Journal of Machine Learning Research 14 (2013) 281-301 Submitted 1/12; Revised 8/12; Published 1/13

Derivative Estimation with Local Polynomial Fitting

Kris De Brabanter KRIS.DEBRABANTER@ESAT.KULEUVEN.BE
Jos De Brabanter* JOS.DEBRABANTER@ESAT.KULEUVEN.BE
Bart De Moor* BART.DEMOOR@ESAT.KULEUVEN.BE
Department of Electrical Engineering SCD-SISTA
KU Leuven
Kasteelpark Arenberg 10
B-3001 Leuven, Belgium

Irène Gijbels IRENE.GIJBELS@WIS.KULEUVEN.BE
Department of Mathematics & Leuven Statistics Research Centre (LStat)
KU Leuven
Celestijnenlaan 200B
B-3001 Leuven, Belgium

Editor: Xiaotong Shen

Abstract

We present a fully automated framework to estimate derivatives nonparametrically without estimating the regression function. Derivative estimation plays an important role in the exploration of structures in curves (jump detection and discontinuities), comparison of regression curves, analysis of human growth data, etc. Hence, the study of estimating derivatives is equally important as regression estimation itself. Via empirical derivatives we approximate the qth order derivative and create a new data set which can be smoothed by any nonparametric regression estimator. We derive L_1 and L_2 rates and establish consistency of the estimator. The new data sets created by this technique are no longer independent and identically distributed (i.i.d.) random variables anymore. As a consequence, automated model selection criteria (data-driven procedures) break down. Therefore, we propose a simple factor method, based on bimodal kernels, to effectively deal with correlated data in the local polynomial regression framework.

Keywords: nonparametric derivative estimation, model selection, empirical derivative, factor rule

1. Introduction

The next section describes previous methods and objectives for nonparametric derivative estimation. Also, a brief summary of local polynomial regression is given.

1.1 Previous Methods And Objectives

Ever since the introduction of nonparametric estimators for density estimation, regression, etc. in the mid 1950s and early 1960s, their popularity has increased over the years. Mainly, this is due to the fact that statisticians realized that pure parametric thinking in curve estimations often does not

*. Bart De Moor and Jos De Brabanter are with IBBT-KU Leuven Future Health Department, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium. Jos De Brabanter is also with the Departement Industrieel Ingenieur, KaHo Sint Lieven (Associatie KU Leuven), G. Desmetstraat 1, B-9000 Gent, Belgium.

meet the need for flexibility in data analysis. Many of their properties have been rigorously investigated and are well understood, see, for example, Fan and Gijbels (1996), Györfi et al. (2002) and Tsybakov (2009). Although the importance of regression estimation is indisputable, sometimes the first or higher order derivatives of the regression function can be equally important. This is the case in the exploration of structures in curves (Chaudhuri and Marron, 1999; Gijbels and Goderniaux, 2004) (jump detection and discontinuities), inference of significant features in data, trend analysis in time series (Rondonotti et al., 2007), comparison of regression curves (Park and Kang, 2008), analysis of human growth data (Müller, 1988; Ramsay and Silverman, 2002), the characterization of submicroscopic nanoparticles from scattering data (Charnigo et al., 2007) and inferring chemical compositions. Also, estimation of derivatives of the regression function is required for plug-in bandwidth selection strategies (Wand and Jones, 1995) and in the construction of confidence intervals (Eubank and Speckman, 1993).

It would be tempting to differentiate the estimated nonparametric estimate $\hat{m}(x)$ w.r.t. the independent variable to obtain the first order derivative of the regression function. However, such a procedure can only work well if the original regression function is extremely well estimated. Otherwise, it can lead to wrong derivative estimates when the data is noisy. Therefore, it can be expected that straightforward differentiation of the regression estimate $\hat{m}(x)$ will result in an accumulation of errors which increase with the order of the derivative.

In the literature there are two main approaches to nonparametric derivative estimation: Regression/smoothing splines and local polynomial regression. In the context of derivative estimation, Stone (1985) has shown that spline derivative estimators can achieve the optimal L_2 rate of convergence. Asymptotic bias and variance properties and asymptotic normality have been established by Zhou and Wolfe (2000). In case of smoothing splines, Ramsay (1998) noted that choosing the smoothing parameter is tricky. He stated that data-driven methods are generally poor guides and some user intervention is nearly always required. In fact, Wahba and Wang (1990) demonstrated that the smoothing parameter for a smoothing spline depends on the integer q while minimizing $\sum_{i=1}^{n}(\hat{m}^{(q)}(x_i) - m^{(q)}(x_i))^2$. Jarrow et al. (2004) suggested an empirical bias bandwidth criterion to estimate the first derivative via semiparametric penalized splines.

Early works discussing kernel based derivative estimation include Gasser and Müller (1984) and Härdle and Gasser (1985). Müller et al. (1987) and Härdle (1990) proposed a generalized version of the cross-validation technique to estimate the first derivative via kernel smoothing using difference quotients. Their cross-validation technique is related to modified cross-validation for correlated errors proposed by Chu and Marron (1991). Although the use of difference quotients may be natural, their variances are proportional to n^2 in case of equispaced design. Therefore, this type of cross-validation will be spoiled due to the large variability. In order to improve on the previous methods, Müller et al. (1987) also proposed a factor method to estimate a derivative via kernel smoothing. A variant of the factor method was also used by Fan and Gijbels (1995).

In case of local polynomial regression (Fan and Gijbels, 1996), the estimation of the qth derivative is straightforward. One can estimate $m^{(q)}(x)$ via the intercept coefficient of the qth derivative (local slope) of the local polynomial being fitted at x, assuming that the degree p is larger or equal to q. Note that this estimate of the derivative is, in general, not equal to the qth derivative of the estimated regression function $\hat{m}(x)$. Asymptotic properties as well as asymptotic normality were established by Fan and Gijbels (1996). Strong uniform consistency properties were shown by Delecroix and Rosa (2007).

As mentioned before, two problems inherently present in nonparametric derivative estimation are the unavailability of the data for derivative estimation (only regression data is given) and bandwidth or smoothing selection. In what follows we investigate a new way to compute derivatives of the regression function given the data $(x_1,Y_1),\ldots,(x_n,Y_n)$. This procedure is based on the creation of a new data set via empirical derivatives. A minor drawback of this approach is the fact the data are correlated and hence poses a threat to classical bandwidth selection methods. In order to deal with correlated data we extend our previous work (De Brabanter et al., 2011) and derive a factor method based on bimodal kernels to estimate the derivatives of the unknown regression function.

This paper is organized as follows. Next, we give a short introduction to local polynomial fitting. Section 2 illustrates the principle of empirical first order derivatives and their use within the local polynomial regression framework. We derive bias and variance of empirical first order derivatives and establish pointwise consistency. Further, the behavior at the boundaries of empirical first order derivatives is described. Section 3 generalizes the idea of empirical first order derivatives to higher order derivatives. Section 4 discusses the problem of bandwidth selection in the presence of correlated data. In Section 5 we conduct a Monte Carlo experiment to compare the proposed method with two often used methods for derivative estimation. Finally, Section 6 states the conclusions.

1.2 Local Polynomial Regression

Consider the bivariate data $(x_1,Y_1),\ldots,(x_n,Y_n)$ which form an independent and identically distributed (i.i.d) sample from a population (x,Y) where x belongs to $\mathcal{X}\subseteq\mathbb{R}$ and $Y\in\mathbb{R}$. If $\mathcal{X}$ denotes the closed real interval $[a,b]$ then $x_i = a+(i-1)(b-a)/(n-1)$. Denote by $m(x)=\mathbf{E}[Y]$ the regression function. The data is regarded to be generated from the model

$$Y = m(x)+e, \tag{1}$$

where $\mathbf{E}[e]=0$, $\mathbf{Var}[e]=\sigma^2<\infty$, x and e are independent and m is twice continuously differentiable on $\mathcal{X}$. Suppose that $(p+1)^{\text{th}}$ derivative of m at the point x_0 exists. Then, the unknown regression function m can be locally approximated by a polynomial of order p. A Taylor expansion yields, for x in a neighborhood of x_0,

$$m(x)\approx\sum_{j=0}^{p}\frac{m^{(j)}(x_0)}{j!}(x-x_0)^j\equiv\sum_{j=0}^{p}\beta_j(x-x_0)^j. \tag{2}$$

This polynomial is fitted locally by the following weighted least squares regression problem:

$$\min_{\beta_j\in\mathbb{R}}\sum_{i=1}^{n}\Big\{Y_i-\sum_{j=0}^{p}\beta_j\,(x_i-x_0)^j\Big\}^2K_h(x_i-x_0), \tag{3}$$

where β_j are the solutions to the weighted least squares problem, h is the bandwidth controlling the size of the local neighborhood and $K_h(\cdot)=K(\cdot/h)/h$ with K a kernel function assigning weights to each point. From the Taylor expansion (2) it is clear that $\hat{m}^{(q)}(x_0)=q!\hat{\beta}_q$ is an estimator for $m^{(q)}(x_0)$, $q=0,1,\ldots,p$. For local polynomial fitting $p-q$ should be taken to be odd as shown in Ruppert and Wand (1994) and Fan and Gijbels (1996). In matrix notation (3) can be written as:

$$\min_{\beta}\{(\mathbf{y}-X\beta)^T\mathbf{W}(\mathbf{y}-X\beta)\},$$

where $\mathbf{y} = (Y_1, \ldots, Y_n)^T$, $\beta = (\beta_0, \ldots, \beta_p)^T$ and

$$\mathbf{X} = \begin{pmatrix} 1 & (x_1 - x_0) & \cdots & (x_1 - x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (x_n - x_0) & \cdots & (x_n - x_0)^p \end{pmatrix},$$

and $\mathbf{W}$ the $n \times n$ diagonal matrix of weights

$$\mathbf{W} = \text{diag}\{K_h(x_i - x_0)\}.$$

The solution vector is given by least squares theory and yields

$$\hat{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}.$$

2. Derivative Estimation

In this section we first illustrate the principle of empirical first order derivatives and how they can be used within the local polynomial regression framework to estimate first order derivatives of the unknown regression function.

