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

Joint work with S. Da Veiga, F. Gamboa, T. Klein, and C. Prieur

Agnès Lagnoux
Institut de Mathématiques de Toulouse
Université Toulouse Jean Jaurès

Journées annuelles de la Fédération Occimath
Montpellier
May, 27-29th 2026

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
- Derivation of our estimator
- Two main ingredients
- Our efficient mirrored high-order kernel-based estimate
- Main results

Numerical applications

Outline of the talk

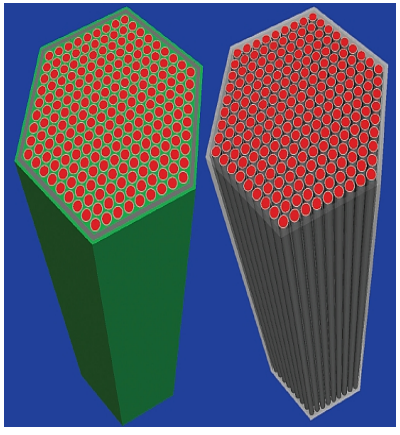
Introduction
- Framework and Sobol' indices
- The classical Pick-Freeze estimation
- Estimation from a single input/output sample

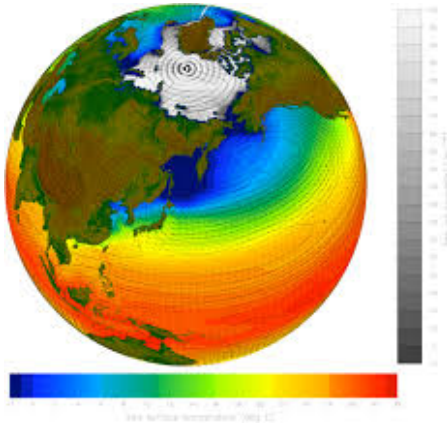
Efficient estimation from a single input/output sample
- Derivation of our estimator
- Two main ingredients
- Our efficient mirrored high-order kernel-based estimate
- Main results

Numerical applications

Motivation : computer models

Computer models have become essential in science and industry !





For clear reasons : cost reduction, possibility to explore hazardous or extreme scenarios...

Computer models as expensive functions

A computer model can be seen as a deterministic function

$$f : \mathbb{R}^p \to \mathbb{R}^k$$
$$v = (v_1, \ldots, v_p) \mapsto y = f(v) = (v_1, \ldots, v_p)$$

- the inputs v_i for $i = 1, \ldots p$ are objects = tunable simulation parameters (e.g. aircraft geometry, wind),
- the output y is a vector = quantity of interest (e.g. energetic efficiency, fuel consumption).

Computer models as expensive functions

Generally, the deterministic function f is

- continuous (at least) and non-linear but not analytically known $\Rightarrow$ black-box model ;
- only available through evaluations $f(v_1, \ldots, v_p)$;
- computationally expensive to evaluate.

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 :

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

Then $f : \mathscr{E}^p \to \mathbb{R}^k$ 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 $V_1, \ldots, V_p \in \mathscr{E}$ are independent.*
2. *The output Y has a finite second moment.*

First toy example

Let have a look on a simple example :

$$(V_1, V_2, V_3) \in \mathbb{R}^3 \mapsto Y = V_1 + V_1 V_2.$$

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 be real-valued and to lie in L^2 in the sequel.

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

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

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

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

From Hoeffding decomposition to Sobol' indices

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

$$1 = \frac{\mathrm{Var}\left(\mathbb{E}[Y|V_{\mathbf{u}}]\right)}{\mathrm{Var}(Y)} + \frac{\mathrm{Var}\left(\mathbb{E}[Y|V_{\sim\mathbf{u}}]\right)}{\mathrm{Var}(Y)} + \frac{\mathrm{Var}(IA)}{\mathrm{Var}(Y)}$$
$$:= S^{\mathbf{u}} + S^{\sim\mathbf{u}} + S^{\mathbf{u},\sim\mathbf{u}} \quad \Rightarrow \quad \text{Sobol}' \text{ indices}$$

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

while $1 - S^{\sim\mathbf{u}} = S^{\mathbf{u}} + S^{\mathbf{u},\sim\mathbf{u}}$ *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

$$(S^1, S^2, S^3) = (1/2, 0, 0)$$

and

$$(S^{1,Tot}, S^{2,Tot}, S^{3,Tot}) = (1, 1/2, 0).$$

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 $V^{\mathbf{u}} = (V_1, V_2, V_3', V_4', V_5')$.

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

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

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
- Derivation of our estimator
- Two main ingredients
- Our efficient mirrored high-order kernel-based estimate
- Main results

Numerical applications

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* $(X_j, Y_j)_{j=1,\ldots,n}$ *of the pair* (X, Y).

Our estimator

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

$$T_{n,\text{oracle}} = \frac{1}{n}\sum_{i=1}^{n}(2Y_i - m(X_i))m(X_i) \quad \text{with } m(x) = \mathbb{E}[Y|X=x].$$

• 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 lemma.

• 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.

