Annals of the Institute of Statistical Mathematics (2023) 75:71–97
https://doi.org/10.1007/s10463-022-00840-8

Nonparametric inference for additive models estimated via simplified smooth backfitting

Suneel Babu Chatla[1]

Received: 10 September 2021 / Revised: 28 April 2022 / Accepted: 6 May 2022 /
Published online: 15 July 2022

Abstract

We investigate hypothesis testing in nonparametric additive models estimated using simplified smooth backfitting (Huang and Yu, Journal of Computational and Graphical Statistics, 28(2), 386–400, 2019). Simplified smooth backfitting achieves oracle properties under regularity conditions and provides closed-form expressions of the estimators that are useful for deriving asymptotic properties. We develop a generalized likelihood ratio (GLR) (Fan, Zhang and Zhang, Annals of statistics, 29(1),153–193, 2001) and a loss function (LF) (Hong and Lee, Annals of Statistics, 41(3), 1166–1203, 2013)-based testing framework for inference. Under the null hypothesis, both the GLR and LF tests have asymptotically rescaled chi-squared distributions, and both exhibit the Wilks phenomenon, which means the scaling constants and degrees of freedom are independent of nuisance parameters. These tests are asymptotically optimal in terms of rates of convergence for nonparametric hypothesis testing. Additionally, the bandwidths that are well suited for model estimation may be useful for testing. We show that in additive models, the LF test is asymptotically more powerful than the GLR test. We use simulations to demonstrate the Wilks phenomenon and the power of these proposed GLR and LF tests, and a real example to illustrate their usefulness.

Keywords Generalized likelihood ratio · Loss function · Hypothesis testing · Local polynomial regression · Wilks phenomenon

✉ Suneel Babu Chatla
sbchatla@utep.edu

[1] Department of Mathematical Sciences, The University of Texas at El Paso, 500 West University Avenue, Texas 79968, USA

1 Introduction

Additive models are popular structural nonparametric regression models and have been widely studied in the literature Friedman and Stuetzle (1981); Hastie and Tibshirani (1990). For a random sample $\{Y_i, X_{i1}, \dots, X_{id}\}_{i=1}^n$, we consider the following additive model:

$$Y_i = \alpha_0 + \sum_{j=1}^{d} m_j(X_{ij}) + \epsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where $\{\epsilon_i, i = 1, \dots, n\}$ is a sequence of i.i.d. random variables with mean zero and finite variance σ^2 and the additive components $m_j(\cdot)$'s are unknown smooth functions which are identifiable subject to the constraints, $E[m_j(\cdot)] = 0$ for $j = 1, \dots, d$.

Additive models do not suffer greatly from the curse of dimensionality because all of the unknown functions are one-dimensional. It is possible to estimate each additive component with the same asymptotic bias and variance of a theoretical estimate which uses the knowledge of other components. Mammen et al. (1999) demonstrates that this oracle property holds true when smooth backfitting is used for estimation. Alternative estimation methods include marginal integration (Tjøstheim and Auestad, 1994; Linton and Nielsen, 1995), backfitting (Buja et al., 1989; Opsomer, 2000), penalized splines (Wood, 2017) and simplified smooth backfitting (Huang and Yu, 2019). In this study, we concentrate on simplified smooth backfitting. Simplified smooth backfitting, in addition to achieving oracle properties under regularity conditions, provides closed-form expressions of the estimators, which are convenient for deriving asymptotic properties.

After fitting an additive model, we are often interested in some hypothesis testing problems, e.g., testing whether a specific additive component in (1) is significant, or whether it may be replaced by a parametric form. For simple hypothesis problems such as component significance, the existing penalized estimation methods (Meier et al., 2009; Lian et al., 2012; Horowitz and Huang, 2013; Lian et al., 2015) provide some quick answers. However, a hypothesis testing framework is necessary for rigorous treatment. The theory for nonparametric hypothesis testing is well developed for univariate nonparametric models ($d = 1$) (Ingster, 1993; Hart, 2013) but is somewhat limited for additive models.

Härdle et al. (2004) propose a bootstrap inference procedure for generalized semiparametric additive models which is based on marginal integration (Linton and Nielsen, 1995). While their test statistic is asymptotically normal, convergence to normality is slow, so they propose a bootstrap approach to calculate its critical values. Roca-Pardiñas et al. (2005) study the testing of second-order interaction terms in generalized additive models. They propose a likelihood ratio test and an empirical-process test based on the deviance under the null and alternative hypotheses. The asymptotic distributions of their test statistics are unknown and are hard to derive, so they use a bootstrap procedure to approximate the null distribution.

Proceeding in this direction, Fan and Jiang (2005) propose a generalized likelihood ratio (GLR) test which is very simple to use. The GLR test compares the likelihood function under the null with the likelihood function under the alternative. Fan and Jiang (2005) derive the asymptotic properties of the GLR test statistic using classical backfitting (Opsomer, 2000) for model estimation. It is known that backfitting does not achieve oracle bias when the covariates are correlated. Moreover, the estimators of backfitting do not have closed-form expressions. Regardless of these drawbacks of backfitting, Fan and Jiang (2005) show that the GLR test exhibits Wilks phenomenon, which means that the null distribution is independent of the nuisance parameters-a much-desired property for likelihood ratio tests. However, their method is still limited in practice because of the disadvantages of backfitting. Better alternatives for model estimation include smooth backfitting, for which the properties of the GLR test still need to be investigated. This motivates us to study the properties of GLR test using simplified smooth backfitting for estimation.

While the GLR test has some appealing features and has been widely used in practice, it is still a nonparametric pseudotest because of the parametric assumptions on error distribution. Another promising alternative is loss function (LF)-based testing framework (Hong and Lee, 2013) which is available for univariate nonparametric models (Model (1) with $d = 1$). A loss function test compares the models under null and alternative by specifying a penalty for their difference. Many times, this is more relevant to decision-making under uncertainty because it provides the flexibility of choosing a loss function that mimics the objective of the decision-maker. Hong and Lee (2013) show that LF test is asymptotically more powerful than GLR test in terms of Pitman's efficiency criterion and possesses both optimal power and Wilks properties. Moreover, all admissible loss functions are asymptotically equally efficient under a general class of local alternatives. In spite of all these advantages, the properties of LF test still need to be investigated for nonparametric additive models ($d > 1$). To fill this void, we propose a LF test for additive model (1) and derive its asymptotic properties. More recently, although in a different context, Mammen and Sperlich (2022) proposed a backfitting test. Their test compares the nonparametric estimators obtained from smooth backftting in the L_2 norm. Using simulations, they show that the backfitting test provides very good performance in finite samples. It is worth mentioning that the proposed LF test in the study takes a similar form asymptotically.

The main contributions from this study are as follows. We develop GLR- and LF-based hypothesis testing frameworks for nonparametric additive model (1) using simplified smooth backfitting (Huang and Yu, 2019) for estimation. In Theorems 1 and 2, we show that both these test statistics follow a rescaled chi-square distribution asymptotically and achieve Wilks phenomenon. Unlike the GLR test in Fan and Jiang (2005), the proposed GLR and LF tests do not require undersmoothing to achieve Wilks phenomenon and the bandwidths that were well suited for model estimation might also be useful for testing. We also construct new F type of tests for additive models and establish the connections between GLR, LF and F-test statistics. Theorem 4 shows that GLR and LF test statistics achieve the optimal rate of convergence for nonparametric testing, $n^{-2\eta/4\eta+1}$ where $\eta = 2(p + 1)$ and p is the order of local polynomial, according to Ingster (1993). Furthermore, in Theorem 5, we show that LF test is asymptotically more powerful than GLR test. Using simulations, we validate our theoretical findings and illustrate that both GLR and LF tests are robust to error distributions to some extent.

The remainder of the paper is organized as follows. In Sect. 2, we introduce smoother matrices which are required for simplified smooth backfitting and outline the estimation algorithm. In Sect. 3, we formulate GLR and LF test statistics for nonparametric additive model. We derive the asymptotic null distributions for both test statistics and discuss their optimal power properties in Sect. 4. In Sect. 5, we evaluate the finite sample performances of both GLR and LF tests using a simulation study and a real example. We include proofs and additional numerical results in Supplementary Material.

2 Simplified smooth backfitting

In this section, we give a brief introduction to local polynomial regression (Fan and Gijbels, 1996) and describe simplified smooth backfitting algorithm, which includes smoother matrices, $\boldsymbol{H}_p^*$, of Huang and Chen (2008).

2.1 Smoother matrix

Suppose (Z_i, Y_i), $i = 1, \dots, n$, are n independent observations generated from the following model

$$Y = m(Z) + \epsilon, \tag{2}$$

where Y is a continuous response variable, Z is a continuous explanatory variable and ϵ denotes an error term with mean zero and finite variance. We choose local polynomial modeling approach (Fan and Gijbels, 1996). To estimate the conditional mean $E(Y|Z = z)$ at a grid point z, it considers a pth order Taylor expansion $m(z) + m^{(1)}(z)(Z - z) + \dots + m^{(p)}(z)(Z - z)^p/p!$, for Z in a neighborhood of z.

Let $\boldsymbol{Z}_z = [\boldsymbol{1} \quad \boldsymbol{z}_1 \cdots \boldsymbol{z}_p]_{n\times(p+1)}$ be a design matrix with $\boldsymbol{1} = (1, \dots, 1)^T$ of length n and $\boldsymbol{z}_r = ((Z_1 - z)^r, \dots, (Z_n - z)^r)^T$ for $r = 1, \dots, p$. Let $\boldsymbol{W}_z = \text{diag}\{K_h(Z_1 - z), \dots, K_h(Z_n - z)\}$ be a weight matrix with $K(\cdot)$ as a nonnegative and symmetric probability density function, and $K_h(\cdot) = K(\cdot/h)/h$ where h is a bandwidth. The local polynomial approach estimates $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)^T$, where $\beta_r = m^{(r)}/r!$, $r = 0, 1, \dots, p$, as

$$\begin{aligned}
&\min_{\boldsymbol{\beta}} \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \sum_{r=0}^{p} \beta_r \left(Z_i - z\right)^r \right)^2 K_h(Z_i - z) \\
&= \min_{\boldsymbol{\beta}} \frac{1}{n} \left(\boldsymbol{y} - \boldsymbol{Z}_z\boldsymbol{\beta}\right)^T \boldsymbol{W}_z \left(\boldsymbol{y} - \boldsymbol{Z}_z\boldsymbol{\beta}\right),
\end{aligned} \tag{3}$$

where $\boldsymbol{y} = (Y_1, \dots, Y_n)^T$. Let $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0, \dots, \widehat{\beta}_p)^T = (\boldsymbol{Z}_z^T \boldsymbol{W}_z \boldsymbol{Z}_z)^{-1} \boldsymbol{Z}_z^T \boldsymbol{W}_z \boldsymbol{y}$, denote the solution vector to (3) and the dependence of $\widehat{\boldsymbol{\beta}}$ on z is suppressed for convenience in notation.

