Computational Statistics
https://doi.org/10.1007/s00180-024-01479-0

ORIGINAL PAPER

A difference-based method for testing no effect in nonparametric regression

Zhijian Li[1] · **Tiejun Tong**[2] · **Yuedong Wang**[3]

Received: 11 June 2023 / Accepted: 21 February 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract
The paper proposes a novel difference-based method for testing the hypothesis of no relationship between the dependent and independent variables. We construct three test statistics for nonparametric regression with Gaussian and non-Gaussian random errors. These test statistics have the standard normal as the asymptotic null distribution. Furthermore, we show that these tests can detect local alternatives that converge to the null hypothesis at a rate close to $n^{-1/2}$ previously achieved only by the residual-based tests. We also propose a permutation test as a flexible alternative. Our difference-based method does not require estimating the mean function or its first derivative, making it easy to implement and computationally efficient. Simulation results demonstrate that our new tests are more powerful than existing methods, especially when the sample size is small. The usefulness of the proposed tests is also illustrated using two real data examples.

Keywords Difference-based test · Asymptotic normality · Locally most powerful test · Nonparametric regression · Permutation · Residual-based test

✉ Zhijian Li
zhijianli@uic.edu.cn

Tiejun Tong
tongt@hkbu.edu.hk

Yuedong Wang
yuedong@pstat.ucsb.edu

1 Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College, Zhuhai, People's Republic of China

2 Department of Mathematics, Hong Kong Baptist University, Hong Kong, People's Republic of China

3 Department of Statistics and Applied Probability, University of California, Santa Barbara, CA, USA

1 Introduction

Consider a regression model of the form

$$y_i = g(x_i) + \epsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where y_i and x_i are the ith observations of the scalar dependent and independent variables, g is a mean function, and ϵ_i are independent and identically distributed (i.i.d.) random errors with mean zero and variance $\sigma^2 > 0$. In regression analysis, we are often interested in testing the null hypothesis of no relationship between the dependent and independent variables:

$$H_0 : g(x) = \text{constant} \quad \text{versus} \quad H_1 : g(x) \neq \text{constant}. \tag{2}$$

For example, it is of interest to investigate whether the COVID-19 incubation period depends on age (Tan et al 2020), whether the adult human gut microbial depends on age (Zhang et al 2021), and to identify genes that show statistically significant changes in expression over time (Storey et al 2005).

When the function g is modeled parametrically by a linear model, the F test is the standard approach for testing no effect as in (2). However, a parametric model for g is often difficult to specify or can be too restrictive in many applications. Nonparametric regression approaches for testing no effect have been considered by many authors (Barry and Hartigan 1990; Raz 1990; Chen 1994; Eubank 2000; Yatchew 2003; Li 2012; Van Keilegom et al 2008; González-Manteiga and Crujeiras 2013). Most of the existing tests are based on a nonparametric fit to the mean function g, or its first derivative g', using nonparametric smoothing techniques. In this paper, we propose a novel difference-based method for testing no effect without needing to estimate g or its first derivative.

The difference-based method was primarily developed for the estimation of error variance σ^2 that does not require an estimate of g (Rice 1984; Gasser et al 1986; Hall et al 1990; Tong and Wang 2005; Tong et al 2013). The idea of differencing has also been used in testing the independence of X and ϵ (Einmahl and Van Keilegom 2008). The ρth-order differencing estimator for σ^2 in (1) is defined as

$$s_d^2 = \frac{1}{n-\rho} \sum_{i=1}^{n-\rho} \Big(\sum_{j=0}^{\rho} d_j y_{j+i} \Big)^2,$$

where the positive integer ρ is the order of differentiation and $(d_0, \dots, d_\rho)$ are the differencing weights that satisfy the regularity conditions $\sum_{j=0}^{\rho} d_j = 0$ and $\sum_{j=0}^{\rho} d_j^2 = 1$. In recent years, the difference-based method has been applied to the estimation of derivatives in nonparametric regression (Brabanter et al 2013; Wang and Lin 2015; Dai et al 2016; Wang et al 2019; Zhang and Dai 2023), the estimation of covariance (Bliznyuk et al 2012), and the estimation of time-varying auto-covariance (Cui et al 2021). With the exception of Yatchew (1999) and Yatchew (2003), the difference-based method has not been used for the purpose of hypothesis testing. Yatchew (2003) proposed a specification test statistic $S_1 = (n/4\delta)^{1/2}(s_{\text{res}}^2 - s_d^2)/s_{\text{res}}^2$, where

$\delta = \sum_{i=1}^{\rho} \left(\sum_{j=0}^{\rho-i} d_j d_{j+i} \right)^2$ and $s_{\text{res}}^2 = n^{-1} \sum_{i=1}^{n} \left(y_i - m(x_i, \hat{\boldsymbol{\gamma}}) \right)^2$, for the null hypothesis $g(x) = m(x, \boldsymbol{\gamma})$ where m is a known function with unknown parameters $\boldsymbol{\gamma}$, against a nonparametric alternative. Under the null hypothesis, $S_1 \xrightarrow{D} N(0,1)$ as $n \to \infty$, where $\xrightarrow{D}$ denotes convergence in distribution. If the weight is the optimal difference sequence (Hall et al 1990), then the test statistic can be simplified as $S_2 = (n\rho)^{1/2}(s_{\text{res}}^2 - s_d^2)/s_{\text{res}}^2$. Under the null hypothesis, $S_2 \xrightarrow{D} N(0,1)$ as $n \to \infty$. To apply the specification test to the hypotheses in (2) under the null hypothesis that $g(x)$ is a constant function, the estimate of $m(x, \hat{\boldsymbol{\gamma}})$ is simply the sample mean of the observed y_i values. The specification test has a convergence rate close to $n^{-1/4}$ for any fixed differencing order r (Yatchew, Yatchew 2003, p. 68), which is far slower than that of the residual-based tests, i.e. $n^{-1/2}$ (Neumeyer and Dette 2003).

To propose a new difference-based test that does not require an estimate of g, we first convert the hypothesis of no effect in (2) as a new hypothesis of zero slope in a linear model for differences. We then construct three difference-based statistics for testing zero slope that are easy to implement and computationally efficient. Our new tests can detect local alternatives that converge to the null hypothesis at a rate close to $n^{-1/2}$, which was previously achieved only by the residual-based tests as the optimal rate. The simulations show that the new tests compare favorably with existing methods. Moreover, we also extend the proposed difference-based method to more general settings.

The remainder of the paper is organized as follows. In Sect. 2, we present three new difference-based test statistics for hypothesis (2) and derive their asymptotic or approximate null distributions. In Sect. 3, we conduct simulation studies to evaluate the finite-sample performance of the proposed tests and compare them with existing methods. In Sect. 4, we apply the difference-based tests to two real data examples to illustrate their usefulness in practice. In Sect. 5, we extend the difference-based testing method to more general problems, including the test for polynomial functions, the test for parallelity of two mean functions, and the test with unequally spaced design points. We present the technical results in the Appendix.

2 Difference-based tests

For simplicity, we consider equally spaced design points with $x_i = i/n$ for $i = 1, \dots, n$. Define the lag-k Rice estimators as

$$s_k = \frac{1}{2(n-k)} \sum_{i=1}^{n-k} \left(y_{i+k} - y_i \right)^2, \quad k = 1, \dots, n-1.$$

We further assume that g has a bounded first derivative. Then by the Taylor expansion,

$$
\begin{aligned}
E(s_k) &= \sigma^2 + \frac{1}{2(n-k)} \sum_{i=1}^{n-k} \left[g(x_{i+k}) - g(x_i)\right]^2 \\
&= \sigma^2 + \frac{1}{2(n-k)} \sum_{i=1}^{n-k} \left[\frac{k}{n} g'(x_i) + o\left(\frac{k}{n}\right)\right]^2 \\
&= \sigma^2 + \beta d_k + o(d_k),
\end{aligned}
$$

where $\beta = \int_0^1 [g'(x)]^2 dx/2$ and $d_k = k^2/n^2$. Now to estimate σ^2, for any $m = o(n)$, Tong and Wang (2005) fitted a linear model as

$$
s_k = \alpha + \beta d_k + \eta_k, \quad k = 1, \dots, m, \tag{3}
$$

and then applied the fitted intercept $\hat{\alpha}$ as the final estimate of σ^2 which can achieve the optimal rate in MSE. Tong et al (2013) further showed that the least squares estimator using the linear model (3) is asymptotically normal, root-n consistent, and reaches the optimal bound in terms of the estimation variance.

We note, however, that there has been restricted attention to the variance estimation in the existing literature, which mainly focused on the estimation of the intercept α. In contrast, the estimate of β is only used as a term in the variance estimate, whereas the statistical inference for the slope itself is largely overlooked. In this paper, we show for the first time that β can indeed play an important role in the hypothesis testing for the mean function. Note that the null hypothesis in (2) holds if and only if $g'(x) = 0$ for all $x \in [0, 1]$, which is equivalent to $\beta = \int_0^1 \left[g'(x)\right]^2 dx/2 = 0$. This shows that the hypotheses in (2) can be converted to the new hypotheses as

$$
H_0 : \beta = 0 \quad \text{versus} \quad H_1 : \beta > 0. \tag{4}
$$

To test the null hypothesis in (4), we first derive a weighted least squares estimator of β and then establish its asymptotic normality. Following from (3), the weighted least squares (WLS) estimator of β is given by

$$
\hat{\beta} = \frac{\sum_{k=1}^{m} w_k s_k \left(d_k - \bar{d}_w\right)}{\sum_{k=1}^{m} w_k \left(d_k - \bar{d}_w\right)^2}, \tag{5}
$$

where $\bar{d}_w = \sum_{k=1}^{m} w_k d_k$, $w_k = (n-k)/N$ and $N = nm - m(m+1)/2$. We choose the weight $w_k = (n-k)/N$ because s_k is the average of $(n-k)$ squared lag-k differences.

