Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

Nonparametric Inferences for Additive Models

Jianqing Fan & Jiancheng Jiang

To cite this article: Jianqing Fan & Jiancheng Jiang (2005) Nonparametric Inferences for Additive Models, Journal of the American Statistical Association, 100:471, 890-907, DOI: 10.1198/016214504000001439

To link to this article: https://doi.org/10.1198/016214504000001439

Nonparametric Inferences for Additive Models

Jianqing FAN and Jiancheng JIANG

Additive models with backfitting algorithms are popular multivariate nonparametric fitting techniques. However, the inferences of the models have not been very well 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 (GLR) tests to additive models, using the backfitting estimator. We demonstrate that under the null models, the newly proposed GLR statistics follow asymptotically rescaled chi-squared 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 GLR 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 GLR. 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.

1. INTRODUCTION

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

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

where $\{\varepsilon_i\}$ is a sequence of iid random variables with mean 0 and finite variance σ^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. Because all of the unknown functions are one-dimensional, the difficulty associated with the so-called "curse of dimensionality" is substantially reduced (for details, see Stone 1985; Hastie and Tibshirani 1990). In fact, Fan, Härdle, and Mammen (1998) have shown that an additive component can be estimated as well as in the case where rest of the 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, including 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 et al. (1999), the Fourier series approximation approach of Amato, Antoniadis, and De Feis (2002), the linear wavelet strategies of Amato and Antoniadis (2001), and the nonlinear wavelet estimation method of Sardy and Tseng (2004) using the block coordinate relaxation algorithm of Sardy, Bruce, and Tseng (2000), among others. Among these methods, the backfitting algorithm is considered a useful fitting tool and has received much attention for its easy of implementation. Haïrdle 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 based on a nonprojection 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 Bayesian backfitting, which is a stochastic generalization of the backfitting algorithm discussed earlier. A simulation study comparing the finite-sample properties of backfitting and marginal integration methods was conducted by Sperlich, Linton, and Härdle (1999).

After fitting the additive model via a backfitting algorithm, one often asks whether a specific additive component in (1) is significant or admits a certain parametric form, such as a polynomial function. This amounts to testing whether the additive component is 0 or of a polynomial form. However, only limited tools are available for such kinds of frequently asked questions. Compared with the studies on estimation, the understanding of such testing problems is limited in the additive model. To our knowledge, the literature contains virtually no formal and theoretical work on testing under the present settings. Recently, Härdle, Sperlich, and Spokoiny (2001) used wavelets along with the adaptive Neyman type of idea (Fan and Gijbels 1996) to test additive components. Although this 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

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 (E-mail: *jqfan@princeton.edu*). Jiancheng Jiang is Associate Professor, Department of Probability & Statistics, Peking University, Beijing 100871, China (E-mail: *jiang@math.pku.edu.cn*). Fan was supported in part by RGC grant CUHK 4262/01P of the HKSAR, National Science Foundation grants DMS-03-55179 and DMS-03-54223, and National Institutes of Health grant R01 HL69720, and Jiang was supported by Chinese National Science Foundation grants 10001004 and 39930160.

Journal of the American Statistical Association
September 2005, Vol. 100, No. 471, Theory and Methods
DOI 10.1198/016214504000001439

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 does exist, 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. Fan et al. (2001) 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, because the distribution of ε_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 ε_i; second, it remains unknown whether a particularly constructed GLR test will follow the Wilks' type of results and share certain optimality. In this article we develop GLR tests and their bias-corrected versions for the additive model to address the foregoing questions. This not only will provide useful tools to address frequently asked questions in additive modeling, but also will enrich the GLR test theory. Our results, together with those of Fan et al. (2001), convincingly show 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) 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 are approximated by generalized quadratic forms, technical challenges include deriving quadratic approximations and the distributions of the quadratic functionals with a backfitting estimator. Because the additive model and local polynomial smoother are widely used in multivariate nonparametric modeling, determined efforts have been made in this article 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 prove that, under general assumptions on the error distribution of ε_i, the proposed GLR tests follow the Wilks type of results and have the asymptotic optimality for nonparametric hypothesis testing. In addition, unlike the classical Wilks type of results and their generalization by Fan et al. (2001), the additivity of degrees of freedom does not hold. The additivity property holds in a more generalized sense (see Thm. 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 1). These types of adaptive results are in line with the oracle property given by Fan et al. (1998) and Mammen et al. (1999). Our theoretical results from the proposed GLR tests shed some light 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. In Section 3 we develop the theoretical framework for the GLR tests. We introduce the bias-corrected GLR tests and a conditional bootstrap method for approximating the null distributions of the GLR statistics in Section 4. In Section 5 we demonstrate the performance of GLR tests on simulated data, and in Section 6 we provide an example of testing on a real dataset. We defer technical proofs to Appendix B.

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) requires choosing bandwidths $\{h_d\}$. The optimal choice of h_d can be obtained as was done by Opsomer and Ruppert (1998) and Opsomer (2000). Here we follow notation introduced by Opsomer (2000). Put $K_{h_d}(x) = h_d^{-1}K(\frac{x}{h_d})$, $K_s(v) = v^{s-1}K(v)$, $\mathbf{H}_d = \operatorname{diag}(1, h_d, \ldots, h_d^{p_d})$, $\mathbf{m}_d = \{m_d(X_{d1}), \ldots, m_d(X_{dn})\}^T$, and $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$. The smoothing matrices for local polynomial regression are

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

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

$$\mathbf{s}_{d,x_d}^T = \mathbf{e}_1^T\left(\mathbf{X}_{x_d}^{dT}\mathbf{K}_{x_d}\mathbf{X}_{x_d}^{d}\right)^{-1}\mathbf{X}_{x_d}^{dT}\mathbf{K}_{x_d}, \tag{2}$$

with $\mathbf{e}_i$ as a vector with a 1 in the ith position and 0's elsewhere, the matrix $\mathbf{K}_{x_d} = \operatorname{diag}\{K_{h_d}(X_{d1} - x_d), \ldots, K_{h_d}(X_{dn} - x_d)\}$ 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

$$\begin{aligned} \begin{bmatrix} \widehat{\mathbf{m}}_1 \\ \widehat{\mathbf{m}}_2 \\ \vdots \\ \widehat{\mathbf{m}}_D \end{bmatrix} &= \begin{bmatrix} \mathbf{I}_n & \mathbf{S}_1^* & \cdots & \mathbf{S}_1^* \\ \mathbf{S}_2^* & \mathbf{I}_n & \cdots & \mathbf{S}_2^* \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_D^* & \mathbf{S}_D^* & \cdots & \mathbf{I}_n \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}_1^* \\ \mathbf{S}_2^* \\ \vdots \\ \mathbf{S}_D^* \end{bmatrix} \mathbf{Y} \\ &\equiv \mathbf{M}^{-1}\mathbf{C}\mathbf{Y}, \end{aligned} \tag{3}$$

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

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

$$\mathbf{W}_d = \mathbf{E}_d\mathbf{M}^{-1}\mathbf{C}, \tag{4}$$

where $\mathbf{E}_d$ is a partioned matrix of dimension $n \times nD$ with an $n \times n$ identity matrix as the dth "block" and 0's 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, $Y_i' = \sum_{k=1,\neq d}^{D} m_k(X_{ki}) + \varepsilon_i$. Denote $\mathbf{m} = \sum_{d=1}^{D} \mathbf{m}_d$ 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 $d \in (1, \ldots, D)$ and a matrix norm $\|\cdot\|$, then by lemma 2.1 of Opsomer (2000), the backfitting estimators exist and are unique and

$$\begin{aligned}\mathbf{W}_d &= \mathbf{I}_n - \left(\mathbf{I}_n - \mathbf{S}_d^* \mathbf{W}_M^{[-d]}\right)^{-1} (\mathbf{I}_n - \mathbf{S}_d^*) \\ &= \left(\mathbf{I}_n - \mathbf{S}_d^* \mathbf{W}_M^{[-d]}\right)^{-1} \mathbf{S}_d^* \left(\mathbf{I}_n - \mathbf{W}_M^{[-d]}\right). \end{aligned} \tag{5}$$

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

$$\limsup_n \left\| \mathbf{S}_d^* \mathbf{W}_M^{[-d]} \right\| < 1 \tag{6}$$

for a matrix norm $\|\cdot\|$. In practice, the smoothing operators $\mathbf{S}_1, \ldots, \mathbf{S}_d$ are conducted over compact sets of design densities. Hence we need to deal only with the case where the design densities have bounded support. In the case of $D = 2$, a sufficient condition for (6) 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) of Opsomer and Ruppert (1997). Then, by Lemma B.2 in Appendix B and direct matrix multiplication,

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

where $\|\mathbf{A}\|_r = \max_{1 \le i \le n} \sum_{j=1}^{n} |\mathbf{a}_{ij}|$ denotes the norm of the maximum row sum. However, for $D > 2$, the condition in (6) is not easily replaced with other conditions. In fact, for the backfitting algorithm using any smoothing technique, condition (6) must be satisfied to ensure the existence of the backfitting estimators. Hence we restrict the design points in $\mathcal{X}$.

