-dimensional predictors, the curse of dimensionality appears and then most testing methods can not efficiently work. Dimension reduction is an important technique to overcome the dimensionality problem for us in high-dimensional data analysis. Dimension reduction contains sufficient dimension reduction (SDR) and numerous methodologies which reduce the dimension of the original predictors X, such as partial least squares and principal component analysis. The purpose of sufficient dimension reduction handles the sparsity of data structure without any pre-specified parametric model structure and without losing any information about the regression of Y over X. The reduction structure is derived by projecting raw predictors onto a lower-dimensional subspace.

Let $\mathcal{S}_{Y|X}$ stand for the central subspace introduced by Cook (1998) and Yin et al. (2008), which is defined as intersection of all subspaces S satisfying

$$Y \perp\!\!\!\perp X | P_S X,$$

where $\perp\!\!\!\perp$ stands for the statistics independence and $P_{(\cdot)}$ indicates a projection operator with respect to the standard inner product. Additionally, the conditional mean $E(Y|X)$ also attracts much attention. Similarly as the definition of central subspace, let $\mathcal{S}_{E(Y|X)}$ denote the central mean subspace (CMS, Cook and Li 2002). It is actually the intersection of all subspaces S such that

$$Y \perp\!\!\!\perp E(Y|X)|P_{\mathcal{S}}X.$$

Under any circumstances, sufficient dimension reduction permits us to find $q \leq p$ new predictors which can be written as linear combinations of the original predictors: $\beta_1^\top X, \ldots, \beta_q^\top X$, where $\{\beta_1, \ldots, \beta_q\}$ is viewed as a basis of $\mathcal{S}_{Y|X}$ or $\mathcal{S}_{E(Y|X)}$.

In the last three decades or so, many promising methods have been developed to issues related to dimension reduction. Three prevalent classes of these estimation methods are respectively: 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 et al. 2010a), direct regression methods (e.g., Li 1992, Härdle et al. 1993, Hristache et al. 2001, Xia et al. 2002, Xia 2006 and Dalalyan et al. 2008) and correlation approaches such as Fourier method (Zhu and Zeng, 2006), KL-distance (Yin and Cook, 2005, Yin et al., 2008).

Inverse regression methods are very easy to implement and widely applied. But those methods need strong conditions over the predictors, for example, a linearity condition (Li 1991) or a constant conditional variance condition (Li 1992), and even they can not achieve a consistent estimator of the central subspace comprehensively (Cook 1998). On the contrary, direct regression methods require weaker conditions about the distribution of the predictors and perform much better in finite samples. However, these methods can not find other directions but the central mean subspace and have heavy computational burden.

1.2.1 Sliced Inverse Regression

Sliced inverse regression (SIR) method, which was developed by Li (1991), is an innovative idea for obtaining $\mathcal{S}_{Y|X}$ in regression. This promising method implicates the inverse regression that X is regressed on Y. To derive SIR estimator, it is usually assumed that the predictors of X satisfy the linearity condition, that is, the conditional mean $E(X|B^\top X)$ is a linear function of $B^\top X$ with the columns of $B \in \mathbb{R}^{p\times q}$

being any basis of the central space $\mathcal{S}_{Y|X}$. Define the standardized predictors X as the new predictors $Z = \Sigma_X^{-1/2}(X - u_X)$, where u_X and Σ_X are the mean and the non-singular covariance matrix of X, respectively. Under this linearity condition, Cook (1998) has justified that $\mathcal{S}_{Y|X} = \Sigma_X^{-1/2}\mathcal{S}_{Y|Z}$ in Proposition 6.1. SIR is due to the fact that for any y, $E(Z|Y = y) \subseteq \mathcal{S}_{Y|Z}$. Define the inverse mean subspace $\mathcal{S}_{E(Z|Y)} = \mathrm{Span}\{\mathrm{E(Z|Y = y), y \in \mathbb{R}}\}$. Then, under the coverage condition $\mathcal{S}_{E(Z|Y)} = \mathcal{S}_{Y|Z}$, the eigenvectors associated with non-zero eigenvalues of $\mathrm{Cov}\{\mathrm{E(Z|Y)}\}$ make up the basis of $\mathcal{S}_{Y|Z}$.

When an i.i.d sample $\{x_i, y_i\}_{i=1}^n$ of (X, Y) is available, from Li (1991), the SIR algorithm is as following steps. Let sample mean and sample covariance matrix of X be denoted as $\bar{X}$ and $\hat{\Sigma}_X$, respectively. Firstly, we partition the range of Y into M intervals, $I_1, \cdots, I_M$, $\mathrm{Cov}\{\mathrm{E(Z|Y)}\}$ is approximately estimated by

$$\widehat{\mathrm{Cov}}\{E(Z|Y)\} = \sum_{k=1}^{M} \hat{p}_k \bar{z}_k \bar{z}_k^\top, \tag{1.1}$$

where for $k = 1, \cdots, M$, $\hat{p}_k$ and $\bar{z}_k$ denote the estimators of $p_k = \mathrm{Pr(Y \in I_k)}$ and $z_k = E(Z|Y \in I_k)$, respectively. Simply, $\hat{p}_k = n_k/n$ with n_k being the number of observations falling in interval I_k and $\bar{z}_k$ can be estimated by the sample mean of Z in each interval I_k, $k = 1, \ldots, M$.

Computing the spectral decomposition of the estimating matrix $\widehat{\mathrm{Cov}}\{E(Z|Y)\}$, we get the eigenvectors $\hat{\alpha}_1, \cdots, \hat{\alpha}_q$ associated with the q largest eigenvalues. Hence, the estimating basis of $\mathcal{S}_{Y|X}$ is constructed as $\hat{\beta}_j = \hat{\Sigma}_X^{-1/2}\hat{\alpha}_j$ for $j = 1, \cdots, q$ and the new predictors are composed as $\hat{\beta}_1^\top X, \cdots, \hat{\beta}_q^\top X$.

Under certain regularity conditions, Li (1991) and Hsing and Carroll (1992) have proved that the SIR estimator enjoys the root-n consistency to the center space $\mathcal{S}_{Y|X}$. Furthermore, Zhu and Ng (1995) has justified that this algorithm is not sensitive to the choice of the number of intervals in theoretical results, which is consistent with the numerical studies in Li (1991).

1.3 Outline of the Thesis

The plan of the remaining thesis is organized as follows. In Chapter 2, we consider the model checking problem for the general parametric models which include generalized linear models and generalized nonlinear models. We suggest a model-adaptive dimension reduction test procedure by extending an existing directional test. Compared with traditional smoothing model checking methodologies, this test not only avoids the curse of dimensionality but also is an omnibus test. The resulting test is omnibus adapting the null and alternative models to fully utilize the dimension-reduction structure under the null hypothesis and can detect fully nonparametric global alternatives and local alternatives distinct from the null hypothesis at a convergence rate as close to square root of the sample size as possible. Finally, both Monte Carlo simulation studies and real data analysis are conducted to compare with existing tests and illustrate the finite sample performance of the new test.

In the nonparametric and semi-parametric regression models, most methods have adverse consequences for the efficiency and can be even inconsistent in the presence of heteroscedasticity. Then heteroscedasticity testing is of importance in regression analysis. Thus, in Chapter 3, we consider heteroscedasticity checks for regression models. Existing local smoothing tests suffer severely from the curse of dimensionality even when the number of covariates is moderate because of use of nonparametric estimation. In this chapter, we propose a dimension reduction-based model adaptive test that behaves like a local smoothing test as if the number of covariates is equal to the number of their linear combinations in the mean regression function, in particular, equal to 1 when the mean function contains a single index. The test statistic is asymptotically normal under the null hypothesis such that critical values are easily determined. The finite sample performances of the test are examined by simulations and a real data analysis.

As nonparametric techniques need much less restrictive conditions than those required for parametric approaches, we consider to check nonparametric regressions with some restrictions under sufficient dimension reduction structure. In Chapter 4, we propose a novel model-adaptive test for the significance testing of a subset of explanatory variables in the context of a nonparametric regression model. Unlike existing local smoothing significance tests, the new test behaves like a local smoothing test as if the number of covariates is just that under the null hypothesis and it can detect local alternatives distinct from the null hypothesis at the rate that is only related to the number of covariates under the null hypothesis. Thus, the curse of dimensionality is largely alleviated when nonparametric estimation is inevitably required. In the cases where there are many insignificant covariates, the improvement of the new test is very significant over existing local smoothing tests on the significance level maintenance and power enhancement. Simulation studies and a real data analysis are conducted to examine the finite sample performance of the proposed test.

Lastly, in Chapter 5, we summarize the main results and discuss future research directions.

Chapter 2

Model Checking for Partially Parametric Single-index Models: A Model-adaptive Approach

2.1 Introduction

Consider the partially parametric single-index model in the form:

$$Y = G(\beta^\top X, W, \theta) + \epsilon, \tag{2.1}$$

where Y is the response variable, the covariate vector (X, W) is in $\mathbb{R}^{p_1+p_2}$, $G(\cdot)$ denotes a known smooth function that depends not only on $\beta^\top X$ but also on the covariate W, β and θ denote the regression parameter vectors and ϵ follows a continuous distribution and is independent with the covariates (X, W). The model (2.1) reduces to the parametric single-index model in the absence of the covariate W and the general parametric model in the absence of the covariate $\beta^\top X$. This structure is often meaningful as in many applications, p_1 is often large while p_2 is not. See the relevant dimension reduction literatures such as Feng et al. (2013).

However, it is not all so clear-cut to know whether real data set marries the above statistical formalization or not. It is worthwhile to perform a suitable and efficient model checking before any further statistical analysis. As we often have no idea about the model structure under alternative hypothesis and thus, the general alternative hypothesis is considered to have the form as:

$$Y = g(X, W) + \epsilon, \tag{2.2}$$

where $g(\cdot)$ denotes an unknown smooth function.

To test the parametric single-index model that absences the covariate W in the model (2.1), and general nonlinear model in the absence of the covariate $\beta^{\top}X$, there are several proposals in the literature. Two prevalent classes of methods are respectively local smoothing and global smoothing tests. As examples, for the former, Härdle and Mammen (1993) suggested a test that is the L_2 distance between the parametric estimate under the null and the nonparametric estimate under the alternative hypothesis. Zheng (1996) and Fan and Li (1996) independently developed tests based on second order conditional moments. Dette (1999) proposed a consistent test depended on the difference between the null parametric variance estimate and the alternative nonparametric variance estimate respectively. Fan et al. (2001) developed a generalized likelihood ratio test. For other developments, see the Neyman threshold test (Fan and Huang 2001), a class of minimum distance tests (Koul and Ni 2004) and the distribution distance test (Van Keilegom et al. 2008). A relevant reference can be seen in Zhang and Dette (2004). However, there exist two obvious shortcomings of local smoothing tests. First, those methodologies have the subjective constraint choice of tuning parameters such as bandwidth. In fact, as Stute and Zhu (2005) pointed out, the optimal bandwidth choice in hypothesis testing is still an open problem. Second, a more serious problem is the typical slow convergence rate of local smoothing tests that is $O(n^{-1/2}h^{-p/4})$ under the null hypothesis, where p is the dimension of the covariates and h is the bandwidth. In the present setup, $p = p_1 + p_2$. Because h must tend to zero at a certain rate, when the dimension p is large, the test statistics tend to their limits very slowly. In other words, local smoothing tests severely suffer from the curse of dimensionality.

For global smoothing tests, examples include Stute (1997) who proposed a nonparametric principal component decomposition-based test that is based on a residual marked empirical process. Stute et al. (1998) introduced the Khmaladze martingale transformation to model checking. Stute and Zhu (2002) developed a relevant asymptotically distribution-free test based on one-dimensional projected covariates to check generalized linear regression models. Khmaladze and Koul (2004) further studied the goodness-of-fit problem for errors when the Khmaladze martingale transformation of the empirical processes is applied to nonparametric regression. A comprehensive review of model checking is González-Manteiga and Crujeiras (2013). The typical convergence rate of existing global smoothing methods is $O(n^{-1/2})$. See e.g. Stute et al. (1998). Thus, it has the theoretical advantage over local smoothing tests. However, practical evidence shows that the high-dimensional covariate X still deteriorates the power of global smoothing tests when p is large. This is particularly the case when alternative model is high-frequent.

Stute and Zhu (2002) thus investigated the parametric single-index model: $Y = G(\beta^\top X)+\epsilon$. Stemming from the fact that under the null hypothesis, $E(Y-G(\beta^\top X)I(X \leq t) = 0$ for all $t \in R^p$ leads to $E(Y - G(\beta^\top X)I(\beta^\top X \leq t) = 0$ for all $t \in R^p$, Stute and Zhu (2002) advised a relevant test as:

$$R_n(x) = n^{-1/2}\sum_{i=1}^{n}(y_i - G(\hat{\beta}^\top x_i))I(\hat{\beta}^\top x_i \leq x),$$

where $\hat{\beta}$ is, under the null hypothesis, a root-n consistent estimate of β. It has been proved to be powerful in many cases. However, this test is not an omnibus test, while a directional test. Thus, the general alternative hypothesis of (2.2) cannot be detected. This phenomenon can be easily illustrated by the following alternative model: $Y = \beta_1^\top X + c\sin(\beta_2^\top X) + \epsilon$, where X is normally distributed $N(0, \Sigma)$ with $p \times p$ identity matrix Σ, and β_1 and β_2 are two orthogonal vectors. The value $c = 0$ corresponds to the null hypothesis. However, for any c, we always have $E(Y - \beta_1^\top X|\beta_1^\top X) = 0$. In other words, this conditional mean cannot distinguish the respective models under the null and alternative hypotheses.

On the other hand, the advantage of Stute and Zhu (2002)'s test under the null hypothesis is very important particularly in high-dimensional paradigms. Guo et al. (2014) recently proposed a model-adaptive dimension-reduction test for the model $Y = G(\beta^\top X, \theta) + \epsilon$ against the general alternative model $Y = g(X) + \epsilon$. The main idea is to fully utilize the dimension reduction structure about X under the null hypothesis as Stute and Zhu (2002) did, but to adapt the alternative model such that the test is still omnibus. Their test is based on local smoothing technique. The improvement over existing local smoothing tests is very significant. The test has a much faster convergence rate of $O(n^{-1/2}h^{-1/4})$ than the typical rate of $O(n^{-1/2}h^{-p/4})$ and can detect local alternatives distinct from the null hypothesis at the rate of $O(n^{-1/2}h^{-1/4})$ that is also much faster than the typical rate of $O(n^{-1/2}h^{-p/4})$. In other words, asymptotically, the test works as if the covariate X was univariate. Thus, this nice property of the model-adaptive dimension reduction test can significantly avoid the curse of dimensionality. The numerical studies in their paper also indicated its advantage in the cases with moderate sizes of sample. It is clearly of interest to see whether the idea of model-adaptivity can be applied to global smoothing test procedures such that a test can be more powerful than the classical global smoothing tests.

In the present chapter, we consider a more general alternative model as

$$Y = g(B^\top X, W) + \epsilon, \tag{2.3}$$

where B is a $p_1 \times q$ matrix with q orthogonal columns for an unknown number q with $1 \leq q \leq p_1$ and g is still an unknown smooth function. For identifiability consideration, assume that the matrix B satisfies $B^\top B = I_q$. This model covers many popularly used models in the literature such as the single-index models with $B = \beta$, the multi-index models with the absence of W, and partial single-index models with the mean function $g_1(\beta^T X) + g_2(W)$. β is considered to be a column of B. When $q = p_1$ and $B = I_{p_1}$, the model (2.3) is reduced to the usual alternative model (2.2). Actually, the model (2.2) can be rewritten in the form of (2.3). When

$q = p_1$, $g(X, W) = g(BB^\top X, W) \equiv: \tilde{g}(B^\top X, W)$, where B is any $p_1 \times p_1$ orthogonal matrix. This persuasively argues that the model (2.2) can be treated as a special case of (2.3). This understanding helps us construct an adaptive estimate of B that can be consistent under both the null and alternative hypotheses. Based on this, a test can be constructed by noticing that under the null hypothesis, $E(Y - g(\beta_1^\top X, W, \theta)I(B^\top X \leq t) = 0$ for all t and under the alternative hypothesis, it is nonzero for some t.

To define an empirical version of this quantity as a test statistic, the estimate of B should have an important feature as that in Guo et al. (2014). When the estimate can be automatically reduce to an estimate of β under the null hypothesis, a test that is based on this estimate can only rely on the dimension-reduced covariate $(\beta^T X, W)$ and is still omnibus to detect the general alternative (2.3). As was mentioned before, when W is absent, Guo et al. (2014)'s test has the adaptiveness property to the alternative model. To identify B and its structural dimension, some dimension reduction approaches such as minimum average variance estimation (MAVE, Xia et al. 2002) and discretization-expectation estimation (DEE, Zhu et al. 2010) were suggested. However, when W is present, these methods fail to work. Also, because of the existence of W, even when the dimension p_1 can be reduced to 1, their local smoothing test has a slower convergence rate of $O(n^{-1/2}h^{-(p_2+1)/4})$ where p_2 is the dimension of W.

In the present chapter, we develop an omnibus test of dimension reduction type that is based on the directional global smoothing test of Stute and Zhu (2002). To make B identifiable, we use the partial sufficient dimension reduction approach (Chiaromonte et al. 2002, Feng et al. 2013). Further, to estimate the structural dimension q of B, we suggest a ridge-type eigenvalue ratio estimate. The detail is presented in the next section.

The materials of the chapter are organized as follows. In Section 2.2, the partial discretization-expectation estimation is reviewed. A ridge-type eigenvalue ratio is suggested to determine the structural dimension q and its asymptotic properties are investigated. Based on these, a test is constructed in Section 2.3. The asymptotic properties under the null and local alternative hypotheses are also presented in this section. As the limiting null distribution is untractable, a Monte Carlo test procedure is described in Section 2.4. In Section 2.5, simulation results are reported and a real data analysis is conducted for illustration. Technical proofs are postponed to Appendix 1.

2.2 Partial discretization-expectation estimation and structural dimension estimation

2.2.1 A brief review on partial discretization-expectation estimation

As discussed above, estimating B is important for constructing an adaptive test. To this end, sufficient dimension reduction techniques can be applied. However, B is not identifiable in effect as we mentioned in Section 2.1. From the sufficient dimension reduction theories, what we can identify is the space spanned by B, or equivalently to say, q basis vectors of the space spanned by B (see, Cook 1998; Chiaromonte et al. 2002). Write $\tilde{B}$ as the $p \times q$ matrix consisting of these q basis vectors. We call $\tilde{B}$ the basis matrix. Note that B is also a basis matrix of the space. Thus it is easy to see that for a $q \times q$ non-singular matrix C, $B = \tilde{B} \times C$. When $q = 1$, C is a constant and thus $\tilde{B}$ is a vector proportional to the vector $B = \beta$. In Section 2.3, when we construct the test statistic, we can see that identifying $\tilde{B}$ is enough. In the following, we simply write $\tilde{B}$ as B without confusion.

In this section, we then focus on identifying a basis matrix B or equivalently to say, identifying the space spanned by B. This space is called the partial central subspace firstly introduced by Chiaromonte et al. (2002). Write it as $S^{(W)}_{Y|X}$. From their definition, it is the intersection of all subspaces S such that

$$Y \perp\!\!\!\perp X | (P_S X, W),$$

where $\perp\!\!\!\perp$ stands for the statistics independence and $P_{(\cdot)}$ indicates a projection operator with respect to the standard inner product. $\dim(S_{Y|X}^{(W)})$ is called the structural dimension of $S_{Y|X}^{(W)}$. In our setup, the structural dimension is q. Chiaromonte et al. (2002) and Wen and Cook (2007) developed estimation methods for $S_{Y|X}^{(W)}$ when W is discrete. Li et al. (2010) proposed groupwise dimension reduction (GDR) that can also deal with this case. Feng et al. (2013) proposed partial discretization-expectation estimation (PDEE), via developing discretization-expectation estimation (DEE) in Zhu et al. (2010). All of those estimations enjoy the root-n consistency to the partial central subspace. In this chapter, we adopt PDEE because PDEE is computational inexpensive and can be easily used to determine the structural dimension q. Also, when W is absent, PDEE can naturally reduce to DEE without any changes in the algorithm.

From Feng et al. (2013), the following are the basic estimation steps.

1. Discretize the covariate $W = (W_1, \cdots, W_{p_2})$ into a set of binary variables by defining $W(\mathbf{t}) = (I\{W_1 \le t_1\}, \cdots, I\{W_{q_2} \le t_{p_2}\})$ where the indicator functions $I\{W_i \le t_i\}$ take value 1 if $W_i \le t_i$ and 0 otherwise, for $i = 1, \cdots, p_2$.

2. Let $S_{Y|X}^{(W(t))}$ denote the partial central subspace of $Y|(X, W(\mathbf{t}))$, and $M(\mathbf{t})$ be a $p_1 \times p_1$ positive semi-definite matrix satisfying that $\mathrm{Span}\{\mathrm{M}(\mathbf{t})\} = S_{\mathrm{Y|X}}^{(\mathrm{W}(\mathbf{t}))}$.

3. Let $T = \tilde{W}$ where $\tilde{W}$ is an independent copy of W. The target matrix is $M = E\{M(\tilde{W})\}$. B consists of the eigenvectors that are associated with the nonzero eigenvalues of $M = E\{M(\tilde{W})\}$.

4. Let $W_1, \cdots, W_n$ be the n observations of W. Define an estimate of M as

$$M_n = \frac{1}{n} \sum_{i=1}^{n} M_n(W_i),$$

where $M_n(W_i)$ is the partial SIR estimate defined in Chiaromonte et al. (2002). Then an estimate $B_n(q)$ of B consists of the eigenvectors that are associated with the q largest eigenvalues of M_n. $B_n(q)$ can be root-n consistent to B. For more details, the reader can refer to Feng et al. (2013).

2.2.2 Structural dimension estimation

To estimate the structural dimension q, Feng et al. (2013) advised the BIC-type criterion. However, penalty selection is an issue. In this chapter, we suggest a ridge-type eigenvalue ratio estimate (RERE) to determine q as:

$$\hat{q} = \arg\min_{1\le j\le p} \left\{ \frac{\hat{\lambda}_{j+1}^2 + c_n}{\hat{\lambda}_j^2 + c_n} \right\}, \tag{2.4}$$

where $\hat{\lambda}_p \le \cdots \le \hat{\lambda}_1$ are the eigenvalues of the matrix M_n. This method is motivated by Xia et al. (2014). The algorithm is very easy to implement.

Theorem 2.1. *Under conditions A1 and A2 in Appendix 1, the estimate $\hat{q}$ of (2.4) with $c_n = \log n/n$ has the following consistency:*

(i) under H_0, $P(\hat{q} = 1) \to 1$;

(ii) under H_1, $P(\hat{q} = q) \to 1$.

From our justification presented in Appendix 1, the choice of c_n can be in a relatively wide range to ensure the consistency under the null and alternative hypotheses. This plays a very important role for the test statistic to be adaptive to the underlying models. Further, we will show in Section 2.3 that under the local alternatives converging to the null hypothesis, the ridge-type eigenvalue ratio estimate $\hat{q}$ can still converge to 1 such that the test can easily detect those local alternatives.

Finally, an estimate of B is $B_n = B_n(\hat{q})$. In the following test statistic construction, this estimate is used.

2.3 A model-adaptive test and its properties

2.3.1 Test statistic construction

The hypotheses of interest can now be restated as: the null hypothesis is

$$H_0: \ E(Y|X,W) = G(\beta^\top X, W, \theta) \quad \text{for some} \quad \beta \in \mathbb{R}^{\mathrm{p_1}}, \ \theta \in \Theta \in \mathbb{R}^{\mathrm{d}},$$

against the alternative hypothesis: for any β and θ

$$H_1: \ E(Y|X,W) = g(B^\top X, W) \neq G(\beta^\top X, W, \theta).$$

Under H_0, $q = 1$, and $B = \kappa\beta$ for some constant κ, then we have:

$$\begin{aligned} E(\epsilon|X,W) = 0 \ &\Leftrightarrow \ E(\epsilon I\{(\beta^\top X, W) \le (u,\omega)\}|\beta^\top X, W) = 0 \\ &\Leftrightarrow \ E(\epsilon I\{(\beta^\top X, W) \le (u,\omega)\}|B^\top X, W) = 0 \end{aligned}$$

for all (u,ω). However, under H_1, since $q \geq 1$ and $E(Y - G(\beta^\top X, W, \theta)|X,W) = g(B^\top X, W) - G(\beta^\top X, W, \theta) \neq 0$, we have:

$$E(Y - G(\beta^\top X, W, \theta)|X,W) \neq 0 \Leftrightarrow E(Y - G(\beta^\top X, W, \theta)|B^\top X, W) \neq 0.$$

Now we identify and estimate the structural dimension q for model adaptiveness to both the null and alternative models. In other words, we define an estimate $\hat{q}$ of q that goes to $q = 1$ under the null hypothesis and automatically to $q \geq 1$ under the alternative hypothesis. The test constructed later can then detect the general alternative model (2.3).

Before proceeding to the test statistic construction, we recall that B is in effect not identifiable: what we can identify is $\tilde{B} = B \times C$ for a $q \times q$ orthogonal matrix C. Thus, we need to make sure this non-identifiability does not affect the equivalence between $E(Y - G(\beta^\top X, W, \theta)|\tilde{B}^\top X, W) \neq 0$ and $E(Y - G(\beta^\top X, W, \theta)|X,W) \neq 0$. But this is easy to check. Note that $\tilde{B} = B \times C$ with C being a non-singular matrix and thus B and $\tilde{B}$ are one-to-one mapping. Then

$$\begin{aligned} E(Y - G(\beta^\top X, W, \theta)|X,W) &= E(g(B^\top X, W) - G(\beta^\top X, W, \theta)|X,W) \\ &= E(\tilde{g}(\tilde{B}^\top X, W) - G(\beta^\top X, W, \theta)|X,W), \end{aligned}$$

where $\tilde{g}(\cdot,\cdot) = g((C^\top)^{-1}\cdot,\cdot)$. It is equivalent between $E(Y - G(\beta^\top X, W, \theta)|B^\top X, W) \neq 0$ and $E(Y - G(\beta^\top X, W, \theta)|\tilde{B}^\top X, W) \neq 0$. Therefore, it is not a must to identify B. As we mentioned before, we simply write $\tilde{B}$ as B.

Now we are in the position to define a residual-marked empirical process. Let

$$V_n(u, \omega) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (y_i - G(\beta_n^\top x_i, w_i, \theta_n)) I\{(B_n^\top x_i, w_i) \leq (u, \omega)\}, \tag{2.5}$$

where β_n and θ_n are the estimates of β and θ, respectively, and B_n is a partial sufficient dimension reduction estimate of B with the structural dimension estimate $\hat{q}$ of q. β_n and θ_n are the nonlinear least squares estimates.

Therefore, we can use V_n as the base to construct a test statistic:

$$T_n = \int V_n^2(B_n^\top x, \omega) dF_n(x, \omega), \tag{2.6}$$

where F_n denotes the empirical distribution based on the samples $\{x_i, w_i\}_{i=1}^n$. Then the null hypothesis can be rejected for large values of T_n.

It is clear that this test statistic is not scale-invariant and thus usually a normalizing constant is required. Such a constant needs to be estimated which involves many unknowns. However, in this chapter, it is not necessary as we in the next section suggest a Monte Carlo test procedure that can make the test procedure scale-invariant automatically. The details are referred to Section 2.4.

2.3.2 Limiting null distribution

To study the properties of the process $V_n(\cdot,\cdot)$ and the test statistic T_n, we define a process here for the purpose of theoretical investigation: for u and ω,

$$V_n^0(u, \omega) = n^{-1/2} \sum_{i=1}^{n} (y_i - G(\beta^\top x_i, w_i, \theta)) I\{(B^\top x_i, w_i) \leq (u, \omega)\}. \tag{2.7}$$

When $E(Y^2) < \infty$, take the conditional variance of Y given $B^\top X = u$ and $W = \omega$,

$$\sigma^2(u, \omega) = Var(Y|B^\top X = u, W = \omega),$$

and put

$$\psi(u,\omega) = \int_{-\infty}^{\omega}\int_{-\infty}^{u} \sigma^2(v_1, v_2)dv_1 dv_2.$$

It is easy to see that under H_0,

$$\mathrm{Cov}\{\mathrm{V_n^0(u_1,\omega_1), V_n^0(u_2,\omega_2)}\} = \psi(\mathrm{u_1 \wedge u_2}, \omega_1 \wedge \omega_2).$$

By Theorem 1.1 in Stute (1997), we can assert that under H_0:

$$V_n^0 \longrightarrow V_\infty \quad \text{in distribution,} \tag{2.8}$$

in the Skorohod space $D[-\infty,\infty]$, where V_∞ is a continuous Gaussian process with mean zero and covariance kernel as follows:

$$K((u_1,\omega_1),(u_2,\omega_2)) = \psi(u_1 \wedge u_2, \omega_1 \wedge \omega_2).$$

Assume the regularity Conditions A3-A4 in Appendix 1 for the limiting null distribution.

Theorem 2.2. *Under H_0 and the regularity conditions A1-A4 in Appendix 1, we have in distribution*

$$V_n \longrightarrow V_\infty - G^\top V \equiv V_\infty^1,$$

where V_∞ is the Gaussian process defined in (2.8) and the vector-valued function $G^\top = (G_1, G_2, \cdots, G_{p+d})$ is defined as

$$G_i(u,\omega) = E\left[m_i(X,W,\beta,\theta)I\{(B^\top X, W) \le (u,\omega)\}\right],$$

where $B = \kappa\beta$, m_i is defined in Appendix 1 and V is a $(p_1+q)-$dimensional normal vector with mean zero and covariance matrix $L(\beta,\theta)$.

Remark 2.1. *From this theorem, we can see that the model-adaptive test statistic has the same convergence rate $1/\sqrt{n}$ to its limit as those of existing global smoothing tests. Thus, this rate can be faster than $n^{-1/2}h^{1/4}$ that Guo et al. (2014)'s model-adaptive test can achieve where h is the bandwidth going to zero at a certain rate. In* *other words, in an asymptotic sense, there is no room for global smoothing tests to improve their performance. This is a very different feature of global smoothing tests from that of local smoothing tests. However, as Stute and Zhu (2002) did, the test can largely avoid dimensionality impact to make the test more powerful when p is large.*

2.3.3 Power Study

To study the power performance, consider the following sequence of alternatives

$$H_{1n}: \; Y = G(\beta^\top X, W, \theta) + C_n g(B^\top X, W) + \varepsilon. \tag{2.9}$$

When C_n is fixed, they are global alternative models; when C_n goes to zero, they are local alternatives.

A very important issue is about estimating q under these local alternatives. Recall that under the global alternative hypothesis in Section 2.2, the estimate $\hat{q}$ converges to q which can be larger than 1 when B contains more than one basis vector. However, under the above local alternatives, when C_n goes to zero, the models converge to the hypothetical model that has one vector β. Thus, it is expected that $\hat{q}$ also converges to 1 even under the alternative models. The following lemma confirms this expectation.

Lemma 2.1. *Under H_{1n} in* (2.9)*, $C_n = 1/\sqrt{n}$ and the regularity conditions in Theorem 2.2, and the estimate $\hat{q}$ in (2.4) satisfies that as $n \to \infty$, $P(\hat{q} = 1) \to 1$.*

To further study the power performance of the test, assume an additional regularity Condition A5 in Appendix 1.

Theorem 2.3. *Under H_{1n} and Conditions A1, A2, A4 and A5,*
(i) when $C_n = 1/\sqrt{n}$, we have in distribution

$$V_n(u,\omega) \longrightarrow V_\infty(u,\omega) + E(g(B^\top X, W) I\{(\kappa\beta_0^\top X, W) \le (u,\omega)\}) + G^\top(\eta - V)(u,\omega),$$

where V_∞, G and V are defined as those in Theorem 2.2. Then T_n has a finite limit.
(ii) when $\sqrt{n}C_n \to \infty$,

$$V_n(u,\omega) \longrightarrow \infty.$$

This implies that $T_n \to \infty$ in distribution.

Remark 2.2. *This theorem clearly states that unlike Stute and Zhu (2002), our test can be omnibus, rather than directional.*

2.4 Monte Carlo Test Process

Since the limiting null distribution of the test statistic T_n is not tractable, the non-parametric Monte Carlo test procedure is suggested to approximate the sampling null distribution, which is in spirit similar to an application of the wild bootstrap, see Stute et al. (1998), Neuhaus and Zhu (1998) and Zhu and Neuhaus (2000). More details can be found in Zhu (2005).

The procedure is as follows:

Step 1 Generate a sequence of iid variables $\mathbf{U} = \{U_i\}_{i=1}^n$ from the standard normal distribution $N(0,1)$. Then construct the following process:

$$\Delta_n(u,\omega,\mathbf{U}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \hat{\varrho}(x_i, w_i, y_i, \beta, \theta) U_i,$$

where $\hat{\varrho}(x_i, w_i, y_i, \beta, \theta)$ is the estimate of $\varrho(x_i, w_i, y_i, \beta, \theta)$ and $\hat{\varrho}$ and ϱ are defined as:

$$\varrho(x_i, w_i, y_i, \beta, \theta) = \epsilon_i I\{(B^\top x_i, w_i) \le (u,\omega)\} - G^\top v_i,$$

$$\hat{\varrho}(x_i, w_i, y_i, \beta, \theta) = \hat{\epsilon}_i I\{(B_n^\top x_i, w_i) \le (u,\omega)\} - \hat{G}^\top \hat{v}_i,$$

with

$$G(u,\omega) = E\left[m(X, W, \beta_0, \theta_0) I\{(B^\top X, W) \le (u,\omega)\}\right],$$

$$v_i = l(x_i, w_i, y_i, \beta_0, \theta_0),$$

and

$$\hat{G} = \frac{1}{n} \sum_{i=1}^n m(x_i, w_i, \beta_n, \theta_n) I\{(B_n^\top X, W) \le (u,\omega)\},$$

$$\hat{\epsilon}_i = y_i - G(\beta_n^\top x_i, w_i, \theta_n),$$

$$\hat{v}_i = l(x_i, w_i, y_i, \beta_n, \theta_n).$$

The resulting Monte Carlo test statistic is

$$\tilde{T}_n(\mathbf{U}) = \int \Delta_n^2(B_n^\top x, w, \mathbf{U}) dF_n(x, w).$$

Step 2 Generate m sets of $\mathbf{U}$, $\mathbf{U}_j$, $j = 1, \cdots, m$, and get m values of $\tilde{T}_n(\mathbf{U})$, say $\tilde{T}_n(\mathbf{U}_j)$, $j = 1, \cdots, m$.

Step 3 The p-value is estimated by

$$\hat{p} = \frac{1}{m} \sum_{j=1}^{m} I(\tilde{T}_n(\mathbf{U}_j) \geq T_n).$$

Whenever $\hat{p} \leq \alpha$, reject H_0, for a given significance level α.

As we mentioned before, this test procedure is scale-invariant although T_n is not scale-invariant. This is because the resampling procedure does not need to involve anything about test statistic standardization and $\hat{p} = \frac{1}{m}\sum_{j=1}^{m} I(\tilde{T}_n(\mathbf{U}_j) \geq T_n) = \frac{1}{m}\sum_{j=1}^{m} I(\tilde{T}_n(\mathbf{U}_j)/c \geq T_n/c)$ for any $c > 0$.

The following theorem states the consistency of the conditional distribution approximation.

Theorem 2.4. *Under the null hypothesis and the conditions in Theorem 2.2, we have that for almost all sequences* $\{(y_1, x_1, w_1), \cdots, (y_n, x_n, w_n), \cdots\}$, *the conditional distribution of* $\tilde{T}_n(\boldsymbol{U})$ *converges to the limiting null distribution of* T_n.

2.5 Numerical Studies

2.5.1 Simulations

In this subsection, the simulations are conducted to examine the finite-sample performance of the proposed test. The simulations are based on a total of 2000 Monte Carlo test replications to compute the critical values or $p-$values. Each experiment is then repeated 1000 times to compute the empirical sizes and powers at the significance level $\alpha = 0.05$. To estimate the central subspace spanned by B, we use the SIR-based PDEE/DEE procedure according to the cases with and without the presence of the variate W in the model. In these two cases, we write the test as T_n^{PDEE} without confusion.

We choose Zheng (1996)'s test and Stute and Zhu (2002)'s test as the respective representatives of local smoothing tests and global smoothing tests to compare with our test. Further, we also compare with Guo et al. (2014)'s test that is also model-adaptive. This is because 1). Zheng (1996)'s test has the explicitly and tractable limiting null distribution that can be used to determine the critical values; 2). Like other local smoothing tests, the re-sampling version helps improve their performance. We then also include the re-sampling version of Zheng (1996)'s test; 3) Stute and Zhu (2002)'s test is asymptotically distribution-free and powerful in many situations, but not an omnibus test. We include Guo et al. (2014)'s test for comparison because it is based on Zheng (1996)'s test and also has model-adaptiveness such that the test can be much more powerful than Zheng (1996)'s test. Write respectively T_n^{PDEE}, T_n^{ZH}, T_n^{SZ} and T_n^{GWZ} as our test, Zheng (1996)'s test, Stute and Zhu (2002)'s test and Guo et al. (2014)'s test.

In this section, we first design four examples to examine the performance in four scenarios without the random variable W. The first has the same projection direction in both the hypothetical and alternative models, the second is to check the adaptiveness of our test to indeed make an omnibus test although dimension reduction structure under the null hypothesis is fully adopted. This example shows that Stute and Zhu (2002)'s test is a directional test and thus with much less power. The third is to check the dimensionality impact from X for local smoothing tests, and Zheng (1996)'s test and Guo et al. (2014)'s test are compared. The fourth is to assess the impact from correlation among the components of X. In the first three examples, the data $\{x_i\}_{i=1}^n$ are generated from the multivariate standard normal distribution

$N(0, I_p)$, independent of the standard normal errors ϵ.

Example 2.1. Consider the following regression models as:

- $Y = \beta_0^\top X + a \times cos(0.6\pi\beta_0^\top X) + 0.5 \times \epsilon$ and $\beta_0 = (0, 0, 1, 1)/\sqrt{2}$.

The value $a = 0$ corresponds the null hypothesis and $\theta \neq 0$ to the alternative hypothesis. The different values $a = (0, 0.2, 0.4, 0.6, 0.8, 1)$ are used. The results are shown in Figure (2.1).

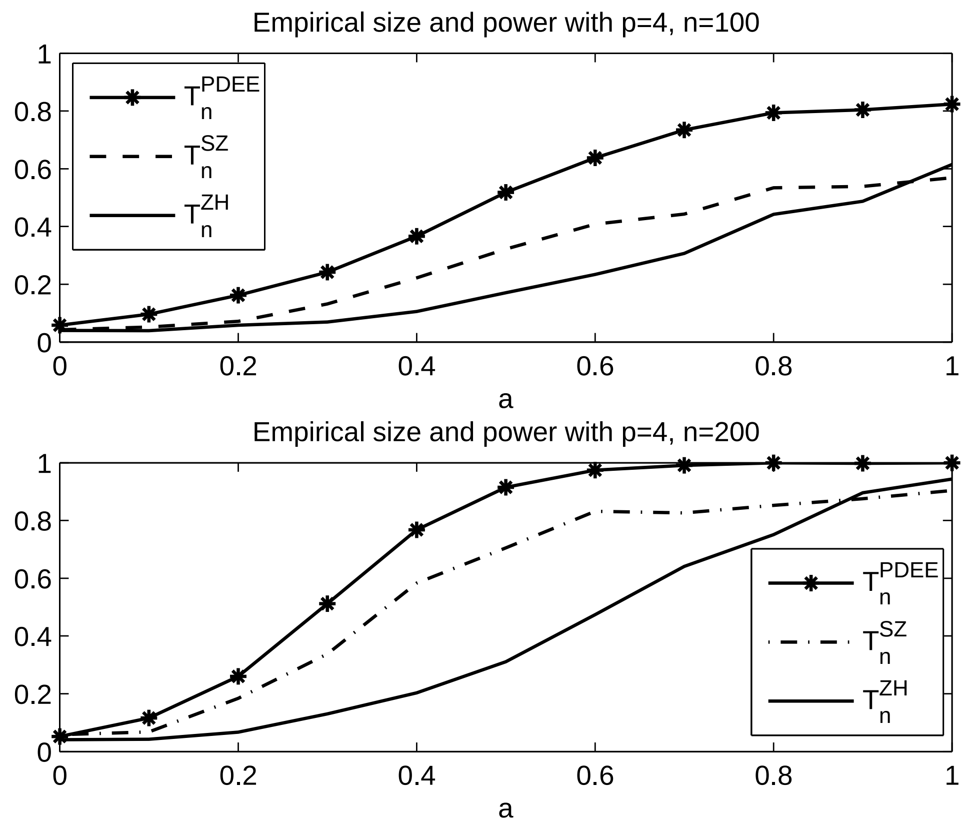

Figure 2.1: The empirical size and power curves of T_n^{PDEE}, T_n^{SZ} and T_n^{ZH} in Example 2.1.