☞ *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,

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

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

For kernel density estimation, bias is (by a multivariate Taylor expansion)

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

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

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 ? (II)

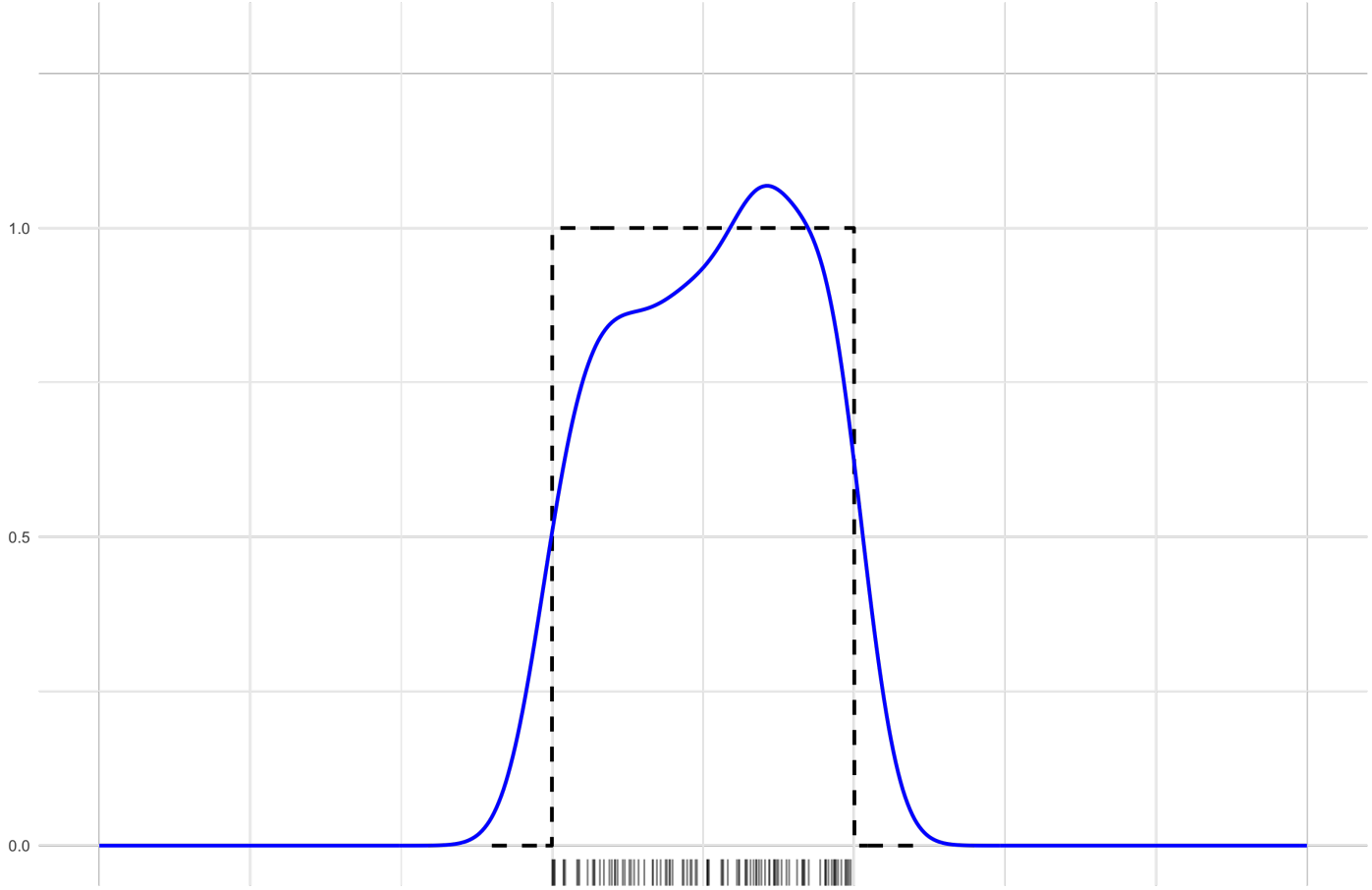
By analysing the variance (skipped here), with a high-order kernel, we finally get

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

☞ kernel with high-enough order.

