The Pennsylvania State University
The Graduate School

HYPOTHESIS TESTING AND VARIABLE SELECTION IN NONPARAMETRIC REGRESSION

A Dissertation in
Statistics
by
Adriano Zanin Zambom

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2012

The dissertation of Adriano Zanin Zambom was reviewed and approved* by the following:

Michael Akritas
Professor of Statistics
Dissertation Advisor, Chair of Committee

Runze Li
Professor of Statistics
Chair of the Graduate Study of Statistics

Bing Li
Professor of Statistics

Adam Smith
Associate Professor - Computer Science and Engineering

Bruce Lindsay
Willaman Professor of Statistics and Department Head

*Signatures are on file in the Graduate School.

Abstract

Let $\mathbf{X}$ be a d dimensional vector of covariates and Y be the response variable. Under the nonparametric model $Y = m(\mathbf{X}) + \sigma(\mathbf{X})\epsilon$ we develop an ANOVA-type test for the null hypothesis that a particular coordinate of $\mathbf{X}$ has no influence on the regression function. The asymptotic distribution of the test statistic, using residuals based on Nadaraya-Watson type kernel estimator is established under the null hypothesis and local alternatives. When using local polynomial regression, it is shown that the theorem holds for higher dimensions under some smooth assumptions. Simulations show that the proposed procedure outperforms existing methods. Moreover, additional simulations suggest that under a sparse model, the applicability of the test extends to arbitrary d through sufficient dimension reduction. Using p-values from this test, a variable selection method based on multiple testing ideas is proposed. Simulations reveal that the proposed variable selection method performs competitively against well established procedures. A real data set is analyzed as an application of the variable selection. The intuitive extension of the test statistic for testing the significance of more than one covariate at a time is developed, and its asymptotic normality is stablished. We investigate the power of the this test under different scenarios, including linear and non-linear regression and logistic regression. There are many situations where the covariates appear in groups, and the selection of the significant groups under the nonparametric regression model is of interest. We propose a group variable selection procedure based on multiple testing ideas. Simulations suggest that under a nonparametric or non-additive model the proposed procedure outperforms linear based group variable selections methods. We also use the ANOVA-type methodology to introduce a test statistic for the additivity of the regression for the homocedastic case.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank very much my advisor Michael G. Akritas for his guidance during these past semesters, for the brilliant insights that contributed enormously to this thesis and for his encouragement in those days of lack of motivation. Also, I would like to thank all the faculty at the Department of Statistics at Pennsylvania State University for the great courses that contributed a lot to my research. I am very thankkful to all those who contributed with ideas, suggestions and support when I needed. Lastly, I want to thank my parents and family for being there for me in each moment.

Adriano

Chapter 1

Introduction

In nonparametric regression analysis, the focus is on studying and exploring the relation between the explanatory variables $\mathbf{X} = (X_1, ..., X_d)$ and the response variable Y through a regression function m. In the simple scenario, $\mathbf{X}$ can be considered as a fixed design, but in the general case we consider $\mathbf{X}$ to be random with density function $f_{\mathbf{X}}(\mathbf{x})$.

The function m is called the regression function, usually considered to be a smooth function such that

$$m(\mathbf{X}) = E(Y|\mathbf{X}). \tag{1.1}$$

It is typically assumed that the conditional variance of Y given $\mathbf{X}$ is constant, but in a more general case, we can relax this assumption and assume it still depends on $\mathbf{X}$. Thus, the general heterocedastic nonparametric regression model is

$$Y_i = m(\mathbf{X}_i) + \sigma(\mathbf{X}_i)\epsilon_i, \ i = 1...n \tag{1.2}$$

where ϵ_i are independent identically distributed random variables with mean 0 and constant variance (without loss of generality equal to 1).

Exploring relations between variables is the very core of applied statistics, and regression is one of the main tools scientists can use. Constantly, we are looking for ways of generalizing models so that they can fit more and more applications, and making these models more flexible, or less restrictive, is of great importance.

With the increasing number of data sets with regression relations that are non trivial, the flexibility of the model becomes of great interest and it will be one

of our main focus. It is one of the fundamental aspects of modeling, together with the dimensionality and interpretability. When focusing in flexibility, if the dimension of the variables is large, fitting the model becomes problematic, the so called curse of dimensionality. Moreover, not only should the model be flexible and deal with higher dimensions, but also have reasonably easy interpretation. Here we focus mainly on flexibility, where model checking and variable selection can be performed without strong restrictions or assumptions on the regression function.

Nonparametric regression is a way of generalizing the classical parametric regression

$$Y_i = m(\theta, \mathbf{X}) + \epsilon_i, \ i = 1...n \tag{1.3}$$

where β are the parameters and $m(\theta, \mathbf{X})$ is of a parametric form. This restricts the model, since it assumes that the family of models $\{m(\theta, \mathbf{x}), \theta \in \Theta\}$ contains the true one.

The first and most commonly used parametric regression model, a particular case of (1.3), is the parametric linear regression, which assumes that the predictors are additive and linear in their effects (a very strong assumption). Even though this is very restrictive, it is widely used and does apply to many situations, there is an increasing amount of data that can not be fit by a linear model. By adding powers of the covariates for a polynomial fit, the number of parameters to estimate will increase and lead to an overfit and lack of interpretability.

In the following example, we will demonstrate the performance of nonparametric regression models (discussed in section 1.1) and the parametric linear regression model.

Example 1: Suppose that we have pairs of observations (Y_i, X_i) from the unknown functions Model 1: $Y = 6X\cos(6\pi X) + \epsilon$ and Model 2: $Y = 4X\cos(5\pi X) + \epsilon$, where $X \sim U(0,1)$ and $\epsilon \sim N(0,1)$. Table 1.1 shows the residual sums of squares for different estimation procedures. The Kernel regression was estimated using Normal kernels, and p is the degree of the polynomial for the parametric linear regression. Clearly, the nonparametric models have much better results compared to the parametric linear regressions, even with a large degree of polynomial (many parameters to estimate) the linear regression does not do well compared to the nonparametric models. This demonstrates the great importance of having a flexible model that does not restrict the estimation to a small range of parametric

families, when the true model can have nonlinear effects on the response.

For high-dimensional vector of covariates, the most used nonparametric re-

Table 1.1. Comparison of RSS for different methods of regression estimation

	SLR (p=1)	SLR (p=2)	SLR (p=5)	Local Linear	Kernel	Splines
Model 1	737.1	703.3	543.3	83.1	118.3	69.5
Model 2	414.4	410.4	210.5	89.9	85.8	61.6

gression model is a generalization of the parametric linear regression. Hastie and Tibshrani (1990) considered this case, known as nonparametric additive models, where

$$Y = m_1(X_1) + m_2(X_2) + ... + m_d(X_d) + \epsilon, \tag{1.4}$$

for the covariates $X_1, ..., X_d$. For identifiability of the model, it is required that $E(X_i) = 0, i = 1, \ldots, d$. In this case, we can clearly see that the additive model is additive in the predictor effects. This relates each explanatory variable X to the response Y in an additive way, but the functions m_i are not parametrically specified and will be determined by the data analytically. Moreover, one fundamental property of the linear model is retained: very easy interpretation. The variation of the fitted response surface due to one covariate (or explanatory variable) does not depend on the values of the other covariates once their values are fixed.

Here again we face the trade off between interpretability with high dimensions and flexibility. The additive model described above can handle high dimensions (as long as n is large) and has somewhat easy interpretation, but it does not account for more complex models, when the predictors interact for example. In the case of interaction, the variation in the response according to one predictor does change depending on the value of the other predictors. Furthermore, sometimes the purpose of the study is to test if additivity holds.

We can find many cases in applied statistics where the additive model may not be the correct one. For example, if we have a study of HDL Cholesterol predicted by BMI (body mass index) and Total Cholesterol, the additive effect of the predictors may not be correct, since BMI probably has a interaction with Total Cholesterol. Another example would be the regression of two different medicines

in a situation where covariates are the levels of different treatments (in medicine or agriculture) and the response is some measure of a reaction against a particular desease.

Example 2: We will make use of a simple example where the data is generated through the model (unknown to us) $Y_i = m_j(X_{1i}, X_{2i}) + \epsilon_i, j = 1, \ldots, 4$, where X_1 and X_2 are generated independently from a $U(0,1)$ distribution, independent of $\epsilon \sim N(0, .5^2)$. Define $m_1 := m_1(X_1, X_2) = X_1X_2, m_2 := m_2(X_1, X_2) = X_1X_2^2, m_3 := m_3(X_1, X_2) = X_1 + X_2 + X_1X_2$ and $m_4 := m_4(X_1, X_2) = X_1/X_2$. Table 1.2 shows the RSS in each case.

The general model (1.2) in this example was estimated by a bivariate kernel

Table 1.2. Comparison of RSS for different methods of nonparametric regression estimation

	m_1	m_2	m_3	m_4
Additive Model	34.7	26.6	27.1	25.2
General Model	29.5	19.2	19.3	21.2

using cross validation for the bandwidth and the additive model was estimated with the backfitting algorithm (which will be described briefly on section 1.1). We see from the table, that the general model outperforms the additive model in residual sum of squares. The results of the additive model are not too far from those of the general model, but and extra concern is that it does not contain the true model. Thus the interpretability can be misleading. Moreover, as the dimensions increase and more interactions come into play, the performance of the additive model may worsen.

Other models can be derived as variations of the additive model, and those include semi-parametric models, varying-coefficient models, partially linear models, varying-coefficient partially linear models, etc. In all of these cases, the effect of each covariate is assumed to be additive, and a more general model is always of interest.

After fitting a model, one of the questions that arise is if the effects of certain covariates are really statistically significant. For that we need to build hypothesis testing procedures, which is one of our main interests in this thesis.

The thesis is organized in the following way. The remainder of Chapter 1 contains an introduction to estimation in nonparametric regression, with some important properties of such estimators. Chapter 2 gives a literature review on hypothesis testing in nonparametric regression, describing the ideas that have been explored in this area, with a more detailed explanation of the generalized likelihood ratio test. Chapter 3 introduces the idea of using ANOVA-type statistics to hypothesis testing in regression models, studied by Akritas and Papadatos (2004) and extended by Wang, Akritas and Van Keilegom (2008). Chapter 4 is the main chapter of this thesis. It contains a very detailed derivation of the hypothesis test for nonparametric regression in the univariate and multivariate case. Asymptotic distribution under the null and local alternatives is given, dimension reduction ideas are introduced to improve the estimation for higher dimensions. This methodology is extended from kernel regression to local polynomial regression. Also in this chapter we extend the test statistic to testing multiple covariates at the same time. Chapter 5 describes the methodology for variable selection using the ANOVA-type hypothesis test in combination with multiple testing correction ideas. Chapter 6 introduces a new idea based on regression to the same hypothesis test. Chapter 7 contains some final discussion on the results obtained and future work.

1.1 Brief Introduction of Estimation

Although estimation is not the focus of this thesis, we will briefly introduce a few methods for estimation of the nonparametric regression function $m(\mathbf{X})$, as we will in many occasions need the estimated function $\hat{m}(\mathbf{X})$.

The usual assumption is that the regression function m is smooth in some sense, hence the classical estimation procedures rely on smoothing methods. To estimate m at a point $\mathbf{x}$, we usually consider a neighborhood of this point and take some weighted average of the response values corresponding to the $\mathbf{X}$ data points in this neighborhood. The obvious dilemma here is how to average the values and how big the neighborhoods should be, leading to a trade-off between the estimation bias and the estimation variance governed by a smoothing parameter, which will adjust the size of the neighborhood.

1.1.1 Local Polynomial Estimation

Consider first the univariate case where X is 1-dimensional. In estimation of $m(x)$ with Local Polynomial Regression, the classical weighted least squares regression is used to fit a q degree polynomial

$$\beta_0 + \beta_1(. - x) + ... + \beta_p(. - x)^q$$

to the data (X_i, Y_i), where the weights are the kernel functions $K_{h_n}(X_i - x) = \frac{1}{h_n}K(\frac{X_i - x}{h_n})$, for a function K such that $\int K(x)dx = 1$ called the kernel function (usually taken to be a density). Therefore, the goal is to minimize

$$\sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1(X_i - x) - \ldots - \beta_p(X_i - x)^q\right)^2 K_{h_n}(X_i - x) \tag{1.5}$$

with respect to $(\beta_0, \ldots, \beta_p)$, and the estimated m will be $\hat{\beta}_0$. Note that one of the assumptions is that the function m has $q + 1$ derivatives and they are continuous at x, since this polynomial tries to approximate (or estimate) m, corresponding to the expansion $m(X_i) \approx m(x) + m'(x)(X_i - x) + \ldots + \frac{m^{(q)}(x)}{q!}(X_i - x)^q$.

For simplicity, assume that the support of f_X is on [0,1] and the kernel is supported on [-1,1]. Also assume that m'', f' σ are continuous, the kernel is symmetric about 0 and the bandwidth h_n is such that $h_n \to 0$ and $nh_n \to \infty$.

In matrix form, let $Y = (Y_1, \ldots, Y_n)'$,

$$X_x = \begin{pmatrix} 1 & X_1 - x & \ldots & (X_1 - x)^q \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_n - x & \ldots & (X_n - x)^q \end{pmatrix}$$

and $W_x = \text{diag}\{K_{h_n}(X_1 - x), \ldots, K_{h_n}(X_n - x)\}$ is the diagonal matrix of weights. It is clear that the solution of this weighted least squares problem is

$$\hat{\beta} = (X_x^T W_x X_x)^{-1} X_x^T W_x Y$$

assuming that $(X_x' W_x X_x)$ is invertible. But because the value we need is just $\hat{\beta}_0$, we then have the estimator of $m(x)$ as

$$\hat{m}(x; q, h_n) = e_1^T (X_x^T W_x X_x)^{-1} X_x^T W_x Y \tag{1.6}$$

where $e_1 = (1, 0, \ldots, 0)$ of size $(q+1) \times 1$.

In the linear case when $q = 1$, using taylor's expansion of $m(x_i)$ around x, the conditional bias is given by

$$E(\hat{m}(x; 1, h_n) - m(x) | X_1, \ldots, X_n) = \frac{1}{2} h_n^2 m''(x) \int z^2 K(z) dz + o_p(h_n^2) \tag{1.7}$$

and the conditional variance is given by

$$Var(\hat{m}(x; 1, h_n) | X_1, \ldots, X_n) = \frac{\int K(z)^2 dz}{n h_n f(x)} \sigma(x)^2 + o_p((n h_n)^{-1}). \tag{1.8}$$

Now consider the case where $\mathbf{X}$ is d-dimensional. The idea here is the same as in the univariate case, only now we have to extend to d dimensions of $\mathbf{X}$ (see Masry, 1996; Ruppert and Wand, 1994). Assume that we have a regression function $m(\mathbf{z})$ such that $q+1$ derivatives exist and are continuous at $\mathbf{x}$. Then we can approximate $m(\mathbf{z})$ by a multivariate polynomial

$$m(\mathbf{z}) \approx \sum_{0 \le |\mathbf{k}| \le q} \frac{1}{\mathbf{k}!} D^{\mathbf{k}} m(\mathbf{y})|_{\mathbf{y}=\mathbf{x}} (\mathbf{z} - \mathbf{x})^{\mathbf{k}} \tag{1.9}$$

where

$$\mathbf{k} = (k_1, \ldots, k_d), \quad \mathbf{k}! = k_1! \times \ldots \times k_d!, \quad |\mathbf{k}| = \sum_{i=1}^{d} k_i$$

$$\mathbf{x}^{\mathbf{k}} = (x_1^{k_1} \times \ldots \times x_d^{k_d})$$

$$\sum_{0 \le |\mathbf{k}| \le q} = \sum_{j=0}^{q} \sum_{\substack{k_1=0 \\ k_1 + \ldots + k_d = j}}^{j} \cdots \sum_{k_d=0}^{j},$$

$$(D^{\mathbf{k}} m)(\mathbf{y}) = \frac{\partial^{\mathbf{k}} m(\mathbf{y})}{\partial y_1^{k_1} \ldots \partial y_d^{k_d}}.$$

In this case, we want the minimizer of the multivariate least squares

$$\sum_{i=1}^{n} \left(Y_i - \boldsymbol{\beta}'(\mathbf{X}_i - \mathbf{x})^{\mathbf{k}}\right)^2 K_{H_n}(\mathbf{X}_i - \mathbf{x}), \tag{1.10}$$

where H_n is the $d \times d$ bandwidth matrix, assumed to be symmetric and positive definite and the kernel K_{H_n} is now a d-variate form of the kernel K. The kernel can have many forms, including product of univariate kernels or d-variate probability density functions, but usual assumptions are that $\int K(\mathbf{u})d\mathbf{u} = 1$ and $K_{H_n}(\mathbf{u}) = |H_n|^{-1/2}K(H_n^{-1/2}\mathbf{u})$.

From the minimization of (1.10), we obtain the estimate $\hat{\boldsymbol{\beta}}$, and using this estimate we can compute the estimated $\mathbf{k}$-th order derivative of $m(\mathbf{x})$ as follows (Zhang and Chan, 2011)

$$\hat{m}^{(\mathbf{k})}(\mathbf{x}) = \prod_{j=1}^{d} k_j! \hat{\beta}_{k_1,\ldots,k_d}(\mathbf{x}),$$

where in particular $m(\mathbf{x})$ is estimated by $\hat{m}(\mathbf{x}) = \hat{m}^{0,\ldots,0}(\mathbf{x}) = \hat{\beta}_{0,\ldots,0}(\mathbf{x})$.

The compute the asymptotic bias and variance of the $\mathbf{k}$-th partial derivative $\hat{m}^{(\mathbf{k})}(\mathbf{x})$ (Masry 1996, Masry 1996, Masry 1999), first let

$$N_i = \begin{pmatrix} i+d-1 \\ d-1 \end{pmatrix}$$

be the number of distinct d-tuples $\mathbf{k}$ with $|\mathbf{k}| = i$, and let g_i^{-1} denote the one-to-one map that arranges vectors in lexographical order (for instance $(0,\ldots,0,i)$ will be maped to $(i,0,\ldots,0)$. Also define

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{0,0} & \mathbf{M}_{0,1} & \ldots & \mathbf{M}_{0,q} \\ \mathbf{M}_{1,0} & \mathbf{M}_{1,1} & & \mathbf{M}_{1,q} \\ \vdots & & & \vdots \\ \mathbf{M}_{q,0} & \mathbf{M}_{q,1} & & \mathbf{M}_{q,q} \end{pmatrix},$$

and

$$\Gamma = \begin{pmatrix} \Gamma_{0,0} & \Gamma_{0,1} & \dots & \Gamma_{0,q} \\ \Gamma_{1,0} & \Gamma_{1,1} & & \Gamma_{1,q} \\ \vdots & & & \vdots \\ \Gamma_{q,0} & \Gamma_{q,1} & & \Gamma_{q,q} \end{pmatrix},$$

where $\mathbf{M}_{i,j}$ is the $N_i \times N_j$ matrix with (ℓ, m) elements $\mu_{g_i(\ell)+g_j(m)}$, and $\Gamma_{i,j}$ is the $N_i \times N_j$ matrix with (ℓ, m) elements $\gamma_{g_i(\ell)+g_j(m)}$. Note that there two matrices are multivariate moments of K and K^2, $M_{i,j} = \int \mathbf{u}^{(i+j)} K(\mathbf{u}) d\mathbf{u}$ and $\Gamma_{i,j} = \int \mathbf{u}^{(i+j)} K^2(\mathbf{u}) d\mathbf{u}$.

For a diagonal bandwidth matrix with elements equal to a single value h_n, Masry (1996) showed that the bias and variance of $\hat{m}^{(\mathbf{k})}(\mathbf{x})$ are

$$\begin{aligned} \text{Bias}\left\{\hat{m}^{(\mathbf{k})}(\mathbf{x})|\mathbf{X}_1, \dots, \mathbf{X}_n\right\} &= \mathbf{k}! \left(\mathbf{M}^{-1}\mathbf{B} m_{q+1}(\mathbf{x})\right)_{|\mathbf{k}|} h_n^{q+1-|k|} \\ \text{Var}\left\{\hat{m}^{(\mathbf{k})}(\mathbf{x})|\mathbf{X}_1, \dots, \mathbf{X}_n\right\} &= \frac{\sigma^2(\mathbf{x})(\mathbf{k}!)^2}{f(\mathbf{x})} \left(\mathbf{M}^{-1}\Gamma\mathbf{M}^{-1}\right)_{|\mathbf{k}|,|\mathbf{k}|} \frac{1}{n h_n^{d+2|\mathbf{k}|}} \end{aligned}$$

where $m_{q+1}(\mathbf{x}) = vech\left(\hat{m}^{(\mathbf{k})}(\mathbf{x})\right)$ for the operator $vech$ being the half-vectorization operator,

$$\mathbf{B} = [\mathbf{M}_{0,q+1}, \dots, \mathbf{M}_{q,q+1}]^T,$$

$\left(\mathbf{M}^{-1}\mathbf{B} m_{q+1}(\mathbf{x})\right)_{|\mathbf{k}|}$ is the $|\mathbf{k}|$-th element of the vector $\mathbf{M}^{-1}\mathbf{B} m_{q+1}(\mathbf{x})$ and $\left(\mathbf{M}^{-1}\Gamma\mathbf{M}^{-1}\right)_{|\mathbf{k}|,|\mathbf{k}|}$ is the $(|\mathbf{k}|, |\mathbf{k}|)$ diagonal element of the matrix $\mathbf{M}^{-1}\Gamma\mathbf{M}^{-1}$.

For a local linear approximation, under the assumptions that $n^{-1}|H_n|$ and each entry of H_n goes to zero, K has compact support, $\int \mathbf{u}\mathbf{u}' K(\mathbf{u}) d\mathbf{u} = \mu_2(K) I$, where I is the identity matrix, and that σ and f are continuously differentiable and all second order derivatives of $m(.)$ are continuous we have that (Ruppert and Wand 1994)

$$E(\hat{m}(\mathbf{x}, H_n) - m(\mathbf{x})|\mathbf{X}_1, \dots, \mathbf{X}_n) = \frac{1}{2}\mu_2(K) tr\{H_n \mathcal{H}_m(\mathbf{x})\} + o_p(tr\{H_n\}) \quad (1.11)$$

$$Var(\hat{m}(\mathbf{x}, H_n)|\mathbf{X}_1, \dots, \mathbf{X}_n) = \{n^{-1}|H_n|^{-1/2} R(K)/f(\mathbf{x})\}\sigma^2(\mathbf{x})\{1 + o_p(1)\} \quad (1.12)$$

where $R(K) = \int K(u)^2 du$ and $\mathcal{H}_m$ is a $d \times d$ Hessian matrix of m at $\mathbf{x}$.

1.1.2 Kernel Estimation

We will call Kernel regression estimator the local polynomial estimator when $p = 0$. That is, what we call kernel regression is a special case of local polynomial regression, and moreover, it is in fact the Nadaraya-Watson Kernel regression estimator.

In the univariate case, i.e. when X has dimension one, we have that

$$\hat{m}(x; 0, h_n) = \sum_{j=1}^{n} \frac{K_{h_n}(X_j - x)}{\sum_{l=1}^{n} K_{h_n}(X_l - x)} Y_j = \sum_{j=1}^{n} w_j(x) Y_j. \tag{1.13}$$

The multivariate case is just an extension of the univariate case, again a special case of the multivariate local linear regression described above:

$$\hat{m}(\mathbf{x}; 0, H_n) = \sum_{j=1}^{n} \frac{K_{H_n}(\mathbf{X}_j - \mathbf{x})}{\sum_{l=1}^{n} K_{H_n}(\mathbf{X}_l - \mathbf{x})} Y_j = \sum_{j=1}^{n} w_j(\mathbf{x}) Y_j. \tag{1.14}$$

Note that a common technique for estimating the regression function for the multivariate case is to make use of a product kernel with a diagonal bandwidth, so that one covariate is considered at a time. The product kernel is defined as

$$K_{H_n}(\mathbf{X}_j - \mathbf{x}) = \prod_{k=1}^{d} K_{h_{k,n}}(X_{jk} - x_k), \tag{1.15}$$

and therefore the estimated regression function is

$$\hat{m}(\mathbf{x}; 0, H_n) = \sum_{j=1}^{n} \frac{\prod_{k=1}^{d} K_{h_{k,n}}(X_{jk} - x_k)}{\sum_{l=1}^{n} \prod_{k=1}^{d} K_{h_{k,n}}(X_{lk} - x_k)} Y_j. \tag{1.16}$$

Sometimes, when estimating the regression function on a specific point of $\mathbf{X}$ which was observed, it is possible to leave that observation out of the estimator, i.e., all the points in the neighborhood h_n are used except for the point itself. It is clear that some information is lost, but when dealing with asymptotic properties

this is negligible since we choose the neighborhood $h_n \to 0$ in a rate such that $nh_n \to \infty$, in other words, the number of observations in the neighborhood goes to infinity. In this case, for example we would have

$$\hat{m}(\mathbf{X}_i; 0, H_n) = \sum_{j \neq i}^{n} \frac{K_{H_n}(\mathbf{X}_j - \mathbf{X}_i)}{\sum_{l \neq i}^{n} K_{H_n}(\mathbf{X}_l - \mathbf{X}_i)} Y_j = \sum_{j=1}^{n} w_j(\mathbf{x}) Y_j. \tag{1.17}$$

1.1.3 Nearest Neighbor

The nearest neighbor method for regression is very simple and in fact is very similar to the kernel method as it is a weighted average of the values on a neighborhood of $\mathbf{x}$, only here we always have the same number of neighbors for each $\mathbf{x}$ to average, while in kernel regression there may be more or less values in the neighborhood of size h_n depending on where $\mathbf{x}$ is located.

Note that in the tails of the density of X, the kernel estimator will probably have less points to average, since there are fewer observations there, while the nearest neighbor method will still average the fixed number of neighbors (say) k. Therefore the bias of the nearest neighbor method is larger in the tails but the variance smaller, and it also has a smoother estimated mean function in the tails.

This method would estimate m with a weighted average of the values of the response corresponding to the k nearest neighbors of $\mathbf{x}$

$$\hat{m}(\mathbf{x}) = \sum_{i=1}^{n} \frac{w\left(\frac{||\mathbf{X}_i - \mathbf{x}||}{||\mathbf{X}_{(k)} - \mathbf{x}||}\right) Y_i}{\sum_{l=1}^{n} w\left(\frac{||\mathbf{X}_l - \mathbf{x}||}{||\mathbf{X}_{(k)} - \mathbf{x}||}\right)} \tag{1.18}$$

for $\int_{\mathbb{R}^d} w(||u||)du = 1$ and $\mathbf{X}_{(k)}$ being the k-th nearest neighbor of $\mathbf{x}$. And as a simple average we would have

$$\hat{m}(\mathbf{x}) = \frac{1}{k} \sum_{i=1}^{n} I(||\mathbf{X}_i - \mathbf{x}|| \leq ||\mathbf{X}_{(k)} - \mathbf{x}||) Y_i. \tag{1.19}$$

Again, the idea of not using the exact observed point $\mathbf{X}_i$ in the estimation of $m(\mathbf{X}_i)$ is possible, where in this case all the k neighbors would be present in the estimation, but not the value of the corresponding observation. The loss of this is negligible asymptotically, similarly to what we described in the kernel method.

1.1.4 Estimation of Additive Models and the Backfitting Algorithm

Consider the additive model 1.4, where we assume that the Y_i have been centered around their mean. Note that in this model, for any $k = 1 \dots d$,

$$E(Y - \sum_{j \neq k}^{d} m_j(X_j)|X_k) = m_k(X_k). \tag{1.20}$$

Following Opsomer 2000, let $\mathbf{m}_j = (m_d(X_{j1}), \dots, m_j(X_{jn}))^T$ be the vector of the function m at the observed points. These additive components are estimated by solving the set of normal equations

$$\begin{bmatrix} I & \mathbf{S}_1 & \dots & \mathbf{S}_1 \\ \mathbf{S}_2 & I & \dots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_d & \mathbf{S}_d & \dots & I \end{bmatrix} \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \vdots \\ \mathbf{m}_d \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_d \end{bmatrix} Y,$$

where $\mathbf{S}_j$ is the linear smoother matrix with respect to the $\mathbf{X}_j = (X_{j1}, \dots, X_{jn})$ covariate vector.

Intuitively, we can think of an iterative algorithm to solve these equations, and it is called the backfitting algorithm, which will converge to the solution

$$\begin{bmatrix} \hat{\mathbf{m}}_1 \\ \hat{\mathbf{m}}_2 \\ \vdots \\ \hat{\mathbf{m}}_d \end{bmatrix} = \begin{bmatrix} I & \mathbf{S}_1 & \dots & \mathbf{S}_1 \\ \mathbf{S}_2 & I & \dots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_d & \mathbf{S}_d & \dots & I \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_d \end{bmatrix} Y = M^{-1}CY,$$

if M^{-1} exists.

Hence, the backfitting algorithm can be described in the following way

---

- 1) Initialize: $\hat{\mu} = \bar{Y}$, set $m_j = 0$, $j = 1 \dots p$
- 2) for $j = 1 \dots d$
  $\hat{m}_k = S_k(Y - \hat{\mu} - \sum_{j \neq k}^{d} m_j|X_k)$

- repeat 2) until convergence

In each step of the algorithm, we "readjust" the estimated m_j, after we remove the effects of the other predictors, taking advantage of the additive effects with the partial residuals in each iteration. This algorithm is similar to the well known Gauss-Seidel algorithm, which iterates the algorithm in the same way, readjusting each step. Here we set the initial functions m_j to be equal to 0, but a better idea would be to start these functions with some previous knowledge or even a simple fit (linear regression for example).

Continuing with Opsomer's notation, in order to have a nice expression for each $\hat{m}_j$, define the additive smoother matrix $W_j = E_j M^{-1} C$, where E_j is a partitioned $n \times nd$ matrix with the j-th block being a nxn identity matrix and zeros elsewhere. In this way, we have $\hat{m}_j = W_j Y$ and $\hat{m} = \sum_{j=1}^{d} \hat{m}_j = WY$, for $W = W_1 + \ldots + W_d$.

1.1.5 Global Approximations

Other smoothing methods can be found in the literature, as smoothing splines, regression splines, penalized splines and orthogonal basis expansion as Fourier regression and wavelets. We will not get into the details of each of these methods here because they are not used in the proof of the theorems we introduce, even though these theorems may hold for other types of estimators.

Chapter 2

Hypothesis Testing in Nonparametric Regression: Literature Review

In multivariate regression analysis, one of the most important questions that arise, is whether a single covariate or a set of covariates is significant for the model. This significance test is applied in parametric or nonparametric regression, as reducing the dimension by excluding predictors leads to better estimators and easier interpretation. This is specially important in nonparametric regression, because of its big flexibility, it may not be easy to interpret models with many covariates. Therefore, the less covariates in the model, the simpler it is to summarize, interpret and visualize (in graphs for instance). In this context, we need to formalize the model simplification, i.e., excluding covariates from the model, with a statistical hypothesis test.

A few authors have considered this type of hypothesis test, investigating the residual sum of squares from kernel regressions, using partial derivatives, or comparing the regression of Y on all covariates with the regression of Y on the covariates under the null hypothesis.

Formally, let $\mathbf{U} = (\mathbf{X}, \mathbf{Z})$ be the vector of d available predictors, with $\mathbf{X}$ being d_1-dimensional and $\mathbf{Z}$ d_2-dimensional. For the nonparametric model

$$Y = m(\mathbf{X}, \mathbf{Z}) + \sigma(\mathbf{X}, \mathbf{Z})\epsilon,$$

the goal is to assess the usefulness of $\mathbf{Z}$, i.e. test the hypothesis $H_0 : m(\mathbf{x}, \mathbf{z}) = m_1(\mathbf{x})$.

Racine (1997) introduced a test based on the following idea

$$E(Y|\mathbf{U}) \perp \mathbf{U}_{(j)} \Leftrightarrow \frac{E(Y|\mathbf{U})}{\partial \mathbf{U}_{(j)}} = 0 \text{ a.s.},$$

where $\mathbf{U}_{(j)} \subset \mathbf{U}$ is a subset of predictors, $\perp$ denotes orthogonality or independence and $E(Y|\mathbf{U})/\partial \mathbf{U}_{(j)} \in \mathbb{R}^j$. Thus, to test the significance of this set of predictors, the null hypothesis can be written as

$$H_0 : \frac{E(Y|\mathbf{U})}{\partial \mathbf{U}_{(j)}} = 0 \text{ for all } \mathbf{u} \in \mathbf{U}.$$

He proceeds by estimating the unknown derivatives with nonparametric methods. However, to conduct tests based on this test statistic, its null distribution has to be obtained through bootstrap, as the asymptotics are not known.

Another class of procedures is based on the weighted differences of the nonparametric regressions under null and alternative hypothesis, that is, the idea that the null hypothesis residuals, $\xi = Y - m_1(\mathbf{X})$, satisfy $E(\xi|\mathbf{U}) = 0$ under H_0 and $E(\xi|\mathbf{U}) = m(\mathbf{U}) - m_1(\mathbf{X})$ under the alternative. Thus, $E(\xi E(\xi|\mathbf{U})) = E([E(\xi|\mathbf{U})]^2) = E\left([m(\mathbf{U}) - m_1(\mathbf{X})]^2\right) > 0$ under the alternative and zero under the null. To avoid the randomness in the denominator of the kernel estimation, Fan and Li (1996) proposed a test statistic with weigths based on the density of $\mathbf{U}$. Their approach is based on estimating $E[\xi f_{\mathbf{X}}(\mathbf{X}) E(\xi f_{\mathbf{X}}(\mathbf{X})|\mathbf{U}) f_{\mathbf{U}}(\mathbf{U})]$ which equals $E[(m(\mathbf{U}) - m_1(\mathbf{X}))^2 f_{\mathbf{X}}(\mathbf{X})^2 f_{\mathbf{U}}(\mathbf{U})]$ under the alternative and zero under the null. Their statistic is defined as

$$\frac{1}{n} \sum_i [\tilde{\xi}_i \tilde{f}_{\mathbf{X}}(\mathbf{X}_i)] \left[\frac{1}{(n-1)h_n^d} \sum_{j \neq i} [\tilde{\xi}_j \tilde{f}_{\mathbf{X}}(\mathbf{X}_j)] K\left(\frac{\mathbf{U}_i - \mathbf{U}_j}{h_n} \right) \right]$$

where $\tilde{f}_{\mathbf{X}}$ is the estimated density of $\mathbf{X}$, $\tilde{\xi}_i$ is the estimated residuals under the null hypothesis, and K is a kernel function. They show that this is asymptotically normal distributed under H_0.

Lavergne and Voung (2000) propose a test statistic based on a different esti-

mator of the same quantity as Fan and Li (1996), which is

$$\frac{(n-4)!}{n!}\sum_a (Y_i - Y_k)(Y_j - Y_l) L_n\left(\frac{\mathbf{X}_i - \mathbf{X}_k}{g_n}\right) L_n\left(\frac{\mathbf{X}_j - \mathbf{X}_l}{g_n}\right) K_n\left(\frac{\mathbf{U}_i - \mathbf{U}_j}{h_n}\right),$$

where $\sum_a$ is the sum over all permutations of 4 distinct elements chosen from n, $L_n = g_n^{-d_1} L$ for a kernel L on $\mathbf{R}^{d_1}$ and $K_n = h_n^{-d} K$ for a kernel K on $\mathbf{R}^d$. Lavergne and Voung (2000) show that their test statistic is also asymptotically normal under H_0.

A related class of procedures is based on direct estimation of $E[(m(\mathbf{U}) - m_1(\mathbf{X}))^2 W(\mathbf{U})]$, for some weight function W. For example, Aït-Sahalia, Bickel and Stoker (2001) propose a test statistic based on

$$\frac{1}{n}\sum_{i=1}^{n} (\hat{m}(\mathbf{U}_i) - \hat{m}_1(\mathbf{X}_i))^2 W(\mathbf{U}).$$

This estimation is more general than that considered by Fan and Li (1996) and Lavergne and Voung (2000), as these are a special case of weights when $W(\mathbf{U}) = f_{\mathbf{X}}(\mathbf{X})^2 f_{\mathbf{U}}(\mathbf{U})$. Choosing the weights permits testing to be performed on analyst-specified subsets of the data, but these weights have to be chosen for each case. The asymptotic distribution of their test statistic is shown to be normal.

Delgado et al. (2001) propose a test that requires only the estimation of the nonparametric regression under the null hypothesis, assuming that $\mathbf{X}$ has a density $f_{\mathbf{X}}(\mathbf{x}) > 0$. The idea is that

$$H_0 : E(Y - m_1(\mathbf{X})|\mathbf{U}) = 0 \text{ a.s.},$$

is equivalent to

$$f_{\mathbf{X}}(\mathbf{X}) E(Y - m_1(\mathbf{X})|\mathbf{U}) = 0 \text{ a.s.},$$

or

$$T(\mathbf{U}) = 0 \text{ a.s.},$$

where $T(\mathbf{x}) = E\{f_{\mathbf{X}}(\mathbf{X})[Y - m_1(\mathbf{X})]\mathbf{1}(\mathbf{U} \leq \mathbf{u})\}$. Thus, their test statistic based

on

$$\begin{aligned} T_n(\mathbf{u}) &= \frac{1}{n}\sum_i \hat{f}_{\mathbf{X}}(\mathbf{X}_i)(Y_i - \hat{m}_1(\mathbf{X}_i))\mathbf{1}(\mathbf{U}_i \leq \mathbf{u}_i) \\ &= \frac{1}{n^2}\sum_i \sum_j \frac{1}{a_1^d} K_{ij}(Y_i - Y_j)\mathbf{1}(\mathbf{U}_i \leq \mathbf{u}_i). \end{aligned}$$

In this case, the asymptotic tests are difficult to implement, since the asymptotic distribution of the test depends on properties of the distribution of $(Y, \mathbf{U})$ that are unknown. For that matter, to perform the test they have make use of bootstrap methods to find critical values for the test.

An important property of hypothesis tests in nonparametric regression, is the Wilks' Phenomenon (Fan, Zhang and Zhang 2001). This means that the asymptotic null distribution of the test statistics does not depend on nuisance parameters and functions. For example, if the regression model is $Y_i = m_1(X_i) + m_2(Z_i) + \epsilon_i$ and we want to test $H_0 : m_2(z) = 0$ vs $H_a : m_2(z) \neq 0$, a statistical test that has asymptotic null distribution that does not depend on the nuisance function $m_1(x)$ would indicate the Wilks' Phenomenon.

In the next section we will introduce in detail one of the most recent methods for model checking in nonparametric regression, which has been frequently used recently.

2.1 The Generalized Likelihood Ratio Test

The Generalized Likelihood Ratio (GLR) test was proposed by Fan, Zhang and Zhang (2001). The test statistic is based on replacing the maximum likelihood estimator of the regression function, that may not exist in the nonparametric regression setting, by a reasonable nonparametric regression estimator.

The idea comes from the well known maximum likelihood ratio test

$$\lambda_n = 2\left(\max_{\theta\in\Theta} \ell_n(\theta) - \max_{\theta\in\Theta_0} \ell_n(\theta)\right) \tag{2.1}$$

where $\ell(\theta)$ is the logarithm of the likelihood function, and the test will reject the null hypothesis when the alternative model has a larger probability of having

generated the data.

The biggest problem in extending this idea to the nonparametric setting is that the maximum likelihood estimators may not exist. In simple scenarios where it is known that these estimators exist, there exists the nonparametric likelihood ratio test, which is just a simple extension of the Likelihood Ratio test. However, this simple extension of the test is not optimal due to the many restrictions on the choice of the smoothing parameters (see Fan et al. (2001)).

Note that under both the null hypothesis and the alternative hypothesis, the regression model is nonparametric, therefore, in the generalized likelihood ratio test, the maximum likelihood estimator (mle) under the alternative hypothesis is replaced by a nonparametric regression estimator, so that the GLR test statistic is

$$\lambda_n = \ell_n(H_1) - \ell_n(H_0). \tag{2.2}$$

One of the most interesting properties of the generalized likelihood ratio tests is that its asymptotic distribution under the null hypothesis is independent of the nuisance parameters or functions. However, it may depend on the nonparametric method that is being used to estimate the functions. Thus the GLR method has the Wilks Phenomenon. In this case, the test and decision to reject the null hypothesis or not can be made by comparing likelihoods, and the critical values are easily extracted from the null asymptotic distribution, which can be simulated by setting the nuisance parameters or functions at reasonable estimates, hence simplifying a lot the computations since theoretical constants of the test are sometimes hard to derive.

2.1.1 Hypothesis Testing in Additive Models with the Generalized Likelihood Ratio test

Fan and Jiang (2005) explored the GLR test for additive models. The idea is to make use of the backfitting algorithm to find the nonparametric estimators under the null and alternative hypothesis and then use them in the likelihood.

Suppose we have the additive model

$$Y = m_1(X_1) + m_2(X_2) + ... + m_d(X_{d_1}) + m_z(Z) + \epsilon,$$

for the covariates $\mathbf{U} = (X_1 \ldots, X_{d_1}, Z)$, and we want to test the following hypothesis

$$H_0 : m_z(z) = 0 \text{ vs } H_1 : m_z(z) \neq 0. \tag{2.3}$$

This hypothesis test is one of the first and most intuitive tests after fitting the regression model. Note that under the alternative hypothesis we have the full model, and under the null hypothesis we have

$$Y = m_1(X_1) + m_2(X_2) + ... + m_{d_1}(X_{d_1}) + \epsilon, \tag{2.4}$$

so that under null and alternative hypothesis the model is nonparametric.