3. GENERALIZED LIKELIHOOD RATIO TESTS

3.1 The Generalized Likelihood Ratio Test

In this section we define the GLR statistics and develop their asymptotic theory under model (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

$$H_0 : m_D(x_D) = 0 \quad \text{vs.} \quad H_1 : m_D(x_D) \neq 0. \tag{7}$$

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

Because the distribution of ε_i is unknown, we do not have a known likelihood function. Pretending that the error distribution is normal, $\mathcal{N}(0, \sigma^2)$, the log-likelihood under model (1) is

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

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

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

where $RSS_1 = \sum_{k=1}^{n} (Y_k - \hat{\alpha} - \sum_{d=1}^{D} \widehat{m}_d(X_{dk}))^2$. Maximizing over the parameter σ^2, we obtain a likelihood of the alternative model,

$$-\frac{n}{2} \log(2\pi/n) - \frac{n}{2} \log(RSS_1) - \frac{n}{2}.$$

Therefore, up to a constant term, the log-likelihood of model (1) is taken as $\ell(H_1) = -\frac{n}{2} \log(RSS_1)$. Similarly, the log-likelihood for H_0 can be taken as $\ell(H_0) = -\frac{n}{2} \log(RSS_0)$, with $RSS_0 = \sum_{k=1}^{n} (Y_k - \hat{\alpha} - \sum_{d=1}^{D-1} \widetilde{m}_d(X_{dk}))^2$ and $\widetilde{m}_d(x_d)$ as the estimator of $m_d(x_d)$ under H_0, using the same backfitting algorithm and bandwidths. Following Fan et al. (2001), we define the following GLR statistic:

$$\begin{aligned}\lambda_n(H_0) &= [\ell(H_1) - \ell(H_0)] = \frac{n}{2} \log \frac{RSS_0}{RSS_1} \\ &\approx \frac{n}{2} \frac{RSS_0 - RSS_1}{RSS_1}, \end{aligned} \tag{8}$$

which compares the likelihood of the nearly best fitting under 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 $\nu_i = \int u^i K(u)\, du$ for $i = 0, 1, \ldots,$ and $\widetilde{\mathbf{S}}_d = (\nu_{i+j-2})$ for $i, j = 1, \ldots, p_d + 1$, 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 $s, t = 1, 2, \ldots$. Put $\mathbf{c}_{j, p_d+j} = (\nu_j, \ldots, \nu_{p_d+j})^T$, $(\tilde{s}_{d,1}, \ldots, \tilde{s}_{d,p_d+1}) = \mathbf{e}_1^T \widetilde{\mathbf{S}}_d^{-1}$, and $C_d^{(j)} = \mathbf{e}_1^T \widetilde{\mathbf{S}}_d^{-1} \mathbf{c}_{j,p_d+j}$ for $j = 0, \ldots, p_d + 1$ and $d = 1, \ldots, D$. Let

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

$$\sigma_n^2 = \frac{2|\Omega_D|}{h_D} \left\| \sum_{t=1}^{p_D+1} \tilde{s}_{D,t}\, K_t - \frac{1}{2} \sum_{s,t=1}^{p_D+1} \tilde{s}_{D,s} \tilde{s}_{D,t} K_s * K_t \right\|_2^2,$$

and

$$r_K \equiv \frac{2\mu_n}{\sigma_n^2} = \frac{\sum_{t=1}^{p_D+1} \tilde{s}_{D,t}\, K_t(0) - \frac{1}{2} \sum_{s,t=1}^{p_D+1} \tilde{s}_{D,s} \tilde{s}_{D,t}\, K_s * K_t(0)}{\| \sum_{t=1}^{p_D+1} \tilde{s}_{D,t} K_t - \frac{1}{2} \sum_{s,t=1}^{p_D+1} \tilde{s}_{D,s} \tilde{s}_{D,t} K_s * K_t \|_2^2},$$

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 1. Suppose that condition A in Appendix A holds. Then, under H_0 for the testing problem (7),

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

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

$$r_K\lambda_n(H_0) \overset{a}{\sim} \chi^2_{r_K\mu_n}.$$

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

Remark 1. When $K(\cdot)$ is a symmetric density kernel and $p_d = 1$ for $d = 1, \ldots, D$, 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 et al. (2001). Therefore, for the additive model, the GLR test has an oracle property in the sense that although the nuisance functions $m_d(x_d)$'s (for $d = 1, \ldots, D-1$) are unknown, the GLR test behaves as though they were known.

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

The asymptotic null distribution offers a method for determining approximately the critical value of the GLR tests, but one cannot 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 given in Section 4.2. An alternative approximation of the null distribution can be obtained by using a calibration idea of Zhang (2003). When $h_D \to \infty$, the local polynomial fitting becomes a global polynomial fitting. Hence one would expect the degrees of freedom to be p_D. This prompted Zhang to use $\chi^2_{r_K\mu_n+p_D}$ to approximate the null distribution.

Now we consider a little more complicated hypothesis testing problem,

$$H_0 : m_{D-d_0}(x_{D-d_0}) = \cdots = m_D(x_D) = 0 \quad \text{vs.} \quad H_1 : m_{D-d_0}(x_{D-d_0}) \neq 0, \ldots, \quad \text{or} \quad m_D(x_D) \neq 0, \tag{9}$$

for some integer d_0. This generalizes problem (7). Let

$$\mu_n' = \sum_{d'=D-d_0}^{D} \frac{|\Omega_{d'}|}{h_{d'}} \times \left[\sum_{t=1}^{p_{d'}+1} \tilde{s}_{d',t}\, K_t(0) - \frac{1}{2}\sum_{s,t=1}^{p_{d'}+1} \tilde{s}_{d',s}\tilde{s}_{d',t}K_s * K_t(0)\right],$$

$$\sigma_n'^2 = \sum_{d'=D-d_0}^{D} \frac{2|\Omega_{d'}|}{h_{d'}} \left\|\sum_{t=1}^{p_{d'}+1} \tilde{s}_{d',t}K_t - \frac{1}{2}\sum_{s,t=1}^{p_{d'}+1} \tilde{s}_{d',s}\tilde{s}_{d',t}K_s * K_t\right\|_2^2,$$

and $r_K' = 2\mu_n'/\sigma_n'^2$.

Theorem 2. For the hypothesis testing problem (9), under the same conditions as in Theorem 1, the results in Theorem 1 continue to hold but with μ_n, σ_n^2, and r_K replaced by μ_n', $\sigma_n'^2$, and r_K', where the condition $nh_d^{2(p_d+1)}h_D \to 0$ for all d's is replaced by $nh_d^{2(p_d+1)}h_{d'} \to 0$, for all d's and any $d' \in \{D - d_0, \ldots, D\}$.

Interestingly, μ_n' and $\sigma_n'^2$ are the summation of the individual μ_n's and σ_n^2's given in Theorem 1. However, the normalization constant r_K' changes with the testing problem, and the degrees of freedom $r_K'\mu_n'$ are no longer the summation of those for testing individual problems such as (7). These mark the difference from those given by Fan et al. (2001). The result is also different from the case of the degrees of freedom of the fit for an additive penalized spline model (see Ruppert, Wand, and Carroll 2003, sec. 8.3). However, when all p_d's 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:

$$H_0 : m_\theta(x_1, \ldots, x_D) \in \mathcal{M}_\Theta \quad \text{vs.} \quad H_1 : m_\theta(x_1, \ldots, x_D) \notin \mathcal{M}_\Theta, \tag{10}$$

where $\mathcal{M}_\Theta = \{m_\theta(x_1, \ldots, x_D) = \sum_{d=1}^{D} m_d(x_d; \theta) : \theta \in \Theta\}$ is a set of functions of parametric forms and the parameter space Θ contains the true parameter value θ_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 $\ell_n(H_0)$. Let $\lambda_n(\mathcal{M}_\Theta)$ denote the GLR statistic for the testing problem (10). To derive the asymptotic null distribution of the test statistic, some conditions on $\mathcal{M}_\Theta$ and Θ are required to render the likelihood ratio test statistic of order $o_p(h_D^{-1/2})$ for the following parametric testing problem:

$$H_0' : m(x_1, \ldots, x_D) = m_{\theta_0}(x_1, \ldots, x_D) \quad \text{vs.} \quad H_1' : m(x_1, \ldots, x_D) \in \mathcal{M}_\Theta.$$

For ease of exposition, we call the required conditions "condition B." Conditions similar to those of Cramér [see, e.g., conditions (C1)–(C5) of Le Cam and Yang 1990, p. 120] are sufficient in the present setting, because the classical Wilks theorem holds, and hence the likelihood ratio statistic is of order $O_p(1)$.

Theorem 3. Suppose that condition A in Appendix A and condition B hold. Then, under H_0 for the testing problem (10),

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

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

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

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

3.3 Power of Generalized Likelihood Ratio Tests

We now consider the power of GLR tests in the framework of Fan et al. (2001). For simplicity, we focus on the null hypothesis in (7).

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 σ_n. As is stated later in Theorem 5, the optimal bandwidth for the testing problem (7) 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 1 leads to an approximate level-α 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 $G_n(X_D) \to 0$ as $n \to \infty$, then the power of the GLR test can be approximated using the following theorem.

Theorem 4. Suppose that condition A in Appendix A holds and that for $d = 1, \ldots, D$, $nh_d^{2(p_d+1)}h_D \to 0$. If

$$E\{G_n(X_D)|X_1, \ldots, X_{D-1}\} = 0 \qquad \text{and}$$

$$h_D \cdot \sum_{i=1}^{n} G_n^2(X_{Di}) \xrightarrow{P} C(G)$$

for some constant $C(G)$, then, under H_{1n} for the testing problem (7),

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

where μ_n is the same as that in Theorem 1,

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

and

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

Remark 2. For testing problem (7), the alternative hypothesis depends on many nuisance functions m_d for $d = 1, \ldots, D-1$. Theorem 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 1 and 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 regularity conditions

$$\begin{aligned} &\operatorname{var}(G_n^2(X_D)) \leq M\big(E[G_n^2(X_D)]\big)^2, \\ &nE[G_n^2(X_D)] > M_n \to \infty, \end{aligned} \tag{11}$$

for some constants $M > 0$ and $M_n \to \infty$. 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 ϕ_h is defined as the smallest ρ_n such that:

(a) 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

(b) for any sequence $\rho_n^* = o(\rho_n)$, there exist $\alpha > 0$ and $\beta > 0$ 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 ϕ_h.

Theorem 5. Under condition A in Appendix A, if $h_d^{p_d+1} = O(h_D^{p_D+1})$ for $d = 1, \ldots, D-1$, then for the testing problem (7), 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. The GLR tests are asymptotically optimal in terms of rates of convergence for nonparametric hypothesis testing according to the formulations of Ingster (1993) and Spokoiny (1996). Although Ingster (1993) and Spokoiny (1996) focused only on the univariate setting, their minimax lower bound is applicable to our additive model with known functions, m_d's, for $d = 1, \ldots, D-1$. Its rate of convergence is the same as the rate of the upper bound given in Theorem 5.

