Efficient estimation of Sobol' indices of any order from a single input/output sample

Joint work with Sébastien Da Veiga, Fabrice Gamboa, Thierry Klein, and Clémentine Prieur

Agnès Lagnoux
Institut de Mathématiques de Toulouse

12ème Biennale Française des Mathématiques Appliquées et Industrielles, Carcans-Maubuisson, June 2-6, 2025

Outline of the talk

Introduction

- Framework and Sobol' indices
- The classical Pick-Freeze estimation
- Estimation from a single input/output sample

Efficient estimation from a single input/output sample

- Two main ingredients
- Our efficient mirrored high-order kernel-based estimate
- Main results

Sketch of the proofs

Numerical applications

- The Bratley function
- The g-Sobol function

Outline of the talk

Introduction
- Framework and Sobol' indices
- The classical Pick-Freeze estimation
- Estimation from a single input/output sample

Efficient estimation from a single input/output sample
- Two main ingredients
- Our efficient mirrored high-order kernel-based estimate
- Main results

Sketch of the proofs

Numerical applications
- The Bratley function
- The g-Sobol function

Framework

Complicated function f valued in $\mathbb{R}^k$ depending on several variables :

$$y = f(v_1, \ldots, v_p) \in \mathbb{R}^k$$

where

1. the inputs v_i pour $i = 1, \ldots p$ are objects ;
2. f is deterministic and unknown. It is called a black-box model.

Aim

Generally,

1. f is not analytically known ;
2. given $(v_1, \ldots, v_p)$, the computer code gives $y = f(v_1, \ldots, v_p)$;
3. computing $y = f(v_1, \ldots, v_p)$ may be costly.

Wishes :

1. Evaluate y for any value of the p-uplet $(v_1, \ldots, v_p)$.
2. Identify the most important variables to be able to fix the less important ones to their nominal value.

Probabilistic frame

In order to quantify the influence of a variable, it is common to assume that the inputs are random :

$$V := (V_1, \ldots, V_p) \in \mathcal{E}^p.$$

Then $f : \mathcal{E}^p \to \mathbb{R}^k$ is a deterministic measurable function evaluable on runs and the output code Y becomes random too :

$$Y = f(V_1, \ldots, V_p).$$

Main assumptions

1. *The inputs $V_1, \ldots, V_p \in \mathcal{E}$ are independent.*
2. *The output Y is scalar with a finite second moment.*

First toy example

Let have a look on a simple example :

$$(V_1, V_2, V_3) \mapsto Y = V_1 + V_1 V_2.$$

Obviously,

1. Y is not depending on V_3 ;
2. V_1 should be more influent than V_2 as it appears once alone (term V_1) and once related to V_2 (term $V_1 V_2$).

An input variable is influent if its variations leads to strong variations on the output.

$\Rightarrow$ *Build an index of influence on the variance of the output*

The so-called Sobol' indices

Quantification of the amount of randomness that a variable or a group of variables bring to Y => so-called Sobol' indices.

Such indices stem from the Hoeffding decomposition of the variance of f (or equivalently Y) that is assumed to lie in L^2.

Let $\mathbf{u}$ be a subset of $\{1,\ldots,p\}$ and $\sim\mathbf{u}$ its complementary in $\{1,\ldots,p\}$: $\sim\mathbf{u} = \{1,\cdots,p\}\setminus\mathbf{u}$.

Let denote $V_{\mathbf{u}} = (V_i, i \in \mathbf{u})$ and $V_{\sim\mathbf{u}} = (V_i, i \in \sim\mathbf{u})$.

From Hoeffding decomposition to Sobol indices

The decomposition of the output Y gives

$$
\begin{aligned}
Y := f(V) = \; & \underbrace{\mathbb{E}[Y]}_{\text{Mean effect}} \\
& + \underbrace{\mathbb{E}[Y|V_{\mathbf{u}}] - \mathbb{E}[Y] + \mathbb{E}[Y|V_{\sim\mathbf{u}}]) - \mathbb{E}[Y]}_{\text{First order effects}} \\
& + \underbrace{Y - (\mathbb{E}[Y] + \mathbb{E}[Y|V_{\mathbf{u}}] - \mathbb{E}[Y] + \mathbb{E}[Y|V_{\sim\mathbf{u}}] - \mathbb{E}[Y])}_{\text{Second order effects or interaction}:=IA}.
\end{aligned}
$$

Factors in the decomposition being orthogonal in L^2, one may compute the variance on both sides,

$$
\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}[Y|V_{\mathbf{u}}]) + \mathrm{Var}(\mathbb{E}[Y|V_{\sim\mathbf{u}}]) + \mathrm{Var}(IA).
$$

From Hoeffding decomposition to Sobol indices

This is the so-called Hoeffding decomposition of f. Dividing by $\mathrm{Var}(Y)$, one gets

$$1 = \frac{\mathrm{Var}\left(\mathbb{E}[Y|V_{\mathbf{u}}]\right)}{\mathrm{Var}(Y)} + \frac{\mathrm{Var}\left(\mathbb{E}[Y|V_{\sim\mathbf{u}}]\right)}{\mathrm{Var}(Y)} + \frac{\mathrm{Var}(IA)}{\mathrm{Var}(Y)}$$

$$:= S^{\mathbf{u}} + S^{\sim\mathbf{u}} + S^{\mathbf{u},\sim\mathbf{u}} \qquad \Rightarrow \quad \text{Sobol indices}$$

☞ $S^{\mathbf{u}} = \frac{\mathrm{Var}(\mathbb{E}[Y|V_{\mathbf{u}}]))}{\mathrm{Var}(Y)}$ *quantifies the first order effect of* $V_{\mathbf{u}}$,

while $S^{\mathbf{u}} + S^{\mathbf{u},\sim\mathbf{u}}$ *quantifies the total effect of* $V_{\mathbf{u}}$.

First toy example (continued)

We consider again

$$Y = f(V) = V_1 + V_1 V_2$$

where $V = (V_1, V_2, V_3) \sim \mathcal{N}_3(0, I_3)$. Then

$$\left(S^1, S^2, S^3, S^{1,2}\right) = (1/2, 0, 0, 1/2).$$

Pick-Freeze estimation of Sobol' indices (I)

To fix ideas assume e.g. $p=5$, $\mathbf{u}=\{1,2\}$ so that $\sim\mathbf{u}=\{3,4,5\}$.
We consider the Pick-Freeze variable $Y^{\mathbf{u}}$ defined as follows :

- draw $V=(V_1,V_2,V_3,V_4,V_5)$,
- build $V^{\mathbf{u}}=\left(V_1,V_2,V_3',V_4',V_5'\right)$.

Then, we compute

- $Y=f(V)$,
- $Y^{\mathbf{u}}=f(V^{\mathbf{u}})$.

A small miracle

$$\mathrm{Var}(\mathbb{E}[Y|V_{\mathbf{u}}])=\mathrm{Cov}(Y,Y^{\mathbf{u}}) \text{ so that } S^{\mathbf{u}}=\frac{\mathrm{Cov}(Y,Y^{\mathbf{u}})}{\mathrm{Var}(Y)}.$$

Pick-Freeze estimation of Sobol' indices (II)

In practice, generate two n-samples :

- one n-sample of V : $(V_j)_{j=1,\dots,n}$,
- one n-sample of $V^{\mathbf{u}}$: $\left(V_j^{\mathbf{u}}\right)_{j=1,\dots,n}$.

