Estimating Sobol indices by combining *pick freeze* estimators and Replicated Latin Hypercube sampling

Clémentine PRIEUR
(joint work with J.Y. Tissot)

University of Grenoble
Laboratoire Jean Kuntzmann
Inria "Project-team" MOISE

SAMO 2013, July 1-4 2013

Introduction

$$\boxed{\text{inputs}} \longrightarrow \boxed{\text{model}} \longrightarrow \boxed{\text{output}}$$

$$\boxed{\left.\begin{matrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_d \end{matrix}\right\}} \longrightarrow \boxed{\mathcal{M}} \longrightarrow \boxed{Y = \mathcal{M}(X_1, \ldots, X_d)}$$

One wishes to quantify the sensitivity of the output Y to the independent inputs $X_1, \ldots, X_d$ by computing Sobol indices.

In this talk, we introduce new *pick & freeze* estimator based on Replicated Latin Hypercube sampling (RLHS).

Outline

I- Our new estimation procedure : notation, definition.

II- Properties.

III- Comparison with randomized QMC approaches.

IV- Conclusion, perspectives.

I- Our new estimation procedure : notation, definition

In this talk, we propose a new estimation procedure for first order Sobol' indices, that is $S_i = \dfrac{\mathrm{Var}\left(\mathbb{E}(Y|X_i)\right)}{\mathrm{Var}(Y)}$, $i = 1, \ldots d$.

We assume (without loss of generality)

$\forall\, i = 1, \ldots, d$ $X_i \sim \mathcal{U}\left([0,1]\right)$, the inputs are independent.

I- Our new estimation procedure : notation, definition

In this talk, we propose a new estimation procedure for first order Sobol' indices, that is $S_i = \dfrac{\mathrm{Var}\left(\mathbb{E}(Y|X_i)\right)}{\mathrm{Var}(Y)}$, $i = 1, \ldots d$.

We assume (without loss of generality)

$\forall\, i = 1, \ldots, d \; X_i \sim \mathcal{U}\left([0,1]\right)$, the inputs are independent.

What about the *pick & freeze* estimation procedure?

Advantages • it is robust (one only needs very soft assumptions on the model), • one can derive asymptotic confidence intervals, • the rate of convergence does not depend on the dimension.

Disadvantages • this rate is rather slow $n^{1/2}$, • with classical sampling strategies, the number of model evaluations needed for estimating all the first order Sobol' indices is linear in the dimension d.

I- Our new estimation procedure : notation, definition

Pick & Freeze procedure : n double evaluations of $\mathcal{M}$ required.

Let $\mathbf{X}$ and $\mathbf{Z}$ be two independent random vectors distributed as $\mathcal{U}\left([0,1]^d\right)$.

• the first of any double evaluation is a realization of the random variable $Y = \mathcal{M}(\mathbf{X})$,

• the complementary evaluation is a realization of the random variable denoted by Y_i defined by $Y_i = \mathcal{M}(\mathbf{X}_i : \mathbf{Z}_{i^c})$ where $\mathbf{X}_i : \mathbf{Z}_{i^c}$ is the d-dimensional random vector defined by

$$(\mathbf{X}_i : \mathbf{Z}_{i^c})_l = \begin{cases} X_i & \text{if } l = i \\ Z_l & \text{if } l \neq i. \end{cases}$$

I- Our new estimation procedure : notation, definition

The i^{th} component of $\mathbf{X}$ has been frozen.

I- Our new estimation procedure : notation, definition

The i^{th} component of $\mathbf{X}$ has been frozen.

We can prove [JKL+12] that

$$S_i = \frac{\mathrm{Cov}(Y, Y_i)}{\mathrm{Var}[Y]} = \frac{\mathbb{E}[YY_i] - \mathbb{E}[Y]\mathbb{E}[Y_i]}{\mathrm{Var}[Y]}.$$

I- Our new estimation procedure : notation, definition

The i^{th} component of $\mathbf{X}$ has been frozen.

