

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

May 25, 2023

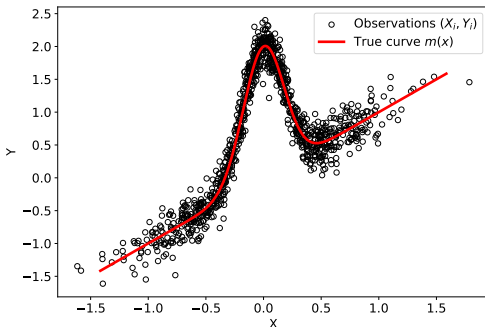
Introduction



Data setting:

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

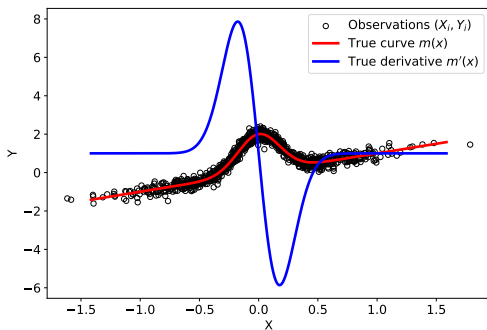
where e_i is independent of X_i and $\mathbb{E}(e_i) = 0$, $\text{Var}(e_i) = \sigma_e^2 < \infty$.



Data setting:

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

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



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

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 inference ([Eubank and Speckman, 1993](#); [Calonico et al., 2018](#); [Cheng and Chen, 2019](#)).

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 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.

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 $\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^n$.

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

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} \sum_{i=1}^n [Y_i - f(X_i; \theta)]^2 \quad \implies \quad \hat{m}^{(1)}(x) = f^{(1)}(x; \hat{\theta}).$$

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

Nonparametric methods: Make no 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).

- **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:

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

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:

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

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:

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

where \hat{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:

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

where $K : \mathbb{R} \rightarrow [0, \infty)$ 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 $X_i, i = 1, \dots, n$:

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

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

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

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

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

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

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

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

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

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

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

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 $\text{Var} \left(\hat{Y}_i^{(1)} | X_{(1)}, \dots, X_{(n)} \right)$.
- The asymptotic rate of convergence given $\{X_{(i)}\}_{i=1}^n$ becomes

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

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

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

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 $\text{Var} \left(\hat{Y}_i^{(1)} | X_{(1)}, \dots, X_{(n)} \right)$.
- The asymptotic rate of convergence given $\{X_{(i)}\}_{i=1}^n$ becomes

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

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

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

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

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

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

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

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

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

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

Insights:

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

Insights:

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

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: $F(X_i) \stackrel{\text{i.i.d.}}{\sim} \text{Uniform}[0, 1]$ for $i = 1, \dots, n$ (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:

$$m^{(1)}(X) = f(X) \cdot r^{(1)}(U),$$

$$m^{(2)}(X) = f^{(1)}(X) \cdot r^{(1)}(U) + [f(X)]^2 r^{(2)}(U), \dots$$

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: $F(X_i) \stackrel{\text{i.i.d.}}{\sim} \text{Uniform}[0, 1]$ for $i = 1, \dots, n$ ([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:

$$m^{(1)}(X) = f(X) \cdot r^{(1)}(U),$$

$$m^{(2)}(X) = f^{(1)}(X) \cdot r^{(1)}(U) + [f(X)]^2 r^{(2)}(U), \dots$$

- Practically, f and F can be estimated by the kernel density estimator (KDE; [Chen 2017](#)).

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

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

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

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

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

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

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

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

where k is a tuning parameter.

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

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

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

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

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

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

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

when $k_1, k_2 \asymp k$.

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

- 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 $e_i, i = 1, \dots, n$.

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

- 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 $e_i, i = 1, \dots, n$.