Moreover, the WLS estimator of β can be written as

$$
\hat{\beta} = \frac{1}{2N} \mathbf{y}^T \boldsymbol{B} \mathbf{y}, \tag{6}
$$

where $\mathbf{y} = (y_1, \dots, y_n)^T$ and $\boldsymbol{B} = (b_{ij})_{n \times n}$ is a symmetric matrix with elements

$$b_{ij} = \begin{cases} \sum_{k=1}^{m} h_k + \sum_{k=1}^{\min(i-1,n-i,m)} h_k, & 1 \le i = j \le n, \\ -h_{|i-j|}, & 0 < |i-j| \le m, \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

$h_0 = 0$, and $h_k = (d_k - \bar{d}_w)/\sum_{k=1}^{m} w_k(d_k - \bar{d}_w)^2$ for $k = 1, \dots, m$. The trace of $\boldsymbol{B}$ is $\mathrm{tr}(\boldsymbol{B}) = 2\sum_{k=1}^{m}(n-k)h_k = 2N\sum_{k=1}^{m} w_k h_k = 0$. Let also $\gamma_4 = E(\epsilon^4)/\sigma^4$, which equals 3 when the errors are normally distributed. In Appendix B, we establish the asymptotic normality for the WLS estimator $\hat{\beta}$. Throughout this paper, we take the bandwidth m to be an integer. We use the ceiling function, and let $\lceil n^r \rceil$ be the smallest integer that is greater than or equal to n^r.

Theorem 1 *Assume that the mean function $g(\cdot)$ has a bounded second derivative and $E(\epsilon^6)$ is finite. For any $m = \lceil n^r \rceil$ with $2/3 < r < 1$, the WLS estimator in (6) has the asymptotic distribution*

$$\sqrt{n^{3r-2}}\left(\hat{\beta} - \beta\right) \xrightarrow{D} N\left(0, \sigma_b^2\right) \quad \text{as } n \to \infty, \tag{8}$$

where $\sigma_b^2 = (15/56)(\gamma_4 - 1)\sigma^4$.

By Theorem 1, the asymptotic variance of $\hat{\beta}$ is $\sigma_\beta^2 = (15/56)n^{2-3r}(\gamma_4 - 1)\sigma^4$. When the errors are normally distributed, a direct estimate of the asymptotic variance is $\hat{\sigma}_\beta^2 = (15/28)n^{2-3r}\hat{\sigma}^4$, where $\hat{\sigma}^2$ is a consistent estimator of the error variance σ^2. However, this direct estimate may not provide an accurate approximation when the sample size n is not large enough. For more details, see Appendix B, in which we also suggest a more accurate estimate of the error variance with a higher-order term:

$$\tilde{\sigma}_\beta^2 = \frac{1}{4N^2}\left[\frac{15n^4}{7m}\hat{\sigma}^4 + \frac{45n^5}{m^3}\hat{\sigma}^4\right] = \frac{1}{N^2}\left[\frac{15n^4}{28m} + \frac{45n^5}{4m^3}\right]\hat{\sigma}^4. \tag{9}$$

When $m = o(n)$, both $\tilde{\sigma}_\beta^2$ and $\hat{\sigma}_\beta^2$ are consistent estimators of σ_β^2. We define the difference-based test (DBT) statistic for the null hypothesis in (4) as

$$T = \frac{\hat{\beta}}{\tilde{\sigma}_\beta}. \tag{10}$$

Theorem 2 *Assume that $\epsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$ and let $\hat{\sigma}^2$ be a consistent estimator of the error variance σ^2. Under the assumptions in Theorem 1 and the null hypothesis in (4), we have $T \xrightarrow{D} N(0,1)$ as $n \to \infty$.*

By Theorem 2, we then reject the null hypothesis that $\beta = 0$ if the observed value of T is greater than z_α, where α is the significance level and z_α is the upper αth percentile of the standard normal distribution. To assess the power of the test, we consider the Pitman local alternative (McManus 1991) that $H_{1n} : \beta = h/a_n$, where $a_n \to \infty$. Under H_{1n}, we have

$$\frac{1}{\tilde{\sigma}_\beta}\left(\hat{\beta} - \frac{h}{a_n}\right) \xrightarrow{D} N(0,1) \quad \text{as } n \to \infty.$$

This yields the power function

$$\pi_n\left(\frac{h}{a_n}\right) = P\left(\frac{1}{\tilde{\sigma}_\beta}(\hat{\beta} - \frac{h}{a_n}) > z_\alpha - \frac{1}{\tilde{\sigma}_\beta}\frac{h}{a_n}\right) = 1 - \Phi\left(z_\alpha - \frac{1}{\tilde{\sigma}_\beta}\frac{h}{a_n}\right) + o(1).$$

If $a_n = \sqrt{n^{3r-2}}$, the power function $\pi_n(h/a_n)$ tends to 1 as $h \to \infty$. This shows that the proposed test statistic T can detect local alternatives that converge to the null hypothesis at a rate of $\sqrt{n^{2-3r}}$. Recall that the specification test in Yatchew (2003) can detect local alternatives that converge to the null hypothesis at a rate close to $n^{-1/4}$. This shows that the convergence rate of our new test is faster than that of the specification test as long as $r > 5/6$. And more importantly, the convergence rate of our new test will approach to the optimal rate at $n^{-1/2}$ as $r \to 1$, which was previously achieved only by the residual-based tests.

When the errors are not normally distributed, γ_4 is also unknown in the asymptotic variance of $\hat{\beta}$. To have a valid test statistic in this case, we propose to replace the whole unknown term $(\gamma_4 - 1)\sigma^4 = E(\epsilon^4) - \sigma^4 = \kappa - (\sigma^2)^2$ by a consistent estimator $\hat{\kappa} - (\hat{\sigma}^2)^2$, where $\hat{\sigma}^2$ is from Tong and Wang (2005) and $\hat{\kappa}$ is from Evans and Jones (2008). To be more specific, we have $\hat{\sigma}^2 = \sum_{k=1}^m w_k s_k - \hat{\beta}\bar{d}_w$ and $\hat{\kappa} = \sum_{i=1}^n [\Pi_{j=1}^4 (y_i - y_{i(j)})]/n$, where $i(j)$ is the index of the jth nearest neighbor of x_i among $x_1, \dots, x_n$.

Theorem 3 *Assume that ϵ_i are i.i.d. random variables with mean zero and variance σ^2. Let $\hat{\sigma}^2$ and $\hat{\kappa}$ be consistent estimators of the error variance σ^2 and the fourth moment κ, respectively. Define the difference-based test statistic*

$$G = \frac{\hat{\beta}}{\check{\sigma}_{\beta g}}, \tag{11}$$

where

$$\check{\sigma}^2_{\beta g} = \frac{1}{4N^2}\left[\frac{15n^4}{14m}\left(\hat{\kappa} - \hat{\sigma}^4\right) + \frac{45n^5}{m^3}\hat{\sigma}^4\right]. \tag{12}$$

Under the assumptions in Theorem 1 and the null hypothesis in (4), we have $G \xrightarrow{D} N(0,1)$ as $n \to \infty$.

The proofs of Theorems 2 and 3 are given in Appendix C. We note that the results in these two theorems hold for any consistent estimators of σ^2 and κ. In addition, for the test statistic G, its power function follows the same structure as that of the test statistic T, with the only change being the standard error in (9) replaced by the one in (12).

Next, we consider the bandwidth m selection in practice. Under the normality assumption, following Lemma 2 (a) and Theorem 1 and 2, it is easy to see that the mean square error (MSE) of $\hat{\beta}$ is approximately $\tilde{\sigma}^2_{\beta} + O((m+1)/(2n-m-1))$. As part of the bias can be computed as $\beta(m+1)/(2n-m-1)$, we can select the m that minimize the value $\widetilde{\text{MSE}}(\hat{\beta}) = \tilde{\sigma}^2_{\beta} + \hat{\beta}^2(m+1)^2/(2n-m-1)^2$. The method for a non-Gaussian random error model is similar. Our simulations show that this method works well for most general cases. However, this bandwidth is too large for a rough mean function because the bias is large. Note that our test statistics (10) and (11) heavily depend on the estimator $\hat{\sigma}^2$. Thus we consider the asymptotic optimal bandwidth $m_{\text{opt}} = \left(28n\sigma^4/\text{var}\left(\epsilon^2\right)\right)^{1/2}$ and $m_{\text{opt}} = (14n)^{1/2}$ for normally distributed random errors, which are given in Tong and Wang (2005). However, as emphasized in Tong and Wang (2005), this bandwidth is still too large for small n or rough g. The adjusted bandwidth $m = \lceil (1 + \lambda - (\lambda^2 + 2\lambda)^{1/2})(14n)^{1/2} \rceil$ is proposed so that the percentage of increase in the higher order terms of MSE of $\hat{\sigma}^2$ using this bandwidth comparing to that of the optimal bandwidth is no more than $100\lambda\%$. Our simulation studies in Sect. 3 indicate that the choice of λ with a small value, say $\lambda = 0.2$ for a small sample or $\lambda = 0$ for a large sample, is enough to make the tests work very well. It is worth mentioning that when the mean function is rough, we need to carefully choose the bandwidth to make the DBT method work, while other methods may perform even worse. See the simulation results in Table 2 in Sect. 3.