We can prove [JKL+12] that

$$S_i = \frac{\mathrm{Cov}(Y, Y_i)}{\mathrm{Var}[Y]} = \frac{\mathbb{E}[YY_i] - \mathbb{E}[Y]\mathbb{E}[Y_i]}{\mathrm{Var}[Y]}.$$

Then the *pick & freeze* approach [Sob93] consists in proposing an empirical estimator for both the numerator and the denominator.

I- Our new estimation procedure : notation, definition

Design of Experiments :

we define $\left\{ \begin{array}{lcl} H(n) & = & \{\mathbf{X}^j,\ 1 \le j \le n\} \\ \tilde{H}(n) & = & \{\mathbf{Z}^j,\ 1 \le j \le n\} \end{array} \right.$

We then define

$$H_i(n) = \{(\mathbf{X}_i : \mathbf{Z}_{i^c})^j,\ 1 \le j \le n\} = \begin{pmatrix} Z_1^1 & \dots & X_i^1 & \dots & Z_d^1 \\ . & . & . & . & . \\ . & . & . & . & . \\ . & . & . & . & . \\ Z_1^n & \dots & X_i^n & \dots & Z_d^n \end{pmatrix}.$$

I- Our new estimation procedure : notation, definition

Design of Experiments :

we define $\begin{cases} H(n) & = & \{\mathbf{X}^j,\ 1 \leq j \leq n\} \\ \tilde{H}(n) & = & \{\mathbf{Z}^j,\ 1 \leq j \leq n\} \end{cases}$

We then define

$$H_i(n) = \{(\mathbf{X}_i : \mathbf{Z}_{i^c})^j,\ 1 \leq j \leq n\} = \begin{pmatrix} Z_1^1 & \dots & X_i^1 & \dots & Z_d^1 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ Z_1^n & \dots & X_i^n & \dots & Z_d^n \end{pmatrix}.$$

Our design of experiments to estimate S_i with the *pick & freeze* approach is $D_i(N) = H(n) \cup H_i(n)$. It is of size $N = 2n$.

I- Our new estimation procedure : notation, definition

Pick & freeze estimator :

For any j in $\{1,\ldots,n\}$, define $\left\{ \begin{array}{lcl} Y^j & = & \mathcal{M}(\mathbf{X}^j) \\ Y_i^j & = & \mathcal{M}((\mathbf{X}_i : \mathbf{Z}_{i^c})^j) \end{array} \right.$

I- Our new estimation procedure : notation, definition

Pick & freeze estimator :

For any j in $\{1, \ldots, n\}$, define $\begin{cases} Y^j &= \mathcal{M}(\mathbf{X}^j) \\ Y_i^j &= \mathcal{M}((\mathbf{X}_i : \mathbf{Z}_{i^c})^j) \end{cases}$

We then introduce [JKL+12]

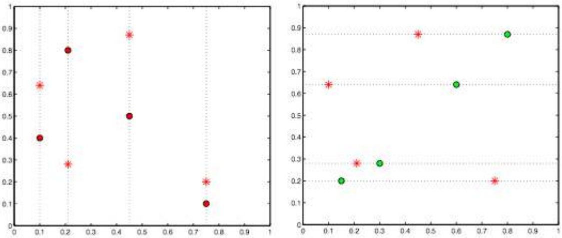
$$\widehat{S}_{i,n} = \frac{\frac{1}{n}\sum_{j=1}^{n} Y^j Y_i^j - \left(\frac{1}{2n}\sum_{j=1}^{n} Y^j + Y_i^j\right)^2}{\frac{1}{2n}\sum_{j=1}^{n}\left((Y^j)^2 + (Y_i^j)^2\right) - \left(\frac{1}{2n}\sum_{j=1}^{n} Y^j + Y_i^j\right)^2}.$$

Other choices for the empirical estimates of the numerator and the denominator are possible (e.g. [Sal02, Mau02, Owe12]).