Suppose the support of Z is [0, 1]. Let

$$K_h(u,v) = \frac{K_h(u-v)}{\int K_h(w-v)dw} I(u,v \in [0,1]), \tag{4}$$

be the boundary corrected kernel function defined in Mammen et al. (1999). It is easy to note that $\int K_h(u,v)du = 1$ for a fixed v. The smoother matrix $\boldsymbol{H}_p^*$ in Huang and Chen (2008) is based on integrating local least squares errors (3)

$$\frac{1}{n}\int \sum_{i=1}^{n}\left(Y_i - \sum_{r=0}^{p}\hat{\beta}_r\left(Z_i - z\right)^r\right)^2 K_h(Z_i,z)dz = \frac{1}{n}\boldsymbol{y}^T\left(\boldsymbol{I} - \boldsymbol{H}_p^*\right), \tag{5}$$

where $\boldsymbol{I}$ is an n- dimensional identity matrix and

$$\boldsymbol{H}_p^* = \int \boldsymbol{W}_z\boldsymbol{Z}_z(\boldsymbol{Z}_z^T\boldsymbol{W}_z\boldsymbol{Z}_z)^{-1}\boldsymbol{Z}_z^T\boldsymbol{W}_z dz, \tag{6}$$

is a smoother matrix in which the integration is taken element by element. Consequently, we may use $\boldsymbol{H}_p^*\boldsymbol{y}$ as a fitted vector for $\boldsymbol{y}$ and its ith element, $\hat{m}(Z_i) = \boldsymbol{e}_i^T\boldsymbol{H}_p^*\boldsymbol{y}$ takes the following form, for $i = 1, \dots, n$,

$$\int \left(\hat{\beta}_0 + \hat{\beta}_1(Z_i - z) + \dots + \hat{\beta}_p\left(Z_i - z\right)^p\right)K_h(Z_i,z)dz, \tag{7}$$

where $\boldsymbol{e}_i$ is the unit vector with 1 as the ith element. The estimator (7) involves double smoothing as it combines the fitted polynomials around Z_i. Therefore, at interior points $[2h, 1-2h]$, the estimator $\hat{m}(Z_i)$ achieves bias reduction. While the bias of traditional local polynomial estimator $\hat{\beta}_0$ is of order $h^{(p+1)}$ for odd p, the bias of $\hat{m}(Z_i)$ is of order $h^{2(p+1)}$ for $p = 0, 1, 2, 3$. In Sect. 2.2, we define simplified smooth backfitting estimators for additive model (1) analogous to (7). Huang and Chen (2008) and Huang and Chan (2014) discuss the properties of $\boldsymbol{H}_p^*$ and they show that it is symmetric, asymptotically idempotent and asymptotically a projection matrix. Moreover, it is nonnegative definite and shrinking.

2.2 Estimation

Huang and Yu (2019)'s simplified smooth backfitting algorithm is analogous to the classical backfitting algorithm of Buja et al. (1989) and Hastie and Tibshirani (1990) in terms of component updates. The key difference is that it uses the univariate matrices $\boldsymbol{H}_p^*$ in (6) as smoothers in backfitting algorithm.

Let $\boldsymbol{m}_j = (m_j(X_{1j}), \dots, m_j(X_{nj}))^T$ and $\boldsymbol{x}_j = (X_{1j}, \dots, X_{nj})^T$ for $j = 1, \dots, d$. Let $\mathbb{X}_j = [\boldsymbol{1}\ \boldsymbol{x}_j\ \cdots\ \boldsymbol{x}_j^{p_j}]$ for $j = 1, \dots, d$, and $\mathbb{X} = [\boldsymbol{1}\ \boldsymbol{x}_1\ \cdots\ \boldsymbol{x}_d\ \cdots\ \boldsymbol{x}_1^{p_1}\ \cdots\ \boldsymbol{x}_d^{p_d}]$, where $\boldsymbol{1}$ is the vector of ones. Let $\mathbb{X}^{[-0]} = [\boldsymbol{x}_1\ \cdots\ \boldsymbol{x}_d\ \cdots\ \boldsymbol{x}_1^{p_1}\ \cdots\ \boldsymbol{x}_d^{p_d}]$ which is same as $\mathbb{X}$ without the column of ones. For any matrix $\boldsymbol{A}$, define $\boldsymbol{A}^\perp = \boldsymbol{I} - \boldsymbol{A}$ and $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T$. Suppose $\boldsymbol{H}_{p_j,j}^*$ is an univariate smoother matrix defined as in (6) for covariate $\boldsymbol{x}_j$ with p_jth order local polynomial approximation and bandwidth h_j for $j = 1, \dots, d$.

We introduce the following spaces before stating the simplified smooth backfitting algorithm for model (1). Let $\mathcal{M}_1(\boldsymbol{H}_{p_j,j}^*)$ be a space spanned by the eigenvectors of $\boldsymbol{H}_{p_j,j}^*$

with eigenvalue 1. It includes polynomials of $\boldsymbol{x}_j$ up to p_jth order because $\boldsymbol{H}^*_{p_j,j}\boldsymbol{x}_j^k = \boldsymbol{x}_j^k$, $k = 0, 1 \ldots, p_j$, and $j = 1, \ldots, d$. Suppose $\boldsymbol{G}$ is an orthogonal projection onto the space $\mathcal{M}_1(\boldsymbol{H}^*_{p_1,1}) + \cdots + \mathcal{M}_1(\boldsymbol{H}^*_{p_d,d})$ and $\boldsymbol{G}_j$ is an orthogonal projection onto the space $\mathcal{M}_1(\boldsymbol{H}^*_{p_j,j})$, $j = 1, \ldots, d$. Then,

$$\boldsymbol{G} = \boldsymbol{P}_{\mathbb{X}} = \boldsymbol{P}_{\mathbf{1}} + \boldsymbol{P}_{\boldsymbol{P}_{\mathbf{1}}^{\perp}\mathbb{X}_{[-0]}}, \qquad \boldsymbol{G}_j = \boldsymbol{P}_{\mathbb{X}_j}, \tag{8}$$

where $\boldsymbol{P}_{\mathbb{X}} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}$ and $\boldsymbol{P}_{\mathbf{1}}$, $\boldsymbol{P}_{\mathbb{X}_j}$ and $\boldsymbol{P}_{\boldsymbol{P}_{\mathbf{1}}^{\perp}\mathbb{X}_{[-0]}}$ are defined similarly. Since the modified smoothers $\boldsymbol{H}^*_{p_j,j} - \boldsymbol{G}_j$, $j = 1, \ldots, d$ have eigenvalues in [0, 1), by Proposition 3 in Buja et al. (1989), we obtain closed form expressions for backfitting estimators. For illustration, we plot the eigenvalues of the smoother and the modified smoother using local constant ($p = 0$) and local linear terms ($p = 1$) in Figure 1. While the local constant smoother has one eigenvalue equal to 1, the local linear smoother has two eigenvalues that are equal to 1. The modified smoother has eigenvalues in [0, 1) for bandwidths that are not too small.

The simplified smooth backfitting algorithm with modified smoothers $\boldsymbol{H}^*_{p_j,j} - \boldsymbol{G}_j$ for $j = 1, \ldots, d$, and $p_j = 0, 1, 2, 3$ is stated as follows:

1. Initialize: $\boldsymbol{m}_j^* = \boldsymbol{m}_j^{(0)}$ with $\boldsymbol{m}_j^{(0)}$ in the space of $\mathcal{M}(\boldsymbol{H}^*_{p_j,j} - \boldsymbol{G}_j)$, $j = 1, \ldots, d$.
2. Cycle: $\boldsymbol{m}_j^{*^{new}} = \boldsymbol{H}^*_{p_j,j}\left(\boldsymbol{G}^{\perp}\boldsymbol{y} - \sum_{l<j}\boldsymbol{m}_l^{*^{new}} - \sum_{l>j}\boldsymbol{m}_l^{*^{old}}\right)$, $j = 1, \ldots, d$, since $\boldsymbol{G}_j\boldsymbol{m}_l^{*^{old}} = \boldsymbol{G}_j\boldsymbol{m}_l^{*^{new}} = \mathbf{0}$ and $\boldsymbol{G}_j\boldsymbol{G}^{\perp} = \mathbf{0}$.
3. Continue step 2 until the individual functions do not change. The final estimator for the overall fit is $\boldsymbol{G}\boldsymbol{y} + \hat{\boldsymbol{m}}_1^* + \ldots + \hat{\boldsymbol{m}}_d^*$.

Furthermore, we can write

$$\boldsymbol{G}\boldsymbol{y} = \left(\boldsymbol{P}_{\mathbf{1}} + \boldsymbol{P}_{\boldsymbol{P}_{\mathbf{1}}^{\perp}\mathbb{X}_{[-0]}}\right) y = \hat{\alpha}_0\mathbf{1} + \hat{\boldsymbol{g}}_1 + \ldots + \hat{\boldsymbol{g}}_d,$$

for $\hat{\boldsymbol{g}}_j = (\hat{g}_{1j}, \ldots, \hat{g}_{nj})^T$ such that $\sum_{i=1}^n \hat{g}_{ij} = 0$, $j = 1, \ldots, d$. Then, the final estimator for jth additive component is $\hat{\boldsymbol{m}}_j = (\hat{m}_j(X_{1j}), \ldots, \hat{m}_j(X_{nj}))^T = \hat{\boldsymbol{g}}_j + \hat{\boldsymbol{m}}_j^*$, for $j = 1, \ldots, d$. Since $\boldsymbol{G}_j\hat{\boldsymbol{m}}_j^* = 0$, it follows that $\sum_{i=1}^n \hat{m}_j(X_{ij}) = 0$ for $j = 1, \ldots, d$. Consequently, the estimators $\hat{\boldsymbol{m}}_j$'s are identifiable.

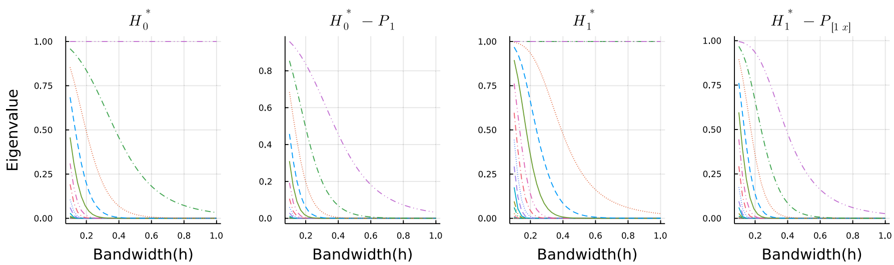

Fig. 1 Eigenvalues of smoother $\boldsymbol{H}^*_{p_j}$ and modified smoother $\boldsymbol{H}^*_{p_j} - \boldsymbol{G}_j$ for different values of bandwidths and for $p_j = 0, 1$. Here $\boldsymbol{G}_j = \boldsymbol{P}_1$ and $\boldsymbol{G}_j = \boldsymbol{P}_{[\mathbf{1}\ \boldsymbol{x}_j]}$ for $p_j = 0$ and 1, respectively

Since the smoothers $\boldsymbol{H}^*_{p_j,j}$, $j = 1, \dots, d$ are symmetric and shrinking, using the results in Buja et al. (1989), Huang and Yu (2019) show that the above algorithm converges. We provide their results in the following:

- It follows from Theorem 2 of Buja et al. (1989) that the normal equations

$$\begin{pmatrix} \boldsymbol{I} & \boldsymbol{H}^*_{p_1,1} & \boldsymbol{H}^*_{p_1,1} & \cdots & \boldsymbol{H}^*_{p_1,1} \\ \boldsymbol{H}^*_{p_2,2} & \boldsymbol{I} & \boldsymbol{H}^*_{p_2,2} & \cdots & \boldsymbol{H}^*_{p_2,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{H}^*_{p_d,d} & \boldsymbol{H}^*_{p_d,d} & \boldsymbol{H}^*_{p_d,d} & \cdots & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{m}_1 \\ \boldsymbol{m}_2 \\ \vdots \\ \boldsymbol{m}_d \end{pmatrix} = \begin{pmatrix} \boldsymbol{H}^*_{p_1,1}\boldsymbol{y} \\ \boldsymbol{H}^*_{p_2,2}\boldsymbol{y} \\ \vdots \\ \boldsymbol{H}^*_{p_d,d}\boldsymbol{y} \end{pmatrix} \tag{9}$$

  are consistent for every $\boldsymbol{y}$.
- Based on Theorem 9 of Buja et al. (1989), the backfitting algorithm converges to some solution of the normal equations (9).
- The solution is unique unless there is an exact concurvity which happens when there is a linear dependence among the eigenspaces corresponding to eigenvalue 1 of the $\boldsymbol{H}_{p_j,j}$'s.