Finally, as the fourth moment of the random error $\kappa = E(\epsilon^4)$ in the test statistic (11) is unknown, we also propose a permutation test which does not require an estimate of κ. We use $\hat{\beta}$ given in (5) as the test statistic and approximate its null distribution using permutations based on the fact that the x labels are exchangeable under the null hypothesis. For each permutation of x labels, we compute the estimate of β. Repeating this process q times, we derive estimates of β denoted as $\hat{\beta}^*_1, \hat{\beta}^*_2, ..., \hat{\beta}^*_q$. We use the empirical distribution of $\hat{\beta}^*$'s as the approximated null distribution and compute the p-value as $\sum_{i=1}^{q} \text{I}(\hat{\beta} < \hat{\beta}^*_i)/q$. We reject the null hypothesis (4) if the p-value is less than $\alpha = 0.05$. We refer to this method as the permutation-based DBT.

3 Simulation studies

In this section, we conduct simulations to evaluate the performance of the proposed DBTs and also compare them with some existing methods. To generate data from the model (1), we consider a factorial design with two choices of g, $g_1(x) = 1 + 5c(x^2 - x)$ and $g_2(x) = 1 + c\sin(4\pi x)$, and three choices of sample sizes, n =30, 50 and 100. For each function, we consider five choices of c =0, 0.2, 0.5, 0.7, and 1.

We first consider Gaussian random errors where $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$ with $\sigma = 0.3$ and $\sigma = 0.5$ for $g_1(x)$ and $g_2(x)$, respectively. For the bandwidth selection, we chose the

m that minimize $\widetilde{\mathrm{MSE}}(\hat{\beta}) = \tilde{\sigma}_\beta^2 + \hat{\beta}^2(m+1)^2/(2n-m-1)^2$ for $g_1(x)$. For the rough mean function $g_2(x)$, we let $m = \lceil(1+\lambda-(\lambda^2+2\lambda)^{1/2})(14n)^{1/2}\rceil$ where $\lambda = (0.2, 0.05, 0)$ for $n = (30, 50, 100)$. We further estimate σ^2 and κ using the estimators proposed by Tong and Wang (2005) and Evans and Jones (2008) as mentioned in Sect. 2.

We calculate the proportions of rejections by counting the number of rejections in 1000 simulations at the significance level $\alpha = 0.05$. For ease of presentation, we denote DBT-Gau, DBT-Gen, and DBT-Perm as the difference-based test T in (10), G in (11), and the permutation method, respectively. For comparison with the residual-based tests, we also consider the locally most powerful (LMP) test by Cox et al (1988), the permutation test using the generalized F-test (F-Perm) by Raz (1990), the specification test with the second order ordinary difference sequence (Spec ord) and with the second order optimal difference sequence (Spec opt) by Yatchew (2003), and the Kolmogorov-Smirnov type statistic (TKS) by Van Keilegom et al (2008). As far as we know, there are no other tests for no effect that can dominate these traditional methods.

Table 1 Proportions of rejection with the mean function $g_1(x)$ and Gaussian random errors with $\sigma = 0.3$

Sample size	Method	$c = 0$	$c = 0.2$	$c = 0.5$	$c = 0.7$	$c = 1$
$n = 30$	F-Perm	0.050	0.089	0.417	0.741	0.937
	LMP	0.046	0.048	0.184	0.364	0.752
	Spec ord	0.046	0.054	0.171	0.428	0.787
	Spec opt	0.050	0.116	0.509	0.848	0.997
	TKS	0.100	0.107	0.346	0.536	0.848
	DBT-Perm	0.053	0.151	0.707	0.949	0.999
	DBT-Gau	0.039	0.153	0.723	0.951	0.999
	DBT-Gen	0.038	0.169	0.737	0.958	1
$n = 50$	F-Perm	0.040	0.110	0.744	0.957	0.984
	LMP	0.050	0.074	0.463	0.862	1
	Spec ord	0.045	0.058	0.298	0.661	0.971
	Spec opt	0.049	0.136	0.732	0.982	1
	TKS	0.068	0.096	0.487	0.798	0.981
	DBT-Perm	0.039	0.265	0.897	0.993	1
	DBT-Gau	0.044	0.221	0.903	0.996	1
	DBT-Gen	0.050	0.234	0.918	0.997	1
$n = 100$	F-Perm	0.049	0.248	0.974	0.992	0.998
	LMP	0.046	0.139	0.963	1	1
	Spec ord	0.048	0.066	0.547	0.931	1
	Spec opt	0.043	0.160	0.937	1	1
	TKS	0.024	0.112	0.778	0.979	1
	DBT-Perm	0.040	0.412	0.986	1	1
	DBT-Gau	0.030	0.388	0.993	1	1
	DBT-Gen	0.033	0.404	0.995	1	1

Table 2 Proportions of rejection with the mean function $g_2(x)$ and Gaussian random errors with $\sigma = 0.5$

Sample size	Method	$c = 0$	$c = 0.2$	$c = 0.5$	$c = 0.7$	$c = 1$
$n = 30$	F-Perm	0.051	0.068	0.200	0.469	0.858
	LMP	0.050	0.072	0.182	0.253	0.369
	Spec ord	0.043	0.055	0.247	0.565	0.927
	Spec opt	0.051	0.114	0.564	0.894	0.999
	TKS	0.097	0.105	0.192	0.273	0.505
	DBT-Perm	0.055	0.129	0.584	0.864	0.994
	DBT-Gau	0.045	0.131	0.600	0.884	0.997
	DBT-Gen	0.047	0.137	0.631	0.902	1
$n = 50$	F-Perm	0.058	0.083	0.539	0.913	0.988
	LMP	0.049	0.096	0.356	0.586	0.872
	Spec ord	0.053	0.068	0.427	0.825	0.993
	Spec opt	0.049	0.153	0.822	0.995	1.000
	TKS	0.062	0.109	0.271	0.558	0.836
	DBT-Perm	0.042	0.153	0.72	0.952	0.998
	DBT-Gau	0.045	0.163	0.739	0.956	1
	DBT-Gen	0.049	0.171	0.767	0.969	1
$n = 100$	F-Perm	0.052	0.181	0.961	0.997	0.999
	LMP	0.059	0.183	0.797	0.989	1
	Spec ord	0.050	0.079	0.722	0.990	1
	Spec opt	0.045	0.202	0.983	1	1
	TKS	0.029	0.092	0.605	0.928	1
	DBT-Perm	0.051	0.36	0.994	1	1
	DBT-Gau	0.051	0.388	0.996	1	1
	DBT-Gen	0.048	0.402	0.998	1	1

The simulation results are given in Table 1 for $g_1(x)$ with $\sigma = 0.3$ and in Table 2 for $g_2(x)$ with $\sigma = 0.5$. It is evident that DBT-Gau and DBT-Gen outperform the other tests in most cases for both mean functions. The superiority of DBTs is more profound when the sample size is small. The power of DBTs increases much faster than the other methods as c increases, especially for highly oscillating functions. In addition, we note that DBT-Gau and DBT-Gen are able to control the type I error rates, while some existing methods have type I error rates that exceed the nominal level. Finally, when the normality assumption holds in our simulations, DBT-Gau has the smallest type I error and performs very well. DBT-Gen has the greatest power in most cases. We have also conducted simulations with other mean functions (not shown to save space), and the comparison results remain the same.

For simulations with non-Gaussian random errors, we generate data from the model (1) with mean function $g_1(x)$ and random errors $\epsilon_i = \tau_i/10\sqrt{3}$, where τ_i follow a t distribution with 3 degrees of freedom. The simulation results are given in Table 3. With the non-Gaussian random errors, DBT-Gau and the specification

Table 3 Proportions of rejection with non-Gaussian random errors

Sample size	Method	$c = 0$	$c = 0.2$	$c = 0.5$	$c = 0.7$	$c = 1$
$n = 30$	F-Perm	0.046	0.091	0.560	0.809	0.945
	LMP	0.048	0.043	0.210	0.495	0.835
	Spec ord	0.036	0.065	0.296	0.537	0.865
	Spec opt	0.076	0.142	0.696	0.904	0.967
	TKS	0.046	0.078	0.392	0.687	0.926
	DBT-Perm	0.059	0.209	0.797	0.934	0.981
	DBT-Gau	0.059	0.238	0.787	0.941	0.986
	DBT-Gen	0.040	0.204	0.755	0.919	0.979
$n = 50$	F-Perm	0.055	0.158	0.799	0.935	0.986
	LMP	0.041	0.096	0.593	0.899	0.983
	Spec ord	0.061	0.067	0.427	0.772	0.945
	Spec opt	0.068	0.147	0.785	0.955	0.991
	TKS	0.029	0.099	0.617	0.903	0.987
	DBT-Perm	0.044	0.292	0.873	0.965	0.995
	DBT-Gau	0.062	0.309	0.905	0.983	0.992
	DBT-Gen	0.032	0.231	0.848	0.960	0.979
$n = 100$	F-Perm	0.034	0.348	0.973	0.990	0.999
	LMP	0.044	0.153	0.943	0.994	0.997
	Spec ord	0.064	0.081	0.629	0.919	0.984
	Spec opt	0.049	0.214	0.911	0.986	0.999
	TKS	0.007	0.100	0.902	0.985	0.997
	DBT-Perm	0.040	0.369	0.918	0.978	0.996
	DBT-Gau	0.099	0.408	0.966	0.992	0.999
	DBT-Gen	0.034	0.294	0.918	0.974	0.991

test with optimal sequence do not work well all the time as their type I error rates are inflated. In contrast, DBT-Gen and DBT-Perm are able to control the type I error rates even when the normality assumption is violated while having larger power than the other tests, in particular when the sample size is small.

