

# Nonparametric Inference for Additive Models \*

Jianqing Fan

Jiancheng Jiang

## Abstract

Additive models with backfitting algorithms are popular multivariate nonparametric fitting techniques. However, the inferences of the models have not been much developed due partially to the complexity of the backfitting estimators. There are few tools available to answer some important and frequently-asked questions, such as whether a specific additive component is significant or admits a certain parametric form. In an attempt to address these issues, we extend the generalized likelihood ratio tests to additive models, using the backfitting estimator. We demonstrate that under the null models the newly proposed generalized likelihood ratio statistics follow asymptotically rescaled  $\chi^2$ -distributions, with the scaling constants and the degrees of freedom independent of the nuisance parameters. This demonstrates that the Wilks phenomenon continues to hold under a variety of smoothing techniques and more relaxed models with unspecified error distributions. We further prove that the generalized likelihood ratio tests are asymptotically optimal in terms of rates of convergence for nonparametric hypothesis testing. In addition, for testing a parametric additive model, we propose a bias corrected method to improve the performance of the generalized likelihood test. The bias corrected test is shown to share the Wilks type of property. Simulations are conducted to demonstrate the Wilks phenomenon and the power of the proposed tests. A real example is used to illustrate the performance of the testing approach.

*KEY WORDS:* Additive models, backfitting algorithm, generalized likelihood ratio, local polynomial regression, Wilks phenomenon.

*AMS1991 subject classifications.* Primary 62G07; secondary 62G10, 62J12.

---

\*Jianqing Fan is Professor, Department of Operation Research and Financial Engineering, Princeton University, Princeton, NJ 08544 and Professor of Statistics, Chinese University of Hong Kong (email: jqfan@princeton.edu); Jiancheng Jiang is Associate Professor, Department of Probability & Statistics, Peking University, Beijing 100871 (email: jiang@math.pku.edu.cn). Fan was partially supported by RGC grant CUHK 4262/01P of the HKSAR, NSF DMS-0355179, DMS-0354223 and NIH R01 HL69720 and Jiang was supported by Chinese NSF Grants 10001004 and 39930160.

# 1 INTRODUCTION

Additive models are an important family of structured multivariate nonparametric models. They model a random sample  $\{(Y_i, \mathbf{X}_i)\}_{i=1}^n$  by

$$Y_i = \alpha + \sum_{d=1}^D m_d(X_{di}) + \varepsilon_i, \quad i = 1, \dots, n, \quad (1.1)$$

where  $\{\varepsilon_i\}$  is a sequence of independent and identically distributed random variables with mean zero and finite variance  $\sigma^2$ . The additive models, which were suggested by Friedman and Stuetzle (1981) and Hastie and Tibshirani (1990), have been widely used in multivariate nonparametric modeling. As all of the unknown functions are one-dimensional, the difficulty that is associated with the so-called ‘‘curse of dimensionality’’ is substantially reduced. For details, see Stone (1985) and Hastie and Tibshirani (1990). In fact, Fan, Härdle and Mammen (1998) have shown that an additive component can be estimated as well as the case where the rest components are known. Similar oracle properties were obtained by Linton (1997) and Mammen, Linton and Nielsen (1999). Several methods for estimating the additive functions have been proposed: for example, the marginal integration estimation methods of Tjøstheim and Auestad (1994) and Linton and Nielsen (1995), the backfitting algorithms of Buja, Hastie and Tibshirani (1989) and Opsomer and Ruppert (1998), the estimating equation methods of Mammen, Linton and Nielsen (1999), the Fourier series approximation approach of Amato *et al.* (2002), the linear wavelet strategies of Amato and Antoniadis (2001), and the nonlinear wavelet estimation method of Sardy and Tseng (2003) using the block coordinate relaxation algorithm of Sardy *et al.* (2000), among others. Among these methods, the backfitting algorithm is regarded as a useful fitting tool and has received much attention for its easy implementation. Härdle and Hall (1993) and Ansley and Kohn (1994) explored the convergence of the algorithm based on projection smoothers. Opsomer and Ruppert (1997) studied asymptotic properties of the backfitting estimators for a bivariate additive model that was based on a non-projection smoother, local polynomial regression, and Wand (1999) and Opsomer (2000) extended the results to general D-dimensional additive models. Recently, Hastie and Tibshirani (2000) considered the Bayesian backfitting which is a stochastic generalization of the backfitting algorithm discussed above. A simulation study comparing the finite sample properties of backfitting and marginal integration methods was conducted in Sperlich, Linton and Härdle (1999).

After fitting the additive model via a backfitting algorithm, one often asks whether a specific additive component in (1.1) is significant or admits a certain parametric form such as a polynomial function. This amounts to testing whether the additive component is zero or of a polynomial form. However, for such kinds of frequently-asked questions, there are limited tools available. Compared with the studies on estimation, the understanding of such testing problems is limited in the additive model. To our knowledge, there is virtually no formal and theoretical work on the testing under the present settings in the literature. Recently, Härdle, Sperlich and Spokoiny (2001) used wavelets

along with the adaptive Neyman (Fan, 1996) type of idea to test additive components. While the procedure is useful, it is tailored to their specific problem and is not easy to comprehend. In contrast, we develop an easily understandable and generally applicable approach to testing problems. The idea is based on comparisons of likelihood functions under null and alternative hypotheses. If the likelihood function for the best model fit under the alternative hypothesis is much larger than that under the null hypothesis, then the null hypothesis looks implausible and should be rejected. How do we determine the critical value? Does the null distribution of the likelihood ratio test depend on nuisance parameters? These questions are poorly understood, particularly for additive models. This motivates us to unveil a new phenomenon for additive models.

Fan, Zhang and Zhang (2001) proposed “generalized likelihood ratio (*GLR*)” tests and showed that the Wilks’ type of results hold for a variety of useful models, including univariate nonparametric regression models and varying-coefficient models and their extensions. The procedure was motivated by the fact that the nonparametric maximum likelihood estimate (*MLE*) usually does not exist and even when it exists, the resulting maximum likelihood ratio test is not optimal. The idea is to replace the *MLE* with a nonparametric estimate, which results in a more relaxed family of tests, called *GLR* tests. They have shown that the resulting tests are optimal. Like the wide applicability of likelihood ratio tests for parametric models, the *GLR* tests should be useful in our setting. However, in general, since the distribution of  $\varepsilon_i$  is unknown, the likelihood function is unavailable. Two important questions that relate to the *GLR* tests arise naturally: first, it is unclear how to construct a *GLR* statistic for a variety of unknown error distributions of  $\varepsilon_i$ ; second, it remains unknown whether a particularly constructed *GLR* test will follow the Wilks’ type of results and share certain optimality. In this paper, we develop generalized likelihood ratio tests and their bias corrected versions for the additive model to address the above questions. This will not only provide useful tools to address frequently-asked questions in additive modeling, but also enrich the *GLR* test theory. Our results, together with those in Fan, Zhang and Zhang (2001) show convincingly the generality of the Wilks phenomenon, and the wide applicability of the *GLR* tests. This will encourage other researchers to apply *GLR* tests to related problems.

The technical derivations of *GLR* tests for the additive model (1.1) based on local polynomial fitting and a backfitting algorithm are very involved due to the lack of simple expressions for the backfitting estimators. Furthermore, the *GLR* statistics involve nonparametric estimators in complicated nonlinear forms. Even though they will be approximated by generalized quadratic forms, technical challenges include deriving quadratic approximations and the distributions of the quadratic functionals with a backfitting estimator. As the additive model and local polynomial smoother are widely used in multivariate nonparametric modeling, determined efforts have been made in this paper to examine the null distribution and powers of the *GLR* tests for the additive model. Such efforts enable us to answer some important questions such as whether the Wilks’ type of results hold for additive models, and whether the intuitively appealing *GLR* tests are powerful

enough.

We will prove that, under general assumptions on the error distribution of  $\varepsilon_i$ , the proposed *GLR* tests follow the Wilks' type of results and possess the asymptotic optimality for nonparametric hypothesis testing. In addition, unlike the classical Wilks' type of results and their generalization by Fan, Zhang and Zhang (2001), the additivity of degrees of freedom does not hold. The additivity property holds in a more generalized sense (see Theorem 3.2). Furthermore, testing a hypothesis on one additive component has the same asymptotic null distribution as the case where the rest of the components are known (Remark 3.1). These types of adaptive results are in line with the oracle property that was given by Fan *et al.* (1998) and Mammen *et al.* (1999). Our theoretical results from the proposed *GLR* tests shed some lights on the validation of the Wilks phenomenon and even future research directions on nonparametric inferences.

This article proceeds as follows. In Section 2, we describe the backfitting estimators based on a local polynomial smoother. Section 3 develops the theoretical framework for the *GLR* tests. The bias corrected *GLR* tests and a conditional bootstrap method for approximating the null distributions of the *GLR* statistics are introduced in Section 4. Section 5 demonstrates the performance of *GLR* tests on simulated data. Section 6 provides an example of testing on a real dataset. Technical proofs are given in Appendix II.

## 2 BACKFITTING ESTIMATORS

To ensure identifiability of the additive component functions  $m_d(x_d)$ , we impose the constraint  $E[m_d(X_{di})] = 0$  for all  $d$ . Fitting the additive component  $m_d(x_d)$  in (1.1) requires choosing bandwidths  $\{h_d\}$ . The optimal choice of  $h_d$  can be obtained as in Opsomer and Ruppert (1998) and Opsomer (2000). We here follow notation that was introduced by Opsomer (2000). Put  $K_{h_d}(x) = h_d^{-1}K(\frac{x}{h_d})$ ,  $K_s(v) = v^{s-1}K(v)$ ,  $\mathbf{H}_d = \text{diag}(1, h_d, \dots, h_d^{p_d})$ ,  $\mathbf{m}_d = \{m_d(X_{d1}), \dots, m_d(X_{dn})\}^T$ , and  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ . The smoothing matrices for local polynomial regression are

$$\mathbf{S}_d = (\mathbf{s}_{d,X_{d1}}, \dots, \mathbf{s}_{d,X_{dn}})^T,$$

where  $\mathbf{s}_{d,x_d}^T$  represents the equivalent kernel (Fan and Gijbels, 1996) for the  $d$ th covariate at the point  $x_d$ :

$$\mathbf{s}_{d,x_d}^T = \mathbf{e}_1^T (\mathbf{X}_{x_d}^d \mathbf{K}_{x_d} \mathbf{X}_{x_d}^d)^{-1} \mathbf{X}_{x_d}^d \mathbf{K}_{x_d}, \quad (2.1)$$

with  $\mathbf{e}_i$  as a vector with a one in the  $i$ th position and zeros elsewhere, the matrix  $\mathbf{K}_{x_d} = \text{diag}\{K_{h_d}(X_{d1} - x_d), \dots, K_{h_d}(X_{dn} - x_d)\}$  for a kernel function  $K(x)$  and bandwidths  $h_d$ ,

$$\mathbf{X}_{x_d}^d = \begin{bmatrix} 1 & (X_{d1} - x_d) & \cdots & (X_{d1} - x_d)^{p_d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & (X_{dn} - x_d) & \cdots & (X_{dn} - x_d)^{p_d} \end{bmatrix},$$

and  $p_d$  is the degree of the local polynomial for fitting  $m_d(x)$ . The intercept  $\alpha = E(Y_i)$  is typically estimated by  $\hat{\alpha} = \sum_{i=1}^n Y_i/n$ . The  $\mathbf{m}_d$ 's can be estimated through the solutions to the set of following normal equations (see Buja *et al.* 1989; Opsomer and Ruppert 1998):

$$\begin{bmatrix} \mathbf{I}_n & \mathbf{S}_1^* & \cdots & \mathbf{S}_1^* \\ \mathbf{S}_2^* & \mathbf{I}_n & \cdots & \mathbf{S}_2^* \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_D^* & \mathbf{S}_D^* & \cdots & \mathbf{I}_n \end{bmatrix} \begin{bmatrix} \mathbf{m}_1 \\ \mathbf{m}_2 \\ \vdots \\ \mathbf{m}_D \end{bmatrix} = \begin{bmatrix} \mathbf{S}_1^* \\ \mathbf{S}_2^* \\ \vdots \\ \mathbf{S}_D^* \end{bmatrix} \mathbf{Y},$$

where  $\mathbf{S}_d^* = (\mathbf{I}_n - \mathbf{1}\mathbf{1}^T/n)\mathbf{S}_d$  is the centered smoother matrix. In practice, the backfitting algorithm (Buja *et al.* 1989) is usually used to solve these equations and the backfitting estimators converge to the solution

$$\begin{bmatrix} \widehat{\mathbf{m}}_1 \\ \widehat{\mathbf{m}}_2 \\ \vdots \\ \widehat{\mathbf{m}}_D \end{bmatrix} = \begin{bmatrix} \mathbf{I}_n & \mathbf{S}_1^* & \cdots & \mathbf{S}_1^* \\ \mathbf{S}_2^* & \mathbf{I}_n & \cdots & \mathbf{S}_2^* \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_D^* & \mathbf{S}_D^* & \cdots & \mathbf{I}_n \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}_1^* \\ \mathbf{S}_2^* \\ \vdots \\ \mathbf{S}_D^* \end{bmatrix} \mathbf{Y} \equiv \mathbf{M}^{-1}\mathbf{C}\mathbf{Y}, \quad (2.2)$$

provided that the inverse of  $\mathbf{M}$  exists.

As in Opsomer (2000), we define the additive smoother matrix as

$$\mathbf{W}_d = \mathbf{E}_d \mathbf{M}^{-1} \mathbf{C}, \quad (2.3)$$

where  $\mathbf{E}_d$  is a partitioned matrix of dimension  $n \times nD$  with an  $n \times n$  identity matrix as the  $d$ th ‘‘block’’ and zeros elsewhere, so that the backfitting estimator for  $\mathbf{m}_d$  is  $\widehat{\mathbf{m}}_d = \mathbf{W}_d \mathbf{Y}$ . Let  $\mathbf{W}_M^{[-d]}$  be the additive smoother matrix for the data generated by the  $(D-1)$ -variate regression model:  $Y_i' = \sum_{k=1, k \neq d}^D m_k(X_{ki}) + \varepsilon_i$ . Denote by  $\mathbf{m} = \sum_{d=1}^D \mathbf{m}_d$  and  $\mathbf{W}_M = \sum_{d=1}^D \mathbf{W}_d$ . The backfitting estimator of  $\mathbf{m}$  is then  $\widehat{\mathbf{m}} = \mathbf{W}_M \mathbf{Y}$ .

If  $\|\mathbf{S}_d^* \mathbf{W}_M^{[-d]}\| < 1$  for some  $d \in (1, \dots, D)$  and a matrix norm  $\|\cdot\|$ , by Lemma 2.1 of Opsomer (2000), the backfitting estimators exist and are unique, and

$$\begin{aligned} \mathbf{W}_d &= \mathbf{I}_n - (\mathbf{I}_n - \mathbf{S}_d^* \mathbf{W}_M^{[-d]})^{-1} (\mathbf{I}_n - \mathbf{S}_d^*) \\ &= (\mathbf{I}_n - \mathbf{S}_d^* \mathbf{W}_M^{[-d]})^{-1} \mathbf{S}_d^* (\mathbf{I}_n - \mathbf{W}_M^{[-d]}). \end{aligned} \quad (2.4)$$

For a finite  $n$  in practice, the existence and uniqueness condition above can numerically be verified. To ensure the existence of the backfitting estimators when  $n$  is sufficiently large, we here consider only the design points, denoted by  $\mathcal{X}$ , such that

$$\limsup_n \|\mathbf{S}_d^* \mathbf{W}_M^{[-d]}\| < 1 \quad (2.5)$$

for a matrix norm  $\|\cdot\|$ . In practice, the smoothing operators  $\mathbf{S}_1, \dots, \mathbf{S}_d$  are conducted over compact sets of design densities. Hence, we need only to deal the case where the design densities have bounded support. In the case of  $D = 2$ , a sufficient condition for (2.5) is

$$\sup_{x_1, x_2} \left| \frac{f_{12}(x_1, x_2)}{f_1(x_1)f_2(x_2)} - 1 \right| < 1,$$

where  $f_d(x_d)$  is the density of  $X_d$  and  $f_{12}(x_1, x_2)$  is the joint density of  $X_1$  and  $X_2$ . This is exactly the restriction (4) in Opsomer and Ruppert (1997). Then by Lemma 2 in the appendix II and direct matrix multiplication

$$\limsup_n \|\mathbf{S}_1^* \mathbf{S}_2^*\|_r < 1,$$

where  $\|\mathbf{A}\|_r = \max_{1 \leq i \leq n} \sum_{j=1}^n |\mathbf{a}_{ij}|$  denotes the norm of the maximum row sum. However for  $D > 2$ , the condition in (2.5) is not easy to be replaced with other conditions. In fact, for backfitting algorithm using any smoothing technique, the condition (2.5) must be satisfied to assure the existence of the backfitting estimators. Hence we restrict the design points in  $\mathcal{X}$ .

### 3 GENERALIZED LIKELIHOOD RATIO TESTS

#### 3.1 The *GLR* Test

In this section we define the *GLR* statistics and develop their asymptotic theory under model (1.1), which is based on the local polynomial smoother and the backfitting algorithm. The Wilks phenomenon and optimality are unveiled in this general setting.

For simplicity, we first consider the hypothesis testing problem:

$$H_0 : m_D(x_D) = 0 \longleftrightarrow H_1 : m_D(x_D) \neq 0. \quad (3.1)$$

This tests whether the  $D$ th variable has any significant contribution to the dependent variable. The testing problem is a nonparametric null hypothesis versus a nonparametric alternative, as the nuisance parameters under  $H_0$  are still nonparametric. Testing significance of more than one variable can be dealt with analogously.

