ABSTRACT

Title of dissertation: Nonparametric Estimation and Testing of Interaction in Generalized Additive Models

Bo Li, Doctor of Philosophy, 2011

Dissertation directed by: Professor Paul J. Smith
Department of Mathematics

The additive model overcomes the "curse of dimensionality" in general nonparametric regression problems, in the sense that it achieves the optimal rate of convergence for a one-dimensional smoother. Meanwhile, compared to the classical linear regression model, it is more flexible in defining an arbitrary smooth functional relationship between the individual regressor and the conditional mean of the response variable Y given X. However, if the true model is not additive, the estimates may be seriously biased by assuming the additive structure.

In this dissertation, generalized additive models (with a known link function) are considered when containing second order interaction terms. We present an extension of the existing marginal integration estimation approach for additive models with the identity link. The corresponding asymptotic normality of the estimators is derived for the univariate component functions and interaction functions. A test statistic for testing significance of the interaction terms is developed. We obtained

the asymptotics for the test functional and local power results. Monte Carlo simulations are conducted to examine the finite sample performance of the estimation and testing procedures. We code our own local polynomial pre-smoother with fixed bandwidths and apply it in the integration method. The widely used *LOESS* function with fixed spans is also used as a pre-smoother. Both methods provide comparable results in estimation and are shown to work well with properly chosen smoothing parameters. With a small and moderate sample size, the implementation of the test procedure based on the asymptotics may produce inaccurate results. Hence a wild bootstrap procedure is provided to get empirical critical values for the test. The test procedure performs well in fitting the correct quantiles under the null hypothesis and shows strong power against the alternative.

Nonparametric Estimation and Testing of Interaction in Generalized Additive Models

by

Bo Li

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Professor Paul J. Smith
Professor Benjamin Kedem
Professor Partha Lahiri
Professor Eric V. Slud
Professor Francis B. Alt

DEDICATION

To my grandmother, parents, and to all my family!

ACKNOWLEDGEMENTS

Finally the journey is about to come to an end. When I look back all these years of my graduate study in College Park, I found that it is such an unforgettable experience and it will be cherished forever. I owe my gratitude to all the people who helped me along the journey.

My deepest appreciation is to my advisor, Professor Paul Smith, for his extraordinary patience, understanding and encouragement throughout all these years. He is the best source of knowledge, inspiration and strength for me. Without his consistent help and support, I would have never completed my thesis successfully. I benefit not only from his broad knowledge in statistics and mathematics, the way of doing rigorous research he taught me, but also from his constantly positive attitude when facing difficulties.

I would like to send my special thanks to Professor Benjamin Kedem, Professor Eric Slud, Professor Partha Lahari and Professor Grace Yang for their kindly advisory and valuable teaching during my doctoral study. I also thank Professor Francis Alt for agreeing to serve on my thesis committee and for sparing his time reviewing the manuscript.

My sincere thanks go to all my friends. There are too many of them and I

can only name a few: Weiran, Huilin, Hua Wei, Huina, Yu-Ru, Qi Zhang, Qianwen, Hanzhu, Min Min, Lingyan, Liyan, Tinghui, Yabing, Wei Hu, Weigang and Huaqiang. Their invaluable friendship accompanies me all the way and pulled me through hardest time.

I would also like to thank some of the staff members in Math Department, Haydee Hidalgo, Linette Berry, Celeste Regalado, Sharon Welton, William Schildknecht, for their selfless help.

Last but not least, I would like to thank my grandmother, my parents, my sister and Linbao for their endless love and unconditional support. They have always stood by me and kept me moving forward. I also would like to thank my daughter Yun-Mo for bringing so much joy to my life. No words can express enough my gratitude towards them. I devoted this dissertation to all my family.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Regression analysis concerns the conditional distribution of a dependent variable Y as a function of one or several explanatory variables X. It is of great interest to estimate the average value of Y, when the predictors X are held fixed $-$ that is the conditional mean $E(Y|X)$. In practice, we will mostly be interested in multiple regression problems, where the regressor X is a $d-$dimensional vector $(X_1, \cdots, X_d)$ with $d > 1$. Classical regression analysis assumes a linear relationship where the mean of Y is related to a set of independent variables $X_1, X_2, \cdots, X_d$ in the following way:

$$E(Y|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d,$$

where $\beta_j,\ j = 1, 2, \ldots, d$ are unknown coefficients. In some cases, if the linearity assumption is not tenable, the expected form of the nonlinear function is known and can be parameterized, often in terms of basis functions. All these types of regression models are referred to as parametric regression, where the estimation of a finite number of parameters is required. In this case, the fitted models can be easily interpreted and estimated accurately if the underlying assumptions are correct. However, if they are violated, then the estimates may be inconsistent and give a misleading conclusion of the regression relationship.

As a result, nonparametric regression has become a rapidly developing field since it avoids restrictive assumptions on the functional form of the regression function. It relaxes the assumption, typically by substituting the weaker assumption that the average value of the response is a smooth function, $m(X)$, of the predictors. The nonparametric regression model assumes

$$Y = m(X) + \sigma(X)\varepsilon \tag{1.1}$$

where the error ε is independent of X with $E(\varepsilon) = 0$ and $Var(\varepsilon) = 1$. Straightforwardly, $\sigma(X)$ corresponds to the conditional variance of Y given X, $Var(Y|X)$. There exist various smoothing techniques to implement nonparametric regression, such as kernel regression, local polynomial regression, smoothing splines, regression splines, and so on. While a nonparametric regression problem involves multiple explanatory variables, it is subject to the well-known "curse of dimensionality:" neighborhoods with a fixed number of points become less local as the dimensions increase. Consequently, estimators based on local averaging perform unsatisfactorily in this situation. Technically, the rate of convergence of the estimator decreases dramatically for higher dimensions d. In addition, the difficulties of interpretation and visualization arises in the case of more than two dimensions. Nonparametric multiple regression will be described in detail in Section 2.1.

For this reason, many methods of dimensionality reduction have been proposed in literature. Among those, Buja, Hastie and Tibshrani (1989) and Hastie and Tibshirani (1990) proposed the additive model,

$$m(x) = c + \sum_{j=1}^{d} f_j(x_j). \tag{1.2}$$

Here c is a constant and $\{f_j(\cdot)\}_{j=1}^{d}$ is a set of unknown functions normalized by $E[f_j(X_j)] = 0$. This type of model structure is useful from a statistical point of view since it achieves a good compromise among flexibility, dimensionality and interpretability. In particular, it is well-known that additive regression models can be estimated with the same rate of estimation error as in the univariate case (Stone, 1985). Additive models are important in both theoretical economics and econometric data analysis. In economic theory, additivity is equivalent to the so-called property of strong separability. These models have a desirable structure allowing empirical data analysis for subsets of regressors. The separability of the input variables is consistent with decentralization in decision making or optimization by stages. In summary, additive models can be easily interpreted. For details, see Deaton and Muellbauer (1980).

Buja, Hastie and Tibshrani (1989) proposed an estimating procedure of backfitting, which estimates the orthogonal projection of the regression function $m(\cdot)$ onto the space of additive functions. The asymptotic theory of backfitting has been developed by Opsomer and Ruppert (1997), Mammen, Linton and Nielsen (1999) and Opsomer (2000). However, many extensions of asymptotic properties of this method remain unknown due to its iterative nature. Several authors have proposed a non-iterative direct method based on marginal integration (Tjøstheim and Auestad, 1994; Linton and Nielsen, 1995). This method generates an alternative projection onto the subspace of additive functions which is not necessarily orthogonal. One advantage of this method is that an explicit asymptotic theory can be constructed. More recently, several modifications of the marginal integration estimator have been proposed (see, for example, Hengartner, 1996; Linton, 1997; Fan, Härdle and Mammen, 1998; Severance-Lossin and Sperlich, 1999; Linton, 2000). For a more detailed discussion on the difference between the backfitting and the marginal integration method, we refer to the work of Nielsen and Linton (1998) and Sperlich, Linton and Härdle (1999). Nevertheless, due to their very different interpretation when the true model is not additive, backfitting and marginal interpretation should not be considered as competitors. We will briefly sketch the difference between the two most popular methods for additive model estimation in Section 2.2.1.4.

Due to the advantages additivity offers to the empirical researchers, the additive model (1.1) should be accompanied by an adequate model check. It was not until recently that the problem of testing additivity gained real interest (see, for example, Hastie and Tibshirani 1990; Barry 1993; Eubank et al. 1995; Chen et al. 1995; Derbort, Dette and Munk 2002; Gozalo and Linton 2001; Dette et al. 2001).

Additivity tests developed in these works have mostly focused on testing whether a regression function $m(x)$ is purely additive or not in the sense of (1.2). However, in case that pure additivity is rejected, one would like to know exactly which interaction terms are present. For this reason, Sperlich, Tjøstheim and Yang (2002) extended model (1.2) to include pairwise interactions, resulting in

$$m(x) = c + \sum_{j=1}^{d} f_j(x_j) + \sum_{1 \leq j < k \leq d} f_{jk}(x_j, x_k). \tag{1.3}$$

Sperlich, Tjøstheim and Yang (2002) showed that all components can be identified and estimated consistently by marginal integration, obtaining the optimal convergence rate in smoothing. Another main point of their paper is to test directly for such interactions based on the estimates of the particular interaction term.

Model (1.1) excludes a wide variety of situations, evidently. To allow for even more flexibility than the additive model does, we extend the model to the class of generalized additive models (GAMs) defined as

$$G\{m(x)\} = c + \sum_{j=1}^{d} f_j(x_j), \tag{1.4}$$

where $G(\cdot)$ is a fixed link function. GAMs were introduced in a series of papers by Hastie and Tibshirani (1986, 1987a, 1987b) and Stone (1986). They are described in detail in Hastie and Tibshirani (1990). Generalized linear models (GLMs) are special cases of GAMs, with the functions $f_j(x_j)$ taken to be linear of the form $\beta_j x_j$. GAMs are appropriate for many situations like binary and survival data, etc.

Widely used link functions include the logit and probit links for binary data, or the logarithm transform for Poisson data. Without loss of generality, we assume that the link function $G(\cdot)$ is known as a priori throughout the dissertation. Härdle et al. (2004) discussed the problem of testing the specification of the link function, but this is beyond the scope of the present research. The local scoring algorithm (Buja, Hastie and Tibshrani, 1989; Hastie and Tibshirani, 1990) was an extension of backfitting for the non-trivial link function $G(\cdot)$. Meanwhile, it is practically identical with the Fisher scoring algorithm used in GLMs, except that the least-square steps used to update the linear coefficients $\hat{\beta}$ are replaced by the backfitting algorithm to update the estimates for the component functions f_j and constant c.

As an alternative, Linton and Härdle (1996) extended the marginal integration idea to the estimation of GAMs and provided the asymptotics of the estimates. Based on this initial investigation, Yang, Sperlich and Härdle (2003) developed a direct estimation procedure for function derivatives, which is a very important matter, especially in economic studies. The GAMs will be described in detail in Section 2.2.2.

Analogous to the case of the identity link function, when the purely additive model (1.2) is rejected by applying some existing tests, for example, the one proposed by Gozalo and Linton (2001), one may want to know further which interaction terms are relevant. For this purpose, we will consider the following model, by allowing for second-order interactions,

$$G\{m(x)\} = c + \sum_{j=1}^{d} f_j(x_j) + \sum_{1 \leq j < k \leq d} f_{jk}(x_j, x_k). \tag{1.5}$$

Some possible estimations of this model have already been existing in the literature. Hastie and Tibshirani (1990) discussed possible algorithms for backfitting with cubic smoothing splines. More recently, Roca-Pardiñas, Cadarso-Suárez and González-Manteiga (2005) studied the estimation and testing problem in the same model, based on the local scoring algorithm with backfitting, where a likelihood ratio-based test and an empirical process-based test were proposed. The resulting tests are useful in practice, but asymptotic distributions of the testing statistics are unknown and hard to derive. It should be mentioned that Coull, Ruppert and Wand (2001) proposed an algorithm based on penalized spline models, which incorporates factor by curve interactions in GAMs, and provides some tests for additivity.

The main objective of this dissertation is to consider estimation of component functions in model (1.5) and testing of the bivariate interaction functions $f_{jk}(\cdot)$ in such models, through the use of marginal integration techniques.

In the sequel we will refer to this model as a GAM with interactions and focus on the interactions between continuous explanatory variables.

Chapter 2 presents a detailed review of the literature on nonparametric regression, particularly on topics related to GAMs.

In Chapter 3, we introduce the technical setting for the problems and describe a marginal integration estimating procedure for the component functions. The pointwise asymptotic properties of the integration estimator are presented. Having these estimates in hand we construct a test statistic with the functional form

$$\int \hat{f}_{jk}^2(x_j, x_k)\pi(x_j, x_k)dx_j dx_k,$$

where π is a nonnegative weight function, to test whether a specific interaction function is present or not. We also derive the asymptotic distribution of the test statistic. All proofs are deferred to the end of the chapter.

Several simulation experiments are carried out for the estimation and testing problems. It is well-known that the first-order asymptotics of test functionals of the previous type do not give a very accurate approximation of the finite sample distribution in practice. Hjellvik, Yao and Tjøstheim (1998) showed that several hundred thousand sample points may be necessary to reach some reasonable accuracy. As a consequence, a wild bootstrap (see, e.g., Liu 1988; Wu 1986) scheme is adopted to construct the null distribution of the test statistic. A detailed introduction, discussion and theory of this method in the context of testing problems combined with a marginal integration method can be found in Gozalo and Linton (2001), Sperlich, Tjøstheim and Yang (2002), Yang, Sperlich and Härdle (2003) and Härdle et al. (2004). We will also sketch the method briefly in Chapter 4, where the details and results of the simulation studies are given.

We summarize this dissertation research in Chapter 5 and describe several possible points of interest for future work.

Chapter 2

Literature Review

2.1 Nonparametric Regression

2.1.1 Simple Nonparametric Regression - Smoothing Scatterplots

In this section, a brief summary is given of the key methods available for estimating univariate nonparametric regression functions. Some of the methods, in particular local polynomial regression, will be discussed in more detail in later subsections.

The nonparametric regression case with a single predictor is also called nonparametric simple regression, or scatterplot smoothing. There now exist various approaches to the problem. Binning, nearest-neighbor and running-mean (or local averaging) are among the simplest smoothing methods. Some of the more popular ones are those based on kernel functions, spline functions and wavelets. Each of these options has its own strengths and weaknesses. Among them, kernel-based regression estimators have the advantage of mathematical and intuitive simplicity. The traditional kernel regression approaches include the famous Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964) and several alternative kernel estimators

(Priestley and Chao, 1972; Gasser and Müller, 1979). We will give a brief overview of these classical estimators in Section 2.1.1.1. Another important class of kernel-type regression estimators is local polynomial regression estimators (Stone, 1977; Cleveland, 1979; Müller, 1987; Fan, 1992; Ruppert and Wand, 1994; Wand and Jones, 1995; Simonoff, 1996; Fan and Gijbels, 1996). They will be discussed in Section 2.1.1.2. Other smoothing methods, e.g., smoothing splines, will be introduced in Section 2.1.1.3.

2.1.1.1 Kernel Estimators

The essential idea of kernel estimation is that in estimating $m(x_0)$, the value of the regression function at $x = x_0$, it is desirable to give greater weight to observations close to the focal x_0 and less weight to those that are remote. It is an improved version of local averaging, which can be thought as local weighted averaging. These weights are defined by a kernel function K, which is usually a symmetric probability density. Let h be a bandwidth, which is a nonnegative number controlling the size of the local neighborhood. The Nadaraya-Watson kernel regression estimator is given by

$$\hat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} K_h(X_i - x)Y_i}{\sum_{i=1}^{n} K_h(X_i - x)}$$

where $K_h(\cdot) = K(\cdot/h)/h$.

If the design is not random, but is rather a fixed set of sorted numbers

$X_1, \cdots, X_n$, a different form of kernel estimator could be considered. The Gasser-Müller estimator is intended for the fixed design case and is defined as follows:

$$\hat{m}_{GM}(x) = \sum_{i=1}^{n} \left[\int_{s_{i-1}}^{s_i} K_h(u-x)du \right] Y_i$$

where $X_{i-1} \leq s_{i-1} \leq X_i$ (a common choice being $s_{i-1} = (X_{i-1} + X_i)/2$, with s_0 and s_n being the upper and lower limits of the range of X, respectively). It is a modification of an earlier version of Priestley and Chao (1972).

Basic calculus shows that $\hat{m}_{NW}$ is the solution to a weighted least squares regression problem,

$$\hat{\beta}_0 = \text{argmin}_{\beta_0} \sum_{i=1}^{n} (Y_i - \beta_0)^2 \omega_i = \sum_{i=1}^{n} \omega_i Y_i / \sum_{i=1}^{n} \omega_i,$$

with $\omega_i = K_h(X_i - x)$. Similarly, the Gasser-Müller estimator is of the same form with weights $\omega_i = \int_{s_{i-1}}^{s_i} K_h(u-x)du$. That is, both estimators use locally constant approximations, giving heavier weight to values of Y_i corresponding to X_is closer to x.

This suggests fitting higher order local polynomials, for example, a local linear fit, since a local constant usually makes sense only over a very small neighborhood. Table 2.1, which is taken from Fan and Gijbels (1996), summarizes the first order asymptotic performance of the Nadaraya-Watson estimator, the Gasser-Müller estimator, and the local linear regression estimator at an interior point of the support of the design density. Note that in the table, $b_n = (1/2)h^2 \int_{-\infty}^{\infty} u^2 K(u)du$,

Table 2.1: Pointwise asymptotic bias and variance of kernel regression smoothers

Method	Bias	Variance
Nadaraya-Watson	$\left(m''(x) + \frac{2m'(x)f'(x)}{f(x)}\right) b_n$	V_n
Gasser-Müller	$m''(x)b_n$	$1.5V_n$
Local Linear	$m''(x)b_n$	V_n

$V_n = (\sigma^2(x)/(f(x)nh)) \int_{-\infty}^{\infty} K^2(u)du$ and $f(x)$ is the design density.

It is easy to see that unlike the Nadaraya-Watson estimator, the bias of the local linear fit is independent of the design and disappears when the true regression curve $m(\cdot)$ is linear. The Gasser-Müller estimator on the other hand corrects the bias of the Nadaraya-Watson estimator but at the expense of increasing its variance. The local linear estimator achieves further improvement in the boundary regions. In the case of Nadaraya-Watson estimates we typically observe problems due to the one-sided neighborhoods at the boundaries. The reason is that in local constant modeling, more or less the same points are used to estimate the curve near the boundary. Local linear fit (also for more general local polynomial regression) corrects this automatically by fitting a higher degree polynomial here.

Comparisons between local linear and local constant fit were discussed in detail by Chu and Marron (1991), Fan (1992), and Hastie and Loader (1993).

2.1.1.2 Local Polynomial Regression

Let us sketch briefly the framework of local polynomial regression in the univariate case. Locally, the regression function m can be approximated by a Taylor expansion of order p,

$$m(z) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}(z-x)^j \equiv \sum_{j=0}^{p} \beta_j (z-x)^j$$

for z in a neighborhood of x, where $m^{(j)}(x)$ denotes the jth derivative of $m(x)$. Now, consider the following locally weighted least squares problem:

Let $\hat{\beta}_j$ $(j = 0, \ldots, p)$ minimize

$$\sum_{i=1}^{n} \left[Y_i - \sum_{j=0}^{p} \beta_j (X_i - x)^j \right]^2 K_h(X_i - x)$$

where K is a kernel function and h is the bandwidth, controlling the size of the local neighborhood. The above exposition suggests that an estimator for $m^{(\nu)}$(x) is

$$\hat{m_\nu}(x) = \nu! \hat{\beta}_\nu.$$

The least squares theory provides the solution

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y},$$

where $\mathbf{X}$ is an $n \times (p+1)$ matrix with the ith row $\{1, (X_i - x), \ldots, (X_i - x)^p\}$, $\mathbf{W}$ is the $n \times n$ diagonal matrix of weights $\mathbf{W} = \text{diag}\{K_h(X_i - x)\}$, and $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$.

Therefore, to estimate the function $m(\cdot)$, the whole curve

$$\hat{m}(x) = \hat{\beta}_0 = \mathbf{e}_1^T (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{Y}$$

is obtained by running the local polynomial regression of order p with x varying in an appropriate estimation domain. Here, $\mathbf{e}_r$ is $(p+1) \times 1$ vector having 1 in the rth entry and all other entries 0.

Moreover, $\nu!\hat{\beta}_\nu = \nu!\mathbf{e}_{\nu+1}^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$ is an estimate of the νth derivative of $m(x)$, $m^{(\nu)}(x)$. It is obvious that when $p = 0$, $\hat{\boldsymbol{\beta}}$ reduces to $\hat{\beta}_0$, which means that the local constant estimator is nothing other than the well-known Nadaraya-Waston estimator.

As in other kernel methods, the bandwidth h determines the degree of smoothness of $\hat{m}(\cdot)$. As $h \to 0$, the resulting estimate essentially interpolates the data, namely at an observation X_i, $\hat{m}(X_i)$ converges to Y_i, while an infinitely large h makes all weights equal, so that we obtain a parametric pth order polynomial fit. As h ranges from 0 to ∞, $\hat{m}$ ranges from the most complex interpolation model to the simplest parametric regression model by polynomials.

In marked contrast to the parametric polynomial regression approach, this technique is local and hence requires a small degree of the local polynomial, typically of order $p = \nu + 1$ or occasionally $p = \nu + 3$. For example, for estimating a regression function itself one often uses a local linear model with $p = 1$.

Besides the advantages described in the last section, there is an absence of boundary effects: the bias at the boundary stays automatically of the same order as that in the interior, without use of specific boundary kernels. Thus no boundary modifications are required with local polynomial fitting. This is an important merit especially when dealing with multidimensional cases for which the boundary effects can be quite substantial (Silverman, 1986; Fan and Marron, 1993).

As argued in Fan and Gijbels (1996), odd order polynomial fits are preferable to even order polynomial fits. As we have seen, the local linear fit performs asymptotically better than the Nadaraya-Watson estimator. On the other hand, for sufficiently smooth regression functions, the asymptotic performance of the local polynomial estimator improves for higher values of p. However, as with higher order kernels, the variance of the estimator becomes larger for higher p and a very large sample may be required for a substantial improvement in practical performance, especially beyond cubic fits. Typically, the order p is taken to be one (local linear) or sometimes three (local cubic) to estimate the regression function.

2.1.1.3 Other Smoothing Methods

The spline smoothing approach is used to estimate the unknown smooth regression function by explicitly trading off fidelity to the data with smoothness of the estimate. A natural measure of "fidelity to the data" for a regression curve m is the residual sum of squares

$$\sum_{i=1}^{n} \left[Y_i - m(X_i)\right]^2 .$$

This distance measure will be reduced to zero by any m that interpolates the data. Such a curve is not acceptable on the grounds that it is too oscillatory and it is not unique. That means we want to produce a curve which fits the data well without too much local variation. To quantify local variation, one could use the measure of roughness based on derivatives, for instance, the roughness penalty $\int [m''(x)]^2 dx$.

Define the penalized residual sum of squares

$$\sum_{i=1}^{n} [Y_i - m(X_i)]^2 + \lambda \int [m''(x)]^2 dx, \tag{2.1}$$

with a smoothing parameter λ, which represents the rate of exchange between residual error and roughness of the curve m. If we restrict the possible minimizing functions to be twice differentiable on the interval $[a, b] = [X_{(1)}, X_{(n)}]$, this problem has a unique solution $\hat{m}_\lambda(x)$, which is a cubic spline with knots at the unique values of X_i. Moreover, it can be argued that $\hat{m}_\lambda$ is linear in the responses:

$$\hat{m}_\lambda(x) = n^{-1} \sum_{i=1}^{n} W_i(x, \lambda; X_1, \cdots, X_n) Y_i,$$

where W_is are the weights. Silverman (1984) pointed out that the smoothing spline is basically a local kernel average with a variable bandwidth.

The larger values of the smoothing parameter λ yields a smoother estimator by penalizing roughness more. One approach in selecting λ is via the minimization of the cross-validation criterion

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} [Y_i - \hat{m}_{\lambda,i}(X_i)]^2, \tag{2.2}$$

where $\hat{m}_{\lambda,i}(X_i)$ is the spline estimator computed without using the i^{th} observation and evaluated at X_i, arising from (2.1). It can be shown that for linear smoothers, $CV(\lambda)$ can be written as a function of fitted values,

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i - \hat{m}_\lambda(X_i)}{1 - A_{ii}(\lambda)}\right]^2, \tag{2.3}$$

where $A_{ii}(\lambda)$ is the ith diagonal element of the smoother matrix. However, this cross-validation criterion is computationally intensive. An alternative version is generalized cross-validation (GCV), proposed by Wahba (1977) and Craven and Wahba (1979). GCV replaces each value $1 - A_{ii}(\lambda)$ in (2.3) with their average, namely, $1 - (1/n)\text{trace}[A(\lambda)]$. Hence the GCV selector of λ is the minimizer of

$$GCV(\lambda) = \frac{\sum_{i=1}^{n}[Y_i - \hat{m}_\lambda(X_i)]^2}{n\{1 - n^{-1}\text{trace}[A(\lambda)]\}^2}.$$

Both quantities $CV(\lambda)$ and $GCV(\lambda)$ are consistent estimates of the MISE (mean integrated squared error) of $\hat{m}_\lambda$. See Wahba and Wang (1990) for a description of other methods for selecting the smoothing parameter.

For comprehensive works on spline smoothing, see Eilers and Marx (1996), Wahba (1990) and Green and Silverman (1994).

Another class of smoothing techniques is called orthogonal series regression. This method uses the fact that under certain conditions, the regression function can be represented by a series of orthogonal basis functions. The coefficients of the basis functions have to be estimated. The smoothing parameter N is the number of terms in the series. For larger N, the estimator will be smoother. The book of Efromovich (1999) provides a detailed discussion of this approach to smoothing.

We close the introduction of scatterplot smoothing methods here. For a comprehensive overview on a variety of smoothing methods, refer to the books of Eubank (1999) and Schimek (2000).

Most of the univariate smoothing techniques discussed above can be generalized to higher dimensions. Since kernel-based nonparametric regression is employed in this dissertation research, we will focus on introducing the multivariate version of the kernel-based methods in next section.

2.1.2 Multiple Nonparametric Regression

2.1.2.1 Kernel-based Regression for Multivariate Data

In practice, researchers will mostly be interested in multiple regression problems, where they need to specify how the response variable Y depends on a vector of predictors $\mathbf{X} = (X_1, \cdots, X_d)^T$. This means we need to estimate the conditional expectation

$$E(Y|\mathbf{X}) = E(Y|X_1, \cdots, X_d) = m(\mathbf{X}).$$

The multivariate generalization of the Nadaraya-Watson estimator is

$$\hat{m}_{NW}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})Y_i}{\sum_{i=1}^{n} \mathcal{K}_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x})},$$

where $\mathcal{K}$ denotes a multivariate kernel function operating on d arguments and $\mathbf{H}$ is a nonsingular bandwidth matrix. An equal bandwidth h in all dimensions corresponds to $\mathbf{H} = h\mathbf{I}_d$ with the $d \times d$ identity matrix $\mathbf{I}_d$. Different bandwidths correspond to $\mathbf{H} = \text{diag}(h_1, \cdots, h_d)$. Hence, the estimator is again a weighted sum of those observed responses Y_i where $\mathbf{X}_i$ lies in a ball or cube around $\mathbf{x}$, depending on the choice of the kernel. Note also that the multivariate Nadaraya-Watson estimator is a local constant fit.

Local polynomial estimation generalizes in a straightforward way to multiple predictors. Let us illustrate this with the simplest case of local linear regression. The minimization problem here is to minimize

$$\sum_{i=1}^{n} \left[Y_i - \beta_0 - \boldsymbol{\beta}_1^T(\mathbf{X}_i - \mathbf{x})\right]^2 \mathcal{K}_{\mathbf{H}}(\mathbf{X}_i - \mathbf{x}),$$

and the solution to the problem can be written as

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}_1^T)^T = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}$$

in which

$$\mathbf{X} = \begin{pmatrix} 1 & (\mathbf{X}_1 - \mathbf{x})^T \\ \vdots & \vdots \\ 1 & (\mathbf{X}_n - \mathbf{x})^T \end{pmatrix}, \quad \mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix},$$

and the weight matrix $\mathbf{W} = \text{diag}(\mathcal{K}_{\mathbf{H}}(\mathbf{X}_1 - \mathbf{x}), \cdots, \mathcal{K}_{\mathbf{H}}(\mathbf{X}_n - \mathbf{x}))$. Hence, $\hat{\beta}_0$ is the estimate of the regression function and $\hat{\boldsymbol{\beta}}_1$ estimates the partial derivatives with respect to the components of $\mathbf{x}$. Hence the multivariate local linear estimator is

$$\hat{m}(\mathbf{x}) = \hat{\beta}_0 = \mathbf{e}_1^T(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}.$$

Under some regularity assumptions, the conditional asymptotic bias and variance of $\hat{m}(\mathbf{x})$ are

$$\text{Bias}\{\hat{m}(\mathbf{x})|\mathbf{X}_1, \cdots, \mathbf{X}_n\} = \frac{1}{2}\mu_2(\mathcal{K})tr\{\mathbf{H}^T\mathcal{H}_m(\mathbf{x})\mathbf{H}\} + o_p\{tr(\mathbf{H}\mathbf{H}^T)\} \tag{2.4}$$

and

$$\text{Var}\{\hat{m}(\mathbf{x})|\mathbf{X}_1, \cdots, \mathbf{X}_n\} = \frac{1}{n\text{det}(\mathbf{H})}\|\mathcal{K}\|_2^2\frac{\sigma^2(\mathbf{x})}{f_{\mathbf{X}}(\mathbf{x})}\{1 + o_p(1)\} \tag{2.5}$$

respectively, in the interior of the support of the density function $f_{\mathbf{X}}$, where $\mu_2(\mathcal{K}) = \int u^2\mathcal{K}(u)du$, $\|\mathcal{K}\|_2^2 = \int \mathcal{K}^2(u)du$, $\mathcal{H}_m(\mathbf{x})$ denotes the $d \times d$ Hessian matrix of the regression function $m(\cdot)$ at $\mathbf{x}$ and $\sigma^2(\mathbf{x}) = Var(Y|\mathbf{X} = \mathbf{x})$.

We refer to Ruppert and Wand (1994) for more details of the derivation. They also pointed out that the local linear estimate has the same order conditional bias in the interior as in the boundary region of the support of $f_{\mathbf{x}}$. Thus it avoids boundary bias problems in the same way as the univariate case.

2.1.2.2 Curse of Dimensionality

Theoretically, regression smoothing for a multi-dimensional predictor can be performed as in the case of a one-dimensional predictor. The local averaging procedure will still provide consistent estimates of the regression surface. Although generalizations of most univariate smoothing techniques to higher dimensions appear to be feasible, there is a serious problem arising: the so-called "curse of dimensionality", as it was termed by Bellman (1961). This problem refers to the fact that a local neighborhood in higher dimensional Euclidean space is no longer local: a neighborhood with a fixed percentage of observations can be very big and far from "local." Or to understand it in another way: if 10 data points are adequate for a one-dimensional nonparametric regression problem, 10^d data points will be required for a d-dimensional problem. As a consequence, much larger data sets are needed even for a moderate d. Unfortunately, in practice, such large data sets are often not available. In another words, the observations in higher dimensions are often sparsely distributed even for large sample sizes, and hence estimators based on local averaging perform unsatisfactory. Technically, we can explain this effect by looking at the AMSE (asymptotic mean squared error) of the estimates. Consider a multiple regression estimator with the identical bandwidth h for all directions, for example, a local linear estimator with bandwidth matrix $\mathbf{H} = h \cdot \mathbf{I}_d$. Based on (2.4) and (2.5), the AMSE will also depend on the number of dimensions d,

$$\text{AMSE} = \frac{1}{nh^d} C_1 + h^4 C_2,$$

where C_1 and C_2 are constants that depend on neither n nor h. If we calculate the optimal bandwidth to minimize the AMSE, we find that $h_{opt} = cn^{-1/(4+d)}$ and hence the fastest possible rate of convergence for AMSE is $n^{-4/(4+d)}$. It is apparent that the rate of convergence drops dramatically for higher dimension d.