In general nonparametric regression, the distribution of the error term ϵ is not necessarily specified, but in order to derive a likelihood test, some kind of distribution is needed, hence we first pretend that the errors are normal distributed with zero mean and variance σ^2. In this case the log-likelihood under the alternative is

$$-\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{d_1} m_j(X_{ji}) - m_z(Z_i)\right)^2.$$

Now, replacing the functions m by the estimated values we have

$$-\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}RSS_1,$$

where $RSS_1 = \sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{d_1}\hat{m}_j(X_{ji}) - \hat{m}_z(Z_i)\right)^2$. The derivative with relation to σ^2 is $-\frac{n}{2\sigma^2} - \frac{RSS_1}{2\sigma^4}$ so that $\hat{\sigma}^2 = \frac{RSS_1}{n}$ and the maximization leads to $-\frac{n}{2}log(2\pi/n) - \frac{n}{2}log(RSS_1) - \frac{n}{2}$, which up to a constant is

$$\ell(H_1) = -\frac{n}{2}log(RSS_1).$$

For the model under the null hypothesis (2.4), the same procedure can be applied and the log likelihood will be

$$\ell(H_0) = -\frac{n}{2}log(RSS_0),$$

where $RSS_0 = \sum_{i=1}^{n}(Y_i - \sum_{j=1}^{d_1} \tilde{m}_j(X_{ji}))$, and the same backfitting algorithm and bandwidths are used to estimate $\tilde{m}_j(X_j)$ under H_0. Therefore, the GLR statistic (Fan and Jiang (2005)) is

$$\lambda_n(H_0) = (\ell(H_1) - \ell(H_0)) = \frac{n}{2} log(\frac{RSS_0}{RSS_1}) = \frac{n}{2}\frac{RSS_0 - RSS_1}{RSS_1}. \tag{2.5}$$

Under some conditions on the kernel, on density of X, on the bandwidths and errors, Fan and Jiang showed that the GLR statistic converges to a normal distribution as n gets large, that is,

$$P(\sigma_n^{-1}(\lambda_n(H_0) - \mu_n - d_{1n}) < t|\mathbf{U}) \to \Phi(t), \tag{2.6}$$

for some constants σ_n^{-1}, μ_n and d_{1n} that depend on n. If an extra condition on the bandwidths holds, then they show that conditionally on $\mathbf{U}$

$$r_K \lambda_n(H_0) \sim^{approx} \chi^2_{r_K \mu_n}, \tag{2.7}$$

for some constant r_K depending on the kernels and bandwidths.

They also show that the Wilks type of phenomenon also occurs in this case, the asymptotic distribution of the statistic does not depend on the nuisance parameters or functions, and therefore they can use bootstrap methods to simulate the null distributions of the GLR tests, avoiding very difficult theoretical results.

It is also possible to extend the GLR test for the case where the dimension of $\mathbf{Z}$ is $d_2 > 1$, i.e.,

$$H_0 : m_{Z_1}(Z_1) = \ldots = m_{Zd_2}(Z_{d_2}) = 0,$$

for the additive model

$$Y = m_1(X_1) + m_2(X_2) + ... + m_d(X_{d_1}) + m_{Z_1}(Z_1) + \ldots m_{Z_{d_2}}(Z_{d_2}) + \epsilon,$$

where the test statistic will have the same type of asymptotic properties for some different constants.

Chapter 3

ANOVA-type Nonparametric Diagnostic Test for Regression Models: Fixed Design

Recently, several authors considered the case of ANOVA hypothesis testing where the number of factor levels a goes to infinity. In this case, it is known that the asymptotic distribution of the test statistic F = MST/MSE can be found by assessing the asymptotic distribution of $a^{1/2}(F-1)$. Hence, it is only necessary to study the asymptotic distribution of $a^{1/2}(MST - MSE)$, for MSE goes in probability to a constant, so that by Slutsky theorem these asymptotic distributions are the same.

Akritas and Papadatos (2004) extended this idea to the heterocedastic case, for weighted and unweighted statistics. In the simple balanced case, which will be interesting for this thesis, they introduced the following theorem

Theorem 3.1. *(Akritas and Papadatos 2004) Let X_{ij}, $i = 1 \dots a, j = 1 \dots n$ be an iid sequence of random variables with $EX_{ij} = \mu$ and $0 < VarX_{ij} = \sigma^2 < \infty$.*

1. *if $n \leq 2$ remains fixed, then $a^{1/2}(F_a - 1) \to^d N(0, \frac{2n}{n-1})$ as $a \to \infty$.*
2. *if $n = n(a) \to \infty$ as $a \to \infty$, then $a^{1/2}(F_a - 1) \to^d N(0, 2)$ as $a \to \infty$.*

Akritas and Papadatos (2004) also introduced a projection method for

quadratic forms in this setting, a very important result which will be used to derive our statistics. Below we briefly describe this method:

Let $\mathbf{X}_i = (X_{i1}, \ldots, X_{in_i})^T, i = 1, \ldots, a$ be independent random vectors with independent components with (for simplicity) mean 0 and variance 1 $\forall i, j$. Let $\mathbf{X} = (X_{11}, \ldots, X_{1n_1}, \ldots, X_{a1}, \ldots, X_{an_a})^T$, so that MST-MSE can be written as $\mathbf{X}^T\mathbf{A}\mathbf{X}$ where

$$A = \begin{pmatrix} \mathbf{B}_1 & -c_3\mathbf{1}_{n_1}\mathbf{1}_{n_2}^T & \ldots & -c_3\mathbf{1}_{n_1}\mathbf{1}_{n_a}^T \\ -c_3\mathbf{1}_{n_2}\mathbf{1}_{n_1}^T & \mathbf{B}_2 & \ldots & -c_3\mathbf{1}_{n_2}\mathbf{1}_{n_a}^T \\ \vdots & \vdots & \ddots & \vdots \\ -c_3\mathbf{1}_{n_a}\mathbf{1}_{n_1}^T & -c_3\mathbf{1}_{n_a}\mathbf{1}_{n_2}^T & \ldots & \mathbf{B}_a \end{pmatrix} \tag{3.1}$$

where $\mathbf{1}_k = (1, \ldots, 1)^T \in \mathbb{R}^k$ and $\mathbf{B}_i$ are $n_i x n_i$ matrices with elements $b_{i,sk}$ given by

$$b_{i,sk} = \begin{cases} \frac{c_1}{n_i} - c_2 - c_3 & if s = k \\ \frac{c_1}{n_i} - c_3 & if s \neq k \end{cases}$$

and $c_1 = \frac{N-1}{(N-a)(a-1)}, c_2 = \frac{1}{N-a}$ and $c_3 = \frac{1}{N(a-1)}$.

Using a variation of Hajek's projection method, they show that the projection in this case is $\mathbf{X}^T\mathbf{A}_D\mathbf{X}$, where $\mathbf{A}_D = diag\{\mathbf{B}_1, \ldots, \mathbf{B}_a\}$. Note that the expected value of the projection $\mathbf{X}^T\mathbf{A}_D\mathbf{X}$ is the same as the one of $\mathbf{X}^T\mathbf{A}\mathbf{X}$, it is not difficult to show that the diffirence between them, after proper scaling, is small enough in probability, so that their asymptotic distribution is the same. Clearly, it is necessary to prove this for each case, whenever an $\mathbf{X}$ is of any different form, but this theory will be very helpful in further chapters of this thesis.

Using this idea, Wang, Akritas and Keilegom (2008) considered hypothesis testing in the heterocedastic nonparametric regression model

$$Y_{ni} = m(x_{ni}) + \sigma(x_{ni})\epsilon_{ni}, \ i = 1, \ldots, n \tag{3.2}$$

where m is the regression function and $\sigma^2(.)$ is an unknown variance function, but they consider only the case where x is of a fixed design in a bounded interval. Their null hypothesis is

$$H_0 : m(x) = C \text{ for all } x, \tag{3.3}$$

for some constant C.

They introduce a new statistic for the hypothesis test in (3.3), based on the ANOVA developments for the F test suggested by Akritas ad Papadatos. Wang, Akritas and Keilegom (2008) create a vector $\mathbf{V} = (Y_j, j \in W_1, \ldots, Y_j, j \in W_n)^T$, where the windows W_i are composed by the k_n covariate values nearest to x_i. In that case, MST-MSE can be written by $\mathbf{V}^T A \mathbf{V}$, and using the projection method suggested by Papadatos and Akritas (2004) they introduce the following theorem

Theorem 3.2. *Under H_0 and Lipschitz continuity of $m(x)$ and $\sigma(x)$, if $E(\epsilon^4)$ is unif. bounded in n and i then*

1. *if $k_n = k$ is fixed, $\sqrt{n}(MST - MSE) \to N\left(0, \frac{2k(2k-1)}{3(k-1)}\tau^2\right)$*
2. *if $k_n \to \infty$ such that $k_n/n \to 0$, $\sqrt{n/k_n}(MST - MSE) \to N\left(0, \frac{4}{3}\tau^2\right)$*

for some variance τ^2.

Note that the larger k_n is, the larger the dependence, and the slower the convergence to the normal distribution, so k_n plays an important role in determining how good the statistic is when applied to a specific dataset.

It is important to notice though, that if we want to test if m is of some specific form, a parametric form for example, we can just fit the desired form on m, and apply the test described above on the residuals. If for example we want to test if the regression function m is of a parametric form, say $G_\Theta = \{f(., \theta), \theta \in \Theta\}$, then the null hypothesis would be

$$H_0 : m(x) \in G_\Theta. \tag{3.4}$$

In this case, we can estimate $m(x, \theta)$ under the null by $m(x, \hat{\theta})$ and compute the residuals $\hat{e}_i = Y_i - m(x_i, \hat{\theta})$. Under the null hypothesis, the residuals of the fit have the same distribution as $\sigma(x_i)\epsilon_i$. Therefore, the testing problem $H_0 : m(x) \in G_\Theta$ is equivalent to the testing problem $H_0 : m_2(x) = C$ for the regression of $\hat{e}_i$ on x_i, using the theory described above for the model to

$$\hat{e}_i = m_2(x_i) + \sigma_2(x_i)\gamma_i. \tag{3.5}$$

Wang, Akritas and Keilegom show that this test has the same asymptotic distribution as the test in Theorem 3.2, except that for the second part the bandwidth must be such that $k_n^{3/2}/n \to 0$.

Chapter 4

ANOVA-type Hypothesis Test for Nonparametric Regression

The idea introduced by Wang, Akritas and Van Keilegom (2008) opens a new perspective for hypothesis testing in nonparametric regression. The restriction of their work is that not only they consider the design of the covariates to be fixed, but they also develop the theory for just one covariate. Here we will extend their approach and develop a methodology of hypothesis testing in a multi-dimensional covariate model where the predictors are random variables.

Specifically, the goal is to develop a hypothesis test for a predictor or a set or predictors that can handle completely nonparametric models. We are interested in models with high flexibility, where not even additivity is assumed. Clearly, the trade off between flexibility and precision may come into play, and we will study the properties of the test we propose here in relation to the already existing tests for a comparison.

In this chapter, a new methodology for hypothesis test is described using a test statistic based on the ideas of the ANOVA-type test described in Chapter 3. The asymptotic distribution of the test statistic is evaluated under the null hypothesis and under local alternatives.

If we consider the additive model (1.4) of Hastie and Tibshrani (1990), a testing hypothesis of the form

$$H_0 : m_1(.) = 0,$$

could be, motivated by distances between estimators under the null and alternative models, tackled by a test statistic based on a discrepancy method: $T = c_1||\hat{m}_1||$. For this, we have to choose the norm $||.||$ and the weight c_1, and yet, the null distribution of T is in general unknown and depends critically on the nuisance functions $m_2, \ldots, m_d$. Moreover, this procedure is based on the assumption that the model is additive in the predictor effects.

We consider the nonparametric regression model in its most general form, and we only make assumptions on m corresponding to its smoothness. In this case we denote the regression function by $m(\mathbf{X}, \mathbf{Z}) = E(Y|\mathbf{X}, \mathbf{Z})$ and we introduce a test statistic to test the significance of $\mathbf{Z}$.

In Section 4.1, we formally describe the model, introduce the hypothesis test and the test statistic for the univariate case.

4.1 ANOVA-type hypothesis test for univariate X and univariate Z

Suppose we have n observations of the random variable Y and predictors $\mathbf{U} = (X, Z)$. Assume also that X and Z have densities $f_X(x)$ and $f_Z(z)$ respectively with support $S_{X,Z}$, and that $f_X(x)$ and $f_Z(z)$ are continuous and bounded away from 0. Denote the expected value of the response variable Y given the predictors by the unknown regression function $m(x, z) = E(Y|X = x, Z = z)$, and define

$$\zeta = Y - m(\mathbf{U}). \tag{4.1}$$

From its definition if follows that $E(\zeta|\mathbf{U}) = E(\zeta) = 0$. Setting $\sigma^2(\mathbf{U}) = \text{Var}(\zeta|\mathbf{U})$, we have the model

$$Y = m(\mathbf{U}) + \sigma(\mathbf{U})\epsilon, \tag{4.2}$$

where ϵ is the standardized error ζ. Based on a sample $(Y_i, \mathbf{U}_i), i = 1, \ldots, n$, of iid observations from model (4.2), without loos of generality, we will consider testing the hypothesis that the regression function does not depend on the Z. Setting

$E(Y|X) = m_1(X)$, the hypothesis we will consider can be written as

$$H_0 : m(x, z) = m_1(x). \tag{4.3}$$

Note that this hypothesis does not require the variance function $\sigma^2(.,.)$ be a function only of x, but requires only the expected value of Y given the covariates be $E(Y|X = x, Z = z) = m_1(x)$.

Let now $m_1(X_i) = E(Y|X_i)$, as before, and define the null hypothesis residuals as

$$\xi_i = Y_i - m_1(X_i). \tag{4.4}$$

Since under the null hypothesis (4.3) $m_1(X_i) = m(U_i)$, it follows that the null hypothesis residuals in (4.4) equal the residuals defined in (4.1) and thus

$$E\left(\xi_i|Z_i\right) = 0. \tag{4.5}$$

Note that $\xi_i = Y_i - m_1(X_i) = \sigma(X_i, Z_i)\epsilon_i$, so that in order to estimate ξ_i we need an estimator of m_1. Consider the Nadaraya-Watson Kernel estimator

$$\hat{m}_1(X_i) = \sum_{j=1}^{n} \left(\frac{K\left(\frac{X_i - X_j}{h_n}\right)}{\sum_{l=1}^{n} K\left(\frac{X_i - X_l}{h_n}\right)} \right) Y_j, \tag{4.6}$$

where h_n is the bandwidth and the kernel function K is bounded, symmetric and has bounded variation. Hence, the estimated residuals are

$$\hat{\xi}_i = Y_i - \hat{m}_1(X_i).$$

It is clear that, as described in Chapter 3, testing H_0 in (4.3) is equivalent to testing if Z has a significant effect on the residuals ξ.

To construct the test statistic, we consider the approach studied by Akritas and Papadatos (2004), Akritas and Arnold (2000), Boos and Brownie (1995) and Wang, Akritas, Van Keilegom (2008). Consider the ANOVA setting where we let the number of factor levels go to infinity, then finding the asymptotic distribution of the

F statistic F = MST/MSE is equivalent to obtaining the asymptotic distribution of $\sqrt{n}(F-1)$, or of $\sqrt{n}(MST-MSE)$, as MSE tends to a constant. Here we consider each value ξ_i as a factor level. But the basic requirement in the ANOVA approach is that there is more than one observation per cell. In this context, we follow the idea of Wang, Akritas and Van Keilegom (2008): augment each cell of the ANOVA by including $p-1$ $\hat{\xi}_\ell$ which correspond to the $(p-1)/2$ Z_ℓ values that are nearest to Z_i. To be specific, we consider the pairs $(\hat{\xi}_i, Z_i)$, $i = 1, \ldots, n$, arranged so that $Z_{i_1} < Z_{i_2}$ whenever $i_1 < i_2$, and for each Z_i, $(p-1)/2 < i \leq n-(p-1)/2$, define the nearest neighbor window W_i as

$$W_i = \left\{ j : |\hat{F}_Z(Z_j) - \hat{F}_Z(Z_i)| \leq \frac{p-1}{2n} \right\}, \tag{4.7}$$

where $\hat{F}_Z$ is the empirical distribution function of Z. W_i defines the augmented cell corresponding to Z_i. Note that the augmented cells are defined as sets of indices rather than as sets of $\hat{\xi}_i$ values. The vector of $(n-p+1)p$ constructed "observations" in the augmented one-way ANOVA design is

$$\hat{\boldsymbol{\xi}}_V = (\hat{\xi}_j, j \in W_{(p-1)/2+1}, \ldots, \hat{\xi}_j, j \in W_{n-(p-1)/2})^T.$$

Note that, the points of the two edges do not have a neighborhood of the same kind as all the other points, but in practice we can use asymmetric windows for these points (or even exclude them from the augmented ANOVA), and that will have negligible effects on the asymptotic distribution of the test statistic. Obviously, with small data sets these points are very important, and the power of the test will be influenced. Further investigation is described in Section 4.1.5.

Let MST = MST($\hat{\boldsymbol{\xi}}_V$), MSE = MSE($\hat{\boldsymbol{\xi}}_V$) denote the balanced one-way ANOVA mean squares due to treatment and error, respectively, computed on the data $\hat{\boldsymbol{\xi}}_V$. The proposed test statistic is based on

$$MST - MSE. \tag{4.8}$$

Here we note that MSE - MSE can be written in the quadratic form $\hat{\boldsymbol{\xi}}_V^T A \hat{\boldsymbol{\xi}}_V$,

where

$$A = \frac{np-1}{n(n-1)p(p-1)}\left(\oplus_{i=1}^{n}\mathbf{J}_p\right) - \frac{1}{n(n-1)p}\mathbf{J}_{np} - \frac{1}{n(p-1)}\mathbf{I}_{np}, \tag{4.9}$$

$\mathbf{I}_d$ is a identity matrix of dimension d, $\mathbf{J}_d$ is a $d \times d$ matrix of 1's and $\oplus$ is the Kronecker sum or direct sum.

The test statistic MST - MSE for no factor level effect is an intuitive test for H_0 in (4.3) and the following theorem gives the asymptotic normal distribution of the test statistic under this hypothesis.

Theorem 4.1. *Assume that the marginal densities f_X, f_Z of X, Z, respectively, are bounded away from zero, the second derivatives of f_X and $m_1(x)$ are uniformly continuous and bounded, that $\sigma^2(.,z) := E(\xi^2|Z = z)$ is Lipschitz continuous, $\sup_{\mathbf{u}} \sigma^2(\mathbf{u}) < \infty$, and $E(\epsilon_i^4) < \infty$. Assume that the bandwidth h_n satisfies*

$$nh_n^8 \to 0 \quad \textit{and} \quad \frac{nh_n^2}{(\log n)^2} \to \infty. \tag{4.10}$$

Then, under H_0 in (4.3), the asymptotic distribution of the test statistic in (4.8) is given by

$$n^{1/2}(MST - MSE) \xrightarrow{d} N\left(0, \frac{2p(2p-1)}{3(p-1)}\tau^2\right),$$

where $\tau = \int \left[\int \sigma^2(x,z) f_{X|Z=z}(x)dx_1\right]^2 f_Z(z)dz$.

The proof of Theorem 4.1 follows from the proof of the multivariate version (see Theorem 4.2).

An estimate of τ^2 can be obtained by modifying Rice's (1984) estimator in the following way:

$$\hat{\tau}^2 = \frac{1}{4(n-3)}\sum_{j=2}^{n-2}(\hat{\xi}_j - \hat{\xi}_{j-1})^2(\hat{\xi}_{j+2} - \hat{\xi}_{j+1})^2. \tag{4.11}$$

Note that the range of the bandwidth is restricted to having asymptotic property $h_n \in (n^{-1/2}, n^{-1/8})$. The general rule of thumb for optimal estimation in the univariate case (X with dimension one) is to use $sd(x)n^{-1/5}$, so the assumptions on the bandwidth that are made in the theorem are flexible enough to cover the optimal rate.

4.1.1 Factorization for two-dimensions

To fully appreciate the hypothesis test and the test statistic for this nonparametric regression, we describe the factorization of the regression function, using ANOVA-type ideas. Recall that, for the case where X and Z are univariate, the regression model is

$$Y_i = m(X_i, Z_i) + \sigma(X_i, Z_i)\epsilon_i, \;\; i = 1...n,$$

where the mean function $m(X, Z) = E(Y|X, Z)$. Consider a factorization of this mean function using a similar approach as the one used in a two-way ANOVA:

$$m(X, Z) = \mu + \tilde{m}_1(X) + \tilde{m}_2(Z) + \tilde{m}_{12}(X, Z). \tag{4.12}$$

Now, similarly to the usual assumptions of a two-way ANOVA, the effects are such that

$$\begin{aligned}
\mu &= \int\int m(x, z)dF_X(x)dF_Z(z), \\
\tilde{m}_1(X) &= \int m(X, z)dF_Z(z) - \mu, \\
\tilde{m}_2(Z) &= \int m(x, Z)dF_X(x) - \mu, \\
\tilde{m}_{12}(X, Z) &= m(X, Z) - \mu - \tilde{m}_1(X) - \tilde{m}_2(Z),
\end{aligned}$$

where

$$\begin{aligned}
\int \tilde{m}_1(x)dF_X(x) &= \int\int m(x, z)dF_Z(z)dF_X(x) - \mu = 0, \\
\int \tilde{m}_2(z)dF_Z(z) &= \int\int m(x, z)dF_X(x)dF_Z(z) - \mu = 0, \\
\int \tilde{m}_{12}(x, Z)dF_X(x) &= 0, \\
\int \tilde{m}_{12}(X, z)dF_Z(z) &= 0.
\end{aligned}$$

Now, the null hypothesis for univariate x and univariate z is

$$H_0 : m(x, z) = m_1(x), \tag{4.13}$$

where we can write $m_1(x) = \mu + \tilde{m}_1(x)$. It is easy to see that under the null

$$E(Y|X,Z) = E(m(X,Z)|X,Z) = E(m_1(X)|X,Z) = m_1(X). \tag{4.14}$$

The goal is to test the dependence of the mean function of Y on Z after taking into account the dependence of Y on X. Therefore, we first want to get rid of all mean dependence of Y on X. This can be accomplished by estimating this mean dependence (nonparametrically in this context), obtaining the residuals of this regression and than testing the mean dependence of these residuals on Z. Thus, we need to consider

$$m(X,.) = E(Y|X) = E(m(X,Z)|X). \tag{4.15}$$

Again, under the null hypothesis, it's easy to see that

$$m(X,.) = E(m(X,Z)|X) = E(m_1(X)|X) = m_1(X).$$

Under the alternative hypothesis however, we have

$$\begin{aligned} m(X,.) &= E(Y|X) = E(m(X,Z)|X) \\ &= \int m(X,z)dF_{Z|X}(z). \end{aligned} \tag{4.16}$$

Using the factorization in (4.12) we have

$$\begin{aligned} m(X,.) &= E(Y|X) = E(m(X,Z)|X) \\ &= \mu + \tilde{m}_1(X) + E(\tilde{m}_2(Z)|X) + E(\tilde{m}_{12}(X,Z)|X), \\ &= \mu + \tilde{m}_1(X) + \int \tilde{m}_2(z)dF_{Z|X}(z) \\ &\quad + \int \tilde{m}_{12}(X,z)dF_{Z|X}(z), \end{aligned} \tag{4.17}$$

and if the mean function is additive, it simplifies to

$$\begin{aligned} m(X,.) &= \mu + \tilde{m}_1(X) + E(\tilde{m}_2(Z)|X) \\ &= \mu + \tilde{m}_1(X) + \int \tilde{m}_2(z)dF_{Z|X}(z). \end{aligned} \tag{4.18}$$

Note that in the previous sections, the factorization was not considered and the Nadaraya-Watson Kernel estimator was used to estimate $m_1(x)$. Interestingly, the ANOVA-type statistic proposed in Section 4.1 estimates $E(Y|X)$, and therefore the residuals that will be used in the proposed test are formed by subtracting the estimate of the four components on the right hand side of (4.17). Under the alternative, such estimation contains some effect of Z, which is subtracted from the residuals, causing a possible reduction in the power of the test.

This situation becomes even clearer when the additive model (4.18) is assumed. In this case, it is obvious that the residuals formed by subtracting $\tilde{m}_1(x)$ would be ideal, as all the effect of Z would be left to the residuals. However, the Nadaraya-Watson kernel regression estimator proposed $\hat{m}_1(x)$ will contain some information of $\tilde{m}_2$, namely $\int \tilde{m}_2(z)dF_{Z|X}(z)$. When X and Z are independent and the model is additive though, this problem does not exist.

Therefore, another approach for estimating the effect of X is to estimate only the function $\tilde{m}_1(x)$, so that no effect of Z will be removed. Here, the situation is posed as a dilemma which needs to be further investigated. Is the power of the test improved by estimating only $\tilde{m}_1(x)$, or is it not? Moreover, how is the level of the test affected by estimating the effect of X using these two different approaches?

Based on simulations, we conjecture that the asymptotic theory of the test statistic remains the same for a wide class of other nonparametric estimators of m_1, such as local linear estimators or, under an additive model, the backfitting estimator. A particular alternative estimator incorporated in our simulations is a version of the estimator proposed by Newey (1994) and Linton and Nielsen (1995), and further studied in Mammen, Linton and Nielsen (1999) and Horowitz and Mammen (2004), computed as

$$\begin{aligned}
\hat{\tilde{m}}_1(x_i) &= \int \hat{m}(x_i, z)d\hat{F}_Z(z) \\
&= \frac{1}{n}\sum_{k=1}^{n} \hat{m}(x_i, z_k) \\
&= \frac{1}{n}\sum_{k=1}^{n}\sum_{j=1}^{n} \frac{K\left(\frac{x_i - x_j}{h_n}, \frac{z_k - z_j}{h_n}\right)}{\sum_{l=1}^{n} K\left(\frac{x_i - x_l}{h_n}, \frac{z_k - z_l}{h_n}\right)} Y_j, \qquad (4.19)
\end{aligned}$$

with $\hat{m}(x,z)$ a Nadaraya-Watson kernel estimator of $m(x,z)$. The estimator $\hat{\tilde{m}}_1(x_i)$ estimates $\tilde{m}_1(x_i)$ in the factorization model. Note that if X and Z are not independent, for a fixed z_k there are several pairs of points (x_i, z_k) where there are no observations in a small neighborhood h_n. In that case, for all these pairs, the regression function can not be estimated. To visualize this, Figure 4.1 shows distribution of points for dependent X and Z. For a fixed observed point $x_i = -.4047303$ (triangle point in the plot is $(x_i, z_i) = (-0.40473030.2708415)$), the estimator in (4.19) computes the Nadaraya-Watson estimates of $m(x_i, z_k), k = 1, \ldots, n$ and takes the average of these values. The solid line represents the x_i value, the square point represents a random point $(x_k, z_k) = (2.248891, 1.759995)$, the intersection of the solid line and dashed line represents an example of a point (x_i, z_k) where the function $m(.)$ has to be estimated, and the black box is the neighborhood defined by a bandwidth (for instance in this example $h_n = .2$) on each covariate. Note that there are no observations within the neighborhood, and therefore the function $m(.)$ can not be estimated. The problem is that once fixed x_i, several of the neighborhoods formed will have no observations, causing a non-reliable estimation of $\tilde{m}_1(x)$.

To overcome this, we propose using a kernel for an adjustment, that is,

$$\hat{\tilde{m}}_1'(x_i) = \sum_{k=1}^{n} \frac{K'\left(\frac{x_i - x_k}{h_n'}\right)}{\sum_{l=1}^{n} K'\left(\frac{x_i - x_l}{h_n'}\right)} \hat{m}(x_i, z_k). \tag{4.20}$$

The estimator in (4.20) assigns weights zero if there are no points to estimate $m(.)$, so that the estimated points that will be averaged in z are only those that can be estimated.

There were a few discussions on other ways to estimate the effect of X on Y to produce residuals that would lead to good level under the null and more power under alternatives. One way suggested by the committee (during the Comprehensive Exam) was to tackle this estimation by first creating the windows W_i described in (4.91) and, using only the points that are contained in each window, estimate the mean function $E(Y|X)$. This estimator will be referred as the "slice" estimator.

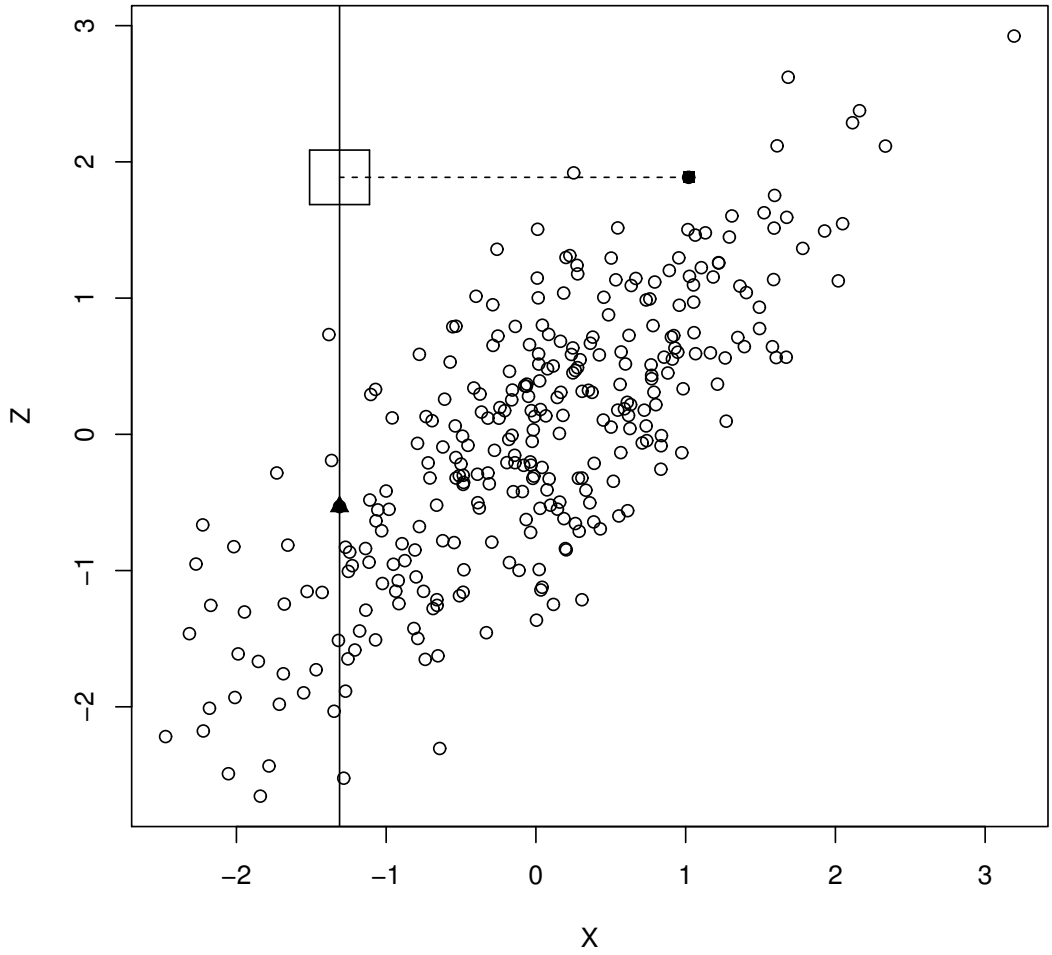

Figure 4.1. Example of a point where $m(.)$ can not be estimated due to non-available observations for normally distributed X and Z with correlation 0.8.

It can be written as

$$\hat{m}_1^s(x_i) = \sum_{j=1}^{n} \frac{K\left(\frac{x_i - x_j}{h_n}\right)}{\sum_{l=1}^{n} K\left(\frac{x_i - x_l}{h_n}\right) \mathbf{1}(l \in W_i)} Y_j \mathbf{1}(j \in W_i). \tag{4.21}$$

Note that this estimator will estimate $m_1(x_i)$ differently in each of the p windows that contain the observation i. To add a little more flexibility and more neighboring observations in the estimation of $m(x_i)$, observations from the closest

windows could be included in the estimation. Thus, such estimator is written as

$$\hat{m}_1^{s_2}(x_i) = \sum_{j=1}^{n} \frac{K\left(\frac{x_i - x_j}{h_n}\right)}{\sum_{l=1}^{n} K\left(\frac{x_i - x_l}{h_n}\right)\mathbf{1}_l} Y_j \mathbf{1}_j. \tag{4.22}$$

where $\mathbf{1}_j = \mathbf{1}\left(j \in W_{(i-\frac{(p-1)}{2})}, \ldots, W_{(i+\frac{(p-1)}{2})}\right)$. Although the idea of estimating $E(Y|X)$ within each window seems to produce a finer estimation specifically for that range of Z, these "slice" estimators fail to take into consideration the rest of points of X that are near x_i but far from z_i (points z_j such that $|z_j - z_j| > h_n$). In that way, they may leave out an important part of the observations. It is clear that asymptotically, as the number of observations tends to infinity, it estimates $m(x, z)$, which leads to no power under the alternative.

Figures 4.2 and 4.3 allow us to visualize how each estimator of m_1 treats the data and which points are used in the estimation process. Figure 4.2 corresponds to the estimator $\hat{m}_1$: for the estimator $\hat{m}_1$ at a point x_i, consider the bandwidth of size h_n, which is the cutpoint of all the observed points, so that we actually only use the points that are within that bandwidth ($j : |x_j - x_i| < h_n$). The estimate $\hat{m}_1(x_i)$ is simply calculated by using the Nadaraya-Watson estimation on these points. Note that in this estimation process, we do not take into account the z axis.

Figure 4.3 shows how to estimate $\hat{\hat{m}}_1(x_i)$. We first consider the bandwidth on the x axis. Now, for each $k = 1, \ldots, n$, consider the point (x_i, z_k) (which is as if the z_k was transported to the x_i line). Building a window around this point on the z axis, we then use only the points within the bandwidth on the x and z axis to estimate $\hat{m}(x_i, z_k), k = 1, \ldots, n$. Finally, the estimator $\hat{\hat{m}}_1(x_i)$ is computed as the mean of the estimated $\hat{m}(x_i, z_k), k = 1, \ldots, n$.

Another possibility is to use the backfitting estimator to estimate m_1, however it also estimates the additive component. In the case that the regression function is not additive, the backfitting estimator will not estimate the correct m_1 function.

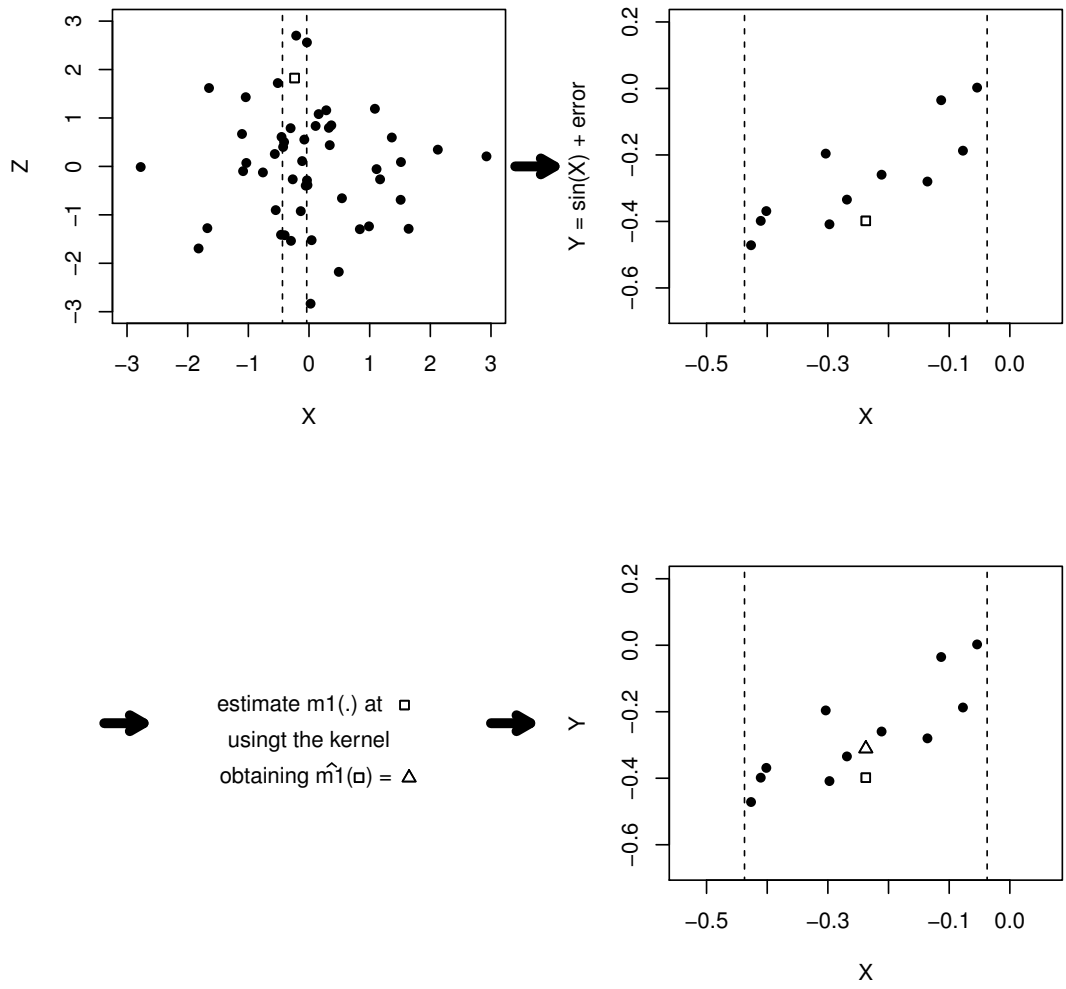

Figure 4.2. Visual aid for the estimation of $\hat{m}_1$

4.1.2 Simulations: ANOVA-type hypothesis test using different types of estimation for the regression function under the null

In this section we explore the different estimators of $m_1(\mathbf{x})$ proposed in Section 4.1.1. We run 2000 simulations of data sets of size $n = 100$ and report the level and power of the ANOVA-type test using such estimators. Table 4.1 reports the results for the model $Y = m(X, Z) + \epsilon$, where $m(X, Z) = X + \theta Z + \gamma XZ$, X, Z are

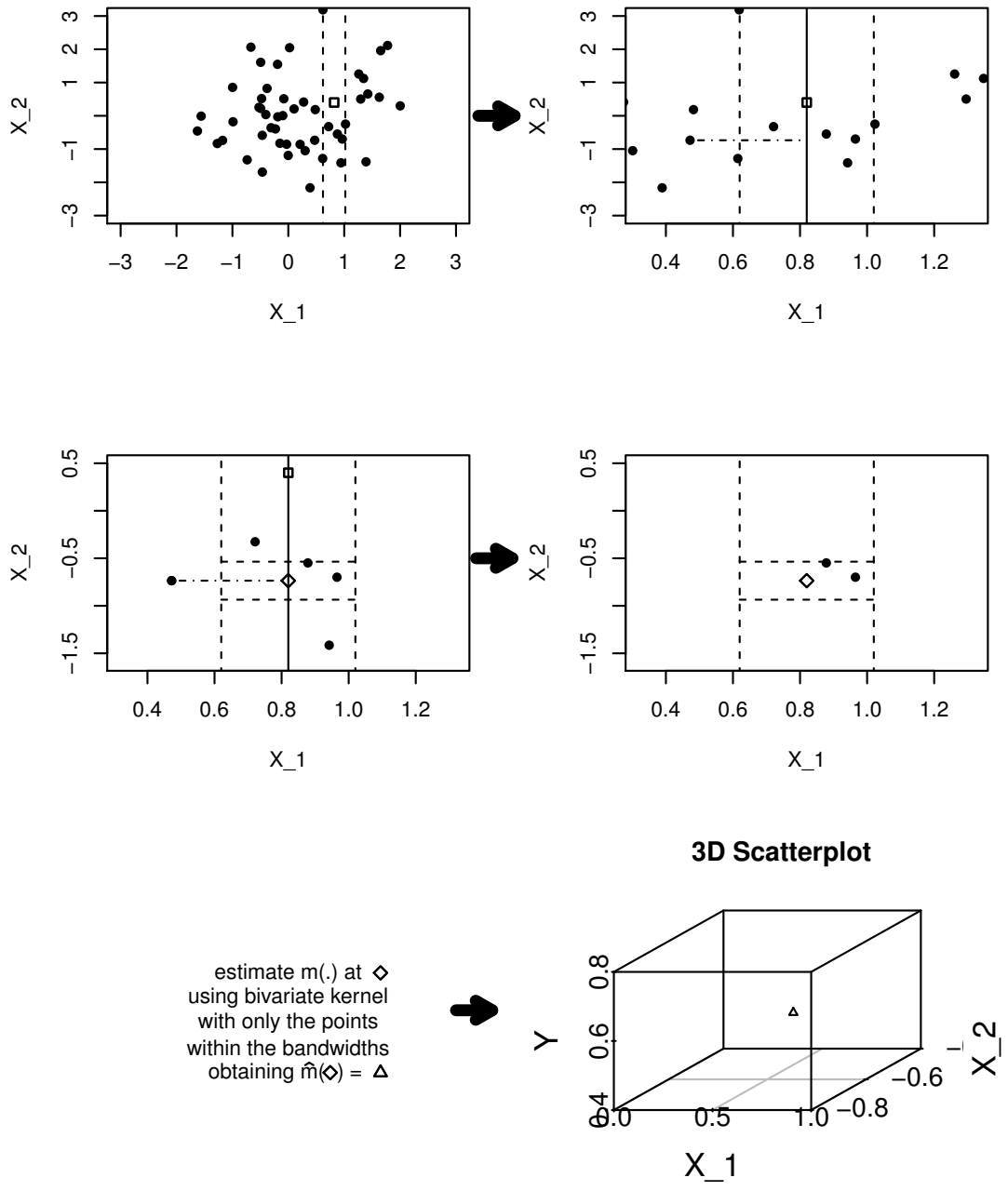

Figure 4.3. Visual aid for the estimation of $\hat{\hat{m}}_1$

independent and identically distributed with distribution $U(0,1)$, and $\epsilon \sim N(0,3^2)$.

We will only include the usual Nadaraya-Watson estimator and the estimator (4.20) of $\tilde{m}_1$. The slice estimator faces problems in estimating the function m_1 in the case of table 4.1 because the windows are formed according to the z points, and the estimation is done by using a kernel regression, which needs points around the estimation point. If the number of observations is not large, the windows constructed may have no points within a bandwidth to estimate m_1, so that the bandwidth would be very large and the estimation not good.

Note that $\hat{m}_1$ estimates

$$\begin{aligned} E(Y|X) &= X + \theta E(Z|X) + \gamma X E(Z|X) \\ &= X + \theta\frac{1}{2} + \gamma\frac{X}{2}. \end{aligned}$$

Following the definition from the ANOVA factorization for two dimensions, we have

$$\begin{aligned} \mu &= \int\int m(X,Z)dF_Z dF_X = E_X\left[E_Z m(X,Z)\right] = \frac{1}{2} + \frac{\theta}{2} + \frac{\gamma}{4} \\ \tilde{m}_1(X) &= X + \theta\frac{1}{2} + \gamma\frac{X}{2} - \left(\frac{1}{2} + \frac{\theta}{2} + \frac{\gamma}{4}\right) = X\left(1 + \frac{\gamma}{2}\right) - \frac{1}{2} - \frac{\gamma}{4} \\ \tilde{m}_2(Z) &= \frac{1}{2} + \theta Z + \frac{\gamma Z}{2} - \left(\frac{1}{2} + \frac{\theta}{2} + \frac{\gamma}{4}\right) = Z\left(\theta + \gamma 2\right) - \frac{\theta}{2} - \frac{\gamma}{4} \\ \tilde{m}_{12}(X,Z) &= \gamma XZ - \frac{\gamma X}{2} - \frac{\gamma Z}{2}. \end{aligned}$$

It is important to emphasize the fact that μ defined in the factorization approach is not always equal to the mean of Y. In the factorization model, $\mu = E_X\left[E_Z m(X,Z)\right]$, while in general, the mean of Y is $E_X\left[E_{Z|X} m(X,Z)\right]$. Obviously, when X and Z are independent, they are the same.