Compute the code on both samples :

- $Y_j = f(V_j)$ for $j = 1, \dots, n$,
- $Y_j^{\mathbf{u}} = f(V_j^{\mathbf{u}})$ for $j = 1, \dots, n$.

Then estimate $S^{\mathbf{u}}$ by

$$S_{n,PF}^{\mathbf{u}} = \frac{\frac{1}{n}\sum_{j=1}^{n} Y_j Y_j^{\mathbf{u}} - \left(\frac{1}{n}\sum_{j=1}^{n} Y_j\right)\left(\frac{1}{n}\sum_{j=1}^{n} Y_j^{\mathbf{u}}\right)}{\frac{1}{n}\sum_{j=1}^{n}(Y_j)^2 - \left(\frac{1}{n}\sum_{j=1}^{n} Y_j\right)^2}$$

Pick-Freeze scheme (III) : some statistical properties

Is the Pick-Freeze estimator of the Sobol' index is "good"?

- Is it consistent ? YES SLLN.
- If yes, at which rate of convergence ? YES CLT (cv in $\sqrt{n}$).
- Is it asymptotically efficient ? YES.
- Is it possible to measure its performance for a fixed n ? YES Berry-Esseen and/or concentration inequalities.

Ref. : A. Janon, T. Klein, A. Lagnoux, M. Nodet, and C. Prieur. " Asymptotic normality et efficiency of a Sobol' index estimator", *ESAIM P&S*, 2013.

F. Gamboa, A. Janon, T. Klein, A. Lagnoux, and C. Prieur. " Statistical Inference for Sobol' Pick Freeze Monte Carlo method", *Statistics*, 2015.

Drawbacks of the Pick-Freeze estimation

- The cost (= number of evaluations of the function f) of the estimation of the p first-order Sobol' indices is quite expensive : $(p+1)n$.
- This methodology is based on a particular design of experiment that may not be available in practice. For instance, when the practitioner only has access to real data.

☞ *We are interested in an estimator based on a n-sample only.*

Mighty estimation based on ranks (I)

Here we assume that

> *the inputs V_i for $i = 1, \ldots, p$ are scalar ($\dim(\mathscr{E}) = d = 1$)*

and we want to estimate the Sobol' index with respect to $X = V_i$:

$$S^i = \frac{\mathrm{Var}(\mathbb{E}[Y|V_i])}{\mathrm{Var}(Y)} = \frac{\mathrm{Var}(\mathbb{E}[Y|X])}{\mathrm{Var}(Y)}.$$

To do so, we consider a n-sample of the input/output pair (X, Y) given by

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n).$$

The pairs $(X_{(1)}, Y_{(1)}), (X_{(2)}, Y_{(2)}), \ldots, (X_{(n)}, Y_{(n)})$ are rearranged in such a way that

$$X_{(1)} < \ldots < X_{(n)}.$$

Mighty estimation based on ranks (II)

We introduce

$$S^i_{n,Rank} = \frac{\frac{1}{n}\sum_{j=1}^{n-1} Y_{(j)}Y_{(j+1)} - \left(\frac{1}{n}\sum_{j=1}^{n} Y_j\right)^2}{\frac{1}{n}\sum_{j=1}^{n} Y_j^2 - \left(\frac{1}{n}\sum_{j=1}^{n} Y_j\right)^2}.$$

Statistical properties - only for $d=1$ and first-order Sobol' indices

Consistency and CLT : OK.

Ref. : S. Chatterjee. "A new coefficient of Correlation", *JASA*, 2020.

F. Gamboa, P. Gremaud, T. Klein, and A. Lagnoux. " Global Sensitivity Analysis : a new generation of mighty estimators based on rank statistics", *Bernoulli.* 2022.

Efficient estimation based on kernels

Here again we assume that the inputs V_i for $i = 1, \ldots, p$ are scalar.

To do so, the initial n-sample is split into two samples of sizes

- $n_1 = \lfloor n/\log n \rfloor \Rightarrow$ *estimation of the joint density of* (V_i, Y)
- $n_2 = n - n_1 \approx n \Rightarrow$ *Monte-Carlo estimation of the integral involved in the quantity of interest.*

Statistical properties - only for $d = 1$ and first-order Sobol' indices
Consistency, CLT, and asymptotic efficiency : OK.

Ref. : S. Da Veiga and F. Gamboa. "Efficient estimation of sensitivity indices", *Journal of Nonparametric Statistics*, 2013.

Estimation based on nearest neighbors

Here the input $X = V_{\mathbf{u}}$, $\mathbf{u} \subset \{1, \cdots, p\}$ with respect we want to compute the Sobol' index is allowed to have dimension $d \geqslant 1$.

To do so, the initial n-sample is split into two samples of sizes

- *$n/2 \Rightarrow$ estim. of the regression function $m(x) = \mathbb{E}[Y|X = x]$ using the first NN of x among the points of the first sample ;*
- *$n/2 \Rightarrow$ plug-in estimator.*

Statistical properties - Consistency and CLT : OK only for $d \leqslant 3$.

Ref. : L. Devroye, L. Györfi, G. Lugosi, and H. Walk. "A nearest neighbor estimate of the residual variance", *EJS*, 2018.

Outline of the talk

Introduction
Framework and Sobol' indices
The classical Pick-Freeze estimation
Estimation from a single input/output sample

Efficient estimation from a single input/output sample
Two main ingredients
Our efficient mirrored high-order kernel-based estimate
Main results

Sketch of the proofs

Numerical applications
The Bratley function
The g-Sobol function

Framework

Recall that

$$S^X = \frac{\mathrm{Var}(\mathbb{E}[Y|X])}{\mathrm{Var}(Y)} = \frac{\mathbb{E}[\mathbb{E}[Y|X]^2] - \mathbb{E}[Y]^2}{\mathrm{Var}(Y)}$$

allowing a multidimensional $X = V_{\mathbf{u}}$ with $\mathbf{u} \subset \{1, \cdots, p\}$:
$X \in \mathscr{D} = [0,1]^d$.

☞ *Thus we focus on the estimation of* $T = \mathbb{E}[\mathbb{E}[Y|X]^2]$ *from the n-sample* $(X_j, Y_j)_{j=1,\dots,n}$ *of the pair* (X, Y).

Our estimator

• Starting point : if the regression function m is known, an asymptotically efficient estimator is (cf. Lagnoux et al. (2024))

$$T_{n,\text{oracle}} = \frac{1}{n}\sum_{i=1}^{n}(2Y_i - m(X_i))m(X_i).$$

• Our goal : build an estimator such that $\widehat{T}_n = T_{n,\text{oracle}} + o_{\mathbb{P}}(n^{-1/2})$.
$\Rightarrow$ CLT/AE through CLT/AE of oracle and Slutsky's theorem.

• Idea : a plug-in estimator of the form

$$\widehat{T}_n = \frac{1}{n}\sum_{i=1}^{n}(2Y_i - \widehat{m}_n(X_i))\widehat{m}_n(X_i).$$

Of course, there is a lot of work to obtain the required control !

