Smoothed Nonparametric Derivative Estimation Using Weighted Difference Quotients

Paper Author: *Yu Liu* and *Kris De Brabanter*

Presented By **Yikun Zhang**

Department of Statistics,
University of Washington

June 16, 2023

Department of
STATISTICS

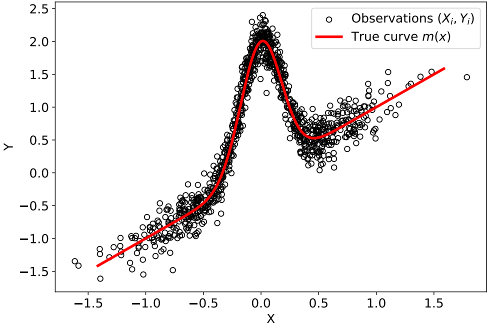
UNIVERSITY *of* WASHINGTON

Introduction

Data setting:

$$Y_i = m(X_i) + e_i, \quad \text{with} \quad X_i \in [a,b] \subset \mathbb{R} \quad \text{for} \quad i = 1, \ldots, n,$$

where e_i is independent of X_i and $\mathbb{E}(e_i) = 0, \mathrm{Var}(e_i) = \sigma_e^2 < \infty$.

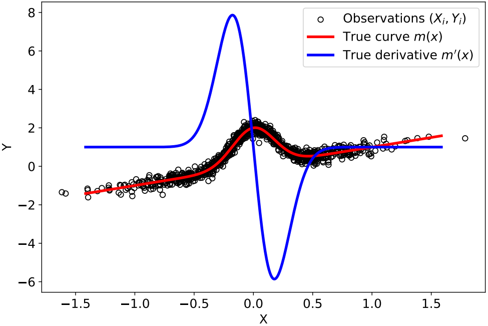


Problem Setting

Data setting:

$$Y_i = m(X_i) + e_i, \quad \text{with} \quad X_i \in [a,b] \subset \mathbb{R} \quad \text{for} \quad i = 1, ..., n,$$

where e_i is independent of X_i and $\mathbb{E}(e_i) = 0, \mathrm{Var}(e_i) = \sigma_e^2 < \infty$.



Question: How do we estimate $m^{(1)}(x) = \lim_{h \to 0} \frac{m(x+h) - m(x)}{h}$ from the data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$?

Motivations of Derivative Estimation

Estimating $m^{(1)}(x)$ has significant impacts within and beyond **Statistics**:

- Explore the structures in curves (Chaudhuri and Marron, 1999) or the changing trend in time series (Rondonotti et al., 2007).
- Correct the bias term for a regression estimator in order to conduct valid statistical inference (Eubank and Speckman, 1993; Calonico et al., 2018; Cheng and Chen, 2019).

Motivations of Derivative Estimation

Estimating $m^{(1)}(x)$ has significant impacts within and beyond **Statistics**:

- Explore the structures in curves (Chaudhuri and Marron, 1999) or the changing trend in time series (Rondonotti et al., 2007).
- Correct the bias term for a regression estimator in order to conduct valid statistical inference (Eubank and Speckman, 1993; Calonico et al., 2018; Cheng and Chen, 2019).
- **Economics:** Quantify the relations between Marginal Propensity to Consume and other labor factors (Haavelmo, 1947).
- **Biomechanics:** Facilitate the kinematic analysis of human movements (Woltring, 1985).

Good news: The data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ from the model

$$Y = m(X) + e$$

are generally available in practice.

Challenges of Derivative Estimation

Good news: The data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ from the model

$$Y = m(X) + e$$

are generally available in practice.

Bad news: We don't have any data directly from the derivative (De Brabanter et al., 2013), *e.g.*, from the model

$$Y^{(1)} = m^{(1)}(X) + e'.$$

Challenge: We need to extract the derivative information from the original data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$.

Existing Methods for Estimating the Derivatives

Parametric methods: Assume $m(x)$ lying in some parametric family $\{g(x;\theta) : \theta \in \Theta\}$ and fit

$$\widehat{\theta} \in \arg\min_{\theta\in\Theta} \sum_{i=1}^{n} \left[Y_i - g(X_i;\theta)\right]^2 \quad \Longrightarrow \quad \widehat{m}^{(1)}(x) = g^{(1)}(x;\widehat{\theta}).$$

- *Drawback:* It is difficult to posit a correct family $\{g(x;\theta) : \theta \in \Theta\}$.

Nonparametric methods: Make no parametric model assumptions on $m(x)$ and estimate $m^{(1)}(x)$ from the data $\mathcal{D}$.

- **Smoothing splines:** Zhou and Wolfe (2000) considered estimating $m^{(q)}(x)$ for $q \geq 1$ through the derivative of smoothing splines (*i.e.*, piecewise polynomial curves).

Nonparametric Methods: Splines and Kernel Methods

- **Smoothing splines:** Zhou and Wolfe (2000) considered estimating $m^{(q)}(x)$ for $q \geq 1$ through the derivative of smoothing splines (*i.e.*, piecewise polynomial curves).

- **Gasser-Müller estimator:** Gasser and Müller (1984) leveraged the derivatives of Gasser-Müller regression estimator as:

$$\widehat{m}_{h,GM}^{(q)}(x) = \frac{1}{h^{q+1}} \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K^{(q)}\left(\frac{x-u}{h}\right) du,$$

where $s_i = \frac{X_{(i)}+X_{(i+1)}}{2}$ with $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$, K is the kernel function, and $h > 0$ is the bandwidth parameter.

- **Smoothing splines:** Zhou and Wolfe (2000) considered estimating $m^{(q)}(x)$ for $q \geq 1$ through the derivative of smoothing splines (*i.e.*, piecewise polynomial curves).

- **Gasser-Müller estimator:** Gasser and Müller (1984) leveraged the derivatives of Gasser-Müller regression estimator as:

$$\widehat{m}_{h,GM}^{(q)}(x) = \frac{1}{h^{q+1}} \sum_{i=1}^{n} Y_i \int_{s_{i-1}}^{s_i} K^{(q)}\left(\frac{x-u}{h}\right) du,$$

where $s_i = \frac{X_{(i)} + X_{(i+1)}}{2}$ with $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$, K is the kernel function, and $h > 0$ is the bandwidth parameter.

- **Nadaraya-Watson estimator:** Mack and Müller (1989) proposed a Nadaraya-Watson-typed derivative estimator as:

$$\widehat{m}_{h,NW}^{(q)}(x) = \frac{1}{nh^{q+1}} \sum_{i=1}^{n} \frac{Y_i \cdot K^{(q)}\left(\frac{x-X_i}{h}\right)}{\widehat{f}_v(X_i)},$$

where $\widehat{f}_v$ is a kernel density estimator for the density of covariate X.

Local polynomial regression (Fan and Gijbels, 1996) solves the weighted least-square problem at each query point x as:

$$\widehat{\boldsymbol{\beta}}(x) \equiv \left(\widehat{\beta}_0(x), ..., \widehat{\beta}_p(x)\right)^T$$
$$= \underset{\boldsymbol{\beta}(x) \in \mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n} \left[Y_i - \sum_{j=0}^{p} \beta_j(x) \cdot (X_i - x)^j\right]^2 K\left(\frac{X_i - x}{h}\right),$$

where $K : \mathbb{R} \to [0, \infty)$ is a symmetric kernel function and $h > 0$ is the bandwidth parameter.

- It estimates the q-th order derivative $m^{(q)}(x)$ as:

$$\widehat{m}^{(q)}(x) = q!\, \widehat{\beta}_q(x)$$

  for any $q \leq p$.

We order the data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ according to the increasing order of $X_i, i = 1, ..., n$:

$$Y_i = m(X_{(i)}) + e_i, \quad i = 1, ..., n.$$

The first-order difference quotients are defined as (Müller et al., 1987; Härdle, 1990):

$$\widehat{q}^{(1)}(X_{(i)}) = \frac{Y_i - Y_{i-1}}{X_{(i)} - X_{(i-1)}}, \quad i = 2, ..., n.$$

We order the data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ according to the increasing order of $X_i, i = 1, ..., n$:

$$Y_i = m(X_{(i)}) + e_i, \quad i = 1, ..., n.$$

The first-order difference quotients are defined as (Müller et al., 1987; Härdle, 1990):

$$\widehat{q}^{(1)}(X_{(i)}) = \frac{Y_i - Y_{i-1}}{X_{(i)} - X_{(i-1)}}, \quad i = 2, ..., n.$$

Drawback: The difference quotient $\widehat{q}^{(1)}(X_{(i)})$ estimates $m^{(1)}(X_{(i)})$ with the conditional variance as:

$$\mathrm{Var}\left[\widehat{q}^{(1)}(X_{(i)}) \middle| X_{(i-1)}, X_{(i)}\right] = O_P\left(n^2\right).$$

Nonparametric Methods: Weighted Difference Quotients

To reduce the variance, Iserles (2009); Charnigo et al. (2011) considered

$$\widehat{Y}_i^{(1)} \equiv \widehat{Y}_i^{(1)}(X_{(i)}) = \sum_{j=1}^{k} w_{i,j} \left(\frac{Y_{i+j} - Y_{i-j}}{X_{(i+j)} - X_{(i-j)}} \right)$$

for $k+1 \leq i \leq n-k$ and $k \leq \frac{(n-1)}{2}$.