The curse of dimensionality has been illustrated clearly in many books, such as Silverman (1986), Härdle (1990), Hastie and Tibshirani (1990), Scott (1992) and Fan and Gijbels (1996).

There is another disadvantage with the multiple regression smoothing. Here the regression function $m(\mathbf{x})$ is a surface in a high dimensional space and since its form can not be easily displayed for $d > 2$, it does not provide a geometric description of the regression relationship between $\mathbf{X}$ and Y. The question is how can we examine the effect of particular variables once we have fitted a complicated surface.

Several multiple nonparametric regression approaches have been investigated, at least partly in response to the dimensionality problem. They all involve some dimensionality reduction process.

Tree based regression is one technique. The regression surface is approximated by a linear combination of step functions

$$m(\mathbf{x}) = \sum_{k=1}^{K} c_k I(\mathbf{x} \in R_k),$$

where the R_k are disjoint hyperrectangles in predictor space with sides parallel to the coordinate axes, and the c_k are coefficients that are estimated by the mean value of Y in region R_k. Recursive partitioning regression carves the predictor space up into disjoint blocks R_k in a binary style and hence the model can be represented by a binary tree. Each terminal node or leaf of the tree represents a hyperrectangle. The fitted values are constants in each leaf. The tree is built sequentially, and a new branch replaces a leaf if a split of that region is warranted in terms of the gain in predictive power. This repeated binary-style splitting process can be terminated when no further splits can significantly improve the homogeneity of the subgroups. However, this strategy can miss effective splits further down in the tree by stopping too soon. A preferable strategy is to build a very large tree and then prune it back to a reasonable size. The pruning can be guided by cross-validation and the final tree selected is the subtree of the original large tree with the smallest estimated prediction error. While this strategy may appear to be computationally formidable, an effective algorithm has been developed in the CART program (Breiman et al., 1993).

Another direct attack on the dimensionality issue is projection pursuit regression, introduced by Friedman and Stuetzle (1981). It models the regression surface as

$$m(\mathbf{x}) = \sum_{k=1}^{K} g_k(\boldsymbol{\beta}_k^T \mathbf{x})$$

where $\boldsymbol{\beta}_k^T \mathbf{x}$ ($\|\boldsymbol{\beta}_k\| = 1, k = 1, \cdots, K$) is a one-dimensional projection of the vector $\mathbf{x}$, and g_k is an arbitrary univariate function of this projection. This model can be thought as an extension of the regression tree model. It builds up the regression surface by estimating the univariate regressions along carefully chosen projections defined by the β_k. The directions β_k and number of terms K are selected to give the best predictive power. Only one-dimensional smoothing is performed to avoid the dimensionality difficulty. Projection pursuit regression models are parsimonious smooth surface estimators but are hard to interpret for K greater than one.

There are some other approaches aimed at overcoming the curse of dimensionality problem, but we prefer to stop here. They all suffer from the difficulty of interpretation. This is not a problem in the linear multiple regression model since the regression function $m(\cdot)$ is assumed to be linear and hence additive in the predictors. The effect of each explanatory variable can be examined separately, which is an important feature of the linear model that has made it so popular for statistical inference. If we drop the linearity assumption and retain the additivity feature, we will get the additive model defined in (1.1), which will be discussed in detail in the next section.

2.2 Additive and Generalized Additive Models

2.2.1 Additive Models

As we just discussed, direct estimation of a multivariate regression surface is limited by difficulties such as curse of dimensionality, interpretation and visualization. A natural way around these problems is to generalize ordinary multiple linear regression to allow arbitrary additive component functions, as in

$$m(\mathbf{x}) = c + \sum_{j=1}^{d} f_j(x_j), \tag{2.6}$$

where c is a constant and the f_j are univariate smooth functions. This model combines flexible nonparametric modeling of multidimensional inputs with a statistical precision that is typical for a one-dimensional predictor. Consider the estimation of the general nonparametric regression function $m(X) = E(Y|X)$. Stone (1985) showed that the optimal convergence rate of estimating $m(\cdot)$ is $n^{-r/(2r+d)}$ with r an index of smoothness of $m(\cdot)$. Thus, a high value of d leads to a slow rate of convergence. He also proved that for an additive regression function, the optimal rate of convergence is identical to that of the one-dimensional smoother, which is $n^{-r/(2r+1)}$. To avoid free constants in the functions and hence to ensure identifiability, we usually require that $E[f_j(X_j)] = 0$ for $1 \leq j \leq d$. This implies that $E(Y) = c$.

Breiman and Friedman (1985) and Buja, Hastie and Tibshirani (1989) proposed the iterative backfitting procedure to estimate the additive components. These methods have been evaluated on numerous data sets and have been refined considerably since their appearance. They iteratively calculate one-dimensional smoothers until some convergence criterion is satisfied. Due to their iterative nature, the theoretical analysis of this approach was eluded until Opsomer and Ruppert (1997) and

Opsomer (2000). The authors provided conditional mean squared error expressions under rather strong conditions on the smoothing matrices and design. In addition, Linton, Mammen and Nielsen (1998) established a central limit theorem for a modified form of backfitting which uses a bivariate integration step as well as the iterative updating of the other methods. The classical backfitting approach is presented in detail in Section 2.2.1.1.

More recently, a noniterative method for estimating marginal effects was introduced by Tjøstheim and Auestab (1994b) and Linton and Nielsen (1995). The idea is to estimate first a multidimensional functional of $m(\cdot)$ and then use marginal integration to obtain the marginal effects. If the regression function $m(\cdot)$ is indeed additive, the marginal integration estimator yields the functions $f_j(\cdot)$ up to a constant. The procedure is explicitly defined and its asymptotic distribution is easily derived. It has been extended to a number of other contexts like estimation of generalized additive models (Linton and Härdle, 1996; Yang, Sperlich and Härdle, 2003), derivative estimation (Severance-Lossin and Sperlich, 1997), dependent variable transformation models (Linton, Chen, Wang and Härdle, 1997), econometric time series models (Masry and Tjøstheim, 1995, 1997), and interactive additive models (Sperlich, Tjøstheim and Yang, 2002), etc.. We will introduce this estimator in Section 2.2.1.2.

Other approaches for fitting additive models were proposed more recently:

the smooth backfitting estimate by Mammen, Linton and Nielsen (1999); the local quasi-differencing approach of Christopeit and Hoderlein (2003) and the two-step procedures of Horowitz, Klemela and Mammen (2006). We will focus on backfitting and marginal integration and omit the details of these latter methods.

2.2.1.1 Backfitting

The backfitting procedures are widely used to estimate the additive models in (2.6). However, the iterative nature of the algorithm leads to additional difficulties for developing asymptotic theory. Moreover, the final estimates may depend on the starting values or the convergence criterion. Since its first introduction, this method has been refined considerably and extended to more complicated models. We will focus on the classical backfitting approach proposed by Buja, Hastie and Tibshirani (1989).

Under identifiability conditions, if the additive model is true, we have

$$E[Y - c - \sum_{j \neq k} f_j(X_j)|X_k] = f_k(X_k), \tag{2.7}$$

for $k = 1, \cdots, d$. This relationship motivates an iterative algorithm for computing all univariate functions $f_1, \cdots, f_d$. For given c and given functions $f_j, j \neq k$, the function f_k can be obtained via a univariate regression fit based on the observations $\{(X_{ik}, Y_i),\ i = 1, \cdots, n\}$. Denote the univariate smoother of f_k by $\mathbf{S}_k$. Note that any univariate regression smoothing technique can be used. The resulting estimate has to be centered to meet the identifiability condition

$$\hat{f}_k^*(\cdot) = \hat{f}_k(\cdot) - \frac{1}{n}\sum_{i=1}^{n} \hat{f}_k(X_{ik}). \tag{2.8}$$

An initial choice of the univariate functions, say f_k^0, is needed as well as an iteration scheme. Then the so-called backfitting algorithm is as follows:

Table 2.2 Backfitting Algorithm

Initialization	$\hat{c} = \bar{Y}$, $\hat{f}_k^{(0)} = f_k^0$, for $k = 1, \ldots, d$
Repeat	for each $k = 1, \cdots d$ the cycles:
	$\hat{f}_k = \mathbf{S}_k\{Y - \hat{c} - \sum_{j \neq k} \hat{f}_j(X_j) \mid X_k\}$
Until	convergence.

The equation (2.7) leads to the matrix representation

$$\begin{pmatrix} \mathcal{I} & \mathcal{P}_1 & \ldots & \mathcal{P}_1 \\ \mathcal{P}_2 & \mathcal{I} & \ldots & \mathcal{P}_2 \\ \vdots & & \ddots & \vdots \\ \mathcal{P}_d & \ldots & \mathcal{P}_d & \mathcal{I} \end{pmatrix} \begin{pmatrix} f_1(X_1) \\ f_2(X_2) \\ \vdots \\ f_d(X_d) \end{pmatrix} = \begin{pmatrix} \mathcal{P}_1 Y \\ \mathcal{P}_2 Y \\ \vdots \\ \mathcal{P}_d Y \end{pmatrix}$$

with the conditional expectation operator $\mathcal{P}_k(\cdot) = E(\cdot|X_k)$. Analogous to above, let $\mathbf{S}_k$ be a $n \times n$ smoother matrix, which yields an $n \times 1$ estimate $\mathbf{S}_k\mathbf{Y}$ of $\{E(Y_1|X_{1k}), \cdots, E(Y_n|X_{nk})\}^T$ when applied to the response vector $\mathbf{Y} = (Y_1, \cdots, Y_n)^T$. Replacing the operator $\mathcal{P}_k$ by the smoother $\mathbf{S}_k$, we obtain a system of equations

$$\begin{pmatrix} \mathbf{I} & \mathbf{S}_1 & \dots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \dots & \mathbf{S}_2 \\ \vdots & & \ddots & \vdots \\ \mathbf{S}_d & \dots & \mathbf{S}_d & \mathbf{I} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{f}}_1 \\ \hat{\mathbf{f}}_2 \\ \vdots \\ \hat{\mathbf{f}}_d \end{pmatrix} = \begin{pmatrix} \mathbf{S}_1\mathbf{Y} \\ \mathbf{S}_2\mathbf{Y} \\ \vdots \\ \mathbf{S}_d\mathbf{Y} \end{pmatrix}$$

which we can write more compactly as

$$\mathbf{A}\hat{\mathbf{f}} = \mathbf{B}\mathbf{Y} \tag{2.9}$$

where $\mathbf{A}$ and $\mathbf{B}$ are block matrices consisting of identity matrices $\mathbf{I}$ and smoothing operators $\mathbf{S}_k$. This system of equations is known as the normal equations of the additive model. In principle, the system (2.9) could be solved exactly, but the exact solution is hardly feasible if nd is large. Furthermore, the matrix $\mathbf{A}$ on the left-hand side is often not regular and thus the system can not be solved directly. As a consequence, the backfitting (Gauss-Seidel) procedure described above is used to solve these equations.

Opsomer and Ruppert (1997) and Opsomer (2000) investigated the statistical properties of backfitting. Mammen, Linton and Nielsen (1999) found a way to modify backfitting and prove consistency and calculate the asymptotics under weaker conditions. However, we do not describe it here since we emphasize the marginal integration method. Backfitting converges fast and is popular in fitting an additive model, partially due to the availability of the software, for example, function $gam()$ in R and S-plus.

2.2.1.2 Marginal Integration

The marginal integration estimator is based on an integration idea, based on the following observation. Let $X_{\underline{j}}$ denote the vector of all independent variables but X_j, i.e., $X_{i\underline{j}} = (X_{i1}, \ldots, X_{i(j-1)}, X_{i(j+1)}, \ldots, X_{id})$ and $\varphi_j, \varphi_{\underline{j}}$ are the marginal densities of X_j and $X_{\underline{j}}$, respectively. When $m(\cdot)$ satisfies the additive structure of (2.6),

$$F_j(x_j) = \int_{R^{d-1}} m(x)\varphi_{\underline{j}}(x_{\underline{j}})dx_{\underline{j}} = c + f_j(x_j), \tag{2.10}$$

by applying the identifiability condition. In case of additivity, the marginal effect on the left-hand side is the additive component function f_j plus a constant. This relation suggests to estimate first the function $m(\cdot)$ with a multidimensional presmoother $\hat{m}$ and then integrate out the other variables different from X_j. The most convenient way to estimate is to replace the integral in (2.10) by averaging over the directions not of interest (i.e., $X_{\underline{j}}$). It results in

$$\hat{F}_j(x_j) = \frac{1}{n}\sum_{i=1}^{n} \hat{m}(x_j, X_{i\underline{j}}).$$

As the constant c can be estimated consistently at rate $n^{-1/2}$ by the sample mean $\bar{Y}$, a possible estimate for f_j is

$$\hat{f}_j(x_j) = \frac{1}{n}\sum_{i=1}^{n} \hat{m}(x_j, X_{i\underline{j}}) - \bar{Y}. \tag{2.11}$$

In principle, the pre-estimator $\hat{m}$ could be any multivariate nonparametric estimator. The literature on marginal integration usually employs the kernel regression estimator (Linton and Nielsen, 1995) or local polynomial regression estimator (Severance-Lossin and Sperlich, 1999). Hence the estimator for the entire regression function is

$$\tilde{m}(x) = \hat{c} + \sum_{j=1}^{d} \hat{f}_j(x_j).$$

The marginal integration estimator was first proposed by Linton and Nielsen (1995), who derived its asymptotic properties for $d = 2$. The asymptotic distribution for arbitrary but finite dimension d was analyzed by Linton and Härdle (1996), based on the Nadaraya-Watson kernel estimation for the pre-smoother $\hat{m}$. Severance-Lossin and Sperlich (1999) proposed a methodology to estimate the derivatives for additive functions. They suggest the application of a local polynomial pilot estimator restricted to the direction of interest with the effect that more information remains in the constant. Also, by using local polynomial regression instead of kernel regression, the estimator is design adaptive in the sense that the bias is independent of the design density.

Linton and Nielsen (1995), Chen et al. (1995), Fan, Härdle and Mammen (1998), and Severance-Lossin and Sperlich (1999) have given conditions under which a variety of estimators based on the marginal integration idea converge at rate $n^{-2/5}$ and are asymptotically normal. However, the marginal integration estimator suffers from a form of the curse of dimensionality in that more derivatives of the f_j are needed to achieve the one-dimensional convergence rate as the dimension of X increases. As an example, the estimator in Fan, Härdle and Mammen (1998), which imposes the weakest smoothness conditions of any existing marginal integration estimator, still requires more than two derivatives when $d \geq 5$.

Linton (1997) argued that the empirical marginal integration map is not an orthogonal projection from the Hilbert space of functions of $\mathbf{X}$ to the subspace of additive functions. This implies that the integration method is not very efficient in estimating m and its components. To overcome this drawback, he suggested calculating starting values via marginal integration and then apply a one-step backfitting iteration. The interpretation of these estimation results is not clear, especially when the additivity assumption is violated. Another unpublished proposal to improve efficiency, also from a computational point of view, is due to Hengartner (1996). He tackles the problem of pilot estimator choice and comes out favoring a so-called internalized estimator.

Hengartner and Sperlich (2005) found a way to modify the marginal integration estimator so as to overcome the curse of dimensionality. To describe the method, let $f_{-1}(x_{-1}) = f_2(x_2) + \cdots + f_d(x_d)$ and let p_1 and p_{-1} be sufficiently smooth density functions on $\mathbb{R}$ and $\mathbb{R}^{d-1}$, respectively. The idea is to use the identifiability conditions

$$\int f_1(x_1)p_1(x_1)dx_1 = 0$$

and

$$\int f_{-1}(x_{-1})p_{-1}(x_{-1})dx_{-1} = 0$$

instead of $E[f_j(X_j)] = 0$. These conditions make it possible to use the smoothness of p_1 and p_{-1} to reduce the bias of the marginal integration estimator instead of using the smoothness of the f_j's. Thus, the f_j's do not need to be as smooth as required in the original integration methods, thereby avoiding the curse of dimensionality.

2.2.1.3 Bandwidth Choice

The choice of an appropriate smoothing parameter is always a crucial and difficult point in nonparametric and semiparametric settings. All the estimation methods for additive models discussed so far work only when smoothing parameter or bandwidth values are selected for each dimension beforehand.

In case of backfitting, the selection of d different smoothing parameters is needed. For $d \leq 2$, this is usually done by optimizing a bandwidth selection criterion, for example, cross validation, by a full grid search. For higher dimensions, however, this grid search approach rapidly becomes computationally intensive. Gu and Wahba (1991) described an efficient algorithm for minimizing GCV with smoothing splines. Unfortunately, their algorithm requires $O(n^3)$ computations and does not carry over directly to kernel-based smoothers. Hastie and Tibshirani (1990) suggested the BRUTO algorithm, which minimizes the GCV criterion over one bandwidth when the others are kept fixed. This is repeated sequentially for all bandwidths and the components are chosen as next step for the algorithm. BRUTO requires $O(n)$ computations but could be slow to converge in practice if the covariates show significant concurvity. There is instead a plug-in approach for the backfitting procedure when local linear fitting is applied (Opsomer and Ruppert, 1998). This method is also computationally difficult, even though the computational burden does not increase as fast as for GCV when the dimension d increases.

Using the marginal integration method instead of iterative procedures does not circumvent the computationally expensive situation. For each dimension, the integration estimator even requires one to choose two bandwidths: h_1 for the direction of interest and h_2 for the nuisance directions. Although cross validation can also be implemented for this method, the standard choices are a simple rule of thumb as in Linton and Nielsen (1995) and plug-in techniques suggested in Severance-Lossin and Sperlich (1999). Both methods give the bandwidth that minimizes MASE (averaged mean squared error) , the former approximating it by means of parametric pre-estimators, the latter by using nonparametric pre-estimators. We give here the formulas for the case of local linear pre-smoothers. The rule of thumb is

$$h_1 = \left\{ \frac{\tilde{\sigma}^2 \|K\|_2^2 (\max - \min)}{\mu_2(K) \left(\sum_{j=1}^d \hat{\beta}_j \right)^2} \right\}^{1/5} n^{-1/5}, \tag{2.12}$$

where $\mu_2(K) = \int u^2 K(u) du$, $\|K\|_2^2 = \int K^2(u) du$ and max and min are the sample maximum and minimum in the direction of interest, $\hat{\beta}_j$ is the coefficient of $x_j^2/2$ from a least squares regression of Y on a constant, x_j, $x_j^2/2$ and $x_j x_k$ for all $j, k = 1, \cdots, d$, $j < k$, while $\tilde{\sigma}^2$ is the average of the squared residuals from the same regression.

The plug-in method uses the following formula to calculate the asymptotically optimal bandwidth:

$$h_1 = \left\{ \frac{\|K\|_2^2 \int \sigma^2 \varphi_{\underline{j}}^2(x_{\underline{j}}) \varphi_j(x_j) \{\varphi(x)\}^{-1} dx_{\underline{j}} dx_j}{\mu_2^2(K) \int \{f_j''(x_j)\}^2 \varphi_j(x_j) dx_j} \right\}^{1/5} n^{-1/5}, \tag{2.13}$$

where $\varphi_{\underline{j}}$ and φ_j, φ are density functions for $X_{\underline{j}}, X_j$ and X respectively, σ^2 is the conditional variance and $f_j(\cdot)$ is the jth true component function. Note that this formula is not valid for h_2, for which the literature recommends undersmoothing. It turns out that this is not essential in practice. The reason is that the multiplicative term corresponding to h_2 is often already very small compared to the bias term corresponding to h_1.

2.2.1.4 Comparison of Backfitting and Marginal Integration

There are two studies comparing the two commonly used approaches. Nielson and Linton (1998) highlighted the theoretical difference and Sperlich, Linton and Härdle (1999) investigated the issue empirically. According to Nielson and Linton (1998), both marginal integration and backfitting can be viewed as an optimization of an integrated mean-squared-error criterion. The integration estimator minimizes the criterion with weighting given by an independent product measure, while backfitting uses weighting based on a joint empirical measure (joint density). The result of marginal integration therefore is correct independently of whether additivity holds or not. The main point is that backfitting is orthogonal projection of the regression into the additive space, whereas the marginal integration estimator always estimates the marginal impact of the explanatory variables taking into account possible correlation among them. The definite advantage of the integration estimator is that it is explicitly defined so it allows extensive studies on the asymptotic properties. However, marginal integration becomes inefficient with increasing correlation among the regressors. Linton (1997, 2000) proposed a two-step procedure that took the integration estimate as a first step and then did one backfitting iteration from that. This procedure is argued to be oracle efficient, i.e., as efficient as the infeasible estimate that is based on knowing all components but the one of interest.

Sperlich, Linton and Härdle (1999) have undertaken the most extensive simulation study so far, in which they tried to trace down the performance differences between the two methods for small samples and bivariate regression functions (in one example, for $d = 4$). They concluded that one cannot declare the superiority of one procedure over the other in general. Both estimators perform poorly in designs with increasing correlation although backfitting does perform slightly better. This is in line with Linton's theory (1997). Backfitting works better at boundary points and under data sparseness. The integration method is more capable of estimating the components as opposed to the regression function itself. Thus they can not be interpreted as competing estimators with the same aim.

Finally, it should be mentioned that choosing the bandwidths and smoothing parameters remains a troublesome issue in the context of additive models, no matter which method is used. This fact is likely to hamper the interpretation of comparative simulation studies.

2.2.2 Generalized Additive Models (GAMs)

Analogous to the way that linear models are extended to Generalized Linear Models (GLMs), the class of generalized additive models was introduced in a series of papers by Hastie and Tibshirani (1986a, 1986b, 1987) and Stone (1986). They are described in detail in Hastie and Tibshirani (1990). GAMs retain an important feature of GLMs, namely, additivity of the predictors, but the predictor effects are modeled by arbitrary smooth functions f_js.

We say that $m(x)$ has a generalized additive structure if

$$G\{m(x)\} = c + \sum_{j=1}^{d} f_j(x_j) \tag{2.14}$$

for some known "link function" G. The assumptions concerning identifiability $E[f_j(X_j)] = 0$ remain same. Additive models are just a special case with the trivial link function $G =$ indentity. Note that (2.14) is a partial model specification without restricting in any way the variance or other aspects of the conditional distribution $\mathcal{L}(Y|X)$. A full model specification is to assume that $\mathcal{L}(Y|X)$ belongs to an exponential family with known link function G and mean m. This class of models is called generalized additive by Hastie and Tibshirani (1990). In some respects, we prefer the partial model specification as this flexibility is a relevant consideration for many data sets when there is overdispersion or heterogeneity.

Stone (1986) showed that for GAMs, the optimal rate of estimating $m(\cdot)$ is the one-dimensional rate of convergence, for example, $n^{-2/5}$ for twice continuously differentiable functions.

The backfitting procedure in conjunction with Fisher scoring is widely used to estimate GAMs. Linton and Härdle (1996) extended the marginal integration method to the context of GAMs. We will give a brief sketch of the two methods in the next two subsections.

2.2.2.1 Backfitting in GAMs (Local Scoring)

In models with a nontrivial link function G, the response Y is not directly related to the index functions. Instead an adjusted dependent variable Z is constructed, and an iteration analogous to the Fisher scoring in the iterative re-weighted least square (IRLS) algorithm for GLMs will be used as an "outer" iteration. The "inner" iteration is a backfitting procedure which fits the index functions instead of linear components in GLMs.

The final algorithm given in Hastie and Tibshirani (1990) is presented in Table 2.3.

The theoretical properties of these iterative procedures are very complicated and intractable. The situation is different for the marginal integration methods, which are sketched in next subsection.

2.2.2.2 Marginal Integration in GAMs

As we have seen for the additive models, the component function $f_j(x_j)$, $j = 1, \cdots, d$, in model (2.14) is equal to the functional

$$F_j(x_j) = \int G\{m(x_j, x_{\underline{j}})\}\varphi_{\underline{j}}(x_{\underline{j}})dx_{\underline{j}} \tag{2.15}$$

up to a constant, due to the additive structure and the identifiability conditions. Linton and Härdle (1996) proposed replacing m in (2.15) by a multivariate Nadaraya-Watson kernel estimator $\hat{m}$ and estimating (2.15) by its sample version

$$\hat{F}_j(X_j) = \frac{1}{n}\sum_{i=1}^{n} G\{\hat{m}(x_j, X_{i\underline{j}}))\}.$$

Thus we obtain an explicit expression for the marginal integration estimator:

$$\tilde{m}(x) = G^{-1}\{\sum_{j=1}^{d} \hat{f}_j(x_j) + \hat{c}\},$$

where

$$\hat{f}_j(x_j) = \frac{1}{n}\sum_{i=1}^{n} G\{\hat{m}(x_j, X_{i\underline{j}})\} - \hat{c}$$

and $\hat{c} = d^{-1}n^{-1}\sum_{j=1}^{d}\sum_{i=1}^{n}\hat{F}_j(X_{ij})$.

Table 2.3 Local Scoring Algorithm

```
Initialize  $c^0 = G(\bar{Y}), f_k^0 = 0$, for $k = 1, \cdots, d$
Repeat      Construct an adjusted dependent variable
            $Z_i = \eta_i^0 + (Y_i - \mu_i^0)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)_0$
            with $\eta_i^0 = c^0 + \sum_{k=1}^d f_k^0(X_{ik})$ and $\mu_i^0 = G^{-1}(\eta_i^0)$.
            Construct weights
            $w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)_0^2 (V_i^0)^{-1}$
            where $V_i = V(\mu_i)$ with $V(\cdot)$ the variance function of $Y$
              in a generalized model.
            Fit a weighted additive model to $Z_i$, to obtain estimated function $f_k^1$,
              additive predictor $\eta^1$, and fitted values $\mu_i^1$.
            Compute the convergence criterion
            $\Delta(\eta^1, \eta^0) = \frac{\sum_{k=1}^d \|f_k^1 - f_k^0\|}{\sum_{k=1}^d \|f_k^0\|}$
Until       $\Delta(\eta^1, \eta^0)$ is below some small threshhold.
```

In contrast to the backfitting procedure, it is easier to analyze the asymptotics of the marginal integration method. Linton and Härdle (1996) showed that the rate of convergence of $\hat{m}$ is not influenced by the curse of dimensionality. The obtained rate of $n^{-2/5}$ is the same as that derived by Stone (1986) for one-dimensional regression function. Yang, Sperlich and Härdle (2003) extended the method and theory for derivative estimation and variable selection problems by using a local polynomial pre-smoother. In practice, one often needs to identify relevant predictors with respect to a response variable. This problem has been addressed by Härdle and Korostelev (1996).

Despite its asymptotic optimality, one has to be cautious about using the algorithm, especially for $d > 2$. In practice, covariates are more or less correlated, and sample sizes are small. However, comprehensive and extensive simulation studies do not yet exist for different combinations of regression and link functions and for various sample sizes. A theoretical comparison with backfitting is not possible due to the lack of asymptotic results for backfitting in GAMs.

Chapter 3

Theory and Methods

A weakness of GAM given in (1.4) is that this model completely ignores the fact that the functional form of the effect of an explanatory variable often varies according to the values of one or more of the remaining variables, i.e., some interaction exists between regressors. In this thesis, we allow for second-order interactions, resulting in a model

$$G\{m(x)\} = c + \sum_{j=1}^{d} f_j(x_j) + \sum_{1 \leq j < k \leq d} f_{jk}(x_j, x_k).$$

In principle, we could also consider interaction terms of higher order, e.g. $f_{j,k,l}(x_j, x_k, x_l), \ldots,$ but this would gradually bring back problems of visualization and interpretation. Furthermore, the advantage of avoiding the curse of dimensionality would get lost step by step. Therefore we will restrict ourselves to the case of only second-order interactions.

3.1 Basic Assumptions and Notations

Let (Y, X) be a random variable with X of dimension d and Y a scalar, and let the regression function be $m(x) = E(Y|X = x)$. We consider the model

$$G\{m(x)\} = c + \sum_{j=1}^{d} f_j(x_j) + \sum_{1 \le j < k \le d} f_{jk}(x_j, x_k) \tag{3.1}$$

for some known and monotone "link function" G, where $x = (x_1, x_2, \dots, x_d)^T$ are the d-dimensional predictor vector, c is an unknown constant, $\{f_j\}_{j=1}^{d}$ are unknown univariate functions, and $\{f_{jk}\}_{1 \le j < k \le d}$ is a set of unknown bivariate functions. Clearly, the representation given in (3.1) is not unique, and constraints must be placed on the main effects f_j and interaction terms f_{jk} by imposing the identifiability conditions

$$E[f_j(x_j)] = \int f_j(x_j)\varphi_j(x_j)dx_j = 0, \text{ for } j = 1, 2, \dots, d, \tag{3.2}$$

and for all $1 \le j < k \le d$,

$$\int f_{jk}(x_j, x_k)\varphi_j(x_j)dx_j = \int f_{jk}(x_j, x_k)\varphi_k(x_k)dx_k = 0, \tag{3.3}$$

with $\{\varphi_j(\cdot)\}_{j=1}^{d}$ being marginal densities of the x_j's.

Note that the conditions (3.2) and (3.3) do not represent restrictions on our model, since if a function in (3.1) does not satisfy these conditions, one can easily shift it in the vertical direction so that it conforms to these same constraints. Moreover, all models of the form (3.1) are equivalent to exactly one model satisfying (3.2) and (3.3). In the sequel we assume each f_j and f_{jk} satisfy (3.2) and (3.3), unless otherwise stated.

Denote the variable $X = (X_j, X_{\underline{j}})$ to highlight a particular direction j, where $X_{\underline{j}} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_d)$ is the $(d-1)$-dimensional random variable obtained by removing X_j from X. Let $X_{\underline{jk}}$ be defined analogously. We write $X = (X_j, X_k, X_{\underline{jk}})$ to highlight the directions represented by the j and k coordinates in a d-dimensional space. The marginal densities of X_j, $X_{\underline{j}}$, $X_{\underline{jk}}$ and the joint density of X are denoted by $\varphi_j(x_j), \varphi_{\underline{j}}(x_{\underline{j}}), \varphi_{\underline{jk}}(x_{\underline{jk}})$ and $\varphi(x)$, respectively.