As the distribution of  $\varepsilon_i$  is unknown, we do not have a known likelihood function. Pretending that error distribution is normal,  $\mathcal{N}(0, \sigma^2)$ , the log-likelihood under model (1.1) is

$$-\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1}^n (Y_k - \alpha - \sum_{d=1}^D m_d(X_{dk}))^2.$$

Replacing the intercept  $\alpha$  and the unknown function  $m_d(\cdot)$  by  $\hat{\alpha}$  and  $\hat{m}_d(\cdot)$  respectively leads to

$$-\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \text{RSS}_1,$$

where  $\text{RSS}_1 = \sum_{k=1}^n (Y_k - \hat{\alpha} - \sum_{d=1}^D \hat{m}_d(X_{dk}))^2$ . Maximizing over the parameter  $\sigma^2$ , we obtain a likelihood of the alternative model:

$$-\frac{n}{2} \log(2\pi/n) - \frac{n}{2} \log(\text{RSS}_1) - \frac{n}{2}.$$

Therefore, up to a constant term, the log-likelihood of model (1.1) is taken as  $\ell(H_1) = -\frac{n}{2} \log(\text{RSS}_1)$ . Similarly, the log-likelihood for  $H_0$  can be taken as  $\ell(H_0) = -\frac{n}{2} \log(\text{RSS}_0)$ , with  $\text{RSS}_0 = \sum_{k=1}^n (Y_k - \hat{\alpha} - \sum_{d=1}^{D-1} \tilde{m}_d(X_{dk}))^2$ , and  $\tilde{m}_d(x_d)$  the estimator of  $m_d(x_d)$  under  $H_0$ , using the same backfitting

algorithm and bandwidths. Following Fan, Zhang and Zhang (2001), we define the following *GLR* statistic:

$$\lambda_n(H_0) = [\ell(H_1) - \ell(H_0)] = \frac{n}{2} \log \frac{\text{RSS}_0}{\text{RSS}_1} \approx \frac{n}{2} \frac{\text{RSS}_0 - \text{RSS}_1}{\text{RSS}_1}, \quad (3.2)$$

which compares the likelihood of the nearly best fitting in the alternative models with that under the null models. The null hypothesis is rejected when  $\lambda_n(H_0)$  is too large.

### 3.2 Asymptotic Null Distribution

Let  $\mu_i = \int u^i K(u) du$ ,  $v_i = \int u^i K^2(u) du$ , for  $i = 0, 1, \dots$ , and  $\tilde{\mathbf{S}}_d = (\mu_{i+j-2})$  (for  $i, j = 1, \dots, p_d + 1$ ) be a  $(p_d + 1) \times (p_d + 1)$  matrix. Denote the convolution of  $K_s(x)$  with  $K_t(x)$  by  $K_s * K_t$ , where  $K_s(x) = x^{s-1} K(x)$  for  $s, t = 1, 2, \dots$ . Put  $\mathbf{c}_{j,p_d+j} = (\mu_j, \dots, \mu_{p_d+j})^T$ ,  $(\tilde{s}_{d,1}, \dots, \tilde{s}_{d,p_d+1}) = \mathbf{e}_1^T \tilde{\mathbf{S}}_d^{-1}$ , and  $C_d^{(j)} = \mathbf{e}_1^T \tilde{\mathbf{S}}_d^{-1} \mathbf{c}_{j,p_d+j}$  for  $j = 0, \dots, p_d + 1$  and  $d = 1, \dots, D$ . Let

$$\begin{aligned} \mu_n &= \frac{|\Omega_D|}{h_D} \left[ \sum_{t=1}^{p_D+1} \tilde{s}_{D,t} K_t(0) - \frac{1}{2} \sum_{s,t=1}^{p_D+1} \tilde{s}_{D,s} \tilde{s}_{D,t} K_s * K_t(0) \right], \\ \sigma_n^2 &= \frac{2|\Omega_D|}{h_D} \left\| \sum_{t=1}^{p_D+1} \tilde{s}_{D,t} K_t - \frac{1}{2} \sum_{s,t=1}^{p_D+1} \tilde{s}_{D,s} \tilde{s}_{D,t} K_s * K_t \right\|_2^2, \\ r_K &\equiv \frac{2\mu_n}{\sigma_n^2} = \frac{\sum_{t=1}^{p_D+1} \tilde{s}_{D,t} K_t(0) - \frac{1}{2} \sum_{s,t=1}^{p_D+1} \tilde{s}_{D,s} \tilde{s}_{D,t} K_s * K_t(0)}{\left\| \sum_{t=1}^{p_D+1} \tilde{s}_{D,t} K_t - \frac{1}{2} \sum_{s,t=1}^{p_D+1} \tilde{s}_{D,s} \tilde{s}_{D,t} K_s * K_t \right\|_2^2}, \end{aligned}$$

where  $|\Omega_d|$  is the length of the support of the density  $f_d(x_d)$  of  $X_d$ . The following theorem describes our generalized Wilks type of results conditional on  $\mathcal{X}$ .

**Theorem 3.1** *Suppose Condition (A) in Appendix I holds. Then, under  $H_0$  for the testing problem (3.1),*

$$P\{\sigma_n^{-1}(\lambda_n(H_0) - \mu_n - d_{1n}) < t \mid \mathcal{X}\} \xrightarrow{\mathcal{L}} \Phi(t),$$

where  $d_{1n} = O_p(1 + \sum_{d=1}^D n h_d^{2(p_d+1)} + \sum_{d=1}^D \sqrt{n} h_d^{p_d+1})$  and  $\Phi(\cdot)$  is the standard normal distribution. Furthermore, if  $n h_d^{2(p_d+1)} h_D \rightarrow 0$  for  $d = 1, \dots, D$ , then conditional on  $\mathcal{X}$

$$r_K \lambda_n(H_0) \stackrel{a}{\sim} \chi_{r_K \mu_n}^2.$$

In Theorem 3.1, asymptotic normality is given with  $d_{1n}$  unspecified. An asymptotic expression for this item is very complicated and unnecessary in our opinion. The theorem gives the asymptotic null distribution, but the  $d_{1n}$  can be negligible under the condition  $n h_d^{2(p_d+1)} h_D \rightarrow 0$  for  $d = 1, \dots, D$ . The above condition holds if  $n h_D^{2p_D+3} \rightarrow 0$  and  $h_d^{p_d+1} = O(h_D^{p_D+1})$ .

**Remark 3.1** When  $K(\cdot)$  is a symmetric density kernel and  $p_d = 1$  for  $d = 1, \dots, D$ , direct computation yields that  $\mu_n = \frac{|\Omega_D|}{h_D} [K(0) - \frac{1}{2} K * K(0)]$ ,  $\sigma_n^2 = \frac{2|\Omega_D|}{h_D} \|K - \frac{1}{2} K * K\|_2^2$ , and  $r_K = \frac{K(0) - \frac{1}{2} K * K(0)}{\|K - \frac{1}{2} K * K\|_2^2}$ . This coincides with the result in the one dimensional nonparametric regression of Fan, Zhang and

Zhang (2001). Therefore, for the additive model, the *GLR* test has an oracle property in the sense that though the nuisance functions  $m_d(x_d)$ 's (for  $d = 1, \dots, D-1$ ) are unknown, the *GLR* test behaves as though they were known.

From Theorem 3.1, under certain conditions the asymptotic null distribution of the *GLR* statistic is independent of the intercept and the nuisance functions  $m_d(\cdot)$  ( $d = 1, \dots, D-1$ ), the nuisance design densities  $f_d(\cdot)$  (for  $d = 1, \dots, D-1$ ), and the nuisance error distributions over a large range of bandwidths. We refer to such a result as the Wilks phenomenon.

The asymptotic null distribution offers a method for determining approximately the critical value of the *GLR* tests, but one can't expect this kind of approximation to be highly accurate unless the bandwidth  $h_D$  is sufficiently small so that the degree of freedom  $r_K \mu_n$  is large. However, the Wilks type of result allows us to simulate the null distributions of the *GLR* tests over a large range of bandwidths with nuisance functions fixed at their estimated values. This justifies the conditional bootstrap method in Section 4.2. An alternative approximation of the null distribution can be obtained by using a calibration idea of Zhang (2003). When  $h_D \rightarrow \infty$ , the local polynomial fitting becomes a global polynomial fitting. Hence, one would expect the degree of freedom to be  $p_D$ . This prompts Zhang to use  $\chi_{r_K \mu_n + p_D}^2$  to approximate the null distribution.

Now we consider a little more complicated hypothesis testing problem:

$$H_0 : m_{D-d_0}(x_{D-d_0}) = \dots = m_D(x_D) = 0 \longleftrightarrow H_1 : m_{D-d_0}(x_{D-d_0}) \neq 0, \dots, \text{ or } m_D(x_D) \neq 0, \quad (3.3)$$

for some integer  $d_0$ . This generalizes problem (3.1). Let

$$\begin{aligned} \mu'_n &= \sum_{d'=D-d_0}^D \frac{|\Omega_{d'}|}{h_{d'}} \left[ \sum_{t=1}^{p_{d'}+1} \tilde{s}_{d',t} K_t(0) - \frac{1}{2} \sum_{s,t=1}^{p_{d'}+1} \tilde{s}_{d',s} \tilde{s}_{d',t} K_s * K_t(0) \right], \\ \sigma_n'^2 &= \sum_{d'=D-d_0}^D \frac{2|\Omega_{d'}|}{h_{d'}} \left\| \sum_{t=1}^{p_{d'}+1} \tilde{s}_{d',t} K_t - \frac{1}{2} \sum_{s,t=1}^{p_{d'}+1} \tilde{s}_{d',s} \tilde{s}_{d',t} K_s * K_t \right\|_2^2, \end{aligned}$$

and  $r'_K = 2\mu'_n / \sigma_n'^2$ .

**Theorem 3.2** *For the hypothesis testing problem (3.3), under the same conditions as in Theorem 3.1, the results in Theorem 3.1 continue to hold but with  $\mu_n$ ,  $\sigma_n^2$  and  $r_K$  replaced by  $\mu'_n$ ,  $\sigma_n'^2$  and  $r'_K$ , where the condition  $nh_d^{2(p_d+1)} h_D \rightarrow 0$  for all  $d$ 's is replaced by for all  $d$ 's and any  $d' \in \{D-d_0, \dots, D\}$ ,  $nh_{d'}^{2(p_{d'}+1)} h_{d'} \rightarrow 0$ .*

Interestingly,  $\mu'_n$  and  $\sigma_n'^2$  are the summation of the individual  $\mu_n$  and  $\sigma_n^2$  given in Theorem 3.1. However, the normalization constant  $r'_K$  changes with the testing problem and the degree of freedom  $r'_K \mu'_n$  are no longer the summation of those for testing individual problem such as (3.1). These mark the difference from those given by Fan, Zhang and Zhang (2001). The result is also



different from the case of the degrees of freedom of the fit for an additive penalized spline model (see Section 8.3 of Ruppert *et al.*, 2003). However, when all  $p_d$  are equal, the additivity of degrees of freedom holds.

The *GLR* tests are also applicable to testing the problems with parametric models as the null hypothesis. Consider the following testing problem with parametric null hypothesis:

$$H_0 : m_\theta(x_1, \dots, x_D) \in \mathcal{M}_\Theta \longleftrightarrow H_1 : m_\theta(x_1, \dots, x_D) \notin \mathcal{M}_\Theta, \quad (3.4)$$

where  $\mathcal{M}_\Theta = \{m_\theta(x_1, \dots, x_D) = \sum_{d=1}^D m_d(x_d; \theta) : \theta \in \Theta\}$  is a set of functions of parametric forms, and the parameter space  $\Theta$  contains the true parameter value  $\theta_0$ . As before, we can use the local polynomial fitting technique and backfitting algorithm to fit the alternative model and obtain the log-likelihood  $\ell_n(H_1)$  for  $H_1$ . By maximizing the likelihood for the fully parametric model under  $H_0$ , we build the log-likelihood  $\ell_n(H_0)$ . Denote by  $\lambda_n(\mathcal{M}_\Theta)$  the *GLR* statistic for the testing problem (3.4). To derive the asymptotic null distribution of the test statistic, some conditions on  $\mathcal{M}_\Theta$  and  $\Theta$  are required to render the likelihood ratio test statistic of order  $o_p(h_D^{-\frac{1}{2}})$  for the following parametric testing problem:

$$H'_0 : m(x_1, \dots, x_D) = m_{\theta_0}(x_1, \dots, x_D) \longleftrightarrow H'_1 : m(x_1, \dots, x_D) \in \mathcal{M}_\Theta.$$

For ease of exposition, the required conditions are referred to as “Condition B”. Conditions similar to those of Cramér [see for example, conditions (C1)-(C5) on page 102 of Le Cam and Yang (1990)] are sufficient in the present setting, as the classical Wilks theorem holds and hence the likelihood ratio statistic is of order  $O_p(1)$ .

**Theorem 3.3** *Suppose that Condition (A) in Appendix I and Condition (B) holds. Then, under  $H_0$  for the testing problem (3.4),*

$$P\{\sigma_n^{*-1}(\lambda_n(\mathcal{M}_\Theta) - \mu_n^* - d_{1n}) < t \mid \mathcal{X}\} \xrightarrow{\mathcal{L}} \Phi(t),$$

where  $d_{1n} = O_p(1 + \sum_{d=1}^D nh_d^{2(p_d+1)} + \sum_{d=1}^D \sqrt{nh_d^{p_d+1}})$ . Furthermore, if  $nh_d^{2(p_d+1)}h_{d'} \rightarrow 0$  for all  $d$ 's and any  $d'$ , then conditioning on  $\mathcal{X}$

$$r_K^* \lambda_n(\mathcal{M}_\Theta) \stackrel{a}{\sim} \chi_{r_K^*}^2 \mu_n^*,$$

where  $\mu_n^*$  and  $\sigma_n^{*2}$  are the same as  $\mu'_n$  and  $\sigma_n'^2$  with  $D - d_0 = 1$  and  $r_K^* = 2\mu_n^*/\sigma_n^{*2}$ .

### 3.3 Power of *GLR* Tests

We now consider the power of *GLR* tests in the framework of Fan, Zhang and Zhang (2001). For simplicity, we focus on the null hypothesis in (3.1).

Assume that  $h_D = o(n^{-1/(2p_D+3)})$  so that the second term in the definition of  $d_{1n}$  is of smaller order than  $\sigma_n$ . As to be seen in Theorem 3.5, the optimal bandwidth for the testing problem (3.1) is

$h_D = O(n^{-2/(4p_D+5)})$ , which satisfies the condition  $h_D = o(n^{-1/(2p_D+3)})$ . Under these assumptions, Theorem 3.1 leads to an approximate level  $\alpha$  test based on the *GLR* statistic:

$$\phi_h = I\{\lambda_n(H_0) - \mu_n \geq z_\alpha \sigma_n\}.$$

If we consider the contiguous alternative of form

$$H_{1n} : m_D(X_D) = G_n(X_D),$$

where  $G_n(X_D) \rightarrow 0$  as  $n \rightarrow \infty$ , then the power of the *GLR* test can be approximated by using the following theorem.

**Theorem 3.4** *Suppose that Condition (A) in Appendix I holds and that  $nh_d^{2(p_d+1)}h_D \rightarrow 0$  for  $d = 1, \dots, D$ . If  $E\{G_n(X_D)|X_1, \dots, X_{D-1}\} = 0$  and  $h_D \cdot \sum_{i=1}^n G_n^2(X_{Di}) \xrightarrow{P} C(G)$  for some constant  $C(G)$ , then under  $H_{1n}$  for the testing problem (3.1)*

$$P\{\sigma_{1n}^{-1}(\lambda_n(H_0) - \mu_n - d_{2n}) < t | \mathcal{X}\} \xrightarrow{\mathcal{L}} \Phi(t),$$

where  $\mu_n$  is the same as that in Theorem 3.1,

$$d_{2n} = \sum_{i=1}^n G_n^2(X_{Di})(1 + o_p(1)),$$

and

$$\sigma_{1n} = \sqrt{\sigma_n^2 + \sigma^{-2} \sum_{i=1}^n G_n^2(X_{Di})}.$$

**Remark 3.2** For testing problem (3.1), the alternative hypothesis depends on many nuisance functions  $m_d$  for  $d = 1, \dots, D - 1$ . Theorem 3.4 shows that the asymptotic alternative distribution of the *GLR* testing statistic is independent of the nuisance functions  $m_d(x_d)$ , for  $d \neq D$ , over a large range of bandwidths. This allows us to compute the power of the test via simulations over a large range of bandwidths with nuisance functions fixed at their estimated values.

Let  $z_{1-\alpha}$  be the  $(1 - \alpha)$ th percentile of  $\mathcal{N}(0, 1)$ . By Theorems 3.1 and 3.4, the power of the test is approximately given by

$$P_{H_{1n}}(W) \approx 1 - \Phi(\sigma_{1n}^{-1} \sigma_n z_{1-\alpha} - \sigma_{1n}^{-1} d_{2n}).$$

To study the optimal property of the *GLR* test, we consider the class of functions  $\mathcal{G}_n$ , satisfying the following regularity conditions:

$$\begin{aligned} \text{Var}(G_n^2(X_D)) &\leq M(E[G_n^2(X_D)])^2 \\ nE[G_n^2(X_D)] &> M_n \rightarrow \infty, \end{aligned} \tag{3.5}$$

for some constants  $M > 0$  and  $M_n \rightarrow \infty$ . For a given  $\rho > 0$ , let

$$\mathcal{G}_n(\rho) = \{G_n \in \mathcal{G}_n : E[G_n^2(X_D)] \geq \rho^2\}.$$

The maximum of the probabilities of type II errors is then given by

$$\beta(\alpha, \rho) = \sup_{G_n \in \mathcal{G}_n(\rho)} \beta(\alpha, G_n),$$

where  $\beta(\alpha, G_n) = P(\phi_h = 0 | m_D = G_n)$  is the probability of type II error at the alternative  $H_{1n} : m_D = G_n$ . The minimax rate of  $\phi_h$  is defined as the smallest  $\rho_n$  such that

- (i) for every  $\rho > \rho_n$ ,  $\alpha > 0$ , and for any  $\beta > 0$ , there exists a constant  $c$  such that  $\beta(\alpha, c\rho) \leq \beta + o(1)$ ; and
- (ii) for any sequence  $\rho_n^* = o(\rho_n)$ , there exist  $\alpha > 0$ ,  $\beta > 0$  such that for any  $c > 0$ ,  $P(\phi_h = 1 | m_D = G_n) = \alpha + o(1)$  and  $\liminf_n \beta(\alpha, c\rho_n^*) > \beta$ .

This measures how close are the alternatives that can be detected by the *GLR* test  $\phi_h$ .