- The weights with $\sum_{j=1}^{k} w_{i,j} = 1$ are chosen to minimize the conditional variance $\mathrm{Var}\left(\widehat{Y}_i^{(1)} | X_{(1)}, ..., X_{(n)}\right)$.
- The asymptotic rate of convergence given $\left\{X_{(i)}\right\}_{i=1}^{n}$ becomes

$$\widehat{Y}_i^{(1)} - m^{(1)}(X_{(i)}) = \underbrace{O_P\left(\frac{k}{n}\right)}_{\text{Bias}} + \underbrace{O_P\left(\frac{n}{k^{\frac{3}{2}}}\right)}_{\sqrt{\text{Variance}}}.$$

To reduce the variance, Iserles (2009); Charnigo et al. (2011) considered

$$\widehat{Y}_i^{(1)} \equiv \widehat{Y}_i^{(1)}(X_{(i)}) = \sum_{j=1}^{k} w_{i,j} \left(\frac{Y_{i+j} - Y_{i-j}}{X_{(i+j)} - X_{(i-j)}} \right)$$

for $k+1 \leq i \leq n-k$ and $k \leq \frac{(n-1)}{2}$.

- The weights with $\sum_{j=1}^{k} w_{i,j} = 1$ are chosen to minimize the conditional variance $\mathrm{Var}\left(\widehat{Y}_i^{(1)} | X_{(1)}, ..., X_{(n)}\right)$.
- The asymptotic rate of convergence given $\left\{X_{(i)}\right\}_{i=1}^{n}$ becomes

$$\widehat{Y}_i^{(1)} - m^{(1)}(X_{(i)}) = \underbrace{O_P\left(\frac{k}{n}\right)}_{\text{Bias}} + \underbrace{O_P\left(\frac{n}{k^{\frac{3}{2}}}\right)}_{\sqrt{\text{Variance}}}.$$

Drawback: It only estimates $m^{(1)}(x)$ at $x = X_{(i)}$ for $k+1 \leq i \leq n-k$.

Contributions of Our Discussed Paper

De Brabanter et al. (2013) proposed using local polynomial regression to smooth out the noisy derivative estimates $\widehat{Y}_i^{(1)}, i = k+1, ..., n-k$.

De Brabanter et al. (2013) proposed using local polynomial regression to smooth out the noisy derivative estimates $\widehat{Y}_i^{(1)}, i = k+1, ..., n-k$.

Drawback: Their method only works for the equispaced design, *i.e.*,

$$X_{(i)} = a + \frac{i \cdot (b-a)}{n-1}, \quad i = 1, ..., n.$$

Contributions of Our Discussed Paper

De Brabanter et al. (2013) proposed using local polynomial regression to smooth out the noisy derivative estimates $\widehat{Y}_i^{(1)}, i = k+1, ..., n-k$.

Drawback: Their method only works for the equispaced design, *i.e.*,

$$X_{(i)} = a + \frac{i \cdot (b-a)}{n-1}, \quad i = 1, ..., n.$$

Main contribution: In this paper (Liu and De Brabanter, 2020), the authors will extend the above framework to the random design.

Methodology

Recall that our i.i.d. data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ are generated from the model

$$Y = m(X) + e,$$

where X has unknown density f and CDF F.

Fact: $F(X_i) \overset{\text{i.i.d.}}{\sim}$ Uniform$[0, 1]$ for $i = 1, ..., n$ (Casella and Berger, 2002).

Insights:

Probability Integral Transform to Uniform$[0,1]$

Recall that our i.i.d. data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ are generated from the model

$$Y = m(X) + e,$$

where X has unknown density f and CDF F.

Fact: $F(X_i) \overset{\text{i.i.d.}}{\sim} \text{Uniform}[0,1]$ for $i = 1, ..., n$ (Casella and Berger, 2002).

Insights:

- Estimate derivatives of the transformed function $r(U) = m(F^{-1}(U))$.

Probability Integral Transform to Uniform$[0,1]$

Recall that our i.i.d. data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ are generated from the model

$$Y = m(X) + e,$$

where X has unknown density f and CDF F.

Fact: $F(X_i) \overset{\text{i.i.d.}}{\sim} \text{Uniform}[0,1]$ for $i = 1, ..., n$ (Casella and Berger, 2002).

Insights:

- Estimate derivatives of the transformed function $r(U) = m(F^{-1}(U))$.
- Refer back to the derivatives of $m(X)$ by the chain rule:

$$\begin{aligned} m^{(1)}(X) &= f(X) \cdot r^{(1)}(U), \\ m^{(2)}(X) &= f^{(1)}(X) \cdot r^{(1)}(U) + [f(X)]^2 \, r^{(2)}(U), ... \end{aligned}$$

Probability Integral Transform to Uniform[0, 1]

Recall that our i.i.d. data $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$ are generated from the model

$$Y = m(X) + e,$$

where X has unknown density f and CDF F.

Fact: $F(X_i) \overset{\text{i.i.d.}}{\sim} \text{Uniform}[0,1]$ for $i = 1, ..., n$ (Casella and Berger, 2002).

Insights:

- Estimate derivatives of the transformed function $r(U) = m(F^{-1}(U))$.
- Refer back to the derivatives of $m(X)$ by the chain rule:

$$\begin{aligned} m^{(1)}(X) &= f(X) \cdot r^{(1)}(U), \\ m^{(2)}(X) &= f^{(1)}(X) \cdot r^{(1)}(U) + [f(X)]^2\, r^{(2)}(U), ... \end{aligned}$$

- Practically, f and F can be estimated by the kernel density estimator $\widehat{f}_v$ (KDE; Chen 2017) with bandwidth parameter $v > 0$.

First-Order Noisy Derivative Estimator

Data Setting: Consider the ordered data $\{(U_{(i)}, Y_i)\}_{i=1}^n$ from the model:

$$Y_i = r(U_{(i)}) + e_i, \quad i = 1, ..., n,$$

where $U_{(1)} \leq \cdots \leq U_{(n)}$ are order statistics from Uniform$[0, 1]$.

First-Order Noisy Derivative Estimator

Data Setting: Consider the ordered data $\{(U_{(i)}, Y_i)\}_{i=1}^n$ from the model:

$$Y_i = r(U_{(i)}) + e_i, \quad i = 1, ..., n,$$

where $U_{(1)} \leq \cdots \leq U_{(n)}$ are order statistics from Uniform$[0,1]$.

First-order noisy derivative estimator at $u = U_{(i)}$:

$$\widehat{Y}_i^{(1)} = \sum_{j=1}^{k} w_{i,j} \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}} \right) \quad \text{for} \quad k+1 \leq i \leq n-k,$$

where k is a tuning parameter.

- The weights are chosen to minimize Var $\left(\widehat{Y}_i^{(1)} | U_{(1)}, ..., U_{(n)}\right)$.
- The asymptotic rate of convergence of $\widehat{Y}_i^{(1)}$ given $\{U_{(i)}\}_{i=1}^n$ is

$$\widehat{Y}_i^{(1)} - r^{(1)}(U_{(i)}) = O_P\left(\frac{k}{n}\right) + O_P\left(\frac{n}{k^{\frac{3}{2}}}\right).$$

Second-order noisy derivative estimator at $u = U_{(i)}$:

$$\widehat{Y}_i^{(2)} = 2\sum_{j=1}^{k_2} w_{ij,2} \cdot \frac{\left(\frac{Y_{i+j+k_1}-Y_{i+j}}{U_{(i+j+k_1)}-U_{(i+j)}} - \frac{Y_{i-j-k_1}-Y_{i-j}}{U_{(i-j-k_1)}-U_{(i-j)}}\right)}{U_{(i+j+k_1)} + U_{(i+j)} - U_{(i-j-k_1)} - U_{(i-j)}},$$

for $k_1 + k_2 + 1 \le i \le n - k_1 - k_2$, where k_1, k_2 are tuning parameters.

- The weights $w_{ij,2}$ are chosen to minimize the asymptotic leading order of $\mathrm{Var}\left(\widehat{Y}_i^{(2)} | U_{(1)}, ..., U_{(n)}\right)$.
- The asymptotic rate of convergence of $\widehat{Y}_i^{(2)}$ given $\{U_{(i)}\}_{i=1}^n$ is

$$\widehat{Y}_i^{(2)} - r^{(2)}(U_{(i)}) = O_P\left(\frac{k}{n}\right) + O_P\left(\frac{n^2}{k^{\frac{5}{2}}}\right)$$

when $k_1, k_2 \asymp k$.

