An optimal test for the additive model with discrete or categorical predictors

Abhijit Mandal[1]

Abstract

In multivariate nonparametric regression, the additive models are very useful when a suitable parametric model is difficult to find. The backfitting algorithm is a powerful tool to estimate the additive components. However, due to complexity of the estimators, the asymptotic p value of the associated test is difficult to calculate without a Monte Carlo simulation. Moreover, the conventional tests assume that the predictor variables are strictly continuous. In this paper, a new test is introduced for the additive components with discrete or categorical predictors, where the model may contain continuous covariates. This method is also applied to the semiparametric regression to test the goodness of fit of the model. These tests are asymptotically optimal in terms of the rate of convergence, as they can detect a specific class of contiguous alternatives at a rate of $n^{-1/2}$. An extensive simulation study and a real data example are presented to support the theoretical results.

Keywords Additive model · Categorical data analysis · Backfitting algorithm · Generalized likelihood ratio test · Semiparametric model · Local polynomial regression

1 Introduction

The additive model is a widely used multivariate smoothing technique. It was originally suggested by Friedman and Stuetzle (1981) and popularized due to extensive discussion in Hastie and Tibshirani (1990). It models a random sample $\{(Y_i, X_i) : i = 1, 2, \ldots, n\}$ by

---

✉ Abhijit Mandal
  abhijit.mandal@wayne.edu

1 Department of Mathematics, Wayne State University, 656 W. Kirby, Detroit, MI 48202, USA

$$Y_i = \alpha + \sum_{p=1}^{P} m_p(X_{pi}) + \epsilon_i,$$

where the random error ϵ_i has mean zero and constant variance σ^2 and the additive component m_p is an unknown smooth function for $p = 1, 2, \ldots, P$. Stone (1985, 1986) have shown that the additive model reduces a full P-dimensional nonparametric regression effectively to a one-dimensional problem by fitting the model with the same asymptotic efficiency, i.e., an optimal convergence rate of $n^{-2/5}$ for twice continuously differentiable functions. So, it has the very desirable property of reducing the "curse of dimensionality" in a satisfactory manner. In this paper, the additive function is estimated using the backfitting algorithm proposed by Buja et al. (1989). Opsomer and Ruppert (1997), Wand (1999) and Opsomer (2000) studied asymptotic properties of the backfitting estimators. In the literature, there are several other algorithms, such as the marginal integration estimation method (Tjøstheim and Auestad 1994), the estimating equation method (Mammen et al. 1999) and the Bayesian backfitting algorithm (Hastie and Tibshirani 2000).

To our knowledge, there are relatively limited theoretical results on the testing problem for the additive models where discrete or categorical (possibly mixed with continuous) explanatory variables are considered. Sperlich et al. (2002) and Yang et al. (2003) considered marginal integration estimators to construct tests for testing the additive components with continuous variables. The asymptotic critical values of these tests are difficult to obtain due to the complicated expressions for the bias and the variance of the test statistic. Moreover, the authors observed that the asymptotic accuracy of their result is limited for small and moderate sample sizes. In the same setup, Fan et al. (2001) and Fan and Jiang (2005) proposed the generalized likelihood ratio (GLR) test for testing the significance of additive components using backfitting estimators. The idea is based on comparison of pseudo-likelihood functions under the null and the alternative hypotheses, which leads to the log ratio of the variance estimators under the null and the alternative. Similar to the maximum likelihood ratio tests in parametric models, the GLR test has an important fundamental property that the asymptotic null distribution of the test is independent of nuisance parameters and functions. This property is referred to as the Wilks phenomenon. The GLR test is asymptotically distribution-free, and it is asymptotically optimal in terms of convergence of the nonparametric hypothesis testing problem (see Ingster 1993a, b, c; Spokoiny 1996). However, the authors mentioned that the GLR test may not be accurate as the test statistic contains an unknown bias term. So, a Monte Carlo simulation or a bootstrap technique is performed to calculate the p value of the test. This somehow restricts the method for being widely applicable among the general practitioners.

In this paper, we propose a GLR test for the additive components having discrete- or categorical-valued predictors, while the model may contain continuous-valued covariates. For categorical predictors, this test may be regarded as the generalized analysis of covariance (ANCOVA), where covariates are modeled by nonparametric functions and the normality assumption on the error term is not required. In this case, the predictors may be referred as treatment or block of the design of experiment.

The rest of the paper is organized as follows. We give an overview of the non-parametric additive model and the semiparametric additive model in Sects. 2 and 3, respectively. In Sect. 4, we introduce the GLR test and present the theoretical properties. Extensive simulation studies are performed in Sect. 5 to explore the behavior of the proposed test. In Sect. 6, we apply our method to analyze a real data containing diamond price and we propose an appropriate model based on quality attributes and physical measurements. Section 7 contains some concluding remarks. The assumptions of the theorems are given in "Appendix A." Online Appendix B presents a brief description of the backfitting estimator. The proofs of the theorems are given in Online Appendix C.

2 The nonparametric additive model

Let us consider a one-dimensional response variable Y, a P-dimensional predictor $X = (X_1, X_2, \ldots, X_P)^T$ and an additional Q-dimensional covariate $Z = (Z_1, Z_2, \ldots, Z_Q)^T$. We assume that X contains only discrete or categorical variables, but Z may contain all types of variables—categorical, discrete or continuous. If X_p is a discrete-valued random variable, then k_p denotes the number of distinct values, $x_{p1}, x_{p2}, \ldots, x_{pk_p}$, where $k_p < \infty$ for all $p = 1, 2, \ldots, P$. If X_p is a categorical variable, then $x_{p1}, x_{p2}, \ldots, x_{pk_p}$ are different levels of the variable. Let $(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)$ be a random sample of size n from (Y, X, Z). The nonparametric additive model is given by

$$Y_i = \alpha + \sum_{p=1}^{P} m_p(X_{pi}) + \sum_{q=1}^{Q} m_{P+q}(Z_{qi}) + \epsilon_i, \tag{1}$$

where the random error ϵ_i has mean zero and a constant variance σ^2. To ensure identifiability of components of the additive model, we set $E(m_p(X_{pi})) = 0$ for all $p = 1, 2, \ldots, P$ and $E(m_{P+q}(Z_{qi})) = 0$ for all $q = 1, 2, \ldots, Q$. The intercept parameter $\alpha = E(Y_i)$ is generally estimated by $\hat{\alpha} = \bar{Y} = \sum_i Y_i/n$. The backfitting estimator is used to estimate the nonparametric functions $m_d(\cdot)$ for $d = 1, 2, \ldots, P + Q$. For the ease of readability, a discussion on the backfitting estimators is given in Online Appendix B. We have divided regressors into two groups—predictor and covariate, as we construct a test of significance for predictors only, whereas their effect is adjusted by "nuisance" covariates. In other words, if we are interested to test the effect of a subset of regressors in the additive model, we name those regressors as predictors and remaining regressors as covariates. We require all predictors to be discrete or categorical variables, but there is no restriction on covariates.

3 The semiparametric additive model

The semiparametric additive model (SAM) is the combination of a parametric model and a nonparametric additive model. Here, some of the additive components are

modeled parametrically, while the remaining ones are unspecified and are estimated nonparametrically. First, we model predictors parametrically and covariates nonparametrically, and then, a generalized model is considered. In general, the predictors are assumed to be discrete-valued random variables. However, if the predictors are ordinal categorical variables, their order or rank may also be modeled parametrically. Let us consider the following SAM model:

$$Y_i = \alpha + \sum_{p=1}^{P} m_{p,\boldsymbol{\theta}_p}(X_{pi}) + \sum_{q=1}^{Q} m_{P+q}(Z_{qi}) + \epsilon_i, \tag{2}$$

where $\mathcal{M}_{p,\Theta_p} = \{m_{p,\boldsymbol{\theta}_p}(X_{pi}), \boldsymbol{\theta}_p \in \Theta_p\}$ is a family of parametric functions for $p = 1, 2, \ldots, P$. We assume that $m_{p,\boldsymbol{\theta}_p}(\cdot)$ is completely known except for the value of the parameter $\boldsymbol{\theta}_p$, $p = 1, 2, \ldots, P$. Opsomer and Ruppert (1999) and Jiang et al. (2007) have studied this model when the parametric models are linear functions. One might be interested to build a SAM when the main interest of study is to precisely quantify the effect of the predictors $X_1, X_2, \ldots, X_P$ on the dependent variable Y, but the relationship is observed in the presence of "nuisance" covariates $Z_1, Z_2, \ldots, Z_Q$. The use of the parametric forms for predictors, if properly specified, allows us to make an easily interpretable inference about their effect on Y. On the other hand, modeling covariates nonparametrically, one may avoid potential introduction of bias in the estimated relationship between predictors and Y. Another possible situation when a SAM would be useful, if someone is fairly confident about the shape of the relationship between predictors and Y, but not about that of the other covariates. It can be shown that by modeling some predictors using appropriate parametric functions, the risk of over-fitting the model is reduced by decreasing the overall degrees of freedom of the test.