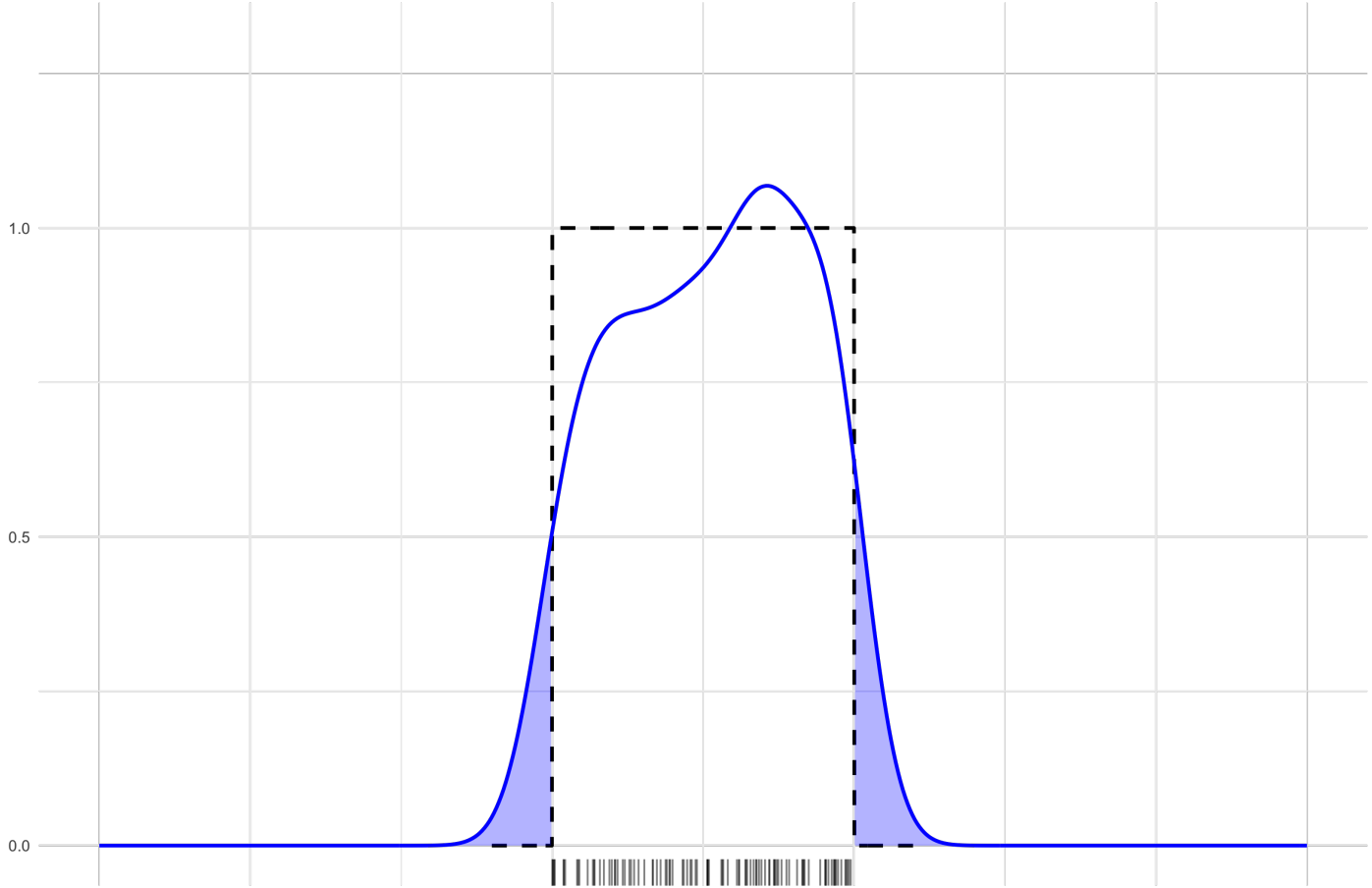
KDE boundary issues (I)

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



KDE boundary issues (II)