Because the distributional property in Theorem 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 well suited for curve estimation may not be the best for testing. The power of the GLR tests depends on the smoothing parameters. In fact, Theorem 5 shows that theoretical optimal bandwidth h_D is $c^* n^{-2/(4p_D+5)}$ for some constant c^*.

4. IMPLEMENTATION OF GENERALIZED LIKELIHOOD RATIO TESTS

The GLR test involves 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 must be sufficiently small 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 (10), we 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 used 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) advocated using the bias-reduction method in the study of testing problems for spectral density.

Consider the testing problem (10). The additive model (1) is equivalent to

$$Y_i^* = m^*(X_{1i}, \ldots, X_{Di}) + \varepsilon_i, \tag{12}$$

where $Y_i^* = Y_i - \hat{\alpha} - m(X_{1i}, \ldots, X_{Di}; \widehat{\theta})$ and

$$m^*(X_{1i}, \ldots, X_{Di}) = \alpha + m(X_{1i}, \ldots, X_{Di}) - \hat{\alpha} - m(X_{1i}, \ldots, X_{Di}; \widehat{\theta}),$$

with $\widehat{\theta}$ being the least squares estimator of θ under the null hypothesis in (10). Therefore, the testing problem (10) is reduced to the following problem:

$$H_0^*: m^*(X_1, \ldots, X_D) \in \mathcal{M}_0 \quad \text{vs.} \quad H_1^*: m^*(X_1, \ldots, X_D) \notin \mathcal{M}_0,$$

where $\mathcal{M}_0 = \{\alpha^* = 0, m_1^* = \cdots = m_D^* = 0\}$. This is the specific case of (10), and hence the GLR test can be applied. Let $\lambda_n^*(\mathcal{M}_\Theta)$ denote the resulting GLR statistic. Because the regression function $m^*(X_1, \ldots, X_D)$ is nearly 0 under H_0, little bias is 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 6. Suppose that condition A in Appendix A and condition B hold. Then, conditioning on $\mathcal{X}$ under H_0 for the testing problem (10),

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

where μ_n^* and r_K^* are the same as those in Theorem 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 we gave the asymptotic distributions of the GLR statistics, demonstrating 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 as follows [to be more specific, consider (3.1)]:

1. Fix the bandwidths at their estimated values $(\hat{h}_1, \ldots, \hat{h}_D)$, 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 $i = 1, \ldots, n$) 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 $Y_i^* = \hat{\alpha} + \hat{m}_1(X_{i1}) + \cdots + \hat{m}_{D-1}(X_{i,D-1}) + \hat{\varepsilon}_i^*$, where $\hat{\alpha}$ and $\hat{m}_j(\cdot)$ $(j \le D-1)$ 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 $(\hat{h}_1, \ldots, \hat{h}_D)$, 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 α, $(m_1, \ldots, m_{D-1})$ and distribution of $\boldsymbol{\varepsilon}$. As shown in Theorem 1, such a dependence is asymptotically negligible. Hence they can be fixed at the values $(\hat{\alpha}, \hat{m}_1, \ldots, \hat{m}_{D-1})$ and the distribution of $\hat{\boldsymbol{\varepsilon}}^*$. The following theorem shows the consistency of the conditional bootstrap method.

Theorem 7. Assume that the conditions in Theorem 1 hold. Then, under H_0 in (7),

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

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 $d = 1, \ldots, D$. 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 use these theoretic optimal bandwidths in our simulation study. For real data examples, we use the automatic bandwidth selection rule of Opsomer and Ruppert (1998). Because of 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 Zhang 2003 for the result on nonparametric regression, which corresponds to $D = 1$.) Thus, for many applications, it suffices to use $h = h_{\text{opt}}/1.5, h_{\text{opt}}, 1.5h_{\text{opt}}$, corresponding to "undersmooth," "right smooth," and "oversmooth," where h_{opt} is the asymptotically optimal bandwidth used for estimation. We follow this idea 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 used.

Example 1. Consider the bivariate additive model

$$Y = m_1(X_1) + m_2(X_2) + \varepsilon, \tag{13}$$

where $m_1(X_1) = .5 - 6X_1^2 + 3X_1^3$, $m_2(X_2) = \sin(\pi X_2)$, and the error ε is distributed as $\mathcal{N}(0, 1)$. The covariates are generated by the following transformation to create correlation:

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} 1 & .5 \\ .5 & 1 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \end{pmatrix}, \tag{14}$$

where $U_i \overset{\text{iid}}{\sim} \mathrm{U}(-.5, .5)$.

We use the optimal bandwidth $h_{d,\text{opt}}$ for the smoother on $m_d(x_d)$ (see Opsomer 2000). To demonstrate the Wilks phenomenon for the GLR test, we evaluate three levels of bandwidth with h_2 fixed at its optimal value, $h_1 = \frac{2}{3}h_{1,\text{opt}}$, $h_{1,\text{opt}}$, or $\frac{3}{2}h_{1,\text{opt}}$. 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 $m_1(X_1)$:

$$m_{1,\beta}(X_1) = \left[1 + \beta\sqrt{\operatorname{var}(.5 - 6X_1^2 + 3X_1^3)}\,\right](.5 - 6X_1^2 + 3X_1^3),$$

where $\beta = -1.5, 0, 1.5$. For the GLR test, we drew 1,000 samples of 200 observations. Based on the 1,000 samples, we obtained 1,000 GLR test statistics. Their distribution is obtained via a kernel estimate with a rule-of-thumb bandwidth, $h = 1.06sn^{-.2}$, 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-squared distributions or, more generally, gamma distributions. Figure 1(a) shows that the null distributions follow chi-squared distributions over a wide range of bandwidth h_1 (with the degree of freedom depending on bandwidth h_2 but not on h_1). Figure 1(b) demonstrates the Wilks type of phenomenon; for three very different choices of nuisance functions, the null distributions are nearly the same.

For the power assessment, we evaluate the power for a sequence of alternative models indexed by θ,

$$H_\theta : m_{2,\theta}(x_2) = \theta \sin(\pi x_2), \qquad 0 \le \theta \le 1, \tag{15}$$

ranging from the null model to reasonably far away from it. Figure 2(a) reports the differences between the null and the alternatives in (15).

For each given value of θ, we use 3,000 Monte Carlo replicates for the calculation of the critical values via the conditional bootstrap method (see Sec. 4.2), and compute the rejection frequencies based on 600 simulations. The parameter θ 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 .05 (or .10) at the .05 (or .10) significance level. This is indeed the case, as shown in Table 1, which again implies that the

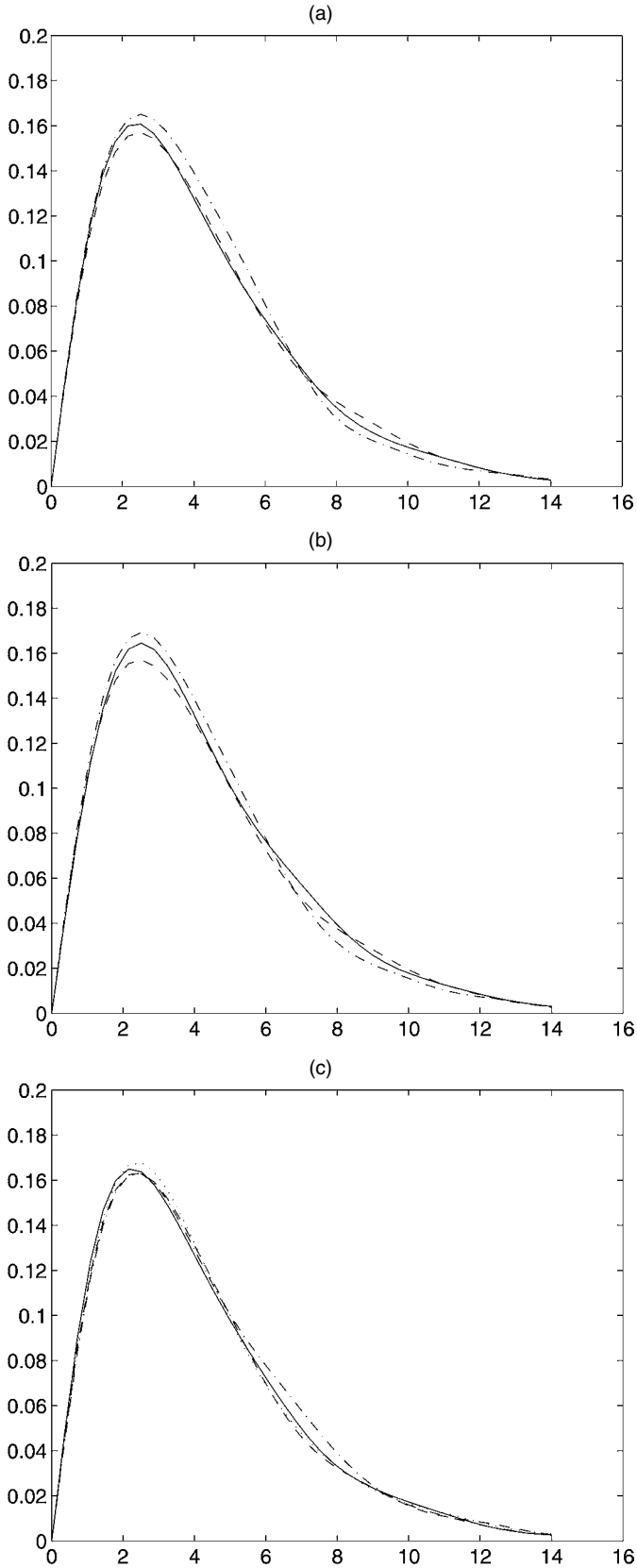

Figure 1. Results for Example 1: Estimated Densities for the GLR Statistics Among 1,000 Simulations. (a) With fixed $h_2 = h_{2,\text{opt}}$, but different bandwidths for h_1 (—— $h_1 = \frac{2}{3}h_{1,\text{opt}}$; − − − $h_1 = h_{1,\text{opt}}$; − · − · $h_1 = \frac{3}{2}h_{1,\text{opt}}$). (b) With different nuisance functions and optimal bandwidths $h_d = h_{d,\text{opt}}$ (—— $\beta = -1.5$; − − − $\beta = 0$; − · − · $\beta = 1.5$). (c) Estimated densities for the GLR statistics under different errors [—— normal; − − − t(5); ····· $\chi^2(5)$; − · − · $\chi^2(10)$].