We now obtain explicit expressions for the estimators $\widehat{\boldsymbol{m}}^*_j$, $j = 1, \dots, d$. Let $\boldsymbol{A}_j = (\boldsymbol{I} - (\boldsymbol{H}^*_{p_j,j} - \boldsymbol{G}_j))^{-1}(\boldsymbol{H}^*_{p_j,j} - \boldsymbol{G}_j)$ and $\boldsymbol{A} = \sum_{j=1}^d \boldsymbol{A}_j$. By Proposition 3 in Buja et al. (1989), we obtain $\widehat{\boldsymbol{m}}^*_j = \boldsymbol{A}_j(\boldsymbol{I} + \boldsymbol{A})^{-1}\boldsymbol{G}^{\perp}\boldsymbol{y}$. While these expressions provide estimators without requiring an iterative procedure, we still favor the backfitting algorithm because of its numerical stability. Furthermore, the backfitting approach is computationally simpler. Assume that a certain number of iterations are sufficient for convergence. In terms of computations, the explicit expressions cost $O(n^3p)$ operations, whereas the backfitting algorithm only costs $O(np)$ operations (Hastie and Tibshirani, 1990). However, this might not be a concern for small sample sizes. In Figure 2, we provide a comparison of the estimated functions $\widehat{m}^*_j$ using backfitting and explicit expressions for the model considered in Sect. 5.1. Both methods provide similar results. The computation times of backfitting and explicit expressions across different sample sizes are presented in Table 1. For small sample sizes, both

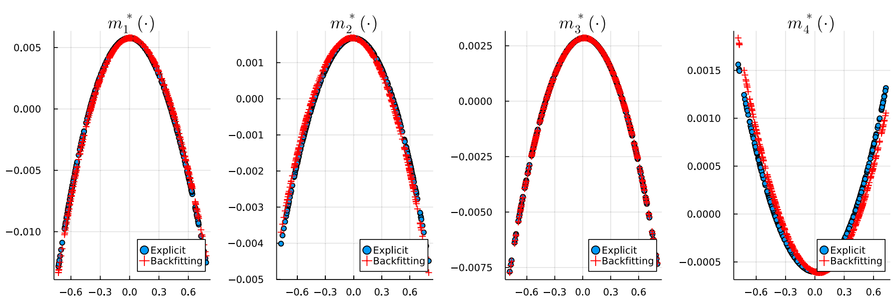

Fig. 2 The estimated additive functions (nonparametric part) for model (30) (shown later in Sect. 5.1) using the explicit expressions and the backfitting algorithm (19 iterations). Here, $n = 400$ and optimal bandwidths $(\widehat{h}_{1,opt}, \widehat{h}_{2,opt}, \widehat{h}_{3,opt}, \widehat{h}_{4,opt}) = (0.74, 1.2, 1.15, 1.16)$

Table 1 Comparison of computation times (in milliseconds) for model (30) (shown later in Sect. 5.1) using backfitting algorithm and explicit expressions

n	Backfitting(ms)	Explicit(ms)
200	664.8	682.6
400	2414	2514
800	8961	8650
1600	40717	40299
3000	169575	194010
6000	856743	1203930

approaches took about the same amount of time. However, solving explicit expressions is computationally costly than backfitting for large sample sizes. More information on the computer facilities can be found in Sect. 5.1.

At the convergence of simplified smooth backfitting algorithm, we obtain smooth backfitting estimates (or estimates at grid points) $\widehat{\beta}_{jr}$ by performing a local polynomial regression of $\boldsymbol{x}_j$ on partial residual $(\boldsymbol{G}^{\perp}\boldsymbol{y} - \sum_{l<j}\widehat{\boldsymbol{m}}_l - \sum_{l>j}\widehat{\boldsymbol{m}}_l)$. Formally,

$$\widehat{\beta}_{jr} = \boldsymbol{e}_r^T(\boldsymbol{X}_{x_j}^{jT}\boldsymbol{W}_{x_j}^j\boldsymbol{X}_{x_j}^j)^{-1}\boldsymbol{X}_{x_j}^{jT}\boldsymbol{W}_{x_j}^j\left(\boldsymbol{G}^{\perp}\boldsymbol{y} - \sum_{l<j}\widehat{\boldsymbol{m}}_l - \sum_{l>j}\widehat{\boldsymbol{m}}_l\right), \tag{10}$$

for $1 \le j \le d$, $0 \le r \le p_j$, where $\boldsymbol{e}_r$ is a unit vector with 1 at the rth position and $\boldsymbol{W}_{x_j}^j$ and $\boldsymbol{X}_{x_j}^j$ are defined similar to $\boldsymbol{W}_z$ and $\boldsymbol{Z}_z$ in (3).

Huang and Yu (2019) discuss the properties of estimators $\widehat{\boldsymbol{m}}_j^*$ and $\widehat{\beta}_{j0}$ for $j = 1, \dots, d$. They show that, estimator $\widehat{\boldsymbol{m}}_j^*$ achieves asymptotic bias of order $\sum_{j=1}^d h_j^{2(p_j+1)}$, for $p_j = 0, 1, 2, 3$, in the interior range $[2h_j, 1 - 2h_j]$ for $j = 1, \dots, d$. Similarly, asymptotic bias of $\widehat{\beta}_{j0}$ is of order $h_j^{(p_j+1)} + \sum_{k\neq j}^d h_k^{2(p_k+1)}$ if $p_j = 1$ or 3 and is of order $h_j^{(p_j+2)} + \sum_{k\neq j}^d h_k^{2(p_k+1)}$ if $p_j = 0$ or 2, for $j = 1, \dots, d$, in the interior range.

3 Proposed test statistics

In this section, we define both GLR and LF test statistics for Model (1) which are computed using the simplified smooth backfitting in Sect. 2. For simplicity in presentation, we consider the following simple hypothesis testing problem:

$$H_0 : m_d(x_d) = 0 \quad \text{vs.} \qquad H_1 : m_d(x_d) \neq 0, \tag{11}$$

which tests whether the dth predictor makes any significant contribution to the dependent variable. This testing problem is a nonparametric null versus a nonparametric alternative. It is also possible to choose more complicated hypothesis testing problems such as composite hypotheses, and nonparametric null versus parametric alternatives. We discuss some of these in our numerical results in Sect. 5.

We now introduce some matrices which will be used in our asymptotic results. Let $\boldsymbol{G}_{[-d]} = \boldsymbol{P}_{\mathbb{X}^{[-d]}}$ where $\mathbb{X}^{[-d]} = [\boldsymbol{1}\, \boldsymbol{x}_1 \, \cdots \, \boldsymbol{x}_{d-1} \, \cdots \, \boldsymbol{x}_1^{p_1} \, \cdots \, \boldsymbol{x}_{d-1}^{p_{d-1}}]$ as in (8). Define

$$\boldsymbol{C} = \boldsymbol{P}_{\boldsymbol{G}^{\perp}_{[-d]}\mathbb{X}_d^{[-0]}} + \boldsymbol{G}^{\perp}\boldsymbol{H}^{*}_{p_d,d} + \boldsymbol{H}^{*}_{p_d,d}\boldsymbol{G}^{\perp} - \boldsymbol{G}^{\perp}\boldsymbol{H}^{*}_{p_d,d}\boldsymbol{H}^{*}_{p_d,d}\boldsymbol{G}^{\perp} + O(n^{-1}h_d^{-1}\boldsymbol{I} + n^{-1}\boldsymbol{J}), \tag{12}$$

$$\boldsymbol{D} = \boldsymbol{G}^{\perp} - \sum_{j=1}^{d}\left\{\boldsymbol{H}^{*}_{p_j,j}\boldsymbol{G}^{\perp} + O(n^{-1}h_j^{-1}\boldsymbol{I} + n^{-1}\boldsymbol{J})\right\}, \tag{13}$$

$$\boldsymbol{E} = \boldsymbol{P}_{\boldsymbol{G}^{\perp}_{[-d]}\mathbb{X}_d^{[-0]}} + \boldsymbol{H}^{*}_{p_d,d}\boldsymbol{G}^{\perp} + O(n^{-1}h_d^{-1}\boldsymbol{I} + n^{-1}\boldsymbol{J}), \tag{14}$$

where $\boldsymbol{P}_{\boldsymbol{G}^{\perp}_{[-d]}\mathbb{X}_d^{[-0]}} = \boldsymbol{G}^{\perp}_{[-d]}\mathbb{X}_d^{[-0]}\left(\mathbb{X}_d^{[-0]^T}\boldsymbol{G}^{\perp}_{[-d]}\mathbb{X}_d^{[-0]}\right)^{-1}\mathbb{X}_d^{[-0]^T}\boldsymbol{G}^{\perp}_{[-d]}$, $\mathbb{X}_d^{[-0]} = [\boldsymbol{x}_d \cdots \boldsymbol{x}_d^{p_d}]_{n\times p_d}$, $\boldsymbol{J}$ is the matrix of ones, and $\boldsymbol{I}$ is an identity matrix of size n.

3.1 The generalized likelihood ratio test

We define the GLR test statistic analogous to Fan and Jiang (2005). Since the distribution of ϵ_i is unknown, pretend that the error distribution is normal, $\mathcal{N}(0, \sigma^2)$, to obtain the likelihood. However, we note that normality assumption is not needed to derive asymptotic properties for GLR statistic. In Sect. 5.1, we show that asymptotic distribution of GLR statistic is robust to error distribution to some extent. Now, the log-likelihood under model (1) is

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \alpha_0 - m_1(X_{i1}) - \cdots - m_d(X_{id})\right)^2.$$

Replacing α_0, $m_k(\cdot)$, $k = 1, \ldots, d$, with their estimates $\widehat{\alpha}_0$, $\widehat{m}_k(\cdot)$ yields

$$\ell(H_1) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}RSS_1,$$

where $RSS_1 = \sum_{i=1}^{n}\left(Y_i - \widehat{\alpha}_0 - \widehat{m}_1(X_{i1}) - \cdots - \widehat{m}_d(X_{id})\right)^2$. This likelihood function attains maximum for $\sigma^2 = \frac{1}{n}RSS_1$ which implies that $\ell(H_1) \approx -\frac{n}{2}\log(RSS_1)$. Similarly, the log-likelihood for H_0 is $\ell(H_0) \approx -\frac{n}{2}\log(RSS_0)$, with $RSS_0 = \sum_{i=1}^{n}\left(Y_i - \widehat{\alpha}_0 - \widetilde{m}_1(X_{i1}) - \cdots - \widetilde{m}_d(X_{i(d-1)})\right)^2$, and $\widetilde{m}_k(\cdot)$, $k = 1, \ldots, d-1$ are the estimators of $m_k(\cdot)$ under H_0, using the simplified smooth backfitting algorithm with the same set of bandwidths. Now, we define the GLR test statistic as

$$\lambda_n(H_0) = [\ell(H_1) - \ell(H_0)] \approxeq \frac{n}{2}\log\frac{RSS_0}{RSS_1} \approx \frac{n}{2}\frac{RSS_0 - RSS_1}{RSS_1}, \tag{15}$$

and reject the null hypothesis when $\lambda_n(H_0)$ is large. From Lemma 1 (Supplementary Material), we obtain that

$$\frac{RSS_0 - RSS_1}{RSS_1} = \frac{\boldsymbol{y}^T\boldsymbol{C}\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{D}\boldsymbol{y}}$$

for $\boldsymbol{C}$ and $\boldsymbol{D}$ defined in (12) and (13), respectively. This motivates us to consider the following F-type of test (Huang and Davidson, 2010)

$$F_\lambda = \frac{\boldsymbol{y}^T\boldsymbol{C}\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{D}\boldsymbol{y}}\frac{\mathrm{tr}(\boldsymbol{D})}{\mathrm{tr}(\boldsymbol{C})}, \tag{16}$$

where $\mathrm{tr}(\cdot)$ denotes the trace. Theorem 1 shows the statistic (16) indeed follows F distribution asymptotically.