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



A partial solution

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

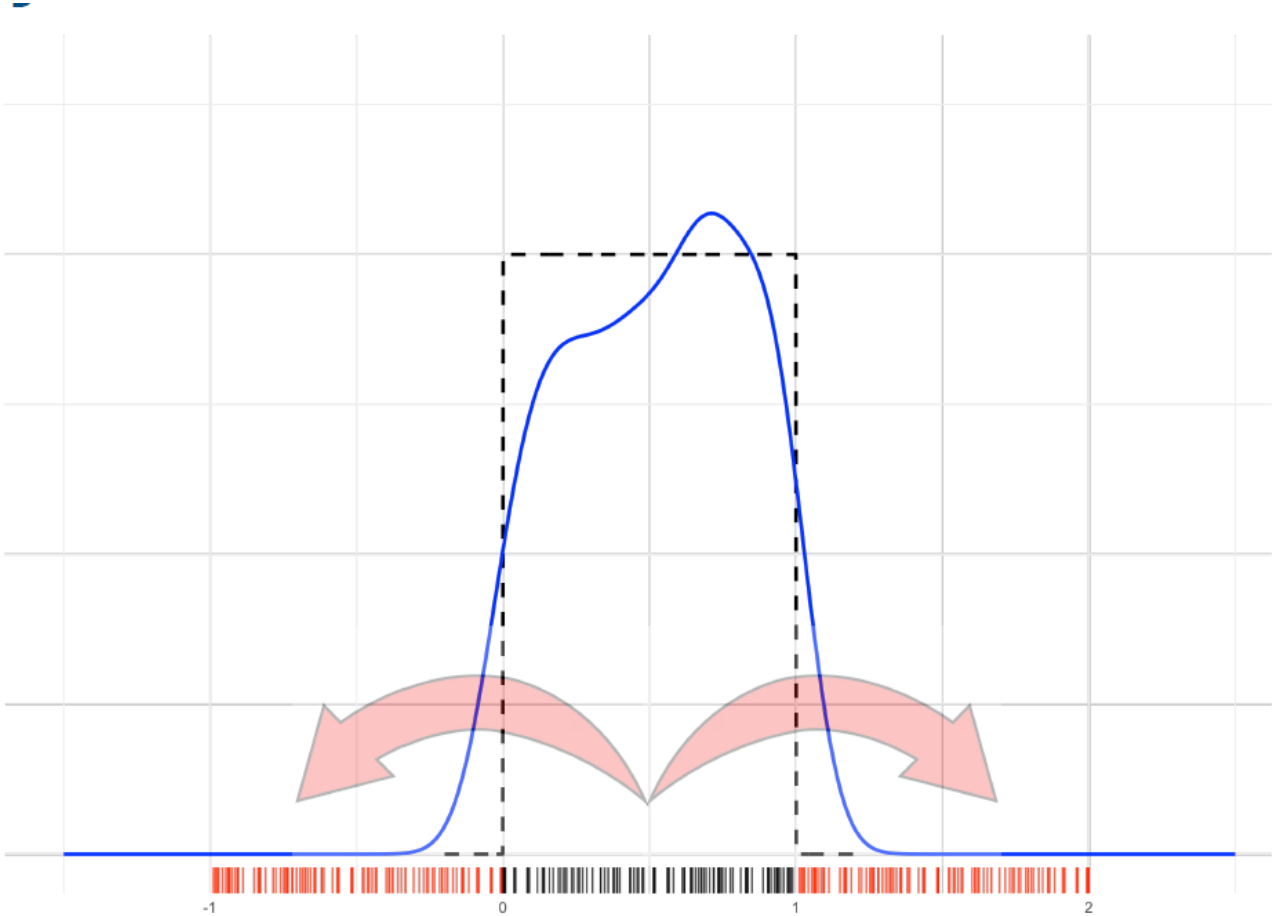
$$T^{\text{trunc},\varepsilon} = \mathbb{E}[\mathbb{E}[Y|X]^2 \mathbb{1}_{X \in (\varepsilon, 1-\varepsilon)^d}].$$

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.

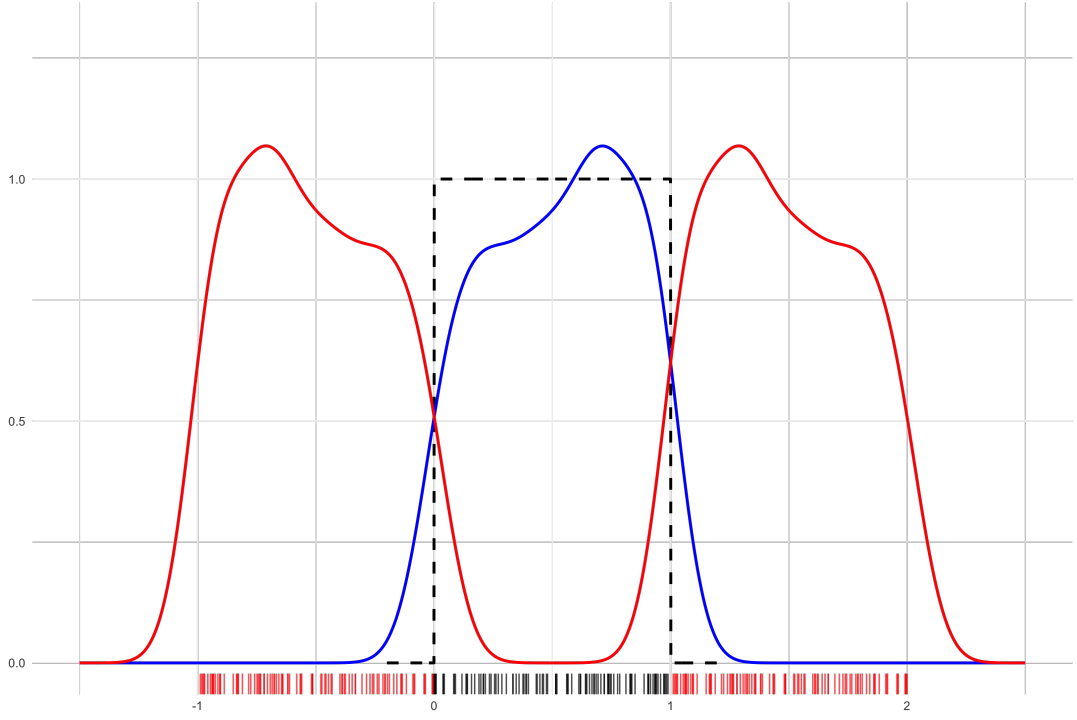
Second ingredient : mirror transformation (I)

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



Second ingredient : mirror transformation (II)