To ensure the identifiability of the model, we assume that the expectation of the parametric term is zero, i.e., $E[m_{p,\boldsymbol{\theta}_p}(X_p)] = 0$ for all $p = 1, 2, \ldots, P$, and $E(m_{P+q}(Z_{qi})) = 0$ for all $q = 1, 2, \ldots, Q$. We consider the case where $m_{p,\boldsymbol{\theta}_p}(\cdot)$ is a polynomial for all $p = 1, 2, \ldots, P$. As X_p takes k_p values, one needs at most $(k_p - 1)$ parameters to completely specify $m_{p,\boldsymbol{\theta}_p}(\cdot)$. So, we assume that $m_{p,\boldsymbol{\theta}_p}(\cdot)$ is a polynomial of degree r_p, where $0 < r_p < k_p - 1$, $p = 1, 2, \ldots, P$. Therefore, with a slight abuse of notation, we write

$$m_{p,\boldsymbol{\theta}_p}(X_p) = \alpha_p + \sum_{s=1}^{r_p} \theta_{ps} X_p^s,$$

where $\boldsymbol{\theta}_p = (\theta_{p1}, \theta_{p2}, \ldots, \theta_{pr_p})^{\mathrm{T}}$, $p = 1, 2, \ldots, P$. Note that α_p is not an independent parameter, it just makes $m_{p,\boldsymbol{\theta}_p}$ centered at zero. Let us define $\boldsymbol{\theta} = (\alpha^*, \boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}}, \ldots, \boldsymbol{\theta}_P^{\mathrm{T}})^{\mathrm{T}}$, where $\alpha^* = \sum_p \alpha_p$. For $a < b$, we define

$$
{}^b_a X_{(p)} = \begin{bmatrix} X^a_{p1} & X^{a+1}_{p1} & \cdots & X^b_{p1} \\ X^a_{p2} & X^{a+1}_{p2} & \cdots & X^b_{p2} \\ \vdots & \vdots & \ddots & \vdots \\ X^a_{pn} & X^{a+1}_{pn} & \cdots & X^b_{pn} \end{bmatrix}, \tag{3}
$$

and if $a = b$ then ${}^b_a X_{(p)}$ is a vector containing the first column of Eq. (3). Let $X^* = (\mathbf{1}_n, {}^{r_1}_1 X_{(1)}, {}^{r_2}_1 X_{(2)}, \ldots, {}^{r_P}_1 X_{(P)})$. Then, following Speckman (1988), the estimates of the additive components are derived from the backfitting algorithm as the solution of the following equations:

$$
\widetilde{\boldsymbol{\theta}} = \left(X^{*\mathrm{T}} X^* \right)^{-1} X^{*\mathrm{T}} \left(Y^* - \sum_{q=1}^{Q} \widetilde{\boldsymbol{m}}_{P+d} \right),
$$

$$
\widetilde{\boldsymbol{m}}_{P+q} = S^*_{P+q} \left(Y^* - X^* \widetilde{\boldsymbol{\theta}} - \sum_{d \neq q} \widetilde{\boldsymbol{m}}_{P+d} \right), \quad \text{for } q = 1, 2, \ldots, Q, \tag{4}
$$

provided $(X^{*\mathrm{T}} X^*)^{-1}$ exists. Here, S^*_{P+q} is the centered smoothing matrix S^*_d for $d = P + q$ as defined after Equation (A3) in Online Appendix B. Suppose W_{P+q} is the additive smoother matrix of $\boldsymbol{m}_{P+q}$, so that the backfitting estimate of $\boldsymbol{m}_{P+q}$ is $\widetilde{\boldsymbol{m}}_{P+q} = W_{P+q}(Y^* - X^* \widetilde{\boldsymbol{\theta}})$ for $q = 1, 2, \ldots, Q$. Let us define $W_{[Z]} = \sum_{q=1}^{Q} W_{P+q}$ and $\widetilde{\boldsymbol{m}}_{[Z]} = \sum_{q=1}^{Q} \widetilde{\boldsymbol{m}}_{P+q}$. Then, the above normal equations are solved non-iteratively as

$$
\widetilde{\boldsymbol{\theta}} = \left(X^{*\mathrm{T}} \left(I_n - W_{[Z]} \right) X^* \right)^{-1} X^{*\mathrm{T}} \left(I_n - W_{[Z]} \right) Y^*,
$$

$$
\widetilde{\boldsymbol{m}}_{[Z]} = W_{[Z]} \left(Y^* - X^* \widetilde{\boldsymbol{\theta}} \right). \tag{5}
$$

Sometimes the experimenter may have a prior knowledge about some of the variables and would like to model them parametrically, whereas keeping other variables in the nonparametric model. In that case, one may consider the following generalized SAM model:

$$
Y_i = \alpha + \sum_{p=1}^{P_1} m_{p, \boldsymbol{\theta}_p}(X_{pi}) + \sum_{p=P_1+1}^{P} m_p(X_{pi}) + \sum_{q=1}^{Q_1} m_{P+q, \boldsymbol{\theta}_{P+q}}(Z_{qi})
$$

$$
+ \sum_{q=Q_1+1}^{Q} m_{P+q}(Z_{qi}) + \epsilon_i, \tag{6}
$$

where $P_1 \leq P$, $Q_1 \leq Q$ and $\mathcal{M}_{p, \Theta_p} = \{m_{p, \boldsymbol{\theta}_p}(X_{pi}), \boldsymbol{\theta}_p \in \Theta_p\}$ for $p = 1, 2, \ldots, P_1$, $\mathcal{M}_{P+q, \Theta_{P+q}} = \{m_{P+q, \boldsymbol{\theta}_{P+q}}(Z_{qi}), \boldsymbol{\theta}_{P+q} \in \Theta_{P+q}\}$ for $q = 1, 2, \ldots, Q_1$ are families of parametric functions. For simplicity, we assume that m_d is a polynomial of degree r_d for all $d = 1, 2, \ldots, P_1$ and $d = P+1, P+2, \ldots, P+Q_1$.

4 The generalized likelihood ratio test

Let us consider the generalized SAM model defined in Eq. (6). For this model, one may be interested mainly in two types of tests based on the predictor variables—a goodness-of-fit test for the parametric function and a model utility test for the nonparametric function. First, we present the generalized test that includes both types of tests, and then, we discuss about the individual test. So, we are now interested in the following null hypothesis:

$$
\begin{aligned}
H_0 : m_p(\cdot) &\in \mathcal{M}_{p,\Theta_p} \text{ for } p = 1, 2, \ldots, P_1, \text{ and } m_p(\cdot) \\
&= 0 \text{ for } p = P_1 + 1, P_1 + 2, \ldots, P.
\end{aligned}
\tag{7}
$$

Under H_0, we define $\widetilde{m}_d(\cdot)$ and $\widetilde{m}_{d,\widetilde{\theta}_d}$ as the backfitting estimators of $m_d(\cdot)$ and m_{d,θ_d}, respectively. Then, the residual sum of squares, under H_0, is given by

$$
\mathrm{RSS}_0 = \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \sum_{p=1}^{P_1} \widetilde{m}_{p,\widetilde{\theta}_p}(X_{pi}) - \sum_{q=1}^{Q_1} \widetilde{m}_{P+q,\widetilde{\theta}_{P+q}}(Z_{qi}) - \sum_{q=Q_1+1}^{Q} \widetilde{m}_{P+q}(Z_{qi}) \right)^2 .
$$

As we are testing only for predictors by keeping covariates unchanged, the unconstrained model is

$$
Y_i = \alpha + \sum_{p=1}^{P} m_p(X_{pi}) + \sum_{q=1}^{Q_1} m_{P+q,\theta_{P+q}}(Z_{qi}) + \sum_{q=Q_1+1}^{Q} m_{P+q}(Z_{qi}) + \epsilon_i.
\tag{8}
$$

Under this model, the residual sum of square is given by

$$
\mathrm{RSS}_1 = \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \sum_{p=1}^{P} \widehat{m}_p(X_{pi}) - \sum_{q=1}^{Q_1} \widehat{m}_{P+q,\widehat{\theta}_{P+q}}(Z_{qi}) + \sum_{q=Q_1+1}^{Q} \widehat{m}_{P+q}(Z_{qi}) \right)^2 ,
$$

where $\widehat{m}_d(\cdot)$ and $\widehat{m}_{d,\widehat{\theta}_d}$ are the backfitting estimators of $m_d(\cdot)$ and m_{d,θ_d}, respectively, under the unconstrained model. The generalized likelihood ratio (GLR) test statistic for testing null hypothesis H_0 is defined as

$$
\lambda_n(H_0) = \frac{n(\mathrm{RSS}_0 - \mathrm{RSS}_1)}{\mathrm{RSS}_1}.
\tag{9}
$$

If the difference between RSS_0 and RSS_1 is small, then the GLR test statistic may be approximated by $n \log(\frac{\mathrm{RSS}_0}{\mathrm{RSS}_1})$. In the parametric model, this is equivalent to the log-likelihood ratio test statistic, where estimators are replaced by the corresponding maximum likelihood estimators. Generally, the nonparametric maximum likelihood estimate does not exist and even when it does exist, the resulting maximum likelihood ratio test is not optimal (see Fan and Jiang 2005; Hall and Marron 1988). So, the GLR test statistic may be regarded as a log ratio of the quasi-likelihoods.

We assumed homoscedasticity of the error term, i.e., error ϵ_i in model (6) has a constant variance. If this assumption is not valid, one may consider a GLR statistic by taking weighted residual sum of squares. The subsequent analysis and the back-fitting algorithm will be modified similar to the weighted likelihood approach for the parametric models.

4.1 Asymptotic distribution

To derive the null distribution of the GLR test statistic, let us define $c_p = (\sqrt{c_{p1}}, \sqrt{c_{p2}}, \ldots, \sqrt{c_{pk_p}})^{\mathrm{T}}$, where $c_{pj} = P(X_p = x_{pj})$ for $j = 1, 2, \ldots, k_p$ and $p = 1, 2, \ldots, P$. We also define ${}_a^b Z_{(P+q)}$ in a similar way of ${}_a^b X_{(p)}$ as defined in Eq. (3) by replacing X with Z. Let us denote $T^* = (\mathbf{1}_n, {}_1^{r_1} X_{(1)}, {}_1^{r_2} X_{(2)}, \ldots,$ ${}_1^{r_{P_1}} X_{(P_1)}, {}_1^{r_{P+1}} Z_{(P+1)}, {}_1^{r_{P+2}} Z_{(P+2)}, \ldots, {}_1^{r_{P+Q_1}} Z_{(P+Q_1)})$. Suppose $I = \sum_{p=1}^{P_1}(k_p - r_p - 1) + \sum_{p=P_1+1}^{P} k_p$, and $\boldsymbol{\Sigma}_1$ is a $I \times I$-dimensional block diagonal matrix whose pth diagonal block is an identity matrix of order $(k_p - r_p - 1)$ if $p \leq P_1$, and $\left(I_{k_p} - c_p c_p^{\mathrm{T}}\right)$ if $p > P_1$. Define another $I \times I$-dimensional block matrix $\boldsymbol{\Sigma}_2$, whose pth diagonal block is the identity matrix of order $(k_p - r_p - 1)$ if $p \leq P_1$, and of order k_p if $p > P_1$. For $p \neq p' \in \{1, 2, \ldots, P\}$, the pp'th off-diagonal block of $\boldsymbol{\Sigma}_2$ is given by

$$\Sigma_{pp'} = \lim_{n \to \infty} \left(R_p^{\mathrm{T}} R_p\right)^{-1/2} R_p^{\mathrm{T}} R_{p'} \left(R_{p'}^{\mathrm{T}} R_{p'}\right)^{-1/2}, \tag{10}$$

where $R_p = {}_{r_p+1}^{k_p-1} X_{(p)}$ if $p \leq P_1$ and $R_p = {}_0^{k_p-1} X_{(p)}$ if $p > P_1$. Then, the following theorem gives the asymptotic null distribution of the GLR statistic.

Theorem 1 *Suppose that regularity conditions (C1)–(C8) in "Appendix A" hold. Further assume that the limit of $n^{-1} T^{*\mathrm{T}} T^*$ exists and it is invertible. Let us consider the unconstrained model (8) and the null hypothesis H_0 in (7), where m_{p,θ_p} is a polynomial of degree r_p and $0 < r_p < (k_p - 1)$, for $p = 1, 2, \ldots, P_1$. Then, under H_0, the asymptotic distribution of the GLR test statistic coincides with $\sum_{i=1}^{s} \lambda_i V_i^2$, where $V_1, V_2, \ldots, V_s$ are independent and identically distributed (i.i.d.) standard normal variables, $\lambda_1, \lambda_2, \ldots, \lambda_s$ are nonzero eigenvalues of $\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1$ and s is the rank of $\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1$.*

The proof of the theorem is given in Online Appendix B. Theorem 1 shows that the asymptotic null distribution of the GLR test statistic is a linear combination of Chi-square variables. The critical region of the test may be calculated using the algorithm proposed by Davies (1980). Note that the null distribution does not depend on the modeling of covariates as we keep covariates unchanged under both the null and the alternative hypotheses. But, in practice, it reduces the possible over-fitting; thus, the finite sample performance of the test improves due to parametric modeling those covariates. However, one must be careful while modeling covariates, and it is important to verify whether such a parametric model is valid or not. A wrong model may cause severe power loss as demonstrated in the simulation studies. A special case of Theorem 1 when the predictor variables are pairwise independent, the null distribution is reduced to a single Chi-square as mentioned in the following corollary.

Corollary 1 *Suppose that the predictor variables are pairwise independent, and the assumptions of Theorem 1 hold. Then, under H_0, the asymptotic distribution of the GLR test statistic is a Chi-square distribution with degrees of freedom $\sum_{p=1}^{P_1}(k_p - r_p - 1) + \sum_{p=P_1+1}^{P}(k_p - 1)$.*

It is shown in the simulation section that even if the predictors are not pairwise independent one may approximate the null distribution using Corollary 1 unless predictors are strongly correlated. Simulation studies show that the approximation makes the test little anti-conservative in small or moderate sample sizes. However, as sample size increases, it gives a good approximation.

It is interesting to note that the asymptotic null distribution of the GLR test statistic does not depend on the nuisance parameters—the design densities of Z, m_d functions for the covariates and the error distribution. But, in general, it depends on the design densities of X as shown in Theorem 1. So, the Wilks phenomenon does not hold in the true sense; however, it holds good if predictors are pairwise independent.

The main advantage of our method is that it is easy to calculate the p values of the test. In fact, the discrete-valued predictors make the test simple. On the other hand, as shown by Fan and Jiang (2005) and Jiang et al. (2007), if the predictors are continuous the GLR test becomes complicated, and the asymptotic null distribution depends on kernel density functions and bandwidth parameters. Moreover, the authors mentioned that the null distribution may not be accurate as the test statistic contains an unknown bias term. So, the null distribution is calculated by Monte Carlo simulation or using a conditional bootstrap method. In our test, we do not need any additional conditions to choose the bandwidth parameter for continuous covariates as long as assumption (C5) holds. Therefore, for simplicity, we may use the same bandwidth parameter which is optimal for estimation (see Sect. 5).