3.2 The loss function test

The LF testing framework (Hong and Lee, 2013) compares models under H_0 and H_1 via a loss function $L : \mathbb{R}^2 \to \mathbb{R}$. This is more relevant to decision-making under uncertainty in some applications. We write the discrepancy between the models as

$$Q_n = \sum_{i=1}^{n} L\Big[\widehat{m}_+(X_{i1}, \dots, X_{id}), \widetilde{m}_+^{(-d)}(X_{i1}, \dots, X_{i(d-1)})\Big], \tag{17}$$

where $\widehat{m}_+(X_{i1}, \dots, X_{id}) = \widehat{\alpha}_0 + \widehat{m}_1(X_{i1}) + \cdots + \widehat{m}_d(X_{id})$ and $\widetilde{m}_+^{(-d)}(X_{i1}, \dots, X_{i(d-1)}) = \widehat{\alpha}_0 + \widetilde{m}_1(X_{i1}) + \cdots + \widetilde{m}_{d-1}(X_{i(d-1)})$ are the ith predicted values for the models under H_1 and H_0, respectively. Similar to Hong and Lee (2013), we consider a specific class of functions called the generalized cost-of-error function defined as $L(u, v) = d(u - v)$, where $d(\cdot)$ is twice continuously differentiable with $d(0) = 0$, $d'(0) = 0$ and $0 < d''(0) < \infty$.

We define the LF test statistic as

$$q_n(H_0) = \frac{Q_n}{n^{-1}RSS_1} \approx \frac{d''(0)/2 \sum_{i=1}^{n}(\widehat{m}_+(X_{i1}, \dots, X_{id}) - \widetilde{m}_+^{(-d)}(X_{i1}, \dots, X_{i(d-1)}))^2 + R}{n^{-1}RSS_1}, \tag{18}$$

where RSS_1 is the residual sum of squares under alternative and R is the remainder term in the Taylor expansion of $d(\cdot)$. We reject the null hypothesis when $q_n(H_0)$ is large.

Interestingly, when the estimated additive functions under H_0 and H_1 are approximately equal, that means, $\widehat{m}_j(X_{ij}) \approxeq \widetilde{m}_j(X_{ij})$ for $i = 1, \dots, n$ and $j = 1, \dots, d-1$, we obtain

$$Q_n \approx \frac{d''(0)}{2} \sum_{i=1}^{n} \widehat{m}_d^2(X_{id}) + R.$$

The form of the statistic Q_n is similar to the backfitting based test statistic proposed in Mammen and Sperlich (2022) as

$$S_n = \int \widehat{m}_d^2(x_d) f_d(x_d) dx_d, \tag{19}$$

where $f_d(\cdot)$ is the distribution of X_d. For more discussion of similar tests, please refer to Mammen and Sperlich (2022).

Based on (18), the arguments analogous to Lemma 1 (Supplementary Material) help us to define the following F-type of test statistic

$$F_q = \frac{\boldsymbol{y}^T\boldsymbol{E}^T\boldsymbol{E}\boldsymbol{y}}{\boldsymbol{y}^T\boldsymbol{D}\boldsymbol{y}}\frac{\mathrm{tr}(\boldsymbol{D})}{\mathrm{tr}(\boldsymbol{E}^T\boldsymbol{E})}, \tag{20}$$

for $\boldsymbol{E}$ defined in (14). This test is discussed in Theorem 2.

4 Asymptotic results

In this section, we develop asymptotic theory for the GLR and the LF test statistics defined in Sect. 3 under model (1). We derive the asymptotic null distributions of these test statistics when the testing problem is of the form (11) and discuss Wilks phenomenon and optimal power properties. For simplicity in theoretical arguments, we assume that all the data points are interior $[2h_j, 1-2h_j]$, $j = 1, \ldots, d$ (Huang and Yu, 2019). However, we remark that, under the conditions of Theorems 1 and 2, the additional bias terms introduced by the boundary points are of smaller order. Therefore, our theory holds when the data include boundary points.

We list some of the assumptions required for our theoretical results in the following.

Assumption 1 The densities $f_j(\cdot)$ of X_j are Lipschitz-continuous and bounded away from 0 and have bounded support Ω_j for $j = 1, \ldots, d$. The joint density of X_j and $X_{j'}$, $f_{j,j'}(\cdot,\cdot)$, for $1 \le j \ne j' \le d$, is also Lipschitz continuous and have bounded support.

Assumption 2 The kernel $K(\cdot)$ is a bounded symmetric density function with bounded support and satisfies Lipschitz condition. The bandwidth $h_j \to 0$ and $nh_j^2/(\ln n)^2 \to \infty$, $j = 1, \ldots, d$, as $n \to \infty$.

Assumption 3 The $(2p_j + 2)$−th derivative of $m_j(\cdot)$, $j = 1, \ldots, d$, exists.

Assumption 4 The error ϵ has mean 0, variance σ^2, and finite fourth moment.

Assumption 5 The loss function $d : \mathbb{R} \to \mathbb{R}^+$ has a unique minimum at 0, and $d(z)$ is monotonically nondecreasing as $|z| \to \infty$. Furthermore, $d(z)$ is twice continuously differentiable at 0 with $d(0) = 0$, $d'(0) = 0$, $M = \frac{1}{2}d''(0) \in (0, \infty)$, and $|d''(z) - d''(0)| \le C|z|$ for any z near 0.

The Assumptions 1, 2 and 4 are standard for additive models in the nonparametric smoothing literature; for example, they are similar to Fan and Jiang (2005); Huang and Chan (2014); Huang and Yu (2019). Assumption 3 is required for simplified smooth backfitting to achieve bias reduction. For example, similar assumptions are found in Huang and Chan (2014); Huang and Yu (2019). Assumption 5 is from Hong and Lee (2013) and it is required for the loss function.

4.1 Asymptotic null distributions of GLR and LF tests

Let $\mu_t = \int u^t K(u)du$ and $v_t = \int u^t K^2(u)du$ for $t = 0, 1, \dots$. Let $\boldsymbol{S} = (\mu_{i+j-2})$, $1 \le i,j \le (p_d + 1)$ be a $(p_d + 1) \times (p_d + 1)$ matrix and $s^{i,j}$ are the elements of $\boldsymbol{S}^{-1}$. Denote the convolution of $K_l(x)$ with $K_m(x)$ by $K_l * K_m$, where $K_l(x) = x^l K(x)$ for $l, m = 0, 1, \dots$. Let,

$$
\begin{aligned}
\mu_n &= \frac{1}{2}E\left(\sum_{i=1}^{n} c_{ii}\right) = \frac{|\Omega_d|}{h_d}(\sum_{l=0}^{p_d}\sum_{m=0}^{p_d} v_{l+m} s^{(m+1),(l+1)} \\
&\quad - \frac{1}{2}\int \{\sum_{l=0}^{p_d}\sum_{m=0}^{p_d} (K_l * K_m)(u)(-1)^m s^{(m+1),(l+1)}\}^2 du) + o_p(h_d^{-1}), \\
\sigma_n^2 &= \sum_{i<j} c_{ij}^2 = \frac{|\Omega_d|}{h_d}\int \{\sum_{l=0}^{p_d}\sum_{m=0}^{p_d} (K_l * K_m)(u)(-1)^m s^{(m+1),(l+1)} \\
&\quad - \frac{1}{2}\int [\sum_{l=0}^{p_d}\sum_{m=0}^{p_d} (K_l * K_m)(u+v)(-1)^m s^{(m+1),(l+1)}] \\
&\quad \times [\sum_{l=0}^{p_d}\sum_{m=0}^{p_d} (K_l * K_m)(v)(-1)^m s^{(m+1),(l+1)}]dv\}^2 du + o_p(h_d^{-1}), \text{ and} \\
r_k &= \frac{2\mu_n}{\sigma_n^2},
\end{aligned}
$$

where c_{ij} is the (i, j)th, $1 \le i,j \le n$, element of $\boldsymbol{C}$ defined in (12), $|\Omega_d|$ is the length of the support of the density $f_d(x_d)$ of X_d. In practice, the above asymptotic expressions are not required to compute the quantities μ_n and σ_n. We can compute them directly from the matrix $\boldsymbol{C}$ defined in (12) which provides a good approximation.

Hereafter, the notations "$\xrightarrow{d}$" and "$\xrightarrow{P}$" stand for convergence in distribution and probability, respectively. The following theorem describes the Wilks type of result for the GLR test conditional on the sample space $\mathcal{X}$.

Theorem 1 (GLR test) *Suppose that conditions 1–4 hold and* $0 \le p_j \le 3$, $j = 1, \dots, d$. *Then, under* H_0 *for the testing problem* (11)

$$
P\left\{\sigma_n^{-1}\left(\lambda_n(H_0) - \mu_n - \frac{1}{2\sigma^2}d_{1n}\right) < t|\mathcal{X}\right\} \xrightarrow{d} \boldsymbol{\Phi}(t), \tag{21}
$$

where $d_{1n} = O_p\left(1 + \sum_{j=1}^{d} nh_j^{4(p_j+1)} + \sum_{j=1}^{d} \sqrt{n}h_j^{2(p_j+1)}\right)$ *and* $\boldsymbol{\Phi}(\cdot)$ *is the standard normal distribution. Furthermore, if* $nh_j^{4(p_j+1)}h_d \to 0$ *for* $j = 1, \dots, d$, *conditional on* $\mathcal{X}$, $r_k\lambda_n(H_0) \to \chi^2_{r_k\mu_n}$ *as* $n \to \infty$. *Similarly,*

$$
F_\lambda = \frac{2\lambda_n(H_0)\mathrm{tr}(\boldsymbol{D})}{n\,\mathrm{tr}(\boldsymbol{C})} \to F_{\mathrm{tr}(C),\mathrm{tr}(D)}, \tag{22}
$$

as $n \to \infty$, *where* tr(C) *and* tr(D) *are the corresponding degrees of freedom.*

Theorem 1 gives the asymptotic null distribution of the GLR test statistic for the testing problem (11) under H_0. In our opinion, the asymptotic expression for d_{1n} is complicated and might not be necessary.

Remark 1 The factors r_k and μ_n in Theorem 1 do not depend on the nuisance parameters and nuisance functions. Therefore, the GLR test statistic λ_n achieves the Wilks phenomenon that its asymptotic distribution does not depend on nuisance parameters and nuisance functions. Theorem 1 is different from Theorem 1 of Fan and Jiang (2005) because it uses simplified smooth backfitting instead of backfitting for estimation of additive components.

Remark 2 Theorem 1 shows that the bias d_{1n} is negligible under the condition $C_1 : nh_j^{4(p_j+1)}h_d \to 0$ which is different from the condition $C_2 : nh_j^{2(p_j+1)}h_d \to 0$ in Theorem 1 of Fan and Jiang (2005). Suppose $h_j^{p_j+1} = O(h_d^{p_d+1})$, then the proposed GLR test achieves Wilks phenomenon for the bandwidths of order $h_j \sim n^{-1/(2p_j+3)}$ for odd p_j, which are the optimal bandwidths used for estimation in Fan and Jiang (2005), while their GLR test statistic does not. To see this, consider $p_j = 1$, then $h_{opt} \sim n^{-1/5}$, the first condition $C_1 : nh^9 \sim nn^{-9/5} = n^{-4/5} = o(1)$ holds where as the second condition $C_2 : nh^5 \sim nn^{-5/5} = O(1)$ does not hold.