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

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

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

$$\hat{r}^{(1)}(u_0) = \epsilon_1^T \hat{\beta}(u_0) = \epsilon_1^T (\mathbf{U}_u^T \mathbf{W}_u \mathbf{U}_u)^{-1} \mathbf{U}_u^T \mathbf{W}_u \hat{\mathbf{Y}}^{(1)},$$

where $\epsilon_1 = (1, 0, \dots, 0)^T \in \mathbb{R}^{p+1}$, $\hat{\mathbf{Y}}^{(1)} = (\hat{Y}_{k+1}^{(1)}, \dots, \hat{Y}_{n-k}^{(1)})^T \in \mathbb{R}^{n-2k}$, and

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

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

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, $\widetilde{e}_i, i = 1, \dots, n$ are correlated in the model

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

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, $\widetilde{e}_i, i = 1, \dots, n$ are correlated in the model

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

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

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

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

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 $\frac{1}{n-2k} \sum_{i=k+1}^{n-k} \left(\hat{r}^{(1)}(U_{(i)}) - \hat{Y}_i^{(1)} \right)^2$ with a correction for applying the unimodal kernel in the final local polynomial smoothing.

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

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

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 $\frac{1}{n-2k} \sum_{i=k+1}^{n-k} \left(\hat{r}^{(1)}(U_{(i)}) - \hat{Y}_i^{(1)} \right)^2$ with a correction for applying the unimodal kernel in the final local polynomial smoothing.
- The asymptotic rate of convergence of $\hat{r}^{(1)}(u_0)$ given $\{U_{(i)}\}_{i=1}^n$ is

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

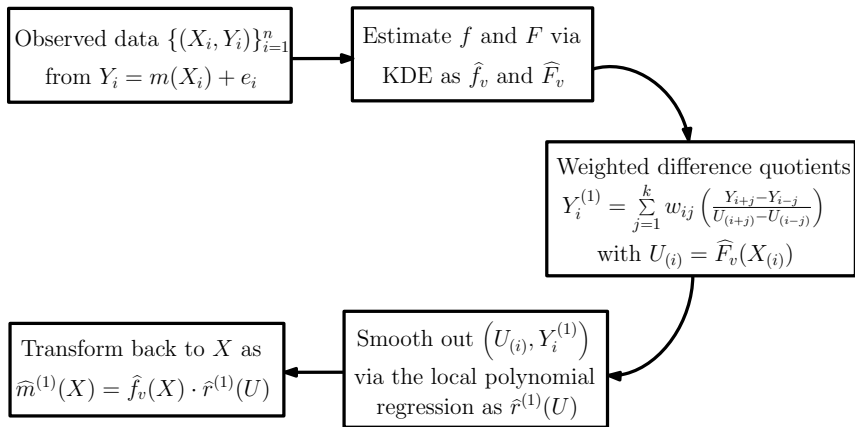
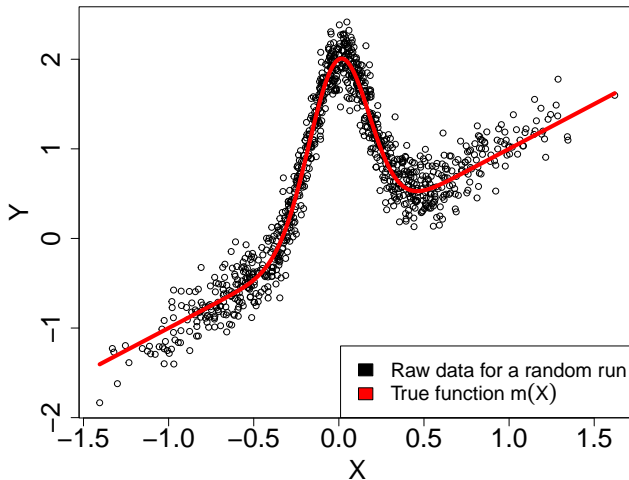
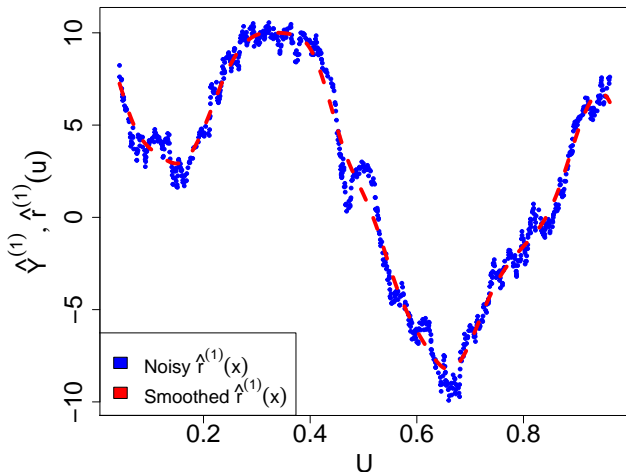