2.1 Empirical Derivatives And Its Properties

Given a local polynomial regression estimate (3), it would be tempting to differentiate it w.r.t. the independent variable. Such a procedure can lead to wrong derivative estimates when the data is noisy and will deteriorate quickly when calculating higher order derivatives. A possible solution to avoid this problem is by using the first order difference quotient

$$Y_i^{(1)} = \frac{Y_i - Y_{i-1}}{x_i - x_{i-1}}$$

as a noise corrupted version of $m'(x_i)$ where the superscript (1) signifies that $\hat{Y}_i^{(1)}$ is a noise corrupted version of the first (true) derivative. Such an approach has been used by Müller et al. (1987) and Härdle (1990) to estimate first order derivatives via kernel smoothing. Such an approach produces a very noisy estimate of the derivative which is of the order $O(n^2)$ and as a result it will be difficult to estimate the derivative function. For equispaced design yields

$$\mathbf{Var}(Y_i^{(1)}) = \frac{1}{(x_i - x_{i-1})^2}(\mathbf{Var}(Y_i) + \mathbf{Var}(Y_{i-1})) = \frac{2\sigma^2}{(x_i - x_{i-1})^2} = \frac{2\sigma^2 (n-1)^2}{d(\mathcal{X})^2},$$

where $d(\mathcal{X}) := \sup \mathcal{X} - \inf \mathcal{X}$. In order to reduce the variance we use a variance-reducing linear combination of symmetric (about i) difference quotients

$$Y_i^{(1)} = Y^{(1)}(x_i) = \sum_{j=1}^{k} w_j \cdot \left(\frac{Y_{i+j} - Y_{i-j}}{x_{i+j} - x_{i-j}} \right), \tag{4}$$

where the weights $w_1, \ldots, w_k$ sum up to one. The linear combination (4) is valid for $k+1 \leq i \leq n-k$ and hence $k \leq (n-1)/2$. For $2 \leq i \leq k$ or $n-k+1 \leq i \leq n-1$ we define $Y_i^{(1)}$ by replacing $\sum_{j=1}^{k}$ in (4) by $\sum_{j=1}^{k(i)}$ where $k(i) = \min\{i-1, n-i\}$ and replacing $w_1, \ldots, w_{k(i)}$ by $w_1 / \sum_{j=1}^{k(i)} w_j, \ldots, w_{k(i)} / \sum_{j=1}^{k(i)} w_j$.

Finally, for $i = 1$ and $i = n$ we define $Y_1^{(1)}$ and $Y_n^{(1)}$ to coincide with $Y_2^{(1)}$ and $Y_{n-1}^{(1)}$. The proportion of indices i falling between $k+1$ and $n-k$ approaches 1 as n increases, so this boundary issue becomes smaller as n becomes larger. Alternatively, one may just leave $Y_i^{(1)}$ undefined for indices i not falling between $k+1$ and $n-k$. This latter approach will be used in the remaining of the paper, except in Figure 1 where we want to illustrate the boundary issues.

Linear combinations such as (4) are frequently used in finite element theory and are useful in the numerical solution of differential equations (Iserles, 1996). However, the weights used for solving differential equations are not appropriate here because of the random errors in model (1). Therefore, we need to optimize the weights so that minimum variance is attained. This result is stated in Proposition 1.

Proposition 1 *Assume model* (1) *holds with equispaced design and let* $\sum_{j=1}^{k} w_j = 1$. *Then, for* $k+1 \le i \le n-k$, *the weights*

$$w_j = \frac{6j^2}{k(k+1)(2k+1)}, \quad j = 1, \ldots, k$$

minimize the variance of $Y_i^{(1)}$ *in* (4).
Proof: see Appendix A. ■

Figure 1a displays the empirical first derivative for $k \in \{2,5,7,12\}$ generated from model (1) with $m(x) = \sqrt{x(1-x)}\sin((2.1\pi)/(x+0.05))$, $x \in [0.25,1]$ for 300 equispaced points and $e \sim \mathcal{N}(0, 0.1^2)$. For completeness the first order difference quotient is also shown. Even for a small k, it can be seen that the empirical first order derivatives are noise corrupted versions of the true derivative m'. In contrast, difference quotients produce an extreme noisy version of the true derivative (Figure 1b). Also, note the large amplitude of the signal constructed by difference quotients. When k is large, empirical first derivatives are biased near local extrema of the true derivative (see Figure 1f). Further, the boundary issues are clearly visible in Figure 1d through Figure 1f for $i \in [1, k+1] \cup [n-k, n]$.

The next two theorems give asymptotic results on the bias and variance and establish pointwise consistency of the empirical first order derivatives.

Theorem 2 *Assume model* (1) *holds with equispaced design and m is twice continuously differentiable on* $\mathcal{X} \subseteq \mathbb{R}$. *Further, assume that the second order derivative* $m^{(2)}$ *is finite on* $\mathcal{X}$. *Then the bias and variance of the empirical first order derivative, with weights assigned by Proposition 1, satisfy*

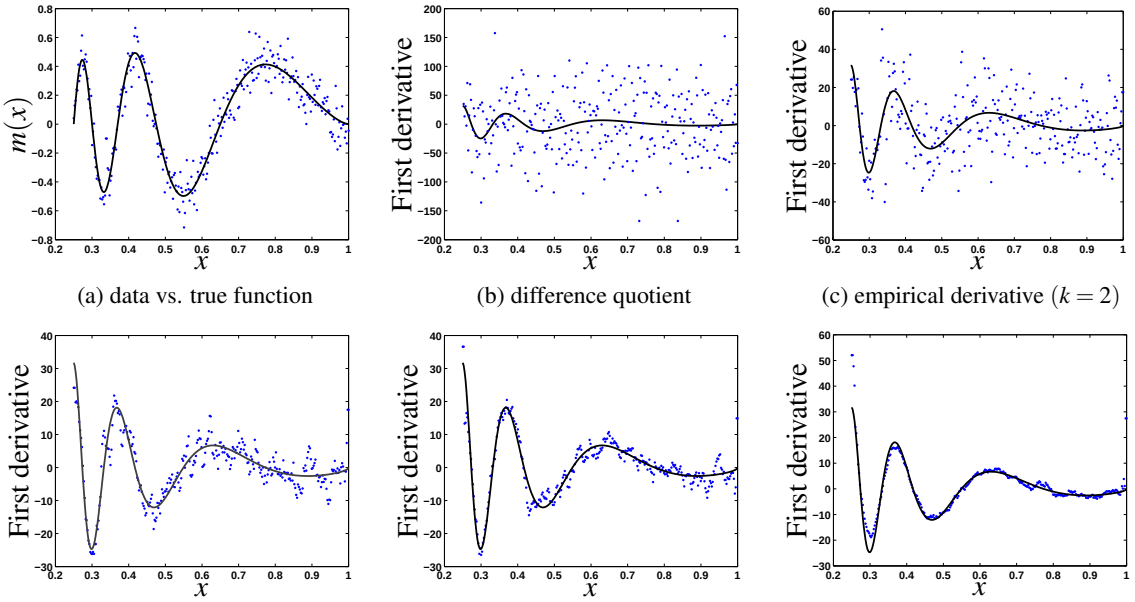
$$\text{bias}(Y_i^{(1)}) = O(n^{-1}k) \quad \textit{and} \quad \mathbf{Var}(Y_i^{(1)}) = O(n^2 k^{-3})$$

uniformly for $k+1 \le i \le n-k$.
Proof: see Appendix B. ■

Theorem 3 (Pointwise consistency) *Assume* $k \to \infty$ *as* $n \to \infty$ *such that* $nk^{-3/2} \to 0$ *and* $n^{-1}k \to 0$. *Further assume that m is twice continuously differentiable on* $\mathcal{X} \subseteq \mathbb{R}$. *Then, for the minimum variance weights given in Proposition 1, we have for any* $\varepsilon > 0$

$$\mathbf{P}\big(|Y_i^{(1)} - m'(x_i)| \ge \varepsilon\big) \to 0.$$

Proof: see Appendix C. ■



(a) data vs. true function (b) difference quotient (c) empirical derivative ($k = 2$)

(d) empirical derivative ($k = 5$) (e) empirical derivative ($k = 7$) (f) empirical derivative ($k = 12$)

Figure 1: (a) Simulated data set of size $n = 300$ equispaced points from model (1) with $m(x) = \sqrt{x(1-x)}\sin((2.1\pi)/(x+0.05))$ and $e \sim \mathcal{N}(0, 0.1^2)$; (b) first order difference quotients which are barely distinguishable from noise. As a reference, the true derivative is also displayed (full line); (c)-(f) empirical first derivatives for $k \in \{2,5,7,12\}$.

