Knowledge-Based Systems 236 (2022) 107781

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Debiased learning and forecasting of first derivative

WenWu Wang [a,*], Jun Lu [b,**], Tiejun Tong [c], Zhonghua Liu [d]

[a] School of Statistics, Qufu Normal University, Shandong, China
[b] College of Arts and Science, National University of Defense Technology, Hunan, China
[c] Department of Mathematics, Hong Kong Baptist University, Hong Kong
[d] Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 11 October 2020
Received in revised form 2 November 2021
Accepted 9 November 2021
Available online 30 November 2021

Keywords:
Boundary problem
Differenced estimation
Equally spaced design
Kernel learning
Nonparametric derivative estimation

ABSTRACT

In the era of big data, there are many data sets recorded in equal intervals of time. To model the change rate of such data, one often constructs a nonparametric regression model and then estimates the first derivative of the mean function. Along this direction, we propose a symmetric two-sided local constant regression for interior points, an asymmetric two-sided local polynomial regression for boundary points, and a one-sided local linear forecasting model for outside points. Specifically, under the framework of locally weighted least squares regression, we derive the asymptotic bias and variance of the proposed estimators, as well as establish their asymptotic normality. Moreover, to reduce the estimation bias for highly-oscillatory functions, we propose debiased estimators based on high-order polynomials and derive their corresponding kernel functions. A data-driven two-step procedure for simultaneous selection of the model and tuning parameters is also proposed. Finally, the usefulness of our proposed estimators is demonstrated by simulation studies and two real data examples.

© 2021 Elsevier B.V. All rights reserved.

1. Introduction

Consider the nonparametric regression model

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

where Y_i's are the response variables, $m(\cdot)$ is an unknown mean function, x_i's are equally spaced design points with $x_i = i/n$, and ϵ_i's are independent and identically distributed random errors with zero mean and variance σ^2. In this paper, we are interested in estimating the first derivative of the mean function, denoted by $m^{(1)}(\cdot)$.

An accurate estimate of the first derivative is often desired and has a wide range of applications in statistics and related areas. It can be applied, for example, to the change point problems for detecting the cellular morphology changes [1], the bump hunting [2], and the trend in time series [3]. Also in the field of pattern recognition, the derivative estimation can be used for time series classification [4], texture classification [5], and word spotting [6]. While for more applications of the first derivative, they include, but not limited to, the following areas: cell biology [7], computer vision [8], medicine [9], machine learning [10], and effect evaluation [11].

In nonparametric regression, the derivative estimate is often a by-product of the nonparametric estimate for the mean function. Taking the local polynomial regression (LPR) [12] as an example, we first fit a local polynomial of degree p as

$$(\hat\beta_0, \dots, \hat\beta_p) = \arg\min_{\beta_j\in\mathbb{R}} \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j (x_i - x_0)^j \right\}^2 K_h(x_i - x_0), \tag{2}$$

where $K_h(\cdot)$ is a kernel function with the bandwidth h. The first derivative of the mean function at a fixed point x_0 is then estimated by $\hat m^{(1)}(x_0) = \hat\beta_1$. In LPR, one is to minimize the mean square errors (MSE) of $\{\hat m(x_i)\}_{i=1}^n$ for selecting the optimal h. For the first derivative estimation, however, a better criterion is to minimize the MSE of $\hat m^{(1)}(x_i)$ as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} \{\hat m^{(1)}(x_i) - m^{(1)}(x_i)\}^2. \tag{3}$$

To be more specific, minimizing Eq. (2) does not necessarily guarantee that the first derivative of $m(\cdot)$ will also be optimally estimated [13], mainly because the convergence rates of the mean function estimator and the first derivative estimator can be largely different. As a consequence, if we follow LPR, the first derivative estimator may not achieve the same optimal rate of convergence as that for the mean function estimator [14].

* Corresponding author at: School of Statistics, Qufu Normal University, Shandong, China.
** Corresponding author at: College of Arts and Science, National University of Defense Technology, Hunan, China.
E-mail addresses: wengewsh@sina.com (W. Wang), jlustat@gmail.com (J. Lu).

https://doi.org/10.1016/j.knosys.2021.107781
0950-7051/© 2021 Elsevier B.V. All rights reserved.

To obtain the optimal convergence rate for the first derivative estimator, Charnigo et al. [13] and De Brabanter et al. [15] employed a variance-minimizing linear combination of symmetric quotients called the empirical first derivative as

$$Y_i^{<1>} = \sum_{l=1}^{k} w_l \left(\frac{Y_{i+l} - Y_{i-l}}{x_{i+l} - x_{i-l}} \right), \quad (4)$$

where $w_l = 6l^2/\{k(k+1)(2k+1)\}$ for $l = 1, \ldots, k$. They further showed that their proposed estimators provide a better estimation efficiency compared to the single difference quotient as in Härdle [16]. Wang and Lin [14] proposed a locally weighted least squares regression and obtained accurate expressions for the main term of bias and variance of their proposed derivative estimators. Dai et al. [17] proposed a general differenced estimation with a fixed bias-reduction level. More recently, Liu and De Brabanter [18] extended the results from the equally spaced design to random designs, and Wang et al. [10] proposed a robust estimation for the derivatives with heavy-tailed errors.

Apart from the theoretical developments, it remains unknown which of the above estimators will provide the best performance in practical applications. To the best of our knowledge, there are few studies in the literature on comparing the finite-sample performance of the optimal estimators with different bias-reduction levels. Moreover, we also note that the derivative estimation for boundary points is less reliable compared to that for interior points. To address those issues, we propose the locally weighted polynomial regression to estimate the first derivative for both interior and boundary points. Meanwhile, for the purpose of forecasting, we propose a one-sided forecasting model for outside points and investigate its statistical properties for the derivative estimation.

The rest of this paper is organized as follows. In Section 2, we propose two-sided high-order polynomial estimators for interior points, and derive the theoretical results including the asymptotic bias, asymptotic variance, and asymptotic normality. We also show that the least squares estimator is equivalent to the variance-minimizing estimator. Under different bias-reduction levels, a data-driven procedure for simultaneously selecting the model and tuning parameters is also proposed. In Section 3, we propose one-sided estimators and asymmetric two-sided estimators for boundary points, which explains the reasons why the two-sided estimators are more efficient than the one-sided estimators. In Section 4, we further propose one-sided local linear estimator for outside points, which can be used for forecasting the variation tendency. In Section 5, we propose a data-driven criterion for choosing the optimal estimator in the sense of minimizing the mean absolute error, and conduct simulation studies to demonstrate its usefulness. Lastly, we apply our new estimators to two real data examples in Section 6, and conclude the paper in Section 7 with some discussions. Technical results and supporting materials are provided in the Appendices.

2. Derivative learning for interior points

In this section, we propose the local polynomial estimators for the first derivative based on the least squares regression, and investigate their theoretical properties. We have three interesting results. First, the local constant estimator based on symmetric difference sequence is equal to the first empirical derivative. Second, the least squares estimator is equivalent to the variance-minimizing estimator. Third, some high-order kernels can be derived according to the local polynomial estimators.

2.1. Local constant estimator

Assume that the mean function is one time continuously differentiable on $[0, 1]$. For any positive integer $k = o(n)$, by Charnigo et al. [13] we define the symmetric difference quotients as

$$Y_{il}^{(1)} = \frac{Y_{i+l} - Y_{i-l}}{x_{i+l} - x_{i-l}} = \frac{m(x_{i+l}) - m(x_{i-l})}{x_{i+l} - x_{i-l}} + \frac{\epsilon_{i+l} - \epsilon_{i-l}}{x_{i+l} - x_{i-l}}, \quad 1 \le l \le k. \quad (5)$$

Then by the first-order Taylor expansion of the mean function at point x_i, we have

$$\frac{m(x_{i+l}) - m(x_{i-l})}{x_{i+l} - x_{i-l}} = m^{(1)}(x_i) + o(1), \quad 1 \le l \le k.$$

Plugging this formula back to (5), the symmetric difference quotients is expressed as

$$Y_{il}^{(1)} = \beta_{i1} + \delta_{il},$$

where $\beta_{i1} = m^{(1)}(x_i)$ and $\delta_{il} = n(\epsilon_{i+l} - \epsilon_{i-l})/(2l) + o(1)$. Note that δ_{il} are independent random errors with mean nearly zero and variance $n^2\sigma^2/(2l^2)$. We then apply the local constant regression to estimate the first derivative as

$$\hat{\beta}_{i1} = \arg\min_{\beta_{i1}} \sum_{l=1}^{k} w_{il}(Y_{il}^{(1)} - \beta_{i1})^2 \quad = (D_1^T W D_1)^{-1} D_1^T W Y_i^{(1)}, \quad (6)$$

where $w_{il} = \sigma^2/(2\mathrm{Var}[\delta_{il}]) = l^2/n^2$, $D_1 = (1, \ldots, 1)_{1\times k}^T$ with T the transpose of a matrix or a vector, $Y_i^{(1)} = (Y_{i1}^{(1)}, \ldots, Y_{ik}^{(1)})^T$, and $W = n^{-2}\mathrm{diag}(1^2, \ldots, k^2)$.