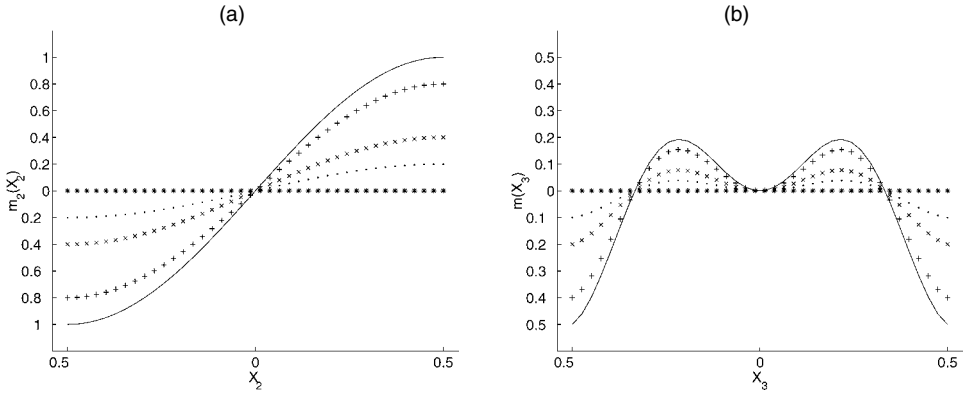

Figure 2. Difference Between the Null and the Alternative Hypotheses for (a) Example 1 and (b) Example 2. (∗ ∗ ∗ $\theta = 0$; ····· $\theta = .2$; × × × $\theta = .4$; + + + $\theta = .8$; —— $\theta = 1.0$.)

Monte Carlo method gives a correct estimator of the null distribution. When θ 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 1.

To investigate the power and the influence of different error distributions on the GLR tests, we now consider the model (13) with different error distributions of ε. 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 has a heavy tail, and that 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 aforementioned four different error distributions are reported in Figure 1(c). 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 1. The powers of the GLR tests for the alternative sequence in (15) 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 for differentiating the null and the smooth alternatives, but also have robustness against error distributions to some extent.

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

$$Y = m(X_1, X_2, X_3) + \varepsilon, \tag{16}$$

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 .25, and X_3 is independent of (X_1, X_2). We rejected all observations in which one of the covariates fell out $[-.5, .5]$ and replaced them with new observations, so that the support of the covariates is bounded. The error ε 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 be easily 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:

$$H_\theta : m_\theta(X_1, X_2, X_3) = m(X_1, X_2, X_3) + \theta X_3 \cdot m_3(X_3), \tag{17}$$

where $0 \le \theta \le 1$. When $\theta = 0$, $H_\theta = H_0$. As θ increases, the alternative model H_θ deviates away from H_0. Figure 2(b) shows the difference between the null and the alternative models.

For each given θ, we simulated data from the alternative model H_θ. The percentages 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 powers of both tests become the test sizes. It is evident from Figure 3 that the bias-corrected test is more powerful than its counterpart. Note that the bandwidths used earlier are optimal for estimation. Setting the bandwidths to be half of their optimal values makes the bias of the backfitting estimator decreases and the relative advantage of the bias-correction method over its counterpart decline. 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.

Table 1. Powers of the Proposed Tests Under Different Error Distributions

α	Error distribution\θ	0	.2	.4	.6	.8	1.0
.05	$\mathcal{N}(0,1)$	.057	.192	.592	.948	.997	1.000
	$t(5)$	.043	.146	.537	.903	.998	1.000
	$\chi^2(5)$	.048	.175	.640	.963	.995	1.000
	$\chi^2(10)$	.053	.230	.657	.952	.995	1.000
.10	$\mathcal{N}(0,1)$	.095	.280	.728	.977	.997	1.000
	$t(5)$	.090	.255	.710	.952	1.000	1.000
	$\chi^2(5)$	.088	.268	.727	.973	.997	1.000
	$\chi^2(10)$	.087	.298	.717	.967	.998	1.000

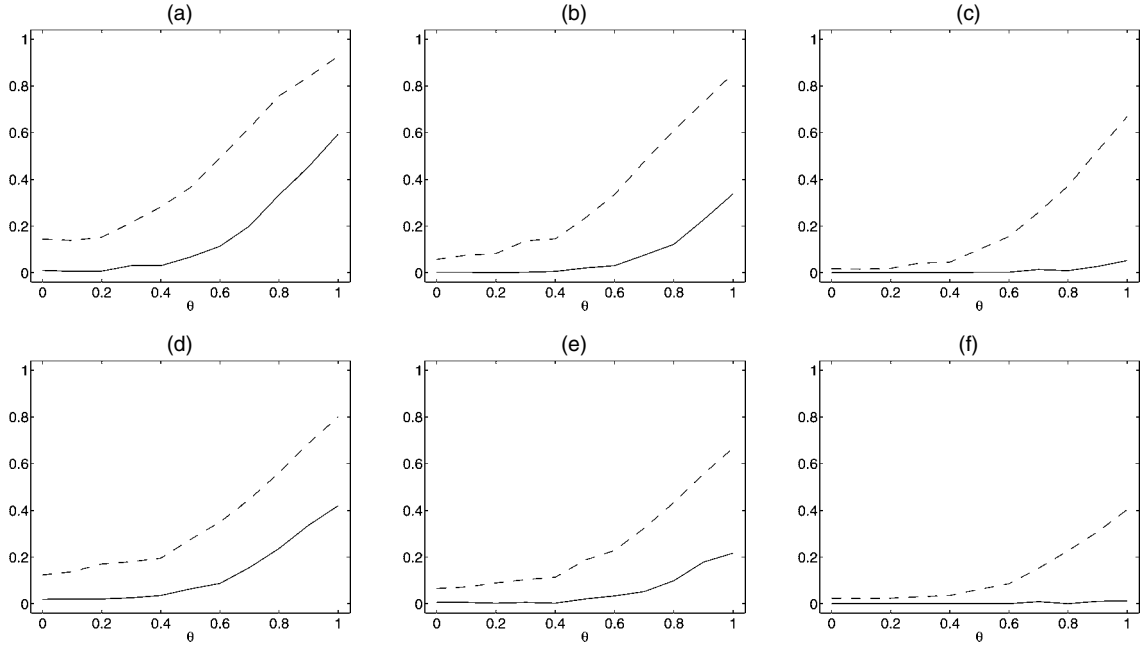

Figure 3. Power Functions of the GLR Tests for Example 2. (a), (b), and (c): $\mathbf{h} = \mathbf{h}_{opt}$; (d), (e), and (f): $\mathbf{h} = \mathbf{h}_{opt}/2$. From left to right, significance levels are $\alpha = .10$ [(a), (d)], .05 [(b), (e)], and .01 [(c), (f)]. The dashed lines represent the bias-corrected method, the solid lines, the tests without bias reduction.

6. REAL DATA EXAMPLE

We use the proposed GLR tests on the Boston Housing dataset to demonstrate their use in applications. The dataset comprises the median value of homes in 506 census tracts in the Boston Standard Metropolitan Statistical Area in 1970, along with 13 accompanying sociodemographic and related variables. It was previously studied by several authors, including Harrison and Rubinfeld (1978), Belsley, Kuh, and Welsch (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,

$$E[MV - \overline{MV}|X_1, X_2, X_3, X_4] = m_1(X_1) + m_2(X_2) + m_3(X_3) + m_4(X_4), \quad (18)$$

where $X_1 = RM$, $X_2 = \log(TAX)$, $X_3 = PTRATIO$, and $X_4 = \log(LSTAT)$. The local linear smoother and a fully automated bandwidth selection method were used after six outliers were removed. Opsomer and Ruppert suggested that the fitted additive components have 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 (18) via the method of Opsomer and Ruppert (1998) gives 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:

$$\begin{aligned} m_1(X_i) &= a_1 + b_1 X_1 + c_1 X_1^2, \\ m_2(X_2) &= a_2 + b_2 X_2, \\ m_3(X_3) &= a_3 + b_3 X_3, \\ m_4(X_4) &= a_4 + b_4 X_4. \end{aligned} \quad (19)$$

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

$$m_i(X_i) = a_i + b_i X_i \quad \text{for } i = 2, 3, 4, \quad (20)$$

where $m_1(X_1)$ is unspecified.

We now use our GLR statistic to test whether the semiparametric null model (20) holds against the additive alternative model (18). 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 given in Section 4.2. The p value of our GLR test is estimated as 0 by using the optimal bandwidth and using 1,000 bootstrap replicates. This does not come as a surprise to us, because the p value depends heavily on the sample size. With a

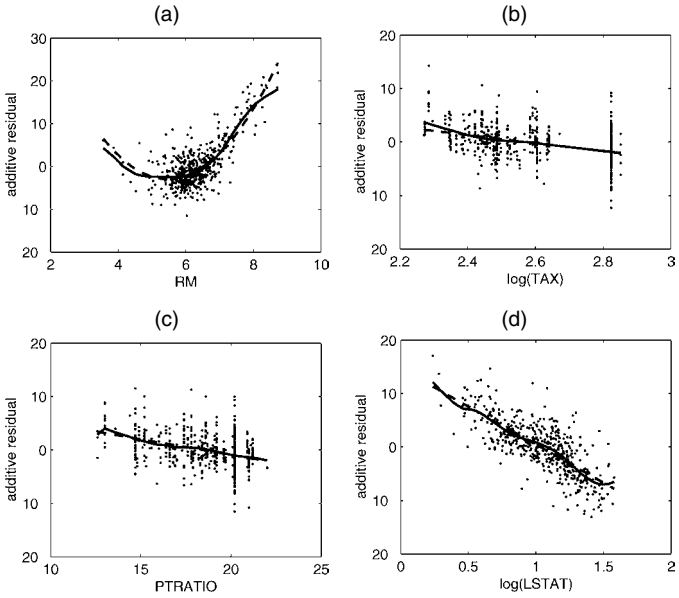

Figure 4. Partial Residual Plots Along With Fitted Regression Curves for the Boston Housing Dataset. (a) RM; (b) log(TAX); (c) PTRATIO; (d) log(LSTAT). The solid lines represent estimated additive functions; dashed lines, empirical regression lines based on model (19).

