An adaptive model checking test for functional linear model

by

Enze Shi

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

in

Statistics

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

Numerous studies have been devoted to the estimation and inference problems for functional linear models (FLM). However, few works focus on model checking problem that ensures the reliability of results. Limited tests in this area do not have tractable null distributions or asymptotic analysis under alternatives. Also, the functional predictor is usually assumed to be fully observed, which is impractical. To address these problems, we propose an adaptive model checking test for FLM. It combines regular moment-based and conditional moment-based tests, and achieves model adaptivity via the dimension of a residual-based subspace. The advantages of our test are manifold. First, it has a tractable chi-squared null distribution and higher powers under the alternatives than its components. Second, asymptotic properties under different underlying models are developed, including the unvisited local alternatives. Third, the test statistic is constructed upon finite grid points, which incorporates the discrete nature of collected data. We develop the desirable relationship between sample size and number of grid points to maintain the asymptotic properties. Besides, we provide a data-driven approach to estimate the dimension leading to model adaptivity, which is promising in sufficient dimension reduction. We conduct comprehensive numerical experiments to demonstrate the advantages the test inherits from its two simple components.

Acknowledgments

First of all, I would like to express my sincere gratitude to my master's supervisors Dr. Linglong Kong and Dr. Lingzhu Li, who patiently provide me with their unreserved help and invaluable guidance in the last two years. I am deeply indebted to their inspirational instructions and tremendous supports during the epidemic. It is my privilege to have worked under the supervision of such two extraordinary academic mentors. They lead by example and make me understand what an outstanding researcher should be like, which strengthens my determination of pursuing a Ph.D. degree.

Moreover, I want to give gratitude to the rest of my committee members, Dr. Adam Kashlak and Dr. Ying Cui for their thorough review, insightful comments, and helpful suggestions.

Furthermore, I also would like to thank all of my friends and colleagues who give me continuous support and encouragement during my study. I really appreciate all of those who have supported and helped me throughout my master's studies.

Last but not the least, I would like to devote my deepest love to my parents. Their encouragement, support and love are always something I can have to fall back on at every stage of my growth.

Table of Contents

1 Introduction **1**

2 Background **5**

2.1 Functional linear model 5

2.1.1 Functional principal component analysis 5

2.1.2 Reproducing kernel Hilbert space 7

2.2 Functional sufficient dimension reduction 9

3 Methodology **11**

3.1 Two-stage slope function estimation 12

3.2 Indicative dimension 14

4 Asymptotic Properties **17**

4.1 Two-stage estimator 18

4.2 Indicative dimension 19

4.3 Test statistics 20

5 Numerical Studies **23**

5.1 Data-driven ridge selection 24

5.2 Simulation results 26

5.2.1 Hybrid effect 28

5.2.2 Number of observations 29

5.3 Real Data Application 31

6 Discussion 34

Bibliography 36

Appendix A: Proof of Lemmas and Theorems 40

List of Tables

5.1 Simulation scenarios and deviations from the null hypothesis 24

5.2 Empirical sizes and powers in percentages for nine scenarios 27

5.3 Empirical sizes and powers for components of the hybrid test. 28

List of Figures

5.1 Empirical sizes (a) and powers (b) versus number of observed points. 30

5.2 Empirical powers versus sample size for different number of observations M in different scenario. Solid line for $M = 32$ and dash line for $M = 16$. 30

5.3 From (a) to (d): Estimated slope functions of gender, age, ADAS11 and ADAS13 for the ADNI DTI data. 32

5.4 (a) Tecator dataset with spectrometric curves; (b) AEMET temperatures for the 73 Spanish weather stations; and (c) The estimated functional coefficient for the AEMET data set by random projection method (black) and RKHS method (red). 33

Chapter 1

Introduction

As an important component of functional data analysis (FDA), FLM is widely adopted in practice to describe the relationship between a functional predictor and a scalar response. It has been actively studied and received increasing attention in recent decades [1–3]. Classical FLM can be formulated as

$$Y = \int_{\mathbb{I}} X(t)\beta(t)dt + \eta, \tag{1.1}$$

where $Y \in \mathbb{R}$ is a scalar response, $X(\cdot) \in L^2(\mathbb{I})$ is a real-valued random process over the interval $\mathbb{I} = [a, b]$, $\beta(\cdot)$ is an unknown slope function in $L^2(\mathbb{I})$, and η is a random noise satisfying $\mathbb{E}(\eta \mid X(\cdot)) = 0$. Without loss of generality, let $\mathbb{I} = [0, 1]$ and assume Y and $X(\cdot)$ are centered. Apart from the scalar-on-function model in (2.1), other forms of FLM include function-on-function regression [4, 5] and function-on-scalar regression [6–8], and also generalized FLM [9, 10].

There are extensive investigations into estimation [11–14] and inference [10, 15–17] problems for FLM. However, most of them assume the model is sufficient. A wrongly specified model could lead to unreliable conclusions, making the model checking procedure an essential step before fitting the data. Even though much attention has been paid to this area, few theoretical results on model checking problems for functional data are developed. Limited tests for FLM include scalar-on-function regression [18] and function-on-function regression [19], both of which are motivated by the residual-marked empirical process proposed in [20]. A recent work [21] considers an efficient empirical process-based test using random projection for scalar-on-function linear regression. It greatly reduces the complexity of calculating the test statistics but at the expense of lower power compared to [18]. Nevertheless, resampling techniques such as wild bootstrap are still required to determine the critical value, which is a computation burden. In addition, the discrete nature of collected data is usually disregarded, which can affect the convergence rate of the test statistic and impair its power. Existing tests considering discretely observed data either lack the theoretical results under local alternatives or cannot provide a reference relationship between the sample size and the number of grid points for the asymptotic properties [13, 19, 22]. The demand for reducing the computation complexity, admitting the discretely observed data and establishing comprehensive theories drives our work.

We propose an adaptive model checking test for FLM and illuminate its asymptotic properties in different underlying models to address the challenges. Our test is motivated by the adaptive-to-model hybrid test proposed in [23]. We are interested in the incomplete nature of collected data. Assume the functional predictor $X(t)$ is observed at M grid points on its support. We then study a desired M to maintain the asymptotic properties rather than assuming $X(t)$ is completely observed as most existing works, which is more practical and realistic. For notation simplicity, denote $X = X(t)$, $\beta = \beta(t)$ and $\int_0^1 X(t)\beta(t)dt$ as $\langle X, \beta\rangle$ without confusion. Our objective is to test the following hypothesis

$$H_0: \quad Y = \langle X, \beta_0\rangle + \eta, \quad \text{for some } \beta_0 \in L^2(\mathbb{I}), \tag{1.2}$$

against

$$H_1: \quad Y = G(X) + \eta, \tag{1.3}$$

where $E(\eta \mid X) = 0$ and $G(X) \neq \langle X, \beta\rangle$ for all $\beta \in L^2(\mathbb{I})$. Our test contains two major components. The first component simply uses a moment-based sum of weighted residuals as a test statistic. It shares asymptotic behaviors with global smoothing methods [24, 25] that can achieve the fastest possible convergence rate [23, 26] while having a tractable null distribution according to the inference results in [10]. The second component adjusts the typical conditional moment-based test proposed by [27] for functional data. It is sensitive to oscillating alternatives and can handling an omnibus test like other local smoothing methods [26, 28]. We can use an indicative dimension induced from a residual-based central mean subspace [29–31] as a bridge to achieve model adaptivity and combine the merits of the two components. The indicative dimension borrows ideas from sufficient dimension reduction (SDR) theory [29–31]. It has been applied to building adaptive-to-model tests [23, 26, 32] which can alleviate the curse of dimensionality. In the past decades, many efforts have been devoted to functional SDR [33–39], paving the way to build the hybrid tests for FLM.

Our main contributions are multifold.

1. The hybrid test has a chi-squared null distribution. Therefore, we do not need a resampling method to obtain the critical value. It reduces the computational burden for functional data analysis. Our test has a fast convergence rate and omnibus property against the alternatives simultaneously, achieved by an adaptive-to-model dimension.

2. We derive the minimum number of grid points to preserve the asymptotic properties of hybrid test by incorporating the discretely observed functional data.

3. We systematically illuminate the asymptotic properties of the hybrid test under the null hypothesis, global alternatives, and local alternatives.

4. We also develop a promising data-driven method for estimating the indicative dimension in practice. The results from various numerical studies show this method is robust to different data generating processes and the underlying models.

The rest of this thesis is organized as follows. In Chapter 2, we give a brief introduction on estimation methods in functional linear regression and functional sufficient dimension reduction. In Chapter 3, we propose the hybrid test statistic for FLM and introduce the estimation procedure for slope function utilizing eigen-system in Hilbert space. The estimation of indicative dimension by SDR in functional space is illustrated as well. Chapter 4 elaborates the asymptotic properties of the test statistic under different hypotheses. In Chapter 5, we present the finite sample powerfulness of the proposed test by various experiments and real data sets.

Chapter 2

Background

2.1 Functional linear model

As has been mentioned in Chapter 1, functional linear model for scalar responses can be formulated as

$$Y = \int_{\mathbb{I}} X(t)\beta_0(t)dt + \eta, \tag{2.1}$$

where η is a random noise satisfying $E(\eta \mid X(\cdot)) = 0$. Furthermore, let $\mathbb{I} = [0,1]$ and assume Y and $X(\cdot)$ are centered. The estimation problem is widely concerned. The main goal is to estimate the slope function β_0 based on a set of training data $\{X_i(t), Y_i\}_{i=1}^n$. There are two popular strategies to retrieve the estimated slope function $\hat{\beta}(t)$. One is based on the functional principal component analysis (FPCA) and the other is to find an optimal solution in the reproducing kernel Hilbert space. The detailed estimation procedure are illustrated as follows.

2.1.1 Functional principal component analysis

First, we formulate the estimation procedure by FPCA. For a square-integrable stochastic process $X(t), t \in [0,1]$, let $\mu(t) = \mathrm{E}(X(t))$ be the expectation function and $C(s,t) = \mathrm{Cov}(X(s), X(t)) = E((X(s) - \mu(s))(X(t) - \mu(t)))$ be the covariance function of $X(t)$. Here $C(s,t)$ is a linear Hilbert-Schmidt operator on $L^2[0,1]$, i.e.

$$C : L^2[0,1] \to L^2[0,1], \quad Cf = \int_0^1 C(s,t)f(s)ds.$$

Notice that $C(s,t)$ is continuous, symmetric, and square integrable, thus admits the spectral decomposition $C(s,t) = \sum_{k=1}^{\infty} \lambda_k \varphi_k(s)\varphi_k(t)$, where $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ are the eigenvalues and $\varphi_1, \varphi_2, \ldots$ are the corresponding orthonormal eigenfunctions of $C(s,t)$. By the Karhunen-Loève theorem, one can express the centered process in the eigenbasis,

$$X(t) - \mu(t) = \sum_{k=1}^{\infty} \xi_k \varphi_k(t)$$

where

$$\xi_k = \int_{\mathcal{T}} (X(t) - \mu(t))\varphi_k(t) dt$$

is the principal component score associated with the k-th eigenfunction φ_k, with the properties

$$\mathrm{E}\left(\xi_k\right) = 0, \mathrm{Var}\left(\xi_k\right) = \lambda_k \text{ and } \mathrm{E}\left(\xi_k \xi_l\right) = 0 \text{ for } k \neq l.$$

A common assumption of FPCA is that $\beta_0(t)$ can be represented by only the first few eigenfunctions. Now we illustrate how to estimate $\beta_0(t)$ based on a set of training data $\{X_i(t), Y_i\}_{i=1}^n$ using a principal components approach.

First, note that in our assumption, both $X(t)$ and Y are centered. Therefore, the mean function $\mu(t) \equiv 0$ and $C(s,t) = E(X(s)X(t))$. Then empirical versions of the covariance function $C(s,t)$ and of its spectral decomposition are

$$\widehat{C}(s,t) \equiv \frac{1}{n} \sum_{i=1}^{n} X_i(s) X_i(t) = \sum_{j=1}^{\infty} \hat{\lambda}_j \hat{\varphi}_j(s) \hat{\varphi}_j(t), \quad s,t \in [0,1].$$

Analogously to the case of $C(s,t)$, $(\hat{\lambda}_j, \hat{\varphi}_j)$ are associated eigenvalue and eigenfunction pairs for the linear operator with kernel $\widehat{C}(s,t)$, ordered such that $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots$. Moreover, $\hat{\lambda}_j = 0$ for $j \geq n+1$. We take $(\hat{\lambda}_j, \hat{\varphi}_j)$ to be our estimator of (λ_j, φ_j). The function $\beta_0(t)$ can be expressed in terms of its Fourier series, as $\beta_0(t) = \sum_{j \geq 1} b_j \varphi_j(t)$, where $b_j = \int_0^1 \beta(t)\varphi_j(t)\mathrm{d}t$. We estimate $\beta_0(t)$ as

$$\hat{\beta}_0(t) = \sum_{j=1}^{m} \hat{b}_j \hat{\varphi}_j(t)$$

where m, lying in the range $1 \leq m \leq n$, denotes a "frequency cut-off" and $\hat{b}_j$ is an estimator of b_j.

To construct $\hat{b}_j$ we note that $b_j = \lambda_j^{-1} g_j$, where g_j denotes the j th Fourier coefficient of $g(t) = \int_0^1 C(s,t)\beta(s)ds$. Notice that we assume the underlying model to be $Y = \int_0^1 X(t)\beta_0(t)dt + \eta$, which can be written as

$$\begin{aligned} X(s)Y = \int_0^1 X(t)X(s)\beta_0(t)dt &\Longleftrightarrow E(X(s)Y) = \int_0^1 E(X(t)X(s))\beta_0(t)dt \\ &\Longleftrightarrow E(X(s)Y) = \int_0^1 C(s,t)\beta_0(t)dt = g(s) \end{aligned}$$

Therefore, a consistent estimator of g is given by its empirical version

$$\hat{g}(t) = \frac{1}{n}\sum_{i=1}^{n} X_i(t)Y_i$$

and so, for $1 \leq j \leq m$, we take $\hat{b}_j = \hat{\lambda}_j^{-1}\hat{g}_j$, where $\hat{g}_j = \int_0^1 \hat{g}(t)\hat{\varphi}_j(t)dt$. Then we can obtain the FPCA-based estimator $\hat{\beta}_0(t)$.

2.1.2 Reproducing kernel Hilbert space

In this subsection, we will illustrate how to estimate the slope function based on reproducing kernel Hilbert space (RKHS). Suppose $X_i(t)$ are fully observed on the interval $[0,1]$. Furthermore, we assume that the slope function $\beta_0(t)$ resides in a RKHS $\mathcal{H}$ which is a subspace of $L^2[0,1]$. The canonical example of $\mathcal{H}$ is the Sobolev spaces with well selected norms, which will be used to illustrate the estimation procedure in the following context. Let $\beta_0 \in \mathcal{H} = H^m[0,1]$, the m-order Sobolev space equipped with the norm $\|\cdot\|_{\mathcal{H}}$ given by

$$\|\beta\|_{\mathcal{H}}^2 = \sum_{i=0}^{m-1}\left(\int_0^1 \beta^{(i)}(t)dt\right)^2 + \int_0^1 \left(\beta^{(m)}(t)\right)^2 dt \tag{2.2}$$

will form a RKHS. The estimator $\hat{\beta}(t)$ are defined to be the solution of a minimization problem over the infinitely dimensional space $\mathcal{H}$. Consider the squared error loss with a penalty term, which associate with the semi-norm of $\mathcal{H}$ that will be used for inducing space decomposition, the regularized estimator of unknown function $\beta(t)$ is given by

$$\hat{\beta}_{n,\lambda} = \underset{\beta \in \mathcal{H}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - \int_0^1 X_i(t)\beta(t)dt \right]^2 + \frac{\lambda}{2} J(\beta, \beta) \right\}, \tag{2.3}$$

where $J(\beta, \widetilde{\beta}) = \int_0^1 \beta^{(m)}(t)\widetilde{\beta}^{(m)}(t)dt$ is a roughness penalty and we use $\lambda/2$ to simplify future expressions. The representation theorem in [40] claims that the minimization problem in (2.3) has an unique solution and an explicit form. We briefly go through this part.