In the following theorem, we derive the asymptotic bias and variance for the first derivative estimator (6). The proof is given in Appendix A.

Theorem 1. *Assume that model (1) holds with the design points equally spaced, and the mean function is one time continuously differentiable on* $[0, 1]$. *Then the variance of estimator (6) is*

$$\mathrm{Var}[\hat{m}^{(1)}(x_i)] \doteq \frac{3\sigma^2}{2}\frac{n^2}{k^3}, \quad k+1 \le i \le n-k,$$

where $\doteq$ means that the higher-order terms are omitted. Further, if the mean function is three times continuously differentiable on $[0, 1]$, *then the bias of estimator (6) is*

$$\mathrm{Bias}[\hat{m}^{(1)}(x_i)] \doteq \frac{m^{(3)}(x_i)}{10}\frac{k^2}{n^2}, \quad k+1 \le i \le n-k.$$

Following Theorem 1, the mean square error (MSE) of the first derivative estimator is given as

$$\mathrm{MSE}[\hat{m}^{(1)}(x_i)] = \frac{3\sigma^2}{2}\frac{n^2}{k^3} + \frac{(m^{(3)}(x_i))^2}{100}\frac{k^4}{n^4}. \quad (7)$$

To minimize the MSE with respect to k, we take the first derivative of Eq. (7) and yield the gradient as

$$\frac{\mathrm{d}}{\mathrm{d}k}\mathrm{MSE}[\hat{m}^{(1)}(x_i)] = \frac{(m^{(3)}(x_i))^2}{25}\frac{k^3}{n^4} - \frac{9\sigma^2}{2}\frac{n^2}{k^4}.$$

Then by letting the first derivative as zero, the optimal k that minimizes the MSE is

$$k_{opt} = \left(\frac{225\sigma^2}{2(m^{(3)}(x_i))^2} \right)^{1/7} n^{6/7}.$$

And consequently, the minimum MSE with the optimal k is

$$\mathrm{MSE}_{k_{opt}}[\hat{m}^{(1)}(x_i)] \approx 0.35 \left(\sigma^8 (m^{(3)}(x_i))^6\right)^{1/7} n^{-4/7}.$$

We also note that the variance and bias for the first derivative estimator (6) require different assumptions on the smoothness of the mean function. In addition, the first derivative estimator (6) is the same as the first empirical derivative (4) in Charnigo et al. [13] and De Brabanter et al. [15]. This shows that the newly proposed estimator is optimal according to the variance-minimizing criterion. Moreover, our method is also very convenient to deduce more accurate bias and variance terms by the least squares theory. Lastly, the asymptotic normality for the first derivative estimator (6) is also established in the following theorem.

Theorem 2 (*Asymptotic Normality*). *Under the assumptions of Theorem 1, if $k = n^{\alpha}$ with $2/3 < \alpha < 1$ as $n \to \infty$, then*

$$\frac{k^{3/2}}{n}\left(\hat{m}^{(1)}(x_i) - m^{(1)}(x_i) - \frac{m^{(3)}(x_i)}{10}\frac{k^2}{n^2}\right) \xrightarrow{d} N\left(0, \frac{3\sigma^2}{2}\right)$$

for $k+1 \leq i \leq n-k$, where $\xrightarrow{d}$ denotes the convergence in distribution. Further, if $k = n^{\alpha}$ with $2/3 < \alpha < 6/7$ as $n \to \infty$, then

$$\frac{k^{3/2}}{n}\left(\hat{m}^{(1)}(x_i) - m^{(1)}(x_i)\right) \xrightarrow{d} N\left(0, \frac{3\sigma^2}{2}\right).$$

By the least squares theory and the fact that $\{\delta_{il}\}_{l=1}^{k}$ are independently distributed with asymptotical mean zeros and variance $\{n^2\sigma^2/(2l^2)\}_{l=1}^{k}$, it can be readily shown that Theorem 2 holds. With a suitable k, the first derivative estimator (6) is asymptotically normal and asymptotically unbiased. The error variance σ^2 can be estimated by a bias-corrected method [19], denoted by $\hat{\sigma}^2$. This consequently yields an approximate $1-\alpha$ confidence interval for $m^{(1)}(x_i)$ as

$$\left(\hat{m}^{(1)}(x_i) - z_{\alpha/2}\sqrt{\frac{3n^2\hat{\sigma}^2}{2k^3}},\ \hat{m}^{(1)}(x_i) + z_{\alpha/2}\sqrt{\frac{3n^2\hat{\sigma}^2}{2k^3}}\right),$$

where z_{α} is the upper αth percentile of the standard normal distribution.

2.2. Local polynomial estimators

To further reduce the bias, we assume that the mean function is p times continuously differentiable on [0, 1], where p is an odd number. By the pth-order Taylor expansion of the mean function at point x_i, we have

$$Y_{il}^{(1)} = \beta_{i1} + \beta_{i3}\frac{l^2}{n^2} + \cdots + \beta_{ip}\frac{l^{p-1}}{n^{p-1}} + \delta_{il}, \quad 1 \leq l \leq k,$$

where $\beta_{iq} = m^{(q)}(x_i)/q!$ for $q = 1, 3, \ldots, p$, and $\delta_{il} = n(\epsilon_{i+l} - \epsilon_{i-l})/(2l) + o(l^{p-1}/n^{p-1})$. Noting also that δ_{il} are independent random errors with a nearly zero mean and variance $n^2\sigma^2/(2l^2)$, we can estimate the coefficients as

$$\begin{aligned}\hat{\beta}_i &= (\hat{\beta}_{i1}, \ldots, \hat{\beta}_{ip})^T \\ &= \arg\min_{\beta_{i1},\ldots,\beta_{ip}} \sum_{l=1}^{k} w_{il}\left(Y_{il}^{(1)} - \beta_{i1} - \cdots - \beta_{ip}\frac{l^{p-1}}{n^{p-1}}\right)^2 \\ &= (D_p^T W D_p)^{-1} D_p^T W Y_i^{(1)},\end{aligned} \tag{8}$$

where $w_{il} = l^2/n^2$, W and $Y_i^{(1)}$ are the same as in Section 2.1, and

$$D_p = \begin{pmatrix} 1 & 1^2 n^{-2} & \cdots & 1^{p-1} n^{-(p-1)} \\ 1 & 2^2 n^{-2} & \cdots & 2^{p-1} n^{-(p-1)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & k^2 n^{-2} & \cdots & k^{p-1} n^{-(p-1)} \end{pmatrix}.$$

By (8), we then define the first derivative estimator as $\hat{m}^{(1)}(x_i) = \hat{\beta}_{i1}$.

Table 1
Asymptotic variances and biases of the first derivative estimators for $p = 1, 3, 5$.

p	Bias	Var	k_{opt}	MSE	$k_p/k_{1,opt}$
1	$\frac{m^{(3)}(x_i)}{10}\frac{k^2}{n^2}$	$\frac{3\sigma^2}{2}\frac{n^2}{k^3}$	$O(n^{6/7})$	$O(n^{-4/7})$	1.00
3	$-\frac{m^{(5)}(x_i)}{504}\frac{k^4}{n^4}$	$\frac{75\sigma^2}{8}\frac{n^2}{k^3}$	$O(n^{10/11})$	$O(n^{-8/11})$	1.84
5	$\frac{m^{(7)}(x_i)}{61776}\frac{k^6}{n^6}$	$\frac{3675\sigma^2}{128}\frac{n^2}{k^3}$	$O(n^{14/15})$	$O(n^{-12/15})$	2.67

In the following theorem, we derive the asymptotic bias and variance for the first derivative estimator, with the proof in Appendix B.