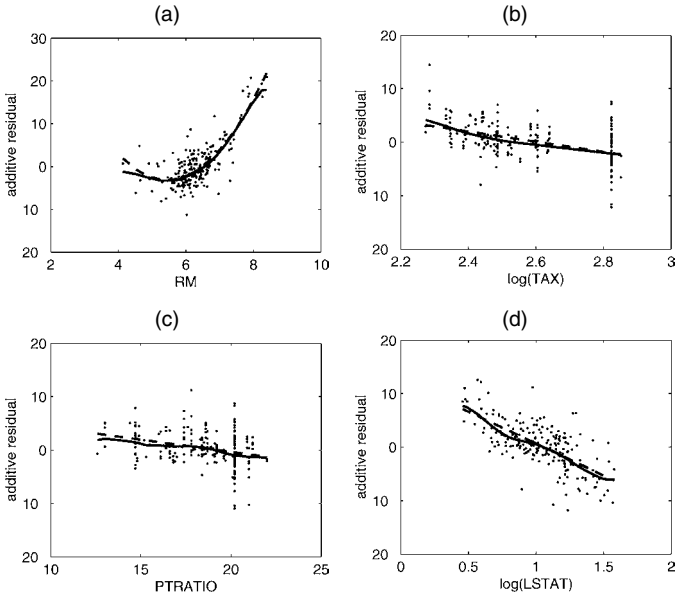

Figure 5. Partial Residual Plots Along With Fitted Regression Curves for a Random Subsample of the Boston Housing Dataset. (a) RM; (b) log(TAX); (c) PTRATIO; (d) log(LSTAT). The solid lines represent estimated additive functions; dashed lines, empirical regression lines based on model (19).

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 (18) for the subsample are reported in Figure 5, which also shows the fitted additive components from models (18) and (19). The optimal bandwidth from the automated bandwidth selection rule of Opsomer and Ruppert (1998) is computed as $\mathbf{h}_{\text{opt}} = (1.1129, .2530, 2.1432, .2315)^T$. Visually, similar parametric forms of the additive components are suggested from Figures 4 and 5. Our interest is to test whether the model (20) is adequate for the subsample. Table 2 reports the results of the GLR tests for five different bandwidths, using 1,000 bootstrap replicates. This provides stark evidence that the semiparametric model is appropriate for this dataset within the additive models at the .01 significance level.

7. DISCUSSION

7.1 Other Tests

Many nonparametric tests have been designed for certain specific problems. Most are in the univariate nonparametric regression setting. (See Fan et al. 2001 for an overview of the literature.) Although these tests can be powerful for the problems for which the tests were designed, extensions to multivariate settings can pose some challenges. Further, these tests are usually not distribution-free when null hypotheses involve nuisance functions. This hampers their applicability.

Table 2. Results of the GLR Tests

Bandwidth	*RSS_0*	*RSS_1*	*GLRT*	*p value*
$\frac{1}{2}\mathbf{h}_{\text{opt}}$	1,974.3	1,721.9	31.0	.097
$\frac{2}{3}\mathbf{h}_{\text{opt}}$	2,044.0	1,812.5	27.0	.034
$\mathbf{h}_{\text{opt}}$	2,091.2	1,904.3	20.8	.016
$\frac{3}{2}\mathbf{h}_{\text{opt}}$	2,158.6	2,042.3	12.0	.046
$2\mathbf{h}_{\text{opt}}$	2,301.7	2,231.6	6.65	.179

NOTE: RSS_0 and RSS_1 are the sum of squared residuals for the GLR test under H_0 and H_1; GLRT is the normalized GLR statistic.

Hypothesis testing for multivariate regression problems is difficult due to the curse of dimensionality. Aerts, Claeskens, and Hart (1999) constructed tests based on orthogonal series for the bivariate regression setting. Fan and Huang (2001) proposed various testing techniques based on the adaptive Neyman test for various alternative models in the multiple regression setting. These problems become conceptually simple by using our generalized likelihood method. Delgado and González-Manteiga (2001) developed a test for selecting explanatory variables in nonparametric regression based on functionals of a U-process, whereas this test can detect a specific class of contiguous alternatives at a rate $n^{-1/2}$. However, this requires that one 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. Neumeyer 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 pseudolikelihood 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 constructed based on the variance estimators from a parametric fitting and a nonparametric kernel smoother (see, e.g., Dette 1999). Our GLR tests are asymptotically distribution-free and yield 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 GLR tests are derived for nonlinear additive models (1) based on the local polynomial smoother and the backfitting algorithm. For heteroscedastic errors, for example, the model

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

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

$$\varepsilon_i^2 = \sigma^2(\mathbf{X}_i) + \sigma^2(\mathbf{X}_i)(u_i^2 - 1). \tag{21}$$

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

1. Fit the regression components by backfitting algorithm, and obtain the residuals $\hat{\varepsilon}_i = Y_i - \bar{Y} - \sum_{d=1}^{D} \hat{m}_d(X_{di})$.
2. Obtain the estimator $\hat{\sigma}(\mathbf{X}_i)$ by fitting the model (21) with ε_i replaced by $\hat{\varepsilon}_i$, and get the $RSS_1 = \sum_{i=1}^{n} \hat{u}_i^2$, where $\hat{u}_i = \hat{\varepsilon}_i / \hat{\sigma}(\mathbf{X}_i)$.
3. Compute $RSS_0 = \sum_{i=1}^{n} (\hat{\varepsilon}_i^0)^2 / \hat{\sigma}^2(\mathbf{X}_i)$ and form the GLR (8), 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 foregoing method still applies, but with other fitting techniques for $\sigma(\mathbf{X}_i)$. The techniques can also be extended to generalized additive models (Hastie and Tibshirani 1990). We would expect similar results to 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 form has the advantage of avoiding the sparsity of design points, but the bandwidths at such regions are large, and hence it can introduce large modeling biases. When the constant span is used, its effective bandwidth depends on the design density and usually is not a constant. Our asymptotic results can be extended to the 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, which 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 result.

APPENDIX A: 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 Ω_d for $d = 1, \ldots, D$.
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 $n \to \infty$, $h_d \to 0$ and $nh_d / \log(n) \to \infty$ for $d = 1, \ldots, D$.
5. The $(p_d + 1)$st derivatives of m_d (for $d = 1, \ldots, D$) exist and are bounded and continuous.
6. $E|\varepsilon_i|^4 < \infty$.

APPENDIX B: 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 $\mathbf{P}_1$ and $\mathbf{P}_2$ of the same dimension. For any constant d, $\mathbf{d}$ is the n-valued vector $(d, \ldots, d)^T$. Denote by $\overline{Z}$ the average of components of any vector $\mathbf{Z}$. To facilitate the exposition of the proofs, we ignore the intercept α 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 B.1. Let assumptions 1–4 in condition A hold. Then

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

uniformly for $x_d \in \Omega_d$.

Proof. The result was derived by Fan and Gijbels (1996, p. 64).

Lemma B.2. Under assumptions 1–4 in condition A, the following asymptotic approximations hold uniformly over all elements of the matrices:

$$\mathbf{S}_d^* = \mathbf{S}_d - \frac{\mathbf{1}\mathbf{1}^T}{n} + o\left(\frac{\mathbf{1}\mathbf{1}^T}{n}\right) \quad \text{a.s.},$$

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

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 B.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 $\mathcal{X}$,

$$RSS_0 - RSS_1 = \mathbf{Y}^T [\mathbf{A}_{n1} - \mathbf{A}_{n2}] \mathbf{Y} \tag{B.1}$$

and

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

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 $1 \le i, j; i_1, j_1; i_2, j_2 \le n$.

Proof. By definition, we have (B.1). Using an argument 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 B.2, and direct matrix multiplications,

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

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

$$\mathbf{W}_M^{[-D]} = \mathbf{S}^{[-D]} + \mathbf{U}^{[-D]} \tag{B.3}$$

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,

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

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 B.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

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

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 B.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). If condition A holds, then

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

$$\mathbf{B} = O\left(\sum_{d=1}^{D} \mathbf{h}_d^{p_d+1}\right) + \sum_{d=1}^{D} \bar{m}_d \cdot O(\mathbf{1}) \quad \text{a.s.}, \tag{B.4}$$

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 B.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 (5) that

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

Note that

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

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

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

Hence (B.4) holds by a recursive argument.