The main contribution of this paper is that the GLR test is constructed for the additive components with discrete or categorical predictors. The asymptotic distribution of the GLR test statistic is derived in Fan and Jiang (2005) with the strict assumption that all the predictors and covariates are continuous; more specifically, their marginal distributions must be Lipschitz continuous on some bounded support. Keeping in mind the real applications, this restriction is modified for the discrete or categorical predictors. For this reason, the proof of the theorem does not directly follow from the previous method, and we used a novel approach to derive the asymptotic distribution.

4.2 Power function

We now consider the power function of the GLR test. Let us take the following contiguous alternative hypothesis:

$$H_1 : m_p(\cdot) = n^{-1/2}m_p^*(\cdot) \text{ for } p = 1, 2, \ldots, P,$$

where m_p^* is an additive function under the alternative hypothesis such that $E(m_p^*) = 0$. We define m_p' as the best fitted polynomial of degree r_p to m_p^* for $p = 1, 2, \ldots, P_1$ and $m_p' = m_p^*$ for $p = P_1 + 1, P_1 + 2, \ldots, P$. Using the following theorem, we get the power function of the GLR test.

Theorem 2 *Let us consider the notations and assumptions of Theorem 1. Then, under H_1, the asymptotic distribution of the GLR test statistic coincides with $\delta^2 + \sum_{i=1}^{s} \lambda_i V_i^2$, where $\delta^2 = \sum_{r,s=1}^{P} E(m_r' m_s')$.*

It is interesting to note that the GLR test detects a specific class of contiguous alternatives at a rate of $n^{-1/2}$. So, the power of the test is asymptotically optimal in terms of the rate of convergence. For a fixed alternative, the power of the GLR test convergences to one, i.e., the test is consistent. In case of pairwise independent predictors, the theorem simplifies as follows.

Corollary 2 *Suppose that the predictor variables are pairwise independent, and the assumptions of Theorem 1 hold. Then, under H_1, the asymptotic distribution of the GLR test statistic is non-central Chi-square with the non-centrality parameter $\sum_{p=1}^{P} E(m_p'^2)$ and the degrees of freedom $\sum_{p=1}^{P_1} (k_p - r_p - 1) + \sum_{p=P_1+1}^{P} (k_p - 1)$.*

It is not surprising that the GLR statistic is $\sqrt{n}$-consistent, whereas most of the non-parametric tests have relatively slower rate. Estimating the m_p function corresponding to a discrete or categorical predictor X_p is a finite-dimensional problem as we assume that k_p, the domain of X_p, is finite. Thus, all m_p functions for $p = 1, 2, \ldots, P$ of the additive models in (6) and (8) are equivalent to the parametric part of the semiparametric model. So, the corresponding convergence rate is consistent with Corollary 1 of Opsomer and Ruppert (1999). The same result is also obtained by Speckman (1988). However, this rate depends on the bandwidth parameter selected for smoothing the continuous covariates (if any). The bandwidth parameter must be selected based on assumption (C5) given in "Appendix A." Even if when the model contains continuous covariates, as the null hypothesis is the significance only for the predictors, the GLR test has a convergence rate similar to a parametric model. Intuitively, in large sample sizes, the effect of smoothers for continuous covariates cancels out when two residual sums of squares are subtracted in the numerator of the GLR statistic in Eq. (9). Meanwhile, the denominator (decided by n) of the GLR statistic converges to σ^2, the error variance of the additive model. So, those continuous smooth functions do not have any major role in the rate of convergence of the GLR test as long as their bandwidth parameters are optimally selected.

We have started with a very general test in H_0 that includes both the goodness-of-fit test and the model utility test. For this reason, the construction of $\mathbf{\Sigma_1}$ and $\mathbf{\Sigma_2}$ used in Theorem 1 looks slightly complicated. However, if we are interested only in one type of test, those matrices become very simple. In these cases, the residual sum of squares RSS_0 and RSS_1 also has simpler expressions. Now, we discuss these special cases. To make the procedure further simple, we assume that all covariates are modeled nonparametrically.

4.3 The goodness-of-fit test for the semiparametric model

Let us consider the full nonparametric additive model given in (1). With respect to that base model, we now construct a goodness-of-fit test for the semiparametric model (2), where all predictors are modeled parametrically. So, the null hypothesis for this

problem is

$$H_0^* : m_p(\cdot) \in \mathcal{M}_{p,\Theta_p}, \text{ for } p = 1, 2, \ldots, P. \tag{11}$$

The residual sum of squares, under H_0^*, is given by

$$\text{RSS}_0^* = \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \sum_{p=1}^{P} \widetilde{m}_{p,\widetilde{\boldsymbol{\theta}}_p}(X_{pi}) - \sum_{q=1}^{Q} \widetilde{m}_{P+q}(Z_{qi}) \right)^2,$$

where $\widetilde{m}_{p,\widetilde{\boldsymbol{\theta}}_p}$ is the backfitting estimator of $m_{p,\boldsymbol{\theta}_p}$ under H_0^* for $p = 1, 2, \ldots, P$ and $\widetilde{m}_{P+q}$ is the backfitting estimator of m_{P+q} under H_0^* for $q = 1, 2, \ldots, Q$. Under the unconstrained nonparametric additive model, the residual sum of square is given by

$$\text{RSS}_1^* = \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \sum_{p=1}^{P} \widehat{m}_p(X_{pi}) - \sum_{q=1}^{Q} \widehat{m}_{P+q}(Z_{qi}) \right)^2, \tag{12}$$

where $\widehat{m}_1(\cdot), \widehat{m}_2(\cdot), \ldots, \widehat{m}_{P+Q}(\cdot)$ are the backfitting estimators under the full model given in (1). Suppose $L = \sum_{p=1}^{P}(k_p - r_p - 1)$, and $\boldsymbol{\Sigma}_2$ is a $L \times L$-dimensional block matrix, whose pth diagonal block is an identity matrix of order $(k_p - r_p - 1)$, and for $p \neq p' \in \{1, 2, \ldots, P\}$ the pp'th off-diagonal block of $\boldsymbol{\Sigma}_2$ is given in Eq. (10) with $\boldsymbol{R}_p = {}_{r_p+1}^{k_p-1}X_{(p)}$. Notice that $\boldsymbol{\Sigma}_1$, defined in Sect. 4.1, becomes an identity matrix in this setup as $P_1 = P$. Moreover, we need existence of $n^{-1}\boldsymbol{X}^{*\text{T}}\boldsymbol{X}^*$ instead of $n^{-1}\boldsymbol{T}^{*\text{T}}\boldsymbol{T}^*$, where $\boldsymbol{X}^*$ is defined in Sect. 3 after Eq. (3). Then, the following result gives the asymptotic null distribution of the GLR goodness-of-fit test statistic $\lambda_n(H_0^*) = \frac{n(\text{RSS}_0^* - \text{RSS}_1^*)}{\text{RSS}_1^*}$.

Corollary 3 *Suppose that the limit of $n^{-1}\boldsymbol{X}^{*\text{T}}\boldsymbol{X}^*$ exists and it is invertible, and regularity conditions (C1)–(C8) in "Appendix A" hold. Let us consider the unconstrained model* (1) *and the null hypothesis H_0^* in* (11)*, where $m_{p,\boldsymbol{\theta}_p}$ is a polynomial of degree r_p and $0 < r_p < (k_p - 1)$ for all $p = 1, 2, \ldots, P$. Then, the asymptotic distribution of the GLR goodness-of-fit test statistic, under H_0^*, coincides with $\sum_{i=1}^{s} \lambda_i V_i^2$, where $V_1, V_2, \ldots, V_s$ are i.i.d. standard normal variables, $\lambda_1, \lambda_2, \ldots, \lambda_s$ are nonzero eigenvalues of $\boldsymbol{\Sigma}_2$ and s is the rank of $\boldsymbol{\Sigma}_2$.*

In general, the asymptotic null distribution of the GLR goodness-of-fit test statistic is a linear combination of Chi-square variables. But, the distribution comes out to be a single Chi-square if the predictor variables are pairwise independent. In this case, the degrees of freedom of the test statistic become $\sum_{p=1}^{P}(k_p - r_p - 1)$.

4.4 Model utility test for the nonparametric additive model

Let us consider the null hypothesis that there is no association between Y and $X_1, \ldots, X_p$, where $Z_1, \ldots, Z_q$ are covariates of the nonparametric additive model

(1). The null hypothesis can be written as

$$H_0^{**} : m_p(\cdot) = 0 \text{ for all } p = 1, 2, \ldots, P. \tag{13}$$

Let $\widetilde{m}_{P+1}(\cdot), \widetilde{m}_{P+2}(\cdot), \ldots, \widetilde{m}_{P+Q}(\cdot)$ be the backfitting estimators under H_0^{**}. Then, the residual sum of squares, under H_0^{**}, is given by

$$\text{RSS}_0^{**} = \sum_{i=1}^{n} \left(Y_i - \hat{\alpha} - \sum_{q=1}^{Q} \widetilde{m}_{P+q}(Z_{qi}) \right)^2.$$

Under the unconstrained nonparametric additive model (1), the residual sum of square is given in Eq. (12). So, the GLR nonparametric test statistic becomes $\lambda_n(H_0^{**}) = \frac{n(\text{RSS}_0^{**} - \text{RSS}_1^*)}{\text{RSS}_1^*}$.

Suppose $\boldsymbol{\Sigma}_1$ is a $K \times K$-dimensional block diagonal matrix whose pth diagonal block is $\left(\boldsymbol{I}_{k_p} - c_p c_p^{\mathrm{T}} \right)$ for $p = 1, 2, \ldots, P$. Define another $K \times K$-dimensional block matrix $\boldsymbol{\Sigma}_2$, whose pth diagonal block is an identity matrix of order k_p, and for $p \neq p' \in \{1, 2, \ldots, P\}$ the ijth element of the pp'th off-diagonal block of $\boldsymbol{\Sigma}_2$ is given by

$$\sigma_{pp',ij} = \frac{1}{\sqrt{c_{pi} c_{p'j}}} P(X_p = x_{pi}, X_{p'} = x_{p'j}),$$

where $i = 1, 2, \ldots, k_p$ and $j = 1, 2, \ldots, k_{p'}$. The following result gives the asymptotic distribution of the GLR nonparametric test statistic for testing the null hypothesis H_0^{**}.

Corollary 4 *Let us assume that regularity conditions (C1)–(C8) in "Appendix A" hold. Let us consider the unconstrained model* (1) *and the null hypothesis H_0^{**} in* (13). *Then, under H_0^{**}, the asymptotic distribution of the GLR nonparametric test statistic coincides with $\sum_{i=1}^{s} \lambda_i V_i^2$, where $V_1, V_2, \ldots, V_s$ are i.i.d. standard normal variables, $\lambda_1, \lambda_2, \ldots, \lambda_s$ are nonzero eigenvalues of $\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1$ and s is the rank of $\boldsymbol{\Sigma}_1 \boldsymbol{\Sigma}_2 \boldsymbol{\Sigma}_1$.*

If the predictor variables are pairwise independent, the GLR nonparametric test statistic follows the Chi-square distribution with degrees of freedom $\sum_{p=1}^{P}(k_p - 1)$ under the null hypothesis. The setup of the GLR nonparametric test is similar to the classical ANCOVA when predictors are categorical variables. In ANCOVA, predictors are called treatments or blocks, and covariates are modeled parametrically. The goal is to test the treatment or block effect in the design of experiment. So, the GLR test is generalization the classical ANCOVA, where covariates are modeled nonparametrically. Moreover, we do not need to assume that the error distribution is normal.

5 Simulation

In the first part of the simulation, we check the null distribution of the GLR nonparametric test statistic for the hypothesis given in (13). Then, we demonstrate the

power of the GLR test and compare it with the F test associated with the nested linear models in the regression analysis. And finally, the performance of the general GLR test for testing H_0 in (7) under the semiparametric model is presented. All numerical examples in this paper are performed using R software. The R code for the GLR test will be provided on request.

5.1 Null distribution

Let us consider the nonparametric additive model defined in (1), where $P = 5$ and $Q = 4$. All five random variables in X and the first two random variables in Z are discrete. The number of discrete values taken by $X_1, X_2, \ldots, X_5$ and Z_1, Z_2 are 3, 4, 5, 4, 3 and 5, 4, respectively, starting from zero with an increment one. The probabilities for each of these variables are generated independently from a uniform (0,1) distribution, and then, they are standardized so that the total probabilities become one. We have discarded very low values of probability (< 0.05) to avoid very small or zero frequencies. To make the situation general, we have taken few independent and few dependent variables. X_1 and X_2 are independent random variables; X_3, X_4, X_5 form a group of dependent variables, but they are independent of X_1 and X_2. Similarly, (Z_1, Z_2) and (Z_3, Z_4) are two independent groups. The covariance matrices for the dependent groups are also generated randomly. Finally, these parameters are kept fixed throughout the entire simulation. To generate a set of dependent discrete variables, first, a random sample is drawn from a multivariate normal distribution with a fixed covariance matrix and then observations are discretized based on their probabilities. Here, we are interested in testing the null hypothesis $H_0^{**} : m_1 = m_2 = \cdots = m_5 = 0$. To show the null distribution, we have taken $m_p = 0$ for all $p = 1, 2, \ldots, 5$. For covariates, the m_d functions are taken as

$$m_6(Z_1) = Z_1, \ \ m_7(Z_2) = Z_2^2, \ \ m_8(Z_3) = Z_3^2, \ \text{and} \ m_9(Z_4) = \sin(\pi Z_4). \quad (14)$$

Notice that these functions are not centered at mean zero. However, it does not violate assumptions of model (1) as constants needed to center those functions contribute to the intercept term α. The smoother using Nadaraya–Watson estimator (Watson 1964) is taken to smooth Z_3 and Z_4. We used the default bandwidth parameter for the kernel density as $h = 1.06 s n^{-1/5}$, where s is the standard deviation of corresponding variable and n is the sample size. We have taken two different types of error distributions— the standard normal distribution and the Chi-square distribution with five degrees of freedom. A sample of size 500 from (Y, X, Z) is generated, and this exercise is replicated 1000 times. The histograms of the observed GLR nonparametric test statistic are presented in Fig. 1, and the corresponding kernel density estimates are also plotted. The plots show that the empirical distributions match with the theoretical null distribution obtained from Corollary 4. The plots give an indication of inflated level, but further simulation studies show that the convergence improves as sample size increases. In the same figure, we have plotted the density of the null distribution of the GLR test under independence assumption on the predictors. It is a single Chi-square distribution with degrees of freedom $\sum_{p=1}^{P}(k_p - 1) = 14$. It gives a good
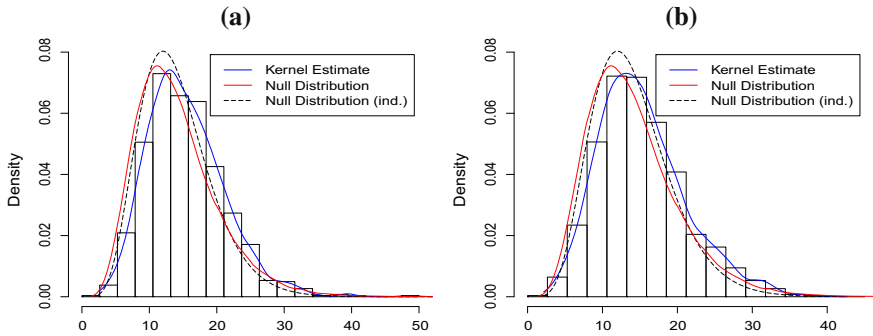
Fig. 1 The observed kernel density estimate of the GLR test, the fitted theoretical null distribution and its approximation assuming all predictors are independent. The error distribution is **a** $N(0, 1)$ and **b** $\chi^2(5)$

approximation of the null distribution. In fact, further simulation studies show that, unless some predictors are strongly correlated, this approximation works reasonably well. In Fig. 1a, b, the error distributions are different in two plots, so they demonstrate that the null distribution of the GLR test does not depend on the choice of the error distribution.