**Theorem 3.5** *Under Condition (A) in Appendix I, if  $h_d^{p_d+1} = O(h_D^{p_D+1})$  for  $d = 1, \dots, D-1$ , then for the testing problem (3.1) the *GLR* test can detect alternatives with rate  $\rho_n = n^{-2(p_D+1)/(4p_D+5)}$  when  $h_D = c_* n^{-2/(4p_D+5)}$  for some constant  $c_*$ .*

**Remark 3.3** The *GLR* tests are asymptotically optimal in terms of rates of convergence for non-parametric hypothesis testing according to the formulations of Ingster (1993) and Spokoiny (1996). While Ingster (1993) and Spokoiny (1996) focus only on the univariate setting, their minimax lower bound is applicable to our additive model with known functions  $(m_1, \dots, m_{D-1})$ . Its rate of convergence is the same as the rate of upper bound given in Theorem 3.5.

Since the distributional property in Theorem 3.1 depends implicitly on the assumption for the bandwidths  $h'_d$ s, in particular,  $nh_d^{2(p_d+1)}h_D = o(1)$  is required to ensure the Wilks properties. This suggests that the bandwidths suited well for curve estimation may not be the best for testing. The power of the *GLR* tests depends on the smoothing parameters. In fact, Theorem 3.5 shows that theoretical optimal bandwidth  $h_D$  is  $c_* n^{-2/(4p_D+5)}$  for some constant  $c_*$ .

## 4 Implementations of GLR tests

The *GLR* test involves the determination of the null distribution and the choice of bandwidth in practice. We now address these two issues.

## 4.1 Bias reduction

The asymptotic null distribution of the  $GLR$  statistic  $\lambda_n(H_0)$  involves a bias term  $d_{1n}$ . The bandwidth has to be small enough to make it negligible. However, in practice, the size of bandwidth that would make the bias negligible is unknown and it is desirable to reduce bias automatically. For the testing problem (3.4), we will demonstrate how this objective can be achieved. The basic idea is inspired by the prewhitening technique of Press and Tukey (1956) in spectral density estimation and the technique that was employed by Härdle and Mammen (1993) for univariate nonparametric testing. The method is also related to the nonparametric estimator that uses a parametric start of Hjort and Glad (1995) and Glad (1998). Recently, Fan and Zhang (2004) advocate the use of the bias reduction method in the study of testing problems for spectral density.

Consider the testing problem (3.4). The additive model (1.1) is equivalent to

$$Y_i^* = m^*(X_{1i}, \dots, X_{Di}) + \varepsilon_i, \quad (4.1)$$

where  $Y_i^* = Y_i - \hat{\alpha} - m(X_{1i}, \dots, X_{Di}; \hat{\theta})$  and  $m^*(X_{1i}, \dots, X_{Di}) = \alpha + m(X_{1i}, \dots, X_{Di}) - \hat{\alpha} - m(X_{1i}, \dots, X_{Di}; \hat{\theta})$  with  $\hat{\theta}$  being the least-squares estimator of  $\theta$  under the null hypothesis in (3.4). Therefore, the testing problem (3.4) is reduced to the following problem:

$$H_0^* : m^*(X_1, \dots, X_D) \in \mathcal{M}_0 \leftrightarrow H_1^* : m^*(X_1, \dots, X_D) \notin \mathcal{M}_0, \quad (4.2)$$

where  $\mathcal{M}_0 = \{\alpha^* = 0, m_1^* = \dots = m_D^* = 0\}$ . This is the specific case of (3.4) and hence the  $GLR$  test can be applied. Denote by  $\lambda_n^*(\mathcal{M}_\Theta)$  the resulting  $GLR$  statistic. As the regression function  $m^*(X_1, \dots, X_D)$  is nearly zero under  $H_0$ , there is little bias involved for the backfitting estimator. This is demonstrated by the following theorem, which allows virtually all of the bandwidths that are used in practice.

**Theorem 4.1** *Suppose that Condition (A) in Appendix I and Condition (B) hold. Then conditioning on  $\mathcal{X}$  under  $H_0$  for the testing problem (3.4)*

$$r_K^* \lambda_n^*(\mathcal{M}_\Theta) \stackrel{a}{\sim} \chi_{r_K^* \mu_n^*}^2,$$

where  $\mu_n^*$  and  $r_K^*$  are the same as those in Theorem 3.3.

## 4.2 Conditional Bootstrap

To implement the  $GLR$  tests, we need to obtain the null distributions of the test statistics. In Section 3.2, the asymptotic distributions of the  $GLR$  statistics were given, which demonstrates that the asymptotic null distributions are independent of nuisance parameters/functions. For a finite sample, this means that the null distributions do not sensitively depend on the nuisance parameters/functions. Therefore, the null distributions can be approximated by simulation methods, via fixing nuisance parameters/functions at their reasonable estimates. This simulation method

is referred to as the conditional bootstrap method, which is detailed below. To be more specific, consider (3.1).

1. Fix the bandwidths at their estimated values  $(\hat{h}_1, \dots, \hat{h}_D)$ , and then obtain the estimators of the additive components under both the null and the unrestricted additive models.
2. Compute the *GLR* test statistic  $\lambda_n(H_0)$  and the residuals  $\hat{\varepsilon}_i$  (for  $i = 1, \dots, n$ ) from the unrestricted model.
3. For each  $\mathbf{X}_i$ , draw a bootstrap residual  $\hat{\varepsilon}_i^*$  from the centered empirical distribution of  $\hat{\varepsilon}_i$  and compute  $Y_i^* = \hat{\alpha} + \hat{m}_1(X_{i1}) + \dots + \hat{m}_{D-1}(X_{i,D-1}) + \hat{\varepsilon}_i^*$ , where  $\hat{\alpha}$  and  $\hat{m}_j(\cdot)$  ( $j \leq D-1$ ) are the estimated regression functions under the unrestricted additive model in step 1. This forms a conditional bootstrap sample  $\{\mathbf{X}_i, Y_i^*\}_{i=1}^n$ .
4. Using the bootstrap sample in step 3 with the bandwidths  $(\hat{h}_1, \dots, \hat{h}_D)$ , obtain the *GLR* statistic  $\lambda_n^*(H_0)$  in the same manner as  $\lambda_n(H_0)$ .
5. Repeat steps 3 and 4 many times to obtain a sample of statistic  $\lambda_n^*(H_0)$ .
6. Use the bootstrap sample in step 5 to determine the quantiles of the test statistic under  $H_0$ . The P-value is the percent of observations from the bootstrap sample of  $\lambda_n^*(H_0)$  whose value exceeds  $\lambda_n(H_0)$ .

Note that the null distribution of  $\lambda_n(H_0)$  depends on  $(\alpha, m_1, \dots, m_{D-1})$  and distribution of  $\varepsilon$ . As shown in Theorem 3.1, such a dependence is asymptotically negligible. Hence, they can be fixed at the values  $(\hat{\alpha}, \hat{m}_1, \dots, \hat{m}_{D-1})$  and the distribution of  $\hat{\varepsilon}^*$ . The following theorem shows the consistency of the conditional bootstrap method.

**Theorem 4.2** *Assume that the conditions in Theorem 3.1 hold. Then, under  $H_0$  in (3.1),*

$$P\{\sigma_n^{-1}(\lambda_n^*(H_0) - \mu_n - d_{1n}) < t | \mathcal{X}, F_n\} \xrightarrow{\mathcal{L}} \Phi(t),$$

where  $F_n$  denotes the empirical distribution of the sample  $\{\mathbf{X}_i, Y_i\}_{i=1}^n$ .

### 4.3 Choice of Bandwidth

The test statistic  $\lambda_n(H_0)$  depends on the choice of the bandwidths  $\{h_d\}$  (for  $d = 1, \dots, D$ ). In fact, it can be regarded as a family of the test statistics indexed by  $h_d$ . The optimal bandwidths for hypothesis testing differ somewhat from those for estimating the additive components, which was elaborated in Section 3.3.

The choice of optimal bandwidths for hypothesis testing has not been seriously explored in the literature, but the optimal bandwidths for estimating the underlying additive components provide a good proxy for those in the testing problem. Opsomer (2000) gave theoretic optimal bandwidths for

a  $D$ -dimensional additive model. We will use these theoretic optimal bandwidths in our simulation study. For real data examples, the automatic bandwidth selection rule of Opsomer and Ruppert (1998) will be used. Due to the difference of the optimal bandwidths between the fitting and testing, it is a good practice for us to explore the sensitivity of the testing results by varying the bandwidths over a relatively large range. The correlation between  $\lambda_n(H_0)$  using bandwidth  $\mathbf{h}_1$  and that using bandwidth  $\mathbf{h}_2$  is expected to be large when  $\mathbf{h}_1 \approx \mathbf{h}_2$ . See the work of Zhang (2003) for the result on nonparametric regression, which corresponds to  $D = 1$ . Thus, for many applications, it suffices to use  $h = h_{opt}/1.5, h_{opt}, 1.5h_{opt}$ , corresponding to “undersmooth”, “right smooth” and “oversmooth”, where  $h_{opt}$  is the asymptotically optimal bandwidth used for estimation. This idea will be followed in our simulations and real data analysis.

## 5 SIMULATIONS

The purpose of the simulations is twofold: demonstrating the Wilks phenomenon and the power of the proposed *GLR* tests. The effect of the error distributions on the performance of the *GLR* tests is also investigated. Numerical results show that the *GLR* tests with bias correction outperform their counterparts. Throughout this section, the Epanechnikov kernel is employed.

**Example 1.** Consider the following bivariate additive model:

$$Y = m_1(X_1) + m_2(X_2) + \varepsilon, \quad (5.1)$$

where  $m_1(X_1) = 0.5 - 6X_1^2 + 3X_1^3$ ,  $m_2(X_2) = \sin(\pi X_2)$ , and the error  $\varepsilon$  is distributed as  $\mathcal{N}(0, 1)$ . The covariates are generated by the following transformation to create correlation:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}, \quad (5.2)$$

where  $U_i$ 's are independent and identically distributed as  $U(-0.5, 0.5)$ .

We employ the optimal bandwidth  $h_{d,opt}$  for the smoother on  $m_d(x_d)$  (see Opsomer 2000). To demonstrate the Wilks phenomenon for the *GLR* test, three levels of bandwidth are evaluated with  $h_2$  fixed at its optimal value:  $h_1 = \frac{2}{3}h_{1,opt}, h_{1,opt},$  or  $\frac{3}{2}h_{1,opt}$ . The null hypothesis is taken as  $H_0 : m_2(x_2) = 0$  (where  $m_1(x_1)$  is a nuisance function). We also use three levels of  $m_1(X_1)$  to demonstrate that the test does not depend on the nuisance function  $m_1(X_1)$ :

$$m_{1,\beta}(X_1) = \left[ 1 + \beta \sqrt{\text{Var}(0.5 - 6X_1^2 + 3X_1^3)} \right] (0.5 - 6X_1^2 + 3X_1^3),$$

where  $\beta = -1.5, 0, 1.5$ . For the *GLR* test, we drew 1000 samples of 200 observations. Based on the 1000 samples, we obtained 1000 *GLR* test statistics. Their distribution is obtained via a kernel estimate with a rule of thumb bandwidth:  $h = 1.06sn^{-0.2}$ , where  $s$  is the standard error of the normalized *GLR* statistics. Figure 1 shows that the estimated densities of the normalized *GLR* statistics,  $r_K \lambda_n(H_0)$ . As expected, they look like densities from  $\chi^2$ -distributions, or more

generally, gamma distributions. The top panel of Figure 1 shows that the null distributions follow  $\chi^2$ -distributions over a wide range of bandwidth  $h_1$  (the degree of freedom depends on bandwidth  $h_2$  but not  $h_1$ ). The middle panel demonstrates the Wilks type of phenomenon: For three very different choices of nuisance functions, the null distributions are nearly the same.

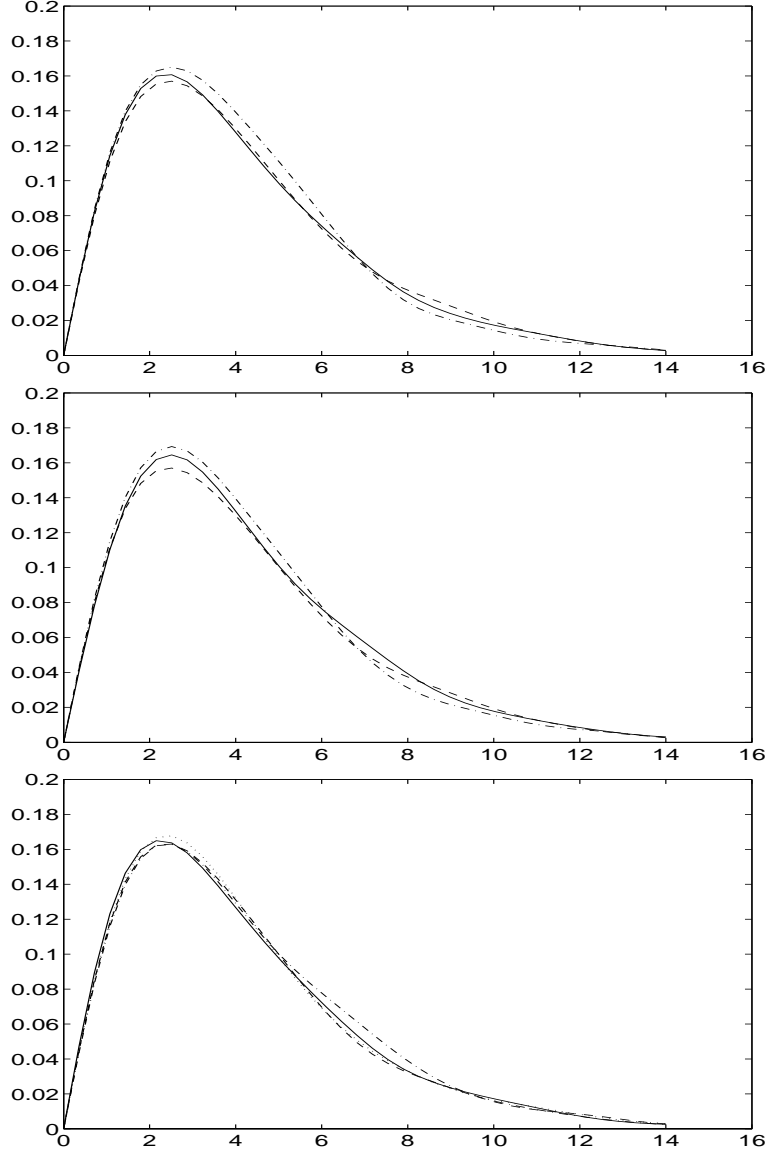


Figure 1: Results for Example 1. Estimated densities for the GLR statistics among 1000 simulations. Top Panel: with fixed  $h_2 = h_{2,opt}$ , but different bandwidths for  $h_1$  (solid -  $h_1 = \frac{2}{3}h_{1,opt}$ ; dashed -  $h_1 = h_{1,opt}$ ; dash-dotted -  $h_1 = \frac{3}{2}h_{1,opt}$ ); Middle Panel: with different nuisance functions and optimal bandwidths  $h_d = h_{d,opt}$  (solid -  $\beta = -1.5$ ; dashed -  $\beta = 0$ ; dotted -  $\beta = 1.5$ ); Bottom Panel: estimated densities for the GLR statistics under different errors (solid - normal; dashed -  $t(5)$ ; dotted -  $\chi^2(5)$ ; dash-dotted -  $\chi^2(10)$ );

For the power assessment, we evaluate the power for a sequence of alternative models that are indexed by  $\theta$ :

$$H_\theta : m_{2,\theta}(x_2) = \theta \sin(\pi x_2), \quad 0 \leq \theta \leq 1, \quad (5.3)$$

ranging from the null model to reasonably far away from it. The left panel of Figure 2 reports the differences between the null and the alternatives in (5.3).

For each given value of  $\theta$ , we use 3000 Monte Carlo replicates for the calculation of the critical values via the conditional bootstrap method (see Section 4.2), and compute the rejection frequencies based on 600 simulations. The parameter  $\theta$  is related to the separation distance between the null and the alternative hypotheses. Note that when  $\theta = 0$ , the alternative is the same as the null hypothesis, so that the power should approximately be 0.05 (or 0.10) at the 0.05 (or 0.10) significance level. This is indeed the case, as shown in Table 1, which again implies that the Monte Carlo method gives a correct estimator of the null distribution. When  $\theta$  increases, the alternative moves further away from the null hypothesis. One would expect the rejection rates of the null hypothesis to get higher and higher, which is evidenced in Table 2.

To investigate the power and the influence of different error distributions on the *GLR* tests, we now consider the model (5.1) with different error distributions of  $\varepsilon$ . In addition to the standard normal distribution, the standardized  $t(5)$  and the standardized  $\chi^2(5)$  and  $\chi^2(10)$  are also used. Note that the  $t(5)$ -distribution is of heavy tail and the chi-squared distributions are asymmetric. They are used to assess the stability of the performance of the *GLR* tests for different error distributions. The sample size is  $n = 200$ . The estimated densities of the normalized *GLR* statistics under the above four different error distributions are reported in the bottom panel of Figure 1. The figure shows that the null distributions of the tests are approximately the same for different error distributions and again exemplifies the Wilks phenomenon stated in Theorem 3.1. The powers of the *GLR* tests for the alternative sequence in (5.3) under different error distributions are given in Table 1, which shows a surprisingly stable performance of the tests for different error distributions with the characteristics of light or heavy tails and symmetric or asymmetric densities. The numerical results here suggest that the *GLR* tests not only have high power to differentiate the null and the smooth alternatives, but also possess robustness, to some extent, against error distributions.