From Figure (2.1), we observe that the power reasonably increases with larger a. Further, the performance of our test T_n^{PDEE} is significantly more powerful than that of T_n^{ZH} and T_n^{SZ} uniformly. When a is not large, T_n^{SZ} works better than T_n^{ZH}, and when a is large, T_n^{ZH} is slightly better than T_n^{SZ} in power.

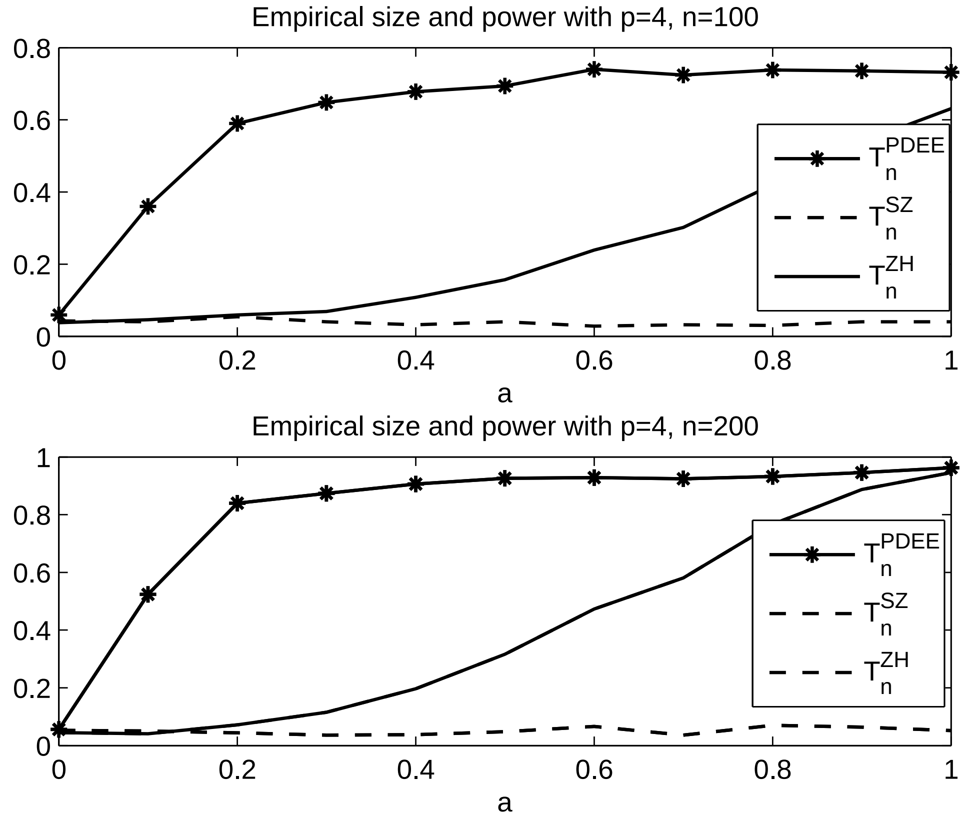

Figure 2.2: The empirical size and power curves of T_n^{PDEE}, T_n^{SZ} and T_n^{ZH} in Example 2.2.

Example 2.2. To further check whether our test is an omnibus test to detect general alternative models rather than a directional test, we again make a comparison with Stute and Zhu (2002)'s test and Zheng (1996)'s test. In this example, we generate the data from the following regression model:

- $Y = \beta_0^\top X + a \times \cos(0.6\pi\beta_1^\top X) + 0.5 \times \epsilon$;

where $\beta_0 = (1, 1, 0, 0)/\sqrt{2}$ and $\beta_1 = (0, 0, 1, 1)/\sqrt{2}$. The value $a = 0$ corresponds the null hypothesis and $\theta \neq 0$ to the alternative hypothesis. The different values $a = (0, 0.2, 0.4, 0.6, 0.8, 1)$ are used. In these models, $B = (\beta_0^\top, \beta_1^\top)^\top$ and $\beta_0^\top X$ is orthogonal to the functions under the alternatives. We can see that Stute and Zhu (2002)'s test cannot detect such alternatives. The results are reported in Figure (2.2).

Figure (2.2) shows clearly that Stute and Zhu (2002)'s test T_n^{SZ} cannot detect the alternative at all. Zheng (1996)'s test can do it, but is not very sensitive to the alternatives.

Example 2.3. To gain further insights into our test, we consider the impact from the dimensionality of X for the tests. When the dimension is large, Zheng (1996)'s test does not work well on significance level maintainance and power performance because of slow convergence. Thus, the wild bootstrap is applied to approximate the sampling null distribution. The re-sampling time is 2000 in this simulation study. The bootstrap version is written as T_n^{ZHB}. Guo et al. (2014)'s test is also compared.

Consider the models as:

- $Y = \beta_0^\top X + a \times \{0.3(\beta_1^\top X)^3 + 0.3(\beta_1^\top X)^2\} + 0.5 \times \epsilon$;

where $\beta_0 = (1, 1, 1, 1, 0, 0, 0, 0)/2$ and $\beta_1 = (0, 0, 0, 0, 1, 1, 1, 1)/2$. Then the dimension $p = 8$. The results are listed in Table (2.1).

Table 2.1: Empirical sizes and powers of T_n^{PDEE}, T_n^{ZHB}, T_n^{ZH} and T_n^{GWZ} for Example 2.3.

a	T_n^{PDEE}			T_n^{ZHB}			T_n^{ZH}			T_n^{GWZ}		
n	50	100	200	50	100	200	50	100	200	50	100	200
0	0.0560	0.0570	0.0460	0.0400	0.0450	0.0580	0.0240	0.0270	0.0430	0.0350	0.0610	0.0510
0.2	0.2210	0.2760	0.3780	0.0480	0.0630	0.0810	0.0260	0.0390	0.0600	0.0760	0.1100	0.1830
0.4	0.3660	0.5100	0.7820	0.0650	0.1140	0.1700	0.0460	0.0810	0.1430	0.1220	0.2750	0.4710
0.6	0.5800	0.7240	0.9380	0.0980	0.1810	0.3860	0.0690	0.1670	0.3590	0.2120	0.4460	0.7680
0.8	0.6120	0.8560	0.9840	0.1230	0.2670	0.5380	0.1070	0.2730	0.5780	0.2840	0.6070	0.9230
1.0	0.7180	0.9100	0.9970	0.1600	0.3510	0.7220	0.1330	0.3820	0.7330	0.3920	0.7440	0.9720

From Table (2.1), we can see that T_n^{ZH} cannot maintain the significance level well, but its bootstrap version T_n^{ZHB} and T_n^{GWZ} can do, and T_n^{PDEE} can work better uniformly. Second, compared with T_n^{ZH}, the model-adaptive dimension reduction test T_n^{GWZ} has a clear advantage over its counterpart T_n^{ZH} in maintaining the significance level and gaining power. However, even though, T_n^{PDEE} still works better uniformly. This seems to suggest that global smoothing test can perform better than local smoothing test. Also compared with those results in Figures (2.1) and (2.2) with $p = 4$, we can see that the dimension p has little impact for T_n^{PDEE}. Yet, it has very significant impact for T_n^{ZHB} and T_n^{ZH}.

Example 2.4. To further assess the performance of the test T_n^{PDEE}, we consider the impact from the correlated covariate X and the distribution of the error term ϵ. Consider the following model:

- $y = \beta_0^\top X + a \times \exp(-(\beta_0^\top X)^2/2)/2 + 0.5 \times \epsilon$;

where X follows normal distribution $N(0, \Sigma)$ with covariance matrix $\Sigma_{ij} = I(i = j) + \rho^{|i-j|} I(i \neq j)$ for $\rho = 0.5$, $i, j = 1, 2, \cdots, p$, $\beta_0 = (1, 1, -1, -1)/2$ and ϵ follows the student's t-distribution with degrees of freedom 4.

Table 2.2: Empirical sizes and powers of T_n^{PDEE}, T_n^{SZ}, T_n^{ZH} and T_n^{GWZ} for Example 2.4.

a	T_n^{PDEE}			T_n^{SZ}			T_n^{ZH}			T_n^{GWZ}		
n	50	100	200	50	100	200	50	100	200	50	100	200
0	0.0620	0.0570	0.0520	0.0310	0.0390	0.0500	0.0380	0.0390	0.0410	0.0510	0.0530	0.0450
0.2	0.1020	0.1670	0.2070	0.0520	0.0600	0.1390	0.0550	0.0690	0.0820	0.0810	0.1110	0.1670
0.4	0.2350	0.4160	0.5870	0.0920	0.1560	0.4060	0.0880	0.1490	0.2720	0.1530	0.2980	0.5220
0.6	0.4310	0.6600	0.8850	0.1560	0.3780	0.7260	0.2150	0.3740	0.5950	0.2980	0.6110	0.8520
0.8	0.5820	0.8540	0.9780	0.2680	0.5420	0.9180	0.3690	0.5860	0.8850	0.5410	0.8280	0.9690
1.0	0.6960	0.9510	0.9960	0.2920	0.7300	0.9800	0.5300	0.7830	0.9640	0.7040	0.9610	0.9990

The results are presented in Table (2.2). Comparing the results in this table with those in Figure (2.1) and Figure (2.2), respectively, we can clearly observe that with the correlated covariate X, the similar conclusions as those in Examples 2.1 and 2.2 are arrived. T_n^{PDEE} can control the type I error very well. We also find that when the structural dimension $q = 1$ under the alternative hypothesis, the power performance of T_n^{GWZ} is very similar to that of T_n^{PDEE}. Comparing Table (2.1) about example 2.3 with Table (2.2) about example 2.4, we can see that the lower structural dimension helps on T_n^{GWZ} with higher empirical power. It seems that the structural dimension q still has negative impact for T_n^{GWZ} although theoretically, T_n^{GWZ} can detect alternatives distinct from the null at a rate as if the dimension of X were one. But, the power of T_n^{PDEE} is not deteriorated by the structural dimension. Further, T_n^{PDEE} can control type I error very well and is significantly more powerful than Zheng (1996)'s test and Stute and Zhu (2002)'s test. It seems evident that T_n^{PDEE} is robust against to the error term.

In summary, the global smoothing-based model-adaptive dimension reduction test inherits the advantages of global smoothing tests and enjoys the model-adaptiveness when dimension reduction structure is adopted. In the following, we consider the parallel models in Examples 2.1-2.4 when the covariate W is included. However, we only present the results about T_n^{PDEE} because based on the results in the above examples and comparisons, the competitors even have worse performance when there are q_1 more dimensions in the model (meaning that q_1 more dimensions are added when W is q_1-dimensional).

Example 2.5. The four models are as:

Case 1) $Y = \beta_0^\top X + W + a \times cos(0.6\pi\beta_0^\top X) + 0.5 \times \epsilon;$

Case 2) $Y = \beta_0^\top X + \sin(W) + a \times (0.25(\beta_1^\top X)^2 + \sin(W)) + 0.5 \times \epsilon;$

Case 3) $Y = \beta_0^\top X + \cos(W) + a \times \{0.3(\beta_1^\top X)^3 + 0.3(\beta_1^\top X)^2\} + 0.5 \times \epsilon;$

Case 4) $y = \beta_0^\top X + \sin(W) + a \times \exp(-(\beta_0^\top X)^2/2) \times W + 0.5\epsilon.$

All the settings are the same as the respective settings in Examples 2.1-2.4 except the additional W following the normal distribution $N(0, 1)$. The results are reported in Tables (2.3). The reported results clearly indicate that when W is presented, T_n^{PDEE} still works well in maintaining the significance level and detecting general alternatives as an omnibus test does.

Table 2.3: Sizes and powers of T_n^{PDEE} for Example 2.5.

	a	n=50	n=100	n=200	n=400
Case 1	0	0.0680	0.0590	0.0560	0.0500
	0.2	0.1720	0.1720	0.2940	0.6420
	0.4	0.2620	0.5120	0.8740	1.0000
	0.6	0.5080	0.8950	1.0000	1.0000
	0.8	0.6520	0.9640	1.0000	1.0000
	1	0.7020	0.9840	1.0000	1.0000
Case 2	0	0.0580	0.0550	0.0530	0.0480
	0.2	0.2710	0.2860	0.4130	0.6300
	0.4	0.4340	0.5650	0.7580	0.9540
	0.6	0.5800	0.7080	0.9160	0.9980
	0.8	0.5870	0.8130	0.9770	1.0000
	1	0.6660	0.8680	0.9950	1.0000
Case 3	0	0.0630	0.0560	0.0550	0.0500
	0.2	0.2810	0.3550	0.4390	0.7400
	0.4	0.5720	0.7200	0.9080	0.9620
	0.6	0.6940	0.8180	0.9840	0.9950
	0.8	0.7700	0.8860	0.9900	1.0000
	1	0.8210	0.9230	1.0000	1.0000
Case 4	0	0.0600	0.0450	0.0470	0.0510
	0.2	0.1020	0.1540	0.2700	0.4960
	0.4	0.2170	0.4270	0.7440	0.9820
	0.6	0.4050	0.7240	0.9730	1.0000
	0.8	0.5410	0.8970	0.9990	1.0000
	1.0	0.6460	0.9560	0.9990	1.0000

2.5.2 Real Data Analysis

In this section, we check the regression modelling of the well-known Boston Housing Data for illustration which was initially studied by Harrison and Rubinfeld (1978). The data set contains 506 observations and 14 variables which are median value of owner-occupied homes in \$1000's (*MEDV*), per capita crime rate by town (*CRIM*), proportion of residential land zoned for lots over 25,000 sq.ft. (*ZN*), proportion of non-retail business acres per town (*INDUS*), Charles River dummy variable (1 if tract bounds river; 0 otherwise) (*CHAS*), nitric oxides concentration (parts per 10 million) (NOX), average number of rooms per dwelling (*RM*), proportion of owner-occupied units built prior to 1940 (*AGE*), weighted distances to five Boston employment centres (*DIS*), index of accessibility to radial highways (*RAD*), full-value property-tax rate per $10,000$ (*TAX*), pupil-teacher ratio by town (*PTRATIO*), the proportion of blacks by town (B) and lower status of the population (*LSTAT*).

As suggested by Feng et al. (2013), we take the logarithm of (*MEDV*) as the predictor, the predictor *CRIM* as W and the other eleven predictors as X except the predictor *CHAS*. This is because *CHAS* has little influence for the housing price as advised by Wang et al. (2010). It is excluded from this data analysis. In this data analysis, we standardize the predictors for ease of explanation. From the plot showed in Feng et al. (2013), a simple linear model is considered to be the hypothetical model. The SIR-based PDEE procedure is applied to determine the partial central subspace $S_{Y|X}^{(W)}$. The structural dimension $\hat{q} = 2$ of the partial central subspace is determined by the method RERE in Section 2.2. A total of 2000 Monte Carlo test replications are implemented to compute the p value that is about zero. Hence, it is reasonable to reject the null hypothesis. Moreover, $\hat{q}$ is estimated to be 2. Thus a partial multi-index modeling is required although the plot seems to suggest a linear model.

2.6 Discussions

In this chapter, we have proposed a model-adaptive dimension reduction test based on residual marked empirical process for partially parametric single-index models. The test can fully utilize the dimension reduction structure to lead the test to suffer less from the curse of dimensionality, and to be still an omnibus test. The comparisons with existing local smoothing tests and global smoothing test suggest 1). Model-adaption enhances the power performance, and at the same time, the capacity of controlling type I error; 2). Global smoothing-based model-adaptive test outperforms local smoothing-based model-adaptive test. Thus, global smoothing test is worthy of recommendation. This method is readily applied to other models and problems when dimension reduction structure is presented. The research is on-going.

2.7 Appendix 1

2.7.1 Regularity Conditions

A1 $M_n(t)$ has the following expansion:

$$M_n(t) = M(t) + E_n\{\psi(X, W, Y, t)\} + R_n(t),$$

where $E_n(\cdot)$ denotes sample averages, $E(\psi(X, W, Y, t)) = 0$ and $\psi(X, W, Y, t)$ has a finite second-order moment.

A2 $\sup_{t \in R^{p_2}} ||R_n(t)||_F = o_p(n^{-1/2})$, where $|| \cdot ||_F$ denotes the Frobenius norm of a matrix.

A3 Under model (2.9), (β_n, θ_n) has an asymptotically linear representation

$$\sqrt{n} \begin{pmatrix} \beta_n - \beta \\ \theta_n - \theta \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} l(X, W, Y, \beta, \theta) + o_p(1),$$

where the vector-function l satisfies:

1. $E(l(X, W, Y, \beta, \theta)) = 0$;
2. the matrix $L(\beta, \theta) = E(l(X, W, Y, \beta, \theta) l^\top (X, W, Y, \beta, \theta))$ is positive definite.

A4 The function $G(\beta^\top x, w, \theta)$ is continuously differentiable at each $(\tilde{\beta}, \tilde{\theta})$ in a neighbourhood of the true parameters (β, θ). The first-order partial derivative

$$m(x, w, \beta, \theta) = \frac{\partial G(\beta^\top x, w, \theta)}{\partial(\beta, \theta)} = (m_1(x, w, \beta, \theta), \cdots, m_{p1+q}(x, w, \beta, \theta))$$

satisfies that there exists an integrable function $M(x, w)$ such that:

$$|m_j(x, w, \tilde{\beta}, \tilde{\theta})| \leq M(x, w) \quad \text{for all} \quad (\tilde{\beta}, \tilde{\theta}) \in R^p \quad \text{and} \quad 1 \leq j \leq p + d.$$

A5 Under (2.9) and $C_n = 1/\sqrt{n}$,

$$\sqrt{n}\begin{pmatrix} \beta_n - \beta \\ \theta_n - \theta \end{pmatrix} = \eta + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} l(x_i, w_i, y_i, \beta, \theta) + o_p(1),$$

with some constant vector η.

Remark 2.3. *Conditions A1 and A2 are necessary for PDEE. Under the linearity condition and constant conditional variance conditions, $PDEE_{SIR}$ satisfies the Conditions A1 and A2. Conditions A3 and A4 are essential for the asymptotic distributions of the nonlinear least squares estimates β_n and θ_n. It is worthwhile to note that Condition A5 is reasonable under the local alternative hypotheses. We can use the nonlinear least squares estimate as an example to explain.*

When the parameter θ is absent, it is easy to derive that

$$\eta = E\{g(B^\top X, W)G'(\beta^\top X, W)\}[E\{X^\top X(G'' + G'G')(\beta^\top X, W)\}]^{-1},$$

and

$$l(x_i, w_i, y_i, \beta) = \{y_i - G(\beta^\top x_i, w_i)\}G'(\beta^\top x_i, w_i)x_i[E\{X^\top X(G'' + G'G')(\beta^\top X, W)\}]^{-1}.$$

When θ exists, we have

$$\eta = \left(E\begin{bmatrix} X^\top X(G_{11} + G_1G_1) & X^\top(G_{12} + G_1G_2) \\ (G_{12} + G_1G_2)^\top X & (G_{22} + G_2G_2) \end{bmatrix}\right)^{-1} E\begin{bmatrix} g(B^\top X, W)XG_1(\beta^\top X, W, \theta) \\ g(B^\top X, W)G_2(\beta^\top X, W, \theta) \end{bmatrix}$$

and

$$l(x_i, w_i, y_i, \beta, \theta) = \left(E\begin{bmatrix} X^\top X(G_{11} + G_1G_1) & X^\top(G_{12} + G_1G_2) \\ (G_{12} + G_1G_2)^\top X & (G_{22} + G_2G_2) \end{bmatrix}\right)^{-1} \begin{pmatrix} (y_i - G(\beta^\top x_i, w_i, \theta)x_iG_1(\beta^\top x_i, w_i, \theta) \\ (y_i - G(\beta^\top x_i, w_i, \theta)G_2(\beta^\top x_i, w_i, \theta) \end{pmatrix},$$

where $G_1 = \frac{\partial G(\beta^\top X, W, \theta)}{\partial(\beta^\top X)}$, $G_2 = \frac{\partial G(\beta^\top X, W, \theta)}{\partial\theta}$, $G_{11} = \frac{\partial^2 G(\beta^\top X, W, \theta)}{\partial(\beta^\top X)^2}$, $G_{12} = \frac{\partial^2 G(\beta^\top X, W, \theta)}{\partial(\beta^\top X)\partial\theta}$, and $G_{22} = \frac{\partial^2 G(\beta^\top X, W, \theta)}{\partial\theta^2}$.

2.7.2 Proof of the theorems

Proof of Theorem 2.1 . Following the arguments for proving Proposition 2 in Feng et al. (2013), we obtain $M_n - M = O_p(n^{-1/2})$. As Zhu and Ng (1995) or Zhu and Fang (1996) showed, the root-n consistency of the eigenvalues of M_n is retained, namely, $\hat{\lambda}_i - \lambda_i = O_p(n^{-1/2})$.

(I) Under H_0, since $\lambda_1 > 0$ and for any $l > 1$, $\lambda_l = 0$, we have that $\hat{\lambda}_1^2 = \lambda_1^2 + O_p(1/\sqrt{n})$ and $\hat{\lambda}_l^2 = \lambda_l^2 + O_p(1/n) = O_p(1/n)$. Taking $c_n = \log n/n$, then for any $l > 1$, we have

$$\begin{aligned} \frac{\hat{\lambda}_2^2 + c_n}{\hat{\lambda}_1^2 + c_n} - \frac{\hat{\lambda}_{(l+1)}^2 + c_n}{\hat{\lambda}_l^2 + c_n} &= \frac{\lambda_2^2 + c_n + O_p(1/n)}{\lambda_1^2 + c_n + O_p(1/\sqrt{n})} - \frac{\lambda_{l+1}^2 + c_n + O_p(1/n)}{\lambda_l^2 + c_n + O_p(1/n)} \\ &\to -1. \end{aligned}$$

Therefor, $P(\hat{q} = 1) \longrightarrow 1$.

(II) Under H_1, since for any $1 < l \leq q$, $\lambda_l > 0$, we have $\hat{\lambda}_l^2 = \lambda_l^2 + O_p(1/\sqrt{n})$. On the other hand, for any $q < l \leq p$, $\lambda_l = 0$, then we have $\hat{\lambda}_l^2 = \lambda_l^2 + O_p(1/n) = O_p(1/n)$. Then for any $1 \leq l_1, l_2 < q$, $\lambda_{l_1} > 0$, $\lambda_{l_2} > 0$, we have

$$\begin{aligned} \frac{\hat{\lambda}_{(l_1+1)}^2 + c_n}{\hat{\lambda}_{l_1}^2 + c_n} - \frac{\hat{\lambda}_{(l_2+1)}^2 + c_n}{\hat{\lambda}_{l_2}^2 + c_n} &= \frac{\lambda_{l_1+1}^2 + c_n + O_p(1/\sqrt{n})}{\lambda_{l_1}^2 + c_n + O_p(1/\sqrt{n})} - \frac{\lambda_{l_2+1}^2 + c_n + O_p(1/\sqrt{n})}{\lambda_{l_2}^2 + c_n + O_p(1/\sqrt{n})} \\ &\to \frac{\lambda_{l_1+1}^2}{\lambda_{l_1}^2} - \frac{\lambda_{l_2+1}^2}{\lambda_{l_2}^2}. \end{aligned}$$

Thus, for any $l < q$, we have $\lambda_l > 0$, $\lambda_{l+1} > 0$ and

$$\begin{aligned} \frac{\hat{\lambda}_{(q+1)}^2 + c_n}{\hat{\lambda}_q^2 + c_n} - \frac{\hat{\lambda}_{(l+1)}^2 + c_n}{\hat{\lambda}_l^2 + c_n} &\to \frac{\lambda_{q+1}^2}{\lambda_q^2} - \frac{\lambda_{l+1}^2}{\lambda_l^2} \\ &= -\frac{\lambda_{l+1}^2}{\lambda_l^2} < 0 \end{aligned}$$

Further, due to the facts that $\lambda_q^2 > 0$ and for any $l > q$, we have $\lambda_l = 0$, we derive

$$\begin{aligned} \frac{\hat{\lambda}_{(q+1)}^2 + c_n}{\hat{\lambda}_q^2 + c_n} - \frac{\hat{\lambda}_{(l+1)}^2 + c_n}{\hat{\lambda}_l^2 + c_n} &= \frac{\lambda_{q+1}^2 + c_n + O_p(1/n)}{\lambda_q^2 + c_n + O_p(1/\sqrt{n})} - \frac{\lambda_{l+1}^2 + c_n + O_p(1/n)}{\lambda_l^2 + c_n + O_p(1/n)} \\ &= \frac{c_n + o_p(c_n)}{\lambda_q^2 + c_n + O_p(1/\sqrt{n})} - \frac{c_n + o_p(c_n)}{c_n + o_p(c_n)} \\ &\to -1 < 0. \end{aligned}$$

Therefor, altogether, we can conclude that $P(\hat{q} = q) \longrightarrow 1$. $\square$

Proof of Theorem 2.2 . Under the null hypothesis, $V_n(u)$ is decomposed to be:

$$\begin{aligned} V_n(u) &= n^{-1/2}\sum_{i=1}^{n}(y_i - G(\beta_n^\top x_i, w_i, \theta_n))I\{(B_n^\top x_i, w_i) \le (u,\omega)\} \\ &= n^{-1/2}\sum_{i=1}^{n}(y_i - G(\beta_n^\top x_i, w_i, \theta_n))I\{(\kappa\beta_n^\top x_i, w_i) \le (u,\omega)\} \\ &\quad + n^{-1/2}\sum_{i=1}^{n}(y_i - G(\beta_n^\top x_i, w_i, \theta_n))[I\{(B_n^\top x_i, w_i) \le (u,\omega)\} \\ &\quad - I\{(\kappa\beta_n^\top x_i, w_i) \le (u,\omega)\}] \\ &= n^{-1/2}\sum_{i=1}^{n}(y_i - G(\beta_n^\top x_i, w_i, \theta_n))I\{(\kappa\beta_n^\top x_i, w_i) \le (u,\omega)\} \\ &\quad + n^{-1/2}\sum_{i=1}^{n}\epsilon_i[I\{(B_n^\top x_i, w_i) \le (u,\omega)\} - I\{(\kappa\beta_n^\top x_i, w_i) \le (u,\omega)\}] \\ &\quad + n^{-1/2}\sum_{i=1}^{n}(G(\beta_0^\top x_i, w_i, \theta_0) - G(\beta_n^\top x_i, w_i, \theta_n)) \\ &\quad [I\{(B_n^\top x_i, w_i) \le (u,\omega)\} - I\{(\kappa\beta_n^\top x_i, w_i) \le (u,\omega)\}] \\ &\equiv: V_{1n} + V_{2n} + V_{3n}, \end{aligned}$$

where

$$\begin{aligned} V_{1n}(u,\omega) &= n^{-1/2}\sum_{i=1}^{n}(y_i - G(\beta_n^\top x_i, w_i, \theta))I\{(\kappa\beta_n^\top x_i, w_i) \le (u,\omega)\}, \\ V_{2n}(u,\omega) &= n^{-1/2}\sum_{i=1}^{n}\epsilon_i[I\{(B_n^\top x_i, w_i) \le (u,\omega)\} - I\{(\kappa\beta_n^\top x_i, w_i) \le (u,\omega)\}], \\ V_{3n}(u,\omega) &= n^{-1/2}\sum_{i=1}^{n}(G(\beta_0^\top x_i, w_i, \theta_0) - G(\beta_n^\top x_i, w_i, \theta_n))[I\{(B_n^\top x_i, w_i) \le (u,\omega)\} \\ &\quad - I\{(\kappa\beta_n^\top x_i, w_i) \le (u,\omega)\}]. \end{aligned}$$

Following the analogous arguments for proving Theorem 1 of Stute and Zhu (2002), we obtain $V_{1n} \longrightarrow V_\infty - G^\top V \equiv: V_\infty^1$.

To study V_{2n}, we first note that for B and β, we get

$$E(\epsilon_i[I\{(B^\top x_i, w_i) \le (u,\omega)\} - I\{(\kappa\beta^\top x_i, w_i) \le (u,\omega)\}]) = 0,$$

and

$$
\begin{aligned}
& Var\left(n^{-1/2}\sum_{i=1}^{n}\epsilon_i[I\{(B^\top x_i, w_i)\le (u,\omega)\} - I\{(\kappa\beta^\top x_i, w_i)\le(u,\omega)\}]\right)\\
= \;& \frac{1}{n}\sum_{i=1}^{n} Var\left(\epsilon_i[I\{(B^\top x_i, w_i)\le (u,\omega)\} - I\{(\kappa\beta^\top x_i, w_i)\le(u,\omega)\}]\right)\\
= \;& Var\left(\epsilon_i[I\{(B^\top x_i, w_i)\le (u,\omega)\} - I\{(\kappa\beta^\top x_i, w_i)\le(u,\omega)\}]\right)\\
\le \;& E(\epsilon_i^2)E[I\{(B^\top x_i, w_i)\le (u,\omega)\} - I\{(\kappa\beta^\top x_i, w_i)\le(u,\omega)\}]^2.
\end{aligned}
$$

Since $B_n = \kappa\beta_n + O_p(1/\sqrt{n}) \longrightarrow \kappa\beta$, the application of Chebyshev's inequality yields that as $n \to 0$,

$$
n^{-1/2}\sum_{i=1}^{n}\epsilon_i[I\{(B_n^\top x_i, w_i)\le (u,\omega)\} - I\{(\kappa\beta_n^\top x_i, w_i)\le(u,\omega)\}] \to 0.
$$

This means V_{2n} converges to zero in probability.

Consider the term V_{3n}. Since $\beta_n = \beta + O_p(n^{-1/2})$, $\theta_n = \theta + O_p(n^{-1/2})$ and $B_n = B + O_p(n^{-1/2})$, we have

$$
\begin{aligned}
V_{3n}(u,\omega) &= n^{-1/2}\sum_{i=1}^{n}(G(\beta_0^\top x_i, w_i, \theta_0) - G(\beta_n^\top x_i, w_i, \theta_n))[I\{(B_n^\top x_i, w_i)\le(u,\omega)\}\\
&\quad -I\{(\kappa\beta_n^\top x_i, w_i)\le(u,\omega)\}]\\
&\le n^{-1/2}\times O_p(n^{-1/2})\times\sum_{i=1}^{n}|I\{(B_n^\top x_i, w_i)\le(u,\omega)\} - I\{(\kappa\beta_n^\top x_i, w_i)\le(u,\omega)\}|\\
&\le O_p(1)\times(E|I\{(B_n^\top x_i, w_i)\le(u,\omega)\} - I\{(\kappa\beta_n^\top x_i, w_i)\le(u,\omega)\}| + o_p(1))\\
&= O_p(n^{-1/2}) = o_p(1).
\end{aligned}
$$

Thus, altogether, the proof is concluded. □

Proof of Lemma 2.1. From the arguments for proving Proposition 2 of Feng et al. (2013), we can see that to obtain $M_n - M = O_p(C_n)$, we only need to prove that $M_n(\mathbf{t}) - M(\mathbf{t}) = O_p(C_n)$ uniformly. In this proof, we only give the details when the partial SIR (Chiaromonte et al. 2002) is used to estimate $M(\mathbf{t})$.

Define

$$\begin{aligned}\Delta &= \big\{(1,0,\cdots,0),(0,1,0,\cdots,0)\cdots(0,1,0,\cdots,0),\\&(1,1,0,\cdots,0),(1,0,1,0,\cdots,0)\cdots(1,0,\cdots,0,1),\\&\cdots,(1,1,\cdots,1)\big\}\end{aligned}$$

Correspondingly, the number of elements in Δ is 2^{p_2}. For any given $\mathbf{t}$, $W(\mathbf{t}) = (I\{W_1 \leq t_1\},\cdots,I\{W_q \leq t_{p_2}\}) \subseteq \Delta$. For notational simplicity, we use (X_u, Y_u) to represent (X,Y) according to $\{W(\mathbf{t}) = u\}$, where $u \in \Delta$. Denote $Z_u(\tilde{t}) = I(Y_u \leq \tilde{t})$. For every $\tilde{t}$, we have

$$\Theta_u(\tilde{t}) = \Sigma_X^{-1} Var(X_u I(Y_u \leq \tilde{t})) = \Sigma_X^{-1}(\nu_1 - \nu_0)(\nu_1 - \nu_0)^\top p_u(1-p_u)$$

where Σ_X is the covariance matrix of X, $\nu_{uj} = E(X_u|Z_u(\tilde{t}) = j)$ with $j = 0$ and 1 and $p_u = E(I(Y_u \leq \tilde{t}))$. Further note that:

$$\begin{aligned}(\nu_{u1} - \nu_{u0}) &= \frac{E(X_u I(Y_u \leq \tilde{t}))}{p_u} - \frac{E(X_u I(Y_u > t))}{1-p_u}\\&= \frac{E(X_u I(Y_u \leq \tilde{t}) - E(X_u))E(I(Y_u \leq \tilde{t}))}{p_u(1-p_u)}.\end{aligned}$$

That is, $\Theta_u(\tilde{t})$ can also be rewritten as:

$$\begin{aligned}\Theta_u(\tilde{t}) &= \Sigma_X^{-1}[E\{(X_u - E(X_u))I(Y_u \leq \tilde{t})\}][E\{(X_u - E(X_u))I(Y_u \leq \tilde{t})\}]^\top\\&=: \Sigma_X^{-1} m_u(\tilde{t}) m_u(\tilde{t})^\top\end{aligned}$$

where $m_u(\tilde{t}) = E\{(X_u - E(X_u))I(Y_u \leq \tilde{t})\}$. Note that $m_u(\tilde{t})$ and $L_n(\tilde{t})$ can be respectively estimated by

$$\begin{aligned}m_{un}(\tilde{t}) &= n^{-1}\sum_{i=1}^{n}(x_{ui} - \bar{x}_u)I(y_{ui} \leq \tilde{t}),\\L_{un}(\tilde{t}) &= m_{un}(\tilde{t})m_{un}(\tilde{t})^\top,\end{aligned}$$

and then $\Theta_u(\tilde{t})$ can be estimated by

$$\Theta_{un}(\tilde{t}) = \hat{\Sigma}_X^{-1} L_{un}(\tilde{t}).$$

Note that the response under the local alternative is related to n. Then we rewrite Y_{un} and y_{uni} as the responses according to $\{W(\mathbf{t}) = u\}$ at the population and sample level. Thus from Chiaromonte et al. (2002),

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n} x_{ui} I(y_{uni} \le \tilde{t}) - E\{X_u I(Y_u \le \tilde{t})\} \\ = \; & [n^{-1}\sum_{i=1}^{n} x_{ui} I(y_{uni} \le \tilde{t}) - E\{X_u I(Y_{un} \le \tilde{t})\}] + [E\{X_u I(Y_{un} \le \tilde{t})\} - E\{X_u I(Y_n \le \tilde{t})\}] \\ = \; & O_p(n^{-1/2}) + E\{X_u I(Y_{un} \le \tilde{t})\} - E\{X_u I(Y_u \le \tilde{t})\}. \end{aligned}$$

Further, it is noted that:

$$E\{X_u I(Y_{un} \le \tilde{t})\} - E\{X_u I(Y_u \le \tilde{t})\} = E[X_u\{P(Y_{un} \le \tilde{t}|X_u)\}] - E[X_u\{P(Y_u \le \tilde{t}|X_u)\}].$$

Under H_{1n}, because $Y_n = Y + C_n G(B^\top X, W)$, we have for all $\tilde{t}$

$$\begin{aligned} P(Y_{un} \le \tilde{t}|X_u) - P(Y_u \le \tilde{t}|X_u) &= F_{Y_u|X_u}(\tilde{t} - C_n G(B^\top X_u, W_u)) - F_{Y_u|X_u}(\tilde{t}) \\ &= -C_n G(B^\top X_u, W_u) f_{Y_u|X_u}(\tilde{t}) + O_p(C_n). \end{aligned}$$

We can conclude that $n^{-1}\sum_{i=1}^{n} x_{ui} I(y_{uni} \le \tilde{t}) - E\{X_u I(Y_u \le \tilde{t})\} = O_p(\max(C_n, n^{-1/2}))$. Using the similar argument, we can prove that $m_{un}(\tilde{t}) - m_u(\tilde{t}) = O_p(C_n)$, $L_{un}(\tilde{t}) - L_{un}(\tilde{t}) = O_p(C_n)$ and $\Theta_{un}(\tilde{t}) - \Theta_u(\tilde{t}) = O_p(C_n)$. Additional, we have $M_n(\mathbf{t}) = \sum_{i=1}^{2^{p_2}} Pr(u_i)\Theta_{un}(\tilde{t})$. Following the argument for proving Theorem 3.2 in Li et al. (2008), we can derive that $M_n(\mathbf{t}) - M(\mathbf{t}) = O_p(C_n)$ uniformly. Thus, $M_n - M = O_p(C_n)$.

As Zhu and Ng (1995) and Zhu and Fang (1996) showed, under certain regularity conditions, the root-n consistency estimate M_n leads to the root-n consistency of the eigenvalues of M_n. Thus, we can get $\hat{\lambda}_i - \lambda_i = O_p(C_n)$.