Lemma B.6. If condition A holds, then, under $H_0 : m_D = 0$,

$$\begin{aligned}d_{1n} &\equiv \mathbf{m}^T(\mathbf{A}_{n1} - \mathbf{A}_{n2})\mathbf{m} + 2\boldsymbol{\varepsilon}^T(\mathbf{A}_{n1} - \mathbf{A}_{n2})\mathbf{m} \\ &= O_p\left(1 + \sum_{d=1}^{D} nh_d^{2(p_d+1)} + \sum_{d=1}^{D} \sqrt{n}h_d^{p_d+1}\right),\end{aligned} \tag{B.6}$$

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

Proof. a. Under H_0, we obtain from Lemma B.5 that

$$\begin{aligned}\mathbf{m}^T(\mathbf{A}_{n1} - \mathbf{A}_{n2})\mathbf{m} &= \mathbf{B}_{-D}^T\mathbf{B}_{-D} - \mathbf{B}^T\mathbf{B} \\ &= O_p\left(1 + \sum_{d=1}^{D} nh_d^{2(p_d+1)}\right).\end{aligned} \tag{B.7}$$

By Lemmas B.3 and B.5, we have, under H_0,

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

$$E[(\mathbf{W}_M - \mathbf{I}_n)\boldsymbol{\varepsilon}] = 0,$$

and

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

By directly computing the mean and variance and using Lemma B.1, we obtain $\mathbf{B}^T\mathbf{S}\boldsymbol{\varepsilon} = O_p(1 + \sum_{d=1}^{D} \sqrt{n}h_d^{p_d+1})$, $\mathbf{B}^T\boldsymbol{\varepsilon} = O_p(1 + \sum_{d=1}^{D} \sqrt{n}h_d^{p_d+1})$, and hence $\mathbf{B}^T(\mathbf{W}_M - \mathbf{I}_n)\boldsymbol{\varepsilon}$ is bounded by $O_p(1 + \sum_{d=1}^{D} \sqrt{n}h_d^{p_d+1})$. By the same argument, $\mathbf{B}_{-D}^T(\mathbf{W}_M^{[-D]} - \mathbf{I}_n)\boldsymbol{\varepsilon} = O_p(1 + \sum_{d=1}^{D} \sqrt{n}h_d^{p_d+1})$. Therefore, the second term in d_{1n} is $O_p(1 + \sum_{d=1}^{D} \sqrt{n}h_d^{p_d+1})$, which, combined with (B.7), leads to (B.6).

b. Assume that $m_d(\cdot)$ is a polynomial of order p_d (for $d = 1, \ldots, D$); then $\mathbf{Q}_d = 0$. Using recursive reasoning, we obtain $\mathbf{B}_{-D} =$

$\sum_{d=1}^{D-1} \bar{m}_d O(\mathbf{1}) = O_p(\frac{\mathbf{1}}{\sqrt{n}})$ and $\mathbf{B} = O_p(\frac{\mathbf{1}}{\sqrt{n}})$. 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 1

The proof consists mainly of the following four steps:

1. Asymptotic expression for $RSS_0 - RSS_1$. By definition, we have

$$RSS_0 - RSS_1 = \left\|\mathbf{W}_M^{[-D]}\mathbf{Y} - \mathbf{Y}\right\|^2 - \|\mathbf{W}_M\mathbf{Y} - \mathbf{Y}\|^2,$$

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

$$\begin{aligned} RSS_0 - RSS_1 &= \mathbf{Y}^T[\mathbf{A}_{n1} - \mathbf{A}_{n2}]\mathbf{Y} \\ &= \boldsymbol{\varepsilon}^T(\mathbf{A}_{n1} - \mathbf{A}_{n2})\boldsymbol{\varepsilon} \\ &\quad + \left[\mathbf{m}^T(\mathbf{A}_{n1} - \mathbf{A}_{n2})\mathbf{m} + 2\boldsymbol{\varepsilon}^T(\mathbf{A}_{n1} - \mathbf{A}_{n2})\mathbf{m}\right] \\ &\equiv \boldsymbol{\varepsilon}^T(\mathbf{A}_{n1} - \mathbf{A}_{n2})\boldsymbol{\varepsilon} + d_{1n}. \end{aligned} \tag{B.8}$$

From Lemma B.6, d_{1n} is bounded by $O_p(1 + \sum_{d=1}^{D} nh_d^{2(p_d+1)} + \sum_{d=1}^{D} \sqrt{n}h_d^{p_d+1})$. In the following, we show that the first term in (B.8) can be approximated as

$$\begin{aligned} &\boldsymbol{\varepsilon}^T[\mathbf{A}_{n1} - \mathbf{A}_{n2}]\boldsymbol{\varepsilon} \\ &\approx \frac{2}{nh_D}\sum_{i<j}\varepsilon_i\varepsilon_j f_D^{-1}(X_{Di}) \\ &\quad \times \left\{2\sum_{t=1}^{p_D+1}\tilde{s}_{D,t}K_t\left(\frac{X_{Dj} - X_{Di}}{h_D}\right)\right. \\ &\quad \left. - \sum_{s,t=1}^{p_D+1}\tilde{s}_{D,s}\tilde{s}_{D,t}K_s * K_t\left(\frac{X_{Dj} - X_{Di}}{h_D}\right)\right\} + \mu_n^* + o_p(h_D^{-1}) \\ &\equiv W_{(n)} + \mu_n^* + o_p(h_D^{-1}), \end{aligned} \tag{B.9}$$

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

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

Then, by (B.8),

$$RSS_0 - RSS_1 \approx W_{(n)} + 2\sigma^2\mu_n + d_{1n} + o_p(h_D^{-1}). \tag{B.10}$$

For readers who are not interested in the proof of (B.9), please skip to step 2. Note that the (j, ℓ)th element of $\mathbf{H}_d^{-1}\mathbf{X}_{X_{dk}}^{dT}\mathbf{K}_{X_{dk}}$ is now

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

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

$$(\mathbf{S}_d)_{ij} \approx \frac{1}{nh_d}f_d^{-1}(X_{di})\sum_{t=1}^{p_d+1}\tilde{s}_{d,t}K_t\left(\frac{X_{dj} - X_{di}}{h_d}\right). \tag{B.11}$$

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

$$\sum_{i=1}^{n}\varepsilon_i^2(\mathbf{S}_d)_{ii} \approx \frac{\sigma^2}{h_d}|\Omega_d|\sum_{t=1}^{p_d+1}\tilde{s}_{d,t}K_t(0) \tag{B.12}$$

and

$$\sum_{i\neq j}\varepsilon_i\varepsilon_j(\mathbf{S}_d)_{ij} \approx \frac{1}{nh_d}\sum_{i\neq j}\varepsilon_i\varepsilon_j f_d^{-1}(X_{di})\sum_{t=1}^{p_d+1}\tilde{s}_{d,t}K_t\left(\frac{X_{dj} - X_{di}}{h_d}\right). \tag{B.13}$$

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

$$\begin{aligned} (\mathbf{S}_d^T\mathbf{S}_{d'})_{ij} &\approx \frac{1}{n}\sum_{s=1}^{p_d+1}\sum_{t=1}^{p_{d'}+1}\tilde{s}_{d,s}\tilde{s}_{d',t} \\ &\quad \times \frac{1}{n}\sum_{k=1}^{n}\left[f_d^{-1}(X_{dk})f_{d'}^{-1}(X_{d'k})\frac{1}{h_d}K_s\left(\frac{X_{di} - X_{dk}}{h_d}\right)\right. \\ &\quad \left. \times \frac{1}{h_{d'}}K_t\left(\frac{X_{d'j} - X_{d'k}}{h_{d'}}\right)\right] \\ &\equiv n^{-1}\sum_{s=1}^{p_d+1}\sum_{t=1}^{p_{d'}+1}\tilde{s}_{d,s}\tilde{s}_{d',t}\frac{1}{n}\sum_{k=1}^{n}P_{dd'ijkh}. \end{aligned} \tag{B.14}$$

Then

$$\begin{aligned} &\sum_{i\neq j}\varepsilon_i\varepsilon_j(\mathbf{S}_d^T\mathbf{S}_{d'})_{ij} \\ &= \frac{1}{n}\sum_{i\neq j}\varepsilon_i\varepsilon_j\sum_{s=1}^{p_d+1}\sum_{t=1}^{p_{d'}+1}\tilde{s}_{d,s}\tilde{s}_{d',t}\frac{1}{n}\sum_{k\neq i,j}P_{dd'ijkh} \\ &\quad + \frac{1}{n}\sum_{i\neq j}\varepsilon_i\varepsilon_j\sum_{s=1}^{p_d+1}\sum_{t=1}^{p_{d'}+1}\tilde{s}_{d,s}\tilde{s}_{d',t}\frac{1}{n}[P_{dd'ijih} + P_{dd'ijjh}] \\ &\equiv L_{ndd'1} + L_{ndd'2}. \end{aligned} \tag{B.15}$$

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

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

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

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

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

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

where the first approximation is from

$$\begin{aligned} &E\left[n^{-1}\sum_{k\neq i,j}\big(P_{dd'ijkh} - E(P_{dd'ijkh}|X_{di}, X_{d'j})\big)\right]^2 \\ &\qquad \leq n^{-2}\sum_{k\neq 1,2}E\big[P_{dd'12kh}^2\big] = O\left(\frac{1}{nh_dh_{d'}}\right). \end{aligned} \tag{B.16}$$

Then, by (B.15), for $d \neq d'$,

$$\sum_{i\neq j}\varepsilon_i\varepsilon_j(\mathbf{S}_d^T\mathbf{S}_{d'})_{ij} = O_p\left(1 + \frac{1}{nh_d\sqrt{h_{d'}}} + \frac{1}{nh_{d'}\sqrt{h_d}}\right). \tag{B.17}$$

By (B.16), 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 $i, j = 1, \ldots, n$, so that

$$(\mathbf{S}_d^T \mathbf{S}_{d'})_{ij} \approx \frac{1}{nh_d} \sum_{s,t=1}^{p_d+1} \tilde{s}_{d,s} \tilde{s}_{d,t} \nu_{s-1} \nu_{t-1} \frac{f_{dd'}(X_{di}, X_{d'j})}{f_d(X_{di}) f_{d'}(X_{d'j})}$$
$$= \frac{1}{n} \frac{f_{dd'}(X_{di}, X_{d'j})}{f_d(X_{di}) f_{d'}(X_{d'j})}. \quad \text{(B.18)}$$

Therefore, for $d \neq d'$,

$$\sum_{i=1}^{n} \varepsilon_i^2 (\mathbf{S}_d^T \mathbf{S}_{d'})_{ii} \approx n^{-1} \sum_{i=1}^{n} \varepsilon_i^2 C_d^{(0)} C_{d'}^{(0)} \frac{f_{dd'}(X_{di}, X_{d'j})}{f_d(X_{di}) f_{d'}(X_{d'j})}$$
$$= O_p(1). \quad \text{(B.19)}$$

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

$$E[P_{ddijkh} | X_{di}, X_{dj}] = \frac{1}{h_d^2} \int f_d^{-1}(u) K_s\left(\frac{X_{di} - u}{h_d}\right) K_t\left(\frac{X_{dj} - u}{h_d}\right) du$$
$$\approx \frac{1}{h_d} f_d^{-1}(X_{di}) K_s * K_t\left(\frac{X_{dj} - X_{di}}{h_d}\right),$$

which, combined with (B.16), leads to

$$n^{-1} \sum_{k \neq i,j} P_{ddijkh} \approx \frac{1}{h_d} f_d^{-1}(X_{di}) K_s * K_t\left(\frac{X_{di} - X_{dj}}{h_d}\right). \quad \text{(B.20)}$$