Table 1: Powers of the proposed tests under different error distributions

$\alpha$	error distr. \ $\theta$	0.0	0.1	0.2	0.4	0.6	0.8	1.0
0.05	$\mathcal{N}(0, 1)$	0.057	0.068	0.192	0.592	0.948	0.997	1.000
	$t(5)$	0.043	0.068	0.146	0.537	0.903	0.998	1.000
	$\chi^2(5)$	0.048	0.077	0.175	0.640	0.963	0.995	1.000
	$\chi^2(10)$	0.053	0.090	0.230	0.657	0.952	0.995	1.000
0.10	$\mathcal{N}(0, 1)$	0.095	0.130	0.280	0.728	0.977	0.997	1.000
	$t(5)$	0.090	0.150	0.255	0.710	0.952	1.000	1.000
	$\chi^2(5)$	0.088	0.135	0.268	0.727	0.973	0.997	1.000
	$\chi^2(10)$	0.087	0.133	0.298	0.717	0.967	0.998	1.000

**Example 2.** Instead of considering a nonparametric null hypothesis against a nonparametric alternative, we deal with parametric null hypothesis to compare the performance of the bias cor-



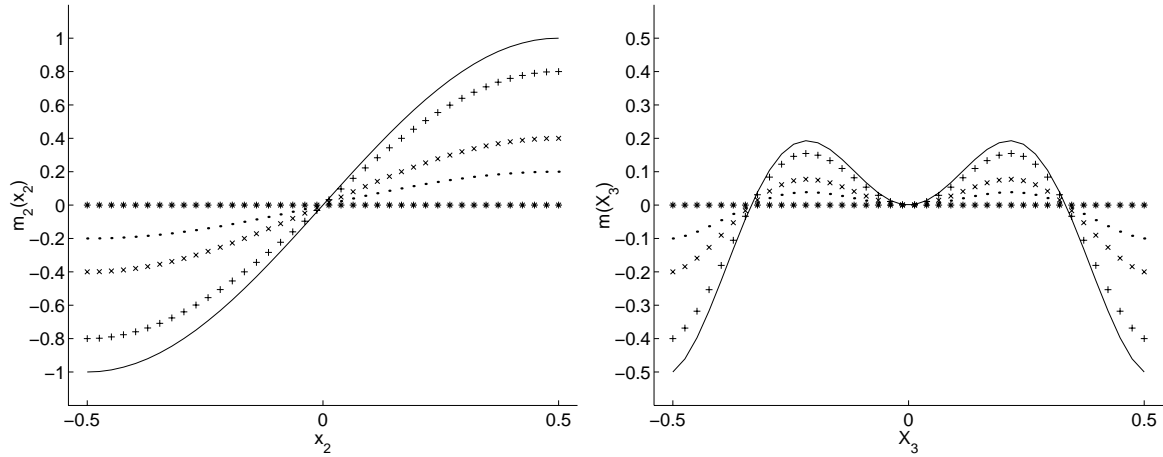


Figure 2: *Difference between the null and the alternative hypotheses. Left panel: for example 1; Right panel: for Example 2 (star -  $\theta = 0$ ; dotted -  $\theta = 0.2$ ; check -  $\theta = 0.4$ ; plus -  $\theta = 0.8$ ; solid -  $\theta = 1.0$ ).*

Table 2: Powers of the proposed tests under different error distributions

$\alpha$	error distr. \ $\theta$	0.0	0.2	0.4	0.6	0.8	1.0
0.05	$\mathcal{N}(0, 1)$	0.055	0.238	0.803	0.995	1.000	1.000
	$t(5)$	0.038	0.206	0.792	0.998	1.000	1.000
	$\chi^2(5)$	0.055	0.193	0.780	0.985	1.000	1.000
	$\chi^2(10)$	0.055	0.235	0.810	0.993	1.000	1.000
0.10	$\mathcal{N}(0, 1)$	0.132	0.328	0.875	1.000	1.000	1.000
	$t(5)$	0.085	0.315	0.870	1.000	1.000	1.000
	$\chi^2(5)$	0.122	0.317	0.872	0.995	1.000	1.000
	$\chi^2(10)$	0.113	0.338	0.888	0.998	1.000	1.000

rected *GLR* test with its counterpart for a testing problem with a parametric null hypothesis. The following three dimensional additive model is used:

$$Y = m(X_1, X_2, X_3) + \varepsilon, \quad (5.4)$$

where  $m(X_1, X_2, X_3) = m_1(X_1) + m_2(X_2) + m_3(X_3)$ ,  $m_1(X_1) = b_1 X_1^3$ ,  $m_2(X_2) = \sin(b_2 X_2)$ , and  $m_3(X_3) = \sin(b_3 X_3)$  with  $\mathbf{b} = (b_1, b_2, b_3) = (9, 3\pi, 3\pi)$ . The covariates are generated from a joint distribution with marginals  $\mathcal{N}(0, 1/9)$ , the correlation between  $X_1$  and  $X_2$  is 0.25, and  $X_3$  is independent of  $(X_1, X_2)$ . We rejected all observations in which one of the covariates fell out  $[-0.5, 0.5]$  and replaced them with new observations, so that the support of the covariates is bounded. The error  $\varepsilon$  is distributed as  $\mathcal{N}(0, 1/4)$ . The null model is taken as  $H_0 : \{(b_1, b_2, b_3) \in \mathcal{R}^3\}$ , which is fully parametric and can easily be fitted by the nonlinear regression function “nlinfit” in Matlab. Throughout this example, the bandwidths are fixed at their optimal values and the sample size is  $n = 200$ .

The power of the GLR test is evaluated at the following sequence of alternative models:

$$H_\theta : m_\theta(X_1, X_2, X_3) = m(X_1, X_2, X_3) + \theta X_3 \cdot m_3(X_3), \quad 0 \leq \theta \leq 1. \quad (5.5)$$

When  $\theta = 0$ ,  $H_\theta = H_0$ . As  $\theta$  increases, the alternative model  $H_\theta$  deviates away from  $H_0$ . The right panel of Figure 2 gives the difference between the null and the alternative models.

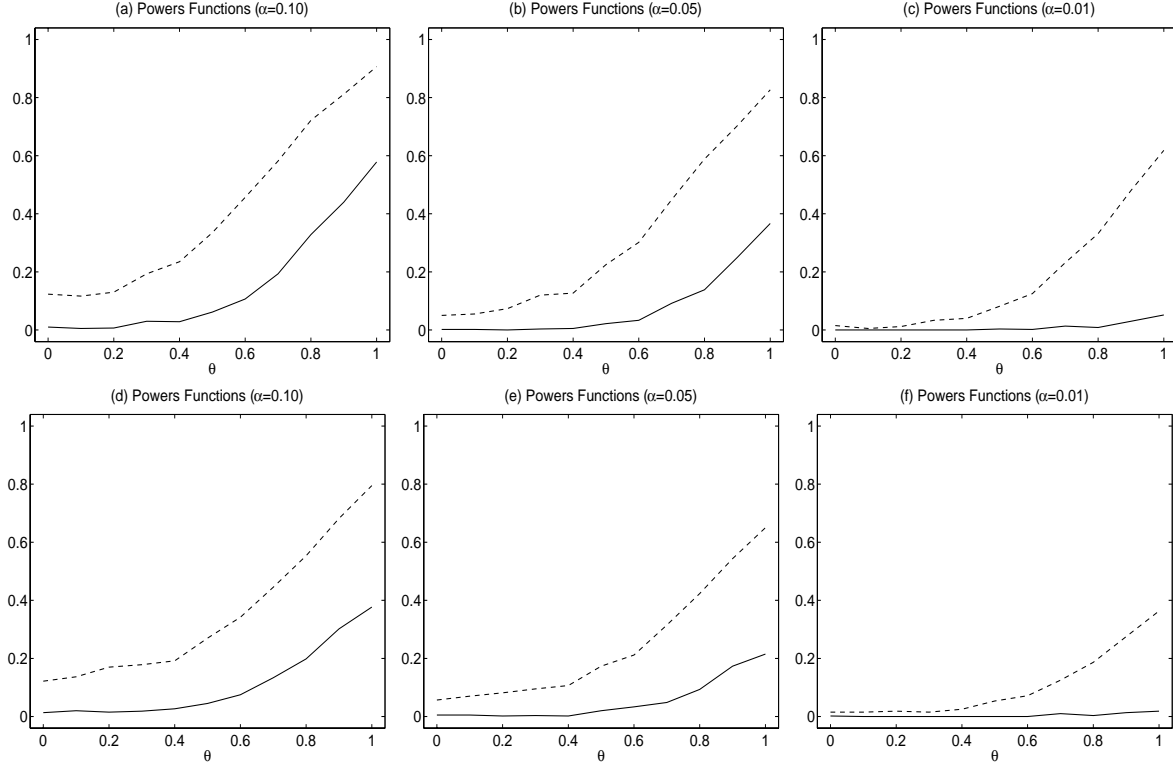


Figure 3: *Power functions of the GLR tests for Example 2. Top panel:  $\mathbf{h} = \mathbf{h}_{opt}$ ; Bottom panel:  $\mathbf{h} = \mathbf{h}_{opt}/2$ . From left to right, significance levels are  $\alpha = 0.10$ ,  $0.05$ , and  $0.01$ , respectively. The dashed lines are for the bias corrected method, and the solid lines are for the tests without bias reduction.*

For each given  $\theta$ , we simulated data from the alternative model  $H_\theta$ . The percents of rejection for  $H_0$  were computed based on the same simulation method as in Example 1. The results are given in Figure 3. When  $\theta = 0$ , the power of both tests become the sizes of the tests. It is evident from Figure 3 that the bias corrected test is more powerful than its counterpart. Note that the bandwidths that are used above are optimal for estimation. By setting the bandwidths to be half of their optimal values, the bias of the backfitting estimator decreases and the relative advantage of the bias correction method over its counterpart declines. This is evidenced in Figure 3, where the power of the bias corrected test increases faster than its counterpart as the bandwidths increase. These are in line with our asymptotic results.

## 6 Real Data Example

We use the proposed *GLR* tests on the Boston Housing dataset to demonstrate their usage in applications. The dataset comprises of the median value of homes in 506 census tracts in the Boston Standard Metropolitan Statistical Area in 1970, and 13 accompanying sociodemographic and related variables. It has previously been studied by several authors, including Harrison and Rubinfeld (1978), Belsley, Kuh, and Welch (1980), Breiman and Friedman (1985), and Opsomer and Ruppert (1998). Of the 13 variables, we use the following dependent variable and covariates of interest to demonstrate how our *GLR* tests work in practice:

MV: median value of owner-occupied homes (in \$1,000)

RM: average number of rooms

TAX: full property tax rate (\$/\$10,000)

PTRATIO: pupil/teacher ratio by town school district

LSTAT: proportion of population that is of “lower status” (%).

The last four covariates were also chosen by Breiman and Friedman (1985) and Opsomer and Ruppert (1998) to investigate the factors that affect the median value of owner-occupied homes.

Opsomer and Ruppert (1998) analyzed the dataset via a four dimensional additive model:

$$E[\text{MV} - \overline{\text{MV}} | X_1, X_2, X_3, X_4] = m_1(X_1) + m_2(X_2) + m_3(X_3) + m_4(X_4), \quad (6.1)$$

where  $X_1 = \text{RM}$ ,  $X_2 = \log(\text{TAX})$ ,  $X_3 = \text{PTRATIO}$ , and  $X_4 = \log(\text{LSTAT})$ . The local linear smoother and a fully automated bandwidth selection method were employed after six outliers were removed. They suggested that the fitted additive components are of apparent features: a linear term for *PTRATIO* and logarithmic terms for *TAX* and *LSTAT*.

We now focus on the model diagnostic problems. Specifically we check whether the fitted functions are of certain parametric forms. Added variable plots (Cook and Weisberg 1982) are useful in this case, see Opsomer and Ruppert (1998). Fitting the data with model (6.1) via the method of Opsomer and Ruppert (1998), we get the partial residuals. Figure 4 reports the partial residual plots along with their simple polynomial regression to indicate their trends and the fitted additive components based on the backfitting algorithm with a local linear smoother. More precisely, the following fully parametric models are fitted to the partial residuals:

$$\begin{aligned} m_1(X_1) &= a_1 + b_1 X_1 + c_1 X_1^2, & m_2(X_2) &= a_2 + b_2 X_2, \\ m_3(X_3) &= a_3 + b_3 X_3, & m_4(X_4) &= a_4 + b_4 X_4. \end{aligned} \quad (6.2)$$

Intuitively, apart from the fitted line for the variable *RM*, these regression lines seem consistent with the data. It is natural to ask whether the additive components apart from the variable *RM*

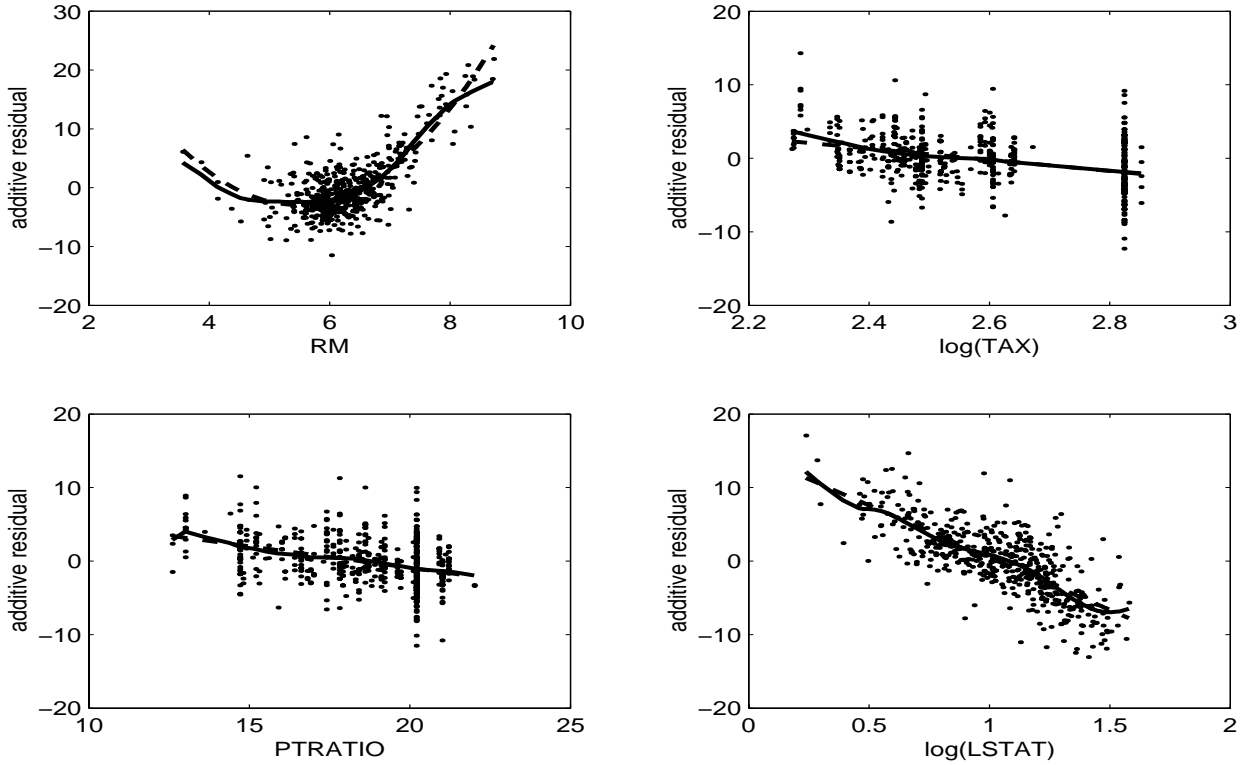


Figure 4: Partial residual plots along with fitted regression curves for the Boston Housing dataset. Solid – estimated additive functions; dashed – empirical regression lines based on model (6.2).

admit these parametric forms, namely, whether the following semiparametric model is consistent with the data:

$$m_2(X_2) = a_2 + b_2X_2, \quad m_3(X_3) = a_3 + b_3X_3, \quad m_4(X_4) = a_4 + b_4X_4, \quad (6.3)$$

where  $m_1(X_1)$  is unspecified.

Table 3: Results of the *GLR* tests.  $RSS_0$  and  $RSS_1$  – sum of squared residuals for the *GLR* test under  $H_0$  and  $H_1$ , respectively; *GLRT* – the normalized *GLR* statistic.

Bandwidth	$RSS_0$	$RSS_1$	<i>GLRT</i>	P-value
$\frac{1}{2}\mathbf{h}_{opt}$	1974.3	1721.9	31.0	0.097
$\frac{3}{3}\mathbf{h}_{opt}$	2044.0	1812.5	27.0	0.034
$\mathbf{h}_{opt}$	2091.2	1904.3	20.8	0.016
$\frac{3}{2}\mathbf{h}_{opt}$	2158.6	2042.3	12.0	0.046
$2\mathbf{h}_{opt}$	2301.7	2231.6	6.65	0.179

We now use our *GLR* statistic to test whether the semiparametric null model (6.3) holds against the additive alternative model (6.1). To compute the P-value of the test statistic, we need to find the null distribution of the *GLR* statistic  $\lambda_n(H_0)$ . This can be estimated by the conditional bootstrap method in Section 4.2. The P-value of our *GLR* test is estimated as zero by employing the optimal

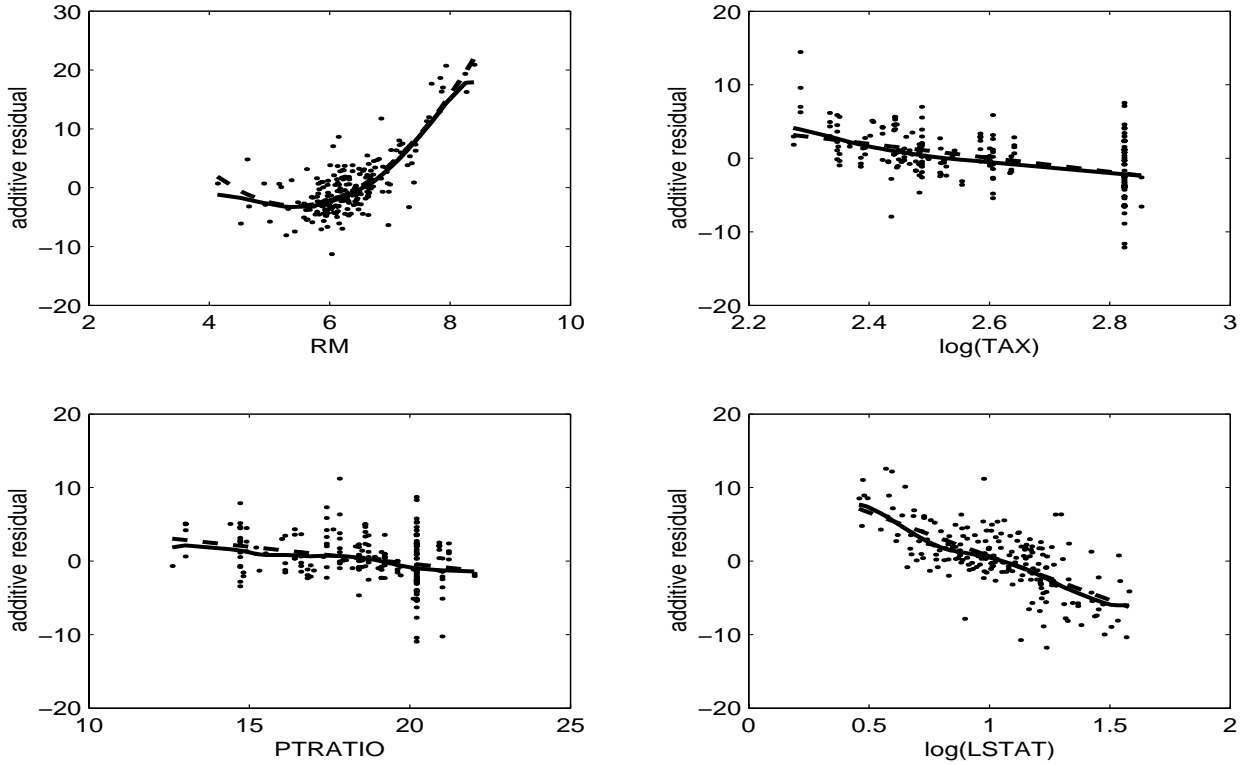


Figure 5: Partial residual plots along with fitted regression curves for a random subsample of the Boston Housing dataset. Solid – estimated additive functions; dashed – empirical regression lines based on model (6.2).

bandwidth and using 1000 bootstrap replicates. This does not come as a surprise to us, as the P-value depends heavily on the sample size. With a sample size as large as 500, a small deviation from the null hypothesis should lead to a tiny P-value. Hence, we take a random subsample of  $n = 200$  for analysis. The partial residuals from model (6.1) for the subsample are reported in Figure 5, where the fitted additive components from models (6.1) and (6.2) are also given. The optimal bandwidth from the automated bandwidth selection rule of Opsomer and Ruppert (1998) is computed to be  $\mathbf{h}_{opt} = (1.1129, 0.2530, 2.1432, 0.2315)^T$ . Visually, similar parametric forms of the additive components are suggested from Figures 4 and 5. Our interest is to test whether the model (6.3) is adequate for the subsample. Table 3 reports the results of the *GLR* tests for five different bandwidths, using 1000 bootstrap replicates. This provides stark evidence that the semiparametric model is appropriate for this dataset within the additive models at the 0.01 significance level.