The simulations reported in Table 4.1 suggest that the test statistic using the residuals $\hat{\xi} = Y - \hat{\tilde{m}}_1(\mathbf{x})$ (ANOVA-type 2 in the table) can have improved power. The level of the test is slightly increased by using the ANOVA-type 2 test statistic, not a significant change that would compromise the estimator.

Table 4.1. Rejection rates with alternative fitting methods

	$\theta = 0$	$\theta = 1$			
	γ	γ			
Method	0	1	2	3	4
ANOVA-type (p=11)	.053	.119	.196	.292	.390
ANOVA-type (p=9)	.052	.121	.191	.296	.388
ANOVA-type 2 (p=11)	.054	.150	.218	.315	.440
ANOVA-type 2 (p=9)	.054	.122	.201	.320	.422

4.1.3 Simulations: Comparison of the ANOVA-type statistic with Lavergne's statistic, Fan and Li's statistic and the Generalized Likelihood Ratio Test

In this section we compare the proposed ANOVA-type and ANOVA-type 2 statistics to the statistics proposed by Lavergne and Vuong (2000) (LV in the tables), Fan and Li (1996) (FL in the tables), and to the generalized likelihood ratio test statistic introduced by Fan, Zhang and Zhang (2001), Fan and Jiang (2005) and Fan and Jiang (2007) (GLR in the tables).

The data is generated according to the models (also used in Lavergne and Vuong, 2000)

$$Y = -X + X^3 + f_j(Z) + \epsilon,\ j = 0, 1, 2, 3, , 4, 5, 6, 7, \tag{4.23}$$

where $\epsilon \sim N(0, 4)$ for 3 different linear alternatives and 4 different non-linear alternatives. Consider the null hypothesis when $f_0(Z) = 0$, the linear alternatives $f_1(Z) = .5Z, f_2(Z) = Z$ and $f_3(Z) = 2Z$, and the nonlinear alternatives $f_4(Z) = sin(2\pi Z), f_5(Z) = sin(\pi Z), f_6(Z) = sin(2/3\pi Z)$ and $f_7(Z) = sin(1/2\pi Z)$. The covariates X and Z are independent and both generated from a standard normal distribution. Note that one of the assumptions of the proposed test statistic is that the support of the density of X is compact, but in this simulation we have a finite sample size (n = 100 and n = 200), hence it is reasonable to apply the proposed methodology.

Figures 4.4 to 4.5 show the plots for the functions we will analyze. The function m_1 is a cubic function, and we will estimate it using the Nadaraya-Watson

kernel, whose estimation will be used to compute the residuals assuming the null hypothesis. The first 3 alternatives are linear and clearly we expect the F test to be more powerful than any other test. To identify significance for testing the sine alternatives on Figure 4.5, one has to overcome the rapidly wiggling of the sine, so the nonparametric test seems to be adequate.

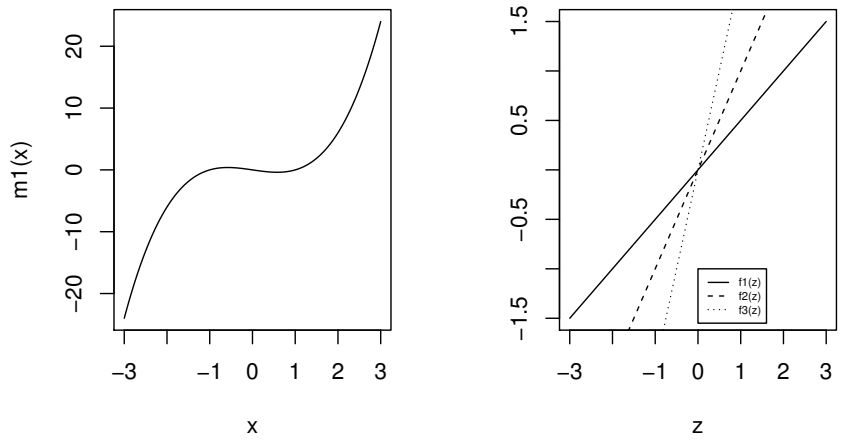

Figure 4.4. $m_1(x)$

The kernel for the Nadaraya-Watson estimation of $m(X)$ is the uniform on $(-0.5, 0.5)$ density, and the bandwidth is selected through leave-one-out cross validation. The rejection rates shown in Table 4.2 for LV, FL, and F tests are taken from the simulation results reported in the LV paper (based on 2000 runs). Also, the p between parenthesis in "ANOVA-type(p)", represents the size of the window W_i. It is important to note that, in each simulation setting, the LV paper reports several rejection rates for the LV and FL tests, each corresponding to different values of smoothing parameters. Since the best performing constants are different for different simulation settings, the rejection rates reported in Table 4.2 represent a) the most accurate alpha level achieved over all constants, and b) the best power achieved overall constants for each alternative. For comparison purposes, the rejection rates for the ANOVA-type tests and the GLR test are also based on 2000 simulation runs.

As expected, the F test achieved the best results for the three linear alternatives

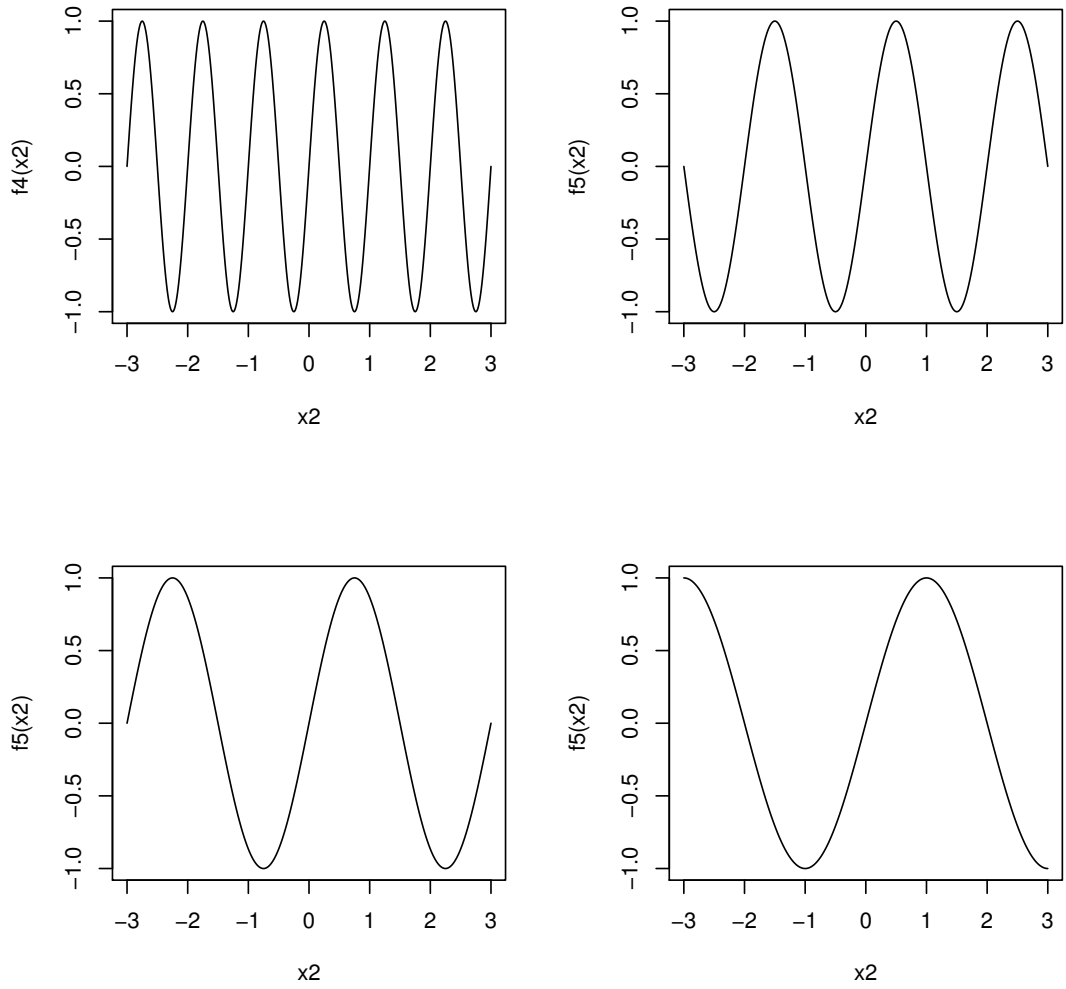

Figure 4.5. f_4, f_4, f_6 and f_7

and the worse results for the three non-linear alternatives. The GLR test has higher power than the ANOVA-type tests against linear alternatives (which is partly explained by the fact it is based on normal likelihood), but is much less powerful against the first of the non-linear alternatives. As the non-linearity decreases (f_5 and f_6) the power of the GLR test improves. For further comparison with the Generalized Likelihood test see Section 4.2.4.

Table 4.2. Rejection rates under H_0, linear and non-linear alternatives

		linear				sine			
n	test	f_0	f_1	f_2	f_3	f_4	f_5	f_6	f_7
100	LV	.041	.098	.482	.991	.182	.266	.319	.352
	FL	.021	.051	.271	.970	.126	.168	.187	.181
	ANOVA-type(9)	.052	.218	.79	.999	.423	.523	.535	.55
	ANOVA-type(7)	.056	.244	.780	.999	.432	.527	.551	.55
	ANOVA-type(5)	.057	.241	.767	1	.418	.486	.485	.517
	ANOVA-type(3)	.079	.208	.675	.998	.343	.392	.416	.408
	ANOVA-type 2	.065	.275	.831	1	.453	.598	.600	612
	GLRT	.044	.365	.951	1	.123	.497	.645	.65
	F-test	.051	.695	.997	1	.046	.055	.222	.608
200	LV	.054	.208	.875	1	.386	.540	.678	.690
	FL	.025	.083	.695	1	.289	.395	.471	.482
	ANOVA-type(9)	.055	.374	.95	.999	.730	.778	.788	.799
	ANOVA-type(7)	.051	.376	.979	1	.746	.82	.82	.821
	ANOVA-type(5)	.055	.356	.961	1	.709	.734	.744	.731
	ANOVA-type(3)	.069	.293	.921	1	.592	.566	.605	.610
	ANOVA-type 2	.069	.487	.999	1	.821	.882	.884	.889
	GLRT	.036	.656	1	1	.188	.877	.936	.945
	F-test	.052	.931	1	1	.051	.053	.340	.856

4.1.4 Simulations: Organizing the windows W_i differently

The power of the test may be improved if we are able to identify first which kind of function is left in the residuals after fitting the null regression function. Obviously, when the data is presented, we do not know what it is, otherwise testing would be unnecessary. The purpose of this simulation is to show that power can be increased by organizing the windows W_i with the correct form of the covariates.

Table 4.3 shows the results for the model $Y = X + \theta XZ + \epsilon$,where X, Z are iid standard normal, and $\epsilon \sim N(0, 0.1)$. We generated 2000 data sets of size $n = 200$ and computed the rejection rates. The ANOVA-type does not do well in this situation, while when sorting the windows W_i by the product $X \times Z$, the power of the ANOVA-type test is a lot larger than the Generalized Likelihood Ratio test.

Table 4.4 shows the results for the model $Y = X + \theta X/Z + \epsilon$, where X, Z are iid standard normal and $\epsilon \sim N(0, 0.1)$. Again 2000 data sets of size $n = 200$ were generated and the rejection rates were computed. Even though the extra term in

Table 4.3. Percentage of rejections

Method — θ	0	.01	.05	.1	.2	.4
ANOVA-type (11)	.047	.051	.052	.045	.048	.06
ANOVA-type (9)	.053	.047	.050	.045	.051	.054
ANOVA-type (11)(by XZ)	.038	.054	.785	.998	1	1
ANOVA-type (9)(by XZ)	.046	.071	.801	.998	1	1
ANOVA-type (7)(by XZ)	.047	.063	.778	.998	1	1
GLRT	.059	.064	.195	.306	.438	.533

Table 4.4. Percentage of rejections

Method — θ	0	.01	.05	.1	.2	.4
ANOVA-type (11)	.042	.091	.104	.098	.106	.117
ANOVA-type (9)	.049	.095	.115	.130	.115	.103
ANOVA-type (11)(by XZ)	.038	.398	.692	.657	.641	.659
ANOVA-type (9)(by XZ)	.047	.395	.650	.689	.659	.686

the model is a ratio of X and Z, organizing the windows W_i with the product gives better results.

Table 4.5 shows the results for the model $Y = X + \theta Z + 2\theta XZ + \epsilon$, where X, Z are iid standard normal, and $\epsilon \sim N(0, 0.1)$. In this situation, the ANOVA-type has performance comparable to the Generalized Likelihood Ratio test.

4.1.5 Edge Effects

The proposed ANOVA-type test statistic is constructed using the windows W_i defined in (4.91). Basically, after ordering the observations z_i, each window W_i represents the point z_i and the set of $(p-1)/2$ points nearest to z_i. Clearly the observations on the edge do not have neighbors on both sides, but the asymptotic results are still valid if we construct asymmetric windows for these points or simply ignore them, i.e., construct the vector $\boldsymbol{\xi}_V$ using only the $n - (p-1)$.

Table 4.5. Percentage of rejections

Method — θ	0	.01	.025	.05	.1	.2
ANOVA-type (p=9)	.060	.078	.335	.736	.907	.945
ANOVA-type (p=9)(by XZ)	.052	.122	.778	.998	1	1
GLRT	.063	.225	.724	.976	1	1

Constructing asymmetric windows or eliminating the edge effects may not affect the asymptotic properties of the test statistic, however it may cause a considerable effect on small data sets. Moreover, the size p of the window W_i can affect the analysis, as it defines how many observations will be on the edge. In this section we will investigate the effects of different ways of dealing with the edge effects on the level and power of the test.

For asymmetric edges, suppose w.l.o.g. that the observation i in on the left edge, i.e., $i \leq (p-1)/2$. One way to construct the windows W_i is to include the observation i, the $(p-1)/2$ observations on the right and $\frac{p-1}{2}$ "new observations", all equal to the mean of the observations to the left of i (Edge in the Tables). Note that we define the first (and last) row in this case to be p values exactly equal to the first (and last) observation. To observe the effect of this in the simulations, we also included a version of Edge without the first and last rows (Edge1 in the tables).

Table 4.6 reports results of 2000 simulations from the linear model $Y = X + \theta Z + \epsilon$, where X and Z are iid $N(0,1)$ and $\epsilon \sim N(0,1)$. Clearly, the level of the test is affected by using the edge observations, where discarding the very edge observation (Edge1) is less affected than using all observations. On the other hand, the power of the test is improved significantly by not discarding the edge observations, since they also contain important information.

4.2 ANOVA-type hypothesis test for multivariate X and univariate Z

In this section we will consider the case where we have n observations of the random variable Y and covariates $\mathbf{X}$ and Z, when $\mathbf{X}$ is of dimension $d_1 > 1$ and Z is univariate. Assume also that $\mathbf{X}$ and Z have densities $f_{\mathbf{X}}$ and f_Z respectively. Denote the expected value of the response variable Y given the covariates by the unknown regression function $m(\mathbf{x}, z) = E(Y|\mathbf{X} = \mathbf{x}, Z = z)$. Therefore, allowing heterocedasticity with the unknown variance function $\sigma^2(.,.)$, where $\sup_{(\mathbf{x},z)\in S_{\mathbf{X},Z}} \sigma^2(\mathbf{x}, z) < M < \infty$, the nonparametric model is

$$Y_i = m(\mathbf{X}_i, Z_i) + \sigma(\mathbf{X}_i, Z_i)\epsilon_i, i = 1...n, \tag{4.24}$$

Table 4.6. Rejection rates for ANOVA-type with and without edge effect modification

		θ						
n	test	0	.2	.3	.4	.5	.6	.7
40	ANOVA-type (3)	.080	.134	.184	.295	.422	.525	.669
	ANOVA-type (5)	.062	.110	.172	.283	.425	.552	.667
	ANOVA-type (7)	.051	.091	.138	.245	.375	.500	.659
	ANOVA-type (9)	.039	.070	.128	.227	.345	.471	.614
	ANOVA-type Edge (3)	.110	.170	.275	.369	.511	.645	.783
	ANOVA-type Edge (5)	.099	.168	.292	.432	.608	.760	.856
	ANOVA-type Edge (7)	.106	.198	.345	.501	.675	.806	.895
	ANOVA-type Edge (9)	.111	.223	.343	.547	.714	.833	.923
	ANOVA-type Edge1 (3)	.082	.128	.170	.200	.298	.395	.561
	ANOVA-type Edge1 (5)	.069	.146	.211	.366	.515	.641	.778
	ANOVA-type Edge1 (7)	.070	.156	.245	.406	.572	.718	.843
	ANOVA-type Edge1 (9)	.071	.165	.299	.446	.628	.765	.881
80	ANOVA-type (3)	.069	.146	.255	.453	.607	.787	.905
	ANOVA-type (5)	.064	.143	.272	.507	.696	.845	.944
	ANOVA-type (7)	.052	.153	.290	.505	.711	.864	.951
	ANOVA-type (9)	.047	.148	.283	.500	.711	.861	.945
	ANOVA-type Edge (3)	.087	.167	.318	.516	.725	.875	.949
	ANOVA-type Edge (5)	.089	.231	.399	.626	.823	.944	.983
	ANOVA-type Edge (7)	.090	.243	.458	.687	.878	.969	.990
	ANOVA-type Edge (9)	.091	.294	.502	.746	.909	.975	.995
	ANOVA-type Edge1 (3)	.075	.133	.250	.427	.628	.811	.900
	ANOVA-type Edge1 (5)	.055	.166	.314	.575	.767	.900	.962
	ANOVA-type Edge1 (7)	.067	.201	.397	.650	.842	.943	.989
	ANOVA-type Edge1 (9)	.067	.222	.430	.686	.890	.963	.992

where ϵ_i is the independent error with zero mean and constant variance and independent of $\mathbf{X}$ and Z.

The goal is to test the null hypothesis that z is significant in the regression, that is,

$$H_0 : m(\mathbf{x}, z) = m_1(\mathbf{x}). \tag{4.25}$$

Similarly to the univariate case, let $m_1(\mathbf{X}_i) = E(Y|\mathbf{X}_i)$, and define

$$\xi_i = Y_i - m_1(\mathbf{X}_i).$$

Under the null hypothesis (4.25), set $\xi_i = \sigma(\mathbf{X}_i, Z_i)\epsilon_i$ and thus,

$$E(\xi_i|Z_i) = 0.$$

An estimator of ξ_i, $\hat{\xi}_i = Y_i - \hat{m}_1(\mathbf{X}_i)$, can be obtained by using the multivariate Nadaraya-Watson estimator of the mean function m_1,

$$\hat{m}_1(\mathbf{X}_i) = \sum_{j=1}^{n} \left(\frac{K_H(\mathbf{X}_i - \mathbf{X}_j)}{\sum_{l=1}^{n} K_H(\mathbf{X}_i - \mathbf{X}_l)} \right) Y_j, \ i = 1...n,$$

with $K_{H_n}(\mathbf{x}) = |H_n|^{-1/2} K(H_n^{-1/2}\mathbf{x})$, where $K(\cdot)$ is a bounded d_1-variate kernel function of bounded variation and with bounded support, and H_n is a symmetric positive definite $d_1 \times d_1$ matrix, and we follow the notation in Ruppert and Wand (1994) calling $H_n^{1/2}$ the bandwidth matrix.

After computing $\hat{\xi}_i, i = 1, \ldots, n$ we set the windows W_i as described in Section 4.1, and calculate the vector of $(n - p + 1)p$ "constructed observations" in the augmented one-way ANOVA design

$$\hat{\boldsymbol{\xi}}_V = (\hat{\xi}_j, j \in W_{(p-1)/2+1}, \ldots, \hat{\xi}_j, j \in W_{n-(p-1)/2})^T. \tag{4.26}$$

Let MST = MST($\hat{\boldsymbol{\xi}}_V$), MSE = MSE($\hat{\boldsymbol{\xi}}_V$) denote the balanced one-way ANOVA mean squares due to treatment and error, respectively, computed on the data $\hat{\boldsymbol{\xi}}_V$. The proposed test statistic is based on

$$MST - MSE = \hat{\boldsymbol{\xi}}_V^T A \hat{\boldsymbol{\xi}}_V. \tag{4.27}$$

The following theorem gives the asymptotic normal distribution of the test statistic under the null hypothesis (4.25).

Theorem 4.2. *Assume that the marginal densities $f_{\mathbf{X}}$, f_Z of $\mathbf{X}$, Z, respectively, are bounded away from zero, the second derivatives of $f_{\mathbf{X}}$ and $m_1(\mathbf{x})$ are uniformly continuous and bounded, that $\sigma^2(., z) := E(\xi^2|Z = z)$ is Lipschitz continuous, $\sup_{\mathbf{u}} \sigma^2(\mathbf{u}) < \infty$, and $E(\epsilon_i^4) < \infty$. Assume that the eigenvalues, λ_i, $i = 1, \ldots, d_1$, of the bandwidth matrix $H_n^{1/2}$ defined in (4.26), converge to zero at the same rate*

and satisfy

$$n\lambda_i^8 \to 0 \quad \textit{and} \quad \frac{n\lambda_i^{2d_1}}{(\log n)^2} \to \infty, \ i = 1, \ldots, d_1. \tag{4.28}$$

Then, under H_0 in (4.25), the asymptotic distribution of the test statistic in (4.63) is given by

$$n^{1/2}(MST - MSE) \xrightarrow{d} N(0, \frac{2p(2p-1)}{3(p-1)}\tau^2),$$

where $\tau = \int \left[\int \sigma^2(\mathbf{x}, z) f_{\mathbf{X}|Z=z}(\mathbf{x}) d\boldsymbol{x}_1\right]^2 f_Z(z) dz$.

An estimate of τ^2 can be obtained by modifying Rice's (1984) estimator as follows

$$\hat{\tau}^2 = \frac{1}{4(n-3)} \sum_{j=2}^{n-2} (\hat{\xi}_j - \hat{\xi}_{j-1})^2 (\hat{\xi}_{j+2} - \hat{\xi}_{j+1})^2. \tag{4.29}$$

Proof of Theorem 4.2. Under H_0 in (4.25) we have

$$\begin{aligned} \hat{\xi}_i &= Y_i - \hat{m}_1(\mathbf{X}_i) + m_1(\mathbf{X}_i) - m_1(\mathbf{X}_i) = \xi_i - (\hat{m}_1(\mathbf{X}_i) - m_1(\mathbf{X}_i)) \\ &= \xi_i - \Delta_{m_1}(\mathbf{X}_i), \end{aligned}$$

where

$$\begin{aligned} \Delta_{m_1}(\mathbf{X}_i) &= \sum_{j=1}^{n} \left(\frac{K_H(\mathbf{X}_i - \mathbf{X}_j)}{\sum_{l=1}^{n} K_H(\mathbf{X}_i - \mathbf{X}_l)} \right) (Y_j - m_1(\mathbf{X}_i)) \\ &= \sum_{j=1}^{n} w(\mathbf{X}_i, \mathbf{X}_j)(Y_j - m_1(\mathbf{X}_i)). \end{aligned}$$

Thus, $\hat{\boldsymbol{\xi}}_V$ of relation (4.62) is decomposed as

$$\begin{aligned} \hat{\boldsymbol{\xi}}_V &= \{\hat{\xi}_j : j \in W_1, ..., \hat{\xi}_j : j \in W_n\}^T \\ &= \{\xi_j - \Delta_{m_1}(\mathbf{X}_j) : j \in W_1, ..., \xi_j - \Delta_{m_1}(\mathbf{X}_j) : j \in W_n\}' \\ &= \boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V}, \end{aligned} \tag{4.30}$$

and $\sqrt{n}$(MST - MSE) can be written as

$$\sqrt{n}\hat{\boldsymbol{\xi}}_V^T A\hat{\boldsymbol{\xi}}_V = \sqrt{n}\boldsymbol{\xi}_V^T A\boldsymbol{\xi}_V - \sqrt{n}2\boldsymbol{\xi}_V^T A\boldsymbol{\Delta}_{m_1V} + \sqrt{n}\boldsymbol{\Delta}_{m_1V}^T A\boldsymbol{\Delta}_{m_1V}. \tag{4.31}$$

The asymptotic normality of $\sqrt{n}\boldsymbol{\xi}_V^T A\boldsymbol{\xi}_V$ follows by arguments similar to those used in Theorem 3.2 of Wang, Akritas and VanKeilegom (2008). It remains to derive its asymptotic variance and to show that the other two terms in (4.66) converge to zero in probability.

Note that

$$\boldsymbol{\xi}_V^T A_d\boldsymbol{\xi}_V = \frac{1}{n(p-1)}\sum_{i=1}^{n}\sum_{j_1\neq j_2}^{n}\xi_{j_1}\xi_{j_2}\mathbf{1}(j_1,j_2\in W_i),$$

and thus has expected value 0.

Recall the notation

$$\begin{aligned}\sigma^2(.,z) &:= E(\xi^2|Z=z) = E(\epsilon^2\sigma^2(\mathbf{X},z)|Z=z)\\ &= \int \sigma^2(\mathbf{x},z)f_{\mathbf{X}|Z=z}(\mathbf{x})d\mathbf{x}.\end{aligned}$$

Thus, using Lemma 4.17 it suffices to find the asymptotic variance of $\sqrt{n}\boldsymbol{\xi}_V^T A_d\boldsymbol{\xi}_V$. To that end, we write

$$\begin{aligned}&E[\boldsymbol{\xi}_V^T A_d\boldsymbol{\xi}_V|\mathbf{Z}=\mathbf{z}]^2\\ &= \frac{1}{n^2(p-1)^2}\sum_{i_1,i_2}^{n}\sum_{j_1\neq l_1}^{n}\sum_{j_2\neq l_2}^{n}E(\xi_{j_1}\xi_{l_1}\xi_{j_2}\xi_{l_2}|\mathbf{Z}=\mathbf{z})I(j_s\in W_{i_s},l_s\in W_{i_s},s=1,2)\\ &= \frac{2}{n^2(p-1)^2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j\neq l}^{n}E(\xi_j^2\xi_l^2|\mathbf{Z}=\mathbf{z})I(j,l\in W_{i_1}\cap W_{i_2})\\ &= \frac{2}{n^2(p-1)^2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j\neq l}^{n}\sigma^2(.,z_j)\sigma^2(.,z_l)I(j,l\in W_{i_1}\cap W_{i_2})\\ &= \frac{2}{n^2(p-1)^2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j\neq l}^{n}\sigma^2(.,z_j)\left(\sigma^2(.,z_j)+O\left(\frac{p}{\sqrt{n}}\right)\right)I(j,l\in W_{i_1}\cap W_{i_2})\end{aligned}$$

$$= \frac{2}{n^2(p-1)^2}\sum_{j=1}^{n}\sigma^4(.,z_j)\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{l\neq j}^{n} I(j,l\in W_{i_1}\cap W_{i_2}) + O\left(\frac{p^2}{n^{3/2}}\right)$$
$$= \frac{2}{n^2(p-1)^2}\sum_{j=1}^{n}\sigma^4(.,z_j)2(1+2^2+3^2+...+(p-1)^2) + O\left(\frac{p^2}{n^{3/2}}\right)$$
$$= \frac{2}{n^2(p-1)^2}\frac{p(p-1)(2p-1)}{3}\sum_{j=1}^{n}\sigma^4(.,z_j) + O\left(\frac{p^2}{n^{3/2}}\right),$$

where, using the assumption that $\sigma^2(.,z)$ is Lipschitz continuous, the third equality follows from Lemma 4.16, and the second last inequality results from the fact that if $1 \leq |j_1 - j_2| = s \leq p-1$, then they are $(p-s)^2$ pairs of windows whose intersection includes j_1 and j_2. Taking limits as $n \to \infty$ it is seen that

$$\begin{aligned} E\left(n^{1/2}\boldsymbol{\xi}_V^T A_d \boldsymbol{\xi}_V \middle| \mathbf{Z}=\mathbf{z}\right)^2 &= \frac{2p(2p-1)}{3(p-1)}\frac{1}{n}\sum_{j=1}^{n}\sigma^4(.,z_j) + O\left(\frac{p}{\sqrt{n}}\right) \\ &\to_{a.s.} \frac{2p(2p-1)}{3(p-1)}E(\sigma^4(.,\mathbf{Z})) \\ &= \frac{2p(2p-1)}{3(p-1)}\tau^2 \end{aligned} \tag{4.32}$$

Hence, $n^{1/2}\xi_V^T A \xi_V$ converges in distribution to the designated normal distribution. That the second and third terms in (4.66) converge in probability to zero are shown in Lemmas 4.1, 4.2 , respectively.

□

Lemma 4.1. *The second term in (4.66) converges in probability to zero, i.e.*

$$T_{2n} := \sqrt{n}\boldsymbol{\xi}_V^T A \boldsymbol{\Delta}_{m_1 V} \xrightarrow{p} 0.$$

Proof. After some algebra it can be seen that

$$\begin{aligned} T_{2n} &= \frac{n^{-1/2}(np-1)}{(n-1)p(p-1)}\sum_{i=1}^{n}\sum_{j\in W_i}\xi_j\sum_{k\in W_i}\Delta_{m_1}(\mathbf{X}_k) \\ &\quad -\frac{n^{-1/2}p}{(n-1)}\sum_{i=1}^{n}\xi_i\sum_{j=1}^{n}\Delta_{m_1}(\mathbf{X}_j) - \frac{n^{-1/2}p}{(p-1)}\sum_{i=1}^{n}\xi_i\Delta_{m_1}(\mathbf{X}_i). \end{aligned} \tag{4.33}$$

We will show that each of the three terms above converge in probability to zero

conditionally on $\mathbf{U} = \{\mathbf{X}, Z\}$, and thus also unconditionally. Note that, because all windows W_i are of finite size (p), the first term on the right hand side of (4.33) can be written as a finite (p^2) sum of terms each of which is similar to the last term in (4.33). Thus, it suffices to show that the last and second terms of (4.33) converge to zero. For notational simplicity, all expectations and variances in this proof are to be understood as conditional on $\mathbf{U} = \{\mathbf{X}, Z\}$. For the last term in (4.33) we have

$$\begin{aligned}
\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\xi_i\Delta_{m_1}(\mathbf{X}_i) &= n^{-1/2}\sum_{i=1}^{n}\xi_i(\hat{m}(\mathbf{X}_i) - m(\mathbf{X}_{1j})) \\
&= n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n} w(\mathbf{X}_i, \mathbf{X}_j)(m(\mathbf{X}_j) + \xi_j - m(\mathbf{X}_i))\xi_i \\
&= n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n} w(\mathbf{X}_i, \mathbf{X}_j)(m(\mathbf{X}_j) - m(\mathbf{X}_i))\xi_i \\
&\quad + n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n} w(\mathbf{X}_i, \mathbf{X}_j)\xi_j\xi_i. \qquad (4.34)
\end{aligned}$$

The first term of the right hand side of (4.34) has zero expectation, so it suffices to show that its variance goes to zero. To this end, we write

$$\begin{aligned}
&Var(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{n}\xi_i w(\mathbf{X}_i, \mathbf{X}_j)(m_1(\mathbf{X}_j) - m_1(\mathbf{X}_i))) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n} w(\mathbf{X}_i, \mathbf{X}_{j_1})w(\mathbf{X}_i, \mathbf{X}_{j_2})\mathsf{x} \\
&\qquad\qquad \mathsf{x}(m_1(\mathbf{X}_{j_1}) - m_1(\mathbf{X}_i))(m_1(\mathbf{X}_{j_2}) - m_1(\mathbf{X}_i))Var(\xi_i) \\
&\leq \frac{M}{n}\sum_{i=1}^{n}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n} w(\mathbf{X}_i, \mathbf{X}_{j_1})w(\mathbf{X}_i, \mathbf{X}_{j_2})\mathsf{x} \\
&\qquad\qquad \mathsf{x}(c||\mathbf{X}_{j_1} - \mathbf{X}_i||c||\mathbf{X}_{j_2} - \mathbf{X}_i||) \\
&= Mc^2 O(||H_n^{1/2}||)O(||H_n^{1/2}||) = o(1),
\end{aligned}$$

for some constants M and c, where the inequality holds because $m_1(\cdot)$ is Lipschitz continuous, and the last equality follows from Lemma 4.19. Thus, by the assumptions of Theorem 4.2 the first term of the right hand side of (4.34) goes

in probability to zero. To show that the second term in (4.34) also goes to 0 in probability since, we will show that its second moment goes to zero. To this end, we write

$$
\begin{aligned}
& E\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{n}\xi_i\xi_j w(\mathbf{X}_i,\mathbf{X}_j)\right]^2 \\
&= E\left[\frac{1}{n}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\xi_{i_1}\xi_{i_2}\xi_{j_1}\xi_{j_2}w(\mathbf{X}_{i_1},\mathbf{X}_{j_1})w(\mathbf{X}_{i_2},\mathbf{X}_{j_2})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}E(\xi_i^2\xi_j^2)w(\mathbf{X}_i,\mathbf{X}_j)^2 \\
&+\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}E(\xi_i^2\xi_j^2)w(\mathbf{X}_i,\mathbf{X}_i)w(\mathbf{X}_j,\mathbf{X}_j) \\
&+\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}E(\xi_i^2\xi_j^2)w(\mathbf{X}_i,\mathbf{X}_j)w(\mathbf{X}_j,\mathbf{X}_i) \\
&+\frac{1}{n}\sum_{i=1}^{n}E(\xi_i^4)w(\mathbf{X}_i,\mathbf{X}_i)w(\mathbf{X}_i,\mathbf{X}_i) \\
\leq\ & \frac{M_1^2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}w(\mathbf{X}_i,\mathbf{X}_j)^2 \\
&+\frac{M_2^2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{c}{n|H_n|^{1/2}\hat{f}(\mathbf{X}_i)}\frac{c}{n|H_n|^{1/2}\hat{f}(\mathbf{X}_j)} \\
&+\frac{M_3^2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{K_{H_n}(\mathbf{X}_i-\mathbf{X}_j)}{n|H_n|^{1/2}\hat{f}(\mathbf{X}_i)}\frac{K_{H_n}(\mathbf{X}_j-\mathbf{X}_i)}{n|H_n|^{1/2}\hat{f}(\mathbf{X}_j)} \\
&+\frac{M_4^4}{n}\sum_{i=1}^{n}\frac{c}{n|H_n|^{1/2}\hat{f}(\mathbf{X}_i)}\frac{c}{n|H_n|^{1/2}\hat{f}(\mathbf{X}_i)} \\
&= \frac{M_1^2}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{K_{H_n}^2(\mathbf{X}_i-\mathbf{X}_j)}{n^2|H_n|(\frac{1}{n|H_n|^{1/2}}\sum_l K_{H_n}(\mathbf{X}_i-\mathbf{X}_l))^2} \\
&+O(\frac{1}{n|H_n|})+O(\frac{1}{n|H_n|^{1/2}})+O(\frac{1}{n^2|H_n|}) \\
&= \frac{M_1^2}{n}\sum_{i=1}^{n}\frac{\hat{f}_1^1(\mathbf{X}_i)}{(\hat{f}_1^2(\mathbf{X}_i))^2}\frac{1}{n|H_n|^{1/2}}+O(\frac{1}{n|H_n|})+O(\frac{1}{n|H_n|^{1/2}})+O(\frac{1}{n^2|H_n|}) \\
&= o(1)+o(1)+o(1)+o(1), \qquad (4.35)
\end{aligned}
$$

for some constants $M_1,\ M_2,\ M_3,\ M_4$ and c, by the fact that $\hat{f}_{\mathbf{X}}$ converges uniformly to $f_{\mathbf{X}}$ a.s. in the compact support $S_{\mathbf{X}}$ (Ruschendorf 1977). Thus, by the assumptions of Theorem 4.2 the second term of the right hand side of (4.34) goes in probability to zero.

Consider now the second term in (4.33). Since $n^{-1/2}\sum_{i=1}^{n}\xi_i$ remains bounded in probability, its convergence to zero will follow if we show that

$$n^{-1}\sum_{k=1}^{n}\Delta_{m_1}(\mathbf{X}_{1k}) \xrightarrow{p} 0$$

. For later use, we will actually show that

$$\frac{1}{n^{3/4}}\sum_{k=1}^{n}\Delta_{m_1}(\mathbf{X}_{1k}) = \frac{1}{n^{3/4}}\sum_{k=1}^{n}(\hat{m}_1(\mathbf{X}_{1k}) - m_1(\mathbf{X}_{1k})) \xrightarrow{p} 0. \tag{4.36}$$

For this we use (cf. Hansen, 2008)

$$\sup_{\mathbf{x}}|\hat{m}_1(\mathbf{x}) - m_1(\mathbf{x})| = O_p(a_n) \ \text{ where } \ a_n = \left(\frac{\log n}{n\lambda^{d-1}}\right)^{1/2} + \lambda^2, \tag{4.37}$$

where $\lambda \to 0$ at the same rate as the eigenvalues $\lambda_i,\ i = 1,\ldots,d-1,$ of H_n. Therefore, the term in the left hand side of (4.36) is of order

$$\frac{1}{n^{3/4}}nO_p\left(\left(\frac{\log n}{n\lambda^{d-1}}\right)^{1/2} + \lambda^2\right) = o_p(1),$$

by the assumed conditions stated in (4.28). This completes the proof of Lemma 4.1.

□

Lemma 4.2. *The third term in (4.66) converges in probability to zero, i.e.*

$$T_{3n} = \sqrt{n}\boldsymbol{\Delta}_{m_1V}^{T}A\boldsymbol{\Delta}_{m_1V} \xrightarrow{p} 0.$$

Proof. In this proof we will use w_{ij} to denote $w(\mathbf{X}_i, \mathbf{X}_j)$. Writing

$$\begin{aligned} T_{3n} &= \frac{\sqrt{n}(np-1)}{n(n-1)p(p-1)} \sum_{i=1}^{n} \left(\sum_{j \in W_i} \Delta_{m_1}(\mathbf{X}_j) \right)^2 \\ &- \frac{\sqrt{n}p}{n(n-1)} \left(\sum_{i=1}^{n} \Delta_{m_1}(\mathbf{X}_i) \right)^2 - \frac{\sqrt{n}p}{n(p-1)} \sum_{i=1}^{n} \Delta_{m_1}^2(\mathbf{X}_i), \end{aligned} \tag{4.38}$$

we have to show that each of the three terms on the right hand side of (4.38) converges to zero in probability. For the first term notice first that

$$\begin{aligned} & E \left(n^{1/4} \sum_{j \in W_i} (\hat{m}_1(\mathbf{X}_j) - m_1(\mathbf{X}_j)) \right)^2 \\ = \; & E \left(n^{1/2} \sum_{j_1, j_2 \in W_i} (\hat{m}_1(\mathbf{X}_{j_1}) - m_1(\mathbf{X}_{j_1}))(\hat{m}_1(\mathbf{X}_{j_2}) - m_1(\mathbf{X}_{j_2})) \right) \\ = \; & E \left(n^{1/2} \sum_{j_1, j_2 \in W_i} \sum_{k_1} \sum_{k_2} (Y_{k_1} - m_1(\mathbf{X}_{j_1})) w_{i_1 k_1} (Y_{k_2} - m_1(\mathbf{X}_{j_2})) w_{j_2 k_2} \right) \\ = \; & E(n^{1/2} \sum_{j_1, j_2 \in W_i} \sum_{k_1} \sum_{k_2} (Y_{k_1} - m_1(\mathbf{X}_{k_1}) + m_1(\mathbf{X}_{k_1}) - m_1(\mathbf{X}_{j_1})) w_{i_1 k_1} \\ & \qquad \times (Y_{k_2} - m_1(\mathbf{X}_{k_2}) + m_1(\mathbf{X}_{k_2}) - m_1(\mathbf{X}_{j_2})) w_{j_2 k_2}) \\ = \; & n^{1/2} \sum_{j_1, j_2 \in W_i} \sum_{k_1 \neq k_2} (m_1(\mathbf{X}_{k_1}) - m_1(\mathbf{X}_{j_1})) w_{j_1 k_1} (m_1(\mathbf{X}_{k_2}) - m_1(\mathbf{X}_{j_2})) w_{j_2 k_2} \\ + \; & n^{1/2} \sum_{j_1, j_2 \in W_i} \sum_{k} \left[\sigma^2(\mathbf{X}_k) + (m_1(\mathbf{X}_k) - m_1(\mathbf{X}_{j_1}))(m_1(\mathbf{X}_k) - m_1(\mathbf{X}_{j_2})) \right] w_{j_1 k} w_{j_2 k} \\ \leq \; & O(n^{1/2} ||H_n^{1/2}||^2) + O(n^{-1/2} |H_n|^{-1/2}) + O(n^{-1/2} ||H_n^{1/2}||^2 |H_n|^{-1/2}), \end{aligned}$$

uniformly in i. Thus, the first term on the right hand side of (4.38), which can be written (up to a multiplicative constant) as

$$\frac{1}{n} \sum_{i=1}^{n} \left(n^{1/4} \sum_{j \in W_i} (\hat{m}(\mathbf{X}_j) - m_1(\mathbf{X}_j)) \right)^2,$$

is an average of terms that converge uniformly to zero in probability, and thus it converges in probability to zero. That the second term on the right hand side of (4.38) converges in probability to zero follows directly from (4.36). Finally,

showing that the third term on the right hand side of (4.38) converges to zero in probability is similar to the first term. □

4.2.1 Factorization for d dimensions

For a more general factorization of the mean regression function, we can still follow the ideas of analisys of variance. Assume we have predictors $\mathbf{X} = (X_1, \ldots, X_d)$. Follow the notation in Abramovich, F. (2009) we write

$$m(\mathbf{X}) = \mu + \sum_{i=1}^{d} m_i(X_i) + m_0(\mathbf{X}), \tag{4.39}$$

where μ is a constant, m_i are one dimensional functions of X_i (main effects) and m_0 involves all second and higher order interactions. For identifiability of the model it is assumed that

$$\int m_i(x_i) dF_{X_i}(x_i) = 0 \tag{4.40}$$

$$\int m_0(\mathbf{x}) dF_{\mathbf{X}}(\mathbf{x}) = 0 \tag{4.41}$$

$$\int m_0(\mathbf{x}) dF_{\mathbf{X}_{-i}}(\mathbf{x}_{-i}) = 0 \tag{4.42}$$

$$\tag{4.43}$$

where $\mathbf{x}_{-i} = (x, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d)$.