Two main ingredients

We propose to estimate the regression function m with a kernel-based estimator.

1. Standard Nadaraya-Watson with usual kernels is doomed by dimensionality.
2. If inputs have compact support, kernels have known boundary issues.

Two main ingredients

We propose to estimate the regression function m with a kernel-based estimator.

1. Standard Nadaraya-Watson with usual kernels is doomed by dimensionality.

☞ *We rely on high-order kernels with regularity assumptions on the output.*

2. If inputs have compact support, kernels have known boundary issues.

☞ *We leverage mirror transformations and derive new useful convergence lemmas.*

First ingredient : high-order kernels

(Symmetric) high-order kernels in a nutshell : $k: [-1,1] \to \mathbb{R}$ bounded : $\|k\|_\infty < \infty$ is a univariate kernel of order $\nu + 1$ if :

$$\int_{-1}^{1} k(u)du = 1,$$
$$\int_{-1}^{1} u^\ell k(u)du = 0, \text{ for any } \ell \in \mathbb{N} \text{ such that } 0 < \ell \leq \nu$$
$$\int_{-1}^{1} u^{\nu+1} k(u)du \neq 0.$$

Commonly used kernels (Gaussian, Epanechnikov,...) are of order 2. Finally,

$$K_h(u) = \frac{1}{h^d} K\left(\frac{u}{h}\right) = \frac{1}{h^d} \prod_{k=1}^{d} k\left(\frac{u_k}{h}\right), \forall u \in [-1,1]^d.$$

First ingredient : why high-order kernels ? (I)

For kernel density estimation, bias is (multivariate Taylor)

$$Bias = \mathbb{E}[\widehat{f}(x)] - f(x) = \sum_{1 \leqslant |\beta| < \nu} \frac{h^{|\beta|}}{\beta\,!} \frac{\partial^\beta f}{\partial x^\beta}(x) \kappa_{1,\beta}(k)$$

$$+ h^\nu \sum_{|\beta| = \nu} \kappa_{2,\beta}(k) \quad as\ h \to 0$$

with the multi-index notation : $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}_+^d$,
$|\beta| = \beta_1 + \cdots + \beta_d$, and $\beta! = \beta_1! \ldots \beta_d!$.

☞ f ($=$ density of X in the sequel) with sufficient regularity.

First ingredient : why high-order kernels ? (II)

For kernel density estimation, bias is (multivariate Taylor)

$$Bias = \mathbb{E}[\widehat{f}(x)] - f(x) = \sum_{1 \leqslant |\beta| < \nu} \frac{h^{|\beta|}}{\beta!} \frac{\partial^\beta f}{\partial x^\beta}(x) \underbrace{\kappa_{1,\beta}(k)}_{\text{with a high-order kernel, this term can cancel}}$$

$$+ h^\nu \sum_{|\beta| = \nu} \underbrace{\kappa_{2,\beta}(k)}_{\text{remainder term}} \quad \text{as } h \to 0$$

with the multi-index notation : $\beta = (\beta_1, \ldots, \beta_d) \in \mathbb{R}_+^d$, $|\beta| = \beta_1 + \cdots + \beta_d$, and $\beta! = \beta_1! \ldots \beta_d!$.

First ingredient : why high-order kernels ? (III)

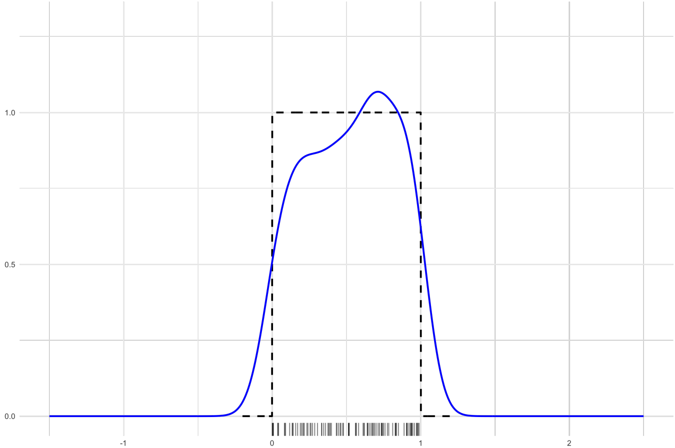
By analysing the variance (skipped here), with a high-order kernel, we finally get

$$
\begin{aligned}
AMISE &= \int_{\mathbb{R}^d} \mathbb{E}[(\widehat{f}(x) - f(x))^2] dx \\
&= O(n^{-\frac{2\nu}{2\nu+d}}) \quad \textit{if } h = O(n^{-\frac{1}{2\nu+d}}) \ \textit{(optimal bandwidth)} \\
&= o(n^{-\frac{1}{2}}) \qquad \textit{if } \nu > d/2
\end{aligned}
$$

☞ kernel with high enough order.

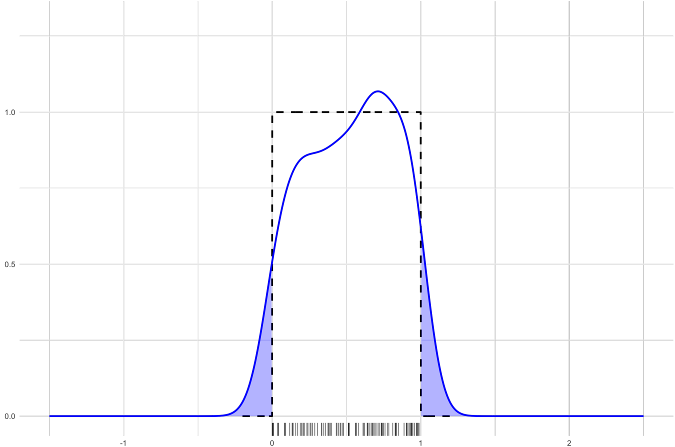
KDE boundary issues (I)

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right) \quad \text{and} \quad \int_{-\infty}^{\infty} \widehat{f}(x) = 1$$



KDE boundary issues (II)

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right) \quad \text{and} \quad \int_0^1 \hat{f}(x) < 1$$



A partial solution

Doksum and Samorov (1995) estimated a truncated version of T defined as

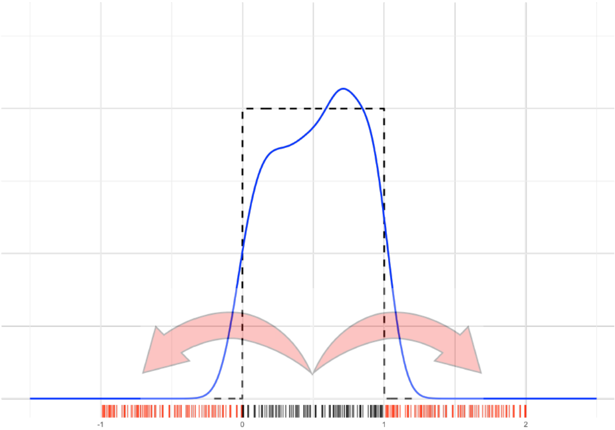
$$T^{\text{trunc},\varepsilon} = \mathbb{E}[\mathbb{E}[Y|X]^2 \mathbb{1}_{X\in(\varepsilon,1-\varepsilon)^d}].$$

Even if $T^{\text{trunc},\varepsilon} \to T$ as $\varepsilon \to 0$ under mild assumptions, the practical tuning of the parameter ε depends on the unknown function f and its choice has a large impact.