I- Our new estimation procedure : notation, definition

We thus need $(1+d)n$ evaluations of the model to compute all the $\widehat{S}_{i,n}$, $i = 1\ldots, d$.

Example with $d = 2$ and $n = 4$:



- On the left hand side $\mathbf{X}$ ($\star$) and $(\mathbf{X}_1 : \mathbf{Z}_2)_{\text{sample}}$ ($\bullet$).
- On the right hand side $\mathbf{X}$ ($\star$) and $(\mathbf{X}_2 : \mathbf{Z}_1)_{\text{sample}}$ ($\bullet$).

I- Our new estimation procedure : notation, definition

Which design of experiments to overcome this issue?

Let D a design of experiments (DoE) of size n defined by

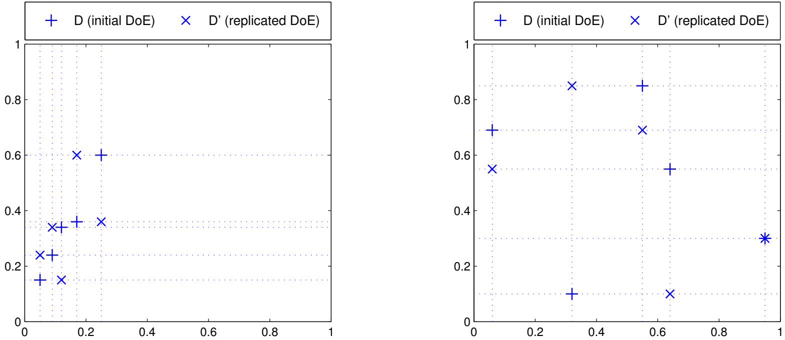
$$D = \left\{ \mathbf{x}^j = (x_1^j, \ldots, x_d^j),\ 1 \leq j \leq n \right\} \cdot$$

The DoE D' is replicated from D if there exist d independent random permutations of $\{1, \ldots, n\}$ — denoted by $\pi_1, \ldots, \pi_d$ — such that

$$D' = \left\{ \mathbf{x}'^j = (x_1^{\pi_1(j)}, \ldots, x_d^{\pi_d(j)}),\ 1 \leq j \leq n \right\} \cdot$$

I- Our new estimation procedure : notation, definition

Then $D \cup D'$ can be used for estimating any first-order Sobol indices using the *pick & freeze* approach (see Figure below).



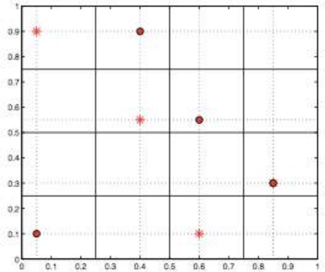
On the left hand side D is an independent sampling.

On the right hand side D is a LHS (thus D' too).

I- Our new estimation procedure : notation, definition

Replicated Latin Hypercube sampling [McK95]

Let $H(n) = \{\mathbf{X}^j,\ 1 \leq j \leq n\}$ and $\tilde{H}(n) = \{\mathbf{X}'^j,\ 1 \leq j \leq n\}$ be two Replicated Latin Hypercubes.



$$j = 1, \ldots, n$$

$$\mathbf{x}^j = \left(\frac{j - U_{1,j}}{n}, \ldots, \frac{j - U_{d,j}}{n}\right)$$

$$\mathbf{x}'^j = \left(\frac{\pi_1(j) - U_{1,\pi_1(j)}}{n}, \ldots, \frac{\pi_d(j) - U_{d,\pi_d(j)}}{n}\right)$$

I- Our new estimation procedure : notation, definition

Define $H_i(n) = \left\{ \mathbf{X}'^{\pi_i^{-1}(j)},\ 1 \le j \le n \right\}$

$$= \begin{pmatrix} X_1'^{\pi_1\circ\pi_i^{-1}(1)} & \dots & X_i^1 & \dots & X_d'^{\pi_d\circ\pi_i^{-1}(1)} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ X_1'^{\pi_1\circ\pi_i^{-1}(n)} & \dots & X_i^n & \dots & X_d'^{\pi_d\circ\pi_i^{-1}(n)} \end{pmatrix}.$$

We then choose $D_i(N) = H(n) \cup H_i(n)$. $D_i(N)$ allows estimating S_i with the *pick freeze* approach.