To conclude, we recommend the DBT-Gen test in (11) for practical use or the DBT-Gau test in (10) when there is strong evidence showing that the random errors are normally distributed.

4 Real data examples

4.1 The adult human gut microbiota and aging

Human gut microbiota is important for modulating host metabolism. Recently, some researchers studied the relationship between age and gut microbial differences (Zhang et al 2021). We consider a dataset consisting of gut microbial characteristics by metagenomic sequencing from 1741 Han Chinese adults aged 26-76. For the

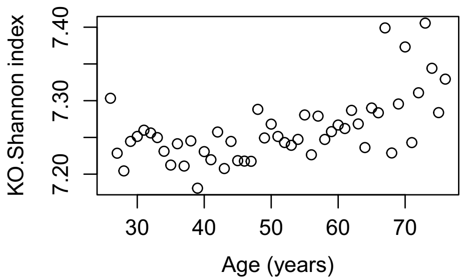

Fig. 1 Age and gut microbial alpha diversity at the KO level

richness and diversity analyses, the Alpha diversity quantified by the Shannon index was calculated on the relative abundance profiles at gene, species, and KEGG (Kyoto Encyclopedia of Genes and Genomes) orthology (KO). Figure 1 shows the average Shannon index at each age.

Consider model (1) with y_i the average Shannon index and x_i age from 26 to 76. Without assuming any parametric form of y_i and x_i, we apply our proposed three tests and get the following results. DBT-Perm has test statistic $\hat{\beta} = 5.85 \times 10^{-3}$ with a p-value 0.01, DBT-Gau has test statistic $T = 5.10$ with a p-value 1.7×10^{-7}, and DBT-Gen has test statistic $G = 3.65$ with a p-value 1.3×10^{-4}. Under the $\alpha = 0.05$ level of significance, we reject the null hypothesis and conclude that there is evidence that age and gut microbiota diversity are related.

4.2 The COVID-19 incubation period and aging

COVID-19, also called SARS-CoV-2, is an infectious disease caused by a newly discovered coronavirus. It has been creating a severe pandemic and panic around the world. One of the most important epidemiological features of COVID-19 is the incubation period, which is important for building up the disease control policies (Lauer et al 2020). Some researchers have investigated the relationship between age and the incubation period of COVID-19. Tan et al (2020) studied the dataset of all confirmed cases admitted to restructured hospitals in Singapore collected from 23 January 2020 to 2 April 2020, and they concluded that elders (age $\geq$ 70 years old) have significantly longer incubation period than those younger people. However, they did not study the relationship for COVID-19 patients under 70 years old. In this

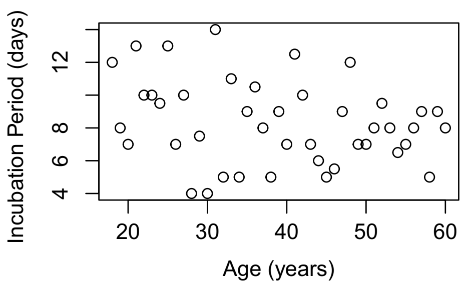

Fig. 2 Age and the incubation period

example, we want to investigate whether the COVID-19 incubation period depends on age for patients aged 18 to 68. The dataset contains 225 documented cases of infection between 1 January 2020 and 16 January 2020 in China (Liu et al 2020). Let y_i be the median incubation period and x_i be the age from 18 to 68. Figure 2 shows the scatter plot of y_i and x_i. Applying the difference-based test, we have the following test results. DBT-Perm has test statistic $\hat{\beta} = -0.793$ with a p-value 0.58, DBT-Gau has test statistic $T = -0.199$ with a p-value 0.579, and DBT-Gen has test statistic $G = -0.235$ with a p-value 0.593. Under the $\alpha = 0.05$ significance level, we fail to reject the null hypothesis and conclude that the relation between the age and the incubation period is not statistically significant for patients under 70 years old.

5 Extension and discussion

We proposed a novel difference-based method to test the hypothesis of no relationship between the dependent and independent variables. The difference-based tests are easy to implement since they do not require an estimate of the mean function or its first derivative. We further derived the null distributions of the new tests by normal approximation or by permutation and showed that they can detect local alternatives that converge to the null at a rate close to $n^{-1/2}$. Simulation results also demonstrated that our new tests compare favorably to existing methods, especially when the sample size is small.

For simplicity, the current paper has focused on testing no effect in nonparametric regression. We note that the method is general and readily extendable to test other hypotheses and settings. We now discuss some future research topics.

5.1 Goodness-of-fit test for polynomial regression

The proposed method can be extended to test the hypothesis that the mean function is a polynomial (Cox et al 1988; Cox and Koh 1989; Chen 1994; Liu and Wang 2004; Eubank et al 2005; Wang 2011a). Specifically, we formulate the null and alternative hypotheses as follows:

$$\begin{aligned} & H_0 \,:\, g(x) = a_0 + a_1 x + \cdots + a_{r-1} x^{r-1}, \\ \text{versus} \quad & H_1 \,:\, g(x) \text{ is not a } (r-1)\text{th or lower order polynomial function,} \end{aligned} \tag{13}$$

where $r \geq 2$. In the special case when $r = 1$, hypothesis (13) reduces to hypothesis (2). To apply the difference-based test, we define the lag-k squared differences of reduced data as

$$s_{rk} = \frac{1}{\binom{2r}{r}(n - rk)} \sum_{i=1}^{n-rk} \left(z_{i+k}^r - z_i^r\right)^2, \quad k = 1, \ldots, m,$$

where $z_i^r = z_{i+k}^{r-1} - z_i^{r-1}$, $z_i^1 = y_i$ and $m = o(n)$ with $m < n/r$. Suppose that the first r derivatives $g'(x), \ldots, g^{(r)}(x)$ are bounded. Then,

$$E(s_{rk}) = \sigma^2 + \beta_r d_{rk} + o(d_{rk}),$$

where $d_{rk} = (d_k)^r = (k/n)^{2r}$ and $\beta_r = \int_0^1 (g^{(r)})^2 dx / \binom{2r}{r}$. This shows that the hypotheses in (13) can be converted to the new hypotheses as

$$H_0 : \beta_r = 0 \quad \text{versus} \quad H_1 : \beta_r > 0. \tag{14}$$

Moreover, following similar arguments as in Sect. 2, we fit the linear regression model

$$s_{rk} = \sigma^2 + \beta_r d_{rk} + \eta_{rk}$$

and derive the WLS estimator of the slope as

$$\hat{\beta}_r = \frac{\sum_{k=1}^{m} w_{rk} s_{rk} \left(d_{rk} - \bar{d}_{rw} \right)}{\sum_{k=1}^{m} w_{rk} \left(d_{rk} - \bar{d}_{rw} \right)^2}, \tag{15}$$

where $\bar{d}_{rw} = \sum_{k=1}^{m} w_{rk} d_{rk}$ and $w_{rk} = (n - rk)/N_r$ with $N_r = nm - rm(m+1)/2$. Finally, we can use $\hat{\beta}_r$ to construct a test statistic and then approximate its null distribution by permutation.

5.2 Test the parallelity of two mean functions

Consider the following nonparametric regression model,

$$y_{ki} = g_k(x_i) + \epsilon_{ki}, \quad k = 1, 2; \; i = 1, \dots n, \tag{16}$$

where g_1 and g_2 are two unknown mean functions, and ϵ_{1i} and ϵ_{2i} are independent random errors with mean zero and constant variances σ_1^2 and σ_2^2, respectively. We are interested in the hypothesis that the two mean functions differ by a constant,

$$H_0 : g_1(x) = g_2(x) + c \quad \text{versus} \quad H_1 : g_1(x) \neq g_2(x) + c, \tag{17}$$

where c is a constant. Consider k as a factor with two levels. The above hypothesis means no interaction exists between x and k under the smoothing spline ANOVA decomposition (Wang 2011b).

Let $\tilde{y}_i = y_{1i} - y_{2i}$, $\tilde{g}(x_i) = g_1(x_i) - g_2(x_i)$, and $\tilde{\epsilon}_i = \epsilon_{1i} - \epsilon_{2i}$. Then $\tilde{y}_i$ follows model (1) with mean function $\tilde{g}$, and hypothesis (17) reduces to hypothesis (2). Consequently, the proposed difference-based method can be applied directly.