Theorem 3. *Assume that model (1) holds with the design points equally spaced, and the mean function is $(p+2)$ times continuously differentiable on [0, 1] with p an odd number. Then for any $k+1 \leq i \leq n-k$, the variance and bias of the first derivative estimator in (8) are*

$$\begin{aligned}&\mathrm{Var}[\hat{m}^{(1)}(x_i)] = \frac{\sigma^2}{2} e_{1,p}^T (D_p^T W D_p)^{-1} e_{1,p}, \\ &\mathrm{Bias}[\hat{m}^{(1)}(x_i)] = e_{1,p}^T (D_p^T W D_p)^{-1} D_p^T W \Delta_{p,i},\end{aligned}$$

where $e_{1,p} = (1, 0, \ldots, 0)^T_{((p+1)/2)\times 1}$ and $\Delta_{p,i} = \frac{m^{(p+2)}(x_i)}{(p+2)!n^{p+1}}(1^{p+1}, \ldots, k^{p+1})^T$.

For comparison, we also list their variances and biases for $p = 1, 3, 5$ in Table 1. It is evident that there is a relatively large increase rate in the variance. For instance, when p increases from 1 to 3, the variance will increase up to 625%. For large σ^2, we suggest to use the lower order p expect that the bias is larger. Empirically, the optimal k increases correspondingly as p increases. We set the optimal $k_{1,opt}$ for $p = 1$, and list the $k_p/k_{1,opt}$ to attain the same variance without considering the bias. It is easy to obtain that the ratios are 1.84 and 2.67 for $p = 3$ and 5, respectively. To trade off between the bias and variance, we will propose a criterion in Section 5 to choose p through simulation study.

2.3. Selection of tuning parameters

For the three estimators with $p = 1, 3, 5$ in Sections 2.1 and 2.2, two questions still remain for practical use. The first question is to choose the tuning parameter k_p for a given p, and the second question is to select the optimal estimator among the three estimators.

For the first question, we can choose k_p globally for interior points by the averaged MSE (AMSE) criterion

$$\mathrm{AMSE}(p, k_p) = \frac{1}{n-2k}\sum_{i=k+1}^{n-k} \mathrm{MSE}[\hat{m}^{(1)}(x_i)], \tag{9}$$

which was used and discussed by Wang and Lin [14]. Specifically for $p = 1$, by (7) we have

$$\begin{aligned}\mathrm{AMSE}(1, k_1) &= \frac{1}{n-2k}\sum_{i=k+1}^{n-k} \mathrm{MSE}[\hat{m}^{(1)}(x_i)] \\ &= \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\left[\frac{3\sigma^2}{2}\frac{n^2}{k^3} + \frac{(m^{(3)}(x_i))^2}{100}\frac{k^4}{n^4}\right].\end{aligned}$$

Note that the parameters σ^2 and $\{m^{(3)}(x_i) | 1 \leq i \leq n\}$ are unknown. We estimate the error variance σ^2 by Hall et al. [20] as

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n-2}(0.809Y_i - 0.5Y_{i+1} - 0.309Y_{i+2})^2,$$

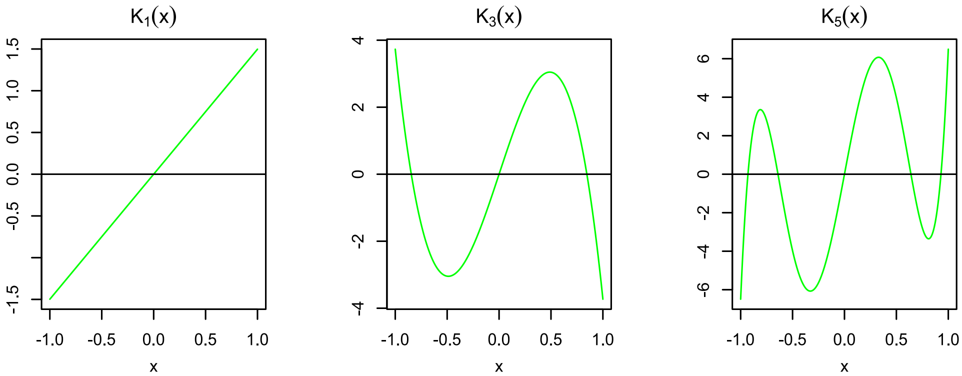

Fig. 1. Kernel functions for the first derivative estimators for $p = 1, 3, 5$.

and estimate the third-order derivatives $\{m^{(3)}(x_i)|1 \le i \le n\}$ by the local polynomial regression of order $p = 3$, which can also be implemented by the R package *locpol* [21]. By minimizing the AMSE, it yields the first derivative estimates $\{\hat{m}^{(1)}_{k_{1,opt}}(x_i)|k+1 \le i \le n-2k\}$ with optimal tuning parameter

$$k_{1,opt} = \arg\min_{k_1 \le n/2} \text{AMSE}(1, k_1).$$

Accordingly, the first derivative estimates for $p = 3$ or 5 can be obtained with the optimal tuning parameters $k_{3,opt}$ and $k_{5,opt}$, respectively. To improve the estimation efficiency of the error variance σ^2, the bias-reduced estimates are also available in the literature [19,22]. The high-order derivatives $\{m^{(p+2)}(x_i)|k+1 \le i \le n-2k\}$ can be estimated by the local $(p+2)$-order or $(p+4)$-order polynomial regression.

For the three estimators, we can select the polynomial order

$$p_{opt} = \arg\min_{p=1,3,5} \text{AMSE}(p, k_{p,opt}),$$

and obtain the optimal first derivative estimates $\{\hat{m}^{(1)}_{p_{opt},opt}(x_i)|k+1 \le i \le n-2k\}$ for interior points. In fact, the above selection procedure can also be generalized to all the design points by adjusting the domain of AMSE. For more details, see Section 5.2.

In real data analysis, when the independent variable is $x_i = i$ rather than $x_i = i/n$, we cannot use the nonparametric method directly. For this scenario, to increase the scope of application, we suggest to transform the independent and dependent variables as $\tilde{x}_i = x_i/\tilde{n} = i/\tilde{n}$ and $\tilde{Y}_i = Y_i/\tilde{n}$, where $\tilde{n}$ is a standardized parameter, e.g., $\tilde{n} = 100, 200, 500$. Moreover in Appendix C, we show that the first derivative estimation remains the same before or after the data transformation for the same k, and the AMSE also remains unchanged as $\tilde{n}$ changes.

2.4. Two-sided kernel learning

In the least squares regression framework, the first derivative estimator is a linear combination of the responses as

$$\hat{m}^{(1)}(x) = \sum_{i=1}^{n} K_p(x_i)Y_i,$$

where $K_p(\cdot)$ is a kernel function with polynomial order p. Also for the uniqueness of the assigned weights, we assume that the kernel function is defined on the support $[-1, 1]$ and satisfies the condition $\int_{-1}^{1} K_p(x)x dx = 1$. Then by Wang and Yu [22], we have the two-sided (TS) kernel functions as follows:

$$K_1(x) = \frac{3}{2}x,$$

$$K_3(x) = \frac{15}{8}(5x - 7x^3),$$

$$K_5(x) = \frac{105}{512}(140x - 504x^3 + 396x^5).$$

Fig. 1 displays the corresponding kernels for $p = 1, 3$ and 5, respectively. It is evident that, as p increases, the kernel function will be more oscillatory. Finally, we refer to the corresponding estimator in (8) as the TS estimator.

3. Derivative learning for boundary points

In this section, we consider the first derivative estimation in the boundary, where an important example can be seen in the regression kink design [23]. For a fixed k, we denote the left and right boundary regions as $\{x_i, 1 \le i \le k\}$ and $\{x_i, n-k+1 \le i \le n\}$, respectively. Then depending on the range of observations, we propose two new estimators: one is the one-sided (OS) local polynomial regression and another is the asymmetric two-sided (ATS) local polynomial regression. Theoretical results show that the ATS estimator is always better than the OS estimator under mild conditions.