Let the null space induced by the semi-norm $J(\beta, \widetilde{\beta})$ on $\mathcal{H}$ be

$$\mathcal{H}_0 := \{\beta \in \mathcal{H} : J(\beta, \beta) = 0\},$$

which is a finite-dimensional linear subspace of $\mathcal{H}$ (polynomials with degree less than m). Denote by $\mathcal{H}_1$ its orthogonal complement in $\mathcal{H}$ such that $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_1$ also forms a RKHS. Let $K(\cdot, \cdot)$ be the corresponding reproducing kernel of $\mathcal{H}_1$. The representation theorem guarantees that the solution of (2.3) can be expressed as

$$\beta(t) = \sum_{k=0}^{m-1} d_k t^k + \sum_{i=1}^{n} c_i \int_0^1 X_i(s)K(t, s)ds \tag{2.4}$$

for some coefficients $d_0, \ldots, d_{m-1}, c_1, \ldots, c_n$. See [40] for further details. If the norm and semi-norm on $\mathcal{H}$ are chosen as $\|\cdot\|_{\mathcal{H}}$ and $J(\beta, \widetilde{\beta})$ defined above, then reproducing kernel K is

$$K(s, t) = \frac{1}{(m!)^2} B_m(s)B_m(t) - \frac{1}{(2m)!} B_{2m}(|s - t|), \quad s, t \in [0, 1], \tag{2.5}$$

where $B_m(\cdot)$ is the m-th Bernoulli polynomial.

Denote $\Sigma = (\Sigma_{ij})$ as an $n \times n$ matrix and $T = (T_{ij})$ an $n \times m$ with elements

$$\Sigma_{ij} = \int_0^1 \int_0^1 X_i(s)K(t, s)X_j(t)dsdt, \quad T_{ij} = \int_0^1 X_i(t)t^{j-1}dt.$$

Set $\mathbf{y} = (Y_1, \ldots, Y_n)'$. Then we can rewrite the minimization problem (2.3) as

$$\hat{\beta}_{n,\lambda} = \underset{\mathbf{d} \in \mathbb{R}^m, \, \mathbf{c} \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{n} \|\mathbf{y} - (T\mathbf{d} + \Sigma\mathbf{c})\|_{\ell_2}^2 + \lambda \mathbf{c}'\Sigma\mathbf{c} \right\},$$

which is quadratic in $\mathbf{c}$ and $\mathbf{d}$, and the explicit form of the solution can be easily obtained for such a problem. Write $W = \Sigma + n\lambda I$, then the coefficient of $\hat{\beta}_{n,\lambda}$ is given by

$$\begin{aligned}\mathbf{d} &= (T'W^{-1}T)^{-1} T'W^{-1}\mathbf{y}, \\ \mathbf{c} &= W^{-1}\left[I - T\,(T'W^{-1}T)^{-1} T'W^{-1}\right]\mathbf{y}.\end{aligned}$$

2.2 Functional sufficient dimension reduction

Classical sufficient dimension reduction is characterized by the conditional independence between the random variables X and the response Y

$$Y \perp\!\!\!\perp X \mid \beta^\top X,$$

where X is a p-dimensional random vector, Y is a random variable, and β is a $p\times d$ matrix ($d \ll p$). The goal is to estimate the space spanned by the columns of β, which is the linear combinations of X that are sufficient to describe the conditional distribution, see [29, 31]. This problem is linear in the sense that the reduced predictor takes the linear form $\beta^\top X$. For this reason, we refer to it as linear sufficient dimension reduction (linear SDR).

The theory of linear SDR was extended to functional data by [33, 34], where the random element X takes values in the Hilbert space $\mathcal{H}$, whose members are functions defined on an interval, say $[0, 1]$. The general framework for linear functional sufficient dimension reduction are developed as

$$Y \perp\!\!\!\perp X \mid \langle \beta_1, X\rangle_{\mathcal{H}_X}, \ldots, \langle \beta_d, X\rangle_{\mathcal{H}_X}, \tag{2.6}$$

where $\beta_1, \ldots, \beta_d$ are members of $\mathcal{H}_X$. It indicates that the conditional distribution only depends on d projections of the random function X on $\beta_1, \ldots, \beta_d$. We use an example to illustrate it

Example 1 *Let $\mathcal{H}_X = L_2[0, 1]$. For any $f, g \in \mathcal{H}_X$, their inner product is defined as $\langle f, g\rangle_{\mathcal{H}_X} = \int_0^1 f(t)g(t)dt$. Let $\beta_1, \ldots, \beta_d$ are elements in $L_2[0, 1]$. Then the following* *model*

$$Y = f\left(\langle \beta_1, X\rangle_{\mathcal{H}_X}, \ldots, \langle \beta_d, X\rangle_{\mathcal{H}_X}, \epsilon\right),$$

is an example of the model satisfying (2.6)*, where* f *is an unknown nonrandom function and* ϵ *is a random function in* $\mathcal{H}_X$ *which is independent of* X*.*

Chapter 3

Methodology

Let $\{X_i, Y_i\}_{i=1}^n$ be a sequence of independent and identically distributed (i.i.d.) random copies of $\{X, Y\}$. Contrary to many other literatures, we assume the functional predictor $X(t)$ is only observed at M grid points $0 = t_1 < t_2 < \ldots < t_M = 1$ satisfying $\max_{1 \le j \le M-1}\{t_{j+1} - t_j\} \le C_0 M^{-1}$ for some constant C_0. For simplicity, we consider the equal distance observations on $X = X(t)$, that is $t_i = (i-1)/(M-1)$, $i = 1, 2, \ldots, M$. Consider the following hybrid test

$$T_n = \gamma_{n,M} \left| V_0^2 I(\hat{q} = 0) + V_1 I(\hat{q} > 0) \right|, \tag{3.1}$$

where V_0 and V_1 are two simple tests, $\gamma_{n,M}$ is the standardizing factor, and $\hat{q}$ is an estimated indicative dimension, which we will elaborate later.

For the test statistics in (3.1), we suggest the following explicit form. Suppose β_0 be the underlying slope function, which can be consistently estimated by $\hat{\beta}$. Let $\epsilon_i = Y - \langle X_i, \beta_0 \rangle$ and $\hat{\epsilon}_i = Y_i - \langle \widehat{X}_i, \hat{\beta} \rangle$ be the estimation. First, we consider

$$V_0 = \sum_{i=1}^n \hat{\epsilon}_i w(\widehat{X}_i)/n$$

where $\hat{X}_i$ is a consistent estimator of X_i and $w(\cdot)$ is a non-linear weight function. Then we use the conditional moment-based test

$$V_1 = \sum_{i=1}^n \sum_{j \ne i, j=1}^n \hat{\epsilon}_i \hat{\epsilon}_j K_h(\langle \widehat{X}_i - \widehat{X}_j, \widehat{\beta} \rangle)/(n(n-1)),$$

where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a one-dimensional kernel function and h is the bandwidth.

Here we illuminate how the hybrid test works. Notice that V_0 is a weighted sum of residuals, which is the moment based test statistics. It converges faster than V_1 under the null hypothesis, indicating that it requires less samples, but its power is relatively low in some alternatives. On the other hand, the conditional moment based test statistics V_1 enjoys is more powerful than V_0 but converges slower than V_0 under the null hypothesis due to its non-parametric property. Our idea is to combine their advantages using the indicative dimension to obtain a more powerful test.

In the following, we illustrate how to estimate the slope function and indicative dimension.

3.1 Two-stage slope function estimation

There are two strategies to estimate the slope function based on discretely observed functional data. One is to directly use the discrete samples $X_i(t)$ for estimation, see [8, 13, 19, 22], but it can be complicated to obtain inference results. The other is the two-stage estimation procedure [13, 35], which will be adopted in this paper. First, non-parametric methods are used to smooth the observations on each curve. Then the closed form of the reproducing kernel-based regularized estimator $\hat{\beta}$ can be derived by smoothed curves. Specifically, in the first stage, we apply the spline smoothing method to $\{X_i(t_j)\}_{j=1}^M$ in an r-th order Sobolev-Hilbert space to obtain

$$\widehat{X}_i = \underset{g \in H^r[0,1]}{\arg\min} \left\{ \frac{1}{M} \sum_{j=1}^{M} (X_i(t_j) - g(t_j))^2 + \lambda_1 \int_0^1 \left[g^{(r)}(t) \right]^2 dt \right\},$$

where λ_1 is the smoothness parameter, H^r is the r-order Sobolev space defined by

$$\begin{aligned} H^r[0,1] = \big\{ & f : [0,1] \mapsto \mathbb{R} \mid f^{(j)}, j = 0, \ldots, r-1 \text{ are absolutely continuous} \\ & \text{and } f^{(r)} \in L^2[0,1] \big\}, \end{aligned} \tag{3.2}$$

where $f^{(j)}$ denotes the j-th derivative of f. The smoothed curves $\{\widehat{X}_i(t)\}_{i=1}^n$ will be used for reproducing kernel-based estimation in the second stage.

Assume that the slope function β^* resides in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H} = H^m[0,1]$, the m-order Sobolev space defined by (3.2) equipped with the norm $\|\cdot\|_{\mathcal{H}}$ given by $\|\beta\|_{\mathcal{H}} = \sum_{i=0}^{m-1}(\int_0^1 \beta^{(i)}(t)dt)^2 + \int_0^1(\beta^{(m)}(t))^2 dt$. We have the following regularized $\hat{\beta}$:

$$\widehat{\beta}_{n,M,\lambda} = \arg\min_{\beta \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[Y_i - \langle \widehat{X}_i, \beta \rangle \right]^2 + \frac{\lambda}{2} J(\beta, \beta) \right\}, \tag{3.3}$$

where $J(\beta, \widetilde{\beta}) = \int_0^1 \beta^{(m)}(t)\widetilde{\beta}^{(m)}(t)dt$ is a roughness penalty. The solution of (3.3) has closed form, see [3, 40] for detailed derivation.

To make valid statistical inference for $\widehat{\beta}_{n,M,\lambda}$, the eigen-system of $\mathcal{H}$ needs to be established. Both [3] and [10] describe the construction procedure. Here we briefly illustrate it. Let the null space induced by the semi-norm $J(\beta, \widetilde{\beta})$ on $\mathcal{H}$ be $\mathcal{H}_0 := \{\beta \in \mathcal{H} : J(\beta, \beta) = 0\}$, which is a finite-dimensional linear subspace of $\mathcal{H}$. Denote by $\mathcal{H}_1$ its orthogonal complement in $\mathcal{H}$ such that $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_1$ also forms a RKHS. Let $K(s,t)$ be the corresponding reproducing kernel of $\mathcal{H}_1$, and $C(s,t)$ be the covariance function of random variable $X(t)$. Then we apply spectral decomposition on both K and C such that

$$K(s,t) = \sum_{\nu=1}^{\infty} \rho_\nu \psi_\nu(s)\psi_\nu(t), \quad C(s,t) = \sum_{\nu=1}^{\infty} \mu_\nu \phi_\nu(s)\phi_\nu(t),$$

where $\rho_1 \geq \rho_2 \geq \ldots$ are the eigenvalues of $K(s,t)$ and ψ_ν the associated eigenfunctions, $\mu_1 \geq \mu_2 \geq \ldots$ the eigenvalues of $C(s,t)$ and ϕ_ν the associated eigenfunctions. Define the new norm on $\mathcal{H}$ by

$$\|\beta\|_{\widetilde{K}}^2 = \langle C\beta, \beta \rangle + J(\beta, \beta),$$

where $\widetilde{K}$ is the new reproducing kernel on $\mathcal{H}$. Let $\widetilde{K}^{1/2}$ be the square-root kernel of $\widetilde{K}$ and $\Omega(s,t) = (\widetilde{K}^{1/2} C \widetilde{K}^{1/2})(s,t)$ be the product kernel. Conduct the spectral decomposition gives $\Omega(s,t) = \sum_{\nu=1}^{\infty} \widetilde{\rho}_\nu \widetilde{\psi}_\nu(s)\widetilde{\psi}_\nu(t)$. Let $\varphi_\nu^* = \widetilde{\rho}_\nu^{-1/2} \widetilde{K}^{1/2} \widetilde{\psi}_\nu$ and $\rho_\nu^* = \widetilde{\rho}_\nu^{-1} - 1$, then we obtain the eigen-system $(\rho_\nu^*, \varphi_\nu^*)_{\nu=1}^{\infty}$.

Remark 2 *Here we list some properties of eigenvalues shown above. Suppose $C(s,t)$ satisfies Sacks-Ylvisaker conditions of order s, see [41], then $\mu_\nu \asymp \nu^{-2(s+1)}$. Recall* *that m is the order of sobolev space in the estimation procedure and thus $\rho_\nu \asymp \nu^{-2m}$. Notice that the eigen-system we construct satisfies Assumption A3 in [10], which implies $\rho_\nu^* \asymp \nu^{2k}$ and $k = m + s + 1$. These orders will determine the convergence rate of $\hat{\beta}_{n,M,\lambda}$ and the standardizing factor $\gamma_{n,M}$ in test statistics.*

3.2 Indicative dimension

In this subsection, we construct the indicative dimension that integrates the components in hybrid tests. First, we introduce some basic notations and definitions about SDR for scalar-on-function model, see [38] for more details. Consider the random variables $X \in L^2[0,1]$ and $Y \in \mathbb{R}$. If there exists a functional vector $B = (\theta_1(t), \dots, \theta_q(t))^T \in \mathcal{H}^q$, such that

$$Y \perp\!\!\!\perp X \mid \langle B, X \rangle, \quad \text{where } \langle B, X \rangle = (\langle \theta_1, X \rangle, \dots, \langle \theta_q, X \rangle)^T$$

Then the space $\text{Span}\{B\}$ is called a sufficient dimension reduction subspace of Y with respect to X. The intersection of all the dimension reduction subspaces is called the central subspace and denoted as $\mathcal{S}_{Y|X}$. The dimension of the central subspace is denoted as $\dim(\mathcal{S}_{Y|X})$. If $\text{Span}\{B\}$ is the central subspace, then $\dim(\mathcal{S}_{Y|X}) = \dim(\text{Span}\{B\}) = q$. The definition mentioned above is a generalization of SDR for finite-dimensional $X \in \mathbb{R}^p$ [29]. When the conditional independence is replaced by $Y \perp\!\!\!\perp E(Y \mid X) \mid \langle B, X \rangle$, the corresponding subspace $\mathcal{S}_{E(Y|X)}$ is called the central mean subspace with dimension $\dim(\mathcal{S}_{E(Y|X)})$. We consider the central mean subspace in this work.

Recall that $\epsilon_i = Y_i - \langle X_i, \beta^* \rangle$ and $\hat{\epsilon}_i = Y_i - \langle \widehat{X}_i, \widehat{\beta}_{n,M,\lambda} \rangle$ is its estimation. Under the null hypothesis, $\epsilon = \eta$ and then $\dim(\mathcal{S}_{E(\epsilon|X)}) = 0$. Under the alternatives, the remainder $\epsilon = G(X) + \eta - \langle X, \beta_0 \rangle = \Delta(X, \eta)$ and $\dim(\mathcal{S}_{E(\epsilon|X)}) > 0$ since $E\{\Delta(X, \eta) \mid X\}$ is a nonconstant function of X. Let $q = \dim(\mathcal{S}_{E(\epsilon|X)})$ and $\hat{q}$ be its estimation. When $\widehat{\beta}_{n,M,\lambda}$ is a consistent estimator, $\hat{q}$ should also be consistent. Under the null hypothesis, $\hat{q}$ equals to 0 with a probability going to 1; under the alternatives, it converges in probability to a positive q. This expected property will perform as a bridge to combine two simple tests together and get a more powerful hybrid test. To ensure the properties mentioned hold, we require some basic assumptions:

A1: $E(Y^2) < \infty$ and $E(\|X\|_{L^2}^4) < \infty$.

A2: The covariance function $C(s,t)$ of X is continuous on $\mathbb{I} \times \mathbb{I}$. Furthermore, for any $\beta \in L^2(\mathbb{I})$ satisfying $C\beta = 0$, we have $\beta = 0$.

A3: There is a bounded linear operator $P_B(C)$: $\mathcal{H} \mapsto \mathcal{H}$ such that the linearity condition $E(X \mid \langle B, X\rangle) = P_B(C)X$ is satisfied.

A4: $\mathrm{Var}(X \mid \langle B, X\rangle)$ is a constant operator on $\mathcal{H}$.