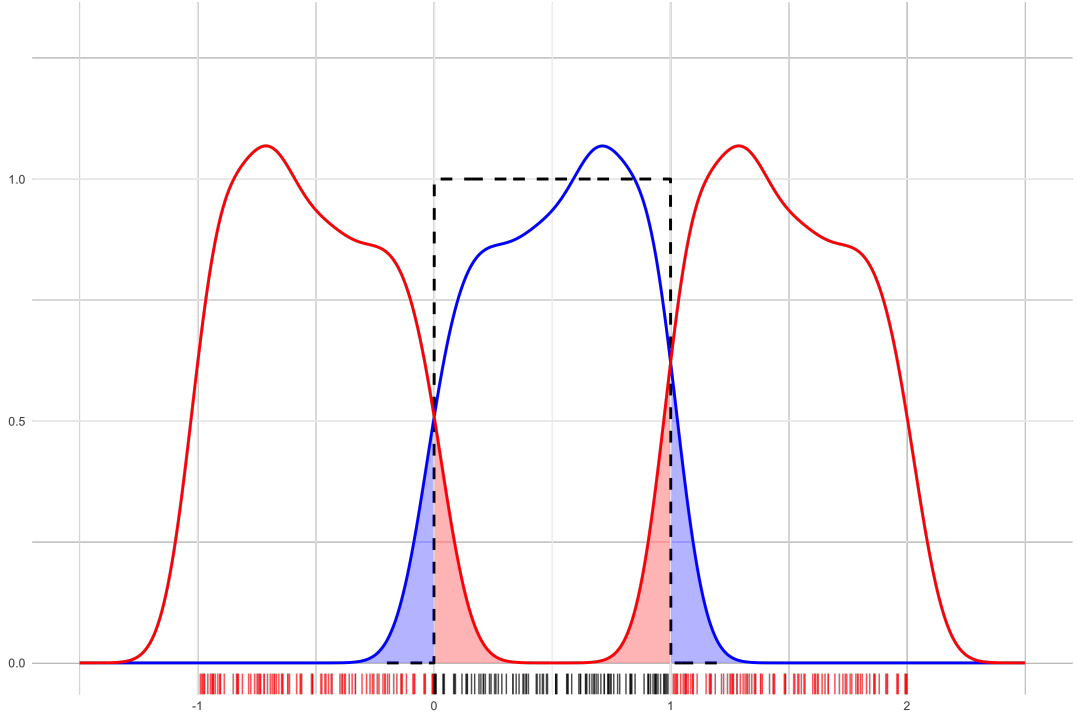
$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right) \quad \text{and} \quad \int_{-\infty}^{\infty} \widehat{f}(x) = 1$$



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

Second ingredient : mirror transformation (III)

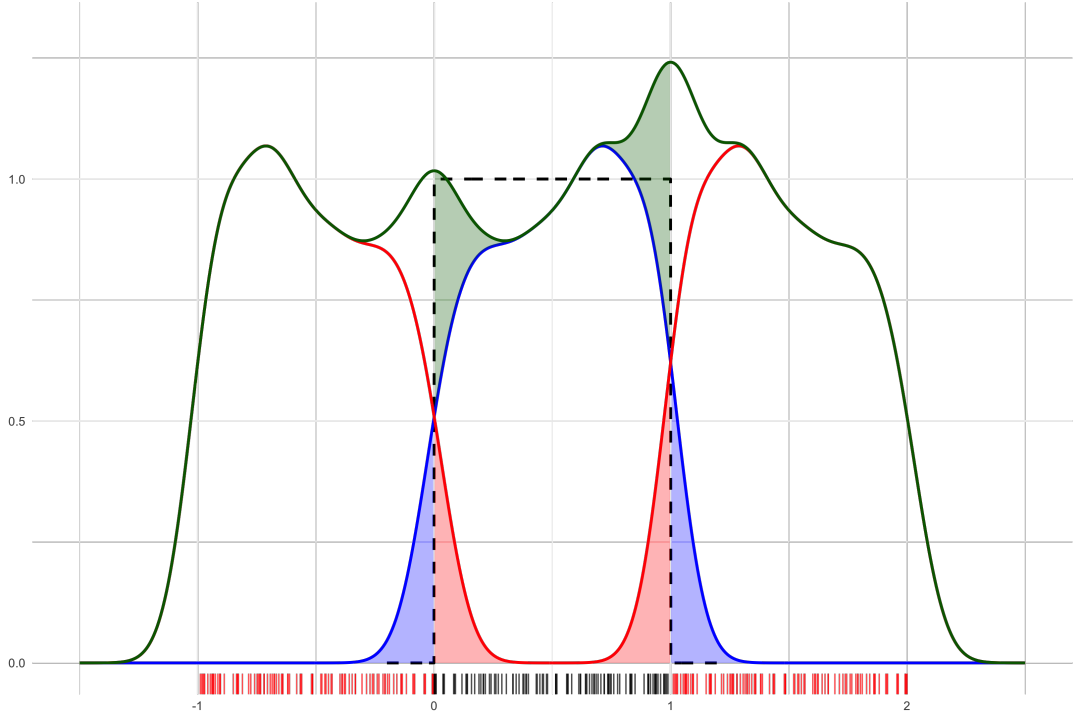
$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right) \quad \text{and} \quad \int_{-\infty}^{\infty} \widehat{f}(x) = 1$$