Here, we therefore focus on mirror-type kernel estimators to estimate T rather than a truncated version of it. Such mirror-type estimators have been proposed recently to efficiently handle boundary effects inherent to kernel estimation.

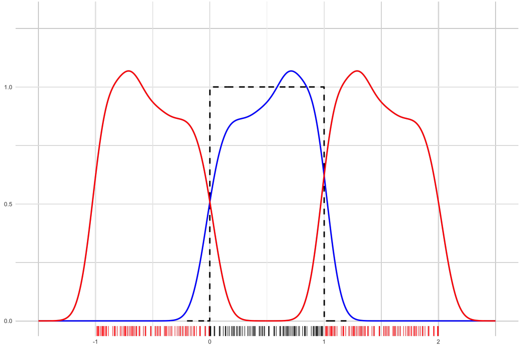
Second ingredient : mirror transformation (I)

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right) \quad \text{and} \quad \int_{-\infty}^{\infty} \widehat{f}(x) = 1$$



Second ingredient : mirror transformation (II)

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right) \quad \text{and} \quad \int_{-\infty}^{\infty} \widehat{f}(x) = 1$$

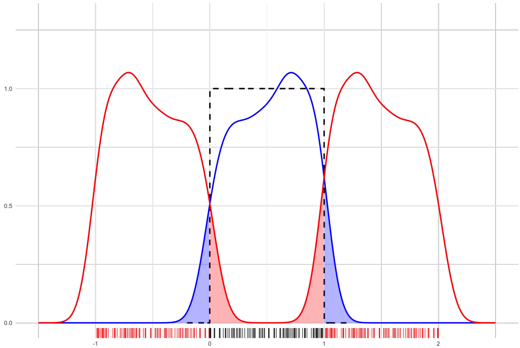


$\widehat{f}_{\text{lower}}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-(-X_i)}{h}\right)$

$\widehat{f}_{\text{upper}}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-(2-X_i)}{h}\right)$

Second ingredient : mirror transformation (III)

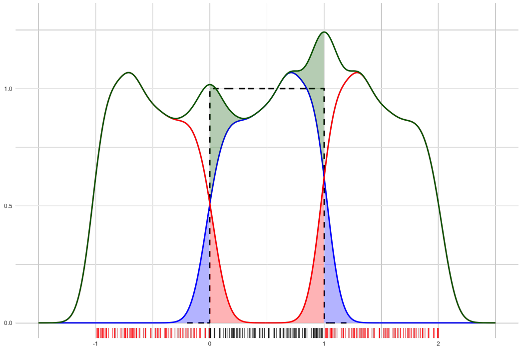
$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right) \quad \text{and} \quad \int_{-\infty}^{\infty} \widehat{f}(x) = 1$$



$$\widehat{f}_{\text{lower}}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-(-X_i)}{h}\right) \qquad \widehat{f}_{\text{upper}}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-(2-X_i)}{h}\right)$$

Second ingredient : mirror transformation (IV)

$$\widehat{f}_{\text{mirror}}(x) = \widehat{f}(x) + \widehat{f}_{\text{lower}}(x) + \widehat{f}_{\text{upper}}(x)$$

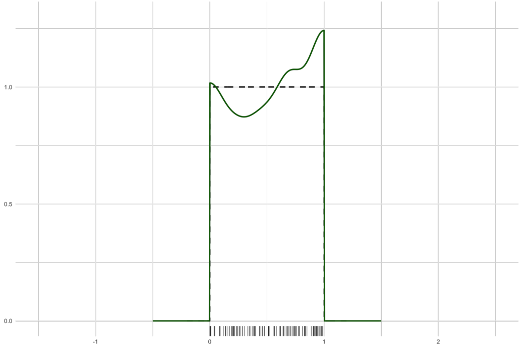


$$\widehat{f}_{\text{lower}}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-(-X_i)}{h}\right)$$

$$\widehat{f}_{\text{upper}}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-(2-X_i)}{h}\right)$$

Second ingredient : mirror transformation (V)

$$\widehat{f}_{\text{mirror}}(x) = \left(\widehat{f}(x) + \widehat{f}_{\text{lower}}(x) + \widehat{f}_{\text{upper}}(x)\right) \times \mathbb{1}_{x \in [0,1]} \quad \text{and} \quad \int_0^1 \widehat{f}_{\text{mirror}}(x) = 1$$



$$\widehat{f}_{\text{lower}}(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x-(-X_i)}{h}\right) \qquad \widehat{f}_{\text{upper}}(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x-(2-X_i)}{h}\right)$$

Our efficient mirrored high-order kernel-based estimate

As Pujol (2022), we consider the following 1D-transformations :

$$\forall z \in [0,1], m^{-1}(z) = -z, \quad m^0(z) = z, \quad and \quad m^1(z) = 2 - z$$

and, for any $a \in \{-1,0,1\}^d$ and $x \in [0,1]^d$, the d-dimensional vector

$$M^a(x) = (m^{a_1}(x_1), \cdots, m^{a_d}(x_d))$$

of mirrors in all possible directions.

Our efficient mirrored high-order kernel-based estimate

The mirrored density estimate of the density f_X of X is

$$\widehat{f}_{mirror}(x) = \frac{1}{nh_n^d} \sum_{j=1}^{n} \sum_{a \in \{-1,0,1\}^d} \prod_{l=1}^{d} k\Big(\frac{x_l - M^a(X_j)_l}{h_n}\Big) = \frac{1}{nh_n^d} \sum_{j=1}^{n} \sum_{a \in \{-1,0,1\}^d} K(x - M^a(X_j))$$

and its leave-one-out version :

$$\widehat{f}_{n,h_n,i}(x) = \frac{1}{nh_n^d} \sum_{j \neq i} \sum_{a \in \{-1,0,1\}^d} K(x - M^a(X_j)).$$

Our efficient mirrored high-order kernel-based estimate

Similarly, the leave-one-out (mirrored) Nadaraya-Watson estimate of the regression function is :

$$\widehat{m}_{n,h_n,i}(X_i) = \frac{\sum_{j\neq i} Y_j \sum_{a\in\{-1,0,1\}^d} K_{h_n}(X_i - M^a(X_j))}{\sum_{j\neq i} \sum_{a\in\{-1,0,1\}^d} K_{h_n}(X_i - M^a(X_j) -)} = \frac{\widehat{g}_{n,h_n,i}(X_i)}{\widehat{f}_{n,h_n,i}(X_i)}.$$

The associated plug-in estimator then becomes :

$$\widehat{T}_{n,h_n} = \frac{1}{n}\sum_{i=1}^{n}(2Y_i - \widehat{m}_{n,h_n,i}(X_i))\widehat{m}_{n,h_n,i}(X_i).$$

Assumptions

($\mathscr{A}1$) Support - The support of $(V_1, \ldots, V_p)$ is $[0,1]^p$ and that of X is $[0,1]^d$.