A1 is commonly required for the consistency and asymptotic properties of $\widehat{\beta}_{n,M,\lambda}$. A2 regularity condition guarantees that $\|\cdot\|_{\tilde{K}}$ is well defined. It also implies that the dimension of a subspace in $\mathcal{H}$ will be preserved after being applied by C. A3 and A4 are usually known as the linearity condition and constant variance condition under the SDR framework, see [33, 34, 38] for more information and see [23] for finite-dimensional case. With these assumptions, We imitate the convex combined matrix proposed in [23] and develop the indicative operator on $\mathcal{H}$ as

$$M^{cc} = E(\epsilon X) \otimes E\left(\epsilon X\right) + HH,$$

where $H = E\left(\epsilon X \otimes X\right)$ and $X(t) \otimes Y(t) = X(s)Y(t)$. The indicative operator can be regarded as a bivariate function defined on $[0,1] \times [0,1]$ and $M^{cc}(s,t) \equiv 0$ under the null hypothesis. Also, $M^{cc}(s,t)$ is continuous, symmetric, and square integrable, thus admits the spectral decomposition $M^{cc}(s,t) = \sum_{\nu=1}^{\infty} \lambda_\nu e_\nu(s) e_\nu(t)$, where λ_ν is its eigenvalue and e_ν is the associated eigenfunction. The lemma below guarantees that $\dim(\mathcal{S}_{E(\epsilon|X)})$ can be obtained from M^{cc}.

Lemma 3 *Let Assumptions A1 through A4 be satisfied, the indicative operator* M^{cc} *satisfies* $\mathrm{Range}(M^{cc}) \subseteq \mathrm{Range}(C\mathcal{S}_{E(\epsilon|X)})$*, where* $\mathrm{Range}(\Gamma) = \{\Gamma\beta : \forall \beta \in \mathcal{H}\}$*. If the number of non-zero eigenvalues of* M^{cc} *is* q*, then we have* $\mathrm{Range}(M^{cc}) = \mathrm{Range}(C\mathcal{S}_{E(\epsilon|X)})$*, indicating* $\dim(\mathcal{S}_{E(\epsilon|X)}) = q$*.*

The indicative operator is estimated by its sample analogue $\widehat{M^{cc}} = \hat{E}(\epsilon X) \otimes \hat{E}(\epsilon X) + \hat{H}\hat{H}$, where $\hat{E}(\epsilon X) = (\sum_{j=1}^{n} \hat{\epsilon}_j \widehat{X}_j)/n$ and $\hat{H} = (\sum_{j=1}^{n} \hat{\epsilon}_j \widehat{X}_j \otimes \widehat{X}_j)/n$. The consistency of $\widehat{\beta}_{n,M,\lambda}$ induces the consistency of the estimated indicative operator, which will be discussed in the next section. As the lemma suggests, the key objective is to determine or estimate the number of non-zero eigenvalues of M^{cc}. The criterion we use is a slight modification of the thresholding double ridge ratio (TDRR) method developed by [42].

Define the eigenvalues of $\widehat{M^{cc}}$ to be $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p \geq \ldots \geq 0$ and let $\hat{s}_j = \hat{\lambda}_j / \left(\hat{\lambda}_j + 1\right)$. Define $\hat{s}_j^* = (\hat{s}_j^2 + c_{1n})/(\hat{s}_{j+1}^2 + c_{1n}) - 1$ and $\hat{r}_j = (\hat{s}_{j+1}^* + c_{2n})/(\hat{s}_j^* + c_{2n})$, where c_{1n} and c_{2n} are the two ridges that converge to 0 in proper rates to be selected later. The criterion of determining indicative dimension can be defined as

$$\hat{q} = \begin{cases} 0, & \text{if } \hat{r}_j > \tau, \quad \forall j \\ \arg\max_j \{j : \hat{r}_j \leq \tau\} & \end{cases}$$

with a threshold $0 < \tau < 1$. Based on the rule of thumb in [42], we also set $\tau = 0.5$.

Remark 4 *According to the above estimation, the standardizing factor $\gamma_{n,M} = n/\sigma_n^2$, where $\sigma_n^2 = \sum_{\nu=1}^{\infty} w_\nu^2/(1 + \lambda\rho_\nu^*)^2$ and $w_\nu = \int_0^1 \widehat{X}_w(t)\varphi_\nu^*(t)dt$. Here $(\rho_\nu^*, \varphi_\nu^*)$ is the eigen-system established on $\mathcal{H}$ and $\widehat{X}_w = (\sum_{i=1}^{n} \widehat{X}_i w(\widehat{X}_i))/n$.*

Remark 5 *With the expected asymptotic properties of $\hat{q}$, T_n reduces to $\gamma_{n,M}V_0^2$ under the null hypothesis, which follows the chi-square distribution χ_1^2. While under the alternatives, T_n jumps to $\gamma_{n,M}|V_1|$. It inherits advantages from the two components. First, it has a tractable null distribution and diverges to infinity faster than using V_1 only. Second, the proposed test can detect local alternatives that converge to the null at a rate slower than $\gamma_{n,M}^{1/2}$. Theorefore, it is more powerful than simply using V_0 or V_1. This is the unique advantage of the adaptive-to-model hybrid test. We will elaborate more details about the asymptotic properties of T_n in the next section.*

Chapter 4

Asymptotic Properties

We now investigate the consistency of $\hat{\beta}_{n,M,\lambda}$ and $\hat{q}$ and provide the asymptotic behaviors of T_n under null and alternatives. Listed below are the assumptions for the theorems.

A5: There exist constants $c_0 \in (0,1)$ and $M_0 > 0$ such that $E(e^{c_0\|X\|_{L^2}}) < \infty$ and for any $\beta \in \mathcal{H}$, $E(|\langle X, \beta\rangle|^4) \leq M_0[E(|\langle X, \beta\rangle|^2)]^2$

A6: Assume $|w_\nu| \asymp 1$, define $M_a = \sum_{\nu=1}^{\infty}(w_\nu^2)/(1+\lambda\rho_\nu^*)^a$ for $a = 1, 2$, then $M_1 \asymp M_2$.

A7: The bandwidth of the kernel h satisfies $h \to 0$, and also $n^{k/(2k+1)}h \to \infty$, $n^{1/(2k+1)}h \to 0$ as $n \to \infty$.

A5 is the regularity condition on process X, which is usually satisfied for Gaussian process; see [3, 10] for details. A6 is required for the asymptotic normality property of our test statistics. This assumption can be easily satisfied according to Proposition 4.2 in [10]. A7 gives the desired order of the bandwidth h for kernel estimation to guarantee the asymptotic properties.

Now we consider the general form of the underlying model:

$$Y = \langle X, \beta\rangle + \delta_n \ell(X) + \eta, \tag{4.1}$$

where $E(\eta \mid X) = 0$ and $\ell(\cdot)$ is a non-linear function. When $\delta_n \equiv 0$, (4.1) refers to models under null hypothesis. It can also represent global and local alternatives when $\delta_n \equiv C \neq 0$ and $\delta_n \to 0$, respectively. The following lemmas and theorems indicate the asymptotic properties of $\hat{\beta}_{n,M,\lambda}$, $\hat{q}$ and T_n with different underlying models.

4.1 Two-stage estimator

Let $\hat{\beta}_{n,\lambda}$ be the regularized estimator based on fully observed functional data

$$\hat{\beta}_{n,\lambda} = \arg\min_{\beta \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} [Y_i - \langle X_i, \beta \rangle]^2 + \frac{\lambda}{2} J(\beta, \beta) \right\}. \tag{4.2}$$

As has been proved in [3], if we take $\lambda = n^{-2k/(2k+1)}$, where k is defined in Remark 2, then $\hat{\beta}_{n,\lambda}$ reaches optimal convergence rate $\|\hat{\beta}_{n,\lambda} - \beta^*\|_{L^2} = O_p(n^{-k/(2k+1)})$, where β^* denotes the limit of $\hat{\beta}$ as n and M goes to infinity. With this property, the consistency of the two-stage estimator under null and global alternatives can be obtained from the theorem below.

Theorem 6 *Let Assumptions A1 through A7 be satisfied, then under null and global alternatives, we have* $\|\hat{\beta}_{n,M,\lambda} - \beta^*\|_{L^2} \leq O_p(n^{-k/(2k+1)}) + O(n^{1/2}M^{-r})$.

The convergence rate of the two-stage estimator consists of two parts. The first term is related to the estimation procedure, which is proved to be optimal under the RKHS framework. The second term reflects the smoothing procedure for discretely observed data. When the number of observations M on a curve becomes large enough, its influence on the estimator will decay quickly. This provides theoretical results for the minimal points we should sample on a curve in practice. We complete this subsection with the following theorem, which indicates the consistency of $\hat{\beta}_{n,M,\lambda}$ under local alternatives.

Theorem 7 *Let Assumptions A1 through A7 be satisfied, under the local alternatives we have*

$$\|\hat{\beta}_{n,M,\lambda} - \beta^*\|_{L^2} \leq O_p(n^{-k/(2k+1)}) + O(n^{1/2}M^{-r}) + O_p(\delta_n)$$

4.2 Indicative dimension

It's easy to derive the convergence rate of estimated indicative operator $\widehat{M^{cc}}$ under null and global alternatives.

Theorem 8 *Let Assumptions A1 through A7 be satisfied, then $\widehat{M^{cc}}$ satisfies:*

(1) under the null hypothesis, $\|\widehat{M^{cc}} - M^{cc}\| = O_p(n^{-2k/(2k+1)}) + O(nM^{-2r})$;

(2) under the global alternatives, $\|\widehat{M^{cc}} - M^{cc}\| = O_p(n^{-k/(2k+1)}) + O(n^{1/2}M^{-r})$.

The convergence rate of estimated indicative operator $\widehat{M^{cc}}$ implies the properties of its eigenvalues. It paves the way to analyze the asymptotic behavior of $\hat{q}$ with TDRR method. The following theorem states the consistency of $\hat{q}$, which is used to indicate the underlying model.

Theorem 9 *Let Assumptions A1 through A7 be satisfied, we have*

(1) if $c_{1n} \to 0, c_{2n} \to 0$ and $c_{1n}c_{2n}/(n^{-2k/(2k+1)} + nM^{-2r})^2 \to \infty$, then under the null hypothesis, $\mathbb{P}(\hat{q} = 0) \to 1$.

(2) if $c_{1n} \to 0, c_{2n} \to 0$ and $c_{1n}c_{2n}/(n^{-2k/(2k+1)} + nM^{-2r}) \to \infty$, then under the alternative hypothesis, $\mathbb{P}(\hat{q} > 0) \to 1$.

From Theorem 9, we can see that $\hat{q}$ indeed has the ability to indicate the type of underlying model. Our hybrid test statistics T_n will degenerate to V_0 and V_1 under null and global alternatives respectively, which means T_n will share their strengths while avoiding their shortcomings. Also we should be aware that the assumptions of two ridges c_{1n} and c_{2n} listed in Theorem 9 is for theoretical analysis. In practice, they will be selected by data-driven approach to be adaptive to different underlying models. We now give the asymptotic behavior of $\hat{q}$ under the local alternatives.

Theorem 10 *Let Assumptions A1 through A7 be satisfied, then under the local alternatives, suppose that $\delta_n = n^{-\alpha}$, then we have*

(1) *when* $\alpha \geq k/(2k+1)$, *let* $c_{1n} \to 0, c_{2n} \to 0$ *and* $c_{1n}c_{2n}/(n^{-2k/(2k+1)} + nM^{-2r})^2 \to \infty$, *then* $\mathbb{P}(\hat{q} = 0) \to 1$.

(2) *when* $0 < \alpha < k/(2k+1)$, *let* $c_{1n} = o\,(\delta_n^4)$, $c_{2n} \to 0$ *and* $c_{1n}c_{2n}/(n^{-8k/(2k+1)} + n^2M^{-4r}) \to \infty$, *then* $\mathbb{P}(\hat{q} = q > 0) \to 1$.

When the order of deviation term δ_n is close to $n^{-k/(2k+1)}$, the estimation $\hat{q}$ will jump from 0 to some positive integer. This property implies that it can detect local alternatives with a deviation term slower than $n^{-k/(2k+1)}$. Unlike the $n^{-1/2}$ threshold shown in [23], there is a shrinkage of critical order of δ_n for functional data, which is the price we have to pay to use reproducing kernel based estimator. However, since $k = m + s + 1$, we can set a larger m to get the critical order closer to $n^{-1/2}$. Finally, we should be aware that these results only have the theoretical meaning, unless we have prior information on the closeness of local alternatives to the null.

4.3 Test statistics

With the results shown above, we now discuss the asymptotic properties of the proposed test in detail under the null, global alternative and local alternative hypothesis.

Theorem 11 *Let Assumptions A1 through A7 be satisfied, if* $Mn^{-1/r} \to \infty$, *then*

(1) *under the null hypothesis,* $T_n \xrightarrow{\mathcal{D}} \chi_1^2$.

(2) *under the global alternative hypothesis,* $T_n/\gamma_{n,M}$ *converges to a constant* $\mu > 0$.

Theorem 11 indicates that the hybrid test statistics T_n will converge in distribution to χ_1^2 under the null hypothesis. Therefore, the critical value of the model checking test can be easily determined without using any resampling techniques, releasing the computation burden. Under the global alternatives, the proposed test will diverge to infinity at order $\gamma_{n,M} = O_p(n^{2k/(2k+1)})$. Suppose all the assumptions mention above hold, we can conclude that the convergence rate of V_0^2 and V_1 under the null hypothesis are $n^{-2k/(2k+1)}$ and $n^{-1}h^{1/2}$, respectively. In simulation studies, the bandwidth h is selected as $O_p(n^{-2/5})$, which is also the choice in [27]. Then the order of V_1 is $n^{-4/5}$ under the null hypothesis. In practice, we usually take the order of Sobolev space $m \geq 2$, and thus $k \geq 3$, which means V_0 converges faster than V_1. Meanwhile, $\gamma_{n,M}V_1$ will diverge faster than $nh^{-1/2}V_1$ under the global alternatives, Therefore, the hybrid test is more powerful than simply use any one of V_0 and V_1. We also obtain the asymptotic distribution of T_n under the local alternatives.

Theorem 12 *Let Assumptions A1 through A7 be satisfied, then under the local alternatives, let $\delta_n = n^{-\alpha}$, if $Mn^{-1/r} \to \infty$, we have*

(1) if $\alpha > k/(2k+1)$, $T_n \xrightarrow{\mathcal{D}} \chi_1^2$.

(2) if $\alpha = k/(2k+1)$, $T_n \xrightarrow{\mathcal{D}} \chi_1^2(\mu_0)$, where $\chi_1^2(\mu_0)$ is a chi-squared distribution with one degree of freedom and noncentrality parameter $\mu_0 \neq 0$.

(3) if $0 < \alpha < k/(2k+1)$, and

(a) $n^{1/2}h^{1/4}\delta_n \to 0$, then $T_n/h^{-1/2}$ converges in distribution to $N(0, \Sigma)$, where

$$\Sigma = 2\int K^2(u)\mathrm{d}u \int \left\{\sigma^2(\mathrm{z})\right\}^2 f^2(\mathrm{z})\mathrm{dz},$$

where $Z = \langle \beta_0, X\rangle, \sigma^2(z) = E\left(\epsilon^2 \mid Z = z\right)$ and $f(z)$ is the probability density function of Z.

(b) $\delta_n = n^{-1/2}h^{-1/4}$, then $T_n/h^{-1/2}$ converges in distribution to $N(E(\ell^2 f), \Sigma)$.

(c) $n^{1/2}h^{1/4}\delta_n \to \infty$, then $T_n/\left(n\delta_n^2\right)$ converges in probability to $E\left(\ell^2 f\right)$.

Theorem 12 states that the hybrid test can detect the alternatives with a deviation term slower than or equal to $n^{-k/(2k+1)}$. When $\delta_n = n^{-k/(2k+1)}$, the estimated indicative dimension $\hat{q}$ will be zero with probability going to 1 according to Theorem 10, reducing T_n to $\gamma_{n,M}V_0^2$. The influence of the deviation term will then be reflected in the asymptotic property of $\gamma_{n,M}^{1/2} V_0$, making it a non-central normal distribution. The third part gives a full picture to show what rate of divergence we can achieve when δ_n is slower than $n^{-k/(2k+1)}$. Until now, we have systematically studied the asymptotic behaviors of T_n with different underlying models. The results demonstrate the expected merits of the hybrid test inherited from moment-based and conditional moment-based tests. Finally, we note that the condition $Mn^{-1/r} \to \infty$ shown in above three theorems implies that if $M = Cn^{1/r}$ for some large enough C, the impact of discrete observations will be eliminated. In practice, we would suggest that $C > 20$.