We introduce here notations for sets of indices. Denote by D_j the subset of $\{1, 2, \ldots, d\}$ with j removed,

$$\begin{aligned} D_{jj} &= \{(l,m) | 1 \leq l < m \leq d, l \in D_j, m \in D_j\} \quad \text{and} \\ D_{jk} &= \left\{(l,m) | 1 \leq l < m \leq d, l \in D_j \bigcap D_k, m \in D_j \bigcap D_k \right\}. \end{aligned}$$

We define by marginal integration

$$F_j(x_j) = \int G\left\{m(x_j, x_{\underline{j}})\right\} \varphi_{\underline{j}}(x_{\underline{j}}) dx_{\underline{j}}, \tag{3.4}$$

for $1 \leq j \leq d$ and

$$F_{jk}(x_j, x_k) = \int G\left\{m(x_j, x_k, x_{\underline{jk}})\right\} \varphi_{\underline{jk}}(x_{\underline{jk}}) dx_{\underline{jk}}, \tag{3.5}$$

for every pair $1 \leq j < k \leq d$. We also define

$$f^*_{jk}(x_j, x_k) = f_{jk}(x_j, x_k) + c_{jk}, \tag{3.6}$$

where $c_{jk} = \int f_{jk}(u,v)\varphi_{jk}(u,v)dudv$, for every pair $1 \leq j < k \leq d$.

Some simple calculations show that $F_j(\cdot)$ is equal to f_j up to an additive constant. Analogously, $(F_{jk} - F_j - F_k)(\cdot)$ is equal to f_{jk} up to an additive constant. Actually, following from the definitions of D_{jj}, D_{jk}, c_{jk} and F_j, F_{jk}, the constraints (3.2) and (3.3) entail the lemma below.

Lemma 3.1.1. *For model (3.1) the following three equations for the marginals hold:*

1. $F_j(x_j) = f_j(x_j) + c + \sum_{(l,m)\in D_{jj}} c_{lm}$,
2. $F_{jk}(x_j, x_k) = f_{jk}(x_j, x_k) + f_j(x_j) + f_k(x_k) + c + \sum_{(l,m)\in D_{jk}} c_{lm}$,
3. $F_{jk}(x_j, x_k) - F_j(x_j) - F_k(x_k) + \int G\{m(x)\}\varphi(x)dx = f_{jk}(x_j, x_k) + c_{jk}$.

By item 3 of Lemma 3.1.1, f^*_{jk} defined in (3.6) satisfies the equation

$$f^*_{jk}(x_j, x_k) = F_{jk}(x_j, x_k) - F_j(x_j) - F_k(x_k) + \int G\{m(x)\}\varphi(x)dx \qquad (3.7)$$

Rather than f_j and f_{jk}, we will work with more convenient quantities F_j and f^*_{jk} and study their limiting behavior. As shown in Lemma 3.1.1, they can be identified with f_j and f_{jk} up to an additive constant.

3.2 Marginal Integration Estimation

Following the idea of Linton and Härdle (1996), we estimate the marginal influence F_j and F_{jk} by replacing the expectations by averages and the function m by an appropriate pre-estimator $\hat{m}$,

$$\hat{F}_j(x_j) = \frac{1}{n}\sum_{i=1}^{n} G\left\{\hat{m}(x_j, X_{i\underline{j}})\right\}, \tag{3.8}$$

$$\hat{F_{jk}}(x_j, x_k) = \frac{1}{n}\sum_{i=1}^{n} G\left\{\hat{m}(x_j, x_k, X_{i\underline{jk}})\right\}, \tag{3.9}$$

where $X_{i\underline{j}}(X_{i\underline{jk}})$ is the ith observation of X with $X_j(X_j$ and $X_k)$ removed.

To compute $\hat{m}$, we employ a special kind of multi-dimensional local polynomial kernel estimator. For details of the estimator, see Ruppert and Wand (1994) for the general case and Severance-Lossin and Sperlich (1999) in the context of marginal integration. The special pre-smoother proposed by Severance-Lossin and Sperlich (1999) is a local polynomial regression of degree p in the direction of interest and degree zero (local constant) in the nuisance directions.

Let $K(\cdot)$ and $L(\cdot)$ be kernel functions and denote $K_{h1}(u) = h_1^{-1}K(u/h_1)$ and $L_{h2}(u) = h_2^{-1}L(u/h_2)$ with bandwidths h_1 and h_2. We use the same letters K and L to represent kernel functions of varying dimensions for ease of notation. It will be clear from the context what the dimensions are in every specific case. For any kernel $K(\cdot)$, define $\mu_q(K) = \int u^q K(u)du$ and $\|K\|_2^2 = \int K^2(u)du$.

For ease of presentation, by setting $p = 1$, we consider the problem of minimizing

$$\sum_{i=1}^{n} \left[Y_i - \beta_0 - \beta_1(X_{ij} - x_j)\right]^2 K_{h1}(X_{ij} - x_j)L_{h2}(X_{i\underline{j}} - X_{l\underline{j}}) \tag{3.10}$$

for each fixed l. With $e_1 = (1, 0, \cdots, 0)^T$, we define the pre-estimator in (3.8):

$$\hat{m}(x_j, X_{l\underline{j}}) = e_1^T (Z_j^T W_{l,j} Z_j)^{-1} Z_j^T W_{l,j} Y, \tag{3.11}$$

where

$$Z_j = \begin{pmatrix} 1 & X_{1j} - x_j \\ \vdots & \vdots \\ 1 & X_{nj} - x_j \end{pmatrix}, \quad W_{l,j} = \begin{pmatrix} \frac{1}{n} K_{h_1}(X_{1j} - x_j) L_{h_2}(X_{1\underline{j}} - X_{l\underline{j}}) & 0 & \dots & 0 \\ 0 & \frac{1}{n} K_{h_1}(X_{2j} - x_j) L_{h_2}(X_{2\underline{j}} - X_{l\underline{j}}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{n} K_{h_1}(X_{nj} - x_j) L_{h_2}(X_{n\underline{j}} - X_{l\underline{j}}) \end{pmatrix}.$$

Analogously, the pre-estimator $\hat{m}(x_j, x_k, X_{l\underline{jk}})$ in (3.9) results from the problem of minimizing

$$\sum_{i=1}^{n} \left[Y_i - \beta_0 - \beta_1(X_{ij} - x_j) - \beta_2(X_{ik} - x_k)\right]^2 K_{h_1}(X_{ij} - x_j, X_{ik} - x_k) L_{h_2}(X_{i\underline{jk}} - X_{l\underline{jk}}) \tag{3.12}$$

for each fixed l.

Accordingly, we define

$$\hat{m}(x_j, x_k, X_{l\underline{jk}}) = e_1^T (Z_{jk}^T W_{l,jk} Z_{jk})^{-1} Z_{jk}^T W_{l,jk} Y, \tag{3.13}$$

where

$$Z_{jk} = \begin{pmatrix} 1 & X_{1j} - x_j & X_{1k} - x_k \\ \vdots & \vdots & \vdots \\ 1 & X_{nj} - x_j & X_{nk} - x_k \end{pmatrix}$$

and

$$W_{l,jk} = diag\left\{\frac{1}{n}K_{h_1}(X_{ij} - x_j, X_{ik} - x_k)L_{h_2}(X_{i\underline{jk}} - X_{l\underline{jk}})\right\}_{i=1}^{n}.$$

It should be noted that $\hat{m}(x_j, X_{l\underline{j}})$ is a locally linear estimator in the direction j and a locally constant one in the other directions. Similarly, $\hat{m}(x_j, x_k, X_{l\underline{jk}})$ gives a locally linear smoother for the directions j, k and a locally constant one for the nuisance directions.

We are now have almost everything at hand to estimate the interaction term $f^*_{jk}(x_j, x_k)$ in (3.7). The term $\int G\{m(x)\}\varphi(x)dx$ can be estimated empirically by $\frac{1}{n}\sum_{i=1}^{n} G\{\hat{m}(X_i)\}$, where $\hat{m}(X_i)$ is a multivariate regression smoother at the ith sample point. Therefore,

$$\hat{f}^*_{jk}(x_j, x_k) = \hat{F}_{jk}(x_j, x_k) - \hat{F}_j(x_j) - \hat{F}_k(x_k) + \frac{1}{n}\sum_{i=1}^{n} G\left\{\hat{m}(X_i)\right\}. \qquad (3.14)$$

Consequently, the combined regression estimator $\tilde{m}(x)$ of $m(x)$ is given by

$$\tilde{m}(x) = G^{-1}\left\{\sum_{j=1}^{d}\hat{F}_j(x_j) + \sum_{1\leq j<k\leq d}\hat{f}^*_{jk}(x_j, x_k) - (d-1)\frac{1}{n}\sum_{i=1}^{n} G\{\hat{m}(X_i)\}\right\}. \qquad (3.15)$$

To establish the asymptotics for the estimators proposed above, we need the following assumptions:

(A01) The kernel functions $K(\cdot)$ and $L(\cdot)$ are symmetric, compactly supported and Lipschitz continuous; $K(\cdot)$ is nonnegative satisfying $\int K(u)du = 1$ while the $(d-1)$-dimensional kernel $L(\cdot)$ is a product of univariate kernels $L(u)$ of order $q \geq 2$.

(A02) Bandwidths satisfy $nh_1h_2^{2(d-1)}/\ln^2 n \to \infty, h_2^q/h_1^2 \to 0$, and $h_1 = \beta n^{-1/5}$, where β is a constant.

(A3) The functions f_j, f_{jk} have bounded Lipschitz continuous derivatives of order q.

(A4) The variance function $\sigma^2(\cdot)$ is bounded and Lipschitz continuous.

(A5) The density functions $\varphi, \varphi_{\underline{j}}, \varphi_{\underline{jk}}$ are uniformly bounded away from zero and infinity and have bounded Lipschitz continuous second derivatives.

(A6) G is uniformly bounded away from zero and infinity over its compact support, and has bounded Lipschitz continuous second derivative.

Theorem 3.2.1. *Under assumptions* $(A01), (A02)$ *and* $(A3)-(A6)$, *for any* $1 \le j \le d$,

$$\sqrt{nh_1}\left\{\hat{F}_j(x_j) - F_j(x_j) - h_1^2 b_j(x_j)\right\} \xrightarrow{\mathcal{D}} N\left\{0, v_j(x_j)\right\}, \tag{3.16}$$

where

$$b_j(x_j) = \frac{\mu_2(K)}{2}\int\left\{(G' \circ m)\frac{\partial^2 m}{\partial x_j^2}\right\}(x_j, x_{\underline{j}})\varphi_{\underline{j}}(x_{\underline{j}})dx_{\underline{j}}, \tag{3.17}$$

and

$$v_j(x_j) = \|K\|_2^2 \int\left\{\frac{(G' \circ m)^2\sigma^2}{\varphi}\right\}(x_j, x_{\underline{j}})\varphi_{\underline{j}}^2(x_{\underline{j}})dx_{\underline{j}}. \tag{3.18}$$

Regarding the asymptotics of the estimator $\hat{f}^*_{jk}(x_j, x_k)$ given in (3.6), we have to adjust the assumptions on kernel functions and bandwidths for the bivariate problem.

(A1) The kernels $K(\cdot)$ and $L(\cdot)$ are symmetric, compactly supported and Lipschitz continuous; the bivariate kernel $K(\cdot)$ is a product kernel such that $K(u, v) = K(u)K(v)$ with $\int K(u)du = 1$ and the $(d-2)$-dimensional kernel $L(\cdot)$ is also a product of $d-2$ univariate kernels $L(u)$ of order $q \geq 2$.

(A2) Bandwidths satisfy $nh_1^2 h_2^{2(d-2)} / \ln^2 n \to \infty, h_2^q/h_1^2 \to 0$, and $h_1 = \beta n^{-1/6}$.

Theorem 3.2.2. *Under assumptions* $(A1) - (A6)$*, for any* $1 \leq j < k \leq d$,

$$\sqrt{nh_1^2}\left\{\hat{f}^*_{jk}(x_j, x_k) - f^*_{jk}(x_j, x_k) - h_1^2 B_{jk}(x_j, x_k)\right\} \xrightarrow{\mathcal{D}} N\{0, V_{jk}(x_j, x_k)\}, \tag{3.19}$$

where

$$\begin{aligned} B_{jk}(x_j, x_k) = \frac{\mu_2(k)}{2} \Bigg[& \int \left\{ (G' \circ m) \left(\frac{\partial^2 m}{\partial x_j^2} + \frac{\partial^2 m}{\partial x_k^2} \right) \right\} (x_j, x_k, x_{\underline{jk}}) \varphi_{\underline{jk}}(x_{\underline{jk}}) dx_{\underline{jk}} \\ & - \int \left\{ (G' \circ m) \frac{\partial^2 m}{\partial x_j^2} \right\} (x_j, x_{\underline{j}}) \varphi_{\underline{j}}(x_{\underline{j}}) dx_{\underline{j}} \\ & - \int \left\{ (G' \circ m) \frac{\partial^2 m}{\partial x_k^2} \right\} (x_k, x_{\underline{k}}) \varphi_{\underline{k}}(x_{\underline{k}}) dx_{\underline{k}} \Bigg] \end{aligned} \tag{3.20}$$

and

$$V_{jk}(x_j, x_k) = (\|K\|_2^2)^2 \int \left\{ \frac{(G' \circ m)^2 \sigma^2}{\varphi} \right\} (x_j, x_k, x_{\underline{jk}}) \varphi^2_{\underline{jk}}(x_{\underline{jk}}) dx_{\underline{jk}}. \tag{3.21}$$

The next result states the limiting distribution of the combined regression estimator $\tilde{m}(x)$. Its proof essentially follows the proofs of the two previous theorems, the delta method and the fact that asymptotically the covariances between individual estimators are of smaller order than the variances of the estimator of each component function and thus negligible.

Theorem 3.2.3. *Under assumptions* $(A1), (A3)-(A6)$ *and choosing bandwidths as in* $(A02)$ *and* $(A2)$ *for the one- and two-dimensional component functions,*

$$\sqrt{nh_1^2}\{\tilde{m}(x) - m(x) - h_1^2 B(x)\} \xrightarrow{\mathcal{D}} N\{0, V(x)\}, \tag{3.22}$$

where h_1 *is as in* $(A2)$,

$$B(x) = (G^{-1})' \left\{ c + \sum_{j=1}^{d} f_j(x_j) + \sum_{1 \le j < k \le d} f_{jk}(x_j, x_k) \right\} \cdot \sum_{1 \le j < k \le d} B_{jk}(x_j, x_k),$$

$$V(x) = \left[(G^{-1})' \left\{ c + \sum_{j=1}^{d} f_j(x_j) + \sum_{1 \le j < k \le d} f_{jk}(x_j, x_k) \right\} \right]^2 \cdot \sum_{1 \le j < k \le d} V_{jk}(x_j, x_k),$$

and $B_{jk}(\cdot)$ *and* $V_{jk}(\cdot)$ *are defined in Theorem 3.2.2.*

3.3 Testing for Interaction

Another main objective of this dissertation is to propose a direct test of second-order interaction. The null hypothesis $H_{jk}^0 : f_{jk} = 0$, namely, that there is no interaction between X_j and X_k for a given fixed pair $(j, k), 1 \le j < k \le d$, is considered for the model (3.1). The interest in testing whether an interaction function is significant at all is obvious as it may be an important step in a model selection procedure. It can also be regarded as a test for a pure GAM.

It has been shown that the function f_{jk} can be related to another function f^*_{jk} up to a constant. It turns out that f^*_{jk} is also a convenient substitute for f_{jk} in the testing problem, since $f^*_{jk}(x_j, x_k) \equiv 0$ is equivalent to $f_{jk}(x_j, x_k) \equiv 0$. This fact suggests the use of the following functional for testing of additivity of the jth and kth directions:

$$T = \int \hat{f}^{*2}_{jk}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k, \tag{3.23}$$

which is an estimate of $\int f^{*2}_{jk}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k$.

The next theorem will show that T is a suitable statistic for testing H^0_{jk}, by establishing its limiting distribution.

Theorem 3.3.1. *Under assumptions* $(A1) - (A6)$,

$$\begin{aligned} & nh_1 T - nh_1 \int f^{*2}_{jk}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k \\ & \quad - \frac{2\{K^{(2)}(0)\}^2}{h_1} \int \left\{ \frac{(G' \circ m)^2 \sigma^2}{\varphi} \right\} (x_j, x_k, x_{\underline{jk}})\varphi^2_{\underline{jk}}(x_{\underline{jk}})\varphi_{jk}(x_j, x_k)dx_j dx_k dx_{\underline{jk}} \\ & \quad - 2nh_1^3 \int f^*_{jk}(x_j, x_k)B_{jk}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j, dx_k \\ & \quad \xrightarrow{\mathcal{D}} N(0, \sigma^2_T), \end{aligned}$$

where

$$\sigma^2_T = 2\|K^{(2)}\|^4_2 \int \left[\int \left\{ \frac{(G' \circ m)^2 \sigma^2}{\varphi} \right\} (x_j, x_k, x_{\underline{jk}})\varphi^2_{\underline{jk}}(x_{\underline{jk}})dx_{\underline{jk}} \right]^2 \varphi^2_{jk}(x_j, x_k)dx_j dx_k,$$

$K^{(2)}$ *is the two-fold convolution of the kernel* K *and* B_{jk} *is defined in Theorem 3.2.2.*

Based on Theorem 3.3.1, the asymptotics of the test statistic under the null hypothesis is given in the following corollary.

Corollary 3.3.2. *Under* H^0_{jk}*,*

$$nh_1T - \frac{2\{K^{(2)}(0)\}^2}{h_1}\int\left\{\frac{(G'\circ m)^2\sigma^2}{\varphi}\right\}(x_j, x_k, x_{\underline{jk}})\varphi^2_{\underline{jk}}(x_{\underline{jk}})\varphi_{jk}(x_j, x_k)dx_jdx_kdx_{\underline{jk}} \xrightarrow{\mathcal{D}} N(0, \sigma^2_T).$$

Thus the test rule with the pre-specified significance level α is to reject H^0_{jk} if

$$nh_1T \geq \frac{2\{K^{(2)}(0)\}^2}{h_1}\int\left\{\frac{(G'\circ m)^2\sigma^2}{\varphi}\right\}(x_j, x_k, x_{\underline{jk}})\varphi^2_{\underline{jk}}(x_{\underline{jk}})\varphi_{jk}(x_j, x_k)dx_jdx_kdx_{\underline{jk}} + z_{1-\alpha}\sigma_T,$$

in which $z_{1-\alpha}$ is the upper $(1-\alpha)$ percentile of the standard normal distribution.

To make the test feasible, we need to get the critical values. Two standard ways are: (1) estimating the asymptotics of the test; (2) applying the wild bootstrap. However, it is well known that the first-order asymptotics derived in Theorem 3.3.1 does not give a very accurate description of the finite sample properties. Hjellvik, Yao and Tjøstheim (1998) showed that as many as several hundred thousand observations might be necessary for the asymptotic formulas to be reasonably accurate. This even gets worse when the limiting expressions also have to be estimated. As a consequence, wild bootstrap is employed for constructing the null distribution of the test functional, in case of a moderate sample size.

In practice, since the density φ_{jk} is unknown, T is approximated by its empirical average

$$\tilde{T} = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{jk}^{*2}(X_{ij}, X_{ik}).$$

The following theorem ensures that replacing T by $\tilde{T}$ does not affect the asymptotics and the test rule.

Theorem 3.3.3. *Under assumptions* $(A1) - (A6)$,

$$\begin{aligned} nh_1\tilde{T} - nh_1 & \int f_{jk}^{*2}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k \\ & - \frac{2\{K^{(2)}(0)\}^2}{h_1} \int \left\{ \frac{(G' \circ m)^2\sigma^2}{\varphi} \right\} (x_j, x_k, x_{\underline{jk}})\varphi^2_{\underline{jk}}(x_{\underline{jk}})\varphi_{jk}(x_j, x_k)dx_j dx_k dx_{\underline{jk}} \\ & - 2nh_1^3 \int f_{jk}^{*}(x_j, x_k)B_{jk}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j, dx_k \\ & \xrightarrow{\mathcal{D}} N(0, \sigma_T^2). \end{aligned}$$

We are also interested in the local power of the test against a sequence of alternatives converging to the null as the sample size grows.

Let S_{jk} be the support of the density function $\varphi_{jk}(\cdot)$. The second order Sobolev seminorm of a bivariate function $f_{jk}(x_j, x_k)$ is defined by

$$\|f_{jk}\|_{H^2(S_{jk})} = \sqrt{\sum_{u=0}^{2} \int_{S_{jk}} \left[\frac{\partial^2 f_{jk}(x_j, x_k)}{\partial^u x_j \partial^{2-u} x_j}\right]^2 dx_j dx_k}.$$

Denote by $\mathcal{B}_{jk}(M)$ the class of functions f_{jk} with bounded second order Sobolev seminorm,

$$\|f_{jk}\|_{H^2(S_{jk})} \leq M,$$

where M is a positive constant.

The next theorem enables one to identify a class of alternatives (indexed by n) such that our test will have asymptotic power one.

Theorem 3.3.4. *Assume G' is bounded away from zero. Under assumptions $(A1)-(A6)$, consider testing the null hypothesis*

$$H_0^{jk}: f_{jk}(x_j, x_k) \equiv 0$$

versus the local alternative

$$H_1^{jk}(a_n): f_{jk,n}(x_j, x_k) \in \mathcal{F}_{jk}(a_n),$$

where $\mathcal{F}_{jk}(a_n)$ is the class of alternatives

$$\left\{ f_{jk} \in \mathcal{B}_{jk}(M): \|f_{jk}\|_{L^2(S_{jk},\varphi_{jk})} = \sqrt{\int_{S_{jk}} f_{jk}^2(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k} \geq a_n \right\}$$

and $\{a_n\}$ is a sequence of constants satisfying $a_n^{-1} = o(nh_1 + h_1^{-2}) = o(n^{5/6})$ as $n \to \infty$. Denote by p_n the probability of rejecting H_0^{jk} in favor of the local alternative $H_1^{jk}(a_n)$. Then $\lim_{n\to\infty} p_n = 1$.

The theorem guarantees that the proposed test procedure is able to detect an interaction term of the magnitude $n^{-5/6}$ with power tending to one.

3.4 Proofs

3.4.1 Proof of Theorem 3.2.1

The following lemma was proved in Fan, Härdle and Mammen (1998). It has also been cited in Severance-Lossin and Sperlich (1999), and Sperlich, Tjøstheim and Yang (2002). Our proof makes use of this lemma.

Lemma 3.4.1. *Let $W_{l,j}, W_{l,jk}, Z_j, Z_{jk}$ be defined as in Section 3.2, then*

(i)

$$(H^{-1}Z_j^T W_{l,j} Z_j H^{-1})^{-1} = \frac{1}{\varphi(x_j, X_{l\underline{j}})} S^{-1}\left\{I + O_p(c_{1n})\right\},$$

and

(ii)

$$(H^{-1}Z_{jk}^T W_{l,jk} Z_{jk} H^{-1})^{-1} = \frac{1}{\varphi(x_j, x_k, X_{l\underline{jk}})} \mathcal{S}^{-1}\left\{I + O_p(c_{2n})\right\},$$

where

$$H = \begin{pmatrix} 1 & 0 \\ 0 & h_1 \end{pmatrix}, \quad S^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & \mu_2^{-1} \end{pmatrix}, \quad \mathcal{S}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \mu_2^{-1} & 0 \\ 0 & 0 & \mu_2^{-1} \end{pmatrix}$$

and $c_{1n} = h_1^2 + \sqrt{\ln n/(nh_1h_2^{d-1})}$, $c_{2n} = h_1^2 + \sqrt{\ln n/(nh_1^2h_2^{d-2})}$.

With this preliminary lemma, the proof of Theorem 3.2.1 starts below.

Proof.

$$
\begin{aligned}
&\hat{F}_j(x_j) - F_j(x_j) \\
&= \frac{1}{n}\sum_{i=1}^{n} G\left\{\hat{m}(x_j, X_{i\underline{j}})\right\} - \frac{1}{n}\sum_{i=1}^{n} G\left\{m(x_j, X_{i\underline{j}})\right\} + \frac{1}{n}\sum_{i=1}^{n} G\left\{m(x_j, X_{i\underline{j}})\right\} - F_j(x_j) \\
&= \frac{1}{n}\sum_{i=1}^{n} \left[G\left\{\hat{m}(x_j, X_{i\underline{j}})\right\} - G\left\{m(x_j, X_{i\underline{j}})\right\}\right] + O_p\left(n^{-\frac{1}{2}}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} G'\left\{\hat{m}(x_j, X_{i\underline{j}})\right\} \cdot \left[\hat{m}(x_j, X_{i\underline{j}}) - m(x_j, X_{i\underline{j}})\right] + R + O_p\left(n^{-\frac{1}{2}}\right)
\end{aligned}
$$

where

$$R = \frac{1}{2n}\sum_{i=1}^{n} G''(m_i^0) \cdot \left[\hat{m}(x_j, X_{i\underline{j}}) - m(x_j, X_{i\underline{j}})\right]^2$$

with m_i^0 intermediate between $\hat{m}(x_j, X_{i\underline{j}})$ and $m(x_j, X_{i\underline{j}})$.

We have the following by the Cauchy-Schwarz inequality:

$$|R| \le \frac{1}{2}\left[\frac{1}{n}\sum_{i=1}^{n} G''(m_i^0)^2\right]^{1/2} \left[\sup_{x_{\underline{j}}}\left|\hat{m}(x_j, x_{\underline{j}}) - m(x_j, x_{\underline{j}})\right|\right]^2.$$

The second term on the right-hand side is of order $O_p(c_{1n}^2)$ with c_{1n} the same as the one defined in Lemma 3.4.1 since $\sup_{x_j}\left|\hat{m}(x_j, x_{\underline{j}}) - m(x_j, x_{\underline{j}})\right| = O_p(h_1^2 + \sqrt{\ln n/(nh_1h_2^{d-1})})$ by the standard theory for nonparametric regression smoothers (Härdle, 1990; Masry, 1996). Hence the reminder R is of order $O_p(c_{1n}^2)$.

By the conditions on the bandwidths, $\sqrt{nh_1}R$ goes to zero as n goes to infinity. Therefore asymptotically,

$$\sqrt{nh_1}(\hat{F}_j(x_j) - F_j(x_j)) = \sqrt{nh_1}\frac{1}{n}\sum_{i=1}^{n} G'\left\{m(x_j, X_{i\underline{j}})\right\}\left[\hat{m}(x_j, X_{i\underline{j}}) - m(x_j, X_{i\underline{j}})\right].$$

This means we only need to consider the expression on the right hand side of the equation.

Define

$$F_i = \begin{pmatrix} m(x_j, X_{i\underline{j}}) \\ \\ (\partial/\partial x_j) m(x_j, X_{i\underline{j}}) \end{pmatrix}$$

and by the argument above,

$$\begin{aligned}
&\hat{F}_j(x_j) - F_j(x_j) \\
&= \frac{1}{n}\sum_{i=1}^{n} G'\left\{m(x_j, X_{i\underline{j}})\right\} \cdot \left[e_1^T (Z_j^T W_{i,j} Z_j)^{-1} Z_j^T W_{i,j} Y - m(x_j, X_{i\underline{j}})\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} G'\left\{m(x_j, X_{i\underline{j}})\right\} \cdot \left[e_1^T (Z_j^T W_{i,j} Z_j)^{-1} Z_j^T W_{i,j} (Y - Z_j F_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} G'\left\{m(x_j, X_{i\underline{j}})\right\} \cdot \left[e_1^T H^{-1} (H^{-1} Z_j^T W_{i,j} Z_j H^{-1})^{-1} H^{-1} Z_j^T W_{i,j} (Y - Z_j F_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} G'\left\{m(x_j, X_{i\underline{j}})\right\} \frac{1}{\varphi(x_j, X_{i\underline{j}})} e_1^T S^{-1} \left\{I + O_p(c_{1n})\right\} \cdot H^{-1} Z_j^T W_{i,j} (Y - Z_j F_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} G'\left\{m(x_j, X_{i\underline{j}})\right\} \frac{1}{\varphi(x_j, X_{i\underline{j}})} \frac{1}{n}\sum_{l=1}^{n} K_{h_1}(X_{lj} - x_j) L_{h_2}(X_{l\underline{j}} - X_{i\underline{j}}) \\
&\quad \times \left\{1 + O_p(c_{1n})\right\} \left[Y_l - m(x_j, X_{i\underline{j}}) - (X_{lj} - x_j) \cdot \frac{\partial m}{\partial x_j}(x_j, X_{i\underline{j}})\right].
\end{aligned}$$

Here we use Lemma 3.4.1 (*i*) in the second last step. Applying Taylor expansion of

$m(X_l)$ around $(x_j, X_{l\underline{j}})$ in $Y_l = m(X_l) + \sigma(X_l)\varepsilon_l$, we obtain

$$\begin{aligned}
&\hat{F}_j(x_j) - F_j(x_j) \\
&= \frac{1}{n}\sum_{i=1}^{n} G'\left\{m(x_j, X_{i\underline{j}})\right\} \frac{1}{\varphi(x_j, X_{i\underline{j}})} \frac{1}{n}\sum_{l=1}^{n} K_{h_1}(X_{lj} - x_j) L_{h_2}(X_{l\underline{j}} - X_{i\underline{j}}) \\
&\quad \times \{1 + O_p(c_{1n})\} \Big[m(x_j, X_{l\underline{j}}) - m(x_j, X_{i\underline{j}}) \\
&\qquad\qquad + (X_{lj} - x_j) \left\{ \frac{\partial m}{\partial x_j}(x_j, X_{l\underline{j}}) - \frac{\partial m}{\partial x_j}(x_j, X_{i\underline{j}}) \right\} \\
&\qquad\qquad + \frac{(X_{lj} - x_j)^2}{2} \frac{\partial^2 m}{\partial x_j^2}(x_j, X_{l\underline{j}}) \\
&\qquad\qquad + O_p((X_{lj} - x_j)^3) + \sigma(X_l)\varepsilon_l \Big]
\end{aligned} \tag{3.24}$$

Note that

$$\begin{aligned}
\hat{a}_i = &\frac{1}{n}\sum_{l=1}^{n} K_{h_1}(X_{lj} - x_j) L_{h_2}(X_{l\underline{j}} - X_{i\underline{j}}) \times \Big[m(x_j, X_{l\underline{j}}) - m(x_j, X_{i\underline{j}}) \\
&+ (X_{lj} - x_j) \left\{ \frac{\partial m}{\partial x_j}(x_j, X_{l\underline{j}}) - \frac{\partial m}{\partial x_j}(x_j, X_{i\underline{j}}) \right\} + \frac{(X_{lj} - x_j)^2}{2} \frac{\partial m^2}{\partial x_j^2}(x_j, X_{l\underline{j}}) \\
&+ O_p\left((X_{lj} - x_j)^3\right) + \sigma(X_l)\varepsilon_l \Big]
\end{aligned}$$

is of $O_p(c_{1n})$ uniformly (Fan, Härdle and Mammen, 1998). Therefore, (3.24) can be written as

$$\begin{aligned}
\hat{F}_j(x_j) - F_j(x_j) &= \frac{1}{n}\sum_{i=1}^{n} G'\left\{m(x_j, X_{i\underline{j}})\right\} \frac{\hat{a}_i}{\varphi(x_j, X_{i\underline{j}})} + O_p(c_{1n}^2) \\
&= \frac{1}{n}\sum_{i=1}^{n} G'\left\{m(x_j, X_{i\underline{j}})\right\} \frac{E_i(\hat{a}_i)}{\varphi(x_j, X_{i\underline{j}})} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} G'\left\{m(x_j, X_{i\underline{j}})\right\} \frac{\hat{a}_i - E_i(\hat{a}_i)}{\varphi(x_j, X_{i\underline{j}})} + O_p(c_{1n}^2)
\end{aligned}$$

with $E_i[W] = E[W|X_i]$ and $E_*[W] = E[W|X_1, \cdots, X_n]$.

It suffices to work with the first two terms on the right hand side, ignoring the remainder term.