It follows from (B.14) and (B.20) that for $i \neq j$,

$$(\mathbf{S}_d^T \mathbf{S}_d)_{ij} \approx \frac{1}{nh_d} \sum_{s,t=1}^{p_d+1} \tilde{s}_{d,s} \tilde{s}_{d,t} f_d^{-1}(X_{di}) K_s * K_t\left(\frac{X_{di} - X_{dj}}{h_d}\right). \quad \text{(B.21)}$$

By the definition of L_{ndd1} and (B.20), we have

$$L_{ndd1} \approx \frac{2(n-2)}{n^2} \sum_{i<j} \varepsilon_i \varepsilon_j \sum_{s,t=1}^{p_d+1} \tilde{s}_{d,s} \tilde{s}_{d,t}$$
$$\times \frac{1}{h_d} f_d^{-1}(X_{di}) K_s * K_t\left(\frac{X_{dj} - X_{di}}{h_d}\right). \quad \text{(B.22)}$$

Observing that $L_{ndd2} = O_p(\frac{1}{nh_d\sqrt{h_d}}) = o_p(h_d^{-1})$, by (B.15) and (B.20) we obtain

$$\sum_{i \neq j} \varepsilon_i \varepsilon_j (\mathbf{S}_d^T \mathbf{S}_d)_{ij} \approx \frac{2(n-2)}{n^2} \sum_{i<j} \varepsilon_i \varepsilon_j \sum_{s,t=1}^{p_d+1} \tilde{s}_{d,s} \tilde{s}_{d,t} \frac{1}{h_d} f_d^{-1}(X_{di})$$
$$\times K_s * K_t\left(\frac{X_{dj} - X_{di}}{h_d}\right) + o_p(h_d^{-1}). \quad \text{(B.23)}$$

By (B.21) and the same argument as that for (B.12), we obtain

$$\sum_{j=1}^{n} \varepsilon_j^2 (\mathbf{S}_d^T \mathbf{S}_d)_{jj} \approx \frac{\sigma^2}{h_d} |\Omega_d| \sum_{s,t=1}^{p_d+1} \tilde{s}_{d,s} \tilde{s}_{d,t} K_s * K_t(0). \quad \text{(B.24)}$$

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

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

This, together with (B.12), (B.13), (B.17), (B.19), (B.23), and (B.24), entails (B.9).

2. Asymptotic normality of $W_{(n)}$. Denote

$$G(x) = 2 \sum_{t=1}^{p_D+1} \tilde{s}_{D,t} K_t(x) - \sum_{s,t=1}^{p_D+1} \tilde{s}_{D,s} \tilde{s}_{D,t} K_s * K_t(x).$$

Then, by the definition of $W_{(n)}$ and direct computation,

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

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

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

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

$$\frac{1}{2\sigma^2} \sigma_n^{-1} W_{(n)} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1). \quad \text{(B.25)}$$

3. Asymptotic expression $RSS_1/n = \sigma^2 + o_p(1)$. By the definition of RSS_1, we have

$$RSS_1 = \boldsymbol{\varepsilon}^T \mathbf{A}_{n2} \boldsymbol{\varepsilon} + \mathbf{m}^T \mathbf{A}_{n2} \mathbf{m} + 2\boldsymbol{\varepsilon}^T \mathbf{A}_{n2} \mathbf{m}$$
$$= \boldsymbol{\varepsilon}^T \mathbf{A}_{n2} \boldsymbol{\varepsilon} + \mathbf{B}^T \mathbf{B} + 2\mathbf{B}^T (\mathbf{W}_M - \mathbf{I}_n) \boldsymbol{\varepsilon}.$$

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

$$RSS_1/n = n^{-1} \boldsymbol{\varepsilon}^T \mathbf{A}_{n2} \boldsymbol{\varepsilon} + o_p(1).$$

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

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

where $\mathbf{R}_{n2} = O(\frac{\mathbf{1}\mathbf{1}^T}{n})$ uniformly over all elements of the matrix. Using an argument similar to that for (B.9), we can obtain

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

4. Conclusion. By step 3, (B.10), and the definition of $\lambda_n(H_0)$, we have

$$\lambda_n(H_0) - \mu_n - \frac{1}{2\sigma^2} d_{1n} + o_p(h_D^{-1}) \approx \frac{1}{2\sigma^2} W_{(n)}. \quad \text{(B.26)}$$

The combination of (B.26) and (B.25) leads to

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

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

Proof of Theorem 2

This follows by the same argument as for Theorem 1.

Proof of Theorem 3

Let $m_{\theta_0}(x_1,\ldots,x_D)$ denote the true function of $m(x_1,\ldots,x_D)$. Then the GLR statistic $\lambda_n(\mathcal{M}_\Theta)$ for testing problem (10) can be decomposed as

$$\lambda_n(\mathcal{M}_\Theta) = \lambda_n\big(m_{\theta_0}\big) - \lambda_n^*(\theta), \tag{B.27}$$

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

$$H_0' : m(x_1,\ldots,x_D) = m_{\theta_0}(x_1,\ldots,x_D) \quad \text{vs.} \quad H_1 : m \neq m_{\theta_0}(x_1,\ldots,x_D),$$

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

$$H_0' : m(x_1,\ldots,x_D) = m_{\theta_0}(x_1,\ldots,x_D) \quad \text{vs.} \quad H_1' : m(x_1,\ldots,x_D) \in \mathcal{M}_\Theta.$$

By the standard parametric hypothesis theory, the second term in (B.27) is $o_p(h_D^{-1/2})$, which is entailed by condition B. This term is negligible in the asymptotic distribution. Hence, by the same argument as for Theorem 1, the result of the theorem holds.

Lemma B.7. Suppose that condition A in Appendix A holds and that $nh_D^{2p_D+3} \to 0$. Then, under H_{1n}, there exists a $\lambda_0 > 0$ such that

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

Proof. Note that by (5),

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

and by (B.5),

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

Then, by Lemma B.4,

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

This entails that

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

Because $\|\mathbf{S}_D^*\mathbf{W}_M^{[-D]}\| < 1$ 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 λ_0 (> 0). Then

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

Using (B.11), for any $n \times 1$ scalar vector $\mathbf{Z}$, we have $\mathbf{S}_D^*\mathbf{Z} \approx \overline{Z}\mathbf{1}$, which implies that

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

Hence,

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

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

$$\begin{aligned} d_{2n} &= d_{2n}^* + \mathbf{G}_n^T\big(\mathbf{I}_n - \mathbf{W}_M^{[-D]}\big)^T(\mathbf{I}_n - \mathbf{S}_D^{*T}\mathbf{S}_D^*)\big(\mathbf{I}_n - \mathbf{W}_M^{[-D]}\big)\mathbf{G}_n \\ &\geq \lambda_0 \sum_{i=1}^n G_n^2(X_{Di}) + o(h_D^{-1}). \end{aligned} \tag{B.28}$$

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

$$d_{2n} = O\left(\sum_{i=1}^n G_n^2(X_{Di})\right) = O(h_D^{-1}). \tag{B.29}$$

Proof of Theorem 4

Write

$$\begin{aligned} &RSS_0 - RSS_1 \\ &\quad = \mathbf{Y}^T[\mathbf{A}_{n1} - \mathbf{A}_{n2}]\mathbf{Y} \\ &\quad = \boldsymbol{\varepsilon}^T[\mathbf{A}_{n1} - \mathbf{A}_{n2}]\boldsymbol{\varepsilon} + 2\boldsymbol{\varepsilon}^T[\mathbf{A}_{n1} - \mathbf{A}_{n2}]\mathbf{m} + \mathbf{m}^T[\mathbf{A}_{n1} - \mathbf{A}_{n2}]\mathbf{m} \\ &\quad \equiv I_{n1} + I_{n2} + I_{n3}. \end{aligned} \tag{B.30}$$

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

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

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

$$\begin{aligned} I_{n3} &= 2\mathbf{B}_{-D}^T\big(\mathbf{W}_M^{[-D]} - \mathbf{I}_n\big)\mathbf{G}_n \\ &\quad + \mathbf{G}_n^T\big(\mathbf{W}_M^{[-D]} - \mathbf{I}_n\big)^T\big(\mathbf{W}_M^{[-D]} - \mathbf{I}_n\big)\mathbf{G}_n \\ &\quad + O_p\left(1 + \sum_{d=1}^D nh_d^{2(p_d+1)}\right). \end{aligned} \tag{B.31}$$

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

$$\begin{aligned} I_{n2} &= 2\varepsilon^T\big(\mathbf{W}_M^{[-D]} - \mathbf{I}_n\big)^T\mathbf{B}_{-D} \\ &\quad + 2\varepsilon^T\big(\mathbf{W}_M^{[-D]} - \mathbf{I}_n\big)^T\big(\mathbf{W}_M^{[-D]} - \mathbf{I}_n\big)\mathbf{G}_n - 2\mathbf{B}^T(\mathbf{W}_M - \mathbf{I}_n)\varepsilon \\ &= 2\varepsilon^T\big(\mathbf{W}_M^{[-D]} - \mathbf{I}_n\big)^T\big(\mathbf{W}_M^{[-D]} - \mathbf{I}_n\big)\mathbf{G}_n \\ &\quad + O_p\left(1 + \sum_{d=1}^D \sqrt{n}h_d^{p_d+1}\right). \end{aligned} \tag{B.32}$$