4.2.2 Dimension Reduction

When testing for the effect of the covariate Z using the proposed ANOVA-type method, it is obvious that the estimation of $m_1(\mathbf{x})$ is crucial: by estimating m_1 accurately, the power of the test will be as large as possible, whereas a bad estimation that leaves features of $\mathbf{x}$ in the mean function may result in lower power for detecting z and may also affect the level of the test. Note that if the dimension d_1 of $\mathbf{x}$ is big, then the estimation of m_1 will suffer from the curse of dimensionality, where the number of observations needed for local smoothing grows exponentially, specially due to the fact that it is assumed fully nonparametric in the methodology

we propose. Local polynomial regression with sufficient dimension smoothness assumptions can help, (see Section 4.3). Here we outline some procedures for dimension reduction, which have been proposed to improve the estimation of the mean function in a nonparametric regression, and in this section we discuss a few of them.

4.2.2.1 Uncorrelated Predictors

In a general model with $d > 1$ covariates, a covariate is partially related to the response Y if there is a significant association even after removing the effect of the other $(d-1)$ covariates. Usually, we measure the association of the covariate of interest with the residuals of the fitted model containing all other covariates. Another correlation type is called marginal correlation, where we measure the association between the response and the covariate of interest, without taking into account the other covariates.

Studying only marginal associations may be misleading, and a model that is not completely correct may be chosen. A covariate that is highly correlated with another covariate but do not have any relation with the response might be added to the model. This happens because it may be correlated to the response through and only through the other covariate, this is called spurious correlation. Therefore, selecting only those covariates with significant marginal relation with the response is not always completely correct, and partial relation needs to be explored. For a linear regression for example, it makes sense to eliminate from the model those covariates whose slopes have p-values greater than some constant, say 0.2 or 0.3. In a nonparametric regression model, we can use the test statistics MST-MSE for each singular covariate X_j using the ANOVA-type test, augmenting (X_j, Y), i.e., test

$$\tilde{H}_0^{(j)} : E(Y|X_j) = C,\ j = 1, \ldots, d,$$

and eliminate those with p-values greater than a cuttof.

In this thesis dissertation, the goal is to test if there is a contribution of the covariate Z to the response on the mean function after taking into account the other possible covariates. Note that if all the other covariates are independent of Z, then testing if the mean function does not depend on Z can be done by simply

descarting all other covariates. Lemma 4.3 formalizes this concept.

Lemma 4.3. *If* $\mathbf{X}$ *and* Z *are independent it follows that*

$$E(Y|\mathbf{X}, Z) = E(Y|\mathbf{X}) \Rightarrow E(Y|Z) = E(Y). \tag{4.44}$$

Proof. Using the decomposition of $E(Y|\mathbf{X}, Z) = m(\mathbf{X}, Z)$ given in (4.12), it is seen that $E(Y|\mathbf{X}, Z) = E(Y|\mathbf{X})$ implies $\tilde{m}_2(Z) = \tilde{m}_{12}(\mathbf{X}, Z) = 0$, so that

$$m(\mathbf{X}, Z) = \mu + \tilde{m}_1(\mathbf{X}).$$

Thus,

$$E(Y|Z) = E[m(\mathbf{X}, Z)|Z] = \mu + E[\tilde{m}_1(\mathbf{X})|Z] = \mu = E(Y).$$

since, by independence of $\mathbf{X}$ and Z, $E[\tilde{m}_1(\mathbf{X})|Z] = E[\tilde{m}_1(\mathbf{X})] = 0$. □

In a more general and realistic case, the covariate to be tested Z might be independent of some covariates but not of other covariates. In that case, it is intuitive to use the ANOVA-type test using only the covariates that are not independent of Z. Lemma 4.4 proves that excluding the independent covariates in the test is adequate, when we have conditional independence.

Lemma 4.4. *Let* $\mathbf{X} = (\mathbf{X}, Z) = (\mathbf{X}^I, \mathbf{X}^D, Z)$*, such that* $\mathbf{X}^I$ *is conditionally independent of* Z *given* $\mathbf{X}^D$*, then it follows that*

$$E(Y|\mathbf{X}, Z) = E(Y|\mathbf{X}) \Rightarrow E(Y|\mathbf{X}^D, Z) = E(Y|\mathbf{X}^D).$$

Proof. Following the ANOVA-type decomposition idea in (4.12), the decomposition of $E(Y|\mathbf{X}, Z) = m(\mathbf{X}, Z)$ is given by

$$\begin{aligned} m(\mathbf{X}^I, \mathbf{X}^D, Z) &= \mu + \tilde{m}_I(\mathbf{X}^I) + \tilde{m}_D(\mathbf{X}^D) + \tilde{m}_2(Z) + \tilde{m}_{ID}(\mathbf{X}^I, \mathbf{X}^D) \\ &\quad + \tilde{m}_{I2}(\mathbf{X}_1^I, Z) + \tilde{m}_{D2}(\mathbf{X}^D, Z) + \tilde{m}_{ID2}(\mathbf{X}^I, \mathbf{X}^D, Z), \end{aligned} \tag{4.45}$$

where

$$\mu = \int\int\int m(\mathbf{x}^I, \mathbf{x}^D, z) dF_{\mathbf{X}^I}(\mathbf{x}^I) dF_{\mathbf{X}^D}(\mathbf{x}^D) dF_Z(z),$$

$$
\begin{aligned}
\tilde{m}_I(\mathbf{x}^I) &= \int\int m(\mathbf{x}^I,\mathbf{x}^D,z)dF_{\mathbf{X}^D}(\mathbf{x}^D)dF_Z(z)-\mu,\\
\tilde{m}_D(\mathbf{x}^D) &= \int\int m(\mathbf{x}^I,\mathbf{x}^D,z)dF_{\mathbf{X}^I}(\mathbf{x}^I)dF_Z(z)-\mu,\\
\tilde{m}_2(z) &= \int\int m(\mathbf{x}^I,\mathbf{x}^D,z)dF_{\mathbf{X}^I}(\mathbf{x}^I)dF_{\mathbf{X}^D}(\mathbf{x}^D)-\mu,\\
\tilde{m}_{ID}(\mathbf{x}^I,\mathbf{x}^D) &= \int m(\mathbf{x}^I,\mathbf{x}^D,z)dF_Z(z)-\tilde{m}_I(\mathbf{x}^I)-\tilde{m}_D(\mathbf{x}^D)-\mu\\
\tilde{m}_{I2}(\mathbf{x}^I,z) &= \int m(\mathbf{x}^I,\mathbf{x}^D,z)dF_{\mathbf{X}^D}(\mathbf{x}^D)-\tilde{m}_I(\mathbf{x}^I)-\tilde{m}_2(z)-\mu\\
\tilde{m}_{D2}(\mathbf{x}^D,z) &= \int m(\mathbf{x}^I,\mathbf{x}^D,z)dF_{\mathbf{X}^I}(\mathbf{x}^I)-\tilde{m}_D(\mathbf{x}^D)-\tilde{m}_2(z)-\mu\\
\tilde{m}_{ID2}(\mathbf{x}^I,\mathbf{x}^D,z) &= m(\mathbf{x}^I,\mathbf{x}^D,z)-\tilde{m}_I(\mathbf{x}^I)-\tilde{m}_D(\mathbf{x}^D)-\tilde{m}_2(z)\\
&\quad -\tilde{m}_{ID}(\mathbf{x}^I,\mathbf{x}^D)-\tilde{m}_{I2}(\mathbf{x}^I,z)-\tilde{m}_{D2}(\mathbf{x}^D,z)-\mu.
\end{aligned}
$$

This definition implies

$$
\begin{aligned}
0 &= \int \tilde{m}_I(\mathbf{x}^I)dF_{\mathbf{X}^I}(\mathbf{x}^I)=\int \tilde{m}_D(\mathbf{x}^D)dF_{\mathbf{X}^D}(\mathbf{x}^D)=\int \tilde{m}_2(z)dF_Z(z),\\
0 &= \int \tilde{m}_{ID}(\mathbf{x}^I,\mathbf{x}^D)dF_{\mathbf{X}^I}(\mathbf{x}^I)=\int \tilde{m}_{ID}(\mathbf{x}^I,\mathbf{x}^D)dF_{\mathbf{X}^D}(\mathbf{x}^D)\\
&= \int \tilde{m}_{I2}(\mathbf{x}^I,z)dF_{\mathbf{X}^I}(\mathbf{x}^I),\\
0 &= \int \tilde{m}_{I2}(\mathbf{x}^I,z)dF_Z(z)=\int \tilde{m}_{D2}(\mathbf{x}^D,z)dF_{\mathbf{X}^D}(\mathbf{x}^D)\\
&= \int \tilde{m}_{D2}(\mathbf{x}^D,z)dF_Z(z),\\
0 &= \int \tilde{m}_{ID2}(\mathbf{x}^I,\mathbf{x}^D,z)dF_{\mathbf{X}^I}(\mathbf{x}^I)=\int \tilde{m}_{ID2}(\mathbf{x}^I,\mathbf{x}^D,z)dF_{\mathbf{X}^D}(\mathbf{x}^D)\\
&= \int \tilde{m}_{ID2}(\mathbf{x}^I,\mathbf{x}^D,z)dF_Z(z).
\end{aligned}
$$

It is seen that $E(Y|\mathbf{X},Z)=E(Y|\mathbf{X})$ implies
$\tilde{m}_2(Z)=\tilde{m}_{I2}(\mathbf{X}^I,Z)=\tilde{m}_{1D2}(\mathbf{X}^D,Z)=\tilde{m}_{ID2}(\mathbf{X}^I,\mathbf{X}^D,Z)=0$, so that

$$
m(\mathbf{X},Z)=\mu+\tilde{m}_I(\mathbf{X}^I)+\tilde{m}_D(\mathbf{X}^D)+\tilde{m}_{ID}(\mathbf{X}^I,\mathbf{X}^D).
$$

Thus, $E(Y|\mathbf{X}^D, Z)$ is equal to

$$\begin{aligned} & \mu + E[\tilde{m}_I(\mathbf{X}^I)|\mathbf{X}^D, Z] + E[\tilde{m}_D(\mathbf{X}^D)|\mathbf{X}^D, Z] + E[\tilde{m}_{ID}(\mathbf{X}^I, \mathbf{X}^D)|\mathbf{X}^D, Z] \\ = \; & \mu + E[\tilde{m}_I(\mathbf{X}^I)|X^D] + \tilde{m}_D(\mathbf{X}^D) + E[\tilde{m}_{ID}(\mathbf{X}^I, \mathbf{X}^D)|X^D] = E(Y|\mathbf{X}^D). \end{aligned}$$

since, by the conditional independence of $\mathbf{X}^I$ and Z given $\mathbf{X}^D$, $dF_{\mathbf{X}^I|\mathbf{X}^D,Z} = dF_{\mathbf{X}^I|\mathbf{X}^D}$. □

4.2.2.2 Principal Components

One of the most used methodologies for dimension reduction is Principal Components (PCA). It uses an orthogonal transformation of the matrix $\mathbf{X}$ to convert it into a set of uncorrelated covariates called principal components. This transformation gives components such that the first accounts for as much of the variability in the data as possible, and each of the succeeding components accounts for as much variability as possible with the restriction that they are orthogonal to the preceding components.

Supervised Principal Components makes use of the covariates that are related with the response to build the regression function. For a more detailed description of this method see Section 4.4.

4.2.2.3 SIR - Sliced Inverse Regression

The Sliced Inverse Regression method proposed by Li (1991) is a data analytic procedure for reducing the dimension of the covariate vector $\mathbf{X}$ without doing any fitting of the model. It is assumed that a small number of linear combinations of the predictors contain all information about y. More specifically, it is assumed that $y = f(\mathbf{x}'\beta_1, \ldots, \mathbf{x}'\beta_K, \epsilon)$, where β's are unknown row vectors and ϵ is independent of $\mathbf{x}$. This is an extension of the so called single index model. The estimation of these directions are based on the inverse regression ox $\mathbf{x}'$ on y coordinate-wise. The procedure is as follows: let $\mathbf{x}_i, i = 1, \ldots, n$ be the p-dimensional vector of the observation i; standardize it getting $\tilde{\mathbf{x}}_i = \hat{\Sigma}^{-1/2}(\mathbf{x}_i - \bar{\mathbf{x}})$; divide the range of y in H slices and let $\hat{p}_h$ be the proportion of y_i that falls in slice h; calculate $\hat{m}_h = \frac{1}{n\hat{p}_h} \sum_{y_i \text{ in slice}h} \tilde{\mathbf{x}}_i$; perform a weighted principal components analysis by

getting the eigenvectors and eigenvalues of $\hat{V} = \sum_{h=1}^{H} \hat{p}_h \hat{m}_h \hat{m}_h'$; the result is $\hat{\beta}_k = \hat{\eta}_k \hat{\Sigma}^{-1/2} (k = 1, \ldots, K)$, where $\hat{\eta}_k$ are the K largest eigenvectors of $\hat{V}$.

The number of eigenvectors can be chosen by a sequential test by testing if the K first eigenvalues of V are significantly different from zero compared to the sampling error, procedure which can be established by the following theorem stated by Li (1991): If $\mathbf{x}$ is normally distributed, then $n(p-K)\tilde{\lambda}_{p-K}$ follows a χ^2 distribution with $(p-K)(H-K-1)$ degrees of freedom asymptotically, where $\tilde{\lambda}_{p-K}$ is the average of the smallest $p-K$eigenvalues of $\hat{V}$.

4.2.2.4 Simple Screening

The variable screening consists of performing the marginal test of Wang, Akritas and Van Keilegom (2008) for the signicance of each variable, and keeping those variables for which the p-value is less than 0.5. This may decrease computational time for variable selection procedures based on backward elimination (see Chapter 5).

4.2.2.5 Simulations: Dimension Reduction Methods

In this section we run simulations in different scenarios to compare how much the ANOVA-type test statistic can be improved by using each of the dimension reduction techniques described above.

We run 2000 simulations in each case, the results are in the tables below. In all simulations in this section, the ANOVA-type test was calculated with $p = 9$. Table 4.7 shows the results of 200 observations generated from the linear model $Y = \mathbf{X}\boldsymbol{\beta} + \theta Z + \epsilon$, where $\epsilon \sim N(0, 3^2)$, $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0)$ and $\mathbf{U} = (\mathbf{X}, Z)$ is generated from a multivariate Normal distribution with mean 0 and variance Σ with entries $\Sigma_{ij} = 0.5^{|i-j|}, i, j = 1, \ldots, 8$. In this table, for the test is $H_0 : m(\mathbf{U}) = m_1(\mathbf{X})$, we report the NOVA-type, and the ANOVA-type using SIR(Sliced Inverse Regression), PC(Principal Components), SPC(Supervised Principal Components) and Partial Correlations. Then numbers between parenthesis represent the number of components used.

Clearly, when the dimensions of $\mathbf{X}$ are large, the level of the ANOVA-type test statistic is not maintained, while using the any of the dimension reduction

Table 4.7. Proportion of rejections for Null and linear alternatives

			θ		
test	0	.5	1	2	3
ANOVA-type	.135	.183	.408	.902	.915
ANOVA-type+SIR	.065	.087	.337	.946	.989
ANOVA-type+PC(1)	.089	.063	.177	.819	.999
ANOVA-type+PC(2)	.051	.139	.397	.973	.999
ANOVA-type+Partial Corr.(.15)	.059	.077	.178	.652	.973
ANOVA-type+SPC(1)	.060	.105	.262	.882	.998
ANOVA-type+SPC(2)	.060	.147	.477	.980	.999

techniques the level is a little liberal. The tests using SPC and SIR perform the best in terms of power, followed by the test using PC. Since the variables are correlated, the test using the partial correlation technique does not do as well as the others, but it mantains an accurate level.

Table 4.8 shows the results of 200 observations generated from the linear model $Y = \mathbf{X}\boldsymbol{\beta} + \theta Z + \epsilon, k = 1, \ldots, 2$, where $\epsilon \sim N(0, 3^2)$,

$$\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0)$$

and $\mathbf{U}_1 = (\mathbf{X}_1, Z)$ and $\mathbf{U}_2 = (\mathbf{X}_2, Z)$ are each generated from a multivariate Normal distribution with mean 0 and variances Σ_1 and Σ_2 with entries $\{\Sigma_1\}_{ij} = 0.5^{|i-j|}, \{\Sigma_2\}_{ij} = 0^{|i-j|}, i, j = 1, \ldots, 25$. In this table the test is $H_0 : m(\mathbf{U}) = m_1(\mathbf{X})$.

The ANOVA-test statistic used alone is not able to detect anything when the number of dimensions is very large. The case where the covariates are dependent with covariance Σ_1, the level of the test with each technique is good, except for SIR, which is a little liberal. In general, the tests using SIR or SPC perform better, achieving more power. In the independent case, again SIR and SPC outperform both principal components and partial correlation by a lot.

Table 4.9 shows the results of 200 observations generated from the non linear model $Y = sin((3/4)\pi X_1) - 3\Phi(-|X_5|^3) + \theta sin(\pi Z) + \epsilon$, where $\epsilon \sim N(0, 0.1^2)$, $\mathbf{U} = (\mathbf{X}, Z)$ is a 8 dimensional random vector, where each dimension is generated from an independent Normal distribution with mean 0 and variance 1. In this table the test is $H_0 : m(\mathbf{U}) = m_1(\mathbf{X})$.

Table 4.8. Proportion of rejections for Null and linear alternatives

	Σ_1				
			θ		
test	0	.5	1	2	3
ANOVA-type	0	0	0	0	0
ANOVA-type+SIR	.072	.069	.224	.847	.965
ANOVA-type+PC(1)	.058	.084	.172	.679	.976
ANOVA-type+PC(2)	.058	.078	.191	.705	.990
ANOVA-type+Partial Corr.	.048	.061	.110	.442	.864
ANOVA-type+SPC(1)	.054	.105	.241	.848	.996
ANOVA-type+SPC(2)	.054	.088	.261	.877	.997
	Σ_2				
			θ		
test	0	.5	1	2	3
ANOVA-type	0	0	0	0	0
ANOVA-type+SIR	.044	.161	.607	.970	.983
ANOVA-type+PC(1)	.061	.100	.204	.771	.993
ANOVA-type+PC(2)	.062	.088	.216	.758	.990
ANOVA-type+Partial Corr.	.056	.085	.187	.750	.994
ANOVA-type+SPC(1)	.059	.105	.224	.825	.994
ANOVA-type+SPC(2)	.057	.105	.266	.861	.998

Table 4.9. Proportion of rejections for Null and non-linear alternatives

			θ		
test	0	.2	.4	.6	.8
ANOVA-type	.301	.290	.341	.361	.362
ANOVA-type+SIR	.054	.184	.747	.992	1
ANOVA-type+PC(1)	.047	.209	.765	.987	1
ANOVA-type+PC(2)	.055	.194	.760	.989	1
ANOVA-type+Partial Corr.	.054	.196	.771	.994	1
ANOVA-type+SPC(1)	.045	.241	.841	.992	1
ANOVA-type+SPC(2)	.059	.532	.921	.997	1

In the case of the non linear model, the ANOVA-type test again fails to mantain the level by itself. When using any dimension reduction technique, it keeps the level near the expected 0.05. For the model used in Table 4.9 SPC seems to have a better power, followed by SIR and PC, which have similar power.

4.2.3 Discussion on the dependence of Z on X

When Z is independent of $\mathbf{X}$, the ANOVA-type test statistic has a good level and power, as was seen in simulations so far. For highly dependent data, say we have X highly correlated with Z, when fitting a model with X, most of the information of Z will be extracted from the response Y and little will be left in the residuals. This challenge is faced by all models when performing regression analysis. When Z is correlated with X but not with Y, we have what is called spurious correlation.

We investigate how this problem affects the proposed test statistic in the following simulation. The data is generated from the model

$$Y = -X + X^3 + f_j(Z) + \epsilon,\ j = 1, 2,$$

where $\epsilon \sim N(0,1)$, $f_0(z) = 0$, $f_1(z) = z$, $f_2(z) = sin(\pi z)$, $X = U_1 + \theta U_2$, $Z = U_2 + \theta U_1$ with U_1 and U_2 uniformly distributed on $(-0.5, 0.5)$. Table 4.10 reports the results for 2000 simulations with data sets of size 100.

Table 4.10. Rejection rates for growing dependence

		θ				
f	test	0	.3	.5	.7	.9
f_0	ANOVA-type(7)	.057	.047	.036	.025	.010
	ANOVA-type(9)	.046	.041	.032	.015	.004
f_1	ANOVA-type(7)	.391	.210	.074	.025	.006
	ANOVA-type(9)	.404	.229	.073	.020	.005
f_2	ANOVA-type(7)	.998	.965	.671	.122	.009
	ANOVA-type(9)	.998	.977	.719	.132	.005

From the table, we see that as the dependence of X with Z increases the level of the test decreases, becoming more and more conservative. This is intuitive, since, for highly dependent covariates X and Z, the estimator $\hat{m}_1(X)$ contains most information of Z, and therefore this will be subtracted from the residuals.

The power of the test is also affected by high dependence. Note that, when θ is larger than 0.5, which corresponds to $corr(X, Z) \sim 0.77$, the power drops drastically.

4.2.4 Simulations: Comparison with Generalized Likelihood Ratio Test

The GLR test is designed for additive models, which is exactly the simulation setting of Table 4.2. Under non-additive alternatives, however, it can perform poorly as indicated by the simulations reported in Table 4.11. These simulations use sample size $n = 200$ with data generated from the model $Y = X_1^Z(1+\theta X_2)+\frac{Z^{(1+\theta X_2)}}{Z}+\epsilon$, where $\epsilon \sim N(0, 0.1)$, and X_1, X_2, Z are i.i.d. $U(0.5, 2.5)$. The hypothesis tested is that $m(X_1, X_2, Z) = m_1(X_1, X_2)$. The residuals for the ANOVA-type test in the first part of Table 4.11 are based on a Nadaraya-Watson fit with kernel the uniform on $(-0.5, 0.5) \times (-0.5, 0.5)$ density and the common bandwidth selected through leave-one-out cross validation.

Table 4.11. Rejection rates for non-additive models

	θ				
test	0	0.02	.04	.06	.08
ANOVA-type (9)	.052	.176	.609	.940	.994
GLR	.048	.082	.110	.189	.304

It should be mentioned that the GLR test does not maintain its level under heteroscedasticity. In simulations, reported in the Table 4.12, under the additive but heteroscedastic model $Y = X^2 + \theta\cos(\pi Z) + Z\epsilon$, X, Z i.i.d. $\mathrm{N}(0, 1)$, $\epsilon \sim \mathrm{N}(0, 0.5)$, using sample size $n = 200$, the GLR test is very liberal while the ANOVA-type test maintains an accurate level.

Table 4.12. Rejection rates for heteroscedastic models

	θ				
test	0	0.02	.04	.06	.08
ANOVA-type (9)	.053	.067	.124	.485	.998
GLR	.465	.511	.624	.908	1

For another example of non-linearity, we simulate 1000 datasets of $n = 200$ observations from the model $Y = X_1X_2X_3X_4(1 + \theta Z) + \epsilon$, where $\epsilon \sim N(0, 0.1)$, X_1, X_2, X_3, X_4 and Z are i.i.d. $U(0.5, 2.5)$. The results on Table 4.13 show that the proposed ANOVA-type procedure performs better than Generalized likelihood

Table 4.13. Percentage of rejections for the model $Y = X_1X_2X_3X_4(1+\theta Z)+\epsilon$

Error Distr.	Method — θ	0	.05	.1	.2	.4
N(0,.1)	ANOVA-type (15)	.041	.123	.360	.900	1
	ANOVA-type (13)	.046	.127	.343	.886	.999
	ANOVA-type (11)	.043	.122	.321	.879	.999
	ANOVA-type (9)	.042	.114	.308	.844	.999
	GLRT	.056	.111	.276	.732	.992

ratio test when the model is not linear and there are more than 2 two covariates, even though the level is slightly conservative.

4.3 ANOVA-type hypothesis test using Local Polynomial Regression

The Nadaraya-Watson kernel estimators are fairly simple, intuitive and easy to interpret. Moreover, their mathematical properties are simple, computer implementation is straightforward and they are consistent for any smooth m, provided the density f satisfies certain assumptions. This estimator has been extensively used, and its properties have been deeply explored, see for example Rosenblatt (1969), Robinson (1983,1986), Collomb and Hardle (1986), Roussas (1990), Roussas and Tran (1992), Fan and Masry (1992).

Stone (1977) first introduced the idea of using a local polynomial to fit the nonparametric regression, this is a special case of the robust local regression estimators in Cleveland (1979). The idea is to use kernel weights for a weighted least squares regression in matrix theory. Stone (1980, 1982) uses the local polynomial fitting and its generalization to higher order polynomials to show the optimal rates of convergence and achievability, assuming certain smooth conditions on the mean regression function m. Cleveland and Devlin (1988) applied the local polynomial regression estimator in multivariate regression.

It has been shown that the local polynomial estimator has several advantages over the simple Nadaraya-Watson kernel regression (local constant regression). Chu and Marron (1991) and Fan (1992) show that it reduces bias. Fan and Gijbels (1992), Hastie and Loader (1993) and Rupert and Wand (1994) point that this

kind of estimator can adapt to the boundary of design points. Its superiority to the local constant estimator includes the estimation of derivatives of the regression function (Fan and Gijbels, 1992).

Ruppert and Wand (1994) explored the case of multivariate predictor variables fitted with a local quadratic approximation using general multivariate kernel weights. They derived the bias and variance of the local linear and quadratic regression approximations, and higher order polynomial fits for the univariate case. Masry (1996) derived the bias and variance terms for the multivariate case using a general local high order q polynomial fitting, and established the asymptotic normality of this estimate of the regression function and its partial derivatives for strongly mixing ρ-mixing processes.

In this section we will extend the methodology described in Section 4.2 for local polynomial fitting. Due to the fact that this type of estimator has better properties, including smaller bias, the goal is to extend Theorem 4.2, so that it can hold with $\mathbf{X}$ of higher dimension.

Consider the same setting described in Section 4.2. Let $\mathbf{X}$ be a vector of covariates with dimensions $d_1 > 1$ and Z be a univariate covariate, where $\mathbf{X}$ and Z have densities $f_{\mathbf{X}}$ and f_Z respectively. For the nonparametric regression model

$$Y_i = m(\mathbf{X}_i, Z_i) + \sigma(\mathbf{X}_i, Z_i)\epsilon_i, i = 1...n, \tag{4.46}$$

where ϵ_i is the independent error with zero mean and constant variance and independent of $\mathbf{X}$ and Z, recall that the goal is to test the hypothesis (4.25)

$$H_0 : m(\mathbf{x}, z) = m_1(\mathbf{x}).$$

To estimate $m_1(\mathbf{x})$ with a local polynomial, we assume that it has $q+1$ derivatives, see Section 1.1.1 for more details. The residual $\xi_i = Y_i - m_1(\mathbf{X}_i)$ is estimated by $\hat{\xi}_i = Y_i - \hat{m}_1(\mathbf{X}_i)$, where $\hat{m}_1$ now will be estimated by the local polynomial of order q regression estimator

$$\hat{m}_1(\mathbf{X}_i) = \mathbf{e}_1^T \left(\mathbb{X}_{\mathbf{X}_i}^T \mathbb{W}_{\mathbf{X}_i} \mathbb{X}_{\mathbf{X}_i}\right)^{-1} \mathbb{X}_{\mathbf{X}_i}^T \mathbb{W}_{\mathbf{X}_i} \mathbf{Y} = \sum_{j=1}^{n} \tilde{w}(\mathbf{X}_i, \mathbf{X}_j) Y_j, \ i = 1...n, \tag{4.47}$$

where

$$\mathbb{X}_{\mathbf{x}} = \begin{pmatrix} 1 & (\mathbf{X}_1 - \mathbf{x})^T & \text{vech}^T\left\{(\mathbf{X}_1 - \mathbf{x})(\mathbf{X}_1 - \mathbf{x})^T\right\} & \dots \\ \vdots & \vdots & \vdots & \dots \\ 1 & (\mathbf{X}_n - \mathbf{x})^T & \text{vech}^T\left\{(\mathbf{X}_n - \mathbf{x})(\mathbf{X}_n - \mathbf{x})^T\right\} & \dots \end{pmatrix}$$

is the $n \times \gamma_d$ design matrix, with

$$\gamma_d = \sum_{j=0}^{q} \sum_{\substack{k_1=0 \\ k_1+\ldots+k_d=j}}^{j} \cdots \sum_{k_d=0}^{j} 1,$$

and $\mathbb{W}_{\mathbf{x}} = diag\{K_H(\mathbf{X}_1 - \mathbf{x}), \ldots, K_H(\mathbf{X}_n - \mathbf{x})\}$, for the kernel function $K_{H_n}(\mathbf{x}) = |H_n|^{-1/2} K(H_n^{-1/2}\mathbf{x})$. We assume that $K(\cdot)$ is a bounded d_1-variate kernel function of bounded variation and with bounded support, and $H_n^{1/2}$ is a symmetric positive definite $d_1 \times d_1$ matrix called the bandwidth matrix. Also, vech is the half-vectorization operator, which for example, for a $d \times d$ matrix A, with entries $(A)_{ij} = a_{ij}$ would compute

$$\text{vech}(A) = [a_{11}, a_{21}, \ldots, a_{d1}, a_{22}, \ldots, a_{d2}, \ldots, a_{(d-1)(d-1)}, a_{(d-1)d}, a_{dd}]^T.$$

After computing $\hat{\xi}_i, i = 1, \ldots, n$ set the windows W_i as described in Section 4.1, and calculate the vector of $(n - p + 1)p$ "constructed observations" in the augmented one-way ANOVA design

$$\hat{\boldsymbol{\xi}}_V = (\hat{\xi}_j, j \in W_{(p-1)/2+1}, \ldots, \hat{\xi}_j, j \in W_{n-(p-1)/2})^T. \tag{4.48}$$

Let MST = MST($\hat{\boldsymbol{\xi}}_V$), MSE = MSE($\hat{\boldsymbol{\xi}}_V$) denote the balanced one-way ANOVA mean squares due to treatment and error, respectively, computed on the data $\hat{\boldsymbol{\xi}}_V$. The proposed test statistic is

$$MST - MSE = \hat{\boldsymbol{\xi}}_V^T A \hat{\boldsymbol{\xi}}_V. \tag{4.49}$$

Consider the following conditions

a) $E|Y|^\rho < \infty$ for some $\rho > 2$.

b) the marginal densities $f_{\mathbf{X}}$, f_Z of $\mathbf{X}$, Z, respectively, are bounded away from

zero.

c) $f_{\mathbf{X}}$ is uniformly continuous and bounded.

d) the $q+1$ derivatives of $m_1(\mathbf{x})$ exist and are Lipschitz uniformly continuous and bounded.

e) $\sigma^2(.,z) := E(\xi^2|Z=z)$ is Lipschitz continuous, $\sup_{\mathbf{u}} \sigma^2(\mathbf{u}) < \infty$, and $E(\epsilon_i^4) < \infty$.

Also, assume that the eigenvalues, $\lambda_i,\ i=1,\ldots,d_1$, of the bandwidth matrix $H_n^{1/2}$ defined in (4.6), converge to zero at the same rate and satisfy:

$$\begin{aligned}
&1) \quad n\lambda_i^{4(q+1)} \to 0 \ i=1,\ldots,d_1.\\
&2) \quad \frac{n\lambda_i^{2d_1}}{(\log n)^2} \to \infty,\ i=1,\ldots,d_1.\\
&3) \quad \frac{n^{1-2/\rho}\lambda_i^{d_1}}{\ln n[\ln n(\ln \ln n)^{1+\delta}]^{2/\rho}} \to \infty,\ i=1,\ldots,d_1.
\end{aligned}$$

The following theorem gives the asymptotic normal distribution of the test statistic under the null hypothesis for the local polynomial fitting.

Theorem 4.3. *Let conditions a)-e) and the conditions on the bandwidth 1)-3) hold. Then, under H_0 in (4.25), the asymptotic distribution of the test statistic in (4.49) is given by*

$$n^{1/2}(MST - MSE) \xrightarrow{d} N\left(0, \frac{2p(2p-1)}{3(p-1)}\tau^2\right),$$

where $\tau = \int \left[\int \sigma^2(\mathbf{x},z) f_{\mathbf{X}|Z=z}(\mathbf{x}) d\boldsymbol{x}\right]^2 f_Z(z)dz$.

In the same way as before, an estimate of τ^2 can be obtained by modifying Rice's (1984) estimator as follows

$$\hat{\tau}^2 = \frac{1}{4(n-3)} \sum_{j=2}^{n-2} (\hat{\xi}_j - \hat{\xi}_{j-1})^2 (\hat{\xi}_{j+2} - \hat{\xi}_{j+1})^2.$$

Note that, in order to use the local polynomial approximation of $m_1(\mathbf{x})$ we have to assume that it has $q+1$ derivatives. That assumption may be a little restrictive.

On the other hand, it is seen from Theorem 4.2 that it is necessary to assume 2 derivatives in order for the kernel (local constant) method to work, and in that case, the number of covariates (dimension of $\mathbf{x}$) is restricted up to 3. By assuming $q+1$ derivatives and using the local polynomial estimator, the restriction on the covariates is to be less than $2(q+1)$,i.e, assuming more smoothness of the function, we can incorporate larger dimensions, at a good "rate" of twice as much.

To understand how more smoothness affects the theorem, note that Masry (1996) showed that

$$\sup_{\mathbf{x}} |\hat{m}_1(\mathbf{x}) - m_1(\mathbf{x})| = O\left(\left(\frac{log(n)}{n\lambda_i^{d_1}}\right)^{1/2}\right) + O\left(\lambda_i^{q+1}\right) \tag{4.50}$$

The two terms in (4.50) correspond to the "variance" and "bias" terms. We see that the variance term does not depend on the number of derivatives, but the bias term does. By assuming more derivatives, the bias of the estimation is smaller, and therefore $\hat{m}_1(\mathbf{x})$ is closer to $m_1(\mathbf{x})$, which allows the convergence to zero of some terms of the test statistic and thus, the convergence of the test statistic to the normal distribution.

Proof of Theorem 4.3. Under H_0 in (4.25) we have

$$\begin{aligned}\hat{\xi}_i &= Y_i - \hat{m}_1(\mathbf{X}_i) + m_1(\mathbf{X}_i) - m_1(\mathbf{X}_i) = \xi_i - (\hat{m}_1(\mathbf{X}_i) - m_1(\mathbf{X}_i)) \\ &= \xi_i - \Delta_{m_1}(\mathbf{X}_i),\end{aligned}$$

where now $\hat{m}_1(\mathbf{X}_i)$ is the local polynomial estimator given by (4.47). Thus, $\hat{\boldsymbol{\xi}}_V$ of relation (4.48) is decomposed as

$$\begin{aligned}\hat{\boldsymbol{\xi}}_V &= \{\hat{\xi}_j : j \in W_1, ..., \hat{\xi}_j : j \in W_n\}' \\ &= \{\xi_j - \Delta_{m_1}(\mathbf{X}_j) : j \in W_1, ..., \xi_j - \Delta_{m_1}(\mathbf{X}_j) : j \in W_n\}' \\ &= \boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V},\end{aligned}$$

and $\sqrt{n}$(MST - MSE) can be written as

$$\sqrt{n}\hat{\boldsymbol{\xi}}_V^T A\hat{\boldsymbol{\xi}}_V = \sqrt{n}\boldsymbol{\xi}_V^T A\boldsymbol{\xi}_V - \sqrt{n}2\boldsymbol{\xi}_V^T A\boldsymbol{\Delta}_{m_1 V} + \sqrt{n}\boldsymbol{\Delta}_{m_1 V}^T A\boldsymbol{\Delta}_{m_1 V}. \tag{4.51}$$

The asymptotic normality of $\sqrt{n}\boldsymbol{\xi}_V^T A\boldsymbol{\xi}_V$ follows by arguments similar to those used in Theorem 3.2 of Wang, Akritas and VanKeilegom (2008). The asymptotic variance of $\sqrt{n}\boldsymbol{\xi}_V^T A\boldsymbol{\xi}_V$, and therefore of the test statistic, is exactly the same as the one derived in Theorem 4.2, since it is the same term composed by the real (not estimated) random vector $\boldsymbol{\xi}_V$.

It remains to show that the other two terms in (4.51) converge to zero in probability. That the second and third terms in (4.51) converge in probability to zero are shown in Lemmas 4.5, 4.6, respectively.

□

Lemma 4.5. *The second term in (4.51) converges in probability to zero, i.e.*

$$T_{2n} := \sqrt{n}\boldsymbol{\xi}_V^T A\boldsymbol{\Delta}_{m_1 V} \xrightarrow{p} 0.$$

Proof. After some algebra it can be seen that

$$\begin{aligned} T_{2n} &= \frac{n^{-1/2}(np-1)}{(n-1)p(p-1)} \sum_{i=1}^{n} \sum_{j \in W_i} \xi_j \sum_{k \in W_i} \Delta_{m_1}(\mathbf{X}_k) \\ &\quad - \frac{n^{-1/2}p}{(n-1)} \sum_{i=1}^{n} \xi_i \sum_{j=1}^{n} \Delta_{m_1}(\mathbf{X}_j) - \frac{n^{-1/2}p}{(p-1)} \sum_{i=1}^{n} \xi_i \Delta_{m_1}(\mathbf{X}_i). \end{aligned} \tag{4.52}$$

We will show that each of the three terms above converge in probability to zero conditionally on $\mathbf{U} = \{\mathbf{X}, Z\}$, and thus also unconditionally. Note that, because all windows W_i are of finite size (p), the first term on the right hand side of (4.52) can be written as a finite (p^2) sum of terms each of which is similar to the last term in (4.52). Thus, it suffices to show that the last and second terms of (4.52) converge to zero. For notational simplicity, all expectations and variances in this proof are to be understood as conditional on $\mathbf{U} = \{\mathbf{X}, Z\}$. For the second term in (4.52), note that $n^{-1/2}\sum_{i=1}^{n} \xi_i$ remains bounded in probability, and therefore, its convergence to zero will follow if we show that $n^{-1}\sum_{k=1}^{n} \Delta_{m_1}(\mathbf{X}_k) \xrightarrow{p} 0$. For later use, we will actually show that

$$\frac{1}{n^{3/4}} \sum_{k=1}^{n} \Delta_{m_1}(\mathbf{X}_k) = \frac{1}{n^{3/4}} \sum_{k=1}^{n} (\hat{m}_1(\mathbf{X}_k) - m_1(\mathbf{X}_k)) \xrightarrow{p} 0. \tag{4.53}$$

By Theorem 6 in Masry (1996), it follows that

$$\sup_{\mathbf{x}} |\hat{m}_1(\mathbf{x}) - m_1(\mathbf{x})| = O\left(\left(\frac{log(n)}{n\lambda_i^{d_1}}\right)^{\frac{1}{2}}\right) + O\left(\lambda_i^{q+1}\right). \tag{4.54}$$

Thus

$$\frac{1}{n^{3/4}}\sum_{k=1}^{n}\Delta_{m_1}(\mathbf{X}_k) = O\left(n^{1/4}\left(\frac{log(n)}{n\lambda_i^{d_1}}\right)^{\frac{1}{2}}\right) + O\left(n^{1/4}\lambda_i^{q+1}\right) = o(1),$$

where the last equality follows from the assumptions of the Theorem 4.3.

Consider now the last term in (4.52). Because the weights $\tilde{w}(\mathbf{X}_i, \mathbf{X}_j)$ of the local polynomial regression sum to 1 (Lemma 4.18), we can write

$$\begin{aligned}
\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\xi_i\Delta_{m_1}(\mathbf{X}_i) &= n^{-1/2}\sum_{i=1}^{n}\xi_i(\hat{m}(\mathbf{X}_i) - m(\mathbf{X}_i)) \\
&= n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{w}(\mathbf{X}_i, \mathbf{X}_j)(m(\mathbf{X}_j) + \xi_j - m(\mathbf{X}_i))\xi_i \\
&= n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{w}(\mathbf{X}_i, \mathbf{X}_j)(m(\mathbf{X}_j) - m(\mathbf{X}_i))\xi_i \\
&\quad + n^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{w}(\mathbf{X}_i, \mathbf{X}_j)\xi_j\xi_i.
\end{aligned} \tag{4.55}$$

The first term of the right hand side of (4.55) has zero expectation, so it suffices to show that its variance goes to zero. To this end, we write

$$\begin{aligned}
&Var(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{n}\xi_i\tilde{w}(\mathbf{X}_i, \mathbf{X}_j)(m_1(\mathbf{X}_j) - m_1(\mathbf{X}_i))) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\tilde{w}(\mathbf{X}_i, \mathbf{X}_{j_1})\tilde{w}(\mathbf{X}_i, \mathbf{X}_{j_2})\mathsf{x} \\
&\qquad\qquad \mathsf{x}(m_1(\mathbf{X}_{j_1}) - m_1(\mathbf{X}_i))(m_1(\mathbf{X}_{j_2}) - m_1(\mathbf{X}_i))Var(\xi_i) \\
&\leq \frac{M}{n}\sum_{i=1}^{n}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\tilde{w}(\mathbf{X}_i, \mathbf{X}_{j_1})\tilde{w}(\mathbf{X}_i, \mathbf{X}_{j_2})\mathsf{x} \\
&\qquad\qquad \mathsf{x}(c||\mathbf{X}_{j_1} - \mathbf{X}_i||c||\mathbf{X}_{j_2} - \mathbf{X}_i||) \\
&= Mc^2O(||H_n^{1/2}||)O(||H_n^{1/2}||) = o(1),
\end{aligned}$$

for some constants M and c, where the inequality holds because $m_1(\cdot)$ is Lipschitz continuous, and the last equality follows from Lemma 4.19. Thus, by the assumptions of Theorem 4.2 the first term of the right hand side of (4.55) goes in probability to zero. To show that the second term in (4.55) also goes to 0 in probability, we will show that its second moment goes to zero. To this end, we write

$$
\begin{aligned}
& E\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{j=1}^{n}\xi_i\xi_j\tilde{w}(\mathbf{X}_i,\mathbf{X}_j)\right]^2 \\
&= E\left[\frac{1}{n}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j_1=1}^{n}\sum_{j_2=1}^{n}\xi_{i_1}\xi_{i_2}\xi_{j_1}\xi_{j_2}\tilde{w}(\mathbf{X}_{i_1},\mathbf{X}_{j_1})\tilde{w}(\mathbf{X}_{i_2},\mathbf{X}_{j_2})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}E(\xi_i^2\xi_j^2)\tilde{w}(\mathbf{X}_i,\mathbf{X}_j)^2 \\
&+\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}E(\xi_i^2\xi_j^2)\tilde{w}(\mathbf{X}_i,\mathbf{X}_i)\tilde{w}(\mathbf{X}_j,\mathbf{X}_j) \\
&+\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}E(\xi_i^2\xi_j^2)\tilde{w}(\mathbf{X}_i,\mathbf{X}_j)\tilde{w}(\mathbf{X}_j,\mathbf{X}_i) \\
&+\frac{1}{n}\sum_{i=1}^{n}E(\xi_i^4)\tilde{w}(\mathbf{X}_i,\mathbf{X}_i)\tilde{w}(\mathbf{X}_i,\mathbf{X}_i) \\
&= O\left(\frac{1}{n|H_n|}\right)+O\left(\frac{1}{n|H_n|}\right)+O\left(\frac{1}{n|H_n|}\right)+O\left(\frac{1}{n^2|H_n|}\right) \\
&= o(1), \qquad (4.56)
\end{aligned}
$$

by the fact that $E(\xi_i^4)$ is bounded and that each weight is of the order $\frac{1}{n|H_n|^{1/2}}$ (Lemma 4.19). Thus, by the assumptions of Theorem 4.2 the second term of the right hand side of (4.55) goes in probability to zero.

