DERIVATIVE ESTIMATION IN RANDOM DESIGN

{ YU LIU AND KRIS DE BRABANTER } DEPARTMENT OF COMPUTER SCIENCE DEPARTMENT OF STATISTICS

OBJECTIVES

Derivatives are important when doing data analysis, e.g. providing curve structure of data, analyzing significant trends and comparing regression functions.
We propose a nonparametric derivative estimation method for random design without having to estimate the regression function.

MODEL ASSUMPTIONS

Random design: Consider the data $(X_1,Y_1),\ldots,(X_n,Y_n)$ which form an independent and identically distributed (i.i.d.) sample from a population (X,Y), where $X_i \in \mathcal{X} = [a,b] \subseteq \mathbb{R}$ and $Y_i \in \mathbb{R}$ for all $i = 1,\ldots,n$. Consider the following model:

$$Y_i = m(X_i) + e_i, \quad i = 1,\ldots,n, \tag{1}$$

where $m(x) = \mathbf{E}[Y|X = x]$ and assume $\mathbf{E}[e] = 0$, $\mathbf{Var}[e] = \sigma_e^2 < \infty$, X and e are independent.

First, consider a special case in random design. n bivariate data (U,Y) are i.i.d and further assume $U \sim \mathcal{U}(0,1)$.
Sort the bivariate data (U,Y) by magnitude of U i.e. $U_{(1)} < U_{(2)} < \ldots < U_{(n)}$. Consider the following model:

$$Y_i = r(U_{(i)}) + e_i, \tag{2}$$

where $r(u) = \mathbf{E}[Y|U = u]$ and assume $\mathbf{E}[e] = 0$, $\mathbf{Var}[e] = \sigma_e^2 < \infty$, U and e are independent.

REFERENCES

[1] De Brabanter et al. Derivative estimation with local polynomial fitting. *Journal of Machine Learning Research*, 2013.

[2] R. Charnigo et al. A generalized c p criterion for derivative estimation. *Technometrics*, 2011.

[3] K. De Brabanter, F. Cao, I. Gijbels, and J. Opsomer. Local polynomial regression with correlated errors in random design and unknown correlation structure, in press. *Biometrika*, 2018.

FIRST ORDER DERIVATIVE BASED ON ORDER STATISTICS

Lemma 1 *Let $U \overset{i.i.d.}{\sim} \mathcal{U}(0,1)$ and arrange the random variables in order of magnitude $U_{(1)} < U_{(2)} < \cdots < U_{(n)}$. Then, for $i > j$*

$$U_{(i+j)} - U_{(i-j)} = \frac{2j}{n+1} + O_p\left(\sqrt{\frac{j}{n^2}}\right)$$

Müller et.al (1987) introduced:

$$\hat{q}_i^{(1)} = \frac{Y_i - Y_{i-1}}{U_{(i)} - U_{(i-1)}}$$

However, the estimator has a large variance.

$$\mathbf{Var}[\hat{q}_i^{(1)}|U_{(i-1)}, U_{(i)}] = O_p(n^2).$$

In order to reduce the variance, we extend the idea of [2] and [1] to random design involving uniform order statistics:

$$\hat{Y}_i^{(1)} = \sum_{j=1}^{k} w_{i,j} \cdot \left(\frac{Y_{i+j} - Y_{i-j}}{U_{(i+j)} - U_{(i-j)}}\right) \tag{3}$$

where the weights $w_{i,1},\ldots,w_{i,k}$ sum up to one and k controls the smoothness.

Proposition 1 *For $k+1 \le i \le n-k$ and under model (2), the weights $w_{i,j}$ that minimize the variance of (3), satisfying $\sum_{j=1}^{k} w_{i,j} = 1$, are given by*

$$w_{i,j} = \frac{(U_{(i+j)} - U_{(i-j)})^2}{\sum_{l=1}^{k}(U_{(i+l)} - U_{(i-l)})^2} \tag{4}$$

In what follows we denote $\mathbb{U} = (U_{(i-j)},\ldots,U_{(i+j)})$ for $i > j$ and $j = 1,\ldots,k$.

Theorem 1 *Under model (2) and assume $r(\cdot)$ is twice continuously differentiable on $[0,1]$ and $k \to \infty$ as $n \to \infty$. Then, for uniform random design on $[0,1]$ and for the weights in Proposition 1, the conditional (absolute) bias and conditional variance of (3) are given by*

$$\begin{aligned}\left|\text{bias}[\hat{Y}_i^{(1)}|\mathbb{U}]\right| &\le \mathcal{B}\frac{3k(k+1)}{4(n+1)(2k+1)} \\ &+ o_p(n^{-1}k)\end{aligned}$$

denote $\mathcal{B} = \sup_{u\in[0,1]}|r^{(2)}(u)|$ and

$$\mathbf{Var}[\hat{Y}_i^{(1)}|\mathbb{U}] = \frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)} + o_p(n^2k^{-3})$$

uniformly for $k+1 \le i \le n-k$.