3.1. One-sided local polynomial estimators

Assume that the mean function is $(p+1)$ times continuously differentiable on $[0, 1]$. For the estimation at point x_i with $1 \le i \le k$, we apply the one-sided observations $\{(x_j, Y_j), i \le j \le i+k\}$ to express model (1) as a local polynomial regression:

$$Y_{i+j} = \beta_{i0} + \beta_{i1}d_j + \cdots + \beta_{ip}d_j^p + \delta_{ij}, \quad 1 \le j \le k,$$

where $\beta_{iq} = m^{(q)}(x_i)/q!$ for $q = 0, 1, \ldots, p$, and $\delta_{ij} = \epsilon_{i+j} + \frac{m^{(p+1)}(x_i)}{(p+1)!}\frac{j^{p+1}}{n^{p+1}} + o(\frac{j^{p+1}}{n^{p+1}})$ are independent random errors. We then estimate the coefficients as

$$\begin{aligned} \hat{\beta}_i &= \arg\min_{\beta_i} \sum_{j=1}^{k} (Y_{i+j} - \beta_{i0} - \beta_{i1}d_j - \cdots - \beta_{ip}d_j^p)^2 \\ &= (\tilde{D}^T_{p;1,k}\tilde{D}_{p;1,k})^{-1}\tilde{D}^T_{p;1,k}\tilde{Y}_i^{(0)}, \end{aligned} \tag{10}$$

where $\hat{\beta}_i = (\hat{\beta}_{i0}, \ldots, \hat{\beta}_{ip})^T$, $\tilde{Y}_i^{(0)} = (Y_{i+1}, \ldots, Y_{i+k})^T$, and

$$\tilde{D}_{p;1,k} = \begin{pmatrix} 1 & 1^1 n^{-1} & \cdots & 1^p n^{-p} \\ 1 & 2^1 n^{-1} & \cdots & 2^p n^{-p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & k^1 n^{-1} & \cdots & k^p n^{-p} \end{pmatrix}.$$

Finally by (10), we define the one-sided first derivative estimator as $\hat{m}^{(1)}(x_i) = \hat{\beta}_{i1}$.

Following Section 2, we present the asymptotic variance and bias for $p = 1, \ldots, 5$ in Table 2. From the results, it is clear that the variance increases rapidly along with p. We set the optimal

Table 2
Statistical properties for the one-sided first derivative estimators for $p = 1, 2, 3, 4, 5$.

p	Bias	Var	k_{opt}	MSE	$k_p/k_{1,opt}$
1	$0.50m^{(2)}(x_i)\frac{k^1}{n^1}$	$12\sigma^2\frac{n^2}{k^3}$	$O(n^{4/5})$	$O(n^{-2/5})$	1.00
2	$-0.10m^{(3)}(x_i)\frac{k^2}{n^2}$	$192\sigma^2\frac{n^2}{k^3}$	$O(n^{6/7})$	$O(n^{-4/7})$	2.52
3	$1.19\times10^{-2}m^{(4)}(x_i)\frac{k^3}{n^3}$	$1200\sigma^2\frac{n^2}{k^3}$	$O(n^{8/9})$	$O(n^{-6/9})$	4.64
4	$-9.92\times10^{-4}m^{(5)}(x_i)\frac{k^4}{n^4}$	$4800\sigma^2\frac{n^2}{k^3}$	$O(n^{10/11})$	$O(n^{-8/11})$	7.37
5	$6.31\times10^{-5}m^{(6)}(x_i)\frac{k^5}{n^5}$	$14700\sigma^2\frac{n^2}{k^3}$	$O(n^{12/13})$	$O(n^{-10/13})$	10.70

Table 3
Ratios of variances between the TS and OS estimators for $p = 1, 2, 3, 4, 5$.

p	1	2	3	4	5
R_p	$\frac{1}{8}$	$\frac{1}{128}$	$\frac{1}{128}$	$\frac{1}{512}$	$\frac{1}{512}$

$k_{1,opt}$ for $p = 1$, and list the $k_p/k_{1,opt}$ to attain the same variance without considering the bias. It is obtained that the ratios are 2.52, 4.64, 7.37, 10.70 for $p = 2, 3, 4, 5$, respectively. In view of the large difference, we recommend to use the lower-order polynomials to estimate the first derivative in the boundary. The corresponding kernels are computed in Appendix D.

To further compare the TS and OS estimators, we define the ratio of their variances as

$$R_p = \frac{\mathrm{Var}(\hat{m}^{(1)}_{TS,p}(x_i))}{\mathrm{Var}(\hat{m}^{(1)}_{OS,p}(x_i))}.$$

While for comparison, we list the ratios for $p = 1, \ldots, 5$ in Table 3. As p increases, the ratio of variances increases dramatically. In particular for $p = 1$, the ratio of variances is 1/8, which is due to that the TS estimator adopts double-sized data compared to the OS estimator. Also by the symmetry of the data, the bias of the TS estimator is of order $O(k^2/n^2)$ which is smaller than the order $O(k/n)$ for the OS estimator.

3.2. Asymmetric two-sided local polynomial estimators

Recall that for the one-sided estimator, only data from x_i to x_{i+k} are used so that the information in the left side is completely ignored. To fully use the information, we propose to use the two-sided data from x_1 to x_{i+k} in this section. Specifically for the estimation at point x_i with $1 \le i \le k$, we consider the following asymmetric two-sided local polynomial estimator:

$$\check{m}^{(1)}(x_i) = e^T_{2,p}(\tilde{D}^T_{p;-(i-1),k}\tilde{D}_{p;-(i-1),k})^{-1}\tilde{D}^T_{p;-(i-1),k}\check{Y}^{(0)}_i, \tag{11}$$

where $e_{2,p} = (0, 1, \ldots, 0)^T_{(p+1)\times1}$, $\check{Y}^{(0)}_i = (Y_1, \ldots, Y_{i+k})^T$, and $\tilde{D}_{p;-(i-1),k}$ is defined similarly as in $\tilde{D}_{p;1,k}$.

To investigate the variation tendency of the ATS estimators from the OS estimator to the TS estimator, we let $\alpha = i/k$ with $0 \le \alpha \le 1$ as the tuning parameter. We then present in Fig. 2 the variance and bias functions for different values of α, and present the bias and variance values in Table 5 in Appendix D for $p = 1, 2, 3, 4, 5$ with $\alpha = 0, 0.1, 0.25, 0.5, 0.75, 0.9, 1$. It is readily known that for $p = 1$, the variance decreases smoothly by a factor of 8, while the bias $O(k/n)$ disappears slowly and decreases to the order $O(k^2/n^2)$ with α from 0 to 1; and for $p = 2$, the variance decreases dramatically by a factor of 128, while the bias $O(k^2/n^2)$ increases from negative value to positive value.

4. Derivative forecasting for outside points

In this section, we are interested to forecast the first derivative at points $\{x_l\}_{n<l\le n+k}$ outside the given observations $\{(x_i, Y_i)\}^n_{i=1}$. And for ease of presentation, we only consider $p = 1$ in what follows. The results for the large p values can be derived accordingly.

Assume that the mean function is one time continuously differentiable on $[0, 1+\delta]$ with $\delta = k/n$. Then for the outside point $x_l > 1$, we consider the following local linear forecasting estimator:

$$\check{m}^{(1)}(x_l) = e^T_{2,p}(\tilde{D}^T_{1;-k,-(l-n)}\tilde{D}_{1;-k,-(l-n)})^{-1}\tilde{D}^T_{1;-k,-(l-n)}\check{Y}^{(0)}_l, \tag{12}$$

where $\check{Y}^{(0)}_l = (Y_{l-k}, \ldots, Y_n)^T$.

Theorem 4. *Assume that model (1) holds with the design points equally spaced, and the smooth function is two times continuously differentiable on $[0, 1+\delta]$. Then for $1 < l/n \le 1+\delta$, the variance and bias of the first derivative estimator in (12) are*

$$\mathrm{Var}[\check{m}^{(1)}(x_l)] \doteq 12\sigma^2\frac{n^2}{(k-t)^3},$$
$$\mathrm{Bias}[\check{m}^{(1)}(x_l)] \doteq \frac{m^{(2)}(x_l)}{2n}\frac{k^3+4k^2t+4kt^2+t^3}{(k-t)^2},$$

where $t = l - n$ with $0 < t \le k$.

The proof of Theorem 4 is given in Appendix E. This theorem shows that the bias and variance both increase along with t. As t approaches to k, the order of variance varies from $O(n^2/k^3)$ to $O(n^2)$, and thus the variance diverges to the infinity; and for the order of bias, it varies from $O(k/n)$ to $O(k^3/n)$. From the tendency of bias and variance, it is also clear that the forecasting will become worse when x_l is far above x_n.

5. Simulation study

In this section, we conduct two simulation studies to evaluate the finite-sample performance of our new method. Specifically, the first study is to show that the debiased estimators with oracle tuning parameters have a better performance for highly-oscillatory functions, and the second study is to compare our method with two existing works in the literature.