This completes the proof of Lemma 4.5. □

Lemma 4.6. *The third term in (4.51) converges in probability to zero, i.e.*

$$T_{3n} = \sqrt{n}\boldsymbol{\Delta}_{m_1V}^T A\boldsymbol{\Delta}_{m_1V} \xrightarrow{p} 0.$$

Proof. Similarly to Lemma 4.5, we can write

$$\begin{aligned} T_{3n} &= \frac{\sqrt{n}(np-1)}{n(n-1)p(p-1)} \sum_{i=1}^{n} \left(\sum_{j \in W_i} \Delta_{m_1}(\mathbf{X}_j) \right)^2 \\ &- \frac{\sqrt{n}p}{n(n-1)} \left(\sum_{i=1}^{n} \Delta_{m_1}(\mathbf{X}_i) \right)^2 - \frac{\sqrt{n}p}{n(p-1)} \sum_{i=1}^{n} \Delta_{m_1}^2(\mathbf{X}_i). \end{aligned} \tag{4.57}$$

we have to show that each of the three terms on the right hand side of (4.57) converges to zero in probability Again, because all windows W_i are of finite size (p), the first term on the right hand side of (4.57) can be written as a finite (p^2) sum of terms each of which is similar to the last term in (4.57). Thus, it suffices to show that the last and second terms of (4.57) converge to zero.

Recall that (Masry, 1996)

$$\sup_{\mathbf{x}} |\hat{m}_1(\mathbf{x}) - m_1(\mathbf{x})| = O\left(\left(\frac{log(n)}{n\lambda_i^{d_1}} \right)^{\frac{1}{2}} \right) + O\left(\lambda_i^{q+1}\right).$$

Replacing Δ_{m_1} by its order, the second term in (4.57) is or order

$$\begin{aligned} & O\left(\frac{\sqrt{n}}{n^2} \left(n \left(\frac{log(n)}{n\lambda_i^{(d-1)}} \right)^{\frac{1}{2}} + n\lambda_i^{q+1} \right)^2 \right) \\ & = O\left(\left(n^{1/4} \left(\frac{log(n)}{n\lambda_i^{d_1}} \right)^{\frac{1}{2}} + n^{1/4}\lambda_i^{q+1} \right)^2 \right) = o(1) \end{aligned}$$

where the last equality follows from the assumptions of the theorem.

Similarly, the third term in (4.57) is or order

$$\begin{aligned} & O\left(\frac{\sqrt{n}}{n} n \left(\left(\frac{log(n)}{n\lambda_i^{d_1}} \right)^{\frac{1}{2}} + \lambda_i^{q+1} \right)^2 \right) \\ & = O\left(\left(n^{1/4} \left(\frac{log(n)}{n\lambda_i^{d_1}} \right)^{\frac{1}{2}} + n^{1/4}\lambda_i^{q+1} \right)^2 \right) = o(1) \end{aligned}$$

This completes the proof of Lemma 4.6. □

4.4 ANOVA-type hypothesis test for multivariate X and Z

Assume we have n observations, $(Y_i, \mathbf{U}_i)$, $i = 1, \ldots, n$, of the random variable Y and covariates $\mathbf{U} = (\mathbf{X}, \mathbf{Z})$, where $\mathbf{X}$ and $\mathbf{Z}$ have dimensions d_1 and d_2 respectively $(r + s = d)$. Let $m(\mathbf{x}, \mathbf{z}) = E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z})$ denote the regression function. The heterocedastic nonparametric model is

$$Y = m(\mathbf{X}, \mathbf{Z}) + \sigma(\mathbf{X}, \mathbf{Z})\epsilon, \tag{4.58}$$

where ϵ has zero mean and constant variance and is independent from $\mathbf{X}$ and $\mathbf{Z}$. The goal is to test the null hypothesis that $\mathbf{Z}$ does not contribute to the regression function, i.e.

$$H_0 : m(\mathbf{x}, \mathbf{z}) = m_1(\mathbf{x}). \tag{4.59}$$

To test this hypothesis, we will adapt the approach of Wang, Akritas and Van Keilegom (2008) and Zambom and Akritas (2012) which constructs a test statistic based on an augmented one-way ANOVA design with levels corresponding to $\mathbf{Z}_i,\ i = 1, \ldots, n$, and response an estimated version of the null hypothesis residuals $\xi_i = Y_i - m_1(\mathbf{x}_i),\ i = 1, \ldots, n$. The idea being that the null hypothesis (4.59) implies the null hypothesis for the augmented one-way ANOVA design. The asymptotic theory for such a statistic relies on the methodology for high dimensional one-way ANOVA of Akritas and Papadatos (2004).

For estimated residuals we will use $\hat{\xi}_i = Y_i - \hat{m}_1(\mathbf{x}_i),\ i = 1, \ldots, n$, where $\hat{m}_1(\mathbf{x})$ is the local polynomial of order q regression estimator defined by

$$\hat{m}_1(\mathbf{X}_i) = \mathbf{e}_1^T \left(\mathbb{X}_{\mathbf{X}_i}^T \mathbb{W}_{\mathbf{X}_i} \mathbb{X}_{\mathbf{X}_i}\right)^{-1} \mathbb{X}_{\mathbf{X}_i}^T \mathbb{W}_{\mathbf{X}_i} \mathbf{Y} = \sum_{j=1}^{n} \tilde{w}(\mathbf{X}_i, \mathbf{X}_j) Y_j,\ i = 1...n, \tag{4.60}$$

where

$$\mathbb{X}_{\mathbf{x}} = \begin{pmatrix} 1 & (\mathbf{X}_1 - \mathbf{x})^T & \text{vech}^T\left\{(\mathbf{X}_1 - \mathbf{x})(\mathbf{X}_1 - \mathbf{x})^T\right\} & \ldots \\ \vdots & \vdots & \vdots & \ldots \\ 1 & (\mathbf{X}_n - \mathbf{x})^T & \text{vech}^T\left\{(\mathbf{X}_n - \mathbf{x})(\mathbf{X}_n - \mathbf{x})^T\right\} & \ldots \end{pmatrix}$$

is the $n \times \gamma_r$ design matrix, with

$$\gamma_r = \sum_{j=0}^{q} \sum_{\substack{k_1=0 \\ k_1+\ldots+k_r=j}}^{j} \cdots \sum_{k_r=0}^{j} 1,$$

and $\mathbb{W}_{\mathbf{x}} = diag\{K_H(\mathbf{X}_1 - \mathbf{x}), \ldots, K_H(\mathbf{X}_n - \mathbf{x})\}$, for a symmetric kernel function $K_{H_n}(\mathbf{x}) = |H_n|^{-1/2} K(H_n^{-1/2}\mathbf{x})$. We assume that $K(\cdot)$ is a bounded d_1-variate kernel function of bounded variation and with bounded support, and $H_n^{1/2}$ is a symmetric positive definite $d_1 \times d_1$ matrix called the bandwidth matrix. Also, vech is the half-vectorization operator.

Augmenting the levels $\mathbf{Z}_i,\ i = 1, \ldots, n$, of the ANOVA design, which is justifiable by the smoothing conditions we will introduce below, is complicated due to the multivariate nature of $\mathbf{Z}$. We propose to circumvent this issue by replacing the $\mathbf{Z}_i$ values by the corresponding values of a suitable univariate surrogate. In particular, we will use a nonlinear version of Bair, Hastie, Paul and Tibshirani's (2004) first supervised principal component, $P_\theta = \mathbf{Z}^T \mathbf{C}_\theta$, to be defined below, as the univariate surrogate of $\mathbf{Z}$.

Having a univariate surrogate of $\mathbf{Z}$, we augment each cell by including additional $p-1$ $\hat{\xi}_\ell$'s which correspond to the $p-1$ values of P_θ that are nearest to $P_{\theta,i} = \mathbf{Z}_i^T \mathbf{C}_\theta$. To be specific, we consider the $(\hat{\xi}_i, P_{\theta,i})$, $i = 1, \ldots, n$, arranged so that $P_{\theta,i_1} < P_{\theta,i_2}$ whenever $i_1 < i_2$, and for each $P_{\theta,i}$, $(p-1)/2 < i \leq n-(p-1)/2$, define the nearest neighbor window W_i as

$$W_i(\mathbf{C}_\theta) = \left\{ j : |\hat{F}_P(P_{\theta,j}) - \hat{F}_P(P_{\theta,i})| \leq \frac{p-1}{2n} \right\}, \tag{4.61}$$

where $\hat{F}_P$ is the empirical distribution function of P_θ. W_i defines the augmented cell corresponding to $P_{\theta,i}$. Note that the augmented cells are defined as sets of indices rather than as sets of $\hat{\xi}_i$ values. The vector of $(n-p+1)p$ constructed "observations" in the augmented one-way ANOVA design is

$$\hat{\boldsymbol{\xi}}_{\mathbf{C}_\theta} = (\hat{\xi}_j, j \in W_{(p-1)/2+1}(\mathbf{C}_\theta), \ldots, \hat{\xi}_j, j \in W_{n-(p-1)/2}(\mathbf{C}_\theta))^T. \tag{4.62}$$

Let MST and MSE denote the balanced one-way ANOVA mean squares due to

treatment and error, respectively, computed on the data $\hat{\boldsymbol{\xi}}_{\mathbf{C}_\theta}$. The proposed test statistic is based on

$$MST - MSE. \tag{4.63}$$

Now we describe the construction of the first non-linearly supervised principal component $P_\theta = \mathbf{Z}^T\mathbf{C}_\theta$. Let $p_j, j = 1, \ldots, d_2$, denote the p-values obtained by applying the test of Zambom and Akritas (2012) for testing the hypothesis H_0^j which specifies that Z_j, the jth coordinate of $\mathbf{Z}$, has no effect on the regression function of the model with response variable Y and covariate vector $(\mathbf{X}, Z_j)$. For a threshold parameter θ define the index set $\mathcal{J} = \{j : p_j < \theta\}$ and let $\mathbf{Z}_\mathcal{J}$ be the vector formed from the $\mathcal{J}$ coordinates of $\mathbf{Z}$. Then, $P_\theta = \mathbf{Z}^T\mathbf{C}_\theta$ is the first principal component of $\mathbf{Z}_\mathcal{J}$. Note that some entries of $\mathbf{C}_\theta$ are equal to 0, corresponding to the coordinates of $\mathbf{Z}$ with p_j greater or equal to θ. It is important to keep in mind that the observable vector of first nonlinear principal components, $\mathbf{P}_\theta = (P_{\theta,1}, \ldots, P_{\theta,n})$, depends on the estimated residuals $\hat{\xi}_i, i = 1, \ldots, n$, to the extend that $\mathcal{J}$, and hence $\mathbf{C}_\theta$, depend on them.

Consider the following conditions

a) $E|Y|^\rho < \infty$ for some $\rho > 2$.

b) $f_\mathbf{Z}$ is continuous and bounded away from zero.

c) $f_\mathbf{X}$ is uniformly continuous, bounded and bounded away from 0.

d) the $q + 1$ derivatives of $m_1(\mathbf{x})$ exist and are Lipschitz uniformly continuous and bounded.

e) $\sigma^2(., \mathbf{z}) := E(\xi^2|\mathbf{Z}^T\mathbf{C})$ is Lipschitz continuous, $\sup_{\mathbf{x},\mathbf{z}} \sigma^2(\mathbf{x}, \mathbf{z}) < \infty$, and $E(\epsilon_i^4) < \infty$.

Also, assume that the eigenvalues, $\lambda_i,\ i = 1, \ldots, d_1$, of the bandwidth matrix $H_n^{1/2}$ defined in (4.6), converge to zero at the same rate and satisfy: for some $0 < \delta < 1$

$$1) \quad n\lambda_i^{4(q+1)} \to 0 \ i = 1, \ldots, d_1.$$

$$2) \quad \frac{n\lambda_i^{2r}}{(\log n)^2} \to \infty,\ i = 1, \ldots, d_1.$$

$$3) \quad \frac{n^{1-2/\rho}\lambda_i^{d_1}}{\ln n[\ln n(\ln \ln n)^{1+\delta}]^{2/\rho}} \to \infty,\ i = 1, \ldots, d_1.$$

Theorem 4.4. *Let conditions a)-e) and the conditions on the bandwidth 1)-3) hold. Then, under H_0 in (4.59), the asymptotic distribution of the test statistic in (4.63) is given by*

$$n^{1/2}(MST - MSE) \xrightarrow{d} N\left(0, \frac{2p(2p-1)}{3(p-1)}\tau^2\right),$$

where $\tau = \int \left[\int \sigma^2(\mathbf{x}, \mathbf{z}) f_{\mathbf{X}|\mathbf{Z}^T\mathbf{C}=\mathbf{z}^T\mathbf{C}}(\mathbf{x}) d\mathbf{x}\right]^2 f_{\mathbf{Z}^T\mathbf{C}}(\mathbf{z}^T\mathbf{C}) d(\mathbf{z}^T\mathbf{C})$.

An estimate of τ^2 can be obtained by modifying Rice's (1984) estimator as follows

$$\hat{\tau}^2 = \frac{1}{4(n-3)} \sum_{j=2}^{n-2} (\hat{\xi}_j - \hat{\xi}_{j-1})^2 (\hat{\xi}_{j+2} - \hat{\xi}_{j+1})^2. \tag{4.64}$$

The next subsection gives the asymptotic theory under local additive and under general local alternatives. As these limiting results show, the asymptotic mean of the test statistic $MST - MSE$ is positive under alternatives. Thus, the test procedure rejects the null hypothesis for "large" values of the test statistic.

Proof of Theorem 4.4. Under H_0 in (4.3) we write

$$\begin{aligned} \hat{\xi}_i &= Y_i - \hat{m}_1(\mathbf{X}_i) + m_1(\mathbf{X}_i) - m_1(\mathbf{X}_i) = \xi_i - (\hat{m}_1(\mathbf{X}_i) - m_1(\mathbf{X}_i)) \\ &= \xi_i - \Delta_{m_1}(\mathbf{X}_i), \end{aligned}$$

where $\Delta_{m_1}(\mathbf{X}_i)$ is defined implicitly in the above relation. Thus, $\hat{\boldsymbol{\xi}}_{\mathbf{C}_\theta}$ of relation (4.62) is decomposed as $\hat{\boldsymbol{\xi}}_{\mathbf{C}_\theta} = \boldsymbol{\xi}_{\mathbf{C}_\theta} - \boldsymbol{\Delta}_{m_1\mathbf{C}_\theta}$, where $\boldsymbol{\xi}_{\mathbf{C}_\theta}$ and $\boldsymbol{\Delta}_{m_1\mathbf{C}_\theta}$ are defined as in (4.62) but using ξ_i and $\Delta_{m_1}(\mathbf{X}_i)$, respectively, instead of $\hat{\xi}_i$. Note that MST-MSE given in (4.63) can be written as a quadratic form $\hat{\boldsymbol{\xi}}^T_{\mathbf{C}_\theta} A \hat{\boldsymbol{\xi}}_{\mathbf{C}_\theta}$ (see Wang, Akritas and Van Keilegom, 2008), where

$$A = \frac{np-1}{n(n-1)p(p-1)} \oplus_{i=1}^{n} \mathbf{J}_p - \frac{1}{n(n-1)p}\mathbf{J}_{np} - \frac{1}{n(p-1)}\mathbf{I}_{np}, \tag{4.65}$$

$\mathbf{I}_d$ is a identity matrix of dimension d, $\mathbf{J}_d$ is a dxd matrix of 1's and $\oplus$ is the Kronecker sum or direct sum. Thus, we can write $\sqrt{n}$(MST - MSE) as

$$\sqrt{n}\hat{\boldsymbol{\xi}}^T_{\mathbf{C}_\theta} A \hat{\boldsymbol{\xi}}_{\mathbf{C}_\theta} = \sqrt{n}\boldsymbol{\xi}^T_{\mathbf{C}_\theta} A \boldsymbol{\xi}_{\mathbf{C}_\theta} - \sqrt{n}2\boldsymbol{\xi}^T_{\mathbf{C}_\theta} A \boldsymbol{\Delta}_{m_1\mathbf{C}_\theta} + \sqrt{n}\boldsymbol{\Delta}^T_{m_1\mathbf{C}_\theta} A \boldsymbol{\Delta}_{m_1\mathbf{C}_\theta}. \tag{4.66}$$

That $\sqrt{n}2\boldsymbol{\xi}_{\mathbf{C}_\theta}^T A\boldsymbol{\Delta}_{m_1\mathbf{C}_\theta}$ and $\sqrt{n}\boldsymbol{\Delta}_{m_1\mathbf{C}_\theta}^T A\boldsymbol{\Delta}_{m_1\mathbf{C}_\theta}$ converge in probability to 0 uniformly follows from arguments similar to those used in Zambom and Akritas (2012).

Using Corolary 4.1, to show the asymptotic normality of $\sqrt{n}\boldsymbol{\xi}_{\mathbf{C}_\theta}^T A\boldsymbol{\xi}_{\mathbf{C}_\theta}$, it is enough to show that

$$\sup_{\mathbf{C}}\left|P\left(\frac{\sqrt{n}\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}}{\frac{2p(2p-1)}{3(p-1)}\tau^2}\leq t\right)-\Phi\left(t\right)\right|\to 0.$$

Let $b_n\sim n^{2/3}$ and $r_n\sim n/b_n\sim n^{1/3}$ and write

$$\begin{aligned}\sqrt{n}\boldsymbol{\xi}_{\mathbf{C}_\theta}^T A_d\boldsymbol{\xi}_{\mathbf{C}_\theta} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{p-1}\sum_{j_1\neq j_2}\xi_{j_1}\xi_{j_2}I\left(j_1,j_2\in W_i(\mathbf{C}_\theta)\right)\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{r_n}U_i(\mathbf{C}_\theta)+\frac{1}{\sqrt{n}}\sum_{i=1}^{r_n}V_i(\mathbf{C}_\theta)\\ &= \frac{1}{\sqrt{n}}S_U(C_\theta)+\frac{1}{\sqrt{n}}S_V(C_\theta),\end{aligned}\tag{4.67}$$

where, with $\gamma_i(\mathbf{C}_\theta)=\frac{1}{p-1}\sum_{j_1\neq j_2}\xi_{j_1}\xi_{j_2}I\left(j_1,j_2\in W_i(\mathbf{C}_\theta)\right)$,

$$\begin{aligned}U_i(\mathbf{C}_\theta) &= \gamma_{(i-1)(b_n+p)+1}(\mathbf{C}_\theta)+\ldots+\gamma_{(i-1)(b_n+p)+b_n}(\mathbf{C}_\theta),\\ V_i(\mathbf{C}_\theta) &= \gamma_{(i-1)(b_n+p)+b_n+1}(\mathbf{C}_\theta)+\ldots+\gamma_{i(b_n+p)}(\mathbf{C}_\theta).\end{aligned}$$

Note that the $U_i(\mathbf{C}_\theta)$ are independent, and the $V_i(\mathbf{C}_\theta)$ are independent.

Now, letting $\text{sd}=\sqrt{\frac{2p(2p-1)}{3(p-1)}\tau^2}$, we have

$$\begin{aligned}&\sup_{\mathbf{C}}\left|P\left(\frac{\sqrt{n}\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}}{\text{sd}}\leq t\right)-\Phi\left(t\right)\right|\\ =\ &\sup_{\mathbf{C}}\left|P\left(\frac{S_U(\mathbf{C})+S_V(\mathbf{C})}{\sqrt{n}\text{sd}}\leq t\right)-\Phi\left(t\right)\right|\\ =\ &\sup_{\mathbf{C}}\left|P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\text{sd}}\leq t-\frac{S_V(\mathbf{C})}{\sqrt{n}\text{sd}},\left|\frac{S_V(\mathbf{C})}{\sqrt{n}\text{sd}}\right|\leq\epsilon\right)\right.\\ &\qquad\left.+P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\text{sd}}\leq t-\frac{S_V(\mathbf{C})}{\sqrt{n}\text{sd}},\left|\frac{S_V(\mathbf{C})}{\sqrt{n}\text{sd}}\right|\geq\epsilon\right)-\Phi\left(t\right)\right|\end{aligned}$$

$$\leq \sup_{\mathbf{C}} \left| P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t - \frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}, \left| \frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \right| \leq \epsilon \right) - \Phi(t) \right| + \sup_{\mathbf{C}} P\left(\left| \frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \right| \geq \epsilon \right) \tag{4.68}$$

That the second in (4.68) term converges to zero follows from Lemma 4.8. That the first term in (4.68) converges to zero follows from Lemma 4.9, provided we show that

$$\mathrm{Var}\left(\frac{S_U(\mathbf{C})}{\sqrt{n}} \right) \to \mathrm{sd}^2, \quad \text{for any} \quad \mathbf{C}. \tag{4.69}$$

By (4.67), and because $\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \xrightarrow{p} 0$, (4.69) follows from $\sup_{\mathbf{C}} \mathrm{Var}(\sqrt{n}\boldsymbol{\xi}_{vC}^T A_d \boldsymbol{\xi}_{vC}) \to \mathrm{sd}^2$. By the definition of $\boldsymbol{\xi}_{\mathbf{C}_\theta}^T A_d \boldsymbol{\xi}_{\mathbf{C}_\theta}$, it is easy to see that $E\left(\boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}\right) = 0$ for any $\mathbf{C}$. To find the variance of $\sqrt{n}\boldsymbol{\xi}_{vC}^T A_d \boldsymbol{\xi}_{vC}$ we first evaluate the conditional second moment $E[(\sqrt{n}\boldsymbol{\xi}_{vC}^T A_d \boldsymbol{\xi}_{vC})^2 | \mathbf{Z}^T\mathbf{C}]$. Recalling the notation $\sigma^2(., \mathbf{z}_j^T\mathbf{C}) = E(\xi_j^2 | \mathbf{Z}^T\mathbf{C} = \mathbf{z}_j^T\mathbf{C})$, we have

$$\begin{aligned}
&\sup_{\mathbf{C}} \frac{1}{n(p-1)^2} \sum_{i_1,i_2}^{n} \sum_{j_1 \neq l_1}^{n} \sum_{j_2 \neq l_2}^{n} E(\xi_{j_1}\xi_{l_1}\xi_{j_2}\xi_{l_2} | \mathbf{Z}^T\mathbf{C}) \\
&\qquad\qquad I(j_s \in W_{i_s}(\mathbf{C}), l_s \in W_{i_s}(\mathbf{C}), s = 1, 2) \\
&= \sup_{\mathbf{C}} \frac{2}{n(p-1)^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{j \neq l}^{n} \sigma^2(., \mathbf{z}_j^T\mathbf{C}) \sigma^2(., \mathbf{z}_l^T\mathbf{C}) I(j, l \in W_{i_1}(\mathbf{C}) \cap W_{i_2}(\mathbf{C})) \\
&= \sup_{\mathbf{C}} \frac{2}{n(p-1)^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{j \neq l}^{n} \sigma^2(., \mathbf{z}_j^T\mathbf{C}) \left(\sigma^2(., \mathbf{z}_j^T\mathbf{C}) + O_p\left(\frac{p}{\sqrt{n}} \right) \right) \\
&\qquad\qquad I(j, l \in W_{i_1}(\mathbf{C}) \cap W_{i_2}(\mathbf{C})) \\
&= \sup_{\mathbf{C}} \frac{2}{n(p-1)^2} \sum_{j=1}^{n} \sigma^4(., \mathbf{z}_j^T\mathbf{C}) \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{l \neq j}^{n} I(j, l \in W_{i_1}(\mathbf{C}) \cap W_{i_2}(\mathbf{C})) \\
&\qquad\qquad + O_p\left(\frac{p^2}{n^{1/2}} \right) \\
&= \sup_{\mathbf{C}} \frac{2}{n(p-1)^2} \sum_{j=1}^{n} \sigma^4(., \mathbf{z}_j^T\mathbf{C}) 2(1 + 2^2 + 3^2 + ... + (p-1)^2) + O_p\left(\frac{p^2}{n^{1/2}} \right) \\
&= \sup_{\mathbf{C}} \frac{2}{n(p-1)^2} \frac{p(p-1)(2p-1)}{3} \sum_{j=1}^{n} \sigma^4(., \mathbf{z}_j^T\mathbf{C}) + O_p\left(\frac{p^2}{n^{1/2}} \right),
\end{aligned}$$

where the third equality follows from Lemma 4.16 using the assumption that $\sigma^2(., \mathbf{z}_j^T\mathbf{C})$ is Lipschitz continuous and the second last inequality results from the fact that if $1 \leq |j_1 - j_2| = s \leq p-1$, then they are $(p-s)^2$ pairs of windows whose intersection includes j_1 and j_2. Taking limits as $n \to \infty$ it is seen that

$$\sup_{\mathbf{C}} E\left(n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}} \middle| \mathbf{Z}^T\mathbf{C}\right)^2 \xrightarrow{a.s.} \frac{2(2p-1)}{3(p-1)} E(\sigma^4(., \mathbf{z}^T\mathbf{C})) = \frac{2(2p-1)}{3(p-1)}\tau^2 \quad (4.70)$$

From relation (4.70) it is easily seen that $\sup_{\mathbf{C}} E[(\sqrt{n}\boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}})^2 | \mathbf{Z}^T\mathbf{C}]$ remains bounded, and thus $\sup_{\mathbf{C}} \mathrm{Var}(n^{1/2}\xi_{\mathbf{C}}^T A \xi_{\mathbf{C}})$ also converges to the same limit by the Dominated Convergence Theorem. □

Lemma 4.7. *If the assumptions of Theorem 4.1 hold, then under H_0 and as $n \to \infty$,*

$$\sup_{\mathbf{C}} P\left(n^{1/2}\left|\boldsymbol{\xi}_{\mathbf{C}_\theta}^T A \boldsymbol{\xi}_{\mathbf{C}_\theta} - \boldsymbol{\xi}_{\mathbf{C}_\theta}^T A_d \boldsymbol{\xi}_{\mathbf{C}_\theta}\right| \geq \epsilon\right) \to 0, \quad (4.71)$$

where $A_d = diag\{B_1, ..., B_n\}$, with $B_i = \frac{1}{n(p-1)}[\mathbf{J}_p - \mathbf{I}_p]$.

Proof. By Chebyshev Inequality, we have that

$$\sup_{\mathbf{C}} P\left(n^{1/2}\left|\boldsymbol{\xi}_{\mathbf{C}}^T A \boldsymbol{\xi}_{\mathbf{C}} - \boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}\right| \geq \epsilon\right) \leq \sup_{\mathbf{C}} \frac{nE\left[\left(\boldsymbol{\xi}_{\mathbf{C}}^T A \boldsymbol{\xi}_{\mathbf{C}} - \boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}\right)^2\right]}{\epsilon^2} \quad (4.72)$$

Since the block diagonal elements of A_d equal those of A, it sufices to show that the off diagonal blocks of A are negligible. For $i_1 \neq i_2$, every element of the block (i_1, i_2) equals $\frac{1}{n(n-1)p}$. We will show that the second moment on the right hand side of (4.72) conditionally on $\mathbf{Z}$ goes to zero, and therefore the unconditional second moment also does. To that end, write

$$\begin{aligned} &\sup_{\mathbf{C}} \frac{nE\left[\left(\boldsymbol{\xi}_{\mathbf{C}}^T A \boldsymbol{\xi}_{\mathbf{C}} - \boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}\right)^2 | \mathbf{Z}\right]}{\epsilon^2} \\ &= n\left(\frac{1}{n(n-1)p}\right)^2 \sup_{\mathbf{C}} E\Bigg(\sum_{i_1 \neq i_2}\sum_{i_3 \neq i_4}\sum_{j_1,j_2,j_3,j_4=1}^{n} \xi_{j_1}\xi_{j_2}\xi_{j_3}\xi_{j_4} \\ &\qquad I(j_k \in W_{i_k}(\mathbf{C}), k = 1, \ldots, 4)|\mathbf{Z}\Bigg) \end{aligned} \quad (4.73)$$

$$= n\left(\frac{1}{n(n-1)p}\right)^2 \sup_{\mathbf{C}} \sum_{i_1\neq i_2}\sum_{i_3\neq i_4}\sum_{j_1,j_2,j_3,j_4=1}^{n} E\left(\xi_{j_1}\xi_{j_2}\xi_{j_3}\xi_{j_4}|\mathbf{Z}\right) I(j_k \in W_{i_k}(\mathbf{C}), k=1,\ldots,4)$$

The expected value in this sum is different from zero, only if $\xi_{j_1},\ldots,\xi_{j_4}$ consists of two pairs of equal observations, or $j_1=j_2=j_3=j_4$. Since there are $O(n^2p^4)$ terms for the former case to happen and $O(np^4)$ for the latter case to happen, and the magnitude of these terms is not affected by $\mathbf{C}$, the order of (4.73) is $O\left(\frac{n}{p}\frac{1}{n^4p^2}n^2p^4\right) = o(1)$, and this completes the proof. □

Corollary 4.1. *Let* $A_d = diag\{B_1,...,B_n\}$, *with* $B_i = \frac{1}{n(p-1)}[\mathbf{J}_p - \mathbf{I}_p]$, $sd = \sqrt{\frac{2p(2p-1)}{3(p-1)}\tau^2}$, *and* $\boldsymbol{\xi}_{\mathbf{C}}$ *be defined in (4.62) with* $\mathbf{C}$ *instead of* $\mathbf{C}_\theta$. *Then, under the assumptions of Theorem 4.1 we have*

$$\sup_{\mathbf{C}}\sup_{t}\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t\right)-\Phi(t)\right|\to 0 \text{ if and only if}$$
$$\sup_{\mathbf{C}}\sup_{t}\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t\right)-\Phi(t)\right|\to 0.$$

Proof. Write

$$\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}}}{sd} = \frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}}{sd} + \frac{n^{1/2}\left(\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}} - \boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}\right)}{sd}.$$

Now, for any t

$$\begin{aligned}
&\sup_{\mathbf{C}}\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t\right)-\Phi(t)\right| \\
&= \sup_{\mathbf{C}}\left|P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t-\frac{n^{1/2}\left(\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}} - \boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}\right)}{sd}, \right.\right. \\
&\qquad\qquad \left.\left.\left|\frac{n^{1/2}\left(\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}} - \boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}\right)}{sd}\right|\le\epsilon\right)\right. \\
&+P\left(\frac{n^{1/2}\boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}}{sd}\le t-\frac{n^{1/2}\left(\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}} - \boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}\right)}{sd}, \right. \\
&\qquad\qquad \left.\left.\left|\frac{n^{1/2}\left(\boldsymbol{\xi}_{\mathbf{C}}^T A\boldsymbol{\xi}_{\mathbf{C}} - \boldsymbol{\xi}_{\mathbf{C}}^T A_d\boldsymbol{\xi}_{\mathbf{C}}\right)}{sd}\right|\ge\epsilon\right)-\Phi(t)\right|
\end{aligned}$$

$$\leq \sup_{\mathbf{C}} \max \left\{ \left| P\left(\frac{n^{1/2} \boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}}{sd} \leq t + \epsilon \right) - \Phi(t) \right|, \right.$$
$$\left. \left| P\left(\frac{n^{1/2} \boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}}{sd} \leq t - \epsilon \right) - \Phi(t) \right| \right\}$$
$$+ \sup_{\mathbf{C}} P\left(\left| \frac{n^{1/2} \left(\boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}} - \boldsymbol{\xi}_{\mathbf{C}}^T A \boldsymbol{\xi}_{\mathbf{C}} \right)}{sd} \right| \geq \epsilon \right). \tag{4.74}$$

The last term in (4.74) goes to zero by Lemma 4.17. Thus,

$$\sup_{\mathbf{C}} \sup_t \left| P\left(\frac{n^{1/2} \boldsymbol{\xi}_{\mathbf{C}}^T A \boldsymbol{\xi}_{\mathbf{C}}}{sd} \leq t \right) - \Phi(t) \right|$$
$$\leq \sup_{\mathbf{C}} \max \left\{ \sup_t \left| P\left(\frac{n^{1/2} \boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}}{sd} \leq t \right) - \Phi(t - \epsilon) \right|, \right.$$
$$\left. \sup_t \left| P\left(\frac{n^{1/2} \boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}}{sd} \leq t \right) - \Phi(t + \epsilon) \right| \right\} + o(1)$$
$$\leq \sup_{\mathbf{C}} \sup_t \left| P\left(\frac{n^{1/2} \boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}}{sd} \leq t \right) - \Phi(t) \right| + \sup_t |\Phi(t) - \Phi(t + \epsilon)| + o(1).$$

Letting $\epsilon \to 0$,

$$\lim_{n \to \infty} \sup_{\mathbf{C}} \sup_t \left| P\left(\frac{n^{1/2} \boldsymbol{\xi}_{\mathbf{C}}^T A \boldsymbol{\xi}_{\mathbf{C}}}{sd} \leq t \right) - \Phi(t) \right|$$
$$\leq \lim_{n \to \infty} \sup_{\mathbf{C}} \sup_t \left| P\left(\frac{n^{1/2} \boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}}{sd} \leq t \right) - \Phi(t) \right|.$$

Using similar steps, it can be shown that

$$\lim_{n \to \infty} \sup_{\mathbf{C}} \sup_t \left| P\left(\frac{n^{1/2} \boldsymbol{\xi}_{\mathbf{C}}^T A_d \boldsymbol{\xi}_{\mathbf{C}}}{sd} \leq t \right) - \Phi(t) \right|$$
$$\leq \lim_{n \to \infty} \sup_{\mathbf{C}} \sup_t \left| P\left(\frac{n^{1/2} \boldsymbol{\xi}_{\mathbf{C}}^T A \boldsymbol{\xi}_{\mathbf{C}}}{sd} \leq t \right) - \Phi(t) \right|,$$

completing the proof. □

Lemma 4.8. *Let $S_V(\mathbf{C})$ be defined as in (4.67). Under the assumptions of Theo-*

rem 4.1,

$$\sup_{\mathbf{C}} P\left(\left|\frac{S_V(\mathbf{C})}{\sqrt{n}\ sd}\right| \geq \epsilon\right) \to 0$$

Proof. For any $\epsilon > 0$, since $V_i(\mathbf{C})$ are independent,

$$\begin{aligned}
&\sup_{\mathbf{C}} P\left(n^{-1/2}|\sum_{i=1}^{r_n} V_i(\mathbf{C})| \geq \epsilon\right) \leq \sup_{\mathbf{C}} \sum_{i=1}^{r_n} P\left(|V_i(\mathbf{C})| \geq \epsilon n^{1/2} r_n^{-1}\right) \\
&\leq \sup_{\mathbf{C}} \sum_{i=1}^{r_n} \frac{E(V_i(\mathbf{C})^4)}{\epsilon^4 n^2 r_n^{-4}} \leq K\epsilon^{-4} n^{-2} r_n^5 (p^2)^2 = o(1),
\end{aligned}$$

where the last inequality follows from the fact that $E(V_i^4(\mathbf{C})) \leq K(p^2)^2$. □

Lemma 4.9. *Let $S_U(\mathbf{C})$ and $S_V(\mathbf{C})$ be defined as in (4.67). Under the assumptions of Theorem 4.1,*

$$\sup_{\mathbf{C}} \left| P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}sd} \leq t - \frac{S_V(\mathbf{C})}{\sqrt{n}sd}, \left|\frac{S_V(\mathbf{C})}{\sqrt{n}sd}\right| \leq \epsilon\right) - \Phi(t)\right| \to 0. \tag{4.75}$$

Proof. Note that, using the Berry Essen bound (see Shorack (Probability for Statisticians)), and the fact that $Var(\frac{S_U(\mathbf{C})}{\sqrt{n}}) \to \mathrm{sd}^2$ as shown in the proof of Theorem 4.1, we have

$$\begin{aligned}
&\sup_{\mathbf{C}} \sup_t \left| P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t\right) - \Phi(t)\right| \leq 9 \sup_{\mathbf{C}} \frac{\sum_{i=1}^{r_n} E\left|U_i(\mathbf{C})\right|^3}{\left[\sum_{i=1}^{r_n} \mathrm{Var}(U_i(\mathbf{C}))\right]^{3/2}} \\
&= O\left(\frac{1}{\sqrt{r_n}}\right) = o(1).
\end{aligned} \tag{4.76}$$

Let $t^* = t - \frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}$, then

$$\begin{aligned}
&\sup_{\mathbf{C}} \left| P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t - \frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}, \left|\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\right| \leq \epsilon\right) - \Phi(t)\right| \\
= &\sup_{\mathbf{C}} \left| P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t^*, \left|\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\right| \leq \epsilon\right) - P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t^*\right)\right. \\
&\qquad \left. + P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t^*\right) - \Phi(t^*) + \Phi(t^*) - \Phi(t)\right|
\end{aligned}$$

$$\leq \sup_{\mathbf{C}} \left| P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t^*, \left|\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\right| \leq \epsilon\right) - P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t^*\right) \right| + \sup_{\mathbf{C}} \left| P\left(\frac{S_U(\mathbf{C})}{\sqrt{n}\mathrm{sd}} \leq t^*\right) - \Phi(t^*) \right| + \left| \Phi(t^*) - \Phi(t) \right|. \quad (4.77)$$

The first term in (4.77) goes to zero by continuity of measures, since by Lemma 4.8 $P\left(\left|\frac{S_V(\mathbf{C})}{\sqrt{n}\mathrm{sd}}\right| \leq \epsilon\right) \to 1$. The second term in (4.77) goes to zero by (4.76), and the third term goes to zero by the continuity of $\Phi(.)$.

□

4.4.1 Simulations: Hypothesis Testing for multivariate X and Z

We compare the proposed ANOVA-type hypothesis test for groups with the generalized likelihood ratio test of Fan and Jiang (2005). The data is generated under three situations: a homoscedastic additive model, a homoscedastic non-additive model, and a heterocedastic non-additive model. All covariates, in all models, are independent standard normal. The homoscedastic additive model is

$$Y = X_1 + \theta(X_2 + X_3 + X_4) + \epsilon, \quad \text{where} \quad \epsilon \sim N(0, 1), \quad (4.78)$$

the homoscedastic non-additive model is

$$Y = X_1^{X_2}(1 + \theta(X_3 + X_4)) + X_2^{\theta(X_3+X_4)} + \epsilon, \quad \text{where} \quad \epsilon \sim N(0, .1^2), \quad (4.79)$$

and the heterocedastic non-additive model is

$$Y = X_1 + \theta \sin(X_2 X_3) + X_2 X_3 \epsilon, \quad \text{where} \quad \epsilon \sim N(0, .5^2). \quad (4.80)$$

In each situation we simulate 2000 data sets ot size $n = 200$. All simulations were performed in R.

In order to evaluate the effect of the threshold parameter θ we applied our test procedure with $\theta = 0.05$ and $\theta = 0.2$. Moreover, in each case we considered two rules to form the set of covariates from which the first supervised principal

component is obtained. Rule 1 consists of using only the covariates with p-value less than θ, and in Rule 2 we consider the set of covariates chosen from Rule 1 and add to the set the covariate with the smallest p-value among the remainder covariates. In each case, if the number of selected covariates is less than two the set is formed from the two with the smallest p-value. Thus the simulations consider four versions of our test statistic: a) Rule 1 with $\theta = 0.05$, b) Rule 1 with $\theta = 0.2$, c) Rule 2 with $\theta = 0.05$, d) Rule 2 with $\theta = 0.2$. All four versions of our test statistic use windows of $p = 11$.

Tables 4.14, 4.15, and 4.16, show the simulation results for models (4.78),(4.79), and (4.80), respectively. It is seen that the proposed test procedure is robust to the choice of the threshold parameter, and to the rules for selecting the set of covariates from which the first supervised principal component is obtained. The Generalized Likelihood Ratio test, which is designed for homoscedastic additive models, achieves better power under model (4.78), but is extremely liberal under heteroscedasticity and its power for the non-additive alternatives of model (4.80) is mainly less than its level; see Table 4.16. Table 4.15 suggests that the GRLT has low power against non-additive alternatives even in the homoscedastic case.

Table 4.14. Rejection rates for the homocedastic additive model

	θ				
Method	0	.2	.4	.6	.8
ANOVA-type-a	.066	.404	.691	.706	.751
ANOVA-type-b	.060	.378	.613	.689	.733
ANOVA-type-c	.066	.396	.600	.692	.749
ANOVA-type-d	.057	.375	.618	.685	.718
GRLT	.048	.883	1	1	1

4.5 Power of the Test under Local Alternatives

For hypothesis test statistics whose power is impossible or difficult to derive analytically, we can consider the probability the test rejects the null hypothesis when the alternative approaches to the null at a certain rate. We investigate the asymptotic power of the test statistic for local additive and local general alternatives.