We remark that $D_i(N)$ as a non ordered set of points does not depend on i, and that's the trick.

II- Properties

A central limit Theorem

If $\mathcal{M}^6$ is integrable then for any $i \in \{1, \ldots, d\}$,

$$\sqrt{n}(\hat{S}_{i,n} - S_i)$$

satisfies a central limit theorem with zero-mean normal limit distribution.

Ideas for the proof : we first prove the result for two independent latin hypercubes, and then control the difference by replacing by replicated latin hypercubes.

Main tools : a SLLN and a CLT for latin hypercube sampling [Loh96], a delta method as in [JKL$^+$12].

The asymptotic variance is smaller than the one in [JKL$^+$12].

III- Comparison with randomized QMC approaches

Model : $Y = f_1(X_1) \times \cdots \times f_d(X_d)$ with $(X_1, \ldots, X_d) \sim \mathcal{U}\left([0,1]^d\right)$ and

$$f_i(X_i) = \frac{|4X_i - 2| + a_i}{1 + a_i}, \quad a_i \geq 0\,,\ i = 1, \ldots d.$$

i) $d = 3$, $\mathbf{a} = (0, 1, 9)$

ii) $d = 12$, $\mathbf{a} = (0, 0, 0, 0, 1, 1, 1, 1, 9, 9, 9, 9)$

iii) $d = 24$, $\mathbf{a} = (\underbrace{0, \ldots, 0}_{8 \text{ times}}, \underbrace{1, \ldots, 1}_{8 \text{ times}}, \underbrace{9, \ldots, 9}_{8 \text{ times}})$.

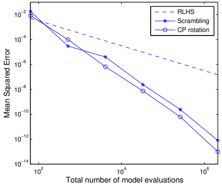
i) $\underline{S}_1 = 0.742$, $\underline{S}_2 = 0.185$, $\underline{S}_3 = 0.007$

ii) $\underline{S}_1 = \cdots = \underline{S}_4 = 0.098$, $\underline{S}_5 = \cdots = \underline{S}_8 = 0.024$, $\underline{S}_9 = \cdots = \underline{S}_{12} = 0.001$,

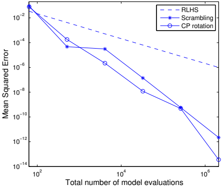
iii) $\underline{S}_1 = \cdots = \underline{S}_8 = 0.018$, $\underline{S}_9 = \cdots = \underline{S}_{16} = 0.004$, $\underline{S}_{17} = \cdots = \underline{S}_{24} = 10^{-4}$.

III- Comparison with randomized QMC approaches

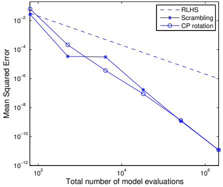
Rand. Sobol' seq. : a) Cranley-Patterson rotation, b) Owen's scrambling [Owe95, Owe97a, Owe97b].