Chapter 5

Numerical Studies

In this section, we perform simulation studies using nine scenarios defined in [21] and compare their powers. The different data generating processes are encoded as follows. For the k-th simulation scenario S_k, with slope function β_k, the deviation from H_0 is measured by a deviation coefficient δ_d, with $\delta_0 = 0$ and $\delta_d > 0$ for $d = 1, 2$. Then, we denote $H_{k,d}$ the data generation from

$$Y = \langle X, \beta_k \rangle + \delta_d \ell_j(X) + \eta,$$

where $j \in \{1, 2, 3\}$ and the deviations from the linear model are defined by the non-linear terms $\ell_1(X) := \|X\|, \ell_2(X) := 25 \int_0^1 \int_0^1 \sin(2\pi ts)s(1-s)t(1-t)X(s)X(t)\mathrm{d}s\ \mathrm{d}t$, and $\ell_3(X) := \left\langle e^{-X}, X^2 \right\rangle$. The error term η follows a centered normal distribution $\mathcal{N}(0, \sigma^2)$, where σ^2 is chosen such that, under H_0, $R^2 = \mathrm{Var}[\langle \mathrm{X}, \beta \rangle]/(\mathrm{Var}[\langle \mathrm{X}, \beta \rangle] + \sigma^2) = 0.95$. The description of the simulation scenarios is given in Table 5.1. We select five types of functional processes $X(t)$, all of them defined on $[0, 1]$:

BM. Brownian motion, denoted by $\mathbf{B}$, with eigenfunctions $\psi_j(t) := \sqrt{2}\sin((j - 0.5)\pi t)$, $j \geq 1$, will be generated by $X(t) = \sum_{j=1}^{100} Z_j \psi_j(t)/((j - 0.5)\pi)$, where Z_j are i.i.d. standard normal random variables.

HHN. The functional process considered in [35], given by $X(t) = \sum_{j=1}^{20} \xi_j \phi_j(t)$, where $\phi_j(t) := \sqrt{2}\cos(j\pi t)$ and ξ_j are i.i.d. random variables distributed as $\mathcal{N}(0, j^{-2l})$, with $l = 1, 2$.

BB. Brownian bridge, defined as $X(t) = \mathbf{B}(t) - t\mathbf{B}(1)$.

Table 5.1: Simulation scenarios and deviations from the null hypothesis

Scenario	Coefficient $\beta(t)$	Process X	Deviation
S1	$(2\psi_1(t) + 4\psi_2(t) + 5\psi_3(t))/\sqrt{2}$	BM	$\ell_1, \delta = (0, 0.25, 0.75)'$
S2	$(2\tilde{\psi}_1(t) + 4\tilde{\psi}_2(t) + 5\tilde{\psi}_3(t))/\sqrt{2}$	BB	$\ell_2, \delta = (0, -2, -7.5)'$
S3	$(2\psi_2(t) + 4\psi_3(t) + 5\psi_7(t))/\sqrt{2}$	BM	$\ell_1, \delta = (0, -0.2, -0.5)'$
S4	$\sum_{j=1}^{20} 2^{3/2}(-1)^j j^{-2}\phi_j(t)$	HHN$(l = 1)$	$\ell_2, \delta = (0, -1, -3)'$
S5	$\sum_{j=1}^{20} 2^{3/2}(-1)^j j^{-2}\phi_j(t)$	HHN$(l = 2)$	$\ell_2, \delta = (0, -1, -3)'$
S6	$\log(15t^2 + 10) + \cos(4\pi t)$	BM	$\ell_1, \delta = (0, 0.2, 1)'$
S7	$\sin(2\pi t) - \cos(2\pi t)$	OU	$\ell_2, \delta = (0, -0.25, -1)'$
S8	$t - (t - 0.75)^2$	OU	$\ell_3, \delta = (0, -0.01, -0.1)'$
S9	$\pi^2(t^2 - 1/3)$	GBM	$\ell_3, \delta = (0, 0.5, 2.5)'$

OU. Ornstein-Uhlenbeck process, defined as the zero-mean Gaussian process with covariance given by $\mathrm{Cov}[X(s), X(t)] = \sigma^2/(2\alpha)e^{-\alpha(s+t)}(e^{2\alpha \min(s,t)} - 1)$. We consider $\alpha = 1/3, \sigma = 1$, and $X(0) \sim \mathcal{N}(0, \sigma^2/(2\alpha))$. It can be generate by $X(t) = (\sigma/\sqrt{2\alpha})e^{-\alpha t}\mathbf{B}(e^{2\alpha t})$.

GBM. Geometric Brownian motion, defined as $X(t) = s_0 \exp\{(\mu - \sigma^2/2)t + \sigma\mathbf{B}(t)\}$. We consider $\sigma = 1, \mu = 0.5$, and $s_0 = 2$

5.1 Data-driven ridge selection

As an important component of hybrid test, indicative dimension needs to be well estimated so that $\hat{q}$ has the desired properties. The ridges c_{1n} and c_{2n} in TDRR method should be carefully selected to satisfy conditions presented in Theorem 9 and Theorem 10. Existing ridge selection methods [23, 26] usually use pre-determined fix numbers based on numerical experiences, which is not adaptive to different underlying models. For instance, the mean and variance of the nine simulation scenarios we considered vary a lot. The fixed ridges cannot achieve satisfying performance in all scenarios, which urges us to develop a data-driven ridge selection method. In the subsection, we will illustrate the construction of c_{1n} and c_{2n} in a data-driven way.

The art of ridge selection should satisfy two requirements simultaneously, (r1): it should achieve the asymptotic properties stated in Theorem 9 and Theorem 10; (r2): it should be adaptive to different process $X(t)$ and slope function $\beta_0(t)$. To satisfy (r1), one good choice is to set $c_{1n} = c_{2n} = 2\hat{s}_1$, where $\hat{s}_1 = \hat{\lambda}_1/(\hat{\lambda}_1 + 1)$ and $\hat{\lambda}_1$ is the largest eigenvalue of $\widehat{M}_{cc}$ under null hypothesis. This will guarantee $\hat{q} = 0$ under the null hypothesis using TDRR method and $\hat{q} > 0$ under alternatives when the deviation is large enough. However, we have no prior information on our underlying model, so we can't tell whether $\widehat{M}^{cc}$ is estimated from the model under the null hypothesis.

To address this problem, we note that $\widehat{M}^{cc}$ only depends on the smoothed functional data $\widehat{X}_i(t)$ and remainders $\hat{\epsilon}_i$. If we assume our data comes from a FLM with error term η such that $E(\eta \mid X) = 0$ and $\mathrm{Var}(\eta) = \sigma^2$, let

$$\widehat{M}_{null} = \hat{E}(\eta X) \otimes \hat{E}\left(\eta X\right) + \hat{H}\hat{H} \tag{5.1}$$

where $\hat{E}(\eta X) = (\sum_{j=1}^{n} \eta_j \widehat{X}_j)/n$, $\hat{H} = (\sum_{j=1}^{n} \eta_j \widehat{X}_j \otimes \widehat{X}_j)/n$ and η_j are drawn i.i.d. from $N(0, \sigma^2)$. Then the eigenvalues of $\widehat{M}_{null}$ should be close to the eigenvalues of $\widehat{M}_{cc}$ under the null hypothesis. Replicate calculating (5.1) for B times we get $\widehat{M}_{null}^{(1)}, \ldots, \widehat{M}_{null}^{(B)}$, then we average them to mitigate the randomness of sampling η_j by setting $\widehat{M}^* = (\sum_{i=1}^{B} \widehat{M}_{null}^{(i)})/B$. In practice, we take $B = 100$ and σ^2 is estimated by the variance of residuals $\hat{\epsilon}_i$. The ridge will be chosen as $c_{1n} = c_{2n} = 2\hat{s}_1^*$, where $\hat{s}_1^* = \hat{\lambda}_1^*/(\hat{\lambda}_1^*+1)$ and $\hat{\lambda}_1^*$ is the largest eigenvalue of $\widehat{M}^*$. The trick of our approach is to replace the remainders under the null hypothesis by i.i.d. samples from normal distributed variables with the same variance. Although we don't know the exact distribution information of the error term under null hypothesis, numerical experiments shows that the difference of $\hat{s}_1^*$ and $\hat{s}_1$ under the null hypothesis can be ignored since η_j shares the same mean and variance with η. Theoretical analysis and discussion of this data-driven selection procedure need further study.

In practice, we also suggest setting a support order to deal with small variance models. When σ^2, the variance of the error term, is small, the largest eigenvalues $\hat{\lambda}_1^*$ would also be small, which makes the TDRR method extremely sensitive to the deviations. In this case, $\hat{q}$ can easily jump to a positive value even under the null hypothesis. To ensure the robustness of our criteria, it is recommended to set $c_{1n} = c_{2n} = \max\{2\hat{s}_1^*, n^{-0.5}\}$.

5.2 Simulation results

In this subsection, the numerical experiments are conducted on nine scenarios defined above using our hybrid test statistics T_n. The results will be compared to the methods proposed in [18] and [21]. The sample size is chosen to be $n = 100, 250$. Throughout the simulations, the stochastic process $X(t)$ is observed at $M = 30$ equidistant points. The weight function $w(X)$ is chosen as $w(X) = 0.01\|X\|$ and the ridge c_{1n} and c_{2n} are chosen from the data driven way mentioned above. The significance level is $\alpha = 0.05$. The bandwidth h is chosen as optimal value $n^{-0.4}$ according to nonparametric kernel theory. The penalty parameter λ is selected by the generalized cross validation (GCV) illustrated in [3]. We replicate the experiments for 1000 times in each setting to calculate the empirical size or empirical power for our test statistics. The performances of CvM_3, KS_3 and $PCvM$ methods are borrowed from [21]. The results are listed in Table 5.2.

We find that the power of T_n is competitive in all scenarios, especially in S_6 and S_8 when $n = 250$. In these two scenarios, the slope function β_0 is not a linear combination of eigenfunctions of the covariance function $C(s, t)$. Therefore, the functional principal component analysis based estimation methods used in [21] will perform badly. Throughout the nine scenarios, our hybrid test T_n has higher powers than CvM_3 and KS_3 when the deviation term is relatively small. Specifically, T_n enjoys highest rejection rate in $H_{1,1}$, $H_{2,1}$, $H_{4,1}$, $H_{5,1}$ $H_{7,1}$ and $H_{9,1}$ among other methods.

Table 5.2: Empirical sizes and powers in percentages for nine scenarios

	n=100				n=250			
$H_{k,\delta}$	CvM_3	KS_3	$PCvM$	T_n	CvM_3	KS_3	$PCvM$	T_n
$H_{1,0}$	3.9	4.6	4.8	5.4	3.9	4.2	4.9	5.1
$H_{2,0}$	4.6	5.1	3.6	5.7	4.8	5.4	4.7	4.9
$H_{3,0}$	4.9	6.0	5.7	3.8	4.1	4.7	5.3	4.4
$H_{4,0}$	4.4	5.0	4.6	5.4	5.2	5.9	4.9	4.7
$H_{5,0}$	4.0	4.3	4.9	5.6	4.2	4.0	5.0	4.7
$H_{6,0}$	4.3	4.9	5.2	5.3	4.3	5.0	4.8	5.2
$H_{7,0}$	3.9	4.7	5.1	6.1	4.1	4.7	5.2	4.5
$H_{8,0}$	3.5	3.7	4.9	4.6	3.9	4.3	5.1	5.1
$H_{9,0}$	4.8	4.7	6.1	6.0	4.4	4.8	5.9	5.4
$H_{1,1}$	59.4	45.0	69.9	98.5	96.3	90.3	98.4	100
$H_{2,1}$	98.5	95.7	99.2	99.5	100	100	100	100
$H_{3,1}$	97.6	93.0	99.2	98.8	100	100	100	100
$H_{4,1}$	35.7	26.8	43.6	48.9	81.8	67.7	88.6	82.9
$H_{5,1}$	43.1	31.8	49.9	51.3	87.9	75.3	91.5	89.1
$H_{6,1}$	22.2	17.0	27.9	23.7	57.0	43.0	66.9	75.6
$H_{7,1}$	99.9	99.8	99.9	99.9	100	100	100	100
$H_{8,1}$	74.8	50.3	74.7	74.4	88.3	76.0	87.7	97.3
$H_{9,1}$	9.2	8.9	12.1	13.4	17.9	16.4	22.3	20.9
$H_{1,2}$	100	99.9	100	100	100	100	100	100
$H_{2,2}$	99.8	99.5	99.9	100	100	100	100	100
$H_{3,2}$	100	100	100	100	100	100	100	100
$H_{4,2}$	96.4	92.0	98.2	97.8	99.9	99.9	100	99.7
$H_{5,2}$	98.9	97.0	99.1	98.1	100	100	100	99.6
$H_{6,2}$	100	99.8	100	99.9	100	100	100	100
$H_{7,2}$	99.9	99.9	99.9	100	100	100	100	100
$H_{8,2}$	76.5	45.8	78.2	99.6	88.0	73.0	88.9	100
$H_{9,2}$	90.5	85.9	93.9	83.1	100	100	100	96.1

The results indicate that T_n has advantages in detecting local alternatives. For larger deviations, the rejection rate of T_n also shows its superiority to CvM_3 and KS_3, and is comparable to $PCvM$. All results above are evidences of the fact that T_n shares the merits from both V_0 and V_1 and thus become powerful under the local and global alternatives while having a chi-square distribution under the null hypothesis.

5.2.1 Hybrid effect

In this part, we examine the expected properties of the hybrid tests through comparisons of their components. The asymptotic distributions of V_0 and V_1 under the null hypothesis and be easily obtained from the proofs of Theorem 11 and Theorem 12. We will demonstrate that the hybrid test T_n is more powerful than simply use any one of V_0 and V_1. Let Q_0 be the percentage of the indicative dimension that is estimated to be zero, that is $P(\hat{q} = 0)$, which will be used to verify the conclusions in Theorem 9 and illustrate the unique property of T_n.

Consider the model as $Y = \langle X, \beta_0 \rangle + \delta \ell(X) + \eta$, where X is Brownian motion, $\beta_0 = \sum_{j=1}^{20} 4(-1)^j j^{-2} \cos(j\pi t)$, $\ell(X) = 0.25 \sin(\langle X, X \rangle)$ and $\eta \sim \mathcal{N}(0, 0.15^2)$. We set sample size $n = 100$, number of observations $M = 200$, order of Sobolev space $m = 2$ and the shift coefficient δ will change from 0 to 1. The empirical sizes and powers are shown in Tables 5.3.

Table 5.3: Empirical sizes and powers for components of the hybrid test.

δ	CvM_3	$PCvM$	T_n	V_0	V_1	Q_0
0	4.7	3.8	5.2	5.1	2.3	99.6
0.2	5.6	5.8	9.2	7.4	3.4	95.4
0.4	23.1	23.8	67.4	38.6	10.2	36.4
0.6	48.2	46.7	88.6	60.4	25.2	12.6
0.8	68.2	70.5	97.6	79.3	54.7	3.5
1	77.9	85.3	99.4	87.5	76.1	0.7

It manifests that the hybrid test is more powerful than either V_0 or V_1 alone. As the shift term becomes larger, the percentage Q_0 converges to zero as expected, which implies our data driven estimation procedure performs well. Assume the bandwidth h satisfies A7, then the convergence rate of V_0 is faster than V_1. Since the standardizing factor $\gamma_{n,M}$ of T_n is determined by V_0, when the indicative dimension is estimated to be positive, T_n will be larger than $nh^{1/2}V_1$ and therefore more powerful.