Figure 1: Summary of the proposed derivative estimation framework in the paper ([Liu and De Brabanter, 2020](#)).

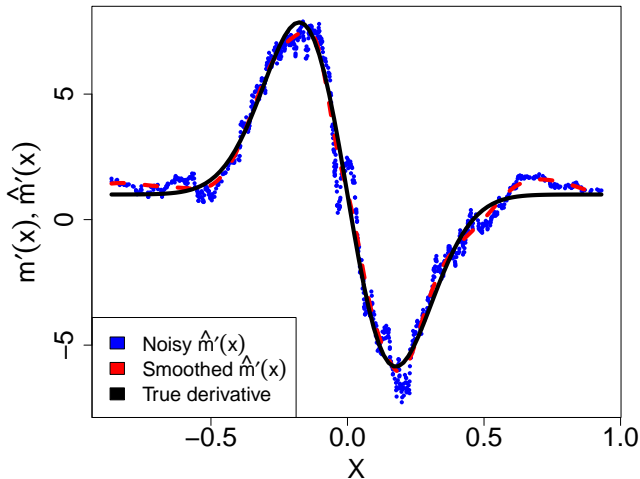
Simulated observations $\{(X_i, Y_i)\}_{i=1}^{1000}$ from $Y = m(X) + e$ with $m(X) = X + 2 \exp(-16X^2)$, $X \sim N(0, 0.5^2)$ and $e \sim N(0, 0.1^2)$



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

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

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

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

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

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

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

$$\hat{m}^{(1)}(x) = \hat{f}_v(x) \cdot \hat{r}^{(1)}(u)$$

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

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

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

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

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

$$\hat{m}^{(1)}(x) = \hat{f}_v(x) \cdot \hat{r}^{(1)}(u)$$

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

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

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 $q = 1, 2$,

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

Experiments



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](#)).

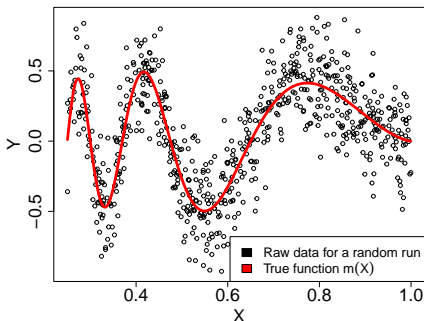
We compare the proposed derivative estimator in the paper ([Liu and De Brabanter, 2020](#)) with other existing derivative estimators as:

- **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.

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

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



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

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

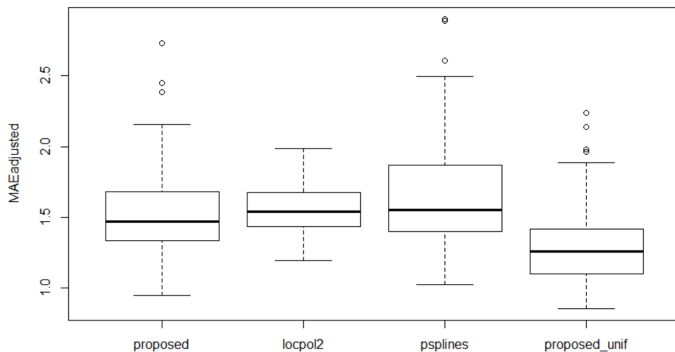


Figure 2: Comparative boxplots of adjusted mean absolute errors for the first-order derivative estimation methods under Monte Carlo simulation studies.