($\mathscr{A}2$) Absolute continuity - X is absolutely continuous with respect to the Lebesgue measure on $[0,1]^d$ with density function f_X and $\exists \delta > 0$ such that $\inf_{x \in [0,1]^d} f_X(x) \geqslant \delta$ for some $\delta > 0$.

($\mathscr{A}3$) Bounded moments - $\mathbb{E}[Y^4] < \infty$ and $\sigma^2(x) = \mathrm{Var}(Y|X = x)$ is bounded on $[0,1]^d$.

($\mathscr{A}4$) Smoothness of f_X - $f_X \in \mathscr{C}^\alpha([0,1]^d)$ for some $\alpha > 0$ and its derivatives of order β $(0 < \beta \leq \lfloor \alpha \rfloor)$ vanish near the boundary.

($\mathscr{A}5$) Smoothness of m - The regression function m belongs to $\mathscr{C}^\alpha([0,1]^d)$.

($\mathscr{A}6$) Kernel - $k : [-1,1] \to \mathbb{R}$ is a bounded univariate kernel of order $(\nu + 1)$ $(\nu = \lfloor \alpha \rfloor)$.

Under the previous assumptions and an additional technical one, for all $i \in \{1, \cdots, d\}$, we get :

- bias and variance controls

$$\left\| \mathbb{E}\left[\widehat{f}_{n,h_n,i}\right] - f_X \right\|_\infty = O\left(h_n^\alpha\right),$$
$$\mathbb{E}\left[\int_{[0,1]^d} (\widehat{f}_{n,h_n,i}(x) - f_X(x))^2 dx\right] = o(n^{-1/2}),$$

- lower control

$$\frac{1}{\inf_{x \in [0,1]^d} \left|\widehat{f}_{n,h_n,i}(x)\right|} = O_{\mathbb{P}}(1),$$

when $nh_n^{2d} \to \infty$ and $nh_n^{4\alpha} \to 0$ as $n \to \infty$.

Theorem (Central Limit Theorem and asymptotic efficiency)

Under the previous assumptions, one has (i)

$$\sqrt{n}\big(\widehat{T}_{n,h_n} - \mathbb{E}[\mathbb{E}[Y|X]^2]\big) \xrightarrow[n\to\infty]{\mathscr{L}} \mathcal{N}\big(0, \mathrm{Var}((2Y - m(X))m(X))\big)$$

as soon as $\alpha > d/2$ *and* $h_n = n^{-\gamma}$ *with* $1/(4\alpha) < \gamma < 1/(2d)$ *;*

(ii) $\widehat{T}_{n,h_n}$ *is asymptotically efficient to estimate* $\mathbb{E}[\mathbb{E}[Y|X]^2]$ *from an i.i.d. sample* $(X_i, Y_i)_{i=1,\cdots,n}$ *of the pair* (X, Y).

Ref. : S. Da Veiga, F. Gamboa, T. Klein, A. Lagnoux, C. Prieur. “Efficient estimation of Sobol’ indices of any order from a single input/output sample.”. Available on Hal and Arxiv (2024). `https://hal.science/hal-04052837v2`.

Using the delta method, we are now able to get the asymptotic behaviour of the estimation of S^X, letting

$$\widehat{S}_{n,h_n} = \frac{\widehat{T}_{n,h_n} - \left(\frac{1}{n}\sum_{j=1}^n Y_j\right)^2}{\frac{1}{n}\sum_{j=1}^n Y_j^2 - \left(\frac{1}{n}\sum_{j=1}^n Y_j\right)^2}.$$

Corollary (CLT & AE for the estimation of the Sobol' indices)

Under all the assumptions of the theorem, one has (i)

$$\sqrt{n}\left(\widehat{S}_{n,h_n} - S^X\right) \xrightarrow[n\to\infty]{\mathscr{L}} \mathcal{N}(0,\sigma^2),$$

where the limit variance σ^2 has an explicit expression.

(ii) $\widehat{S}_{n,h_n}$ is asymptotically efficient to estimate S^X from an i.i.d. sample $(X_i, Y_i)_{i=1,\cdots,n}$ of the pair (X, Y).

Let us denote S^i the first-order Sobol index associated to the i-th input and its estimator $\widehat{S}^i$ given by :

$$\widehat{S}^i_{n,h_n} = \frac{\widehat{T}^i_{n,h_n} - \left(\frac{1}{n}\sum_{j=1}^n Y_j\right)^2}{\frac{1}{n}\sum_{j=1}^n Y_j^2 - \left(\frac{1}{n}\sum_{j=1}^n Y_j\right)^2}.$$

Corollary (CLT & AE for the global estimation of the p first-order Sobol' indices)

Under all the assumptions of the theorem, one has

$$\sqrt{n}\Big((\widehat{S}^1_{n,h_n},\ldots,\widehat{S}^p_{n,h_n})^T - (S^1,\ldots,S^p)^T\Big) \xrightarrow[n\to\infty]{\mathscr{D}} \mathscr{N}(0,\Sigma),$$

where the limit variance Σ has an explicit expression. Furthermore, such estimation is asymptotically efficient.

Outline of the talk

Introduction
- Framework and Sobol' indices
- The classical Pick-Freeze estimation
- Estimation from a single input/output sample

Efficient estimation from a single input/output sample
- Two main ingredients
- Our efficient mirrored high-order kernel-based estimate
- Main results

Sketch of the proofs

Numerical applications
- The Bratley function
- The g-Sobol function

Sketch of the proof : CLT

Following the same lines as in the proof of Theorem 2.1 in Doksum (1995), we aim at proving that

$$\widehat{T}_{n,h} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(2Y_i - m(X_i))m(X_i)}_{=T_{n,oracle}} + o_{\mathbb{P}}(n^{-1/2}). \quad (1)$$

The conclusion of the theorem will then follow directly applying the standard central limit theorem for the sum of i.i.d. random variables to the right-hand side of the previous display together with Slutsky's lemma.

Sketch of the proof : asymptotic efficiency

The influence efficient function of ψ at P, as stated in Doksum (1995), is given by (see Klein (2024) for the details) :

$$\widetilde{\psi}_P(x,y) = (2y - m(x))m(x) - \mathbb{E}[Ym(X)].$$

Moreover, we deduce from (1) that

$$\widehat{T}_{n,h} = \psi(P) + \frac{1}{n}\sum_{i=1}^{n} \widetilde{\psi}_P(X_i, Y_i) + o_{\mathbb{P}}(n^{-1/2}) = T_{n,oracle} + o_{\mathbb{P}}(n^{-1/2})$$

and conclude using Condition (25.22) of Van der Vaart (2000).

Outline of the talk

Introduction
Framework and Sobol' indices
The classical Pick-Freeze estimation
Estimation from a single input/output sample

Efficient estimation from a single input/output sample
Two main ingredients
Our efficient mirrored high-order kernel-based estimate
Main results

Sketch of the proofs

Numerical applications
The Bratley function
The g-Sobol function

For all test cases :

- first-order and total-order Sobol' indices for each input variable V_i (i.e. $X = V_i$ and $X = V_{\sim i}$ resp.) ;
- mirror-type estimator with an Epanechnikov kernel of order 2 and 4 (kernel bandwidth optimized via LOO on m) ;
- concurrent estimators :
  - PF estimator studied (Janon'12) ("PF1")
  - replicated PF estimator (Tissot'15) ("PF2")
  - rank estimator (Gamboa'20) ("Rank") for 1st-order indices
  - lag estimator (Klein'24) ("Lag") for 1st-order indices
  - nearest-neighbour estimator (Devroye 2018) ("NN") ;
- we generate a n-sample $(X_1, Y_1), \cdots, (X_n, Y_n)$ (except for PF) ;
- each experiment is repeated 50 times with $n = 500$;
- the reference value is obtained from a PF estimation with very large sample size.