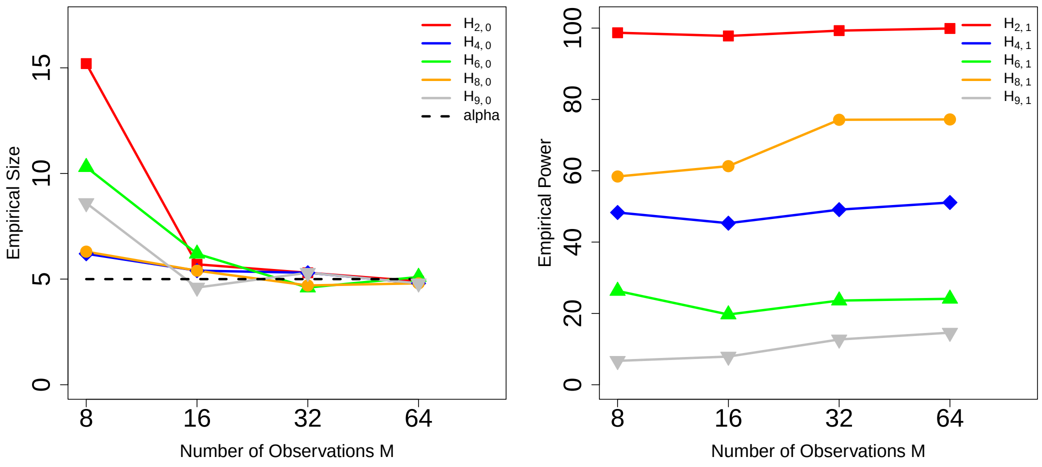
5.2.2 Number of observations

In the subsection, we investigate how the number of observations M on a curve influences the power of T_n. As has been discussed in the asymptotic property part, the error from smoothing is vanishing with a scale of M^{-r} when M is large enough. That is, when M exceeds some threshold, its increase will not help to improve the performance. On the other hand, if we do not sample enough observations on a curve, the power of T_n will not increase as simple size n increases. Our theorems provide guidance to balance these two folds and give an economical and efficient sampling criterion.

To verity these two patterns, we design the experiments as follows. In the first experiment, we set $M = 2^k$, $k = 3, 4, 5, 6$, $n = 100$, $\alpha = 0.05$ and select S_2, S_4, S_6, S_8 and S_9 scenarios (five different stochastic processes $X(t)$). Other parameters are the same as previous subsection. The results are shown in Figure 5.1.

From Figure 5.1, we notice that $M = 8$ is obviously not enough to describe a curve, because we can't control the empirical size to be close to α. This makes the powers under alternatives not reliable. As M increases, the empirical size converges to α and the empirical powers are improved gradually. When M exceeds 32, its effect is not significant any more, indicating that $M = 32$ can capture enough information on a curve with support $[0, 1]$.

In the second experiment, we compare the trends of empirical powers under two settings $M = 16$ and $M = 32$ as sample size n increase from 100 to 300. We still set $\alpha = 0.05$ and select S_2, S_4, S_6, S_8 and S_9 scenarios. The deviation coefficient δ is set



(a) (b)

Figure 5.1: Empirical sizes (a) and powers (b) versus number of observed points.

to be −0.5, −1, 0.2, −0.01 and 2.5 respectively. Other parameters are the same as above. The results are shown in Figure 5.2.

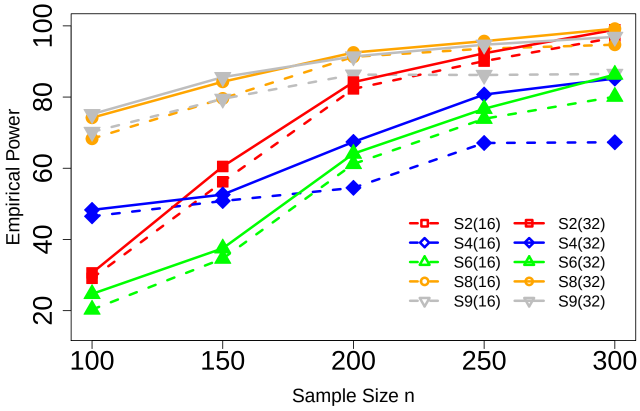


Figure 5.2: Empirical powers versus sample size for different number of observations M in different scenario. Solid line for $M = 32$ and dash line for $M = 16$.

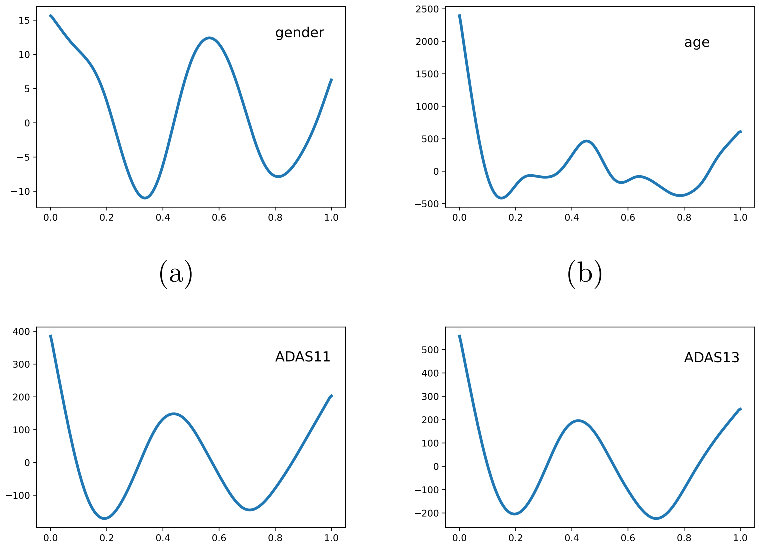
It can be concluded from Figure 5.2 that for S_2, S_6 and S_8, solid lines and dash lines follow nearly the same trends with solid lines a little bit above the dash lines. In these three scenarios, the number of observations $M = 16$ may already be enough to smooth the discrete functional data. If we choose $M = 32$, it is more powerful in detecting alternatives, which coincides with the conclusion in the first experiment. In contrast, for S_4 and S_9, the power will gradually reach a bound and stop increasing even if we set a larger sample size. In this case, the term $M^{-1/2r}$ will dominate the convergence rate of two-stage estimator $\hat{\beta}_{n,M,\lambda}$ no matter how large the sample size n is. It will influence the order of V_0 and V_1 and impair the power of the hybrid test. These two experiments well validate our theoretical analysis for the number of points observed on a curve, which will provide guidance for us to choose an optimal M that balances the sample costs as well as the quality of functional data.

5.3 Real Data Application

We apply the hybrid test to three real datasets to examine whether FLM is sufficient to describe the data. First, we analyze the diffusion tensor imaging (DTI) data in the Alzheimer's Disease Neuroimaging Initiative (ADNI) study, which has also been studied in [8]. In this study, there are 217 subjects in total, with four outliers. The outliers can be identified by the plot and will be eliminated before the test. The fractional anisotropy (FA), a scalar measure of the degree of anisotropy, along the corpus callosum (CC) with 83 equally spaced grid points can be regarded as the discretely observed functional variable $X(t)$ in our model. Other demographic, clinical and genotype variables, such as gender and ADAS11, can be viewed as the scalar response Y in our model. See [8] for a detailed explanation and discussion of the ADNI DTI data. We select gender, age, ADAS11 and ADAS13 as our scalar response candidates and test them separately. The estimated slope functions $\hat{\beta}(t)$ for each response are plotted in Figure 5.3.

The test statistics are 4.46×10^{-5}, 0.0619, 0.0113 and 0.2133 with p-values 0.9946, 0.8032, 0.9157 and 0.8838, respectively. Therefore, at the significance leverl 0.05, we do not reject the null hypothesis that FLM can be used to describe the data for the four responses.

The second and third examples we considered are the same as described in [21].

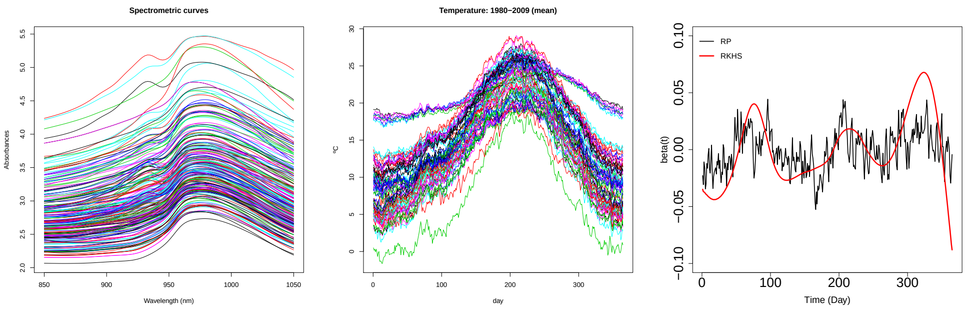


(a) (b)

(c) (d)

Figure 5.3: From (a) to (d): Estimated slope functions of gender, age, ADAS11 and ADAS13 for the ADNI DTI data.

Both datasets are provided in the R package **fda.usc**. Our proposed test returns the same conclusions. The second dataset is the classical Tecator data, a well known example in FDA for nonlinear regression. The data record the absorbance of light at some particular wavelengths by 215 spectrometric curves collected from some chopped meat samples. The fat, water and protein contents of the meat samples are also included in the data set. We test whether these contents can be modeled by FLM using the spectrometric curves. Before the test, six outliers are removed using the Fraiman and Muniz depth. We test the adequacy of FLM for the data with the proposed T_n. The test statistics are $7.977 \times 10^3, 4.998 \times 10^4$ and 9.563×10^3 for fat, water and protein. We reject the null hypothesis as the correspding p-values are all very close to 0. At the significance level 0.05, there is strong statistical evidence against the FLM. The third data we study is the Spanish weather stations data. The data contain yearly profiles of temperature from 73 weather stations of the AEMET (Spanish Meteorological Agency; Spanish acronym) network and other meteorological variables. Our goal is to test whether the mean of the wind speed at each location



(a) (b) (c)

Figure 5.4: (a) Tecator dataset with spectrometric curves; (b) AEMET temperatures for the 73 Spanish weather stations; and (c) The estimated functional coefficient for the AEMET data set by random projection method (black) and RKHS method (red).

can be described by FLM using the average yearly temperature curves. We remove two outliers using the Fraiman and Muniz depth. The test statistics is 1.149 and the p-value is 0.2856. Therefore, at the level $\alpha = 0.05$, there is no evidence against the FLM. The estimated slope functions for the two data sets are plotted in Figure 5.4.

Chapter 6

Discussion

We propose an adaptive hybrid test for FLM. It is a powerful detector under the local and global alternatives with a tractable null distribution. To achieve the adaptivity, we adopt an indicative dimension in an SDR subspace, which combines two simple but less powerful tests. We develop an entirely data-driven method to determine the ridges in the ratio methods like TDRR to estimate the indicative dimension. Besides, we consider the effect of the number of observations on a curve, which provides solid theoretical guarantees for real data applications. Moreover, extensive numerical experiments demonstrate that our test can detect the local alternatives much better than existing methods and is competitive to its competitors under the global alternatives. The proposed test is feasible to real data sets, and its validity is supported by the motivating example data set for nonlinear functional model.

We conclude the paper with several directions for future study. First, although our method is proposed under the assumption of a functional linear model, it is possible to extend the results to other scenarios like a functional partial linear model or a nonlinear functional model. For the other models, there have been works on parameter estimation. However, many of them lack the study of asymptotic properties. This results in challenges in deriving the asymptotic behaviours of the corresponding test statistic. Second, the proposed test can achieve a faster convergence rate under the null if skipping the smoothing procedure for the discretely observed functional data. In the literature, the slope function can be estimated using the discretely observed functional data directly without smoothing them. In this case, the estimator may have a faster convergence rate, leading to a more powerful test under the alternatives. A thorough investigation of estimation and inference problems based on discretely observed functional data is needed to achieve this. Third, the data-driven method proposed for ridge selection is promising as existing approaches tend to underestimate the underlying dimension or rely on some manually determined ridges. An underestimated dimension results in the weaker ability to detect local and global alternatives. TDRR is proposed to solve this issue but still needs to select ridges. Based on our simulation experience, the dimension estimation is related to the data generating process due to the subspace construction. Therefore, to fully use the data structure and reduce the possible power loss caused by an underestimated dimension, we propose the data-driven method.

Bibliography

[1] J. Ramsay and B. Silverman, *Functional Data Analysis*. Springer Series in Statistics, 2005.

[2] F. Ferraty and P. Vieu, *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 2006.

[3] M. Yuan and T. T. Cai, "A reproducing kernel hilbert space approach to functional linear regression," *The Annals of Statistics*, vol. 38, no. 6, pp. 3412–3444, 2010.

[4] F. Yao, H.-G. Müller, and J.-L. Wang, "Functional linear regression analysis for longitudinal data," *The Annals of Statistics*, pp. 2873–2903, 2005.

[5] X. Sun, P. Du, X. Wang, and P. Ma, "Optimal penalized function-on-function regression under a reproducing kernel hilbert space framework," *Journal of the American Statistical Association*, vol. 113, no. 524, pp. 1601–1611, 2018.

[6] P. T. Reiss, L. Huang, and M. Mennes, "Fast function-on-scalar regression with penalized basis expansions," *The international journal of biostatistics*, vol. 6, no. 1, 2010.

[7] J. Goldsmith and T. Kitago, "Assessing systematic effects of stroke on motorcontrol by using hierarchical function-on-scalar regression," *Journal of the Royal Statistical Society. Series C, Applied statistics*, vol. 65, no. 2, p. 215, 2016.

[8] Z. Zhang, X. Wang, L. Kong, and H. Zhu, "High-dimensional spatial quantile function-on-scalar regression," *Journal of the American Statistical Association*, pp. 1–16, 2021.

[9] H.-G. Müller and U. Stadtmüller, "Generalized functional linear models," *the Annals of Statistics*, vol. 33, no. 2, pp. 774–805, 2005.

[10] Z. Shang and G. Cheng, "Nonparametric inference in generalized functional linear models," *The Annals of Statistics*, vol. 43, no. 4, pp. 1742–1773, 2015.

[11] T. T. Cai and P. Hall, "Prediction in functional linear regression," *The Annals of Statistics*, vol. 34, no. 5, pp. 2159–2179, 2006.

[12] T. T. Cai and M. Yuan, "Minimax and adaptive prediction for functional linear regression," *Journal of the American Statistical Association*, vol. 107, no. 499, pp. 1201–1216, 2012.

[13] T. T. Cai and M. Yuan, “Optimal estimation of the mean function based on discretely sampled functional data: Phase transition,” *The annals of statistics*, vol. 39, no. 5, pp. 2330–2355, 2011.

[14] P. Hall and J. L. Horowitz, “Methodology and convergence rates for functional linear regression,” *The Annals of Statistics*, vol. 35, no. 1, pp. 70–91, 2007.

[15] J. Lei, “Adaptive global testing for functional linear models,” *Journal of the American Statistical Association*, vol. 109, no. 506, pp. 624–634, 2014.

[16] H. Cardot, F. Ferraty, A. Mas, and P. Sarda, “Testing hypotheses in the functional linear model,” *Scandinavian Journal of Statistics*, vol. 30, no. 1, pp. 241–255, 2003.

[17] K. Xue and F. Yao, “Hypothesis testing in large-scale functional linear regression,” *Statistica Sinica*, vol. 31, pp. 1101–1123, 2021.

[18] E. García-Portugués, W. González-Manteiga, and M. Febrero-Bande, “A goodness-of-fit test for the functional linear model with scalar response,” *Journal of Computational and Graphical Statistics*, vol. 23, no. 3, pp. 761–778, 2014.

[19] F. Chen, Q. Jiang, Z. Feng, and L. Zhu, “Model checks for functional linear regression models based on projected empirical processes,” *Computational Statistics & Data Analysis*, vol. 144, p. 106 897, 2020.

[20] J. C. Escanciano, “A consistent diagnostic test for regression models using projections,” *Econometric Theory*, vol. 22, no. 6, pp. 1030–1051, 2006.

[21] J. A. Cuesta-Albertos, E. García-Portugués, M. Febrero-Bande, and W. González-Manteiga, “Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes,” *The Annals of Statistics*, vol. 47, no. 1, pp. 439–467, 2019.

[22] H. Zhu, R. Li, and L. Kong, “Multivariate varying coefficient model for functional responses,” *Annals of statistics*, vol. 40, no. 5, p. 2634, 2012.

[23] L. Li, X. Zhu, and L. Zhu, “Adaptive-to-model hybrid of tests for regressions,” *Journal of the American Statistical Association*, pp. 1–10, 2021.

[24] W Stute, W. G. Manteiga, and M. P. Quindimil, “Bootstrap approximations in model checks for regression,” *Journal of the American Statistical Association*, vol. 93, no. 441, pp. 141–149, 1998.

[25] H. Dette, N. Neumeyer, and I. V. Keilegom, “A new test for the parametric form of the variance function in non-parametric regression,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 5, pp. 903–917, 2007.