5.1. Debiased estimators are better for highly-oscillatory functions

For the mean function, we consider the sine function:

$$m(x) = A\sin(2\pi x), \quad x \in [0, 1],$$

where $A \in [0, 10]$ controls the magnitude of oscillation. The error distribution is $N(0, 0.1^2)$ and the sample size is $n = 100$. We generate a total of 100 data sets. For each data set, we obtain three oracle first derivative estimators with $p = 1, 3, 5$ as follows. Assume that the true first derivative function is known, we obtain the oracle parameter k_p by minimizing the mean absolute error (MAE), which is to illustrate the possible best performance of each estimator [17]. Specifically, the oracle parameter k_p is chosen as

$$k_p = \arg\min_k \mathrm{MAE}(k) = \arg\min_k \frac{1}{n}\sum_{i=1}^{n}|\hat{m}^{(1)}_{p,k}(x_i) - m^{(1)}(x_i)|,$$

where $\hat{m}_{p,k}(x_i)$ depends on the polynomial order p and the tuning parameter k. Note that this version of MAE includes the estimates at interior and boundary points.

In Table 4, we apply $m\bar{e}an(s\bar{d}, \bar{k_p})$ to denote the finite-sample performance, where $m\bar{e}an$ and $s\bar{d}$ denote the mean and standard deviation of the 100 MAEs, and $\bar{k_p}$ denotes the mean of the corresponding optimal k_p. It can be shown that the MAE increases and the optimal k_p decreases as A increases with a fixed p, and the optimal k_p increases as p increases. We can choose the optimal p

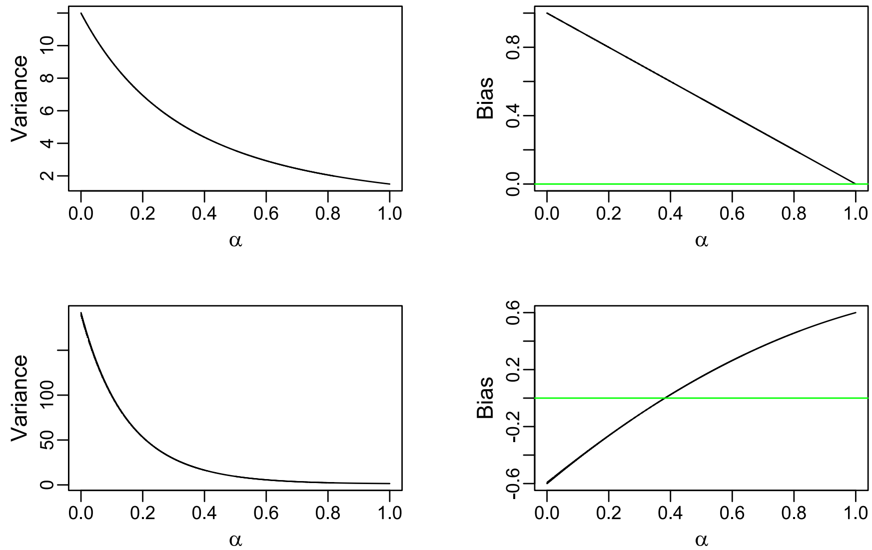

Fig. 2. Variance and bias functions for $p = 1$ (the top two panels) and $p = 2$ (the bottom two panels) with $\alpha \in [0, 1]$.

Table 4
Finite-sample performance of the first derivative estimators for $p = 1, 3, 5$.

n	$p\backslash A$	0	0.1	0.2	0.5
100	1	0.04 (0.02, 49)	0.15 (0.05, 21)	0.19 (0.05, 17)	0.28 (0.07, 13)
	3	0.15 (0.05, 47)	0.18 (0.06, 45)	0.20 (0.05, 44)	0.24 (0.07, 34)
	5	0.31 (0.10, 46)	0.31 (0.08, 47)	0.33 (0.12, 47)	0.32 (0.10, 48)
	$p\backslash A$	1	2	5	10
	1	0.36 (0.07, 11)	0.48 (0.08, 9)	0.71 (0.09, 7)	0.95 (0.10, 6)
	3	0.30 (0.07, 27)	0.35 (0.10, 24)	0.42 (0.08, 20)	0.50 (0.13, 18)
	5	0.31 (0.08, 46)	0.33 (0.10, 46)	0.35 (0.09, 43)	0.42 (0.12, 36)

as $p_{opt} = \arg\min_p \text{MAE}(k_p)$. To summarize, the estimator with $p = 1$ is the best for $A \in [0, 0.5)$, the estimator with $p = 3$ is the best for $A \in [0.5, 2)$, and the estimator with $p = 5$ is the best for $A \in [2, 10]$. These results demonstrate that the estimation bias increases rapidly as A increases, and the debiased estimators are needed for the mean function with a high-level oscillation.

5.2. Comparison with related methods

For further evaluation, we also compare our new method with two related methods including the local polynomial regression (LPR) and the penalized smoothing spline. For the local polynomial regression, we adopt the R package *locpol* with $p = 3$ [21]; and for the penalized smoothing spline, we adopt the R package *pspline* with *norder* $= 2$ (cubic smoothing spline) and *method* $= 4$ (the ordinary cross-validation for tuning parameter) [24]. For our method, we minimize the adjusted AMSE including all the design points

$$\text{AMSE}(p, k_p) = \frac{1}{n}\sum_{i=1}^{n} \text{MSE}[\hat{m}^{(1)}(x_i)],$$

and choose the optimal tuning parameters

$$k_{1,opt} = 1.96\left(\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(\hat{m}^{(3)}(x_i))^2/n}\right)^{1/7} n^{6/7},$$

$$k_{3,opt} = 3.84\left(\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(\hat{m}^{(5)}(x_i))^2/n}\right)^{1/11} n^{10/11},$$

$$k_{5,opt} = 4.96\left(\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(\hat{m}^{(7)}(x_i))^2/n}\right)^{1/15} n^{14/15},$$

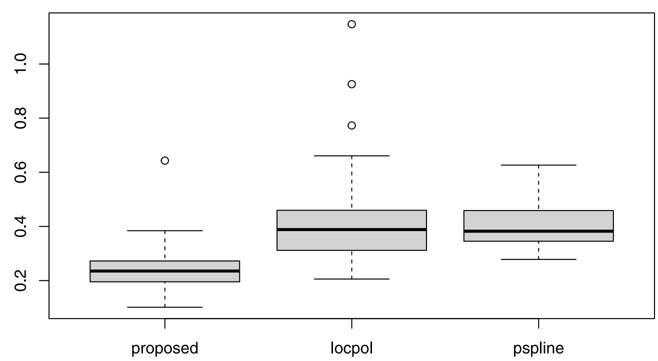

Fig. 3. MAEs for the proposed debiased estimator, the local polynomial estimator, and the penalized smoothing spline estimator.

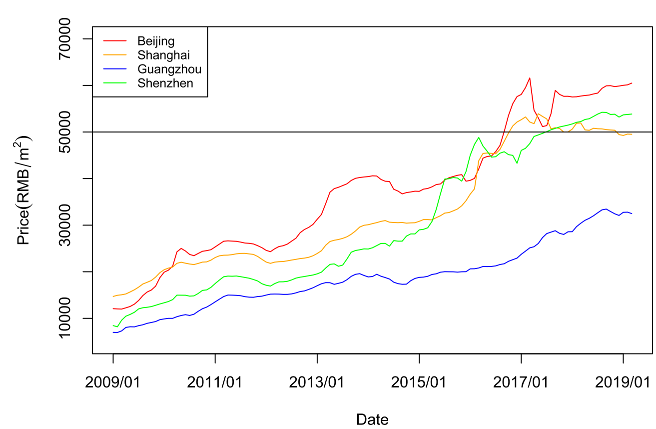

Fig. 4. Monthly house prices of the four first-tier cities in China with size 123 from January 2009 to March 2019.