According to Theorem 2 and Theorem 3, the bias and variance of the empirical first order derivative tends to zero and $k \to \infty$ faster than $O(n^{2/3})$ but slower than $O(n)$. The optimal rate at which $k \to \infty$ such that the mean squared error (MSE) of the empirical first order derivatives will tend to zero at the fastest possible rate is a direct consequence of Theorem 2. This optimal L_2 rate is achieved for $k = O(n^{4/5})$ and consequently, the $\text{MSE}(Y_i^{(1)}) = \mathbf{E}(Y_i^{(1)} - m'(x_i))^2 = O(n^{-2/5} + n^{-1/5})$. Similar, one can also establish the rate of the mean absolute deviation (MAD) or L_1 rate of the estimator, that is, $\mathbf{E}\,|Y_i^{(1)} - m'(x_i)|$. By Jensen's inequality

$$\begin{aligned} \mathbf{E}\,|Y_i^{(1)} - m'(x_i)| &\leq \mathbf{E}\,|Y_i^{(1)} - \mathbf{E}(Y_i^{(1)})| + |\,\mathbf{E}(Y_i^{(1)}) - m'(x_i)| \\ &\leq \sqrt{\mathbf{Var}(Y_i^{(1)})} + \text{bias}(Y_i^{(1)}) = O(n^{-1/5}), \end{aligned}$$

for the optimal L_1 rate of $k = O(n^{4/5})$ (equal to the optimal L_2 rate). Under the same conditions as Theorem 3, it is easy to show that $\mathbf{E}\,|Y_i^{(1)} - m'(x_i)| \to 0$. Even though we know the optimal asymptotic order of k, the question still remains how to choose k in practice. In many data analyses, one would like to get a quick idea what the value of k should be. In such a case a rule of thumb can be very suitable. Such a rule can be somewhat crude but it possesses simplicity and is easily computable. In order to derive a suitable expression for the MSE, we start from the bias and variance expressions for the empirical derivatives. An upperbound for the MSE is given by (see also the proof

of Theorem 2)

$$\begin{aligned} \mathrm{MSE}(Y_i^{(1)}) &= \mathrm{bias}^2(Y_i^{(1)}) + \mathbf{Var}(Y_i^{(1)}) \\ &\leq \frac{9k^2(k+1)^2\mathcal{B}^2 d(\mathcal{X})^2}{16(n-1)^2(2k+1)^2} + \frac{3\sigma^2(n-1)^2}{k(k+1)(2k+1)d(\mathcal{X})^2}, \end{aligned} \tag{5}$$

where $\mathcal{B} = \sup_{x\in\mathcal{X}} |m^{(2)}(x)|$. Setting the derivative of (5) w.r.t. k to zero yields

$$3\mathcal{B}^2 d(\mathcal{X})^4 k^3 (1+k)^3 (1+2k+2k^2) = 8(1+8k+18k^2+12k^3)(n-1)^4\sigma^2. \tag{6}$$

Solving (6), with the constraint that $k > 0$, can be done by means of any root finding algorithm and will result in the value k for which the MSE is lowest. However, a much simpler rule of thumb and without much loss of accuracy is obtained by only considering the highest order terms yielding

$$k = \left(\frac{16\sigma^2}{\mathcal{B}^2 d(\mathcal{X})^4}\right)^{1/5} n^{4/5}.$$

The above quantity contains some unknown quantities and need to be estimated. The error variance σ^2 can be estimated by means of Hall's $\sqrt{n}$-consistent estimator (Hall et al., 1990)

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

For the second unknown quantity $\mathcal{B}$ one can use the local polynomial regression estimate of order $p = 3$ leading to the following (rough) estimate of the second derivative $\hat{m}^{(2)}(x_0) = 2\hat{\beta}_2$ (see also Section 1). Consequently, a rule of thumb selector for k is given by

$$\hat{k} = \left(\frac{16\hat{\sigma}^2}{(\sup_{x_0\in\mathcal{X}} |\hat{m}^{(2)}(x_0)|)^2 d(\mathcal{X})^4}\right)^{1/5} n^{4/5}. \tag{7}$$

The result of the rule of thumb (7) is a value for k which is real. In practice we round the obtained k value closest to the next integer value. As an alternative, one could also consider cross-validation or complexity criteria in order to find an optimal value for k.

2.2 Behavior At The Boundaries

Recall that for the boundary region ($2 \leq i \leq k$ and $n-k+1 \leq i \leq n-1$) the weights in the derivative (4) and the range of the sum are slightly modified. Such a modification allows for an automatic bias correction at the boundaries. This can be seen as follows. Let the first $(q+1)$ derivatives of m be continuous on $\mathcal{X}$. Then a Taylor series of m in a neighborhood of x_i yields

$$m(x_{i+j}) = m(x_i) + \sum_{l=1}^{q}\frac{1}{l!}\left(\frac{jd(\mathcal{X})}{n-1}\right)^l m^{(l)}(x_i) + O\big((j/n)^{q+1}\big)$$

and

$$m(x_{i-j}) = m(x_i) + \sum_{l=1}^{q}\frac{1}{l!}\left(\frac{-jd(\mathcal{X})}{n-1}\right)^l m^{(l)}(x_i) + O\big((j/n)^{q+1}\big).$$

From the above series it follows that

$$\begin{aligned}\mathbf{E}(Y_i^{(1)}) &= \sum_{j=1}^{k} w_j \frac{m(x_{i+j}) - m(x_{i-j})}{x_{i+j} - x_{i-j}} \\ &= \frac{n-1}{2d(\mathcal{X})} \sum_{j=1}^{k} w_j \frac{\sum_{l=1}^{q} \frac{1}{l!} \left(\frac{jd(\mathcal{X})}{n-1}\right)^l m^{(l)}(x_i) - \sum_{l=1}^{q} \frac{1}{l!} \left(\frac{-jd(\mathcal{X})}{n-1}\right)^l m^{(l)}(x_i) + O\big((j/n)^{q+1}\big)}{j}.\end{aligned}$$

By noticing that all even orders of the derivative cancel out, the previous result can be written as

$$\begin{aligned}\mathbf{E}(Y_i^{(1)}) &= \frac{n-1}{2d(\mathcal{X})} \sum_{j=1}^{k} \frac{w_j}{j} \left[\frac{2jd(\mathcal{X})}{n-1} m'(x_i) + \sum_{l=3,5,\ldots}^{q} \frac{2}{l!} \left(\frac{jd(\mathcal{X})}{n-1}\right)^l m^{(l)}(x_i) + O\big((j/n)^{q+1}\big) \right] \\ &= m'(x_i) \sum_{j=1}^{k} w_j + \sum_{l=3,5,\ldots}^{q} m^{(l)}(x_i) \sum_{j=1}^{k} \frac{w_j}{l!} \frac{j^{l-1} d(\mathcal{X})^{l-1}}{(n-1)^{l-1}} + O\big((j/n)^q\big).\end{aligned}$$

For $2 \leq i \leq k$, the sum in the first term is not equal to 1. This immediately follows from the definition of the derivative in (4). Therefore, the length of the sum k has to be replaced with $k(i) = i-1$. Let $0 \leq \kappa = \sum_{j=1}^{k(i)} w_j < 1$ for $2 \leq i \leq k$. Then, the bias of the derivative (4) is given by

$$\text{bias}(Y_i^{(1)}) = (\kappa - 1) m'(x_i) + \sum_{l=3,5,\ldots}^{q} m^{(l)}(x_i) \sum_{j=1}^{k} \frac{w_j}{l!} \frac{j^{l-1} d(\mathcal{X})^{l-1}}{(n-1)^{l-1}} + O\big(n^{-q/5}\big),$$

where $\sum_{j=1}^{k} \frac{w_j}{l!} \frac{j^{l-1} d(\mathcal{X})^{l-1}}{(n-1)^{l-1}} = O(n^{-(l-1)/5})$ since $k = O(n^{4/5})$. However, in order to obtain an automatic bias correction at the boundaries, we can make $\kappa = 1$ by normalizing the sum leading to the following estimator

$$Y_i^{(1)} = \sum_{j=1}^{k(i)} \frac{w_j}{\sum_{j=1}^{k(i)} w_j} \left(\frac{Y_{i+j} - Y_{i-j}}{x_{i+j} - x_{i-j}} \right) \tag{8}$$

at the boundaries. Also notice that the bias at the boundaries is of the same order as in the interior.