My extended experimental results are

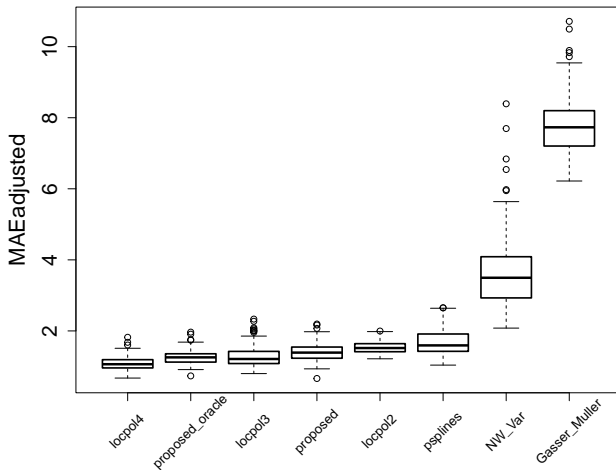
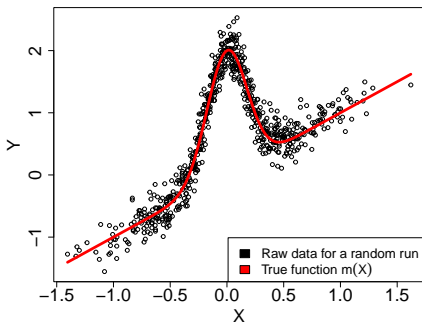


Figure 2: Comparative boxplots of adjusted mean absolute errors for the first-order derivative estimation methods under Monte Carlo simulation studies.

Beyond the uniform distribution of X , we also consider 100 Monte Carlo repeated simulations 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

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



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

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

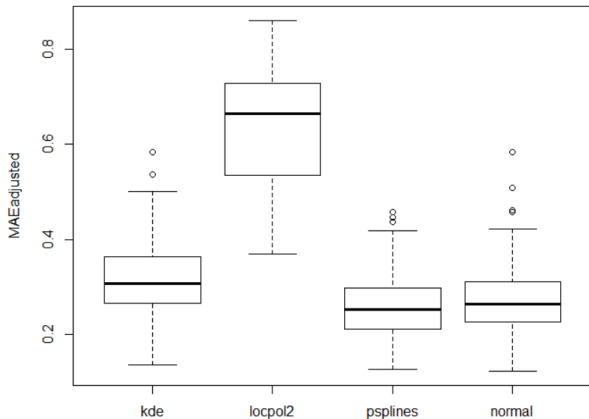


Figure 3: Comparative boxplots of adjusted mean absolute errors for the first-order derivative estimation methods under Monte Carlo simulation studies.

My extended experimental results are

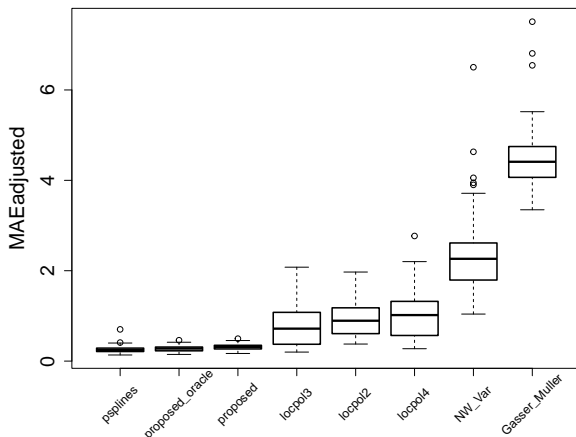
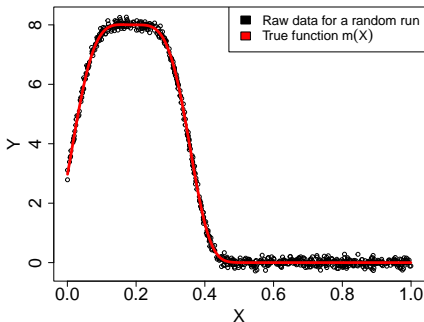


Figure 3: Comparative boxplots of adjusted mean absolute errors for the first-order derivative estimation methods under Monte Carlo simulation studies.