where $\hat{\sigma}^2$ and $\hat{m}^{(3)}(x_i)$ can be obtained by the methods in Section 2.3. Furthermore, we select the optimal polynomial order

$$p_{opt} = \arg\min_{p=1,3,5} \text{AMSE}(p, k_{p,opt}).$$

Also for simplicity, we consider the mean function

$$m(x) = 10\sin(2\pi x), \quad x \in [0, 1].$$

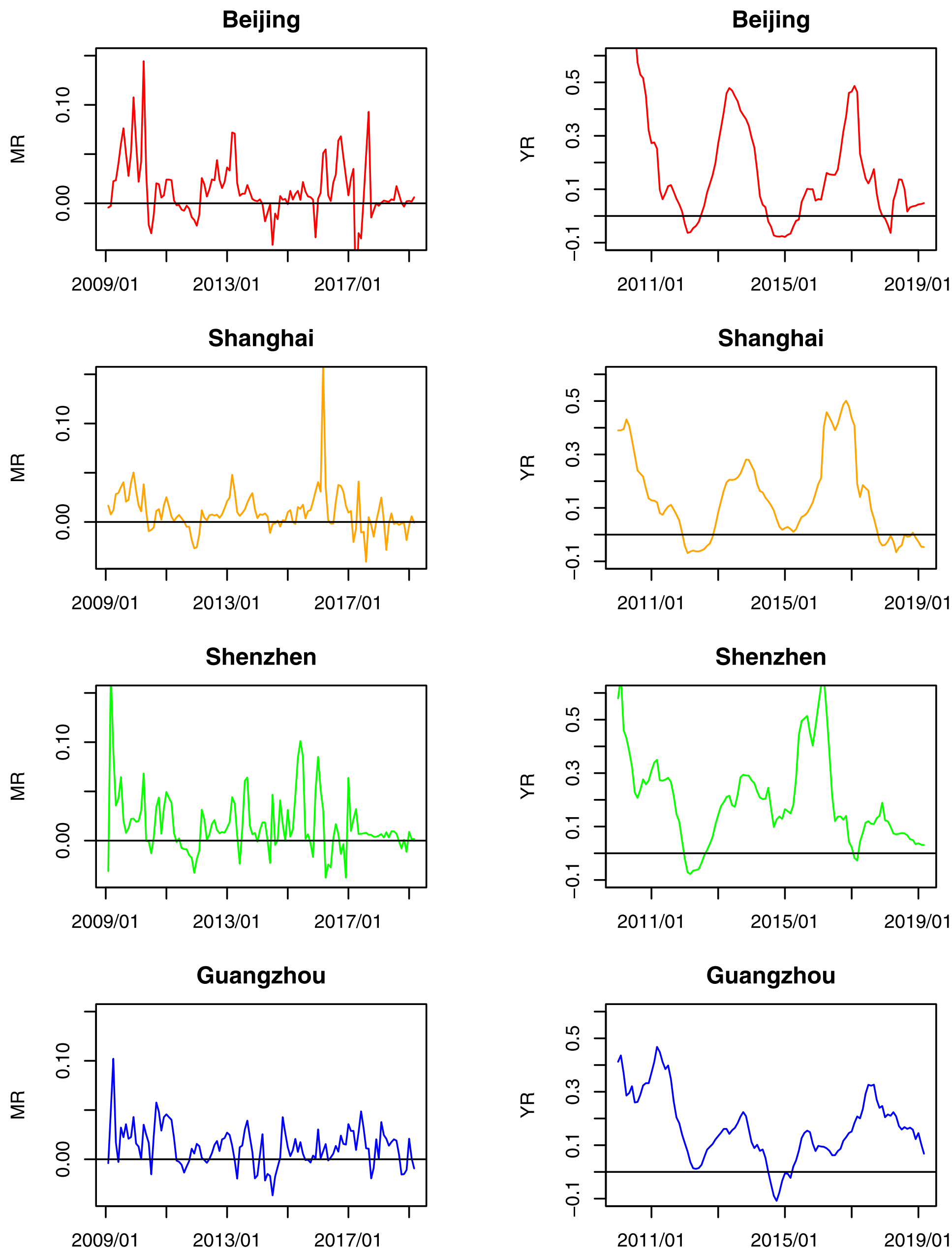

Fig. 5. Monthly (the left four panels) and Yearly (the right four panels) growth rates of house prices in the four first-tier cities of China.

The sample size is $n = 500$, and the repetition number is 100. Based on our procedure, we compute the optimal tuning parameters (mean, standard deviation) for $p = 1, 3, 5$: $k_{1,opt}$ (26.4, 1.4), $k_{3,opt}$ (83.6, 11.9), and $k_{5,opt}$ (169.7, 12.9). And we then select the optimal polynomial order $p = 5$ for all 100 repetitions and the corresponding optimal first derivative estimation, which matches the oracle estimator in Section 5.1. Simulation results show that our new method does have the ability to select the optimal first derivative estimation, and that the debiased estimator selected by our procedure has a better performance for highly-oscillatory functions in Fig. 3, which thus demonstrates the main advantage of our new method.

6. Real data analysis

In this section, we apply our proposed method to analyze two real data sets for estimating and forecasting the first derivatives. The first data set is the housing price data in the four first-tier cities of China in the past ten years, and the second data set is the daily return data of stock market index in USA and Hong Kong.

6.1. House price of first-tier cities in China

In the past ten years, China's economy has been developed rapidly. With the increased income, there is a huge demand for a higher housing quality, and consequently, it also increased the house price dramatically. We collect the monthly data of house price from the webpage: http://www.anjuke.com, from January 2009 to March 2019 for the four first-tier cities of China: Beijing, Shanghai, Guangzhou and Shenzhen. Fig. 4 indicates that the house prices have grown about five times in the past ten years, and are currently 50,000 RMB/m^2 or above except for Guangzhou.

There are two definitions for describing the growth trend: the monthly growth rate and the yearly growth ratio. Specifically, the monthly growth rate is defined as $\text{MR}_i = (Y_i - Y_{i-1})/Y_{i-1}$, and the yearly growth rate is $\text{YR}_i = (Y_i - Y_{i-12})/Y_{i-12}$.

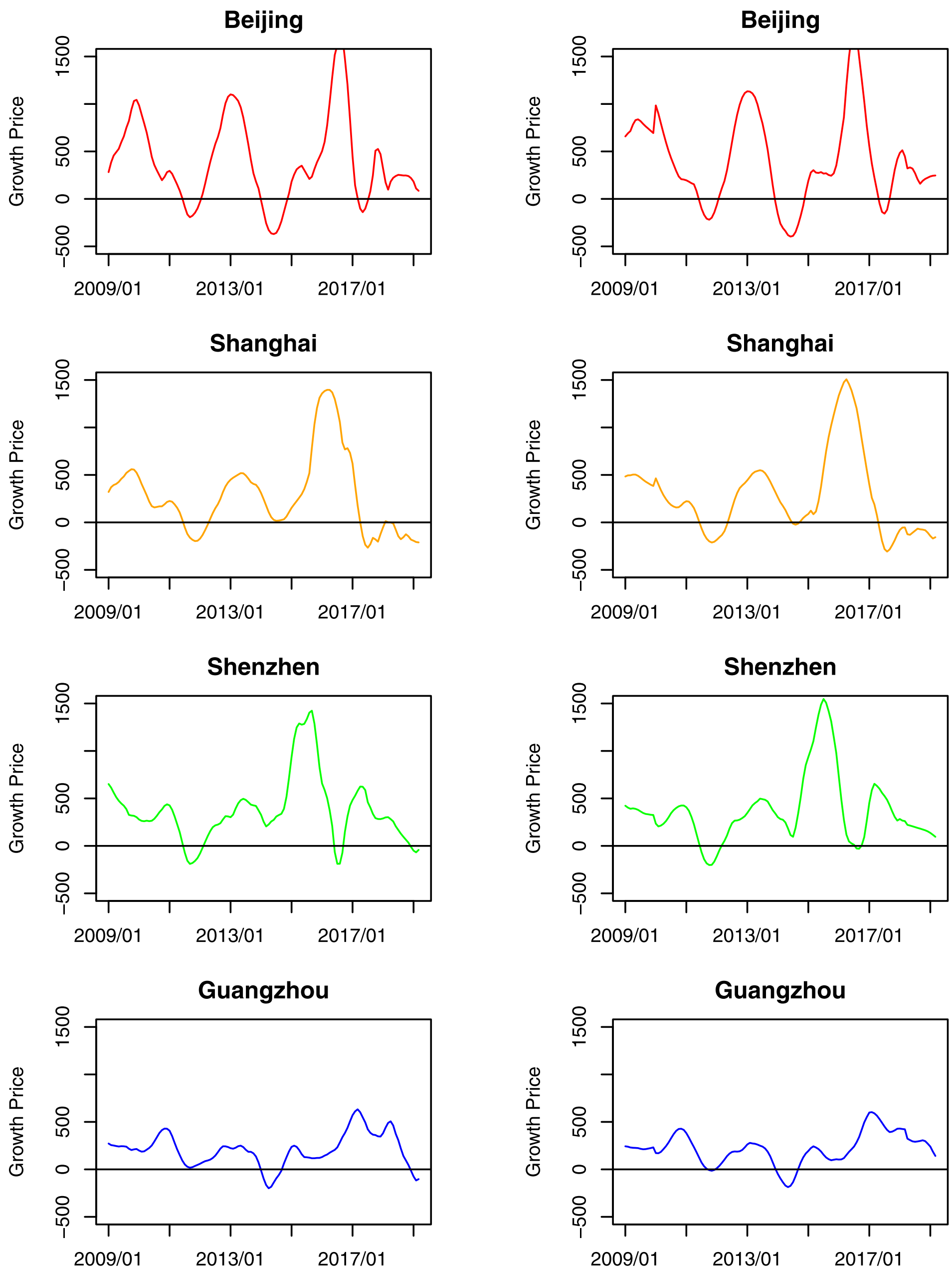

Fig. 6. The estimated average monthly growth prices with $p = 1$ (the left four panels) and with $p = 3$ (the right four panels) in the four first-tier cities of China.