Since in the result, $c_n = \log n/n$ and under the local alternative $C_n = 1/\sqrt{n}$, we have $C_n^2 = o_p(c_n)$. We have the similar statements as Theorem 2.2. Note that $\lambda_1 > 0$ and for any $l > 1$, we have $\lambda_l = 0$. The RERE criterion is computed to be

$$\begin{aligned}
\frac{\hat\lambda_2^2 + c_n}{\hat\lambda_1^2 + c_n} - \frac{\hat\lambda_{(l+1)}^2 + c_n}{\hat\lambda_l^2 + c_n} &= \frac{\lambda_2^2 + c_n + O_p(C_n^2)}{\lambda_1^2 + c_n + O_p(C_n)} - \frac{\lambda_{l+1}^2 + c_n + O_p(C_n^2)}{\lambda_l^2 + c_n + O_p(C_n^2)} \\
&= \frac{\lambda_2^2 + c_n + o_p(c_n)}{\lambda_1^2 + c_n + O_p(C_n^2)} - \frac{\lambda_{l+1}^2 + c_n + o_p(c_n)}{\lambda_l^2 + c_n + o_p(c_n)} \\
&= \frac{c_n + o_p(c_n)}{\lambda_1^2 + c_n + O_p(C_n^2)} - \frac{c_n + o_p(c_n)}{c_n + o_p(c_n)}.
\end{aligned}$$

Thus, we have

$$\frac{\hat\lambda_2^2 + c_n}{\hat\lambda_1^2 + c_n} - \frac{\hat\lambda_{(l+1)}^2 + c_n}{\hat\lambda_l^2 + c_n} \to -1$$

Therefor, we can conclude that $P(\hat q = 1) \longrightarrow 1$. □

Proof of Theorem 2.3 . Under the local alternatives (2.9), we have:

$$\begin{aligned}
V_n(u) &= n^{-1/2}\sum_{i=1}^n (y_i - G(\beta_n^\top x_i, w_i, \theta_n)) I\{(B_n^\top x_i, w_i) \le (u,\omega)\} \\
&= n^{-1/2}\sum_{i=1}^n \big(G(\beta^\top x_i, w_i, \theta) + C_n g(B^\top x_i, w_i) + \varepsilon_i \\
&\qquad - G(\beta_n^\top x_i, w_i, \theta_n)\big) I\{(B_n^\top x_i, w_i) \le (u,\omega)\} \\
&= n^{-1/2}\sum_{i=1}^n \varepsilon_i I\{(B_n^\top x_i, w_i) \le (u,\omega)\} \\
&\quad + n^{1/2} C_n n^{-1}\sum_{i=1}^n g(B^\top x_i, w_i) I\{(B_n^\top x_i, w_i) \le (u,\omega)\} \\
&\quad + n^{-1/2}\sum_{i=1}^n \{G(\beta^\top x_i, w_i, \theta) - G(\beta_n^\top x_i, w_i, \theta_n)\} I\{(B_n^\top x_i, w_i) \le (u,\omega)\} \\
&\equiv: V_{1n} + V_{2n} + V_{3n}.
\end{aligned}$$

By the Lindeberg-Levy central limit theorem, and $B_n \to \kappa\beta_0$, we have $V_{1n} \to V_\infty$, where V_∞ is given by (2.8). Let $M(X, W, u, \omega) = g(B_0^\top X, W) I\{(B_0^\top X, W) \le (u, \omega)\}$. It is easy to prove that the class of functions $\{M(X, W, u, \omega) :$ for any real valued vetor $(u, \omega)\}$ is a VC-class. An application of the uniform law of large numbers yields that in probability when $C_n = 1/\sqrt{n}$,

$$V_{2n} \to E(g(B^\top X, W) I\{(\kappa\beta_0^\top X, W) \le (u, \omega)\}).$$

Finally, under condition (A3), the similar argument for proving Theorem 1 in Stute and Zhu (2002) leads to

$$V_{3n} \to G^{\top}(\eta - V).$$

In summary, we have

(i) when $C_n = 1/\sqrt{n}$ and condition A3 holds, we have

$$\begin{aligned} V_n(u,\omega) \quad \to \quad & V_{\infty}(u,\omega) + E(g(B^{\top}X, W)I\{(\kappa\beta_0^{\top}X, W) \le (u,\omega)\}) \\ & + G^{\top}(\eta - V)(u,\omega) \end{aligned}$$

where V_{∞}, G and V are defined as Theorem 2.2;

(ii) when $\sqrt{n}C_n \to \infty$, it is obvious that $V_{2n} \to \infty$ in probability and then

$$V_n(u,\omega) \to \infty.$$

Theorem 2.3 is proved. □

Proof of Theorem 2.4 . Define

$$\tilde{\Delta}_n(u,\omega,\mathbf{U}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varrho(x_i, w_i, y_i, \beta, \theta)U_i.$$

The difference between the terms $\Delta_n(u,\omega,\mathbf{U})$ and $\tilde{\Delta}_n(u,\omega,\mathbf{U})$ is:

$$\begin{aligned} \Delta_n(u,\omega,\mathbf{U}) \quad - \quad & \tilde{\Delta}_n(u,\omega,\mathbf{U}) \\ = \quad & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\varrho}(x_i, w_i, y_i, \beta, \theta) - \varrho(x_i, w_i, y_i, \beta, \theta))U_i \\ = \quad & \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\epsilon}_i I\{(B_n^{\top}x_i, w_i) \le (u,\omega)\} - \epsilon_i I\{(B^{\top}x_i, w_i) \le (u,\omega)\})U_i \\ & + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(G^{\top}v_i - \hat{G}^{\top}\hat{v}_i)U_i \\ \equiv: \quad & \Delta_{1n} + \Delta_{2n}. \end{aligned}$$

Consider the term Δ_{1n}. Since $\hat{\epsilon}_i = y_i - G(\beta_n^\top x_i, w_i, \theta_n)$, we have

$$\begin{aligned}
\Delta_{1n}(u,\omega,\mathbf{U}) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\hat{\epsilon}_i I\{(B_n^\top x_i, w_i) \le (u,\omega)\} - \epsilon_i I\{(B^\top x_i, w_i) \le (u,\omega)\})U_i \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i[I\{(B_n^\top x_i, w_i) \le (u,\omega)\} - I\{(B^\top x_i, w_i) \le (u,\omega)\}]U_i \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(G(\beta^\top x_i, w_i, \theta) - G(\beta_n^\top x_i, w_i, \theta_n))I\{(B_n^\top x_i, w_i) \le (u,\omega)\}U_i \\
&\equiv \Delta_{11n} + \Delta_{12n}.
\end{aligned}$$

It is obvious that $E(\Delta_{11n}) = 0$. Then to obtain the convergence rate to zero, compute the second order moment of Δ_{11n} for any fixed B_n to be:

$$E(\Delta_{11n}^2) = E(\epsilon_i^2)E([I\{(B_n^\top x_i, w_i) \le (u,\omega)\} - I\{(B^\top x_i, w_i) \le (u,\omega)\}]^2)$$

Since $||B_n^\top - B|| = o_p(1)$, Chebyshiev's inequality yields $\Delta_{11n} = o_p(1)$.

As also $E(\Delta_{12n}) = 0$, we again compute the second order moment of Δ_{11n} for any fixed β_n, θ_n and B_n. It is clear that

$$E(\Delta_{12n}^2) = E\big[(G(\beta^\top x_i, w_i, \theta) - G(\beta_n^\top x_i, w_i, \theta_n))^2 I\{(B_n^\top x_i, w_i) \le (u,\omega)\}\big].$$

Since $||\beta_n - \beta|| = o_p(1)$,$||\theta_n - \theta|| = o_p(1)$ and $||B_n - B|| = o_p(1)$, Chebyshiev's inequality again implies $\Delta_{12n} = o_p(1)$.

Now deal with the term Δ_{2n}. The similar argument can be used to prove that $\Delta_{2n} = o_p(1)$. The details are omitted.

Therefor, altogether, we have

$$\Delta_n(u,\omega,\mathbf{U}) = \tilde{\Delta}_n(u,\omega,\mathbf{U}) + o_p(1).$$

Note that $E(\tilde{\Delta}_n(u,\omega,\mathbf{U})) = 0$. For any (u_1,ω_1) and (u_2,ω_2), the covariance is $\mathrm{Cov}(\tilde{\Delta}_\mathrm{n}(\mathrm{u}_1,\omega_1,\mathbf{U}), \tilde{\Delta}_\mathrm{n}(\mathrm{u}_2,\omega_2,\mathbf{U})) = \psi(\mathrm{u}_1 \wedge \mathrm{u}_2, \omega_1 \wedge \omega_2)$.

Let $M_1(X,W,u,\omega) = \varrho(X,W,Y,\beta,\theta)U$, where $\varrho(X,W,Y,\beta,\theta) = \epsilon I\{(B^\top X, W) \le (u,\omega)\} - G^\top V$, and $V = l(X,W,Y,\beta,\theta)$. It is easy to prove that the class of functions $\{M_1(X,W,u,\omega) : \text{for any real valued vetor}\,(u,\omega)\}$ is a VC-class.

An application of Functional Central Limit Theorem 10.6 of Pollard (1990) yields that the empirical process $\Delta_n(u, \omega, \mathbf{U})$ converges weakly to a mean zero Gaussian process whose covariance kernel is given by $\psi(u_1 \wedge u_2, \omega_1 \wedge \omega_2)$. By Theorem VII 21 of Pollard (1984, p. 157), the theorem is proved. □

Chapter 3

Heteroscedasticity Checks for Regression Models: A Dimension Reduction-based Model Adaptive Approach

3.1 Introduction

As heteroscedasticity structure would make a regression analysis more different than that under homoscedasticity structure, a heteroscedasticity check is required to accompany before stepping to any further analysis since ignoring the presence of heteroscedasticity may result in inaccurate inferences, say, inefficient or even inconsistent estimates. Consider a regression model with the nonparametric variance model:

$$Var(Y|X)) = Var(\varepsilon|X), \tag{3.1}$$

where Y is the response variable with the vector of covariates $X \in \mathbb{R}^p$ and the error term ϵ satisfies $E(\varepsilon|X) = 0$. Heteroscedasticity testing for the regression model has received much attention in the literature. Cook and Weisberg (1983) and Tsai (1986) proposed respectively two score tests for a parametric structure variance function under linear regression model and first-order autoregressive model. Simonoff and Tsai (1994) further developed a modified score test under linear models. Zhu et al. (2001) suggested a test that is based on squared residual-marked empirical process. Liero (2003) advised a consistent test for heteroscedasticity in nonparametric regression models, which is based on the L^2-distance between the underlying and hypothetical variance function. This test is analogous to the one proposed by Dette and Munk (1998). Dette (2002), Zheng (2009) and Zhu et al. (2015a), extended the idea of Zheng (1996), which was primitively used for testing mean regressions, to heteroscedasticity check under several different regression models. Further, Lin and Qu (2012) extended the idea of Dette (2002) to semi-parametric regressions. Moreover, Dette et al. (2007) studied a more general problem of testing the parametric form of the conditional variance under nonparametric regression model. For other references, see You and Chen (2005) and Dette and Marchlewski (2008).

The hypotheses of interest are:

$$\begin{aligned} &H_0 : \exists\ \sigma^2 > 0 \text{ s.t. } P\{Var(\varepsilon|X) = \sigma^2\} = 1 \\ &\text{against} \\ &H_1 : P\{Var(\varepsilon|X) = \sigma^2\} < 1, \forall\ \sigma^2. \end{aligned} \tag{3.2}$$

To motivate the test statistic construction, we comment on Zhu et al. (2001)'s test and Zheng (2009)'s test as the representatives of global smoothing tests and local smoothing tests, respectively. Thanks to the fact that under the null hypothesis,

$$E(\varepsilon^2 - \sigma^2|X) = 0 \Leftrightarrow E\left\{(\varepsilon^2 - \sigma^2)I(X \le t)\right\} = 0 \quad \text{for all} \quad \mathrm{t} \in \mathbb{R}^{\mathrm{p}},$$

Zhu et al. (2001) then developed a squared residual-marked empirical process as

$$V_n(x) = n^{-1/2}\sum_{i=1}^{n} \hat{\varepsilon}_i^2\{I(x_i \le x) - F_n(x)\},$$

where $\hat{\varepsilon}_i^2 = \{y_i - \hat{g}(x_i)\}^2$ and $\hat{g}(\cdot)$ is an estimate of the regression mean function. A quadratic functional form such as the Crämer-von Mises type test can be constructed. But there exist two obvious disadvantages of this global smoothing test though it works well even when the local alternative hypotheses converge to the null hypothesis at a rate of $O(1/\sqrt{n})$. First, it may be invalid in numerical studies of finite samples when the dimension of X is high. This is because the residual-marked empirical process for over heteroscedasticity involves nonparametric estimation of the mean function g and thus, the curse of dimensionality severely affects the estimation efficiency. As a local smoothing-based test, Zheng (2009)'s test can work in the scenario where the local alternative models converge to the hypothetical model at the rate of $O(n^{-1/2}h^{-p/4})$, where p denotes the dimension of the covariate X and h is a bandwidth in kernel estimation. Note that the bandwidth h converges to zero at a certain rate, which can be very slow when the dimension p is large. Local smoothing tests severely suffer from the curse of dimensionality. To illustrate those disadvantages, Figure 3.1 in Section 3.4 depicts the empirical powers of Zheng (2009)'s test and Zhu et al. (2001)'s test across the 2000 simulations with the sample size of $n = 400$ against the dimension $p = 2, 4, 6, 8, 10, 12$. This figure clearly suggests a very significant and negative impact from the dimension for the power performance of Zheng (2009)'s test and Zhu et al (2001)'s test: when p is getting larger, the power is getting down to a very low level at around 0.1 no matter the mean regression function $g(\cdot)$ is fully nonparametric or semiparametric with $\beta^\top X$ in the lieu of X. The details are presented in Section 3.4.

Therefore, how to handle the serious problem caused by dimensionality is of great importance in constructing efficient tests. The goal of the present chapter is to propose a new test procedure.

Consider a general regression model in the following form:

$$Y = g(B_1^\top X) + \delta(B_2^\top X)e, \tag{3.3}$$

where $\varepsilon = \delta(B_2^\top X)e$, B_1 is a $p \times q_1$ matrix with q_1 orthonormal columns and q_1 is a known number satisfying $1 \le q_1 \le p$, B_2 is a $p \times q_2$ matrix with q_2 orthonormal columns, q_2 is an unknown number satisfying $1 \le q_2 \le p$, e is independent of X with $E(e|X) = 0$ and the functions g and δ are unknown. This model is semiparametric

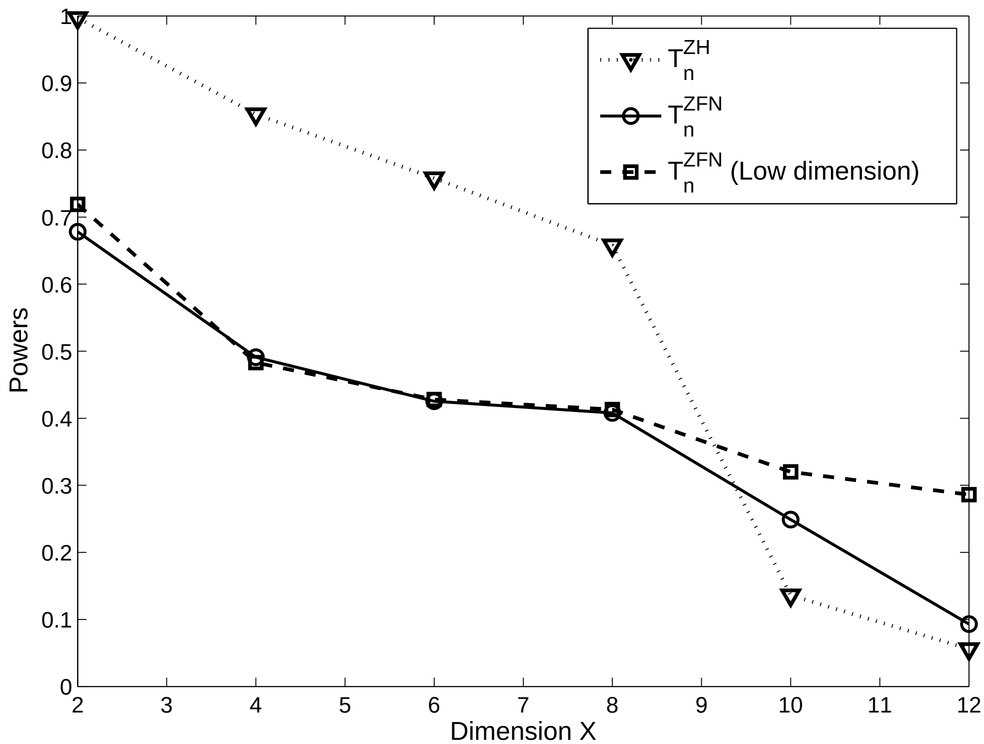

Figure 3.1: The empirical power curves of Zheng (2009)'s test and Zhu et al. (2001)'s test against the dimension of X with sample size 400 and $a = 1$ in Example 3.1. Here T_n^{ZFN} (low)' denotes Zhu et al. (2001)'s test by replacing X by $\beta^\top X$ to estimate the function $g(\cdot)$.

in the mean regression function. We assume that under the null hypothesis, the function $\delta(\cdot)$ is a constant. It is worth noting that because the functions g and δ are unknown, the following model with nonparametric variance function $\delta(\cdot)$ can also be reformulated in this form:

$$\begin{aligned} Y &= g(B_1^\top X) + \delta(X)\varepsilon = g(B_1^\top X) + \delta(B_2 B_2^\top X)e \\ &\equiv g(B_1^\top X) + \tilde{\delta}(B_2^\top X)e, \end{aligned}$$

where B_2 is any orthogonal $p \times p$ matrix. That is, $q_2 = p$. In other words, any nonparametric variance model (3.1), up to the mean function, can be reformulated as a special multi-index model with $q_2 = p$. This model covers many popularly used models in the literature, including the single-index models, the multi-index models and the partially linear single index models. When the model (3.3) is a single index model or partially linear single index model, the corresponding number of the index becomes one or two, respectively.

In this chapter, we propose a dimension reduction-based model adaptive test (DRMAT). The basic idea is to construct a test that is based on the local smoothing test proposed by Zheng (2009) when a model-adaptive strategy is utilized to adapt the structures under hypothetical model and alternatives. The method is motivated by Guo et al. (2014) who considered model checks for mean regression function. However, the construction is very different as the test not only use the model structure of conditional variance, but also the dimension reduction structure of the mean function. The advantages of this method include: (1) DRMAT computes critical values by simply applying its limiting null distribution without heavy computational burden, which is often an inherent property of local smoothing testing methodologies; (2) the embedded dimension reduction procedure is model-adaptive, that is, it is automatically adaptive to the underlying model (3.3) by using more information on data such that the test can still be omnibus; more importantly, (3) under the null hypothesis, DRMAT has a significant faster convergence rate of $O(n^{1/2}h^{q_1/4})$ to its limit than $O(n^{1/2}h^{p/4})$ in existing tests when $q_1 \ll p$; and (4) DRMAT can also detect the local alternative hypotheses converging to the null hypothesis at a much faster rate of $O(n^{-1/2}h^{-q_1/4})$ than the typical rate of $O(n^{-1/2}h^{-p/4})$. More details are presented in the next section.

The rest of the chapter is organized as follows: In Section 3.2, we will give a brief review for the discretization-expectation estimation and suggest a minimum ridge-type eigenvalue ratio to determine the structural dimension of the model. Moreover, the dimension reduction adaptive heteroscedasticity test is also constructed in Section 3.2. The asymptotic properties of the proposed test statistic under the null and alternative hypotheses are investigated in Section 3.3. In Section 3.4, the simulation results are reported and a real data analysis is carried out for illustration. Because when there are no more specific conditional variance structure assumed, the convergence rate $O(n^{-1/2}h^{-q_1/4})$ is optimal for local smoothing tests, thus, in Section 3.5, we discuss how to further improve the performance of DRMAT when there are some specific conditional variance structures. The proofs of the theoretical results are postponed to Appendix 2.

3.2 Test statistic construction

3.2.1 Basic construction

Under the model (3.3), the null hypothesis is

$$H_0 : P\{Var(\varepsilon|X) = Var(\varepsilon|B_1^\top X) = \sigma^2\} = 1 \quad \text{for some } \sigma^2$$

and the alternative hypothesis is

$$H_1 : P\{Var(\varepsilon|B_2^\top X) = \sigma^2\} < 1, \quad \text{for all } \sigma^2.$$

Write B to be a $p \times q$ matrix where q orthogonal columns contained in the matrix (B_1, B_2). Then, under the null hypothesis, we have the following moment condition:

$$E(\varepsilon^2 - \sigma^2|X) = 0.$$

All existing local smoothing tests are based on this equation when the left hand side is estimated by a chosen nonparametric smoother such as kernel estimator. As was mentioned before, this severely suffers from the curse of dimensionality. Note that, under the null hypothesis, it is unnecessary to use this technique because $E(\varepsilon^2 - \sigma^2|X) = E(\varepsilon^2) - \sigma^2$. Thus, how to sufficiently use the information provided by the hypothetical model is a key to improve the efficiency of a test. It is clear that we cannot simply use two estimates in lieu of $E(\varepsilon^2)$ and σ^2 respectively to construct a test. Therefore, we consider the following idea. Note that under the null hypothesis, B_2 needs not to consider and thus, B is reduced to B_1 and

$$E(\varepsilon^2 - \sigma^2|X) = E(\varepsilon^2 - \sigma^2|B^\top X) = 0,$$

and then

$$E[(\varepsilon^2 - \sigma^2)E(\varepsilon^2 - \sigma^2|B^\top X)W(B^\top X)] = E[E^2(\varepsilon^2 - \sigma^2|B^\top X)W(B^\top X)] = 0, \quad (3.4)$$

where $W(\cdot)$ is some positive weight function which will be specified latter. Under the alternative hypothesis,

$$E(\varepsilon^2 - \sigma^2|X) = E(\varepsilon^2 - \sigma^2|B^\top X) \neq 0.$$

and then the left hand side of (3.4) is greater than zero. Thus, its empirical version, as a base, can be devoted to constructing a test statistic. The null hypothesis is rejected for large values of the test statistic. We note that there are two identifiability issues.

1. First, for any $q \times q$ orthogonal matrix C (3.4)holds true when the matrix B is replaced by $BC^\top$. This means B is not identifiable while $BC^\top$ for an orthogonal matrix C is. But such an unidentifiabiliy problem does not affect the properties under the null and alternative hypotheses as for any $q \times q$ orthogonal matrix C, under the null hypothesis,

$$E(\varepsilon^2 - \sigma^2|X) = E(\varepsilon^2 - \sigma^2|CB^\top X) = 0,$$

and under the alternative hypothesis, this function is not equal to a zero function. Thus, in the following, we write $BC^\top$ as B in the estimation procedure without notational confusion.

2. Second, under the null hypothesis, B is reduced to B_1 with q_1 columns and under the alternative hypothesis, B has q columns. Ideally, B can be identified to be B_1 under the null hypothesis such that we can reduce the dimension of B from q to q_1 and the nonparametric estimation can be lower-dimensional. On the other hand, under the alternative hypothesis, we wish to keep B such that the criterion can fully use the information provided by the alternative hypothesis and the constructed test can be omnibus. To achieve this goal, we need a test that can automatically adapt the projected $B^\top X$ with the respective dimension q_1 and q.

In the following estimation procedure, we introduce a model-adaptive approach. Let $\{(x_1, y_1), \cdots (x_n, y_n)\}$ denote an $i.i.d$ sample from (X, Y) and $\varepsilon_i = y_i - g(B^\top x_i)$. Then $E(\varepsilon^2 - \sigma^2 | B^\top X)$ can be estimated by the following form:

$$\hat{E}(\varepsilon^2 - \sigma^2 | \hat{B}_{\hat{q}}^\top x_i) = \frac{\frac{1}{(n-1)} \sum_{j \neq i, j=1}^{n} K_h\left(\hat{B}_{\hat{q}}^\top x_j - \hat{B}_{\hat{q}}^\top x_i\right)(\hat{\varepsilon}_j^2 - \hat{\sigma}^2)}{\frac{1}{(n-1)} \sum_{j \neq i, j=1}^{n} K_h\left(\hat{B}_{\hat{q}}^\top x_j - \hat{B}_{\hat{q}}^\top x_i\right)},$$

where $\hat{\varepsilon}_i^2 = (y_i - \hat{g}(\hat{B}^\tau x_i))^2$, $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \hat{\varepsilon}_i^2$, $K_h(\cdot) = K(\cdot/h)/h^{\hat{q}}$ with a $\hat{q}-$dimensional multivariate kernel function $K(\cdot)$, h is a bandwidth and $\hat{B}_{\hat{q}}$ is an estimate of B with an estimate $\hat{q}$ of q, which will be discussed later. We choose the weight function $W(\cdot)$ to be the density function $p(\cdot)$ of $\hat{B}_{\hat{q}}^\top X$, and for any $\hat{B}_{\hat{q}}^\top X$, we can estimate the density function $p(\cdot)$ as the following form:

$$\hat{p}(\hat{B}_{\hat{q}}^\top x_i) = \frac{1}{(n-1)} \sum_{j \neq i, j=1}^{n} K_h\left(\hat{B}_{\hat{q}}^\top x_j - \hat{B}_{\hat{q}}^\top x_i\right).$$

Therefore, a non-standardized test statistic can be constructed as S_n by:

$$S_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i, j=1}^{n} K_h\left(\hat{B}_{\hat{q}}^\top x_i - \hat{B}_{\hat{q}}^\top x_j\right)(\hat{\varepsilon}_i^2 - \hat{\sigma}^2)(\hat{\varepsilon}_j^2 - \hat{\sigma}^2). \tag{3.5}$$

The resulting test statistic is

$$nh^{\frac{q_1}{2}} S_n. \tag{3.6}$$

Remark 3.1. *From the construction, it seems that except an estimate of* B*, the test statistic has no difference in spirit from that by Zheng (2009) as follows:*

$$\tilde{S}_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i, j=1}^{n} \tilde{K}_h\left(x_i - x_j\right)(\hat{\varepsilon}_i^2 - \hat{\sigma}^2)(\hat{\varepsilon}_j^2 - \hat{\sigma}^2), \tag{3.7}$$

where $\tilde{K}_h(\cdot) = \tilde{K}(\cdot/h)/h^p$ *with a* p*-dimensional multivariate kernel function* $\tilde{K}(\cdot)$*. Our test statistic is even more complicated in the case where* $q = p$ *because even* q *is given, our test still involves an estimate of* q*. However, we note that if we do not have this estimate that can automatically adapt to* q_1 *and* q*, we do not have the chance to construct a test that can use the standardizing constant* $nh^{\frac{q_1}{2}}$*. Zheng (2009)'s test statistic must use, actually all existing local smoothing tests, the standardizing constant* $nh^{\frac{p}{2}}$*, otherwise, the limit goes to infinity if* $nh^{\frac{q_1}{2}}$ *is used. Comparing (3.5) with (3.7), we observe that the different dimensions of the kernel estimators in* S_n *and* $\tilde{S}_n$ *make this significant improvement of* S_n *than* $\tilde{S}_n$*. Clearly, under the null hypothesis, the curse of dimensionality is largely avoided. As we will see in Section 3.3,* S_n *is asymptotically normal at the rate of order* $nh^{q_1/2}$ *under the null hypothesis, whereas* $\tilde{S}_n$ *has the asymptotic normality at the rate of order* $nh^{p/2}$*. Particularly, when the model (3.3) is a single-index model or generalized linear model,* $q_1 = 1$*. More importantly, in Section 3.3, we will show that our test can be much more sensitive than existing local smoothing tests in the sense that it can detect local alternatives converging to the null hypothesis at the rate of* $1/(\sqrt{n}h^{q_1/4})$ *that is a much faster rate than* $1/(\sqrt{n}h^{p/4})$ *existing local smoothing tests can achieve. This gain is again due to the adaptive estimation of the matrix* B *under the null and alternative hypotheses respectively. In the next section, we will use sufficient dimension reduction technique to construct an estimate of* q*.*

Remark 3.2. *In the construction, we consider the conditional expectation given* $B^\top X$ *where* B *consists of both* B_1 *and* B_2*. Under the null hypothesis,* B *is reduced to* B_1*,* *but under the alternative hypothesis, the dimension q may be greater than q_2 and thus, we have to use higher order kernel in local smoothing step. A natural idea is to estimate B_1 and B_2 separately. However, we find that such a procedure makes the implementation more complicated. Thus, in the present chapter, we do not use this idea and a further research is ongoing to see how to improve the performance of the test.*

3.2.2 A review on discretization-expectation estimation

As illustrated in Subsection 3.2.1, estimating B plays an important role for the test statistic construction. To this end, we consider sufficient dimension reduction technique. Since B is not identifiable in the model (3.3) and the functions $g(\cdot)$ and $\delta(\cdot)$ are unknown, what we can identify is $BC^\top$ because for any $q \times q$ orthogonal matrix C, $g(B^\top X)$ and $\delta(B^\top X)$ can be further rewritten as $\tilde{g}(C^\top B^\top X)$ and $\tilde{\delta}(C^\top B^\top X)$. Sufficient dimension reduction technique helps us identify the subspace spanned by B called the central subspace (Cook 1998). More precisely, from the definition of the central subspace, it is the intersection of all subspaces S such that

$$Y \perp\!\!\!\perp X | (P_S X),$$

where $\perp\!\!\!\perp$ denotes the statistics independence and $P_{(\cdot)}$ stands for a projection operator with respect to the standard inner product. $\dim(\mathrm{P}_S\mathrm{X})$ is called the structure dimension of $P_S X$ and is q in our setup. In other words, P_S is equal to $CB^\top$ for some $q \times q$ orthogonal matrix C. As we mentioned, we still use B without confusion.

There exist several promising dimension reduction proposals available in the literature. For example, Li (1991) proposed sliced inverse regression (SIR), Cook and Weisberg (1991) advised sliced average variance estimation (SAVE), Xia et al. (2002) discussed minimum average variance estimation (MAVE), and Zhu et al. (2010a) suggested discretization-expectation estimation (DEE). As DEE does not need to select the number of slices and has been proved in the simulation studies to have a good performance, we then adopt it to estimate B for the test statistic construction. From Zhu et al. (2010a), the SIR-based DEE can be carried out by the following estimation steps.

1. Discretize the response variable Y into a set of binary variables by defining $Z(t) = I\{Y \leq t\}$, where the indicator function $I\{Y \leq t\}$ takes value 1 if $Y \leq t$ and 0 otherwise.

2. Let $S_{Z(t)|X}$ denote the central subspace of $Z(t)|X$. When SIR is used, the related SIR matrix $M(t)$ is an $p \times p$ positive semi-definite matrix satisfying that $\mathrm{Span\{M(t)\}} = S_{\mathrm{Z(t)|X}}$.

3. Let $\tilde{Y}$ be an independent copy of Y. The target matrix is $M = E\{M(\tilde{Y})\}$. B consists of the eigenvectors associated with the nonzero eigenvalues of M.

4. Obtain an estimate of M as:

$$M_n = \frac{1}{n}\sum_{i=1}^{n} M_n(y_i),$$

   where $M_n(y_i)$ is the estimate of the SIR matrix $M(y_i)$. When q is given, an estimate $\hat{B}_q$ of B consists of the eigenvectors associated with the largest q eigenvalues of M_n. $\hat{B}_q$ can be root-n consistent to B. More details can be referred to Zhu et al. (2010a).

3.2.3 The estimation of structural dimension

To completely estimate B, we also need to estimate the structural dimension q. Thus, an estimate of q is essential for the test statistic. The BIC-type criterion was suggested by Zhu et al. (2010a). However, choosing an appropriate tuning parameter is an issue. Thus, we suggest another method that is very easy to implement. In the aforementioned sufficient dimension reduction procedure, the estimating matrix M_n is a root-n consistent estimation of the target matrix M. Let $\hat{\lambda}_p \leq \hat{\lambda}_{(p-1)} \leq \cdots \leq \hat{\lambda}_1$

be the eigenvalues of the estimating matrix M_n. In spirit similar to that in Xia et al. (2014), we advise a ridge-type eigenvalue ratio estimate (RERE) to determine q as:

$$\hat{q} = \arg\min_{1\leq j\leq p}\left\{\frac{\hat{\lambda}_{j+1}^2 + c_n}{\hat{\lambda}_j^2 + c_n}\right\}. \tag{3.8}$$

The following theorem shows that the structure dimension q can be consistently determined by RERE criterion.

Theorem 3.1. *Under Conditions A1 and A2 in Appendix 2, the estimate $\hat{q}$ of (3.8) with $\frac{\log n}{n} \leq c_n \to 0$ satisfies that as $n \to 0$ in a probability going to* 1*,*

(i) under H_0, $\hat{q} \to q_1$;

(ii) under H_1, $\hat{q} \to q$.

This theorem implies that, in the test statistic construction, the structure dimension estimate $\hat{q}$ can be automatically adaptive to the model (3.3) rather than the nonparametric model (3.1). An consistent estimate of B is denoted by $\hat{B}_{\hat{q}}$. In the above test statistic construction, this estimate plays a crucial role as $\hat{B}_{\hat{q}}$ converges to a $p \times q$ matrix B under H_1 and to a $p \times q_1$ matrix B_1 under H_0.

3.3 Asymptotic properties

3.3.1 Limiting null distribution

Define two notations first. Let

$$s^2 = 2\int K^2(u)du E\{[Var(\varepsilon^2|B^\top X)]^2 p(B^\top X)\}, \tag{3.9}$$

and

$$\hat{s}^2 = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i, j=1}^{n} K_h^2\left(\hat{B}_{\hat{q}}^\top x_i - \hat{B}_{\hat{q}}^\top x_j\right)(\hat{\varepsilon}_i^2 - \hat{\sigma}^2)^2(\hat{\varepsilon}_j^2 - \hat{\sigma}^2)^2. \tag{3.10}$$

We will prove that $\hat{s}^2$ is a consistent estimate of s^2 under the null and local alternative hypotheses. Further, we have the following asymptotic properties of the test statistic under the null hypothesis.

Theorem 3.2. *Given Conditions A1-A8 in Appendix 2 and under H_0, we have*

$$nh^{\frac{q_1}{2}} S_n \xrightarrow{\text{d}} N(0, s^2),$$

where the notation $\xrightarrow{\text{d}}$ denotes convergence in distribution and s^2 is defined by (3.9). Further, s^2 can be consistently estimated by $\hat{s}^2$ given by (3.10).

According to Theorem 3.2, by standardizing S_n, we then get an standardized test statistic T_n as:

$$T_n = nh^{q_1/2} S_n / \hat{s}. \tag{3.11}$$

Furthermore, using the Slusky theorem yields the following corollary.

Corollary 3.1. *Under the conditions in Theorem 3.2 and H_0, we have*

$$T_n \xrightarrow{\text{d}} N(0, 1),$$

and then

$$T_n^2 \xrightarrow{\text{d}} \chi_1^2,$$

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

Based on Theorem 3.2 and Corollary 3.1, it is easy to calculate $p-$values by its limiting null distribution of T_n^2. As a popularly used approach, Monte Carlo simulation can also be employed.

Summarizing all the aforementioned constructing procedure gives the following steps:

Step 1: Use a sufficient dimension reduction method such as DEE (Zhu et al. 2010) to detain the estimators $\hat{B}_{\hat{q}}$ with $\hat{q}$ determined by MRRE criterion, and then apply the nonparametric kernel method to estimate the mean function $g(\cdot)$ as follows:

$$\hat{g}(\hat{B}_{\hat{q}}^\top x_i) = \frac{\sum_{i=1}^n Q_{h_1}\left(\hat{B}_{\hat{q}}^\top x_j - \hat{B}_{\hat{q}}^\top x_i\right) y_i}{\sum_{i=1}^n Q_{h_1}\left(\hat{B}_{\hat{q}}^\top x_j - \hat{B}_{\hat{q}}^\top x_i\right)},$$

where $Q_{h_1}(\cdot) = Q(\cdot/h_1)/h_1$ with $Q(\cdot)$ being a $\hat{q}-$dimensional kernel function and h_1 being a bandwidth.

Step 2: Calculate the test statistic $T_n = nh^{q_1/2}S_n/\hat{s}$ with S_n and $\hat{s}$ given by (3.5) and (3.10), respectively.

3.3.2 Power study

To study the power performance of the proposed test statistic, consider the sequence of local alternative hypotheses with the following form:

$$H_{1n} : Y = g(B_1^\top X) + \eta, Var(\eta|X) = \sigma^2 + C_n f(B^\top X) \tag{3.12}$$

where $f(\cdot)$ is some continuously differentiable function satisfying $E[f^2(X)] < \infty$ and the columns of B_1 can be the linear combination of the columns of B.

Under the global alternative with fixed C_n, Theorem 3.1 shows that $\hat{q}$ converges to q in a probability going to 1. Under the local alternative hypotheses, C_n goes to zero and the part B_2 vanishes as $n \to \infty$, we may expect that $\hat{q}$ also tends to q_1 although at the population level, the structural dimension is still q. In other words, the estimate $\hat{q}$ is not consistent. But this is just what we want because it will make the test more sensitive to the local alternative hypotheses. The following states this result.

Lemma 3.1. *Under the local alternative hypotheses H_{1n} in (3.12) with $C_n = n^{-\frac{1}{2}}h^{-\frac{q_1}{4}}$ and the same conditions in Theorem 3.1 except that $C_n^2 \log n \le c_n \to 0$, the estimate $\hat{q}$ given by (3.8) satisfies that $\hat{q} \to q_1$ in probability as $n \to 0$.*

Now, we state the power performance of the test.

Theorem 3.3. *Under Conditions A1-A8 in Appendix 2, we have the following results.*

(I) Under the global alternative hypothesis H_1, we have

$$S_n \xrightarrow{p} E\{[Var(\varepsilon|B^\top X) - Var(\varepsilon)]^2 p(B^\top X)\},$$

and

$$\hat{s}^2 \xrightarrow{\mathrm{p}} 2\int K^2(u)duE\{[Var(\varepsilon^2|B^\top X) + (Var(\varepsilon|B^\top X) - \sigma^2)^2]^2 p(B^\top X)\},$$

where the notation $\xrightarrow{\mathrm{p}}$ *denotes convergence in probability and* $\hat{s}^2$ *is defined in (3.10). Thus,*

$$T_n/(nh^{q_1/2}) \xrightarrow{\mathrm{p}} Constant.$$

(II) *Under the local alternative hypotheses* H_{1n} *in (3.12) with* $C_n = n^{-\frac{1}{2}}h^{-\frac{q_1}{4}}$, *we have*

$$T_n \xrightarrow{\mathrm{d}} N(m, 1),$$

and

$$T_n^2 \xrightarrow{\mathrm{d}} \chi_1^2(m^2),$$

where $m = E\{[E\{f(B^\top X)|B_1^\top X\}]^2 p(B_1^\top X)\}/s$ *with* s *given by (3.9) and* $\chi_1^2(m^2)$ *is a noncentral chi-squared random variable with one degree of freedom and the noncentrality parameter* m.

Remark 3.3. *The above results confirm the claims that we made in Section 3.1. Unlike the convergence rates in Zheng (2009), our test can have the following rates under the null and alternative hypotheses. (I) Under the null hypothesis,* S_n *converges to its limit at the rate of order* $nh^{q_1/2}$ *whereas* $\tilde{S}_n$ *has a much slower rate of order* $nh^{p/2}$. *Particularly, when the null hypothesis belongs to the single index models or the generalized linear models,* S_n *has a fastest convergence rate as if the dimension* $p = 1$, *namely,* $nh^{1/2}$. *(II)* T_n *can detect the local alternative models converging to the hypothetical model at the rate of order* $n^{-\frac{1}{2}}h_1^{-\frac{q_1}{4}}$ *rather than* $n^{-\frac{1}{2}}h_1^{-\frac{p}{4}}$ *that Zheng (2009)'s test can achieve.*

3.4 Numerical Studies

3.4.1 Simulations

In this subsection, we conduct the following simulations to illustrate the performance of the proposed test. We choose the product of $\hat{q}$ Quartic kernel function as $K(u) = Q(u) = 15/16(1-u^2)^2$, if $|u| \le 1$ and 0 otherwise. In the test statistic, B is estimated by the SIR-based DEE procedure and q is determined by the RERE criterion (3.8) with $c_n = \log n/(nh^{q_1/2})$. Let T_n^{DEE}, T_n^{ZH} and T_n^{ZFN} denote the proposed test in the present chapter, Zhu et al. (2002)'s test and Zheng (2009)'s test, respectively. We focus on the performance of these tests under different settings of the dimension and the correlation structure of the covariate vector X and the distribution of the error term ε. The sample sizes are $50, 200, 400$. The empirical sizes and powers are computed through 2000 replications for each experiment at the significance level $\alpha = 0.05$.

The observations x_i, for $i = 1, 2, \cdots, n$ are i.i.d. from multivariate normal distribution $N(0, \Sigma_1)$ or $N(0, \Sigma_2)$ and independent of the standard normal errors ε, where $\Sigma_1 = (\sigma_{ij}^{(1)})_{p\times p}$ and $\Sigma_2 = (\sigma_{ij}^{(2)})_{p\times p}$ with the elements respectively

$$\sigma_{ij}^{(1)} = I(i=j) + 0.5^{|i-j|} I(i \ne j) \quad \text{and} \quad \sigma_{ij}^{(2)} = I(i=j) + 0.3 I(i \ne j).$$

Example 3.1. Consider the following single-index model:

$$Y = \beta^\top X + \exp(-(\beta^\top X)^2) + 0.5(1 + a \times |\beta^\top X|) \times \epsilon,$$

where X follows normal distribution $N(0, \Sigma_1)$, independent of the standard normal errors ϵ and $\beta = (\underbrace{1, \cdots, 1}_{p/2}, 0, \cdots, 0)^\tau / \sqrt{p/2}$ and p is set to be 2, 4 and 8 to reveal the impact from dimension. Moreover, $a = 0$ and $a \ne 0$ respectively, correspond to the null and the alternative hypotheses. First, we investigate the impact from bandwidth selection. We choose different bandwidths $(0.5 + 0.25 \times i) n^{-1/(4+\hat{q})}$ for $i = 0, \cdots, 5$ and obtain the empirical sizes and powers when the dimension of X is relatively high, namely, $p = 4, 8$.