[26] X. Guo, T. Wang, and L. Zhu, “Model checking for parametric single-index models: A dimension reduction model-adaptive approach,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 78, no. 5, pp. 1013–1035, 2016.

[27] J. X. Zheng, “A consistent test of functional form via nonparametric estimation techniques,” *Journal of Econometrics*, vol. 75, no. 2, pp. 263–289, 1996.

[28] W. Hardle and E. Mammen, “Comparing nonparametric versus parametric regression fits,” *The Annals of Statistics*, pp. 1926–1947, 1993.

[29] R. D. Cook and B. Li, “Dimension reduction for conditional mean in regression,” *The Annals of Statistics*, vol. 30, no. 2, pp. 455–474, 2002.

[30] L. Zhu, T. Wang, L. Zhu, and L. Ferré, “Sufficient dimension reduction through discretization-expectation estimation,” *Biometrika*, vol. 97, no. 2, pp. 295–304, 2010.

[31] B. Li, *Sufficient dimension reduction: Methods and applications with R*. CRC Press, 2018.

[32] F. Tan, X. Zhu, and L. Zhu, “A projection-based adaptive-to-model test for regressions,” *Statistica Sinica*, pp. 157–188, 2018.

[33] L. Ferré and A.-F. Yao, “Functional sliced inverse regression analysis,” *Statistics*, vol. 37, no. 6, pp. 475–488, 2003.

[34] L. Ferré and A.-F. Yao, “Smoothed functional inverse regression,” *Statistica Sinica*, pp. 665–683, 2005.

[35] P. Hall, H.-G. Müller, and J.-L. Wang, “Properties of principal component methods for functional and longitudinal data analysis,” *The annals of statistics*, pp. 1493–1517, 2006.

[36] F. Ferraty, A. Goia, E. Salinelli, and P. Vieu, “Functional projection pursuit regression,” *Test*, vol. 22, no. 2, pp. 293–320, 2013.

[37] H. Lian, “Functional sufficient dimension reduction: Convergence rates and multiple functional case,” *Journal of Statistical Planning and Inference*, vol. 167, pp. 58–68, 2015.

[38] J. Song, “On sufficient dimension reduction for functional data: Inverse moment-based methods,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 11, no. 4, e1459, 2019.

[39] J. Song and B. Li, “Nonlinear and additive principal component analysis for functional data,” *Journal of Multivariate Analysis*, vol. 181, p. 104 675, 2021.

[40] G. Wahba, *Spline models for observational data*. SIAM, 1990.

[41] K. Ritter, G. W. Wasilkowski, and H. Woźniakowski, “Multivariate integration and approximation for random fields satisfying sacks-ylvisaker conditions,” *The Annals of Applied Probability*, pp. 518–540, 1995.

[42] X. Zhu, X. Guo, T. Wang, and L. Zhu, “Dimensionality determination: A thresholding double ridge ratio approach,” *Computational Statistics & Data Analysis*, vol. 146, p. 106 910, 2020.

[43] R. A. DeVore and G. G. Lorentz, *Constructive approximation*. Springer Science & Business Media, 1993, vol. 303.

[44] G. W. Stewart, “Stochastic perturbation theory,” *SIAM review*, vol. 32, no. 4, pp. 579–610, 1990.

Appendix A: Proof of Lemmas and Theorems

Proof of Lemma 3. Suppose $B = (\theta_1(t), \ldots, \theta_q(t))$ is a basis of $S_{E(\epsilon|X)}$, that is $E(\epsilon|X) = E\left(\epsilon|\langle B, X\rangle\right)$, where $\langle B, X\rangle$ is defined as $(\langle\theta_1, X\rangle, \ldots, \langle\theta_q, X\rangle)$.

Under the linearity condition $E\left(X \mid \langle B, X\rangle\right) = P_B(C)X$, where $P_B(C)$ is the constant linear operator on $\mathcal{H}$. Then we have

$$\begin{aligned} E(\epsilon X) &= E(E(\epsilon \mid X)X) = E\left(E\left(\epsilon \mid \langle B, X\rangle\right)X\right) = E\left(E\left(\epsilon \mid \langle B, X\rangle\right)E\left(X|\langle B, X\rangle\right)\right) \\ &= E\left(E\left(\epsilon \mid \langle B, X\rangle\right)P_B(C)X\right) = P_B(C)E\left(E\left(\epsilon \mid \langle B, X\rangle\right)X\right) \\ &= P_B(C)E(\epsilon X). \end{aligned}$$

That is, $E(\epsilon X) \subset \text{Range}(CB)$. For the pHd-matrix, it is sufficient to show that $H \subset \text{Range}(CS_{\epsilon|X})$. Similar to the pervious analysis, we have

$$\begin{aligned} H &= E\left(E(\epsilon \mid X)X \otimes X\right) = E\left(\epsilon E\left(X \otimes X \mid \langle B, X\rangle\right)\right) \\ &= E\left(\epsilon \operatorname{Var}\left(X \mid \langle B, X\rangle\right)\right) + E\left(\epsilon E\left(X \mid \langle B, X\rangle\right)E\left(X \mid \langle B, X\rangle\right)\right) \\ &= E(\epsilon)\operatorname{Var}\left(X \mid \langle B, X\rangle\right) + P_B(C)E\left(\epsilon X \otimes X\right)P_B(C) \\ &= P_B(C)HP_B(C) \subset \text{Range}(CB). \end{aligned}$$

Therefore, $\text{Range}(M^{cc}) \subset \text{Range}(C\mathcal{S}_{E(\epsilon|X)})$. According to assumption A2, we have $\text{Range}(C\mathcal{S}_{E(\epsilon|X)}) = \text{Range}(\mathcal{S}_{E(\epsilon|X)})$. If the number of non-zero eigenvalues of M^{cc} is q, then we can conclude that $\dim(M^{cc}) = q = \dim(\mathcal{S}_{E(\epsilon|X)})$.

■

Proof of Theorem 6. The representation theorem guarantees that the solution of the two stage estimator can be expressed as

$$\hat{\beta}_{n,M,\lambda}(t) = \sum_{k=0}^{m-1} d_k t^k + \sum_{i=1}^{n} c_i \int_0^1 \widehat{X}_i(s)K(t,s)ds$$

for some coefficients $d_0, \dots, d_{m-1}, c_1, \dots, c_n$. Denote $\Sigma = (\Sigma_{ij})$ as an $n \times n$ matrix with element $\Sigma_{ij} = \int_0^1 \int_0^1 \widehat{X}_i(s) K(t,s) \widehat{X}_j(t) dsdt$, and $T = (T_{ij})$ an $n \times m$ with element $T_{ij} = \int_0^1 \widehat{X}_i(t) t^{j-1} dt$. Set $\mathbf{y} = (Y_1, \dots, Y_n)'$. Then we can rewrite the minimization problem as

$$\hat{\beta}_{n,M,\lambda} = \underset{\mathbf{d} \in \mathbb{R}^m,\, \mathbf{c} \in \mathbb{R}^n}{\arg\min} \left\{ \frac{1}{n} \| \mathbf{y} - (T\mathbf{d} + \Sigma \mathbf{c}) \|_{\ell_2}^2 + \lambda \mathbf{c}' \Sigma \mathbf{c} \right\},$$

which is quadratic in $\mathbf{c}$ and $\mathbf{d}$, and the explicit form of the solution can be easily obtained for such a problem. Write $W = \Sigma + n\lambda I$, then the coefficients of $\hat{\beta}_{n,\lambda}$ are given by

$$\mathbf{d} = (T'W^{-1}T)^{-1} T'W^{-1} \mathbf{y},$$
$$\mathbf{c} = W^{-1}[I - T(T'W^{-1}T)^{-1} T'W^{-1}] \mathbf{y}.$$

When we ignore the discrete nature of $X_i(t)$, the slope function can be expressed as

$$\hat{\beta}_{n,\lambda}(t) = \sum_{k=0}^{m-1} d_k^* t^k + \sum_{i=1}^{n} c_i^* \int_0^1 X_i(s) K(t,s) ds.$$

Denote $\Sigma^* = (\Sigma_{ij}^*)$ where $\Sigma_{ij}^* = \int_0^1 \int_0^1 X_i(s) K(t,s) X_j(t) dsdt$, and $T^* = (T_{ij}^*)$ where $T_{ij}^* = \int_0^1 X_i(t) t^{j-1} dt$. Set $\mathbf{y} = (Y_1, \dots, Y_n)'$. Write $W^* = \Sigma^* + n\lambda I$, then the coefficients of $\hat{\beta}_{n,\lambda}$ are given by

$$\mathbf{d}^* = (T^{*\prime} W^{*-1} T^*)^{-1} T^{*\prime} W^{*-1} \mathbf{y},$$
$$\mathbf{c}^* = W^{*-1}[I - T^*(T^{*\prime} W^{*-1} T^*)^{-1} T^{*\prime} W^{*-1}] \mathbf{y}.$$

The distance between the smoothed function $\widehat{X}_i$ and the true X_i has been well studied in the numerical approximation theory. It is well known that $\|\widehat{X}_i - X_i\|_{L^2}^2 \leq O(M^{-2r})$ for $i = 1, 2, \dots, n$, see Chapter 12 Theorem 2.4 in [43] for detailed proof. As has been proved in [3], if we take $\lambda = n^{-2k/(2k+1)}$, then $\hat{\beta}_{n,\lambda}$ satisfies $\|\hat{\beta}_{n,\lambda} - \beta^*\|_{L^2} = O_p(n^{-k/(2k+1)})$. Now we consider the distance between $\hat{\beta}_{n,\lambda}$ and $\hat{\beta}_{n,M,\lambda}$. Since we already have $\|\widehat{X}_i - X_i\|_{L^2}^2 \leq O(M^{-2r})$, then a directly application of Hölder's inequality gives

$$\begin{aligned}
T^*_{ij} - T_{ij} =& \int_0^1 \left(\widehat{X}_i(t) - X_i(t)\right) t^j dt \leq \left(\int_0^1 \left(\widehat{X}_i(t) - X_i(t)\right)^2 dt\right)^{1/2} \left(\int_0^1 t^{2j} dt\right)^{1/2} \\
=& \frac{1}{\sqrt{2j+1}} \|\widehat{X}_i - X_i\|_{L^2} \leq O(M^{-r}) \\
\Sigma^*_{ij} - \Sigma_{ij} =& \int_0^1 \int_0^1 \left(\widehat{X}_i(s)K(s,t)\widehat{X}_j(t) - X_i(s)K(s,t)X_j(t)\right) dtds \\
\leq& \left(\int_0^1 \int_0^1 \left(\widehat{X}_i(s)\widehat{X}_j(t) - X_i(s)X_j(t)\right)^2 dtds\right)^{1/2} \left(\int_0^1 \int_0^1 (K(s,t))^2 dtds\right)^{1/2} \\
\leq& c_0 \left(\int_0^1 \int_0^1 \left((\widehat{X}_i(s) - X_i(s))(\widehat{X}_j(t) + X_j(t))\right)^2 dtds\right)^{1/2} \\
&+ c_0 \left(\int_0^1 \int_0^1 \left(X_i(s)(\widehat{X}_j(t) - X_j(t)) - X_j(t)(\widehat{X}_i(s) - X_i(s))\right)^2 dtds\right)^{1/2} \\
\leq& c_1 \|\widehat{X}_i - X_i\|_{L^2} + c_2 \|\widehat{X}_j - X_j\|_{L^2} + c_3 \|\widehat{X}_i - X_i\|_{L^2} \leq O(M^{-r}).
\end{aligned}$$

Therefore, we can get

$$\Sigma^* = \Sigma + M^{-r} A_1, \quad T^* = T + M^{-r} A_2, \quad W^* = W + M^{-r} A_1,$$

where A_1 and A_2 are two matrices.

According to the matrix perturbation theory, for an arbitrary nonsingular matrix A and a given presumed small matrix E, we have $(A+E)^{-1} = A^{-1} + A^{-1}EA^{-1} + O(\|E\|^2)$, see [44]. Therefore, we can further get

$$W^{*-1} = W^{-1} + M^{-r}(W^{-1}A_1W^{-1}) + O(M^{-2r}A_1^2) = W^{-1} + M^{-r}A_3$$

We can also obtain $(T^{*\prime}W^{*-1}T^*)^{-1} = (T'W^{-1}T)^{-1} + M^{-r}A_4$, and finally we have

$$\mathbf{d}^* = \mathbf{d} + M^{-r}A_5\mathbf{y}, \quad \mathbf{c}^* = \mathbf{c} + M^{-r}A_6\mathbf{y}.$$

Therefore we have

$$\begin{aligned}
\|\hat{\beta}_{n,\lambda} - \hat{\beta}_{n,M,\lambda}\|^2_{L^2} =& \int_0^1 \left(\sum_{k=0}^{m-1}(d^*_k - d_k)t^k + \sum_{i=1}^n \left(c^*_i \int_0^1 X_i(s)K(t,s)ds - c_i \int_0^1 \widehat{X}_i(s)K(t,s)ds\right)\right)^2 dt \\
\leq& \sum_{k=0}^{m-1}(d^*_k - d_k)^2 \int_0^1 t^{2k} dt + \sum_{i=1}^n (c^*_i - c_i)^2 \int_0^1 \left(\int_0^1 (X_i(s) + \widehat{X}_i(s))K(t,s)ds\right)^2 dt \\
&+ \sum_{i=1}^n c_i \int_0^1 \left(\int_0^1 (X_i(s) - \widehat{X}_i(s))K(t,s)ds\right)^2 dt \\
&+ \sum_{i=1}^n (c^*_i - c_i)^2 \int_0^1 \left(\int_0^1 \widehat{X}_i(s)K(t,s)ds\right)^2 dt \\
\leq& mO(M^{-2r}) + nO(M^{-2r}) + nO(M^{-2r}) + nO(M^{-2r}) \leq O(nM^{-2r})
\end{aligned}$$

Applying the triangle inequality gives

$$\|\hat{\beta}_{n,M,\lambda} - \beta^*\|_{L^2} = O_p(n^{-k/(2k+1)}) + O(n^{1/2}M^{-r}).$$

■

Proof of Theorem 7.