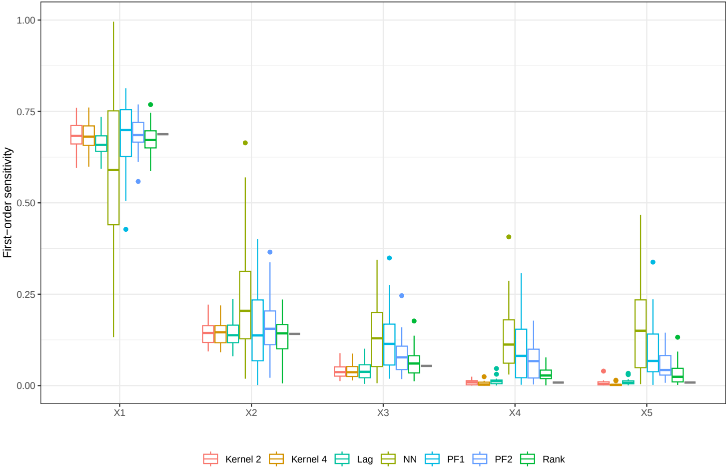
The Bratley function

First, we consider the Bratley function defined by :

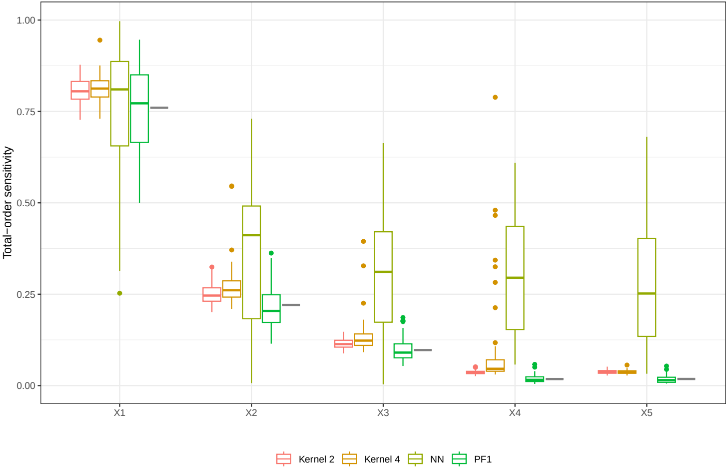
$$g_{\text{Bratley}}(V_1,\ldots,V_p) = \sum_{i=1}^{p} (-1)^i \prod_{j=1}^{i} V_j,$$

with $V_i \sim \mathscr{U}([0,1])$ i.i.d. and $p = 5$.

The Bratley function - first-order indices - $n = 500$



The Bratley function - total-order indices - $n = 500$



The g-Sobol function

We investigate the g-Sobol function defined by

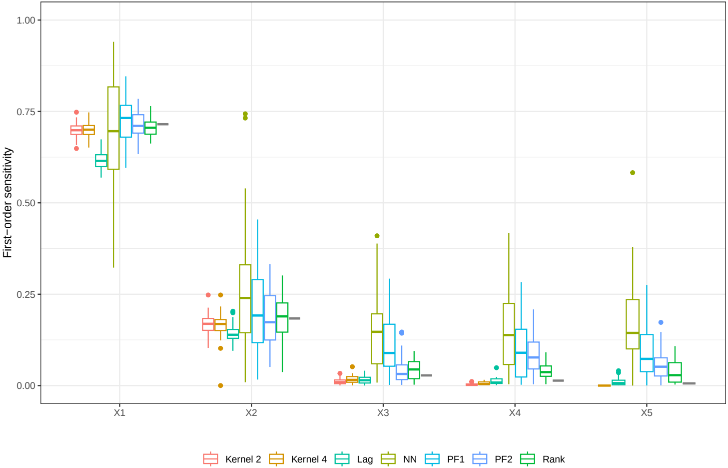
$$g_{\text{g-Sobol}}(V_1,\dots,V_p) = \prod_{i=1}^{p} \frac{|4V_i - 2| + a_i}{1 + a_i},$$

with $V_i \sim \mathscr{U}([0,1])$ i.i.d., $p = 5$ and $a = (0,1,4.5,9,99)$.

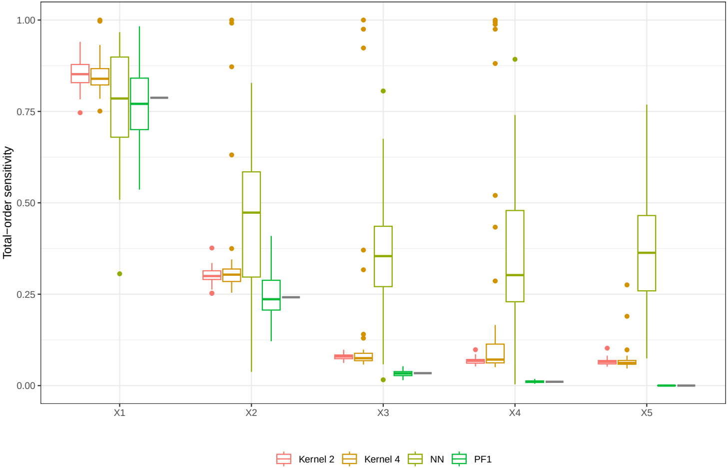
Notice that it is non-differentiable at any input value with a component equal to 0.5, but the impact on our estimator performance is negligible for first-order indices.

Except for the degraded performance of the lag estimator, the conclusions are the same as for the Bratley function, even for total indices.

The g-Sobol function - first-order indices - $n = 500$



The g-Sobol function - total-order indices - $n = 500$



Tuning of parameter ϵ

We illustrate numerically that the choice of the ϵ tuning parameter of the estimator proposed in Doksum (1995) is very sensitive, thus limiting its practical use as opposed to our mirror-type estimator.

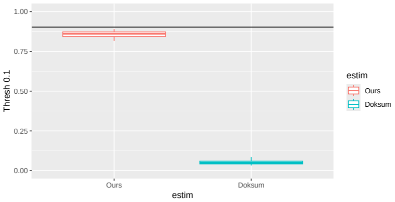
We consider Example 3.2 from Doksum and Samarov (1995) :

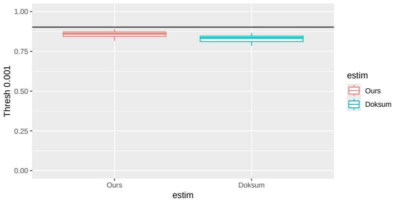
$$Y = \frac{1}{2} + 4X_1 + 4(X_2 - \frac{1}{2})^2 + 4X_3^{1/2} + \tau e,$$

with X_1, X_2, and X_3 i.i.d. $\sim \mathscr{U}([0,1])$ and $e \sim \mathscr{N}(0,1)$.