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

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



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

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

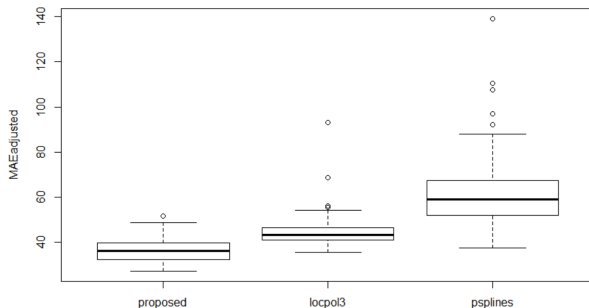


Figure 4: Comparative boxplots of adjusted mean absolute errors for the second-order derivative estimation methods under Monte Carlo simulation studies.

My extended experimental results are

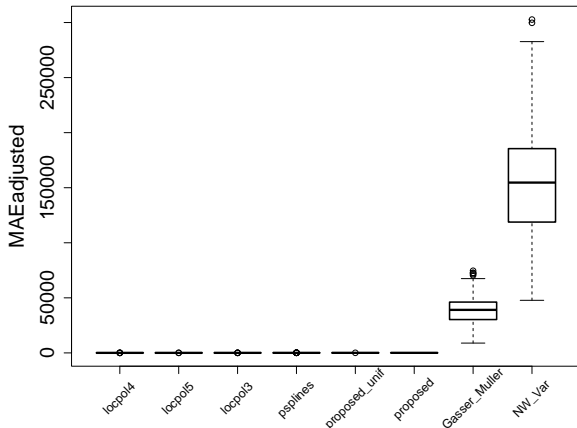


Figure 4: Comparative boxplots of adjusted mean absolute errors for the second-order derivative estimation methods under Monte Carlo simulation studies.

My extended experimental results are

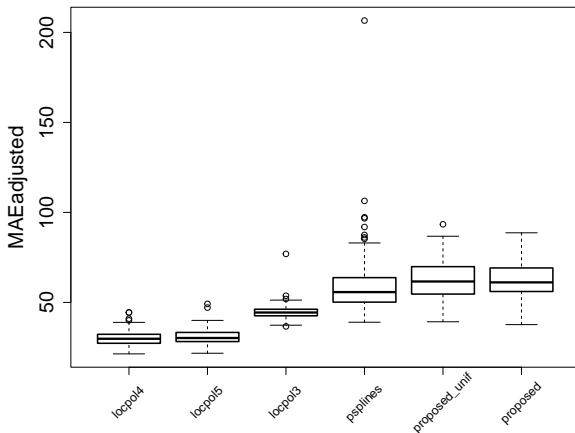


Figure 4: Comparative boxplots of adjusted mean absolute errors for the second-order derivative estimation methods under Monte Carlo simulation studies.

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.

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?

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 “Nope” 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 “Nope” 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?

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

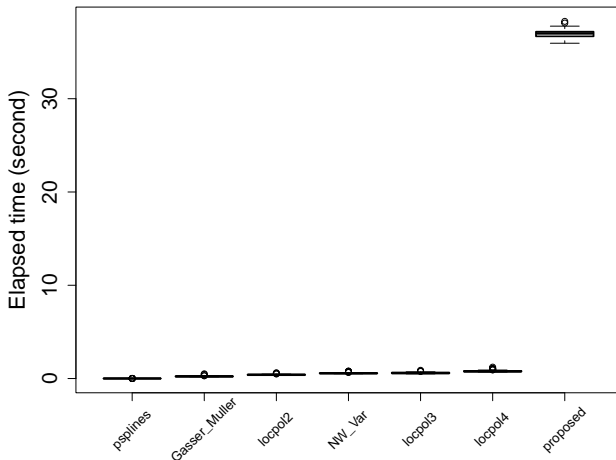


Figure 5: Time comparisons for different first-order derivative estimation methods under 100 repeated experiments.