Let $\mathbf{y}$ and $\tilde{\mathbf{y}}$ be the response under the null hypothesis and local alternatives respectively, then it is obvious that $\tilde{\mathbf{y}} - \mathbf{y} = O(\delta_n)$. Under local alternatives, the coefficients can be obtained from

$$\begin{aligned}\tilde{\mathbf{d}} &= (T'W^{-1}T)^{-1}T'W^{-1}\tilde{\mathbf{y}},\\ \tilde{\mathbf{c}} &= W^{-1}[I - T(T'W^{-1}T)^{-1}T'W^{-1}]\tilde{\mathbf{y}}.\end{aligned}$$

Therefore we have $\tilde{\mathbf{d}} - \mathbf{d} = O(\delta_n)$ and $\tilde{\mathbf{c}} - \mathbf{c} = O(\delta_n)$, which implies

$$\|\hat{\beta}_{n,M,\lambda} - \beta^*\|_{L^2} = O_p(n^{-k/(2k+1)}) + O(n^{1/2}M^{-r}) + O(\delta_n)$$

under the local alternatives. ■

Proof of Theorem 8. For notation simplicity, let $a_n = n^{-k/(2k+1)}$ and $\hat{\beta} = \hat{\beta}_{n,M,\lambda}$ First we have

$$\begin{aligned}\hat{H} &= \frac{1}{n}\sum_{j=1}^{n} \hat{\epsilon}_j \widehat{X}_j \otimes \widehat{X}_j = \frac{1}{n}\sum_{j=1}^{n} (\epsilon_j + \langle \widehat{X}_j, \hat{\beta}\rangle - \langle X_j, \beta_0\rangle)\widehat{X}_j \otimes \widehat{X}_j\\ &= \frac{1}{n}\sum_{j=1}^{n} \epsilon_j X_j \otimes X_j + \frac{1}{n}\sum_{j=1}^{n}\langle \hat{\beta} - \beta_0, X_j\rangle X_j \otimes X_j + O_p(M^{-r})\\ &= E(\epsilon X \otimes X) + O_p(n^{-1/2}) + O_p(n^{1/2}M^{-r}) + O_p(a_n)\\ &= H + O_p(a_n + n^{1/2}M^{-r}).\end{aligned}$$

Then under the null and the alternatives. We have

$$\begin{aligned}\hat{H}\hat{H} &= (H + O_p(a_n + n^{1/2}M^{-r}))(H + O_p(a_n + n^{1/2}M^{-r}))\\ &= HH + HO_p(a_n + n^{1/2}M^{-r}) + O_p(a_n^2 + nM^{-2r}).\end{aligned}$$

Similarly, we have

$$\begin{aligned}\hat{E}(\epsilon X) &= \frac{1}{n}\sum_{j=1}^{n} \hat{\epsilon}_j \widehat{X}_j = \frac{1}{n}\sum_{j=1}^{n} (\epsilon_j + \langle \widehat{X}_j, \hat{\beta}\rangle - \langle X_j, \beta_0\rangle)\widehat{X}_j\\ &= \frac{1}{n}\sum_{j=1}^{n} \epsilon_j X_j + \frac{1}{n}\sum_{j=1}^{n}\langle \hat{\beta} - \beta_0, X_j\rangle X_j + O_p(n^{1/2}M^{-r})\\ &= E(\epsilon X) + O_p(a_n + n^{1/2}M^{-r}).\end{aligned}$$

Under the null hypothesis, $H = E\left(\epsilon X \otimes X\right) = 0$ and $E(\epsilon X) = 0$, therefore $\hat{M}^{cc} = O_p(a_n^2 + nM^{-2r})$. Under the alternatives, we have $H \neq 0$ and $\hat{M}^{cc} = M + O_p(a_n + n^{1/2}M^{-r})$. Altogether, the proof is finished. ■

Proof of Theorem 9. For notation simplicity, let $a_n = n^{-k/(2k+1)}$. Under the null hypothesis, $\hat{M}^{cc} = O_p(a_n^2 + nM^{-2r})$. All eigenvalues of the target operator M^{cc} are $\lambda_1 = \cdots = \lambda_n = \ldots = 0$. Thus the eigenvalues of $\hat{M}^{cc}$, denoted as $\hat{\lambda}_j, j = 1, 2, \ldots$, satisfy $\hat{\lambda}_j = O_p(a_n^2 + M^{-2r})$. At the population level,

$$s_j^* = 0 \text{ and } r_j = 1, \text{ for } \forall j$$

At the sample level, $\hat{s}_j = \frac{\hat{\lambda}_j}{\hat{\lambda}_{j+1}} = O_p(a_n^2 + nM^{-2r})$,

$$\hat{s}_j^* = \frac{\hat{s}_j^2 + c_{1n}}{\hat{s}_{j+1}^2 + c_{1n}} - 1 = \frac{\hat{s}_j^2 - \hat{s}_{j+1}^2}{\hat{s}_{j+1}^2 + c_{1n}} = \frac{O_p((a_n^2 + nM^{-2r})^2)}{O_p((a_n^2 + nM^{-2r})^2 + c_{1n}} \to 0$$

if $c_{1n} \to 0$ and $c_{1n}/(a_n^2 + nM^{-2r})^2) \to \infty$. Thus $\hat{s}_j^* = O_p((a_n^2 + nM^{-2r})^2)/c_{1n})$ for $\forall j$. The ratio satisfies

$$\hat{r}_j = \frac{\hat{s}_{j+1}^* + c_{2n}}{\hat{s}_j^* + c_{2n}} \to 1, \text{ for } \forall j$$

if $c_{2n} \to 0$ and $c_{1n}c_{2n}/(a_n^2 + nM^{-2r})^2) \to \infty$. Therefore, the estimator $\hat{q} = 0$ with a probability going to 1 as $n \to \infty$.

Under the alternatives, the dimension $q = \dim\left(\mathcal{S}_{\epsilon|X}\right) > 0$ and it is easy to show that $\|\hat{M}^{cc} - M\| = O_p(a_n + n^{1/2}M^{-r})$. The eigenvalues of M^{cc} and $\hat{M}^{cc}$ satisfy $\lambda_1 \geq \cdots \geq \lambda_q > 0 = \lambda_{q+1} = \cdots$ and $\hat{\lambda}_i - \lambda_i = O_p(a_n + n^{1/2}M^{-r})$. Then,

$$\hat{s}_j^2 = \begin{cases} s_j^2 + O_p(a_n + n^{1/2}M^{-r}), \text{ for } 1 \leq j \leq q \\ O_p(a_n^2 + nM^{-2r}), \text{ for } j > q \end{cases}$$

According to the definition of $\hat{s}_j^*$, for $1 \leq j < q$

$$\hat{s}_j^* = \frac{\hat{s}_j^2 + c_{1n}}{\hat{s}_{j+1}^2 + c_{1n}} - 1 = \frac{s_j^2}{s_{j+1}^2} - 1 + O_p\left(\max\left\{a_n + n^{1/2}M^{-r}, c_{1n}\right\}\right)$$

For $j = q$,

$$\hat{s}_j^* = \frac{\hat{s}_q^2 + c_{1n}}{\hat{s}_{q+1}^2 + c_{1n}} - 1 = O_p\left(\frac{1}{c_{1n}}\right)$$

if $c_{1n}/(a_n^2 + nM^{-2r}) \to \infty$.

For $j > q$,

$$\hat{s}_j^* = \frac{\hat{s}_j^2 + c_{1n}}{\hat{s}_{j+1}^2 + c_{1n}} - 1 = \frac{\hat{s}_j^2 - \hat{s}_{j+1}^2}{\hat{s}_{j+1}^2 + c_{1n}} = O_p\left((a_n^2 + nM^{-2r})/c_{1n}\right).$$

That is,

$$\hat{s}_j^* = \begin{cases} s_j^* + O_p\left(\max\left\{a_n + n^{1/2}M^{-r}, c_{1n}\right\}\right), & \text{for } 1 \leq j < q \\ O_p\left(1/c_{1n}\right), & \text{for } j = q, \\ O_p\left((a_n^2 + nM^{-2r})/c_{1n}\right), & \text{for } j > q \end{cases}$$

with $c_{1n}/(a_n^2 + nM^{-2r}) \to \infty$.

For the estimated ratios with $c_{1n}c_{2n}/(a_n^2 + nM^{-2r}) \to \infty$ if $1 \leq j < q - 1$,

$$\hat{r}_j = \frac{\hat{s}_{j+1}^* + c_{2n}}{\hat{s}_j^* + c_{2n}} \to r_j,$$

for $j = q - 1$,

$$\hat{r}_j = \frac{O_p\left(1/c_{1n}\right)}{\hat{s}_{q-1}^* + c_{2n}} \to \infty,$$

for $j = q$

$$\hat{r}_j = \frac{O_p\left((a_n^2 + nM^{-2r})/c_{1n}\right) + c_{2n}}{O_p\left(1/c_{1n}\right) + c_{2n}} \to 0,$$

and for $j > q$,

$$\hat{r}_j = \frac{O_p\left((a_n^2 + nM^{-2r})/c_{1n}\right) + c_{2n}}{O_p\left(1/nc_{1n}\right) + c_{2n}} \to 1.$$

In summary,

$$\lim_{n\to\infty} \hat{r}_j = \begin{cases} r_j > 0, \text{ for } 1 \leq j < q - 1 \\ \infty, \text{ for } j = q - 1 \\ 0, \text{ for } j = q, \\ 1, \text{ for } j > q. \end{cases}$$

Therefore, under the alternative hypothesis, under the constraints $c_{1n} \to 0, c_{2n} \to 0$ and $c_{1n}c_{2n}/(a_n^2 + nM^{-2r}) \to \infty$, the estimated dimension satisfies $\mathbb{P}(\hat{q} = q > 0) \to 1$. ■

Proof of Theorem 10. According to Theorem 7, we have

$$\|\hat{\beta} - \beta_0\|_{L^2} = O_p(a_n + n^{1/2}M^{-r} + \delta_n).$$

under local alternatives, where $a_n = n^{-k/(2k+1)}$ and $\hat{\beta} = \hat{\beta}_{n,M,\lambda}$. Then we have

$$\begin{aligned}
\hat{H} &= \frac{1}{n}\sum_{j=1}^{n} \hat{\epsilon}_j \widehat{X}_j \otimes \widehat{X}_j = \frac{1}{n}\sum_{j=1}^{n} (\epsilon_j + \langle \widehat{X}_j, \hat{\beta}\rangle) - \langle X_j, \beta_0\rangle))\widehat{X}_j \otimes \widehat{X}_j \\
&= \frac{1}{n}\sum_{j=1}^{n} \epsilon_j X_j \otimes X_j + \frac{1}{n}\sum_{j=1}^{n} \langle \hat{\beta} - \beta, X_j\rangle X_j \otimes X_j + O_p(M^{-r}) \\
&= \frac{1}{n}\sum_{j=1}^{n} \eta_j X_j \otimes X_j + \delta_n \frac{1}{n}\sum_{j=1}^{n} X_j \otimes X_j + \frac{1}{n}\sum_{j=1}^{n} \langle \hat{\beta} - \beta, X_j\rangle X_j \otimes X_j + O_p(M^{-r}) \\
&= E(\eta X \otimes X) + O_p(\delta_n) + O_p(\delta_n + a_n + n^{1/2}M^{-r}) \\
&= O_p(\delta_n + a_n + n^{1/2}M^{-r}).
\end{aligned}$$

Similarly, we can also get $\hat{E}(\epsilon X) = O_p(\delta_n + a_n + n^{1/2}M^{-r})$. It follows that $\widehat{M}^{cc} = O_p(\delta_n^2 + a_n^2 + nM^{-2r})$. Similar to the proof of Theorem 3.4, when $\alpha > k/(2k+1)$, let $c_{1n} \to 0, c_{2n} \to 0$ and $c_{1n}c_{2n}/(a_n + nM^{-2r})^2 \to \infty$, then $\mathbb{P}(\hat{q} = 0) \to 1$.

When $0 < \alpha < k/(2k+1)$, then $\|\widehat{M}^{cc} - \delta_n^2 \widetilde{M}_n\| = O_p(a_n^2 + nM^{-2r})$ for some operator $\widetilde{M}_n$. Suppose the eigenvalues of $\widehat{M}^{cc}$ and $\widetilde{M}_n$ are $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_n \geq \ldots$ and $\tilde{\lambda}_1 \geq \ldots \geq \tilde{\lambda}_q \geq 0 = \tilde{\lambda}_{q+1} = \tilde{\lambda}_{q+2} = \ldots$ in descending order. It's easy to show that $\hat{\lambda}_j - \delta_n^2 \tilde{\lambda}_j = O_p(a_n^2 + nM^{-2r})$ and thus

$$\hat{s}_j = \frac{\hat{\lambda}_j}{\hat{\lambda}_j + 1} = \begin{cases} O_P(\delta_n^2), & j = 1, \ldots, q \\ O_p(a_n^2 + nM^{-2r}), & j > q \end{cases}$$

Then if $c_{1n} = o(\delta_n^4)$ and $c_{1n}/(a_n^2 + nM^{-2r}) \to \infty$, then

$$\hat{s}_j^* = \frac{\hat{s}_j^2 + c_{1n}}{\hat{s}_{j+1}^2 + c_{1n}} - 1 = \frac{\hat{s}_j^2 - \hat{s}_{j+1}^2}{\hat{s}_{j+1}^2 + c_{1n}} \to \begin{cases} (\tilde{\lambda}_j^2 - \tilde{\lambda}_{j+1}^2)/\tilde{\lambda}_{j+1}^2, & j = 1, \ldots, q-1 \\ O_p(\delta_n^4/c_{1n}), & j = q \\ O_p(a_n^4 + n^2M^{-4r})/c_{1n}), & j > q \end{cases}$$

Let C_j represent some constants that may vary hereafter. We have

$$\lim_{n\to\infty} \hat{s}_j^* = \begin{cases} C_j, & j = 1, \ldots, q-1 \\ \infty, & j = q \\ 0, & j > q \end{cases}$$

For $j = 1, \ldots, q-2$, the ratios are

$$\hat{r}_j = \frac{\hat{s}_{j+1}^* + c_{2n}}{\hat{s}_j^* + c_{2n}} = \frac{O_p(1) + c_{2n}}{O_p(1) + c_{2n}}$$

for $j = q-1$,

$$\hat{r}_j = \frac{O_p(\delta_n^4/c_{1n}) + c_{2n}}{O_p(1) + c_{2n}}$$

for $j = q$,

$$\hat{r}_j = \frac{O_p((a_n^4 + n^2M^{-4r})/c_{1n}) + c_{2n}}{O_p(\delta_n^4/c_{1n}) + c_{2n}}$$

for $j > q$,

$$\hat{r}_j = \frac{O_p((a_n^4 + n^2M^{-4r})/c_{1n}) + c_{2n}}{O_p(a_n^4 + n^2M^{-4r})/c_{1n}) + c_{2n}}$$

In summary, with $c_{1n} = o(\delta_n^4)$ and $c_{1n}c_{2n}/(a_n^4 + n^2M^{-4r}) \to \infty$, we have

$$\lim_{n\to\infty} \hat{r}_j = \begin{cases} C_j, & j = 1, \dots, q-2 \\ \infty, & j = q-1 \\ 0, & j = q; \\ 1, & j > q, \dots, p \end{cases}$$

The proof is done. ■

Lemma 13 *Under the null hypothesis, let assumptions be satisfied and $Mn^{-1/r} \to \infty$, then we have*

$$W_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i}^{n} K_h\{\langle \hat{\beta}_{n,M,\lambda}, X_i - X_j \rangle\}\epsilon_i M(X_j) = O_p(\frac{1}{\sqrt{n}}),$$

where $M(\cdot)$ is continuously differentiable and $E\{M^2(X)|\langle \beta_0, X\rangle\} \leq b(\langle \beta_0, X\rangle)$ for $X \in L^2[0,1]$ and $E\{b(\langle \beta_0, X\rangle)\} < +\infty$.

Proof of Lemma 13. For notation simplicity, let $\hat{\beta}$ denote $\hat{\beta}_{n,M,\lambda}$. Denote $b_{ij} = \langle \beta_0, X_i - X_j\rangle$ and $\hat{b}_{ij} = \langle \hat{\beta}, X_i - X_j \rangle$. Note that

$$\begin{aligned} W_n =& \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i}^{n} \frac{1}{h} K(\frac{b_{ij}}{h})\epsilon_i M(X_j) \\ &+ \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i}^{n} \frac{1}{h}[K(\frac{\hat{b}_{ij}}{h}) - K(\frac{b_{ij}}{h})]\epsilon_i M(X_j) \\ =& W_{n1} + W_{n2} \end{aligned}$$

Let $t_i = \{Y_i, X_i\}$, then W_{n1} can be written in a U-statistic with the kernel

$$H_n(t_i, t_j) = \frac{1}{2h} K(\frac{b_{ij}}{h})\{\epsilon_i M(X_j) + \epsilon_j M(X_i)\}$$

To apply the theory for non-degenerate U-statistic, we need to show $E[H_n^2(\mathrm{t}_i, \mathrm{t}_j)] = o(n)$. Let $Z = \langle \beta_0, X \rangle$, $f(z)$ be the probability density function of Z and $\sigma^2(z) =$

$E(\epsilon^2 \mid Z = z)$. It can be verified that

$$\begin{aligned}
&E[H_n^2(\mathrm{t}_i, \mathrm{t}_j)] \\
\le& 2E[\frac{1}{2h}K(\frac{b_{ij}}{h})\epsilon_i M(X_j)]^2 + 2E[\frac{1}{2h}K(\frac{b_{ij}}{h})\epsilon_j M(X_i)]^2 \\
=& \int \frac{1}{h^2}\sigma^2(z_i)E\{M^2(X_j) \mid z_j\}K^2(\frac{z_i - z_j}{h})f(z_i)f(z_j)dz_i dz_j \\
\le& \int \frac{1}{h}\sigma^2(z_i)b(z_i - hu)K^2(u)f(z_i)f(z_i - hu)dz_i du \\
=& \int \frac{1}{h}\sigma^2(z)b(z)f^2(z)dz \cdot \int K^2(u)du + o(1/h) \\
=& O(1/h) = o(n)
\end{aligned}$$

The last equation holds under the assumption that $nh \to \infty$. Since $E(\epsilon | X) = 0$, it can be derived that $E\{H_n(t_i, t_j)\} = 0$. Now, consider the conditional expectation of $H_n(\mathrm{t}_i, \mathrm{t}_j)$. Also, it is easy to compute that

$$\begin{aligned}
r_n(t_i) = E\{H_n(t_i, t_j) \mid t_i\} &= \frac{\epsilon_i}{2h}E[K(\frac{z_i - Z}{h})E\{M(X_i) \mid Z\}] \\
&= \frac{\epsilon_i}{2}\int E\{M(X_i) \mid z_i - hu\}f(z_i - hu)K(u)du \\
&= \frac{\epsilon_i f(z_i)E\{M(X_i) \mid z_i\}}{2} + l_n(t_i).
\end{aligned}$$

The last equation comes from Taylor expansion and it implies that $l_n(t_i) = O(h)$.