5.2 Power function

In the power calculation for testing $H_0^{**} : m_1 = m_2 = \cdots = m_5 = 0$, we have taken the same setup of the previous example, except for the distribution of X_1 and the corresponding m_1 function. Here, we have taken

$$m_1(x_1) = \beta(x_1 - 0.75)^2, \qquad (15)$$

and the distribution of X_1 is given by $P(X_1 = 0) = p^2$, $P(X_1 = 1) = 2pq$ and $P(X_1 = 2) = q^2$, where $p = 1 - q = 0.75$. So, it violates the null hypothesis H_0^{**} if $\beta \neq 0$. As we compare the power of the GLR test and the classical F test associated with the nested linear models, the error term is generated from the standard normal distribution to make all results comparable. For each value of β, we simulated 500 samples and repeated it 1000 times to calculate the observed power. The observed power is the proportion of the test statistics greater than the corresponding critical value at 5% level of significance obtained from Corollary 4. In Fig. 2a, the GLR test shows a good power, but the F test completely fails in this situation. The GLR test is slightly anti-conservative at null, which is also reflected in Fig. 1. In the same plot, we presented the observed powers of few other tests including the GLR test under independence assumption, abbreviated as GLR (ind.). Its performance is very similar to the original GLR test although predictors were not pairwise independent. We also plotted the theoretical power function of the GLR test calculated from Corollary 2 under the independence assumption. The observed and the theoretical power functions are very close to each other.
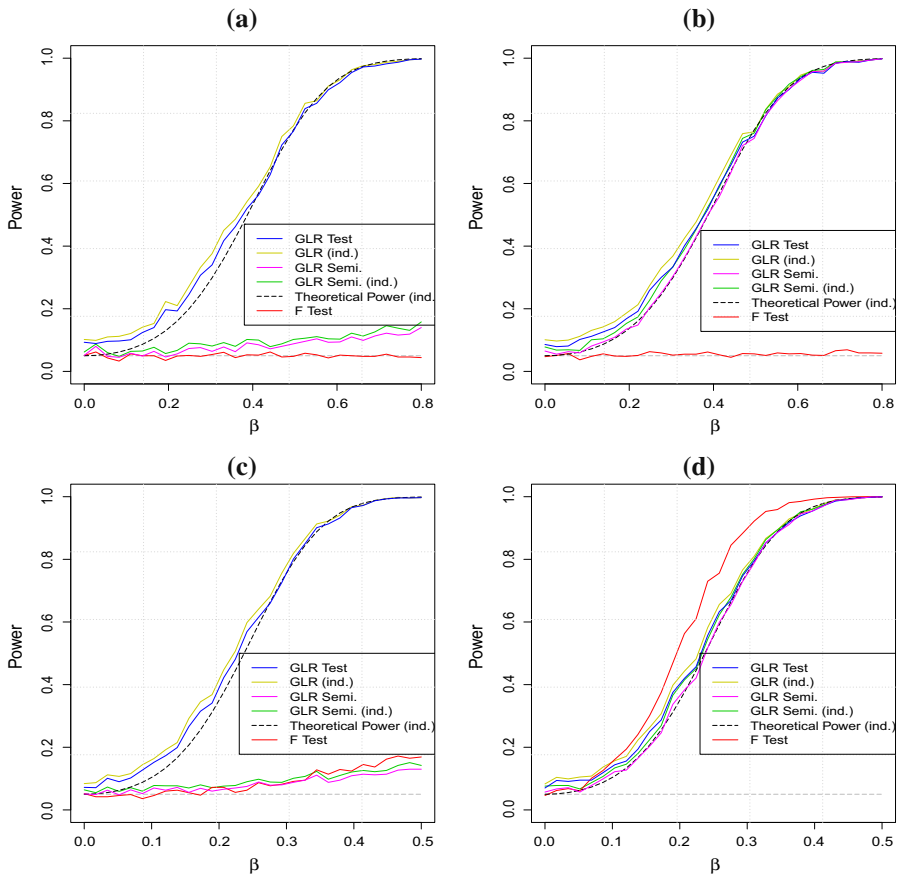
(a) **(b)**

(c) **(d)**

Fig. 2 Simulated powers of the F test and different GLR tests when—**a** both Y, X and Y, Z are nonlinear, **b** Y, X is nonlinear but Y, Z linear, **c** Y, X is linear but Y, Z nonlinear, and **d** both Y, X and Y, Z are linear

If we fit a linear regression model in this setup, the breakdown situation of the F test is apparent as the least square estimates of the regression coefficients vanish in the large sample sizes. Conditioning on the other variables the expected value of the least square estimate of the regression coefficient corresponding to X_1 turns out to be

$$\beta_{LS} = \frac{\text{Cov}(X_1, Y)}{\text{Var}(X_1)} = -pm_1(0) + (1 - 2q)m_1(1) + qm_1(2),$$

which simplifies to zero for all values of β in Eq. (15) when $p = 0.75$. Thus, H_0^{**} seems to be true for all values of β with respect to a linear model, and the power function for the F test centers around the nominal level of the test. On the other hand, the full additive model successfully captures the nonlinear relationship, and the power of the GLR test tends to one as β increases.

In the same plot, we presented the observed power function of the GLR semi-parametric test (abbreviated as GLR Semi.), assuming that all covariates are linearly

related to Y (although it is not true), where the predictors are modeled nonparametrically. As the covariates are highly nonlinear, including a sin function, the semiparametric test breaks down. Notice that the distribution of the GLR test statistic does not change if the covariates are modeled parametrically, so the theoretical critical value of the GLR semiparametric test is same as the GLR nonparametric test, and it is obtained from Corollary 4. The GLR Semi. (ind.) test in Fig. 2a is the approximation of the GLR semiparametric test under independence assumption on the predictor variables. The performance of these two tests is similar to the F test. It shows that a wrong model of the covariates may cause severe power loss.

We have investigated few more cases by generating different relationships between Y, X and Y, Z. Figure 2b presents power functions where Y, X is nonlinear as given in Equation (15), but Y, Z is linear. The functions corresponding to the linear relationships between Y and Z are taken simply as $m_d(z) = z$ for $d = 6, 7, 8, 9$ instead of nonlinear functions in Eq. (14). Here, we get the similar results from the GLR and F test. The GLR semiparametric test gives almost equal power as described by its theoretical power function derived from Corollary 2. In fact, in this case, the theoretical power of the GLR semiparametric test is same as the GLR nonparametric test. But the advantage of semiparametric modeling is that its finite sample performance is better than the full nonparametric test, if the modeling of the parametric part is correct. For this reason, the nonparametric GLR test is showing slightly inflated level, whereas the GLR semiparametric test properly maintains the level of the test.

In Fig. 2c, we plot the power functions when Y, X is linear by taking $m_1(x) = \beta x$ instead of Eq. (15), but Y, Z is nonlinear as given in Eq. (14). Even if the F test successfully models the relationship between Y and X, its power is almost unchanged as it fails to model the relationship between Y and Z. Similarly, the GLR semiparametric test is also showing poor power.

Finally, the power functions of these tests, when both the relationships between Y, X and Y, Z are linear, are plotted in Fig. 2d. Here, all assumptions of the F test are satisfied, so it is the most powerful among all unbiased tests. It is interesting to notice that the power of all GLR tests is very competitive with the F test. Therefore, even if the both relationships are linear, we do not expect to lose a significance amount of power by conducting the GLR test. These simulation results show that it is better to use the GLR test unless we are confident of a suitable parametric model. If some assumptions of the parametric model are violated, the F test may break down. On the other hand, the GLR test produces very high power almost in all situations.