Let

$$T_{1n} = \frac{1}{n}\sum_{i=1}^{n} G'\left\{m(x_j, X_{i\underline{j}})\right\}\frac{E_i(\hat{a}_i)}{\varphi(x_j, X_{i\underline{j}})},$$

$$\text{and} \quad T_{2n} = \frac{1}{n}\sum_{i=1}^{n} G'\left\{m(x_j, X_{i\underline{j}})\right\}\frac{\hat{a}_i - E_i(\hat{a}_i)}{\varphi(x_j, X_{i\underline{j}})},$$

where T_{1n} is a systematic "bias" term and T_{2n} is a stochastic "variance" term.

We next prove the theorem by showing

I. $T_{1n} = h_1^2 b_j(x_j) + o_p((nh_1)^{-1/2})$,

II. $T_{2n} = \sum_{i=1}^{n} w_{ij}\varepsilon_i + o_p((nh_1)^{-1/2})$,

with $\quad w_{ij} = n^{-1}K_{h_1}(X_{ij} - x_j)\sigma(X_i)G'\{m(x_j, X_{i\underline{j}})\}\varphi_{\underline{j}}(X_{i\underline{j}})/\varphi(x_j, X_{i\underline{j}})$.

First we analyze T_{1n}. Since $E_*(\varepsilon_l) = 0$,

$$\begin{aligned}
\frac{E_i(\hat{a}_i)}{\varphi(x_j, X_{i\underline{j}})} = \frac{1}{\varphi(x_j, X_{i\underline{j}})} E_i \Bigg[& \frac{1}{n}\sum_{l=1}^{n} K_{h_1}(X_{ij} - x_j)L_{h_2}(X_{l\underline{j}} - X_{i\underline{j}}) \times [m(x_j, X_{l\underline{j}}) \\
& - m(x_j, X_{i\underline{j}}) + (X_{lj} - x_j)\left\{\frac{\partial m}{\partial x_j}(x_j, X_{l\underline{j}}) - \frac{\partial m}{\partial x_j}(x_j, X_{i\underline{j}})\right\} \\
& + \frac{(X_{lj} - x_j)^2}{2}\frac{\partial^2 m}{\partial x_j^2}(x_j, X_{l\underline{j}}) + O_p((X_{lj} - x_j)^3)]\Bigg] \\
= \frac{1}{\varphi(x_j, X_{i\underline{j}})} \int & K_{h_1}(z - x_j)L_{h_2}(w - X_{i\underline{j}})\varphi(z, w) \times [m(x_j, w) \\
& - m(x_j, X_{i\underline{j}}) + (z - x_j)\left\{\frac{\partial m}{\partial x_j}(x_j, w) - \frac{\partial m}{\partial x_j}(x_j, X_{i\underline{j}})\right\} \\
& + \frac{(z - x_j)^2}{2}\frac{\partial^2 m}{\partial x_j^2}(x_j, w) + O_p((z - x_j)^3)]dwdz.
\end{aligned}$$

Changing variables to $u = (z - x_j)/h_1$ and $v = (w - X_{i\underline{j}})/h_2$, where v and w are $(d-1)$-dimensional vectors, we have

$$\begin{aligned}\frac{E_i(\hat{a}_i)}{\varphi(x_j, X_{i\underline{j}})} = \frac{1}{\varphi(x_j, X_{i\underline{j}})} \int K(u)L(v)\varphi(x_j + h_1u, X_{i\underline{j}} + h_2v) \times \Big[& m(x_j, X_{i\underline{j}} + h_2v) \\ & - m(x_j, X_{i\underline{j}}) + h_1u\left\{\frac{\partial m}{\partial x_j}(x_j, X_{i\underline{j}} + h_2v) - \frac{\partial m}{\partial x_j}(x_j, X_{i\underline{j}})\right\} \\ & + \frac{(h_1u)^2}{2}\frac{\partial^2 m}{\partial x_j^2}(x_j, X_{i\underline{j}} + h_2v) + O_p((h_1u)^3)\Big] dvdu \\ = \frac{1}{2}h_1^2\mu_2(K)\frac{\partial^2 m}{\partial x_j^2}(x_j, X_{i\underline{j}}) + o_p(h_1^2) + o_p(h_2^q), &\end{aligned}$$

by assumptions $(A1) - (A3)$ and $(A5)$.

Since the random variables $G'\{m(x_j, X_{i\underline{j}})\}\varphi(x_j, X_{i\underline{j}})^{-1}E_i(\hat{a}_i), i = 1, \ldots, n$, are independent and bounded, the single sum T_{1n} converges to its population mean by Chebyshev's Law of Large Numbers,

$$\begin{aligned}T_{1n} &= \int G'\{m(x_j, x_{\underline{j}})\}\frac{1}{2}h_1^2\mu_2(K)\frac{\partial^2 m}{\partial x_j^2}(x_j, x_{\underline{j}})\varphi_{\underline{j}}(x_{\underline{j}})dx_{\underline{j}} + o_p(h_1^2) + O_p(n^{-1/2}) \\ &= h_1^2\frac{\mu_2(K)}{2}\int\{(G' \circ m)\frac{\partial^2 m}{\partial x_j^2}\}(x_j, x_{\underline{j}})\varphi_{\underline{j}}(x_{\underline{j}})dx_{\underline{j}} + o_p(h_1^2) + O_p(n^{-1/2}).\end{aligned}$$

We now turn to the stochastic term

$$T_{2n} = \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, X_{i\underline{j}})\}\frac{\hat{a}_i - E_i(\hat{a}_i)}{\varphi(x_j, X_{i\underline{j}})}$$

and separate further as

$$\hat{a}_i - E_i(\hat{a}_i) = \hat{a}_i - E_*(\hat{a}_i) + E_*(\hat{a}_i) - E_i(\hat{a}_i).$$

We first show that

$$\frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, X_{i\underline{j}})\}\frac{\hat{a}_i - E_*(\hat{a}_i)}{\varphi(x_j, X_{i\underline{j}})} = \sum_{i=1}^{n} w_{ij}\varepsilon_i + o_p((nh_1)^{-1/2}).$$

Note that

$$\hat{a}_i - E_*(\hat{a}_i) = \frac{1}{n}\sum_{l=1}^{n} K_{h_1}(X_{lj} - x_j)L_{h_2}(X_{l\underline{j}} - X_{i\underline{j}})\sigma(X_l)\varepsilon_l.$$

Hence,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, X_{i\underline{j}})\}\frac{\hat{a}_i - E_*(\hat{a}_i)}{\varphi(x_j, X_{i\underline{j}})} \\
&= \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, X_{i\underline{j}})\}\frac{1}{\varphi(x_j, X_{i\underline{j}})}\frac{1}{n}\sum_{l=1}^{n} K_{h_1}(X_{lj} - x_j)L_{h_2}(X_{l\underline{j}} - X_{i\underline{j}})\sigma(X_l)\varepsilon_l \\
&= \frac{1}{n}\sum_{l=1}^{n} K_{h_1}(X_{lj} - x_j)\sigma(X_l)\varepsilon_l\left[\frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, X_{i\underline{j}})\}\frac{1}{\varphi(x_j, X_{i\underline{j}})}L_{h_2}(X_{l\underline{j}} - X_{i\underline{j}})\right].
\end{aligned}$$

Let $\eta_l = \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, X_{i\underline{j}})\}\frac{1}{\varphi(x_j, X_{i\underline{j}})}L_{h_2}(X_{l\underline{j}} - X_{i\underline{j}})$, and separate η_l into $E_l(\eta_l) + [\eta_l - E_l(\eta_l)]$. Then

$$\begin{aligned}
E_l(\eta_l) &= \int G'\{m(x_j, z)\}\frac{1}{\varphi(x_j, z)}L_{h_2}(X_{l\underline{j}} - z)\varphi_{\underline{j}}(z)dz \\
&= \int G'\{m(x_j, X_{l\underline{j}} + h_2 v)\}\frac{1}{\varphi(x_j, X_{l\underline{j}} + h_2 v)}L(v)\varphi_{\underline{j}}(X_{l\underline{j}} + h_2 v)dv \\
&= G'\{m(x_j, X_{l\underline{j}})\}\frac{\varphi_{\underline{j}}(X_{l\underline{j}})}{\varphi(x_j, X_{l\underline{j}})} + O_p(h_2^q).
\end{aligned}$$

Further, we have

$$\begin{aligned}
&E_l[\eta_l - E_l(\eta_l)]^2 \\
&= \frac{1}{n}\int\left[\frac{G'\{m(x_j, z)\}}{\varphi(x_j, z)}L_{h_2}(X_{l\underline{j}} - z) - G'\{m(x_j, X_{l\underline{j}})\}\frac{\varphi_{\underline{j}}(X_{l\underline{j}})}{\varphi(x_j, X_{l\underline{j}})} + O_p(h_2^q)\right]^2\varphi_{\underline{j}}(z)dz \\
&= \frac{1}{n}\int\left[\frac{G'\{m(x_j, z)\}}{\varphi(x_j, z)}L_{h_2}(X_{l\underline{j}} - z) - G'\{m(x_j, X_{l\underline{j}})\}\frac{\varphi_{\underline{j}}(X_{l\underline{j}})}{\varphi(x_j, X_{l\underline{j}})}\right]^2\varphi_{\underline{j}}(z)dz + O_p(\frac{h_2^{2q}}{n}) \\
&= \frac{1}{n}\int\left[\frac{G'\{m(x_j, z)\}}{\varphi(x_j, z)}L_{h_2}(X_{l\underline{j}} - z)\right]^2\varphi_{\underline{j}}(z)dz + O_p(n^{-1}).
\end{aligned}$$

By a change of variable, we get

$$
\begin{aligned}
E_l[\eta_l - E_l(\eta_l)]^2 &= \frac{1}{nh_2^{d-1}} \int \left[\frac{G'\{m(x_j, X_{l\underline{j}} + h_2 v)\}}{\varphi(x_j, X_{l\underline{j}} + h_2 v)} L(v) \right]^2 \varphi_{\underline{j}}(X_{l\underline{j}} + h_2 v) dv \\
&= \frac{1}{nh_2^{d-1}} \left[G'\{m(x_j, X_{l\underline{j}})\} \right]^2 \frac{\varphi_{\underline{j}}(X_{l\underline{j}})}{\varphi^2(x_j, X_{l\underline{j}})} \|L\|_2^2 + O_p(n^{-1}) \\
&= o_p(h_1),
\end{aligned}
$$

by the assumptions $(A1), (A2)$ and $(A5)$.

Therefore,

$$
\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} G'\{m(x_j, X_{i\underline{j}})\} \frac{\hat{a}_i - E_*(\hat{a}_i)}{\varphi(x_j, X_{i\underline{j}})} \\
&= \frac{1}{n} \sum_{l=1}^{n} K_{h_1}(X_{lj} - x_j) \sigma(X_l) \varepsilon_l \eta_l \\
&= \sum_{l=1}^{n} \frac{1}{n} K_{h_1}(X_{lj} - x_j) \sigma(X_l) \varepsilon_l \left[G'\{m(x_j, X_{l\underline{j}})\} \frac{\varphi_{\underline{j}}(X_{l\underline{j}})}{\varphi(x_j, X_{l\underline{j}})} + O_p(h_2^q) + o_p(h_1^{1/2}) \right] \\
&= \sum_{l=1}^{n} w_{lj} \varepsilon_l \left[1 + o_p(1) \right],
\end{aligned}
$$

where w_{lj} is defined previously in this section.

By Linton and Härdle (1996), the term $\sum_{l=1}^{n} w_{lj} \varepsilon_l$ provides the asymptotic variance of the estimator; it is $O_p((nh_1)^{-1/2})$, since only smoothing with respect to X_j is present. Furthermore, $\sqrt{nh_1} \sum_{l=1}^{n} w_{lj} \varepsilon_l$ obeys a central limit theorem with limiting variance as stated in Theorem 3.2.1.

Note that $w_{lj}\varepsilon_l, l = 1, \ldots, n$ are i.i.d. with mean zero. The variance is

$$
\begin{aligned}
Var(w_{lj}\varepsilon_l) &= E(w_{lj}^2\varepsilon_l^2) \\
&= E(w_{lj}^2) \\
&= \frac{1}{n^2}\int K_{h_1}^2(x_j - z)\sigma^2(z, w)G'^2\{m(x_j, w)\}\frac{\varphi_{\underline{j}}^2(w)}{\varphi^2(x_j, w)}\varphi(z, w)dzdw \\
&= \frac{1}{n^2}\int \frac{1}{h_1}K^2(u)\sigma^2(x_j + h_1u, w) \\
&\quad \times G'^2\{m(x_j, w)\}\frac{\varphi_{\underline{j}}^2(w)}{\varphi^2(x_j, w)}\varphi(x_j + h_1u, w)dudw \\
&= n^{-2}h_1^{-1}\|K\|_2^2\int \frac{G'^2\{m(x_j, w)\}\sigma^2(x_j, w)}{\varphi(x_j, w)}\varphi_{\underline{j}}^2(w)dw + o_p(n^{-2}h_1^{-1}) \\
&= n^{-2}h_1^{-1}\|K\|_2^2\int \left\{\frac{(G' \circ m)^2\sigma^2}{\varphi}\right\}(x_j, w)\varphi_{\underline{j}}^2(w)dw + o_p(n^{-2}h_1^{-1}).
\end{aligned}
\tag{3.25}
$$

Since w_{lj}^2, $l = 1, \ldots, n$, are bounded under the assumptions on $K(\cdot)$, $\sigma(\cdot)$, $G(\cdot)$ and the density functions, $w_{lj}^2/E(w_{lj}\varepsilon_l^2)$ is bounded and we further have

$$
\frac{w_{lj}^2\varepsilon_l^2}{E(w_{lj}^2\varepsilon_l^2)}I\left\{\frac{w_{lj}^2\varepsilon_l^2}{E(w_{lj}^2\varepsilon_l^2)} \geq \delta n\right\} \leq C\varepsilon_l^2,
$$

for some constant C and any $\delta > 0$.

Below we will show that the Lindeberg condition, required for the Central

Limit Theorem, follows from the Lebesgue Dominated Convergence Theorem:

$$
\begin{aligned}
&\lim_{n\to\infty} \frac{1}{S_n^2} \sum_{l=1}^{n} E\left[w_{lj}^2 \varepsilon_l^2 I\left\{w_{lj}^2 \varepsilon_l^2 \geq \delta S_n^2\right\}\right] \\
&= \lim_{n\to\infty} \frac{1}{nE(w_{lj}^2 \varepsilon_l^2)} \sum_{l=1}^{n} E\left[w_{lj}^2 \varepsilon_l^2 I\left\{\frac{w_{lj}^2 \varepsilon_l^2}{E(w_{lj}^2 \varepsilon_l^2)} \geq \delta n\right\}\right] \\
&= \lim_{n\to\infty} \frac{1}{n} \sum_{l=1}^{n} E\left[\frac{w_{lj}^2 \varepsilon_l^2}{E(w_{lj}^2 \varepsilon_l^2)} I\left\{\frac{w_{lj}^2 \varepsilon_l^2}{E(w_{lj}^2 \varepsilon_l^2)} \geq \delta n\right\}\right] \\
&= \lim_{n\to\infty} E\left[\frac{w_{lj}^2 \varepsilon_l^2}{E(w_{lj}^2 \varepsilon_l^2)} I\left\{\frac{w_{lj}^2 \varepsilon_l^2}{E(w_{lj}^2 \varepsilon_l^2)} \geq \delta n\right\}\right] \\
&= E\left[\frac{w_{lj}^2 \varepsilon_l^2}{E(w_{lj}^2 \varepsilon_l^2)} \lim_{n\to\infty} I\left\{\frac{w_{lj}^2 \varepsilon_l^2}{E(w_{lj}^2 \varepsilon_l^2)} \geq \delta n\right\}\right] \\
&= 0.
\end{aligned}
$$

As a result of the Central Limit Theorem,

$$\frac{\sum_{l=1}^{n} w_{lj}\varepsilon_l}{\sqrt{nE(w_{lj}^2 \varepsilon_l^2)}} \xrightarrow{\mathcal{D}} N(0,1).$$

Therefore, by (3.25), $\sqrt{nh_1}\sum_{l=1}^{n} w_{lj}\varepsilon_l$ obeys a Central Limit Theorem with asymptotic variance $\|K\|_2^2 \int \left\{(G' \circ m)^2 \sigma^2/\varphi\right\}(x_j, w)\varphi_{\underline{j}}^2(w)dw$.

Next we show that

$$\frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, X_{i\underline{j}})\}\frac{E_*(\hat{a}_i) - E_i(\hat{a}_i)}{\varphi(x_j, X_{i\underline{j}})} = o_p((nh_1)^{-1/2}).$$

Let

$$
\begin{aligned}
U_n &= \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, X_{i\underline{j}})\}\frac{E_*(\hat{a}_i) - E_i(\hat{a}_i)}{\varphi(x_j, X_{i\underline{j}})} \\
&= \sum_{i=1}^{n}\sum_{k=1}^{n} \tilde{\zeta}_{ik} \\
&= \sum_{i=1}^{n} \tilde{\zeta}_{ii} + \sum\sum_{i\neq k} \tilde{\zeta}_{ik},
\end{aligned}
$$

where $\tilde{\zeta}_{ik} = \zeta_{ik} - E_i(\zeta_{ik})$ and

$$\begin{aligned}\zeta_{ik} = \frac{1}{n^2}\frac{G'\{m(x_j, X_{i\underline{j}})\}}{\varphi(x_j, X_{i\underline{j}})} & K_{h_1}(X_{kj} - x_j)L_{h_2}(X_{k\underline{j}} - X_{i\underline{j}}) \\ & \times \left[m(x_j, X_{k\underline{j}}) - m(x_j, X_{i\underline{j}}) + (X_{kj} - x_j)\left\{\frac{\partial m}{\partial x_j}(x_j, X_{k\underline{j}}) - \frac{\partial m}{\partial x_j}(x_j, X_{i\underline{j}})\right\}\right. \\ & \quad \left. + \frac{(X_{kj} - x_j)^2}{2}\frac{\partial^2 m}{\partial x_j^2}(x_j, X_{k\underline{j}}) + O((X_{kj} - x_j)^3)\right]. \end{aligned} \tag{3.26}$$

When $i = k$, $E_i(\zeta_{ii}) = \zeta_{ii}$. Hence $\tilde{\zeta}_{ii} = 0$ and $U_n = \sum\sum_{i \neq k}\tilde{\zeta}_{ik}$. The double sum U_n has mean zero. In order to calculate the variance $E(\sum\sum_{i \neq k}\tilde{\zeta}_{ik})^2$, we need the calculations for the following three terms,

$$\text{(i)}\sum\sum_{i \neq k} E(\tilde{\zeta}_{ik}^2), \quad \text{(ii)}\sum\sum_{i \neq k} E(\tilde{\zeta}_{ik}\tilde{\zeta}_{ki}), \quad \text{(iii)}\sum_{i \neq k,}\sum_{i \neq l,}\sum_{k \neq l} E(\tilde{\zeta}_{ik}\tilde{\zeta}_{lk}),$$

since all other terms have mean zero by a conditioning argument.

We have

$$\begin{aligned} E_i(\zeta_{ik}) = \frac{1}{n^2}E_i & \left\{\frac{G'\{m(x_j, X_{i\underline{j}})\}}{\varphi(x_j, X_{i\underline{j}})}K_{h_1}(X_{kj} - x_j)L_{h_2}(X_{k\underline{j}} - X_{i\underline{j}})\right. \\ & \times \left[m(x_j, X_{k\underline{j}}) - m(x_j, X_{i\underline{j}})\right. \\ & \quad + (X_{kj} - x_j)\left\{\frac{\partial m}{\partial x_j}(x_j, X_{k\underline{j}}) - \frac{\partial m}{\partial x_j}(x_j, X_{i\underline{j}})\right\} \\ & \quad \left.\left. + \frac{(X_{kj} - x_j)^2}{2}\frac{\partial^2 m}{\partial x_j^2}(x_j, X_{k\underline{j}}) + O((X_{kj} - x_j)^3)\right]\right\} \\ = \frac{1}{n^2} & O(h_1^2 + h_2^q). \end{aligned}$$

Next we work with the three terms one by one as follows,

(i) Each individual term

$$E(\tilde{\zeta}_{ik}^2) = E(\zeta_{ik}^2) - E[E_i^2(\zeta_{ik})]$$

$$= E[E_i(\zeta_{ik}^2)] - E[E_i^2(\zeta_{ik})].$$

Note that $E_i(\zeta_{ik}^2) = (n^4 h_2^{d-1} h_1)^{-1} O(h_1^4 + h_2^2)$. Therefore,

$$\begin{aligned} E(\tilde{\zeta}_{ik}^2) &= \frac{1}{n^4 h_2^{d-1} h_1} O(h_1^4 + h_2^2) + \frac{1}{n^4} O(h_1^4 + h_2^{2q}) \\ &= \frac{1}{n^4 h_2^{d-1} h_1} O(h_1^4 + h_2^2) \end{aligned}$$

and $\sum\sum_{i \neq k} E(\tilde{\zeta}_{ik}^2) = \frac{1}{n^2 h_2^{d-1} h_1} O(h_1^4 + h_2^2)$.

(ii) By the Cauchy-Schwartz inequality,

$$[E(\tilde{\zeta}_{ik}\tilde{\zeta}_{ki})]^2 \leq E(\tilde{\zeta}_{ik}^2) E(\tilde{\zeta}_{ki}^2) = [E(\tilde{\zeta}_{ik}^2)]^2.$$

From (i), $E(\tilde{\zeta}_{ik}\tilde{\zeta}_{ki}) = (n^4 h_2^{d-1} h_1)^{-1} O(h_1^4 + h_2^2)$. Thus

$$\sum\sum_{i \neq k} E(\tilde{\zeta}_{ik}\tilde{\zeta}_{ki}) = \frac{1}{n^2 h_2^{d-1} h_1} O(h_1^4 + h_2^2).$$

(iii) Consider the term with three different indices:

$$E(\tilde{\zeta}_{ik}\tilde{\zeta}_{lk}) = E(\zeta_{ik}\zeta_{lk}) - E(\zeta_{ik})E(\zeta_{lk})$$

$$= E[E_k^2(\zeta_{ik})] - [E(\zeta_{ik})]^2.$$

We have

$$\begin{aligned}
E_k(\zeta_{ik}) = & \frac{1}{n^2} K_{h_1}(X_{kj} - x_j) \int \frac{G'\{m(x_j, w)\}}{\varphi(x_j, w)} L_{h_1}(X_{k\underline{j}} - w) \varphi_{\underline{j}}(w) \Big[m(X_{kj}, X_{k\underline{j}}) \\
& \qquad - m(x_j, w) - (X_{kj} - x_j) \frac{\partial m}{\partial x_j}(x_j, w) \Big] dw \\
= & \frac{1}{n^2} K_{h_1}(X_{kj} - x_j) \int \frac{G'\{m(x_j, X_{k\underline{j}} + h_2 v)\}}{\varphi(x_j, X_{k\underline{j}} + h_2 v)} L(v) \varphi_{\underline{j}}(X_{k\underline{j}} + h_2 v) \\
& \qquad \times \Big[m(X_{kj}, X_{k\underline{j}}) - m(x_j, X_{k\underline{j}} + h_2 v) \\
& \qquad - (X_{kj} - x_j) \frac{\partial m}{\partial x_j}(x_j, X_{k\underline{j}} + h_2 v) \Big] dv \\
= & \frac{1}{n^2} K_{h_1}(X_{kj} - x_j) \Bigg\{ \frac{G'\{m(x_j, X_{k\underline{j}})\}}{\varphi(x_j, X_{k\underline{j}})} \cdot \Big[m(X_{kj}, X_{k\underline{j}}) - m(x_j, X_{k\underline{j}}) \\
& \qquad - (X_{kj} - x_j) \frac{\partial m}{\partial x_j}(x_j, X_{k\underline{j}}) \Big] \varphi_{\underline{j}}(X_{k\underline{j}}) + O(h_2^q) \Bigg\},
\end{aligned}$$

and

$$\begin{aligned}
E[E_k^2(\zeta_{ik})] \leq & 2E \Bigg\{ \frac{1}{n^4} K_{h_1}^2(X_{kj} - x_j) \frac{G'^2\{m(x_j, X_{k\underline{j}})\}}{\varphi^2(x_j, X_{k\underline{j}})} \cdot \Big[m(X_{kj}, X_{k\underline{j}}) \\
& \qquad - m(x_j, X_{k\underline{j}}) - (X_{kj} - x_j) \frac{\partial m}{\partial x_j}(x_j, X_{k\underline{j}}) \Big]^2 \varphi_{\underline{j}}^2(X_{k\underline{j}}) \Bigg\} \\
& + O(h_2^{2q}) E \left[\frac{1}{n^4} K_{h_1}^2(X_{kj} - x_j) \right] \\
= & 2 \int \frac{1}{n^4} \frac{1}{h_1} K^2(u) \frac{G'^2\{m(x_j, w)\}}{\varphi^2(x_j, w)} \Big[m(x_j + h_1 u, w) - m(x_j, w) \\
& \qquad - h_1 u \frac{\partial m}{\partial x_j}(x_j, w) \Big]^2 \cdot \varphi_{\underline{j}}^2(w) \varphi(x_j + h_1 u, w) du dw + \frac{1}{n^4 h_1} O(h_2^{2q}) \\
= & \frac{1}{n^4 h_1} O(h_1^4 + h_2^{2q})
\end{aligned}$$

by a change of variable and Taylor expansion.

We also note that

$$E(\zeta_{ik}) = \frac{1}{n^2} O(h_1^2 + h_2^q).$$

Therefore

$$\begin{aligned} E(\tilde{\zeta}_{ik}\tilde{\zeta}_{lk}) &= \frac{1}{n^4 h_1} O(h_1^4 + h_2^{2q}) + \frac{1}{n^4} O(h_1^4 + h_2^{2q}) \\ &= \frac{1}{n^4 h_1} O(h_1^4 + h_2^{2q}) \end{aligned}$$

and

$$\sum_{i\neq k,} \sum_{i\neq l,} \sum_{k\neq l} E(\tilde{\zeta}_{ik}\tilde{\zeta}_{lk}) = \frac{1}{nh_1} O(h_1^4 + h_2^{2q}).$$

Combining the calculations of the three terms (i)-(iii),

$$E(\sum_{i\neq k}\sum \tilde{\zeta}_{ik})^2 = \frac{1}{n^2 h_2^{d-1} h_1} O(h_1^4 + h_2^2) + \frac{1}{nh_1} O(h_1^4 + h_2^{2q}) = o_p((nh_1)^{-1}).$$

So we proved that

$$U_n = \frac{1}{n} \sum_{i=1}^{n} G'\{m(x_j, X_{i\underline{j}})\} \frac{E_*(\hat{a}_i) - E_i(\hat{a}_i)}{\varphi(x_j, X_{i\underline{j}})} = o_p((nh_1)^{-1/2}).$$

This establishes II and thus completes the proof of Theorem 3.2.1. □

3.4.2 Proof of Theorem 3.2.2

Proof. Analogous to our analysis of the univariate function $\hat{F}_j$, we proceed to the bivariate case considering $\hat{F}_{jk}$. Recall that $X_{i\underline{jk}}$ is the ith observation vector with components j and k removed. We have

$$\begin{aligned}
&\hat{F}_{jk}(x_j, x_k) - F_{jk}(x_j, x_k)\\
&= \frac{1}{n}\sum_{i=1}^{n} G\{\hat{m}(x_j, x_k, X_{i\underline{jk}})\} - \frac{1}{n}\sum_{i=1}^{n} G\{m(x_j, x_k, X_{i\underline{jk}})\}\\
&\quad + \frac{1}{n}\sum_{i=1}^{n} G\{m(x_j, x_k, X_{i\underline{jk}})\} - F_{jk}(x_j, x_k)\\
&= \frac{1}{n}\sum_{i=1}^{n} \Big[G\{\hat{m}(x_j, x_k, X_{i\underline{jk}})\} - G\{m(x_j, x_k, X_{i\underline{jk}})\}\Big] + O_p(n^{-1/2})\\
&= \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, x_k, X_{i\underline{jk}})\} \cdot \Big[\hat{m}(x_j, x_k, X_{i\underline{jk}}) - m(x_j, x_k, X_{i\underline{jk}})\Big] + R + O_p(n^{-1/2}),
\end{aligned}$$

where $R = (2n)^{-1}\sum_{i=1}^{n} G''(m_i^*)\Big[\hat{m}(x_j, x_k, X_{i\underline{jk}}) - m(x_j, x_k, X_{i\underline{jk}})\Big]^2$ and m_i^* lies between $\hat{m}(x_j, x_k, X_{i\underline{jk}})$ and $m(x_j, x_k, X_{i\underline{jk}})$.

Similar to the univariate case, we can show that

$$|R| = O_p(c_{2n}^2), \quad \text{with} \quad c_{2n} = h_1^2 + \sqrt{\frac{\ln n}{nh_1^2 h_2^{d-2}}}$$

The constraints $(A2)$ on the bandwidths guarantee that $\sqrt{nh_1^2}R \to 0$ as $n \to \infty$. Therefore, $\sqrt{nh_1^2}\left(\hat{F}_{jk}(x_j, x_k) - F_{jk}(x_j, x_k)\right)$ is asymptotically equivalent to

$$\sqrt{nh_1^2}\frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, x_k, X_{i\underline{jk}})\} \cdot \Big[\hat{m}(x_j, x_k, X_{i\underline{jk}}) - m(x_j, x_k, X_{i\underline{jk}})\Big].$$

Define the vector

$$\mathcal{F}_i = \begin{pmatrix} m(x_j, x_k, X_{i\underline{jk}}) \\ (\partial/\partial x_j) m(x_j, x_k, X_{i\underline{jk}}) \\ (\partial/\partial x_k) m(x_j, x_k, X_{i\underline{jk}}) \end{pmatrix}.$$

Going through similar steps as for the one-dimensional case and applying Lemma 3.4.1 (ii), we have

$$
\begin{aligned}
&\hat{F}_{jk}(x_j, x_k) - F_{jk}(x_j, x_k) \\
&= \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, x_k, X_{i\underline{jk}})\} \cdot \big[e_1^T H^{-1}(H^{-1}Z_{jk}^T W_{i,jk} Z_{jk} H^{-1})^{-1} H^{-1} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \cdot Z_{jk}^T W_{i,jk}(Y - Z_{jk}\mathcal{F}_i)\big] \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{G'\{m(x_j, x_k, X_{i\underline{jk}})\}}{\varphi(x_j, x_k, X_{i\underline{jk}})} e_1^T \mathcal{S}^{-1}\left\{I + O_p(c_{2n})\right\} \cdot H^{-1} Z_{jk}^T W_{i,jk}(Y - Z_{jk}\mathcal{F}_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{G'\{m(x_j, x_k, X_{i\underline{jk}})\}}{\varphi(x_j, x_k, X_{i\underline{jk}})} \cdot \frac{1}{n}\sum_{l=1}^{n} K_{h_1}(X_{lj} - x_j, X_{lk} - x_k) L_{h_2}(X_{l\underline{jk}} - X_{i\underline{jk}}) \\
&\quad \times \Bigg[m(X_{lj}, X_{lk}, X_{l\underline{jk}}) - m(x_j, x_k, X_{i\underline{jk}}) - (X_{lj} - x_j)\frac{\partial m}{\partial x_j}(x_j, x_k, X_{i\underline{jk}}) \\
&\qquad - (X_{lk} - x_k)\frac{\partial m}{\partial x_k}(x_j, x_k, X_{i\underline{jk}}) + \sigma(X_l)\varepsilon_l\Bigg] \cdot \{1 + O_p(c_{2n})\}.
\end{aligned}
$$

As in the previous proof, we separate this expression into a "bias" term and a "variance" term:

$$
\begin{aligned}
&\hat{F}_{jk}(x_j, x_k) - F_{jk}(x_j, x_k) \\
&= \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, x_k, X_{i\underline{jk}})\} \frac{E_i(\hat{a}_i)}{\varphi(x_j, x_k, X_{i\underline{jk}})} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, x_k, X_{i\underline{jk}})\} \frac{\hat{a}_i - E_i(\hat{a}_i)}{\varphi(x_j, x_k, X_{i\underline{jk}})} + O_p(c_{2n}^2)
\end{aligned}
$$

where

$$
\begin{aligned}
\hat{a}_i = &\frac{1}{n}\sum_{l=1}^{n} K_{h_1}(X_{lj} - x_j) K_{h_1}(X_{lk} - x_k) L_{h_2}(X_{l\underline{jk}} - X_{i\underline{jk}}) \\
&\times \Bigg[m(X_{lj}, X_{lk}, X_{l\underline{jk}}) - m(x_j, x_k, X_{i\underline{jk}}) - (X_{lj} - x_j)\frac{\partial m}{\partial x_j}(x_j, x_k, X_{i\underline{jk}}) \\
&\qquad - (X_{lk} - x_k)\frac{\partial m}{\partial x_k}(x_j, x_k, X_{i\underline{jk}}) + \sigma(X_l)\varepsilon_l\Bigg].
\end{aligned}
$$

Again, it suffices to work with the first-order approximations. Let

$$\mathcal{T}_{1n} = \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, x_k, X_{i\underline{jk}})\}\frac{E_i(\hat{a}_i)}{\varphi(x_j, x_k, X_{i\underline{jk}})},$$

$$\text{and} \quad \mathcal{T}_{2n} = \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, x_k, X_{i\underline{jk}})\}\frac{\hat{a}_i - E_i(\hat{a}_i)}{\varphi(x_j, x_k, X_{i\underline{jk}})}.$$

We will consider the two terms separately as follows.