We test $\epsilon = 10^{-1}$ and 10^{-3}.

Tuning of parameter ϵ





When ϵ is equal to 10^{-3}, the performance of both estimators are similar. However when $\epsilon = 10^{-1}$, the bias of Doksum and Samarov (1995) can be very large. Since in practice such an estimation problem is unsupervised, the tuning of ϵ seems highly difficult and the non-robustness of the final estimator with respect to this parameter limits its practical use.

Thanks for your attention !
Questions ?

Reference

S. Da Veiga, F. Gamboa, T. Klein, A. Lagnoux, C. Prieur.
“Efficient estimation of Sobol’ indices of any order from a single input/output sample.”. Available on Hal and Arxiv (2024).
https://hal.science/hal-04052837v2.

Advertising

For your book project, think about the collection

Lecture Notes on Applied Deterministic and Stochastic Mathematics

published in open access, at no cost to authors and readers, under a CC-BY-NC licence retained by the authors.

Efficient influence function and asymptotic efficiency

Let $\mathscr{P}$ be the set of absolutely continuous probability distributions on $[0,1]^d \times \mathbb{R}$ and $P_0 \in \mathscr{P}$ be the probability distribution of (X, Y), such that we can write our target $T = \psi(P_0)$ where $\psi : \mathscr{P} \to \mathbb{R}$.

If ψ is differentiable at all $P \in \mathscr{P}$, the efficient influence function $\widetilde{\psi}_P : [0,1]^d \times \mathbb{R} \to \mathbb{R}$ is the gradient with smallest variance among all gradients of ψ at P with zero mean w.r.t. to P.

The link with efficient estimators is the following : a sequence of estimators T_n of $T = \psi(P_0)$ is asymptotically efficient iif

$$T_n - T = T_n - \psi(P_0) = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\psi}_{P_0}(X_i, Y_i) + o_{P_0}\left(\frac{1}{\sqrt{n}}\right),$$

See Eq.(25.22) in van der Vaart (2000).

Efficient influence function and asymptotic efficiency

In our case, the efficient influence function at any $P \in \mathscr{P}$ writes

$$\widetilde{\psi}_P(x,y) = (2y - m(x))m(x) - \psi(P).$$

where m is the regression function under P : $m(x) = \mathbb{E}_P[Y|X=x]$, see Klein, Lagnoux, Rochet (2024).

Then, if m under P_0 is known, taking

$$T_{n,oracle} = \frac{1}{n}\sum_{i=1}^{n}(2Y_i - m(X_i))m(X_i)$$

leads to an asymptotically efficient estimator of T.

Plug-in estimation

A first point of view consists in seeing

$$\widehat{T}_n = \frac{1}{n}\sum_{i=1}^{n}(2Y_i - \widehat{m}_n(X_i))\widehat{m}_n(X_i),$$

as a plug-in version of

$$T_{n,oracle} = \frac{1}{n}\sum_{i=1}^{n}(2Y_i - m(X_i))m(X_i)$$

where the difference $m - \widehat{m}_n$ needs to be controlled to still have

$$\widehat{T}_n = \psi(P_0) + \frac{1}{n}\sum_{i=1}^{n}\widetilde{\psi}_{P_0}(X_i, Y_i) + o_{P_0}\Big(\frac{1}{\sqrt{n}}\Big).$$

One-step estimation

A second point of view relies on one-step estimators, that consider a first-order bias correction of an initial estimator $\psi(\widehat{P})$ where $\widehat{P}$ is a smoothed estimate of P_0.

More precisely, a simple Taylor expansion of $\psi(P_0)$ around $\psi(\widehat{P})$ involves the efficient influence function $\widetilde{\psi}$ at $\widehat{P}$:

$$\psi(P_0) - \psi(\widehat{P}) = \mathbb{E}_{P_0}[\widetilde{\psi}_{\widehat{P}}] - \overbrace{\mathbb{E}_{\widehat{P}}[\widetilde{\psi}_{\widehat{P}}]}^{=0} + r_2(\widehat{P}, P) = \mathbb{E}_{P_0}[\widetilde{\psi}_{\widehat{P}}] + r_2(\widehat{P}, P)$$

since by definition, $\mathbb{E}_P[\widetilde{\psi}_P] = 0$ for all P. Thus, if $r_2(\widehat{P}, P) = o(1)$,

$$\psi(\widehat{P}) + \mathbb{E}_{P_0}[\widetilde{\psi}_{\widehat{P}}] \sim \psi(P_0).$$

One-step estimation

Thus it is possible to improve $\psi(\widehat{P})$ by considering an estimate of this first-order bias $\mathbb{E}_{P_0}[\widetilde{\psi}_{\widehat{P}}]$: for instance, $\mathbb{E}_{P_n}[\widetilde{\psi}_{\widehat{P}}]$ where P_n is the empirical distribution of the observations $(X_i, Y_i)_{i=1,\ldots,n}$.

In our particular case, this induces an estimator given by

$$\widehat{T}_n = \psi(\widehat{P}) + \mathbb{E}_{P_n}[\widetilde{\psi}_{\widehat{P}}] = \frac{1}{n}\sum_{i=1}^{n}(2Y_i - \widehat{m}(X_i))\widehat{m}(X_i)$$

where $\widehat{m}$ is the regression function under $\widehat{P}$, that is precisely a smoothing estimate of m. We can then hope that $\widehat{T}_n$ will be asymptotically efficient if the difference $\widehat{P} - P_0$ converges to 0 at an appropriate rate.

Construction of high-order kernels

The kernel k is typically chosen as a symmetric second-order kernel (Epanechnikov, Gaussian, ...) with the following properties :

$$\int k(u)du = 1, \quad \int uk(u)du = 0, \quad \int u^2 k(u) > 0.$$

The terminology second-order refers to the fact that the first non-zero moment of k is the second one (except for the zero-th order one which ensures the kernel is normalized).

Construction of high-order kernels

More generally, a high-order kernel of order r satisfies

$$\int k(u)du = 1, \quad \int u^j k(u)du = 0, \ \forall j = 1, \ldots, r-1, \quad \int u^r k(u) > 0.$$

Here, we will focus on high-order kernels with compact support, which are used together with mirror-type transformations to avoid boundary effects appearing when the domain is compact.

In particular, we will study symmetric kernels on $[-1,1]$ and non-symmetric ones on $[0,1]$.

Construction of high-order kernels

In order to build a kernel of order r with compact support $[-1,1]$, there are at least two approaches, which are described below.

Legendre orthonormal polynomials. The first construction relies on the (normalized) Legendre orthonormal polynomials on $[-1,1]$ denoted by $\{P_m(\cdot)\}_{m\in\mathbb{N}}$. Then we define the kernel k as

$$k(u) = \sum_{m=0}^{r+1} P_m(0) P_m(u) \mathbb{1}_{u\in[-1,1]}, \tag{2}$$

see Comte (2017).