We now derive the asymptotic null distribution of the LF test statistic. Let $\boldsymbol{e}_k = (e_{1k}, \ldots, e_{nk})^T$ where e_{ij} is the (i,j)th, $1 \le i,j \le n$, element of $\boldsymbol{E}$ defined in (14). Define

$$\begin{aligned}
v_n = E(\mathrm{tr}(\boldsymbol{E}^T\boldsymbol{E})) = E\left(\sum_{i=1}^{n} \boldsymbol{e}_i^T \boldsymbol{e}_i\right) \\
&= \frac{|\Omega_d|}{h_d} \int \left\{ \sum_{l=0}^{p_d} \sum_{m=0}^{p_d} (K_l * K_m)(u)(-1)^m s^{(m+1),(l+1)} \right\}^2 \\
&\quad du + o_p(h_d^{-1}),
\end{aligned}$$

$$\begin{aligned}
\delta_n^2 = \sum_{j \ne j'}^{n} (\boldsymbol{e}_j^T \boldsymbol{e}_{j'})^2 = \frac{|\Omega_d|}{h_d} \int \Bigg\{ \int \left[\sum_{l=0}^{p_d} \sum_{m=0}^{p_d} (K_l * K_m)(u+v)(-1)^m s^{(m+1),(l+1)} \right] \\
\times \left[\sum_{l=0}^{p_d} \sum_{m=0}^{p_d} (K_l * K_m)(v)(-1)^m s^{(m+1),(l+1)} \right] dv \Bigg\}^2 \\
du + o_p(h_d^{-1}),
\end{aligned}$$

$$\text{and } s_k = \frac{2v_n}{\delta_n^2}.$$

Denote $M = d''(0)/2$ where $d(\cdot)$ is the loss function given in Sect. 3.2. The following theorem describes the Wilks type of result for the LF test statistic conditional on the sample space $\mathcal{X}$.

Theorem 2 (LF test) *Suppose that conditions 1–5 hold and* $0 \le p_j \le 3$, $j = 1, \ldots, d$. *Then, under* H_0 *for the testing problem* (11)

$$P\left\{ \delta_n^{-1} \left(\frac{q_n(H_0)}{M} - v_n - \frac{1}{\sigma^2} b_{1n} \right) < t | \mathcal{X} \right\} \xrightarrow{d} \boldsymbol{\Phi}(t), \tag{23}$$

where $b_{1n} = O_p\left(1 + \sum_{j=1}^d nh_j^{4(p_j+1)}\right)$. *Furthermore, if* $nh_j^{4(p_j+1)} h_d \to 0$ *for* $j = 1, \ldots, d$, *conditional on* $\mathcal{X}$, $s_k M^{-1} q_n(H_0) \to \chi^2_{s_k v_n}$ *as* $n \to \infty$. *Similarly,*

$$F_q = \frac{q_n(H_0)}{Mn} \frac{\mathrm{tr}(\boldsymbol{D})}{\mathrm{tr}(\boldsymbol{E}^T \boldsymbol{E})} \to F_{\mathrm{tr}(\boldsymbol{E}^T\boldsymbol{E}), \mathrm{tr}(\boldsymbol{D})}, \tag{24}$$

as $n \to \infty$.

Remark 3 Theorem 2 shows that the factors s_k and v_n do not depend on the nuisance parameters and nuisance functions. Therefore, like the GLR statistic, the LF test statistic also enjoys Wilks phenomenon that its asymptotic distribution does not depend on nuisance parameters and nuisance function. Further, since Wilks phenomenon is achieved for optimal bandwidths $h_j \sim n^{-1/(2p_j+3)}$, for odd p_j, $j = 1, \ldots, d$, undersmoothing may not be necessary.

Remark 4 The LF test statistic includes an extra scaling constant M, which is the curvature of the loss function. When M is correctly specified, the asymptotic distribution of the scaled statistic does not depend on it. However, the choice of M is irrelevant if the conditional bootstrap method (Supplementary Material) is used to simulate the null distribution. We further validate this using simulations in Sect. 5.1. We also provide more discussion on the efficiency of loss functions in Theorem 3.

Unlike GLR test statistic, the LF test statistic includes only the second-order term in the Taylor expansion because the first-order term vanishes to 0 under H_0. For univariate nonparametric model (1) with $d = 1$, Hong and Lee (2013) argue that not having first-order term could be one reason for LF test to be asymptotically more powerful than GLR test. In Theorem 5, we show that similar result holds for the nonparametric additive model (1). We now discuss the optimal power properties of the proposed test statistics in the following section.

4.2 Power of GLR and LF tests

We consider the framework of Fan et al. (2001) and Fan and Jiang (2005) to study the power of GLR and LF tests. Assume that $h_d = o(n^{-1/(4p_d+5)})$, so that the second term in both d_{1n} and b_{1n} is of smaller order than σ_n and δ_n, respectively. We note that the optimal bandwidth for the testing problem (11) is $h_d = O(n^{-2/(8p_d+9)})$ (to be shown later in Theorem 4), which satisfies the condition $h_d = o(n^{-1/(4p_d+5)})$. Under these assumptions, Theorems 1 and 2 lead to the following approximate level α tests for GLR and LF test statistics, respectively

$$\phi_{\lambda_n} = I\{\lambda_n(H_0) - \mu_n \geq z_\alpha \sigma_n\}, \qquad \phi_{q_n} = I\left\{\frac{q_n(H_0)}{M} - \nu_n \geq z_\alpha \delta_n\right\}.$$

Let $\mathcal{M}_n$ be a class of functions such that any $M_n \in \mathcal{M}_n$ satisfy the following regularity conditions as stated in Fan and Jiang (2005):

$$\text{var}(M_n^2(X_d)) \leq K(E[M_n^2(X_d)])^2, \qquad nE[M_n^2(X_d)] > K_n \to \infty, \tag{25}$$

for some constants $K > 0$ and $K_n \to \infty$. Let $\eta = 2(p_d + 1)$ with $0 \leq p_d \leq 3$. Define a class of functions,

$$\mathcal{M}_n(\rho;\eta) = \left\{M_n \in \mathcal{M}_n : E[M_n^2(X_d)] \geq \rho^2, E[\nabla^r M_n(X_d)]^2 \leq R_*^2 \text{ with } r \leq \eta\right\}, \tag{26}$$

for a given $\rho > 0$, where $\nabla^r M_n(X_d)$ is the rth derivative of M_n and R_* is some positive constant. Consider the contiguous alternative of the form

$$H_{1n} : m_d(X_d) = M_n(X_d), \tag{27}$$

where $M_n(X_d) \to 0$ and $M_n \in \mathcal{M}_n(\rho;\eta)$.

The following theorem is useful to approximate the power of GLR and LF tests under the contiguous alternative (27).

Theorem 3 *Suppose* $E\{M_n(X_d)|X_1, \dots, X_{d-1}\} = 0$ *and* $h_d \cdot \sum_{i=1}^n M_n^2(X_{id}) \xrightarrow{P} C_M$ *for some constant* C_M. *Suppose* $0 \leq p_j \leq 3$, *for* $j = 1, \dots, d$.

(i) [*GLR test*] *Suppose that conditions* 1–4 *hold. Under* H_{1n} *for the testing problem* (11),

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

where μ_n, d_{1n} and σ_n are same as those in Theorem 1 and

$$d_{2n} = \sum_{i=1}^n M_n^2(X_{id})(1 + o_p(1)).$$

(ii) [*LF test*] *Suppose that conditions* 1–5 *hold. Under* H_{1n} *for the testing problem* (11),

$$P\left\{\delta_n^{-1}\left(\frac{q_n(H_0)}{M} - \nu_n - \frac{b_{1n} + b_{2n}}{\sigma^2}\right) < t|\mathcal{X}\right\} \xrightarrow{d} \boldsymbol{\Phi}(t),$$

where ν_n, b_{1n} and δ_n are same as those in Theorem 2 and

$$b_{2n} = \sum_{i=1}^n M_n^2(X_{id})(1 + o_p(1)).$$

Theorem 3 shows that when $nh_j^{4(p_j+1)}h_d \to 0$, $j = 1, \ldots, d$, the alternative distributions are independent of the nuisance functions $m_j(x_j)$, $j \neq d$, and this helps us to compute the power of the tests via simulations over a large range of bandwidths with nuisance functions fixed at their estimated values.

It is interesting to note that the noncentrality parameters in part (ii) of Theorem 3 are independent of the curvature parameter $M = d''(0)/2$ of the loss function $d(\cdot)$. This implies that, as discussed in Hong and Lee (2013), all loss functions satisfying Assumption 5 are asymptotically equally efficient under H_1 in terms of Pitman's efficiency criterion [Pitman (2018), Chapter 7].

The maximum of the probabilities of type II errors is

$$\beta_{\lambda_n}(\alpha,\rho) = \sup_{M_n \in \mathcal{M}_n(\rho;\eta)} \beta_{\lambda_n}(\alpha, M_n), \qquad \beta_{q_n}(\alpha,\rho) = \sup_{M_n \in \mathcal{M}_n(\rho;\eta)} \beta_{q_n}(\alpha, M_n), \tag{28}$$

where $\beta_{\lambda_n}(\alpha, M_n) = P(\phi_{\lambda_n} = 0 | m_d = M_n)$ and $\beta_{q_n}(\alpha, M_n) = P(\phi_{q_n} = 0 | m_d = M_n)$ are the probabilities of type II errors at the alternative $H_{1n} : m_d = M_n$. Use $\beta(\alpha,\rho)$ to denote either $\beta_{\lambda_n}(\alpha,\rho)$ or $\beta_{q_n}(\alpha,\rho)$ and ϕ to denote ϕ_{λ_n} or ϕ_{q_n}. As mentioned in Fan et al. (2001) and Fan and Jiang (2005), the minimax rate of ϕ_{λ_n} or ϕ_{q_n} 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 exists $\alpha > 0$ and $\beta > 0$ such that for any $c > 0$, $P(\phi = 1 | m_d = M_n) = \alpha + o(1)$ and $\liminf_n \beta(\alpha, c\rho_n^*) > \beta$.

The following theorem provides the rate with which the alternatives can be detected by GLR (ϕ_{λ_n}) and LF (ϕ_{q_n}) tests. The convergence rate depends on bandwidth.

Theorem 4 *Under conditions 1–5, if* $h_k^{2(p_k+1)} = O(h_d^{2(p_d+1)})$ *and* $0 \leq p_k \leq 3$, *for* $k = 1, \ldots, d-1$, *then for the testing problem* (11), *both GLR and LF tests can detect alternatives with rate* $\rho_n = n^{-\frac{4(p_d+1)}{8p_d+9}}$ *when* $h_d = c_* n^{-\frac{2}{8p_d+9}}$ *for some constant* c_*.

Remark 5 For the class of alternatives $\mathcal{M}_n(\rho;\eta)$ in (26), the rate of convergence for nonparametric hypothesis testing according to the formulations of Ingster (1993) and Spokoiny (1996) is $n^{-\frac{2\eta}{4\eta+1}}$ where η is the smoothness parameter. Since $\eta = 2(p_d + 1)$ in this study, the GLR and LF tests are asymptotically optimal based on their rates given in Theorem 4. Our rates are different from the rates in Theorem 5 of Fan and Jiang (2005) because of different smoothness parameter $\eta = (p_d + 1)$ considered in their study. For this reason, the optimal bandwidth for testing in our study $n^{-\frac{2}{8p_d+9}}$ which is also different from $n^{-\frac{2}{4p_d+5}}$ in Fan and Jiang (2005).

Remark 6 Based on Theorems 1 and 2, the assumption on bandwidths $nh_j^{4(p_j+1)}h_d = o(1)$, $j = 1, \ldots, d$, is required to ensure Wilks property for both GLR and LF tests. This is true for a collection of bandwidths $h_j \in (0, n^{-\frac{1}{4p_j+4}}]$ which includes the optimal bandwidths $n^{-\frac{1}{2p_j+3}}$ used in backfitting (Fan and Jiang, 2005).