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

Second ingredient : mirror transformation (IV)

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

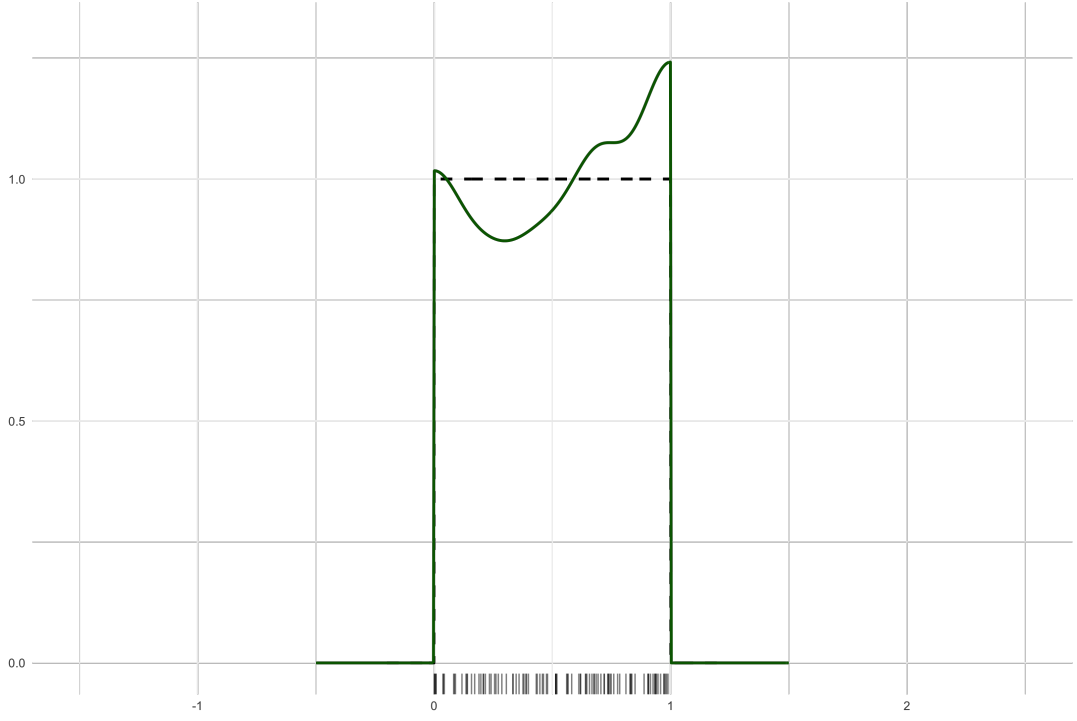


$$\widehat{f}_{\text{lower}}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x-(-X_i)}{h}\right)$$

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

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 :

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

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

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

and its leave-one-out version :

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

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 - $k\colon[-1,1]\to\mathbb{R}$ 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

$$\left\| \mathbb{E}\left[\widehat{f}_{n,h_n,i}\right] - f_X \right\|_\infty = O\left(h_n^\alpha\right),$$

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

- 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)*.*

☞ The same guarantees hold for the estimation $\widehat{S}_{n,h_n}$ of S^X applying the delta method.

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`.

Sketch of the proofs

Following the same lines as in the proof of Theorem 2.1 in Doksum and Samarov (1995), we prove that

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

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

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.

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
- Derivation of our estimator
- Two main ingredients
- Our efficient mirrored high-order kernel-based estimate
- Main results

Numerical applications

For all test cases :

- first-order and total 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.

A realistic flood model

The flood model used is a simplification of the 1D Saint-Venant hydrodynamic equations, assuming constant and uniform flows and very wide rectangular cross-sections. It consists of an equation involving the characteristics of the river section upstream of the industrial site :