Unfortunately, this bias correction comes at a prize, that is, increased variance at the boundaries. The variance of (8), for $k(i) = i-1$, is given by

$$\mathbf{Var}(Y_i^{(1)}) = \frac{\sigma^2 (n-1)^2}{2d(\mathcal{X})^2} \sum_{j=1}^{k(i)} \frac{w_j^2}{\left(\sum_{j=1}^{k(i)} w_j\right)^2} \frac{1}{j^2} = \frac{3\sigma^2 (n-1)^2}{d(\mathcal{X})^2} \frac{1}{i(i-1)(2i-1)}.$$

Then, at the boundary (for $2 \leq i \leq k$), it follows that an upper bound for the variance is given by

$$\mathbf{Var}(Y_i^{(1)}) \leq \frac{\sigma^2 (n-1)^2}{2d(\mathcal{X})^2}$$

and a lower bound by

$$\begin{aligned}\mathbf{Var}(Y_i^{(1)}) &\geq \frac{3\sigma^2 (n-1)^2}{d(\mathcal{X})^2} \frac{1}{k(k-1)(2k-1)} \\ &\geq \frac{3\sigma^2 (n-1)^2}{d(\mathcal{X})^2} \frac{1}{k(k+1)(2k+1)}.\end{aligned}$$

Hence, the variance will be largest (but limited) for $i = 2$ and will decrease for growing i till $i = k$. Also, from the last inequality it follows that variance at the boundaries will always be larger or equal than the variance of the interior. An analogue calculation shows the same result for $n-k+1 \leq i \leq n-1$ by setting $k(i) = n-i$.

3. Higher Order Empirical Derivatives

In this section, we generalize the idea of first order empirical derivatives to higher order derivatives. Let q denote the order of the derivative and assume further that $q \geq 2$, then higher order empirical derivatives can be defined inductively as

$$Y_i^{(l)} = \sum_{j=1}^{k_l} w_{j,l} \cdot \left(\frac{Y_{i+j}^{(l-1)} - Y_{i-j}^{(l-1)}}{x_{i+j} - x_{i-j}} \right) \qquad \text{with} \quad l \in \{2, \ldots, q\}, \tag{9}$$

where $k_1, k_2, \ldots, k_q$ are positive integers (not necessary equal), the weights at each level l sum up to one and $Y_i^{(0)} = Y_i$ by definition. As with the first order empirical derivative, a boundary issue arises with expression (9) when $i < \sum_{l=1}^{q} k_l + 1$ or $i > n - \sum_{l=1}^{q} k_l$. Similar to (4), a boundary correction can be used. Although, the qth order derivatives are linear in the weights at level q, they are not linear in the weights at all levels. As such, no simple formulas for variance minimizing weights exist. Fortunately, simple weight sequences exist which control the asymptotic bias and variance quite well assuming that $k_1, \ldots, k_q$ increase appropriately with n (see Theorem 4).

Theorem 4 *Assume model* (1) *holds with equispaced design and let* $\sum_{j=1}^{k_l} w_{j,l} = 1$. *Further assume that the first* $(q+1)$ *derivatives of* m *are continuous on the interval* $\mathcal{X}$. *Assume that there exist* $\lambda \in (0,1)$ *and* $c_l \in (0, \infty)$ *such that* $k_l n^{-\lambda} \to c_l$ *for* $n \to \infty$ *and* $l \in \{1, 2, \ldots, q\}$. *Further, assume that*

$$w_{j,1} = \frac{6j^2}{k_1(k_1+1)(2k_1+1)} \quad \textit{for} \quad j = 1, \ldots, k_1,$$

and

$$w_{j,l} = \frac{2j}{k_l(k_l+1)} \quad \textit{for} \quad j = 1, \ldots, k_l \quad \textit{and} \quad l \in \{2, \ldots, q\}.$$

Then the asymptotic bias and variance of the empirical q*th order derivative are given by*

$$\text{bias}(Y_i^{(q)}) = O(n^{\lambda - 1}) \quad \textit{and} \quad \mathbf{Var}(Y_i^{(q)}) = O(n^{2q - 2\lambda(q+1/2)})$$

uniformly for $\sum_{l=1}^{q} k_q + 1 < i < n - \sum_{l=1}^{q} k_q$.
Proof: see Appendix C. ■

An interesting consequence of Theorem 4 is that the order of the bias of the empirical derivative estimator does not depend on the order of the derivative q. The following two corollaries are a direct consequence of Theorem 4. Corollary 5 states that the L_2 rate of convergence (and L_1 rate) will be slower for increasing orders of derivatives q, that is, higher order derivatives are progressively more difficult to estimate. Corollary 5 suggests that the MSE of the qth order empirical derivative will tend to zero for $\lambda \in (\frac{2q}{2q+1}, 1)$ prescribing, for example, $k_q = O(n^{2(q+1)/(2q+3)})$. Similar results can be obtained for the MAD. Corollary 6 proves L_2 and L_1 consistency.

Corollary 5 *Under the assumptions of Theorem 4, for the weight sequences defined in Theorem 4, the asymptotic mean squared error and asymptotic mean absolute deviation are given by*

$$\mathbf{E}(Y_i^{(q)} - m^{(q)}(x_i))^2 = O(n^{2(\lambda-1)} + n^{2q-2\lambda(q+1/2)}) \quad \textit{and} \quad \mathbf{E}\,|Y_i^{(q)} - m^{(q)}(x_i)| = O(n^{\lambda-1} + n^{q-\lambda(q+1/2)}).$$

Corollary 6 *Under the assumptions of Theorem 4, for the weight sequences defined in Theorem 4 and* $\lambda \in (\frac{2q}{2q+1}, 1)$*, it follows that*

$$\mathbf{E}(Y_i^{(q)} - m^{(q)}(x_i))^2 \to 0 \quad \textit{and} \quad \mathbf{E}\,|Y_i^{(q)} - m^{(q)}(x_i)| \to 0, \quad n \to \infty.$$

4. Bandwidth Selection For Correlated Data

From (4), it is clear that for the newly generated data set the i.i.d. assumption is no longer valid since it is a weighted sum of differences of the original data set. In such cases, it is known that data-driven bandwidth selectors and plug-ins break down (Opsomer et al., 2001; De Brabanter et al., 2011). In this paper we extend the idea of De Brabanter et al. (2011) and develop a factor rule based on bimodal kernels to determine the bandwidth. They showed, under mild conditions on the kernel function and for equispaced design, that by using a kernel satisfying $K(0) = 0$ the correlation structure is removed without any prior knowledge about its structure. Further, they showed that bimodal kernels introduce extra bias and variance yielding in a slightly wiggly estimate. In what follows we develop a relation between the bandwidth of a unimodal kernel and the bandwidth of a bimodal kernel. Consequently, the estimate based on this bandwidth will be smoother than the one based on a bimodal kernel.

Assume the following model for the qth order derivative

$$Y^{(q)}(x) = m^{(q)}(x) + \varepsilon$$

and assume that m has two continuous derivatives. Further, let $\mathbf{Cov}(\varepsilon_i, \varepsilon_{i+l}) = \gamma_l < \infty$ for all l and assume that $\sum_{l=1}^{\infty} l|\gamma_l| < \infty$. Then, if $h \to \infty$ and $nh \to \infty$ as $n \to \infty$, the bandwidth h that minimizes the mean integrated squared error (MISE) of the local polynomial regression estimator (3) with p odd under correlation is given by (Simonoff, 1996; Fan and Gijbels, 1996)

$$\hat{h} = C_p(K) \left[\frac{(\sigma^2 + 2\sum_{l=1}^{\infty} \gamma_l)\, d(\mathcal{X})}{\int \{m^{(p+1)}(u)\}^2 \, du} \right]^{1/(2p+3)} n^{-1/(2p+3)}, \tag{10}$$

where

$$C_p(K) = \left[\frac{\{(p+1)!\}^2 \int K_p^{\star 2}(u)\, du}{2(p+1)\{\int u^{p+1} K_p^{\star}(u)\, du\}^2} \right]^{1/(2p+3)}$$

and $K_p^{\star}$ denotes the equivalent kernel defined as

$$K_p^{\star}(u) = (1\ 0\ \cdots\ 0) \begin{pmatrix} \mu_0 & \mu_1 & \cdots & \mu_p \\ \mu_1 & \mu_2 & \cdots & \mu_{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ \mu_p & \mu_{p+1} & \cdots & \mu_{2p} \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ u \\ \vdots \\ u^p \end{pmatrix} K(u),$$

with $\mu_j = \int u^j K(u)\,du$. Since the bandwidth h_b based on a symmetric bimodal kernel $\overline{K}$ has a similar expression as (10) for a unimodal kernel, one can express h as a function of h_b resulting into a factor method. It is easily verified that

$$\hat{h} = C_p(K,\overline{K})\hat{h}_b,$$

where

$$C_p(K,\overline{K}) = \left[\frac{\int K_p^{\star 2}(u)\,du\,\{\int u^{p+1}\overline{K}_p^{\star}(u)\,du\}^2}{\int \overline{K}_p^{\star 2}(u)\,du\,\{\int u^{p+1}K_p^{\star}(u)\,du\}^2}\right]^{1/(2p+3)}.$$