Drawbacks of the proposed noisy derivative estimators $\widehat{Y}_i^{(1)}$ and $\widehat{Y}_i^{(2)}$:

1. They are only defined at the (interior) design points $U_{(i)}$ for $k+1 \leq i \leq n-k$.

2. They contain noises from the unknown error $e_i, i = 1, ..., n$.

Drawbacks of the proposed noisy derivative estimators $\widehat{Y}_i^{(1)}$ and $\widehat{Y}_i^{(2)}$:

1. They are only defined at the (interior) design points $U_{(i)}$ for $k+1 \leq i \leq n-k$.

2. They contain noises from the unknown error $e_i, i = 1, ..., n$.

Solution: Apply the local polynomial regression to smoothing out these noisy derivative estimators.

Smoothing the Noisy Derivative Estimators

Take the first-order derivative data $\{(U_{(i)}, \widehat{Y}_i^{(1)})\}_{i=k+1}^{n-k}$ as an example.

At any point $u_0 \in [0,1]$, the solution of the local polynomial regression is

$$\widehat{r}^{(1)}(u_0) = \boldsymbol{\epsilon}_1^T \widehat{\boldsymbol{\beta}}(u_0) = \boldsymbol{\epsilon}_1^T \left(\boldsymbol{U}_u^T \boldsymbol{W}_u \boldsymbol{U}_u\right)^{-1} \boldsymbol{U}_u^T \boldsymbol{W}_u \widehat{\boldsymbol{Y}}^{(1)},$$

where $\boldsymbol{\epsilon}_1 = (1,0,...,0)^T \in \mathbb{R}^{p+1}$, $\widehat{\boldsymbol{Y}}^{(1)} = \left(\widehat{Y}_{k+1}^{(1)}, ..., \widehat{Y}_{n-k}^{(1)}\right)^T \in \mathbb{R}^{n-2k}$, and

$$\boldsymbol{U}_u = \begin{pmatrix} 1 & \left(U_{(k+1)} - u_0\right) & \cdots & \left(U_{(k+1)} - u_0\right)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \left(U_{(n-k)} - u_0\right) & \cdots & \left(U_{(n-k)} - u_0\right)^p \end{pmatrix},$$

$$\boldsymbol{W}_u = \begin{pmatrix} K\left(\frac{U_{(k+1)} - u_0}{h}\right) & & \\ & \ddots & \\ & & K\left(\frac{U_{(n-k)} - u_0}{h}\right) \end{pmatrix}.$$

Caveat: $\left\{\widehat{Y}_i^{(1)}\right\}_{i=k+1}^{n-k}$ are no longer independent even when we condition on $\{U_{(i)}\}_{i=1}^{n}$. Equivalently, $\widetilde{e}_i, i = 1, ..., n$ are correlated in the model

$$\widehat{Y}_i^{(1)} = r^{(1)}(U_{(i)}) + \widetilde{e}_i, \quad i = 1, ..., n.$$

Smoothing the Noisy Derivative Estimators

Caveat: $\left\{\widehat{Y}_i^{(1)}\right\}_{i=k+1}^{n-k}$ are no longer independent even when we condition on $\{U_{(i)}\}_{i=1}^{n}$. Equivalently, $\widetilde{e}_i, i = 1, ..., n$ are correlated in the model

$$\widehat{Y}_i^{(1)} = r^{(1)}(U_{(i)}) + \widetilde{e}_i, \quad i = 1, ..., n.$$

Solution: Use a bimodal kernel $\bar{K}$ with $\bar{K}(0) = 0$ in the local polynomial regression to tackle the correlated errors (De Brabanter et al., 2013).

- Gaussian bimodal kernel: $\bar{K}(u) = \frac{2u^2}{\sqrt{\pi}} \exp(-u^2)$.

Smoothing the Noisy Derivative Estimators

Caveat: $\left\{\widehat{Y}_i^{(1)}\right\}_{i=k+1}^{n-k}$ are no longer independent even when we condition on $\{U_{(i)}\}_{i=1}^n$. Equivalently, $\widetilde{e}_i, i = 1, ..., n$ are correlated in the model

$$\widehat{Y}_i^{(1)} = r^{(1)}(U_{(i)}) + \widetilde{e}_i, \quad i = 1, ..., n.$$

Solution: Use a bimodal kernel $\bar{K}$ with $\bar{K}(0) = 0$ in the local polynomial regression to tackle the correlated errors (De Brabanter et al., 2013).

- Gaussian bimodal kernel: $\bar{K}(u) = \frac{2u^2}{\sqrt{\pi}} \exp(-u^2)$.
- Bandwidth h is selected by minimizing $\frac{1}{n-2k} \sum_{i=k+1}^{n-k} \left(\widehat{r}^{(1)}(U_{(i)}) - \widehat{Y}_i^{(1)}\right)^2$ with a correction for applying the unimodal kernel in the final local polynomial smoothing.

Smoothing the Noisy Derivative Estimators

Caveat: $\left\{\widehat{Y}_i^{(1)}\right\}_{i=k+1}^{n-k}$ are no longer independent even when we condition on $\{U_{(i)}\}_{i=1}^n$. Equivalently, $\widetilde{e}_i, i = 1, ..., n$ are correlated in the model

$$\widehat{Y}_i^{(1)} = r^{(1)}(U_{(i)}) + \widetilde{e}_i, \quad i = 1, ..., n.$$

Solution: Use a bimodal kernel $\bar{K}$ with $\bar{K}(0) = 0$ in the local polynomial regression to tackle the correlated errors (De Brabanter et al., 2013).

- Gaussian bimodal kernel: $\bar{K}(u) = \frac{2u^2}{\sqrt{\pi}} \exp(-u^2)$.
- Bandwidth h is selected by minimizing $\frac{1}{n-2k} \sum_{i=k+1}^{n-k} \left(\widehat{r}^{(1)}(U_{(i)}) - \widehat{Y}_i^{(1)}\right)^2$ with a correction for applying the unimodal kernel in the final local polynomial smoothing.
- The asymptotic rate of convergence of $\widehat{r}^{(1)}(u_0)$ given $\{U_{(i)}\}_{i=1}^n$ is

$$\widehat{r}^{(1)}(u_0) - r^{(1)}(u_0) = O_P\left(h^{p+1}\right) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n}{k^3 h}}\right).$$

Summary of the Derivative Estimation Framework

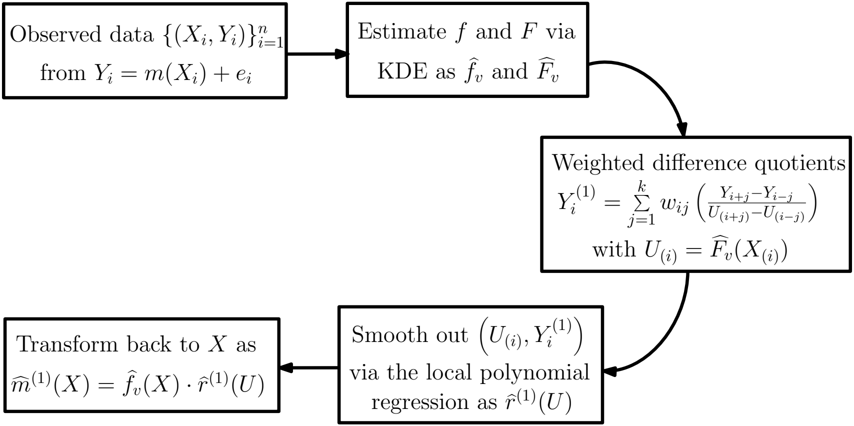
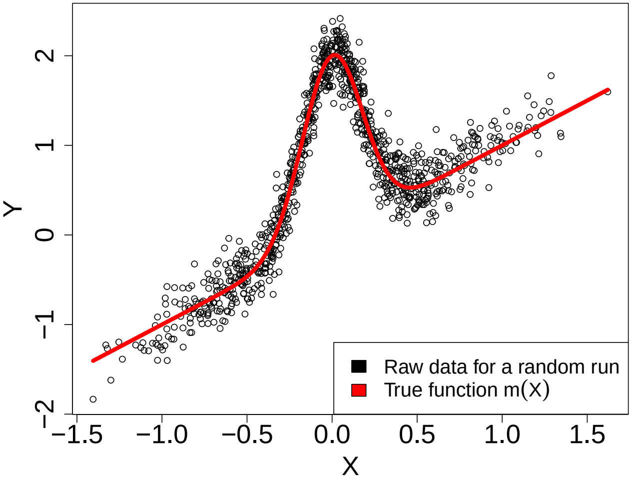


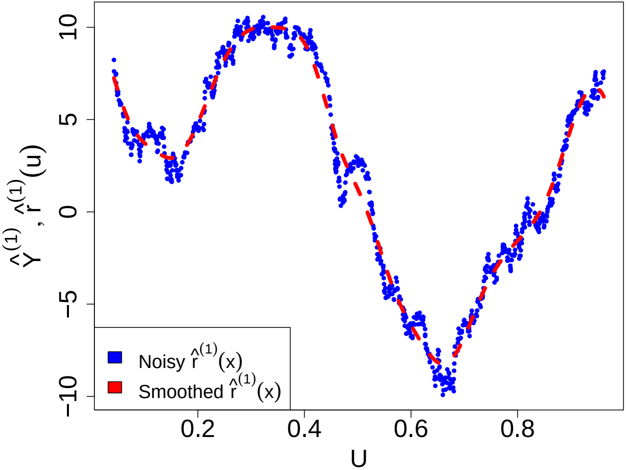
Figure 1: Summary of the proposed derivative estimation framework in the paper (Liu and De Brabanter, 2020).