5.3 The goodness-of-fit test

In a similar setup, we have studied the general GLR test including the goodness-of-fit testing problem for the semiparametric model. We test the null hypothesis (7) that there exists a linear relationship between Y and X_2, but there is no effect of other components, i.e., $m_1 = m_3 = m_4 = m_5 = 0$. We did not assume any parametric
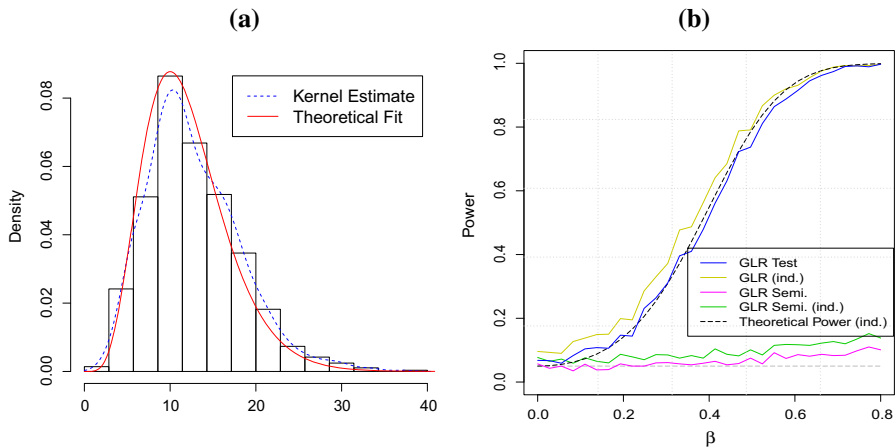
Fig. 3 **a** The observed and theoretical null distributions of the GLR semiparametric test statistic for the goodness-of-fit problem. **b** The power functions of different GLR semiparametric test statistics

model for the covariates. In this simulation, we have taken $m_2(x_2) = \frac{1}{2}x_2$ and the same m_1 as given in Eq. (15). So, the null hypothesis is true when $\beta = 0$, and power of the GLR test should increase as β deviates from zero. The other setup for the simulation is same as the previous simulations including the functions for the covariates as given in Eq. (14). Figure 3a shows that the observed and theoretical null distributions are close to each other. The power functions of different tests are shown in Fig. 3b. All tests in this plot are semiparametric tests; however, to make similarity with the previous simulation, we denote "GLR Semi." when all covariates are assumed to be linear. "GLR Test" refers to the main semiparametric test whose null distribution is derived from Theorem 1 without assuming that covariates are linearly modeled. Similarly, "GLR (ind.)" is the approximation of this test by assuming that all predictors are pairwise independent. The black dotted line in the plot is the approximate power function derived for Corollary 2 that assumes all predictors are pairwise independent. The main GLR test maintains the nominal level of the test and gives good power when the null hypothesis is not true. GLR (ind.) shows slight inflated power; however, it gives good approximation of the original test. GLR Semi. and GLR Semi. (ind.) fail to produce any significant power as the linearity assumptions on covariates are not satisfied.

As a whole, these simulation results in Sect. 5 give enough justification for the theoretical results derived in the paper. These GLR tests are simple to calculate and produce a good power. As a virtue of the nonparametric method, the tests do not depend on the error distribution of the model. For the model utility test, the GLR test gives better power than the classical F test when the parametric modeling is not appropriate. And even if the parametric model holds good, the GLR test produces a comparative power. The GLR test can be further simplified if we assume that the predictors are pairwise independent. In this case, the test is slightly anti-conservative, but overall the approximation is good unless some predictors are strongly correlated.

6 Real data example

In this section, we apply the GLR test to analyze the diamonds dataset used in Wickham (2009). This dataset contains the price (in 2008 US dollars) and other attributes of 53,940 diamonds. The attributes include the four C's of diamond quality—cut, color, clarity and carat. There are three main physical measurements x, y and z—the largest length, width and height of a diamond, respectively. The dataset has other two physical measurements—depth and table, but we have not included them in this analysis as they are functions of x, y and z. Carat is a unit of mass used for measuring gemstones and pearls. Cut is an objective measure of a diamond's light performance what we generally think of as sparkle. Cut, color and clarity are categorical variables, and other variables are continuous. There are five categories of cut—Fair, Good, Very Good, Premium and Ideal, and the percentage of each diamond in this dataset is 2.98, 9.10, 22.30, 25.57 and 39.95, respectively. Color has seven categories D (best) to J (worst) with 12.56%, 18.16%, 17.69%, 20.93%, 15.40%, 10.05% and 5.21%, respectively. Clarity contains eight categories I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1 and IF (best) with 1.37%, 17.04%, 24.22%, 22.73%, 15.15%, 9.39%, 6.77% and 3.32%, respectively.

As cut, color and clarity are categorical variables, we apply the GLR test for testing their effect in diamond price, whereas carat, x, y and z serve as covariates. The sample size is huge, so a slight deviation from a null hypothesis may cause rejection of the null hypothesis. For this reason, we have taken a random subsample of 1000 observations from the dataset. At first, we test the null hypothesis (13) that there is no effect of cut, color and clarity. So, $H_0^{**} : m_1 = m_2 = m_3 = 0$. The p value of the GLR test comes out to be zero. Then, we have conducted three different tests to check the main effect of cut, color and clarity, separately, by taking others as covariates (e.g., $H_0^{**} : m_1 = 0$). Even in these cases p values of the GLR tests are 6.45×10^{-5} for cut and zero for color and clarity. This is not surprising as there is a strong effect of cut, color and clarity in the price of diamonds.

All categorical variables, cut, color and clarity, maintain an order in their quality. So, in the next step, we have conducted some goodness-of-fit tests to know the effect of different levels of those variables. We are particularly interested to know whether there is any specific pattern in the order of different levels of a category. The levels are ranked according to the increasing order of quality. In our first goodness-of-fit test, we have taken the order of cut as a single predictor ($P = 1$) and six other variables including the ranks of color and clarity as covariates ($Q = 6$). We tested the null hypothesis that the diamond price is linear with the rank of the levels of cut. The p value of the GLR test is 0.5476. This indicates that the diamond price may be modeled as a linear function of the rank of its cutting quality. The box plots in Fig. 4a show the partial effect of cut after eliminating the effects of other variables. The middlemost line of median in the box plot is replaced by the corresponding mean, so it gives the mean effect of cut for that category. The red line in the plot is the least square regression line of the price difference on the rank of the cut effect. The plot also shows that the ranks in cut maintain a proper linear relationship in modeling the diamond price.

In the second semiparametric test, we checked the linearity effect of the rank of color in diamond price. The GLR test gave a p value of 2.02×10^{-8}, so we considered
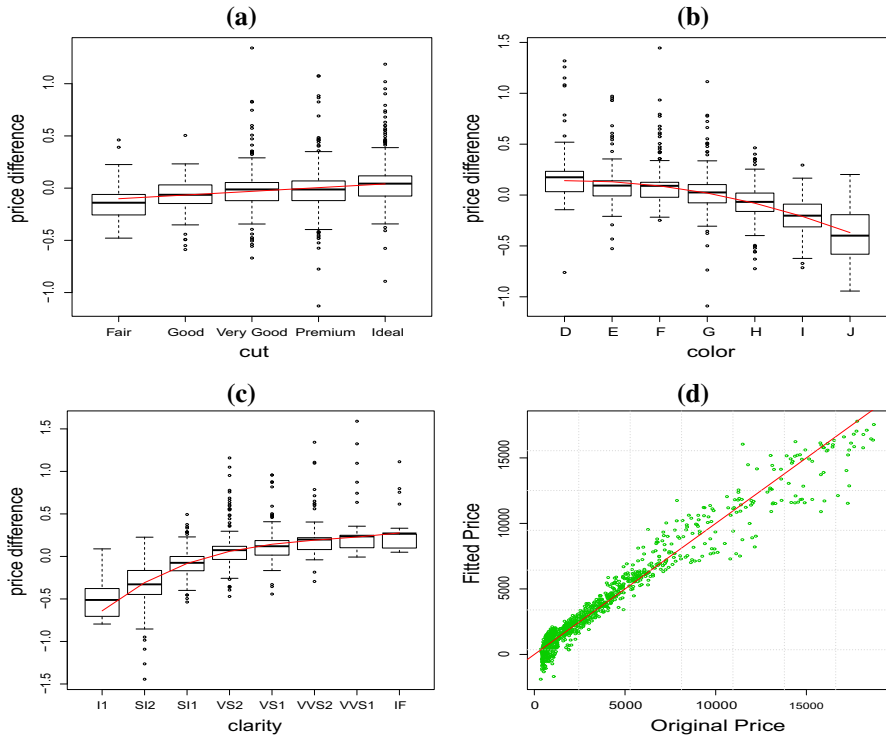
Fig. 4 Box plots showing the partial effect of **a** cut, **b** color and **c** clarity on the diamond price. The red line gives the fitted least square estimates, where x-axis is converted to the corresponding rank. **d** Plot of fitted price from the semiparametric model against the original price. The red line indicates the perfect fit (color figure online)

a quadratic model for the rank of color. The p value of the test came out to be 0.0857, so the quadratic effect of color is not rejected at 5% level of significance (see Fig. 4b). Similarly, the p values of corresponding to the linear, quadratic and cubic models of the rank of clarity are 10^{-15}, 0.0412 and 0.4233, respectively. So, the effect of clarity in diamond price may be modeled as a cubic function of its rank (see Fig. 4c).