The factor $C_p(K,\overline{K})$ is easy to calculate and Table 1 lists some of these factors for different unimodal kernels and for various odd orders of polynomials p. We take $\overline{K}(u) = (2/\sqrt{\pi})u^2\exp(-u^2)$ as bimodal kernel.

p	Gaussian	Uniform	Epanechnikov	Triangular	Biweight	Triweight
1	1.16231	2.02248	2.57312	2.82673	3.04829	3.46148
3	1.01431	2.45923	2.83537	2.98821	3.17653	3.48541
5	0.94386	2.79605	3.09301	3.20760	3.36912	3.62470

Table 1: The factor $C_p(K,\overline{K})$ for different unimodal kernels and for various odd orders of polynomials p with $\overline{K}(u) = (2/\sqrt{\pi})u^2\exp(-u^2)$ as bimodal kernel.

5. Simulations

In what follows, we evaluate the proposed method for derivative estimation with several other methods used in the literature.

5.1 First Order Derivative Estimation

We evaluate the proposed method for derivative estimation with several other methods used in the literature, that is, via the local slope in local polynomial regression with $p = 3$ (*R* package `locpol` (Cabrera, 2009)) and penalized smoothing splines (*R* package `pspline` (Ramsey and Ripley, 2010)). For the latter we have used quintic splines (Newell and Einbeck, 2007) to estimate the first order derivative. All smoothing parameters were determined by weighted generalized cross-validation (WGCV$^{(q)}$) defined as

$$\text{WGCV}^{(q)} = \frac{1}{n}\sum_{i=1}^{n} s_i\left(\frac{Y_i^{(q)} - \hat{m}_n^{(q)}(x_i)}{1-\text{trace}(L)/n}\right)^2,$$

with $s_i = \mathbf{1}\{\sum_{l=1}^{q} k_l + 1 \le i \le n - \sum_{l=1}^{q} k_l\}$ and let L be the smoother matrix of the local polynomial regression estimate. The Gaussian kernel has been used for all kernel methods. The proposed method uses $\overline{K}(u) = (2/\sqrt{\pi})u^2\exp(-u^2)$ as bimodal kernel. The corresponding sets of bandwidths of the bimodal kernel h_b were $\{0.04, 0.045, \ldots, 0.095\}$ and k_1 was determined in each run by (7). Consider the following two functions

$$m(x) = \sin^2(2\pi x) + \log(4/3 + x) \quad \text{for} \quad x \in [-1,1] \tag{11}$$

and

$$m(x) = 32e^{-8(1-2x)^2}(1-2x) \quad \text{for} \quad x \in [0,1], \tag{12}$$

In a first simulation we show a typical result for the first order derivative ($q = 1$) of (11) and (12), its first order empirical derivative (see Figure 2). The data sets are of size $n = 1000$ and are generated from model (1) with $e \sim N(0,\sigma^2)$ for $\sigma = 0.03$ (regression function (11)) and $\sigma = 0.1$ (regression function (12)). To smooth the noisy derivative data we have chosen a local polynomial regression estimate of order $p = 3$. For the Monte Carlo study, we constructed data sets size with $n = 500$ and

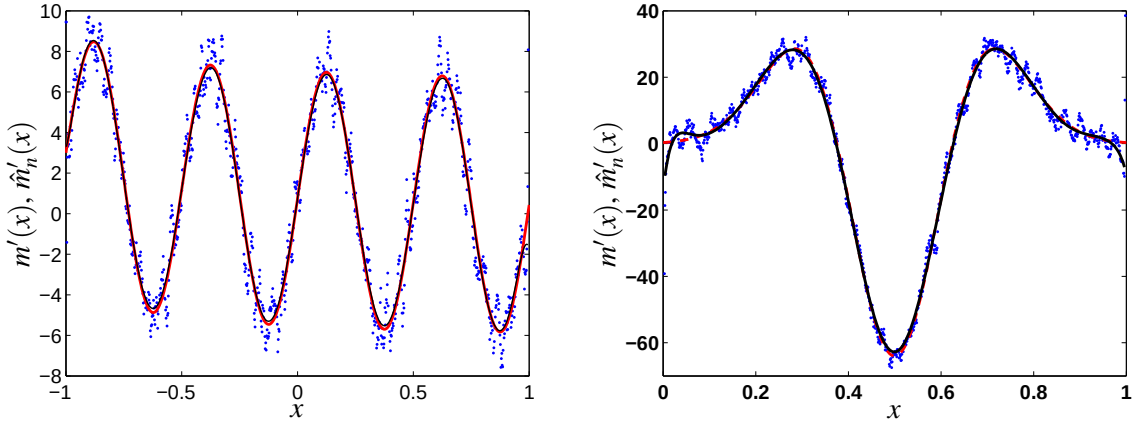


Figure 2: Illustration of the noisy empirical first order derivative (data points), smoothed empirical first order derivative based on a local polynomial regression estimate of order $p = 3$ (bold line) and true derivative (bold dashed line). (a) First order derivative of regression function (11) with $k_1 = 7$; (b) First order derivative of regression function (12) with $k_1 = 12$.

generated the function

$$m(x) = \sqrt{x(1-x)}\sin\left(\frac{2.1\pi}{x+0.05}\right) \quad \text{for} \quad x \in [0.25,1]$$

100 times according to model (1) with $e \sim N(0,\sigma^2)$ and $\sigma = 0.1$. As measure of comparison we chose the adjusted mean absolute error defined as

$$\text{MAEadjusted} = \frac{1}{481}\sum_{i=10}^{490} |\hat{m}'_n(x_i) - m'(x_i)|.$$

This criterion was chosen to ignore boundary effects in the estimation for the three methods. The result of the Monte Carlo study for (12) is given in Figure 3. From the Monte Carlo experiment, it is clear that all three methods yield similar results and no method supersedes the other.

5.2 Second Order Derivative Estimation

As before, all smoothing parameters were determined by weighted generalized cross-validation (WGCV$^{(q)}$) for $q = 2$. A typical result for the second order derivative ($q = 2$) of (11) and (12) and

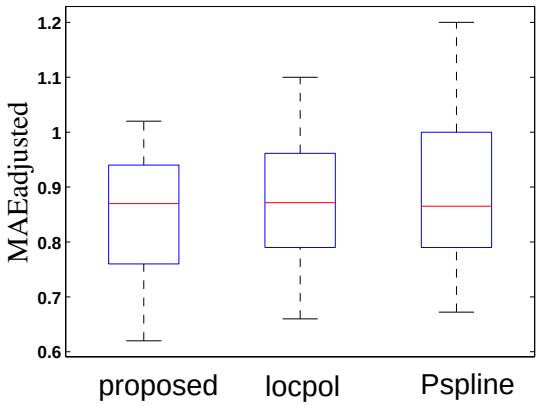


Figure 3: Result of the Monte Carlo study for the proposed method and two other well-known methods for first order derivative estimation.

its second order empirical derivative is shown in Figure 4. To smooth the noisy derivative data we have chosen a local polynomial regression estimate of order $p = 3$. The question that arises is the following: How to tune k_1 and k_2 for second order derivative estimation? Consider a set of candidate values of k_1 and k_2, for example, $\{5,\ldots,40\}$. Note that, according to Corollary 5, the order of k_q should increase with q. The size of the set is determined both by the computational time that one is willing to invest and by the maximum fraction of the observation weights $s_1,\ldots,s_n$ that one is willing to set to 0 in order to circumvent the aforementioned boundary issues. In order to have a fair comparison among the values of k_1 and k_2, one should use the same observation weights for all candidate values. Therefore, the largest value determines the weights. To choose the value k_1 and k_2 from the candidate set, we can take k_1 and k_2 that minimize $\mathrm{WGCV}^{(2)}$. A similar strategy can be used to determine k_q. We have chosen to tune k_1 according to the way described above and not via (7) because the optimal k_1 for first derivatives is not necessarily the optimal one to be used for estimating second derivatives. From the simulations, it is clear that the variance is larger for increasing q for $\lambda \in (\frac{2q}{2q+1}, 1)$ (the order of the bias remains the same). This was already confirmed by Theorem 4.