Simulated observations $\{(X_i, Y_i)\}_{i=1}^{1000}$ from $Y = m(X) + e$ with $m(X) = X + 2\exp(-16X^2)$, $X \sim N(0, 0.5^2)$ and $e \sim N(0, 0.1^2)$

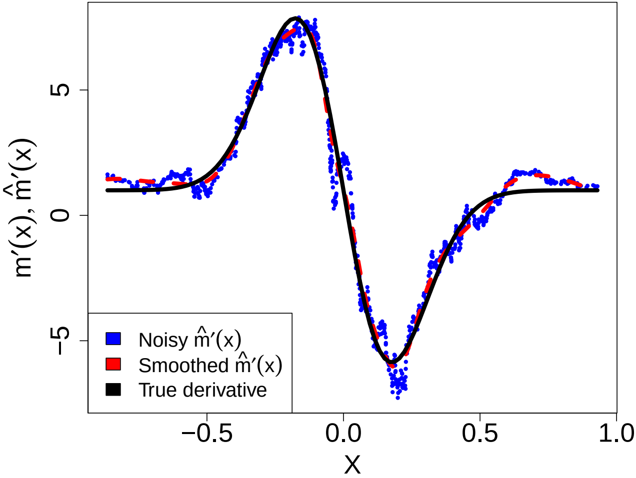


Graphical Illustration of the Proposed Derivative Estimator

The proposed first-order noisy derivatives and the smoothed ones by local polynomial regression on $[0, 1]$.



The proposed first-order derivative estimates back-transformed to the original space of X with the true derivative.



Extensions

Limitation: All the asymptotic properties and consistency results in the paper (Liu and De Brabanter, 2020) are developed after the probability integral transform $U = F(X)$, *i.e.*, it assumes that

$$Y_i = r(U_i) + e_i \quad \text{with} \quad U_i \sim \text{Uniform}[0,1] \quad \text{for} \quad i = 1, ..., n,$$

and study the derivative estimators $\widehat{r}^{(1)}(u)$ and $\widehat{r}^{(2)}(u)$.

Limitation of the Theoretical Results in the Paper

Limitation: All the asymptotic properties and consistency results in the paper (Liu and De Brabanter, 2020) are developed after the probability integral transform $U = F(X)$, *i.e.*, it assumes that

$$Y_i = r(U_i) + e_i \quad \text{with} \quad U_i \sim \text{Uniform}[0,1] \quad \text{for} \quad i = 1, ..., n,$$

and study the derivative estimators $\widehat{r}^{(1)}(u)$ and $\widehat{r}^{(2)}(u)$.

Actual Estimators: In reality, the proposed final derivative estimators are

$$\begin{aligned}
\widehat{m}^{(1)}(x) &= \widehat{f}_v(x) \cdot \widehat{r}^{(1)}(u) \\
\widehat{m}^{(2)}(x) &= \widehat{f}_v^{(1)}(x) \cdot \widehat{r}^{(1)}(u) + \left[\widehat{f}_v(x)\right]^2 \widehat{r}^{(2)}(u).
\end{aligned}$$

Limitation of the Theoretical Results in the Paper

Limitation: All the asymptotic properties and consistency results in the paper (Liu and De Brabanter, 2020) are developed after the probability integral transform $U = F(X)$, *i.e.*, it assumes that

$$Y_i = r(U_i) + e_i \quad \text{with} \quad U_i \sim \text{Uniform}[0,1] \quad \text{for} \quad i = 1, ..., n,$$

and study the derivative estimators $\widehat{r}^{(1)}(u)$ and $\widehat{r}^{(2)}(u)$.

Actual Estimators: In reality, the proposed final derivative estimators are

$$\widehat{m}^{(1)}(x) = \widehat{f}_v(x) \cdot \widehat{r}^{(1)}(u)$$
$$\widehat{m}^{(2)}(x) = \widehat{f}_v^{(1)}(x) \cdot \widehat{r}^{(1)}(u) + \left[\widehat{f}_v(x)\right]^2 \widehat{r}^{(2)}(u).$$

Question: What are the rates of convergence for $\widehat{m}^{(1)}(x)$ and $\widehat{m}^{(2)}(x)$?

Consistency of the Proposed Derivative Estimators

By leveraging convergence theories for KDE (Giné and Guillou, 2002; Einmahl and Mason, 2005; Chacón et al., 2011) and local polynomial regression (Francisco-Fernández et al., 2003) of order p, we derive that

- **Pointwise consistency:** for $q = 1, 2$,

$$\left|\widehat{m}^{(q)}(x) - m^{(q)}(x)\right| = \underbrace{O\left(h^{p+1}\right) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n^{2q-1}}{k^{2q+1}h}}\right)}_{\text{Original Rates for } \widehat{r}^{(q)}(u)} + \underbrace{O(v^2) + O_P\left(\sqrt{\frac{1}{nv^{2q-1}}}\right)}_{\text{Additional rates from KDE } \widehat{f}_v},$$

  - h is the bandwidth parameter of local polynomial regression;
  - k is the tuning parameter in constructing noisy derivative estimators;
  - v is the bandwidth parameter for KDE.

Consistency of the Proposed Derivative Estimators

By leveraging convergence theories for KDE (Giné and Guillou, 2002; Einmahl and Mason, 2005; Chacón et al., 2011) and local polynomial regression (Francisco-Fernández et al., 2003) of order p, we derive that

- **Pointwise consistency:** for $q = 1, 2$,

$$\left|\widehat{m}^{(q)}(x) - m^{(q)}(x)\right| = \underbrace{O\left(h^{p+1}\right) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n^{2q-1}}{k^{2q+1}h}}\right)}_{\text{Original Rates for } \widehat{r}^{(q)}(u)} + \underbrace{O(v^2) + O_P\left(\sqrt{\frac{1}{nv^{2q-1}}}\right)}_{\text{Additional rates from KDE } \widehat{f}_v},$$

  - h is the bandwidth parameter of local polynomial regression;
  - k is the tuning parameter in constructing noisy derivative estimators;
  - v is the bandwidth parameter for KDE.

- **Uniform consistency:** for $q = 1, 2$,

$$\sup_{x\in[a,b]}\left|\widehat{m}^{(q)}(x) - m^{(q)}(x)\right| = O\left(h^{p+1}\right) + O_P\left(\frac{k}{n}\right) + O_P\left(\sqrt{\frac{n^{2q-1}\log n}{k^{2q+1}h}}\right) + O(v^2) + O_P\left(\sqrt{\frac{\log n}{nv^{2q-1}}}\right).$$

Comparative Experiments

Derivative Estimation Methods For Comparisons

We compare the proposed derivative estimator in the paper (Liu and De Brabanter, 2020) with other existing derivative estimators as:

- **Penalized smoothing cubic splines:** It is implemented in R package `pspline` (Ramsey and Ripley, 2022).
- **Local polynomial regression:** It is implemented in R package `locpol` (Ojeda Cabrera, 2022).

We compare the proposed derivative estimator in the paper (Liu and De Brabanter, 2020) with other existing derivative estimators as:

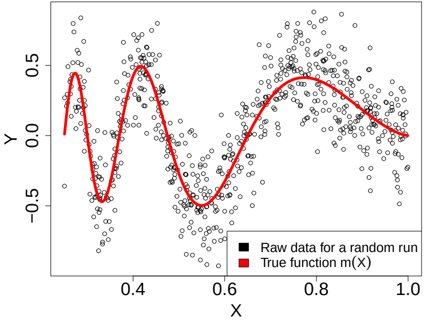
- **Penalized smoothing cubic splines:** It is implemented in `R` package `pspline` (Ramsey and Ripley, 2022).
- **Local polynomial regression:** It is implemented in `R` package `locpol` (Ojeda Cabrera, 2022).
- **Gasser-Müller estimator:** We implement it in `R` with Gaussian kernel and an optimal cross-validated bandwidth under the local polynomial regression with $p = 0$.
- **Nadaraya-Watson estimator:** We implement it in `R` with Gaussian kernel, a two-stage plug-in bandwidth for KDE, and the same cross-validated bandwidth for the regression component.

Experiments on First-Order Derivative Estimation (I)

We repeat the following procedure 100 times for each first-order derivative estimation method:

1. Sample i.i.d. observations $\{(X_i, Y_i)\}_{i=1}^{700}$ from $Y = m(X) + e$ with

$$m(X) = \sqrt{X(1-X)} \cdot \sin\left(\frac{2.1\pi}{X+0.05}\right) \text{ for } \boldsymbol{X \sim \text{Unif}(0.25, 1)} \text{ and } e \sim N(0, 0.2^2).$$



2. Compute an adjusted mean absolute error $\frac{1}{650}\sum_{i=26}^{675}\left|\widehat{m}^{(1)}(X_{(i)}) - m^{(1)}(X_{(i)})\right|$.