## 7 Discussion

### 7.1 Other tests

There are many nonparametric tests designed for certain specific problems. Most of them are in univariate nonparametric regression setting. See Fan, Zhang and Zhang (2001) for an overview

of the literature. While they can be powerful for their problems where the tests were designed, extensions of these tests to multivariate setting can pose some challenges. Further, these tests are usually not distribution free, when null hypotheses involve nuisance functions. This would hamper their applicability.

Hypothesis testing for multivariate regression problems is difficult due to the curse of dimensionality. In bivariate regression, Aerts et al. (1999) constructed tests based on orthogonal series. Fan and Huang (2001) proposed various testing techniques based on the adaptive Neyman test for various alternative models in multiple regression setting. These problems become conceptually simple by using our generalized likelihood method. Delgado and González-Manteiga (2001) developed a test to select explanatory variables in nonparametric regression based on functionals of a  $U$ -process, while this test can detect a specific class of contiguous alternatives at a rate  $n^{-\frac{1}{2}}$ . However, this requires one to estimate the joint density of the significant variables and the regression function. In addition, Gozalo and Linton (2001) studied several tests for additivity in generalized nonparametric regression based on the integration estimation method and the generalized method of moments. Neumeier and Sperlich (2003) developed a test for difference of impacts from a specific covariate on the regression curve in two independent samples, via comparing a distance of the fitted curves based on the integration estimation approach.

Our  $GLR$  tests are motivated by comparing the pseudo-likelihood of the nearly best fitting in the null and alternative models, which leads to the log ratio of the variance estimators under the null and the alternative. This lends further support to the widely used goodness-of-fit test for a parametric regression that was constructed based on the variance estimators from a parametric fitting and a nonparametric kernel smoother, see for example Dette (1999). Our  $GLR$  tests are asymptotically distribution free, and possess the Wilks' type of results. They are asymptotically optimal in terms of convergence for nonparametric hypothesis testing according to the formulations of Ingster (1993) and Spokoiny (1996).

## 7.2 Extension

Under *iid* errors, the generalized likelihood ratio tests are derived for nonlinear additive models (1.1) based on the local polynomial smoother and the backfitting algorithm. For the heteroscedastic errors, for example the model

$$Y_i = \alpha + \sum_{d=1}^D m_d(X_{di}) + \varepsilon_i, \quad i = 1 \cdots, n,$$

where  $\varepsilon_i = \sigma(\mathbf{X}_i)u_i$  and  $\sigma^2(\mathbf{X}_i) = \sum_{d=1}^D \sigma_d^2(X_{di})$  is also of additive form to assuage the curse of dimensionality. Note that

$$\varepsilon_i^2 = \sigma^2(\mathbf{X}_i) + \sigma^2(\mathbf{X}_i)(u_i^2 - 1). \quad (7.1)$$

Our method continues to apply by considering the generalized likelihood ratio statistic for  $\{u_i\}$  which consists of the following three steps:

- (a) Fit the regression components by backfitting algorithm, and obtain the residuals  $\hat{\varepsilon}_i = Y_i - \bar{Y} - \sum_{d=1}^D \hat{m}_d(X_{di})$ ;
- (b) Obtain the estimator  $\hat{\sigma}(\mathbf{X}_i)$  by fitting the model (7.1) with  $\varepsilon_i$  replaced by  $\hat{\varepsilon}_i$ , and get the  $RSS_1 = \sum_{i=1}^n \hat{u}_i^2$ , where  $\hat{u}_i = \hat{\varepsilon}_i / \hat{\sigma}(\mathbf{X}_i)$ .
- (c) Compute  $RSS_0 = \sum_{i=1}^n (\hat{\varepsilon}_i^0)^2 / \hat{\sigma}^2(\mathbf{X}_i)$  and form the *GLR* (3.2), where  $\hat{\varepsilon}_i^0$  is the residual under  $H_0$ .

The conditional bootstrap approximation in Section 4.2 can also be adapted to this situation, if one draws bootstrap residuals from the centered empirical distribution of  $\{\hat{u}_i\}_{i=1}^n$ . For other forms of the conditional standard deviation  $\sigma(\mathbf{X}_i)$ , the above method still applies but with other fitting techniques for  $\sigma(\mathbf{X}_i)$ . The techniques can also be extended to the generalized additive models (Hastie and Tibshirani, 1990). We would expect that similar results continue to hold.

In implementation, two forms of bandwidths have been introduced: constant bandwidth and constant span (see e.g. LOWESS in Cleveland, 1979). The constant span has an advantage of avoiding the sparsity of design points, but the bandwidths at such regions are large and hence it can introduce large modelling biases. When the constant span is used, its effective bandwidth depends on the design density and is usually not a constant. Our asymptotic results can be extended to such a constant span case, but its normalization constant and degree of freedom will depend on nuisance functions under the null hypothesis. In other words, the Wilks phenomenon does not hold and this makes estimating the null distribution harder. The situation is very much like using the ordinary GLR tests in the heteroscedastic model. The asymptotic results can be extended but the normalization constant and degrees of freedom depend on the unknown variance function. See Remark 4.2 of Fan *et al.* (2001) for this kind of results.

### APPENDIX I: Condition (A)

To derive the asymptotic distributions of the testing statistics, we make the following technical assumptions and use the following notation.

- (1) The kernel function  $K(x)$  is bounded and Lipschitz continuous with a bounded support.
- (2) The densities  $f_d(x_d)$  of  $X_d$  are Lipschitz continuous and bounded away from 0, and have bounded supports  $\Omega_d$  for  $d = 1, \dots, D$ .
- (3) The joint density of  $X_d$  and  $X_{d'}$ ,  $f_{dd'}(x_d, x_{d'})$ , is Lipschitz continuous on its support  $\Omega_d \times \Omega_{d'}$ .
- (4) As  $n \rightarrow \infty$ ,  $h_d \rightarrow 0$  and  $nh_d / \log(n) \rightarrow \infty$  for  $d = 1, \dots, D$ .
- (5) The  $(p_d + 1)$ th derivatives of  $m_d$  ( for  $d = 1, \dots, D$  ) exist and are bounded and continuous.
- (6)  $E|\varepsilon_i|^4 < \infty$ .

## APPENDIX II: PROOFS

In this appendix, we give technical proofs of the theorems. Let  $\mathbf{P}_1 \approx \mathbf{P}_2$  denote  $\mathbf{P}_1 = \mathbf{P}_2(1 + o(1))$  *a.s.*, componentwise for any matrices  $\mathbf{P}_1, \mathbf{P}_2$  of the same dimension. For any constant  $d$ ,  $\mathbf{d}$  is the  $n$ -valued vector  $(d, \dots, d)^T$ . Denote by  $\bar{Z}$  the average of components of any vector  $\mathbf{Z}$ . To facilitate the exposition of the proofs, we will ignore the intercept  $\alpha$  and introduce the following technical lemmas. Because  $\hat{\alpha}$  is root- $n$  consistent, the same arguments can be used for the case with the unknown intercept.

**Lemma 1** *Let assumptions (1)-(4) in Condition (A) hold. Then*

$$\mathbf{s}_{d,x_d}^T \approx n^{-1} f_d^{-1}(x_d) \mathbf{e}_1^T \tilde{\mathbf{S}}_d^{-1} \mathbf{H}_d^{-1} \mathbf{X}_{x_d}^d{}^T \mathbf{K}_{x_d},$$

*uniformly for  $x_d \in \Omega_d$ .*

**Proof.** The result is derived on page 64 in Fan and Gijbels (1996).

**Lemma 2** *Under assumptions (1)-(4) in Condition (A), the following asymptotic approximations hold uniformly over all elements of the matrices:*

$$\begin{aligned} \mathbf{S}_d^* &= \mathbf{S}_d - \frac{\mathbf{1}\mathbf{1}^T}{n} + o\left(\frac{\mathbf{1}\mathbf{1}^T}{n}\right) \quad \text{a.s.}, \\ \mathbf{S}_d^* \mathbf{S}_{d'}^* &= \mathbf{T}_{dd'}^* + o\left(\frac{\mathbf{1}\mathbf{1}^T}{n}\right) \quad \text{a.s.}, \end{aligned}$$

where  $\mathbf{T}_{dd'}^*$  is a matrix with  $(i, j)$ th element,

$$[\mathbf{T}_{dd'}^*]_{ij} = \frac{1}{n} \left[ \frac{f_{dd'}(X_{di}, X_{d'j})}{f_d(X_{di})f_{d'}(X_{d'j})} - 1 \right].$$

**Proof.** This is shown in Lemma 3.1 of Opsomer and Ruppert (1997).

**Lemma 3** *Denote by  $\mathbf{A}_{n1} = (\mathbf{W}_M^{[-D]} - \mathbf{I}_n)^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n)$  and  $\mathbf{A}_{n2} = (\mathbf{W}_M - \mathbf{I}_n)^T (\mathbf{W}_M - \mathbf{I}_n)$ . If assumptions (1)-(4) in Condition (A) hold, then conditional on  $\mathcal{X}$*

$$\text{RSS}_0 - \text{RSS}_1 = \mathbf{Y}^T [\mathbf{A}_{n1} - \mathbf{A}_{n2}] \mathbf{Y} \tag{A.1}$$

and

$$\mathbf{A}_{n1} - \mathbf{A}_{n2} = \mathbf{S}_D + \mathbf{S}_D^T - \mathbf{S}_D^T \mathbf{S}_D - \left( \sum_{d=1}^{D-1} \mathbf{S}_d \right)^T \mathbf{S}_D - \mathbf{S}_D^T \left( \sum_{d=1}^{D-1} \mathbf{S}_d \right) + \mathbf{R}_n, \tag{A.2}$$

where  $\mathbf{R}_n$  is a matrix whose  $(i, j)$ th element is  $[\mathbf{R}_n]_{ij}$  such that  $E\{[\mathbf{R}_n]_{i_1 j_1} [\mathbf{R}_n]_{i_2 j_2}\} = O(1/n^2)$  and  $[\mathbf{R}_n]_{ij} = O(\frac{1}{n})$  *a.s.* uniformly for  $1 \leq i, j; i_1, j_1; i_2, j_2 \leq n$ .



**Proof.** By definition, we have (A.1). Using an argument which is similar to that in the proof for Theorem 3.1 of Opsomer (2000), we obtain  $\mathbf{S}_d^* \mathbf{W}_M^{[-d]} = O(\frac{\mathbf{1}\mathbf{1}^T}{n})$  *a.s.*, and  $(\mathbf{I}_n - \mathbf{S}_d^* \mathbf{W}_M^{[-d]})^{-1} = \mathbf{I}_n + O(\frac{\mathbf{1}\mathbf{1}^T}{n})$ , uniformly over all elements of the matrix. Throughout the proof of this lemma, the term  $O(\mathbf{1}\mathbf{1}^T/n)$  means that each element is of order  $O(1/n)$ . Then by (2.4), Lemma 2 and direct matrix multiplications

$$\begin{aligned} \mathbf{W}_M &= \sum_{d=1}^D \mathbf{W}_d = \sum_{d=1}^D (\mathbf{I}_n - \mathbf{S}_d^* \mathbf{W}_M^{[-d]})^{-1} \mathbf{S}_d^* (\mathbf{I}_n - \mathbf{W}_M^{[-d]}) \\ &= \mathbf{S} + \mathbf{U}, \end{aligned}$$

where  $\mathbf{S} = \sum_{d=1}^D \mathbf{S}_d$  and  $\mathbf{U} = O(\frac{\mathbf{1}\mathbf{1}^T}{n})$  *a.s.*. Hence,

$$\mathbf{A}_{n2} = \mathbf{S}^T \mathbf{S} - \mathbf{S} - \mathbf{S}^T + \mathbf{I}_n + \mathbf{R}_{n2},$$

where  $\mathbf{R}_{n2} = O(\frac{\mathbf{1}\mathbf{1}^T}{n})$  *a.s.*. Similarly, we have

$$\mathbf{W}_M^{[-D]} = \mathbf{S}^{[-D]} + \mathbf{U}^{[-D]} \quad (\text{A.3})$$

and

$$\mathbf{A}_{n1} = \mathbf{S}^{[-D]T} \mathbf{S}^{[-D]} - \mathbf{S}^{[-D]} - \mathbf{S}^{[-D]T} + \mathbf{I}_n + \mathbf{R}_{n1},$$

where  $\mathbf{S}^{[-D]} = \sum_{d=1}^{D-1} \mathbf{S}_d$ ,  $\mathbf{U}^{[-D]} = O(\frac{\mathbf{1}\mathbf{1}^T}{n})$  *a.s.*, and  $\mathbf{R}_{n1} = O(\frac{\mathbf{1}\mathbf{1}^T}{n})$  *a.s.*. Therefore,

$$\mathbf{A}_{n1} - \mathbf{A}_{n2} = \mathbf{S}_D + \mathbf{S}_D^T - \mathbf{S}_D^T \mathbf{S}_D - \left( \sum_{d=1}^{D-1} \mathbf{S}_d \right)^T \mathbf{S}_D - \mathbf{S}_D^T \left( \sum_{d=1}^{D-1} \mathbf{S}_d \right) + \mathbf{R}_n,$$

with  $\mathbf{R}_n = O(\frac{\mathbf{1}\mathbf{1}^T}{n})$  *a.s.*. Furthermore, by assumption (2) in Condition (A), we complete the proof of the lemma.