With our method, the bandwidths well suited for curve estimation might also be useful for testing.

We now show that the LF test is asymptotically more powerful than the GLR test. For ease of exposition, we assume that $p_j = 0,\ j = 1, \dots, d$. Without loss of generality, let $M_n(x_d) = n^{-1/2} h_d^{-1/2} g(x_d)$ which satisfy the condition in Theorem 3. We now compare the relative efficiency between the LF test statistic q_n and the GLR test statistic λ_n under the class of local alternatives

$$H_n\ :\ m_d(x_d) = n^{-1/2} h_d^{-1/2} g(x_d), \tag{29}$$

where $E(g(X_d)|X_1, \dots, X_{d-1}) = 0$ and $\sum_{i=1}^n g^2(X_{id}) = O_p(h_d^{-1})$. While Theorem 4 shows that the GLR and the LF tests achieve optimal rate of convergence in the sense of Ingster (1993) and Spokoiny (1996), Theorem 5 provides that under the same set of regularity conditions, the LF test is asymptotically more powerful than the GLR test under H_n in (29).

Theorem 5 [*Relative efficiency*] *Suppose Conditions 1–5 hold,* $h \propto n^{-\omega}$ *for* $\omega \in (0, 1/(4p_d + 5))$ *and* $p_j = 0$ *for* $j = 1, \dots, d$*. Then Pitman's relative efficiency of the LF test over the GLR test under* H_n *in* (29) *is given by*

$$\mathrm{ARE}(q_n, \lambda_n) = \left[\frac{\int \left\{ 2(K_0 * K_0)(u) - \int (K_0 * K_0)(u+v)(K_0 * K_0)(v) dv \right\}^2 du}{\int \left\{ \int (K_0 * K_0)(u+v)(K_0 * K_0)(v) dv \right\}^2 du} \right]^{1/(2-3\omega)}.$$

The asymptotic relative efficiency $ARE(q_n, \lambda_n)$ *is larger than* 1 *for any kernel satisfying Condition 2 and* $K(\cdot) \le 1$.

Remark 7 Theorem 5 shows that the Pitman's relative efficiency of the LF test over the GLR test is larger than 1 for $p_j = 0,\ j = 1, \dots, d$, which means that the LF test is asymptotically more efficient than the GLR test. Given the complicated expressions of σ_n and δ_n, the extension of Theorem 5 to general $p_j = 1, 2, 3$, is not straightforward. We defer this for future research.

Remark 8 The result in Theorem 5 does not imply that the GLR test is not useful. The GLR test is a natural extension of classical Likelihood Ratio test with many desirable features and has been widely used in the literature. As stated in Hong and Lee (2013), same bandwidths and same kernel functions $K(\cdot)$ are required for the relative efficiency of q_n over λ_n to hold. Therefore, it might be possible that both test statistics achieve similar efficiencies under different bandwidths and kernel functions. In our simulations in Sect. 5.1, we observe that the statistic q_n achieves larger powers than the statistic λ_n.

Remark 9 The result in Theorem 5 is new to the literature. While Theorem 4 in Hong and Lee (2013) discusses the asymptotic relative efficiency of q_n over λ_n for

Nadaraya-Watson estimator in an univariate model, the proposed Theorem 5 discusses the relative efficiency for similar type of estimators using $\boldsymbol{H}^*_{0,j}$, $j = 1, \dots, d$, in additive models.

5 Numerical comparison of GLR and LF tests

In this section, we evaluate the performance of GLR and LF tests in finite samples. Using simulations, we demonstrate the Wilks phenomenon and examine the effect of error distribution on the performances of GLR and LF tests. Local linear smoothing with Gaussian kernel is considered in all the simulations. We use software Julia (Bezanson et al., 2017) to carry out simulations and data analysis. We also illustrate the usefulness of the proposed statistics using Boston housing data. Additional results are presented in Section S3.2 of supplementary Material.

5.1 Simulations

We mimic the simulation designs in Fan and Jiang (2005) and Huang and Yu (2019). Consider the additive model,

$$Y = m_1(X_1) + m_2(X_2) + m_3(X_3) + m_4(X_4) + \epsilon, \tag{30}$$

where $m_1(X_1) = 0.5 - X_1^2 + 3X_1^3$, $m_2(X_2) = \sin(\pi X_2)$, $m_3(X_3) = X_3(1 - X_3)$, $m_4(X_4) = \exp(2X_4 - 1)$, and ϵ is distributed as $\mathcal{N}(0, 1)$. For the covariates X_1, X_2, X_3 and X_4, we first simulate normally distributed random variables Z_1, Z_2, Z_3 and Z_4 with mean [0, 0, 0, 0] and covariance $0.4\boldsymbol{I}_4 + 0.6\mathbf{1}\mathbf{1}^T$ and project them back on to $[-1, 1]$ using the transformation $X_i = 2\tan^{-1}(Z_i)/\pi, i = 1, 2, 3, 4$.

We consider the null hypothesis $H_0 : m_2(x_2) = 0$ and treat $m_1(x_1)$, $m_3(x_3)$ and $m_4(x_4)$ as nuisance functions. For comparison purpose, we implement the GLR test in Fan and Jiang (2005) which is based on classical backfitting; hereafter, denoted as GLR(FJ) and the corresponding test statistic as $\lambda_n(FJ)$. We also implement the backfitting-based test (19) proposed in Mammen and Sperlich (2022); hereafter, denoted as SB. The SB statistic in (19) is approximated using Riemann sum as

$$S_n = \sum_{i=1}^{n} \hat{m}_2^2(X_{(i)2})\hat{f}_2(X_{(i)2})(X_{(i)2} - X_{(i-1)2}), \text{ with } X_{(0)2} := X_{(1)2},$$

where $X_{(i)2}$ is the ith order statistic of X_2. An R (R Core Team, 2021) package *wsbackfit* (Roca-Pardinas et al., 2021) is used to estimate the additive component $\hat{m}_2$ and kernel density estimate $\hat{f}_2$ which uses the bandwidth considered for $\hat{m}_2(\cdot)$.

We compute the optimal bandwidths using the following cross-validation procedure which is defined similar to Nielsen and Sperlich (2005).

1. Fit the additive model for initial values of bandwidths.
2. For any direction (covariate), consider the corresponding partial residual as a response variable and use an Akaike Information Criterion-based smoothing

parameter selection method (xHurvich et al., 1998) to determine the optimal bandwidth in univariate local linear regression.
3. Refit the model with the updated bandwidths and proceed to the step 2 choosing a different direction.
4. Obtain the optimal bandwidths at the convergence of the above procedure.

To demonstrate Wilks phenomenon for GLR, LF, and F tests, we choose three levels of bandwidths $h_1 = h_{1,\text{opt}}/3, h_{1,\text{opt}}, 1.5h_{1,\text{opt}}$ and $h_2 = h_{2,\text{opt}}, h_3 = h_{3,\text{opt}}$ and $h_4 = h_{4,\text{opt}}$. Similarly, we consider three levels of $m_1(X_1)$ to show that the proposed tests do not depend on the nuisance function $m_1(X_1)$:

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

where $\beta = -1.5, 0, 1.5$. For LF statistic, we consider the following class of LINEX functions (Hong and Lee, 2013):

$$d(z) = \frac{t}{s^2}\left[\exp(sz) - (1 + sz)\right], \tag{31}$$

where $d(z)$ is an asymmetric loss function for each pair of parameters (s, t). The magnitude of s, which is a shape parameter, controls the degree of asymmetry. The parameter t is a scale factor.

We draw 1000 samples of 100 observations from (30) and for each sample, we compute the scaled GLR, LF, and F test statistics. The distributions of scaled GLR, LF, and F test statistics among 1000 simulations are obtained via a kernel estimate using the rule of thumb bandwidth $h = 1.06sn^{-.2}$, where s is the standard deviation of the test statistics. Figure 3 shows the estimated densities for the scaled GLR

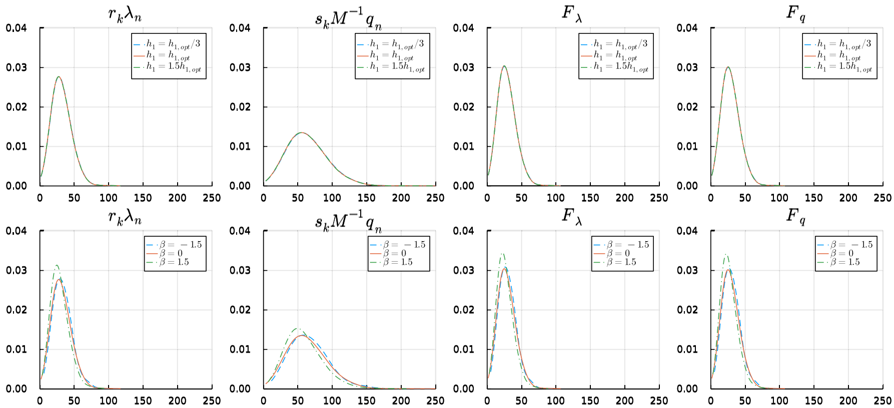

Fig. 3 Estimated densities for the scaled GLR, LF, and F test statistics among 1000 simulations. (Top row) With fixed $(h_2, h_3, h_4) = \{h_{2,\text{opt}}, h_{3,\text{opt}}, h_{4,\text{opt}}\}$, but different bandwidths for h_1 (- - $h_1 = h_{1,\text{opt}}/3$; —$h_1 = h_{1,\text{opt}}$; $- \cdot -h_1 = 1.5h_{1,\text{opt}}$). (Bottom row) With different nuisance functions and optimal bandwidths $h_j = h_{j,\text{opt}}$, $j = 1, 2, 3, 4$, (- - $\beta = -1.5$; —$\beta = 0$; $- \cdot -\beta = 1.5$). The LF test considers the class of LINEX functions (31)with $s = 0$ and $t = 1$

and LF, and F test statistics. Plots in the top row show that the null distributions of scaled GLR and LF statistics follow a chi-squared distribution over a wide range of bandwidth values for h_1. It is interesting to note that $r_k \approx s_k$, $\mu_n \approx \nu_n$ and $M = 1/2$. Due to the extra scaling constant $M = 1/2$, the distribution of $s_k M^{-1} q_n$ seems like a scaled version of the distribution of $r_k \lambda_n$. Both F statistics, F_λ and F_q, are computed using the $\boldsymbol{C}$, $\boldsymbol{D}$ and $\boldsymbol{E}$ matrices defined in (12,13,14); the results also illustrate that they provide a good approximation. Similarly, plots from the bottom row demonstrate the Wilks phenomenon for scaled GLR, LF, and F test statistics, as their null distributions are nearly the same for three different choices of the nuisance functions for $m_1(\cdot)$. For LF test, we consider the LINEX loss function (31) with the choice $s = 0$ and $t = 1$.

For power comparison among GLR, LF, F, GLR(FJ), and SB tests, we evaluate the power for a sequence of alternative models indexed by θ,

$$H_\theta : m_{2,\theta} = \theta \sin(\pi X_2), \qquad 0 \le \theta \le 1, \tag{32}$$

where $E(\sin(\pi X_2)|X_1, X_3, X_4) = 0$, $\theta = 0$ gives the null model and $\theta > 0$ makes the alternative reasonably far away from the null model. For each given value of θ, we consider 2000 Monte Carlo replicates for calculation of the critical values via conditional bootstrap method which is described in Section S3 of Supplementary Material. The rejection percentage values are computed based on 500 simulations. When $\theta = 0$, the alternative is identical to the null hypothesis and the power is approximately equal to the significance level $\alpha = 0.05$ or 0.01. Furthermore, to illustrate the influence of different error distributions on the power of GLR, LF, and F tests, we consider model (30) with different error distributions of ϵ, namely $\mathcal{N}(0, 1)$, $t(5)$, $\chi^2(5)$ and $\chi^2(10)$. The distributions of test statistics among 1000 simulations are provided in Figure S4.1 in Supplementary Material. The estimated densities are approximately similar across different error distributions.