The original experimental results in the paper (Liu and De Brabanter, 2020) are

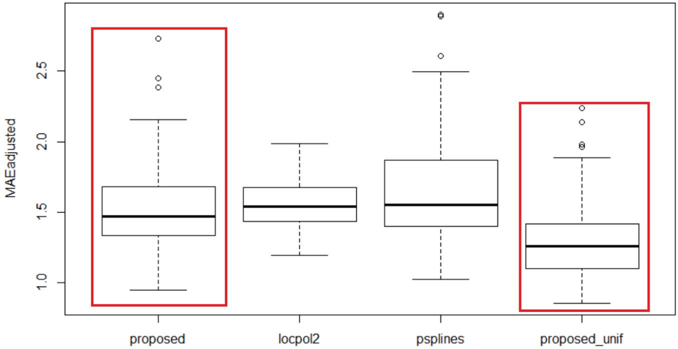


Figure 2: Comparative boxplots of adjusted mean absolute errors for the first-order derivative estimation methods under Monte Carlo simulation studies.

My extended experimental results are

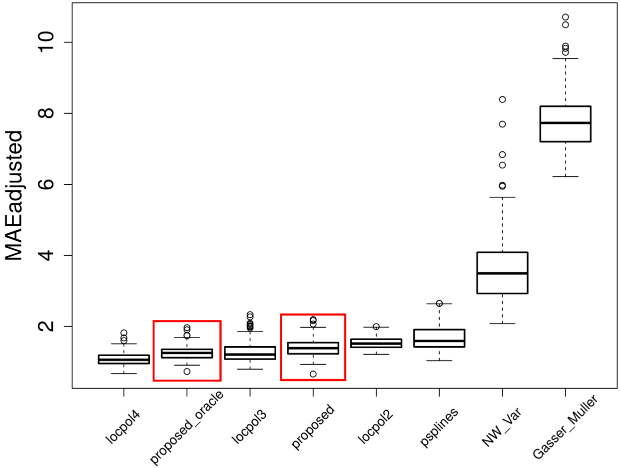
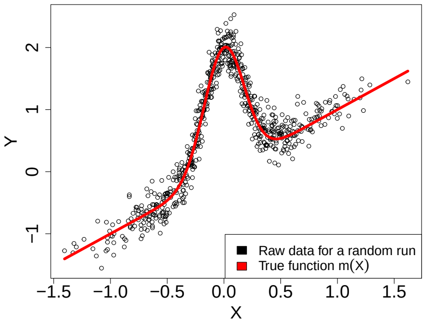


Figure 2: Comparative boxplots of adjusted mean absolute errors for the first-order derivative estimation methods under Monte Carlo simulation studies.

Beyond the uniform distribution of X, we also consider the following repeated simulations 100 times for each derivative estimation method:

1. Sample i.i.d. observations $\{(X_i, Y_i)\}_{i=1}^{700}$ from $Y = m(X) + e$ with

$$m(X) = X + 2\exp(-16X^2) \text{ for } \boldsymbol{X \sim N\left(0, 0.5^2\right)} \text{ and } e \sim N(0, 0.2^2).$$



2. Compute an adjusted mean absolute error $\frac{1}{650}\sum_{i=26}^{675}\left|\widehat{m}^{(1)}(X_{(i)}) - m^{(1)}(X_{(i)})\right|$.

The original experimental results in the paper (Liu and De Brabanter, 2020) are

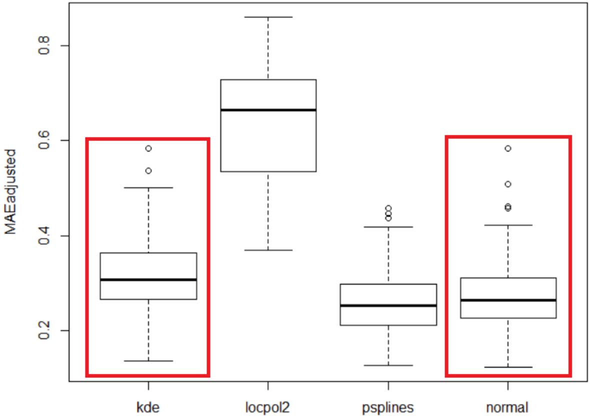


Figure 3: Comparative boxplots of adjusted mean absolute errors for the first-order derivative estimation methods under Monte Carlo simulation studies.

My extended experimental results are

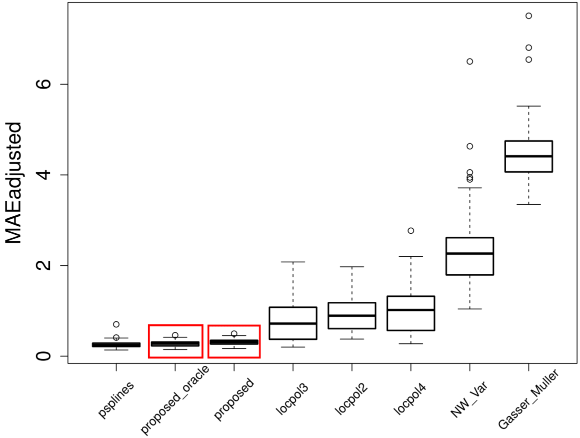


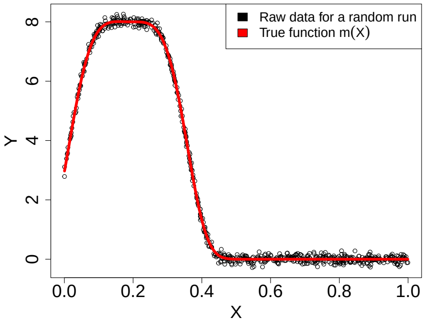
Figure 3: Comparative boxplots of adjusted mean absolute errors for the first-order derivative estimation methods under Monte Carlo simulation studies.

Experiments on Second-Order Derivative Estimation (I)

We repeat the following procedure 100 times for each second-order derivative estimation method:

1. Sample i.i.d. observations $\{(X_i, Y_i)\}_{i=1}^{700}$ from $Y = m(X) + e$ with

$$m(X) = 8e^{-(1-5x)^3(1-7x)} \quad \text{for} \quad X \sim \text{Unif}(0,1) \quad \text{and} \quad e \sim N(0, 0.1^2).$$



2. Compute an adjusted mean absolute error $\frac{1}{640}\sum_{i=31}^{670}\left|\widehat{m}^{(2)}(X_{(i)}) - m^{(2)}(X_{(i)})\right|$.

Experiments on Second-Order Derivative Estimation (I)

The original experimental results in the paper (Liu and De Brabanter, 2020) are

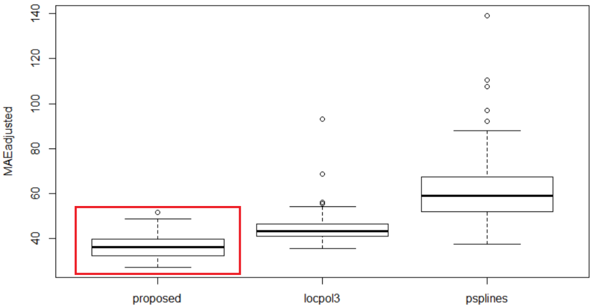


Figure 4: Comparative boxplots of adjusted mean absolute errors for the second-order derivative estimation methods under Monte Carlo simulation studies.

Experiments on Second-Order Derivative Estimation (I)

My extended experimental results are

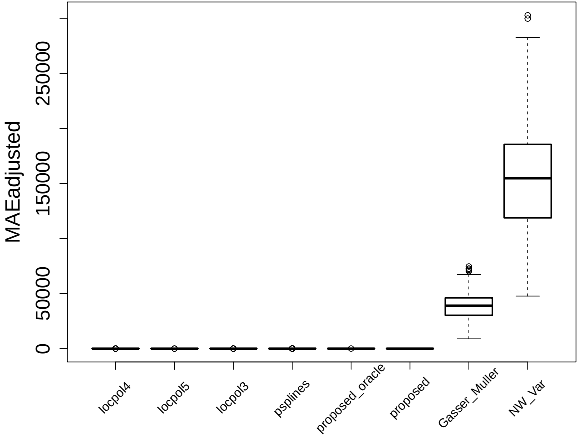


Figure 4: Comparative boxplots of adjusted mean absolute errors for the second-order derivative estimation methods under Monte Carlo simulation studies.