5.3 DBT with unequally spaced design

We now provide a brief overview of how to adapt the proposed method for situations involving unequally spaced designs. Assume that we have a sequence of ordered design points $x_1 < \cdots < x_n$ such that for each i we have some $k = 1, \dots, m_i$, satisfying $\Omega = \{(i,k) : x_{i+k} - x_i < L, i + m_i \le n\}$ with $L = o(1)$. Then by letting $z_{ik} = (y_{i+k} - y_i)^2/2$, we have

$$E[z_{ik}] = \sigma^2 + \frac{1}{2}(x_{i+k} - x_i)^2(g'(x_i))^2,$$

(see supplement S4 in Dai et al (2017)). This suggests that we can fit the linear model

$$z_{ik} = \sigma^2 + d_{ik}^2\beta_i + \tilde{\epsilon}_{ik}, \quad (i,k) \in \Omega,$$

where $d_{ik} = (x_{i+k} - x_i)$ and $\beta_i = (g'(x_i))^2/2$ are constant for each i. We further derive

$$\hat{\beta}_i = \frac{\sum_{k=1}^{m_i} w_k z_{ik}(d_{ik} - \bar{d}_{iw})}{\sum_{k=1}^{m_i} w_k (d_{ik} - \bar{d}_{iw})^2}$$

as the WLS estimator of β_i, where $\bar{d}_{iw} = \sum_{k=1}^{m_i} w_k d_{ik}$ is the weighted average of d_{ik}. Finally, by taking the average or the Riemann sum of $\hat{\beta}_i$, we have a test statistic for hypothesis (2) and can apply the permutation method to generate the null distribution. Further research is required to derive the asymptotic null distributions and the statistical properties of these new test statistics, which is outside the scope of this paper.

Appendix 1: Some lemmas and their proofs

Lemma 1 *Assume that $m \to \infty$ and $m = o(n)$. We have*

(a) $\sum_{k=1}^{m} h_k = \frac{15}{16}n + o(n)$;
(b) $\sum_{k=1}^{m} k^2 h_k = n^2 m + o(n^2 m)$;
(c) $\sum_{k=1}^{i-1} h_k = \frac{15n^2}{4m^4}(i^3 - m^2 i) + O\left(\frac{n^2}{m^2}\right) + o\left(\frac{n^2 i}{m^2}\right)$;
(d) $\sum_{k=i}^{m} k h_k = O(n^2)$;
(e) $\sum_{k=1}^{i-1} k^2 h_k = O(\frac{n^2 i^3}{m^2})$;
(f) $\sum_{k=1}^{m} h_k^2 = \frac{45n^4}{4m^3} + o(\frac{n^4}{m^3})$;
(g) $\sum_{k=1}^{m} k h_k^2 = \frac{225n^4}{32m^2} + o(\frac{n^4}{m^2})$.

Proof Following the Appendix in Tong and Wang (2005), we have

$$\sum_{k=1}^{m}(d_k - \bar{d}_w) = \frac{m^4}{12n^3} + o\left(\frac{m^4}{n^3}\right), \tag{A1}$$

$$\sum_{k=1}^{m} w_k(d_k - \bar{d}_w)^2 = \frac{4m^4}{45n^4} + o\left(\frac{m^4}{n^4}\right), \tag{A2}$$

$$\bar{d}_w = \frac{m^2}{3n^2} + o\Big(\frac{m^2}{n^2}\Big). \tag{A3}$$

(a) By (A1) and (A2), we have

$$\sum_{k=1}^{m} h_k = \frac{\sum_{k=1}^{m}(d_k - \bar{d}_w)}{\sum_{k=1}^{m} w_k(d_k - \bar{d}_w)^2} = \frac{\frac{m^4}{12n^3} + o(\frac{m^4}{n^3})}{\frac{4m^4}{45n^4} + o(\frac{m^4}{n^4})} = \frac{15n}{16} + o(n).$$

(b) By (A2) and (A3), we have

$$\sum_{k=1}^{m} k^2 h_k = \frac{\sum_{k=1}^{m} k^2(d_k - \bar{d}_w)}{\sum_{k=1}^{m} w_k(d_k - \bar{d}_w)^2} = \frac{\frac{4m^5}{45n^2} + o\big(\frac{m^5}{n^2}\big)}{\frac{4m^4}{45n^4} + o(\frac{m^4}{n^4})} = n^2 m + o(n^2 m),$$

where

$$\begin{aligned}\sum_{k=1}^{m} k^2(d_k - \bar{d}_w) &= \frac{1}{n^2}\Big(\frac{m^5}{5} + O(m^4)\Big) - \Big(\frac{m^3}{3} + O(m^2)\Big)\Big[\frac{m^2}{3n^2} + o\Big(\frac{m^2}{n^2}\Big)\Big] \\ &= \frac{4m^5}{45n^2} + o\Big(\frac{m^5}{n^2}\Big).\end{aligned}$$

(c) For $1 \leq i \leq m$, by (A2) and (A3) we have

$$\begin{aligned}\sum_{k=1}^{i-1} h_k &= \frac{\sum_{k=1}^{i-1}(d_k - \bar{d}_w)}{\sum_{k=1}^{m} w_k(d_k - \bar{d}_w)^2} \\ &= \frac{\frac{1}{3n^2}(i^3 - m^2 i) + O\big(\frac{m^2}{n^2}\big) + o\big(\frac{m^2 i}{n^2}\big)}{\frac{4m^4}{45n^4} + o(\frac{m^4}{n^4})} \\ &= \frac{15n^2}{4m^4}(i^3 - m^2 i) + O\Big(\frac{n^2}{m^2}\Big) + o\Big(\frac{n^2 i}{m^2}\Big),\end{aligned}$$

where

$$\sum_{k=1}^{i-1}(d_k - \bar{d}_w) = \sum_{k=1}^{i-1}(\frac{k}{n})^2 - (i-1)\bar{d}_w = \frac{1}{3n^2}(i^3 - m^2 i) + O\Big(\frac{m^2}{n^2}\Big) + o\Big(\frac{m^2 i}{n^2}\Big).$$

(d) For $1 \leq i \leq m$, by (A2) we have

$$\sum_{k=i}^{m} k h_k = \frac{\sum_{k=i}^{m} k(d_k - \bar{d}_w)}{\sum_{k=1}^{m} w_k(d_k - \bar{d}_w)^2} = \frac{\sum_{k=i}^{m} \frac{k^3}{n^2} - \bar{d}_w \sum_{k=i}^{m} k}{\sum_{k=1}^{m} w_k(d_k - \bar{d}_w)^2} = O(n^2).$$

(e) For $1 \leq i \leq m$, by (A2) we have

$$\sum_{k=1}^{i-1} k^2 h_k = \frac{\sum_{k=1}^{i-1} k^2(d_k - \bar{d}_w)}{\sum_{k=1}^{m} w_k(d_k - \bar{d}_w)^2}$$
$$= \frac{\frac{1}{n^2}(\frac{i^5}{5} + O(i^4)) - [\frac{m^2}{3n^2} + o(\frac{m^2}{n^2})][\frac{i^3}{3} + O(i^2)]}{\frac{4m^4}{45n^4} + o(\frac{m^4}{n^4})}$$
$$= O\left(\frac{n^2 i^3}{m^2}\right).$$

(f) By (A2), we have

$$\sum_{k=1}^{m} h_k^2 = \frac{\sum_{k=1}^{m}(d_k - \bar{d}_w)^2}{(\sum_{k=1}^{m} w_k(d_k - \bar{d}_w)^2)^2}$$
$$= \frac{\sum_{k=1}^{m} d_k^2 - 2\bar{d}_w \sum_{k=1}^{m} d_k + m(\bar{d}_w)^2}{(\sum_{k=1}^{m} w_k(d_k - \bar{d}_w)^2)^2}$$
$$= \frac{\frac{m^5}{n^4}(\frac{1}{5} - \frac{2}{9} + \frac{1}{9}) + o(\frac{m^5}{n^4})}{[\frac{4m^4}{45n^4} + o(\frac{m^4}{n^4})]^2}$$
$$= \frac{45n^4}{4m^3} + o\left(\frac{n^4}{m^3}\right).$$

(g) By (A2), we have

$$\sum_{k=1}^{m} kh_k^2 = \frac{\sum_{k=1}^{m} k(d_k - \bar{d}_w)^2}{(\sum_{k=1}^{m} w_k(d_k - \bar{d}_w)^2)^2}$$
$$= \frac{\sum_{k=1}^{m} kd_k^2 - 2\bar{d}_w \sum_{k=1}^{m} kd_k + (\bar{d}_w)^2 \sum_{k=1}^{m} k}{(\sum_{k=1}^{m} w_k(d_k - \bar{d}_w)^2)^2}$$
$$= \frac{\frac{m^6}{n^4}(\frac{1}{6} - \frac{1}{6} + \frac{1}{18}) + o(\frac{m^6}{n^4})}{[\frac{4m^4}{45n^4} + o(\frac{m^4}{n^4})]^2}$$
$$= \frac{225n^4}{32m^2} + o\left(\frac{n^4}{m^2}\right).$$

□

Lemma 2 *Assume that $m \to \infty$ and $m = o(n)$, and let $\mathbf{g} = (g(x_1), \ldots, g(x_n))^T$. We have*

(a) $\mathbf{g}^T \boldsymbol{B} \mathbf{g} = 2\beta mn + O(m^2)$;
(b) $\mathbf{g}^T \boldsymbol{B}^2 \mathbf{g} = O(n^2 m)$.