I. Bias term $\mathcal{T}_{1n}$.

Denote the expression in the bracket of the above formula of $\hat{a}_i$ by a_{il}. Using the Taylor expansion of $m(X_l)$ around $(x_j, x_k, X_{l\underline{jk}})$, we obtain

$$\begin{aligned}
a_{il} &= m(x_j, x_k, X_{l\underline{jk}}) - m(x_j, x_k, X_{i\underline{jk}}) \\
&\quad + (X_{lj} - x_j)\left(\frac{\partial m}{\partial x_j}(x_j, x_k, X_{l\underline{jk}}) - \frac{\partial m}{\partial x_j}(x_j, x_k, X_{i\underline{jk}})\right) \\
&\quad + (X_{lk} - x_k)\left(\frac{\partial m}{\partial x_k}(x_j, x_k, X_{l\underline{jk}}) - \frac{\partial m}{\partial x_k}(x_j, x_k, X_{i\underline{jk}})\right) \\
&\quad + \frac{(X_{lj} - x_j)^2}{2}\frac{\partial^2 m}{\partial x_j^2}(x_j, x_k, X_{l\underline{jk}}) + \frac{(X_{lk} - x_k)^2}{2}\frac{\partial^2 m}{\partial x_k^2}(x_j, x_k, X_{l\underline{jk}}) \\
&\quad + (X_{lj} - x_j)(X_{lk} - x_k)\frac{\partial^2 m}{\partial x_j \partial x_k}(x_j, x_k, X_{l\underline{jk}}) \\
&\quad + O_p((X_{lj} - x_j)^3) + O_p((X_{lk} - x_k)^3) + O_p\{(X_{lj} - x_j)(X_{lk} - x_k)\} \\
&\quad + \sigma(X_l)\varepsilon_l.
\end{aligned}$$

Replacing the expectation with an integral, we get

$$
\begin{aligned}
&\frac{E_i(\hat{a}_i)}{\varphi(x_j, x_k, X_{i\underline{jk}})} \\
&= \frac{1}{\varphi(x_j, x_k, X_{i\underline{jk}})} \int K_{h_1}(z_j - x_j) K_{h_1}(z_k - x_k) L_{h_2}(w - X_{i\underline{jk}}) \varphi(z_j, z_k, w) \\
&\qquad\qquad \times \Big[m(x_j, x_k, w) - m(x_j, x_k, X_{i\underline{jk}}) \\
&\qquad\qquad\quad + (z_j - x_j) \left(\frac{\partial m}{\partial x_j}(x_j, x_k, w) - \frac{\partial m}{\partial x_j}(x_j, x_k, X_{i\underline{jk}}) \right) \\
&\qquad\qquad\quad + (z_k - x_k) \left(\frac{\partial m}{\partial x_k}(x_j, x_k, w) - \frac{\partial m}{\partial x_k}(x_j, x_k, X_{i\underline{jk}}) \right) \\
&\qquad\qquad\quad + \frac{(z_j - x_j)^2}{2} \frac{\partial^2 m}{\partial x_j^2}(x_j, x_k, w) + \frac{(z_k - x_k)^2}{2} \frac{\partial^2 m}{\partial x_k^2}(x_j, x_k, w) \\
&\qquad\qquad\quad + (z_j - x_j)(z_k - x_k) \frac{\partial^2 m}{\partial x_j \partial x_k}(x_j, x_k, w) + O_p((z_j - x_j)^3) \\
&\qquad\qquad\quad + O_p((z_k - x_k)^3) + O_p\{(z_j - x_j)(z_k - x_k)\} \Big] \, dz_j dz_k \\
&= \frac{1}{\varphi(x_j, x_k, X_{i\underline{jk}})} \int K(u_j) K(u_k) L(v) \varphi(x_j + h_1 u_j, x_k + h_1 u_k, X_{i\underline{jk}} + h_2 v) \\
&\qquad\qquad \times \Big[m(x_j, x_k, X_{i\underline{jk}} + h_2 v) - m(x_j, x_k, X_{i\underline{jk}}) \\
&\qquad\qquad\quad + h_1 u_j \left(\frac{\partial m}{\partial x_j}(x_j, x_k, X_{i\underline{jk}} + h_2 v) - \frac{\partial m}{\partial x_j}(x_j, x_k, X_{i\underline{jk}}) \right) \\
&\qquad\qquad\quad + h_1 u_k \left(\frac{\partial m}{\partial x_k}(x_j, x_k, X_{i\underline{jk}} + h_2 v) - \frac{\partial m}{\partial x_k}(x_j, x_k, X_{i\underline{jk}}) \right) \\
&\qquad\qquad\quad + \frac{h_1^2 u_j^2}{2} \frac{\partial^2 m}{\partial x_j^2}(x_j, x_k, X_{i\underline{jk}} + h_2 v) \\
&\qquad\qquad\quad + \frac{h_1^2 u_k^2}{2} \frac{\partial^2 m}{\partial x_k^2}(x_j, x_k, X_{i\underline{jk}} + h_2 v) \\
&\qquad\qquad\quad + h_1^2 u_j u_k \frac{\partial^2 m}{\partial x_j \partial x_k}(x_j, x_k, X_{i\underline{jk}} + h_2 v) + O_p(h_1^3) \Big] \, dv du \\
&= \frac{1}{2} h_1^2 \mu_2(K) \left[\frac{\partial^2 m}{\partial x_j^2}(x_j, x_k, X_{i\underline{jk}}) + \frac{\partial^2 m}{\partial x_k^2}(x_j, x_k, X_{i\underline{jk}}) \right] + o_p(h_1^2) + O_p(h_2^q)
\end{aligned}
$$

by the substitutions $u_j = (z_j - x_j)/h_1,\ u_k = (z_k - x_k)/h_1$, and $v = (w - X_{i\underline{jk}})/h_2$ where v and w are $(d-2)$-dimensional vectors.

Again, as the random variables $G'\{m(x_j, x_k, X_{i\underline{jk}})\}\varphi(x_j, x_k, X_{i\underline{jk}})^{-1}E_i(a_i)$, $i = 1, \ldots, n$, are independent and bounded, we have

$$\begin{aligned}\mathcal{T}_{1n} &= \int G'\{m(x_j, x_k, X_{i\underline{jk}})\}\frac{1}{2}h_1^2\mu_2(K)\left[\frac{\partial^2 m}{\partial x_j^2}(x_j, x_k, x_{\underline{jk}}) + \frac{\partial^2 m}{\partial x_k^2}(x_j, x_k, x_{\underline{jk}})\right]\\ &\quad \times \varphi_{\underline{jk}}(x_{\underline{jk}})dx_{\underline{jk}} + o_p(h_1^2) + O_p(n^{-1/2})\\ &= h_1^2\frac{\mu_2(k)}{2}\int\left\{(G'\circ m)\left(\frac{\partial^2 m}{\partial x_j^2} + \frac{\partial^2 m}{\partial x_k^2}\right)\right\}(x_j, x_k, x_{\underline{jk}})\varphi_{\underline{jk}}(x_{\underline{jk}})dx_{\underline{jk}}\\ &\quad + o_p(h_1^2) + O_p(n^{-1/2}).\end{aligned}$$

Thus, combining with the bias formula obtained for univariate functions $\hat{F}_j(x_j)$ and $\hat{F}_k(x_k)$, the bias of $\hat{F}_{jk}(x_j, x_k) - \hat{F}_j(x_j) - \hat{F}_k(x_k)$ is as claimed in Theorem 3.2.2:

$$\begin{aligned}h_1^2 B_{jk}(x_j, x_k) = h_1^2\frac{\mu_2(k)}{2}&\left[\int\left\{(G'\circ m)\left(\frac{\partial^2 m}{\partial x_j^2} + \frac{\partial^2 m}{\partial x_k^2}\right)\right\}(x_j, x_k, x_{\underline{jk}})\varphi_{\underline{jk}}(x_{\underline{jk}})dx_{\underline{jk}}\right.\\ &\quad - \int\left\{(G'\circ m)\frac{\partial^2 m}{\partial x_j^2}\right\}(x_j, x_{\underline{j}})\varphi_{\underline{j}}(x_{\underline{j}})dx_{\underline{j}}\\ &\quad \left. - \int\left\{(G'\circ m)\frac{\partial^2 m}{\partial x_k^2}\right\}(x_k, x_{\underline{k}})\varphi_{\underline{k}}(x_{\underline{k}})dx_{\underline{k}}\right].\end{aligned}$$

II. Let us turn to the variance part $\mathcal{T}_{2n}$. We will show that

$$\mathcal{T}_{2n} = \sum_{i=1}^{n} w_{ijk}\varepsilon_i + o_p\{(nh_1^2)^{-1/2}\},$$

where

$$w_{ijk} = \frac{1}{n}K_{h_1}(X_{ij} - x_j)K_{h_1}(X_{ik} - x_k)\sigma(X_i)G'\{m(x_j, x_k, X_{i\underline{jk}})\}\frac{\varphi_{\underline{jk}}(X_{i\underline{jk}})}{\varphi(x_j, x_k, X_{i\underline{jk}})},$$

and the term $\sum_{i=1}^{n} w_{ijk}\varepsilon_i$ is of order $O_p\{(nh_1^2)^{-1/2}\}$.

The proof of the part is very similar to that of Theorem 3.2.1, except the dimension of the functions. So we only sketch the proof here.

As before, we have

$$\hat{a}_i - E_i(\hat{a}_i) = \hat{a}_i - E_*(\hat{a}_i) + E_*(\hat{a}_i) - E_i(\hat{a}_i).$$

First consider

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, x_k, X_{i\underline{jk}})\}\frac{\hat{a}_i - E_*(\hat{a}_i)}{\varphi(x_j, x_k, X_{i\underline{jk}})} \\
&= \frac{1}{n}\sum_{l=1}^{n} K_{h_1}(X_{lj} - x_j)K_{h_1}(X_{lk} - x_k)\sigma(X_l)\varepsilon_l \\
&\qquad \times \left[\frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, x_k, X_{i\underline{jk}})\}\frac{1}{\varphi(x_j, x_k, X_{i\underline{jk}})} L_{h_2}(X_{l\underline{jk}} - X_{i\underline{jk}})\right]
\end{aligned}$$

Let $\eta_l = n^{-1}\sum_{i=1}^{n} G'\{m(x_j, x_k, X_{i\underline{jk}})\}\frac{1}{\varphi(x_j,x_k,X_{i\underline{jk}})} L_{h_2}(X_{l\underline{jk}} - X_{i\underline{jk}})$ and consider it separately as $E_l(\eta_l) + [\eta_l - E_l(\eta_l)]$. Then

$$E_l(\eta_l) = G'\{m(x_j, x_k, X_{l\underline{jk}})\}\frac{\varphi_{\underline{jk}}(X_{l\underline{jk}})}{\varphi(x_j, x_k, X_{l\underline{jk}})} + O_p(h_2^q),$$

and we have

$$\begin{aligned}
E_l[\{\eta_l - E_l(\eta_l)\}^2] &= \frac{1}{nh_2^{d-2}}\left[G'\{m(x_j, x_k, X_{l\underline{jk}})\}\right]^2 \frac{\varphi_{\underline{jk}}(X_{l\underline{jk}})}{\varphi^2(x_j, x_k, X_{l\underline{jk}})}\|L\|_2^2 + O_p(n^{-1}) \\
&= o_p(h_1^2).
\end{aligned}$$

Hence

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, x_k, X_{i\underline{jk}})\}\frac{\hat{a}_i - E_*(\hat{a}_i)}{\varphi(x_j, x_k, X_{i\underline{jk}})} \\
&= \sum_{l=1}^{n} \frac{1}{n} K_{h_1}(X_{lj} - x_j) K_{h_1}(X_{lk} - x_k)\sigma(X_l)\varepsilon_l \\
&\qquad \times \left[G'\{m(x_j, x_k, X_{l\underline{jk}})\}\frac{\varphi_{\underline{jk}}(X_{l\underline{jk}})}{\varphi(x_j, x_k, X_{l\underline{jk}})} + O_p(h_2^q) + o_p(h_1) \right] \\
&= \sum_{l=1}^{n} w_{ljk}\varepsilon_l\{1 + o_p(1)\}.
\end{aligned}$$

Again, $w_{ljk}\varepsilon_l$, $l = 1, \ldots, n$, are of mean zero and i.i.d. with variance

$$\begin{aligned}
\mathrm{Var}(w_{ljk}\varepsilon_l) &= E(w_{ljk}^2) \\
&= (nh_1)^{-2}(\|K\|_2^2)^2 \int \frac{G'^2\{m(x_j, x_k, w)\}\sigma^2(x_j, x_k, w)}{\varphi(x_j, x_k, w)} \varphi_{\underline{jk}}^2(w)dw \\
&\quad + o_p(n^{-2}h_1^{-2}) \\
&= (nh_1)^{-2}(\|K\|_2^2)^2 \int \left\{ \frac{(G' \circ m)^2\sigma^2}{\varphi} \right\}(x_j, x_k, w)\varphi_{\underline{jk}}^2(w)dw + o_p(n^{-2}h_1^{-2}).
\end{aligned}$$

Thus $\mathrm{Var}(\sum_{l=1}^{n} w_{ljk}\varepsilon_l) = (nh_1^2)^{-1}(\|K\|_2^2)^2 \int \cdots dw + o_p\{(nh_1^2)^{-1}\}$. Therefore the term $\sum_{l=1}^{n} w_{ljk}\varepsilon_l = O_p\{(nh_1^2)^{-1/2}\}$ dominates the corresponding stochastic term $\sum_{i=1}^{n} w_{ij}\varepsilon_i = O_p\{(nh_1)^{-1/2}\}$ from the proof of normality of the univariate functions $\hat{F}_j$ and $\hat{F}_k$. Analogously, the asymptotic normality of $\sqrt{nh_1^2}\sum_{l=1}^{n} w_{ljk}\varepsilon_l$ follows by applying the Central Limit Theorem with variance

$$\left(\|K\|_2^2\right)^2 \int \left\{ \frac{(G' \circ m)^2\sigma^2}{\varphi} \right\}(x_j, x_k, w)\varphi_{\underline{jk}}^2(w)dw.$$

Next we consider

$$U_n = \frac{1}{n}\sum_{i=1}^{n} G'\{m(x_j, x_k, X_{i\underline{jk}})\}\frac{E_*(\hat{a}_i) - E_i(\hat{a}_i)}{\varphi(x_j, x_k, X_{i\underline{jk}})}.$$

We will show that $U_n = o_p\{(nh_1^2)^{-1/2}\}$ using a similar technique. Let

$$U_n = \sum_{i=1}^{n} \tilde{\zeta}_{ii} + \sum\sum_{i \neq l} \tilde{\zeta}_{il}$$

where $\tilde{\zeta}_{il} = \zeta_{il} - E_i(\zeta_{il})$ with

$$\begin{aligned}\zeta_{il} = \frac{1}{n^2} \frac{G'\{m(x_j, x_k, X_{i\underline{jk}})\}}{\varphi(x_j, x_k, X_{i\underline{jk}})} & K_{h_1}(X_{lj} - x_j) K_{h_1}(X_{lk} - x_k) L_{h_2}(X_{l\underline{jk}} - X_{i\underline{jk}}) \\ \times & \left[m(X_l) - m(x_j, x_k, X_{i\underline{jk}}) - (X_{lj} - x_j) \frac{\partial m}{\partial x_j}(x_j, x_k, X_{i\underline{jk}}) \right. \\ & \left. - (X_{lk} - x_k) \frac{\partial m}{\partial x_k}(x_j, x_k, X_{i\underline{jk}}) \right].\end{aligned}$$

Since $\tilde{\zeta}_{ii} = 0$, $U_n = \sum\sum_{i \neq l} \tilde{\zeta}_{il}$. It is obvious that $E(U_n) = 0$. We now calculate the variance of $\sum\sum_{i \neq l} \tilde{\zeta}_{il}$, which involves calculation of the following three quantities:

$$\sum\sum_{i \neq l} E(\tilde{\zeta}_{il}^2), \quad \sum\sum_{i \neq l} E(\tilde{\zeta}_{il} \tilde{\zeta}_{li}), \quad \sum_{i \neq l,} \sum_{i \neq m,} \sum_{m \neq l} E(\tilde{\zeta}_{il} \tilde{\zeta}_{ml}).$$

First, $E_i(\zeta_{il}) = n^{-2} O(h_1^2 + h_2^q)$.

Next, it can be shown that

$$\sum\sum_{i \neq l} E(\tilde{\zeta}_{il}^2) = \frac{1}{n^2 h_2^{d-2} h_1^2} O(h_1^4 + h_2^2),$$

$$\sum\sum_{i \neq l} E(\tilde{\zeta}_{il} \tilde{\zeta}_{li}) = \frac{1}{n^2 h_2^{d-2} h_1^2} O(h_1^4 + h_2^2),$$

and

$$\sum_{i \neq l,} \sum_{i \neq m,} \sum_{m \neq l} E(\tilde{\zeta}_{il} \tilde{\zeta}_{ml}) = \frac{1}{n h_1^2} O(h_1^4 + h_2^{2q}).$$

Then the total contribution to the variance of U_n from these terms is

$$\mathrm{Var}(\sum\sum_{i\neq l}\tilde{\zeta}_{il}) = \frac{1}{n^2h_2^{d-2}h_1^2}O(h_1^4+h_2^2) + \frac{1}{nh_1^2}O(h_1^4+h_2^{2q}) = o_p\{(nh_1^2)^{-1}\}.$$

Thus U_n is of order $o_p\{(nh_1^2)^{-1/2}\}$ and is asymptotically negligible compared to the term $\sum_{i=1}^n w_{ijk}\varepsilon_i$. So it is sufficient to consider $\sum_{i=1}^n w_{ijk}\varepsilon_i$. As stated previously, this stochastic term has a slower convergence rate than that of the terms from the univariate function estimator. Consequently, $\hat{f}_{jk}^*(x_j, x_k)$ is asymptotically normal, as described in Theorem 3.2.2.

□

3.4.3 Proof of Theorem 3.2.3

Proof. : In order to prove the theorem we need to show that the asymptotic covariance between the two function estimates is of smaller order than the variances of each component function.

From the proof of Theorem 3.2.2, we know that the stochastic term $\sum_{i=1}^n w_{ijk}\varepsilon_i$ dominates the variance. It will be sufficient to look at the covariance between two such terms $\sum_{i=1}^n w_{ijk}\varepsilon_i$ and $\sum_{i=1}^n w_{ilm}\varepsilon_i$, $1 \leq j < k \leq d$, $1 \leq l < m \leq d$, $(j,k) \neq (l,m)$. Thus we need to show that

$$E\left[\left\{\sum_{i=1}^n w_{ijk}\varepsilon_i\right\}\left\{\sum_{i=1}^n w_{ilm}\varepsilon_i\right\}\right] = o(n^{-1}h_1^{-2}).$$

Since $E(\varepsilon_i \varepsilon_p) = 0$ for $i \neq p$ and $w_{ijk} w_{ilm} \varepsilon_i^2$ are i.i.d.,

$$E\left[\left\{\sum_{i=1}^{n} w_{ijk}\varepsilon_i\right\}\left\{\sum_{i=1}^{n} w_{ilm}\varepsilon_i\right\}\right] = nE\left[w_{1jk}w_{1lm}\varepsilon_1^2\right]$$

and

$$\begin{aligned}
&E\left[w_{1jk}w_{1lm}\varepsilon_1^2\right] \\
&= \frac{1}{n^2}E\Bigg[K_{h_1}(X_{1j}-x_j)K_{h_1}(X_{1k}-x_k)\sigma(X_1)G'\{m(x_j,x_k,X_{1\underline{jk}})\}\frac{\varphi_{\underline{jk}}(X_{1\underline{jk}})}{\varphi(x_j,x_k,X_{1\underline{jk}})} \\
&\qquad \times K_{h_1}(X_{1l}-x_l)K_{h_1}(X_{1m}-x_m)\sigma(X_1)G'\{m(x_l,x_m,X_{1\underline{lm}})\}\frac{\varphi_{\underline{lm}}(X_{i\underline{lm}})}{\varphi(x_l,x_m,X_{1\underline{lm}})}\Bigg] \\
&= \frac{1}{n^2}\int K_{h_1}(z_j-x_j)K_{h_1}(z_k-x_k)\sigma^2(z_j,z_k,z_l,z_m,w) \\
&\qquad \times G'\{m(x_j,x_k,z_l,z_m,w)\}\frac{\varphi_{\underline{jk}}(z_l,z_m,w)}{\varphi(x_j,x_k,z_l,z_m,w)} \\
&\qquad \times K_{h_1}(z_l-x_l)K_{h_1}(z_m-x_m)G'\{m(x_l,x_m,z_j,z_k,w)\} \\
&\qquad \times \frac{\varphi_{\underline{lm}}(z_j,z_k,w)}{\varphi(x_l,x_m,z_j,z_k,w)}\cdot\varphi(z_j,z_k,z_l,z_m,w)dz_jdz_kdz_ldz_mdw \\
&= \frac{1}{n^2}\int K(u_j)K(u_k)\sigma^2(x_j+h_1u_j,x_k+h_1u_k,x_l+h_1ul,x_m+h_1u_m,w) \\
&\qquad \times G'\{m(x_j,x_k,x_l+h_1u_l,x_m+h_1u_m,w)\} \\
&\qquad \times \frac{\varphi_{\underline{jk}}(x_l+h_1u_l,x_m+h_1u_m,w)}{\varphi(x_j,x_k,x_l+h_1u_l,x_m+h_1u_m,w)} \\
&\qquad \times K(u_l)K(u_m)G'\{m(x_l,x_m,x_j+h_1u_j,x_k+h_1u_k,w)\} \\
&\qquad \times \frac{\varphi_{\underline{lm}}(x_j+h_1u_j,x_k+h_1u_k,w)}{\varphi(x_l,x_m,x_j+h_1u_j,x_k+h_1u_k,w)} \\
&\qquad \times \varphi(x_j+h_1u_j,x_k+h_1u_k,x_l+h_1u_l,x_m+h_1u_m,w)du_jdu_kdu_ldu_mdw.
\end{aligned}$$

It is easy to see that the expression above is of order $O(n^{-2}h_1^{-1})$ if exactly one of the indices (j,k) equals one of the indices (l,m) and is of order $O(n^{-2})$ if none of the indices j, k, l, m are equal. This establishes the negligible asymptotic covariance of $\hat{f}^*_{jk}$ and $\hat{f}^*_{lm}$, thus proves the theorem. $\square$

3.4.4 Proof of Theorem 3.3.1

By Theorem 3.2.2, we can write

$$\hat{f}^*_{jk}(x_j, x_k) = f^*_{jk}(x_j, x_k) + h_1^2 B_{jk}(x_j, x_k) + \sum_{i=1}^{n}(w_{ijk} - w_{ij} - w_{ik})\varepsilon_i + o_p(h_1^2),$$

where the weights w_{ijk}, w_{ij} and w_{ik} are defined in Sections 3.4.1 and 3.4.2. Thus

$$\begin{aligned}
&\int \hat{f}^*_{jk}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k \\
&= \int \left[\sum_{i=1}^{n}(w_{ijk} - w_{ij} - w_{ik})\varepsilon_i\right]^2 \varphi_{jk}(x_j, x_k)dx_j dx_k + \int f^{*2}_{jk}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k \\
&\quad + 2h_1^2 \int f^*_{jk}(x_j, x_k)B_{jk}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k + o_p(h_1^2).
\end{aligned}$$

Let Q be the quadratic term $\int \left[\sum_{i=1}^{n}(w_{ijk} - w_{ij} - w_{ik})\varepsilon_i\right]^2 \varphi_{jk}(x_j, x_k)dx_j dx_k$ and write it as $\sum\sum_{i,l=1}^{n} \varepsilon_i\varepsilon_l A(X_i, X_l)$, where

$$A(X_i, X_l) = \frac{1}{n^2}\int (w_{ijk} - w_{ij} - w_{ik})(w_{ljk} - w_{lj} - w_{lk})\varphi_{jk}(x_j, x_k)dx_j dx_k.$$

Separating its diagonal and cross terms, one gets $Q = Q_1 + Q_2$ with

$$Q_1 = \sum_{i=1}^{n} \varepsilon_i^2 A(X_i, X_i), \quad \text{and} \quad Q_2 = \sum\sum_{i \neq l} \varepsilon_i\varepsilon_l A(X_i, X_l). \tag{3.27}$$

We then calculate the asymptotics of Q_1 and Q_2 separately and put the results together to get the limiting distribution of the test statistic itself.

Plugging in the formula for w_{ijk}, w_{ij} and w_{ik}, we have

$$\begin{aligned}
A(X_i, X_l) = \frac{1}{n^2} \int \Bigg[& K_{h_1}(X_{ij} - x_j) K_{h_1}(X_{ik} - x_k) G'\{m(x_j, x_k, X_{i\underline{jk}})\} \frac{\varphi_{\underline{jk}}(X_{i\underline{jk}})}{\varphi(x_j, x_k, X_{i\underline{jk}})} \\
& - K_{h_1}(X_{ij} - x_j) G'\{m(x_j, X_{i\underline{j}})\} \frac{\varphi_{\underline{j}}(X_{i\underline{j}})}{\varphi(x_j, X_{i\underline{j}})} \\
& -K_{h_1}(X_{ik} - x_k) G'\{m(x_k, X_{i\underline{k}})\} \frac{\varphi_{\underline{k}}(X_{i\underline{k}})}{\varphi(x_k, X_{i\underline{k}})} \Bigg] \\
\times \Bigg[& K_{h_1}(X_{lj} - x_j) K_{h_1}(X_{lk} - x_k) G'\{m(x_j, x_k, X_{l\underline{jk}})\} \frac{\varphi_{\underline{jk}}(X_{l\underline{jk}})}{\varphi(x_j, x_k, X_{l\underline{jk}})} \\
& - K_{h_1}(X_{lj} - x_j) G'\{m(x_j, X_{l\underline{j}})\} \frac{\varphi_{\underline{j}}(X_{l\underline{j}})}{\varphi(x_j, X_{l\underline{j}})} \\
& -K_{h_1}(X_{lk} - x_k) G'\{m(x_k, X_{l\underline{k}})\} \frac{\varphi_{\underline{k}}(X_{l\underline{k}})}{\varphi(x_k, X_{l\underline{k}})} \Bigg] \\
\times\, & \sigma(X_i)\sigma(X_l)\varphi_{jk}(x_j, x_k) dx_j dx_k.
\end{aligned}$$

Make the change of variables $u = (x_j - X_{ij})/h_1$ and $v = (x_k - X_{ik})/h_1$. Then

$$\begin{aligned}
A(X_i, X_l) = \frac{1}{n^2} \int \Bigg[& K(u)K(v) G'\{m(X_i)\} \frac{\varphi_{\underline{jk}}(X_{i\underline{jk}})}{\varphi(X_i)h_1} \\
& -K(u) G'\{m(X_i)\} \frac{\varphi_{\underline{j}}(X_{i\underline{j}})}{\varphi(X_i)} - K(v) G'\{m(X_i)\} \frac{\varphi_{\underline{k}}(X_{i\underline{k}})}{\varphi(X_i)} \Bigg] \\
\times \Bigg[& K(u + \frac{X_{ij} - X_{lj}}{h_1}) K(v + \frac{X_{ik} - X_{lk}}{h_1}) \\
& \times G'\{m(X_{ij}, X_{ik}, X_{l\underline{jk}})\} \frac{\varphi_{\underline{jk}}(X_{l\underline{jk}})}{\varphi(X_{ij}, X_{ik}, X_{l\underline{jk}})h_1} \\
& - K(u + \frac{X_{ij} - X_{lj}}{h_1}) G'\{m(X_{ij}, X_{l\underline{j}})\} \frac{\varphi_{\underline{j}}(X_{l\underline{j}})}{\varphi(X_{ij}, X_{l\underline{j}})} \\
& -K(v + \frac{X_{ik} - X_{lk}}{h_1}) G'\{m(X_{ik}, X_{l\underline{k}})\} \frac{\varphi_{\underline{k}}(X_{l\underline{k}})}{\varphi(X_{ik}, X_{l\underline{k}})} \Bigg] \\
\times\, & \sigma(X_i)\sigma(X_l)\varphi_{jk}(X_{ij}, X_{ik}) du dv [1 + o_p(1)].
\end{aligned}$$