$$S = Z_v + H - H_d - C_b$$

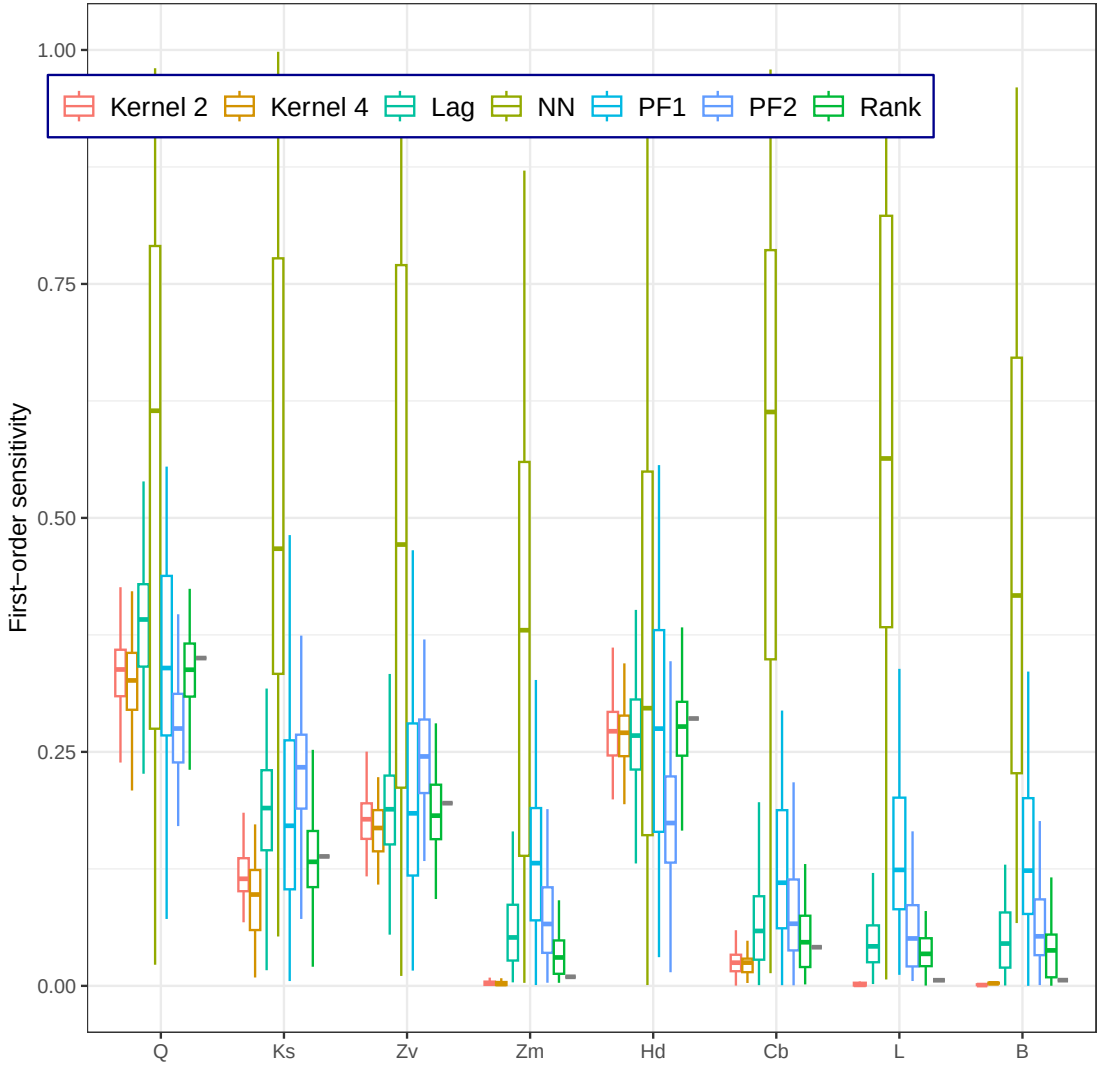
where H is calculated as :

$$H = \left(\frac{Q}{BK_s \sqrt{\frac{Z_m - Z_v}{L}}} \right)^{0.6}.$$

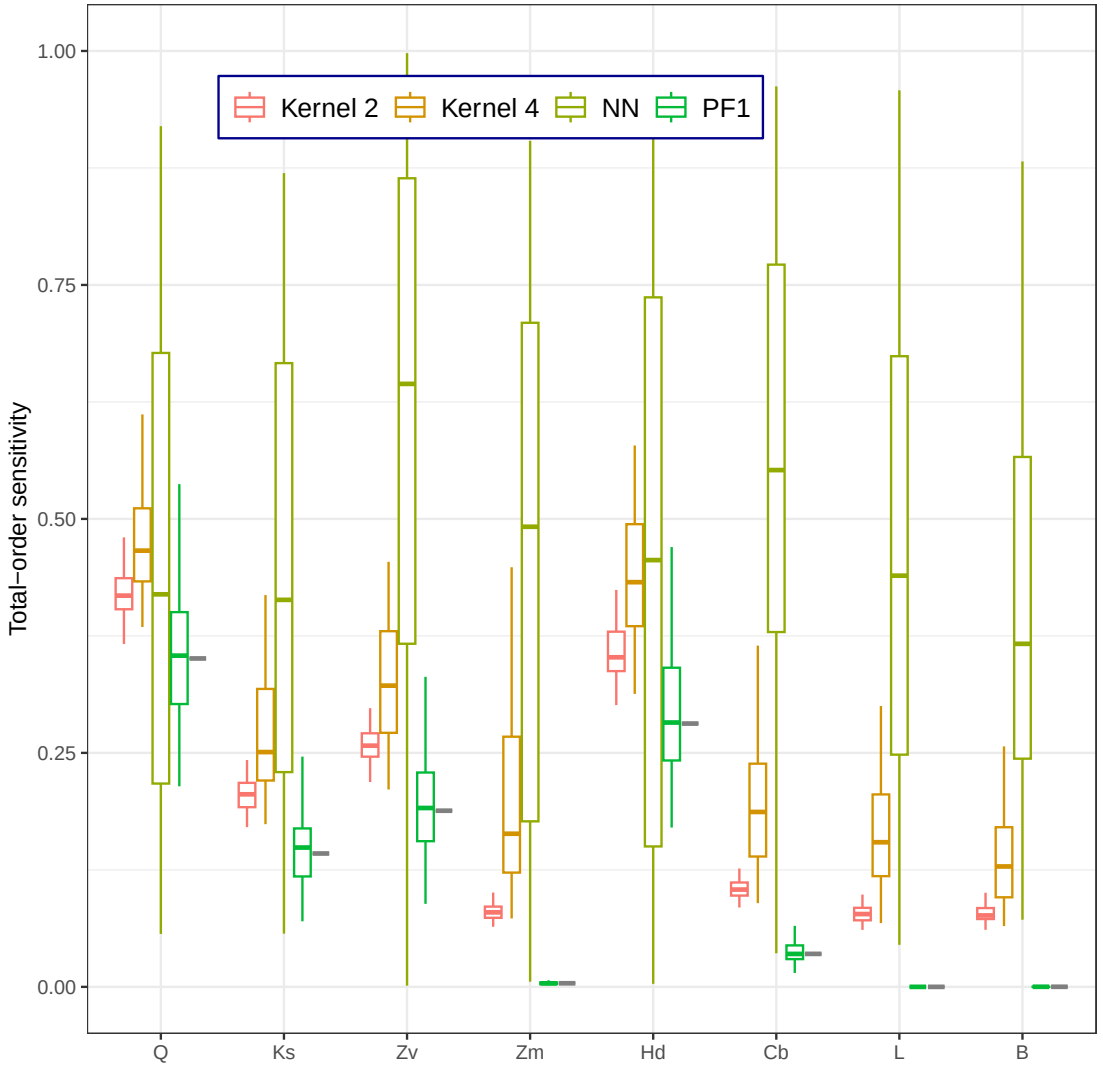
A realistic flood model - model parameters