Proof

(a) Let $\boldsymbol{A} = (a_{ij})_{n\times n}$ be a symmetric matrix with a_{ij} having the same form as b_{ij} in (7) but $h_0 = 0$ and $h_k = 1$ for $k = 1, \dots, m$. Let $\boldsymbol{D} = (d_{ij})_{n\times n}$ is the matrix defined in Theorem 1 of Tong and Wang (2005). Then,

$$\mathbf{g}^T\boldsymbol{B}\mathbf{g} = \frac{\mathbf{g}^T(\boldsymbol{A}-\boldsymbol{D})\mathbf{g}}{\bar{d}_w}.$$

To simplify the notation, we let $g_i = g(x_i)$. We can show that

$$\begin{aligned}
\mathbf{g}^T\boldsymbol{A}\mathbf{g} &= \sum_{k=1}^{m}\sum_{i=1}^{n-k}(g_{i+k}-g_i)^2 \\
&= \sum_{k=1}^{m}\sum_{i=1}^{n-k}\left[\frac{k^2}{n^2}(g_i')^2 + O\Big(\frac{k^3}{n^3}\Big)\right] \\
&= \sum_{k=1}^{m}\frac{k^2}{n^2}\sum_{i=1}^{n-k}(g_i')^2 + \sum_{k=1}^{m}O\Big(\frac{(n-k)k^3}{n^3}\Big) \\
&= \sum_{k=1}^{m}\frac{k^2}{n}\left[\frac{1}{n}\sum_{i=1}^{n}(g_i')^2 - \frac{1}{n}\sum_{i=n-k+1}^{n}(g_i')^2\right] + O\Big(\frac{m^4}{n^2}\Big) \\
&= \sum_{k=1}^{m}\frac{k^2}{n}\left[2\beta + O\Big(\frac{k}{n}\Big)\right] + O\Big(\frac{m^4}{n^2}\Big) \\
&= \frac{2\beta m^3}{3n} + O\Big(\frac{m^4}{n^2}\Big),
\end{aligned}$$

where $\beta = \int_0^1 (g'(x))^2\,dx/2$. Note also that $\mathbf{g}^T\boldsymbol{D}\mathbf{g} = O(m^4/n^2)$ by Lemma 2 in Tong et al (2013). Then by (A3), we have

$$\mathbf{g}^T\boldsymbol{B}\mathbf{g} = \frac{\frac{2\beta m^3}{3n} + O(\frac{m^4}{n^2})}{\frac{m^2}{3n^2} + o(\frac{m^2}{n^2})} = 2\beta mn + O(m^2).$$

(b) Noting that $\boldsymbol{B}$ is a symmetric matrix, we let $\mathbf{g}^T\boldsymbol{B}^2\mathbf{g} = (\boldsymbol{B}\mathbf{g})^T(\boldsymbol{B}\mathbf{g}) = \boldsymbol{q}^T\boldsymbol{q}$, where $\boldsymbol{q} = \boldsymbol{B}\mathbf{g} = (q_1, \dots, q_n)^T$. For $i \in [1, m]$, by parts (b), (d) and (e) of Lemma 1, we have

$$
\begin{aligned}
q_i &= \sum_{k=1}^{i-1} h_k(g_i - g_{i-k}) - \sum_{k=1}^{m} h_k(g_{i+k} - g_i) \\
&= \sum_{k=1}^{i-1} h_k\Big(\frac{k}{n}g_i' - \frac{k^2}{2n^2}g_i'' + o(\frac{k^2}{n^2})\Big) - \sum_{k=1}^{m} h_k\Big(\frac{k}{n}g_i' + \frac{k^2}{2n^2}g_i'' + o(\frac{k^2}{n^2})\Big) \\
&= -\frac{g_i'}{n}\sum_{k=i}^{m} kh_k - \Big[\frac{g_i''}{2n^2}(\sum_{k=1}^{i-1} k^2h_k + \sum_{k=1}^{m} k^2h_k)\Big] + o\Big(\frac{1}{n}\sum_{k=i}^{m} k^2h_k\Big) \\
&= O(n) + O(m) + O\Big(\frac{i^3}{m^2}\Big) + o(m) \\
&= O(n).
\end{aligned}
$$

Similary, we can show that $q_i = O(n)$ for $i \in [n-m+1, n]$. While for $i \in [m+1, n-m]$, by Lemma 1(b) we have

$$
\begin{aligned}
q_i &= \sum_{k=1}^{m} h_k(g_i - g_{i-k}) - \sum_{k=1}^{m} h_k(g_{i+k} - g_i) \\
&= \sum_{k=1}^{m} h_k\Big(\frac{k}{n}g_i' - \frac{k^2}{2n^2}g_i'' + o\Big(\frac{k^2}{n^2}\Big)\Big) - \sum_{k=1}^{m} h_k\Big(\frac{k}{n}g_i' + \frac{k^2}{2n^2}g_i'' + o\Big(\frac{k^2}{n^2}\Big)\Big) \\
&= -\frac{1}{n^2}g_i''\sum_{k=1}^{m} k^2h_k + o\Big(\frac{\sum_{k=1}^{m} k^2h_k}{n^2}\Big) \\
&= O(m).
\end{aligned}
$$

Taken together the above results, it yields that

$$
\mathbf{g}^T\boldsymbol{B}^2\mathbf{g} = \sum_{i=1}^{m} q_i^2 + \sum_{i=m+1}^{n-m} q_i^2 + \sum_{i=n-m+1}^{n} q_i^2 = O(n^2m).
$$

□

Lemma 3 *Assume that $m \to \infty$ and $m = o(n)$. We have*

(a) $\sum_{i=1}^{n} b_{ii}^2 = \frac{15n^4}{14m} + o(\frac{n^4}{m})$;
(b) $\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} b_{ij}^2 = \frac{45n^5}{2m^3} + o(\frac{n^5}{m^3})$.

Proof

(a) By parts (a) and (c) of Lemma 1, we have

$$
\begin{aligned}
\sum_{i=1}^{n} b_{ii}^2 &= 2\sum_{i=1}^{m}\Big(\sum_{k=1}^{m} h_k + \sum_{k=1}^{i-1} h_k\Big)^2 + \sum_{i=m+1}^{n-m}\Big(2\sum_{k=1}^{m} h_k\Big)^2 \\
&= 2\sum_{i=1}^{m}\Big[\frac{15}{16}n + o(n) + \frac{15n^2}{4m^4}(i^3 - m^2 i) + O\Big(\frac{n^2}{m^2}\Big) + o\Big(\frac{n^2 i}{m^2}\Big)\Big]^2 \\
&\quad + 4(n-2m)\Big[\frac{15}{16}n + o(n)\Big]^2 \\
&= 2\Big(\frac{15}{16}\Big)^2 n^2 m + \Big(\frac{15n^2}{4m^4}\Big)^2\Big(\sum_{i=1}^{m} i^6 + m^4\sum_{i=1}^{m} i^2 - 2m^2\sum_{i=1}^{m} i^4\Big) \\
&\quad + \frac{15^2 n^3}{32m^4}\Big(\sum_{i=1}^{m} i^3 - m^2\sum_{i=1}^{m} i\Big) + o\Big[\frac{n^4}{m^6}\Big(\sum_{i=1}^{m} i^4 - m^2\sum_{i=1}^{m} i^2\Big)\Big] \\
&\quad + 4\Big[\Big(\frac{15}{16}\Big)^2 n^3 + o(n^3)\Big] \\
&= \frac{15n^4}{14m} + o\Big(\frac{n^4}{m}\Big).
\end{aligned}
$$

(b) By parts (f) and (g) of Lemma 1, we have

$$
\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} b_{ij}^2 &= 2\sum_{k=1}^{m}(n-k)h_k^2 \\
&= 2n\Big[\frac{45n^4}{4m^3} + o\Big(\frac{n^4}{m^3}\Big)\Big] - 2\Big[\frac{225n^4}{32m^2} + o\Big(\frac{n^4}{m^2}\Big)\Big] \\
&= \frac{45n^5}{2m^3} + o\Big(\frac{n^5}{m^3}\Big).
\end{aligned}
$$

□

Appendix 2: Proof of Theorem 1

Proof Let $\mathbf{g} = (g(x_1), \ldots, g(x_n))^T$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$. By (1) and (6), we have

$$
\hat{\beta} = \frac{\mathbf{g}^T \boldsymbol{B}\mathbf{g}}{2N} + \frac{\mathbf{g}^T \boldsymbol{B}\boldsymbol{\epsilon}}{N} + \frac{\boldsymbol{\epsilon}^T \boldsymbol{B}\boldsymbol{\epsilon}}{2N}.
$$

From Lemma 2(a) we have

$$
\frac{\mathbf{g}^T \boldsymbol{B}\mathbf{g}}{2N} = \frac{2\beta mn + O(m^2)}{2N} = \beta + O\Big(\frac{m}{n}\Big).
$$