My extended experimental results are

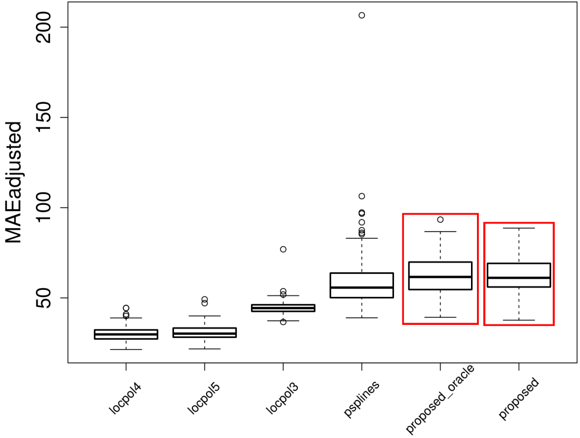


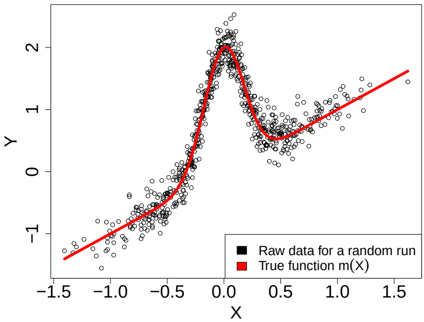
Figure 4: Comparative boxplots of adjusted mean absolute errors for the second-order derivative estimation methods under Monte Carlo simulation studies.

Experiments on Second-Order Derivative Estimation (II)

Beyond the uniform distribution of X, we also consider the following repeated experiments 100 times for each derivative estimation method:

1. Sample i.i.d. observations $\{(X_i, Y_i)\}_{i=1}^{700}$ from $Y = m(X) + e$ with

$$m(X) = X + 2\exp(-16X^2) \text{ for } \boldsymbol{X \sim N\left(0, 0.5^2\right)} \text{ and } e \sim N(0, 0.2^2).$$



2. Compute an adjusted mean absolute error $\frac{1}{640}\sum_{i=31}^{670}\left|\widehat{m}^{(2)}(X_{(i)}) - m^{(2)}(X_{(i)})\right|$.

My extended experimental results are

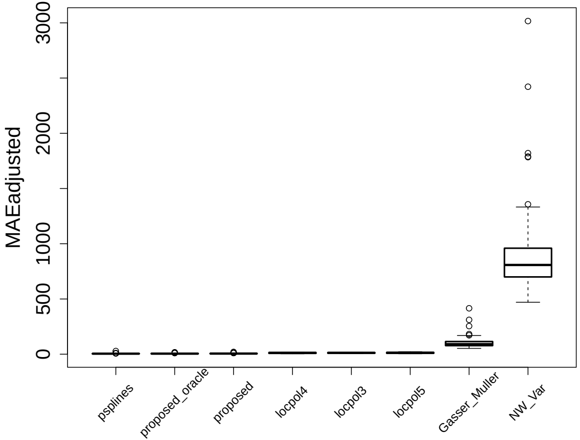


Figure 5: Comparative boxplots of adjusted mean absolute errors for the second-order derivative estimation methods under Monte Carlo simulation studies.

My extended experimental results are

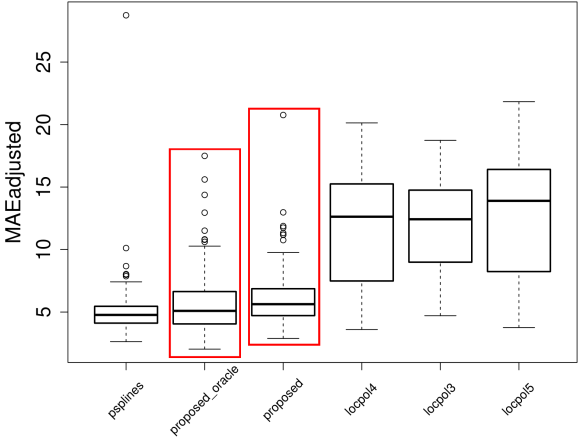


Figure 5: Comparative boxplots of adjusted mean absolute errors for the second-order derivative estimation methods under Monte Carlo simulation studies.

Discussions

Summary: The paper (Liu and De Brabanter, 2020) proposed a data-driven method for estimating the first and second order derivatives via

- Weighted difference quotients.
- Local polynomial regression.

Main contribution: It develops asymptotic properties for the proposed estimators under the random design.

Discussions for the Paper

Summary: The paper (Liu and De Brabanter, 2020) proposed a data-driven method for estimating the first and second order derivatives via

- Weighted difference quotients.
- Local polynomial regression.

Main contribution: It develops asymptotic properties for the proposed estimators under the random design.

Question: Are the proposed estimators useful in practice?

Discussions for the Paper

Summary: The paper (Liu and De Brabanter, 2020) proposed a data-driven method for estimating the first and second order derivatives via

- Weighted difference quotients.
- Local polynomial regression.

Main contribution: It develops asymptotic properties for the proposed estimators under the random design.

Question: Are the proposed estimators useful in practice?

Answer: Our answer may be "No!" based on their estimation errors in the simulation studies.

What's worse: It is difficult to generalize the proposed framework to the higher order derivative estimation.

Summary: The paper (Liu and De Brabanter, 2020) proposed a data-driven method for estimating the first and second order derivatives via

- Weighted difference quotients.
- Local polynomial regression.

Main contribution: It develops asymptotic properties for the proposed estimators under the random design.

Question: Are the proposed estimators useful in practice?

Answer: Our answer may be “No!” based on their estimation errors in the simulation studies.

What’s worse: It is difficult to generalize the proposed framework to the higher order derivative estimation.

Follow-up: How about its running time?

Time Comparisons for Derivative Estimation Methods

Sad news for the proposed methods in the running time comparisons!

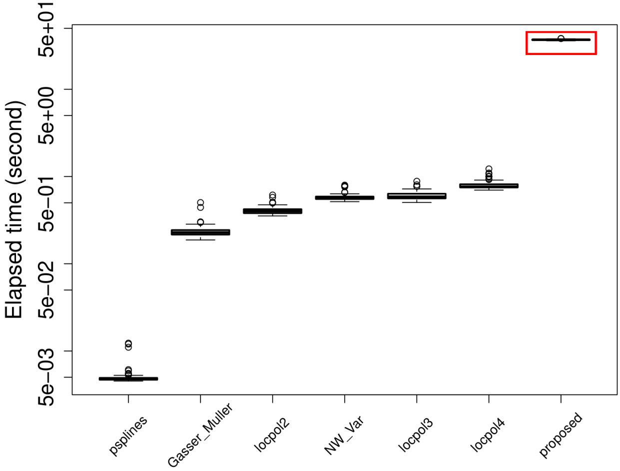


Figure 6: Time comparisons for different first-order derivative estimation methods under 100 repeated experiments.

Future Directions for the Paper

1. **Improving accuracy:** Smooth the data first by penalized smoothing splines (or other regression methods) before taking the noisy derivatives:

$$\widehat{Y}_i^{(1)} = \sum_{j=1}^{k} w_{i,j} \left(\frac{\widehat{m}(X_{(i+j)}) - \widehat{m}(X_{(i-j)})}{X_{(i+j)} - X_{(i-j)}} \right) \quad \text{for} \quad k+1 \leq i \leq n-k.$$

Future Directions for the Paper

1. **Improving accuracy:** Smooth the data first by penalized smoothing splines (or other regression methods) before taking the noisy derivatives:

$$\widehat{Y}_i^{(1)} = \sum_{j=1}^{k} w_{i,j} \left(\frac{\widehat{m}(X_{(i+j)}) - \widehat{m}(X_{(i-j)})}{X_{(i+j)} - X_{(i-j)}} \right) \quad \text{for} \quad k+1 \leq i \leq n-k.$$

2. **Generalization to multivariate data:** Dang (2021) considered generalizing the proposed framework to multivariate data $\{(X_{i1}, ..., X_{id}, Y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ under the independence assumption between covariates.
   - **Open problem:** How can we estimate the (partial) derivatives of a multivariate regression function with $\{(X_{i1}, ..., X_{id}, Y_i)\}_{i=1}^n$ when the covariates are not independent?

Thank you!

More details can be found in
https://github.com/zhangyk8/NonDeriDQ.

S. Calonico, M. D. Cattaneo, and M. H. Farrell. On the effect of bias estimation on coverage accuracy in nonparametric inference. *Journal of the American Statistical Association*, 113(522):767–779, 2018.

G. Casella and R. Berger. *Statistical Inference*. Duxbury advanced series. Thomson Learning, 2nd edition, 2002.

J. E. Chacón, T. Duong, and M. Wand. Asymptotics for general multivariate kernel density derivative estimators. *Statistica Sinica*, pages 807–840, 2011.

R. Charnigo, B. Hall, and C. Srinivasan. A generalized c_p criterion for derivative estimation. *Technometrics*, 53(3):238–253, 2011.

P. Chaudhuri and J. S. Marron. Sizer for exploration of structures in curves. *Journal of the American Statistical Association*, 94(447):807–823, 1999.

Y.-C. Chen. A tutorial on kernel density estimation and recent advances. *Biostatistics & Epidemiology*, 1 (1):161–187, 2017.

G. Cheng and Y.-C. Chen. Nonparametric inference via bootstrapping the debiased estimator. *Electronic Journal of Statistics*, 13(1):2194 – 2256, 2019.