Table 4.15. Rejection rates for the homocedastic non-additive model

	θ				
Method	0	.02	.04	.06	.08
ANOVA-type-a	.051	.202	.522	.693	.724
ANOVA-type-b	.047	.192	.560	.710	.739
ANOVA-type-c	.050	.193	.520	.679	.711
ANOVA-type-d	.047	.161	.510	.676	.733
GRLT	.052	.059	.117	.235	.379

Table 4.16. Rejection rates for the heterocedastic non-additive model

	θ				
Method	0	.3	.6	1	2
ANOVA-type-a	.035	.168	.503	.654	.789
ANOVA-type-b	.040	.172	.529	.663	.767
ANOVA-type-c	.037	.161	.501	.651	.757
ANOVA-type-d	.036	.190	.520	.657	.742
GRLT	.584	.585	.535	.439	.297

4.5.1 Power of the test under additive alternatives

Consider the case of additive alternatives for the regression model with covariates $\mathbf{U} = (\mathbf{X}, Z)$. Local alternatives can be stated in the following way

$$\begin{aligned} H_0 : m(\mathbf{x}, z) &= m_1(\mathbf{x}) \\ H_1^A : m(\mathbf{x}, z) &= m_1(\mathbf{x}) + \rho_n \tilde{m}_2(z), \end{aligned} \tag{4.81}$$

for a sequence of constants ρ_n converging to 0. We assume that $E(\tilde{m}_2(z)) = 0$ and that $\tilde{m}_2(z)$ is Lipschitz continuous. It is important to note that if ρ_n goes to 0 too fast, the null hypothesis will be confounded with the alternative, and the test will not be able to detect the difference, having an impact on the power. But for ρ_n going to 0 too slow, the null hypothesis will always be rejected for a large enough n. For the alternative hypothesis considered in this section, we set ρ_n to be $a(n)^{-1/4}$, for a constant a.

Note that under H_1^A we have

$$Y_i = m_1(\mathbf{X}_i) + \rho_n \tilde{m}_2(Z_i) + \xi_i.$$

By estimating $m_1(\mathbf{x})$ by the local polynomial estimator (4.47)

$$\hat{m}_1(\mathbf{x}) = \mathbf{e}_1^T \left(\mathbb{X}_\mathbf{x}^T \mathbb{W}_\mathbf{x} \mathbb{X}_\mathbf{x}\right)^{-1} \mathbb{X}_\mathbf{x}^T \mathbb{W}_\mathbf{x} \mathbf{Y} = \sum_{j=1}^{n} \tilde{w}(\mathbf{x}, \mathbf{X}_j) Y_j, \ i = 1...n,$$

we construct the vector

$$\hat{\boldsymbol{\xi}}_V = (\hat{\xi}_j, j \in W_{(p-1)/2+1}, \ldots, \hat{\xi}_j, j \in W_{n-(p-1)/2})^T. \tag{4.82}$$

where the windows W_i are defined according to Z as described in (4.91). The test statistic is the same proposed in (4.49), but we will study its properties under the local additive alternative (4.81).

Note that, under the alternative, we can write $\hat{\xi}_j$ as

$$\begin{aligned}
\hat{\xi}_j &= Y_j - \hat{m}_1(\mathbf{X}_{1j}) \\
&= Y_j - m_1(\mathbf{X}_j) - \rho_n \tilde{m}_2(Z_j) - [\hat{m}_1(\mathbf{X}_j) - m_1(\mathbf{X}_j)] + \rho_n \tilde{m}_2(Z_j) \\
&= \xi_j - \Delta_{m_1}(\mathbf{X}_j) + \rho_n \tilde{m}_2(Z_j),
\end{aligned}$$

and therefore

$$\begin{aligned}
\hat{\boldsymbol{\xi}}_V &= (\xi_j - \Delta_{m_1}(\mathbf{X}_j) + \rho_n \tilde{m}_2(Z_j), j \in W_1, ..., \xi_j - \Delta_{m_1}(\mathbf{X}_j) + \rho_n \tilde{m}_2(Z_j), j \in W_n)^T \\
&= \boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V} + \rho_n \tilde{\mathbf{m}}_{2V}.
\end{aligned}$$

Theorem 4.5. *Consider the notation and assumptions of Theorem 4.3. Moreover, assume that $\tilde{m}_2(x)$ is Lipschitz continuous. Then, under H_1^A in (4.81), as $n \to \infty$,*

$$n^{1/2}(MST - MSE) \xrightarrow{d} N\left(a^2 p \, Var(\tilde{m}_2(Z)), \frac{2p(2p-1)}{3(p-1)} \tau^2\right).$$

Proof. Note that we can write the test statistic as

$$\begin{aligned}
\sqrt{n}(MST - MSE) &= \sqrt{n} \hat{\boldsymbol{\xi}}_V^T A \hat{\boldsymbol{\xi}}_V = \sqrt{n} (\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V})^T A (\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V}) \\
&\quad + \sqrt{n} 2 \rho_n (\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V})^T A \tilde{\mathbf{m}}_{2V} \\
&\quad + \sqrt{n} \rho_n^2 \tilde{\mathbf{m}}_{2V}^T A \tilde{\mathbf{m}}_{2V}.
\end{aligned} \tag{4.83}$$

We know by Theorem 4.3 that

$$\sqrt{n}(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1V})' A(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1V}) \xrightarrow{d} N\left(0, \frac{2p(2p-1)}{3(p-1)}\tau^2\right).$$

It is left to show that

$$\sqrt{n}2\rho_n(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1V})^T A\tilde{\mathbf{m}}_{2V} \xrightarrow{p} 0$$

and

$$\sqrt{n}\rho_n^2\tilde{\mathbf{m}}_{2V}^T A\tilde{\mathbf{m}}_{2V} \xrightarrow{p} a^2 pV(\tilde{m}_2(Z)).$$

The convergence to zero in probability of these two terms is shown in Lemma 4.10 and Lemma 4.11, respectively. □

Lemma 4.10. *The second term in (4.83) converges in probability to zero, i.e.*

$$\sqrt{n}2\rho_n(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1V})^T A\tilde{\mathbf{m}}_{2V} \xrightarrow{p} 0.$$

Proof. By the definition of the matrix A, we can write

$$\begin{aligned}
& (\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1V})' A\tilde{\mathbf{m}}_{2V} \\
= \; & \frac{np-1}{n(n-1)p(p-1)} \sum_{i=1}^{n} \left[\sum_{j=1}^{n} \tilde{m}_2(Z_j) I(j \in W_i)\right] \left[\sum_{k=1}^{n} (\xi_k - \Delta_{m_1}(\mathbf{X}_k)) I(k \in W_i)\right] \\
& - \frac{1}{n(n-1)p} \left[p \sum_{i=1}^{n} \tilde{m}_2(Z_i)\right] \left[p \sum_{i=1}^{n} (\xi_i - \Delta_{m_1}(\mathbf{X}_i))\right] \\
& - \frac{p}{n(p-1)} \sum_{i=1}^{n} \tilde{m}_2(Z_i)(\xi_i - \Delta_{m_1}(\mathbf{X}_i)).
\end{aligned}$$

Using Lemma 4.16 and the fact that $\tilde{m}_2(\cdot)$ is Lipschitz continuous, the sum in the first term can be expressed as

$$p \sum_{i=1}^{n} [\tilde{m}_2(Z_i) + O(n^{-1/2})] \left[\sum_{k=1}^{n} (\xi_k - \Delta_{m_1}(\mathbf{X}_k)) I(k \in W_i)\right]$$

$$
\begin{aligned}
&\leq p\sum_{k=1}^{n}\left[\sum_{i=1}^{n}\tilde{m}_2(Z_i)I(i\in W_k)\right](\xi_k-\Delta_{m_1}(\mathbf{X}_k))\\
&\quad +p^2O(n^{-1/2})\sum_{k=1}^{n}|(\xi_k-\Delta_{m_1}(\mathbf{X}_k))|\\
&= p^2\sum_{k=1}^{n}\tilde{m}_2(Z_k)(\xi_k-\Delta_{m_1}(\mathbf{X}_k))+O_p(p^2n^{1/2}),
\end{aligned}
$$

so that

$$
\begin{aligned}
\sqrt{n}\rho_n\tilde{\mathbf{m}}_{2V}^TA(\boldsymbol{\xi}_V-\boldsymbol{\Delta}_{m_1V}) &= \frac{an^{1.25}p}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{m}_2(Z_i)(\xi_i-\Delta_{m_1}(\mathbf{X}_i))\right]\\
&\quad -\frac{an^{1.25}p}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{m}_2(Z_i)\right]\left[\frac{1}{n}\sum_{i=1}^{n}(\xi_i-\Delta_{m_1}(\mathbf{X}_i))\right]\\
&\quad +O_p\left(\frac{1}{n^{1/4}}\right).
\end{aligned}
$$

Using the fact that $E\left(\tilde{m}_2(Z_i)\right)=E\left(\tilde{m}_2(Z_i)\xi_i\right)=E(\xi_i)=0$, relation (4.36) and also that $n^{-3/4}\sum_{i=1}^{n}\tilde{m}_2(Z_i)\Delta_{m_1}(\mathbf{X}_i)\xrightarrow{p}0$, as is shown in a similar way to (4.36), completes the proof of the lemma. □

Lemma 4.11. *The third term in (4.83) converges in probability to $a^2pV(\tilde{m}_2(Z))$, i.e.*

$$
\sqrt{n}\rho_n^2\tilde{\mathbf{m}}_{2V}^TA\tilde{\mathbf{m}}_{2V}\xrightarrow{p}a^2pV(\tilde{m}_2(Z)).
$$

Proof. Writing

$$
\begin{aligned}
\tilde{\mathbf{m}}_{2V}^TA\tilde{\mathbf{m}}_{2V} &= \frac{np}{n-1}\left\{\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{m}_2^2(Z_i)\right]-\left[\frac{1}{n}\sum_{i=1}^{n}\tilde{m}_2(Z_i)\right]^2\right\}+O\left(\frac{1}{n^{1/2}}\right)\\
&= p\left\{E\tilde{m}_2^2(Z)-[E\tilde{m}_2(Z)]^2\right\}+O_p\left(\frac{1}{n^{1/2}}\right),
\end{aligned}
$$

it follows that

$$
\sqrt{n}\rho_n^2\tilde{\mathbf{m}}_{2V}^TA\tilde{\mathbf{m}}_{2V} = \sqrt{n}\frac{np\rho_n^2}{(n-1)}Var(\tilde{m}_2(Z))+O_p\left(\sqrt{n}\frac{\rho_n^2}{n^{1/2}}\right)
$$

$$\begin{aligned} &= \frac{n}{n-1}a^2 p\mathrm{Var}(\tilde{m}_2(Z)) + o_p(1) \\ &\to^P a^2 p\mathrm{Var}(\tilde{m}_2(Z)), \end{aligned}$$

which completes the proof. □

4.5.2 Power of the test under general alternatives

Now, let us consider more general alternatives. In order to study the behavior of the test we use a similar factorization of the regression function described in Section 4.1.1, but extended to a vector in the first entry, i.e.,

$$m(\mathbf{x}, z) = \mu + \tilde{m}_1(\mathbf{x}) + \tilde{m}_2(z) + \tilde{m}_{12}(\mathbf{x}, z). \tag{4.84}$$

Thus, the null and alternative hypothesis are

$$\begin{aligned} H_0 : m(\mathbf{x}, z) &= m_1(\mathbf{x}) \\ H_1^G : m(\mathbf{x}, z) &= m_1(\mathbf{x}) + \rho_{2n}\tilde{m}_2(z) + \rho_{12n}\tilde{m}_{12}(\mathbf{x}, z), \end{aligned} \tag{4.85}$$

for the sequences of constants ρ_{2n} and ρ_{12n} converging to 0. Besides the usual assumptions based on the factorization model, we need to assume $\tilde{m}_2(z)$ Lipschitz continuous and $\tilde{m}_{12}(\mathbf{x}, z)$ Lipschitz continuous on z uniformly on $\mathbf{x}$. Again, if ρ_{2n} or ρ_{12n} go to 0 too fast or too slow, we will have the same problems described in the previous section. For this alternative hypothesis, we set $\rho_{1n} = an^{-1/4}$ and $\rho_{2n} = bn^{-1/4}$, for some constants a and b.

Note that under H_1^G we have

$$Y_i = m_1(\mathbf{X}_i) + \rho_{1n}\tilde{m}_2(Z_i) + \rho_{2n}\tilde{m}_{12}(\mathbf{X}_i, Z_i) + \xi_i,$$

and hence

$$\begin{aligned} \hat{\xi}_j &= Y_j - \hat{m}_1(\mathbf{X}_j) \\ &= Y_j - m_1(\mathbf{X}_j) - \rho_{1n}\tilde{m}_2(Z_j) - \rho_{2n}\tilde{m}_{12}(\mathbf{X}_j, Z_i) - [\hat{m}_1(\mathbf{X}_j) - m_1(\mathbf{X}_j)] \\ &\quad + \rho_{1n}\tilde{m}_2(Z_j) + \rho_{2n}\tilde{m}_{12}(\mathbf{X}_j, Z_i) \\ &= \xi_j - \Delta_{m_1}(\mathbf{X}_j) + \rho_{1n}\tilde{m}_2(Z_j) + \rho_{2n}\tilde{m}_{12}(\mathbf{X}_j, Z_i). \end{aligned}$$

for the estimated regression function m_1 using local polynomial regression.

Note that under this general local alternative

$$\begin{aligned}\hat{\boldsymbol{\xi}}_V &= \begin{pmatrix} \xi_j - \Delta_{m_1}(\mathbf{X}_j) + \rho_{1n}\tilde{m}_2(Z_j) + \rho_{2n}\tilde{m}_{12}(\mathbf{X}_j, Z_i), j \in W_1 \\ \vdots \\ \xi_j - \Delta_{m_1}(\mathbf{X}_j) + \rho_{1n}\tilde{m}_2(Z_j) + \rho_{2n}\tilde{m}_{12}(\mathbf{X}_j, Z_i), j \in W_n \end{pmatrix} \\ &= \boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V} + \rho_{1n}\tilde{\mathbf{m}}_{2V} + \rho_{2n}\tilde{\mathbf{m}}_{12V},\end{aligned}$$

and we have the following theorem.

Theorem 4.6. *Consider the notation and assumptions of Theorem 4.3. Moreover, assume that $\tilde{m}_2(z)$ is Lipschitz continuous and that $\tilde{m}_{12}(\mathbf{x}, z)$ is Lipschitz continuous on z uniformly on $\mathbf{x}$. Then, under H_1^G in (4.85), as $n \to \infty$,*

$$n^{1/2}(MST - MSE) \xrightarrow{d} N\left(\mu^G, \frac{2p(2p-1)}{3(p-1)}\tau^2\right),$$

where

$$\mu^G = pa^2\, Var(\tilde{m}_2(Z)) + pb^2 Var(\tilde{m}_{12}(\mathbf{X}, Z)) + 2pabCov(\tilde{m}_2(Z), \tilde{m}_{12}(\mathbf{X}, Z)).$$

If $a = b$ the formula simplifies to

$$\mu^G = pa^2\, Var(\tilde{m}_2(Z) + \tilde{m}_{12}(\mathbf{X}, Z)).$$

Proof. Note that the test statistic $\sqrt{n}(MST - MSE)$ can be written as

$$\begin{aligned}\sqrt{n}\hat{\boldsymbol{\xi}}_V^T A \hat{\boldsymbol{\xi}}_V &= \sqrt{n}(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V} - \rho_{1n}\tilde{\mathbf{m}}_{2V})^T A(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V} - \rho_{1n}\tilde{\mathbf{m}}_{2V}) \\ &\quad + \sqrt{n}2\rho_{2n}(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V} - \rho_{1n}\tilde{\mathbf{m}}_{2V})^T A\tilde{\mathbf{m}}_{12V} \\ &\quad + \sqrt{n}\rho_{2n}^2 \tilde{\mathbf{m}}_{12V}^T A\tilde{\mathbf{m}}_{12V}. \qquad (4.86)\end{aligned}$$

By Theorem 4.5, $\sqrt{n}(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V} - \rho_{1n}\tilde{\mathbf{m}}_{2V})^T A(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V} - \rho_{1n}\tilde{\mathbf{m}}_{2V})$ converges in distribution to

$$N(a^2 pVar(m_2(Z)), \frac{2p(2p-1)}{3(p-1)}\tau^2).$$

Hence, it is enough to show that

$$\sqrt{n}2\rho_{2n}(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1V} - \rho_{1n}\tilde{\mathbf{m}}_{2V})^T A\tilde{\mathbf{m}}_{12V} \xrightarrow{p} 2pabCov(\tilde{m}_2(Z), \tilde{m}_{12}(\mathbf{X}, Z)).$$

and

$$\sqrt{n}\rho_{2n}^2\tilde{\mathbf{m}}_{12V}^T A\tilde{\mathbf{m}}_{12V} \xrightarrow{p} pb^2Var(\tilde{m}_{12}(\mathbf{X}, Z))$$

These are shown in Lemmas 4.13 and 4.12, respectively. □

Lemma 4.12. *The second term in (4.86) converges in probability to* $2pabCov(\tilde{m}_2(Z), \tilde{m}_{12}(\mathbf{X}, Z))$*, i.e.*

$$\sqrt{n}2\rho_{2n}(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1V} - \rho_{1n}\tilde{\mathbf{m}}_{2V})^T A\tilde{\mathbf{m}}_{12V} \xrightarrow{p} 2pabCov(\tilde{m}_2(Z), \tilde{m}_{12}(\mathbf{X}, Z)).$$

Proof. By the definition of the matrix A, we can write

$$\begin{aligned}
&\sqrt{n}\rho_{2n}(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1V} - \rho_{1n}\tilde{\mathbf{m}}_{2V})^T A\tilde{\mathbf{m}}_{12V} = \sqrt{n}\rho_{2n}\frac{np-1}{n(n-1)p(p-1)} \times \\
&\times \sum_{i=1}^{n}\left[\sum_{j=1}^{n}\tilde{m}_{12}(\mathbf{X}_j, Z_j)I(j \in W_i)\right]\left[\sum_{k=1}^{n}(\xi_k - \Delta_{m_1}(\mathbf{X}_k) - \rho_{1n}\tilde{m}_2(Z_k))I(k \in W_i)\right] \\
&-\sqrt{n}\rho_{2n}\frac{1}{n(n-1)p}\left[p\sum_{i=1}^{n}\tilde{m}_{12}(\mathbf{X}_i, Z_i)\right]\left[p\sum_{i=1}^{n}(\xi_i - \Delta_{m_1}(\mathbf{X}_i) - \rho_{1n}\tilde{m}_2(Z_i))\right] \\
&-\sqrt{n}\rho_{2n}\frac{p}{n(p-1)}\sum_{i=1}^{n}\tilde{m}_{12}(\mathbf{X}_i, Z_i)(\xi_i - \Delta_{m_1}(\mathbf{X}_i) - \rho_{1n}\tilde{m}_2(Z_i)). \qquad (4.87)
\end{aligned}$$

Noting that

$$n^{-3/4}\sum_{i=1}^{n}\xi_i\tilde{m}_{12}(\mathbf{X}_i, Z_i) \xrightarrow{p} 0$$

and

$$\frac{1}{n^{3/4}}\sum_{i=1}^{n}\Delta_{m_1}(\mathbf{X}_i)\tilde{m}_{12}(\mathbf{X}_i, Z_i) \xrightarrow{p} 0,$$

which follows by arguments similar to (4.36), the third term in (4.87) goes in probability to

$$\frac{pab}{(p-1)}E(\tilde{m}_2(Z)\tilde{m}_{12}(\mathbf{X}, Z)).$$

Also, using (4.36), and the facts $E(\tilde{m}_{12}(\mathbf{X}, Z)) = 0$, and $n^{-3/4}\sum_{i=1}^{n}\xi_i = o_p(1)$, the second term in (4.87) goes in probability to

$$pabE(\tilde{m}_2(Z))E(\tilde{m}_{12}(\mathbf{X}, Z)).$$

Next, the component of the first term in (4.87) that corresponds to

$$\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{i=1}^{n}\tilde{m}_{12}(\mathbf{X}_j, Z_j)(\xi_k - \Delta_{m_1}(\mathbf{X}_k))I(j \in W_i)I(k \in W_i)$$

goes to zero in probability by arguments similar to those used for the last term in (4.87).

Set

$$\begin{aligned}\bar{m}_2^i(Z_i) &= \frac{1}{p}\sum_{j=1}^{n}\tilde{m}_2(Z_j)I(j \in W_i)\\ \bar{m}_{12}^i(., Z_i) &= \frac{1}{p}\sum_{j=1}^{n}\tilde{m}_{12}(\mathbf{X}_j, Z_i)I(j \in W_i),\end{aligned}$$

so that

$$\begin{aligned}&\frac{1}{p}\sum_{j=1}^{n}\tilde{m}_{12}(\mathbf{X}_j, Z_j)I(j \in W_i) = \bar{m}_{12}^i(., Z_i) + o_p(1),\\ &\frac{1}{p}\sum_{j=1}^{n}\tilde{m}_2(Z_j)I(j \in W_i) = \bar{m}_2^i(Z_i) + o_p(1).\end{aligned}$$

The remaining component of the first term in (4.87) can be written as

$$\begin{aligned}&\frac{(np-1)ab}{n(n-1)p(p-1)}\sum_{j=1}^{n}\sum_{k=1}^{n}\sum_{i=1}^{n}\tilde{m}_{12}(\mathbf{X}_j, Z_j)\tilde{m}_2(Z_k)I(j \in W_i)I(k \in W_i)\\ = \;&\frac{(np-1)pab}{(n-1)(p-1)}\frac{1}{n}\sum_{i=1}^{n}\bar{m}_{12}^i(., Z_i)\bar{m}_2^i(Z_i) + o_p(1)\\ \xrightarrow{p}\;&\frac{p^2b^2}{p-1}E\left[\tilde{m}_{12}(\mathbf{X}, Z)\tilde{m}_2(Z)\right],\end{aligned}$$

completing the proof. □

Lemma 4.13. *The third term in (4.86) converges in probability to* $pb^2Var(\tilde{m}_{12}(\mathbf{X}, Z))$*, i.e.*

$$\sqrt{n}\rho_{2n}^2\tilde{\mathbf{m}}_{12V}^T A\tilde{\mathbf{m}}_{12V} \xrightarrow{p} pb^2Var(\tilde{m}_{12}(\mathbf{X}, Z)).$$

Proof. Note that we can write $\sqrt{n}\rho_{2n}^2\tilde{\mathbf{m}}_{12V}^T A\tilde{\mathbf{m}}_{12V}$ as

$$\begin{aligned}
&\frac{(np-1)b^2}{n(n-1)p(p-1)}\sum_{i=1}^{n}\left[\sum_{j=1}^{n}\tilde{m}_{12}(\mathbf{X}_j, Z_j)I(j\in W_i)\right]\left[\sum_{k=1}^{n}\tilde{m}_{12}(\mathbf{X}_k, Z_k))I(k\in W_i)\right]\\
&-\frac{pb^2}{n(n-1)}\left[\sum_{i=1}^{n}\tilde{m}_{12}(\mathbf{X}_i, Z_i)\right]\left[\sum_{i=1}^{n}\tilde{m}_{12}(\mathbf{X}_i, Z_i)\right]\\
&-\frac{pb^2}{n(p-1)}\sum_{i=1}^{n}\tilde{m}_{12}(\mathbf{X}_i, Z_i)^2. \qquad (4.88)
\end{aligned}$$

Clearly, the third term in (4.88) goes to $[pb^2/(p-1)]E[\tilde{m}_{12}(\mathbf{X}, Z)^2]$ in probability, and the second term in (4.88) goes to $pb^2[E(\tilde{m}_{12}(\mathbf{X}, Z))]^2$ in probability. Using the same notation as in lemma 4.12, the first term in (4.88) is equal to

$$\frac{(np-1)pb^2}{n(n-1)(p-1)}\sum_{i=1}^{n}\left[\bar{m}_{12}^i(., Z_i)\right]^2 + o_p(1) \xrightarrow{p} \frac{p^2b^2}{p-1}E[(\tilde{m}_{12}(\mathbf{X}_i, Z_i))^2],$$

completing the proof. □

4.6 Test for Additivity

In this section we introduce a hypothesis test for additivity of the regression function based on the ANOVA-type methodology. Stone (1985) and Hastie and Tibshirani (1990) emphasized the importance, flexibility and usefulness of considering an additive model. There are many methods for estimating the additive functions, such as the backfitting algorithm (Hastie and Tibishrani, 1990) combined with splines or smoothers, or marginal integration (Auestad and Tjostheim, 1991, Linton and Nielsen, 1995, Newey, 1994, Tjostheim and Auestad, 1994). A comparison of backfitting and marginal integration can be found in Sperlich, Linton and Hardle (1999).

We consider the following mean function

$$m(\mathbf{x}) = \mu + \sum_{i=1}^{d} m_i(x_i) + \sum_{1\le i\le j\le d} m_{ij}(x_i, x_j),$$

where $\mathbf{X}$ is the vector of available covariates. Here we consider the homocedastic case

$$Y = \mu + \sum_{i=1}^{d} m_i(x_i) + \sum_{1\le i\le j\le d} m_{ij}(x_i, x_j) + \xi,$$

where ξ is the independet error with zero mean, constant variance and independent of $\mathbf{X}$.

The null and local alternative hypothesis for additivity are

$$H_0: \quad m(\mathbf{x}) = \mu + \sum_{i=1}^{d} m_i(x_i), \tag{4.89}$$

$$H_1: \quad m(\mathbf{x}) = \mu + \sum_{i=1}^{d} m_i(x_i) + \sum_{1\le i\le j\le d} m_{ij}(x_i, x_j) \tag{4.90}$$

Denote the null hypothesis residuals by $\xi_k = Y_k - m_1(X_{1k}) - \ldots - m_d(X_{dk})$ so that $\hat{\xi}_k = Y_k - \hat{m}_1(X_{1k}) - \ldots - \hat{m}_d(X_{dk})$.

As described before, there are many ways of estimating the additive components. For the asymptotic theory of this section to hold a few assumptions are made on the set of weights that compute such estimators. Let $\hat{m}_i(x) = S_i Y = \sum_{j=1}^{n} w_{ij}(x) Y_j$, and note that the weights w_{ij} depend on the observations $x_{i1}, \ldots, x_{in}$.

(A) $\displaystyle\sum_{j=1}^{n} w_{ij}(x) = 1.$

(B) $\displaystyle\sum_{j=1}^{n} w_{ij}(x)|X_{ij} - x| = o(1).$

(C) $w_{ij}(x) = o(1/\sqrt{n}).$

(D) The weights w_{ij} are such that $\sup_x |\hat{m}_i(x) - m_i(x)| = o(n^{-1/4})$.

Because under the alternative, the effect left on the residuals from fitting the

additive model depend on $\mathbf{X}$, we can not simply use one of the covariates to form the windows W_i. It is necessary to define the windows taking all covariates into account. Here we follow the idea of testing multiple covariates in Section 4.4. Define $\mathbf{P}$ the first principal component obtained by the SVD decomposition of the $n \times d(d-1)/2$ matrix whose columns are the observations corresponding to the cross-product of the centered covariates, i.e., $\Big[(X_1-\bar{X}_1)(X_2-\bar{X}_2), (X_1-\bar{X}_1)(X_3-\bar{X}_3), \ldots, (X_1-\bar{X}_1)(X_d-\bar{X}_d), \ldots, (X_{d-1}-\bar{X}_{d-1})(X_d-\bar{X}_d)\Big]$. Then, obtain the vector

$$\hat{\boldsymbol{\xi}}_V = (\hat{\xi}_j, j \in W_{(p-1)/2+1}, \ldots, \hat{\xi}_j, j \in W_{n-(p-1)/2})^T.$$

where the window W_i is defined by

$$W_i = \left\{ j : |\hat{F}_P(P_j) - \hat{F}_P(P_i)| \leq \frac{p-1}{2n} \right\}, \tag{4.91}$$

where $\hat{F}_P$ is the empirical distribution function of P.

Define the test statistic as $MST - MSE = \hat{\boldsymbol{\xi}}_V^T A \hat{\boldsymbol{\xi}}_V$.

Theorem 4.7. *Assume that the marginal densities f_X and f_Z of X and Z respectively are bounded away from zero, $E(\epsilon_k^4) < \infty$ and $m_1(x), m_2(z)$ are Lipschitz continuous. If assumptions (A) - (D) hold, then, under H_1 in (4.90), the asymptotic distribution of the test statistic is given by*

$$n^{1/2}(MST - MSE) \xrightarrow{d} N\left(0, \frac{2p(2p-1)}{3(p-1)}\sigma^4\right).$$

Note that the optimal rate of convergence for dimension d is $(n^{-1}log(n))^{q/2q+d}$ (Stone, 1982), where q is a measure of smoothness (number of derivatives). For additive models, the optimal rate is $n^{-2(q/2q+1)}$ for each component. The backfitting algorithm and the marginal integreation estimation have convergence rate $n^{-2/5}$, which is faster than the one we require. There is no literature to our knowledge studying the theory for uniform convergence rates for these algorithms.

An estimate of τ^2 can be obtained by modifying Rice's (1984) estimator in the

folowing way:

$$\hat{\tau}^2 = \frac{1}{4(n-3)} \sum_{j=2}^{n-2} (\hat{\xi}_j - \hat{\xi}_{j-1})^2 (\hat{\xi}_{j+2} - \hat{\xi}_{j+1})^2. \tag{4.92}$$

Proof of Theorem 4.7

Under H_0 we have

$$\begin{aligned} \hat{\xi}_k &= Y_k - \sum_{i=1}^{d} \hat{m}_i(X_{ik}) + \sum_{i=1}^{d} m_i(X_{ik}) - \sum_{i=1}^{d} m_i(X_{ik}) \\ &= \xi_k - \sum_{i=1}^{d} (\hat{m}_i(X_{ik}) - m_i(X_{ik})) \\ &= \xi_k - \sum_{i=1}^{d} \Delta_i(X_{ik}). \end{aligned} \tag{4.93}$$

Now, we can write $\hat{\boldsymbol{\xi}}_V$ as

$$\begin{aligned} \hat{\boldsymbol{\xi}}_V &= \{\hat{\xi}_k : k \in W_1, ..., \hat{\xi}_k : k \in W_n\}^T \\ &= \{\xi_k - \sum_{i=1}^{d} \Delta_i(X_{ik}) : k \in W_1, ..., \xi_k - \sum_{i=1}^{d} \Delta_i(X_{ik}) : k \in W_n\}^T \\ &= \boldsymbol{\xi}_V - \sum_{i=1}^{d} \boldsymbol{\Delta}_{V,i}, \end{aligned}$$

and therefore MST - MSE can be written as

$$\hat{\boldsymbol{\xi}}_V^T A \hat{\boldsymbol{\xi}}_V = \boldsymbol{\xi}_V^T A \boldsymbol{\xi}_V - 2\boldsymbol{\xi}_V^T A \left(\sum_{i=1}^{d} \boldsymbol{\Delta}_{V,i} \right) + \left(\sum_{i=1}^{d} \boldsymbol{\Delta}_{V,i} \right)^T A \left(\sum_{i=1}^{d} \boldsymbol{\Delta}_{V,i} \right). \tag{4.94}$$

By Teorem 4.1, we know that $\boldsymbol{\xi}_V^T A \boldsymbol{\xi}_V$ converges to a normal distribution. That the second term of (4.94) goes to zero in probability is proven in Lemma 4.14. Using assumption (D) and arguments similar to those in the proof of Lemma 4.5, it is easy to show that the last term in (4.94) goes in probability to 0.

It remains to find the asymptotic variance. Note that

$$E[\boldsymbol{\xi}_V' A_d \boldsymbol{\xi}_V | \mathbf{X} = \mathbf{x}]^2$$

$$
\begin{aligned}
&= \frac{1}{n^2(p-1)^2}\sum_{i_1,i_2}^{n}\sum_{j_1\neq l_1}^{n}\sum_{j_2\neq l_2}^{n} E(\xi_{j_1}\xi_{l_1}\xi_{j_2}\xi_{l_2}|\mathbf{X}=\mathbf{x})I(j_s\in W_{i_s}, l_s\in W_{i_s}, s=1,2)\\
&= \frac{2}{n^2(p-1)^2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j\neq l}^{n} E(\xi_j^2\xi_l^2|\mathbf{X}=\mathbf{x})I(j,l\in W_{i_1}\cap W_{i_2})\\
&= \frac{2}{n^2(p-1)^2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j\neq l}^{n} \sigma^2\sigma^2 I(j,l\in W_{i_1}\cap W_{i_2})\\
&= \frac{2}{n^2(p-1)^2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{j\neq l}^{n} \sigma^4 I(j,l\in W_{i_1}\cap W_{i_2})\\
&= \frac{2}{n^2(p-1)^2}\sum_{j=1}^{n}\sigma^4\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\sum_{l\neq j}^{n} I(j,l\in W_{i_1}\cap W_{i_2})\\
&= \frac{2}{n^2(p-1)^2}\sum_{j=1}^{n}\sigma^4 2(1+2^2+3^2+...+(p-1)^2)\\
&= \frac{2}{n(p-1)^2}\frac{p(p-1)(2p-1)}{3}\sigma^4,
\end{aligned}
$$

where the second to last equality results from the fact that if $1 \leq |j_1 - j_2| = s \leq p-1$, then they are $(p-s)^2$ pairs of windows whose intersection includes j_1 and j_2.

Thus,

$$
E\left(\left(\frac{n}{p}\right)^{1/2}\boldsymbol{\xi}_V^T A_d \boldsymbol{\xi}_V|\mathbf{X}=\mathbf{x}\right)^2 = \frac{2(2p-1)}{3(p-1)}\sigma^4
$$

Lemma 4.14. *The second term in (4.94) goes to 0 in probability, i.e.,*

$$
\sqrt{n}\boldsymbol{\xi}_V^T A\left(\sum_{i=1}^{d}\boldsymbol{\Delta}_V\right)\xrightarrow{p} 0.
$$

Proof. Write

$$
\begin{aligned}
\sqrt{n}\boldsymbol{\xi}_V^T A\left(\sum_{i=1}^{d}\boldsymbol{\Delta}_{V,i}\right) &= \sqrt{n}\sum_{i=1}^{d}\boldsymbol{\xi}_V^T A\boldsymbol{\Delta}_{V,i}\\
&= \frac{n^{-1/2}(np-1)}{(n-1)p(p-1)}\sum_{i=1}^{d}\sum_{k=1}^{n}\sum_{j\in W_k}\xi_j\sum_{\ell\in W_k}\Delta_i(\mathbf{X}_{i\ell})
\end{aligned}
$$

$$-\frac{n^{-1/2}p}{(n-1)}\sum_{i=1}^{d}\sum_{k=1}^{n}\xi_k\sum_{j=1}^{n}\Delta_i(\mathbf{X}_{ij})$$
$$-\frac{n^{-1/2}p}{(p-1)}\sum_{i=1}^{d}\sum_{k=1}^{n}\xi_k\Delta_i(\mathbf{X}_{ik}). \tag{4.95}$$

For the second term in (4.95), note that $n^{-1/2}\sum_{i=1}^{n}\xi_i$ remains bounded in probability, and $n^{-1}\sum_{k=1}^{n}\Delta_{m_1}(\mathbf{X}_k)\xrightarrow{p}0$, therefore it convergences to zeroin probability. Using the fact that and arguments similar to those in the proof of Lemma 4.5, the first and last terms in (4.95) go in probability to 0. □

4.6.1 Simulations: Test for Additivity

The simulation settings used in this section are borrowed from Debort, Dette and Munk (2002). The data is generated with a uniform grid $\{(\frac{i}{n_1},\frac{j}{n_2})\}_{j=1,\ldots,n_2}^{i=1,\ldots,n_1}$ with $\sigma = 1, 0.5$, or 0.1. Table 4.17 shows the level of the test at $\alpha = 0.05$ for different additive functions (with $\sigma = 1$). The ANOVA-type test with backfitting algorithm, referred as A(p) in the tables, mantain its level in almost all cases. We included the test proposed by Debort, Dette and Munk (2002), DDM in the tables.

$$g_1(x_1,x_2) = 0$$
$$g_2(x_1,x_2) = x_1 + x_2$$
$$g_3(x_1,x_2) = e^{x_1} + sin(\pi x_2)$$
$$g_4(x_1,x_2) = sin(\pi x_1) + sin(\pi x_2)$$
$$g_5(x_1,x_2) = e^{x_1} + e^{x_2}.$$

To study the power under alternatives, the following functions are tested:

$$g_6(x_1,x_2) = x_1x_2$$
$$g_7(x_1,x_2) = \frac{exp(5(x_1+x_2))+1}{exp(5(x_1+x_2))-1}$$
$$g_8(x_1,x_2) = 0.5(1+sin(2\pi(x_1+x_2)))$$

$$g_9(x_1, x_2) = 64(x_1x_2)^3(1 - x_1x_2)^3$$
$$g_{10}(x_1, x_2) = I\{x_1 > 0.5, x_2 > 0.5\}.$$

Table 4.17. Percent Rejection rates for the models representing the null hypothesis

	$(n_1, n_2) = (5, 5)$			$(n_1, n_2) = (10, 10)$			$(n_1, n_2) = (5, 20)$		
test	A(7)	A(9)	DDM	A(7)	A(9)	DDM	A(7)	A(9)	DDM
g_1	6.2	3.7	6.2	5.6	4.3	4.4	2.6	2.2	4.2
g_2	5.7	4.8	6.2	5.5	4.6	4.6	2.8	2.4	4.2
g_3	5.5	4.4	5.7	5.5	4.6	4.5	3.4	2.0	4.4
g_4	5.4	4.6	4.8	5.3	5.1	4.9	2.8	2.5	4.4
g_5	6.1	4.5	6.0	5.3	4.8	4.3	2.9	2.2	4.4

Table 4.18 shows the results for the ANOVA-type statistic using the backfitting algorithm, for DDM and for the test proposed Barry (1993), referred as BA in the table. Between parenthesis are the results for the ANOVA-type when using the marginal integration for the estimation of the regression function under the null hypothesis.

4.6.2 Power of the test under local alternatives when testing for additivity

In this section we will study the asymptotic behavior of the test for additivity under local alternatives. Here we only state the theorem in terms of two covariates, but the extension is straight forward. Let

$$\begin{aligned} H_0 : m(x, z) &= m_1(x) + m_2(z) \\ H_1 : m(x, z) &= m_1(x) + m_2(z) + \rho_n m_{12}(x, z), \end{aligned} \tag{4.96}$$

for a sequence of constants ρ_n converging to 0. Again, if ρ_n goes to 0 too fast, the null hypothesis will be confounded with the alternative, and the test will not be able to detect the diference, having impact on the power. But for ρ_n going to 0 too slow, the null hypothesis will always be rejected for a large enough n.

Table 4.18. Percent Rejection rates for the models representing alternative hypothesis

		$(n_1, n_2) = (5, 5)$			$(n_1, n_2) = (20, 5)$			$(n_1, n_2) = (20, 20)$		
g	σ	A(7)	DDM	BA	A(9)	DDM	BA	A(11)	DDM	BA
	.1	67.5(68.2)	64.8	98.4	99.9(1)	93.9	1	1	1	1
g_6	.5	7.9(7.9)	7.2	11.3	9.0(8.8)	8.3	36.2	32.0	74.7	89.2
	1	6.5(6.3)	6.3	7.3	3.5(3.9)	6.8	10.5	9.7	31.8	34.0
	.1	12.6	11.8	65.6	40.7	33.9	1	1	87.4	1
g_7	.5	6.6	7.9	8.3	3.2	15.3	18.2	11.7	44.7	68.9
	1	5.5	5.2	5.9	2.1	5.9	8.0	6.2	15.1	21.1
	.1	72.3	1	1	1	1	1	0	1	1
g_8	.5	24.1	36.7	30.5	68.7	1	99.8	5.2	1	1
	1	11.9	15.3	9.9	16.2	64.7	59.5	5.3	97.7	99.9
	.1	1	72.0	18.4	1	1	66.6	1	1	1
g_9	.5	1	20.5	6.9	1	45.5	7.4	1	89.1	11.5
	1	99.6	8.3	5.7	1	11.9	6.3	1	21.9	5.4
	.1	1	45.6	1	1	1	1	1	1	1
g_{10}	.5	52.2	26.7	47.0	91.2	58.6	98.8	1	91.2	1
	1	17.1	13.4	22.1	27.5	39.2	56.7	80.5	59.2	99.5

Under H_1 we have

$$Y_i = m_1(X_i) + m_2(Z_i) + \rho_n m_{12}(X_i, Z_i) + \xi_i. \tag{4.97}$$

Recall that $\boldsymbol{\xi}_V$ is defined as

$$\boldsymbol{\xi}_V = (\hat{\xi}_j, j \in W_1, ..., \hat{\xi}_j, j \in W_n)^T,$$

where the windows W_i are defined according to $(X_i - \bar{X})(Z_i - \bar{Z})$ as described previously.

Note that we can write $\hat{\xi}_j$ as

$$\begin{aligned}
\hat{\xi}_j &= Y_j - \hat{m}_1(X_j) - \hat{m}_2(Z_j) \\
&= Y_j - m_1(X_j) - m_2(Z_j) - \rho_n m_{12}(X_j, Z_j) \\
&\quad -[\hat{m}_1(X_j) - m_1(X_j)] - [\hat{m}_2(Z_j) - m_2(Z_j)] + \rho_n m_{12}(X_j, Z_j)
\end{aligned}$$

$$= \xi_j - \Delta_{m_1}(X_j) - \Delta_{m_2}(Z_j) + \rho_n m_{12}(X_j, Z_j), \tag{4.98}$$

and therefore

$$\begin{aligned} \boldsymbol{\xi}_V &= \begin{pmatrix} \xi_j - \Delta_{m_1}(X_j) - \Delta_{m_2}(Z_j) + \rho_n m_{12}(X_j, Z_j), j \in W_1 \\ \vdots \\ \xi_j - \Delta_{m_1}(X_j) - \Delta_{m_2}(Z_j) + \rho_n m_{12}(X_j, Z_j), j \in W_n \end{pmatrix} \\ &= \boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V} - \boldsymbol{\Delta}_{m_2 V} + \rho_n \mathbf{m}_{12V}. \end{aligned} \tag{4.99}$$

With this set up, we have the following theorem

Theorem 4.8. *Consider the notation and assumptions of Theorem 4.7. Moreover, assume that $m_{12}(x, z)$ is Lipschitz continuous uniformly on x and z. Then, under H_1 in (4.96), as $n \to \infty$,*

$$\left(\frac{n}{p}\right)^{1/2} (MST - MSE) \to N\left(Var(m_{12}(X, Z)), \frac{2(2p-1)}{3(p-1)}\sigma^4\right).$$

Proof of Theorem 4.8

Note that the test statistic

$$\begin{aligned} &\sqrt{n}(MST - MSE) = \sqrt{n}\boldsymbol{\xi}_V^T A \boldsymbol{\xi}_V \\ = \; &\sqrt{n}(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V} - \boldsymbol{\Delta}_{m_2 V} - \rho_n \mathbf{m}_{12V})^T A (\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V} - \boldsymbol{\Delta}_{m_2 V} - \rho_n \mathbf{m}_{12V}). \end{aligned}$$

We know by Theorem 4.7 that $\sqrt{n}(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V} - \boldsymbol{\Delta}_{m_2 V})^T A(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V} - \Delta_{m_2 V})$ is asymptotically Normal distributed with mean 0 and variance $\frac{2p(2p-1)}{3(p-1)}\sigma^4$. Hence, it is enough to show that $\sqrt{n}2\rho_n(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V} - \boldsymbol{\Delta}_{m_2 V})^T A \mathbf{m}_{12V} \to^P 0$ and $\sqrt{n}\rho_n^2 \mathbf{m}_{12V}^T A \mathbf{m}_{12V} \to^P p^{1/2} V(m_{12}(X, Z))$.