Denote $\hat{W}_n$ as the projection of the statistic W_{n1} as:

$$\sqrt{n}\hat{W}_n = \frac{2}{\sqrt{n}}\sum_{i=1}^{n} r_n(t_i) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \epsilon_i f(z_i)E\{M(X_i) \mid z_i\} + \frac{2}{\sqrt{n}}\sum_{i=1}^{n} l_n(t_i) = O_p(1)$$

The last equation holds because first term follows the central limit theorem and the fact that $E\{l_n^2(t_i)\} = O(h^2) \to 0$. As a result, we have $W_{n1} = O_p(\hat{W}_n) = O_p(1/\sqrt{n})$. Denote

$$W_{n2}^* = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j \neq i}^{n} \frac{1}{h}K'(\frac{b_{ij}}{h})^T \epsilon_i M(X_j)\frac{\langle \hat{\beta} - \beta_0, X_i - X_j \rangle}{h}$$

Then for the term W_{n2}, we have $W_{n2} = W_{n2}^* + o_p(W_{n2}^*)$. Notice that here K' denote the gradient of K, and $\langle \hat{\beta} - \beta_0, X_i - X_j \rangle$ is a scalar. Apply Cauchy's inequality we have

$$\langle \hat{\beta} - \beta_0, X_i - X_j \rangle \le \|\hat{\beta} - \beta_0\|_{L_2} \cdot \|X_i - X_j\|_{L^2} = O_p(n^{-k/(2k+1)} + n^{1/2}M^{-r}).$$

Since $K(\cdot)$ is spherically symmetric, similar to W_{n1}, the following term

$$\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j \neq i}^{n} \frac{1}{h}K'(\frac{b_{ij}}{h})^T \epsilon_i M(X_j)\|X_i - X_j\|$$

can be rewritten as a U-statistic. Then we can similarly show that this term is also of order $O_p(1/\sqrt{n})$. Thus we can obtain that $W_{n2} = o_p(1/\sqrt{n})$ if $(n^{-k/(2k+1)} + n^{1/2}M^{-r})/h \to 0$. Then we can conclude that $W_n = O_p(1/\sqrt{n})$. The proof is completed. ■

Lemma 14 *Under the null hypothesis, let assumptions be satisfied and $Mn^{-1/r} \to \infty$, then we have*

$$nh^{1/2}V_1 \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \Sigma\right),$$

where

$$\Sigma = 2\int K^2(u)\mathrm{d}u \int \left\{\sigma^2(\mathrm{z})\right\}^2 f^2(\mathrm{z})\mathrm{d}\mathrm{z},$$

in which $Z = \langle \beta_0, X\rangle, \sigma^2(z) = E\left(\epsilon^2 \mid Z = z\right)$.

Proof of Lemma 14. For notational convenience, denote $\langle\hat{\beta}-\beta_0, X_i - X_j\rangle$ as A_{ij}, $\langle\hat{\beta}-\beta_0, X_i\rangle$ as C_i and $a_n = n^{-k/(2k+1)} + n^{1/2}M^{-r}$. Also denote $b_{ij} = \langle\beta_0, X_i - X_j\rangle$ and $\hat{b}_{ij} = \langle\hat{\beta}, X_i - X_j\rangle$. First, noting the symmetry of $K_h(\cdot)$, then V_1 can be decomposed as

$$\begin{aligned} V_1 =& \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K_h(\hat{b}_{ij})\epsilon_i\epsilon_j - \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K_h(\hat{b}_{ij})\epsilon_i C_j \\ &+ \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K_h(\hat{b}_{ij})C_iC_j + o_p(V_n^*) \\ =:& V_{11} - V_{12} + V_{13} + o_p(V_n^*) \end{aligned}$$

where V_n^* denotes the term $V_{11} - V_{12} + V_{13}$.

For V_{12}, a direct application of Lemma 13 and the fact that $C_i = O_p(a_n)$ yield $V_{12} = O_p(a_n/\sqrt{n})$. Thus $nh^{1/2}V_{12} = o_p(1)$ if $\sqrt{n}h^{1/2}a_n \to 0$. For V_{13}, it's easy to see $V_{13} = o_p(1/n)$ by the rate of C_i. Thus $nh^{1/2}V_{13} = o_p(1)$.

Finally, for term V_{11}, consider the decomposition

$$\begin{aligned} V_{11} =& \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} K_h(b_{ij})\epsilon_i\epsilon_j \\ &+ \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}[K_h(\hat{b}_{ij}) - K_h(b_{ij})]\epsilon_i\epsilon_j \\ =:& V_{11,1} + V_{11,2} \end{aligned}$$

For the term $V_{11,1}$, using the similar argument can derive the asymptotic normality $nh^{1/2}V_{11,1} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma)$. For the term $V_{11,2}$, an application of Taylor expansion yields

$$V_{11,2} = V_{11,2}^* + o_p(V_{11,2}^*),$$

where

$$V_{11,2}^{*}=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\frac{1}{h}K'(\frac{b_{ij}}{h})\epsilon_i\epsilon_j\cdot\frac{A_{ij}}{h},$$

which can be considered as an U-statistic. Together with $A_{ij}=O_p(a_n)$ and $a_n/h\to 0$, we can conclude that $nh^{1/2}V_{11,2}^{*}=o_p(1)$. Combine all the results above conclude the proof. ■

Proof of Theorem 11. Part 1): Under the null hypothesis, $\mathbb{P}(\hat{q}=0)\to 1$, we only need to work with $V_0=\sum_{i=1}^{n}\hat{\epsilon}_i w(\widehat{X}_i)=\sum_{i=1}^{n}\hat{\epsilon}_i w_i$. Based on previous analysis, β_0 coincides with β_0. Therefore

$$\begin{aligned}V_0&=\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i w_i=\frac{1}{n}\sum_{i=1}^{n}\epsilon_i w_i+\frac{1}{n}\sum_{i=1}^{n}(\hat{\epsilon}_i-\epsilon_i)w_i\\&=\frac{1}{n}\sum_{i=1}^{n}\epsilon_i w_i+\frac{1}{n}\sum_{i=1}^{n}\langle\hat{\beta}_{n,M,\lambda}-\beta_0,\widehat{X}_i\rangle w_i+o_p(\frac{1}{\sqrt{n}})\\&=\frac{1}{n}\sum_{i=1}^{n}\epsilon_i w_i+\frac{1}{n}\sum_{i=1}^{n}\langle\hat{\beta}_{n,M,\lambda}-\hat{\beta}_{n,\lambda},\widehat{X}_i\rangle w_i+\frac{1}{n}\sum_{i=1}^{n}\langle\hat{\beta}_{n,\lambda}-\beta_0,\widehat{X}_i\rangle w_i+o_p(\frac{1}{\sqrt{n}})\\&=\frac{1}{n}\sum_{i=1}^{n}\epsilon_i w_i+\frac{1}{n}\sum_{i=1}^{n}\langle\hat{\beta}_{n,\lambda}-\beta_0,\widehat{X}_i\rangle w_i+O_p(n^{1/2}M^{-r})+o_p(\frac{1}{\sqrt{n}})\\&=\frac{1}{n}\sum_{i=1}^{n}\epsilon_i w_i+\langle\hat{\beta}_{n,\lambda}-\beta_0,\widehat{X}_w\rangle+O_p(n^{1/2}M^{-r})+o_p(\frac{1}{\sqrt{n}}),\end{aligned}$$

where $\widehat{X}_w=\sum_{i=1}^{n}w_i\widehat{X}_i/n$. Suppose all assumptions are satisfied, according to Theorem 5.1 in [10], we can get $\sqrt{n}\langle\hat{\beta}_{n,\lambda}-\beta_0,\widehat{X}_w\rangle/\sigma_n\xrightarrow{\mathcal{D}}N(0,1)$.

Here it's easy to show that $\sum_{\nu=1}^{\infty}w_v^2/(1+\lambda\rho_v^*)^2\asymp\lambda^{-1/(2k)}$ based on the proof of Proposition 4.2 in [10]. When we take $\lambda=n^{-k/(2k+1)}$ as in estimation procedure, it gives $\sigma_n\asymp n^{1/(2(2k+1))}$ and $V_0=O_p(n^{-k/(2k+1)})$, which coincides with the estimation convergence rate of $\hat{\beta}_{n,\lambda}$. By central limit theorem, $\sum_{i=1}^{n}\epsilon_i w_i/\sqrt{n}\xrightarrow{\mathcal{D}}N(0,\sigma_0^2)$, where $\sigma^2=E(\epsilon^2w^2)$. Notice that $\sqrt{n}/\sigma_n=o_p(\sqrt{n})$ and $M^r=O_p(1/n)$. Therefore, we have $\sqrt{n}/(\sigma_n n^{1/2}M^r)=o(1)$ under the null hypothesis, which implies $\sqrt{n}V_0/\sigma_n\xrightarrow{\mathcal{D}}N(0,1)$.

Part 2): Under the global alternative hypothesis, $\mathbb{P}(\hat{q}>0)\to 1$, then we only need to deal with V_1. Let $\Delta_i=G(X_i)-\langle\beta_0,X_i\rangle$, then $\hat{\epsilon}_i=\eta_i+\Delta_i+\langle\beta_0-\hat{\beta},X_i\rangle$. For notational convenience, denote $\langle\hat{\beta}-\beta_0,X_i-X_j\rangle$ as A_{ij}, $\langle\hat{\beta}-\beta_0,X_i\rangle$ as C_i and $a_n=n^{-k/(2k+1)}+n^{1/2}M^{-r}$. Denote $b_{ij}=\langle\beta_0,X_i-X_j\rangle$ and $\hat{b}_{ij}=\langle\hat{\beta},X_i-X_j\rangle$. Then

V_1 can be decomposed as

$$\begin{aligned}
V_1 =& \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i,j=1}^{n} (\eta_i + C_i + \Delta_i)(\eta_j + C_j + \Delta_j) K_h(\hat{b}_{ij}) \\
=& \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i,j=1}^{n} (\eta_i + C_i)(\eta_j + C_j) K_h(\hat{b}_{ij}) \\
&+ \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i,j=1}^{n} (\eta_i + C_i)\Delta_j K_h(\hat{b}_{ij}) \\
&+ \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i,j=1}^{n} \Delta_i \Delta_j K_h(\hat{b}_{ij}) + o_p(1) \\
=:& V_{11} + V_{12} + V_{13} + o_p(1).
\end{aligned}$$

First, it's easy to show that $V_{11} = o_p(1)$, which is from the proofs in Lemma 14. For V_{12}, a direct application of U-statistic theory combined with $C_i = O_p(a_n)$ and $a_n/h \to 0$ implies that $V_{12} = o_p(1)$. Again use the U-statistic theory, we conclude $V_{13} \xrightarrow{\mathcal{D}} \mu = E(\Delta^2\langle\beta_0, X\rangle)$, which complete the proof.

■

Proof of Theorem 12. Under the local alternatives, we have $\hat{\epsilon}_i = \eta_i + \delta_n l(X_i) + \langle\beta_0 - \hat{\beta}, X_i\rangle$. If $\alpha \geq k/(2k+1)$, $\hat{q} \to 0$. The working test statistic is reduced to V_0. Then it follows

$$\begin{aligned}
V_0 &= \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i w_i = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i w_i + \frac{1}{n}\sum_{i=1}^{n} (\hat{\epsilon}_i - \epsilon_i) w_i \\
&= \frac{1}{n}\sum_{i=1}^{n} \epsilon_i w_i + \frac{1}{n}\sum_{i=1}^{n} \langle\hat{\beta}_{n,M,\lambda} - \beta_0, \widehat{X}_i\rangle w_i + o_p(\frac{1}{\sqrt{n}}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \epsilon_i w_i + \frac{1}{n}\sum_{i=1}^{n} \langle\hat{\beta}_{n,M,\lambda} - \hat{\beta}_{n,\lambda}, \widehat{X}_i\rangle w_i + \frac{1}{n}\sum_{i=1}^{n} \langle\hat{\beta}_{n,\lambda} - \beta_0, \widehat{X}_i\rangle w_i + o_p(\frac{1}{\sqrt{n}}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \epsilon_i w_i + \frac{1}{n}\sum_{i=1}^{n} \langle\hat{\beta}_{n,\lambda} - \beta_0, \widehat{X}_i\rangle w_i + O_p(n^{1/2}M^{-r}) + o_p(\frac{1}{\sqrt{n}}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \epsilon_i w_i + \langle\hat{\beta}_{n,\lambda} - \beta_0, \widehat{X}_w\rangle + O_p(n^{1/2}M^{-r}) + o_p(\frac{1}{\sqrt{n}}), \\
&= \frac{1}{n}\sum_{i=1}^{n} \eta_i w_i + \frac{\delta_n}{n}\sum_{i=1}^{n} l(X_i) w_i + \langle\hat{\beta}_{n,\lambda} - \beta_0, \widehat{X}_w\rangle + O_p(n^{1/2}M^{-r}) + o_p(\frac{1}{\sqrt{n}}).
\end{aligned}$$

Thus when $\alpha > k/(2k+1)$, we have $\sqrt{n}\delta_n/\sigma_n \to 0$ and $T_n \xrightarrow{\mathcal{D}} \chi_1^2$. When $\alpha =$

$k/(2k+1)$, $\sqrt{n}\delta_n/\sigma_n = O_p(1)$, then $\sqrt{n}V_0/\sigma_n \xrightarrow{\mathcal{D}} N(\mu_0, 1)$, where $\mu = E(l(X)w(X))$, and thus $T_n \xrightarrow{\mathcal{D}} \chi_1^2(\mu_0)$.

If $0 < \alpha < k/(2k+1)$, $\hat{q} \to q > 0$. The working test statistic is reduced to V_1. Then it follows

$$\begin{aligned}
V_1 =& \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i, j=1}^{n} (\eta_i + C_i + \delta_n l_i)(\eta_j + C_j + \delta_n l_j) K_h(\hat{b}_{ij}) + o_p(1) \\
=& \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i, j=1}^{n} (\eta_i + C_i)(\eta_j + C_j) K_h(\hat{b}_{ij}) \\
&+ \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i, j=1}^{n} (\eta_i + C_i)\delta_n l_j K_h(\hat{b}_{ij}) \\
&+ \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i, j=1}^{n} \delta_n^2 l_i l_j K_h(\hat{b}_{ij}) + o_p(1) \\
=:& \widetilde{V}_{11} + \widetilde{V}_{12} + \widetilde{V}_{13} + o_p(1).
\end{aligned}$$

Based on the results of previous lemmas and theorems, it's easy to conclude that $\widetilde{V}_{11} = O_p(n^{-1}h^{-1/2})$, $\widetilde{V}_{12} = O_p(\delta_n/\sqrt{n})$ and $\widetilde{V}_{13} = O_p(\delta_n^2)$. Therefore, the leading term of V_1 is actually $\widetilde{V}_{11}$ and $\widetilde{V}_{13}$. Therefore, we can have the following results.

(a) If $n^{1/2}h^{1/4}\delta_n \to 0$, then $nh^{1/2}V_1 \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma)$, where Σ is defined in Lemma 14. Then $T_n = O_p(h^{-1/2})$ and $T_n/h^{-1/2}$ converges in distribution to a centered normal distribution.

(b) If $\delta_n = n^{1/2}h^{1/4}$, $nh^{1/2}\widetilde{V}_{11} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma)$ and $nh^{1/2}\widetilde{V}_{13} \xrightarrow{\mathcal{D}} E(l^2 f)$, then $T_n/h^{-1/2}$ converges in distribution to a non-central normal distribution with a shift $E(l^2 f)$ from situation (a).

(c) If $n^{1/2}h^{1/4}\delta_n \to \infty$, then V_1/δ_n^2 converges to $E(l^2 f)$ in probability, thus $T_n/n\delta_n^2$ converges in probability to $E(l^2 f)$.

■