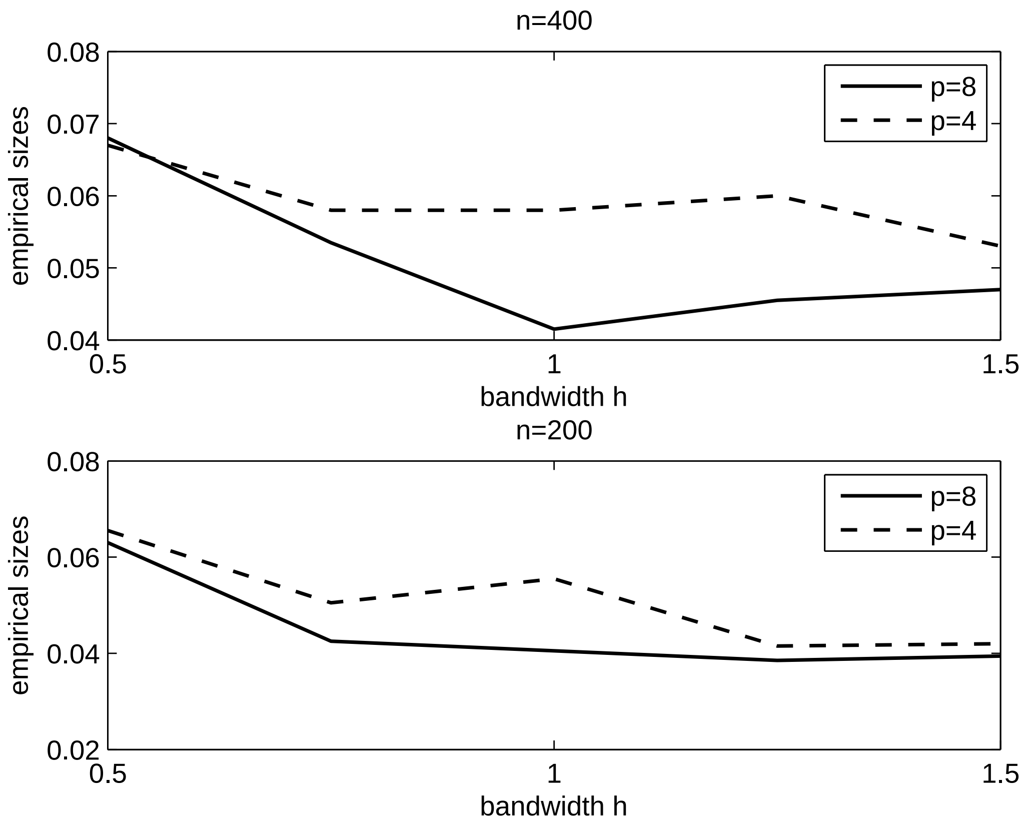

Figure 3.2: The empirical size curves of T_n^{DEE} against the bandwidth and sample size 200 (the below panel) and 400 (the above panel) with $a = 0$ in Example 3.1.

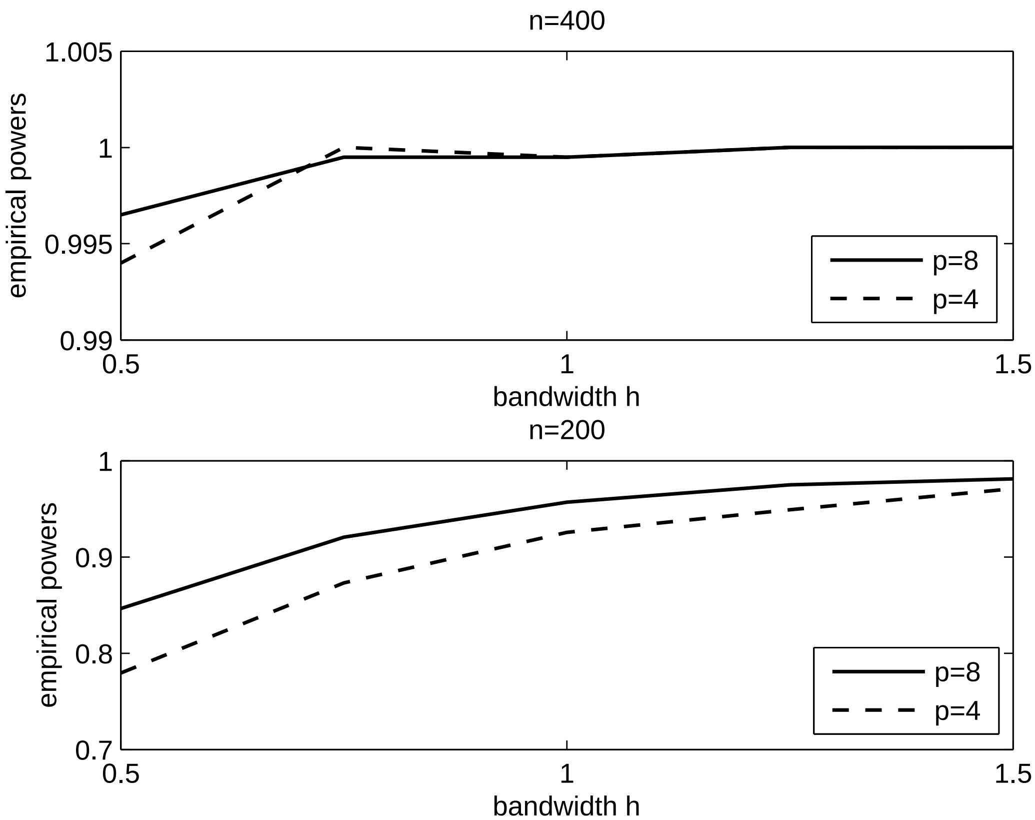

Figure 3.3: The empirical power curves of T_n^{DEE} against the bandwidth and sample size 200 (the below panel) and 400 (the above panel) with $a = 1$ in Example 3.1.

From Figures 3.2 and 3.3, it can be clearly observed that the test is robust against different bandwidths and the type I error can be controlled well. On the other hand, when the sample size is small, the bandwidth selection has little impact for the power performance when the bandwidth becomes small. Therefore, we recommend $h = 1.25 \times n^{-1/(4+\hat{q})}$ in the simulations.

Now we turn to compare the empirical sizes (type I errors) and powers of all the three tests under different combinations of sample sizes and dimensions of covariate vector X. The results are reported in Table 3.1.

Table 3.1: Empirical sizes and powers of T_n^{DEE}, T_n^{ZH} and T_n^{ZFN} for Example 3.1.

		T_n^{DEE}			T_n^{ZH}			T_n^{ZFN}		
	a/n	50	200	400	50	200	400	50	200	400
p=2	0	0.0370	0.0545	0.0525	0.0365	0.0415	0.0445	0.0455	0.0450	0.0550
	0.2	0.0535	0.1540	0.2680	0.0370	0.0490	0.1130	0.0620	0.0645	0.0865
	0.4	0.0900	0.3975	0.7675	0.0430	0.1610	0.5125	0.0710	0.1010	0.1790
	0.6	0.1530	0.6600	0.9870	0.0515	0.3595	0.8645	0.0720	0.1320	0.3815
	0.8	0.1975	0.8665	0.9960	0.0640	0.5605	0.9790	0.0780	0.1785	0.5385
	1.0	0.2520	0.9790	1.0000	0.0665	0.7260	0.9970	0.0820	0.1865	0.6780
p=4	0	0.0450	0.0535	0.0525	0.0105	0.0320	0.0700	0.0670	0.0650	0.0590
	0.2	0.0540	0.1570	0.3750	0.0377	0.0670	0.1920	0.0795	0.0750	0.1210
	0.4	0.0935	0.4990	0.8835	0.0305	0.0865	0.3525	0.0800	0.1160	0.1670
	0.6	0.1420	0.7895	0.9930	0.0265	0.3105	0.6810	0.0840	0.1230	0.2390
	0.8	0.2110	0.9265	1.0000	0.0365	0.4500	0.8050	0.1030	0.1970	0.3800
	1.0	0.2580	0.9650	1.0000	0.0565	0.5545	0.8535	0.1050	0.2900	0.4910
p=8	0	0.0460	0.0425	0.0535	0.0145	0.0225	0.0345	0.0420	0.0640	0.0580
	0.2	0.0620	0.1790	0.4380	0.0110	0.0550	0.1675	0.0500	0.0700	0.0975
	0.4	0.1270	0.4220	0.9530	0.0115	0.1175	0.3105	0.0580	0.0930	0.2040
	0.6	0.1740	0.8840	0.9870	0.0230	0.1910	0.4245	0.0555	0.1250	0.2605
	0.8	0.2090	0.9630	1.0000	0.0260	0.2490	0.5225	0.0660	0.1485	0.3575
	1.0	0.2390	0.9850	1.0000	0.0270	0.3640	0.6580	0.0715	0.1515	0.4080

It is clear that the empirical power increases as a gets larger. Further, our test T_n^{DEE} is significantly more powerful than T_n^{ZH} and T_n^{ZFN}. In addition, T_n^{DEE} can control the size very well for the different dimensions of X and the different sample sizes. However, for T_n^{ZH} and T_n^{ZFN}, the impact from the dimension of X is very significant. When the dimension of X becomes larger, the empirical power of our test slightly changes whereas the empirical powers of T_n^{ZH} and T_n^{ZFN} drop down quickly. Even when the sample size is $n = 400$, the situation does not become significantly better. This implies that the dimensionality is a big obstacle for these two tests to perform well.

Next, we use the following example that has two different B under the null and alternative hypotheses to check the usefulness of model-adaptiveness in promoting the performance of test. Further, we examine the robustness of the test against different error ε.

Example 3.2. Consider the following model in which the dimension of B_1 and that of B are different under H_1:

- $Y = \beta_1^\top X + 0.5[a \times \{(\beta_1^\top X)^2 + (\beta_2^\top X)^2\} + 1] \times \epsilon.$

where $\beta_1 = (1, 1, 0, 0)^\tau/\sqrt{2}$ and $\beta_2 = (0, 0, 1, 1)^\tau/\sqrt{2}$. Further, ϵ follows the Student's t-distribution $t(6)$ with degrees of freedom 6. Two cases are investigated, where X follows $N(0, \Sigma_1)$ and $N(0, \Sigma_2)$, respectively. In this example, under H_0 with $a = 0$, $B = \beta_1$ and under H_1 with $a \neq 0$, $B = (\beta_1, \beta_2)$. The results are presented in Table 3.2.

From Table 3.2, we have the following findings. First, the comparison between the two cases of this example shows that the correlation structure of X would not deteriorate the power performance. Second, in the limited simulations, the heavy tail of the error term does not have a significant impact on the performance of our test. Third, we can observe that although B has higher structure dimension $q = 2$ under the alternative hypothesis than $q_1 = 1$ under the null hypothesis in this example, the results are still similar to those in Example 3.1 that has the same B in the hypothetical and alternative model. These findings suggest that the proposed test is robust against the correlation structure of X and the different error ε. The power performance is less negatively affected by the dimension under the alternative model.

Example 3.3. To further examine the performance of te proposed test, consider

Table 3.2: Empirical sizes and powers of T_n^{DEE}, T_n^{ZH} and T_n^{ZFN} for Example 3.2.

		T_n^{DEE}			T_n^{ZH}			T_n^{ZFN}		
	a/n	50	200	400	50	200	400	50	200	400
$X \sim N(0, \Sigma_1)$	0	0.0580	0.0540	0.0535	0.0210	0.0565	0.0590	0.0730	0.0780	0.0900
	0.2	0.1115	0.5145	0.8985	0.0225	0.2295	0.5028	0.0850	0.1530	0.3415
	0.4	0.1635	0.8190	0.9775	0.0238	0.3045	0.5440	0.1055	0.2220	0.5540
	0.6	0.2120	0.8515	0.9835	0.0535	0.4047	0.5875	0.1215	0.2505	0.6310
	0.8	0.2615	0.8750	0.9830	0.1080	0.5005	0.6865	0.1300	0.2620	0.6685
	1.0	0.2820	0.9095	0.9930	0.1340	0.5925	0.7785	0.1360	0.2980	0.6740
$X \sim N(0, \Sigma_2)$	0	0.0560	0.0450	0.0465	0.0375	0.0425	0.0605	0.0680	0.0590	0.0560
	0.2	0.1135	0.5655	0.9245	0.0520	0.1205	0.1800	0.0830	0.1855	0.4090
	0.4	0.1970	0.8575	0.9725	0.0980	0.3530	0.5300	0.0935	0.2955	0.5450
	0.6	0.2505	0.9060	0.9860	0.1085	0.5325	0.7165	0.1010	0.3000	0.6425
	0.8	0.2790	0.9205	0.9900	0.1390	0.6185	0.7875	0.1335	0.3120	0.6590
	1.0	0.3155	0.9335	0.9940	0.1615	0.6750	0.8105	0.1220	0.3310	0.6645

the following model:

- $Y = \beta_1^\top X + 2\sin(\beta_2^\top X/2) + 0.5[a \times \{(\beta_1^\top X)^2 + (\beta_2^\top X)^2\} + 1]^{0.5} \times \epsilon.$

where the values of β_1, β_2 and p are set to be the same as those in example 3.2 and ϵ is from normal distribution. In this example, B is identical under the null and alternative hypotheses as that in Example 3.1, but $q = q_1 = 2$. Because the results of T_n^{ZH} and T_n^{ZFN} are similar to those in Example 3.2, we omit to present the detailed results and only present the results of our test T_n^{DEE} to save the space. By the comparison between Tables 3.2 and 3.3, we can find that the test power with $q_1 = 2$ in Example 3.3 is lower than that with $q_1 = 1$ in Example 3.2.

These numerical results support the aforementioned theoretical results indicating that DRMAT has significantly improved the performance of existing local smoothing tests. The empirical sizes also show that, in our test, critical values computed by simply applying the limiting null distribution is reliable. Hence, the computational workload of DRMAT is not heavy.

Table 3.3: Empirical sizes and powers of T_n^{DEE} for Example 3.3.

		T_n^{DEE}		
	a/n	50	200	400
$X \sim N(0, \Sigma_1)$	0	0.0490	0.0530	0.0465
	0.2	0.0550	0.3280	0.6490
	0.4	0.1035	0.5965	0.9240
	0.6	0.1360	0.7250	0.9780
	0.8	0.1570	0.8110	0.9895
	1.0	0.1930	0.8665	0.9955
$X \sim N(0, \Sigma_2)$	0	0.0615	0.0460	0.0480
	0.2	0.0615	0.2815	0.6445
	0.4	0.0965	0.5920	0.9300
	0.6	0.1395	0.7750	0.9910
	0.8	0.1475	0.8425	0.9920
	1.0	0.1830	0.9015	0.9980

3.4.2 Real Data Analysis

We consider the well-known 1984 Olympic records data on various track events, which has been analyzed by Naik and Khattree (1996) using the method of principal component analysis for the investigation of their athletic excellence and the relative strength on certain countries at the different running. Further, Zhu (2003) once analyzed this data in checking certain parametric structure. The data for men consists of 55 countries with eight running events presented, which are the 100 meters, 200 meters, 400 meters, 800 meters, 1,500 meters, 5,000 meters, 10,000 meters and the Marathon distance, see Naik and Khattree (1996).

As argued by Naik and Khattree (1996), it may be more tenable to use the speed rather than the winning time for the study. Here, what we are interested in is to examine whether the performance of a nation in running long distances has a significant effect on that in short running speed, see Zhu (2003). We also take the speed of the 100 meters running event as the response and the speed of the 1,500 meters, 5,000 meters, 10,000 meters and the Marathon distance as covariates. Figure (3.4) presents the plots of the residuals versus $B^{\top}X$ with the different bandwidth

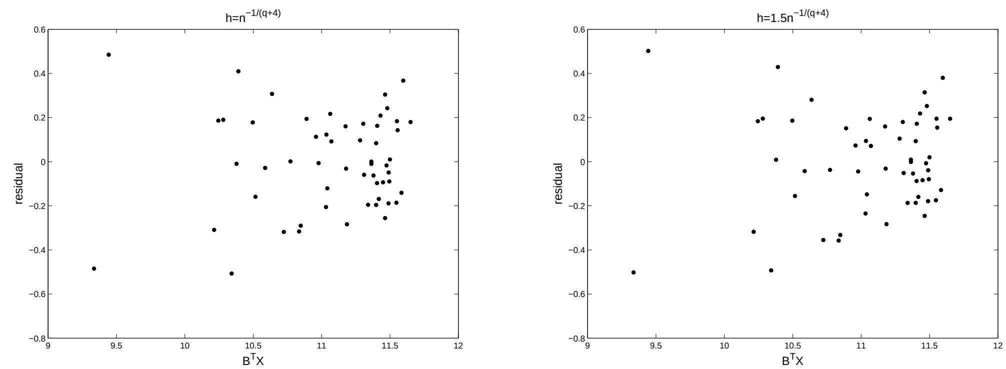

Figure 3.4: The residual plots from the regression model (3.3) against the single-indexing direction obtained from DEE in the real data analysis.

$h = n^{-1/(4+\hat{q})}$ and $h = 1.5n^{-1/(4+\hat{q})}$, where the estimator of B is constructed by DEE and q by RERE. From the plots we see that a heteroscedasticity structure may exist. The values of the test statistic are computed as $T_1 = 3.4230$ and $T_2 = 3.9028$ with $h = n^{-1/(4+\hat{q})}$ and $h = 1.5n^{-1/(4+\hat{q})}$, respectively and the corresponding $p-$values are 0.0003 and 0.0000. Thus, a heteroscedasticity model may be tenable for this data set. The result of this analysis implies that the volatility of the performance in running short distances depends on the performance in running long distances.

3.5 Discussions

Heteroscedasticity checking is an important step in regression analysis. In this chapter, we develop a dimension reduction model-adaptive test. The critical ingredient in the test statistic construction is that the test embeds the dimension reduction structure under the null hypothesis to overcome the curse of dimensionality and adopts to model structure under the alternative hypothesis such that it is still an omnibus test. The test statistic has the limit at the rate as if the number of covariates was the number of linear combinations in the mean regression function. Note that under the null hypothesis, the number of covariates is 0. Thus could we further improve our test to have a faster rate? Looking at the construction procedure, it seems not possible if we do not have any other extra assumptions on the conditional variance structure. However, if we have prior information that under the alternative hypothesis, an improvement seems possible. For instance, when we know that q is greater than q_1, the consistency of the estimator $\hat{q}$ gives us the chance to have idea whether the underlying model is hypothetical or alternative model. Of course, we cannot simply use this information to be a test as type I and II errors cannot be determined. We then use an estimate $\hat{B}_{\tilde{q}}$ where $\tilde{q} = I(\hat{q} = q_1) + \hat{q}I(\hat{q} > q_1)$ and the test statistic is based on $\hat{B}_{\tilde{q}}$. This means that under the null hypothesis, with a probability going to 1, the test statistic is only with one linear combination of the covariates, rather than q_1 linear combinations. The standardizing constant is $nh^{1/2}$ rather than $nh^{q_1/2}$. It is expectable to have the asymptotic normality under the null hypothesis. It can also detect the local alternatives distinct from the null at the rate of $1/\sqrt{nh^{1/2}}$. The study is ongoing.

Further, this method can also be extended to handle other conditional variance models such as single-index and multi-index models. The relevant research is ongoing.

3.6 Appendix 2

3.6.1 Regularity Conditions

To investigate the asymptotic properties in Sections 3.2 and 3.3, the following regularity conditions are designed.

A1 $M_n(t)$ has the following expansion:

$$M_n(t) = M(t) + E_n\{\psi(X, Y, t)\} + R_n(t),$$

where $E_n(\cdot)$ denotes sample averages, $E(\psi(X, Y, t)) = 0$ and $\psi(X, Y, t)$ has a finite second-order moment.

A2 $\sup_{t \in R^{p_2}} ||R_n(t)||_F = o_p(n^{-1/2})$, where $|| \cdot ||_F$ denotes the Frobenius norm of a matrix.

A3 $(B^\top x_i, y_i)_{i=1}^n$ follows a probability distribution $F(B^\top x, y)$ on $\mathbb{R}^q \times \mathbb{R}$. $E(\varepsilon^8 | B^\top X = B^\top x)$ is continuously differentiable and $E(\varepsilon^8 | B^\top X = B^\top x) \le b(B^\top x)$ almost surely, where $b(B^\top x)$ is a measurable function satisfying $E(b^2(B^\top X)) < \infty$.

A4 The density function $p(\cdot)$ of $B^\top X$ exists with support $\mathbb{C}$ and has a continuous and bounded second-order derivative on the support $\mathbb{C}$. The density $p(\cdot)$ satisfies

$$0 < \inf_{B^\top X \in \mathbb{C}} p(B^\top X) \le \sup_{B^\top X \in \mathbb{C}} p(B^\top X) < \infty.$$

A5 For some positive integer r, the rth derivative of $g(\cdot)$ is bounded.

A6 $Q(\cdot)$ is a bounded, symmetric and twice continuously differentiable kernel function such that $\int Q(u)du = 1$, $\int u^i Q(u)du = 0$ and $\int u^r Q(u)du \neq 0$ for $0 < i < r$, where i is a nonnegative integer and r is given by Condition A5.

A7 $K(\cdot)$ is a bounded, symmetric and twice continuously differentiable kernel function satisfying $\int K(u)du = 1$.

A8 $n \to \infty$, $h_1 \to 0$, $h \to 0$,

1) under the null or local alternative hypotheses, $nh^{q_1} \to \infty$, $h_1 = O(n^{-\frac{1}{4+q_1}})$, $nh_1 \to \infty$ and $nh^{q_1/2} h_1^{4r} \to 0$;

2) under global alternative hypothesis H_1, $nh^q \to \infty$, $h_1 = O(n^{-1/(4+q)})$ and $nh^{q/2} h_1^{4r} \to 0$,

where η is given by Condition A6.

Remark 3.4. *It is needed for DEE to assume Conditions A1 and A2. Under the linearity condition and constant conditional variance condition, DEE_{SIR} satisfies Conditions A1 and A2. See Zhu et al. (2010a). Conditions A3, A4, A5 and A6 are widely used for nonparametric estimation in the literature and are also needed for obtaining uniform convergence of $\hat{p}(\cdot)$ and $\hat{g}(\cdot)$. Conditions A4 and A7 guarantee* *the asymptotic normality of our test statistic. Applying a higher order kernel in A6 guarantees that the estimator $\hat{g}$ and $\hat{\sigma}^2$ have sufficiently small biases, respectively, see Powell et al. (1989) and Hall and Marron (1990). To be specific, $\hat{\sigma}^2$ has a convergence rate as $O_p(h_1^r)$. Note that the density estimator $\hat{p}(\cdot)$ appears in the denominator of $\hat{g}(\cdot)$ and small values of $\hat{p}(\cdot)$ may cause the estimator $\hat{g}(\cdot)$ and then the test statistic to be ill-behaved. Thus, Condition A4 can evade this problem. Thus, Conditions A4, A5 and A6 are needed for the test to be well-behaved. Condition A8 is similar to that in Fan and Li (1996), which was originally for model checking about the mean regression. We note that Zheng (2009) used a single bandwidth. Actually, in our case, we could also use a single bandwidth. However, we found that when we respectively use different bandwidths h and h_1 for estimating the mean function $\hat{g}$ and constructing the test statistic T_n, the final test statistic is less sensitive to the bandwidth selection. This phenomenon has been discussed in the literature such as Stute and Zhu (2005) pointing out that the optimal bandwidth for estimation is different from that for test statistic construction.*

3.6.2 Proofs of the theorems

Proof of Theorem 3.1. Under the assumptions designed in Zhu et al. (2010a), their Theorem 2 shows that $M_n - M = O_p(n^{-1/2})$. Following the similar arguments used in Zhu and Ng (1995) or Zhu and Fang (1996), we can gain the root-n consistency of the eigenvalues of M_n, namely, $\hat{\lambda}_i - \lambda_i = O_p(n^{-1/2})$.

Prove (i). It is obvious that under H_0, for any l with $1 < l \leq q_1$, $\lambda_l > 0$. Therefore, we have $\hat{\lambda}_l^2 = \lambda_l^2 + O_p(1/\sqrt{n})$. Since for any l with $q_1 < l \leq p$, $\lambda_l = 0$, we have $\hat{\lambda}_l^2 = O_p(1/n) = O_p(1/n)$. For any $l < q_1$, we have $\lambda_l > 0$, $\lambda_{l+1} > 0$ and

$$\frac{\hat{\lambda}^2_{(q_1+1)} + c_n}{\hat{\lambda}^2_{q_1} + c_n} - \frac{\hat{\lambda}^2_{(l+1)} + c_n}{\hat{\lambda}^2_{l} + c_n} = \frac{\lambda^2_{(q_1+1)} + c_n + O_p(1/n)}{\lambda^2_{q_1} + c_n + O_p(1/\sqrt{n})} - \frac{\lambda^2_{(l+1)} + c_n + O_p(1/\sqrt{n})}{\lambda^2_{l} + c_n + O_p(1/\sqrt{n})}$$

$$= \frac{c_n + O_p(1/n)}{\lambda^2_{q_1} + c_n + O_p(1/\sqrt{n})} - \frac{\lambda^2_{(l+1)} + c_n + O_p(1/\sqrt{n})}{\lambda^2_{l} + c_n + O_p(1/\sqrt{n})}.$$

Taking $\frac{\log n}{n} \le c_n \to 0$, we can obtain

$$\frac{\hat\lambda^2_{(q_1+1)} + c_n}{\hat\lambda^2_{q_1} + c_n} - \frac{\hat\lambda^2_{(l+1)} + c_n}{\hat\lambda^2_{l} + c_n} \to \frac{0}{\lambda^2_{q_1}} - \frac{\lambda^2_{(l+1)}}{\lambda^2_l} = -\frac{\lambda^2_{(l+1)}}{\lambda^2_l} < 0.$$

Further, since for any $l > q_1$, we have $\lambda_l = 0$ and $\lambda^2_{q_1} > 0$, we have

$$\begin{aligned}
\frac{\hat\lambda^2_{(q_1+1)} + c_n}{\hat\lambda^2_{q_1} + c_n} - \frac{\hat\lambda^2_{(l+1)} + c_n}{\hat\lambda^2_{l} + c_n} &= \frac{\lambda^2_{q_1+1} + c_n + O_p(1/n)}{\lambda^2_{q_1} + c_n + O_p(1/\sqrt{n})} - \frac{\lambda^2_{l+1} + c_n + O_p(1/n)}{\lambda^2_{l} + c_n + O_p(1/n)} \\
&= \frac{c_n + o_p(c_n)}{\lambda^2_{q_1} + c_n + O_p(1/\sqrt{n})} - \frac{c_n + o_p(c_n)}{c_n + o_p(c_n)} \\
&\to -1 < 0.
\end{aligned}$$

Therefore, altogether, we can conclude that $\hat q \to q_1$.

For part (ii), by replacing q_1 by q, and using the same arguments as the above we can obtain $\hat q \to q$ in probability. □

Proof of Theorem 3.2. For notational convenience, denote $z_i = B^\top x_i$ $g_i = g(B^\top x_i)$, $\hat g_i = \hat g(\hat B_{\hat q}^\top x_i)$, $\mu_i = (y_i - g_i)^2 - \sigma^2$, $\epsilon_i = y_i - g_i$, $K_{Bij} = K(B^\top(x_i - x_j)/h)$. Under the null hypothesis, without loss of generality, write $B_1 = B$.

Since $\hat\mu_i \equiv: (y_i - \hat g_i)^2 - \hat\sigma^2 = \mu_i - 2\epsilon_i(\hat g_i - g_i) + (\hat g_i - g_i)^2 - (\hat\sigma^2 - \sigma^2)$, we decompose the term S_n to be:

$$\begin{aligned}
S_n &= \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{i\neq j}\frac{1}{h^{q_1}}K_{\hat B ij}\mu_i\mu_j + 4\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{i\neq j}\frac{1}{h^{q_1}}K_{\hat B ij}\epsilon_i\epsilon_j(\hat g_i - g_i)(\hat g_j - g_j) \\
&+\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{i\neq j}\frac{1}{h^{q_1}}K_{\hat B_{\hat q} ij}(\hat g_i - g_i)^2(\hat g_j - g_j)^2 + \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{i\neq j}\frac{1}{h^{q_1}}K_{\hat B_{\hat q} ij}(\hat\sigma^2 - \sigma^2) \\
&-4\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{i\neq j}\frac{1}{h^{q_1}}K_{\hat B_{\hat q} ij}\mu_i\epsilon_j(\hat g_j - g_j) + 2\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{i\neq j}\frac{1}{h^{q_1}}K_{\hat B_{\hat q} ij}\mu_i(\hat g_j - g_j)^2 \\
&-2\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{i\neq j}\frac{1}{h^{q_1}}K_{\hat B_{\hat q} ij}\mu_i(\hat\sigma^2_j - \sigma^2_j) - 4\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{i\neq j}\frac{1}{h^{q_1}}K_{\hat B_{\hat q} ij}\epsilon_i(\hat g_i - g_i)(\hat g_j - g_j)^2 \\
&4\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{i\neq j}\frac{1}{h^{q_1}}K_{\hat B_{\hat q} ij}\epsilon_i(\hat g_i - g_i)(\hat\sigma^2_j - \sigma^2_j) \\
&-2\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{i\neq j}\frac{1}{h^{q_1}}K_{\hat B_{\hat q} ij}\epsilon_i(\hat g_i - g_i)^2(\hat\sigma^2 - \sigma^2) + o_p(n^{-1}h^{-q_1/2}) \\
&\equiv: \sum_{i=1}^{10} Q_{in} + o_p(n^{-1}h^{-q_1/2}).
\end{aligned}$$

The final equation is derived by applying Lemma 2 of Guo et al. (2014), where $\hat{q} = q_1$. We now deal with the terms. First, consider the term Q_{1n}. By Taylor expansion for Q_{1n} with respect to B, we have

$$Q_{1n} \equiv: Q_{11n} + Q_{12n} + Q_{13n},$$

where Q_{11n}, Q_{12n} and Q_{13n} have following forms:

$$\begin{aligned} Q_{11n} &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j} \frac{1}{h^{q_1}} K_{Bij} \mu_i \mu_j, \\ Q_{12n} &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j} \frac{1}{h^{2q_1}} K'_{Bij} \mu_i \mu_j (\hat{B}_{\hat{q}} - B)^\top (x_i - x_j) \end{aligned}$$

and

$$Q_{13n} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j} \frac{1}{h^{3q_1}} K''_{\tilde{B}ij} \mu_i \mu_j (\hat{B}_{\hat{q}} - B)^\top (x_i - x_j)(x_i - x_j)^\top (\hat{B}_{\hat{q}} - B).$$

Here $\tilde{B} = \{\tilde{B}_{ij}\}_{p \times q_1}$ with $\tilde{B}_{ij} \in [\min\{\hat{B}_{ij}, B_{ij}\}, \max\{\hat{B}_{ij}, B_{ij}\}]$. Due to the two facts that $||\hat{B}_{\hat{q}} - B|| = O_p(1/\sqrt{n})$ and the second-order differential function of $K_B(\cdot)$ is a bounded continuous function of B, we assert that replacing $\tilde{B}$ by $\hat{B}_{\hat{q}}$ does not affect the convergence rate of Q_{13n}.

By Theorem 1 in Zheng (2009), we obtain that:

$$nh^{q_1/2} Q_{11n} \to N(0, s^2).$$

Since $E(\mu_i) = 0$, we have $E(Q_{21n}) = 0$. Then we compute the second order moment of Q_{12n} as follows:

$$\begin{aligned} E(Q_{12n}^2) &= E\Big[\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j} \frac{1}{h^{2q_1}} K'_{Bij} \mu_i \mu_j (\hat{B}_{\hat{q}} - B)^\top (x_i - x_j)\Big]^2 \\ &= E\Big[\frac{1}{n^2(n-1)^2} \frac{1}{h^{4q_1}} \sum_{i=1}^{n} \sum_{i' \neq j'} \sum_{i=1}^{n} \sum_{i' \neq j'} K'_{Bij} K'_{Bi'j'} \\ &\quad \mu_i \mu_j \mu_{i'} \mu_{j'} (\hat{B}_{\hat{q}} - B)^\top (x_i - x_j)(x_{i'} - x_{j'})^\top (\hat{B}_{\hat{q}} - B)\Big] \end{aligned}$$

Noting that $E(\mu_i\mu_j\mu_{i'}\mu_{j'}) \neq 0$ only if $i = i'$, $j = j'$ or $i = j'$, $j = i'$, we have

$$\begin{aligned}
E(Q_{12n}^2) &= \frac{n(n-1)}{n^2(n-1)^2}\frac{1}{h^{4q_1}}E\big[(K'_{Bij})^2\mu_i^2\mu_j^2(\hat{B}_{\hat{q}}-B)^\top(X_i-X_j)(x_i-x_j)^\top(\hat{B}_{\hat{q}}-B)\big] \\
&= \frac{1}{n(n-1)}\frac{1}{h^{4q_1}}E\big[(K'_{Bij})^2\mu_i^2\mu_j^2(\hat{B}_{\hat{q}}-B)^\top(x_i-x_j)(x_i-x_j)^\top(\hat{B}_{\hat{q}}-B)\big] \\
&= \frac{1}{n(n-1)}\frac{1}{h^{4q_1}}\{E(\mu_i^2)\}^2(\hat{B}_{\hat{q}}-B)^\top E\big\{(K'_{Bij})^2(x_i-x_j)(x_i-x_j)^\top\big\}(\hat{B}_{\hat{q}}-B).
\end{aligned}$$

By a variable transformation as $u_1 = (x_i - x_j)/h$, the above value is as

$$\begin{aligned}
E(Q_{12n}^2) &= \frac{1}{n(n-1)}\frac{1}{h^{4q_1}}\{E(\mu_i^2)\}^2\int\int(K'_{Bij})^2(\hat{B}_{\hat{q}}-B)^\top(x_i-x_j) \\
&\quad (x_i-x_j)^\top(\hat{B}_{\hat{q}}-B)p(B^\top x_i)p(B^\top x_j)dx_idx_j \\
&= \frac{1}{n(n-1)}\frac{1}{h^{q_1}}\{E(\mu_i^2)\}^2\int\int(K'(u))^2(\hat{B}_{\hat{q}}-B)^\top uu^\top(\hat{B}_{\hat{q}}-B) \\
&\quad p(B^\top x_i)p(B^\top(x_i-hu))dx_idu.
\end{aligned}$$

By Taylor expansion of $p(B^\top(x_i - hu))$ about x_i and under Conditions A3-A7 in Appendix 2, we have

$$\begin{aligned}
E(Q_{12n}^2) &= \frac{1}{n(n-1)}\frac{1}{h^{q_1}}\{E(\mu_i^2)\}^2\int\int(K'(u))^2(\hat{B}_{\hat{q}}-B)^\top uu^\top(\hat{B}_{\hat{q}}-B) \\
&\quad p(B^\top x_i)p(B^\top(x_i-hu))dx_idu \\
&= \frac{1}{n(n-1)}\frac{1}{h^{q_1}}\{E(\mu_i^2)\}^2\int\int(K'(u))^2(\hat{B}_{\hat{q}}-B)^\top uu^\top(\hat{B}_{\hat{q}}-B) \\
&\quad (p^2(B^\top x_i)+p(B^\top x_i)p'(B^\top x_i)h^pu)dx_idu + o_p(\frac{1}{n(n-1)}) \\
&= \frac{1}{n(n-1)}\frac{1}{h^{q_1}}\{E(\mu_i^2)\}^2\int\int(K'(u))^2(\hat{B}_{\hat{q}}-B)^\top uu^\top(\hat{B}_{\hat{q}}-B)p^2(B^\top x_i)dx_idu \\
&\quad +\frac{1}{n(n-1)}\frac{1}{h^{q_1}}\{E(\mu_i^2)\}^2\int\int(K'(u))^2(\hat{B}_{\hat{q}}-B)^\top uu^\top(\hat{B}_{\hat{q}}-B) \\
&\quad p(B^\top x_i)p'(B^\top x_i)h^{q_1}u)dx_idu + o_p(\frac{1}{n(n-1)})O(\frac{1}{n}) \\
&= O_p(\frac{1}{n^2(n-1)h^{q_1}}).
\end{aligned}$$

The application of Chebyshiev's inequality yields that $|Q_{12n}| = o_p(n^{-1}h^{-q_1/2})$. Similarly, we can prove the term Q_{13n} to have the rate: $Q_{13n} = o_p(n^{-1}h^{-q_1/2})$. Therefore, the above decomposition term Q_{1n} convergences to a normal distribution:

$$nh^{q_1/2}Q_{1n} \xrightarrow{\text{d}} N(0, s_1^2).$$

To obtain the results of the theorem, it remains to prove that $nh^{q_1/2}Q_{in} = o_p(1)$, $i = 2, 3, \cdots, 10$.