**Lemma 4** *Let*

$$\mathbf{Q}_d = \begin{bmatrix} \mathbf{s}_{d, X_{d1}}^T \mathbf{Q}_{m_d}(X_{d1}) \\ \vdots \\ \mathbf{s}_{d, X_{dn}}^T \mathbf{Q}_{m_d}(X_{dn}) \end{bmatrix}$$

and

$$\mathbf{Q}_d^* = \left( \mathbf{I}_n - \frac{\mathbf{1}\mathbf{1}^T}{n} \right) \mathbf{Q}_d,$$

where

$$\mathbf{Q}_{m_d}(x_d) = \begin{bmatrix} (X_{d1} - x_d)^{p_d+1} \\ \vdots \\ (X_{dn} - x_d)^{p_d+1} \end{bmatrix} \frac{\partial^{p_d+1} m_d(x_d)}{\partial x_d^{p_d+1}}.$$

If assumptions (1)-(5) in Condition (A) hold, then

$$\mathbf{Q}_d = C_d^{(p_d+1)} h_d^{p_d+1} \mathcal{D}^{p_d+1} \mathbf{m}_d + o(h_d^{p_d+1}) \quad \textit{a.s.},$$

where

$$\mathcal{D}^{p_d+1} \mathbf{m}_d = \begin{bmatrix} \frac{\partial^{p_d+1} m_d(X_{d1})}{\partial x_d^{p_d+1}} \\ \vdots \\ \frac{\partial^{p_d+1} m_d(X_{dn})}{\partial x_d^{p_d+1}} \end{bmatrix}.$$

**Proof.** The lemma follows by Taylor's expansion.

**Lemma 5** Put  $\mathbf{B}^{(d)} = E[\mathbf{W}_d \mathbf{Y} - \mathbf{m}_d | \mathbf{X}]$  and  $\mathbf{B} = E[\mathbf{W}_M \mathbf{Y} - \mathbf{m} | \mathbf{X}] = (\mathbf{W}_M - \mathbf{I}_n) \mathbf{m}$ , where  $\mathbf{B}$  is the conditional bias in estimation of  $\mathbf{m}$  by the model (1.1). If Condition (A) holds, then

$$\begin{aligned} \mathbf{B}^{(D)} &= (\mathbf{I}_n - \mathbf{S}_D^* \mathbf{W}_M^{[-D]})^{-1} \left[ \frac{1}{(p_D + 1)!} \mathbf{Q}_D^* - \mathbf{S}_D^* \mathbf{B}_{-D} \right] + \bar{m}_D O(1) + o(\mathbf{h}_D^{p_D+1}) \quad a.s., \\ \mathbf{B} &= O\left(\sum_{d=1}^D \mathbf{h}_d^{p_d+1}\right) + \sum_{d=1}^D \bar{m}_d \cdot O(1) \quad a.s., \end{aligned} \quad (\text{A.4})$$

uniformly over all elements of the vector, where  $\mathbf{B}_{-D} = (\mathbf{W}_M^{[-D]} - \mathbf{I}_n) \mathbf{m}_{(-D)}$  is the conditional bias in estimation of  $\mathbf{m}_{(-D)}$  by the  $(D-1)$ -variate regression model

$$Y_i' = \sum_{d=1}^{D-1} m_d(X_{di}) + \varepsilon_i.$$

**Proof.** Applying the same Taylor expansion approximations as in Theorem 2.1 of Ruppert and Wand (1994), we obtain

$$\mathbf{S}_d \mathbf{m}_d = \mathbf{m}_d + \frac{1}{(p_d + 1)!} \mathbf{Q}_d + o(\mathbf{h}_d^{p_d+1}).$$

Then by Lemma 2

$$(\mathbf{I}_n - \mathbf{S}_d^*) \mathbf{m}_d = \bar{m}_d \mathbf{1} - \frac{1}{(p_d + 1)!} \mathbf{Q}_d^* + o(\mathbf{h}_d^{p_d+1}).$$

It follows from (2.4) that

$$\begin{aligned} (\mathbf{I}_n - \mathbf{W}_D) \mathbf{m}_D &= (\mathbf{I}_n - \mathbf{S}_D^* \mathbf{W}_M^{[-D]})^{-1} (\mathbf{I}_n - \mathbf{S}_D^*) \mathbf{m}_D \\ &= (\mathbf{I}_n - \mathbf{S}_D^* \mathbf{W}_M^{[-D]})^{-1} \left[ \bar{m}_D \mathbf{1} - \frac{1}{(p_D + 1)!} \mathbf{Q}_D^* \right] + o(\mathbf{h}_D^{p_D+1}) \quad a.s. \end{aligned} \quad (\text{A.5})$$

Note that

$$\begin{aligned} (\mathbf{I}_n - \mathbf{W}_D) \mathbf{m}_{(-D)} &= (\mathbf{I}_n - \mathbf{S}_D^* \mathbf{W}_M^{[-D]})^{-1} (\mathbf{I}_n - \mathbf{S}_D^*) \mathbf{m}_{(-D)} \\ &= \mathbf{m}_{(-D)} + (\mathbf{I}_n - \mathbf{S}_D^* \mathbf{W}_M^{[-D]})^{-1} \mathbf{S}_D^* \mathbf{B}_{-D}. \end{aligned} \quad (\text{A.6})$$

This, together with (A.5), leads to

$$\mathbf{B}^{(D)} = \bar{m}_D O(1) + (\mathbf{I}_n - \mathbf{S}_D^* \mathbf{W}_M^{[-D]})^{-1} \left[ \frac{1}{(p_D + 1)!} \mathbf{Q}_D^* - \mathbf{S}_D^* \mathbf{B}_{-D} \right] + o(\mathbf{h}_D^{p_D+1}) \quad a.s.$$

Hence, (A.4) holds by a recursive argument.

**Lemma 6** *If Condition (A) holds, then under  $H_0 : m_D = 0$*

$$d_{1n} \equiv \mathbf{m}^T(\mathbf{A}_{n1} - \mathbf{A}_{n2})\mathbf{m} + 2\varepsilon^T(\mathbf{A}_{n1} - \mathbf{A}_{n2})\mathbf{m} = O_p\left(1 + \sum_{d=1}^D nh_d^{2(p_d+1)} + \sum_{d=1}^D \sqrt{nh_d^{p_d+1}}\right), \quad (\text{A.7})$$

where  $\mathbf{A}_{n1}$  and  $\mathbf{A}_{n2}$  are defined in Lemma 3. Furthermore,  $d_{1n} \equiv O_p(1)$  if  $m_d(\cdot)$  is a polynomial of order  $p_d$  for  $d = 1, \dots, D$ .

**Proof.** (i) Under  $H_0$ , we obtain from Lemma 5 that

$$\mathbf{m}^T(\mathbf{A}_{n1} - \mathbf{A}_{n2})\mathbf{m} = \mathbf{B}_{-D}^T \mathbf{B}_{-D} - \mathbf{B}^T \mathbf{B} = O_p\left(1 + \sum_{d=1}^D nh_d^{2(p_d+1)}\right). \quad (\text{A.8})$$

By Lemmas 3 and 5, we have under  $H_0$

$$\begin{aligned} \mathbf{m}^T(\mathbf{A}_{n1} - \mathbf{A}_{n2})^T \varepsilon &= \mathbf{B}_{-D}^T \left( \mathbf{W}_M^{[-D]} - \mathbf{I}_n \right) \varepsilon - \mathbf{B}^T (\mathbf{W}_M - \mathbf{I}_n) \varepsilon, \\ E[(\mathbf{W}_M - \mathbf{I}_n)\varepsilon] &= 0, \end{aligned}$$

and

$$(\mathbf{W}_M - \mathbf{I}_n)\varepsilon = \mathbf{W}_M \varepsilon - \varepsilon = \sum_{d=1}^D \mathbf{S}_d \varepsilon - \varepsilon + O_p\left(\sum_{d=1}^D nh_d^{2(p_d+1)}\right).$$

By directly computing the mean and variance and using Lemma 1, we obtain  $\mathbf{B}^T \mathbf{S} \varepsilon = O_p(1 + \sum_{d=1}^D \sqrt{nh_d^{p_d+1}})$ ,  $\mathbf{B}^T \varepsilon = O_p(1 + \sum_{d=1}^D \sqrt{nh_d^{p_d+1}})$ , and hence  $\mathbf{B}^T (\mathbf{W}_M - \mathbf{I}_n) \varepsilon$  is bounded by  $O_p(1 + \sum_{d=1}^D \sqrt{nh_d^{p_d+1}})$ . With the same argument,  $\mathbf{B}_{-D}^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n) \varepsilon = O_p(1 + \sum_{d=1}^D \sqrt{nh_d^{p_d+1}})$ . Therefore, the second term in  $d_{1n}$  is  $O_p(1 + \sum_{d=1}^D \sqrt{nh_d^{p_d+1}})$ , which combined with (A.8) leads to (A.7).

(ii) Assume that  $m_d(\cdot)$  is a polynomial of order  $p_d$  (for  $d = 1, \dots, D$ ), then  $\mathbf{Q}_d = 0$ . Using recursive reasoning, we obtain  $\mathbf{B}_{-D} = \sum_{d=1}^{D-1} \bar{m}_d O(\mathbf{1}) = O_p(\frac{1}{\sqrt{n}})$  and  $\mathbf{B} = O_p(\frac{1}{\sqrt{n}})$ . Hence, the first term in  $d_{1n}$  is  $O_p(1)$ . Similarly, the second term in  $d_{1n}$  is  $O_p(1)$ , which completes the proof of the lemma.

**Proof of Theorem 3.1.** The proof mainly consists of the following four steps:

(i). **Asymptotic expression for  $\text{RSS}_0 - \text{RSS}_1$ :** By definition, we have

$$\text{RSS}_0 - \text{RSS}_1 = \|\mathbf{W}_M^{[-D]} \mathbf{Y} - \mathbf{Y}\|^2 - \|\mathbf{W}_M \mathbf{Y} - \mathbf{Y}\|^2,$$

which can be written, using the notation of Lemma 3, as

$$\begin{aligned} \text{RSS}_0 - \text{RSS}_1 &= \mathbf{Y}^T [\mathbf{A}_{n1} - \mathbf{A}_{n2}] \mathbf{Y} \\ &= \varepsilon^T (\mathbf{A}_{n1} - \mathbf{A}_{n2}) \varepsilon + [\mathbf{m}^T (\mathbf{A}_{n1} - \mathbf{A}_{n2}) \mathbf{m} + 2\varepsilon^T (\mathbf{A}_{n1} - \mathbf{A}_{n2}) \mathbf{m}] \\ &\equiv \varepsilon^T (\mathbf{A}_{n1} - \mathbf{A}_{n2}) \varepsilon + d_{1n}. \end{aligned} \quad (\text{A.9})$$

From Lemma 6,  $d_{1n}$  is bounded by  $O_p(1 + \sum_{d=1}^D nh_d^{2(p_d+1)} + \sum_{d=1}^D \sqrt{nh_d^{p_d+1}})$ . In the following, we will show that the first term in (A.9) can be approximated as:

$$\begin{aligned} \varepsilon^T[\mathbf{A}_{n1} - \mathbf{A}_{n2}]\varepsilon &\approx \frac{2}{nh_D} \sum_{i < j} \varepsilon_i \varepsilon_j f_D^{-1}(X_{Di}) \left\{ 2 \sum_{t=1}^{p_D+1} \tilde{s}_{D,t} K_t \left( \frac{X_{Dj} - X_{Di}}{h_D} \right) \right. \\ &\quad \left. - \sum_{s,t=1}^{p_D+1} \tilde{s}_{D,s} \tilde{s}_{D,t} K_s * K_t \left( \frac{X_{Dj} - X_{Di}}{h_D} \right) \right\} + \mu_n^* + o_p(h_D^{-1}) \\ &\equiv W_{(n)} + \mu_n^* + o_p(h_D^{-1}), \end{aligned} \quad (\text{A.10})$$

where  $\mu_n^* = 2\sigma^2 \mu_n$  with

$$\mu_n = h_D^{-1} |\Omega_D| \left( \sum_{t=1}^{p_D+1} \tilde{s}_{D,t} K_t(0) - \frac{1}{2} \sum_{s,t=1}^{p_D+1} \tilde{s}_{D,s} \tilde{s}_{D,t} K_s * K_t(0) \right).$$

Then, by (A.9)

$$\text{RSS}_0 - \text{RSS}_1 \approx W_{(n)} + 2\sigma^2 \mu_n + d_{1n} + o_p(h_D^{-1}). \quad (\text{A.11})$$

For readers who are not interested in the proof of (A.10), please skip to (ii). Note that the  $(j, \ell)$ th element of  $\mathbf{H}_d^{-1} \mathbf{X}_{X_{dk}}^d \mathbf{K}_{X_{dk}}$  is now

$$\frac{1}{h_d} K_j \left( \frac{X_{d\ell} - X_{dk}}{h_d} \right).$$

By Lemma 1, direct matrix multiplications give the  $(i, j)$ th element of  $\mathbf{S}_d$ :

$$(\mathbf{S}_d)_{ij} \approx \frac{1}{nh_d} f_d^{-1}(X_{di}) \sum_{t=1}^{p_d+1} \tilde{s}_{d,t} K_t \left( \frac{X_{dj} - X_{di}}{h_d} \right). \quad (\text{A.12})$$

Then by directly computing the mean and variance, we obtain from the Chebychev inequality

$$\sum_{i=1}^n \varepsilon_i^2 (\mathbf{S}_d)_{ii} \approx \frac{\sigma^2}{h_d} |\Omega_d| \sum_{t=1}^{p_d+1} \tilde{s}_{d,t} K_t(0) \quad (\text{A.13})$$

and

$$\sum_{i \neq j} \varepsilon_i \varepsilon_j (\mathbf{S}_d)_{ij} \approx \frac{1}{nh_d} \sum_{i \neq j} \varepsilon_i \varepsilon_j f_d^{-1}(X_{di}) \sum_{t=1}^{p_d+1} \tilde{s}_{d,t} K_t \left( \frac{X_{dj} - X_{di}}{h_d} \right). \quad (\text{A.14})$$

Similarly, using Lemma 1 we have the  $(i, j)$ th element of  $\mathbf{S}_d^T \mathbf{S}_{d'}$ :

$$\begin{aligned} (\mathbf{S}_d^T \mathbf{S}_{d'})_{ij} &\approx \frac{1}{n} \sum_{s=1}^{p_d+1} \sum_{t=1}^{p_{d'}+1} \tilde{s}_{d,s} \tilde{s}_{d',t} \frac{1}{n} \sum_{k=1}^n \left[ f_d^{-1}(X_{dk}) f_{d'}^{-1}(X_{d'k}) \frac{1}{h_d} K_s \left( \frac{X_{di} - X_{dk}}{h_d} \right) \frac{1}{h_{d'}} K_t \left( \frac{X_{d'j} - X_{d'k}}{h_{d'}} \right) \right] \\ &\equiv n^{-1} \sum_{s=1}^{p_d+1} \sum_{t=1}^{p_{d'}+1} \tilde{s}_{d,s} \tilde{s}_{d',t} \frac{1}{n} \sum_{k=1}^n P_{dd'ijkh}. \end{aligned} \quad (\text{A.15})$$

Then

$$\begin{aligned}
\sum_{i \neq j} \varepsilon_i \varepsilon_j (\mathbf{S}_d^T \mathbf{S}_{d'})_{ij} &= \frac{1}{n} \sum_{i \neq j} \varepsilon_i \varepsilon_j \sum_{s=1}^{p_d+1} \sum_{t=1}^{p_{d'}+1} \tilde{s}_{d,s} \tilde{s}_{d',t} \frac{1}{n} \sum_{k \neq i,j} P_{dd'ijkh} \\
&\quad + \frac{1}{n} \sum_{i \neq j} \varepsilon_i \varepsilon_j \sum_{s=1}^{p_d+1} \sum_{t=1}^{p_{d'}+1} \tilde{s}_{d,s} \tilde{s}_{d',t} \frac{1}{n} [P_{dd'ijih} + P_{dd'ijjh}] \\
&\equiv L_{n d d' 1} + L_{n d d' 2}.
\end{aligned} \tag{A.16}$$

It can easily be shown that  $E(L_{n d d' 2}) = 0$  and

$$\text{Var}(L_{n d d' 2}) = O\left(\frac{1}{n^2 h_d^2 h_{d'}} + \frac{1}{n^2 h_{d'}^2 h_d}\right),$$

which implies  $L_{n d d' 2} = O_p\left(\frac{1}{n h_d \sqrt{h_{d'}}} + \frac{1}{n h_{d'} \sqrt{h_d}}\right)$ . By the definition of  $P_{dd'ijkh}$  and taking the iterative expectation, we get for  $d \neq d'$  and  $k \neq i, j$  ( $i \neq j$ )

$$E[P_{dd'ijkh} | X_{di}, X_{d'j}] = \mu_{s-1} \mu_{t-1} \frac{f_{dd'}(X_{di}, X_{d'j})}{f_d(X_{di}) f_{d'}(X_{d'j})} + o_p(1),$$

uniformly for  $i, j = 1, \dots, n$ . Hence for  $d \neq d'$

$$\begin{aligned}
L_{n d d' 1} &\approx \frac{2(n-2)}{n^2} \sum_{i < j} \varepsilon_i \varepsilon_j \sum_{s=1}^{p_d+1} \sum_{t=1}^{p_{d'}+1} \tilde{s}_{d,s} \tilde{s}_{d',t} E[P_{dd'ijkh} | X_{di}, X_{d'j}] \\
&= \frac{2(n-2)}{n^2} \sum_{i < j} \varepsilon_i \varepsilon_j C_d^{(0)} C_{d'}^{(0)} \frac{f_{dd'}(X_{di}, X_{d'j})}{f_d(X_{di}) f_{d'}(X_{d'j})} (1 + o_p(1)) \\
&= O_p(1),
\end{aligned}$$

where the first approximation is from

$$E[n^{-1} \sum_{k \neq i,j} (P_{dd'ijkh} - E(P_{dd'ijkh} | X_{di}, X_{d'j}))^2] \leq n^{-2} \sum_{k \neq 1,2} E[P_{dd'12kh}^2] = O\left(\frac{1}{n h_d h_{d'}}\right). \tag{A.17}$$

Then, by (A.16), for  $d \neq d'$

$$\sum_{i \neq j} \varepsilon_i \varepsilon_j (\mathbf{S}_d^T \mathbf{S}_{d'})_{ij} = O_p\left(1 + \frac{1}{n h_d \sqrt{h_{d'}}} + \frac{1}{n h_{d'} \sqrt{h_d}}\right) \tag{A.18}$$

By (A.17), we have

$$n^{-1} \sum_{k \neq i,j} P_{dd'ijkh} = E(P_{dd'ijkh} | X_{di}, X_{d'j}) + o_p(1),$$

uniformly for  $i, j = 1, \dots, n$ , so that

$$\begin{aligned}
(\mathbf{S}_d^T \mathbf{S}_{d'})_{ij} &\approx \frac{1}{n h_d} \sum_{s,t=1}^{p_d+1} \tilde{s}_{d,s} \tilde{s}_{d',t} \mu_{s-1} \mu_{t-1} \frac{f_{dd'}(X_{di}, X_{d'j})}{f_d(X_{di}) f_{d'}(X_{d'j})} \\
&= \frac{1}{n} \frac{f_{dd'}(X_{di}, X_{d'j})}{f_d(X_{di}) f_{d'}(X_{d'j})}.
\end{aligned} \tag{A.19}$$

Therefore, for  $d \neq d'$

$$\sum_{i=1}^n \varepsilon_i^2 (\mathbf{S}_d^T \mathbf{S}_{d'})_{ii} \approx n^{-1} \sum_{i=1}^n \varepsilon_i^2 C_d^{(0)} C_{d'}^{(0)} \frac{f_{dd'}(X_{di}, X_{d'j})}{f_d(X_{di}) f_{d'}(X_{d'j})} = O_p(1). \quad (\text{A.20})$$

By the definition of  $P_{ddijkh}$  and using a change of variable, we obtain for  $i \neq j$  and  $k \neq i, j$

$$\begin{aligned} E[P_{ddijkh} | X_{di}, X_{dj}] &= \int f_d^{-2}(u) \frac{1}{h_d} K_s \left( \frac{X_{di} - u}{h_d} \right) \frac{1}{h_d} K_t \left( \frac{X_{dj} - u}{h_d} \right) f_d(u) du \\ &\approx \frac{1}{h_d} f_d^{-1}(X_{di}) \int K(v) v^{s-1} K \left( v + \frac{X_{dj} - X_{di}}{h_d} \right) \left( v + \frac{X_{dj} - X_{di}}{h_d} \right)^{t-1} dv \\ &= \frac{1}{h_d} f_d^{-1}(X_{di}) K_s * K_t \left( \frac{X_{dj} - X_{di}}{h_d} \right), \end{aligned}$$

which combined with (A.17) leads to

$$n^{-1} \sum_{k \neq i, j} P_{ddijkh} \approx \frac{1}{h_d} f_d^{-1}(X_{di}) K_s * K_t \left( \frac{X_{di} - X_{dj}}{h_d} \right). \quad (\text{A.21})$$