For the Monte Carlo study, we constructed data sets are of size $n = 1500$ and generated the function

$$m(x) = 8e^{-(1-5x)^3(1-7x)} \quad \text{for} \quad x \in [0, 0.5]$$

100 times according to model (1) with $e \sim N(0, \sigma^2)$ and $\sigma = 0.1$. As measure of comparison we chose the adjusted mean absolute error defined as

$$\mathrm{MAEadjusted} = \frac{1}{1401} \sum_{i=50}^{1450} |\hat{m}_n^{(2)}(x_i) - m^{(2)}(x_i)|.$$

This criterion was chosen to ignore boundary effects in the estimation. We evaluate the proposed method for derivative estimation with the local slope in local polynomial regression with $p = 5$ and penalized smoothing splines. For the latter we have used septic splines (Newell and Einbeck, 2007) to estimate the second order derivative. The result of the Monte Carlo study is shown in Figure 5. As before, all three methods perform equally well and show similar variances.

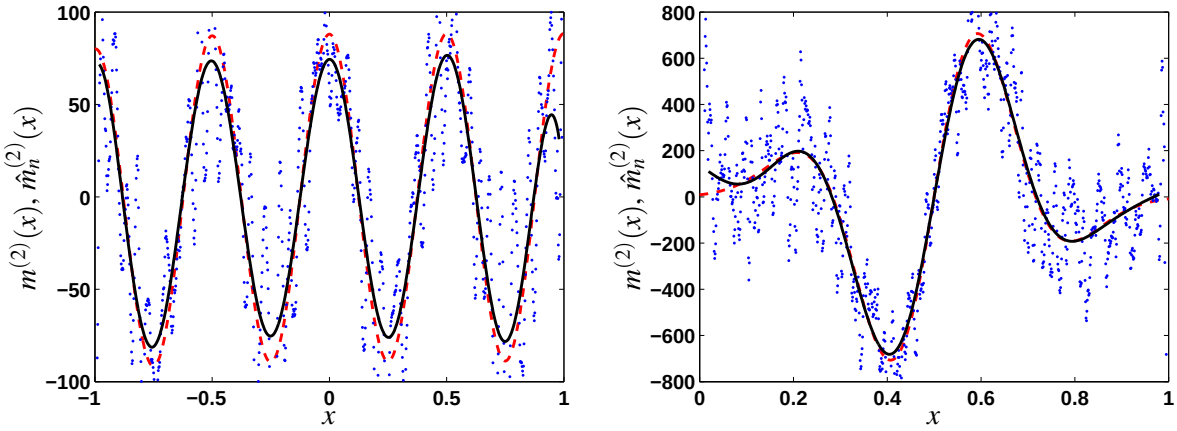


Figure 4: Illustration of the noisy empirical second order derivative (data points), smoothed empirical second order derivative based on a local polynomial regression estimate of order $p = 3$ (bold line) and true derivative (bold dashed line). (a) Second order derivative of regression function (11) with $k_1 = 6$ and $k_2 = 10$; (b) Second order derivative of regression function (12) with $k_1 = 3$ and $k_2 = 25$.

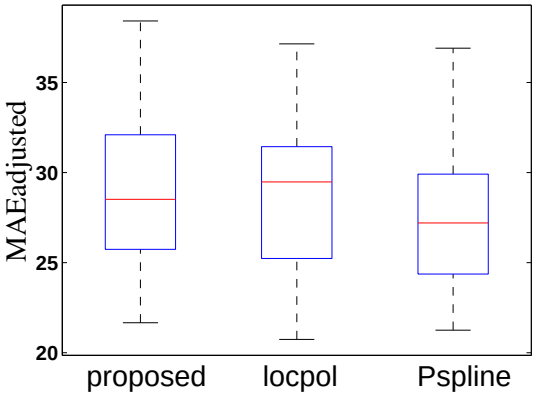


Figure 5: Result of the Monte Carlo study for the proposed method and two other well-known methods for second order derivative estimation.

6. Conclusion

In this paper we proposed a methodology to estimate derivatives nonparametrically without estimating the regression function. We derived L_1 and L_2 rates and established consistency of the estimator. The newly created data sets based on empirical derivatives are no longer independent and identically distributed (i.i.d.) random variables. In order to effectively deal with the non-i.i.d. nature of

the data, we proposed a simple factor method, based on bimodal kernels, for the local polynomial regression framework. Further, we showed that the order bias of the empirical derivative does not depend on the order of the derivative q and that slower rates of convergence are to be expected for increasing orders of derivatives q. However, our technique has also a drawback w.r.t. the design assumptions. All our results have been derived for equispaced design. In many practical applications and data coming from industrial sensors (e.g., process industry, robotics, nanoparticles, growth data) equispaced data is often available since sensors are measuring at predefined times, see, for example, Charnigo et al. (2007) and Patan (2008). However, our approach does not cover all possible applications, that is, application with inherent random design. In this case the weight sequence would depend on the design density, which in practice has to be estimated.

Acknowledgments

Kris De Brabanter is a postdoctoral researcher supported by an FWO fellowship grant. BDM is full professor at the Katholieke Universiteit Leuven, Belgium. Research supported by Onderzoeksfonds KU Leuven/Research Council KUL: GOA/11/05 Ambiorics, GOA/10/09 MaNet , CoE EF/05/006 Optimization in Engineering (OPTEC) en PFV/10/002 (OPTEC), IOF-SCORES4CHEM, several PhD/postdoc & fellow grants; Flemish Government:FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0320.08 (convex MPC), G.0558.08 (Robust MHE), G.0557.08 (Glycemia2), G.0588.09 (Brain-machine) research communities (WOG: ICCoS, ANMMM, MLDM); G.0377.09 (Mechatronics MPC) IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011);IBBT EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940), FP7-SADCO (MC ITN-264735), ERC HIGHWIND (259 166) Contract Research: AMINAL Other: Helmholtz: viCERP ACCM. IG is a full professor at the Katholieke Universiteit Leuven, Belgium. GOA/07/04 en GOA/12/014, IUAP: P6/03, FWO-project G.0328.08N. Interreg IVa 07-022-BE i-MOCCA. The scientific responsibility is assumed by its authors.

Appendix A. Proof Of Proposition 1

Using the fact that $x_{i+j} - x_{i-j} = 2j(n-1)^{-1}d(\mathcal{X})$, where $d(\mathcal{X}) := \sup\mathcal{X} - \inf\mathcal{X}$, yields

$$\begin{aligned}
\mathbf{Var}(Y_i^{(1)}) &= \mathbf{Var}\left(\sum_{j=1}^{k} w_j \cdot \left(\frac{Y_{i+j} - Y_{i-j}}{x_{i+j} - x_{i-j}}\right)\right) \\
&= \mathbf{Var}\left(\left(1 - \sum_{j=2}^{k} w_j\right)\frac{Y_{i+1} - Y_{i-1}}{x_{i+1} - x_{i-1}} + \sum_{j=2}^{k} w_j \cdot \left(\frac{Y_{i+j} - Y_{i-j}}{x_{i+j} - x_{i-j}}\right)\right) \\
&= \frac{\sigma^2(n-1)^2}{2\,d(\mathcal{X})^2}\left\{\left(1 - \sum_{j=2}^{k} w_j\right)^2 + \sum_{j=2}^{k}\frac{w_j^2}{j^2}\right\}.
\end{aligned}$$

Setting the partial derivatives to zero gives

$$2\left(1 - \sum_{j=2}^{k} w_j\right) = \frac{2w_j}{j^2}, \quad j = 2, \ldots, k,$$

and hence $j^2 w_1 = w_j$. Normalizing such that the weights sum up to one yields

$$w_j = \frac{j^2}{\sum_{i=1}^{k} i^2} = \frac{6j^2}{k(k+1)(2k+1)} \quad j = 1, \ldots, k.$$

Appendix B. Proof Of Theorem 2

Since m is twice continuously differentiable, the following Taylor expansions are valid for $m(x_{i+j})$ and $m(x_{i-j})$ round x_i:

$$m(x_{i+j}) = m(x_i) + (x_{i+j} - x_i)m'(x_i) + \frac{(x_{i+j} - x_i)^2}{2} m^{(2)}(\zeta_{i,i+j})$$

and

$$m(x_{i-j}) = m(x_i) + (x_{i-j} - x_i)m'(x_i) + \frac{(x_{i-j} - x_i)^2}{2} m^{(2)}(\zeta_{i-j,i}),$$

where $\zeta_{i,i+j} \in]x_i, x_{i+j}[$ and $\zeta_{i-j,i} \in]x_{i-j}, x_i[$. Using the above Taylor series and the fact that $x_{i+j} - x_{i-j} = 2j(n-1)^{-1} d(\mathcal{X})$ and $(x_{i+j} - x_i) = \frac{1}{2}(x_{i+j} - x_{i-j})$, it follows that the absolute value of the bias of $Y_i^{(1)}$ is given by