After some tedious algebra, one obtains

$$Q_2 = \sum\sum_{i \neq l} \varepsilon_i \varepsilon_l A(X_i, X_l) = \sum\sum_{1 \leq i < l \leq n} (A_1 + A_2 + A_3 + A_4 + A_5)[1 + o_p(1)],$$

where

$$\begin{aligned}
A_1 = {} & \frac{\varepsilon_i \varepsilon_l \sigma(X_i)\sigma(X_l)}{n^2 h_1^2} K^{(2)}\left(\frac{X_{ij} - X_{lj}}{h_1}\right) K^{(2)}\left(\frac{X_{ik} - X_{lk}}{h_1}\right) \varphi_{\underline{jk}}(X_{i\underline{jk}}) \varphi_{\underline{jk}}(X_{l\underline{jk}}) \\
& \times \left\{ G'\{m(X_i)\} G'\{m(X_{ij}, X_{ik}, X_{l\underline{jk}})\} \frac{\varphi_{jk}(X_{ij}, X_{ik})}{\varphi(X_i)\varphi(X_{ij}, X_{ik}, X_{l\underline{jk}})} \right. \\
& \qquad \left. + G'\{m(X_l)\} G'\{m(X_{lj}, X_{lk}, X_{i\underline{jk}})\} \frac{\varphi_{jk}(X_{lj}, X_{lk})}{\varphi(X_l)\varphi(X_{lj}, X_{lk}, X_{i\underline{jk}})} \right\},
\end{aligned}$$

$$\begin{aligned}
A_2 = {} & -\frac{\varepsilon_i \varepsilon_l \sigma(X_i)\sigma(X_l)}{n^2 h_1} G'\{m(X_i)\} \frac{\varphi_{\underline{jk}}(X_{i\underline{jk}})}{\varphi(X_i)} \\
& \times \left\{ K^{(2)}\left(\frac{X_{ij} - X_{lj}}{h_1}\right) G'\{m(X_{ij}, X_{l\underline{j}})\} \frac{\varphi_{\underline{j}}(X_{l\underline{j}})}{\varphi(X_{ij}, X_{l\underline{j}})} \right. \\
& \qquad \left. + K^{(2)}\left(\frac{X_{ik} - X_{lk}}{h_1}\right) G'\{m(X_{ik}, X_{l\underline{k}})\} \frac{\varphi_{\underline{k}}(X_{l\underline{k}})}{\varphi(X_{ik}, X_{l\underline{k}})} \right\} \varphi_{jk}(X_{ij}, X_{ik}) \\
& - \frac{\varepsilon_i \varepsilon_l \sigma(X_i)\sigma(X_l)}{n^2 h_1} G'\{m(X_l)\} \frac{\varphi_{\underline{jk}}(X_{l\underline{jk}})}{\varphi(X_l)} \\
& \times \left\{ K^{(2)}\left(\frac{X_{lj} - X_{ij}}{h_1}\right) G'\{m(X_{lj}, X_{i\underline{j}})\} \frac{\varphi_{\underline{j}}(X_{i\underline{j}})}{\varphi(X_{lj}, X_{i\underline{j}})} \right. \\
& \qquad \left. + K^{(2)}\left(\frac{X_{lk} - X_{ik}}{h_1}\right) G'\{m(X_{lk}, X_{i\underline{k}})\} \frac{\varphi_{\underline{k}}(X_{i\underline{k}})}{\varphi(X_{lk}, X_{i\underline{k}})} \right\} \varphi_{jk}(X_{lj}, X_{lk}),
\end{aligned}$$

$$\begin{aligned}
A_3 = {} & -\frac{\varepsilon_i \varepsilon_l \sigma(X_i)\sigma(X_l)}{n^2 h_1} G'\{m(X_i)\} G'\{m(X_{ij}, X_{ik}, X_{l\underline{jk}})\} \frac{\varphi_{\underline{jk}}(X_{l\underline{jk}})}{\varphi(X_i)\varphi(X_{ij}, X_{ik}, X_{l\underline{jk}})} \\
& \times \left\{ K^{(2)}\left(\frac{X_{ij} - X_{lj}}{h_1}\right) \varphi_{\underline{j}}(X_{i\underline{j}}) + K^{(2)}\left(\frac{X_{ik} - X_{lk}}{h_1}\right) \varphi_{\underline{k}}(X_{i\underline{k}}) \right\} \varphi_{jk}(X_{ij}, X_{ik}) \\
& - \frac{\varepsilon_i \varepsilon_l \sigma(X_i)\sigma(X_l)}{n^2 h_1} G'\{m(X_l)\} G'\{m(X_{lj}, X_{lk}, X_{i\underline{jk}})\} \frac{\varphi_{\underline{jk}}(X_{i\underline{jk}})}{\varphi(X_l)\varphi(X_{lj}, X_{lk}, X_{i\underline{jk}})} \\
& \times \left\{ K^{(2)}\left(\frac{X_{lj} - X_{ij}}{h_1}\right) \varphi_{\underline{j}}(X_{l\underline{j}}) + K^{(2)}\left(\frac{X_{lk} - X_{ik}}{h_1}\right) \varphi_{\underline{k}}(X_{l\underline{k}}) \right\} \varphi_{jk}(X_{lj}, X_{lk}),
\end{aligned}$$

$$\begin{aligned}
A_4 = {}& \frac{\varepsilon_i \varepsilon_l \sigma(X_i)\sigma(X_l)}{n^2} G'\{m(X_i)\} \\
& \times \left\{ K^{(2)}\left(\frac{X_{ij} - X_{lj}}{h_1}\right) \frac{G'\{m(X_{ij}, X_{l\underline{j}})\}\varphi_{\underline{j}}(X_{i\underline{j}})\varphi_{\underline{j}}(X_{l\underline{j}})}{\varphi(X_i)\varphi(X_{ij}, X_{l\underline{j}})} \right. \\
& \qquad \left. + K^{(2)}\left(\frac{X_{ik} - X_{lk}}{h_1}\right) \frac{G'\{m(X_{ik}, X_{l\underline{k}})\}\varphi_{\underline{k}}(X_{i\underline{k}})\varphi_{\underline{k}}(X_{l\underline{k}})}{\varphi(X_i)\varphi(X_{ik}, X_{l\underline{k}})} \right\} \varphi_{jk}(X_{ij}, X_{ik}) \\
& + \frac{\varepsilon_i \varepsilon_l \sigma(X_i)\sigma(X_l)}{n^2} G'\{m(X_l)\} \\
& \times \left\{ K^{(2)}\left(\frac{X_{lj} - X_{ij}}{h_1}\right) \frac{G'\{m(X_{lj}, X_{i\underline{j}})\}\varphi_{\underline{j}}(X_{l\underline{j}})\varphi_{\underline{j}}(X_{i\underline{j}})}{\varphi(X_l)\varphi(X_{lj}, X_{i\underline{j}})} \right. \\
& \qquad \left. + K^{(2)}\left(\frac{X_{lk} - X_{ik}}{h_1}\right) \frac{G'\{m(X_{lk}, X_{i\underline{k}})\}\varphi_{\underline{k}}(X_{l\underline{k}})\varphi_{\underline{k}}(X_{i\underline{k}})}{\varphi(X_l)\varphi(X_{lk}, X_{i\underline{k}})} \right\} \varphi_{jk}(X_{lj}, X_{lk}),
\end{aligned}$$

and

$$\begin{aligned}
A_5 = {}& \frac{\varepsilon_i \varepsilon_l \sigma(X_i)\sigma(X_l)}{n^2} G'\{m(X_i)\} \\
& \times \left\{ G'\{m(X_{ik}, X_{l\underline{k}})\} \frac{\varphi_{\underline{j}}(X_{i\underline{j}})\varphi_{\underline{k}}(X_{l\underline{k}})}{\varphi(X_i)\varphi(X_{ik}, X_{l\underline{k}})} + G'\{m(X_{ij}, X_{l\underline{j}})\} \frac{\varphi_{\underline{j}}(X_{l\underline{j}})\varphi_{\underline{k}}(X_{i\underline{k}})}{\varphi(X_i)\varphi(X_{ij}, X_{l\underline{j}})} \right\} \\
& \times \varphi_{jk}(X_{ij}, X_{ik}) \\
& + \frac{\varepsilon_i \varepsilon_l \sigma(X_i)\sigma(X_l)}{n^2} G'\{m(X_l)\} \\
& \times \left\{ G'\{m(X_{lk}, X_{i\underline{k}})\} \frac{\varphi_{\underline{j}}(X_{l\underline{j}})\varphi_{\underline{k}}(X_{i\underline{k}})}{\varphi(X_l)\varphi(X_{lk}, X_{i\underline{k}})} + G'\{m(X_{lj}, X_{i\underline{j}})\} \frac{\varphi_{\underline{j}}(X_{i\underline{j}})\varphi_{\underline{k}}(X_{l\underline{k}})}{\varphi(X_l)\varphi(X_{lj}, X_{i\underline{j}})} \right\} \\
& \times \varphi_{jk}(X_{lj}, X_{lk}).
\end{aligned}$$

All the A_i, $i = 1, \ldots, 5$, are symmetric functions. Note that the random vectors (X_i, ε_i), $i = 1, \ldots n$, are i.i.d. and

$$\begin{aligned}
E_i\left[\varepsilon_i \varepsilon_l A(X_i, X_l)\right] &= E\left[E\left[\varepsilon_i \varepsilon_l A(X_i, X_l) | \varepsilon_i, X_i, X_l\right] | \varepsilon_i, X_i\right] \\
&= E\left[\varepsilon_i A(X_i, X_l) E(\varepsilon_l | X_l) | \varepsilon_i, X_i\right] = 0,
\end{aligned}$$

where here $E_i = E(\cdot|\varepsilon_i, X_i)$. Therefore A_1 to A_5 and thus Q_2 are all symmetric and nondegenerate U-statistics. We will derive the asymptotic variance of A_1. It will be seen in the process that A_2 through A_5 are of higher order and hence negligible. We will consider A_1 in the following.

By Hall (1984), to apply a central limit theorem to this U-statistic, the following three quantities need to be calculated:

1. The variance of one term: $B_n = E[A_1^2(X_1, \varepsilon_1, X_2, \varepsilon_2)]$,

2. The fourth moment of one term: $C_n = E[A_1^4(X_1, \varepsilon_1, X_2, \varepsilon_2)]$,

3. The quantity $D_n = E[J_1^2(X_1, \varepsilon_1, X_2, \varepsilon_2)]$, where

$$J_1(x, \varepsilon, y, \delta) = E[A_1(X_1, \varepsilon_1, x, \varepsilon) A_1(X_1, \varepsilon_1, y, \delta)].$$

Here ε and δ are independent and disributed as ε_1.

Then one must verify that $[D_n + (1/n)C_n]/B_n^2 \to 0$ as $n \to \infty$. The following lemmas will address the orders of these quantities when n goes to infinity.

Lemma 3.4.2. *As $n \to \infty$, it holds that*

$$B_n = \frac{4}{n^4 h_1^2} \|K^{(2)}\|_2^4 \int \left[\int \left\{ \frac{(G' \circ m)^2 \sigma^2}{\varphi} \right\} (z_j, z_k, z_{\underline{jk}}) \varphi_{\underline{jk}}^2 (z_{\underline{jk}}) dz_{\underline{jk}} \right]^2 \times \varphi_{jk}^2 (z_j, z_k) dz_j dz_k \left[1 + o_p(1)\right].$$

Proof.

$$\begin{aligned}B_n = \frac{1}{n^4h_1^4}\int & (K^{(2)})^2\left(\frac{z_{1j}-z_{2j}}{h_1}\right)(K^{(2)})^2\left(\frac{z_{1k}-z_{2k}}{h_1}\right)\varphi^2_{\underline{jk}}(z_{1\underline{jk}})\varphi^2_{\underline{jk}}(z_{2\underline{jk}}) \\ & \times \left\{G'\{m(z_1)\}G'\{m(z_{1j},z_{1k},z_{2\underline{jk}})\}\frac{\varphi_{jk}(z_{1j},z_{1k})}{\varphi(z_1)\varphi(z_{1j},z_{1k},z_{2\underline{jk}})}\right. \\ & \left. +G'\{m(z_2)\}G'\{m(z_{2j},z_{2k},z_{1\underline{jk}})\}\frac{\varphi_{jk}(z_{2j},z_{2k})}{\varphi(z_2)\varphi(z_{2j},z_{2k},z_{1\underline{jk}})}\right\}^2 \\ & \times \sigma^2(z_1)\sigma^2(z_2)\varphi(z_1)\varphi(z_2)dz_1dz_2.\end{aligned}$$

Using a substitution $z_{2j} = z_{1j} - h_1u$ and $z_{2k} = z_{1k} - h_1v$, one obtains

$$\begin{aligned}B_n = \frac{1}{n^4h_1^2}\int & (K^{(2)})^2(u)(K^{(2)})^2(v)\varphi^2_{\underline{jk}}(z_{1\underline{jk}})\varphi^2_{\underline{jk}}(z_{2\underline{jk}})\sigma^2(z_1)\sigma^2(z_{1j},z_{1k},z_{2\underline{jk}}) \\ & \times \left\{G'\{m(z_1)\}G'\{m(z_{1j},z_{1k},z_{2\underline{jk}})\}\frac{\varphi_{jk}(z_{1j},z_{1k})}{\varphi(z_1)\varphi(z_{1j},z_{1k},z_{2\underline{jk}})}\right. \\ & \left. +G'\{m(z_{1j},z_{1k},z_{2\underline{jk}})\}G'\{m(z_1)\}\frac{\varphi_{jk}(z_{1j},z_{1k})}{\varphi(z_{1j},z_{1k},z_{2\underline{jk}})\varphi(z_2)}\right\}^2 \\ & \times \varphi(z_1)\varphi(z_{1j},z_{1k},z_{2\underline{jk}})dz_1dudvdz_{2\underline{jk}}\left[1+o_p(1)\right] \\ = \frac{4}{n^4h_1^2}\|K^{(2)}\|_2^4\int & \frac{\varphi^2_{\underline{jk}}(z_{1\underline{jk}})\varphi^2_{\underline{jk}}(z_{2\underline{jk}})\varphi^2_{jk}(z_{1j},z_{1k})}{\varphi(z_1)\varphi(z_{1j},z_{1k},z_{2\underline{jk}})}G'^2\{m(z_1)\} \\ & \times G'^2\{m(z_{1j},z_{1k},z_{2\underline{jk}})\}\sigma^2(z_1)\sigma^2(z_{1j},z_{1k},z_{2\underline{jk}})dz_1dz_{2\underline{jk}} \\ & \times [1+o_p(1)] \\ = \frac{4}{n^4h_1^2}\|K^{(2)}\|_2^4\int & \left[\int\left\{\frac{(G'\circ m)^2\sigma^2}{\varphi}\right\}(z_j,z_k,z_{\underline{jk}})\varphi^2_{\underline{jk}}(z_{\underline{jk}})dz_{\underline{jk}}\right]^2 \\ & \times \varphi^2_{jk}(z_j,z_k)dz_jdz_k\left[1+o_p(1)\right].\end{aligned}$$

Note that $B_n = (2\sigma_T^2/n^4h_1^2)[1+o_p(1)]$, where σ_T^2 is defined in Theorem 3.3.1.

□

Lemma 3.4.3. *As $n \to \infty, n^{-1}C_n = O(n^{-9}h_1^{-6}) = O(B_n^2)$.*

Proof. Similar to the case of the second moment, the fourth moment of $A_1(X_1, \varepsilon_1, X_2, \varepsilon_2)$ is

$$\begin{aligned}
C_n &= E[A_1^4(X_1, \varepsilon_1, X_2, \varepsilon_2)] \\
&= \frac{1}{n^8 h_1^6} \int (K^{(2)})^4(u)(K^{(2)})^4(v)\varphi^4_{\underline{jk}}(z_{1\underline{jk}})\varphi^4_{\underline{jk}}(z_{2\underline{jk}})\sigma^4(z_1)\sigma^4(z_{1j}, z_{1k}, z_{2\underline{jk}}) \\
&\qquad \times \Bigg\{ G'\{m(z_1)\}G'\{m(z_{1j}, z_{1k}, z_{2\underline{jk}})\}\frac{\varphi_{jk}(z_{1j}, z_{1k})}{\varphi(z_1)\varphi(z_{1j}, z_{1k}, z_{2\underline{jk}})} \\
&\qquad\quad + G'\{m(z_{1j}, z_{1k}, z_{2\underline{jk}})\}G'\{m(z_1)\}\frac{\varphi_{jk}(z_{1j}, z_{1k})}{\varphi(z_{1j}, z_{1k}, z_{2\underline{jk}})\varphi(z_2)} \Bigg\}^4 \\
&\qquad \times \varphi(z_1)\varphi(z_{1j}, z_{1k}, z_{2\underline{jk}})dz_1 du dv dz_{2\underline{jk}}[1 + o_p(1)].
\end{aligned}$$

This calculation implies that

$$n^{-1}C_n = O(n^{-9}h_1^{-6}) = B_n^2 \cdot O(n^{-1}h_1^{-2})$$

and completes the proof the lemma. □

Lemma 3.4.4. *As $n \to \infty$,*

$$
\begin{aligned}
&J_1(x, \varepsilon, y, \delta) \\
&= \frac{2\varepsilon\delta\varphi_{jk}(x_j, x_k)\varphi_{\underline{jk}}(x_{\underline{jk}})\varphi_{\underline{jk}}(y_{\underline{jk}})\sigma(x)\sigma(y)}{n^4 h_1^2 \varphi(x)} K^{(4)}\left(\frac{x_j - y_j}{h_1}\right) K^{(4)}\left(\frac{x_k - y_k}{h_1}\right) \\
&\quad \times G'\{m(x)\} \int G'\{m(x_j, x_k, z_{\underline{jk}})\} \\
&\qquad\qquad \times \left\{ G'\{m(x_j, x_k, z_{\underline{jk}})\} G'\{m(x_j, x_k, y_{\underline{jk}})\} \frac{\varphi_{jk}(x_j, x_k)}{\varphi(x_j, x_k, z_{\underline{jk}})\varphi(x_j, x_k, y_{\underline{jk}})} \right. \\
&\qquad\qquad\quad \left. + \, G'\{m(y)\} G'\{m(y_j, y_k, z_{\underline{jk}})\} \frac{\varphi_{jk}(y_j, y_k)}{\varphi(y)\varphi(y_j, y_k, z_{\underline{jk}})} \right\} \\
&\qquad\qquad \times \varphi_{\underline{jk}}^2(z_{\underline{jk}})\sigma^2(x_j, x_k, z_{\underline{jk}}) dz_{\underline{jk}} [1 + o_p(1)].
\end{aligned}
$$

Proof. By definition,

$$
\begin{aligned}
J_1(x, \varepsilon, y, \delta) = {}& \frac{\varepsilon\delta\varphi_{\underline{jk}}(x_{\underline{jk}})\varphi_{\underline{jk}}(y_{\underline{jk}})\sigma(x)\sigma(y)}{n^4 h_1^4} \\
& \times \int \varphi_{\underline{jk}}^2(z_{\underline{jk}})\sigma^2(z) K^{(2)}\left(\frac{z_j - x_j}{h_1}\right) K^{(2)}\left(\frac{z_k - x_k}{h_1}\right) \\
& \quad \times \left\{ G'\{m(z)\} G'\{m(z_j, z_k, x_{\underline{jk}})\} \frac{\varphi_{jk}(z_j, z_k)}{\varphi(z)\varphi(z_j, z_k, x_{\underline{jk}})} \right. \\
& \qquad \left. + G'\{m(x)\} G'\{m(x_j, x_k, z_{\underline{jk}})\} \frac{\varphi_{jk}(x_j, x_k)}{\varphi(x)\varphi(x_j, x_k, z_{\underline{jk}})} \right\} \\
& \quad \times K^{(2)}\left(\frac{z_j - y_j}{h_1}\right) K^{(2)}\left(\frac{z_k - y_k}{h_1}\right) \\
& \quad \times \left\{ G'\{m(z)\} G'\{m(z_j, z_k, y_{\underline{jk}})\} \frac{\varphi_{jk}(z_j, z_k)}{\varphi(z)\varphi(z_j, z_k, y_{\underline{jk}})} \right. \\
& \qquad \left. + G'\{m(y)\} G'\{m(y_j, y_k, z_{\underline{jk}})\} \frac{\varphi_{jk}(y_j, y_k)}{\varphi(y)\varphi(y_j, y_k, z_{\underline{jk}})} \right\} \varphi(z) dz,
\end{aligned}
$$

which, by the change of variables $z_j = x_j + h_1 u$ and $z_k = x_k + h_1 v$, becomes

$$
\begin{aligned}
&J_1(x,\varepsilon,y,\delta)\\
&=\frac{\varepsilon\delta\varphi_{\underline{jk}}(x_{\underline{jk}})\varphi_{\underline{jk}}(y_{\underline{jk}})\sigma(x)\sigma(y)}{n^4h_1^2}\\
&\quad\times\int\varphi^2_{\underline{jk}}(z_{\underline{jk}})\sigma^2(x_j,x_k,z_{\underline{jk}})K^{(2)}(u)K^{(2)}(v)\\
&\qquad\times\left\{G'\{m(x_j,x_k,z_{\underline{jk}})\}G'\{m(x)\}\frac{\varphi_{jk}(x_j,x_k)}{\varphi(x_j,x_k,z_{\underline{jk}})\varphi(x)}\right.\\
&\qquad\qquad\left.+\,G'\{m(x)\}G'\{m(x_j,x_k,z_{\underline{jk}})\}\frac{\varphi_{jk}(x_j,x_k)}{\varphi(x)\varphi(x_j,x_k,z_{\underline{jk}})}\right\}\\
&\qquad\times K^{(2)}\left(u+\frac{x_j-y_j}{h_1}\right)K^{(2)}\left(v+\frac{x_k-y_k}{h_1}\right)\\
&\qquad\times\left\{G'\{m(x_j,x_k,z_{\underline{jk}})\}G'\{m(x_j,x_k,y_{\underline{jk}})\}\frac{\varphi_{jk}(x_j,x_k)}{\varphi(x_j,x_k,z_{\underline{jk}})\varphi(x_j,x_k,y_{\underline{jk}})}\right.\\
&\qquad\qquad\left.+\,G'\{m(y)\}G'\{m(y_j,y_k,z_{\underline{jk}})\}\frac{\varphi_{jk}(y_j,y_k)}{\varphi(y)\varphi(y_j,y_k,z_{\underline{jk}})}\right\}\\
&\qquad\times\varphi(x_j,x_k,z_{\underline{jk}})dudvdz_{\underline{jk}}[1+o_p(1)]\\
&=\frac{2\varepsilon\delta\varphi_{jk}(x_j,x_k)\varphi_{\underline{jk}}(x_{\underline{jk}})\varphi_{\underline{jk}}(y_{\underline{jk}})\sigma(x)\sigma(y)}{n^4h_1^2\varphi(x)}\\
&\quad\times G'\{m(x)\}K^{(4)}\left(\frac{x_j-y_j}{h_1}\right)K^{(4)}\left(\frac{x_k-y_k}{h_1}\right)\\
&\quad\times\int G'\{m(x_j,x_k,z_{\underline{jk}})\}\\
&\qquad\times\left\{G'\{m(x_j,x_k,z_{\underline{jk}})\}G'\{m(x_j,x_k,y_{\underline{jk}})\}\frac{\varphi_{jk}(x_j,x_k)}{\varphi(x_j,x_k,z_{\underline{jk}})\varphi(x_j,x_k,y_{\underline{jk}})}\right.\\
&\qquad\qquad\left.+\,G'\{m(y)\}G'\{m(y_j,y_k,z_{\underline{jk}})\}\frac{\varphi_{jk}(y_j,y_k)}{\varphi(y)\varphi(y_j,y_k,z_{\underline{jk}})}\right\}\\
&\qquad\times\varphi^2_{\underline{jk}}(z_{\underline{jk}})\sigma^2(x_j,x_k,z_{\underline{jk}})dz_{\underline{jk}}[1+o_p(1)].
\end{aligned}
$$

□

Lemma 3.4.5. *As $n \to \infty$, it holds that*

$$D_n = O(n^{-8}h_1^{-2}) = o(B_n^2).$$

Proof. By Lemma 3.4.4 and techniques used in the two previous lemmas, this can be easily shown. The tedious calculation is omitted here. □

So far, we have established that $B_n \propto 1/(n^4h_1^2)$, $C_n \propto 1/(n^8h_1^6)$ and $D_n \propto 1/(n^8h_1^2)$. Hence

$$\frac{D_n + n^{-1}C_n}{B_n^2} = O\left(h_1^2 + \frac{1}{nh_1^2}\right) \to 0 \quad \text{as} \quad n \to \infty.$$

Therefore, the central limit theorem for a nondegenerate U-statistic, stated as in Theorem 1 of Hall (1984), implies the following proposition regarding the limiting distribution of Q_2.

Proposition 3.4.6. *As $n \to \infty$,*

$$nh_1Q_2 \xrightarrow{\mathcal{D}} N(0, \sigma_T^2).$$

The next proposition provides an approximation to the "diagonal" term $Q_1 = \sum_{i=1}^n \varepsilon_i^2 A(X_i, X_i)$.

Proposition 3.4.7. *As $n \to \infty$,*

$$Q_1 = \frac{2\{K^{(2)}(0)\}^2}{nh_1^2} \int \left\{\frac{(G' \circ m)^2\sigma^2}{\varphi}\right\}(z_j, z_k, z_{\underline{jk}})\varphi_{\underline{jk}}^2(z_{\underline{jk}})\varphi_{jk}(z_j, z_k)dz_jdz_kdz_{\underline{jk}}$$
$$+ O_p(n^{-1}h_1^{-1}).$$

Proof. We need to calculate the mean and variance of Q_1. We have

$$\begin{aligned}
EQ_1 &= nE[\varepsilon_1^2 A(X_1, X_1)] \\
&= nE[A_1(X_1, \varepsilon_1, X_1, \varepsilon_1)][1 + O_p(h)] \\
&= n\frac{\{K^{(2)}(0)\}^2}{n^2 h_1^2}\int \frac{2G'^2\{m(z)\}\sigma^2(z)}{\varphi^2(z)}\varphi^2_{\underline{jk}}(z_{\underline{jk}})\varphi_{jk}(z_j, z_k)\varphi(z)dz[1 + O_p(h)] \\
&= \frac{2\{K^{(2)}(0)\}^2}{nh_1^2}\int \left\{\frac{(G' \circ m)^2\sigma^2}{\varphi}\right\}(z_j, z_k, z_{\underline{jk}})\varphi^2_{\underline{jk}}(z_{\underline{jk}})\varphi_{jk}(z_j, z_k)dz[1 + O_p(h)]
\end{aligned}$$

and

$$\begin{aligned}
Var(Q_1) &= nVar[\varepsilon_1^2 A(X_1, X_1)] \\
&\le nE[\varepsilon_1^4 A^2(X_1, X_1)] \\
&= n\frac{\{K^{(2)}(0)\}^4}{n^4 h_1^4}\int \frac{4G'^4\{m(z)\}\sigma^4(z)}{\varphi^4(z)}\varphi^4_{\underline{jk}}(z_{\underline{jk}})\varphi^2_{jk}(z_j, z_k)\varphi(z)dz[1 + O_p(h)] \\
&= O_p(\frac{1}{n^3 h_1^4}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
Q_1 = &\frac{2\{K^{(2)}(0)\}^2}{nh_1^2}\int \left\{\frac{(G' \circ m)^2\sigma^2}{\varphi}\right\}(z_j, z_k, z_{\underline{jk}})\varphi^2_{\underline{jk}}(z_{\underline{jk}})\varphi_{jk}(z_j, z_k)dz_j dz_k dz_{\underline{jk}} \\
&+ O_p\left(\frac{1}{nh_1} + \frac{1}{n^{3/2}h_1^2}\right) \\
= &\frac{2\{K^{(2)}(0)\}^2}{nh_1^2}\int \left\{\frac{(G' \circ m)^2\sigma^2}{\varphi}\right\}(z_j, z_k, z_{\underline{jk}})\varphi^2_{\underline{jk}}(z_{\underline{jk}})\varphi_{jk}(z_j, z_k)dz_j dz_k dz_{\underline{jk}} \\
&+ O_p(n^{-1}h_1^{-1}).
\end{aligned}$$

Now putting the results on Q_1 and Q_2 together, based on the derivation at the beginning of the proof, we obtain Theorem 3.3.1. □

3.4.5 Proof of Theorem 3.3.4

Proof. Analogous to f^*_{jk}, the definition of $f^*_{jk,n}(x_j, x_k)$ is

$$f^*_{jk,n}(x_j, x_k) = f_{jk,n}(x_j, x_k) + c_{jk}$$

with $c_{jk} = \int f_{jk,n}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k$.

Thus

$$\begin{aligned} \|f^*_{jk,n}\|^2_{L^2(\mathcal{S}_{jk},\varphi_{jk})} &= \int f^{*2}_{jk,n}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k \\ &= \int f^2_{jk,n}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k + 3c^2_{jk} \\ &\geq \int f^2_{jk,n}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k \\ &\geq a^2_n. \end{aligned}$$

Taking the first partial derivative with respect to x_j on both sides of the equation of model (3.1), we get

$$G'\{m(x)\}\frac{\partial m}{\partial x_j} = f^{(1)}_j(x_j) + \sum_{\gamma \in D_j} f^{(1,0)}_{j\gamma}(x_j, x_\gamma). \tag{3.28}$$

With another partial differentiation on both sides of (3.28),

$$G''\{m(x)\}\left(\frac{\partial m}{\partial x_j}\right)^2 + G'\{m(x)\}\frac{\partial^2 m}{\partial x_j^2} = f^{(2)}_j(x_j) + \sum_{\gamma \in D_j} f^{(2,0)}_{j\gamma}(x_j, x_\gamma).$$

Hence,

$$G'\{m(x)\}\frac{\partial^2 m}{\partial x_j^2} = f^{(2)}_j(x_j) + \sum_{\gamma \in D_j} f^{(2,0)}_{j\gamma}(x_j, x_\gamma) - G''\{m(x)\}\left(\frac{\partial m}{\partial x_j}\right)^2.$$

Substituting the expression on the right hand side within the integral below, we have

$$
\begin{aligned}
&\int \left\{ (G' \circ m) \frac{\partial^2 m}{\partial x_j^2} \right\} (x_j, x_k, x_{\underline{jk}}) \varphi_{\underline{jk}}(x_{\underline{jk}}) dx_{\underline{jk}} \\
&= \int \left\{ f_j^{(2)}(x_j) + \sum_{\gamma \in D_j} f_{j\gamma}^{(2,0)}(x_j, x_\gamma) - G''\{m(x)\} \left(\frac{\partial m}{\partial x_j} \right)^2 (x_j, x_k, x_{\underline{jk}}) \right\} \varphi_{\underline{jk}}(x_{\underline{jk}}) dx_{\underline{jk}} \\
&= f_j^{(2)}(x_j) + f_{jk}^{(2,0)}(x_j, x_k) + \sum_{\gamma \in D_j \bigcap D_k} \int f_{j\gamma}^{(2,0)}(x_j, x_\gamma) \varphi_{\underline{jk}}(x_{\underline{jk}}) dx_{\underline{jk}} \\
&\quad - \int \left\{ (G'' \circ m) \left(\frac{\partial m}{\partial x_j} \right)^2 \right\} (x_j, x_k, x_{\underline{jk}}) \varphi_{\underline{jk}}(x_{\underline{jk}}) dx_{\underline{jk}}.
\end{aligned}
$$

The property of the third term is stated in the following lemma.