It follows from (A.15) and (A.21) that for  $i \neq j$

$$(\mathbf{S}_d^T \mathbf{S}_d)_{ij} \approx \frac{1}{nh_d} \sum_{s,t=1}^{p_d+1} \tilde{s}_{d,s} \tilde{s}_{d,t} f_d^{-1}(X_{di}) K_s * K_t \left( \frac{X_{di} - X_{dj}}{h_d} \right). \quad (\text{A.22})$$

By the definition of  $L_{ndd1}$  and (A.21), we have

$$L_{ndd1} \approx \frac{2(n-2)}{n^2} \sum_{i < j} \varepsilon_i \varepsilon_j \sum_{s,t=1}^{p_d+1} \tilde{s}_{d,s} \tilde{s}_{d,t} \frac{1}{h_d} f_d^{-1}(X_{di}) K_s * K_t \left( \frac{X_{dj} - X_{di}}{h_d} \right). \quad (\text{A.23})$$

Observing that  $L_{ndd2} = O_p\left(\frac{1}{nh_d \sqrt{h_d}}\right) = o_p(h_d^{-1})$ , by (A.16) and (A.21) we obtain

$$\sum_{i \neq j} \varepsilon_i \varepsilon_j (\mathbf{S}_d^T \mathbf{S}_d)_{ij} \approx \frac{2(n-2)}{n^2} \sum_{i < j} \varepsilon_i \varepsilon_j \sum_{s,t=1}^{p_d+1} \tilde{s}_{d,s} \tilde{s}_{d,t} \frac{1}{h_d} f_d^{-1}(X_{di}) K_s * K_t \left( \frac{X_{dj} - X_{di}}{h_d} \right) + o_p(h_d^{-1}). \quad (\text{A.24})$$

By (A.22) and the same argument as that for (A.13), we obtain

$$\sum_{j=1}^n \varepsilon_j^2 (\mathbf{S}_d^T \mathbf{S}_d)_{jj} \approx \frac{\sigma^2}{h_d} |\Omega_d| \sum_{s,t=1}^{p_d+1} \tilde{s}_{d,s} \tilde{s}_{d,t} K_s * K_t(0). \quad (\text{A.25})$$

Applying Lemma 3, we obtain  $\varepsilon^T \mathbf{R}_n \varepsilon = o_p(h_D^{-1})$  and

$$\varepsilon^T [\mathbf{A}_{n1} - \mathbf{A}_{n2}] \varepsilon = 2\varepsilon^T \mathbf{S}_D \varepsilon - \varepsilon^T (\mathbf{S}_D^T \mathbf{S}_D) \varepsilon - 2\varepsilon^T \left( \sum_{d=1}^{D-1} \mathbf{S}_d \right)^T \mathbf{S}_D \varepsilon + o_p(h_D^{-1}).$$

This, together with (A.13), (A.14), (A.18), (A.20), (A.24) and (A.25) entails (A.10).

**(ii). Asymptotic normality of  $W_{(n)}$ :**

Denote by  $G(x) = 2 \sum_{t=1}^{p_D+1} \tilde{s}_{D,t} K_t(x) - \sum_{s,t=1}^{p_D+1} \tilde{s}_{D,s} \tilde{s}_{D,t} K_s * K_t(x)$ . Then by the definition of  $W_{(n)}$  and direct computation

$$\text{Var}[W_{(n)} | \mathcal{X}] = 4\sigma^4 \frac{1}{n^2 h_D^2} \sum_{i < j} [f^{-1}(X_{Di}) G\left(\frac{X_{Dj} - X_{Di}}{h_D}\right)]^2 \equiv 4\sigma^4 \sigma_n^{*2}.$$

Applying Proposition 3.2 of de Jong (1987), we obtain

$$\frac{1}{2\sigma^2}\sigma_n^{*-1}W_{(n)}|\mathcal{X} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Note that

$$\begin{aligned} \sigma_n^{*2} &\approx \frac{1}{2h_D^2}E[f^{-1}(X_{D1})G(\frac{X_{D2} - X_{D1}}{h_D})]^2 \\ &\approx 2h_D^{-1}|\Omega_D| \cdot \left\| \sum_{t=1}^{p_D+1} \tilde{s}_{D,t}K_t - \frac{1}{2} \sum_{s,t=1}^{p_D+1} \tilde{s}_{D,s}\tilde{s}_{D,t}K_s * K_t \right\|_2^2 \equiv \sigma_n^2. \end{aligned}$$

It follows that conditional on  $\mathcal{X}$

$$\frac{1}{2\sigma^2}\sigma_n^{-1}W_{(n)} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1). \quad (\text{A.26})$$

(iii). **Asymptotic expression**  $\text{RSS}_1/n = \sigma^2 + o_p(1)$ : By the definition of  $\text{RSS}_1$ , we have

$$\begin{aligned} \text{RSS}_1 &= \varepsilon^T \mathbf{A}_{n2}\varepsilon + \mathbf{m}^T \mathbf{A}_{n2}\mathbf{m} + 2\varepsilon^T \mathbf{A}_{n2}\mathbf{m} \\ &= \varepsilon^T \mathbf{A}_{n2}\varepsilon + \mathbf{B}^T \mathbf{B} + 2\mathbf{B}^T (\mathbf{W}_M - \mathbf{I}_n)\varepsilon. \end{aligned}$$

Referring to the results in the proof of Lemma 6, we obtain

$$\text{RSS}_1/n = n^{-1}\varepsilon^T \mathbf{A}_{n2}\varepsilon + o_p(1).$$

It remains to show that  $n^{-1}\varepsilon^T \mathbf{A}_{n2}\varepsilon = \sigma^2 + o_p(1)$ . Note that from the proof of Lemma 3

$$\mathbf{A}_{n2} = \mathbf{I}_n + \mathbf{S}^T \mathbf{S} - \mathbf{S} - \mathbf{S}^T + \mathbf{R}_{n2},$$

where  $\mathbf{R}_{n2} = O(\frac{11^T}{n})$ , uniformly over all elements of the matrix. By using an argument similar to that for (A.10), we can obtain

$$\begin{aligned} n^{-1}\varepsilon^T \mathbf{A}_{n2}\varepsilon &= n^{-1}\varepsilon^T \mathbf{I}_n\varepsilon + o_p(1) \\ &= \sigma^2 + o_p(1). \end{aligned}$$

(iv) **Conclusion**: By (iii), (A.11) and the definition of  $\lambda_n(H_0)$ , we have

$$\lambda_n(H_0) - \mu_n - \frac{1}{2\sigma^2}d_{1n} + o_p(h_D^{-1}) \approx \frac{1}{2\sigma^2}W_{(n)}. \quad (\text{A.27})$$

The combination of (A.27) and (A.26) leads to

$$P\{\sigma_n^{-1}(\lambda_n(H_0) - \mu_n - \frac{1}{2\sigma^2}d_{1n}) < t|\mathcal{X}\} \xrightarrow{\mathcal{L}} \Phi(t),$$

which reduces to the first result of the theorem. If  $nh_d^{2(p_d+1)}h_D \rightarrow 0$  for  $d = 1, \dots, D$ , then  $d_{1n} = o_p(h_D^{-1})$ , which is dominated by  $\mu_n$ . Then  $r_K\lambda_n(H_0)|\mathcal{X} \stackrel{a}{\sim} \chi_{r_K\mu_n}^2$ .

**Proof of Theorem 3.2.** By the same argument as in Theorem 3.1.

**Proof of Theorem 3.3.** Let  $m_{\theta_0}(x_1, \dots, x_D)$  denote the true function of  $m(x_1, \dots, x_D)$ . Then the *GLR* statistic  $\lambda_n(\mathcal{M}_\Theta)$  for testing problem (3.4) can be decomposed as

$$\lambda_n(\mathcal{M}_\Theta) = \lambda_n(m_{\theta_0}) - \lambda_n^*(\theta), \quad (\text{A.28})$$

where  $\lambda_n(m_{\theta_0})$  is the *GLR* statistic for the fabricated testing problem with the simple null hypothesis

$$H'_0 : m(x_1, \dots, x_D) = m_{\theta_0}(x_1, \dots, x_D) \longleftrightarrow H_1 : m \neq m_{\theta_0}(x_1, \dots, x_D)$$

and  $\lambda_n^*(\theta)$  is the *GLR* statistic for another fabricated testing problem with simple null hypothesis

$$H'_0 : m(x_1, \dots, x_D) = m_{\theta_0}(x_1, \dots, x_D) \longleftrightarrow H'_1 : m(x_1, \dots, x_D) \in \mathcal{M}_\Theta.$$

By the standard parametric hypothesis theory, the second term in (A.28) is  $o_p(h_D^{-\frac{1}{2}})$ , which is entailed by Condition B. This term is negligible in the asymptotic distribution. Hence by the same argument as that for Theorem 3.1 the result of the theorem holds.

**Lemma 7** *Suppose Condition (A) in Appendix I holds and that  $nh_D^{2p_D+3} \rightarrow 0$ . Then under  $H_{1n}$ , there exists a  $\lambda_0 > 0$  such that*

$$d_{2n} \equiv \mathbf{G}_n^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n)^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n) \mathbf{G}_n \geq \lambda_0 \sum_{i=1}^n G_n^2(X_{Di}) + o(h_D^{-1}).$$

**Proof.** Note that by (2.4)

$$\mathbf{S}_D^* (\mathbf{I}_n - \mathbf{W}_M^{[-D]}) = (\mathbf{I}_n - \mathbf{S}_D^* \mathbf{W}_M^{[-D]}) \mathbf{W}_D$$

and by (A.5)

$$\mathbf{W}_D \mathbf{G}_n = \mathbf{G}_n - (\mathbf{I}_n - \mathbf{S}_D^* \mathbf{W}_M^{[-D]})^{-1} [\bar{G}_n \mathbf{1} - \frac{1}{(p_D + 1)!} \mathbf{Q}_D^*] + o(\mathbf{h}_D^{p_D+1}) \quad a.s..$$

Then by Lemma 4

$$\mathbf{S}_D^* (\mathbf{I}_n - \mathbf{W}_M^{[-D]}) \mathbf{G}_n = (\mathbf{I}_n - \mathbf{S}_D^* \mathbf{W}_M^{[-D]}) \mathbf{G}_n + O(\mathbf{h}_D^{p_D+1}) + o\left(\frac{1}{\sqrt{n}}\right) \quad a.s..$$

This entails that

$$\begin{aligned} d_{2n}^* &\equiv \mathbf{G}_n^T (\mathbf{I}_n - \mathbf{W}_M^{[-D]})^T \mathbf{S}_D^{*T} \mathbf{S}_D^* (\mathbf{I}_n - \mathbf{W}_M^{[-D]}) \mathbf{G}_n \\ &= \mathbf{G}_n^T (\mathbf{I}_n - \mathbf{S}_D^* \mathbf{W}_M^{[-D]})^T (\mathbf{I}_n - \mathbf{S}_D^* \mathbf{W}_M^{[-D]}) \mathbf{G}_n + o(h_D^{-1}) \quad a.s.. \end{aligned}$$

As  $\|\mathbf{S}_D^* \mathbf{W}_M^{[-D]}\| < 1$  when  $n$  is large enough, the matrix  $(\mathbf{I}_n - \mathbf{S}_D^* \mathbf{W}_M^{[-D]})^T (\mathbf{I}_n - \mathbf{S}_D^* \mathbf{W}_M^{[-D]})$  is positive definite. Denote its minimum eigenvalue by  $\lambda_0 (> 0)$ . Then,

$$d_{2n}^* \geq \lambda_0 \sum_{i=1}^n G_n^2(X_{Di}) + o(h_D^{-1}).$$



Using (A.12), for any  $n \times 1$  scalar vector  $\mathbf{Z}$ , we have  $\mathbf{S}_D^* \mathbf{Z} \approx \bar{Z} \mathbf{1}$ , which implies that

$$\mathbf{Z}^T \mathbf{S}_D^{*T} \mathbf{S}_D^* \mathbf{Z} \approx n \bar{Z}^2.$$

Hence,

$$\begin{aligned} \|\mathbf{S}_D^*\|_2 &\equiv \sup_{\|\mathbf{Z}\|=1} \sqrt{\mathbf{Z}^T \mathbf{S}_D^* \mathbf{S}_D^* \mathbf{Z}} \\ &\approx \sup_{\|\mathbf{Z}\|=1} \sqrt{n \bar{Z}^2} \\ &\leq 1, \end{aligned}$$

by using the Cauchy-Schwartz inequality. Therefore, the matrix  $\mathbf{I}_n - \mathbf{S}_D^{*T} \mathbf{S}_D^*$  is asymptotically non-negative definite, and its eigenvalues are in  $[0, 1]$ . It follows that

$$\begin{aligned} d_{2n} &= d_{2n}^* + \mathbf{G}_n^T (\mathbf{I}_n - \mathbf{W}_M^{[-D]})^T (\mathbf{I}_n - \mathbf{S}_D^{*T} \mathbf{S}_D^*) (\mathbf{I}_n - \mathbf{W}_M^{[-D]}) \mathbf{G}_n \\ &\geq \lambda_0 \sum_{i=1}^n G_n^2(X_{Di}) + o(h_D^{-1}). \end{aligned} \quad (\text{A.29})$$

Furthermore, if  $h_D \sum_{i=1}^n G_n^2(X_{Di}) = O(1)$  a.s., then by direct but tedious algebra, we obtain

$$d_{2n} = O\left(\sum_{i=1}^n G_n^2(X_{Di})\right) = O(h_D^{-1}). \quad (\text{A.30})$$

**Proof of Theorem 3.4.** Write

$$\begin{aligned} RSS_0 - RSS_1 &= \mathbf{Y}^T [\mathbf{A}_{n1} - \mathbf{A}_{n2}] \mathbf{Y} \\ &= \varepsilon^T [\mathbf{A}_{n1} - \mathbf{A}_{n2}] \varepsilon + 2\varepsilon^T [\mathbf{A}_{n1} - \mathbf{A}_{n2}] \mathbf{m} + \mathbf{m}^T [\mathbf{A}_{n1} - \mathbf{A}_{n2}] \mathbf{m} \\ &\equiv I_{n1} + I_{n2} + I_{n3}. \end{aligned} \quad (\text{A.31})$$

Under  $H_{1n}$ , by the definition of  $\mathbf{B}$  and  $\mathbf{B}_{-D}$ ,

$$I_{n3} = \mathbf{B}_{-D}^T \mathbf{B}_{-D} - \mathbf{B}^T \mathbf{B} + 2\mathbf{B}_{-D}^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n) \mathbf{G}_n + \mathbf{G}_n^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n)^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n) \mathbf{G}_n.$$

Note that from Lemma 5 both  $\mathbf{B}$  and  $\mathbf{B}_{-D}$  are of order  $O_p(\sum_{d=1}^D \mathbf{h}_d^{p_d+1} + \frac{1}{\sqrt{n}})$ . It follows that

$$I_{n3} = 2\mathbf{B}_{-D}^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n) \mathbf{G}_n + \mathbf{G}_n^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n)^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n) \mathbf{G}_n + O_p\left(1 + \sum_{d=1}^D n h_d^{2(p_d+1)}\right). \quad (\text{A.32})$$

By the definitions of  $\mathbf{A}_{n1}$  and  $\mathbf{A}_{n2}$  in Lemma 3 and the result in the proof of Lemma 6, we obtain

$$\begin{aligned} I_{n2} &= 2\varepsilon^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n)^T \mathbf{B}_{-D} + 2\varepsilon^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n)^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n) \mathbf{G}_n - 2\mathbf{B}^T (\mathbf{W}_M - \mathbf{I}_n) \varepsilon \\ &= 2\varepsilon^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n)^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n) \mathbf{G}_n + O_p\left(1 + \sum_{d=1}^D \sqrt{n} h_d^{p_d+1}\right). \end{aligned} \quad (\text{A.33})$$

The combination of (A.31) with (A.32) and (A.33) leads to

$$\begin{aligned}
RSS_0 - RSS_1 &= I_{n1} + \varepsilon^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n)^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n) \mathbf{G}_n + 2\mathbf{B}_{-D}^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n) \mathbf{G}_n \\
&\quad + \mathbf{G}_n^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n)^T (\mathbf{W}_M^{[-D]} - \mathbf{I}_n) \mathbf{G}_n + O_p \left( 1 + \sum_{d=1}^D \sqrt{n} h_d^{p_d+1} + \sum_{d=1}^D n h_d^{2(p_d+1)} \right) \\
&\equiv I_{n1} + C_n + D_n + d_{2n} + O_p \left( 1 + \sum_{d=1}^D \sqrt{n} h_d^{p_d+1} + \sum_{d=1}^D n h_d^{2(p_d+1)} \right). \tag{A.34}
\end{aligned}$$

We now assess each of the above terms. Note that by (A.10)

$$I_{n1} = W_{(n)} + 2\sigma^2 \mu_n + o_p(h_D^{-1}). \tag{A.35}$$

Using (A.3), we obtain

$$D_n = 2\mathbf{B}_{-D}^T \mathbf{S}^{[-D]} \mathbf{G}_n - 2\mathbf{B}_{-D}^T \mathbf{G}_n + 2\mathbf{B}_{-D}^T \mathbf{U}^{[-D]} \mathbf{G}_n.$$

By the Cauchy-Schwartz inequality, we have

$$|\mathbf{B}_{-D}^T \mathbf{G}_n| \leq \|\mathbf{B}_{-D}\| \|\mathbf{G}_n\| = o_p(1)$$

and

$$\mathbf{B}_{-D}^T \mathbf{U}^{[-D]} \mathbf{G}_n = O(n^{-1}) \sum_{i=1}^n |(B_{-D})_i| \sum_{j=1}^n |G_n(X_{Dj})| = o_p(h_D^{-1}).$$

Using (A.12), one gets

$$\mathbf{B}_{-D}^T \mathbf{S}^{[-D]} \mathbf{G}_n = o_p(h_D^{-1}).$$

Hence,

$$D_n = o_p(h_D^{-1}). \tag{A.36}$$

Observing that both  $\mathbf{S}_d$  and  $\mathbf{S}_{d'}^T \mathbf{S}_d$  (for  $d, d' \neq D$ ) are of order  $\mathbf{R}_{3n} \equiv O(\frac{11^T}{nh_D})$ , we obtain from (A.3) that

$$d_{2n} = \mathbf{G}_n^T \mathbf{G}_n + \mathbf{G}_n^T \mathbf{R}_{3n} \mathbf{G}_n.$$

Note that  $\mathbf{R}_{3n}$  does not involve  $X_{D1}, \dots, X_{Dn}$ . By conditioning argument and directly computing the mean and variance, the second term above is  $o_p(h_D^{-1})$ . Hence,

$$d_{2n} = \sum_{i=1}^n G_n^2(X_{Di}) + o_p(h_D^{-1}). \tag{A.37}$$

Similarly, we have

$$C_n = \varepsilon^T \mathbf{G}_n + \varepsilon^T \mathbf{R}_{3n} \mathbf{G}_n = \varepsilon^T \mathbf{G}_n + o_p(h_D^{-1}),$$

which conditional on  $\mathcal{X}$  is asymptotically identically distributed as  $\mathcal{N}(0, \frac{C(G)}{h_D})$ . This together with (A.26) and (A.34)-(A.37) yields the result of the theorem.