- S : overflow (in meters), model output ;
- H : maximum annual water level (in meters) ;
- Q : maximum annual flow rate (in m^3/s) with Gumbel max distribution $Gu(1013, 558)$, truncated to be in $[500, 3000]$;
- K_s : strickler coefficient with normal distribution $\mathcal{N}(30, 8)$, truncated below at 15 ;
- Z_v : downstream riverbed elevation (in meters) with triangular distribution $T(49, 50, 51)$;
- Z_m : upstream riverbed elevation (in meters) with triangular distribution $T(54, 55, 56)$;
- H_d : dike height (in meters) with uniform distribution $\mathcal{U}(7, 9)$;
- C_b : bank elevation (in meters) with triangular distribution $T(55, 55.5, 56)$;
- L : length of the river section (in meters) with triangular distribution $T(4990, 5000, 5010)$;
- B : river width (in meters) with triangular distribution $T(295, 300, 305)$.

A realistic flood model - first-order indices - $n = 500$



A realistic flood model - total 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.

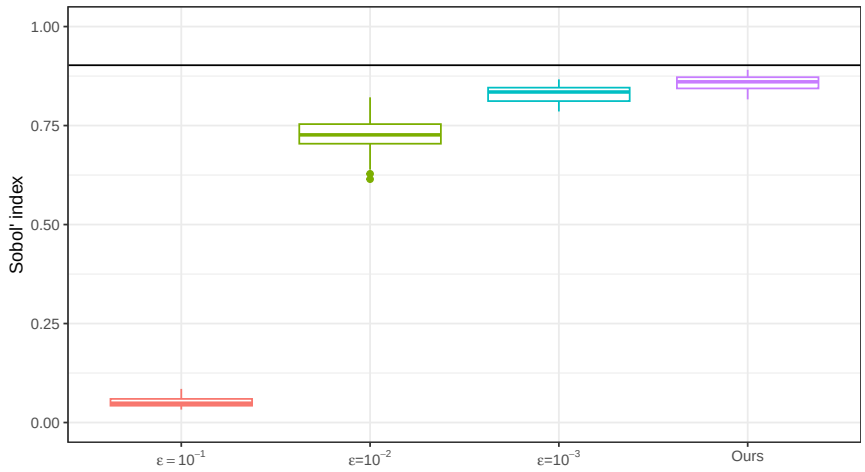
We consider Example 3.2 from Doksum and Samarov (1995) :

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

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 $\epsilon = 10^{-1}$, $\epsilon = 10^{-2}$ and 10^{-3}.

Tuning of parameter ϵ



When ϵ is equal to 10^{-3} (3rd), the performance of both estimators are similar. However when $\epsilon = 10^{-1}$ (1st), 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 wrt this parameter limits its practical use.

Advertising

For your book project, think about the collection

Lecture Notes on Applied Deterministic and Stochastic Mathematics

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". In revision at Information and Inference. Available on Hal and Arxiv (2025). `https://hal.science/hal-04052837v2`.

T. Klein, A. Lagnoux, T.M.N. Nguyen, P. Rochet. "Asymptotic efficiency for Sobol' and Cramér-von Mises indices under two designs of experiments". In revision at ESAIM P&S. `https://hal.science/hal-03477991`.