$$\begin{aligned}
\left| \sum_{j=1}^{k} w_j \frac{m(x_{i+j}) - m(x_{i-j})}{x_{i+j} - x_{i-j}} - m'(x_i) \right| &= \left| \sum_{j=1}^{k} w_j \frac{(x_{i+j} - x_{i-j})[m^{(2)}(\zeta_{i,i+j}) - m^{(2)}(\zeta_{i-j,i})]}{8} \right| \\
&\leq \sup_{x \in \mathcal{X}} |m^{(2)}(x)| \left| \sum_{j=1}^{k} w_j \frac{(x_{i+j} - x_{i-j})}{4} \right| \\
&= \frac{\sup_{x \in \mathcal{X}} |m^{(2)}(x)| (n-1)^{-1} d(\mathcal{X})}{2} \sum_{j=1}^{k} \frac{j^3}{\sum_{i=1}^{k} i^2} \\
&= \frac{3k(k+1) \sup_{x \in \mathcal{X}} |m^{(2)}(x)| d(\mathcal{X})}{4(n-1)(2k+1)} \\
&= O(kn^{-1})
\end{aligned}$$

uniformly over i. Using Proposition 1, the variance of $Y_i^{(1)}$ yields

$$\begin{aligned}
\mathbf{Var}(Y_i^{(1)}) &= \frac{\sigma^2 (n-1)^2}{2 d(\mathcal{X})^2} \left\{ \left(1 - \sum_{j=2}^{k} w_j \right)^2 + \sum_{j=2}^{k} \frac{w_j^2}{j^2} \right\} \\
&= \frac{\sigma^2 (n-1)^2}{2 d(\mathcal{X})^2} \sum_{j=1}^{k} \frac{w_j^2}{j^2} \\
&= \frac{\sigma^2 (n-1)^2}{2 d(\mathcal{X})^2} \sum_{j=1}^{k} \frac{36 j^2}{k^2 (k+1)^2 (2k+1)^2} \\
&= \frac{3\sigma^2 (n-1)^2}{k(k+1)(2k+1) d(\mathcal{X})^2} = O(n^2 k^{-3})
\end{aligned}$$

uniformly over i.

Appendix C. Proof of Theorem 3

Due to Chebyshev's inequality, it suffices to show that the mean squared error (MSE) goes to zero, that is,

$$\lim_{n\to\infty} \mathrm{MSE}(Y_i^{(1)}) \to 0. \tag{13}$$

Under the conditions $k \to \infty$ as $n \to \infty$ such that $n^{-1}k \to 0$ and $nk^{-3/2} \to 0$, the bias and variance go to zero (see Theorem 2). Hence, condition (13) is fulfilled.

Appendix D. Proof Of Theorem 4

The first step is to notice that there exist $\lambda \in (0,1)$ and $c_1 \in (0,\infty)$ (see Theorem 3) so that the bias and variance of the first order empirical derivative can be written as $\mathrm{bias}(Y_i^{(1)}) = O(n^{\lambda-1})$ and $\mathbf{Var}(Y_i^{(1)}) = O(n^{2-3\lambda})$ uniformly over i for $k_1 n^{-\lambda} \to c_1$ as $n \to \infty$. Next, we continue the proof by induction. For the bias, assume that the first $(q+1)$ derivatives of m are continuous on the compact interval $\mathcal{X}$. Hence, all $O(\cdot)$-terms are uniformly over i. For any $l \in \{0,1,\ldots,q\}$, a Taylor series yields

$$m^{(l)}(x_{i\pm j}) = m^{(l)}(x_i) + \sum_{p=1}^{q-l} \frac{\left(\pm\frac{jd(\mathcal{X})}{n-1}\right)^p}{p!} m^{(p+l)}(x_i) + O\left((j/n)^{q-l+1}\right). \tag{14}$$

The expected value of the first order empirical derivative is given by (see Section 2)

$$\mathbf{E}(Y_i^{(1)}) = m'(x_i) + \sum_{p=3,5,\ldots}^{q} m^{(p)}(x_i) \sum_{j=1}^{k_1} \frac{w_{j,1}}{p!} \frac{j^{p-1} d(\mathcal{X})^{p-1}}{(n-1)^{p-1}} + O\left(n^{q(\lambda-1)}\right),$$

with

$$\theta_{p,1} = \sum_{j=1}^{k_1} \frac{w_{j,1}}{p!} \frac{j^{p-1} d(\mathcal{X})^{p-1}}{(n-1)^{p-1}} = O\left(n^{(p-1)(\lambda-1)}\right),$$

for $k_1 n^{-\lambda} \to c_1$ as $n \to \infty$. Suppose that for $l \in \{2,\ldots,q\}$ and $k_l n^{-\lambda} \to c_l$, where $c_l \in (0,\infty)$, as $n \to \infty$

$$\mathbf{E}(Y_i^{(l-1)}) = m^{(l-1)}(x_i) + \sum_{p=l+1,l+3,\ldots}^{q} \theta_{p,l-1} m^{(p)}(x_i) + O\left(n^{(q-l+2)(\lambda-1)}\right) \tag{15}$$

for $\theta_{p,l-1} = O\left(n^{(p-l+1)(\lambda-1)}\right)$. We now prove that

$$\mathbf{E}(Y_i^{(l)}) = m^{(l)}(x_i) + \sum_{p=l+2,l+4,\ldots}^{q} \theta_{p,l} m^{(p)}(x_i) + O\left(n^{(q-l+1)(\lambda-1)}\right)$$

for $\theta_{p,l} = O\left(n^{(p-l)(\lambda-1)}\right)$. Using (14) and (15) yields for $\Delta = \mathbf{E}(Y_{i+j}^{(l-1)}) - \mathbf{E}(Y_{i-j}^{(l-1)})$

$$\begin{aligned}
\Delta &= m^{(l-1)}(x_{i+j}) + \sum_{p=l+1,l+3,\ldots}^{q} \theta_{p,l-1} m^{(p)}(x_{i+j}) - m^{(l-1)}(x_{i-j}) - \sum_{p=l+1,l+3,\ldots}^{q} \theta_{p,l-1} m^{(p)}(x_{i-j}) + O\left(n^{(q-l+2)(\lambda-1)}\right) \\
&= \sum_{p=1}^{q-l+1} \frac{\left(\frac{jd(\mathcal{X})}{n-1}\right)^p}{p!} m^{(p+l-1)}(x_i) + O\left((j/n)^{q-l+2}\right) \\
&\quad + \sum_{p=l+1,l+3,\ldots}^{q} \theta_{p,l-1} \left[m^{(p)}(x_i) + \sum_{s=1}^{q-p} \frac{\left(\frac{jd(\mathcal{X})}{n-1}\right)^s}{s!} m^{(p+s)}(x_i) + O\left((j/n)^{q-p+1}\right) \right] \\
&\quad - \sum_{p=1}^{q-l+1} \frac{\left(\frac{-jd(\mathcal{X})}{n-1}\right)^p}{p!} m^{(p+l-1)}(x_i) + O\left((j/n)^{q-l+2}\right) \\
&\quad - \sum_{p=l+1,l+3,\ldots}^{q} \theta_{p,l-1} \left[m^{(p)}(x_i) + \sum_{s=1}^{q-p} \frac{\left(\frac{-jd(\mathcal{X})}{n-1}\right)^s}{s!} m^{(p+s)}(x_i) + O\left((j/n)^{q-p+1}\right) \right] + O\left(n^{(q-l+2)(\lambda-1)}\right).
\end{aligned}$$

Rearranging and grouping term gives

$$\begin{aligned}
\frac{\Delta}{x_{i+j} - x_{i-j}} &= m^{(l)}(x_i) + \sum_{p=3,5,\ldots}^{q-l+1} \frac{\left(\frac{jd(\mathcal{X})}{n-1}\right)^{p-1}}{p!} m^{(p+l-1)}(x_i) + O\left((j/n)^{q-l+1}\right) \\
&\quad + \sum_{p=l+1,l+3,\ldots}^{q} \theta_{p,l-1} \left[\sum_{s=1,3,\ldots}^{q-p} \frac{\left(\frac{jd(\mathcal{X})}{n-1}\right)^{s-1}}{s!} m^{(p+s)}(x_i) + O\left((j/n)^{q-p}\right) \right] \\
&\quad + \frac{n-1}{2jd(\mathcal{X})} O\left(n^{(q-l+2)(\lambda-1)}\right).
\end{aligned}$$