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

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

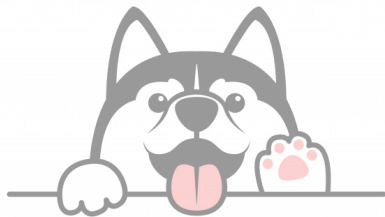
- 1 **Improving accuracy:** Smooth the data first by penalized smoothing splines (or other regression methods) before taking the noisy derivatives:

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

- 2 **Generalization to multivariate data:** [Dang \(2021\)](#) considered generalizing the proposed framework to multivariate data $\{(X_{i1}, \dots, X_{id}, Y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$ under the independence assumption between covariates.
 - **Open problem:** How can we estimate the (partial) derivatives of a multivariate regression function given the data $\{(X_{i1}, \dots, X_{id}, Y_i)\}_{i=1}^n$ 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 ed. 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.

- 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. Guillaou. Rates of strong uniform consistency for multivariate kernel density estimators. *Annales de l'Institut Henri Poincaré (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.
- 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.
- 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.

Recall that the proposed first-order noisy derivative estimator

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

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 $n-k+1 \leq i \leq n-1$.

Naive Solution:

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

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

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

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:

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

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

The paper adopts a two-step procedure proposed by [De Brabanter et al. \(2018\)](#) to select the final bandwidth \hat{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 \hat{h}_b by minimizing

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

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:

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

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

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

Assume that

- 1 The kernel function $K : \mathbb{R} \rightarrow [0, \infty)$ is bounded, symmetric, and Lipschitz continuous at 0. Furthermore, it satisfies $\lim_{|u| \rightarrow \infty} |u|^\ell K(u) < \infty$ for $\ell = 0, \dots, p$.
- 2 The correlation function ρ_n of the error terms $\tilde{e}_i, i = 1, \dots, n$ 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

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

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

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

$$\text{RSS}(h) = \text{SSE}(h) + \frac{1}{n-2k} \sum_{i=k+1}^{n-k} \tilde{e}_i^2 - \frac{2\sigma_e^2 \cdot K(0) \cdot (\mathbf{S}^{-1})_{11} \cdot (b-a+\rho_c)}{nh} + o_p \left(n^{-\frac{2p+2}{2p+3}} \right),$$

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

Here,

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

- 1 The kernel function for KDE $K_{\text{kde}} : \mathbb{R} \rightarrow [0, \infty)$ 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 $M(\mathcal{K}, L_2(Q), \epsilon)$ is the ϵ -covering number of the normed space $(\mathcal{K}, \|\cdot\|_{L_2(Q)})$, 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 differentiable 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 ρ'_n for the error terms \tilde{e}_i, \dot{e}_i in the local polynomial smoothing come from a first-order autoregressive process with $\mathbb{E}(|\tilde{e}_i|^\delta) < \infty, \mathbb{E}(|\dot{e}_i|^\delta) < \infty$ and are α -mixing with mixing coefficients $\alpha(k)$ such that $\sum_{k=1}^{\infty} k \cdot \alpha(k)^{1-\frac{2}{\delta}} < \infty$ for some $\delta > 2$. Moreover, define the sequence

$M_n = (n \log n (\log \log n)^{1+\gamma})^{\frac{1}{\delta}}$ for some $0 < \gamma < 1$. Then, the

bandwidth $h = h_n$ satisfies that $\gamma_n = \left(\frac{nM_n^2}{h_n^3 \log n} \right)^{\frac{1}{2}} \rightarrow \infty$ and

$b_n = \left(\frac{nh_n}{M_n^2 \log n} \right)^{\frac{1}{2}} \rightarrow \infty$ as $n \rightarrow \infty$. 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. \(2003\)](#).