Lemma 3.4.8. *For any $\gamma \in D_j \bigcap D_k$, it holds that $\int f_{j\gamma}^{(2,0)}(x_j, x_\gamma) \varphi_{\underline{jk}}(x_{\underline{jk}}) dx_{\underline{jk}} = 0$.*

Proof.

$$
\begin{aligned}
&\int f_{j\gamma}^{(2,0)}(x_j, x_\gamma) \varphi_{\underline{jk}}(x_{\underline{jk}}) dx_{\underline{jk}} \\
&= \frac{\partial^2}{\partial x_j^2} \int f_{j\gamma}(x_j, x_\gamma) \varphi_{\underline{jk}}(x_{\underline{jk}}) dx_{\underline{jk}} \\
&= \frac{\partial^2}{\partial x_j^2} \int \left\{ \int f_{j\gamma}(x_j, x_\gamma) \left[\int \varphi(x_j, x_k, x_{\underline{jk}}) dx_j dx_k \right] dx_{\underline{jk\gamma}} \right\} dx_\gamma \\
&= \frac{\partial^2}{\partial x_j^2} \int f_{j\gamma}(x_j, x_\gamma) \left[\iint \varphi(x_j, x_k, x_\gamma, x_{\underline{jk\gamma}}) dx_j dx_k dx_{\underline{jk\gamma}} \right] dx_\gamma \\
&= \frac{\partial^2}{\partial x_j^2} \int f_{j\gamma}(x_j, x_\gamma) \varphi_\gamma(x_\gamma) dx_\gamma \\
&= 0
\end{aligned}
$$

by the side condition (3.3). □

Therefore,

$$\begin{aligned}&\int \left\{(G' \circ m)\frac{\partial^2 m}{\partial x_j^2}\right\}(x_j, x_k, x_{\underline{jk}})\varphi_{\underline{jk}}(x_{\underline{jk}})dx_{\underline{jk}} \\ &= f_j^{(2)}(x_j) + f_{jk}^{(2,0)}(x_j, x_k) - \int \left\{(G'' \circ m)\left(\frac{\partial m}{\partial x_j}\right)^2\right\}(x_j, x_k, x_{\underline{jk}})\varphi_{\underline{jk}}(x_{\underline{jk}})dx_{\underline{jk}}.\end{aligned} \tag{3.29}$$

Similarly,

$$\begin{aligned}&\int \left\{(G' \circ m)\frac{\partial^2 m}{\partial x_k^2}\right\}(x_j, x_k, x_{\underline{jk}})\varphi_{\underline{jk}}(x_{\underline{jk}})dx_{\underline{jk}} \\ &= f_k^{(2)}(x_k) + f_{jk}^{(0,2)}(x_j, x_k) - \int \left\{(G'' \circ m)\left(\frac{\partial m}{\partial x_k}\right)^2\right\}(x_j, x_k, x_{\underline{jk}})\varphi_{\underline{jk}}(x_{\underline{jk}})dx_{\underline{jk}},\end{aligned} \tag{3.30}$$

$$\begin{aligned}&\int \left\{(G' \circ m)\frac{\partial^2 m}{\partial x_j^2}\right\}(x_j, x_{\underline{j}})\varphi_{\underline{j}}(x_{\underline{j}})dx_{\underline{j}} \\ &= f_j^{(2)}(x_j) - \int \left\{(G'' \circ m)\left(\frac{\partial m}{\partial x_j}\right)^2\right\}(x_j, x_{\underline{j}})\varphi_{\underline{j}}(x_{\underline{j}})dx_{\underline{j}},\end{aligned} \tag{3.31}$$

$$\begin{aligned}&\int \left\{(G' \circ m)\frac{\partial^2 m}{\partial x_k^2}\right\}(x_k, x_{\underline{k}})\varphi_{\underline{k}}(x_{\underline{k}})dx_{\underline{k}} \\ &= f_k^{(2)}(x_k) - \int \left\{(G'' \circ m)\left(\frac{\partial m}{\partial x_k}\right)^2\right\}(x_k, x_{\underline{k}})\varphi_{\underline{k}}(x_{\underline{k}})dx_{\underline{k}}.\end{aligned} \tag{3.32}$$

Combining (3.29) − (3.32), the bias term $B_{jk}(x_j, x_k)$ defined in Theorem 3.2.2

becomes

$$\begin{aligned}B_{jk}(x_j, x_k) = \frac{\mu_2(K)}{2} \Big[& f_{jk}^{(2,0)}(x_j, x_k) + f_{jk}^{(0,2)}(x_j, x_k) \\ & + \int \left\{ (G'' \circ m) \left(\frac{\partial m}{\partial x_j} \right)^2 \right\} (x_j, x_{\underline{j}}) \varphi_{\underline{j}}(x_{\underline{j}}) dx_{\underline{j}} \\ & + \int \left\{ (G'' \circ m) \left(\frac{\partial m}{\partial x_k} \right)^2 \right\} (x_k, x_{\underline{k}}) \varphi_{\underline{k}}(x_{\underline{k}}) dx_{\underline{k}} \\ & - \int \left\{ (G'' \circ m) \left[\left(\frac{\partial m}{\partial x_j} \right)^2 + \left(\frac{\partial m}{\partial x_k} \right)^2 \right] \right\} (x_j, x_k, x_{\underline{jk}}) \\ & \times \varphi_{\underline{jk}}(x_{\underline{jk}}) dx_{\underline{jk}} \Big] .\end{aligned} \tag{3.33}$$

Since by (3.28), we have

$$\begin{aligned}\left\{ (G'' \circ m) \left(\frac{\partial m}{\partial x_j} \right)^2 \right\} (x) &= (G'' \circ m)(x) \left[\frac{f_j^{(1)}(x_j) + \sum_{\gamma \in D_j} f_{j\gamma}^{(1,0)}(x_j, x_\gamma)}{(G' \circ m)(x)} \right]^2 \\ &= \frac{(G'' \circ m)(x)}{(G' \circ m)^2(x)} \left[f_j^{(1)}(x_j) + \sum_{\gamma \in D_j} f_{j\gamma}^{(1,0)}(x_j, x_\gamma) \right]^2 ,\end{aligned}$$

the last three integrals in (3.33) are all bounded under the condition that G' is bounded away from zero, in addition to assumptions $(A3)$ and $(A6)$. Therefore the sum of the three integrals are bounded.

For any $f_{jk,n} \in \mathcal{B}_{jk}(M)$, with compact support $\mathcal{S}_{jk}$ (the support of φ_{jk}), there exists a constant $b > 0$ such that

$$\|B_{jk}\|_{L^2(\mathcal{S}_{jk}, \varphi_{jk})} = \sqrt{\int B_{jk}^2(x_j, x_k) \varphi_{jk}(x_j, x_k) dx_j dx_k} \le bM.$$

Note that although we are dealing with a sequence of functions $(f_{jk,n})_{n=1}^\infty$, the limiting distribution in Theorem 3.3.1 still holds because all the main effects $\{f_j\}_{j=1}^d$

and other interactions $\{f_{\gamma\delta}\}_{1\leq\gamma<\delta\leq d,(\gamma,\delta)\neq(j,k)}$ remain fixed and the second Sobolev seminorm of $f_{jk,n}$ is bounded uniformly for each n. In other words, as $n \to \infty$,

$$\begin{aligned}
T_n' &= nh_1 \int \hat{f}_{jk,n}^{*2}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k - nh_1 \int f_{jk,n}^{*2}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k \\
&\quad - \frac{2\{K^{(2)}(0)\}^2}{h_1} \int \left\{\frac{(G' \circ m)^2\sigma^2}{\varphi}\right\}(z)\varphi_{\underline{jk}}^2(z_{\underline{jk}})\varphi_{jk}(z_j, z_k)dz \\
&\quad - 2nh_1^3 \int f_{jk,n}^{*}(x_j, x_k)B_{jk,n}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k \\
&\xrightarrow{\mathcal{D}} N(0, \sigma_T^2)
\end{aligned}$$

where $B_{jk,n}$ is the bias function associated with $f_{jk,n}$.

Note that

$$\begin{aligned}
t_n &= nh_1 \int f_{jk,n}^{*2}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k \\
&\quad + 2nh_1^3 \int f_{jk,n}^{*}(x_j, x_k)B_{jk,n}(x_j, x_k)\varphi_{jk}(x_j, x_k)dx_j dx_k \\
&\geq nh_1 \|f_{jk,n}^*\|_{L^2(\mathcal{S}_{jk},\varphi_{jk})}^2 - 2nh_1^3 \|f_{jk,n}^*\|_{L^2(\mathcal{S}_{jk},\varphi_{jk})} \|B_{jk,n}\|_{L^2(\mathcal{S}_{jk},\varphi_{jk})} \\
&= nh_1 \|f_{jk,n}^*\|_{L^2(\mathcal{S}_{jk},\varphi_{jk})} \left[\|f_{jk,n}^*\|_{L^2(\mathcal{S}_{jk},\varphi_{jk})} - 2h_1^2 \|B_{jk,n}\|_{L^2(\mathcal{S}_{jk},\varphi_{jk})}\right] \\
&\geq nh_1 a_n(a_n - 2h_1^2 bM).
\end{aligned}$$

Thus $t_n \to \infty$ as $n \to \infty$ if $a_n^{-1} = o(nh_1 + h_1^{-2})$.

The rejection probability is

$$p_n = P(T_n' + t_n \geq z_{1-\alpha}\sigma_T)$$

It is obvious that as $n \to \infty, t_n \to \infty$ makes $\lim_{n\to\infty} p_n = 1$.

□

Chapter 4

Simulation Studies

In this chapter, we report the results of our finite sample simulation studies. The purpose of the numerical studies is twofold: first to investigate the computational performance of the proposed estimating procedure with finite samples and then to investigate the power of the test statistic for significance of the individual second order interaction term. Although our study is limited in scope, the results do demonstrate the computational feasibility and the power of the test proposed in this dissertation.

4.1 Function Estimation

Sperlich, Tjøstheim and Yang (2002) examined the small sample behavior of the estimators for the identity link function $G(\cdot)$. Although the introduction of a nontrivial link function looks straightforward for the marginal integration method, in practice it can bring on some numerical difficulties and negative effects on small sample performance. In addition, the classical marginal integration estimator entails long computing time. Thus we tried two different R functions to obtain the pre-smoother: one (coded by the author) strictly follows the definition of the multivariate local polynomial regression estimator as defined in (3.11) and (3.13): that is, locally linear in the direction of interest and locally constant in nuisance directions, while the other one is the well-known R function *loess*(), which may be the most commonly available software for a local polynomial surface fit. The main difference of the two implementations is that the former uses fixed bandwidths and the latter is a nearest-neighbor smoother which requires one to select a span. An advantage of *loess* is its short computing time. On the other hand, it loses the flexibility of controlling bandwidth in different directions. Our Monte Carlo experiments employed both methods to see if they will provide comparable results.

Another question is the choice of bandwidth (or span) which is very important in kernel-based nonparametric regression problems. However, there does not yet exist a really complete and practically useful guidance on how to choose the smoothing parameters in our problems. Cross-Validation (CV) seems to be a commonly used bandwidth selector but it aims to minimize the mean squared error of the entire estimated regression function, not of any particular component function. As discussed in section 2.2.1.3, the plug-in methods suggested by Severance-Lossin and Sperlich (1999) might be more appropriate in our current setting.

Let us consider the estimation of the univariate function $F_j(\cdot)$. Recall assumption (A02) that $h_1 = \beta n^{-1/5}$. Following Theorem 3.2.1, the asymptotically optimal bandwidth constant β, with respect to the integrated mean squared error (MISE), is given by

$$\begin{aligned}\beta &= \left[\frac{\int \nu_j(x_j)\varphi_j(x_j)dx_j}{4\int b_j^2(x_j)\varphi_j(x_j)dx_j}\right]^{1/5}\\ &= \left[\frac{\|K\|_2^2 \int\{(G'\circ m)^2\sigma^2/\varphi\}(x_j, x_{\underline{j}})\varphi_{\underline{j}}^2(x_{\underline{j}})dx_{\underline{j}}dx_j}{\mu_2^2(K)\int(\int\{(G'\circ m)(\partial^2/\partial x_j^2)m\}(x_j, x_{\underline{j}})\varphi_{\underline{j}}(x_{\underline{j}})dx_{\underline{j}})^2\varphi_j(x_j)dx_j}\right]^{1/5}\end{aligned}$$

after substituting the bias function $b_j(x_j)$ and variance function $\nu_j(x_j)$ by their formulas defined in (3.17) and (3.18), respectively. Although the expression above could theoretically identify the asymptotically optimal bandwidth, in practice, it is difficult to have an accurate guess either for the parametric regression function or for the nonparametric estimators. In addition, it is not necessarily the best for any given set of data. Therefore, in our simulation studies, we chose smoothing parameters based on experimentation. The final bandwidths or spans (for *loess*) were selected as a fair compromise between a reasonable degree of smoothness and numerical feasibility.

Here, we consider a logistic GAM with interactions, which takes the following form:

$$\log\left(\frac{m(x)}{1-m(x)}\right) = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_{12}(x_1, x_2). \tag{4.1}$$

Note that we use the logit link function $G(u) = \log(u/(1-u))$. Given the covariate vector $X = (X_1, X_2, X_3)^T$, the binary outcome Y was generated from the distribution $Y \sim \mathit{Bernoulli}(m(X))$, with X_1, X_2 and X_3 drawn as independent random variables distributed as $U[-2,2]$. We examined two sample sizes, $n = 500$ and $n = 1000$, and considered the component functions:

$$f_1(x_1) = 2x_1, \quad f_2(x_2) = 1.5\sin(-1.5x_2),$$
$$f_3(x_3) = -x_3^2 + \frac{4}{3}, \quad \text{and} \quad f_{12}(x_1, x_2) = x_1 x_2.$$

This choice of component functions was made in previous simulation studies (Sperlich, Tjøstheim and Yang, 2002) and enables us to make comparisons with other work.

For the kernel in all estimators, we used the quartic kernel

$$K(t) = \frac{15}{16}(1 - t^2)^2 I[|t| \leq 1],$$

and a product of two kernels of this type as a two-dimensional kernel. When estimating the component functions, we used $h_1 = 1.2$ for $n = 500$ and $h_1 = 1.07$ for the larger sample size $n = 1000$. For simplification we set $h_2 = h_1$ for the directions not of interest.

For 100 Monte Carlo replications we estimated the functions on an equally spaced grid. Figures 4.1 – 4.4 show the performance of the proposed estimator with two samples sizes. In Figures 4.1 and 4.3, the true univariate data generating functions are given as dashed lines together with the 90% confidence bands (solid lines) for the estimator resulting from the 100 simulation runs. The average results of the estimated interaction surface $\hat{f}_{12}$ are depicted in Figures 4.2 and 4.4. The corresponding heat map and contour map of the estimated interaction are also given in the same figure. To get an impression of how the *loess* function in R works, the corresponding results were given along the right column. We used a span= 0.10 for $n = 500$ and a span= 0.05 for $n = 1000$. Table 4.1 summarizes averaged squared bias, variance and MASE (averaged mean squared error). Due to the poor performance of the estimators near the boundaries, all the graphs and numerical results are presented over a trimmed region of data on $[-1.9, 1.9]$.

The procedure seems to work reasonably well. *Loess* provided comparable results to those obtained from the fixed bandwidth local polynomial regression estimator. Not surprisingly, the bias can be seen clearly when the link function is not trivial. We can further recognize the boundary effects. As discussed before, the chosen bandwidths are not optimal but the results are quite reasonable. One of our findings from experimentation is that the performance of the estimators is sensitive to the different choices of bandwidths. Again, this constitutes an open problem and needs more intensive investigation and computation.

4.2 Interaction Testing

We present the simulation results for the testing problem in this section. As indicated in Section 3.3, we have to be cautious when using the asymptotic distribution with small or moderate sample sizes. In addition, asymptotic critical values are hard to calculate for the complicated expressions of the bias and variance terms of the test

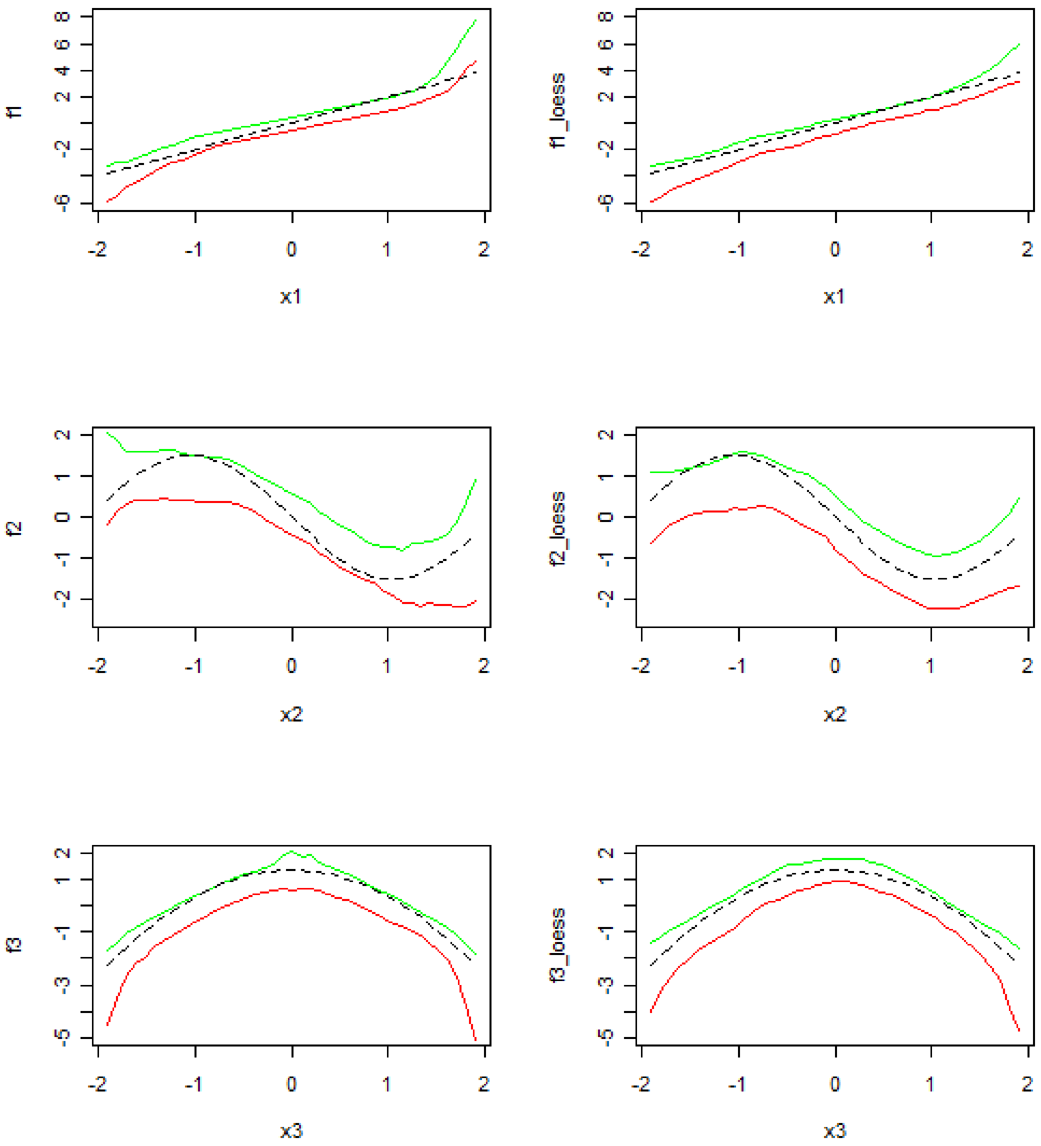

Figure 4.1: Model (4.1) with sample size n=500. Dashed lines are the data generating functions, solid lines are the 90% confidence bands after 100 simulation runs. Results with fixed bandwidth code are on the left and results with fixed span (*loess*) are on the right.

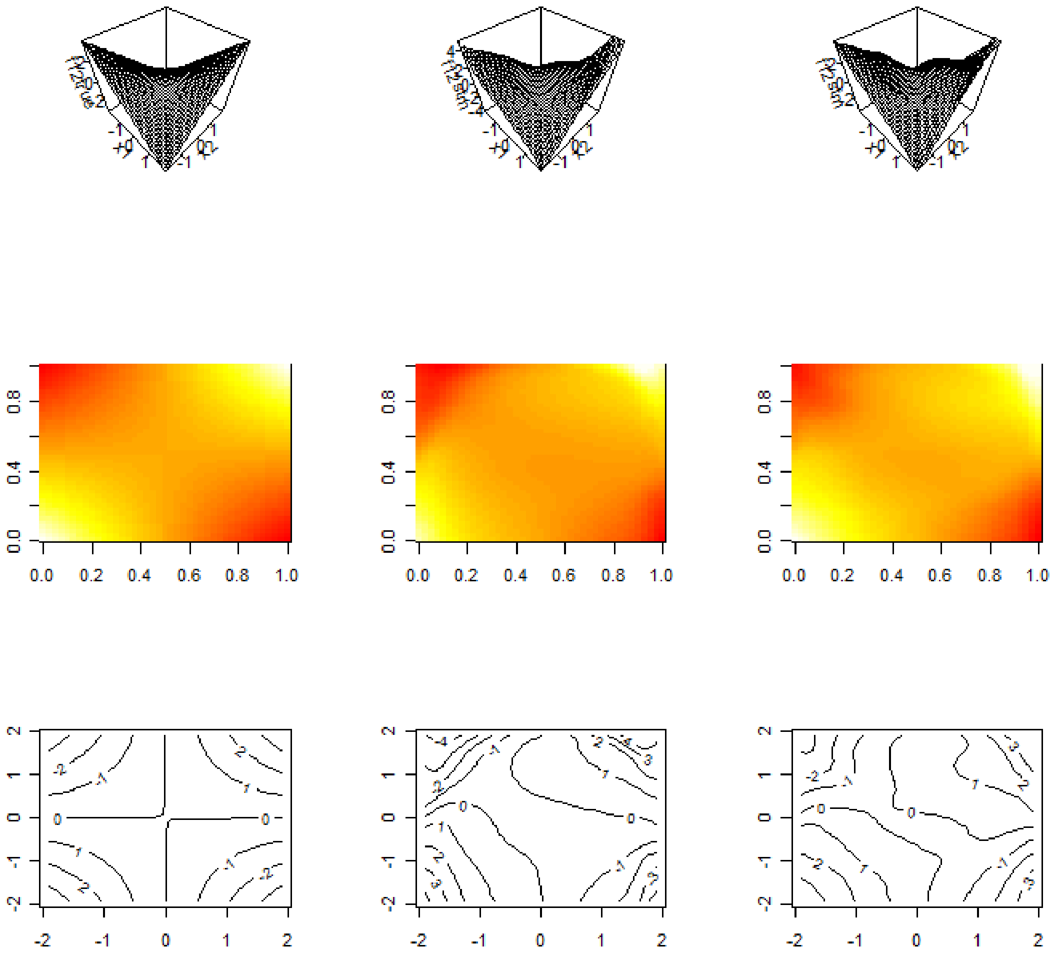

Figure 4.2: Model (4.1) with sample size n=500. Grid plot (upper row), heat map (middle row) and contour map (lower row) for interaction function. True interaction function is on the left, estimator with fixed bandwidth code is in the middle and estimator with fixed span (*loess*) is on the right.

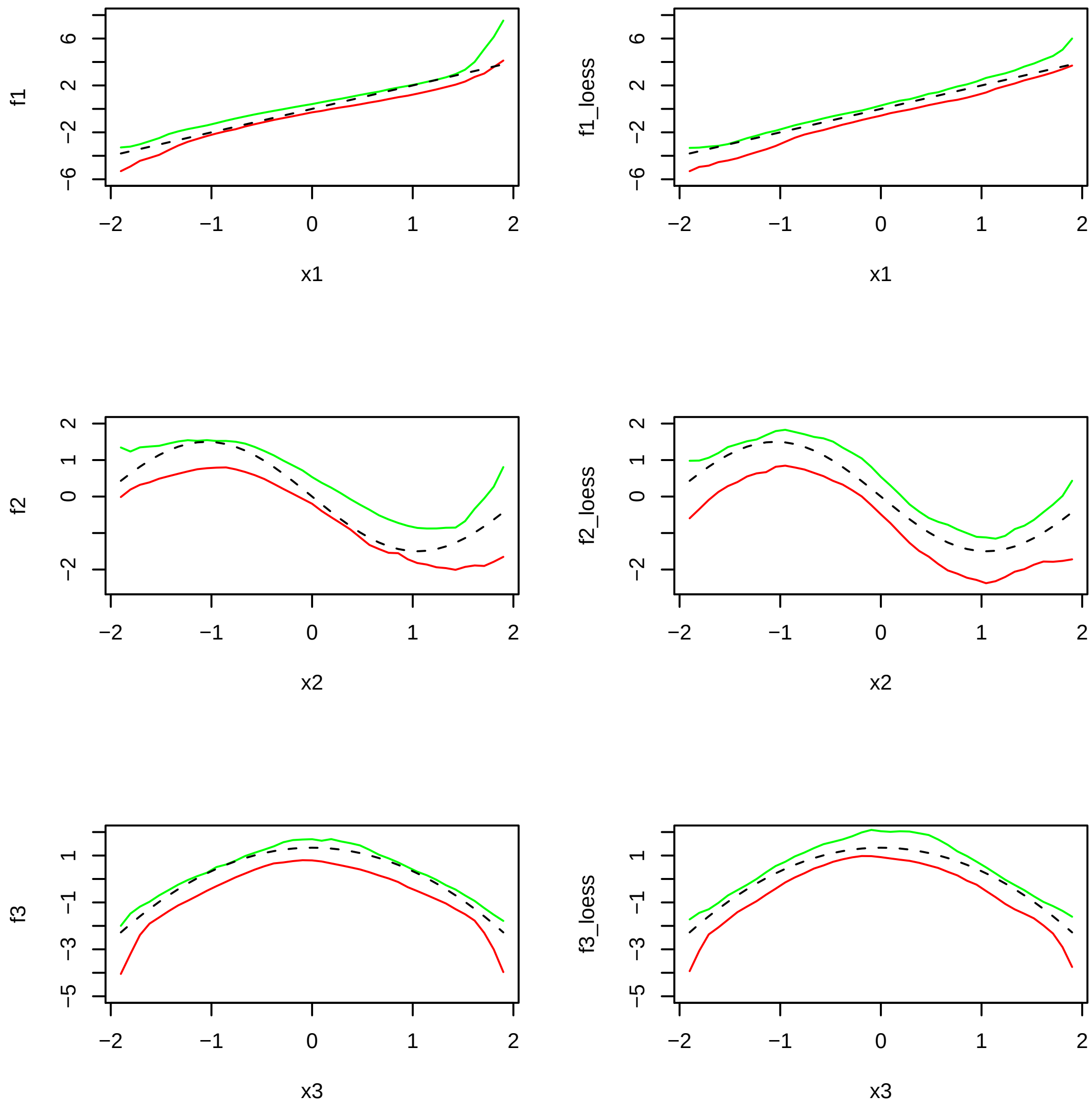

Figure 4.3: Model (4.1) with sample size n=1000. Dashed lines are the data generating functions, solid lines are the 90% confidence bands after 100 simulation runs. Results with fixed bandwidth code are on the left and results with fixed span (*loess*) are on the right.

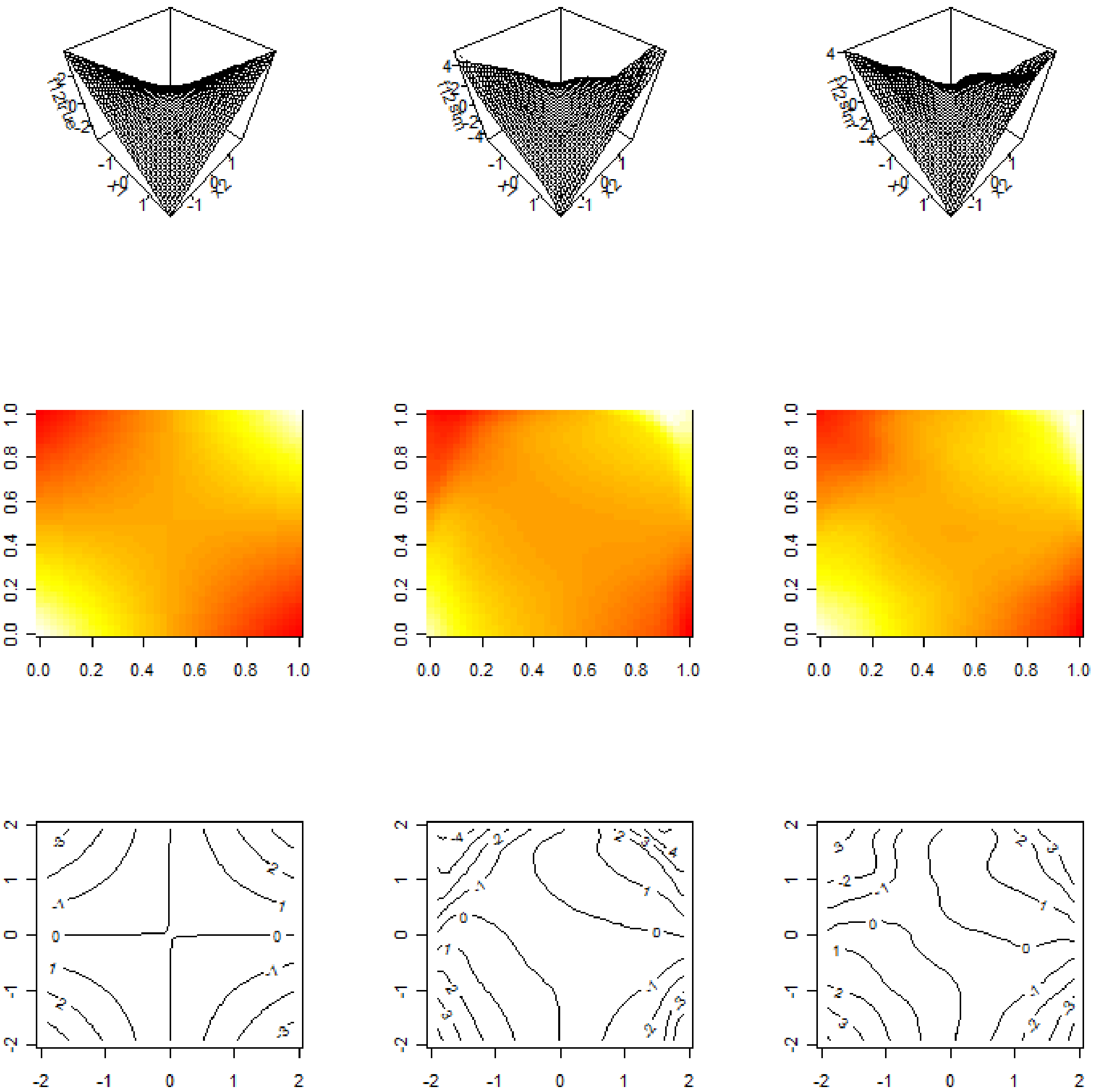

Figure 4.4: Model (4.1) with sample size n=1000. Grid plot (upper row), heat map (middle row) and contour map (lower row) for interaction function. True interaction function is on the left, estimator with fixed bandwidth code is on the middle and estimator with fixed span (*loess*) is on the right.

Table 4.1: Averaged squared bias, averaged variance and MASE, using the fixed bandwidth code and fixed span function (*loess*)

		$n = 500$			$n = 1000$		
Function	Method	Bias^2	Var	MSE	Bias^2	Var	MSE
$\hat{f}_1$	fixed bw	0.329	0.211	0.540	0.224	0.126	0.350
	fixed span	0.146	0.212	0.358	0.150	0.133	0.283
$\hat{f}_2$	fixed bw	0.068	0.169	0.237	0.048	0.094	0.142
	fixed span	0.091	0.184	0.274	0.034	0.117	0.151
$\hat{f}_3$	fixed bw	0.122	0.170	0.292	0.075	0.093	0.168
	fixed span	0.058	0.184	0.242	0.048	0.117	0.165
$\hat{f}_{12}$	fixed bw	0.175	0.473	0.647	0.158	0.293	0.451
	fixed span	0.120	0.362	0.483	0.089	0.327	0.416
$\hat{G}\{\tilde{m}\}$	fixed bw	0.577	0.824	1.401	0.433	0.510	0.943
	fixed span	0.305	0.771	1.076	0.300	0.642	0.942

statistic, as seen in Theorem 3.3.1. Moreover, it is known that the distribution of a similar test functional (Hjellvik, Yao and Tjøstheim, 1998) is poorly approximated by its asymptotic distribution in moderate sized sample.