Using the assumptions of the theorem and steps similar to those in Theorem 4.5, it is easy to show that $\sqrt{n}2\rho_n(\boldsymbol{\xi}_V - \boldsymbol{\Delta}_{m_1 V})^T A \mathbf{m}_{12V} \to^P 0$. A similar proof will show that $\sqrt{n}2\rho_n \boldsymbol{\Delta}_{m_2 V}^T A \mathbf{m}_{12V} \to^P 0$. Using steps similar to those in Lemma 4.13, we find

$$\sqrt{n}\rho_{12n}^2 \mathbf{m}_{12V}^T A \mathbf{m}_{12V} \to^P p^{1/2} V(m_{12}(X, Z)). \tag{4.100}$$

4.7 Auxiliary Results

In this section we introduce lemmas that will be necessary to the proof of the theorems in this Chapter.

Lemma 4.15. *Let $\hat{F}_n(x)$ be the empirical distribution function of X with cdf F based on a sample of size n. Then, for any constant c,*

$$sup_{x_i,x_j}\left\{|F(x_i)-F(x_j)|I\left[|\hat{F}(x_i)-\hat{F}(x_j)|\leq\frac{c}{n}\right]\right\}=O_p\left(\frac{1}{\sqrt{n}}\right).$$

Proof. By Dvoretzky, Kiefer and Wolfowitz (1956), we have that

$$\forall\epsilon\geq 0, P\left(\sup_x|\hat{F}_n(x)-F(x)|\geq\epsilon\right)\leq Ce^{-2n\epsilon^2}.$$

Therefore, $|\hat{F}(x)-F(x)|=O_p\left(\frac{1}{\sqrt{n}}\right)$ uniformly on x. Hence, writing

$$|F(x_i)-F(x_j)| \;=\; |F(x_i)-\hat{F}_n(x_i)+\hat{F}_n(x_i)-F(x_j)+\hat{F}_n(x_j)-\hat{F}_n(x_j)|,$$

we have

$$\begin{aligned}
& sup_{x_i,x_j}\left\{|F(x_i)-F(x_j)|I\left[|\hat{F}(x_i)-\hat{F}(x_j)|\leq\frac{p-1}{2n}\right]\right\}\\
\leq\;& sup_{x_i,x_j}\left\{|F(x_i)-\hat{F}_n(x_i)|+|\hat{F}_n(x_j)-F(x_j)|\right\}\\
& +sup_{x_i,x_j}\left\{|\hat{F}_n(x_i)-\hat{F}_n(x_j)|\right\}I\left[|\hat{F}_n(x_i)-\hat{F}_n(x_j)|\leq\frac{p-1}{2n}\right]\\
& =O_p\left(\frac{1}{\sqrt{n}}\right)+O_p\left(\frac{1}{\sqrt{n}}\right)+O_p\left(\frac{1}{n}\right).
\end{aligned}$$

□

Lemma 4.16. *For any Lipschitz continuous function $g(x)$, we have*

$$\frac{1}{p}\sum_{j=1}^{n}g(z_j)I(j\in W_i)-g(z_i)=O_p\left(\frac{1}{\sqrt{n}}\right).$$

Proof. First note that by the Lipschitz continuity and the Mean Value Theorem

we have, for some constant M,

$$|g(z_j) - g(z_i)| \leq M|z_j - z_i| \leq M|F_Z(z_j) - F_Z(z_i)|/f_Z(\tilde{x}_{ij})$$

where $\tilde{x}_{ij}$ is between z_j and z_i. Thus,

$$\begin{aligned} & |\frac{1}{p}\sum_{j=1}^{n} g(z_j)I(j \in W_i) - g(z_i)| \\ \leq & \frac{1}{p}\sum_{j=1}^{n} |g(z_j) - g(z_i)| I\left[|\hat{F}_Z(z_i) - \hat{F}_Z(z_j)| \leq \frac{p-1}{2n}\right] \\ \leq & \frac{M}{p}\sum_{j=1}^{n} \frac{|F_Z(z_j) - F_Z(z_i)|}{f_Z(\tilde{x}_{ij})} I\left[|\hat{F}_Z(z_i) - \hat{F}_Z(z_j)| \leq \frac{p-1}{2n}\right] = O_p\left(\frac{1}{\sqrt{n}}\right), \end{aligned}$$

where the last equality follows from Lemma 4.15 and the assumption that f_Z remains bounded away from zero.

□

Lemma 4.17. *Assume $\sigma^2(., \mathbf{z})$ is Lipschitz continuous and $E(\epsilon_i^4) < \infty$. Then, under H_0 as $n \to \infty$,*

$$n^{1/2}[\boldsymbol{\xi}_V' A \boldsymbol{\xi}_V - \boldsymbol{\xi}_V' A_d \boldsymbol{\xi}_V]$$

goes in probability to 0, where A_d is the block diagonal matrix

$$A_d = diag\{B_1, ..., B_n\} \text{ with } B_i = \frac{1}{n(p-1)}[\mathbf{J}_p - \mathbf{I}_p].$$

The proof of this lemma follows from Lemma 3.1 in Wang, Akritas and Van Keilegom (2008).

Lemma 4.18. *Let Y is the response variable, $\tilde{\mathbf{X}} = (\mathbf{1}, \mathbf{X})$, where $\mathbf{X}$ is the $n \times d$ matrix of covariates and $\mathbf{1}$ is the vector of 1, and let e_j denote the $(d+1) \times 1$ vector having 1 in the j-th entry and all other entries 0. In the least squares regression estimation of $(\alpha, \boldsymbol{\beta})^T$ of the model*

$$\mathbf{Y} = \tilde{\mathbf{X}} \begin{pmatrix} \alpha \\ \boldsymbol{\beta} \end{pmatrix} + \epsilon,$$

the weights

$$w_j = e_1^T(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T e_j, \quad j = 1, \ldots, n$$

from the estimator $\hat{\alpha} = e_1^T(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^T Y$ *are such that*

$$\sum_{j=1}^{n} w_j = 1.$$

Proof. Note first that $\hat{\boldsymbol{\beta}} = (\bar{\mathcal{X}}^T\bar{\mathcal{X}})^{-1}\bar{\mathcal{X}}^T Y$, where the columns of $\bar{\mathcal{X}}$ are the centered columns of the design matrix.

Since

$$\begin{aligned} \hat{\alpha} &= \bar{Y} - \hat{\beta}_1\bar{X} - \ldots \hat{\beta}_d\bar{X}_d \\ &= \left(\frac{1}{n}\mathbf{1}^T - (\bar{X}, \ldots, \bar{X}_d)(\bar{\mathcal{X}}^T\bar{\mathcal{X}})^{-1}\bar{\mathcal{X}}^T\right)\mathbf{Y}. \end{aligned}$$

The lemma follows from the fact that the weights $(\bar{X}, \ldots, \bar{X}_d)(\bar{\mathcal{X}}^T\bar{\mathcal{X}})^{-1}\bar{\mathcal{X}}^T$ sum to zero because $\bar{\mathcal{X}}^T\mathbf{1} = \mathbf{0}$. □

Lemma 4.19. *For the local polynomial regression estimator (4.47)*

$$\hat{m}_1(\mathbf{x}) = \mathbf{e}_1^T\left(\mathbb{X}_\mathbf{x}^T\mathbb{W}_\mathbf{x}\mathbb{X}_\mathbf{x}\right)^{-1}\mathbb{X}_\mathbf{x}^T\mathbb{W}_\mathbf{x},$$

each of the weights denoted by

$$\tilde{w}(\mathbf{x}, \boldsymbol{X}_j) = \mathbf{e}_1^T\left(\mathbb{X}_\mathbf{x}^T\mathbb{W}_\mathbf{x}\mathbb{X}_\mathbf{x}\right)^{-1}\mathbb{X}_\mathbf{x}^T\mathbb{W}_\mathbf{x}\mathbf{e}_j, \quad j = 1, \ldots, n,$$

is of order $O_p\left(\frac{1}{n|H_n|^{1/2}}\right)$.

Proof. Recall from (4.47) that

$$\mathbb{X}_\mathbf{x} = \begin{pmatrix} 1 & (\mathbf{X} - \mathbf{x})^T & \text{vech}^T\left\{(\mathbf{X} - \mathbf{x})(\mathbf{X} - \mathbf{x})^T\right\} & \ldots \\ \vdots & \vdots & \vdots & \ldots \\ 1 & (\mathbf{X}_n - \mathbf{x})^T & \text{vech}^T\left\{(\mathbf{X}_n - \mathbf{x})(\mathbf{X}_n - \mathbf{x})^T\right\} & \ldots \end{pmatrix},$$

$\mathbb{W}_\mathbf{x} = diag\{K_H(\mathbf{X}_1 - \mathbf{x}), \ldots, K_H(\mathbf{X}_n - \mathbf{x})\}$, and that the dimensions of $\mathbb{X}_\mathbf{x}$ are

(4.61) $n \times \gamma_d$.

Now, it is easy to see that $\frac{1}{n}\mathbb{X}_{\mathbf{x}}^T\mathbb{W}_{\mathbf{x}}$ is a $\gamma_d \times n$ matrix with column j given by

$$\mathbf{V}_j := \frac{1}{n}\begin{pmatrix} K_{H_n}(\mathbf{X}_j - \mathbf{x}) \\ K_{H_n}(\mathbf{X}_j - \mathbf{x})(\mathbf{X}_j - \mathbf{x}) \\ K_{H_n}(\mathbf{X}_j - \mathbf{x})\text{vech}\{(\mathbf{X}_j - \mathbf{x})(\mathbf{X}_j - \mathbf{x})^T\} \\ \vdots \end{pmatrix}. \tag{4.101}$$

Let $\mathbf{1}$ be the vector of ones (of dimension defined by the context) and $D_f(\mathbf{x})$ the vector of partial derivatives of $f(\mathbf{x})$. As in Ruppert and Wand (1994), define

$$\begin{aligned} N_{\mathbf{x}} &= \begin{pmatrix} \nu_{\mathbf{x},11} & \nu_{\mathbf{x},12} & \nu_{\mathbf{x},13} & \dots \\ \nu_{\mathbf{x},21} & \nu_{\mathbf{x},22} & \nu_{\mathbf{x},23} & \dots \\ \nu_{\mathbf{x},31} & \nu_{\mathbf{x},32} & \nu_{\mathbf{x},33} & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \\ &= \int \begin{bmatrix} 1 \\ \mathbf{u} \\ \text{vech}(\mathbf{u}\mathbf{u}^T) \\ \vdots \end{bmatrix} \begin{bmatrix} 1 & \mathbf{u} & \text{vech}^T(\mathbf{u}\mathbf{u}^T) & \dots \end{bmatrix} K(\mathbf{u})d\mathbf{u}, \end{aligned}$$

and

$$Q_{\mathbf{x}} = \int K(\mathbf{u}) \begin{bmatrix} 0 & \mathbf{u}^T & 0 & \dots \\ \mathbf{u} & 0 & \mathbf{u}\text{vech}^T(\mathbf{u}\mathbf{u}^T) & \dots \\ 0 & \text{vech}(\mathbf{u}\mathbf{u}^T)\mathbf{u}^T & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \{D_f^T(\mathbf{x})H_n^{1/2}\mathbf{u}\}d\mathbf{u}.$$

It is known (Ruppert and Wand, 1994) that $Q_{\mathbf{x}} = O(tr(H_n^{1/2}))$, where $tr(H_n)$ is the trace of matrix H_n.

For $\ell = 2, \dots, q$, let C_ℓ be a matrix whose each element is of the same order of a product of ℓ elements of $H_n^{1/2}$. For example, C_2 can be defined (see proof of Theorem 3.1 Ruppert and Wand, 1994) as the $\frac{1}{2}d(d+1) \times \frac{1}{2}d(d+1)$ matrix such that

$$\text{vech}(H_n^{1/2}\mathbf{u}\mathbf{u}^T H_n^{1/2}) = C_2\text{vech}(\mathbf{u}\mathbf{u}^T),$$

for all d-vectors $\mathbf{u}$.

Extending the formulas of $\left(n^{-1}\mathbb{X}_{\mathbf{x}}^T\mathbb{W}_{\mathbf{x}}\mathbb{X}_{\mathbf{x}}\right)^{-1}$ in Ruppert and Wand (1994) to a more general case, we have that

$$\mathbf{e}_1^T\left(n^{-1}\mathbb{X}_{\mathbf{x}}^T\mathbb{W}_{\mathbf{x}}\mathbb{X}_{\mathbf{x}}\right)^{-1} = O_p\left(\mathbf{1}^T\text{diag}\{1, H_n^{-1/2}, C_2^{-1}, C_3^{-1}, \ldots\}\right).$$

Also, each column V_j of $\frac{1}{n}\mathbb{X}_{\mathbf{x}}^T\mathbb{W}_{\mathbf{x}}$ defined in (4.101) is of order

$$O_p\left(\frac{1}{n|H_n|^{1/2}}\text{diag}\{1, H_n^{1/2}, C_2, C_3, \ldots\}\mathbf{1}\right).$$

Therefore, noting that $\mathbf{1}$ is a $\gamma_d \times 1$ vector and that each weight $w(\mathbf{x}, \mathbf{X}_j)$ is computed by $\mathbf{e}_1^T\left(n^{-1}\mathbb{X}_{\mathbf{x}}^T\mathbb{W}_{\mathbf{x}}\mathbb{X}_{\mathbf{x}}\right)^{-1}\mathbf{V}_j$, we have that

$$\begin{aligned}
& \mathbf{e}_1^T\left(n^{-1}\mathbb{X}_{\mathbf{x}}^T\mathbb{W}_{\mathbf{x}}\mathbb{X}_{\mathbf{x}}\right)^{-1}\mathbf{V}_j \\
= \;& O_p\left(\mathbf{1}^T\text{diag}\{1, H_n^{-1/2}, C_2^{-1}, C_3^{-1}, \ldots\}\right) O_p\left(\frac{1}{n|H_n|^{1/2}}\text{diag}\{1, H_n^{1/2}, C_2, C_3, \ldots\}\mathbf{1}\right) \\
= \;& O_p\left(\frac{1}{n|H_n|^{1/2}}\right),
\end{aligned}$$

completing the proof. □

Lemma 4.20. *For a symmetric, positive definite bandwidth matrix $H_n^{1/2}$, we have*

1. *The determinant of $H_n^{1/2}$ is equal to the product of the eigen-values of $H_n^{1/2}$.*
2. *Define the norm $||H_n^{1/2}||$ to be the maximum of its eigenvalues. Given $\mathbf{X}$, for the weights used in local polynomial regression (4.47)*

$$w(\boldsymbol{X}_{1i}, \boldsymbol{X}_{1j}) = \mathbf{e}_1^T\left(\mathbb{X}_{\mathbf{X}_{1i}}^T\mathbb{W}_{\mathbf{X}_{1i}}\mathbb{X}_{\mathbf{X}_{1i}}\right)^{-1}\mathbb{X}_{\mathbf{X}_{1i}}^T\mathbb{W}_{\mathbf{X}_{1i}}\mathbf{e}_j,$$

we have that

$$\sum_{j=1}^{n} w(\boldsymbol{X}_{1i}, \boldsymbol{X}_{1j})||\boldsymbol{X}_{1j} - \boldsymbol{X}_{1i}|| = O(||H_n^{1/2}||). \tag{4.102}$$

Proof. 1. Note that the bandwidthmatrix $H_n^{1/2}$ is positive definite, therefore there exists an eigen value decomposition $H_n^{1/2} = V\Lambda V^{-1}$ such that V is a

orthogonal matrix with columns corresponding to the eigen-vectors of $H_n^{1/2}$ and Λ is a diagonal matrix with elements corresponding to the eigen-values of $H_n^{1/2}$. Thus,

$$|H_n^{1/2}| = |V\Lambda V^{-1}| = |V||\Lambda||V^{-1}| = |VV^{-1}||\Lambda| = |\Lambda|.$$

2. Let b be such that $K(\mathbf{x}) = K(\mathbf{x})I(||\mathbf{x}|| \leq \sqrt{d-1}b)$. Such a b exists by the assumption that the density K has bounded support. By noting that

$$\begin{aligned} & K\left(H_n^{-1/2}(\mathbf{X}_{1i} - \mathbf{X}_{1j})\right)||\mathbf{X}_{1j} - \mathbf{X}_{1i}|| \\ = \; & K\left(H_n^{-1/2}(\mathbf{X}_{1i} - \mathbf{X}_{1j})\right)||H_n^{1/2}H_n^{-1/2}(\mathbf{X}_{1i} - \mathbf{X}_{1j})|| \\ \leq \; & K\left(H_n^{-1/2}(\mathbf{X}_{1i} - \mathbf{X}_{1j})\right)||H_n^{1/2}||||H_n^{-1/2}(vX_{1i} - \mathbf{X}_{1j})|| \\ \leq \; & K\left(H_n^{-1/2}(\mathbf{X}_{1i} - \mathbf{X}_{1j})\right)||H_n^{1/2}||\sqrt{d-1}b, \end{aligned}$$

we have that

$$\begin{aligned} & K_{H_n}\left(H_n^{-1/2}(\mathbf{X}_{1i} - \mathbf{X}_{1j})\right)||\mathbf{X}_{1j} - \mathbf{X}_{1i}|| \\ = \; & O\left(K_{H_n}\left(H_n^{-1/2}(\mathbf{X}_{1i} - \mathbf{X}_{1j})\right)||H_n^{1/2}||\sqrt{d-1}b\right) \\ = \; & K_{H_n}\left(H_n^{-1/2}(\mathbf{X}_{1i} - \mathbf{X}_{1j})\right)O\left(||H_n^{1/2}||\right). \end{aligned} \tag{4.103}$$

Let $V_{\mathbf{X}_{1i}}$ be the $n \times 1$ vector with j-th entry $||\mathbf{X}_{1i} - \mathbf{X}_{1j}||O\left(||H_n^{1/2}||\right)$. From the definition of the weights, the left hand side of (4.102) is equal to

$$\begin{aligned} & \mathbf{e}_1^T\left(\mathbb{X}_{\mathbf{X}_{1i}}^T\mathbb{W}_{\mathbf{X}_{1i}}\mathbb{X}_{\mathbf{X}_{1i}}\right)^{-1}\mathbb{X}_{\mathbf{X}_{1i}}^T\mathbb{W}_{\mathbf{X}_{1i}}V_{\mathbf{X}_{1i}} \\ = \; & O\left(||H_n^{1/2}||\right)\mathbf{e}_1^T\left(\mathbb{X}_{\mathbf{X}_{1i}}^T\mathbb{W}_{\mathbf{X}_{1i}}\mathbb{X}_{\mathbf{X}_{1i}}\right)^{-1}\mathbb{X}_{\mathbf{X}_{1i}}^T\mathbb{W}_{\mathbf{X}_{1i}}\mathbf{1} \\ = \; & O\left(||H_n^{1/2}||\right). \end{aligned}$$

The first equality follows from the fact that the each entry j of the $n \times 1$ vector $\mathbb{W}_{\mathbf{X}_{1i}}V_{\mathbf{X}_{1i}}$ is equal to (4.103), and the last equality follows from the fact that the weights sum to 1, proved in Lemma 4.18. This completes the proof.

□

Chapter 5

Nonparametric Variable Selection

5.1 Introduction to Variable Selection in Regression and Literature Review

In regression analysis, the main problem is to select the set of predictors to include in the model, i.e, given a dependent variable Y and a set of d potencial predictors $\mathbf{X} = X_1, \ldots, X_d$, the goal is to select the subset of $\mathbf{X}$ that, according to some creteria, gives the best model. By using the selected predictors in the model, we expect to have a good fit, with a reasonable prediction and we would like to avoid overfitting. Therefore, selecting only the (few) covariates that are really important in the model is crucial.

It is of great importance to stablish the goal of the variable selection in terms of the model. One can define variable selection by selecting the predictors that have any influence in the response variable, but this is not the goal of this chapter. Our objective is to identify those predictors that influence the mean regression function specifically, and this is an issue that has not been well explored in the nonparametric context.

The regression model most often considered is the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$, where ϵ has mean zero and variance σ^2. It is natural to choose the subset of predictors by looking at all the subsets possible and comparing them with a criteria. Many of the available methods are based on the Residual Sum of Squares (RSS), on the Prediction Sum of Squares (PRESS), Adjusted R-squared, and the number of

covariates in the subset. The first and very well known methods proposed for this problem are the Mallows Cp (Mallows 1973), the AIC (Akaike 1974) and the BIC (Schwarz 1978). Mallows Cp is based purely on RSS and the number of covariates, while AIC and BIC are based on the likelihood and the number of covariates. One can use various forms of selection algorithms to choose the model based on these criteria: forward, backward or stepwise selection.

The dual problems of testing for the significance of a particular covariate, and identification of the set of relevant covariates are very common both in applied research and in methodological investigations. Due to readily available software, these tasks are often performed under the assumption of a linear model, $m(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$. Model checking fits naturally in the methodological context of hypothesis testing, while variable selection is typically addressed through minimization of a constrained or penalized objective function.

Recently, the idea of Penalized Least Squares has changed variable selection in regression to a whole new angle and here we will describe the most recent methods. The idea is to estimate the parameters $\boldsymbol{\beta}$ by minimizing the penalized function

$$\frac{1}{2}||Y - \mathbf{X}\boldsymbol{\beta}||^2 + n\sum_{j=1}^{d} p_\lambda(|\beta_j|),$$

where p_λ is a penalty function with tunning parameter λ. Penalty functions are expected to have three main effects on the estimators: the bias of the estimator is small when the true unknown parameter is large (unbiasedness), it sets to zero those coefficients that are small leading to a simpler model (sparsity) and it returns continuous estimators to avoid instability of the model.

It is easy to see that Cp, AIC and BIC can be written in the Penalized Least Squares form:

$$Cp: \quad \frac{1}{2}||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + \sigma^2\sum_{j=1}^{d} I(|\beta_j| \neq 0),$$

$$AIC: \quad \frac{1}{2}||\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}||^2 + n\frac{(\sigma\sqrt{2/n})^2}{2}\sum_{j=1}^{d} I(|\beta_j| \neq 0), \text{asymptotically},$$

$$BIC: \qquad \frac{1}{2}||\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}||^2 + n\frac{(\sigma\sqrt{log(n)/n})^2}{2}\sum_{j=1}^{d} I(|\beta_j| \neq 0), \text{asymptotically.}$$

Other criteria as AICc, CIC, RIC and others can also be considered in this class of estimators. These are called L_0 penalty estimators.

Hoerl and Kennard (1970) considered the L_2 type of penalty: $\frac{1}{2}||Y-\mathbf{X}\boldsymbol{\beta}||^2 + \frac{n\lambda}{2}||\boldsymbol{\beta}||^2$. The resulting estimator is $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X} + n\lambda I_d)^{-1}\mathbf{X}'Y$, and it is called the Ridge Regression estimator, which can be used for dealing with colinearity in predictors and outliers, but does not produce sparse estimators, and therefore should not be used for variable selection.

Frank and Friedman (1993) introduced the following estimator:

$$\frac{1}{2}||\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}||^2 + \frac{n\lambda}{q}\sum_{j=1}^{d}|\beta_j|^q,$$

it is called Bridge Regression estimator, and is referred to PLS with L_q penalty with $0 \leq q \leq 2$, bridging L_0 to $L2$.

Some of the most recent and most used methods are Least Absolute Shrinkage and Selection Operator (LASSO) by Tibshirani (1996), Adaptive LASSO by Zou (2006), Smoothly Clipped Absolute Deviation (SCAD) by Fan and Li (2001), Non-negative Garrote by Breiman (1995), Elastic Net by Zou and Hastie (2005), among others.

For Penalized Least Squares regression, it is common to standardize the covariates subtracting the mean and deviding by the standard deviation, so that the estimation of the intercept is not needed. LASSO is a special case of Bridge regressors: the estimators are the $\boldsymbol{\beta}$ that minimize $\frac{1}{2}||Y-\mathbf{X}\boldsymbol{\beta}||^2$ subject to $\sum_j |\beta_j| \leq s$, where s is a tunning parameter that can be found by cross-validation. It is equivalent to the penalized least squares with L_1 penalty $\frac{1}{2}||Y-\mathbf{X}\boldsymbol{\beta}||^2 + \lambda_n \sum_{j=1}^{d}|\beta_j|$, where λ_n is a tunning parameter that varies with n. The shrinkage applied to the estimates is controlled by the tunning parameter s: values of $t < \sum|\hat{\beta}_j^0|$, where β_j^0 is the full least squares estimates, will cause shrinkage of the solutions toward 0.

The LASSO procedure followed the Non-negative Garrote. The Non-negative Garrote works in the following way: the estimates of the coefficients are $\hat{\beta}_j(s) = c_j\tilde{\beta}_j(s)$, where $\tilde{\beta}_j(s)$ are the ordinary least squares estimates and c_j are the mini-

mizers of $\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{d} c_j \tilde{\beta}_j x_j)^2$ under the constraints $c_j \geq 0$ and $\sum_j c_j \leq s$. The garrote estimates depend heavily on the OLS estimates, signs and magnitude, and therefore may behave poorly whenever the OLS behave poorly (multicolinearity, overfitting).

The $\sqrt{n}$ consistency of the LASSO estimates is achieved when $\lambda_n = O(1/\sqrt{n})$, and the LASSO would have oracle properties when $\sqrt{n}\lambda_n \to \infty$, but these two conditions can never hold at the same time, therefore Fan and Li (2001) showed that the oracle property does not hold for the LASSO (actually L_1 penalty). To overcome this problem, Zou (2006) proposed the Adaptive LASSO, which is the solution of the weighted version of the LASSO:

$$\frac{1}{2}||Y - \mathbf{X}\boldsymbol{\beta}||^2 + n\sum_{j=1}^{d} w_j|\beta_j|,$$

where $\mathbf{w}$ is a known vector of weights. Zou showed that for the choice of weights $w_j = |\beta_{LS,j}|^{-\gamma}$, if $\lambda_n/\sqrt{n} \to 0$ and $\lambda_n n^{(\gamma-1)/2} \to \infty$, the Adaptive LASSO is consistent in variable selection (the probability of the selected set be the correct set tends to 1) and the estimates are asymptotically normal.

Fan and Li's SCAD improves the properties of the L_1 penalty and the hard thresholding penalty function (see Antoniadis 1997 and Fan 1997). Denoting $\mathbf{z} = \mathbf{X}'\mathbf{Y}$ and $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{X}'\mathbf{Y}$ we can write

$$\frac{1}{2}||Y - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda\sum_{j=1}^{d} p_j(|\beta_j|) = \frac{1}{2}||\mathbf{Y} - \hat{\mathbf{Y}}||^2 + \frac{1}{2}\sum_{j=1}^{d}(z_j - \beta_j)^2 + \lambda\sum_{j=1}^{d} p_j(|\beta_j|) \quad (5.1)$$

Since minimizing (5.1) is equivalent to minimizing componentwise, letting $p_\lambda(|.|) = \lambda p(|.|)$, consider the penalized least squares problem

$$\frac{1}{2}(z - \theta)^2 + p_\lambda(|\theta|).$$

With the hard thresholding penalty function $p_\lambda(|\theta|) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda)$, one can obtain the hard thresholding rule $\hat{\theta} = zI(|z| < \lambda)$. Note that the L_q penalty in this case $p_\lambda(|\theta|) = \lambda|\theta|^q$. Therefore, the SCAD continuous differentiable penalty

is defined by

$$p'_\lambda(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\}$$

for some $a > 2$ and $\theta > 0$.

The Elastic Net proposed by Zou and Hastie (2005) improves the LASSO when the number of covariates is larger the the number of observations, and the LASSO does not have good results. It is basically a convex combination of the LASSO and the ridge penalty, where one has to find the minimizer of the function

$$||Y - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda_2 \sum_{j=1}^{d} \beta_j^2 + \lambda_1 \sum_{j=1}^{d} |\beta_j|,$$

or equivalently, it can be seen as a penalized least squares problem by the optimization problem (let $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$):

$$\hat{\boldsymbol{\beta}} = argmin_{\boldsymbol{\beta}} ||Y - \mathbf{X}\boldsymbol{\beta}||^2 \text{ subject to } \alpha \sum_{j=1}^{d} \beta_j^2 + (1-\alpha) \sum_{j=1}^{d} |\beta_j| \leq t, \text{ for some } t.$$

A review of recent methods for variable selection needs to mention the LARS (Least Angle Regression) algorithm. Proposed by Efron, Hastie and Tibishrani (2004), it is a less greedy version of the famous and widely used forward regression. It basically starts with all coefficients set to zero and finds the predictor that has the highest correlation with the reponse; then it takes the largest step possible in the directions of that predictor until some other predictor has as much correlation with the current residual; after that it proceeds in a direction equiangular between the two predictors until another predictor is in the set of most correlated; etc. Modifications of this algorithm implements the LASSO or the Forward Stagewise methods.

At a conceptual level, however, the two problems are intimately connected: dropping variable j from the model is equivalent to not rejecting the null hypothesis $H_0^j : \beta_j = 0$. Abramovich, Benjamini, Donoho and Johnstone (2006) bridged the methodological divide by showing that application of the false discovery rate (FDR) controlling procedure of Benjamini and Hockberg (1995) on p values resulting from testing each H_0^j can be translated into minimizing a model selection criterion of

the form

$$\sum_{i=1}^{n}\left(Y_i-\sum_{j\in S}\widehat{\beta}_j^S x_{ij}\right)^2+\sigma^2|S|\lambda, \tag{5.2}$$

where S is a subset of $\{1,2,\ldots,d\}$ specifying the model, $\widehat{\beta}_i^S$ denotes the least squares estimator from fitting model S, $|S|$ is the cardinality of the subset S, and the penalty parameter λ depends both on d and $|S|$. This is similar to penalty parameters used in Tibshirani and Knight (1999), Birge and Massart (2001) and Foster and Stine (2004), which also depend on both d and $|S|$, and more flexible than the proposal in Donoho and Johnstone (1994) which uses λ depending only on d, as well as AIC and Mallow's C_p which use constant λ.

Working with orthogonal designs, Abramovich et al. (2006) showed that the global minimum of the penalized least squares (5.2) with the FDR penalty parameter is asymptotically minimax for ℓ^r loss, $0<r\leq 2$, simultaneously throughout a range of sparsity classes, provided the level q for the FDR is set to $q<0.5$. Generalizations of this methodology to non-orthogonal designs differ mainly in the generation of the p values for testing $H_0^j:\beta_j=0$, and the FDR method employed. Bunea, Wegkamp and Auguste (2006) use p values generated from the standardized regression coefficients resulting from fitting the full model and employ Benjamini and Yekuteli's (2001) method for controlling FDR under dependency, while Benjamini and Gavrilov (2009) use p values from a forward selection procedure where the ith stage p-to-enter is the ith stage constant in the multiple-stage FDR procedure in Benjamini, Krieger and Yekutieli (2006).

Model checking and variable selection procedures based on the assumption that the regression function is linear may fail to discern the relevance of covariates whose effect on $m(\mathbf{x})$ is nonlinear. For example, in Table ???4.2 the power of the F-test for testing, at level $\alpha=0.05$, the significance of a covariate with nonlinear (sinusoidal) effect is 0.051. Similarly, in the simulations reported in Tables ???**??** and ???5.4, the SCAD, LASSO and adaptive LASSO procedures fail to detect the contribution of covariates with nonlinear effects almost always. Because of this, procedures for both model checking and variable selection have been developed under more general/flexible models such as semiparametric models (cf. Li and

Liang, 2008), varying coefficient models (cf. Wang and Xia, 2008), Stone's (1985) additive model (Huang, Horowitz and Wei (2010) and references therein), and smoothing spline ANOVA models which expand the class of additive models by including interaction terms (Friedman, 1991, Zhang et al., 2004, Lin and Zhang, 2006, and Storlie et al. 2011).

However, the methodological approaches for variable selection under these more flexible models have been distinct from those of model checking. Our aim is at showing that a suitably flexible and powerful nonparametric model checking procedure can be used to construct a competitive nonparametric variable selection procedure by exploiting the aforementioned conceptual connection between model checking and variable selection.

5.2 Nonparametric variable selection using multiple testing: The Procedure

Suppose we have n observations of the random variable Y and covariates $\mathbf{X}$ with dimension $d > 1$. The nonparametric regression model is

$$Y_i = m(\mathbf{X}_i) + \varepsilon_i, i = 1...n, \tag{5.3}$$

where ε_{ni} is the independet error with zero mean and constant variance σ^2 and independent of $\mathbf{X}$.

The nonparametric model checking described in Chapter 4 can be used to contruct a nonparametric variable selection procedure by exploiting the conection explained above. Consider the hypothesis tests:

$$H_{0j} : m(\mathbf{x}) = m_1(\mathbf{x}_{(-j)}), j = 1, \ldots, d \tag{5.4}$$

where $\mathbf{x}_{(-j)} = (x_1, ..., x_{j-1}, x_{j+1}, ..., x_d)$.

Here, we follow Zambom and Akritas (2011): let $\varepsilon_i = Y_i - m_1(\mathbf{X}_{(-j)i})$ so that $\hat{\varepsilon}_i = Y_i - \hat{m}_1(\mathbf{X}_{(-j)i})$, where $\hat{m}_1(\mathbf{X}_{(-j)i})$ is the local polinomial estimator of order r of $m_1(\mathbf{X}_{(-j)i})$.

Assume a sparse regression model in the sense that there exits a sub-

set of indices $I_0 = \{j_1, \ldots, j_{d_0}\} \subset \{1, \ldots, d\}$ such that only the covariates X_j with $j \in I_0$ influence the regression function. Moreover, we will assume the dimension reduction Sliced Inversed Regression model of Li (1991), i.e. $m(\mathbf{x}) = g(\mathbf{Bx})$, where $\mathbf{B}$ is a $K \times d$ matrix. In this context we will describe the following variable selection procedure using backward elimination based on the Benjamini and Yekuteli (2001) method for controlling the false discovery rate (FDR):

1. Apply the variable screening procedure described in Section 4.2.2.4. With a slight abuse of notation, the vector of the remaining covariates and its dimension will be denoted by $\mathbf{x}$ and d.

2. Use SIR to obtain the estimator $\widehat{\mathbf{B}}$.

3. Obtain p-values from testing each of the hypotheses:

$$H_0^j : m(\mathbf{x}) = m_1(\mathbf{x}_{(-j)}),\ j = 1, \ldots, d, \tag{5.5}$$

   where $\mathbf{x}_{(-j)} = (x_1, ..., x_{j-1}, x_{j+1}, ..., x_d)$:

   (a) Compute the test statistic (see Theorem 4.2)

$$z_j = \sqrt{n}(MST_j - MSE_j)/\sqrt{\frac{2p(2p-1)}{3(p-1)}\hat{\tau}_j^2}$$

   using residuals formed by a kernel estimator on the variables $\widehat{\mathbf{B}}_{(-j)}\mathbf{x}_{(-j)}$, where $\widehat{\mathbf{B}}_{(-j)}$ is the $K \times (d-1)$ matrix obtained by omitting the jth column of $\widehat{\mathbf{B}}$.

   (b) Compute the p-value for H_0^j as $\pi_j = 1 - \Phi(z_j)$.

4. Compute

$$k = \max\left\{ j : \pi_{(j)} \leq \frac{j}{d}\frac{\alpha}{\sum_{l=1}^{d} l^{-1}} \right\} \tag{5.6}$$

   for a choice of level α, where $\pi_{(1)}, \ldots, \pi_{(d)}$ are the ordered p-values. If $k = d$ stop and retain all variables. If $k < d$

   (a) update $\mathbf{x}$ by eliminating the covariate corresponding to $\pi_{(d)}$,

(b) update $\widehat{\mathbf{B}}$ by eliminating the column corresponding to the deleted variable, and

(c) proceed to the next step.

5. Repeat steps 2 and 2b, with the updated vx and $\widehat{\mathbf{B}}$.

Remarks. 1) Another approach for constructing a variable selection procedure is to use a single application of the Benjamini and Yekuteli (2001) method for controlling the false discovery rate (FDR). This is similar to one of the two procedures proposed in Bunea et al. (2006). However, this did not perform well in simulations and is not recommended. A backward elimination approach was Li, Cook and Nachtsheim (2005), but they did not use multiple testing ideas.

2) Based on our simulation results, the variable screening part (Step 1) of the variable selection procedure does not improve the performance.

5.3 Simulations: Variable Selection

In this section we will study the behavior of the proposed variable selection procedure in comparison to the well known methods LASSO (Tibshirani, 1996), Adaptive LASSO (Zhou, 2006), SCAD (Fan and Li, 2001) and the test proposed by Bunea, Wegkamp and Auguste (2005) (BWA in the tables). For LASSO we found that the R code in in http://cran.r-project.org/web/packages/glmnet/index.html, with the lambda.lse option for selecting lambda, gave the best results; for adaptive LASSO we used the R code from http://www4.stat.ncsu.edu/∼boos/var.select /lasso.adaptive.html; for SCAD we used the function scadglm of the package SIS in R.

For the first simulation, we simulate 1000 data sets of size $n = 40$ of the linear model, for which the methods mentioned above are developed, with coefficients $\beta_1 = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and $\beta_2 = (3, 1.5, 0, 0, 0.1, 0, 0, 0)^T$. The error is normal with mean 0 and variance $\sigma^2 = 3$. Consider the covariates generate from

$$\mathbf{X} \sim N(\mathbf{0}, \mathbf{\Sigma}) \text{ where } \mathbf{\Sigma} = \mathbf{0.5^{|i-j|}}, \mathbf{i}, \mathbf{j} = \mathbf{1}, \ldots, \mathbf{8}. \tag{5.7}$$

We included in the table the simple proposed ANOVA-type procedure, the

Table 5.1. Number of coefficients set to 0

β_1		
	Correct	Incorrect
SCAD	4.31	0.27
LASSO	3.80	0.1
Adaptive LASSO	4.38	0.3
BWA	4.77	0.52
ANOVA-type (p = 7)	4.74	1.36
ANOVA-type+SIR (p = 7)	4.69	1.37
ANOVA-type+Partial Corr.+SIR (p = 7)	4.71	1.52
ANOVA-type (p = 5)	4.70	1.33
ANOVA-type+SIR (p = 5)	4.66	1.35
ANOVA-type+Partial Corr.+SIR (p = 5)	4.69	1.48
ANOVA-type (p = 3)	4.66	1.41
ANOVA-type+SIR (p = 3)	4.63	1.35
ANOVA-type+Partial Corr.+SIR (p = 3)	4.70	1.69
β_2		
	Correct	Incorrect
SCAD	4.77	1.16
LASSO	4.41	1.0
Adaptive LASSO	4.61	1.24
BWA	4.81	1.32
ANOVA-type (p = 7)	4.81	1.78
ANOVA-type+SIR (p = 7)	4.76	1.71
ANOVA-type+Partial Corr.+SIR (p = 7)	4.82	1.70
ANOVA-type (p = 5)	4.74	1.71
ANOVA-type+SIR (p = 5)	4.73	1.72
ANOVA-type+Partial Corr.+SIR (p = 5)	4.75	1.73
ANOVA-type (p = 3)	4.72	1.72
ANOVA-type+SIR (p = 3)	4.71	1.71
ANOVA-type+Partial Corr.+SIR (p = 3)	4.66	1.67

ANOVA-type using SIR to estimate the null regression function and te ANOVA-type using the Partial Correlations technique described in the Dimension Reduction Section. In all cases, the ANOVA-type uses backward elimination.

Table 5.1 shows the number of correct and incorrect coefficients set to 0. The methods designed for the linear model have comparable power in identifying the important covariates, but the ANOVA-type selects more predictors incorrectly.

Next, we study the behavior of these procedures for nonlinear regression func-

Table 5.2. Number of coefficients set to 0

$\sigma = .3$		
	correct	incorrect
SCAD	6.76	0.94
LASSO	6.75	0.95
Adaptive LASSO	6.66	0.94
BWA	6.64	0.94
ANOVA-type (p=7)	6.84	0.005
ANOVA-type+SIR (p=7)	6.84	0
ANOVA-type+Partial Corr.+SIR (p=7)	6.82	0
ANOVA-type (p=5)	6.83	0
ANOVA-type+SIR (p=5)	6.76	0
ANOVA-type+Partial Corr.+SIR (p=5)	6.77	0
ANOVA-type p=3)	6.75	0
ANOVA-type+SIR p=3)	6.7	0
ANOVA-type+Partial Corr.+SIR p=3)	6.72	0
$\sigma = 1$		
	correct	incorrect
SCAD	6.77	0.95
LASSO	6.75	0.964
Adaptive LASSO	6.65	0.95
BWA	6.67	0.946
ANOVA-type (p=7)	6.74	0.23
ANOVA-type+SIR (p=7)	6.66	0.26
ANOVA-type+Partial Corr.+SIR (p=7)	6.69	0.21
ANOVA-type (p=5)	6.69	0.25
ANOVA-type+SIR (p=5)	6.59	0.25
ANOVA-type+Partial Corr.+SIR (p=5)	6.59	0.23
ANOVA-type (p=3)	6.58	0.29
ANOVA-type+SIR (p=3)	6.53	0.25
ANOVA-type+Partial Corr.+SIR (p=3)	6.53	0.26

tions. Table 5.2 shows the results 1000 data sets of size $n = 40$ for the non linear model $Y = sin(\pi X_1) + \epsilon$, where $\epsilon \sim N(0, \sigma)$ and the covariates are generated as in (5.7).

For the sine curve, the linear-based methods fail to identify the only predictor that is significant. In this case, they end up selecting almost no covariate, while the ANOVA-type method correctly sets most of the insignificant covariates to 0, and incorrectly sets to 0 only when the variance is too large.