Using Lemma 2(b), we have $E(\mathbf{g}^T \boldsymbol{B}\boldsymbol{\epsilon}/N)^2 = \sigma^2 \mathbf{g}^T \boldsymbol{B}^2 \mathbf{g}/N^2 = O(1/m)$. This leads to

$$
\frac{\mathbf{g}^T \boldsymbol{B}\boldsymbol{\epsilon}}{N} = O_p\Big(\frac{1}{\sqrt{m}}\Big).
$$

Let $\boldsymbol{\epsilon}^T\boldsymbol{B}\boldsymbol{\epsilon}/(2N) = \boldsymbol{\epsilon}^T\boldsymbol{C}\boldsymbol{\epsilon} - \boldsymbol{\epsilon}^T\boldsymbol{U}\boldsymbol{\epsilon}$, where the elements of matrix $\boldsymbol{C}$ are

$$c_{ij} = \begin{cases} \sum_{k=1}^{m} h_k/N, & 1 \le i = j \le n, \\ -h_{|i-j|}/(2N), & 0 < |i-j| \le m, \\ 0, & \text{otherwise}, \end{cases}$$

and $\boldsymbol{U} = \text{diag}(u_1, \cdots, u_n)$ with $u_i = \sum_{k=min\{i,n+1-i,m+1\}}^{m+1} h_k/(2N)$, for $i = 1, \dots, n$ and $h_{m+1} = 0$. Let $c_0 = \sum_{k=1}^{m} h_k/N$, $c_{i-j} = c_{j-i} = -h_{|i-j|}/(2N)$ for $1 \le |i-j| \le m$, and $c_{i-j} = c_{j-i} = 0$ for $|i-j| > m$. Then $\boldsymbol{\epsilon}^T\boldsymbol{C}\boldsymbol{\epsilon} = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{i-j}\epsilon_i\epsilon_j$, where ϵ_i are i.i.d. with mean zero. Thus by parts (a) and (f) of Lemma 1,

$$\sum_{-\infty}^{\infty} c_k^2 = \frac{(\sum_{k=1}^{m} h_k)^2}{N^2} + 2\sum_{k=1}^{m}\frac{h_k^2}{4N^2} = \frac{O(n^2)}{O(n^2m^2)} + \frac{O(n^4/m^3)}{O(n^2m^2)} = O\left(\frac{1}{m^2}\right) + O\left(\frac{n^2}{m^5}\right) < \infty,$$

as $m = \lceil n^r \rceil$ with $2/5 \le r < 1$. Assuming $E(\epsilon^6) < \infty$, by Theorem 2 in Whittle (1962), $\boldsymbol{\epsilon}^T\boldsymbol{C}\boldsymbol{\epsilon}$ is asymptotically normally distributed.

We have $\boldsymbol{\epsilon}^T\boldsymbol{U}\boldsymbol{\epsilon} = \sum_{i=1}^{n} u_i\epsilon_i^2$. Let $X_i = u_i\epsilon_i^2$, then $X_1, X_2, \dots, X_n$ are independent random variables, where $X_i = \sum_{k=i}^{m} h_k\epsilon_i^2/(2N)$ for $1 \le i \le m$, $X_i = \sum_{k=n-i+1}^{m} h_k\epsilon_i^2/(2N)$ for $n-m+1 \le i \le n$, and $X_i = 0$ for $m+1 \le i \le n-m$. For $1 \le i \le m$, using parts (a) and (c) of Lemma 1 we have

$$\begin{aligned} E[X_i] &= \frac{\sigma^2}{2N}\sum_{k=i}^{m} h_k = \frac{\sigma^2}{2N}\left(\sum_{k=1}^{m} h_k - \sum_{k=1}^{i-1} h_k\right) \\ &= \frac{15\sigma^2}{8}\left(\frac{1}{4m} - \frac{ni^3}{m^5} + \frac{ni}{m^3}\right) + O\left(\frac{n}{m^3}\right) + o\left(\frac{1}{m}\right) + o\left(\frac{ni}{m^3}\right) < \infty, \end{aligned}$$

as $m = \lceil n^r \rceil$ with $1/2 < r < 1$. For $n-m+1 \le i \le n$, the results are similar. It is intuitive to show that for $1 \le i \le m$, the variance of X_i is

$$\begin{aligned} \text{Var}(X_i) &= E(X_i^2) - E(X_i)^2 = \left(\frac{\sum_{k=i}^{m} h_k}{2N}\right)^2 [E(\epsilon_i^4) - \sigma^4] \\ &= (\gamma_4 - 1)\sigma^4\left[\frac{15}{8}\left(\frac{1}{4m} - \frac{ni^3}{m^5} + \frac{ni}{m^3}\right) + O\left(\frac{n}{m^3}\right) + o\left(\frac{1}{m}\right) + o\left(\frac{ni}{m^3}\right)\right]^2 < \infty, \end{aligned}$$

as $n \to \infty$ and $m = \lceil n^r \rceil$ with $1/2 < r < 1$. We have similar results for $n-m+1 \le i \le n$, and $\text{Var}(X_i) = 0$ for $m+1 \le i \le n-m$. Noting also that $\sum_{i=1}^{m} \text{Var}(X_i) = \sum_{i=n-m+1}^{n} \text{Var}(X_i)$, we can derive the sum of variance as

$$
\begin{aligned}
s_n^2 &= \sum_{i=1}^{n} \mathrm{Var}(X_i) = 2\sum_{i=1}^{m} \mathrm{Var}(X_i) \\
&= 2(\gamma_4 - 1)\sigma^4 \sum_{i=1}^{m} \left[\frac{15}{8}\Big(\frac{1}{4m} - \frac{ni^3}{m^5} + \frac{ni}{m^3} \Big) + O\Big(\frac{n}{m^3}\Big) + o\Big(\frac{1}{m}\Big) + o\Big(\frac{ni}{m^3}\Big) \right]^2 \\
&= 2(\gamma_4 - 1)\sigma^4 \Big[\frac{225}{64} \sum_{i=1}^{m} \Big(\frac{1}{16m^2} + \frac{n^2 i^6}{m^{10}} + \frac{n^2 i^2}{m^6} - \frac{ni^3}{2m^6} + \frac{ni}{2m^4} - \frac{2n^2 i^4}{m^8} \Big) + o\Big(\frac{n^2}{m^3}\Big) \Big] \\
&= 2(\gamma_4 - 1)\sigma^4 \cdot \frac{225}{64} \Big(\frac{1}{7} + \frac{1}{3} - \frac{2}{5} \Big) \frac{n^2}{m^3} + o\Big(\frac{n^2}{m^3}\Big) \\
&= \frac{15}{28}(\gamma_4 - 1)\sigma^4 \frac{n^2}{m^3} + o\Big(\frac{n^2}{m^3}\Big).
\end{aligned}
$$

Thus s_n^2 is finite as $m = \lceil n^r \rceil$ with $2/3 \le r < 1$. Moreover, we have

$$
\begin{aligned}
\sum_{i=1}^{n} E\big[|X_i - \mu_i|^3\big] &= 2\sum_{i=1}^{m} E\Big[\Big| \frac{\sum_{k=i}^{m} h_k}{2N} (\epsilon_i^2 - \sigma^2) \Big|^3 \Big] \\
&= 2\sum_{i=1}^{m} \Big(\frac{\sum_{k=i}^{m} h_k}{2N} \Big)^3 E\big[|\epsilon_i^2 - \sigma^2|^3\big] \\
&= \tau_0 \sum_{i=1}^{m} \Big[O\Big(\frac{1}{m}\Big) + O\Big(\frac{ni^3}{m^5}\Big) + O\Big(\frac{ni}{m^3}\Big) \Big]^3 \\
&= O\Big(\frac{1}{m^2}\Big) + O\Big(\frac{n^3}{m^5}\Big) = O\Big(\frac{n^3}{m^5}\Big),
\end{aligned}
$$

and

$$
s_n^3 = \tau_1 \frac{n^3}{m^{9/2}} + o\Big(\frac{n^3}{m^{9/2}}\Big) = O\Big(\frac{n^3}{m^{9/2}}\Big),
$$

where τ_0 and τ_1 are some constants and $m \to \infty$ with $n \to \infty$. Thus

$$
\lim_{n\to\infty} \frac{\sum_{i=1}^{n} E[|X_i - \mu_i|^3]}{s_n^3} = \lim_{n\to\infty} O\Big(\frac{1}{\sqrt{m}}\Big) = 0.
$$

By the Lyapunov CLT, $\boldsymbol{\epsilon}^T \boldsymbol{U} \boldsymbol{\epsilon}$ is asymptotically normally distributed. Therefore, $\boldsymbol{\epsilon}^T \boldsymbol{B} \boldsymbol{\epsilon}/(2N)$ is asymptotically normally distributed. The mean of $\boldsymbol{\epsilon}^T \boldsymbol{B} \boldsymbol{\epsilon}/(2N)$ can be shown to be

$$
E\Big[\frac{\boldsymbol{\epsilon}^T \boldsymbol{B} \boldsymbol{\epsilon}}{2N} \Big] = \frac{1}{2N} E\Big[\sum_{i=1}^{n} \sum_{j=1}^{n} b_{ij} \epsilon_i \epsilon_j \Big] = \frac{\sigma^2}{2N} \mathrm{tr}(\boldsymbol{B}) = 0,
$$

and the variance is

$$
\mathrm{Var}\Big(\frac{\boldsymbol{\epsilon}^T \boldsymbol{B} \boldsymbol{\epsilon}}{2N} \Big) = E\Big[\Big(\frac{\boldsymbol{\epsilon}^T \boldsymbol{B} \boldsymbol{\epsilon}}{2N} \Big)^2 \Big] = \frac{1}{4N^2} \Big[\sum_{i=1}^{n} b_{ii}^2 (E(\epsilon_i^4) - \sigma^4) + 2\sum_{i=1}^{n} \sum_{j=1, j\ne i}^{n} b_{ij}^2 \sigma^4 \Big].
$$