Finally, we have constructed a semiparametric model where the ranks of cut, color and clarity were taken as linear, quadratic and cubic functions, respectively. Carat, x, y and z serve as covariates for this test, and they are modeled nonparametrically. So, according to our notations $P = 3$ and $Q = 4$. The GLR test (11) for the goodness-of-fit of the model gave a p value of 0.1104 for this semiparametric model. Figure 4d displays the scatter plot of the original price of diamonds and their fitted price from the semiparametric model; the correlation between the two prices is 0.9632. The points in the plot cluster around the red line which represents the perfect fit. These results indicate that if we build a semiparametric model for these data the appropriate choices of the additive components for the rank of cut, color and clarity are linear, quadratic and cubic functions, respectively.

The estimate of the overall mean price for the dataset is $\hat{\alpha} = 3931.3760$, and the standard deviation is $\hat{\sigma} = 4149.5376$. Let us consider an additive model as $y =$

$\hat{\alpha} + \hat{\sigma} \sum_{d=1}^{7} m_d$, where $m_1, m_2, \ldots, m_7$ being the additive effects of the rank of cut, color, clarity, carat, x, y and z, respectively. Then, the parametric parts of the fitted semiparametric model are given by

$$m_1(x_1) = -0.1330 + 0.0344x_1, \quad m_2(x_2) = 0.1271 + 0.0312x_2 - 0.0146x_2^2,$$
$$m_3(x_3) = -1.1149 + 0.5466x_3 - 0.0801x_3^2 + 0.0042x_3^3. \tag{16}$$

7 Discussion

In an additive model, a novel method is derived for testing the main effect of the predictor variables which take discrete or categorical values. The main effect is adjusted by covariates possibly containing continuous-valued random variables. The predictors and covariates are modeled nonparametrically using an additive model, so the test avoids loss of power because model misspecification often arises in classical parametric tests. This method is further extended to the semiparametric model, and a goodness-of-fit test is derived. The simulation results show that the GLR test may outperform the parametric test when the model for just one of the components fails, and at the same time, it produces a comparable power as the conventional tests if the assumed parametric model holds good. The power of the GLR test is asymptotically optimal in terms of the rate of convergence, and it can detect a specific class of contiguous alternatives at a rate of $n^{-1/2}$. In case of categorical predictors, the GLR test generalizes the classical ANCOVA by modeling covariates nonparametrically and without assuming normality of the error term. So, it is a development of the basic statistical theory, and the methodology can be widely useful in practice.

Appendix A: Regularity conditions

To derive the asymptotic distribution of the GLR test statistic, we need the following assumptions:

(C1) Suppose $c_{pj} = P(X_p = x_{pj})$, then $c_{pj} \in (0, 1)$ for all $j = 1, 2, \ldots, k_p$ and $p = 1, 2, \ldots, P$, where $\sum_{j=1}^{k_p} c_{pj} = 1$.

(C2) The kernel function $K(z)$ is bounded and Lipschitz continuous with a bounded support.

(C3) If Z_q is continuous, then the density f_q of Z_q is Lipschitz continuous and bounded away from 0 and has bounded supports Ω_q for $q \in \{1, 2, \ldots, Q\}$.

(C4) If both Z_q and $Z_{q'}$ are continuous, then the joint density $f_{qq'}$ of Z_q and $Z_{q'}$ is Lipschitz continuous on its support $\Omega_q \times \Omega_{q'}$ for $q \neq q' \in \{1, 2, \ldots, Q\}$.

(C5) $nh_q / \log(n) \to \infty$ as $n \to \infty$ and $h_q \to 0$ for $q = 1, 2, \ldots, Q$.

(C6) If Z_q is continuous and d_q is the degree of the polynomial used for smoothing of Z_q, then the $(d_q + 1)$th derivative of m_{P+q}, for $q \in \{1, 2, \ldots, Q\}$, exists and is bounded and continuous.

(C7) $\sigma^2 = \text{Var}[\epsilon] = E[\epsilon^2] < \infty$.

(C8) $E[m_q(Z_q)|X_p = x_{pj}] = 0$ for all $j = 1, 2, \ldots, k_p$, $p = 1, 2, \ldots, P$ and $q = 1, 2, \ldots, Q$.

References

Buja, A., Hastie, T., Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, *17*(2), 453–555.

Davies, R. B. (1980). The distribution of a linear combination of χ^2 random variables. Algorithm AS155. *Applied Statistics*, *29*, 323–333.

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. *The 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.

Hall, P., Marron, J. S. (1988). Variable window width kernel estimates of probability densities. *Probability Theory and Related Fields*, *80*(1), 37–49.

Hastie, T., Tibshirani, R. (2000). Bayesian backfitting (with discussion). *Statistical Science. A Review Journal of the Institute of Mathematical Statistics*, *15*(3), 196–223.

Hastie, T. J., Tibshirani, R. J. (1990). *Generalized additive models, volume 43 of monographs on statistics and applied probability*. London: Chapman and Hall Ltd.

Ingster, Y. I. (1993a). Asymptotically minimax hypothesis testing for nonparametric alternatives. I. *Mathematical Methods of Statistics*, *2*, 85–114.

Ingster, Y. I. (1993b). Asymptotically minimax hypothesis testing for nonparametric alternatives. II. *Mathematical Methods of Statistics*, *3*, 171–189.

Ingster, Y. I. (1993c). Asymptotically minimax hypothesis testing for nonparametric alternatives. III. *Mathematical Methods of Statistics*, *4*, 249–268.

Jiang, J., Zhou, H., Jiang, X., et al. (2007). Generalized likelihood ratio tests for the structure of semiparametric additive models. *The Canadian Journal of Statistics*, *35*(3), 381–398.

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.

Opsomer, J. D. (2000). Asymptotic properties of backfitting estimators. *Journal of Multivariate Analysis*, *73*(2), 166–179.

Opsomer, J. D., Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *The Annals of Statistics*, *25*(1), 186–211.

Opsomer, J. D., Ruppert, D. (1999). A root-n consistent backfitting estimator for semiparametric additive modeling. *Journal of Computational and Graphical Statistics*, *8*(4), 715–732.

Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society. Series B. Methodological*, *50*(3), 413–436.

Sperlich, S., Tjøstheim, D., Yang, L. (2002). Nonparametric estimation and testing of interaction in additive models. *Econometric Theory*, *18*(2), 197–251.

Spokoiny, V. G. (1996). Adaptive hypothesis testing using wavelets. *The Annals of Statistics*, *24*(6), 2477–2498.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, *13*(2), 689–705.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, *14*(2), 590–606.

Tjøstheim, D., Auestad, B. H. (1994). Nonparametric identification of nonlinear time series: Projections. *Journal of the American Statistical Association*, *89*(428), 1398–1409.

Wand, M. P. (1999). A central limit theorem for local polynomial backfitting estimators. *Journal of Multivariate Analysis*, *70*(1), 57–65.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā (Statistics). The Indian Journal of Statistics. Series A*, *26*, 359–372.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York: Springer.

Yang, L., Sperlich, S., Härdle, W. (2003). Derivative estimation and testing in generalized additive models. *Journal of Statistical Planning and Inference*, *115*(2), 521–542.