Second, we consider the term Q_{2n}. Since

$$\begin{aligned}
\frac{Q_{2n}}{4} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{h^{q_1}}K_{\hat{B}_{\hat{q}}ij}\varepsilon_i\varepsilon_j(\hat{g}_i - g_i)(\hat{g}_j - g_j)\frac{\hat{p}_i}{p_i}\frac{\hat{p}_j}{p_j} \\
&\quad + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{h^{q_1}}K_{\hat{B}_{\hat{q}}ij}\varepsilon_i\varepsilon_j(\hat{g}_i - g_i)(\hat{g}_j - g_j)\left(\frac{\hat{p}_i - p_i}{p_i}\frac{\hat{p}_j - p_j}{p_j} - 2\frac{(\hat{p}_i - p_i)\hat{p}_j}{p_ip_j}\right) \\
&\equiv \tilde{Q}_{2n} + o_p(\tilde{Q}_{2n}).
\end{aligned}$$

Substituting the kernel estimates $\hat{g}$ and $\hat{p}$ into $\tilde{Q}_{2n}$, we have

$$\begin{aligned}
\tilde{Q}_{2n} &= \frac{1}{n^3(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{1}{h^{q_1}h_1^{2q_1}}\frac{1}{p_ip_j}K_{\hat{B}_{\hat{q}}ij}Q_{\hat{B}_{\hat{q}}il}Q_{\hat{B}_{\hat{q}}jk}\epsilon_i\epsilon_j \\
&\quad \times(y_l - g(B^\top x_i))(y_k - g(B^\top x_j)).
\end{aligned}$$

By the two order Taylor expansion for $\tilde{Q}_{2n}$ with respect to B, we can have

$$\tilde{Q}_{2n} \equiv Q_{21n} + Q_{22n} + Q_{23n},$$

where Q_{21n}, Q_{22n} and Q_{23n} have following forms:

$$\begin{aligned}
Q_{21n} &= \frac{1}{n^3(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{1}{h^{q_1}h_1^{2q_1}}\frac{1}{p_ip_j}K_{Bij}Q_{Bil}Q_{Bjk}\varepsilon_i\varepsilon_j \\
&\quad (y_l - g(B^\top x_i))(y_k - g(B^\top x_j));
\end{aligned}$$

$$Q_{22n} \equiv: (\hat{B}_{\hat{q}} - B)^\top Q_{221n} + (\hat{B}_{\hat{q}} - B)^\top Q_{222n} + (\hat{B}_{\hat{q}} - B)^\top Q_{223n};$$

and

$$Q_{23n} \equiv: (\hat{B}_{\hat{q}} - B)^\top(Q_{231n} + Q_{232n} + Q_{233n} + Q_{234n} + Q_{235n} + Q_{236n})(\hat{B}_{\hat{q}} - B);$$

with $\{Q_{22in}\}_{i=1}^3$ and $\{Q_{23jn}\}_{j=1}^6$ being following forms:

$$\begin{aligned}
Q_{221n} &= \frac{1}{n^3(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{1}{h^{q_1}h_1^{3q_1}}\frac{1}{p_ip_j}K_{Bij}Q'_{Bil}Q_{Bjk}\varepsilon_i\varepsilon_j \\
&\quad (y_l - g(B^\top x_i))(y_k - g(B^\top x_j))(x_i - x_l),
\end{aligned}$$

$$Q_{222n} = \frac{1}{n^3(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{1}{h^{q_1} h_1^{3q_1}} \frac{1}{p_i p_j} K_{Bij} Q_{Bil} Q'_{Bjk} \varepsilon_i \varepsilon_j$$
$$(y_l - g(B^\top x_i))(y_k - g(B^\top x_j))(x_j - x_k),$$

$$Q_{223n} = \frac{1}{n^3(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{1}{h^{q_1} h_1^{3q_1}} \frac{1}{p_i p_j} K'_{Bij} Q_{Bil} Q_{Bjk} \varepsilon_i \varepsilon_j$$
$$(y_l - g(B^\top x_i))(y_k - g(B^\top x_j))(x_i - x_j)$$

and

$$Q_{231n} = \frac{1}{n^3(n-1)} \frac{2}{h^{q_1} h_1^{4q_1}} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{1}{p_{1i} p_{1j}} K_{\tilde{B}ij} Q'_{\tilde{B}il} Q'_{\tilde{B}jk} \varepsilon_i \varepsilon_j$$
$$(y_l - B^\top x_i))(y_k - g(B^\top x_j))(x_i - X_l)(x_j - x_k)^\top,$$

$$Q_{232n} = \frac{1}{n^3(n-1)} \frac{2}{h^{q_1} h_1^{4q_1}} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{1}{p_i p_j} K'_{\tilde{B}ij} Q'_{\tilde{B}il} Q_{\tilde{B}jk} \varepsilon_i \varepsilon_j$$
$$(y_l - B^\top x_i))(y_k - g(B^\top x_j))(x_i - x_j)(x_i - x_l)^\top,$$

$$Q_{233n} = \frac{1}{n^3(n-1)} \frac{2}{h^{q_1} h_1^{4q_1}} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{1}{p_i p_j} K'_{\tilde{B}ij} Q_{\tilde{B}il} Q'_{\tilde{B}jk} \varepsilon_i \varepsilon_j$$
$$(y_l - B^\top x_i))(y_k - g(B^\top x_j))(x_i - x_j)(x_j - x_k)^\top,$$

$$Q_{234n} = \frac{1}{n^3(n-1)} \frac{1}{h^{q_1} h_1^{4q_1}} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{1}{p_{1i} p_{1j}} K''_{\tilde{B}ij} Q_{\tilde{B}il} Q_{\tilde{B}jk} \varepsilon_i \varepsilon_j$$
$$(y_l - B^\top x_i))(y_k - g(B^\top x_j))(x_i - x_j)(x_i - x_j)^\top,$$

$$Q_{235n} = \frac{1}{n^3(n-1)} \frac{1}{h^{q_1} h_1^{4q_1}} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{1}{p_{1i} p_{1j}} K_{\tilde{B}ij} Q''_{\tilde{B}il} Q_{\tilde{B}jk} \varepsilon_i \varepsilon_j$$
$$(y_l - B^\top x_i))(y_k - g(B^\top x_j))(x_i - x_l)(x_i - x_l)^\top,$$

$$Q_{236n} = \frac{1}{n^3(n-1)} \frac{1}{h^{q_1} h_1^{4q_1}} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{n} \sum_{l=1}^{n} \frac{1}{p_{1i} p_{1j}} K_{\tilde{B}ij} Q_{\tilde{B}il} Q''_{\tilde{B}jk} \varepsilon_i \varepsilon_j$$
$$(y_l - B^\top x_i))(y_k - g(B^\top x_j))(x_j - x_k)(x_j - x_k)^\top,$$

and where $\tilde{B} = \{\tilde{B}_{ij}\}_{p \times q_1}$ with $\tilde{B}_{ij} \in [\min\{\hat{B}_{ij}, B_{ij}\}, \max\{\hat{B}_{ij}, B_{ij}\}]$. As described for the term Q_{23n}, we also assert that that replacing $\tilde{B}$ by $\hat{B}_{\hat{q}}$ does not affect the convergence rate.

For the term $\tilde{Q}_{2n}$, we first consider Q_{21n}. Since for any fixed $\hat{B}_{\hat{q}}$, $E(Q_{21n})=0$, we compute its second order moment as follows:

$$\begin{aligned}
E(Q_{21n}^2) &= E\Big[\frac{1}{n^3(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{1}{h^{q_1}h_1^{2q_1}}\frac{1}{p_ip_j}K_{Bij}Q_{Bil}Q_{Bjk}\varepsilon_i\varepsilon_j \\
&\quad (y_l-g(B^\top x_i))(y_k-g(B^\top x_j))\Big]^2 \\
&= E\Big[\frac{1}{n^6(n-1)^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\sum_{l=1}^{n}\sum_{i'=1}^{n}\sum_{j'\neq i'}\sum_{k'=1}^{n}\sum_{l'=1}^{n}\frac{1}{h^{2q_1}h_1^{4q_1}}\frac{1}{p_ip_{i'}p_jp_{j'}} \\
&\quad K_{Bij}Q_{Bil}Q_{Bjk}K_{Bi'j'}Q_{Bi'l'}Q_{Bj'k'}\varepsilon_i\varepsilon_j\varepsilon_i'\varepsilon_j'(Y_l-g(B^\top X_i)) \\
&\quad (y_{l'}-g(B^\top x_{i'}))(y_k-g(B^\top x_j))(y_{k'}-g(B^\top x_{j'}))\Big].
\end{aligned}$$

Noting that $E(\varepsilon_i\varepsilon_j\varepsilon_{i'}\varepsilon_{j'})\neq 0$ only if $i=i'$, $j=j'$ or $i=j'$, $j=i'$, we have

$$\begin{aligned}
& E(Q_{21n}^2) \\
= \; & \frac{1}{n^6(n-1)^2h^{2q_1}h_1^{4q_1}}n(n-1)(n-2)^2(n-3)^2E(\frac{1}{p_i^2}\frac{1}{p_j^2}K_{Bij}^2Q_{Bil}Q_{Bjk}Q_{Bil'}Q_{Bjk'} \\
& (g_l-g_i)(g_k-g_j)(g_{l'}-g_i)(g_{k'}-g_j)\delta^4+o((n^2h)^{-q}).
\end{aligned}$$

By transforming variables as $u_1=(z_i-z_j)/h$, $u_2=(z_i-z_l)/h_1$, $u_3=(z_j-z_k)/h_1$, $u_4=(z_i-z_{l'})/h_1$ and $u_5=(z_j-z_{k'})/h_1$, we can have

$$\begin{aligned}
E(Q_{21n}^2) &= \frac{h^{q_1}h_1^{4q_1}}{n^6(n-1)^2h^{2q_1}h_1^{4q_1}}n(n-1)(n-2)^2(n-3)^2 \\
&\quad \int\int\int\int\int\int\frac{1}{p^2(Z_i)p^2(Z_i-hu_1)}K_{Bij}^2(u_1)Q_{Bil}(u_2)Q_{Bjk}(u_3)Q_{Bil'}(u_4) \\
&\quad Q_{Bjk'}(u_5)[g(z_i-h_1u_2)-g(z_i)][g(z_i-hu_1-h_1u_3)-g(z_i-hu_1] \\
&\quad [g(z_i-h_1u_4)-g(Z_i)][g(z_i-hu_1-h_1u_5)-g(z_i-hu_1]p(z_i)p(z_i-hu_1) \\
&\quad p_1(z_i-h_1u_2)p_1(z_i-hu_1-h_1u_3)p_1(z_i-h_1u_4)p_1(z_i-hu_1-h_1u_5) \\
&\quad dz_idu_1du_2du_3du_4du_5+o((n^2h^q)^{-1}).
\end{aligned}$$

By taking Taylor expansions of $g(z_i-hu_1)-g(z_i)$ and similar terms at z_i and using Conditions A4, A5 and A6, we have

$$E(Q_{21n}^2)=O_p(\frac{h^{q_1}h_1^{4q_1}h_1^{4r}}{n^2h^{2q_1}h_1^{4q_1}})=O_p(\frac{h_1^{4r}}{n^2h^{q_1}}).$$

the application of Chebyshiev's inequality leads to $|Q_{21n}| = o_p(1/(nh^{q_1/2}))$.

Similarly, the terms Q_{22in} and Q_{23jn} for $\{i = 1, 2, 3, j = 1, \cdots, 6\}$ can be proved to have the following rates: $Q_{22in} = o_p(\frac{h_1^r}{nh^{q_1/2}} \cdot \frac{1}{h_1^{q_1/2}}) = o_p(\frac{h_1^r}{nh^{q_1/2}h_1^{q_1/2}})$ and $Q_{23jn} = o_p(\frac{h_1^{r/2}}{nh^{q_1/2}} \cdot \frac{1}{h_1^{q_1}}) = o_p(\frac{h_1^{r/2}}{nh^{q_1/2}h_1^{q_1}})$. As $h_1 = O(n^{-1/(4+q_1)})$, we can get $|Q_{22n}| = o_p(\frac{h_1^r}{nh^{q_1/2}h_1^{q_1/2}} \cdot \frac{1}{\sqrt{n}}) = o_p(1/(nh^{q_1/2}))$ and $|Q_{23n}| = o_p(\frac{h_1^{r/2}}{nh^{q_1/2}h_1^{q_1}} \cdot \frac{1}{n}) = o_p(1/(nh^{q_1/2}))$. Thus we arrive at the result that $nh^{q_1/2}Q_{2n} = o_p(1)$.

Now we consider Q_{3n}. Following the similar argument for proving Theorem 3 of Collomb and Härdle (1986), we have

$$\sup_{t \in C^q} |\hat{g}(t) - g(t)| = O_p\Big(\sqrt{\frac{\ln n}{nh_1^{q_1}}}\Big).$$

Further,

$$\begin{aligned} Q_{3n} &\leq \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{h^{q_1}} K_{Bij} (\sup_{t \in C^q} |\hat{g}_i(t) - g_i(t)|^4) \\ &= O_p\Big(\frac{\ln^2 n}{n^2 h_1^{2q_1}}\Big) = o_p(\frac{1}{nh^{q_1/2}}), \end{aligned}$$

we can obtain $nh^{q_1/2}Q_{3n} = o_p(1)$.

Similarly as Hall and Marron (1990), we can easily obtain that under the null hypothesis, $\hat{\sigma}^2 = \sigma^2 + O_p(h_1^{2r})$. Since $nh^{q_1/2}h_1^{4r} \to 0$, we have $Q_{4n} = O_p(h_1^{4r}) = o_p((nh^{q_1/2})^{-1})$. Using the same argument as the above, we can prove $nh^{q_1/2}Q_{5n} = o_p(1), \cdots, nh^{q_1/2}Q_{10n} = o_p(1)$. Hence, we can conclude that

$$nh^{q_1/2} S_{1n} \xrightarrow{\text{d}} N(0, s^2).$$

Second, we also need to prove $\hat{s}^2 \xrightarrow{\text{p}} s^2$. Note that $\hat{\beta}$, $\hat{B}_{\hat{q}}$ and $\hat{g}$ are respectively the uniform consistency estimators of β, B and g. Thus,

$$\hat{s}^2 = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \frac{1}{h^{q_1}} K_{Bij}^2 (\varepsilon_i^2 - \sigma^2)^2 (\varepsilon_i^2 - \sigma^2)^2 + o_p(1) \equiv: s_n + o_p(1),$$

where s_n is an U-statistic with the kernel as:

$$H_n(w_i, w_j) = \frac{1}{h^{q_1}} K_{Bij}^2 (\varepsilon_i^2 - \sigma^2)^2 (\varepsilon_j^2 - \sigma^2)^2,$$

with $w_i = (x_i, \varepsilon_i^2)$ for $i = 1, \cdots, n$. It can be computed to obtain that

$$\begin{aligned} E(H_n(w_i, w_j)) &= \int\int \frac{1}{h^{q_1}} K^2(\frac{z_i - z_j}{h}) Var(\varepsilon_i^2|z_i) Var(\varepsilon_j^2|z_j) p(z_i) p(z_j) dz_i dz_j \\ &= \frac{1}{h^{q_1}} \int\int K^2(u) Var(\varepsilon_i^2|z_i) Var(\varepsilon_j^2|z_i - hu) p(z_i) p(z_i - hu) h^{q_1} dz_i du \\ &= \int K^2(u) du \int [Var(\varepsilon_i^2|z_i)]^2 p^2(z_i) dz_i + o_p(1) = s_1^2 + o_p(1). \end{aligned}$$

Here the variable transformation $u = (z_i - z_j)/h$ is used. Using the similar argument used to prove Lemma 3.1 of Zheng (1996), we have $s_n = E(H_{1n}(w_i, w_j)) + o_p(1) = s^2 + o_p(1)$. Thus,

$$\hat{s}_1^2 \xrightarrow{\text{p}} s^2.$$

Finally, Slutsky lemma is applied to gain

$$T_n \xrightarrow{\text{d}} N(0, 1).$$

The proof of Theorem 3.2 is concluded. □

Proof of Lemma 3.1. Consider MRRE criterion when SIR-based DEE is used. To derive $M_n - M = O_p(C_n)$, we only need to prove that $M_n(t) - M(t) = O_p(C_n)$ uniformly, where $M(t) = \Sigma^{-1} Var(E(X|I(Y \leq t)) = \Sigma^{-1}(\nu_1 - \nu_0)(\nu_1 - \nu_0)^\top p_t(1 - p_t)$, Σ is the covariance matrix of X, $\nu_0 = E(X|I(Y \leq t) = 0)$, $\nu_1 = E(X|I(Y \leq t) = 1)$ and $p_t = E(I(Y \leq t))$. It is easy to see that

$$\begin{aligned} \nu_1 - \nu_0 &= \frac{E(XI(Y \leq t))}{p_t} - \frac{E(XI(Y > t))}{1 - p_t} \\ &= \frac{E(XI(Y \leq t) - E(X))E(I(Y \leq t))}{p_t(1 - p_t)}. \end{aligned}$$

Thus, $M(t)$ can also be rewritten as

$$\begin{aligned} M(t) &= \Sigma^{-1}[E\{(X - E(X))I(Y \leq t)\}][E\{(X - E(X))I(Y \leq t)\}]^\top \\ &=: \Sigma^{-1} m(t) m(t)^\top, \end{aligned}$$

where $m(t) = E\{(X - E(X))I(Y \leq t)\}$. Therefore, $m(t)$ can be estimated by:

$$m_n(t) = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x}) I(y_i \leq t),$$

and then $M(t)$ can be estimated by

$$M_n(t) = \hat{\Sigma}^{-1} L_n(t),$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, $L_n(t) = m_n(t)m_n(t)^\top$ and $\hat{\Sigma}$ is the sample version of Σ. Since the response under the local alternative hypotheses is related to n, we write the response under the null and local alternative hypotheses as Y and Y_n respectively. Further, it is noted that:

$$E\{XI(Y_n \le t)\} - E\{XI(Y \le t)\} = E[X\{P(Y_n \le t|X)\}] - E[X\{P(Y \le t|X)\}].$$

Under H_{1n}, because $Var(\varepsilon|X) = \sigma^2 + C_n f(B^\top X)$, we rewrite the local alternative model as $Y = g(B_1^\top X) + \varepsilon(1 + C_n f(B^\top X)/2)$. Thus, we have for all t,

$$\begin{aligned}
& P(Y_n \le t|X) - P(Y \le t|X) \\
= \; & P(\varepsilon \le \frac{t - g(B_1^\top X)}{1 + C_n f(B^\top X)/2}|X) - P(\varepsilon \le t - g(B_1^\top X)|X) \\
= \; & P(\varepsilon \le t - g(B_1^\top X) + C_n(t - g(B_1^\top X))f(B^\top X)/2|X) \\
& -P(\varepsilon \le t - g(B_1^\top X)|X) + o_p(C_n) \\
= \; & F_{Y|X}(t - C_n(t - g(B_1^\top X))f(B^\top X)/2) - F_{Y|X}(t) \\
= \; & -C_n(t - g(B_1^\top X))f(B^\top X)/2 f_{Y|X}(t) + o_p(C_n).
\end{aligned}$$

Therefore, under Condition A2, we can conclude that

$$\begin{aligned}
& n^{-1}\sum_{i=1}^n x_i I(y_{ni} \le t) - E\{XI(Y \le t)\} \\
= \; & n^{-1}\sum_{i=1}^n x_i I(y_{ni} \le t) - E\{XI(Y_n \le t)\} + \{E\{XI(Y_n \le t) - E\{XI(Y \le t)\} \\
= \; & O_p(\max{(C_n, n^{-1/2})}).
\end{aligned}$$

Using the similar arguments used for proving Theorem 3.2 of Li et al. (2008), we can derive that $M_n(t) - M(t) = O_p(\max{(C_n, n^{-1/2})})$ uniformly. Thus, $M_n - M = O_p(\max{(C_n, n^{-1/2})})$.

As Zhu and Fang (1996) and Zhu and Ng (1995) demonstrated, since $M_n - M = O_p(\max{(C_n, n^{-1/2})})$, we conclude $\hat{\lambda}_i - \lambda_i = O_p(\max{(C_n, n^{-1/2})})$ where $\hat{\lambda}_p \le \hat{\lambda}_{(p-1)} \le \cdots \le \hat{\lambda}_1$ are the eigenvalues of the matrix M_n.

Note that under the null hypothesis, we have $\lambda_p = \cdots = \lambda_{p-q_1} = 0$ and $0 < \lambda_{q_1} \leq \cdots \leq \lambda_1$ to be the eigenvalues of the matrix M. Since $\frac{\log n}{nh^{q_1/2}} = c_n \to 0$ and under the local alternative hypotheses H_{1n} with $C_n = 1/(n^{1/2}h^{q_1/4})$, we have $C_n^2 = o_p(c_n)$. Similarly as the proof for Theorem 3.1. It is clear that for any $l \leq q_1$, we have $\lambda_l > 0$. Then we have $\hat{\lambda}_l^2 = \lambda_l^2 + O_p(C_n)$. On the other hand, for any $q_1 < l \leq p$, as $\lambda_l = 0$, we have $\hat{\lambda}_l^2 = \lambda_l^2 + O_p(C_n^2) = O_p(C_n^2)$. When $l > q_1$, MRRE is computed to be

$$\begin{aligned}
\frac{\hat{\lambda}_{q_1+1}^2 + c_n}{\hat{\lambda}_{q_1}^2 + c_n} - \frac{\hat{\lambda}_{(l+1)}^2 + c_n}{\hat{\lambda}_l^2 + c_n} &= \frac{\lambda_{q_1+1}^2 + c_n + O_p(C_n^2)}{\lambda_{q_1}^2 + c_n + O_p(C_n)} - \frac{\lambda_{l+1}^2 + c_n + O_p(C_n^2)}{\lambda_l^2 + c_n + O_p(C_n^2)} \\
&= \frac{\lambda_{q_1+1}^2 + c_n + o_p(c_n)}{\lambda_{q_1}^2 + c_n + O_p(C_n^2)} - \frac{\lambda_{l+1}^2 + c_n + o_p(c_n)}{\lambda_l^2 + c_n + o_p(c_n)} \\
&= \frac{c_n + o_p(c_n)}{\lambda_{q_1}^2 + c_n + O_p(C_n^2)} - \frac{c_n + o_p(c_n)}{c_n + o_p(c_n)}.
\end{aligned}$$

Thus, we have

$$\frac{\hat{\lambda}_{q_1+1}^2 + c_n}{\hat{\lambda}_{q_1}^2 + c_n} - \frac{\hat{\lambda}_{(l+1)}^2 + c_n}{\hat{\lambda}_l^2 + c_n} \to -1 < 0.$$

When $1 \leq l < q_1$, MRRE is computed to be:

$$\begin{aligned}
\frac{\hat{\lambda}_{q_1+1}^2 + c_n}{\hat{\lambda}_{q_1}^2 + c_n} - \frac{\hat{\lambda}_{(l+1)}^2 + c_n}{\hat{\lambda}_l^2 + c_n} &= \frac{\lambda_{q_1+1}^2 + c_n + O_p(C_n^2)}{\lambda_{q_1}^2 + c_n + O_p(C_n)} - \frac{\lambda_{l+1}^2 + c_n + O_p(C_n)}{\lambda_l^2 + c_n + O_p(C_n)} \\
&= \frac{c_n + o_p(c_n)}{\lambda_{q_1}^2 + c_n + O_p(C_n^2)} - \frac{\lambda_{l+1}^2 + c_n + o_p(c_n)}{\lambda_l^2 + c_n + o_p(c_n)}.
\end{aligned}$$

Then

$$\frac{\hat{\lambda}_{q_1+1}^2 + c_n}{\hat{\lambda}_{q_1}^2 + c_n} - \frac{\hat{\lambda}_{(l+1)}^2 + c_n}{\hat{\lambda}_l^2 + c_n} \to -\frac{\lambda_{l+1}^2}{\lambda_l^2} < 0.$$

Therefor, altogether, we can conclude that $\hat{q} \to q_1$. □

Proof of Theorem 3.3. First, we prove Part (I). Applying the same decomposition technique as that in Theorem 3.2, S_n can be decomposed in the following form:

$$S_n \equiv: \sum_{i=1}^{10} Q_{in},$$

where $\{Q_{in}\}_{i=1}^{10}$ is defined in Theorem 3.2. Note that $\hat{B}_{\hat{q}}$, $\hat{g}$ and $\hat{\sigma}^2$ are respectively uniform consistent estimators of B, g and σ^2. Then we have $S_n = Q_{1n} + o_p(1)$. It is clear that Q_{1n} is a $U-$statistic with kernel $H_n = \frac{1}{h^q} K_{Bij} u_i u_j$. Denote $w_i = (x_i, \varepsilon_i^2)$ for $i = 1, \cdots, n$. Under the alternative hypothesis, due to the fact $E(u_i | B^\top x_i) = Var(\varepsilon_i^2 | B^\top x_i) - Var(\varepsilon_i^2)$, we have

$$\begin{aligned} E[H_n(w_i, w_i)] &= \int\int K_h(B^\top x_i - B^\top x_j)[Var(\varepsilon_i | B^\top x_i) - Var(\varepsilon_i)] \\ &\quad [Var(\varepsilon_j | B^\top x_j) - Var(\varepsilon_j)] p(B^\top x_i) p(B^\top x_j) dB^\top x_i dB^\top x_j. \end{aligned}$$

using the transformed variable $u = (z_i - z_j)/h$, we have

$$\begin{aligned} E[H_n(Z_i, Z_j)] &= \frac{1}{h^p} \int\int K(u)[Var(\varepsilon_i | B^\top x_i) - Var(\varepsilon_i)] \\ &\quad [Var(\varepsilon_j | B^\top x_i - hu) - Var(\varepsilon_j)] p(B^\top x_i) p(B^\top x_i - hu) dB^\top x_i du \\ &= E\big([Var(\varepsilon | B^\top x) - Var(\varepsilon)]^2 p(B^\top x)\big). \end{aligned}$$

Lemma 3.1 of Zheng (1996) yields that

$$S_n = Q_{1n} + o_p(1) = E[H_n(w_i, w_j)] + o_p(1) = E\big\{[Var(\varepsilon | B^\top X) - Var(\varepsilon)]^2 p(B^\top X)\big\} + o_p(1).$$

Similarly, we can prove that $\hat{s}^2 \xrightarrow{\text{p}} s^2$, and then

$$T_n/(nh^{\frac{q}{2}}) \xrightarrow{\text{d}} E\big\{[Var(\varepsilon | B^\top X) - Var(\varepsilon)]^2 p(B^\top X)\big\}/s.$$

Prove Part (II). Under the local alternative hypotheses H_{1n}, similar arguments used for proving Theorem 3.2, we can show that $S_n = Q_{1n} + o_p((nh^{q_1})^{-1})$. Let $\varepsilon_{2i}^2 = \varepsilon_i^2 - C_n f(B^\top x_i)$. Under the local alternative, $E(\varepsilon_{2i}^2 | x_i) = \sigma^2$. Q_{1n} is then decomposed as:

$$\begin{aligned} Q_{1n} &= \{\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} \frac{1}{h^{q_1}} K_{\hat{B}_{\hat{q}} ij} u_i u_j\} \\ &= \{\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} \frac{1}{h^{q_1}} K_{\hat{B}_{\hat{q}} ij} (\varepsilon_{2i}^2 - \sigma^2)(\varepsilon_{2j}^2 - \sigma^2)\} \\ &\quad + C_n \{\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} \frac{1}{h^{q_1}} K_{\hat{B}_{\hat{q}} ij} f(B^\top x_i)(\varepsilon_{2j}^2 - \sigma^2)\} \\ &\quad + C_n^2 \{\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i} \frac{1}{h^{q_1}} K_{\hat{B}_{\hat{q}} ij} f(B^\top x_i) f(B^\top x_j)\} \\ &\equiv W_{1n} + C_n W_{2n} + C_n^2 W_{3n}. \end{aligned}$$

W_{1n} has the following decomposition by Taylor expansion:

$$W_{1n} \equiv: W_{11n} + (\hat{B}_{\hat{q}} - B_1)^\top W_{12n} + (\hat{B}_{\hat{q}} - B_1)^\top W_{13n}(\hat{B}_{\hat{q}} - B_1),$$

with $\{W_{1in}\}_{i=1}^3$ being following forms:

$$W_{11n} = \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}\frac{1}{h^{q_1}}K_{B_1ij}(\varepsilon_{2i}^2-\sigma^2)(\varepsilon_{2j}^2-\sigma^2);$$

$$\begin{aligned} W_{12n} &= \{\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}\frac{1}{h^{2q_1}}K'_{B_1ij}(\varepsilon_{2i}^2-\sigma^2)(\varepsilon_{2j}^2-\sigma^2)(x_i-x_j)\}\}; \\ W_{13n} &= \{\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}\frac{1}{h^{3q_1}}K''_{\tilde{B}ij}(\varepsilon_{2i}^2-\sigma^2)(\varepsilon_{2j}^2-\sigma^2)(x_i-x_j)(x_{i'}-x_{j'})^\top\}. \end{aligned}$$

where $\tilde{B} = \{\tilde{B}_{ij}\}_{p\times q_1}$ with $\tilde{B}_{ij} \in [\min\{\hat{B}_{ij}, B_{1ij}\}, \max\{\hat{B}_{ij}, B_{1ij}\}]$. Here for the terms W_{12n} and W_{12n}, using the similar argument of terms Q_{21n} and Q_{23n} in Theorem 3.2, respectively, we prove to have the following rates: $W_{22n} = o_p(\frac{1}{\sqrt{n}h^{q_1/2}})$, $W_{23n} = (\frac{1}{\sqrt{n}h^{q_1/2}})$. On the other hand, in the same way as that for proving Theorem 1 in Zheng (2009), we can easily derive that $nh^{\frac{q_1}{2}}W_{11n} \xrightarrow{\text{d}} N(0, s^2)$. Thus, we have

$$nh^{\frac{q_1}{2}}W_{1n} \xrightarrow{\text{d}} N(0, s^2).$$

According to Lemma 3.1 of Zheng (1996), it is easy to prove that $\sqrt{n}W_{2n} = O_p(1)$. Thus, when $C_n = n^{-\frac{1}{2}}h^{-\frac{q_1}{4}}$, $nh^{\frac{q_1}{2}}W_{2n} = o_p(1)$.

Finally, consider the term W_{3n}. Also by Taylor expansion, we have

$$\begin{aligned} W_{3n} &= \{\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}\frac{1}{h^p}K_{B_1ij}f(B^\top x_i)f(B^\top x_j) \\ &\quad +(\hat{B}_{\hat{q}} - B_1)^\top\{\frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j\neq i}\frac{1}{h^{2q}}K'_{\tilde{B}ij}f(B^\top x_i)f(B^\top x_j)(x_i-x_j)\}\} \\ &\equiv: W_{31n} + (\hat{B}_{\hat{q}} - B_1)^\top W_{32n}, \end{aligned}$$

where $\tilde{B} = \{\tilde{B}_{ij}\}_{p\times q_1}$ with $\tilde{B}_{ij} \in [\min\{\hat{B}_{ij}, B_{1ij}\}, \max\{\hat{B}_{ij}, B_{1ij}\}]$. We can also asset that replacing $\tilde{B}$ by B_1 can not impact the converging rate of the term W_{32n}. Note that W_{32n} can be written as an $U-$Statistic with the kernel:

$$H_n(x_i, x_j) = \frac{1}{h^{2q_1}}K'_{B_1ij}f(B^\top x_i)f(B^\top x_j)(x_i-x_j) + \frac{1}{h^{2q_1}}K'_{B_1ji}f(B^\top x_i)f(B^\top x_j)(x_j-x_i)$$

Using $U-$Statistics theory (e. g. Serfling 1980), we have $W_{32n} = O_p(1)$. Additionally, W_{31n} is also an $U-$Statistic with the kernel:

$$H_n(w_i, w_j) = \frac{1}{h^{q_1}} K_{B_1 ij} f(B^\top x_i) f(B^\top x_j),$$

where $w_i = (x_i, \varepsilon_i)$ for $i = 1, \cdots, n$. First, we compute the first moment of $H_n(w_i, w_j)$ as

$$E(H_n(w_i, w_j)) = E\{\frac{1}{h^{q_1}} K_{B_1 ij} E[f(B^\top x_i)|B_1^\top x_i] E[f(B^\top x_j)|B_1^\top x_j]\}.$$

For notational convenience, we assume $M(B_1^\top x_i) = E[f(B^\top x_i)|B_1^\top x_i]$ and $z_i = B_1^\top x_i$. Further, $E[H_n(w_i, w_j)]$ can be computed as

$$\begin{aligned} E(H_n(w_i, w_j)) &= E\{\frac{1}{h^{q_1}} K_{B_1 ij} M(B_1^\top x_i) M(B_1^\top x_j)\} \\ &= \int\int \frac{1}{h^{q_1}} K(\frac{z_i - z_j}{h}) M(z_i) M(z_j) p(z_i) p(z_j) dz_i dz_j \\ &= \int\int K(u) M(z_i) M(z_i - hu) p(z_i) p(z_i - hu) dz_i du \\ &= \int K(u) du \int M(z_i) M(z_i) p(z_i) p(z_i) dz_i + o_p(h) \\ &= E\{[E\{f(B^\top X)|B_1^\top X\}]^2 p(B_1^\top X)\}, \end{aligned}$$

where $p(\cdot)$ denotes the density function of $B_1^\top X$. Similarly, we have the consistency of W_{3n} as it goes to $E\{[E\{f(B^\top X)|B_1^\top X\}]^2 p(B_1^\top X)\}$ in probability. Additionally, similarly as the above proof for Part (I) of Theorem 3.2, it is easy to prove $\hat{s}^2 \xrightarrow{\text{p}} s^2$.

Thus, invoking Slutsky theorem, we can conclude that

$$T_n \xrightarrow{\text{d}} N(E\{[E\{f(B^\top X)|B_1^\top X\}]^2 p(B_1^\top X)\}/s, 1).$$

□

Chapter 4

Dimension Reduction-based Significance Testing in Nonparametric Regression

4.1 Introduction

Consider the nonparametric regression model:

$$Y = m(Z) + \epsilon, \tag{4.1}$$

where Y is a scale dependent variable with the covariates $Z = (X^\top, W^\top)^\top$, $X = (X_1, \cdots, X_{p_1})^\top \in \mathbb{R}^{p_1}$, $W = (W_1, \cdots, W_{p_2})^\top \in \mathbb{R}^{p_2}$ and $p_1 + p_2 = d$, the regression function $g(\cdot) : \mathbb{R}^d \to \mathbb{R}$ is unknown in its form and ϵ is the error term with zero conditional expectation when Z is given: $E(\epsilon|Z) = 0$. As well known, the success of any further statistical analysis hinges on the correction of working model. Note that in regression modeling building, there are often part of the covariates to be redundant. A subset of the covariates W is said to be insignificant for the response variable Y given X if

$$E(Y|X, W) = E(Y|X). \tag{4.2}$$

The equality (4.2) means that W does not provide more information to predict Y. W should be removed from the regression model (4.1), otherwise, such redundant variables cause statistical analysis more complicated and less accurate and efficient. This step, particularly in the first stage of regression analysis is necessary. In the literature, relevant testing problem called significance testing has attracted much attention. There exist several proposals that are based on prevalent local smoothing and global smoothing methodologies in the literature. For the former, Lavergne and Vuong (2000) extended the idea introduced by Fan and Li (1996), proposed a test based on a second conditional moment to check the significance of a subset of covariates. Further, Li (1999) developed a nonparametric significance test that is based on the idea of Fan and Li (1996) for nonparametric and semiparametric time-series models. Racine et al. (2006) suggested a test for the significance of categorical variables in fully nonparametric regression models. Lavergne et al. (2014) devised a new kernel-based test that is based on a suitable equivalent Fourier transformation. It is noted that these local smoothing-based test statistics converge to the respective limiting null distributions at the typical rate $O_p(n^{-1/2}h^{-d/4})$, where d is the number of all the covariates and h is the bandwidth in kernel estimation. Further, these tests can also only detect local alternatives distinct from the null hypothesis at the rate $O_p(n^{-1/2}h^{-d/4})$. When $d = p_1 + p_2$ is large, the convergence rate is very slow because h converges to zero at a certain rate. This implies that these local smoothing methodologies severely suffer from the curse of dimensionality. This problem is caused by using nonparametric estimation for the models under both the null and alternative hypotheses that assumes the significance of all the covariates. However, this is clearly not reasonable. An ideal situation is that a test can benefit from the number p_1 of the significant covariates under the null hypothesis such that we can have a much faster convergence rate and the test can well maintain the significance level and enhance the power performance. For global smoothing tests, examples include Racine (1997) who advised a test based on nonparametric estimation of partial derivative. Delgado and Manteiga (2001) introduced a consistent test based on a stochastic process. Both the tests have a fast rate of order $\sqrt{n}$ that they can detect local alternative hypotheses distinct from the null hypothesis. Racine (1997)'s test involves the non-parametric estimation of partial derivative $\partial E(Y|X,W)/\partial W$ and thus, estimation inefficiency can greatly deteriorate the performance of the test. Delgado and Manteiga (2001)'s test involves the multivariate nonparametric estimation of $E(Y|X)$ and the data sparseness in high-dimensional space still causes negative impact for their power performance of this high-dimensional stochastic process. We can see from the following example that the power drops down very quickly as d increases. Finally, when the dimension of the complete set of covariates d is large, the computational burden is also an issue because Monte Carlo approximation to their sampling null distributions is a computational intensive method.

We now present the simulation results for the following illustrative example. The model is

$$Y = (X_1 + X_2)/\sqrt{2} + 2W_1 + \epsilon,$$

where $X = (X_1, X_2, \cdots, X_{p_1})$ and $W = (W_1, W_2, \cdots, W_{p_2})$ follow multiple standard normal distributions, respectively, the sample size is $n = 200$ and the dimension of covariates X is set to be $p_1 = 2, 3, 4$ and the dimension of covariates W varies from 1 to 8 in this numerical simulations. The hypothetical regression function is $(X_1+X_2)/\sqrt{2}$. In other words, we use this example to show how the powers of existing tests decrease with increasing dimension p_2. Here we choose Fan and Li (1996)'s test and Delgado and Manteiga (2001)'s test as the respective representatives of local and global smoothing tests to demonstrate this.

Figures 4.1 and 4.2 depict the curves of the empirical powers at the significance level $\alpha = 0.05$. More details for this simulation can be found in Section 4.4. Clearly, the empirical powers of these two tests rapidly decrease as p_1 and p_2 increase. This indicates that both the tests seriously suffer from the curse of dimensionality. Section 4.1 will also include relevant discussions on the maintenance of the significance

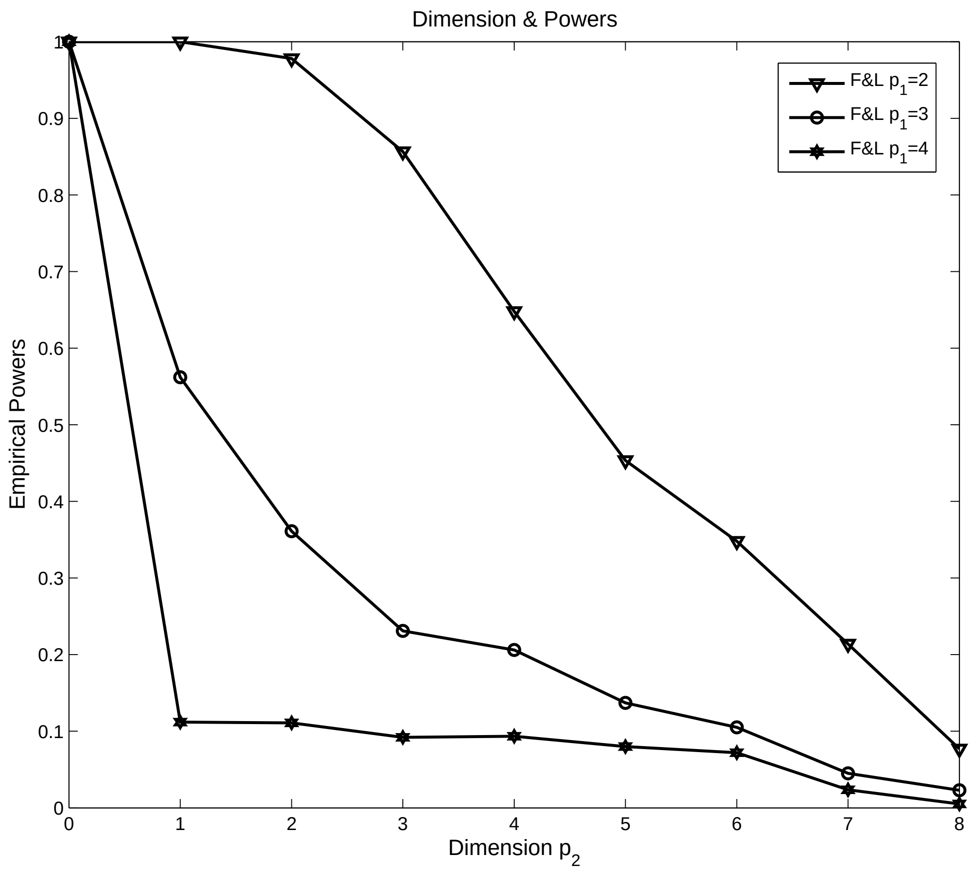

Figure 4.1: The empirical power curve of Fan and Li (1996)'s test against the dimensions of X and W with sample size 200 at the significance level $\alpha = 0.05$.

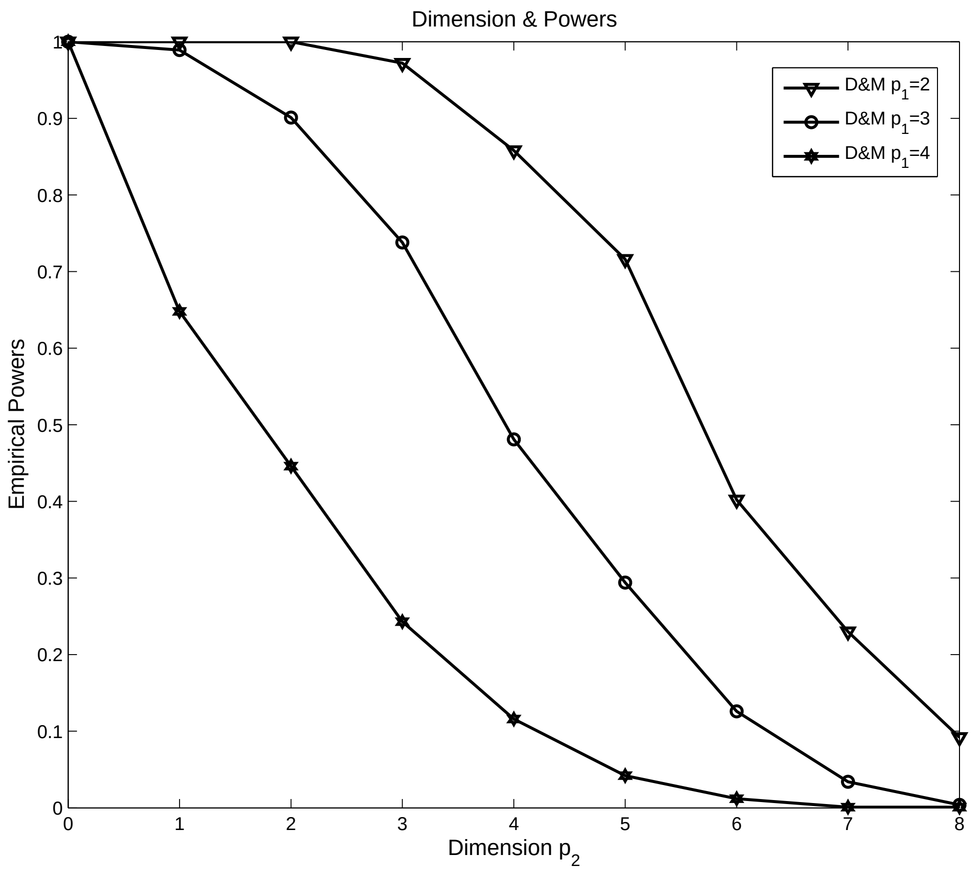

Figure 4.2: The empirical power curve of Delgado and Manteiga (2001)'s test against the dimensions of X and W with sample size 200 at the significance level $\alpha = 0.05$.

level. To this end, how to overcome the aforementioned problems caused by dimensionality is of great importance.