The power of GLR, LF, F, GLR(FJ), and SB tests for the alternative model sequence in (32) at the significance level $\alpha = 0.05$ is provided in Figure 4 for $n = 100$ and $n = 400$. Figure 4 illustrates that both GLRT and LFT differentiate the null and alternative hypotheses with high power while not being sensitive to error distributions. When $\theta = 0$, the alternative is identical to the null, and hence, the power should be approximately equal to α (0.05); this is evident from the results. This gives an indication that Monte Carlo approach yields the correct estimator of the null distribution. Based on Theorem 4, we consider the bandwidths that are optimal for testing $h_j = S_{X_j} n^{-2/17}$, $j = 1, 2, 3, 4$. Here, S_{X_j} is the standard deviation of X_j. One important observation is that the results for the statistic F_λ exhibit some variation. However, in Fig. 5 where the optimal bandwidths for model estimation (cross-validation) are considered, we observe that F_λ performs very similar to other statistics. We note that the optimal bandwidths are larger than $S_{X_j} n^{-2/17}$; it seems F_λ is not stable for smaller bandwidths due to approximation. Overall, Figures 4 and 5 illustrate that the proposed methods in the study work well with the finite samples and comparable to other existing methods in the literature.

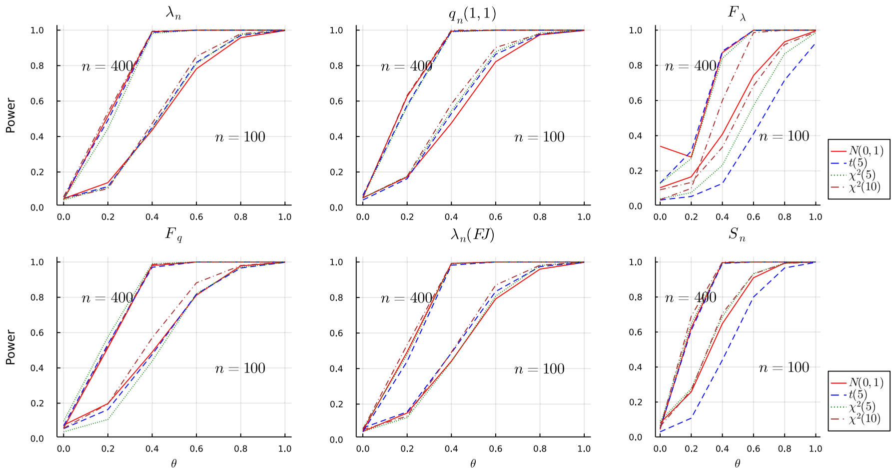

Fig. 4 Power of the tests under alternative model sequence (32) using optimal bandwidths for testing, $S_X n^{-2/17}$, at 5% level of significance. Only the LF test with LINEX loss function (31) for $s = 1, t = 1$ is reported. The power values are similar for other choices of s and t

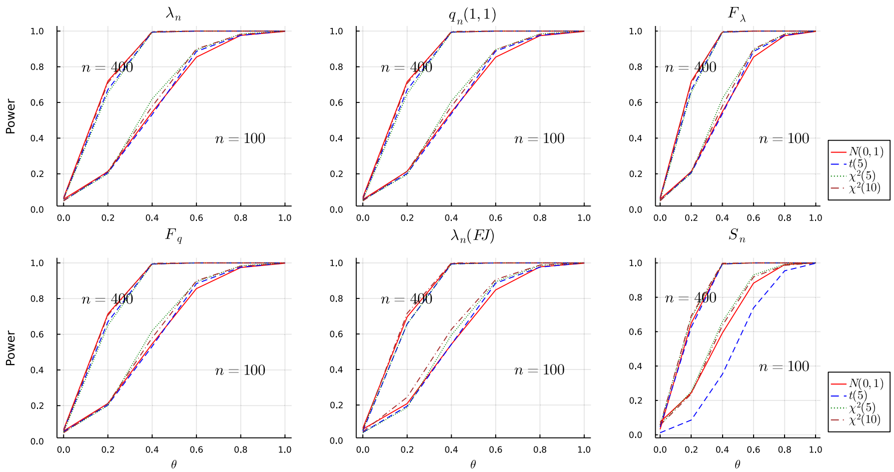

Fig. 5 Power of the tests under alternative model sequence (32) using optimal bandwidths for estimation (cross-validation) at 5% level of significance. Only the LF test with LINEX loss function (31) for $s = 1, t = 1$ is reported. The power values are similar for other choices of s and t

We also provide the power comparison of the above methods at 1% level of significance. The results are available in Figures S4.2 and S4.3 in Supplementary Material. The findings remain similar.

5.2 Boston housing data analysis

To demonstrate the usefulness of the proposed GLR and LF tests, we consider the Boston housing data. These data include the information collected by the U.S Census Service regarding housing in the area of Boston Mass, and originally published in Harrison and Rubinfeld (1978). It contain the median values of 506 homes along with 13 sociodemographic and related variables. These data have been previously used in the literature to benchmark different algorithms and to illustrate different methodologies. For example, please see Belsley et al. (2005), Breiman and Friedman (1985), Opsomer and Ruppert (1998) and Fan and Jiang (2005). For the sake of easy comparison, we consider the same dependent and independent variables used in Fan and Jiang (2005).

- *MV*, median value of owner-occupied homes in \$1000's
- *RM*, average number of rooms per dwelling
- *TAX*, full-value property tax rate ($\$/\$10,000$)
- *PTRATIO*, pupil/teacher ratio by town school district
- *LSTAT*, proportion of population that is of "lower status"(%).

Opsomer and Ruppert (1998) and Fan and Jiang (2005) analyze these data by considering the following four-dimensional additive model,

$$E[MV \mid X_1, X_2, X_3, X_4] = m_0 + m_1(X_1) + m_2(X_2) + m_3(X_3) + m_4(X_4), \tag{33}$$

where $X_1 = RM$, $X_2 = \log(TAX)$, $X_3 = PTRATIO$, and $X_4 = \log(LSTAT)$. We use simplified smooth backfitting algorithm in Sect. 2.2 with local linear smoothing to estimate model (33) after six outliers ($\hat{\epsilon}_i < -11$ or $\hat{\epsilon}_i > 12$) are removed. To alleviate the effect of sample size on p-value, we take a random sample $n = 200$ observations for hypothesis testing. The optimal bandwidths are selected using the cross-validation procedure described in Sect. 5.1 which uses the AIC to find optimal bandwidth in each direction. For comparison, we also fit model (33) using classical backfitting in Fan and Jiang (2005), smooth backfitting (Mammen and Sperlich, 2022; Roca-Pardinas et al., 2021) with *wsbackfit* package in R and penalized splines approach with *mgcv* package in R (Wood and Wood, 2015; R Core Team, 2021).

Figure 6 shows the estimated additive functions along with the partial residuals. The simplified smooth backfitting algorithm estimates the additive functions $\hat{m}_j$ as a sum of two functions, i.e., $\hat{m}^*_j$ and $\hat{g}_j$, where $\hat{m}^*_j$ is the purely nonparametric part and $\hat{g}_j$ is the parametric part corresponding to the eigenvectors of eigenvalue 1 of the smoother $\boldsymbol{H}^*_{1,j}$. For comparison, Figure 7 includes the fits from gam() function in *mgcv* package in R, from sback() function in *wsbackfit* package in R, and from the classical backfitting of Fan and Jiang (2005). Figure 7 shows that the fits from the these methods are very similar. From both Figures 6 and 7, we find that the additive components for all the variables except *RM* exhibit the following parametric forms:

$$m_i(X_i) = a_i + b_i X_i \qquad \text{for } i = 2, 3, 4. \tag{34}$$

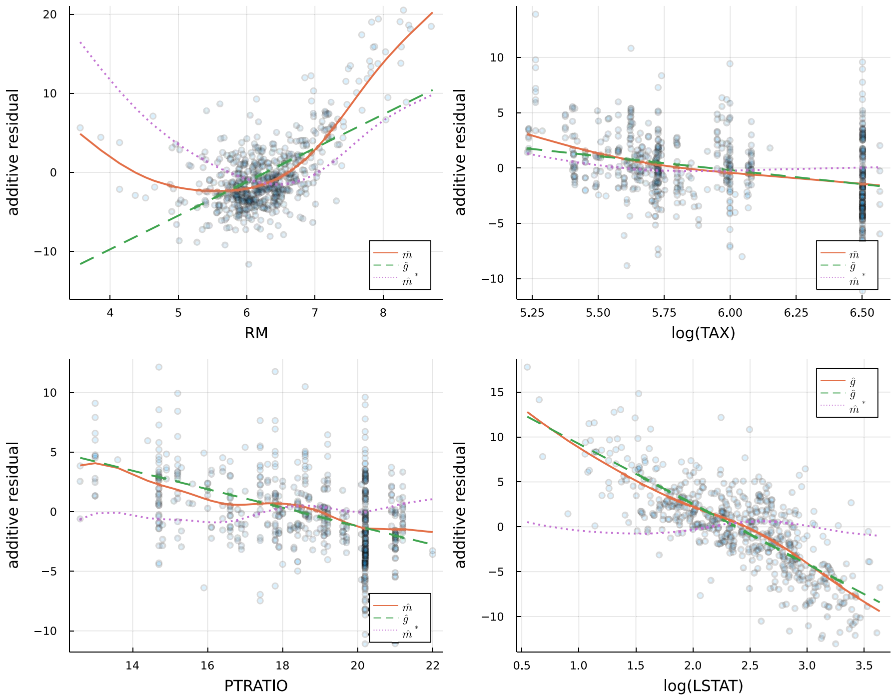

Fig. 6 Partial residual plots along with fitted regression curves for the Boston housing dataset. The solid lines represent the estimated additive functions $\widehat{m}_j = \widehat{m}_j^* + \widehat{g}_j$, $j = 1, 2, 3, 4$; dotted lines indicate the purely nonparametric functions $\widehat{m}_j^*$, dashed lines represent the parametric part $\widehat{g}_j$, for $j = 1, 2, 3, 4$

This confirms with the observations of Opsomer and Ruppert (1998) and Fan and Jiang (2005).

We use the proposed GLR and LF statistics to test whether the semiparametric null model (34) holds against the additive alternative model (33). For the LF test statistic, we consider the family of LINEX loss functions (31) with parameters ($s = \{0, 0.2, 0.5, 1\}, t = 1$). For comparison, we include the results for the GLR test in Fan and Jiang (2005), which we refer as GLR(FJ). Further, we also include the results from the backfit test (SB) defined in Mammen and Sperlich (2022). For convenience, the test statistic for SB method is computed as

$$S_n = \sum_{j=2}^{4} \sum_{i=1}^{n} \{\widehat{m}_j(X_{(i)j}) - \widehat{g}_j(X_{(i)j})\}^2 \widehat{f}_j(X_{(i)j})(X_{(i)j} - X_{(i-1)j}), \text{ with } X_{(0)j} := X_{(1)j},$$

where $X_{(i)j}$ is the ith order statistic of X_j, $\widehat{g}_j(\cdot)$ is the corresponding parametric part. The null distributions of the test statistics λ_n, q_n, F_λ, F_q, $\lambda_n(FJ)$, and S_n are necessary to compute their p-values. Therefore, we use the conditional bootstrap method described in Supplementary Material to obtain the null distributions of the test statistics. The optimal bandwidths $\boldsymbol{h}_{\text{opt}} = (0.40, 0.20, 0.59, 0.39)^T$ are computed using the procedure described in Sect. 5.1.