The monthly growth rate in Fig. 5 shows that the house price increases very rapidly, and the largest monthly growth rate reaches up to 15% for Beijing, Shanghai and Shenzhen, and up to 10% for Guangzhou. The yearly growth rate in Fig. 5 shows that the house price goes through three cycles with the largest yearly growth rate more than 50% except for Guangzhou.

We apply our method to estimate the average monthly growth prices. First, we normalize the month variable into the domain [0, 1], that is $x_i = i/n$ with $n = 123$. Second, we consider two cases, $p = 1$ with tuning parameter $k_1 = 6$ and $p = 3$ with tuning parameter $k_3 = 12$, and obtain two first derivative estimates under different bias-reduction levels. Third, we compute the monthly growth prices using the first derivative estimates divided by n.

Fig. 6 shows that the two new estimators have a very similar overall performance, which indicates that the house price data set is with low-level oscillation. The averaged monthly growth price in China goes through three times fast-increasing and slow-decreasing. In the past three circles, Beijing was the leader of the price growth that had the largest growth amplitude (over 1500 RMB/m^2) and had the three obvious oscillation periods; Guangzhou was smooth and steady in the price growth, the largest growth amplitude of which was just a half of the other three major cities (about 750 RMB/m^2). It is also noteworthy that Guangzhou has recently achieved the highest growth price in the past ten years, which would have the biggest growth potential among the four first-tier cities.

6.2. Stock market index

The second data set contains the daily returns of two stock market indexes from 02 January 2008 to 30 June 2016, both for the S&P 500 index (SPI) and the Hang Seng index (HSI) [25]. As a routine approach, one often applies the time series model to analyze the daily log returns. Following this, we plot the time plot

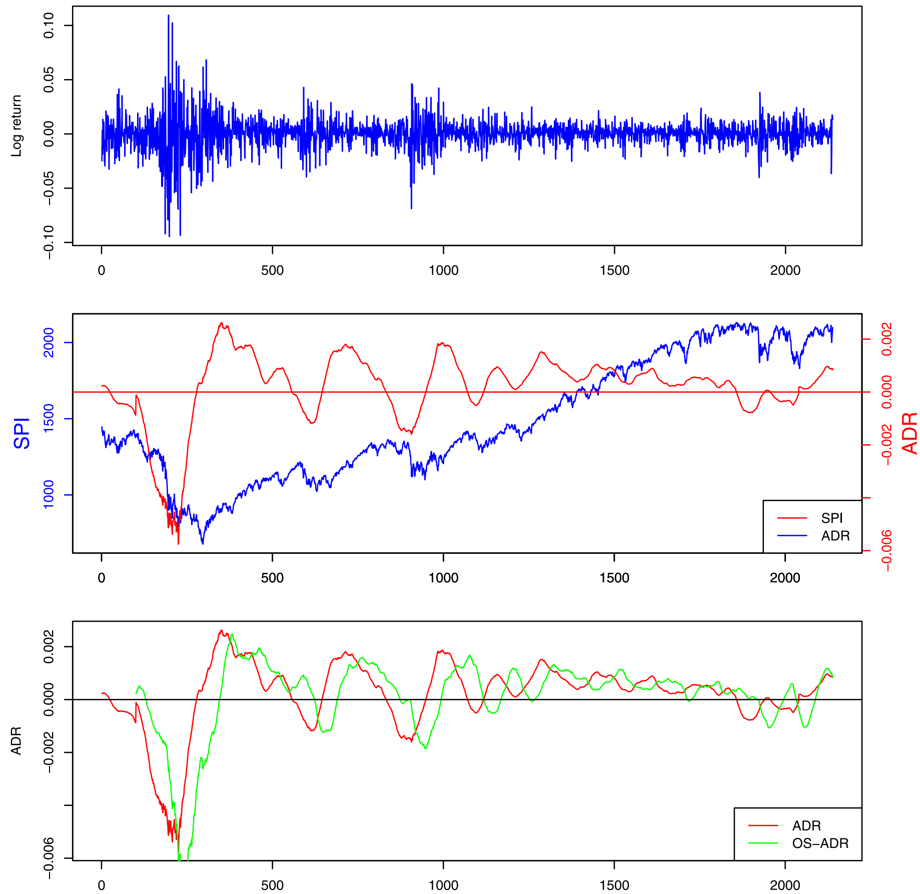

Fig. 7. The top panel shows the daily log returns of SPI data with $n = 2139$. The middle panel shows the SPI data and its ADR with the optimal tuning parameters $p = 3$ and $k = 95$. The bottom panel provides a comparison between the ADR and OS-ADR for SPI data with $p = 1$ and $k = 100$.

of the log returns $\{\log(Y_i) - \log Y_{i-1}\}$ in the top panel of Fig. 7. However, it does not present a clear pattern for the variation tendency of the stock market index using the daily log returns.

For the SPI data, we adopt the TS estimator in (8) and the ATS estimator in (11) to understand the past. To evaluate the performance, we compare our estimator with the empirical derivative [15] and the LowLS estimator [14] by the AMSE criterion, in which the standardized parameter is set as $\tilde{n} = 100, 200, 500$. By comparison, we have three interesting findings as summarized in Table 5. The first finding is that the standardized parameter has little effect on the selection of the optimal estimator, which coincides with the theoretical result in Appendix C that the AMSE remains nearly unchanged as $\tilde{n}$ varies. The second finding is that the optimal estimator with $p = 5$ selected by our procedure has the least AMSE among the three estimators, which supports the conclusion about the data set low-or-high-level oscillation. The third finding is that the information of 9 days observations should be used in the tendency analysis of the SPI data, which serves as a useful result for the security analysis.

To learn the long-term variation tendency, we choose the tuning parameters $p = 3$ and $k_3 = 95$. We obtain the average daily returns (ADR) in the middle panel of Fig. 7, which provides a more accurate pattern of the variation tendency for the ADR. In the past eight years, the ADR goes through four stages: descending dramatically stage, oscillatory ascending stage, robust increasing stage, and periodic oscillatory stage, which matches the variation tendency of the SPI. Furthermore, we adopt the OS method to forecast the future. Based on the OS method with $p = 1$ and $k_1 = 100$, we obtain the one-step average daily returns forecasting (OS-ADR). For comparison, we also plot the ADR and OS-ADR in the bottom panel of Fig. 7. It is evident that the both estimators have a similar variation tendency, which can be a valuable information for the investigation of the future, even though the OS-ADR is delayed for a few days. Finally, for the HSI data, we also perform a similar analysis and plot the average daily returns with $p = 1$ and $k = 100$ in Appendix F.

Table 5
Comparison of AMSEs among the three estimators for $\tilde{n} = 100, 200, 500$.

$\tilde{n}$	Method	k_{opt}	Bias^2	Variance	AMSE	p_{opt}
100	Empirical	3	4.70	11.44	16.14	
	LowLS	6	3.97	8.94	12.91	
	Proposed	9	1.74	8.11	9.85	5
200	Empirical	3	12.34	11.44	23.78	
	LowLS	6	3.93	8.94	12.87	
	Proposed	9	3.06	8.11	11.17	5
500	Empirical	3	4.74	11.44	16.18	
	LowLS	6	3.87	8.94	12.81	
	Proposed	9	1.76	8.11	9.87	5

7. Discussion

In this paper, we proposed several new estimators for the first derivative of the mean function. For interior points, the local constant estimator based on the difference sequence in essence is the same as the first empirical derivative in Charnigo et al. [13]. To solve the peak-valley problem caused by the estimation bias, we proposed local polynomial estimators and computed their corresponding two-sided kernels, which are equivalent to the variance-minimizing kernels in Gasser et al. [26].