To the best of our knowledge, Guo et al. (2014) recently devised a dimension-reduction local smoothing test which used to test generalized linear regression models. The basic idea is to utilize the dimension reduction structure to adapt the true regression models. This approach improves existing local smoothing tests for checking generalized linear regression models. Under the null hypothesis, the test statistic converges to its limit at a faster convergence rate $O(n^{-1/2}h^{-1/4})$ than the typical rate $O(n^{-1/2}h^{-d/4})$ of local smoothing tests and detect the sequences of local alternative hypotheses distant apart from the null hypothesis at a faster convergence rate $O(n^{-1/2}h^{-1/4})$. Zhu et al. (2015b), following the similar idea in Guo et al. (2014), developed a dimension reduction global smoothing test for more general regression models. Both of these adaptive methods can overcome the curse of dimensionality. However, to identify the projected covariates in the hypothetical and alternative models, they assumed that the projected covariates in the hypothetical models are contained in the alternative models. In the present chapter, we do not impose this assumption and thus, need to modify the identification procedure in the model adaption step. The details will be presented in the next section.

It is of interest to apply the idea to construct a dimension reduction-based test in significance testing. As is known, the objective of significance testing focuses on choosing the significant covariates X in the nonparametric regression. Let $g(x) = E(Y|X = x)$. Then the significance testing becomes the following hypothesis:

$$H_0: \ E(Y|X, W) = g(X) \quad \text{versus} \quad H_1: \ E(Y|X, W) \neq g(X) \tag{4.3}$$

Let $U = Y - g(X)$. Recall $Z = (X^\top, W^\top)^\top$. Then we have under the null hypothesis H_0, $E(U|Z) = m(Z) - g(X) = 0$, but under the alternative hypothesis H_1, $E(U|Z) \neq 0$. To facilitate a more general formulation of model structure we want to test, consider the following reformulation of the above model. Note that $g(\cdot)$ is an unknown function, the nonparametric regression model $Y = g(X) + U$ can be reformulated as:

$$Y = g(X) + U = g(B_1 B_1^\top X) + U \equiv: \tilde{g}(B_1^\top X) + U,$$

where B_1 is an orthogonal $p_1 \times p_1$ matrix. This means that the above regression model can be viewed as a special multi-index regression model with p_1 index corresponding to the covariates X. Thus, in the present chapter, we consider the more general null hypothesis as:

$$H_0 : E(Y|Z) = g(B_1^\top X) = E(Y|X) \quad \text{a.s,} \tag{4.4}$$

where B_1 is a $p_1 \times q_1$ matrix which includes q_1 orthogonal columns with a known number q_1 satisfied $1 \leq q_1 \leq p_1$. For identifiability consideration, assume $B_1^\top B_1 = I_{q_1 \times q_1}$. The hypothetical regression model covers many popularly used models in the literature, including the single-index models, the multi-index models and the partially linear single-index models. When the above regression model is a single-index model or partially linear single index model, the corresponding number of the indexes becomes one or two, respectively. Particularly, when $q_1 = p_1$, the hypothetical model becomes the classical model of (4.3). The alternative hypothesis becomes

$$E(Y|Z) = m(B_1^\top X, W) \neq g(B_1^\top X) = E(Y|X).$$

The term $m(B_1^\top X, W)$ can be reformulated as:

$$E(Y|Z) = m(B_1^\top X, W) = m(B^\top Z),$$

where

$$B = \begin{pmatrix} B_1 & 0_{p_1 \times p_2} \\ 0_{p_2 \times q_1} & I_{p_1 \times p_1} \end{pmatrix}.$$

From this reformulation, we similarly consider a more general alternative hypothesis as:

$$H_1 : E(Y|Z) = m(B^\top Z) \neq g(B_1^\top X) = E(Y|X), \tag{4.5}$$

where B is a $d \times q$ matrix with $q \leq d$ being an unknown value and $E(\varepsilon|Z) = E(\varepsilon|B^\top Z) = 0$. For identifiability consideration, assume $B^\top B = I_{q\times q}$. Therefore, in this chapter, we test the null hypothesis (4.4) against the alternative hypothesis (4.5).

To this end, we develop a test that uses the advantage with less dimension under H_0 and automatically adapt to H_1 such that the test is omnibus. We will call it a dimension reduction model adaptive test (DRMAT). Compared with existing local smoothing tests, DRMAT converges to their limit at the rate of order $O(n^{-1/2}h^{-q_1/4})$ and can detect the local alternatives distinct from the null hypothesis at also this rate, rather than at the rate of $O(n^{-1/2}h^{-d/4})$. Further, the critical values of the new test can be determined by its limiting null distribution. It is worthwhile to mention that almost all existing local smoothing tests require the assistance of Monte Carlo approximation to determine critical values otherwise, the significance level is difficult to maintain. However, the dimension reduction structure of the test alleviates this difficulty. More details will be discussed in the next sections.

This chapter is organized as follows. In Section 4.2, the test statistic construction is described. Because dimension reduction technique plays a very important role, we first briefly review a promising method: the discretization-expectation estimation. To make the test have the model-adaption property, a ridge-type eigenvalue ratio estimate (RERE) for the dimension q of B is recommended, which can estimate q_1 and q consistently under the null and alternative hypotheses accordingly. The asymptotic properties of the test statistic are presented in Section 4.3. Further, the test statistic tends to infinity at the certain rate under the global alternative hypothesis in Section 4.3. In Section 4.4, we examine the finite sample performance of our test and also apply our test to a real data analysis for illustration. All the technical conditions and the proofs of the theoretical results are postponed to Appendix 3.

4.2 A dimension reduction-based model adaptive test

4.2.1 Basic test statistic construction

It is worth noticing that in the above formulations, B and B_1 are usually not identifiable. Actually, for any $q \times q$ orthogonal matrix C, $E(Y|Z) = E(Y|B^\top Z) = E(Y|CB^\top Z)$ and for any $q_1 \times q_1$ orthogonal matrix C_1, $E(Y|X) = E(Y|B_1^\top X) = E(Y|C_1 B_1^\top X)$. Hence, what we can identify is just $BC^\top$ for a matrix C and $B_1 C_1^\top$ for a matrix C_1. Under H_0, since $E(Y|Z) = E(Y|B^\top Z) = E(Y|B_1^\top X)$, B is automatically reduced to a $d \times q_1$ matrix $\tilde{B} = (B_1^\top, O_{q_1 \times p_2}^\top)^\top$ and $q = q_1$. Therefore, it is enough to have such a weaker identification because $CB^\top Z = C(B_1^\top, O_{q_1 \times p_2}^\top)^\top Z = CB_1^\top X$ does not involve W. In the following subsection, we will briefly introduce a method to identify $BC^\top$ for a matrix C and $B_1 C_1^\top$ for a matrix C_1. Without notational confusion, we use B and B_1 to write $BC^\top$ and $B_1 C_1^\top$, respectively, throughout the rest of the present chapter.

Then under H_0, we have

$$E(U|Z) = 0 \Longrightarrow E(U|Z) = E(\epsilon|B^\top Z) = 0.$$

Therefore, we can obtain

$$E(UE(U|B^\top Z)W(B^\top Z)) = E(E^2(\epsilon|B^\top Z)W(B^\top Z)) = 0, \tag{4.6}$$

where $W(B^\top Z)$ is some positive weight function that is discussed below. Under the alternative hypothesis H_1, since

$$E(U|B^\top Z) = m(B^\top Z) - g(B_1^\top X) \neq 0,$$

we have

$$E(UE(U|B^\top Z)W(B^\top Z)) = E(E^2(U|B^\top Z)W(B^\top Z)) > 0. \tag{4.7}$$

The above argument implies that the empirical version of the left hand side in (4.6) can be viewed as a base to construct a test statistic. Further, the null hypothesis H_0 is rejected for large values of the test statistic. This motivates a naive construction as any one in the literature. However, we also note that under the null hypothesis,

$$E(UE(U|B^\top Z)W(B^\top Z)) = E(E^2(\epsilon|B_1^\top X)W(B_1^\top X)) = 0. \tag{4.8}$$

This means that under the null hypothesis, the dimension is reduced from q to q_1.

The key is that how to construct a test statistic that fully uses this piece of information and can automatically adapt the model structure under the alternative hypothesis such that the test is omnibus. We will present our idea in the following construction.

When a sample $\{(z_1, y_1), \cdots, (z_n, y_n)\}$ is available, the residual term u_i is estimated as $\hat{u}_i = y_i - \hat{g}(\hat{B}_1^\top x_i)$, where $\hat{g}(\hat{B}_1^\top x_i)$ is a kernel estimate of $g(B_1^\top x_i)$ as following form:

$$\hat{g}(\hat{B}_1^\top x_i) = \frac{\frac{1}{n-1}\sum_{j\neq j}^{n} Q_{h_1}(\hat{B}_1^\top x_j - \hat{B}_1^\top x_i)y_j}{\frac{1}{n-1}\sum_{j\neq j}^{n} Q_{h_1}(\hat{B}_1^\top x_j - \hat{B}_1^\top x_i)},$$

and $\hat{p}_{\hat{B}_1 i}$ is a kernel estimate of the density function $p_{B_1}(\cdot)$ of $B_1^\top x_i$ given by

$$\hat{p}_{\hat{B}_1 i} = \frac{1}{n-1}\sum_{j=1}^{n} Q_{h_1}(\hat{B}_1^\top x_j - \hat{B}_1^\top x_i),$$

and where $Q_{h_1} = Q(\cdot/h_1)/h_1^{q_1}$ with $Q(\cdot)$ being a q_1-dimensional product kernel function from the univariate kernel $\tilde{Q}(\cdot)$, h_1 being a bandwidth and $\hat{B}_1$ being a estimate of the central mean space $S_{E(Y|X)}$. Then we obtain the following kernel estimator $E(U|B^\top Z)$ as:

$$\hat{E}(\hat{u}_i|\hat{B}^\top z_i) = \frac{\frac{1}{n-1}\sum_{j\neq i}^{n} K_{\hat{q}h}(\hat{B}^\top z_j - \hat{B}^\top z_i)\hat{u}_j}{\frac{1}{n-1}\sum_{j\neq i}^{n} K_{\hat{q}h}(\hat{B}^\top z_j - \hat{B}^\top z_i)}.$$

In this formula, $\hat{B}$ is a sufficient dimension reduction estimator of B with an estimated structural dimension $\hat{q}$ in a certain sense that will be specified in the next subsection, where $K_{\hat{q}h} = K(\cdot/h)/h^{\hat{q}}$ with $K(\cdot)$ being a $\hat{q}$-dimensional kernel function and h being a bandwidth. If we choose the weight $W(\cdot)$ to be the density function $p_B(\cdot)$ of $B^\top Z$, for any $\hat{B}^\top z_i$, we can estimate $p_B(\cdot)$ as the following form:

$$\hat{p}_{\hat{B}i} = \frac{1}{n-1}\sum_{j\neq i}^{n} K_{\hat{q}h}(\hat{B}^\top z_j - \hat{B}^\top z_i).$$

Therefor, a non-standardized test statistic is defined by

$$V_n = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \hat{u}_i\hat{u}_j K_{\hat{q}h}(\hat{B}^\top z_j - \hat{B}^\top z_i). \tag{4.9}$$

Remark 4.1. *Note that the test statistic developed by Fan and Li (1996) is:*

$$\tilde{V}_n = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} \hat{u}_i\hat{f}_{1i}\hat{u}_j\hat{f}_{1j}\tilde{K}_h(z_i - z_j), \tag{4.10}$$

where $\tilde{K}_h(\cdot) = K(\cdot/h)/h^p$ with $K(\cdot)$ being a (p_1+p_2)-dimensional kernel function and $\hat{f}_1$ is an estimator of density function $f_1(\cdot)$ of X no matter whether the underlying model is under the null or alternative hypothesis. Compared the formula (4.9) with (4.10), the difference is that our test uses $\hat{B}^\top Z$ in lieu of Z and applies $K_h(\cdot)$ in V_n instead of $\tilde{K}_h(\cdot)$. It is clear that to make the test model adaptive, we need an estimator of B having the property that under H_0, it estimates $\tilde{B} = (B_1^\top, O_{q_1\times p_2}^\top)^\top$ and under H_1, it automatically estimates B. To achieve this goal, we also need an estimator $\hat{q}$ that can be consistent to q_1 under H_0 and to q under H_1. We will see that this model adaptive estimation plays a very important role in the test statistic construction that uses a standardizing constant $nh^{q_1/2}$ in the test statistic that diverges to infinity much faster than the classical standardizing constant $nh^{(p_1+p_2)/2}$ in Fan and Li (1996)'s test statistic.

4.2.2 A brief review on discretization-expectation estimation (DEE)

As we commented above, we need to identify $BC^\top$ and $B_1C_1^\top$. To this end, a method is discussed in this subsection. The method is to identify the spaces spanned by B and B_1 automatically under the null and alternative hypotheses. In other words, the method is to identify the basis vectors in the respective subspaces. This is an estimation problem for the central subspaces in sufficient dimension reduction (e.g. Cook 1998). The respective central mean subspaces are respectively denoted as $S_{E(Y|X)}$ and $S_{E(Y|Z)}$. Also, $q_1 = \dim(S_{\mathrm{E(Y|X)}})$ and $q = \dim(S_{\mathrm{E(Y|Z)}})$ are respectively called the structural dimensions of $S_{E(Y|X)}$ and $S_{E(Y|Z)}$. Here we assume that q_1 is known, but q is unknown.

There exist several dimension reduction proposals available in the literature. For example, Li (1991) proposed the sliced inverse regression (SIR), Cook and Weisberg (1991) advised sliced average variance estimation (SAVE), Xia et al. (2002) discussed minimum average variance estimation (MAVE), Li and Wang (2007) presented directional regression (DR). Cook and Forzani (2009) developed likelihood acquired directions (LAD), Zhu et al. (2010a) suggested discretization-expectation estimation (DEE) and Zhu et al. (2010b) provided average partial mean estimation (APME). In this chapter, we adopt DEE because it is computationally inexpensive without any tuning parameter selection that is required by SIR, SAVE or DR, and can be easily used to construct a criterion for determining the structural dimensions.

From Zhu et al. (2010a), to identify and estimate B, the DEE estimation procedure can be summarized as the following steps.

1. Define the set of binary variables $\Upsilon(t) = I\{Y \leq t\}$ by discretizing the response variable Y, where the indicator function $I\{Y \leq t\} = 1$ if $Y \leq t$ and 0, otherwise.

2. Let $S_{\Upsilon(t)|Z}$ denote the central subspace of $\Upsilon(t)|Z$, and $M(t)$ be an $d \times d$ positive semi-definite matrix satisfying $\mathrm{Span\{M(t)\}} = S_{\Upsilon\mathrm{(t)|Z}}$.

3. Let $\tilde{Y}$ denote an independent copy of Y. Taking the expectation over the random variable $\tilde{Y}$, the target matrix becomes $M = E\{M(\tilde{Y})\}$. B consists of the eigenvectors associated with the nonzero eigenvalues of M.

4. Get an estimation of the target matrix M as:

$$M_n = \frac{1}{n}\sum_{i=1}^{n} M_n(y_i),$$

where $M_n(y_i)$ is the estimating matrix of $M(y_i)$ by some certain sufficient dimension reduction method such that SIR. Then the estimate $\hat{B}$ of B consists of the eigenvectors associated with the largest q eigenvalues of M_n. Virtually, $\hat{B}$ is root-n consistency of the matrix B when q is given.

In this chapter, the target matrices $M(t)$ and M are based on sliced inverse regression (SIR). More details can be referred to Li (1991) or Zhu et al. (2010a). Similarly, we can also utilize the DEE procedure to estimate the matrix B_1 as $\hat{B}_1$.

The following proposition states the consistency of the estimated matrix $\hat{B}$ under H_0.

Proposition 4.1. *Under H_0, we have $q = q_1$. Further, under H_0 and Conditions A1 and A2 in Appendix 3, the estimator $\hat{B}$ by the DEE procedure is consistent to $\tilde{B} = (B_1^\top C_1, O_{q_1 \times p_2}^\top)^\top$ with C_1 to be some $q_1 \times q_1$ orthogonal matrix.*

Proposition 4.1 indicates that when we construct the test statistic, under H_0, we can only use those variables that are significant. The curse of dimensionality can then be largely alleviated when nonparametric estimation is inevitably required.

However, we do not know whether the underlying model is under the null or alternative hypothesis and q is unknown. Thus, we need to estimate q in general even under the null hypothesis with a given $q = q_1$ as we can then define a final estimate of B. The following subsection provides an estimate and its consistency.

4.2.3 Structural dimension estimation

To estimate q and q_1 in an automatic manner, we need a criterion. Although Zhu et al. (2010a) devised the BIC-type criterion that was motivated from Zhu et al. (2006), choosing an appropriate penalty is an issue in the BIC criterion.

In this chapter, we recommend a ridge-type eigenvalue ration estimate (RERE). Based on our experience in practice, it is less sensitive to the tuning parameter selection. Let $\hat{\lambda}_d \leq \hat{\lambda}_{d-1} \leq \cdots \leq \hat{\lambda}_1$ be the eigenvalues of the estimating matrix M_n. Define

$$\lambda_i^* = \frac{\hat{\lambda}_i - \frac{1}{\sqrt{n}}}{\hat{\lambda}_i - \frac{1}{\sqrt{n}} + 1}, \text{ for } 1 \leq i \leq d.$$

The structure dimension q can be estimated as:

$$\hat{q} = \arg\min_{1 \leq j \leq d} \left\{ j, \frac{(\lambda_{j+1}^*)^2 + c_n}{(\lambda_j^*)^2 + c_n} \right\}. \tag{4.11}$$

This method is motivated by Xia et al. (2014) and Zhu et al. (2015b). This algorithm is easily implemented.

The following proposition states the estimation consistency.

Proposition 4.2. *Under Conditions A1, A2 and A3 in Appendix 3, assume* $c \times \log n/n \leq c_n \to 0$ *with some fixed* $c > 0$*, then the estimator* $\hat{q}$ *by (4.11) is consistent to* q_1 *under* H_0 *and to* q *under* H_1*.*

From Proposition 4.2, the choice of c_n can be in a relatively wide range to guarantee the estimation consistency under the null and global alternative hypotheses. Further, we will prove that $\hat{q}$ converges to q_1 as the sequence of local alternative hypotheses converges to the null model, rather than q. In other words, under the local alternative hypotheses that we will consider below, the estimate $\hat{q}$ is not consistent to the true dimension q. However, this inconsistency has a positive impact for detecting local alternative models. We will discuss this below.

Combining the estimation of B and the relevant structural dimension q, we can see that the estimate $\hat{B}_{\hat{q}}$ can have the model adaptive property in the sense that under H_0, it is consistent to $\tilde{B} = (B_1^\top C_1, O_{q_1 \times p_2}^\top)^\top$ with C_1 to be some $q_1 \times q_1$ orthogonal matrix and under H_1, to B.

4.3 Asymptotic properties

4.3.1 Limiting null distribution

First, define some notations. Let

$$s^2 = 2\int K^2(u)duE\{[Var(\varepsilon^2|B^\top Z)]^2 p(B^\top Z)\}, \tag{4.12}$$

and

$$\hat{s}^2 = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i,j=1}^{n} K_h^2\left(\hat{B}^\top z_i - \hat{B}^\top z_j\right) \hat{u}_i^2 \hat{u}_j^2. \tag{4.13}$$

Theorem 4.1. *Under H_0 and the regularity conditions in Appendix 3, we have*

$$nh^{q_1/2} V_n \xrightarrow{\mathrm{d}} N(0, s^2).$$

Further, s^2 can be consistently estimated by $\hat{s}^2$.

Therefor, according to Theorem 4.1, we can get the standardized test statistic as:

$$T_n = nh^{q_1/2} V_n / \hat{s}.$$

Further, applying the Slusky theorem yields that T_n is asymptotically normal under H_0:

$$T_n \xrightarrow{\mathrm{d}} N(0, 1).$$

4.3.2 Power study

We now study the power performance of the test statistic T_n. Consider the following sequence of local alternative hypotheses as:

$$H_{1n}: \ Y = g(B_1^\top X) + C_n G(B^\top Z) + \varepsilon. \tag{4.14}$$

Fixed C_n corresponds to the global alternative model and when C_n goes to zero, the sequences are local alternative hypotheses.

To obtain the main results about the power performance under H_{1n} of (4.14), we first present the asymptotic behavior of the estimator $\hat{q}$ when $C_n \to 0$. Since the sequence of the local alternative hypotheses converges to the null regression model as $n \to \infty$, the following lemma states that the estimator $\hat{q}$ converges to q_1 rather than the true dimension q.

Lemma 4.1. *Under H_{1n} of (4.14) with $C_n = n^{-\frac{1}{2}} h^{-\frac{q_1}{4}}$ and the regularity conditions in Proposition 4.2 except that $c \times C_n^2 \log n \le c_n \to 0$ for some fixed $c > 0$, as $n \to 0$, $\hat{q}$ determined by (4.11) converges to q_1 with a probability going to 1.*

Now, we turn to state the results about the power performance of our test.

Theorem 4.2. *Under the regularity conditions in Appendix 3, we have the following results.*

(i) Under the global alternative hypothesis H_1, we have

$$T_n/(nh^{q_1/2}) \xrightarrow{\text{p}} Constant > 0.$$

(ii) Under the local alternative hypothesis H_{1n} with $C_n = n^{-\frac{1}{2}} h^{-\frac{q_1}{4}}$, we have

$$nh^{q_1/2} V_n \xrightarrow{\text{d}} N(u, s^2),$$

where $u = E\big([E\{G(B^\top Z)|B_1^\top X\}]^2 p_{B_1}(B_1^\top X)\big)$ and s^2 is given by (4.12). Further, s^2 can be consistently estimated by $\hat{s}^2$.

Remark 4.2. *The results in this theorem confirm our claim in the first section. The convergence rate of the test statistic is $nh^{q_1/2}$ and the test can detect local alternative models converging to the hypothetical model also at the rate of order $C_n = n^{-\frac{1}{2}} h^{-\frac{q_1}{4}}$. Fan and Li (1996)'s test, which is also the case for existing local smoothing tests, can have the respective rates where q_1 is replaced by d, that is a much slower rate.*

4.4 Numerical Studies

4.4.1 Simulations

In this subsection, we conduct the simulations to investigate the finite sample performance of our proposed test. The empirical sizes and powers are computed via 2000 replications of the experiments at the significance level $\alpha = 0.05$. Write our test as T_n^{DEE}. For comparison, we use Fan and Li (1996)'s test and Delgado and Manteiga (2001)'s test as the representatives of existing tests. Write them as T_n^{FL} and T_n^{DM}.

Delgado and Manteiga (2001)'s test is defined as:

$$\tilde{V}_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} [\sum_{j \neq i}^{n} \hat{u}_j \hat{f}_{1j} I(x_j < x_i) I(w_j < w_i)]^2.$$

The critical values are determined by the wild bootstrap. The bootstrap observations are from : $y_i^* = \hat{g}(x_i) + \hat{u}_i \times V_i$, where $\{V_i\}_{i=1}^n$ is a sequence of i.i.d. random variables from the two-point distribution as:

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

The bootstrap critical values are computed across 1000 bootstrap replications. The Gaussian-based kernel of order 4, $Q(u) = (u^4 - 7u^2 + 6)\phi(u)/2$, is used to estimate the nonparametric function $g(\cdot)$, where $\phi(\cdot)$ denotes the standard normal density, see Fan and Hu (1992). For both our test and Fan and Li (1996)'s test, we use the Quartic kernel function as $K(u) = \frac{15}{16}(1-u^2)^2$, if $|u| \leq 1$ and 0, otherwise, in constructing the test statistic such as that in (4.9). To determine the structure dimension q, $c_n = 0.1 \log n / nh^{\frac{q_1}{2}}$ is used.

The observations $\{x_i\}_{i=1}^n$ and $\{w_i\}_{i=1}^n$ are i.i.d., respectively, from multivariate normal distribution $N(0, \Sigma_1)$, $N(0, \Sigma_2)$, or $N(0, \Sigma_3)$ and independent of the standard normal errors, in which $\Sigma_1 = (\sigma_{ij}^{(1)})$, $\Sigma_2 = (\sigma_{ij}^{(2)})$ and $\Sigma_3 = (\sigma_{ij}^{(3)})$ with $\sigma_{ij}^{(1)} = I(i = j) + 0I(i \neq j)$, $\sigma_{ij}^{(2)} = I(i = j) + 0.5^{|i-j|} I(i \neq j)$ and $\sigma_{ij}^{(3)} = I(i = j) + 0.2I(i \neq j)$.

In the simulations, we consider several configurations of the dimensions and the correlation structures of the covariate vectors X and W. The details are in the

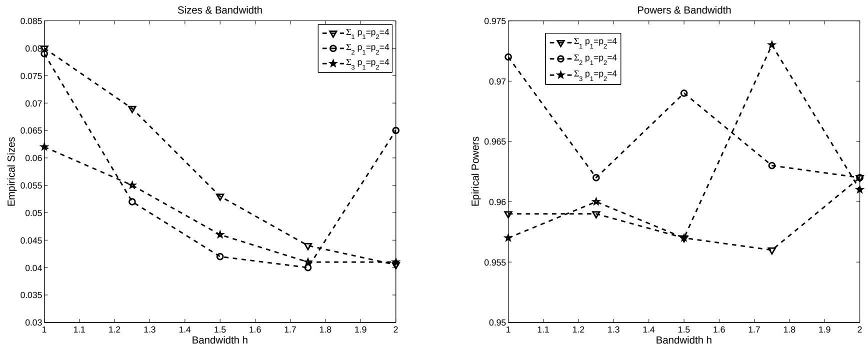

Figure 4.3: The empirical sizes and power curves of our test against the bandwidth h with sample size 200 at the significance level $\alpha = 0.05$ for Example 4.1.

examples.

Example 4.1. Consider the following linear regression model:

- $Y = 2\beta_1^\top X + 2a \times \beta_2^\top W + 0.5 \times \epsilon,$

where $\beta_1 = (\underbrace{1, \cdots, 1}_{p_1/2}, 0, \cdots, 0)^\top / \sqrt{p_1/2}$, $\beta_2 = (0, \cdots, 0, \underbrace{1, \cdots, 1}_{p_2/2})^\top / \sqrt{p_2/2}$. In this example, $p_1 = p_2 = 2$ and $p_1 = p_2 = 4$ are discussed, where the hypothetical and alternative models respond to $a = 0$ and $a \neq 0$. To check the sensitivity of the bandwidth selection, we choose the different bandwidths $h = c \times n^{-1/4+\hat{q}}$ for $c = 1, 1.25, 1.5, 1.75, 2$.

Figure 4.3 reports the empirical sizes and powers with the above bandwidths when $n = 200$. The empirical power is robust against the different bandwidths that we use. The empirical size is sensitive to the bandwidth, but not much. Thus, the bandwidth $h = 1.75 \times n^{-1/(4+\hat{q})}$ is recommended and also used throughout the simulations.

The results of the three tests under different combinations of sample sizes, dimensions of covariate vector X and W and covariance matrixs Σ are reported in Tables 4.1 and 4.2.

From Table 4.1, we can clearly observe that when the dimensions p_1 and p_2 are lower, all the tests have similar empirical powers and can control the empirical sizes well. It is noteworthy that compared to the power performance of the other tests, the

Table 4.1: Empirical sizes and powers of T_n^{DEE}, T_n^{FL} and T_n^{DM} with $p_1 = 2$ and $p_2 = 2$ in Example 4.1.

		T_n^{DEE}			T_n^{FL}			T_n^{DM}		
	a/n	50	100	200	50	100	200	50	100	200
$X \sim N(0, \Sigma_1)$	0	0.0515	0.0485	0.0515	0.0570	0.0595	0.0550	0.0490	0.0520	0.0545
	0.4	0.1060	0.1535	0.4530	0.1560	0.2465	0.4560	0.3665	0.8490	0.9710
$W \sim N(0, \Sigma_1)$	0.8	0.4060	0.7320	0.9230	0.3075	0.5880	0.9335	0.4420	0.8935	0.9860
	1.2	0.6020	0.8770	0.9570	0.4550	0.7750	0.9655	0.4485	0.9075	0.9805
	1.6	0.6680	0.9045	0.9695	0.5115	0.8700	0.9780	0.4590	0.9050	0.9890
	2.0	0.7320	0.9195	0.9785	0.5415	0.8945	0.9890	0.4375	0.9135	0.9935
$X \sim N(0, \Sigma_2)$	0	0.0400	0.0635	0.0515	0.0615	0.0790	0.0580	0.0475	0.0470	0.0460
	0.4	0.0935	0.1450	0.3600	0.1610	0.2345	0.4230	0.3670	0.8520	0.9770
$W \sim N(0, \Sigma_2)$	0.8	0.3435	0.6770	0.9015	0.3780	0.5430	0.9295	0.4400	0.9055	0.9880
	1.2	0.5655	0.8365	0.9420	0.5265	0.7085	0.9510	0.4410	0.9005	0.9930
	1.6	0.6650	0.8985	0.9520	0.6020	0.8340	0.9795	0.4480	0.9015	0.9890
	2.0	0.7110	0.9015	0.9675	0.6145	0.8980	0.9910	0.4645	0.9035	0.9960
$X \sim N(0, \Sigma_3)$	0	0.0515	0.0525	0.0520	0.0590	0.0640	0.0555	0.0455	0.0535	0.0465
	0.4	0.0985	0.1155	0.3825	0.1225	0.2440	0.3270	0.3660	0.8495	0.9750
$W \sim N(0, \Sigma_3)$	0.8	0.3640	0.7455	0.9380	0.3035	0.5905	0.8015	0.4275	0.8935	0.9870
	1.2	0.5930	0.8675	0.9470	0.4150	0.8190	0.9355	0.4455	0.9095	0.9865
	1.6	0.6830	0.9070	0.9640	0.5185	0.8685	0.9660	0.4605	0.8950	0.9875
	2.0	0.7495	0.9175	0.9700	0.5495	0.9055	0.9765	0.4925	0.9045	0.9880

advantage of our test becomes more pronounced when n is relatively small ($n = 50$). From Table 4.2, with increasing the dimensions p_1 and p_2, T_n^{DEE} is significantly and uniformly more powerful than T_n^{FL} and T_n^{DM}. Meanwhile, T_n^{DEE} can still well maintain the significance level, while the competitors can not. Comparing Table 4.1 with Table 4.2, we can see that the dimensions of X and W have litter influence for T_n^{DEE}, but have a significant impact for T_n^{FL} and T_n^{DM}. Fan and Li (1996)'s test completely fails to detect the alternative hypotheses with a power similar to the significance level even when $n = 200$. Under the different covariance matrices Σ of this example, the comparison shows that our test is robust against the correlation structure of (X, Z) whereas it significantly influences the power performance of Delgado and Manteiga

Table 4.2: Empirical sizes and powers of T_n^{DEE}, T_n^{FL} and T_n^{DM} with $p_1 = 4$ and $p_2 = 4$ in Example 4.1.

		T_n^{DEE}			T_n^{FL}			T_n^{DM}		
	a/n	50	100	200	50	100	200	50	100	200
$X \sim N(0, \Sigma_1)$	0	0.0545	0.0435	0.0520	0.0310	0.0620	0.0670	0.0020	0.0045	0.0240
	0.4	0.0730	0.1690	0.3770	0.0520	0.0615	0.0805	0.0060	0.0325	0.2205
$W \sim N(0, \Sigma_1)$	0.8	0.3420	0.7355	0.9190	0.0550	0.0695	0.0955	0.0055	0.0480	0.2645
	1.2	0.5635	0.8660	0.9465	0.0545	0.0700	0.0950	0.0090	0.0425	0.2485
	1.6	0.6290	0.8860	0.9610	0.0530	0.0790	0.0960	0.0065	0.0520	0.2545
	2.0	0.6960	0.9115	0.9825	0.0550	0.0870	0.1150	0.0025	0.0485	0.2610
$X \sim N(0, \Sigma_2)$	0	0.0620	0.0550	0.0525	0.0330	0.0600	0.0565	0.0250	0.0440	0.0800
	0.4	0.0815	0.1550	0.3320	0.0685	0.0820	0.0930	0.1260	0.2005	0.5100
$W \sim N(0, \Sigma_2)$	0.8	0.3555	0.6950	0.9165	0.0810	0.1005	0.1205	0.1410	0.3930	0.7040
	1.2	0.5390	0.8295	0.9540	0.0900	0.1065	0.1415	0.1270	0.4260	0.7420
	1.6	0.6330	0.8795	0.9615	0.1000	0.1215	0.1430	0.1400	0.4285	0.7600
	2.0	0.6965	0.9090	0.9795	0.1070	0.1360	0.1595	0.1410	0.4180	0.7720
$X \sim N(0, \Sigma_3)$	0	0.0400	0.0420	0.0555	0.0320	0.0705	0.06255	0.0070	0.0280	0.0715
	0.4	0.0680	0.1325	0.3900	0.0660	0.0720	0.07500	0.0410	0.1265	0.6300
$W \sim N(0, \Sigma_3)$	0.8	0.3325	0.6960	0.9170	0.0815	0.0775	0.09505	0.0520	0.2320	0.6640
	1.2	0.5500	0.8345	0.9490	0.0830	0.0915	0.10005	0.0530	0.2315	0.6545
	1.6	0.6245	0.8845	0.9535	0.0865	0.1040	0.11850	0.0630	0.2320	0.6520
	2.0	0.6740	0.8890	0.9775	0.0875	0.1030	0.11655	0.0560	0.2500	0.6430

(2001)'s test, particularly when $p_1 = p_2 = 4$.

Example 4.2. In this example, consider a nonlinear high-frequency regression model as:

- $Y = 2\sin(\beta_1^\top X) + 2a \times \sin(\beta_2^\top W) + 0.5 \times \epsilon$,

where $\beta_1 = (\underbrace{1, \cdots, 1}_{p_1/2}, 0, \cdots, 0)^\top/\sqrt{p_1/2}$ and $\beta_2 = (0, \cdots, 0, \underbrace{1, \cdots, 1}_{p_2/2})^\top/\sqrt{p_2/2}$. We also consider two cases of dimensions: $p_1 = p_2 = 4$ and $p_1 = p_2 = 6$. Again $a = 0$ responds to the hypothetical model.

The results are presented in Tables 4.3 and 4.4. Comparing Table 4.2 with Table 4.3, we find that when the dimensions of X and W are added up to $p_1 = 6$ and $p_2 = 6$, the empirical powers of T_n^{FL} and T_n^{DM} are close to 0. This result means both

Table 4.3: Empirical sizes and powers of T_n^{DEE}, T_n^{FL} and T_n^{DM} with $p_1 = 4$ and $p_2 = 4$ in Example 4.2.

		T_n^{DEE}			T_n^{FL}			T_n^{DM}		
	a/n	50	100	200	50	100	200	50	100	200
$X \sim N(0, \Sigma_1)$	0	0.0575	0.0520	0.0495	0.0360	0.0320	0.0695	0.0060	0.0100	0.0180
	0.4	0.0805	0.2225	0.6120	0.0665	0.0780	0.0910	0.0080	0.0520	0.1225
$W \sim N(0, \Sigma_1)$	0.8	0.2910	0.7200	0.9190	0.0775	0.0875	0.0915	0.0040	0.0560	0.1790
	1.2	0.5465	0.8295	0.9405	0.0885	0.0820	0.1075	0.0200	0.0655	0.2290
	1.6	0.6215	0.8445	0.9685	0.0775	0.0915	0.1110	0.0160	0.0805	0.2415
	2.0	0.6565	0.8695	0.9730	0.0830	0.0960	0.1220	0.0260	0.0940	0.2615
$X \sim N(0, \Sigma_2)$	0	0.0435	0.0435	0.0470	0.0640	0.0695	0.0760	0.0240	0.0380	0.0470
	0.4	0.1145	0.3035	0.6545	0.0745	0.0860	0.1005	0.0620	0.1255	0.5870
$W \sim N(0, \Sigma_2)$	0.8	0.3735	0.6780	0.8755	0.0820	0.1065	0.1230	0.0845	0.2380	0.7190
	1.2	0.5570	0.7850	0.8830	0.0780	0.1000	0.1420	0.1070	0.2565	0.7660
	1.6	0.6445	0.8130	0.9070	0.0905	0.1090	0.1545	0.1150	0.3070	0.7610
	2.0	0.6745	0.8455	0.9300	0.0890	0.1265	0.1620	0.1115	0.3330	0.7900
$X \sim N(0, \Sigma_3)$	0	0.0475	0.0490	0.0525	0.0360	0.0820	0.0705	0.0130	0.0230	0.0405
	0.4	0.0970	0.2465	0.6405	0.0590	0.0845	0.0920	0.0370	0.1140	0.4645
$W \sim N(0, \Sigma_3)$	0.8	0.3360	0.7120	0.9120	0.0615	0.0850	0.1045	0.0580	0.1835	0.5715
	1.2	0.5600	0.7995	0.9265	0.0520	0.0925	0.1180	0.0690	0.2000	0.6370
	1.6	0.6625	0.8405	0.9370	0.0680	0.0990	0.1260	0.0700	0.2020	0.6125
	2.0	0.6760	0.8605	0.9500	0.0715	0.0950	0.1215	0.0730	0.2225	0.6240

the competitors completely fail to detect the alternative hypothesis. We can also find that the empirical power of our test has little change. This seems that the dimensions of X and W have little impact for T_n^{DEE}, but it has a very serious impact for T_n^{FL} and T_n^{DM}. Again our test can also control the empirical size well.

The next example is to confirm that our test is omnibus rather than directional.

Example 4.3. The data are generated from the following model:

- $Y = 2\sin(\beta_1^\top X) + \exp(\beta_2^\top X/2) + 2a \times \sin(\beta_2^\top W) + 0.5 \times \epsilon$,

where $p_1 = 4$, $p_2 = 4$, $\beta_1 = (1, 1, 0, 0)^\top/\sqrt{2}$ and $\beta_2 = (0, 0, 1, 1)^\top/\sqrt{2}$. Thus, $q_1 = 2$, and $q = 3$. In this model, the conditional expectation of $E(Y|B_1^T X)$ is the same under the null and alternative hypotheses. If we simply used this function to define

Table 4.4: Empirical sizes and powers of T_n^{DEE}, T_n^{FL} and T_n^{DM} with $p_1 = 6$ and $p_2 = 6$ in Example 4.2.