J. Dang. Smoothed nonparametric derivative estimation using random forest based weighted difference quotients. 2021.

K. De Brabanter, J. De Brabanter, I. Gijbels, and B. De Moor. Derivative estimation with local polynomial fitting. *Journal of Machine Learning Research*, 14(1):281–301, 2013.

K. De Brabanter, F. Cao, I. Gijbels, and J. Opsomer. Local polynomial regression with correlated errors in random design and unknown correlation structure. *Biometrika*, 105(3):681–690, 2018.

U. Einmahl and D. M. Mason. Uniform in bandwidth consistency of kernel-type function estimators. *The Annals of Statistics*, 33(3):1380–1403, 2005.

R. L. Eubank and P. L. Speckman. Confidence bands in nonparametric regression. *Journal of the American Statistical Association*, 88(424):1287–1301, 1993.

Reference

J. Fan and I. Gijbels. *Local polynomial modelling and its applications*, volume 66. Chapman & Hall/CRC, 1996.

M. Francisco-Fernández, J. M. Vilar-Fernández, and J. A. Vilar-Fernández. On the uniform strong consistency of local polynomial regression under dependence conditions. *Communications in Statistics-Theory and Methods*, 32(12):2415–2440, 2003.

T. Gasser and H.-G. Müller. Estimating regression functions and their derivatives by the kernel method. *Scandinavian journal of statistics*, pages 171–185, 1984.

E. Giné and A. Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, 38(6):907–921, 2002.

T. Haavelmo. Methods of measuring the marginal propensity to consume. *Journal of the American Statistical Association*, 42(237):105–122, 1947.

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

W. Härdle. *Applied nonparametric regression*. Number 19. Cambridge university press, 1990.

A. Iserles. *A first course in the numerical analysis of differential equations*. Number 44. Cambridge university press, 2009.

Y. Liu and K. De Brabanter. Smoothed nonparametric derivative estimation using weighted difference quotients. *Journal of Machine Learning Research*, 21(1):2438–2482, 2020.

Y. Mack and H.-G. Müller. Derivative estimation in nonparametric regression with random predictor variable. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 59–72, 1989.

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. L. Ojeda Cabrera. *locpol: Kernel Local Polynomial Regression*, 2022. URL https://CRAN.R-project.org/package=locpol. R package version 0.8.0 [Online; accessed 3-April-2023].

J. Ramsey and B. Ripley. *pspline: Penalized Smoothing Splines*, 2022. URL https://CRAN.R-project.org/package=pspline. R package version 1.0-19 [Online; accessed 3-April-2023].

V. Rondonotti, J. S. Marron, and C. Park. SiZer for time series: A new approach to the analysis of trends. *Electronic Journal of Statistics*, 1(none):268 – 289, 2007.

C. J. Stone. Optimal rates of convergence for nonparametric estimators. *The annals of Statistics*, pages 1348–1360, 1980.

C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

H. J. Woltring. On optimal smoothing and derivative estimation from noisy displacement data in biomechanics. *Human Movement Science*, 4(3):229–245, 1985.

S. Zhou and D. A. Wolfe. On derivative estimation in spline regression. *Statistica Sinica*, pages 93–108, 2000.

Optimal Rate of Convergence for Derivative Estimation

Consider a p-times differentiable regression function $m : \mathcal{X} \to \mathbb{R}$ with $\mathcal{X}$ being a compact subset of $\mathbb{R}^d$. Or, we can assume that

$$\left| m^{(\boldsymbol{\alpha})}(\boldsymbol{x}) - m^{(\boldsymbol{\alpha})}(\boldsymbol{y}) \right| \leq C \left|\left| \boldsymbol{x} - \boldsymbol{y} \right|\right|_2^{\zeta}$$

for some constants $C > 0, \zeta \in (0, 1]$ and take $p = [\boldsymbol{\alpha}] + \zeta$, where

$$m^{(\boldsymbol{\alpha})}(\boldsymbol{x}) = \frac{\partial^{[\boldsymbol{\alpha}]}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}} m(\boldsymbol{x}) \quad \text{with } \boldsymbol{\alpha} = (\alpha_1, ..., \alpha_d) \text{ and } [\boldsymbol{\alpha}] = \sum_{i=1}^{d} \alpha_i.$$

- Let $\widehat{m}^{(\boldsymbol{\alpha})}(\boldsymbol{x})$ be an estimator of $m^{(\boldsymbol{\alpha})}(\boldsymbol{x})$ based on the i.i.d. data $\{(\boldsymbol{X}_i, Y_i)\}_{i=1}^n$ from $Y = m(\boldsymbol{X}) + e$ with $\boldsymbol{X} \perp e$.
- Assume also that the density f of $\boldsymbol{X}$ is bounded away from 0 in an open subset of $\mathbb{R}^d$ that covers $\mathcal{X}$.

Optimal Rate of Convergence for Derivative Estimation

Definition

A sequence $\{b_n\}_{n=1}^{\infty}$ of positive constants is said to be an *optimal rate of convergence* if there exist constants $c_1, c_2 > 0$ such that

$$\lim_{n\to\infty} \inf_{\widehat{m}} \sup_{m} \mathrm{P}\left(\left|\left| \widehat{m}^{(\boldsymbol{\alpha})} - m^{(\boldsymbol{\alpha})} \right|\right|_q \geq c_1 b_n \right) = 1$$

and there exists some derivative estimator $\widetilde{m}^{(\boldsymbol{\alpha})}$ such that

$$\lim_{n\to\infty} \sup_{m} \mathrm{P}\left(\left|\left| \widetilde{m}^{(\boldsymbol{\alpha})} - m^{(\boldsymbol{\alpha})} \right|\right|_q \geq c_2 b_n \right) = 0,$$

where $||g||_q = \left(\int_{\mathcal{X}} |g(\boldsymbol{x})| d\boldsymbol{x} \right)^{\frac{1}{q}}$ if $0 < q < \infty$ and $||g||_\infty = \sup_{\boldsymbol{x}\in\mathcal{X}} |g(\boldsymbol{x})|$.

Under the definition and conditions, the optimal rate of convergence is given by (Stone, 1980, 1982):

- $\left\{ n^{-\frac{p-[\boldsymbol{\alpha}]}{2p+d}} \right\}$ if $0 < q < \infty$; and $\left\{ \left(\frac{\log n}{n} \right)^{\frac{p-[\boldsymbol{\alpha}]}{2p+d}} \right\}$ if $q = \infty$.

Recall that the proposed first-order noisy derivative estimator

$$\widehat{Y}_i^{(1)} = \sum_{j=1}^{k} w_{i,j} \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}} \right)$$

is only defined at $U_{(i)}$ for $k+1 \le i \le n-k$.

Issue: There are not enough pairs of observations within the left and right boundary regions $2 \le i \le k$ and $n-k+1 \le i \le n-1$.

Naive Solution:

$$\widehat{Y}_i^{(1)} = \sum_{j=1}^{k(i)} w_{i,j} \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}} \right),$$

where $k(i) = i-1$ for the left boundary and $k(i) = n-i$ for the right boundary.

Recall that the proposed first-order noisy derivative estimator

$$\widehat{Y}_i^{(1)} = \sum_{j=1}^{k} w_{i,j} \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}} \right)$$

is only defined at $U_{(i)}$ for $k+1 \leq i \leq n-k$.

Issue: There are not enough pairs of observations within the left and right boundary regions $2 \leq i \leq k$ and $n-k+1 \leq i \leq n-1$.

Proposed boundary correction:

$$\widehat{Y}_i^{(1)} = \sum_{j=1}^{k(i)} w_{i,j} \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}} \right) + \sum_{j=k(i)+1}^{k} w_{i,j} \left[\left(\frac{Y_{i+j} - Y_i}{U_{(i+j)} - U_{(i)}} \right) \mathbb{1}_{\{2 \leq i \leq k\}} + \left(\frac{Y_i - Y_{i-j}}{U_{(i)} - U_{(i-j)}} \right) \mathbb{1}_{\{n-k < i < n\}} \right],$$

where $k(i) = i-1$ for the left boundary and $k(i) = n-i$ for the right boundary.

Selecting the Tuning Parameter k in Practice

Assume that the regression function r is twice continuously differentiable on $[0,1]$ under the model

$$Y_i = r(U_{(i)}) + e_i, \quad i = 1, ..., n.$$

Let $\mathcal{B} = \sup_{u \in [0,1]} \left| r^{(2)}(u) \right|$. Then, the tuning parameter k that minimizes the asymptotic upper bound of the conditional MISE is given by

$$k_{\text{opt}} = \mathop{\arg\min}_{k=1,2,...,\lfloor \frac{n-1}{2} \rfloor} \left[\mathcal{B}^2 \frac{9k^2(k+1)^2}{16(n+1)^2(2k+1)^2} + \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} \right],$$

where $\mathbb{U} = \left(U_{(1)}, ..., U_{(n)}\right)$ and $\sigma_e^2 = \text{Var}(e_i) < \infty$. In practice,

- $\mathcal{B}$ can be approximated by the second-order local slope of a local polynomial regression of order $p = 3$ fitted to the data $\{(U_{(i)}, Y_i)\}_{i=1}^n$.
- σ_e^2 can be estimated by Hall's $\sqrt{n}$-consistent estimator with the optimal second-order difference sequence (Hall et al., 1990) as

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

The paper adopts a two-step procedure proposed by De Brabanter et al. (2018) to select the final bandwidth $\widehat{h}$ for the local polynomial regression.