Construction of high-order kernels

High-order Epanechnikov kernel. Hansen (2005) proposes a high-order generalization of smooth and second-order kernels on $[-1,1]$ including the uniform, biweight, and Epanechnikov ones. Focusing on the latter, the kernel

$$k(u) = B_r(u) k_e(u) \tag{3}$$

where $k_e(u) = \frac{3}{4}(1-u^2)\mathbb{1}_{u \in [-1,1]}$ and

$$B_r(u) = \frac{\left(\frac{3}{2}\right)_{r/2-1}\left(\frac{5}{2}\right)_{r/2-1}}{(2)_{r/2-1}} \sum_{k=0}^{r/2-1} \frac{(-1)^k \left(\frac{r+3}{2}\right)_k u^{2k}}{k!(r/2-1-k)!\left(\frac{3}{2}\right)_k}$$

is of order r for odd r where $(x)_a$ is the Pochhammer's symbol.

Construction of high-order kernels

As for kernels with compact support $[0,1]$, the two following methods can be envisioned.

Shifted Legendre orthonormal polynomials. Similarly to the first construction above, we can also consider the shifted Legendre orthonormal polynomials on $[0,1]$, denoted by $\{Q_m(\cdot)\}_{m\in\mathbb{N}}$, leading to

$$k(u) = 2\sum_{m=0}^{r+1} Q_m(0)Q_m(u)\mathbb{1}_{u\in[0,1]}. \tag{4}$$

Construction

Dilatation. Another approach, due to Kerkyacharian (2001), relies on dilatations of an integrable function $g : \mathbb{R} \to \mathbb{R}$:

$$k(u) = \sum_{k=1}^{r} \binom{r}{k} (-1)^{k+1} \frac{1}{k} g\left(\frac{u}{k}\right). \tag{5}$$

If g has support $[a, b]$, then k has support $[a, rb]$ and is of order r.

To obtain a kernel with support $[0, 1]$, one can for example take a shifted Epanechnikov kernel k_{shift} on $[0, 1/r]$:

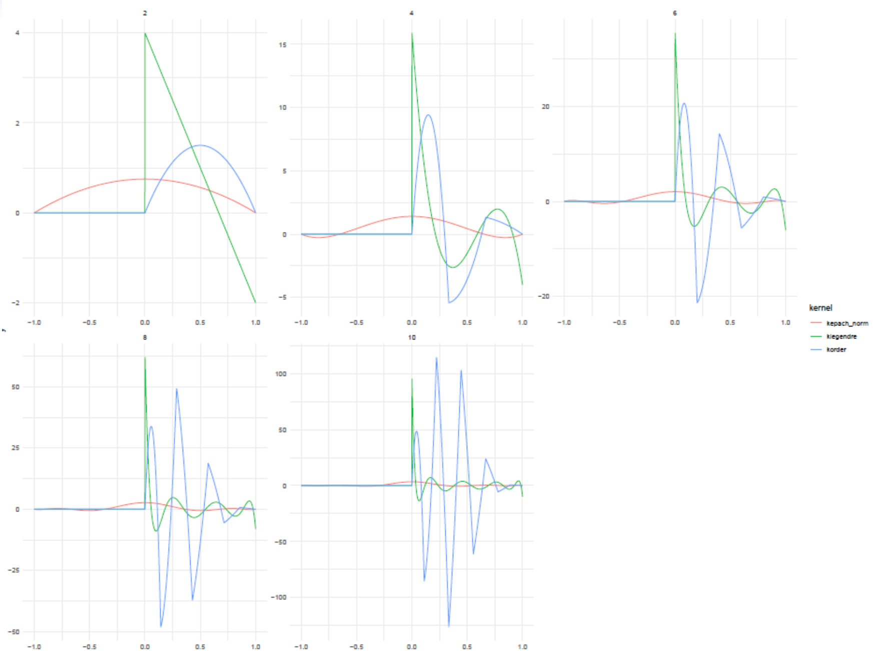
$$k_{\mathrm{shift}}(u) = 6u(1 - ru)r^2 \mathbb{1}_{u \in [0,1/r]}.$$

Numerical stability - Kernel values versus order

In what follows, we investigate numerically the high-order kernels introduced above.

Since kernels in (2) and (4) are identical up to a shift, we only focus on kernels as defined in (3) for $[-1,1]$ and (4) and (5) for $[0,1]$.

They are coded below, note that they all take as input a parameter h which corresponds to the kernel bandwidth.


2
4
6
8
10
kernel
kepach_norm
klegendre
korder

Numerical stability - Kernel values versus order

It appears clearly that non-symmetric kernels with support $[0,1]$ exhibit large variations which increase with the order, as opposed to the symmetric kernel on $[-1,1]$. This implies that numerical instabilities when computing estimators are to be expected, as illustrated below on a simple regression case.

Regression with mirror transformations

Now we consider a standard regression setting : we have access to a n-sample (X_i, Y_i) for $i = 1, \ldots, n$ with

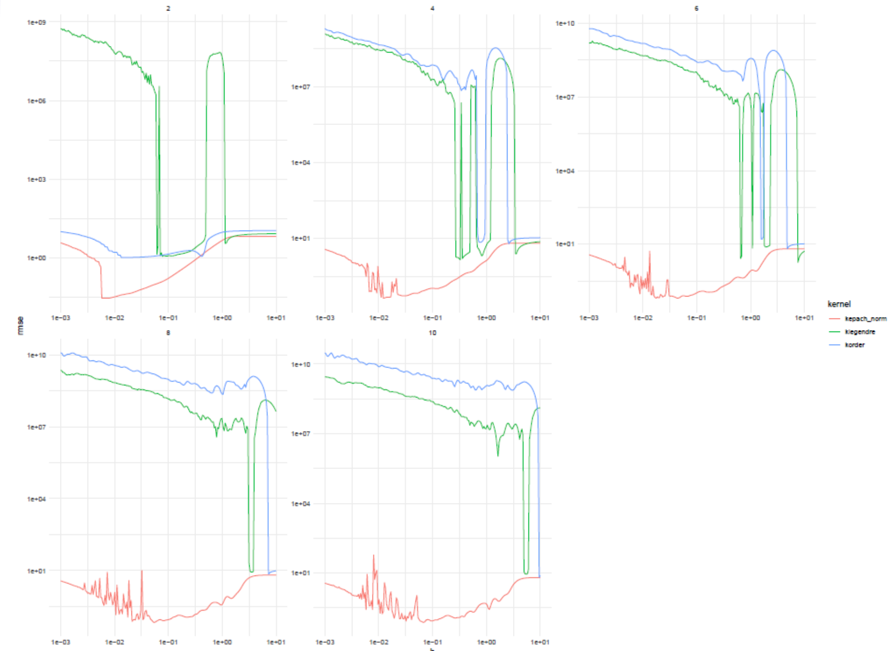
$$Y_i = m(X_i) + \epsilon_i$$

where the X_i's are i.i.d. r.v. on $[0,1]$ and ϵ_i is a centred noise.

We consider regression estimators denoted by $\widehat{m}^1$ on $[0,1]$ and $\widehat{m}^2$ on $[-1,1]$ and the Bratley function.

The only parameter which needs to be tuned is the bandwidth h.

We consider a grid of evenly-spaced values on a logarithmic scale and compute the leave-one-out mean square error for each of them.


2
4
6
8
10
rmse
h
kernel
kepach_norm
klegendre
korder

Regression with mirror transformations

We clearly see a very high numerical instability for the first estimator with kernels supported on $[0,1]$, even on a simple regression example in dimension 1.