		T_n^{DEE}			T_n^{FL}			T_n^{DM}		
	a/n	50	100	200	50	100	200	50	100	200
$X \sim N(0, \Sigma_1)$	0	0.0405	0.0455	0.0465	0.0005	0.0005	0.0055	0	0	0
	0.4	0.0720	0.1860	0.5515	0.0005	0.0025	0.0035	0	0	0
$W \sim N(0, \Sigma_1)$	0.8	0.2445	0.6240	0.9045	0	0.0010	0.0030	0	0	0
	1.2	0.4575	0.7925	0.9190	0	0.0015	0.0035	0	0	0
	1.6	0.5595	0.8260	0.9240	0	0.0005	0.0035	0	0	0
	2.0	0.6255	0.8455	0.9445	0	0.0015	0.0040	0	0	0
$X \sim N(0, \Sigma_2)$	0	0.0425	0.0470	0.0485	0.0035	0.0135	0.0215	0.0020	0.0110	0.0195
	0.4	0.1335	0.2490	0.5550	0.0020	0.0080	0.0270	0.0020	0.0250	0.1040
$W \sim N(0, \Sigma_2)$	0.8	0.3160	0.6215	0.8120	0.0020	0.0185	0.0275	0.0030	0.0270	0.1550
	1.2	0.5025	0.7645	0.8580	0.0030	0.0100	0.0250	0.0030	0.0290	0.1695
	1.6	0.5770	0.7950	0.8850	0.0050	0.0105	0.0290	0.0020	0.0280	0.1725
	2.0	0.6375	0.8290	0.9105	0.0070	0.0140	0.0205	0.0020	0.0320	0.1685
$X \sim N(0, \Sigma_3)$	0	0.0455	0.0460	0.0475	0	0.0025	0.0100	0	0.0035	0.0115
	0.4	0.0985	0.2430	0.5830	0.0005	0.0035	0.0090	0	0.0040	0.0520
$W \sim N(0, \Sigma_3)$	0.8	0.3225	0.6515	0.8750	0.0030	0.0020	0.0095	0.0010	0.0075	0.0605
	1.2	0.4915	0.7695	0.8940	0.0015	0.0015	0.0080	0.0040	0.0080	0.0780
	1.6	0.5880	0.8030	0.9135	0.0010	0.0005	0.0060	0.0020	0.0120	0.0720
	2.0	0.6265	0.8380	0.9395	0.0010	0.0035	0.0085	0.0050	0.0135	0.0895

a test, the alternative hypothesis cannot be detected at all from the theoretical point of view. However, our test can adapt to the alternative model with the matrix $B^T Z$, we want to use this simulation experiment to show the power of the test.

By the comparison between Tables 4.2, 4.3 and 4.5, we observe that the results for Example 4.3 are parallel to those in Examples 4.1 and 4.2 with high power. This means the test has the advantage of dimension reduction and is still an omnibus test.

The following example considers higher dimensional $B^\top Z$ in a model.

Example 4.4 The data are generated from the following model:

- $Y = X_1 + 0.2\exp(X_2) + a \times \frac{1.5(W_1+W_2)}{0.5+(1.5W_3+0.5)^{1.5}} + 0.75\sin(W_4+1)^{1.5} + 0.2 \times \epsilon,$

where $p_1 = 4$, $p_2 = 4$, $X \sim N(0, 4\Sigma_1)$ and $W \sim N(0, 4\Sigma_1)$ and $\epsilon \sim N(0, 1)$. Thus, B_1

Table 4.5: Empirical sizes and powers of T_n^{DEE}, T_n^{FL} and T_n^{DM} with $p_1 = 4$ and $p_2 = 4$ in Example 4.3.

		T_n^{DEE}			T_n^{FL}			T_n^{DM}		
	a/n	50	100	200	50	100	200	50	100	200
$X \sim N(0, \Sigma_1)$	0	0.0420	0.0530	0.0525	0.0575	0.0770	0.0760	0.0015	0.0080	0.0190
	0.4	0.0655	0.0985	0.3755	0.0510	0.0740	0.0765	0.0015	0.0280	0.1495
$W \sim N(0, \Sigma_1)$	0.8	0.1470	0.5635	0.8930	0.0600	0.0845	0.1070	0.0040	0.0350	0.2085
	1.2	0.3690	0.8045	0.9390	0.0510	0.0840	0.0980	0.0035	0.0420	0.2525
	1.6	0.5730	0.8765	0.9485	0.0555	0.0840	0.1080	0.0090	0.0330	0.2510
	2.0	0.6480	0.8960	0.9645	0.0565	0.1000	0.1225	0.0030	0.0280	0.2595
$X \sim N(0, \Sigma_2)$	0	0.0510	0.0545	0.0510	0.0720	0.0745	0.0810	0.0180	0.0440	0.0580
	0.4	0.0755	0.1165	0.2820	0.0685	0.0915	0.0945	0.0715	0.2460	0.3295
$W \sim N(0, \Sigma_2)$	0.8	0.2075	0.5265	0.8490	0.0800	0.1175	0.1315	0.0925	0.3525	0.7430
	1.2	0.4245	0.7610	0.9365	0.0860	0.1040	0.1575	0.0950	0.3740	0.7615
	1.6	0.5595	0.8600	0.9480	0.0810	0.1095	0.1470	0.1145	0.3885	0.7630
	2.0	0.6315	0.8730	0.9560	0.0885	0.1205	0.1610	0.1140	0.3910	0.8185
$X \sim N(0, \Sigma_3)$	0	0.0513	0.0550	0.0495	0.0515	0.0405	0.0790	0.0110	0.0220	0.0340
	0.4	0.0750	0.1255	0.3135	0.0625	0.0720	0.0915	0.0200	0.1320	0.4925
$W \sim N(0, \Sigma_3)$	0.8	0.1945	0.5490	0.8840	0.0625	0.0875	0.1060	0.0340	0.1870	0.6035
	1.2	0.4330	0.7795	0.9300	0.0655	0.0945	0.1270	0.0230	0.1990	0.6100
	1.6	0.5955	0.8370	0.9460	0.0580	0.0960	0.1225	0.0470	0.1960	0.6035
	2.0	0.6405	0.8610	0.9535	0.0595	0.0990	0.1290	0.0460	0.2200	0.6280

is a 4×2 matrix with $\beta_1 = (1, 0, 0, 0)^\top$ and $\beta_2 = (0, 1, 0, 0)^\top$ and B is a 8×5 matrix with low-right block in which the columns are $b_1 = (1, 1, 0, 0)^\top$, $b_2 = (0, 0, 1, 0)^\top$, and $b_3 = (0, 0, 0, 1)^\top$. The results are summarized in Tables 4.6. Comparing with the results in examples 4.1-4.3, we can see that the empirical power of T_n^{DEE} reasonably becomes lower, but is still higher than those of T_n^{FL} and T_n^{DM}. Also, the significance level maintainance works well.

In summary, the above simulations sustain the aforementioned theoretical properties that the proposed test is significantly superior to existing tests among which Fan and Li (1996)'s test and Delgado and Manteiga (2001)'s test are regarded as representatives of existing tests.

Table 4.6: Empirical sizes and powers of T_n^{DEE} and the frequency of structure dimension $\hat{q}$ with $p_1 = 4$ and $p_2 = 4$ in Example 4.4.

n	a	T_n^{DEE}	T_n^{FL}	T_n^{DM}
$n = 200$	0	0.0555	0.0600	0.0290
	1	0.3315	0.1455	0.1870
$n = 400$	0	0.0535	0.0645	0.0465
	1	0.4910	0.1710	0.3205

4.4.2 Baseball hitters' salary data

We now analyze the well-known Baseball hitters' salary data, which is originally published for the 1988 ASA Statistical Graphics and Computing Data Exposition and is available at `http://euclid.psych.yorku.ca/ftp/sas/sssg/data/baseball.sas`. The data set consists of information on salary and 16 performance measures of 263 major league baseball hitters. As always, the question of main interest is whether salary reflects performance. As displayed by Friendly (2002), the 16 measures naturally belong to three performance categories: the season hitting statistics, which include the numbers of times at bat (X_1), hits (X_2), home runs (X_3), runs (X_4), runs batted in (X_5), and walks (X_6) in 1986; the career hitting statistics, which include the numbers of years in the major leagues (X_7), times at bat (X_8), hits (X_9), home runs (X_{10}), runs (X_{11}), runs batted in (X_{12}) and walks (X_{13}) during the players' entire career up to 1986; and the fielding variables, which include the numbers of putouts (X_{14}), assists (X_{15}) and errors (X_{16}) in 1986.

Further, the predictors from different groups have weak correlations. The logarithm of annual salary in 1987 is used to be the response variable (Y) and the new predictors from the career totals by dividing totals by years in the major leagues are constructed. Let $X_j^* = X_j/X_7$ for $j = 8, \cdots, 13$. As remarked by Hoaglin and Velleman (1995), the analyses working with ln(salary) and with the annual rate predictors fared better than those worked with the raw forms of these variables. Below, we use X_j^*'s instead. All the predictors are standardized to have mean zero and unit length. We write $V_1 = (X1, \cdots, X6)$, $V_2 = (X_7, X_8^*, \cdots, X_{13}^*)$ and $V_3 = (X_{14}, X_{15}, X_{16})$. In this application, we consider two cases:

Case (I): $X = (V_1, V_2)$ and $W = V_3$;

Case (II): $X = V_1$ and $W = (V_2, V_3)$;

The values of the test statistics are 6.5831 and 1.1171 under the two cases and the corresponding $p-$values are respectively 0.0000 and 0.1320.

From these results under Cases (I) and (II), we can conclude that the career hitting statistics of the group V_2 leads some support to the annal salary. The results are consistent with those advised by Xia et al. (2002) who found that the variables X_7, X_9 and X_{13} in the group V_2 are prominently to affect the annal salary. However, the coefficients of the fielding variables in the group V_3 are closed to 0 in the estimated directions suggested by Xia et al. (2002). Therefore, for the annual salary, the group V_2 contains the significance variables but the group V_3 does not.

4.5 Conclusions

In this chapter, we have developed a dimension reduction model-adaptive test to determine significant covariates in a nonparametric regression framework. The approach employs dimension reduction technique to reduce the dimension such that the constructed test can well maintain the significance level and more powerful than existing local smoothing tests in the literature. This methodology can be applicable to check other semi-parametric regression models, for example partially linear models, single-index models and partially linear single-index models. The research is on-going.

4.6 Appendix 3

4.6.1 Regularity Conditions

To prove the asymptotic properties in Sections 4.2 and 4.3, we provide the following regularity conditions:

A1 $M_n(t)$ has the following expansion:

$$M_n(t) = M(t) + E_n\{\psi(Z, Y, t)\} + R_n(t),$$

where $E_n(\cdot)$ denotes the average over all sample points, $E(\psi(Z, Y, t)) = 0$ and $E\{\psi^2(Z, Y, t)\} < \infty$.

A2 $\sup_t ||R_n(t)||_F = o_p(n^{-1/2})$, where $||\cdot||_F$ denotes the Frobenius norm of a matrix.

A3 The estimator $\tilde{M}_n(t)$ has the following expansion:

$$\tilde{M}_n(t) = \tilde{M}(t) + E_n\{\tilde{\psi}(X, Y, t)\} + \tilde{R}_n(t),$$

where $\tilde{M}_n(y_i)$ is an estimator of the $p_1 \times p_1$ positive semi-definite matrix $\tilde{M}(t)$ satisfying $\text{Span}\{\text{M(t)}\} = S_{\Upsilon\text{(t)}|\text{X}}$, $E(\tilde{\psi}(X, Y, t)) = 0$, $E\{\tilde{\psi}^2(X, Y, t)\} < \infty$ and $\sup_t ||\tilde{R}_n(t)||_F = o_p(n^{-1/2})$. Corresponding, $\tilde{M}_n = \frac{1}{n}\sum_{i=1}^{n} \tilde{M}_n(y_i)$ is an estimator of the target matrix $\tilde{M}$ satisfying $\text{Span}\{\tilde{\text{M}}\} = S_{\text{Y}|\text{E(Y|X)}}$.

A4 $(B^\top z_i, y_i)_{i=1}^n$ is from the probability distribution $F(B^\top z, y)$ on $\mathbb{R}^q \times \mathbb{R}$. The error $\varepsilon = Y - m(B^\top Z)$ satisfies that $E(\varepsilon^8|B^\top Z = B^\top z)$ is continuous and $E(\varepsilon^8|B^\top Z = B^\top z) \leq b(B^\top z)$ almost surely, where $b(B^\top z)$ is a measurable function such that $E(b^2(B^\top Z)) < \infty$.

A5 The density function $p_{B_1}(\cdot)$ of $B_1^\top X$ exists with support $\mathbb{C}$ and has a continuous and bounded first-order derivative on the support $\mathbb{C}$. The density $p_{B_1}(\cdot)$ satisfies

$$0 < \inf_{B_1^\top X \in \mathbb{C}} p_{B_1}(B_1^\top X) \leq \sup_{B_1^\top X \in \mathbb{C}} p_{B_1}(B_1^\top X) < \infty.$$

A6 The function $g(\cdot)$ is η-order partially differentiable for some positive integer η, and the ηth partially derivative of $g(\cdot)$ is bounded.

A7 $\tilde{Q}(\cdot)$ is a symmetric and twice order continuously differentiable kernel function satisfying

$$\int u^i \tilde{Q}(u)du = \delta_{i0}, \quad (i = 0, 1, \cdots, \eta - 1),$$
$$\tilde{Q}(u) = O[(1 + |u|^{\eta+1+\epsilon})^{-1}], \quad \text{some} \quad \epsilon > 0,$$

where δ_{ij} is the Kronecker's delta and η is given in Condition A6.

A8 $K(\cdot)$ is a bounded, symmetric kernel function and it is a first order continuously differentiable kernel function satisfying $\int K(u)du = 1$.

A9 $n \to \infty$, $h_1 \to 0$, $h \to 0$,

1) under the null or local alternative hypotheses, $nh_1^{q_1} \to \infty$, $nh^{q_1} \to \infty$ and $nh_1^{2\eta}h^{q_1/2} \to 0$;

2) under global alternative hypothesis H_1, $nh_1^{q_1} \to \infty$, $nh^q \to \infty$ and $nh_1^{2\eta}h^{q/2} \to 0$,

where η is given in Condition A6.

Remark 4.3. *Conditions A1, A2 and A3 are necessary for DEE to estimate the matrixes B and B_1. Under the linearity condition and constant conditional variance condition, DEE_{SIR} satisfies the Conditions A1, A2 and A3. Conditions A4, A5, A6 and A7 are widely used for nonparametric estimation in the literature. It is worth pointing out that Condition A6 about the higher-order kernel plays an important roles in bias reduction, see Fan and Li (1996). Conditions A5 and A8 guarantee the asymptotic normality of our test statistic and make the test well-behaved. Condition A9 about the choice of bandwidth h is reasonable because the estimation $\hat{q}$ is different under the null and alternative hypotheses.*

4.6.2 Proof of the theorems

Proof of Proposition 4.1 Note that under the null hypothesis, $E(Y|Z) = g(B_1^\top X)$. Then, we have

$$Y \perp\!\!\!\perp E(Y|Z)|B_1^\top X,$$

where the notation $\perp\!\!\!\perp$ stands for independence. This is equivalent to

$$Y \perp\!\!\!\perp E(Y|Z)|\tilde{B}^\top Z,$$

where $\tilde{B} = (B_1^\top, O_{q_1 \times p_2}^\top)^\top$. From the definition of central mean subspace, $S_{E(Y|Z)}$ is the intersection of all the linear spaces spanned respectively by the columns of any $d \times q$ orthogonal matrix Γ with $1 \leq q \leq d$ such that the above conditional independence holds. Thus, $S_{E(Y|Z)} \subseteq \text{Span}(\tilde{\text{B}})$ where $\text{Span}(\tilde{\text{B}})$ is the linear space spanned by the columns of $\tilde{B}$. Let $\{\gamma_1, \cdots, \gamma_{q_1}\}$ the eigenvectors associated with the nonzero eigenvalue eigenvalues of M. As Zhu et al. (2010a) argued that $\{\gamma_1, \cdots, \gamma_{q_1}\} \in S_{E(Y|Z)}$, we have $\{\gamma_1, \cdots, \gamma_{q_1}\} \in \text{Span}(\tilde{\text{B}})$. This implies that γ_j for $j = 1, \cdots q_1$ can be denoted as a linear combination of the columns of $\tilde{B}$. Thus, for $j = 1, \cdots q_1$, γ_j has the similar form as $\gamma_j = (\tilde{\gamma}_j, 0_{1\times p_2}^\top)^\top$ with $\tilde{\gamma}_j$ being a $p_1 \times 1$ vector. This implies that any element in $S_{E(Y|Z)}$ can also be written as $\gamma_j = (\tilde{\gamma}_j, 0_{1\times p_2}^\top)^\top$. Further, the structural dimension of $S_{E(Y|Z)}$ is smaller than or equal to q_1. Further, we note that under H_0, $E(Y|Z) = E(Y|B^\top Z) = E(Y|B_1^\top X)$ and $E(Y|Z) = E(Y|X)$. Thus, $q = q_1 = \dim(S_{\text{E(Y|X)}})$.

Under Conditions A1 and A2, Theorem 2 in Zhu et al. (2010a) shows that $M_n - M = O_p(n^{-1/2})$. From the arguments in Zhu and Fang (1996) and Zhu and Ng (1995), under some regularity conditions, $\hat{\lambda}_i - \lambda_i = O_p(n^{-1/2})$, where $\hat{\lambda}_d \leq \hat{\lambda}_{d-1} \leq \cdots \leq \hat{\lambda}_1$ are the eigenvalues of the matrix M_n and λ_i are the eigenvalues of the matrix M. The estimate $\hat{B}$ that consists of the eigenvectors associated with the largest q_1 eigenvalues of M_n is consistent to $\tilde{B} = (C_1 B_1^\top, O_{q_1\times p_2}^\top)^\top$ for a $q_1 \times q_1$ orthogonal matrix C_1. □

Proof of Proposition 4.2. From the proof of Proposition 4.1, $\hat{\lambda}_i - \lambda_i = O_p(n^{-1/2})$, where $\hat{\lambda}_d \leq \hat{\lambda}_{d-1} \leq \cdots \leq \hat{\lambda}_1$ are the eigenvalues of the matrix M_n.

Define

$$\tilde{\lambda}_i = \frac{\lambda_i}{\lambda_i + 1}, \text{ for } 1 \le i \le d,$$

where $\lambda_d = \cdots = \lambda_{d-q} = 0$ and $0 < \lambda_q \le \cdots \le \lambda_1$ are the eigenvalues of the target matrix M.

Recall the definition of λ_l^* in Subsection 4.2.3. For any $1 < l \le q$, since $\lambda_l > 0$ and $\hat{\lambda}_l^2 = \lambda_l^2 + O_p(1/\sqrt{n})$, we have $(\lambda_l^*)^2 = (\tilde{\lambda}_l)^2 + O_p(1/\sqrt{n})$. On the other hand, for any $q < l \le d$, as $\lambda_l = 0$ and $\hat{\lambda}_l^2 = \lambda_l^2 + O_p(1/n) = O_p(1/n)$, and then $(\lambda_l^*)^2 = O_p(1/n)$.

For any $l < q$, because $\tilde{\lambda}_l^2 > 0$ and $\tilde{\lambda}_{l+1}^2 > 0$, we have

$$\begin{aligned}
\frac{(\lambda_{q+1}^*)^2 + c_n}{(\lambda_q^*)^2} - \frac{(\lambda_{l+1}^*)^2 + c_n}{(\lambda_l^*)^2} &= \frac{\tilde{\lambda}_{q+1}^2 + c_n + O_p(1/n)}{\tilde{\lambda}_q^2 + c_n + O_p(1/\sqrt{n})} - \frac{\tilde{\lambda}_{l+1}^2 + c_n + O_p(1/\sqrt{n})}{\tilde{\lambda}_l^2 + c_n + O_p(1/\sqrt{n})} \\
&= \frac{c_n + O_p(1/n)}{\tilde{\lambda}_q^2 + c_n + O_p(1/\sqrt{n})} - \frac{\tilde{\lambda}_{l+1}^2 + c_n + O_p(1/\sqrt{n})}{\tilde{\lambda}_l^2 + c_n + O_p(1/\sqrt{n})}.
\end{aligned}$$

Since $c\frac{\log n}{n} \le c_n \to 0$ with some fixed $c > 0$, we get

$$\frac{(\lambda_{q+1}^*)^2 + c_n}{(\lambda_q^*)^2} - \frac{(\lambda_{l+1}^*)^2 + c_n}{(\lambda_l^*)^2} \to \frac{0}{\tilde{\lambda}_{q_1}^2} - \frac{\tilde{\lambda}_{(l+1)}^2}{\tilde{\lambda}_l^2} = -\frac{\tilde{\lambda}_{(l+1)}^2}{\tilde{\lambda}_l^2} < 0.$$

For any $l > q$, $\tilde{\lambda}_l = 0$ and $\tilde{\lambda}_q^2 > 0$, then we have

$$\begin{aligned}
\frac{(\lambda_{q+1}^*)^2 + c_n}{(\lambda_q^*)^2} - \frac{(\lambda_{l+1}^*)^2 + c_n}{(\lambda_l^*)^2} &= \frac{\tilde{\lambda}_{q+1}^2 + c_n + O_p(1/n)}{\tilde{\lambda}_q^2 + c_n + O_p(1/\sqrt{n})} - \frac{\tilde{\lambda}_{l+1}^2 + c_n + O_p(1/n)}{\tilde{\lambda}_l^2 + c_n + O_p(1/n)} \\
&= \frac{c_n + o_p(c_n)}{\tilde{\lambda}_q^2 + c_n + O_p(1/\sqrt{n})} - \frac{c_n + o_p(c_n)}{c_n + o_p(c_n)} \\
&\to -1 < 0.
\end{aligned}$$

Therefor, altogether, it is concluded that $\hat{q} = q$ in probability. □

Proof of Theorem 4.1. For notational convenience, denote $z_i = (x_i, w_i)$, $g_i = g(B_1^\top x_i)$, $\hat{g}_i = \hat{g}(\hat{B}_1^\top x_i)$, $u_i = y_i - g_i$, $\hat{u}_i = y_i - \hat{g}_i$, $K_{Bij} = K(B^\top(z_i - z_j)/h)$, $p_i = p_{B_1 i}$ and $\hat{p}_i = \hat{p}_{\hat{B}_1 i}$. Throughout Appendix 3, let $E_i(\cdot) = E(\cdot|z_i)$.

Noted that $\hat{u}_i \equiv: y_i - \hat{g}_i = u_i - (\hat{g}_i - g_i)$. We then decompose the term V_n as:

$$\begin{aligned} V_n &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j} \frac{1}{h^{q_1}} K_{\hat{B}ij} \mu_i \mu_j + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j} \frac{1}{h^{q_1}} K_{\hat{B}ij} (\hat{g}_i - g_i)(\hat{g}_j - g_j) \\ &\quad -2 \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j} \frac{1}{h^{q_1}} K_{\hat{B}ij} u_i (\hat{g}_j - g_j) + o_p(n^{-1} h^{-q_1/2}) \\ &\equiv: Q_{1n} + Q_{2n} - 2Q_{3n} + o_p(n^{-1} h^{-q_1/2}). \end{aligned}$$

The first equality is got by using Lemma 2 in Guo et al. (2014), where $\hat{q} = q$. First, we consider the term Q_{1n}. By taking the one order Taylor expansion for Q_{1n} with respect to B, we have

$$Q_{1n} \equiv: Q_{11n} + Q_{12n},$$

where Q_{11n} and Q_{12n} have following forms:

$$\begin{aligned} Q_{11n} &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j} \frac{1}{h^{q_1}} K_{Bij} \mu_i \mu_j, \\ Q_{12n} &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j} \frac{1}{h^{2q_1}} K'_{\tilde{B}ij} \mu_i \mu_j (\hat{B} - B)^\top (z_i - z_j), \end{aligned}$$

and where $\tilde{B} = \{\tilde{B}_{ij}\}_{p \times q_1}$ with $\tilde{B}_{ij} \in [\min\{\hat{B}_{ij}, B_{ij}\}, \max\{\hat{B}_{ij}, B_{ij}\}]$. Because $||\hat{B} - B|| = O_p(1/\sqrt{n})$ and the one-order differential function of $K_B(\cdot)$ respect to B is a bounded continuity function of B, we conclude that replacing $\tilde{B}$ by B does not affect the convergence rate of Q_{12n}.

In the present chapter, we suppose the dimension of $B^\top Z$ is fixed, the term Q_{11n} is an $U-$statistic. Since under H_0, $q = q_1$, following a similar argument as that for Lemma 3.3 in Zheng (1996), it is easy to obtain:

$$n h^{q_1/2} Q_{11n} \xrightarrow{\text{d}} N(0, s),$$

where

$$s = 2 \int K_B^2(u) du \cdot \int (\sigma^2(B^\top Z))^2 p^2(B^\top Z) dB^\top Z$$

with $\sigma^2(B^\top z) = E(u^2 | B^\top Z = B^\top z)$. We then omit the details.

We turn to discuss the term Q_{12n}. Since $E(u_i|z_i) = 0$, we have $E(Q_{12n}) = 0$. We then calculate the second order moment of Q_{12n} as follows:

$$\begin{aligned} E(Q_{12n}^2) &= E\big[\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{i\neq j}\frac{1}{h^{2q_1}}K'_{Bij}\mu_i\mu_j(\hat{B}-B)^\top(z_i-z_j)\big]^2 \\ &= E\big[\frac{1}{n^2(n-1)^2}\frac{1}{h^{4q_1}}\sum_{i=1}^{n}\sum_{i'\neq j'}\sum_{i=1}^{n}\sum_{i'\neq j'}K'_{Bij}K'_{Bi'j'} \\ &\quad \mu_i\mu_j\mu_{i'}\mu_{j'}(\hat{B}-B)^\top(z_i-z_j)(z_{i'}-z_{j'})^\top(\hat{B}-B)\big]. \end{aligned}$$

Since $E(u_iu_ju_{i'}u_{j'}) \neq 0$ if and only if $i = i'$, $j = j'$ or $i = j'$, $j = i'$, we have

$$\begin{aligned} E(Q_{12n}^2) &= \frac{2n(n-1)}{n^2(n-1)^2}\frac{1}{h^{4q_1}}E\big[(K'_{B12})^2\mu_1^2\mu_2^2(\hat{B}-B)^\top(z_1-z_2)(z_1-z_2)^\top(\hat{B}_{\hat{q}}-B)\big] \\ &= \frac{2}{n(n-1)}\frac{1}{h^{4q_1}}E\big[(K'_{B12})^2\mu_1^2\mu_2^2(\hat{B}-B)^\top(z_1-z_2)(z_1-z_2)^\top(\hat{B}_{\hat{q}}-B)\big] \\ &= \frac{2}{n(n-1)}\frac{1}{h^{4q_1}}\{E(u_1^2)\}^2(\hat{B}-B)^\top E\big\{(K'_{B12})^2(z_1-z_2)(z_1-z_2)^\top\big\}(\hat{B}-B). \end{aligned}$$

By changing variables as $v_1 = (z_1 - z_2)/h$, we further compute to have

$$\begin{aligned} E(Q_{12n}^2) &= \frac{1}{n(n-1)}\frac{1}{h^{4q_1}}\{E(u_1^2)\}^2\int\int(K'_{B12})^2(\hat{B}-B)^\top(z_1-z_2) \\ &\quad (z_1-z_2)^\top(\hat{B}_{\hat{q}}-B)p(B^\top z_1)p(B^\top z_2)dz_1dz_2 \\ &= \frac{1}{n(n-1)}\frac{1}{h^{q_1}}\{E(u_1^2)\}^2\int\int(K'(u))^2(\hat{B}-B)^\top uu^\top(\hat{B}-B) \\ &\quad p(B^\top z_1)p(B^\top(z_1-hu))dz_1du. \end{aligned}$$

By taking Taylor expansions of $p(B^\top(z_1 - hu))$ around z_1 and using Conditions A4, A7, A8 and A9 in Appendix 3, we have

$$\begin{aligned} E(Q_{12n}^2) &= \frac{1}{n(n-1)}\frac{1}{h^{q_1}}\{E(u_1^2)\}^2\int\int(K'(u))^2(\hat{B}-B)^\top uu^\top(\hat{B}-B) \\ &\quad [p^2(B^\top z_1)+p(B^\top z_1)p'(B^\top z_1)h^pu]dz_1du+o_p(\frac{1}{n(n-1)}) \\ &= \frac{1}{n(n-1)}\frac{1}{h^{q_1}}\{E(u_1^2)\}^2\int\int(K'(u))^2(\hat{B}-B)^\top uu^\top(\hat{B}-B)p^2(B^\top z_1)dz_1du \\ &\quad +\frac{1}{n(n-1)}\frac{1}{h^{q_1}}\{E(u_1^2)\}^2\int\int(K'(u))^2(\hat{B}-B)^\top uu^\top(\hat{B}-B) \\ &\quad p(B^\top z_1)p'(B^\top z_1)h^{q_1}B^\top u)dz_1du+o_p(\frac{1}{n(n-1)})O(\frac{1}{n}) \\ &= O_p(\frac{1}{n^2(n-1)h^{q_1}}) = o_p(\frac{1}{n^2}). \end{aligned}$$

Using Chebyshiev's inequality, we can derive $|Q_{12n}| = o_p(n^{-1}h^{-q_1/2})$. Combined the above results for the terms Q_{11n} and Q_{12n}, it is deduced that $nh^{q_1/2}Q_{1n} \xrightarrow{\text{d}} N(0, s^2)$.

Now we turn to consider the term Q_{2n}. Since

$$\begin{aligned}
Q_{2n} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{h^q}K_{\hat{B}_{\hat{q}}ij}(\hat{g}_i - g_i)(\hat{g}_j - g_j)\frac{\hat{p}_i}{p_i}\frac{\hat{p}_j}{p_j} \\
&\quad + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{h^q}K_{\hat{B}_{\hat{q}}ij}(\hat{g}_i - g_i)(\hat{g}_j - g_j)\left(\frac{\hat{p}_i - p_i}{p_i}\frac{\hat{p}_j - p_j}{p_j} - 2\frac{(\hat{p}_i - p_i)\hat{p}_j}{p_ip_j}\right) \\
&\equiv: \tilde{Q}_{2n} + o_p(\tilde{Q}_{2n}).
\end{aligned}$$

Substituting kernel estimates $\hat{g}$ and $\hat{p}$ into $\tilde{Q}_{2n}$, we have

$$\begin{aligned}
\tilde{Q}_{2n} &= \frac{1}{n^3(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{1}{h^{q_1}h^{2q_1}}\frac{1}{p_ip_j}K_{\hat{B}ij}Q_{\hat{B}_1il}Q_{\hat{B}_1jk} \\
&\quad \times(y_l - g(B_1^\top x_i))(y_k - g(B_1^\top x_j)).
\end{aligned}$$

By an application of the Taylor expansion for $\tilde{Q}_{2n}$ with respect to B and B_1, we can have

$$\tilde{Q}_{2n} \equiv Q_{21n} + Q_{22n},$$

where Q_{21n} and Q_{22n} have following forms:

$$\begin{aligned}
Q_{21n} &= \frac{1}{n^3(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{1}{h^{q_1}h_1^{2q_1}}\frac{1}{p_ip_j}K_{Bij}Q_{B_1il}Q_{B_1jk} \\
&\quad (y_l - g(B_1^\top x_i))(y_k - g(B_1^\top x_j)); \\
Q_{22n} &\equiv (\hat{B}_1 - B_1)^\top Q_{221n} + (\hat{B}_1 - B_1)^\top Q_{222n} + (\hat{B} - B)^\top Q_{223n};
\end{aligned}$$

with Q_{221n}, Q_{222n} and Q_{223n} being the following forms:

$$\begin{aligned}
Q_{221n} &= \frac{1}{n^3(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{1}{h^{q_1}h_1^{3q_1}}\frac{1}{p_ip_j}K_{\tilde{B}ij}Q'_{\tilde{B}_1il}Q_{\tilde{B}_1jk} \\
&\quad (y_l - g(B_1^\top x_i))(y_k - g(B_1^\top x_j))(x_i - x_l), \\
Q_{222n} &= \frac{1}{n^3(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{1}{h^{q_1}h_1^{3q_1}}\frac{1}{p_ip_j}K_{\tilde{B}ij}Q_{\tilde{B}_1il}Q'_{\tilde{B}_1jk} \\
&\quad (y_l - g(B_1^\top x_i))(y_k - g(B_1^\top x_j))(x_j - x_k)
\end{aligned}$$

and

$$
\begin{aligned}
Q_{223n} &= \frac{1}{n^3(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{1}{h^{2q_1}h_1^{2q_1}}\frac{1}{p_ip_j}K'_{\tilde{B}ij}Q_{\tilde{B}_1il}Q_{\tilde{B}_1jk} \\
&\quad (y_l - g(B_1^\top x_i))(y_k - g(B_1^\top x_j))(z_i - z_j).
\end{aligned}
$$

Here $\tilde{B} = \{\tilde{B}_{ij}\}_{d\times q_1}$ with $\tilde{B}_{ij} \in [\min\{\hat{B}_{ij}, B_{ij}\}, \max\{\hat{B}_{ij}, B_{ij}\}]$ and $\tilde{B}_1 = \{\tilde{B}_{1ij}\}_{p_1\times q_1}$ with $\tilde{B}_{1ij} \in [\min\{\hat{B}_{1ij}, B_{1ij}\}, \max\{\hat{B}_{1ij}, B_{1ij}\}]$. As illustrated for the term Q_{12n}, we also assert that replacing $\tilde{B}$ and $\tilde{B}_1$ by B and B_1, respectively, can not influence the converging rate of the term Q_{22n}.

To finished the proof of this theorem, we need to prove that $E(Q_{21n}^2) = o_p(n^{-2}h^{q_1})$. Here we use the similar argument as the proof of Proposition A.1 in Fan and Li (1996). It is obvious that a straightforward calculation of $E(Q_{21n}^2)$ would be very tedious. First, we consider the term $E(Q_{21n})$ by two different cases.

Case I: where i, j, l, k are all different from each other and denote the resulting expression as $\tilde{Q}_{21n}$. Under the assumption that $nh_1^{2\eta}h^{q_1/2} = o_p(1)$, by applying the Lemma B.1, Lemmas 2 and 3 in Robinson (1988), we have

$$
\begin{aligned}
E(\tilde{Q}_{21n}) &= \frac{1}{h^{q_1}h_1^{2q_1}}E\{E_1\{\frac{1}{p_1}Q_{B_112}(g(B_1^\top x_2) - g(B_1^\top x_1))\} \\
&\quad E_3\{\frac{1}{p_3}Q_{B_134}(g(B_1^\top x_4) - g(B_1^\top x_3))\}K_{B13}\}; \\
&\leq C\frac{h_1^{2\eta}}{h^{q_1}}E[D_g(B_1^\top x_1)D_g(B_1^\top x_3)K_{B13}] \\
&= O_p(h_1^{2\eta}) = o_p((nh^{q_1/2})^{-1}).
\end{aligned}
$$

Case II: where i, j, l, k take no more than three different values and denote as $\tilde{Q}'_{21n}$. It is easy to derive that $E(\tilde{Q}'_{21n}) = o_p((nh^{q_1/2})^{-1})$.

Hence, altogether, we have $E(Q_{21n}) = o_p((nh^{q_1/2})^{-1})$.

Now we turn to compute $E(Q^2_{21n})$ as follows:

$$\begin{aligned}
E(Q^2_{21n}) &= E\big\{\frac{1}{n^3(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\sum_{l=1}^{n}\frac{1}{h^{q_1}h_1^{2q_1}}\frac{1}{p_ip_j}K_{Bij}Q_{B_1il}Q_{B_1jk} \\
&\quad [y_l - g(B_1^\top x_i)][y_k - g(B_1^\top x_j)]\big\}^2 \\
&= \frac{1}{n^6(n-1)^2}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\sum_{l=1}^{n}\sum_{i'=1}^{n}\sum_{j'\neq i'}\sum_{k'=1}^{n}\sum_{l'=1}^{n}\frac{1}{h^{2q_1}h_1^{4q_1}} \\
&\quad E\big[\{\frac{1}{p_ip_j}K_{Bij}Q_{B_1il}Q_{B_1jk}[g(B_1^\top x_l) - g(B_1^\top x_i)][g(B_1^\top x_k) - g(B_1^\top x_j)]\} \\
&\quad \{\frac{1}{p_{i'}p_{j'}}K_{Bi'j'}Q_{B_1i'l'}Q_{B_1j'k'}[g(B_1^\top x_{l'}) - g(B_1^\top x_{i'})][g(B_1^\top x_{k'}) - g(B_1^\top x_{j'})]\}\big] \\
&= LA
\end{aligned}$$

When i, j, l, k are all different from i', j', l', k', the two parts in two different braces are independent of each other. Applying the same description as that of $E(Q_{21n}) = o_p((nh^{q_1/2})^{-1})$, we derive $E(Q^2_{21n}) = o_p((n^2h^{q_1})^{-1})$.

Next we consider the case where exactly one index from i, j, l, k equals one of subscripts i', j', l', k'. By symmetry, we only need to compute the case (I) $i = i'$, case (II) $i = l'$, case (III) $l = l'$.

Under case (I), by apply the Lemma B.1, Lemmas 2 and 3 in Robinson (1988), we have

$$\begin{aligned}
LA &= \frac{1}{n^6(n-1)^2h^{2q_1}h_1^{4q_1}}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\sum_{l=1}^{n}E\big[\{\frac{1}{p_ip_j}K_{Bij}Q_{B_1il}Q_{B_1jk} \\
&\quad [g(B_1^\top x_l) - g(B_1^\top x_i)][g(B_1^\top x_k) - g(B_1^\top x_j)]\} \\
&\quad \times\sum_{j'\neq i'}\sum_{k'=1}^{n}\sum_{l'=1}^{n}\{\frac{1}{p_{i'}p_{j'}}K_{Bi'j'}Q_{B_1i'l'}Q_{B_1j'k'} \\
&\quad [g(B_1^\top x_{l'}) - g(B_1^\top x_{i'})][g(B_1^\top x_{k'}) - g(B_1^\top x_{j'})]\}\big] \\
&= \frac{1}{nh^{2q_1}h_1^{4q_1}}E\big\{[E_1(g(B_1^\top x_3) - g(B_1^\top x_1))Q_{B_113}][\frac{1}{p_1p_2}E_2(g(B_1^\top x_4) - g(B_1^\top x_2))Q_{B_124}] \\
&\quad \times\frac{1}{p_1p_5}[E_1(g(B_1^\top x_6) - g(B_1^\top x_1))Q_{B_116}][E_5(g(B_1^\top x_7) - g(B_1^\top x_5))Q_{B_157}]\big\} \\
&\leq \frac{(n-1)^4h_1^{4\eta}}{n^5h^{2q_1}}E\{\frac{1}{p_1p_2p_1p_5}D_g(B_1^\top x_1)D_g(B_1^\top x_2)K_{B12}D_g(B_1^\top x_1)D_g(B_1^\top x_5)K_{B15}\} \\
&= O_p(h_1^{4\eta}n^{-1}) = o_p((n^2h^{q_1})^{-1}).
\end{aligned}$$

Under case (II), we have

$$
\begin{aligned}
LA &= \frac{1}{n^6(n-1)^2h^{2q_1}h_1^{4q_1}}\sum_{i=1}^n\sum_{j\neq i}\sum_{k=1}^n\sum_{l=1}^n E\big[\{\frac{1}{p_ip_j}K_{Bij}Q_{B_1il}Q_{B_1jk} \\
&\quad [g(B_1^\top x_l)-g(B_1^\top x_i)][g(B_1^\top x_k)-g(B_1^\top x_j)]\} \\
&\quad \times\sum_{i'\neq i}\sum_{j'\neq i'}^n\sum_{k'=1}^n\{\frac{1}{p_{i'}p_{j'}}K_{Bi'j'}Q_{B_1i'i}Q_{B_1j'k'} \\
&\quad [g(B_1^\top x_i)-g(B_1^\top x_{i'})][g(B_1^\top x_{k'})-g(B_1^\top x_{j'})]\}\big] \\
&= \frac{1}{nh^{2q_1}h_1^{4q_1}}E\{\frac{1}{p_1p_2}K_{B12}Q_{B_113}Q_{B_124}[g(B_1^\top x_3)-g(B_1^\top x_1)][g(B_1^\top x_4)-g(B_1^\top x_2)] \\
&\quad \frac{1}{p_5p_6}K_{B56}Q_{B_115}Q_{B_167}[g(B_1^\top x_1)-g(B_1^\top x_5)][g(B_1^\top x_7)-g(B_1^\top x_6)]\} \\
&= \frac{1}{nh^{2q_1}h_1^{4q_1}}E\big\{K_{B12}[E_1(g(B_1^\top x_3)-g(B_1^\top x_1))Q_{B_113}][\frac{1}{p_1p_2}E_2(g(B_1^\top x_4)-g(B_1^\top x_2))Q_{B_124}] \\
&\quad \times\frac{1}{p_5p_6}K_{B56}E_1[(g(B_1^\top x_1)-g(B_1^\top x_5))Q_{B_115}][E_6(g(B_1^\top x_7)-g(B_1^\top x_6))Q_{B_167}]\big\} \\
&\leq \frac{(n-1)^4h_1^{4\eta}}{n^5h^{2q_1}}E\{\frac{1}{p_1p_2p_1p_5}D_g(B_1^\top x_1)D_g(B_1^\top x_2)K_{B12}D_g(B_1^\top x_1)D_g(B_1^\top x_6)K_{B56}\} \\
&= O_p(h_1^{3\eta}n^{-1}) = o_p((n^2h^{q_1})^{-1}).
\end{aligned}
$$

Similarly, it is easy to prove that for case (III), $LA = o_p((n^2h^{q_1})^{-1})$. Combining all cases, we get $E(Q_{21n}^2) = o_(n^{-2}h^{q_1})$. Using Chebyshiev's inequality produces $Q_{21n} = o_p(n^{-1}h^{-q_1/2})$.