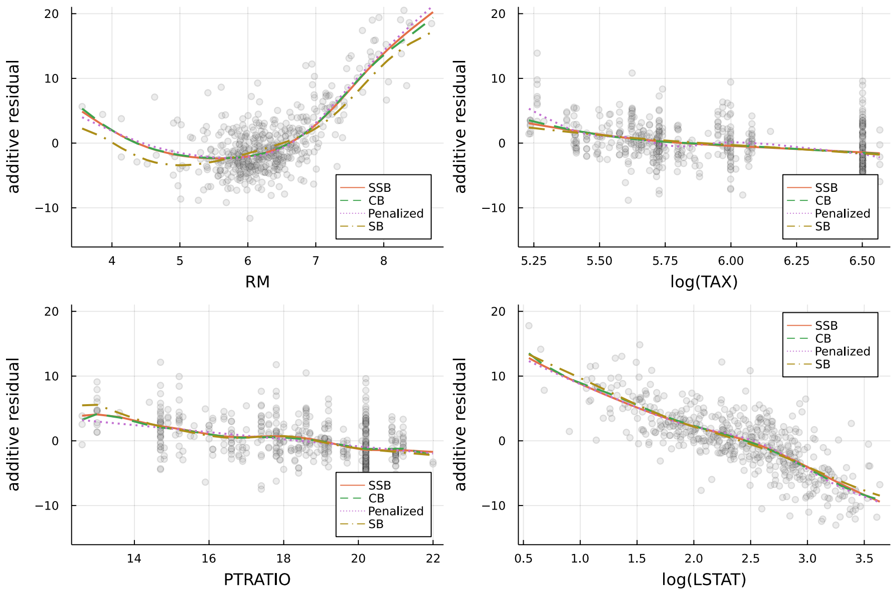

Fig. 7 Comparison of fits for the Boston housing dataset. SSB: (solid) $\hat{m}_j$, $j = 1, 2, 3, 4$; Penalized: (dotted) fits from gam() function in *mgcv* package in R; CB: (dashed) fits from the classical backfitting of Fan and Jiang (2005); SB: (dash dot) fits from sback() function in *wsbackfit* package in R

Table 2 provides the p values for statistics λ_n, q_n, F_λ, F_q, $\lambda_n(FJ)$, and S_n with the following five different bandwidths $(\frac{1}{2}\boldsymbol{h}_{\text{opt}}, \frac{2}{3}\boldsymbol{h}_{\text{opt}}, \boldsymbol{h}_{\text{opt}}, \frac{3}{2}\boldsymbol{h}_{\text{opt}}, 2\boldsymbol{h}_{\text{opt}})^T$ and using 1000 bootstrap replications to compute null distributions. These results indicate that the semiparametric model (34) is appropriate for this dataset within the additive models. For smaller bandwidths (undersmoothing), there is some evidence to reject the null hypothesis which is not surprising. For larger bandwidths, the estimated additive functions look more like parametric models and therefore the evidence is in favor of the null hypothesis. For the optimal bandwidths considered for estimation, the proposed GLR and LF tests conclude that semiparametric additive model is appropriate at 0.01 and 0.1 significance levels, respectively. This result also validates our finding that the LF test

Table 2 P values of statistics λ_n, q_n, F_λ, F_q, $\lambda_n(FJ)$, and S_n test statistics for a random sample of 200 observations from the Boston housing data

Bandwidth	λ_n	$q_n(0,1)$	$q_n(0.2,1)$	$q_n(0.5,1)$	$q_n(1,1)$	F_λ	F_q	$\lambda_n(FJ)$	S_n
$\frac{1}{2}\boldsymbol{h}_{\text{opt}}$	0.001	0.007	0.01	0.022	0.024	0.101	0.125	0.0	0.044
$\frac{2}{3}\boldsymbol{h}_{\text{opt}}$	0.005	0.02	0.032	0.05	0.047	0.183	0.232	0.0	0.237
$\boldsymbol{h}_{\text{opt}}$	0.023	0.09	0.092	0.104	0.111	0.218	0.228	0.01	0.574
$\frac{3}{2}\boldsymbol{h}_{\text{opt}}$	0.08	0.134	0.15	0.171	0.205	0.194	0.155	0.055	0.418
$2\boldsymbol{h}_{\text{opt}}$	0.124	0.081	0.086	0.102	0.129	0.15	0.091	0.079	0.191

The LF test statistic uses the family of LINEX loss functions (31). We consider 1000 bootstrap replications to compute the null distributions of respective statistics.

is asymptotically more powerful than the GLR test. The p values of the statistic $\lambda_n(FJ)$ are the smallest among all. We note that the optimal bandwidths are computed using simplified smooth backfitting and the same are used for GLR(FJ) as well.

To sum up, the results in this section indicate that both GLR and LF tests are very useful in practical applications. While their performances are sensitive to the choice of bandwidth, it is not straightforward to find optimal bandwidths for these statistics. Additionally, the finite sample performance of LF test is mildly sensitive to the choice of the loss function. Therefore, in practice, it is advisable to use both frameworks for a given hypothesis testing problem. This helps minimizing errors associated with hypothesis testing.

6 Summary and conclusions

In this study, we develop a hypothesis testing framework for additive models using GLR and LF tests where simplified smooth backfitting is used for model estimation. While the properties of GLR test are available in the literature for additive models estimated via classical backfitting (Opsomer, 2000), it is not the case for additive models estimated with simplified smooth backfitting (Huang and Yu, 2019). Similarly, the results for the LF test are not available for additive models. We fill this void by proposing inference methods using GLR and LF tests when a model uses simplified smooth backfitting for estimation. Under some regularity conditions, we show both the test statistics achieve Wilks phenomenon and have optimal power properties. Furthermore, LF test is asymptotically more powerful than GLR test. This result is a new addition to the existing literature. The numerical performance of test statistics is also very similar across different bandwidths and robust to different error distributions to some extent.

One possible direction for future research is to propose similar testing frameworks for generalized additive models. The LF test is asymptotically more powerful than the GLR test in linear additive models. It will be interesting to see whether the same result holds in generalized additive models.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-022-00840-8.

Acknowledgements The author would like to thank the associate editor and two anonymous referees for their constructive feedback, which resulted in major changes to the paper's presentation. The author would also like to thank Prof. Li-Shan Huang for offering a postdoctoral position and for sharing her research on simplified smooth backfitting, which provided the required framework for this article. The author gratefully acknowledges the support from grants 107-2811-M-007-014, 105-2118-M-007-006-MY2, and 107-2811-M-007-1047 by the Ministry Of Science and Technology (MOST) in Taiwan (R.O.C).

References

Belsley, D. A., Kuh, E., Welsch, R. E. (2005). *Regression diagnostics*: *Identifying influential data and sources of collinearity*. New Jersey: Wiley.

Bezanson, J., Edelman, A., Karpinski, S., Shah, V. B. (2017). Julia: A fresh approach to numerical computing. *SIAM Review, 59*(1), 65–98.
Breiman, L., Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association, 80*(391), 580–598.
Buja, A., Hastie, T., Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics, 17*(2), 453–510.
Fan, J., Gijbels, I. (1996). *Local polynomial modelling and its applications*: *Monographs on statistics and applied probability 66*. Boca Raton: CRC Press.
Fan, J., Jiang, J. (2005). Nonparametric inferences for additive models. *Journal of the American Statistical Association, 100*(471), 890–907.
Fan, J., Zhang, C., Zhang, J. (2001). Generalized likelihood ratio statistics and wilks phenomenon. *Annals of statistics, 29*(1), 153–193.
Friedman, J. H., Stuetzle, W. (1981). Projection pursuit regression. *Journal of the American statistical Association, 76*(376), 817–823.
Härdle, W., Huet, S., Mammen, E., Sperlich, S. (2004). Bootstrap inference in semiparametric generalized additive models. *Econometric Theory, 20*(2), 265–300.
Harrison, D., Jr., Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management, 5*(1), 81–102.
Hart, J. (2013). *Nonparametric smoothing and lack-of-fit tests*. New York: Springer Science & Business Media.
Hastie, T., Tibshirani, R. (1990). *Generalized Additive Models, Chapman & Hall/CRC Monographs on Statistics & Applied Probability*. Taylor & Francis.
Hong, Y., Lee, Y.-J. (2013). A loss function approach to model specification testing and its relative efficiency. *Annals of Statistics, 41*(3), 1166–1203.
Horowitz, J. L., Huang, J. (2013). Penalized estimation of high-dimensional models under a generalized sparsity condition. *Statistica Sinica, 23*, 725–748.
Huang, L.-S., Chan, K.-S. (2014). Local polynomial and penalized trigonometric series regression. *Statistica Sinica, 24*, 1215–1238.
Huang, L.-S., Chen, J. (2008). Analysis of variance, coefficient of determination and f-test for local polynomial regression. *The Annals of Statistics, 36*(5), 2085–2109.
Huang, L.-S., Davidson, P. W. (2010). Analysis of variance and f-tests for partial linear models with applications to environmental health data. *Journal of the American Statistical Association, 105*(491), 991–1004.
Huang, L.-S., Yu, C.-H. (2019). Classical backfitting for smooth-backfitting additive models. *Journal of Computational and Graphical Statistics, 28*(2), 386–400.
Hurvich, C. M., Simonoff, J. S., Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved akaike information criterion. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), *60*(2), 271–293.
Ingster, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives. i, ii, iii. *Mathematical Methods Statistics, 2*(2), 85–114.
Lian, H., Chen, X., Yang, J.-Y. (2012). Identification of partially linear structure in additive models with an application to gene expression prediction from sequences. *Biometrics, 68*(2), 437–445.
Lian, H., Liang, H., Ruppert, D. (2015). Separation of covariates into nonparametric and parametric parts in high-dimensional partially linear additive models. *Statistica Sinica, 25*, 591–607.
Linton, O., Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika, 82*(1), 93–100.
Mammen, E., Sperlich, S. (2022). Backfitting tests in generalized structured models. *Biometrika, 109*(1), 137–152.
Mammen, E., Linton, O., Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *The Annals of Statistics, 27*(5), 1443–1490.
Meier, L., Van de Geer, S., Bühlmann, P. (2009). High-dimensional additive modeling. *The Annals of Statistics, 37*(6B), 3779–3821.
Nielsen, J. P., Sperlich, S. (2005). Smooth backfitting in practice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67*(1), 43–61.
Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis, 73*(2), 166–179.
Opsomer, J. D., Ruppert, D. (1998). A fully automated bandwidth selection method for fitting additive models. *Journal of the American Statistical Association, 93*(442), 605–619.

Pitman, E. J. (2018). *Some basic theory for statistical inference*: *Monographs on applied probability and statistics*. Boca Raton: CRC Press.
R Core Team (2021). *R*: *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
Roca-Pardiñas, J., Cadarso-Suárez, C., González-Manteiga, W. (2005). Testing for interactions in generalized additive models: Application to so 2 pollution data. *Statistics and Computing, 15*(4), 289–299.
Roca-Pardinas, J., Rodriguez-Alvarez, M. X., Sperlich, S. (2021). *wsbackfit: Weighted Smooth Backfitting for Structured Models*. R package version 1.0-5. https://CRAN.R-project.org/package=wsbackfit
Spokoiny, V. G., et al. (1996). Adaptive hypothesis testing using wavelets. *Annals of Statistics, 24*(6), 2477–2498.
Tjøstheim, D., Auestad, B. H. (1994). Nonparametric identification of nonlinear time series: Projections. *Journal of the American Statistical Association, 89*(428), 1398–1409.
Wood, S. N. (2017). *Generalized additive models*: *An introduction with R*. Boca Raton: CRC Press.
Wood, S., Wood, M. S. (2015). Package 'mgcv'. *R package version, 1*, 29.