1. Fit a local polynomial regression with $p = 3$ using a bimodal kernel $\bar{K}(u) = \frac{2u^2}{\sqrt{\pi}} \exp(-u^2)$ and compute a pilot bandwidth $\widehat{h}_b$ by minimizing

$$\widehat{h}_b = \arg\min_{h_b > 0} \mathrm{RSS}(h_b) = \arg\min_{h_b > 0} \left\{ \frac{1}{n-2k} \sum_{i=k+1}^{n-k} \left(\widehat{r}^{(1)}(U_{(i)}) - \widehat{Y}_i^{(1)} \right)^2 \right\},$$

given the tuning parameter k is chosen a priori.

2. Correct the bandwidth for the unimodal kernel $K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right)$ as:

$$\widehat{h} = \left\{ \frac{\int \left(K_p^\star(t)\right)^2 dt \left[\int t^{p+1} \bar{K}_p^\star(t) dt\right]^2}{\int \left(\bar{K}_p^\star(t)\right)^2 dt \left[\int t^{p+1} K_p^\star(t) dt\right]^2} \right\}^{\frac{1}{2p+2}} \widehat{h}_b = 1.01431 \widehat{h}_b,$$

where $K_p^\star(u), \bar{K}_p^\star(u)$ are equivalent kernels defined by $\bar{K}(u)$ and $K(u)$.

The final smoothed derivative estimator $\widehat{r}^{(1)}(u_0)$ is computed with the unimodal kernel K and selected bandwidth $\widehat{h}$.

Assume that

1. The kernel function $K : \mathbb{R} \to [0, \infty)$ is bounded, symmetric, and Lipschitz continuous at 0. Furthermore, it satisfies $\lim_{|u|\to\infty} |u|^{\ell} K(u) < \infty$ for $\ell = 0, ..., p$.

2. The correlation function ρ_n of the error terms $\widetilde{e}_i, i = 1, ..., n$ is an element of a sequence $\{\rho_n\}_{n=1}^{\infty}$ with the following properties for all $n \geq 1$: there exists constants $\rho_{\max}, \rho_c > 0$ such that

$$n \int |\rho_n(x)| dx < \rho_{\max} \quad \text{and} \quad \lim_{n\to\infty} n \int \rho_n(x) dx = \rho_c.$$

   In addition, for any sequence $\epsilon_n > 0$ with $n\epsilon_n \to \infty$, it holds that $n \int_{|x|\geq\epsilon_n} |\rho_n(x)| dx \to 0$ as $n \to \infty$.

Lemma (Theorem 2 in De Brabanter et al. 2018)

Under the above assumptions and a $(p+2)$ times continuously differentiable function $r(\cdot)$, if $n^{\delta}\int|\rho_n(t)|dt < \rho_{\delta}$ for $\delta > 1$, p is odd, and $h \in \mathcal{H}_n$ with $\mathcal{H}_n = \left[c_1 n^{-\frac{1}{2p+3}}, c_2 n^{-\frac{1}{2p+3}}\right]$ for some constants $0 < c_1 < c_2 < \infty$, then

$$\mathrm{RSS}(h) = \mathrm{SSE}(h) + \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\widetilde{e}_i^2 - \frac{2\sigma_{\widetilde{e}}^2 \cdot K(0) \cdot \left(\mathbf{S}^{-1}\right)_{11} \cdot (1+\rho_c)}{nh} + o_P\left(n^{-\frac{2p+2}{2p+3}}\right),$$

recalling that the domain of $r^{(1)}$ is $[0,1]$ and $\left(\mathbf{S}^{-1}\right)_{11}$ is the first element in the first row of $\mathbf{S}^{-1}$, where $\mathbf{S} = \left(\mu_{i+j-2}\right)_{1\leq i,j\leq p+1}$ with $\mu_j = \int u^j K(u)du$.

Here,

$$\mathrm{RSS}(h) = \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\left(\widehat{r}^{(1)}(U_{(i)}) - \widehat{Y}_i^{(1)}\right)^2 \quad \text{and} \quad \mathrm{SSE}(h) = \frac{1}{n-2k}\sum_{i=k+1}^{n-k}\left(\widehat{r}^{(1)}(U_{(i)}) - r^{(1)}(U_{(i)})\right)^2.$$

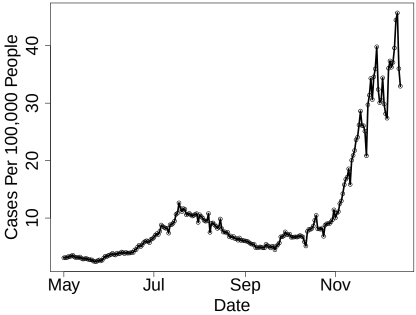
Assumptions for Consistency Results of $\widehat{m}^{(1)}(x), \widehat{m}^{(2)}(x)$

1. The kernel function for KDE $K_{\text{kde}} : \mathbb{R} \to [0, \infty)$ is bounded, symmetric, and differentiable (almost everywhere) with $\int u^2 K_{\text{kde}}(u) du < \infty$ and $\int K_{\text{kde}}^{(\alpha)}(u)^2 du < \infty$ for $\alpha = 0, 1$.

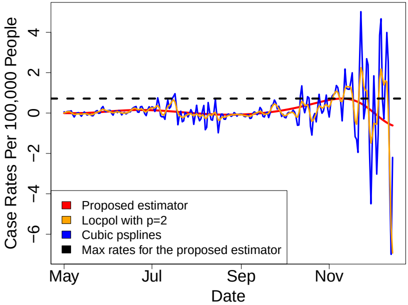
2. Let $\mathcal{K} = \left\{ y \mapsto K_{\text{kde}}^{(\alpha)}\left(\frac{x-y}{v}\right) : x \in \mathbb{R}, v > 0, \alpha = 0, 1 \right\}$. We assume that $\mathcal{K}$ is a bounded VC (subgraph) class of measurable functions on $\mathbb{R}$, *i.e.*, there exist absolute constants $A, \nu > 0$ such that for any $\epsilon \in (0, 1)$, $\sup_Q N\left(\mathcal{K}, L_2(Q), \epsilon \,||F||_{L_2(Q)}\right) \leq \left(\frac{A}{\epsilon}\right)^{\nu}$, where $M\left(\mathcal{K}, L_2(Q), \epsilon\right)$ is the ϵ-covering number of the normed space $\left(\mathcal{K}, ||\cdot||_{L_2(Q)}\right)$, Q is any probability measure on $\mathbb{R}$, and F is an envelope function of $\mathcal{K}$. Here, the norm $||F||_{L_2(Q)}$ is defined as $\left[\int_{\mathbb{R}} |F(x)|^2 dQ(x)\right]^{\frac{1}{2}}$; see Giné and Guillou (2002); Einmahl and Mason (2005).

3. The regression function $m(\cdot)$ is $(p+3)$ times continuously differetiable within $[a,b]$, and the density f of X is at least three times continuously differentiable with $\inf_{x\in[a,b]} f(x) > c > 0$ for some constant c.

4. The stationary correlation functions ρ_n and $\acute{\rho}_n$ for the error terms $\widetilde{e}_i, \acute{e}_i$ in the local polynomial smoothing come from a first-order autoregressive process with $\mathbb{E}\left(|\widetilde{e}_i|^\delta\right) < \infty, \mathbb{E}\left(|\acute{e}_i|^\delta\right) < \infty$ and are α-mixing with mixing coefficients $\alpha(k)$ such that $\sum_{k=1}^{\infty} k \cdot \alpha(k)^{1-\frac{2}{\delta}}$ for some $\delta > 2$. Moreover, define the sequence $M_n = \left(n \log n (\log\log n)^{1+\gamma}\right)^{\frac{1}{\delta}}$ for some $0 < \gamma < 1$. Then, the bandwidth $h = h_n$ satisfies that $\gamma_n = \left(\frac{nM_n^2}{h_n^3 \log n}\right)^{\frac{1}{2}} \to \infty$ and $b_n = \left(\frac{nh_n}{M_n^2 \log n}\right)^{\frac{1}{2}} \to \infty$ as $n \to \infty$. Finally, the α-mixing sequence $\alpha(k)$ satisfies $\sum_{n=1}^{\infty} \frac{n\gamma_n}{b_n} \left(\frac{nM_n^2}{h_n \log n}\right)^{\frac{1}{2}} \alpha(b_n) < \infty$; see Francisco-Fernández et al. (2003).



(a) Reported COVID-19 cases.



(b) Estimated case rates for different methods.

Figure 7: Estimated COVID-19 case rates at the Washington State between "2020-05-01" and "2020-12-15" by the proposed first-order derivative estimator ("proposed"), local polynomial regression of order $p = 2$ ("locpol2"), and penalized smoothing cubic splines ("psplines").