Following similar arguments as those for the terms Q_{22n}, under Condition A4−A9 in the Appendix 3, Q_{221n}, Q_{222n}, Q_{223n} can be proved to have the following converging rates:

$$
\begin{aligned}
E(Q_{221n}^2) &= O_p(\max\{h_1^{2\eta}, h_1^{2\eta}n^{-1}, h_1^{\eta}n^{-1}\}) = o_p(n^{-1}h^{q_1/2}), \\
E(Q_{222n}^2) &= O_p(\max\{h_1^{2\eta}, h_1^{2\eta}n^{-1}, h_1^{\eta}n^{-1}\}) = o_p(n^{-1}h^{q_1/2}), \\
E(Q_{223n}^2) &= O_p(\max\{h_1^{4\eta}, h_1^{4\eta}n^{-1}, h_1^{3\eta}n^{-1}\}) = o_p(n^{-1}h^{q_1/2}).
\end{aligned}
$$

Here we delete the details. Duo to the facts $||\hat{B}-B|| = O_p(1/\sqrt{n})$ and $||\hat{B}_1-B_1|| = O_p(1/\sqrt{n})$, we derive $Q_{22n} = o_(n^{-1}h^{q_1/2})$ by an application of Chebyshiev's inequality. Therefor, altogether, we have $Q_{2n} = o_p(n^{-1}h^{q_1/2})$.

Lastly, we consider the terms Q_{3n}. Since

$$\begin{aligned} Q_{3n} &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{i\neq j}\frac{1}{h^{q_1}}K_{\hat{B}ij}u_i(\hat{g}_j-g_j)\frac{\hat{p}_j}{p_j} \\ &\quad+\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{i\neq j}\frac{1}{h^{q_1}}K_{\hat{B}ij}u_i(\hat{g}_j-g_j)\left(\frac{\hat{p}_j-p_j}{p_j}\right) \\ &\equiv: \tilde{Q}_{3n}+o_p(\tilde{Q}_{3n}), \end{aligned}$$

substituting the kernel estimates $\hat{g}$ and $\hat{p}$ into $\tilde{Q}_{3n}$, we have

$$\tilde{Q}_{3n} = \frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\frac{1}{h^{q_1}h_1^{q_1}}\frac{1}{p_j}K_{\hat{B}ij}u_iQ_{\hat{B}_1jk}(y_k-g(B_1^\top x_j)).$$

By using the Taylor expansion for $\tilde{Q}_{3n}$ with respect to B and B_1, we can have

$$\tilde{Q}_{3n} \equiv: Q_{31n}+Q_{32n},$$

where Q_{31n} and Q_{32n} have following forms:

$$Q_{31n} = \frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\frac{1}{h^{q_1}h_1^{q_1}}\frac{1}{p_j}K_{Bij}u_iQ_{B_1jk}(y_k-g(B_1^\top x_j))$$

and

$$\begin{aligned} Q_{32n} &= \frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\frac{1}{h^{q_1}h_1^{q_1}}\frac{1}{p_j}K_{\tilde{B}ij}u_iQ'_{\tilde{B}_1jk} \\ &\quad (y_k-g(B_1^\top x_j))(\hat{B}_1-B_1)^\top(x_i-x_l) \\ &\quad+\frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\frac{1}{h^{q_1}h_1^{q_1}}\frac{1}{p_j}K'_{\tilde{B}ij}u_iQ_{\tilde{B}_1jk} \\ &\quad (y_k-g(B_1^\top x_j))(\hat{B}-B)^\top(z_i-z_j) \\ &\equiv (\hat{B}_1-B_1)^\top Q_{321n}+(\hat{B}-B)^\top Q_{322n}; \end{aligned}$$

and where $\tilde{B} = \{\tilde{B}_{ij}\}_{d\times q_1}$ with $\tilde{B}_{ij} \in [\min\{\hat{B}_{ij}, B_{ij}\}, \max\{\hat{B}_{ij}, B_{ij}\}]$ and $\tilde{B}_1 = \{\tilde{B}_{1ij}\}_{p_1\times q_1}$ with $\tilde{B}_{1ij} \in [\min\{\hat{B}_{1ij}, B_{1ij}\}, \max\{\hat{B}_{1ij}, B_{1ij}\}]$. Similarly, replacing $\tilde{B}$ and $\tilde{B}_1$ by B and B_1, respectively, does not influence the convergence rate of the term Q_{32n}.

Because $E(u_i|z_i)=0$, we have $E(Q_{31n})=0$. Then we compute the second order moment of Q_{31n} as follows:

$$
\begin{aligned}
E(Q_{31n}^2) &= E\Big[\frac{1}{n^2(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\sum_{k=1}^{n}\frac{1}{h^{q_1}h_1^{q_1}}\frac{1}{p_j}K_{Bij}u_iQ_{B_1jk}(y_k-g(B_1^\top x_j))\Big]^2 \\
&= E\Big[\frac{1}{n^4(n-1)^2}\frac{1}{h^{2q_1}h_1^{2q_1}}\sum_{i=1}^{n}\sum_{i\neq j}\sum_{k=1}^{n}\sum_{i'=1}^{n}\sum_{i'\neq j'}\sum_{k'=1}^{n}\frac{1}{p_jp_{j'}}K_{Bij}Q_{B_1jk} \\
&\quad K_{Bi'j'}Q_{B_1j'k'}u_iu_{i'}(y_k-g(B_1^\top x_j))(y_k'-g(B_1^\top x_j'))\Big] \\
&= E\Big[\frac{1}{n^4(n-1)^2}\frac{1}{h^{2q_1}h_1^{2q_1}}\sum_{i=1}^{n}\sum_{i\neq j}\sum_{k=1}^{n}\sum_{i'=1}^{n}\sum_{i'\neq j'}\sum_{k'=1}^{n}\frac{1}{p_jp_{j'}}K_{Bij}Q_{B_1jk}K_{Bi'j'}Q_{B_1j'k'} \\
&\quad u_iu_{i'}(g(B_1^\top x_k)-g(B_1^\top x_j))(g(B_1^\top x_{k'})-g(B_1^\top x_j'))\Big]+o_p((n^2h^{q_1})^{-1}).
\end{aligned}
$$

Since $E(u_iu_{i'})\neq 0$ if and only if $i=i'$, we have

$$
\begin{aligned}
E(Q_{31n}^2) &= \frac{1}{n}\frac{1}{h^{2q_1}h_1^{2q_1}}E(u_1^2)E\Big[\frac{1}{p_2p_4}K_{B12}Q_{B_123}K_{B14}Q_{B_145} \\
&\quad (g(B_1^\top x_3)-g(B_1^\top x_2))(g(B_1^\top x_5)-g(B_1^\top x_4))\Big] \\
&= \frac{1}{n}\frac{1}{h^{2q_1}h_1^{2q_1}}E(u_1^2)E\Big[\frac{1}{p_2p_4}K_{B12}K_{B14}E_2[Q_{B_123}(g(B_1^\top x_3)-g(B_1^\top x_2))] \\
&\quad \times E_4[Q_{B_145}(g(B_1^\top x_5)-g(B_1^\top x_4))]\Big] \\
&\leq \frac{1}{n}\frac{h_1^{2\eta}}{h^{2q_1}}E\{\frac{1}{p_2p_4}K_{B12}K_{B14}D_g(B_1^\top x_2)D_g(B_1^\top x_4)\} \\
&= O_p(h_1^{2\eta}n^{-1})=o_p((n^2h^{q_1})^{-1}),
\end{aligned}
$$

by applying the Lemma B.1, Lemmas 2 and 3 in Robinson (1988). Therefore, we have $Q_{31n}=o_p((nh^{q_1/2})^{-1})$.

Following the similar way as the proof of the term Q_{31n}, since $||\hat{B}-B||=O_p(1/\sqrt{n})$ and $||\hat{B}_1-B_1||=O_p(1/\sqrt{n})$, we get $Q_{32n}=o_p((nh^{q_1/2})^{-1})$. Hence, $Q_{3n}=o_p((nh^{q_1/2})^{-1})$.

In summary, we conclude that:

$$
nh^{q_1/2}V_n\xrightarrow{\text{d}}N(0,s^2).
$$

Because s is unknown, it is needed to estimate and it is estimated by the following form:

$$\hat{s}^2 = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i, j=1}^{n} K_h^2 \left(\hat{B}^\top z_i - \hat{B}^\top z_j\right) \hat{u}_i^2 \hat{u}_j^2.$$

Since the proof is rather straightforward, we only present a very brief description. Under the null hypothesis, since the estimators $\hat{B}$ and $\hat{g}$ are consistent to B and g, using some elementary computations results in an asymptotic presentation as:

$$\hat{s}^2 = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i, j=1}^{n} K_h^2 \left(\hat{B}^\top z_i - \hat{B}^\top z_j\right) u_i^2 u_j^2 + o_p(1).$$

Applying a similar way as the justification of Theorem 1 in Guo et al. (2014), one can deriver

$$\begin{aligned} \hat{s}^2 &= \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i, j=1}^{n} K_h^2 \left(B^\top z_i - B^\top z_j\right) u_i^2 u_j^2 + o_p(1) \\ &\equiv: \tilde{s}^2 + o_p(1). \end{aligned}$$

By an application of $U-$statistic properties, it is easy to detain $\tilde{s}^2 \xrightarrow{\text{p}} s^2$. The more details can be referred to Zheng (1996). Hence, we have finished the proof of Theorem 4.1. □

Proof of Lemma 4.1. Here we adopt the similar justification of Lemma 3.1 in Zhu et al. (2015b).

Firstly, we present some details that RERE criterion for DEE_{SIR} process estimating the structure dimension $\hat{q}$. From the justification of Theorem 3.2 in Li et al. (2008), we see that to detain $M_n - M = O_p(C_n)$, it is only needed to prove $M_n(t) - M(t) = O_p(C_n)$ uniformly, where $M(t) = \Sigma^{-1} Var(E(Z|I(Y \leq t)) = \Sigma^{-1}(\nu_1 - \nu_0)(\nu_1 - \nu_0)^\top p_t(1-p_t)$, Σ is the covariance matrix of Z, $\nu_0 = E(Z|I(Y \leq t) = 0)$, $\nu_1 = E(Z|I(Y \leq t) = 1)$ and $p_t = E(I(Y \leq t))$.

We further compute to get

$$\begin{aligned} \nu_1 - \nu_0 &= \frac{E(ZI(Y \leq t))}{p_t} - \frac{E(ZI(Y > t))}{1-p_t} \\ &= \frac{E(ZI(Y \leq t) - E(Z))E(I(Y \leq t))}{p_t(1-p_t)}. \end{aligned}$$

Therefore, the matrix $M(t)$ can also be reformulated as

$$\begin{aligned} M(t) &= \Sigma^{-1}[E\{(Z-E(Z))I(Y\le t)\}][E\{(Z-E(Z))I(Y\le t)\}]^\top \\ &=: \Sigma^{-1}\tilde{m}(t)\tilde{m}(t)^\top, \end{aligned}$$

where $\tilde{m}(t) = E\{(Z-E(Z))I(Y\le t)\}$. Naturally, $\tilde{m}(t)$ can be simply estimated as:

$$\tilde{m}_n(t) = n^{-1}\sum_{i=1}^{n}(z_i-\bar{z})I(y_i\le t).$$

Then $M(t)$ can be estimated as following form:

$$M_n(t) = \hat{\Sigma}^{-1}L_n(t),$$

where $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$, $L_n(t) = \tilde{m}_n(t)\tilde{m}_n(t)^\top$ and $\hat{\Sigma}$ is the estimator of Σ.

However, under the local alternative hypotheses H_{1n}, the response is associated with n, so denote the response under the null and local alternative hypotheses as Y and Y_n, respectively. Under H_{1n}, we have

$$E\{ZI(Y_n\le t)\} - E\{ZI(Y\le t)\} = E[Z\{P(Y_n\le t|Z)\}] - E[Z\{P(Y\le t|Z)\}],$$

where $Y_n = g(B_1^\top X) + C_nG(B^\top Z) + \varepsilon \equiv: Y + C_nG(B^\top Z)$, then for all t, we have

$$\begin{aligned} P(Y_n\le t|Z) - P(Y\le t|Z) &= F_{Y|Z}(t - C_nG(B^\top Z)) - F_{Y|Z}(t) \\ &= -C_nG(B^\top Z)f_{Y|Z}(t) + O_p(C_n). \end{aligned}$$

Thus, under the condition A2, we can conclude that $n^{-1}\sum_{i=1}^{n} z_iI(y_{ni}\le t) - E\{ZI(Y\le t)\} = O_p(\max(C_n, n^{-1/2})) = O_p(C_n)$. By the parallel argument for the justifying Theorem 3.2 of Li et al. (2008), we can deduce that $M_n(t) - M(t) = O_p(C_n)$ uniformly. Thus, we conclude $M_n - M = O_p(C_n)$.

Using the similar statements in Zhu and Fang (1996) and Zhu and Ng (1995), under some certain conditions, we have that $\hat{\lambda}_i - \lambda_i = O_p(C_n)$, where $\hat{\lambda}_d \le \hat{\lambda}_{d-1} \le \cdots \le \hat{\lambda}_1$ are the eigenvalues of the matrix M_n. Thus, we can use the similar statements as those in Proposition 4.2. It is obvious that under H_0, we have

$\lambda_d = \cdots = \lambda_{d-q} = 0$ and $0 < \lambda_q \le \cdots \le \lambda_1$ to be the eigenvalues of the matrix M. Since $cC_n^2 \log n \le c_n \to 0$ with some fixed $c > 0$ and $C_n = 1/(n^{1/2}h^{q_1/4})$, we have $C_n^2 = o_p(c_n)$. It is clear that for any $l > q$, $\lambda_l = 0$, so we get $(\lambda_l^*)^2 = O_p(C_n^2)$. For any $1 \le l \le q$, we have $(\lambda_l^*)^2 = (\tilde{\lambda}_l)^2 + O_p(C_n)$.

Thus, when $l > q$, we have

$$\begin{aligned}
\frac{(\lambda_{q+1}^*)^2 + c_n}{(\lambda_q^*)^2} - \frac{(\lambda_{l+1}^*)^2 + c_n}{(\lambda_l^*)^2} &= \frac{\tilde{\lambda}_{q+1}^2 + c_n + O_p(C_n^2)}{\tilde{\lambda}_q^2 + c_n + O_p(C_n)} - \frac{\tilde{\lambda}_{l+1}^2 + c_n + O_p(C_n^2)}{\tilde{\lambda}_l^2 + c_n + O_p(C_n^2)} \\
&= \frac{\tilde{\lambda}_{q+1}^2 + c_n + o_p(c_n)}{\tilde{\lambda}_q^2 + c_n + O_p(C_n^2)} - \frac{\tilde{\lambda}_{l+1}^2 + c_n + o_p(c_n)}{\tilde{\lambda}_l^2 + c_n + o_p(c_n)} \\
&= \frac{c_n + o_p(c_n)}{\tilde{\lambda}_q^2 + c_n + O_p(C_n^2)} - \frac{c_n + o_p(c_n)}{c_n + o_p(c_n)}.
\end{aligned}$$

Therefor, we have

$$\frac{(\lambda_{q+1}^*)^2 + c_n}{(\lambda_q^*)^2} - \frac{(\lambda_{l+1}^*)^2 + c_n}{(\lambda_l^*)^2} \to -1 < 0.$$

When $1 \le l < q$, we derive that:

$$\begin{aligned}
\frac{(\lambda_{q+1}^*)^2 + c_n}{(\lambda_q^*)^2} - \frac{(\lambda_{l+1}^*)^2 + c_n}{(\lambda_l^*)^2} &= \frac{\tilde{\lambda}_{q+1}^2 + c_n + O_p(C_n^2)}{\tilde{\lambda}_q^2 + c_n + O_p(C_n)} - \frac{\tilde{\lambda}_{l+1}^2 + c_n + O_p(C_n)}{\tilde{\lambda}_l^2 + c_n + O_p(C_n)} \\
&= \frac{c_n + o_p(c_n)}{\tilde{\lambda}_q^2 + c_n + O_p(C_n^2)} - \frac{\tilde{\lambda}_{l+1}^2 + c_n + o_p(c_n)}{\tilde{\lambda}_l^2 + c_n + o_p(c_n)}.
\end{aligned}$$

Then we have

$$\frac{(\lambda_{q+1}^*)^2 + c_n}{(\lambda_q^*)^2} - \frac{(\lambda_{l+1}^*)^2 + c_n}{(\lambda_l^*)^2} \to -\frac{\tilde{\lambda}_{l+1}^2}{\tilde{\lambda}_l^2} < 0.$$

Therefor, altogether, we can deduce that $\hat{q} \to q$.

Proof of Theorem 4.2. Prove the part (I). Since the details of this justification is similar to that of the proof of Theorem 4.1, we will only sketch it. By using $||\hat{B}_1 - B_1|| = O_p(1/\sqrt{n})$ and Conditions A6 and A8 in the Appendix 3, $\hat{g}(\hat{B}_1^\top x)$ is an uniform consistent estimator of $g(B_1^\top x) = E(Y|B_1^\top X = B_1^\top x)$, see Powell et al. (1989) or Robinson (1988). We then have

$$\begin{aligned}
V_n &= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i \ne j} \frac{1}{h^{\hat{q}}} K_{\hat{B}ij} \hat{u}_i \hat{u}_j \\
&= \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i \ne j} \frac{1}{h^{q}} K_{Bij} u_i u_j + o_p(1),
\end{aligned}$$

where $u_i = y_i - g(B_1^\top x_i)$ with $g(B_1^\top x_i) = E(y_i|x_i)$. Let $\triangle(z_i) = m(B^\top z_i) - g(B_1^\top x_i)$. Therefore, by using the $U-$statistics theory, we get that

$$V_n = E\{K_{B12}u_1u_2\} + o_p(1) = E\{\triangle^2(Z)p(B^\top Z)\} + o_p(1).$$

Similarly, we can also prove that in probability $\hat{s}$ converges to some positive value which may be different from s defined by 4.12. Therefore, we can obtain $T_n/(nh^{q_1/2}) \xrightarrow{\text{p}}$ $Constant > 0$ in probability.

Consider Part (II). Following the similar arguments used to prove Theorem 4.1, we can show that:

$$\begin{aligned} V_n &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{i\neq j}\frac{1}{h^{\hat{q}}}K_{\hat{B}ij}\hat{u}_i\hat{u}_j \\ &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{i\neq j}\frac{1}{h^{q}}K_{\tilde{B}ij}u_iu_j + o_p((nh^{q_1})^{-1}) \equiv: Q_n + o_p((nh^{q_1})^{-1}), \end{aligned}$$

where $u_i = y_i - g(B_1^\top x_i) = C_nG(B^\top z_i) + \varepsilon_i$ Then under the local alternative hypotheses, $E(\varepsilon_i|z_i) = 0$. Q_n is further decomposed as:

$$\begin{aligned} Q_n &= \{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{h^{q_1}}K_{\tilde{B}ij}(C_nG(B^\top z_i) + \varepsilon_i)(C_nG(B^\top z_j) + \varepsilon_j)\} \\ &= \{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{h^{q_1}}K_{\tilde{B}ij}\varepsilon_i\varepsilon_j\} + C_n\{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{h^{q_1}}K_{\tilde{B}ij}G(B^\top z_i)\varepsilon_j\} \\ &\quad + C_n^2\{\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}\frac{1}{h^{q_1}}K_{\tilde{B}ij}G(B^\top z_i)G(B^\top z_j)\} \\ &\equiv W_{1n} + C_nW_{2n} + C_n^2W_{3n}, \end{aligned}$$

where $\tilde{B} = (B_1^\top, O_{p_2\times q_1}^\top)^\top$. Again following a similar argument as that for Lemma 3.3 in Zheng (1996), we can easily derive that $nh^{\frac{q_1}{2}}W_{1n} \xrightarrow{\text{d}} N(0, s^2)$. By Lemma 3.1 of Zheng (1996), we get that $\sqrt{n}W_{2n} = O_p(1)$. Since $C_n = n^{-\frac{1}{2}}h^{-\frac{q_1}{4}}$, it is deduced that $nh^{\frac{q_1}{2}}C_nW_{2n} = o_p(1)$. Lastly, we consider the term W_{3n}. It is obvious that the term W_{3n} is $U-$statistic with the kernel as:

$$H(z_i, z_j) = \frac{1}{h^{q_1}}K_{\tilde{B}ij}G(B^\top z_i)G(B^\top z_j).$$

We firstly calculate the expectation of $H(z_i, z_j)$ as

$$\begin{aligned} E\{H(z_i,z_j)\} &= \frac{1}{h^{q_1}} E\{K_{\tilde{B}ij} G(B^\top z_i) G(B^\top z_j)\} \\ &= \frac{1}{h^{q_1}} E\big[K_{\tilde{B}ij} E\{G(B^\top z_i)|\tilde{B}^\top z_i\} E\{G(B^\top z_j)|\tilde{B}^\top z_j\}\big]. \end{aligned}$$

In order to conveniently write, suppose $M(\tilde{B}^\top z_i) = E\{G(B^\top z_i)|\tilde{B}^\top z_i\}$ and $t_i = \tilde{B}^\top z_i$. The expectation of $H(z_i, z_j)$ can be further calculated as

$$\begin{aligned} E\{H(z_i,z_j)\} &= \frac{1}{h^{q_1}} E\big[K(\frac{t_i - t_j}{h}) M(t_i) M(t_j)\big] \\ &= \int\int \frac{1}{h^{q_1}} K(\frac{t_i - t_j}{h}) M(t_i) M(t_j) p_1(t_i) p_1(t_j) dt_i dt_j. \end{aligned}$$

Further, we apply the changing variables $u = \frac{t_i - t_j}{h}$ to get that:

$$\begin{aligned} E\{H(z_i,z_j)\} &= \int\int K(u) M(t_i) M(t_i - hu) p_1(t_i) p_1(t_i - hu) du dt_i \\ &= \int K(u) du \int M^2(t_i) p_1^2(t_i) dt_i + o_p(h) \\ &= E\big([E\{G(B^\top Z)|\tilde{B}^\top Z\}]^2 p_{\tilde{B}}(\tilde{B}^\top Z)\big) + o_p(1) \\ &= E\big([E\{G(B^\top Z)|B_1^\top X\}]^2 p_{B_1}(B_1^\top X)\big) + o_p(1). \end{aligned}$$

Again using the element characteristics of $U-$statistic, we derive that

$$W_{3n} \xrightarrow{\mathrm{p}} E\big([E\{G(B^\top Z)|B_1^\top X\}]^2 p_{B_1}(B_1^\top X)\big).$$

Thus, we can deduce that

$$V_n \xrightarrow{\mathrm{d}} N(E\big([E\{G(B^\top Z)|B_1^\top X\}]^2 p_{B_1}(B_1^\top X)\big), s^2).$$

Additionally, following the similar arguments as the proof of $\hat{s} \xrightarrow{\mathrm{p}} s$ in Theorem 4.1, we get the expected result.

In summary, invoking Slutsky lemma, it is concluded that

$$T_n \xrightarrow{\mathrm{d}} N(E\big([E\{G(B^\top Z)|B_1^\top X\}]^2 p_{B_1}(B_1^\top X)\big)/s, 1).$$

We have completely finished the proof of Theorem 4.2 $\square$.

Chapter 5

Summary

5.1 Conclusion

In this thesis, we have developed several model-adaptive dimension-reduction tests for different testing problems. In the first part of this thesis, we firstly proposed a model-adaptive dimension reduction test procedure based on residual marked empirical process for partially parametric single-index models which are very general parametric models and include generalized linear models and generalized nonlinear models. The resulting test is omnibus adapting the null and alternative models to fully utilize the dimension-reduction structure under the null hypothesis. The comparisons between existing local smoothing tests and global smoothing test suggest: 1). Model-adaption enhances the power performance, and at the same time, the capacity of controlling type I error; 2). Global smoothing-based model-adaptive test outperforms local smoothing-based model-adaptive test. Thus, global smoothing test is worth to be recommended.

It is an important step to check heteroscedasticity in regression analysis. Thus, in the second part of this thesis, we propose a consistent dimension reduction model-adaptive test to check heteroscedasticity. The critical ingredient in the test statistic construction is that the test embeds the dimension reduction structure under the null hypothesis to overcome the curse of dimensionality and adopts to model structure under the alternative hypothesis such that it is still an omnibus test. The test statistic has the limit at the rate as if the number of covariates was the number of linear combinations in the mean regression function. Further, the test statistic is asymptotically normally distributed under the null hypothesis such that critical values are easily determined.

Lastly, as nonparametric techniques need much less restrictive conditions than those required for parametric approaches, in the third part of this thesis, we built a novel dimension reduction adaptive test for the significance of a subset of explanatory variables in the context of a nonparametric regression model. In its construction, only those variables that are significant under the null hypothesis are used such that the curse of dimensionality is largely alleviated when nonparametric estimation is inevitably required. Asymptotic normality is proved under the null hypothesis and power study is conducted under a sequence of local alternative hypotheses and the global alternative hypothesis.

5.2 Future Work

For each kind of problems researched in this thesis, there exists some work needed further to study. For model checking, in chapter 2, our proposed method is readily applied to other models and problems when dimension reduction structure is presented. Secondly, the limiting null distribution of this test is not directly tractable, which is a limitation in real data analysis. The problem how to construct a model-adaptive test which is asymptotically distribution-free, is also ongoing study.

For heteroscedasticity checks, in chapter 3, our proposed model-adaptive method can be extended to other models, for example, partial linear model, partial linear single index model. What's more, this test can be used to to handle other conditional variance models such as single-index and multi-index models. The relevant research is ongoing.

Lastly, our proposed model-adaptive test, in chapter 4, is probably the most widely used method to check different restrictions for nonparametric regression, for example, partially linear model, single-index model and partially linear single-index model.

All these problems are obviously promising and worth studying.

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] Chaudhuri, P., Huang, M. C., Loh, W. and Yao, R. (1994) Piecewise-polynomial regression trees. *Statistica Sinica*, **4**, 143167.

[5] Chiaromonte, F., Cook, R. D. and Li, B. (2002). Sufficient dimension reduction in regressions with categorical predictors. *The Annals of Statistics*. **30**, 475-497.

[6] Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, **89**, 177-189.

[7] Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics.* New York: Wiley.

[8] Cook, R. D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, **104**, 197-208.

[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. (1983). Diagnostics for heteroscedasticity in regression, *Biometrika*, **70**, 1-10.

[12] 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.

[13] Dalalyan, A., Juditsky, A. and Spokoiny, V. (2008). A new algorithm for estimating the effective dimension-reduction subspace. *Journal of Machine Learning Research*, **9**, 1647-1678.

[14] Delgado, M. A., and Manteiga, W. G. (2001). Significance testing in nonparametric regression based on the bootstrap. *Annals of Statistics*, **29**, 1469-1507.

[15] Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variance estimates. *The Annals of Statistics*. **27**, 1012-1050.

[16] 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.

[17] Dette, H. and Marchlewski, M. (2008). A test for the parametric form of the variance function in a partial linear regression model, *Journal of Statistical Planning and Inference*, **138(10)**, 3005-3021.

[18] Dette, H. and Munk, A. (1998). Testing heteroscedasticity in nonparametric regression, *Journal of the Royal Statistical Society: Series B*, **60**, 693-708.

[19] Dette, H., Neumeyer, N. and 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(5)**, 903-917.

[20] Diebolt, J. (1995). A nonparametric test for the regression function: asymptotic theory. *Journal of statistical planning and inference*, **44(1)**, 1-17.

[21] Fan, J. Q., and Hu, T. C. (1992). Bias correction and higher order kernel functions. *Statistics & probability letters*, bf13, 235-243.

[22] Fan, J. Q. and Huang, L. S. (2001). Goodness-of-fit tests for parametric regression models, *Journal of the American Statistical Association*, **96**, 640-652.

[23] Fan, J. and Jiang, J. (2005). Nonparametric inference for additive models. *Journal of the American Statistical Association*, **100**, 890-907.

[24] Fan, J. and Jiang, J. (2007). Nonparametric inference with generalized likelihood ratio tests. *Test*, **16**, 409-444.

[25] Fan, J., Zhang, C. and Zhang, J. (2001) Generalized likelihood ratio statistics and Wilks phenomenon. *The Annals of Statistics*, **29**, 153-193.

[26] Fan, Y. (1994). Testing the goodness of fit of a parametric density function by kernel method. *Econometric Theory*, **10**, 316-316.

[27] Fan, Y. (1998). Goodness-of-fit tests based on kernel density estimators with fixed smoothing parameters. *Econometric Theory*, **14(05)**, 604-621.

[28] Fan, Y. and Li, Q., (1996). Consistent model specication tests: omitted variables and semiparametric functional forms. *Econometrica*, **64**, 865-890.

[29] Fan, Y. and Li, Q. (2000). Consistent Model Specification Tests: Kernel-Based Tests Versus Bierens' ICM Tests. *Econometric Theory*, **16**, 1016-1041.

[30] Feng, Z., Wen, X., Yu Z. and Zhu, L. X. (2013). On Partial Sufficient Dimension Reduction With Applications to Partially Linear Multi-Index Models. *Journal of the American Statistical Association*. **501**, 237-246.

[31] Friendly, M. (2002) Corrgrams: Exploratory displays for correlation matrices. *The American Statistician*, **56**, 316324.

[32] González-Manteiga, W. and Cao, R. (1993) Testing the hypothesis of a general linear model using nonparametric regression estimation. *Test*, **2**, 161-188.

[33] González-Manteiga, W. and Crujeiras, R. M. (2013). An updated review of Goodness-of-Fit tests for regression models. *TEST*, **22**, 361-411.

[34] Guo, X., Wang, T. and Zhu, L. X. (2014). Model checking for generalized linear models: a dimension-reduction model-adaptive approach. `http://arxiv.org/abs/1405.2134`

[35] Hall, P. and Marron, J. S. (1990). On Variance Estimation in Nonparametric Regression, *Biometrika*, **77**, 415-419.

[36] Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *The Annals of Statistics*, **21**, 1926-1947.

[37] Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, **21**, 157-178.

[38] Harrison, D. and Rubinfeld, D. L. (1978). Hedonic prices and the demand for clean air. *Journal of Environmental Economics and Management*, **5**, 81-102.

[39] Hart, J. (1997). *Nonparametric smoothing and lack-of-fit tests.* Springer, Berlin.

[40] Hausman, J., (1978), Specification Tests in Econometrics, *Econometrica*, **46**, 1251-1271.

[41] Hoaglin, D. C. and Velleman, P. F. (1995) A critical look at some analyses of major league baseball salaries. *The American Statistician*, **49**, 277-285.

[42] 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.

[43] Hsing, T., and Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics*, **20**, 1040-1061.

[44] Huskova, M. and Meintanis, S. (2010). Test for the error distribution in nonparametric possibly heterocedastic regression models. *Test*, **19**, 92-112.

[45] Khmadladze, E. V. and Koul, H. L. (2004) Martingale transforms goodness-of-fit tests in regression models. *The Annals of Statistics*, **37**, 995-1034

[46] Koul, H. L. and Ni, P. P. (2004). Minimum distance regression model checking. *Journal of Statistical Planning and Inference*, **119**, 109-141.

[47] Koul, H. L. and Stute, W. (1998) Lack of fit tests in regression with non-random design. *Journal of Applied Statistical Science*, **3**, 5369.

[48] Lavergne, P., and Vuong, Q. (2000). Nonparametric significance testing. *Econometric Theory*, **16(04)**, 576-601.

[49] Lavergne, P., Maistre, S., and Patilea, V. (2014). A significance test for covariates in nonparametric regression. *arXiv preprint arXiv:1403.7063.*

[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, K. C. (1991). Sliced inverse regression for dimension reduction, *Journal of the American Statistical Association*, **86**, 316-327.

[54] Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association*, **87(420)**, 1025-1039.

[55] Li, L., Li, B. and Zhu, L. X. (2010). Groupwise dimension reduction. *Journal of the American Statistical Association*. **105**, 1188-1201.

[56] Li, Q. (1999). Consistent model specification tests for time series econometric models. *Journal of Econometrics*, **92(1)**, 101-147.

[57] Liero, H. (2003). Testing homoscedasticity in nonparametric regression, *Journal of Nonparametric Statistics*, **15(1)**, 31-51.

[58] Lin, J. G. and Qu, X. Y. (2012). A consistent test for heteroscedasticity in semi-parametric regression with nonparametric variance function based on the kernel method, *Statistics*, **46**, 565-576.

[59] Ma, Y. and Zhu, L. (2013). A review on dimension reduction. *International Statistical Review*, **81**, 134-150.

[60] Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability & Its Applications*, **9(1)**, 141-142.

[61] Naik, D. N. and Khattree, R. (1996). Revisiting Olympic track records: Some practical considerations in the principal component analysis. *The American Statistician* **50**, 140-144.

[62] Pollard D (1984). *Convergence of stochastic processes.* Springer, New York.

[63] Pollard, D. (1990). *Empirical Processes: Theory and Applications.* NFS-CBMS Regional Conference Series 2. Institute of Mathematical Statistics, Hayward, CA.

[64] Powell, J. L., Stock, J. H. and Stoker, T. M. (1989) Semiparametric Estimation of Index Coefficients, *Econometrica*, **57**, 1403-1430.

[65] Racine, J. (1997). Consistent Significance Testing for Nonparametric Regression, *Journal of Business & Economic Statistics*, **15**, 369-378.

[66] Racine, J. S., Hart, J., and Li, Q. (2006). Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Reviews*, **25(4)**, 523-544.

[67] Robinson, P. M. (1988). Root-N-consistent semiparametric regression. *Econometrica*, **56**, 931-954.

[68] Simonoff, J. S. and Tsai, C. L. (1994) Improved tests for nonconstant variance in regression based on the modified profile likelihood, *Journal of Applied Statistics*, **43**, 357-370.

[69] Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics.* John Wiley, New York.

[70] Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics*. **25**, 613-641.

[71] Stute, W., Gonzáles-Manteiga, W. and Presedo-Quindimil, M. (1998). Bootstrap approximation in model checks for regression. *Journal of the American Statistical Association*. **93**, 141-149.

[72] Stute, W., Thies, S. and Zhu, L. X. (1998). Model checks for regression: An innovation process approach. *The Annals of Statistics*. **26**, 1916-1934.

[73] Stute, W. and Zhu, L. X. (2002). Model checks for generalized linear models. *Scandinavian Journal of Statistics* . **29**, 535-545.

[74] Stute, W. and Zhu, L. X. (2005). Nonparametric checks for single-index models, *The Annals of Statistics*, **33**, 1048-1083.

[75] Su, J. Q., and 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.

[76] Su, L. and Ullah, A. (2013). A nonparametric goodness-of-fit-based test for conditional heteroskedasticity. *Econometric Theory*, **29(01)**, 187-212.

[77] Tsai, C. L. (1986). Score test for the first-order autoregressive model with heteroscedasticity, *Biometrika*, **73**, 455-460.

[78] 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.

[79] You, J. H. and Chen, G. M. (2005). Testing heteroscedasticity in partially linear regression models, *Statistics and Probability Letters*, **73**, 61-70.

[80] 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.

[81] Wang, J. L., Xue, L., Zhu, L. X. and Chong, Y. S. (2010). Estimation for a partial linear single-index model. *The Annals of Statistics*, **30**, 475-497.

[82] Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A,* 359-372.

[83] Wen, X. and Cook, R. D. (2007). Optimal sufficient dimension reduction in regressions with categorical predictors. *Journal of Statistical Planning and Inference*. **137**, 1961-1978.

[84] Xia, Y. C. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory*, **22**, 1112-1137.

[85] Xia, Q., Xu, W. L. and Zhu. L. X (2014). Consistently determining the number of factors in multivariate volatility modelling, *Statistica Sinica*, accepted.

[86] 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.

[87] Yin, X. and Cook, R. D. (2005). Direction estimation in single-index regressions. *Biometrika*, **92**, 371-384.

[88] 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.

[89] Zhang, C. and Dette, H. (2004). A power comparison between nonparametric regression tests. *Statistics and Probability Letters*, **66**, 289-301.

[90] Zhang, J., Zhu, L. P. and Zhu, L. X. (2012). On a dimension reduction regression with covariate adjustment, *Journal of Multivariate Analysis*, **104**, 39-55.

[91] Zheng, J. X. (1996). A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, **75**, 263-289.

[92] Zheng, J. X. (2009). Testing heteroscedasticity in nonlinear and nonparametric regressions, *Canadian Journal of Statistics*, **37**, 282-300.

[93] Zhu, L. P., Zhu, L.X. , Ferré, L. and Wang, T. (2010a). Sufficient dimension reduction through discretization-expectation estimation. *Biometrika*, **97**, 295-304.

[94] Zhu, L. P., Zhu, L. X. and Feng, Z. H. (2010b). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association* , **105**, 1455-1466.

[95] Zhu, L. X. (2003). Model checking of dimension-reduction type for regression. *Statistica Sinica*, **13**, 283-296.

[96] Zhu, L. X. (2005). Nonparametric Monte Carlo tests and their applications. *Springer, New York*.

[97] Zhu, L. X. and Fang, K. T. (1996). Asymptotics for the kernel estimates of sliced inverse regression. *Annals of Statistics*, **24**, 1053-1067.

[98] 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.

[99] Zhu, L. X. and Ng, K. W. (1995). Asymptotics for sliced inverse regression. *Statistica Sinica*, **5**, 727-736.

[100] Zhu, L. X., Fujikoshi, Y. and Naito, K. (2001). Heteroscedasticity test for regression models. *Sci. China Ser. A* **44**, 1237-1252.

[101] Zhu, L. X. and Neuhaus, G. (2000). Nonparametric Monte Carlo tests for multivariate distributions. *Biometrika*, **87**, 919-928.

[102] Zhu, L. X. and Ng, K. W. (1995). Asymptotics for sliced inverse regression. *Statistica Sinica*, **5**, 727-736.

[103] 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.

[104] Zhu, X. H., Guo, X., Lin, L. and Zhu, L. X. (2015a). Heteroscedasticity Checks for Single Index Models. *Journal of Multivariate Analysis*, **136**, 41-55.

[105] Zhu, X. H., Guo, X. and Zhu, L. X. (2015b). Model checking for partially parametric single-index models: A model-adaptive approach. *Working paper.*

CURRICULUM VITAE

Academic qualifications of the thesis author, Mr. ZHU Xuehu:

- Received the degree of Bachelor of Science in Mathematics from Zhengzhou University, June 2010.

August 2015