Table 5.3. Number of coefficients set to 0

	correct	incorrect
SCAD	5.78	1.95
LASSO	5.87	1.97
Adaptive LASSO	5.75	1.95
ANOVA-type	5.8	.7
ANOVA-type+SIR	5.83	.44
ANOVA-type+Partial Corr.	5.87	.606

Table 5.3 shows the results for the simulation of 1000 data sets of size $n = 60$ of the model $Y = sin(3/4\pi X_1) - 3\Phi(-|X_5|) + \epsilon$, where ϵ is normal with mean 0 and variance 0.1^2 and the covariates are as in (5.7).

The ANOVA-type procedure (with $p = 7$ in this table) again outperforms the linear methods by not setting to zero the important covariates in most of the time.

For a more difficult setup to the linear based methods, we simulate data sets of size $n = 40$ from the model $Y = sin(3/4\pi X_1) - 3\Phi(-|X_5|^3) + \epsilon$, where ϵ is normal with mean 0 and variance 0.1^2 and the covariates are generated as in (5.7). The results are shown in Table 5.4, where we also included a modification of the method proposed by Bunea et al., by using the backward elimination. The

Table 5.4. Number of coefficients set to 0

	correct	incorrect
SCAD	5.67	1.75
LASSO	5.73	1.79
Adaptive LASSO	5.66	1.75
BWA	5.99	1.99
BWA - Backward Elimination	5.72	1.71
ANOVA-type (p=7)	5.88	0.04
ANOVA-type+SIR (p=7)	5.78	0.06
ANOVA-type+Partial Corr.+SIR (p=7)	5.8	0.08
ANOVA-type (p=5)	5.82	0.03
ANOVA-type+SIR (p=5)	5.72	0.04
ANOVA-type+Partial Corr.+SIR (p=5)	5.76	0.06
ANOVA-type (p=3)	5.75	0.04
ANOVA-type+SIR (p=3)	5.67	0.065
ANOVA-type+Partial Corr.+SIR (p=3)	5.72	0.05

ANOVA-type method again identifies the two important predictors, non linear

regression function, and does not set them to zero, what happens most of the time for the linear based procedures.

In Table 5.5, data sets of size $n = 110$ were generated from the linear model $Y = \boldsymbol{\beta}^T\mathbf{X} + \epsilon$, where $\epsilon \sim N(0, 3^2)$, the dimension of $\mathbf{X}$ is $d = 25$, and

$$\boldsymbol{\beta}^T = (3, 1.5, 0, 0, 2, 0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0, 0, 0).$$

The covariates are generated from a multivariate normal distribution with marginal means zero and covariances as shown in the table. It is seen that the proposed nonparametric variable selection procedures correctly exclude, on average, about 19.5 out of the 20 nonsignificant predictors. This is about as good as the procedures designed for linear models. The proposed procedures incorrectly exclude, on average, about 0.5 of the 5 significant predictors, which is more than the other procedures (with the exception of BWA).

Table 5.5. Number of coefficients set to 0

	$\Sigma = I$		$\Sigma = (0.5^{\lvert i-j \rvert})$	
test	correct	incorrect	correct	incorrect
SCAD	19.48	.026	19.37	.023
LASSO	18.29	.005	18.28	.004
Adaptive LASSO	19.28	.005	19.26	.025
BWA	19.99	1.02	19.97	1.41
BWA+BE	19.55	.001	19.49	.041
ANOVA-type(p=7)	19.46	.63	19.30	.44
ANOVA-type(p=9)	19.52	.65	19.40	.36

5.4 Real Data Example: Body Fat Dataset

The Body Fat data is supplied by Dr. A. Garth Fisher for non-commercial purposes, and it can be found at "http://lib.stat.cmu.edu/datasets/bodyfat". The data set contains measurements of percent body fat (using Siri's (1956) method), Age (years), Weight (lbs), Height (inches), circunferences of Neck (cm), Chest (cm), Abdomen (cm), Hip (cm), Thigh (cm), Knee (cm), Ankle (cm), Biceps (cm), Forearm (cm) and Wrist (cm), from 252 men. The response variable is the percentage of body fat.

We compare the results of SCAD, LASSO, Adaptive LASSO and BWA with backward elimination to the ANOVA-type procedure with variable screening and SIR, where screening consists of performing the marginal test of Wang, Akritas and Van Keilegom (2008) for the significance of each variable, and keeping those variables for which the p-value is less than 0.5. Table 5.6 shows the results for LASSO, SCAD, Adaptive LASSO and BWA.

Table 5.6. Results for LASSO, Adaptive LASSO, SCAD, BWA

Predictor	LASSO	Adpt. LASSO	SCAD	BWA
Age	.06499	0	.001061	0
Weight	0	-.09511	-.11688	-.1356
Height	-.1591	0	-.05818	0
Neck	-.2579	0	0	0
Chest	0	0	0	0
Abdomen	.7079	.9113	.9052	.9958
Hip	0	0	0	0
Thigh	0	0	0	0
Knee	0	0	0	0
Ankle	0	0	0	0
Biceps	0	0	0	0
Forearm	.21756	0	0	.4729
Wrist	-1.5353	-.9871	0	-1.5056

It is seen that Weight and Abdomen are selected by all except LASSO. This can be explained by the fact that LASSO does not perform well in the presence of highly correlated variables, which is the case with this data set. The Adaptive LASSO and BWA give almost the same results but differ considerably from those of SCAD.

For the ANOVA-type method we used SIR with the number of slices ranging from 2 to 100. Abdomen, Weight, Biceps and Knee were selected with 99, 87, 88 and 23 of the 99 different numbers of slices, respectively. All other variables were selected less than 15 times. On the basis of these results we recommend a model based on Abdomen, Weight and Biceps.

As an explanation of the fact that Biceps was not selected by any of the other methods, we investigated possible violations of the modeling assumptions on which they are based. Marginal plots of the response versus each of the important variables reveal both hetoroscedasticity and nonlinearity. Moreover, the 99 applications of SIR yielded more than one linear combination (i.e. $K > 1$) 50 times. To put this number into perspective, we generated a single set of responses, using the same covariate values with coefficients those from Adaptive LASSO and normal errors using the residual variance. Application of SIR with the number of slices ranging from 2 to 100 on this data set yielded $K = 1$ 92 out of the 99 times. This casts serious doubts on the validity of the assumption of a linear model.

5.5 Nonparametric Group Variable Selection using multiple testing: The Procedure

Suppose we have n observations of the random variable Y and covariates $\mathbf{X}$, and that the covariates come in d groups indexed by J, where $J_i = \{j : X_j \text{ belongs to group i}\}, i = 1, \ldots, d$, and let s_i denote the size of group J_i. Consider the nonparametric regression model

$$Y_i = m(\mathbf{X}_i) + \varepsilon_i, i = 1...n, \tag{5.8}$$

where ε_i is the independet error with zero mean and constant variance σ^2 and independent of $\mathbf{X}$.

The nonparametric model checking for groups described in Chapter 4 can be used to contruct a nonparametric group variable selection procedure by exploiting the conection explained above. Consider the hypothesis tests:

$$H_{0i} : m(\mathbf{x}) = m_1(\mathbf{x}_{(-J_i)}), i = 1, \ldots, d \tag{5.9}$$

where $\mathbf{x}_{(-J_i)}$ is the set of all covariates except those whose index are in J_i.

Here, we follow Zambom and Akritas (2011): let $\varepsilon_k = Y_k - m_1(\mathbf{X}_{(-J_i)k})$ so that $\hat{\varepsilon}_k = Y_k - \hat{m}_1(\mathbf{X}_{(-J_i)k})$, where $\hat{m}_1(\mathbf{X}_{(-J_i)k})$ is the local polinomial estimator of order r of $m_1(\mathbf{X}_{(-J_i)k})$.

Assume a sparse regression model in the sense that there exits a subset of group indices $I_0 = \{J_1, \ldots, J_{d_0}\} \subset \{J_1, \ldots, J_d\}$ such that only groups of covariates X_J with $J \in I_0$ influence the regression function. Moreover, we will assume the dimension reduction Sliced Inversed Regression model of Li (1991), i.e. $m(\mathbf{x}) = g(\mathbf{B}\mathbf{x})$, where $\mathbf{B}$ is a $K \times (\sum s_i)$ matrix. In this context we will describe the following group variable selection procedure using backward elimination based on the Benjamini and Yekuteli (2001) method for controlling the false discovery rate (FDR):

1. Use SIR to obtain the estimator $\widehat{\mathbf{B}}$.

2. Obtain p-values from testing each of the hypotheses:

$$H_0^i : m(\mathbf{x}) = m_1(\mathbf{x}_{(-J_i)}),\ i = 1, \ldots, d, \tag{5.10}$$

   where $\mathbf{x}_{(-J_i)}$ is the set of all covariates except those whose index are in J_i.

   (a) Compute the test statistic (see Theorem 4.4)

$$z_i = \sqrt{n}(MST_i - MSE_i)/\sqrt{\frac{2p(2p-1)}{3(p-1)}\hat{\sigma}_j^4}$$

   using the residuals formed by a kernel estimator on the variables $\widehat{\mathbf{B}}_{(-J_i)}\mathbf{x}_{(-J_i)}$, where $\widehat{\mathbf{B}}_{(-J_i)}$ is the $K \times (d - s_i)$ matrix obtained by omitting the J_ith columns of $\widehat{\mathbf{B}}$.

   (b) Compute the p-value for H_0^i as $\pi_i = 1 - \Phi(z_i)$.

3. Compute

$$k = \max\left\{ i : \pi_{(i)} \leq \frac{i}{d}\frac{\alpha}{\sum_{\ell=1}^{d} \ell^{-1}} \right\} \tag{5.11}$$

for a choice of level α, where $\pi_{(1)}, \ldots, \pi_{(d)}$ are the ordered p-values. If $k = d$ stop and retain all variables. If $k < d$

(a) update $\mathbf{x}$ by eliminating the covariates corresponding to $\pi_{(d)}$,

(b) update $\widehat{\mathbf{B}}$ by eliminating the columns corresponding to the deleted variables, and

(c) proceed to the next step.

4. Repeat steps 2 and 2b, with the updated $\mathbf{x}$ and $\widehat{\mathbf{B}}$.

Remark 1 Another approach for constructing a variable selection procedure is to use a single application of the Benjamini and Yekuteli (2001) method for controlling the false discovery rate (FDR). This is similar to one of the two procedures proposed in Bunea et al. (2006). However, this did not perform well in simulations and is not recommended. A backward elimination approach was Li, Cook and Nachtsheim (2005), but they did not use multiple testing ideas.

5.5.1 Simulations: Group Variable Selection Procedure

In this section we compare the variable selection based on the ANOVA-type test to the Group Lasso proposed by Yuan and Lin (2006). We study the behavior of the selection for two different scenarios, one with a continuous response and another with a binary response.

For the continuous response scenario the data is generated according to the models

$$\begin{aligned}
\text{Model 1} : Y &= X_3^3 + X_3^2 + X_3 + (1/3)X_6^3 - X_6^2 + (2/3)X_6 + \epsilon \\
\text{Model 2} : Y &= sin(X_3^3 + X_3^2 + X_3) + (1/3)X_6^3 - X_6^2 + (2/3)X_6 + \epsilon \\
\text{Model 3} : Y &= 10sin(X_3^3 + X_3^2 + X_3) + 5sin((1/3)X_6^3 - X_6^2 + (2/3)X_6) + \epsilon
\end{aligned}$$

where $X_i = (Z_i + W)/\sqrt{2}$, $Z_i, i = 1, \ldots, 16$ and W iid $N(0,1)$, and $\epsilon \sim N(0, 2^2)$. Thus, for Models 1, 2 and 3 there are 16 groups of three covariates each, represented by the polynomial terms. The only groups that are significant are groups 3 and 6. We run 1000 simulations of data sets of size $n = 100$. Table 5.7 shows the

Table 5.7. Results for the ANOVA-type and Group Lasso

Model	Method	Corr.Selected	Incorr.Selected
Model 1	ANOVA-type(9)	1.80	.55
	Group LASSO	2	4.7
Model 2	ANOVA-type(9)	1.15	.81
	Group LASSO	1.59	4.21
Model 3	ANOVA-type(9)	1.84	0.64
	Group LASSO	1.80	6.75

mean number of correct and incorrect groups selected by the ANOVA-type variable selection and Group Lasso using the C_p criterion.

For the second scenario, we consider the following three logistic regression models.

$$\text{Models 1 and 2}: \quad p_j(\mathbf{X}) = \frac{1}{1 + exp(-\boldsymbol{\beta}_j^T (1, \mathbf{X})^T)}, \ j = 1, 2,$$

where $\mathbf{X} = (X_1, \ldots, X_{15})$ are iid $U(0, 1)$, grouped sequentially in 5 groups of 3 covariates each, and

$$\begin{aligned} \boldsymbol{\beta}_1 &= (1, -2.2, 2, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 0, 0, 0, 0)^T \\ \boldsymbol{\beta}_2 &= (1, -2.2, 3, 0, 0, 0, 0, 0, 0, 0, 0, 1, 3, 0, 0, 0, 0)^T. \end{aligned}$$

$$\text{Model 3}: \quad p_3(\mathbf{X}) = \frac{1}{1 + exp\left(-18 \sin(\pi X_2) - 18 \sin(\pi X_8)\right)},$$

where $\mathbf{X} = (X_1, \ldots, X_{12})$, with $X_1, \ldots, X_{11}$ iid $U(0, 3)$ and $X_{12} \sim N(-3, 1)$ independent of the others, are grouped sequentially in 4 groups of 3 covariates each.

The results in Table 5.8 are based on 1000 simulation runs using $n = 100$ for Models 1 and 2, and $n = 200$ for Model 3. It is seen that for Models 1 and 2 the number of correctly selected covariates by either procedure is low. This is probably due to the smaller sample size and the larger number of covariates. For Model 3, the Group Lasso fails to select covariates, while the ANOVA-type procedure seems to perform very well. In summary, the simulation results suggest that the ANOVA-type variable selection procedure outperforms the Group Lasso

Table 5.8. Results for logistic regression

Model	Method	Corr.Selected	Incorr.Selected
Model 1	ANOVA-type(9)	.340	.261
	Group LASSO	.197	.032
Model 2	ANOVA-type(9)	.287	.312
	Group LASSO	.100	.021

Table 5.9. Results for non-linear logistic regression

Model	Method	Corr.Selected	Incorr.Selected
Model 3	ANOVA-type(9)	1.223	0.080
	Group LASSO	0.040	0.039

when the logistic regression model involves a non-linear function of the covariates, and has competitive performance in the other cases.

Chapter 6

A Regression-Type Statistic for Hypothesis Testing in Nonparametric Regression

It is of great interest in this thesis to find ways of testing the significance of covariates in nonparametric regression without making strong assumptions on the nonparametric regression model. As in Chapter 4, we want to develop a testing hypothesis without assuming additivity or homocedasticity.

In this chapter we will develop a new hypothesis test for the simple nonparametric regression model with only one covariate, and in future work we will try to extend this result for higher dimensions. The model is allowed to have heterocedasticity, and as in Chapter 4 we assume we have n observations of the response Y and the covariate X and that the density of X has bounded support. The model then is $m(X) = E(Y|X)$ with

$$Y_i = m(X_i) + \sigma(X_i)\epsilon_i \tag{6.1}$$

for a variance function $\sigma^2(X)$ such that $\sup_{x \in S_{X_1}} \sigma^2(x) < M < \infty$, and ϵ_i independent error with zero mean and constant variance independent of X.

The goal is to test the significance of X on the regression, that is, the null hypothesis is

$$H_0 : m(x_i) = \mu. \tag{6.2}$$

It is clear that under the null hypothesis $Y_i = \mu + \sigma(X_i)\epsilon_i$, so that estimating the regression function is the same as estimating the mean

$$\hat{\mu} = \bar{Y}.$$

In a simple linear regression model

$$Y_i = \beta_0 + \beta_1 W_i + \gamma_i, \tag{6.3}$$

where γ_i has mean 0 and constant variance, the estimated slope is

$$\hat{\beta} = \frac{\sum W_i(Y_i - \bar{Y})/N}{\sum (W_i - \bar{W})^2/N} = \frac{\mathrm{Cov}(Y, W)}{Var(W)}. \tag{6.4}$$

Note that the denominator is a function of the variance of W and the numerator is the covariance of Z and W. Setting $W_i = m(X_i)$, equation 6.3 with $\beta_1 = 1$ and $\beta_0 = 0$ becomes model 6.1. Thus we propose a test statistic based on

$$\lambda_n(H_0) = \sum_{i=1}^{n} \hat{m}(X_i)\,(Y_i - \hat{\mu}) = \sum_{i=1}^{n} \hat{m}(X_i)\left(Y_i - \bar{Y}\right), \tag{6.5}$$

which is a regression/correlation relation between $m(X)$ and Y.

Without loss of generality, consider the observations ordered in the X variable (and Y accordingly). We will consider here the simple case where we estimate the regression function at an observed x using the nearest neighbor technique without using the point itself, so that

$$\hat{m}(X_i) = \frac{1}{k_n} \sum_{j \neq i}^{n} Y_j \mathbf{1}(|X_i - X_j| \leq |X_i - X_{(k_n)}|), \tag{6.6}$$

where $X_{(k_n)}$ is the k_n nearest neighbor of X_i. Equivalently, we could use the $k_n/2$ neighbors of each side of X_i, in this case

$$\hat{m}(X_i) = \frac{1}{k_n} \sum_{j \neq i}^{n} Y_j \mathbf{1}(|i - j| \leq k_n/2), \tag{6.7}$$

therefore estimating $m(\hat{X}_i)$ with the $k_n/2$ neighbors to the right and to the left of

X_i, and the asymptotics will be the same.

The following theorem formalizes the asymptotic distribution of our test statistic:

Theorem 6.1. *Assume $\sigma(x)$ is Lipschitz continuous, $E(Y_i^4)$ are uniformly bounded in n and i. Then, under H_0 in (6.2)*

$$\sqrt{h_n}\lambda_n(H_0) \to^d N(0, \gamma^2),$$

where $\gamma^2 = 2E(\sigma^4(X))$.

Proof of Theorem 6.1

Proof. In this proof, we will make use of the simple version of the nearest neighbor estimator, where we consider $k_n/2$ neighbors on each side of X_i, but the same proof is valid for the usual nearest neighbor method. Consider the ordered X vector $X_{(1)}, X_{(2)}, \ldots X_{(n)}$, and the corresponding Y (we will omit the subscript). In order to derive the asymptotic distribution, centralizing the variables by subtracting the mean under the null will lead to

$$\begin{aligned}
\lambda_n(H_0) &= \sum_{i=1}^{n} \hat{m}(X_i)\left(Y_i - \bar{Y}\right) = \sum_{i=1}^{n}\left(\frac{1}{k_n}\sum_{j\neq i}^{n} Y_j \mathbf{1}(|i-j| \leq k_n/2)\right)\left(Y_i - \bar{Y}\right) \\
&= \frac{1}{k_n}\sum_{i=1}^{n}\sum_{j\neq i}^{n} Y_j(Y_i - \bar{Y})\mathbf{1}(|i-j| \leq k_n/2) \\
&= \frac{1}{k_n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(Y_j - \mu + \mu)(Y_i - \mu - (\bar{Y} - \mu))\mathbf{1}(|i-j| \leq k_n/2) \\
&= \frac{1}{k_n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(Y_j - \mu)(Y_i - \mu)\mathbf{1}(|i-j| \leq k_n/2) \\
&\quad -\frac{1}{k_n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\mu(\bar{Y} - \mu)\mathbf{1}(|i-j| \leq k_n/2) \\
&\quad -\frac{1}{k_n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(Y_j - \mu)(\bar{Y} - \mu)\mathbf{1}(|i-j| \leq k_n/2)
\end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{k_n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\mu(Y_i-\mu)\mathbf{1}(|i-j|\leq k_n/2)\\
=\;& \frac{1}{k_n}S_n-\frac{1}{k_n}\mu(\bar{Y}-\mu)n(k_n-1)\\
&-\left[\frac{(\bar{Y}-\mu)^2nk_n}{k_n}-\frac{(\bar{Y}-\mu)n(\bar{Y}-\mu)}{k_n}\right]+\frac{1}{k_n}\mu n(\bar{Y}-\mu)(k_n-1)\\
=\;& \frac{1}{k_n}S_n-\left[\frac{(\bar{Y}-\mu)^2nk_n}{k_n}-\frac{(\bar{Y}-\mu)n(\bar{Y}-\mu)}{k_n}\right]. \qquad (6.8)
\end{aligned}
$$

Note that we have approximately k_n neighbors for each Y_i using the simple nearest neighbor method, and asymptotically $k_n \sim nh_n$ for a bandwidth h_n.
Let $b_n \sim n^{2/3}(2k_n)^{1/3}, l_n \sim k_n$ and $r_n \sim b_n^{-1}n \sim n^{1/3}k_n^{-1/3}$ and assume n is a multiple of $b_n + l_n$ and let

$$S_n = \sum_{i=1}^{r_n} U_{ni} + \sum_{i=1}^{r_n} V_{ni} \qquad (6.9)$$

where

$$
\begin{aligned}
U_{ni} &= A_{(i-1)(b_n+l_n)+1} + \ldots + A_{(i-1)(b_n+l_n)+b_n},\\
V_{ni} &= A_{(i-1)(b_n+l_n)+b_n+1} + \ldots + A_{i(b_n+l_n)},
\end{aligned}
$$

and $A_i = \sum_{j\neq i}^{n}(Y_j-\mu)(Y_i-\mu)\mathbf{1}(|i-j|\leq k_n/2)$.

As in the paper of Wang, Ingrid, Akritas

$$
\begin{aligned}
&E(S_n^4) \qquad (6.10)\\
=\;& \sum_{i_1=1}^{n}\ldots\sum_{i_4=1}^{n}\sum_{j_1\neq i_1}^{n}\ldots\sum_{j_4\neq i_4}^{n} E\left((Y_{i_1}-\mu)\ldots(Y_{i_4}-\mu)(Y_{j_1}-\mu)\ldots(Y_{j_4}-\mu)\right)\\
&\mathbf{1}(|i_\ell - j_\ell|\leq k_n/2, \ell = 1\ldots 4). \qquad (6.11)
\end{aligned}
$$

For the null hypothesis of interest $H_0 : m(x) = \mu$, the nonzero terms in $E(S_n^4)$ are of one of the following forms: $E((Y_i-\mu)^4(Y_j-\mu)^4)$ which is of order $O(nk_n)$, $E((Y_i-\mu)^3(Y_j-\mu)^3(Y_k-\mu)^2)$ of order $O(nk_n^2)$, $E((Y_i-\mu)^4(Y_j-\mu)^2(Y_k-\mu)^2)$ of order $O(nk_n^2)$ and $E((Y_i-\mu)^2(Y_j-\mu)^2(Y_k-\mu)^2(Y_l-\mu)^2)$ of order $O(n^2k_n^2)$.

Therefore, $E(S_n^4) \leq B'n^2k_n^2$ for some positive constant B'. In a similar way, $E(V_{ni})$ is of the order $O(l_n^2k_n^2)$ and $E(V_{ni}^4) \leq Bl_n^2k_n^2$ for some constant B.

Next, note that

$$P\left(\frac{1}{\sqrt{nk_n}}|\sum_{i=1}^{r_n} V_{ni}| \geq \epsilon\right) \leq \sum_{i=1}^{r_n} P\left(|V_{ni}| \geq \epsilon\sqrt{nk_n}r_n^{-1}\right)$$
$$\leq \frac{B}{\epsilon^4}\frac{r_n^5}{n^2k_n^2}(l_nk_n)^2 = O(\frac{r_n^5k_n^2}{n^2}) = O(k_n^{1/3}n^{-1/3}) = o(1), \tag{6.12}$$

so that $\sum_{i=1}^{r_n} V_{ni} = o_p((nk_n)^{-1/2})$.

For the statistic λ_n to converge in distribution, we only need to show that the U_{ni} converge in distribution. By construction, U_{ni} are independent and $E(U_{ni}^2) \sim O(b_nk_n)$ and $E(U_{ni}^4) \sim O((b_nk_n)^2)$, so that

$$\sum_{i=1}^{r_n} \frac{E(U_{ni}^4)}{\sum_{j=1}^{r_n} E(U_{nj}^2)} \sim O(r_n\frac{b_n^2k_n^2}{r_n^2b_n^2k_n^2}) = O(r_n^{-1}) = o(1) \tag{6.13}$$

and the Lyapunov condition is satisfied, hence $\frac{1}{\sqrt{nk_n}}S_n \to N(0, variance)$.

Therefore

$$\begin{aligned}\sqrt{h_n}\lambda_n(H_0) &= \frac{\sqrt{h_n}}{k_n}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(Y_j-\mu)(Y_i-\mu)\mathbf{1}(|i-j| \leq k_n/2) \\ &\quad -\sqrt{h_n}\left[\frac{(\bar{Y}-\mu)^2nk_n}{k_n} - \frac{(\bar{Y}-\mu)n(\bar{Y}-\mu)}{k_n}\right] \to^d N(0,\gamma^2),\end{aligned} \tag{6.14}$$

since $\sqrt{n}(\bar{Y}-\mu) \to^d N(0,.)$ so that the second term goes to 0 in probability and

γ^2 is the asymptotic variance of the test statistic, and because $E(S_n) = 0$, it can be written as

$$\begin{aligned}\gamma_n^2 &= \frac{1}{nk_n}E(S_n^2) \\ &= \frac{1}{nk_n}E\sum_{i_1,i_2}\sum_{j_1\neq i_1, j_2\neq i_2}(Y_{i_1}-\mu)(Y_{i_2}-\mu)(Y_{j_1}-\mu)(Y_{j_2}-\mu) \\ &\quad \mathbf{1}(|i_1-j_1| \leq k_n/2, |i_2-j_2| \leq k_n/2)]\end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{nk_n} E\left[\sum_i \sum_{j\neq i} (Y_i-\mu)^2 (Y_j-\mu)^2 \mathbf{1}(|i-j|\leq k_n/2)\right] \\
&= \frac{2}{nk_n} \sum_i \sum_{j\neq i} \sigma^2(X_i)\sigma^2(X_j)\mathbf{1}(|i-j|\leq k_n/2) \\
&= \frac{2}{nk_n} \sum_i \sum_{j\neq i} \sigma^2(X_i)(\sigma^2(X_i)+O(\frac{k_n}{n}))\mathbf{1}(|i-j|\leq k_n/2) \\
&= \frac{2}{nk_n} \sum_i \sum_{j\neq i} \sigma^4(X_i)\mathbf{1}(|i-j|\leq k_n/2) + O(\frac{k_n n k_n}{n n k_n}) \\
&= \frac{2(k_n-1)}{nk_n} \sum_i \sigma^4(X_i) + o(1) \to 2E(\sigma^4(X)) = \gamma^2, \qquad (6.15)
\end{aligned}
$$

completing the proof. □

6.1 Simulations: A comparison between the Correlation/Regression-type statistic with the one of Wang, Akritas and Keilegom

In this section we will compare our correlation/regression-type statistics to the one proposed in Wang, Akritas and Keilegom and with Munk's test (2000). Note that their statistic is valid only in the case where the values of X are of a fixed design, while our test statistic is valid for random X.

The setting of the simulation is the following: with a sample size of $n = 60$, we simulate 2000 replications of the model

$$Y = m_j(x) + \epsilon,\ j = 0, 1, 2, 3, \qquad (6.16)$$

where $\epsilon \sim N(0,1)$ independend random errors, and x is a fixed equally spaced design on (0,1]. There are three different alternatives to consider when looking at the power: $m_0(x) = c$ (Null hypothesis), $m_1(x) = 2x$, $m_2(x) = 64x^3(1-x)^3$ and $m_4(x) = (1 + sin(3\pi x))/2$. Table 6.1 shows the proportion of rejections for these four functions.

Note In parenteses is the rejection rate we found when we simulate their statistic, it seems that their simulation does not compute the right power, it is in fact

Table 6.1. Proportion of rejections for Null and alternatives

test		m_0	m_1	m_2	m_3
WAK	$k_n = 7$	.05	.87(.825)	.45(.40)	.46(.39)
	$k_n = 9$	.04	.87(.83)	.45(.41)	.47(.41)
	$k_n = 11$	.04	.86(.815)	.43(.38)	.46(.41)
Munk	r = 2	.07	.78	.40	.32
	r = 4	.05	.83	.52	.36
	r = 6	.04	.82	.56	.35
	r = 8	.03	.80	.62	.31
our		.05	.82	.455	.352
F-test	.05	.99	.038	.039	

a little lower. For WAK, k_n is the size of the window created in the ANOVA-type statistic, and for Munk's test, r is the band size of the matrix.

Our test statistic seems to have a resonable competitive performance with resonable power at these alternatives.

Chapter 7

Future Work

7.1 Nonparametric variable selection using multiple testing: Asymptotic Properties

In Chapter 5, a variable selection procedure based on the ANOVA-type hypothesis test was proposed. This procedure uses the idea of multiple testing corrections for controlling the false discovery rate. In summary, a modified backward elimination algorithm was used where the false discovery rate was controlled in each step.

For the linear regression model, Bunea, Wegkamp and Auguste (2006) view the variable selection as the problem of estimating an index set I_0 containing the indices of the coefficients that are not equal to zero. With this idea, we outline here a discussion on how to prove the consistency of the variable selection method proposed in Chapter 5.

For higher dimensions, Munk et. al (2005) suggest the following estimator of σ^2

$$\frac{1}{\nu}S^2 = \frac{1}{\nu}\sum_{l=1}^{n}\left(Y_l - \sum_{k=1}^{n} w_{lk}Y_k\right)^2,$$

where $\nu = n - 2\sum_l w_{ll} + \sum_{l,k} w_{lk}^2$ and $w_{lk} = \left(\frac{K_{H_n}(\mathbf{X}_l - \mathbf{X}_k)}{\sum_{\ell=1}^{n} K_{H_n}(\mathbf{X}_l - \mathbf{X}_\ell)}\right)$ (see also Hall and Marron (1990)). While this is a consistent estimator for higher dimensions, the usual estimator based on the differences is not consistent for more than 4 dimensions. Assume that the regression function m has $r+1$ derivatives, the kernel

is d-dimensional or order r and that the assumptions of Theorem 4.3 hold. The variable selection problem in this case, is equivalent to identifying/selecting the index set $I_0 \in \{1, \ldots, d\} \stackrel{def}{=} I_d$ corresponding to the variables that are significant in the nonparametric regression.

Denote the p-value of each test $j = 1, \ldots, d$ in (5.10) by $\pi_j = 1 - \Phi(z_j)$, for $z_j = \frac{\sqrt{n}(MST - MSE)}{\sqrt{\frac{2p(2p-1)}{3(p-1)}S^2}}$. Note that MST and MSE depend on j, i.e., on the hypothesis we are testing, but we will ommit the subscript. The FDR procedure of Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001) consider the ordered p-values $\pi_{(1)} \leq \ldots \leq \pi_{(d)}$ and compute

$$k = \max \left\{ j : \pi_{(j)} \leq \frac{i}{d} \frac{q}{\sum_{l=1}^{d} l^{-1}} \right\} \tag{7.1}$$

for a choice of level q, and reject all $H_{0(j)}, j = 1, \ldots, k$ where $H_{0(j)}$ is the null hypothesis corresponding to the ordered p-value $\pi_{(j)}$. If no such k exists, do not reject any hypothesis.

Consider the estimator $\hat{I}$ of I_0 defined by the indices corresponding to the first k ordered p-values. $\hat{I}$ is considered to be a consistent estimator of I_0 if $P(\hat{I} = I_0) \xrightarrow{P} 1$.

Benjamini and Yekutieli (2001) showed that this procedure controls the false discovery rate at level q, that is,

$$E(Q) \leq \frac{d - d_0}{d} q \leq q, \tag{7.2}$$

where

$$Q = \begin{cases} V/R & \text{if } R > 0 \\ 0 & \text{otherwise,} \end{cases}$$

d_0 is the cardinality of I_0, $0 \leq R \leq p$ is the total number of rejected hypothesis and $0 \leq V \leq R$ the number of falsely rejected hypothesis.

Here we follow the steps of Bunea, Wegkamp and Auguste (2006). Note that if the estimator $\hat{I}$ is equal to the set I_0, we have exactly d_0 rejections ($R = d_0$) none

of them erroneously ($V = 0$). Therefore, consistency of $\hat{I}$ is verified by proving

$$P(\hat{I} = I_0) = P(R = d_0, V = 0) \to 1, \text{ as } n \to \infty, \tag{7.3}$$

i.e., showing that both $P(R \neq d_0)$ and $P(V \geq 1)$ are asymptotic negligible.

Lemma 2.1 in Bunea, Wegkamp and Auguste (2006) stablishes that

$$P(V \geq 1) \leq P(R \neq d_0) + \frac{d_0(d - d_0)}{d} q, \tag{7.4}$$

so that in order to show consistency of $\hat{I}$ we only need to show that $P(R \neq d_0) \to 0$, provided $q \to 0$.

Note that Bunea fits the full regression model, i.e., with all covariates and then uses t-tests for testing each $\beta_i, i = 1, \ldots, d$ separately. In fact, the estimate of the coefficient being tested under the full model is equivalent to the estimate of the regression when fitting that covariate on the residuals of all the other covariates.

Define the event $B_n = \{(\pi_{(1)}, \ldots, \pi_{(d_0)}) = (\pi_{j_1}, \ldots, \pi_{j_{d_0}})\}$ for $I_0 = \{j_1, \ldots, j_{d_0}\}$. This is the event that the d_0 smallest p-values are those that correspond to the predictors indexed by the set I_0. The idea is to prove that

$$\lim_{n \to \infty} P(B_n) = 1,$$

and use this to prove that $P(R \neq d_0) \to 0$.

7.2 Asymptotic Properties of the ANOVA-type Test Statistic when using Dimension Reduction

The asymptotic theorems shown in this dissertation are based on the estimators of m_1 that use the whole set of predictors $\mathbf{X}$. When using local polynomial estimators, the theorems holds when increasing smoothness assumptions on m are made. Simulations presented throughout this dissertation suggest that, the dimension reduction techniques (as SIR for example) provide reasonable estimators of m_1 for moderate dimension of $\mathbf{X}$.

By studying the asymptotic properties of the estimator of m_1 when using SIR as dimension reduction, if the bias is small enough, it may be possible to assess the asymptotic distribution of the proposed ANOVA-type test statistic as the sample size increases.

Bibliography

[1] Ait-Sahalia, Y., Bickel, P. J. and Stoker, T.M. (2001). Goodness-of-fit tests for kernel regression with an application to option implied volatilities. *Journal of Econometrics*, 105, 363-412.

[2] Abramovich, F.,Benjamini, Y., Donoho, D.L. and Johnstone,I.M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics*, 34, 584-653.

[3] Akaike, H. (1974). A new look at the statistical model identification.*IEEE Transactions on Automatic Control*,vol. 6, 716723.

[4] Akritas, M. G. and Papadatos, N. (2004). Heterocedastic One-Way ANOVA and Lack-of-Fit Tests. *Journal of the American Statistical Association*, 99, Theory and Methods.

[5] Benjamini, Y.; Gavrilov, Y. (2009). A Simple Forward Selection Procedure Based on False Discovery Rate Control. *The Annals of Applied Statistics*, 3, 179-198.

[6] Benjamini, Y.; Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57 (1): 289300.

[7] Benjamini, Y., Krieger, A.M., Yekutieli, D. (2006). Adaptive Linear Step-up False Discovery Rate controlling procedures. *Biometrika*, 93 (3): 491-507.

[8] Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statististics*, 29, 1165-1188.

[9] Birge, L. and Massart, P. (2001) A generalized Cp criterion for Gaussian model. *Technical report*, Lab. De Probabilities, Univ. Paris VI. (http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2001)

[10] Breiman, L. (1995). Better Subset Regression Using the Nonnegative Garrote. *Technometrics*, 37, 373384.

[11] Bunea, F., Wegkamp, M. and Auguste, A. (2006). Consistent variable selection in high dimensional regression via multiple testing. *Journal of Statistical Planning and Inference,* 136, 4349-4364.

[12] Cacoullos, T. (1964). Estimation of a Multivariate Density. *Ann. Inst. Statist. Math.*, 18, 179189.

[13] Candes, E., and Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35, 2313-2351.

[14] Delgado, M. A. and Manteiga, W. G. (2001). Significance Testing in Nonparametric Regression Based on the Bootstrap *The Annals of Statistics*, 29, 14691507.

[15] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81, 42555.

[16] Dvoretzky, A.; Kiefer, J. and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27 (3), 642669.

[17] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32, 407-499.

[18] Eubank, R. L. (1999). Nonparametric Regression and Spline Smoothing, *Statistics: A Series of Textbooks and Monographs*, Second Edition, Vol. 157.

[19] Fan, J. and Gijebels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman & Hall.

[20] Fan, J. and Jiang, J.(2005). Nonparametric Inferences for Additive Models. *Journal of the American Statistical Association*, 100, 890-907.

[21] Fan, J. and Jiang, J. (2007) Nonparametric inference with generalized likelihood ratio tests, *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research, Springer(2007)*, vol. 16, pages 409-444.

[22] Fan, Y. and Li, Q. (1996). Consistent model specification tests: Omitted variables and semiparametric functional forms. *Econometrica*, 64, 865-890.

[23] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96, 1348-1360.

[24] Fan, J., Zhang, C. M., and Zhang, J. (2001), Generalized Likelihood Ratio Statistics and Wilks Phenomenon. *The Annals of Statistics*, 29, 153193.

[25] Foster, D. P. and Stine, R. A. (2004). Variable selection in data mining: building a predictive model for bankruptcy. *Journal of the American Statistical Association*, 99, 303-313.

[26] Green, P.J and Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman & Hall.

[27] Hansen, B. E., (2008). Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24, 726-748.

[28] Hart, J. D. (1997). *Nonparametric Smoothing and Lack of fit tests*, Springer.

[29] Hastie, T.J. and Tibshirani, R.J. (1990). *Generalized Additive Models*, Chapman & Hall.

[30] Horowitz, J.L. and Mammen, E. (2004). Nonparametric estimation of an additive model with a link function. *The Annals of Statistics*, 32, 2412-2443.

[31] Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable Selection in Nonparametric Additive Models. Available at http://faculty.wcas.northwestern.edu/ jlh951/papers/HHW-npam.pdf

[32] Lavergne, P. and Vuong, Q. (2000). Nonparametric Significance Testing. *Econometric Theory*, 16, 576-601.

[33] Li, K. C., (1991). Sliced Inverse Regression for Dimension Reduction. *Journal of the American Statistical Association*, 86, 316-327.

[34] Li, R. and Liang, H. (2008). Variable selection in Semiparametric Regression Modeling. *The Annals of Statistics*, 36, 261-286.

[35] Lin, Y. and Zhang, H. (2006). Component Selection and Smoothing in Multivariate Nonparametric Regression. *The Annals of Statistics*, 34, 2272-2297.

[36] Linton, O. and Nielsen, J.P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82, 93-100.

[37] Mallows, C. L. (1973). Some Comments on Cp. *Technometrics*, 4, 661675.

[38] Mammen, E., Linton, O. and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics*, 27, 1443-1490.

[39] Nadaraya, E. A. (1965). On nonparametric estimates of density functions and regression curves. *Theory Prob. Appl.*, 10, 186-190.

[40] Newey, W. K. (1994). Kernel estimation of partial means. *Econom. Theory*, 10, 233-253.

[41] Racine, J. (1997). Consistent Significance Testing for Nonparametric Regression. *Journal of Business & Economic Statistics*,15, 369-378.

[42] Racine, J., Hart, J.D. and Li, Q. (2006). Testing the significance of categorical predictor variables in nonparametric regression models. *Econometric Reviews*, 25, 523-544.

[43] Rice, J. (1984). Bandwidth choice for nonparametric regression. *The Annals of Statistics*, 12, 1215-1230.

[44] Ruschendorf, L. (1977). Consistency of Estimators for Multivariate Density Functions and For the Mode. *Sankhya: The Indian Journal of Statistics*, 39, 243-250.

[45] Ryzin, J. Van (1969). On Strong Consistency of Density Estimates, *Ann. Math. Statist.*, Vol. 40, pp. 1765-1772.

[46] Stone, C. (1982). Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, 10, 1040-1053.

[47] Stone, C. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, 13, 689-705.

[48] Storlie, C. B, Bondell, H. D, Reich, B. J, Zhang, H. H. (2011). Surface Estimation, Variable Selection, and the Nonparametric Oracle Property. *Statistica Sinica*, 21(2), 679-705.

[49] Stute, W. (1984). Asymptotic Normality of Nearest Neighbor Regression Function Estimates. *The Annals of Statistcs*, 12, 917-926.

[50] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statististical Society B*, 58, 267-288.

[51] Tibshirani, R. and Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society B*, 61, 529-546.

[52] Wahba, G. (1990). Spline models for observational data. *Regional Conference Series in Applied Mathematics*, 59.

[53] Wand, M. P. and Jones, M. C. (1995) Kernel Smoothing, *Monographs on Statistics and Applied Probability*, Chapman & Hall, Vol. 60

[54] Wang, H. and Xia, Y. (2008). Shrinkage estimation of the varying coefficient model. Journal of the American Statistical Association, 104,747-757.

[55] Wang, L., Akritas, M. G. and Keilegom, I.V. (2008). An ANOVA-type Nonparametric Diagnostic Test for Heterocedastic Regression Models. *Journal of Nonparametric Statistics*, 20, 365-382.

[56] Wegman, E.J. (1972). Nonparametric Probability Density Estimation: I. A Summary of Available Methods, *Technometrics, American Statistical Association and Ameican Society for Quality*, Vol.14, pp.533-546.

[57] Zhang, H., Wahba, G., Lin, Y., Voelker, M., Ferris, M., Klein, R. and Klein, B. (2004). Variable selection and model building via likelihood basis pursuit. *Journal of American Statistical Association*, 99, 659-672.

[58] Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, 67, Part 2, 301-320.

[59] Zou, H. (2006). The Adaptive Lasso and its Oracle Properties. *Journal of the American Statistical Association*, 101(476), 1418-1429.

Vita

Adriano Zanin Zambom

Vita Adriano Zanin Zambom EDUCATION

- Ph.D. Statistics, Pennsylvania State University.
- Ms.Sci. Statistics, State University of Campinas (UNICAMP), Brazil. Feb 2008
- B.S. Statistics, State University of Campinas (UNICAMP), Brazil. Dec 2005

PROFESSIONAL EXPERIENCE

- 2011 - 2012: Research Assistant/Consultant, Pennsylvania State University
- 2008 - 20010: Teaching Assistant, Pennsylvania State University
- 2007: Teaching Assistant, State University of Campinas (UNICAMP), Brazil

AWARDS and SCHOLARSHIPS

- 2008 - 2012: Ph.D. Scholarship. Brazilian Ministry of Education - CAPES
- 2007: Ms.Sci. Scholarship. FAPESP - Brazil