The combination of (B.30) with (B.31) and (B.32) leads to

$$\begin{aligned} &RSS_0 - RSS_1 \\ &\quad = I_{n1} + \varepsilon^T\big(\mathbf{W}_M^{[-D]} - \mathbf{I}_n\big)^T\big(\mathbf{W}_M^{[-D]} - \mathbf{I}_n\big)\mathbf{G}_n \\ &\qquad + 2\mathbf{B}_{-D}^T\big(\mathbf{W}_M^{[-D]} - \mathbf{I}_n\big)\mathbf{G}_n \end{aligned}$$

$$+\mathbf{G}_n^T\big(\mathbf{W}_M^{[-D]}-\mathbf{I}_n\big)^T\big(\mathbf{W}_M^{[-D]}-\mathbf{I}_n\big)\mathbf{G}_n$$
$$+O_p\left(1+\sum_{d=1}^{D}\sqrt{n}h_d^{p_d+1}+\sum_{d=1}^{D}nh_d^{2(p_d+1)}\right)$$
$$\equiv I_{n1}+C_n+D_n+d_{2n}$$
$$+O_p\left(1+\sum_{d=1}^{D}\sqrt{n}h_d^{p_d+1}+\sum_{d=1}^{D}nh_d^{2(p_d+1)}\right). \quad \text{(B.33)}$$

We now assess each of the foregoing terms. Note that by (B.9),

$$I_{n1}=W_{(n)}+2\sigma^2\mu_n+o_p(h_D^{-1}). \quad \text{(B.34)}$$

Using (B.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

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

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 (B.11), we get

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

Hence

$$D_n=o_p(h_D^{-1}). \quad \text{(B.35)}$$

Observing that both $\mathbf{S}_d$ and $\mathbf{S}_{d'}^T\mathbf{S}_d$ (for $d,d'\neq D$) are of order $\mathbf{R}_{3n}\equiv O(\frac{\mathbf{1}\mathbf{1}^T}{nh_D})$, we obtain from (B.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 $X_{D1},\ldots,X_{Dn}$. By conditioning arguments and directly computing the mean and variance, the second term above is $o_p(h_D^{-1})$. Hence we have

$$d_{2n}=\sum_{i=1}^{n}G_n^2(X_{Di})+o_p(h_D^{-1}). \quad \text{(B.36)}$$

Similarly, we have

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

which, conditional on $\mathcal{X}$, is asymptotically identically distributed as $\mathcal{N}(0,\frac{C(G)}{h_D})$. This, together with (B.25) and (B.33)–(B.36), yields the result of the theorem.

Proof of Theorem 5

The argument used here is similar to that for theorem 8 of Fan et al. (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 (B.33), (B.34), and (B.35) that for $h_D\to 0$,

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

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

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

with

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

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

$$b_{1n}=\big(\sqrt{n}h_D^{(2p_D+3)/2}\sigma_n\big)^{-1}O_p\left(1+\sum_{d=1}^{D}\sqrt{n}h_d^{p_d+1}\right)=O_p(1),$$

$$b_{2n}=\big(nh_D^{(4p_D+5)/2}\sigma_n\big)^{-1}O_p\left(\sum_{d=1}^{D}nh_d^{2(p_d+1)}\right)=O_p(1),$$

and

$$b_{3n}=\big(\sqrt{h_D}\sigma_n\sigma^2\big)^{-1}\frac{1}{2}[d_{2n}+C_n].$$

Note that $E[C_n|\mathcal{X}]=0$ and

$$\operatorname{var}(C_n|\mathcal{X})=\sigma^2\mathbf{G}_n^T\mathbf{A}_{n1}^T\mathbf{A}_{n1}\mathbf{G}_n=O\left(\sum_{i=1}^{n}G_n^2(X_{Di})\right).$$

Hence we have $C_n=O_p(\sqrt{d_{2n}})$. This, together with (B.28) and (B.29), yields

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

When $h_D\le c_0^{-1/(p_D+1)}n^{-1/(2(p_D+1))}$, we have $\sqrt{n}h_D^{(2p_D+3)/2}\ge c_0nh_D^{(4p_D+5)/2}$, $\sqrt{n}h_D^{(2p_D+3)/2}\to 0$, and $nh_D^{(4p_D+5)/2}\to 0$. Thus for $h_D\to 0$ and $nh_D\to\infty$, it follows that $\beta(\alpha,\rho)\to 0$ only when $n\sqrt{h_D}\rho^2\to+\infty$. This implies that $\rho_n^2=n^{-1}h_D^{-1/2}$, and the possible minimum value of ρ_n in this setting is $n^{-(4p_D+3)/(8(p_D+1))}$. When $nh_D^{2(p_D+1)}\to\infty$, 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

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

Note that $\sup_{\mathcal{G}_n(\rho)}P_{1n}\to 0$ only when $B(h_D)\equiv nh_D^{(4p_D+5)/2}M-nh_D^{1/2}\rho^2\to-\infty$. Because $B(h_D)$ attains the minimum value

$$-\frac{4(p_D+1)}{4p_D+5}[(4p_D+5)M]^{-1/(4(p_D+1))}n\rho^{(4p_D+5)/(2(p_D+1))}$$

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

Proof of Theorem 6

Using exactly the same argument as that for the proof of Theorem 3, and noting that from Lemma B.6, $d_{1n} = O(1)$ in the current situation, we obtain the result of the theorem.

Proof of Theorem 7

Let RSS_0^* and RSS_1^* be defined similarly as RSS_0 and RSS_1, 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,

$$P\{\sigma_n^{-1}(\lambda_n^*(H_0) - \mu_n - d_{1n}) < t | \mathcal{X}, F_n\} \xrightarrow{\mathcal{L}} \Phi(t), \tag{B.37}$$

which is proven through the following three steps:

1. Noting that $\mathbf{Y}^* = \hat{\mathbf{m}}_{(-D)} + \hat{\boldsymbol{\varepsilon}}^*$, it follows that

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

2. Using the same argument as for (B.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 defined similarly as $W_{(n)}$ but with ε_i replaced by $\hat{\varepsilon}_i^*$. By an argument similar to that for Lemma B.6 [note that $\hat{\mathbf{m}}_{(-D)} = \mathbf{m}_0(1 + o_p(1))$], we have

$$\begin{aligned} d_{1n}^* &\approx \mathbf{m}_0(A_{n1} - A_{n2})\mathbf{m}_0 + 2\hat{\boldsymbol{\varepsilon}}^{*T}(A_{n1} - A_{n2})\mathbf{m}_0 \\ &\approx O_p\left(1 + \sum_d nh_d^{2(p_d+1)} + \sum_d \sqrt{n}h_d^{p_d+1}\right). \end{aligned}$$

3. 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

$$\begin{aligned} \operatorname{var}[W_{(n)}^*|F_n] &= \frac{4\sigma^4}{n^2h_D^2}\sum_{i<j}\left[\frac{1}{f(X_{Di})}G\left(\frac{X_{Dj} - X_{Di}}{h_D}\right)\right]^2 \\ &\approx 4\sigma^4\sigma_n^2, \end{aligned}$$

where $G(\cdot)$ is defined as in the proof of Theorem 1. Then, applying proposition 3.2 of de Jong (1987), we get

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

Combining steps 1–3 yields (B.37). Note that $\hat{h}_d, d = 1, \ldots, D$, satisfy the bandwidth restriction in condition A. Consistency of the bootstrap estimate of the conditional null distribution is obtained.

[Received March 2004. Revised July 2004.]

REFERENCES

Aerts, M., Claeskens, G., and Hart, J. D. (1999), "Testing the Fit of a Parametric Function," *Journal of the American Statistical Association*, 94, 869–879.

Amato, U., and Antoniadis, A. (2001), "Adaptive Wavelet Series Estimation in Separable Nonparametric Regression Models," *Statistics in Computing*, 11, 373–394.

Amato, U., Antoniadis, A., and De Feis, I. (2002), "Fourier Series Approximation of Separable Models," *Journal of Computation and Applied Mathematics*, 146, 459–479.

Ansley, C. F., and Kohn, R. (1994), "Convergence of the Backfitting Algorithm for Additive Models," *Journal of the Australian Mathematical Society*, Ser. 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," *Journal of the American Statistical Association*, 74, 829–836.

Cook, R. D., and Weisberg, S. (1982), *Residuals and Influence in Regression*, New York: Chapman & Hall.

Delgado, M. A., and González-Manteiga, W. (2001), "Significance Testing in Nonparametric Regression Based on the Bootstrap," *The Annals of Statistics*, 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," *The Annals of Statistics*, 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," *Journal of the American Statistical Association*, 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.

Gozalo, 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," *Statistica 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 & Hall.

——— (2000), "Bayesian Backfitting" (with discussion), *Statistical Science*, 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," *Mathematical Methods in Statistics*, 2, 85–114; 3, 171–189; 4, 249–268.

Le Cam, L., and Yang, G. L. (1990), *Asymptotic in Statistics: Some Basic Concepts*, New York: Springer-Verlag.

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," *The Annals of Statistics*, 27, 1443–1490.

Neumeyer, N., and Sperlich, S. (2003), "Comparison of Separable Components in Different Samples," unpublished manuscript.

Opsomer, J.-D. (2000), "Asymptotic Properties of Backfitting Estimators," *Journal of Multivariate Analysis*, 73, 166–179.

Opsomer, J.-D., and Ruppert, D. (1997), "Fitting a Bivariate Additive Model by Local Polynomial Regression," *The Annals of Statistics*, 25, 186–211.

——— (1998), "A Fully Automated Bandwidth Selection Method for Fitting Additive Models," *Journal of the American Statistical Association*, 93, 605–619.

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, D., and Wand, M. P. (1994), "Multivariate Locally Weighted Least Squares Regression," *The Annals of Statistics*, 22, 1346–1370.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge, U.K.: Cambridge University Press.

Sardy, S., Bruce, A. G., and Tseng, P. (2000), "Block Coordinate Relaxation Methods for Nonparametric Wavelet Denoising," *Journal of Computational and Graphical Statistics*, 9, 361–379.

Sardy, S., and Tseng, P. (2004), "AMlet, RAMlet, and GAMlet: Automatic Nonlinear Fitting of Additive Models, Robust and Generalised, With Wavelets," *Journal of Computational and Graphical Statistics*, 13, 283–309.

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," *The Annals of Statistics*, 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," *Journal of Multivariate Analysis*, 70, 57–65.

Zhang, C. M. (2003), " Calibrating the Degrees of Freedom for Automatic Data Smoothing and Effective Curve Checking," *Journal of the American Statistical Association*, 98, 609–628.