For boundary and outside points, we proposed the asymmetric two-sided estimators for reducing the bias and variance, and proposed one-sided estimator for forecasting the variation tendency in the future. The debiased estimators are optimal under different bias-reduction levels, and are fully boundary adaptive and automatic without requiring specific data modification. Also to increase the scope of application, the data transformation can be recommended for real data analysis.

Recall that all results in this paper were derived under the equally spaced design with independent and identically distributed random errors. In the areas of machine learning and artificial intelligence, more and more data sets are nowadays recorded in equal intervals of time, e.g., the monthly house price and the daily return of stock market index as in our real data analysis. This indicates that our method may have the potential to be widely used. From another perspective, we note that the equally spaced data may not necessarily be independent, but rather be correlated with each other [27]. Some recent advances in this direction include, for example, the study of regression function with correlated errors in De Brabanter et al. [28], and the study of continuous-time regression models with correlated errors in Dette et al. [29]. To our knowledge, there is little theoretical development on the derivative learning with correlated errors, and hence it may warrant further research.

Besides the first derivative of the mean function, it is noteworthy that the density functions are also widely applied to clustering in pattern recognition. To name a few, Xie et al. [30] applied the density methodology to analyze the extended-range electric city bus. Ikonomakis et al. [31] proposed a content driven clustering algorithm by combining density and distance functions. Li et al. [32] further provided a nearest neighbor graph algorithm. Now since the density function can be interpreted and estimated as the first derivative of the empirical distribution function [33], we expect that our newly proposed methods may also be readily applied in the density function estimation to further improve the estimation efficiency.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

WenWu Wang was supported by National Natural Science Foundation of China (No. 12071248), National Bureau of Statistics of China (No. 2020LZ26).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.knosys.2021.107781.

References

[1] T. Duong, B. Goud, K. Schauer, Closed-form density-based framework for automatic detection of cellular morphology changes, Proc. Natl. Acad. Sci. USA 109 (2012) 8382–8387.
[2] J. Chacón, T. Duong, Data-driven density derivative estimation, with applications to nonparametric clustering and bump hunting, Electron. J. Stat. 7 (2013) 499–532.
[3] M. Khismatullina, M. Vogt, Multiscale inference and long-run variance estimation in non-parametric regression with time series errors, J. R. Stat. Soc. Ser. B Stat. Methodol. 82 (2020) 5–37.
[4] T. Górecki, Using derivatives in a longest common subsequence dissimilarity measure for time series classification, Pattern Recognit. Lett. 45 (2014) 99–105.
[5] K. Wang, C. Bichot, Y. Li, B. Li, Local binary circumferential and radial derivative pattern for texture classification, Pattern Recognit. 67 (2017) 213–229.
[6] T. Mondal, N. Ragot, J. Ramel, U. Pal, Comparative study of conventional time series matching techniques for word spotting, Pattern Recognit. 73 (2018) 47–64.
[7] P.S. Swain, K. Stevenson, A. Leary, L.F. Montano-Gutierrez, I.B.N. Clark, J. Vogel, T. Pilizota, Inferring time derivatives including cell growth rates using Gaussian process, Nature Commun. 7, 13766 (2016).
[8] D. Chen, J.M. Mirebeau, L.D. Cohen, Global minimum for a finsler elastica minimal path approach, Int. J. Comput. Vis. 122 (2017) 458–483.
[9] A.J. Simpkin, M. Durban, D.A. Lawlor, C. MacDonald-Wallis, M.T. May, C. Metcalfe, K. Tilling, Derivative estimation for longitudinal data analysis: Examining features of blood pressure measured repeatedly during pregnancy, Stat. Med. 37 (2018) 2836–2854.
[10] W. Wang, P. Yu, L. Lin, T. Tong, Robust estimation of derivatives using locally weighted least absolute deviation regression, J. Mach. Learn. Res. 20 (60) (2019) 1–49.
[11] G.L. Page, M.X. Rodrí guez Álvarez, D.J. Lee, Bayeslan hierarchical modelling of growth curve derivatives via sequences of quotient differences, J. R. Stat. Soc. Ser. C. Appl. Stat. 69 (2020) 459–481.
[12] J. Fan, I. Gijbels, Local Polynomial Modelling and Its Applications, Chapman & Hall, London, 1996.
[13] R. Charnigo, B. Hall, C. Srinivasan, A generalized C_p criterion for derivative estimation, Technometrics 53 (2011) 238–253.
[14] W. Wang, L. Lin, Derivative estimation based on difference sequence via locally weighted least squares regression, J. Mach. Learn. Res. 16 (2015) 2617–2641.
[15] K. De Brabanter, J. De Brabanter, B. De Moor, I. Gijbels, Derivative estimation with local polynomial fitting, J. Mach. Learn. Res. 14 (2013) 281–301.
[16] W. Härdle, Applied Nonparametric Regression, Cambridge University Press, Cambridge, 1990.
[17] W. Dai, T. Tong, M.G. Genton, Optimal estimation of derivatives in nonparametric regression, J. Mach. Learn. Res. 17 (164) (2016) 1–25.
[18] Y. Liu, K. De Brabanter, Derivative estimation in random design, in: the 32nd Conference on Neural Information Processing Systems, Montréal, Canada, 2018.
[19] W. Wang, L. Lin, L. Yu, Optimal variance estimation based on lagged second-order difference in nonparametric regression, Comput. Statist. 32 (2017) 1047–1063.
[20] P. Hall, J. Kay, D. Titterington, Asymptotically optimal difference-based estimation of variance in nonparametric regression, Biometrika 77 (1990) 521–528.
[21] J.L.O. Cabrera, Locpol: Kernel local polynomial regression, 2018, R packages version 0.7-0. https://cran.r-project.org/web/packages/locpol.
[22] W. Wang, P. Yu, Asymptotically optimal differenced estimators of error variance in nonparametric regression, Comput. Statist. Data Anal. 105 (2017) 125–143.
[23] D. Card, D.S. Lee, Z. Pei, A. Weber, Inference on causal effects in a generalized regression kink design, Econometrica 83 (2015) 2453–2483.
[24] J.O. Ramsay, B. Ripley, Pspline: Penalized smoothing splines, 2017, R packages version 1.0-18. https://cran.r-project.org/web/packages/pspline.
[25] Y. Zheng, Q. Zhu, G. Li, Z. Xiao, Hybrid quantile regression estimation for time series models with conditional heteroscedasticity, J. R. Stat. Soc. Ser. B Stat. Methodol. 80 (2018) 975–993.
[26] T. Gasser, H.G. Müller, V. Mammitzsch, Kernels for nonparametric curve estimation, J. R. Stat. Soc. Ser. B Stat. Methodol. 47 (1985) 238–252.
[27] G. Wang, W. Li, K. Zhu, New HSIC-based tests for independence between two stationary multivariate time series, Statist. Sinica 31 (2020) 269–300.
[28] K. De Brabanter, F. Cao, I. Gijbels, J. Opsomer, Local polynomial regression with correlated errors in random design and unknown correlated structure, Biometrika 105 (2018) 681–690.
[29] H. Dette, A. Pepelyshev, A. Zhigljavsky, The BLUE in continuous-time regression models with correlated errors, Ann. Statist. 47 (2019) 1928–1959.
[30] H. Xie, G. Tian, H. Chen, J. Wang, Y. Huang, A distribution density-based methodology for driving data cluster analysis: A case study for an extended-range electric city bus, Pattern Recognit. 73 (2018) 131–143.
[31] E.K. Ikonomakis, G.M. Spyrou, M.N. Vrahatis, Content driven clustering algorithm combining density and distance functions, Pattern Recognit. 87 (2019) 190–202.
[32] H. Li, X. Liu, T. Li, R. Gan, A novel density-based clustering algorithm using nearest neighbor graph, Pattern Recognit. 102 (2020) 107206.
[33] M.D. Cattaneo, M. Jansson, X. Ma, Simple local polynomial density estimators, J. Amer. Statist. Assoc. 115 (2020) 1449–1455.