**Proof of Theorem 3.5.** The argument used here is similar to that for Theorem 8 of Fan, Zhang and Zhang (2001), but the technical details are much more complex. Under  $H_{1n} : m_D(X_D) = G_n(X_D)$  and under Condition (A), it follows from (A.34), (A.35), and (A.36) that for  $h_D \rightarrow 0$ ,

$$-\lambda_n(H_0)\sigma^2 = -\mu_n\sigma^2(1 + o_p(1)) - W_{(n)}/2 - d_{2n}/2 + O_p(1 + \sum_{d=1}^D \sqrt{n}h_d^{p_d+1} + \sum_{d=1}^D nh_d^{2(p_d+1)}) - C_n/2,$$

uniformly in  $G_n \in \mathcal{G}_n$ . Thus, by definition

$$\begin{aligned} \beta(\alpha, G_n) &= P\{\sigma_n^{-1}(-\lambda_n(H_0) + \mu_n) \geq z_\alpha | \mathcal{X}\} \\ &= P\left\{\sigma_n^{-1}\left[-\frac{W_{(n)}}{2\sigma^2} - \frac{d_{2n}}{2\sigma^2} - \frac{C_n}{2\sigma^2} + O_p\left(1 + \sum_{d=1}^D \sqrt{n}h_d^{p_d+1} + \sum_{d=1}^D nh_d^{2(p_d+1)}\right)\right] \geq z_\alpha \mid \mathcal{X}\right\} \\ &= P_{1n} + P_{2n} \end{aligned}$$

with

$$\begin{aligned} P_{1n} &= P\left\{\sigma_n^{-1}\left(-\frac{W_{(n)}}{2\sigma^2}\right) + \sqrt{nh_D} \frac{2p_D+3}{2} b_{1n} + nh_D \frac{4p_D+5}{2} b_{2n} - \sqrt{h_D} b_{3n} \geq z_\alpha, |b_{1n}| \leq M, |b_{2n}| \leq M \mid \mathcal{X}\right\}, \\ P_{2n} &= P\left\{\sigma_n^{-1}\left(-\frac{W_{(n)}}{2\sigma^2}\right) + \sqrt{nh_D} \frac{2p_D+3}{2} b_{1n} + nh_D \frac{4p_D+5}{2} b_{2n} - \sqrt{h_D} b_{3n} \geq z_\alpha, |b_{1n}| \geq M, |b_{2n}| \geq M \mid \mathcal{X}\right\}, \end{aligned}$$

and

$$\begin{aligned} b_{1n} &= (\sqrt{nh_D} \frac{2p_D+3}{2} \sigma_n)^{-1} O_p(1 + \sum_{d=1}^D \sqrt{n}h_d^{p_d+1}) = O_p(1), \\ b_{2n} &= (nh_D \frac{4p_D+5}{2} \sigma_n)^{-1} O_p(\sum_{d=1}^D nh_d^{2(p_d+1)}) = O_p(1), \\ b_{3n} &= (\sqrt{h_D} \sigma_n \sigma^2)^{-1} \frac{1}{2} [d_{2n} + C_n]. \end{aligned}$$

Note that  $E[C_n | \mathcal{X}] = 0$  and  $Var(C_n | \mathcal{X}) = \sigma^2 \mathbf{G}_n^T \mathbf{A}_{n1}^T \mathbf{A}_{n1} \mathbf{G}_n = \sigma^2 \mathbf{G}_n O(\frac{\mathbf{1}\mathbf{1}^T}{n}) \mathbf{G}_n = O(\sum_{i=1}^n G_n^2(X_{Di}))$ . Hence  $C_n = O_p(\sqrt{d_{2n}})$ . This together with (A.29) and (A.30) in Lemma 7 yield

$$\sqrt{h_D} b_{3n} \rightarrow \infty \text{ only when } n\sqrt{h_D} \rho^2 \rightarrow \infty.$$

When  $h_D \leq c_0^{-\frac{1}{p_D+1}} n^{-\frac{1}{2(p_D+1)}}$ , we have  $\sqrt{nh_D} \frac{2p_D+3}{2} \geq c_0 nh_D \frac{4p_D+5}{2}$ ,  $\sqrt{nh_D} \frac{2p_D+3}{2} \rightarrow 0$ , and  $nh_D \frac{4p_D+5}{2} \rightarrow 0$ . Thus for  $h_D \rightarrow 0$  and  $nh_D \rightarrow \infty$ , it follows that  $\beta(\alpha, \rho) \rightarrow 0$  only when  $n\sqrt{h_D} \rho^2 \rightarrow +\infty$ . It implies that  $\rho_n^2 = n^{-1} h_D^{-\frac{1}{2}}$ , and the possible minimum value of  $\rho_n$  in this setting is  $n^{-\frac{4p_D+3}{8(p_D+1)}}$ . When  $nh_D^{2(p_D+1)} \rightarrow \infty$ , for any  $\delta > 0$ , there exists a constant  $M > 0$  such that  $P_{2n} < \frac{\delta}{2}$  uniformly in  $G_n \in \mathcal{G}_n$ . Then

$$\beta(\alpha, \rho) \leq \frac{\delta}{2} + P_{1n}.$$

Note that  $\sup_{G_n(\rho)} P_{1n} \rightarrow 0$  only when  $B(h_D) \equiv nh_D \frac{4p_D+5}{2} M - nh_D \frac{1}{2} \rho^2 \rightarrow -\infty$ . Since  $B(h_D)$  attains the minimum value  $-\frac{4(p_D+1)}{4p_D+5} [(4p_D+5)M]^{-\frac{1}{4(p_D+1)}} n\rho^{\frac{4p_D+5}{2(p_D+1)}}$  at  $h_D = [\rho^2 / ((4p_D+5)M)]^{\frac{1}{2(p_D+1)}}$ .

Now simple algebra shows that in this setting the corresponding minimum value of  $\rho_n$  is  $n^{-\frac{2(p_D+1)}{4p_D+5}}$  with  $h_D = c_* n^{-\frac{2}{4p_D+5}}$  for some constant  $c_*$ .

**Proof of Theorem 4.1.** Using exactly the same argument as that for the proof of Theorem 3.3, and noticing that from Lemma 6 the  $d_{1n} = O(1)$  in the current situation, we obtain the result of the theorem.

**Proof of Theorem 4.2.** Let  $RSS_0^*$  and  $RSS_1^*$  be similarly defined as  $RSS_0$  and  $RSS_1$  respectively, based on a bootstrap sample  $\{\mathbf{X}_i, Y_i^*\}_{i=1}^n$ . We use the superscript  $*$  of a quantity as its bootstrap analog. Then

$$\lambda_n^*(H_0) \approx \frac{n}{2} \frac{RSS_0^* - RSS_1^*}{RSS_1^*}.$$

It can be shown that under  $H_0$  for given bandwidths satisfying Condition (A)

$$P\{\sigma_n^{-1}(\lambda_n^*(H_0) - \mu_n - d_{1n}) < t | \mathcal{X}, F_n\} \xrightarrow{\mathcal{L}} \Phi(t), \quad (\text{A.38})$$

which will be proven through the following three steps:

(i) Note that  $\mathbf{Y}^* = \hat{\mathbf{m}}_{(-D)} + \hat{\varepsilon}^*$ , it follows that

$$\begin{aligned} RSS_0^* - RSS_1^* &= \hat{\varepsilon}^{*T}(A_{n1} - A_{n2})\hat{\varepsilon}^* + [\hat{\mathbf{m}}_{(-D)}(A_{n1} - A_{n2})\hat{\mathbf{m}}_{(-D)} + 2\hat{\varepsilon}^{*T}(A_{n1} - A_{n2})\hat{\mathbf{m}}_{(-D)}] \\ &\equiv \hat{\varepsilon}^{*T}(A_{n1} - A_{n2})\hat{\varepsilon}^* + d_{1n}^*. \end{aligned}$$

(ii) Using the same argument as that for (A.11), conditional on  $F_n$  we have

$$RSS_0^* - RSS_1^* \approx W_{(n)}^* + 2\sigma^2\mu_n + d_{1n}^* + o_p(h_D^{-1}),$$

where  $W_{(n)}^*$  is similarly defined as  $W_{(n)}$  but with  $\varepsilon_i$  replaced by  $\hat{\varepsilon}_i^*$ , and by argument similar to that for Lemma 6 [note that  $\hat{\mathbf{m}}_{(-D)} = \mathbf{m}_0(1 + o_p(1))$ ]

$$d_{1n}^* \approx \mathbf{m}_0(A_{n1} - A_{n2})\mathbf{m}_0 + 2\hat{\varepsilon}^{*T}(A_{n1} - A_{n2})\mathbf{m}_0 \approx O_p(1 + \sum_d n h_d^{2(p_d+1)} + \sum_d \sqrt{n} h_d^{p_d+1}).$$

(iii) Note that  $RSS_1^*/n \approx \sigma^2$ ,  $E[\hat{\varepsilon}_i^* | F_n] = 0$ ,  $E[\hat{\varepsilon}_i^{*2} | F_n] = \sigma^2$ , and

$$\text{Var}[W_{(n)}^* | F_n] = 4\sigma^4 \frac{1}{n^2 h_D^2} \sum_{i < j} [f^{-1}(X_{Di})G(\frac{X_{Dj} - X_{Di}}{h_D})]^2 \approx 4\sigma^4 \sigma_n^2,$$

where  $G(\cdot)$  is defined in the proof of Theorem 3.1. Then applying Proposition 3.2 of de Jong (1987), one gets

$$\frac{1}{2\sigma^2} \sigma_n^{-1} W_{(n)}^* \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Combination of (i)-(iii) yields (A.38). Note that  $\hat{h}_d$ ,  $d = 1, \dots, D$  satisfy the bandwidth restriction in Condition (A). The consistency of the bootstrap estimate of the conditional null distribution is obtained.

## REFERENCES

- Aerts, M., Claeskens, G. & Hart, J.D. (1999), "Testing the fit of a parametric function," *Jour. Amer. Statist. Assoc.*, **94**, 869–879.
- Amato, U. and Antoniadis, A. (2001), "Adaptive wavelet series estimation in separable nonparametric regression models," *Statist. Comput.*, **11**, 373–394.
- Amato, U., Antoniadis, A. and De Feis, I. (2002), "Fourier series approximation of separable models," *J. Computat. Appl. Math.*, **146**, 459–479.
- Ansley, C. F., and Kohn, R. (1994), "Convergence of the Backfitting Algorithm for Additive Models," *Journal of the Australian Mathematical Society (Series A)*, **57**, 316 – 329.
- Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: Wiley.
- Breiman, L., and Friedman, J.H. (1985), "Estimating Optimal Transformations for Multiple Regression and Correlation," *Journal of the American Statistical Association*, **80**, 580 – 619
- Buja, A., Hastie, T.J., and Tibshirani, R. J. (1989), "Linear Smoothers and Additive Models," *The Annals of Statistics*, **17**, 453 – 555.
- Cleveland, W.S. (1979), "Robust locally weighted regression and smoothing scatterplots", *J. Amer. Statist. Assoc.*, **74**, 829–836.
- Cook, R.D. and Weisberg, S. (1982), *Residuals and Influence in Regression*, New York: Chapman and Hall.
- Delgado, M.A. and González-Manteiga, W. (2001), "Significance testing in nonparametric regression based on the bootstrap," *Ann. Statist.*, **29**, 1469-1507.
- de Jong, P. (1987), "A Central Limit Theorem for Generalized Quadratic Forms," *Probability Theory and Its Related Fields*, **75**, 261 – 277.
- Dette, H. (1999), "A consistent test for the functional form of a regression based on a difference of variance estimators," *Ann. Statist.*, **27**, 1012-1040.
- Fan, J., and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, New York: Chapman & Hall.
- Fan, J., Härdle, W. and Mammen, E. (1998), "Direct estimation of additive and linear components for high dimensional data," *The Annals of Statistics*, **26**, 943-971.
- Fan, J. and Huang, L. (2001), "Goodness-of-fit test for parametric regression models," *Jour. Amer. Statist. Assoc.*, **96**, 640–652.
- Fan, J., Zhang, C.M., and Zhang, J. (2001), "Generalized Likelihood Ratio Statistics and Wilks Phenomenon," *The Annals of Statistics*, **29**, 153 – 193.
- Fan, J., and Zhang, W. (2004), "Generalized likelihood ratio tests for spectral density," *Biometrika*, **91**, 195-209.
- Friedman, J. H., and Stuetzle, W. (1981), "Projection Pursuit Regression," *Journal of the American Statistical Association*, **76**, 817 – 823.
- Glad, I.K. (1998), "Parametrically Guided Non-parametric Regression," *Scandinavian Journal of Statistics*, **25**, 649 – 668.

- Gozaló, P. L., and Linton, O. B. (2001), "Testing additivity in generalized nonparametric regression models with estimated parameters," *Journal of Econometrics*, **104**, 1–48.
- Härdle, W., and Hall, P. (1993), "On the Backfitting Algorithm for Additive Regression Models," *Statist. Neerlandica*, **47**, 43 – 57.
- Härdle, W., and Mammen, E. (1993), "Comparing Nonparametric Versus Parametric Regression Fits," *The Annals of Statistics*, **21**, 1926 – 1947.
- Härdle, W., Sperlich, S., and Spokoiny, V. (2001), "Structural Tests in Additive Regression," *Journal of the American Statistical Association*, **96**, 1333 – 1347.
- Harrison, D., and Rubinfeld, D.L. (1978), "Hedonic Housing Prices and the Demand for Clean Air," *Journal of Economics and Management*, **5**, 81 – 102.
- Hastie, T. J., and Tibshirani, R. J. (1990), *Generalized Additive Models*, New York: Chapman and Hall.
- Hastie, T. J., and Tibshirani, R. J. (2000), "Bayesian backfitting (with discussion)," *Statist. Sci.*, **15**, 196–223.
- Hjort, N.L., and Glad, I.K. (1995), "Nonparametric density estimation with a parametric start," *The Annals of Statistics*, **23**, 882 – 904.
- Ingster, Yu. I. (1993), "Asymptotic Minimax Hypothesis Testing for Nonparametric Alternatives I-III," *Math. Methods Statist.*, **2**, 85 – 114; **3**, 171 – 189; **4**, 249 – 268.
- Le Cam, L. and Yang, G.L. (1990), *Asymptotic in Statistics: Some Basic Concepts*, Springer-Verlag, New York.
- Linton, O. B. (1997), "Efficient estimation of additive nonparametric regression models," *Biometrika*, **84**, 469-473.
- Linton, O. B., and Nielsen, J. P. (1995), "A Kernel Method of Estimating Regressing Structured Nonparametric Regression Based on Marginal Integration," *Biometrika*, **82**, 93 – 100.
- Mammen, E., Linton, O., and Nielsen, J. (1999), "The existence and asymptotic properties of a backfitting projection algorithm under weak conditions," *Ann. Statist.*, **27**, 1443–1490.
- Neumeyer, N. and Sperlich, S. (2003), "Comparison of separable components in different samples," manuscript.
- Opsomer, J.-D., and Ruppert D. (1997), "Fitting a Bivariate Additive Model by Local Polynomial Regression," *The Annals of Statistics*, **25**, 186 – 211.
- Opsomer, J.-D., and Ruppert D. (1998), "A Fully Automated Bandwidth Selection Method for Fitting Additive Models," *Journal of the American Statistical Association*, **93**, 605 – 619.
- Opsomer, J.-D. (2000), "Asymptotic Properties of Backfitting Estimators," *Journal of Multivariate Analysis*, **73**, 166 – 179.
- Press, H. and Tukey, J.W. (1956), *Power Spectral Methods of Analysis and Their Application to Problems in Airplane Dynamics*. Bell Telephone System Monograph 2606.
- Ruppert, Wand and Carroll (2003), *Semiparametric Regression*, Cambridge: Cambridge University Press.

- Sardy, S., Bruce, A.G. and Tseng, P. (2000), "Block coordinate relaxation methods for nonparametric wavelet denoising," *J. Comp. Graph. Statist.*, **9**, 361–379.
- Sardy, S. and Tseng, P. (2003), "Automatic nonlinear fitting of additive and generalised additive models with wavelets," *J. Comp. Graph. Statist.*, **12** (to appear).
- Sperlich, S., Linton, O.B. and Härdle, W. (1999), "Integration and backfitting methods in additive models – finite sample properties and comparison," *Test*, **8**, 419–458.
- Spokoiny, V. G. (1996), "Adaptive Hypothesis Testing Using Wavelets," *Ann. Statist.*, **24**, 2477 – 2498.
- Stone, C. J. (1985), "Additive Regression and Other Nonparametric Models," *The Annals of Statistics*, **13**, 689 – 705.
- Tjøstheim, D., and Auestad, B. (1994), "Nonparametric Identification of Nonlinear Time Series: Projection," *Journal of the American Statistical Association*, **89**, 1398 – 1409.
- Wand, M.P. (1999), "A central limit theorem for local polynomial backfitting estimators," *J. Multivariate Anal.*, **70**, 57–65.
- Zhang, C.M. (2003), "Calibrating the degrees of freedom for automatic data smoothing and effective curve checking," *Jour. Amer. Statist. Assoc.*, **98**, 609–628.