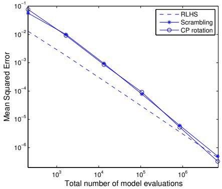
mean squared error for $\underline{S}_1$



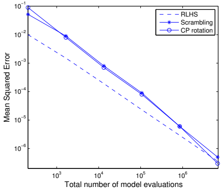
mean squared error for $\underline{S}_2$



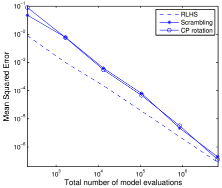
mean squared error for $\underline{S}_3$



mss for $\underline{S}_1, \ldots, \underline{S}_4$

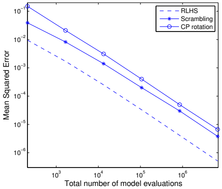


mse for $\underline{S}_5, \ldots, \underline{S}_8$

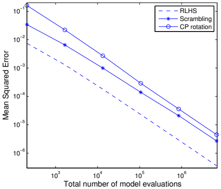


mse for $\underline{S}_9, \ldots, \underline{S}_{12}$

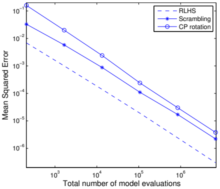
III- Comparison with randomized QMC approaches



mss for $\underline{S}_1, \ldots, \underline{S}_8$



mse for $\underline{S}_9, \ldots, \underline{S}_{16}$



mse for $\underline{S}_{17}, \ldots, \underline{S}_{24}$

IV- Conclusion, perspectives

We have proposed a new *pick-freeze* estimator, based on replicated latin hypercube sampling, that allows estimating all the first order Sobol' indices with a coat independent of the dimension.

Remarks, perspectives :

- the asymptotic variance in the CLT can be estimated (work in progress),
- the estimation procedure can be generalized with replicated latin hypercube sampling based on orthogonal arrays (strength 2= second order Sobol' indices, ...) [TP12],
- one probably can adapt ideas in [GJK+13] for deriving non asymptotic properties (work in progress),
- ...

Some references I

[GJK+13] F. Gamboa, A. Janon, T. Klein, A. Lagnoux, and C. Prieur. Statistical inference for sobol pick freeze monte carlo method. *Preprint available at http://hal.archives-ouvertes.fr/hal-00804668*, 2013+.

[JKL+12] A. Janon, T. Klein, A. Lagnoux, M. Nodet, and C. Prieur. Asymptotic normality and efficiency of two Sobol' index estimators. *To appear in ESAIM P&S*, 2012.

[Loh96] W. L. Loh. On Latin hypercube sampling. *The Annals of Statistics*, 24(5):2058–2080, 1996.

[Mau02] W. Mauntz. Global sensitivity analysis of general nonlinear systems. *Master's Thesis, Imperial College. Supervisors: C. Pantelides and S. Kucherenko*, 2002.

[McK95] M. D. McKay. Evaluating prediction uncertainty. *Technical Report NUREG/CR-6311, US Nuclear Regulatory Commission and Los Alamos National Laboratory*, pages 1–79, 1995.

[Owe95] A. B. Owen. Randomly permuted (t,m,s)-nets and (t,s)-sequences. In H. Niederreiter and P. J.-S. Shiue, editors, *Monte Carlo and Quasi-Monte CArlo Methods in Scientific Computing*, pages 299–317. Springer-Verlag, New York, 1995.

Some references II

[Owe97a] A. B. Owen. Monte Carlo variance of scrambled equidistribution quadrature. *SIAM Journal of Numerical Analysis*, 34(5):1884–1910, 1997.

[Owe97b] A. B. Owen. Scrambled net variance for integral of smooth functions. *Annals of Statistics*, 25(4):1541–1562, 1997.

[Owe12] A. B. Owen. Variance components and generalized Sobol' indices. *Preprint available at* *http://www-stat.stanford.edu/∼owen/reports/effdim-may.pdf*, 2012+.

[Sal02] A. Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145:280–297, 2002.

[Sob93] I. M. Sobol'. Sensitivity analysis for nonlinear mathematical models. *Mathematical Modeling and Computational Experiment*, 1:407–414, 1993.

[TP12] J. Y. Tissot and C. Prieur. Variance-based sensitivity analysis using harmonic analysis. *Preprint available at* *http://hal.archives-ouvertes.fr/docs/00/68/07/25/PDF/FAST_RBD_revisited.pdf*, 2012+.

Thanks for your attention