For these reasons most authors propose the application of the wild bootstrap in this context (see, for example, Gozalo and Linton, 2001; Sperlich, Tjøstheim and Yang, 2002; Yang, Sperlich and Härdle, 2003; Härdle, et al., 2004). The wild bootstrap was first introduced by Wu (1986) and Liu (1988). Härdle and Mammen (1993) introduced it into the context of nonparametric hypothesis testing. We will provide the details of the wild bootstrap in next subsection.

We conducted small simulations on two models with logit and log link functions, respectively. The components of the three-dimensional explanatory variable X were again drawn independently from $U[-2,2]$. The two models are:

$$\textit{Model 1}: \quad \log\left(\frac{m(x)}{1-m(x)}\right) = f_1(x_1) + f_2(x_2) + f_3(x_3) + f_{12}(x_1, x_2),$$

where $f_1(x_1) = \sin(\pi(x_1+2)/2)$, $\quad f_2(x_2) = x_2^2 - (4/3)$, $\quad f_3(x_3) = x_3$, $\quad$ and $f_{12}(x_1, x_2) = ax_1x_2$ with a being a constant. For Model 1, the binary outcome variable Y was generated from $Bernoulli(m(X))$ for a given vector X. Two hundred and five hundred independent samples $\{(X_i, Y_i)\}_{i=1}^n$ were drawn based on this model definition. A very similar model was used in a previous simulation study by Roca-Pardiñas, Cadarso-Suárez and González-Manteiga (2005). Model 2 is:

$$\textit{Model 2}: \quad \log(m(x)) = c + f_1(x_1) + f_2(x_2) + f_3(x_3) + f_{12}(x_1, x_2),$$

where $f_1(x_1) = x_1, \quad f_2(x_2) = 3\sin(-3x_2/2)/4, \quad f_3(x_3) = -x_3^2/2 + (2/3)$, $f_{12}(x_1, x_2) = ax_1x_2$ with a being a constant, and $c = 3/2$. We work with Poisson data generated from $Y \sim Poisson(m(X))$ in Model 2, where a single sample size of $n = 200$ was used to study the performance of the test. It should be noted that the value $a = 0$ corresponds to the null hypothesis of no interaction ($H_{12}^0 : f_{12} = 0$), and that the more the constant a shifts from zero, the greater the degree of interaction.

For all computations in this section, the quartic kernel was used as before. We used bandwidths $h_1 = 1.25$ and $h_1 = 1.07$ for sample sizes $n = 200$ and $n = 500$, respectively, in Model 1, while $h_1 = 1.45$ was applied with the Poisson data generated from Model 2. We again let $h_1 = h_2$ in the local linear smoother for simplicity.

Let us first look at the asymptotics based on 200 Monte Carlo simulations of Model 1. In Figure 4.5, a density estimate for the standardized test statistic $\tilde{T}$ was plotted together with the curve of standard normal density. It is obviously seen from this figure that the normal approximation is quite inaccurate for a sample size of $n = 500$. The normal q-q plot given in Figure 4.6 shows consistent skewed behavior of the distribution of the test statistic. Hence, even though we could estimate bias and variance of the test statistics well, its asymptotic distribution is hardly useful in testing where there is a small or moderate sample size. Our conclusion is consistent with those of Sperlich, Tjøstheim and Yang (2002) and Hjellvik, Yao and Tjøstheim (1998). A possible reason for the poor approximation is that unlike a standard parametric situation, the next order terms in the Edgeworth expansion of our statistics are very close to the leading normal approximation terms (Hjellvik, Yao and Tjøstheim, 1998). Thus very many observations are needed for the dominance of the first-order term to yield normality.

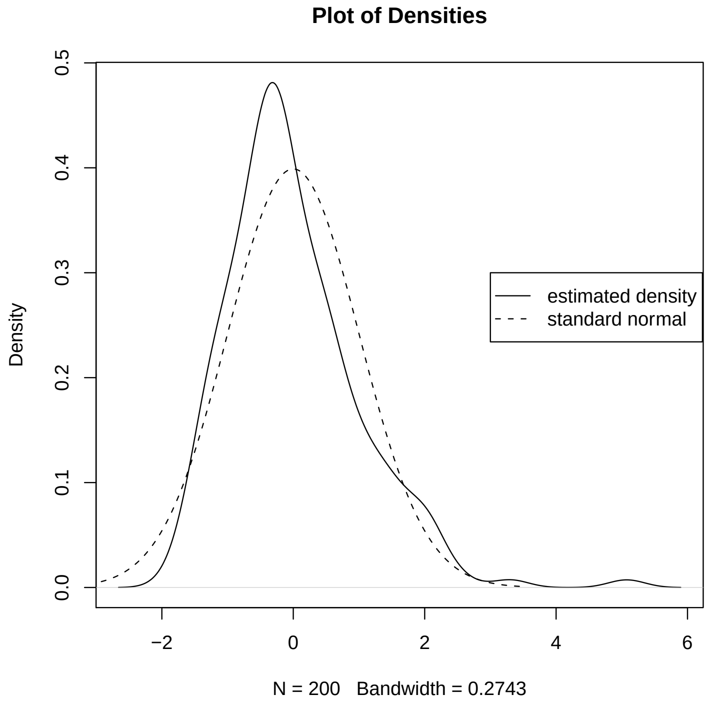

Figure 4.5: Density of the test statistic (solid) and standard normal density (dashed).

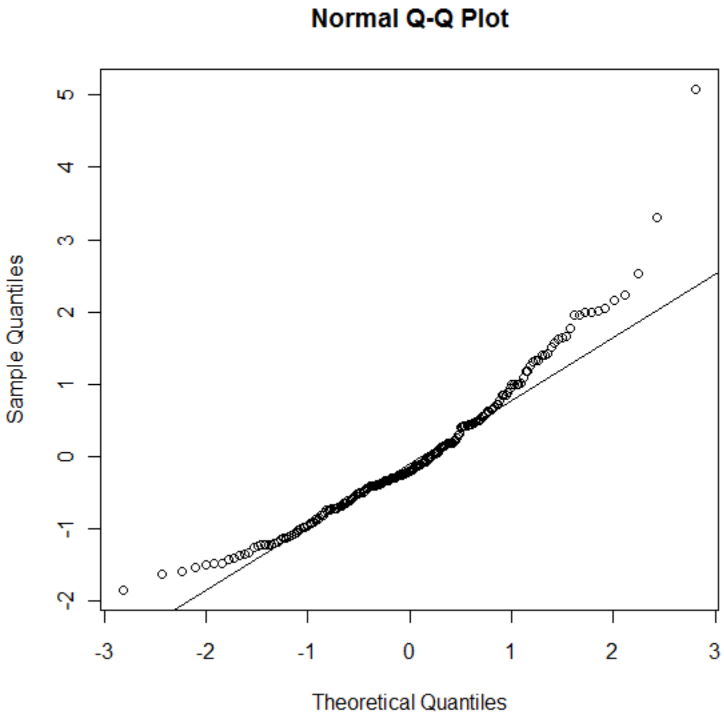

Figure 4.6: Q-Q plot of the test statistic (dots) and Q-Q line of the theoretical normal distribution (solid line) with $n = 500$.

4.2.1 The Wild Bootstrap

The basic idea of wild bootstrap is to draw each bootstrap residual u_i^*, $i = 1, \ldots, n$, from a distribution ${F_i}^W$ such that

$E_{{F_i}^W}(u_i^*) = 0, \quad E_{{F_i}^W}(u_i^{*2}) = \hat{u}_i^2 \quad \text{and} \quad E_{{F_i}^W}(u_i^{*3}) = \hat{u}_i^3,$

where the residual $\hat{u}_i$ is estimated under the null hypothesis. Then $Y_i^* = \tilde{m}_0(X_i)+u_i^*$. This approach is used when the range of Y is the real line $\mathbb{R}$.

When the range of Y is restricted to a subset of $\mathbb{R}$, for instance $\{0, 1\}$, a different approach is needed. If Y is binary, then Y_i^* is chosen from the Bernoulli distribution with parameter $p_i^W = \tilde{m}_0(X_i)$. If Y is Poisson, Y_i^* is chosen from the Poisson distribution with parameter $\lambda_i^W = \tilde{m}_0(X_i)$.

In either case the distribution of u_i^* depends only on one value of the estimated regression, leading to the name "wild bootstrap." See Gozalo and Linton (2001) for details.

For each bootstrap sample, a bootstrap test statistic $\tilde{T}^*$ is calculated. Since the $\tilde{T}^*$ are distributed as T under H_0, repeating this procedure many times one gets a simulated critical value under the null hypothesis and a simulated p - value for T. The bootstrap steps used in our simulations are therefore:

Step 1: Calculate the estimated regression function $\tilde{m}_0(X_i)$, $i = 1, \ldots, n$, under the hypothesis $H_{12}^0 : f_{12} = 0$.

Step 2: Generate Y_i^*, $i = 1, \ldots, n$ as a random draw from a Bernoulli $(\tilde{m}_0(X_i))$ for Model 1 or a Poisson $(\tilde{m}_0(X_i))$ for Model 2.

Step 3: Calculate the bootstrap test statistic $\tilde{T}^*$, in the identical way that $\tilde{T}$ was computed from the original sample.

Step 4: Repeat steps $2 - 3$ B times and use the B values of $\tilde{T}^*$ to determine the quantiles of the test statistic under H_{12}^0 and subsequently the critical values or $p-$values.

4.2.2 Simulations

For the wild bootstrap, we drew Y_i^*, $i = 1, \ldots, n$, from the estimated data generating process under H_{12}^0, given $\{X_i\}_{i=1}^n$ and calculated the corresponding test statistic $\tilde{T}^* = n^{-1} \sum_{i=1}^n \hat{f}_{12}^{*2}(X_1, X_2)$. To get $\hat{f}_{12}^*$, we used the local linear smoother with fixed bandwidths h_1 and h_2. Only $B = 200$ bootstrap iterations were implemented to approximate the distribution of $\tilde{T}$, due to the limitation of computing power. In practice, one should certainly draw about 1000 bootstrap samples to get a satisfactory approximation. All rejection probabilities were determined by performing 500 replications.

For both models, we tested $H_{12}^0 : f_{12} = 0$ for data generated with various values of the constant a on the range $[0, 1]$. Table 4.2 shows the percentage of rejections of the alternative hypothesis based on the 1%, 5%, 10% and 15% empirical critical values under the first model with sample size $n = 200$. Table 4.3 gives the results for the same model but with a larger sample size of $n = 500$. The relative rejection frequencies for the Poisson data generated based on Model 2 are presented in Table 4.4. The corresponding power functions are depicted in Figures 4.7 and 4.8, for different models and sample sizes. The tables also display the average bootstrap p-values, averaged over the 500 Monte Carlo replications.

As can be seen from the tables, the proposed test provided satisfactory results overall, with a type I error very close to the nominal levels in evidence for $a = 0$ and well fitted p-values. The figures show how fast the probability of rejection rises in response to an increase in the value of the constant a.

We found that, for $n = 200$, the bandwidth choice can be very crucial to obtaining accurate control of the error of the first kind with the aid of the wild bootstrap. Although the wild bootstrap appears to work reasonably well, we chose bandwidths somewhat arbitrary. It is obvious that a much more intensive simulation study would be of interest to investigate the performance of the test, particularly concerning the interplay between model complexity and choice of bandwidth.

Table 4.2: Percentage of rejection and p-values for testing $H_{12}^0 : f_{12} = 0$, using local linear smoother with $h = g = 1.25$, based on Model 1 (binary data and logit link): $M = 500$, $n = 200$, $B = 200$

a	Significance level (%)				
	1	5	10	15	mean p-value
0.00	1.4	5.6	12.4	18.6	50.0
0.25	11.0	23.8	32.8	40.2	32.4
0.50	46.0	64.6	74.6	80.6	10.4
0.75	85.2	93.8	95.8	96.6	2.1
1.00	98.0	99.6	100	100	0.08

Table 4.3: Percentage of rejection and p-values for testing $H_{12}^0 : f_{12} = 0$, using local linear smoother with $h = g = 1.07$, based on Model 1 (binary data and logit link): $M = 500$, $n = 500$, $B = 200$

a	Significance level (%)				
	1	5	10	15	mean p-value
0.00	1.6	6.2	11.6	14.8	51.3
0.25	20.4	36.2	45.6	53.6	25.0
0.50	78.2	91.0	93.8	94.4	2.4
1.00	100	100	100	100	0

Table 4.4: Percentage of rejection and p-values for testing $H_{12}^0 : f_{12} = 0$, using local linear smoother with $h = g = 1.45$, based on Model 2 (Poisson data and log link): $M = 500$, $n = 200$, $B = 200$

a	Significance level (%)				
	1	5	10	15	mean p-value
0.00	3.2	6.6	12.0	16.4	55.4
0.25	7.2	15.6	21.8	26.4	50.7
0.50	77.8	84.6	89.6	91.8	4.6
0.75	99.8	100	100	100	0
1.00	100	100	100	100	0

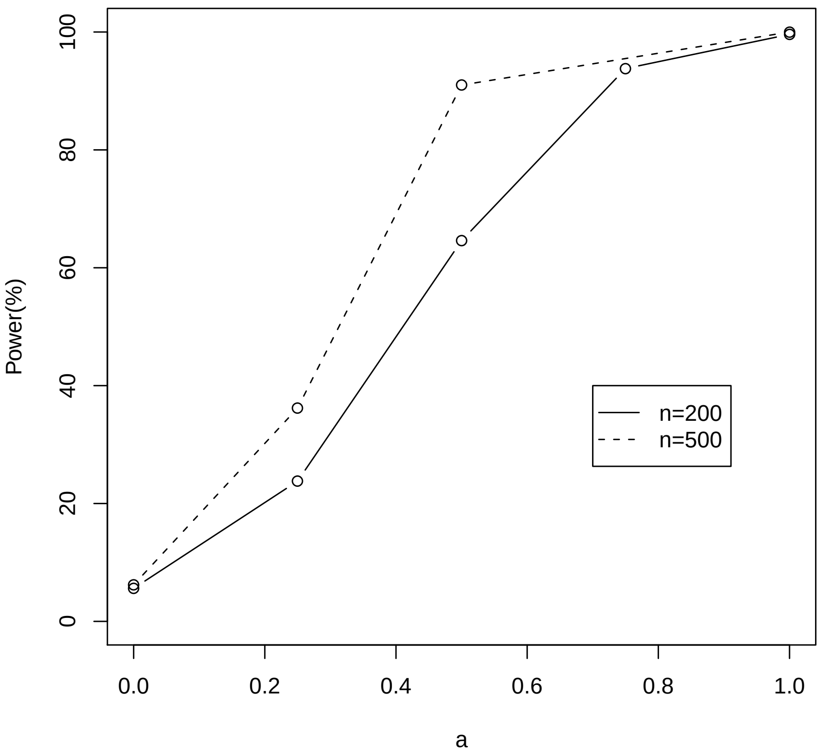

Figure 4.7: Power function at 5% significance level for $Model$ 1 (binary). Sample sizes n=200 (solid) and n=500 (dashed).

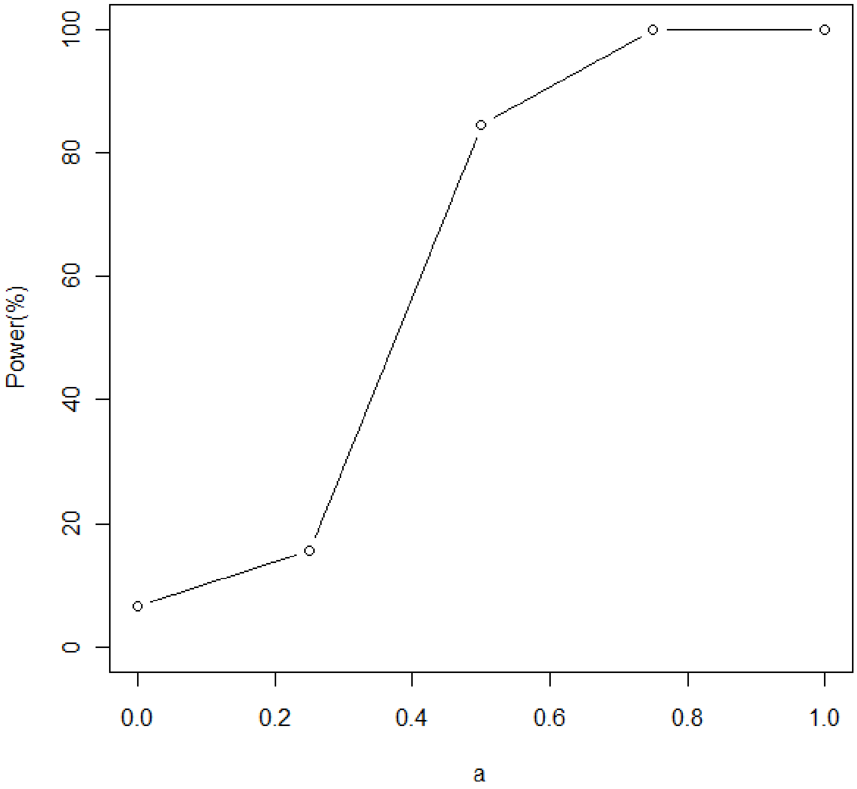

Figure 4.8: Power function at 5% significance level for *Model* 2 (Poisson). Sample sizes n=200.

Chapter 5

Conclusions and Future Research

5.1 Conclusions

In this dissertation, we have proposed a non-iterative marginal integration approach for estimaing generalized additive models with second order interaction terms. We derived the asymptotic normality for the estimators of individual univariate and bivariate component functions and also of the entire regression function. A test procedure to check for significance of the interactions was also introduced and its asymptotics were investigated. This test procedure was shown to be able to detect an interaction term of the order greater than $n^{-5/6}$ with limiting probability 1.

The finite sample performance of the estimation procedure was investigated through Monte Carlo simulations. We examined a model which generates binary responses with logit link function. We coded our own program to apply local polynomial regression with fixed bandwidth and compared the results with those by using $loess()$ available in R. Both provided reasonable and comparable results. For the model, design and sample size used in our study, the estimation could be erratic and irregular unless the bandwidth or span was selected properly. The results are sensitive to the choice of bandwidth. However, choosing a reasonable bandwidth is not an easy job in an applied context. Making it work needs a lot of patience, experimentation and intensive computation. We employed wild bootstrap to investigate the performance of the proposed test statistic. The test keeps the level well and shows reasonable power with two different data generating models: binary data (logit link) and Poisson data (log link). Again, it is very crucial to select a proper bandwidth to have the test work well.

5.2 Future Work

The optimal choice of bandwidth is crucial but is still a challenging open problem in the GAM context. In the absence of an optimal procedure for choosing a bandwidth, we chose the bandwidths somewhat arbitrarily. One future interest is to develop a feasible procedure for bandwidth selection. Cross-validation may be a choice but implies a high computational cost. Some acceleration techniques may be considered to speed up the estimation procedure. Obviously a much more detailed and thorough simulation study would be of interest, in particular concerning the interplay between model complexity and choice of bandwidth. Roca-Pardiñas, Cadarso-Suárez and González-Manteiga (2005) developed a local scoring algorithm (with backfitting) to estimate the GAM with second order interaction terms. Another possible future work is to compare the finite sample performance of the backfitting and integration method in the presence of interaction and non-trivial link function. This work will be an extension of the extensive simulation study of Sperlich, Linton and Härdle (1999).

Wild bootstrap works well in our simulation examples. Another meaningful future research problem is to construct the consistency of the wild bootstrap for our test procedure. Other types of test statistics are certainly a topic of interest and we plan to explore this as future research. It would be interesting to compare the local power of different test statistics against some sequence of local alternatives. In particular, we might estimate the case where the interaction converges to zero at the exact rate $O(n^{-5/6})$. In this case, it is possible that the asymptotic power is strictly between zero and one.

Bibliography

[1] Barry, D. (1993), “Testing for additivity of a regression function,” Annals of Statistics, 21, pp. 235–254.

[2] Bellman, R. E. (1961), *Adaptive Control Processes*, Princeton University Press: Princeton, NJ.

[3] Breiman, L. and Friedman, J. H. (1985), “Estimating optimal transformations for multiple regression and correlation,” With discussion and with a reply by the authors. Journal of the American Statistical Association, 80, pp. 580–619.

[4] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1993), *CART: Classification and Regression Trees* (first edition, 1984), Wadsworth Advanced Books and Software: Belmont, CA.

[5] Buja, A., Hastie, T., Tibshirani, R. (1989), “Linear smoothers and additive models,” Annals of Statistics, 17, pp. 453–555.

[6] Chen, R., Härdle, W., Linton, O. B. and Severance-Lossin, E. (1995), "Estimation and variable selection in additive nonparametric regression models," Discussion paper, SFB 373, Humboldt-Universität zu Berlin.

[7] Cheng, M. Y., Fan, J. and Marron, J. S. (1993), "Minimax efficiency of local polynomial fit estimators at boundaries," Mimeo Series 2098, Institute of Mathematical Statistics, University of North Carolina, Chapel Hill.

[8] Chu, C.-K. and Marron, J. S. (1991), "Choosing a kernel regression estimator," With comments and a rejoinder by the authors. Statistical Science, 6, pp. 404–436.

[9] Cleveland, W. S. (1979), "Robust locally weighted regression and smoothing scatterplots," Journal of the American Statistical Association, 74, pp. 829–836.

[10] Coull, B. A., Ruppert, D. and Wand, M. P. (2001), "Simple incorporation of interactions into additive models," Biometrics, 57, pp. 539–545.

[11] Craven, P. and Wahba, G. (1978/79), "Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation," Numerical Mathematics, 31, pp. 377–403.

[12] Deaton, A. and Muellbauer, J. (1980), *Economics and Consumer Behavior*, Cambridge University Press: New York.

[13] Derbort, S., Dette, H. and Munk, A. (2002), “A test for additivity in nonparametric regression,” Annals of the Institute of Statistical Mathematics, 54, pp. 60–82.

[14] Dette, H. and C. von Lieres and Wilkau (2001), “Testing additivity by kernel based methods – what is a reasonable test?,” Bernoulli, 7, pp. 669-697.

[15] Efromovich, S. (1999), *Nonparametric Curve Estimation. Methods, Theory, and Applications*, Springer-Verlag: New York.

[16] Eilers, P. H. C. and Marx, B. D. (1996), “Flexible smoothing with B-splines and penalties,” With comments and a rejoinder by the authors. Statistical Science, 11, pp. 89–121.

[17] Eubank, R. L., Hart J. D., Simpson D. G. and Stefanski, L. A. (1995), “Testing for additivity in nonparametric regression,” Annals of Statistics, 23, pp. 1896-1920.

[18] Fan, J. (1992), “Design-adaptive nonparametric regression,” Journal of the American Statistical Association, 87, 998–1004.

[19] Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications,* Chapman and Hall: London.

[20] Fan, J., Härdle, W. and Mammen, E. (1998), "Direct estimation of low-dimensional components in additive models," Annals of Statistics, 26, pp. 943–971.

[21] Friedman, J. H. and Stuetzle, W. (1981), "Projection pursuit regression," Journal of the American Statistical Association, 76, pp. 817–823.

[22] Gasser, T. and Müller, H. (1979), "Kernel estimation of regression functions. Smoothing techniques for curve estimation," pp. 23–68, Lecture Notes in Mathematics, 757, Springer: Berlin.

[23] Green, P. J. and Silverman, B. W. (1994), *Nonparametric Regression and Generalized Linear Models. A Roughness Penalty Approach,* Chapman and Hall: London.

[24] Gozalo, P. L. and Linton, O. B. (2001), "Testing additivity in generalized nonparametric regression models with estimated parameters," Journal of Econometrics, 104, pp. 1–48.

[25] Gu, C. and Wahba, G. (1991), "Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method," SIAM Journal on Scientific and Statistical Computing, 12, pp. 383–398.

[26] Hall, P. (1984), “Central limit theorem for integrated square error of multivariate nonparametric density estimators,” Journal of Multivariate Analysis, 14, pp. 1-16.

[27] Härdle, W. (1990), *Applied Nonparametric Regression,* Cambridge University Press: Cambridge.

[28] Härdle, W., Huet, S., Mammen, E. and Sperlich, S. (2004). “Bootstrap inference in semiparametric generalized additive models,” Econometric Theory, 20, pp. 265-300.

[29] Härdle, W. and Korostelev, A. (1996), “Search for significant variables in nonparametric additive regression,” Biometrika, 83, pp. 541–549.

[30] Härdle, W. and Mammen E. (1993), “Comparing nonparametric versus parametric regression fits,” Annals of Statistics, 21, pp. 1926-1947.

[31] Hastie, T. and Loader, C. (1993), “Local regression: automatic kernel carpentry,” (with discussion). Statistical Science, 8, pp. 120–143.

[32] Hastie, T. and Tibshirani, R. (1986a), “Generalized additive models,” With discussion. Statistical Science, 1, pp. 297–318.

[33] Hastie, T. and Tibshirani, R. (1986b), "Generalized additive models, cubic splines and penalized likelihood," Technical Report, Division of Biostatistics, University of Toronto.

[34] Hastie, T. and Tibshirani, R. (1987), "Generalized additive models: Some applications," Journal of the American Statistical Association, 82, pp. 371–386.

[35] Hastie, T. and Tibshirani, R. (1990), *Generalized Additive Models,* Chapman and Hall: London.

[36] Hengartner, N. W. (1996), "Rate optimal estimation of additive regression via the integration method in the presence of many covariates," Preprint, Department of Statistics, Yale University.

[37] Hengartner, N. W. and Sperlich, S. (2005), "Rate optimal estimation with the integration method in the presence of many covariates," Journal of Multivariate Analysis, 95, pp. 246–272.

[38] Hjellvik, V., Yao, Q. and Tjøstheim, D. (1998), "Linearity testing using local polynomial approximation," Journal of Statistical Planning and Inference, 68, pp. 295-321.

[39] Linton, O. B. (1997), "Efficient estimation of additive nonparametric regression models," Biometrika, 84, pp. 469–473.

[40] Linton, O. B. (2000), "Efficient estimation of generalized additive nonparametric regression models," Econometric Theory, 16, pp. 502-523.

[41] Linton, O. B. and Härdle, W. (1996), "Estimation of additive regression models with known links," Biometrika, 83, pp. 529-540.

[42] Linton, O. B. and Nielsen, J. P. (1995), "A kernel method of estimating structured nonparametric regression based on marginal integration," Biometrika, 82, pp. 93-100.

[43] Liu, R. Y. (1988), "Bootstrap procedures under some non-i.i.d. models," Annals of Statistics, 16, pp. 1696–1708.

[44] Mammen, E., Linton, O. B. and Nielsen, J. P. (1999), "The existence and asymptotic properties of backfitting projection algorithm under weak conditions," Annals of Statistics, 27, pp. 1443-1490.

[45] Masry, E. (1996), "Multivariate local polynomial regression for time series: uniform strong consistency and rates," Journal of Time Series Analysis, 17, pp. 571–599.

[46] Müller, H. G. (1987), "Weighted local regression and kernel methods for nonparametric curve fitting," Journal of the American Statistical Association, 82, pp. 231-238.

[47] Nadaraya, E. A. (1964), "On estimating regression," Theory of Probability and Its Applications, 9, pp. 141-142.

[48] Nielsen, J. P. and Linton, O. B. (1998), "An optimization interpretation of integration and backfitting estimators for separable nonparametric models," Journal of the Royal Statistical Society: Series B, 60, pp. 217–222.

[49] Opsomer, J. D. (2000), "Asymptotic properties of backfitting estimators," Journal of Multivariate Analysis, 73, pp. 166-179.

[50] Opsomer, J. D. and Ruppert, D. (1997), "Fitting a bivariate additive model by local polynomial regression," Annals of Statistics, 25, pp. 186-211.

[51] Opsomer, J. D. and Ruppert, D. (1998), "A fully automated bandwidth selection method for fitting additive models," Journal of the American Statistical Association, 93, pp. 605–619.

[52] Priestley, M. B. and Chao, M. T. (1972), "Non-parametric function fitting," Journal of the Royal Statistical Society: Series B, 34 , pp. 385–392.

[53] Roca-Pardiñas, J., Cadarso-Suárez, C. and González-Manteiga, W. (2005), "Testing for interactions in generalized additive models: applications to SO2 pollution data," Statistics and Computing, 15, pp. 289-299.

[54] Ruppert, D. and Wand, M. P. (1994), “Multivariate locally weighted least squares regression,” Annals of Statistics, 22, pp. 1346–1370.

[55] Schimek, M. G. (2000), *Smoothing and Regression. Approaches, Computation, and Application,* John Wiley and Sons: New York.

[56] Scott, D. W. (1992), *Multivariate Density Estimation. Theory, Practice, and Visualization,* John Wiley and Sons: New York.

[57] Severance-Lossin, E. and Sperlich, S. (1999), “Estimation of derivatives for additive separable models,” Statistics, 33, pp. 241-265.

[58] Silverman, B. W. (1984), “Spline smoothing: the equivalent variable kernel method,” Annals of Statistics, 12, , pp. 898–916.

[59] Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis,* Chapman and Hall: London.

[60] Simonoff, J. S. (1996), *Smoothing Methods in Statistics,* Springer-Verlag: New York.

[61] Sperlich, S., Linton, O. B. and Härdle, W. (1999), “Integration and backfitting methods in additive models: finite sample properties and comparison”, Test, 8, pp. 419-458.

[62] Sperlich, S., Tjøstheim, D. and Yang, L. (2002), "Nonparametric estimation and testing of interaction in additive models," Econometric Theory, 18, pp. 197–251.

[63] Stone, C. J. (1977), "Consistent nonparametric regression (with discussion)," Annals of Statistics, 5, pp. 595–620.

[64] Stone, C. J. (1985), "Additive regression and other nonparametric models," Annals of Statistics, 13, pp. 689-705.

[65] Stone, C. J. (1986), "The dimensionality reduction principle for generalized additive models," Annals of Statistics, 14, pp. 592–606.

[66] Tjøstheim, D. and Auestad, B. H. (1994), "Nonparametric identification of nonlinear time series: Projections," Journal of the American Statistical Association, 89, pp. 1398-1409.

[67] Wahba, G. (1977), "A survey of some smoothing problems and the method of generalized cross-validation for solving them," Applications of Statistics, pp. 507–523. North-Holland: Amsterdam.

[68] Wahba, G. (1990), *Spline Models for Observational Data,* Society for Industrial and Applied Mathematics (SIAM): Philadelphia, PA.

[69] Wahba, G. and Wang, Y. H. (1990), “When is the optimal regularization parameter insensitive to the choice of the loss function?” Communications in Statistics - Theory and Methods, 19, pp. 1685–1700.

[70] Wand, M. P. and Jones, M. C. (1995), *Kernel Smoothing,* Chapman and Hall: London.

[71] Watson, G. S. (1964), ”Smooth regression analysis,” Sankhya Series A 26, pp. 359–372.

[72] Wu, C. F. J. (1986), “Jackknife, bootstrap and other resampling methods in regression analysis,” With discussion and a rejoinder by the author. Annnal of Statistics, 14, pp. 1261–1350.

[73] Yang, L., Sperlich, S. and Härdle, W. (2003), “Derivative estimation and testing in generalized additive models,” Journal of Statistical Planning and Inference, 115, pp. 521–542.