Multiplying all the above terms by $w_{j,l} = \dfrac{j}{\sum_{i=1}^{k_l} i}$ and summing over $j = 1, 2, \ldots, k_l$ results in

$$\begin{aligned}
\mathbf{E}(Y_i^{(l)}) &= m^{(l)}(x_i) \\
&\quad + \sum_{j=1}^{k_l} \frac{j}{\sum_{i=1}^{k_l} i} \sum_{p=3,5,\ldots}^{q-l+1} \frac{\left(\frac{jd(\mathcal{X})}{n-1}\right)^{p-1}}{p!} m^{(p+l-1)}(x_i) && (16) \\
&\quad + \sum_{j=1}^{k_l} \frac{j}{\sum_{i=1}^{k_l} i} O\left((j/n)^{q-l+1}\right) && (17) \\
&\quad + \sum_{j=1}^{k_l} \frac{j}{\sum_{i=1}^{k_l} i} \sum_{p=l+1,l+3,\ldots}^{q} \theta_{p,l-1} \sum_{s=1,3,\ldots}^{q-p} \frac{\left(\frac{jd(\mathcal{X})}{n-1}\right)^{s-1}}{s!} m^{(p+s)}(x_i) && (18) \\
&\quad + \sum_{j=1}^{k_l} \frac{j}{\sum_{i=1}^{k_l} i} \sum_{p=l+1,l+3,\ldots}^{q} \theta_{p,l-1} O\left((j/n)^{q-p}\right) && (19) \\
&\quad + \sum_{j=1}^{k_l} \frac{j}{\sum_{i=1}^{k_l} i} \frac{n-1}{2jd(\mathcal{X})} O\left(n^{(q-l+2)(\lambda-1)}\right). && (20)
\end{aligned}$$

The terms (17), (19) and (20) all yield $O(n^{(q-l+1)(\lambda-1)})$ for $\theta_{p,l-1} = O(n^{(p-l+1)(\lambda-1)})$. Similar, the terms (16) and (18) yield $\sum_{p=l+2,l+4,\ldots}^{q} \theta_{p,l} m^{(p)}(x_i)$ for $\theta_{p,l} = O\left(n^{(p-l)(\lambda-1)}\right)$ for $k_l n^{-\lambda} \to c_l$ as $n \to \infty$. As a consequence, the bias of $Y_i^{(l)}$ is given by

$$\text{bias}(Y_i^{(l)}) = \mathbf{E}(Y_i^{(l)}) - m^{(l)}(x_i) = \sum_{p=l+2,l+4,\ldots}^{q} \theta_{p,l} m^{(p)}(x_i) + O(n^{(\lambda-1)}) = O(n^{\lambda-1}).$$

For the variance, we proceed in a similar way. Note that $\mathbf{Var}(Y_i^{(1)}) = O(n^{2-3\lambda})$ uniformly over i. Assume that $\mathbf{Var}(Y_i^{(l-1)}) = O(n^{2(l-1)-2\lambda(l-1/2)})$ uniformly over i for $l \in \{2, 3, \ldots, q\}$. The proof will be complete if we show that $\mathbf{Var}(Y_i^{(l)}) = O(n^{2l-2\lambda(l+1/2)})$. The variance of $Y_i^{(l)}$ is given by

$$\begin{aligned}
\mathbf{Var}(Y_i^{(l)}) &= \frac{(n-1)^2}{4d(\mathcal{X})^2} \mathbf{Var}\left(\sum_{j=1}^{k_l} \frac{w_{j,l}}{j}\left(Y_{i+j}^{(l-1)} - Y_{i-j}^{(l-1)}\right)\right) \\
&\leq \frac{(n-1)^2}{2d(\mathcal{X})^2}\left[\mathbf{Var}\left(\sum_{j=1}^{k_l} \frac{w_{j,l}}{j} Y_{i+j}^{(l-1)}\right) + \mathbf{Var}\left(\sum_{j=1}^{k_l} \frac{w_{j,l}}{j} Y_{i-j}^{(l-1)}\right)\right].
\end{aligned}$$

For $a_j \in \mathbb{N} \setminus \{0\}, j = 1, \ldots, k_l$, the variance is upperbounded by

$$\mathbf{Var}(Y_i^{(l)}) \leq \frac{(n-1)^2}{d(\mathcal{X})^2}\left(\sum_{j=1}^{k_l} a_j \frac{w_{j,l}^2}{j^2}\right) O(n^{2(l-1)-2\lambda(l-1/2)}).$$

As in the proof of the bias, the choice of the weights become clear. If we choose $w_{j,l} = \frac{j}{\sum_{i=1}^{k_l} i}$ for $l \geq 2$ then $\sum_{j=1}^{k_l} a_j \frac{w_{j,l}^2}{j^2} = O(n^{-2\lambda})$. Then, for $k_l n^{-\lambda} \to c_l$ as $n \to \infty$, it readily follows that $\mathbf{Var}(Y_i^{(l)}) = O(n^{2l-2\lambda(l+1/2)})$.

References

J.L.O. Cabrera. *locpol: Kernel Local Polynomial Regression*, 2009. URL `http://CRAN.R-project.org/package=locpol`. R package version 0.4-0.

R. Charnigo, M. Francoeur, M.P. Mengüç, A. Brock, M. Leichter, and C. Srinivasan. Derivatives of scattering profiles: tools for nanoparticle characterization. *J. Opt. Soc. Am. A*, 24(9):2578–2589, 2007.

P. Chaudhuri and J.S. Marron. SiZer for exploration of structures in curves. *J. Amer. Statist. Assoc.*, 94(447):807–823, 1999.

C.K. Chu and J.S. Marron. Comparison of two bandwidth selectors with dependent errors. *Ann. Statist.*, 19(4):1906–1918, 1991.

K. De Brabanter, J. De Brabanter, J.A.K. Suykens, and B. De Moor. Kernel regression in the presence of correlated errors. *J. Mach. Learn. Res.*, 12:1955–1976, 2011.

M. Delecroix and A.C. Rosa. Nonparametric estimation of a regression function and its derivatives under an ergodic hypothesis. *J. Nonparametr. Stat.*, 6(4):367–382, 2007.

R.L. Eubank and P.L. Speckman. Confidence bands in nonparametric regression. *J. Amer. Statist. Assoc.*, 88(424):1287–1301, 1993.

J. Fan and I. Gijbels. Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. R. Stat. Soc. Ser. B*, 57(2):371–394, 1995.

J. Fan and I. Gijbels. *Local Polynomial Modeling and Its Applications*. Chapman & Hall, 1996.

T. Gasser and H.-G. Müller. Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.*, 11(3):171–185, 1984.

I. Gijbels and A.-C. Goderniaux. Data-driven discontinuity detection in derivatives of a regression function. *Communications in Statistics–Theory and Methods*, 33:851–871, 2004.

L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

P. Hall, J.W. Kay, and D.M. Titterington. Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, 77(3):521–528, 1990.

W. Härdle. *Applied Nonparametric Regression*. Cambridge University Press, 1990.

W. Härdle and T. Gasser. On robust kernel estimation of derivatives of regression functions. *Scand. J. Statist.*, 12(3):233–240, 1985.

A. Iserles. *A First Course in the Numerical Analysis of Differential Equations*. Cambridge University Press, 1996.

R. Jarrow, D. Ruppert, and Y. Yu. Estimating the term structure of corporate debt with a semiparametric penalized spline model. *J. Amer. Statist. Assoc.*, 99(465):57–66, 2004.

H.-G. Müller. *Nonparametric Regression Analysis of Longitudinal Data*. Springer-Verlag, 1988.

H.-G. Müller, U. Stadtmüller, and T. Schmitt. Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika*, 74(4):743–749, 1987.

J. Newell and J. Einbeck. A comparative study of nonparametric derivative estimators. *Proc. of the 22nd International Workshop on Statistical Modelling*, 2007.

J. Opsomer, Y. Wang, and Y. Yang. Nonparametric regression with correlated errors. *Statist. Sci.*, 16(2):134–153, 2001.

C. Park and K.-H. Kang. SiZer analysis for the comparison of regression curves. *Comput. Statist. Data Anal.*, 52(8):3954–3970, 2008.

K. Patan. *Artificial Neural Networks for the Modelling and Fault Diagnosis of Technical Processes*. Springer-Verlag, 2008.

J. Ramsay. Derivative estimation. StatLib – S-News, Thursday, March 12, 1998: `http://www.math.yorku.ca/Who/Faculty/Monette/S-news/0556.html`, 1998.

J.O. Ramsay and B.W. Silverman. *Applied Functional Data Analysis*. Springer-Verlag, 2002.

J. Ramsey and B. Ripley. *pspline: Penalized Smoothing Splines*, 2010. URL `http://CRAN.R-project.org/package=pspline`. R package version 1.0-14.

V. Rondonotti, J.S. Marron, and C. Park. SiZer for time series: A new approach to the analysis of trends. *Electron. J. Stat.*, 1:268–289, 2007.

D. Ruppert and M.P. Wand. Multivariate locally weighted least squares regression. *Ann. Statist.*, 22 (3):1346–1370, 1994.

J.S. Simonoff. *Smoothing Methods in Statistics*. Springer-Verlag, 1996.

C. Stone. Additive regression and other nonparametric models. *Ann. Statist.*, 13(2):689–705, 1985.

A.B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

G. Wahba and Y. Wang. When is the optimal regularization parameter insensitive to the choice of loss function? *Comm. Statist. Theory Methods*, 19(5):1685–1700, 1990.

M.P. Wand and M.C. Jones. *Kernel Smoothing*. Chapman & Hall, 1995.

S. Zhou and D.A. Wolfe. On derivative estimation in spline regression. *Statist. Sinica*, 10(1): 93–108, 2000.