Using parts (a) and (b) of Lemma 3 and combining the above results, we have

$$\begin{aligned}\mathrm{Var}\left(\frac{\boldsymbol{\epsilon}^T\boldsymbol{B}\boldsymbol{\epsilon}}{2N}\right) &= \frac{1}{4N^2}\left[\frac{15n^4}{14m}(E(\epsilon_i^4)-\sigma^4)+\frac{45n^5}{m^3}\sigma^4\right]+o\left(\frac{n^2}{m^3}\right)\\ &= \frac{15}{56}(\gamma_4-1)\sigma^4\frac{n^2}{m^3}+o\left(\frac{n^2}{m^3}\right),\end{aligned}$$

where $m=\lceil n^r\rceil$ with $2/3<r<1$. This then leads to

$$\sqrt{n^{3r-2}}(\hat{\beta}-\beta)\xrightarrow{D}N(0,\sigma_b^2),$$

as $n\to\infty$, where $\sigma_b=\sqrt{15(\gamma_4-1)\sigma^4/56}$. □

Appendix 3: Proofs of Theorem 2 and Theorem 3

Proof of Theorem 2 The estimated error variance of $\hat{\beta}$ given in (9) can be written as $\tilde{\sigma}_\beta^2=\tau_n\hat{\sigma}^4$. As $n\to\infty$, $\tau_n\to(15/28)n^{2-3r}$ with $m=\lceil n^r\rceil$ in (9). Let $\hat{\sigma}^2$ be a consistent estimator of σ^2, and $\sigma_\beta^2=(15/28)n^{2-3r}\sigma^4$. Under Theorem 1 and the null hypothesis H_0 in (4), we have $\hat{\beta}/\sigma_\beta\xrightarrow{D}N(0,1)$ when the random errors are normally distributed. In addition, we have $\sigma_\beta/\tilde{\sigma}_\beta\to 1$ as $n\to\infty$. Thus by Slutsky's theorem,

$$T=\frac{\hat{\beta}}{\tilde{\sigma}_\beta}=\frac{\hat{\beta}}{\sigma_\beta}\cdot\frac{\sigma_\beta}{\tilde{\sigma}_\beta}\xrightarrow{D}N(0,1)\quad\text{as } n\to\infty.$$

□

Proof of Theorem 3 Given that $\hat{\kappa}$ and $\hat{\sigma}^2$ are consistent estimators of κ and σ^2 respectively, we note that $\check{\sigma}_{\beta g}^2$ in (12) is also a consistent estimator of $\sigma_\beta^2=(15/56)n^{2-3r}(\kappa-(\sigma^2)^2)$. Therefore under Theorem 1 and the null hypothesis H_0 in (4), by Slutsky's theorem we have

$$G=\frac{\hat{\beta}}{\check{\sigma}_{\beta g}}=\frac{\hat{\beta}}{\sigma_\beta}\cdot\frac{\sigma_\beta}{\check{\sigma}_{\beta g}}\xrightarrow{D}N(0,1)\quad\text{as } n\to\infty.$$

□

Acknowledgements Zhijian Li was supported in part by the Guangdong Provincial Key Laboratory of Interdisciplinary Research and Application for Data Science, BNU-HKBU United International College, project code 2022B1212010006 and UIC research grant R0400001-22, and UIC Start-up Research Fund/

Grant UICR0700024-22. Tiejun Tong was supported in part by the General Research Fund of Hong Kong (HKBU12300123, HKBU12303421) and the National Natural Science Foundation of China (12071305). The authors thank the editor, the associate editor and two reviewers for their constructive comments and suggestions that have led to a substantial improvement in the paper and thank the authors of Zhang et al (2021) and Liu et al (2020) for providing the real data sets.

References

Barry D, Hartigan JA (1990) An omnibus test for departures from constant mean. Ann Stat 18:1340–1357
Bliznyuk N, Carroll R, Genton M et al (2012) Variogram estimation in the presence of trend. Stat Interface 5(2):159–168
Brabanter KD, Brabanter JD, Moor BD et al (2013) Derivative estimation with local polynomial fitting. J Mach Learn Res 14:281–301
Chen JC (1994) Testing for no effect in nonparametric regression via spline smoothing techniques. Ann Inst Stat Math 46:251–265
Cox D, Koh E (1989) A smoothing spline based test of model adequacy in polynomial regression. Ann Inst Stat Math 41:383–400
Cox D, Koh E, Wahba G et al (1988) Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models. Ann Stat 16:113–119
Cui Y, Levine M, Zhou Z (2021) Estimation and inference of time-varying auto-covariance under complex trend: a difference-based approach. Electr J Stat 15(2):4264–4294
Dai W, Tong T, Genton M (2016) Optimal estimation of derivatives in nonparametric regression. J Mach Learn Res 17:1–25
Dai W, Tong T, Zhu L (2017) On the choice of difference sequence in a unified framework for variance estimation in nonparametric regression. Stat Sci 32:455–468
Einmahl JH, Van Keilegom I (2008) Tests for independence in nonparametric regression. Stat Sin 18:601–615
Eubank RL (2000) Testing for no effect by cosine series methods. Scand J Stat 27:747–763
Eubank RL, Li CS, Wang S (2005) Testing lack-of-fit of parametric regression models using nonparametric regression techniques. Stat Sin 15:135–152
Evans D, Jones AJ (2008) Nonparametric estimation of residual moments and covariance. Proc Royal Soc A 464:2831–2846
Gasser T, Sroka L, Jennen-Steinmetz C (1986) Residual variance and residual pattern in nonlinear regression. Biometrika 73:625–633
González-Manteiga W, Crujeiras RM (2013) An updated review of goodness-of-fit tests for regression models. TEST 22:361–411
Hall P, Kay JW, Titterington DM (1990) Asymptotically optimal difference-based estimation of variance in nonparametric regression. Biometrika 77:521–528
Lauer SA, Grantz KH, Bi Q et al (2020) The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. Ann Int Med 172:577–582
Li CS (2012) Testing for no effect via splines. Computat Stat 27:343–357
Liu A, Wang Y (2004) Hypothesis testing in smoothing spline models. J Statist Computat Simul 74:581–597
Liu X, He Y, Ma X et al (2020) Statistical data analysis on the incubation and suspected period of COVID-19 based on 2172 confirmed cases outside Hubei province. Acta Math Appl Sin 43:278–294
McManus DA (1991) Who invented local power analysis? Econom Theory 7:265–268
Neumeyer N, Dette H (2003) Nonparametric comparison of regression curves: an empirical process approach. Ann Stat 31:880–920
Raz J (1990) Testing for no effect when estimating a smooth function by nonparametric regression: a randomization approach. J Am Stat Assoc 85:132–138
Rice J (1984) Bandwidth choice for nonparametric regression. Ann Stat 12:1215–1230
Storey JD, Xiao W, Leek JT et al (2005) Significance analysis of time course microarray experiments. Proc Natl Acad Sci 102(36):12837–12842
Tan WYT, Wong LY, Leo YS et al (2020) Does incubation period of COVID-19 vary with age? A study of epidemiologically linked cases in Singapore. Epidemiol Infection 148:e197

Tong T, Wang Y (2005) Estimating residual variance in nonparametric regression using least squares. Biometrika 92:821–830

Tong T, Ma Y, Wang Y (2013) Optimal variance estimation without estimating the mean function. Bernoulli 19:1839–1854

Van Keilegom I, González Manteiga W, Sánchez Sellero C (2008) Goodness-of-fit tests in parametric regression based on the estimation of the error distribution. TEST 17:401–415

Wang W, Lin L (2015) Derivative estimation based on difference sequence via locally weighted least squares regression. J Mach Learn Res 16:2617–2641

Wang W, Yu P, Lin L et al (2019) Robust estimation of derivatives using locally weighted least absolute deviation regression. J Mach Learn Res 20:1–49

Wang Y (2011) Smoothing splines: methods and applications. Chapman and Hall, New York, pp 12–45

Wang Y (2011b) Smoothing splines: methods and applications. CRC Press

Whittle P (1962) On the convergence to normality of quadratic forms in independent variables. Theory Probab Appl 9:103–108

Yatchew A (1999) An elementary nonparametric differencing test of equality of regression functions. Econom Lett 62:271–278

Yatchew A (2003) Semiparametric regression for the applied econometrician. Cambridge University Press, Cambridge, pp 10–25

Zhang M, Dai W (2023) On difference-based gradient estimation in nonparametric regression. Sci J Stat Anal Data Mining. https://doi.org/10.1002/sam.11644

Zhang X, Zhong H, Li Y et al (2021) Sex- and age-related trajectories of the adult human gut microbiota shared across populations of different ethnicities. Nature Aging 1:87–100

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.