Corollary 1 *Under the assumptions of Theorem 1 and denote $\mathcal{B} = \sup_{u\in[0,1]}|r^{(2)}(u)|$, then the k that minimizes asymptotic upper bound of MISE is*

$$\begin{aligned}k_{\text{opt}} &= \underset{k\in\mathbb{N}^+\setminus\{0\}}{\arg\min}\left\{\mathcal{B}^2\frac{9k^2(k+1)^2}{16(n+1)^2(2k+1)^2}\right. \\ &+ \left.\frac{3\sigma_e^2(n+1)^2}{k(k+1)(2k+1)}\right\} = O(n^{4/5}).\end{aligned}$$

Corollary 2 *Under the assumptions of Theorem 1, $k \to \infty$ as $n \to \infty$ such that $n^{-1}k \to 0$ and $n^2k^{-3} \to 0$. Then, for $\sigma_e^2 < \infty$ and the weights given in Proposition 1, we have for any $\varepsilon > 0$*

$$\mathbf{P}(|\hat{Y}_i^{(1)} - r^{(1)}(U_{(i)})| \ge \varepsilon) \to 0$$

for $k+1 \le i \le n-k$.

OTHER DISCUSSIONS

- Boundary Correction:
  $k(i) = i-1$ for $i < k+1$ and $k(i) = n-i$ for $i > n-k$.
- Smoothing empirical derivatives:
  Use local polynomial regression with a bimodal kernel K such that $K(0) = 0$ to remove dependence among the errors [3].
- Generalizing first order derivatives to any continuous distribution:

$$\begin{aligned}F(X) &\sim U(0,1) \\ m^{(1)}(X) &= f(X)r^{(1)}(U)\end{aligned}$$

SIMULATIONS

Consider the following function

$$m(X) = \cos^2(2\pi X), \quad X \sim \text{beta}(2,2),$$

with sample size $n = 700$ and $e \sim N(0, 0.2^2)$.

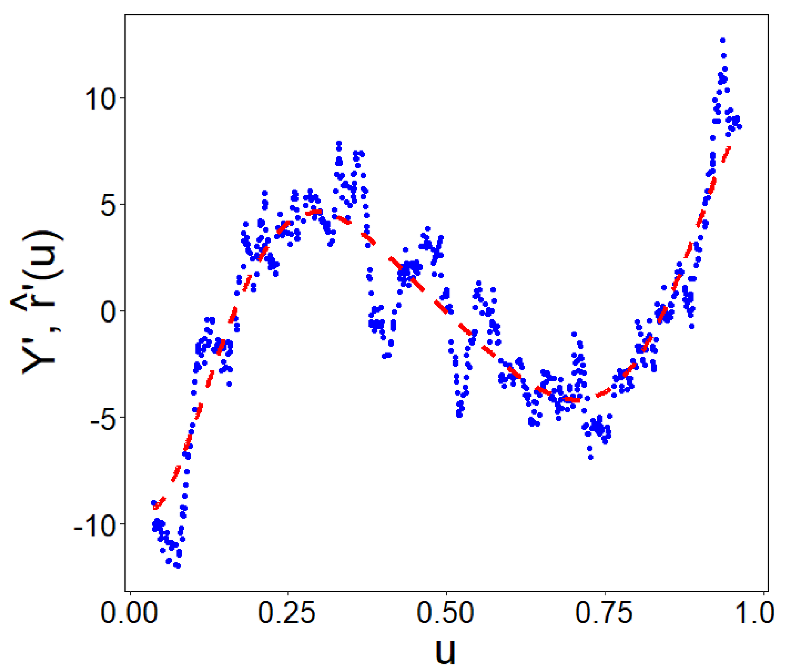

Figure 1: noisy derivative (dots), smoothed derivative (dash line)

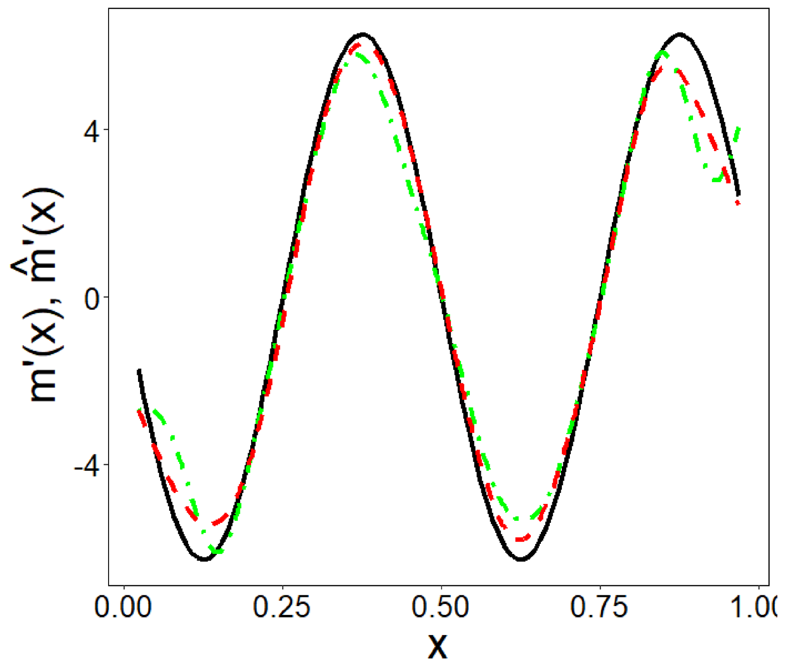

Figure 2: true derivative (full line), proposed method (dashed line) and localpol(dash-dotted line)

Consider the following function:

$$m(X) = \sqrt{X(1-X)}\sin\{(2.1\pi)/(X+0.05)\}$$

with $X \sim \mathcal{U}(0.25,1)$, sample size $n = 700$ and $e \sim N(0, 0.2^2)$. Rpeat 100 times and compare with other methods: