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
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).
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
\[ S_i = \frac{\text{Var}(\mathbb{E}(Y|X_i))}{\text{Var}(Y)}, \quad i = 1, \ldots, d. \]

We assume (without loss of generality)
\[ \forall i = 1, \ldots, d \; X_i \sim \mathcal{U}([0,1]), \] the inputs are independent.

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

We assume (without loss of generality) $\forall i = 1, \ldots, d \; X_i \sim \mathcal{U}([0, 1])$, 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 $d$. 

Clémentine PRIEUR (joint work with J.Y. Tissot)
**Pick & Freeze** procedure: $n$ double evaluations of $\mathcal{M}$ required.

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

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

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

$$
(\mathbf{X}_i : \mathbf{Z}_{i^c})_l = \begin{cases} 
\mathbf{X}_i & \text{if } l = i \\
\mathbf{Z}_i & \text{if } l \neq i.
\end{cases}
$$
I- Our new estimation procedure: notation, definition

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

We can prove [JKL+12] that

$$S_i = \frac{\text{Cov}(Y, Y_i)}{\text{Var}[Y]} = \frac{\mathbb{E}[YY_i] - \mathbb{E}[Y]\mathbb{E}[Y_i]}{\text{Var}[Y]}.$$
The $i^{th}$ component of $X$ has been frozen.

We can prove [JKL$+12$] that

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

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

\[ H(n) = \{ X^j, 1 \leq j \leq n \} \]
\[ \tilde{H}(n) = \{ Z^j, 1 \leq j \leq n \} \]

We then define

\[ H_i(n) = \{ (X_i : Z_i^c)^j, 1 \leq j \leq n \} = \begin{pmatrix}
Z_1^1 & \ldots & X_i^1 & \ldots & Z_d^1 \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
Z_1^n & \ldots & X_i^n & \ldots & Z_d^n
\end{pmatrix} \]
I- Our new estimation procedure: notation, definition

Design of Experiments:
we define

\[ H(n) = \{ X^j, 1 \leq j \leq n \} \]
\[ \tilde{H}(n) = \{ Z^j, 1 \leq j \leq n \} \]

We then define

\[ H_i(n) = \{ (X_i : Z_{i^c})^j, 1 \leq j \leq n \} = \begin{pmatrix} Z_1^1 & \cdots & X_i^1 & \cdots & Z_d^1 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ Z_n^1 & \cdots & X_i^n & \cdots & Z_d^n \end{pmatrix} \]

Our design of experiments to estimate \( S_i \) with the \textit{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

\[
\begin{align*}
Y_j^i &= \mathcal{M}(X_j^i) \\
Y_j &= \mathcal{M}((X_i : Z_{ic})^j)
\end{align*}
\]
I- Our new estimation procedure: notation, definition

*Pick & freeze estimator*:

For any $j$ in $\{1, \ldots, n\}$, define

\[
\begin{align*}
Y_j &= M(X_j) \\
Y_i^j &= M((X_i : Z_{i_c})^j)
\end{align*}
\]

We then introduce [JKL$^+$12]

\[
\hat{S}_{i,n} = \frac{1}{n} \sum_{j=1}^{n} Y_j Y_i^j - \left( \frac{1}{2n} \sum_{j=1}^{n} Y_j + Y_i^j \right)^2
\]

\[
= \frac{1}{2n} \sum_{j=1}^{n} \left( (Y_j)^2 + (Y_i^j)^2 \right) - \left( \frac{1}{2n} \sum_{j=1}^{n} Y_j + Y_i^j \right)^2.
\]

Other choices for the empirical estimates of the numerator and the denominator are possible (e.g. [Sal02, Mau02, Owe12]).
We thus need \((1 + d)n\) evaluations of the model to compute all the \(\hat{S}_{i,n}, i = 1 \ldots, d\).

Example with \(d = 2\) and \(n = 4\):

- On the left hand side \(X (\star)\) and \((X_1 : Z_2)_{\text{sample}} (\bullet)\).
- On the right hand side \(X (\star)\) and \((X_2 : Z_1)_{\text{sample}} (\bullet)\).
Which design of experiments to overcome this issue?

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

$$D = \{ x^j = (x_1^j, \ldots, x_d^j), \; 1 \leq j \leq n \}.$$

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

$$D' = \{ x'^j = (x_1^{\pi_1(j)}, \ldots, x_d^{\pi_d(j)}), \; 1 \leq j \leq n \}.$$
Then \( D \cup D' \) can be used for estimating any first-order Sobol indices using the \textit{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).

Clémentine PRIEUR (joint work with J.Y. Tissot)
I- Our new estimation procedure: notation, definition

Replicated Latin Hypercube sampling [McK95]

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

$j = 1, \ldots, n$

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

$x'^j = \left(\frac{\pi_1(j)-U_{1,\pi_1(j)}}{n}, \ldots, \frac{\pi_d(j)-U_{d,\pi_d(j)}}{n}\right)$
I- Our new estimation procedure: notation, definition

Define $H_i(n) = \{X'_{\pi_i^{-1}(j)}, \ 1 \leq j \leq n\}$

$$H_i(n) = \left\{ X'_{\pi_i^{-1}(j)}, \ 1 \leq j \leq n \right\}$$

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.
A central limit Theorem
If $M^6$ 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 \((X_1, \ldots, X_d) \sim \mathcal{U}([0,1]^d)\)

and

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

i) \( d = 3, \ a = (0,1,9) \)

ii) \( d = 12, \ a = (0,0,0,0,1,1,1,1,9,9,9,9) \)

iii) \( d = 24, \ a = (0, \ldots, 0, 1, \ldots, 1, 9, \ldots, 9). \)

\[ \text{8 times} \quad \text{8 times} \quad \text{8 times} \]

i) \( S_1 = 0.742, \ S_2 = 0.185, \ S_3 = 0.007 \)

ii) \( S_1 = \cdots = S_4 = 0.098, \ S_5 = \cdots = S_8 = 0.024, \)
\( S_9 = \cdots = S_{12} = 0.001, \)

iii) \( S_1 = \cdots = S_8 = 0.018, \ S_9 = \cdots = S_{16} = 0.004, \)
\( S_{17} = \cdots = S_{24} = 10^{-4}. \)
III- Comparison with randomized QMC approaches

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

Clémentine PRIEUR (joint work with J.Y. Tissot)
III- Comparison with randomized QMC approaches

- mss for $S_1, \ldots, S_8$
- mse for $S_9, \ldots, S_{16}$
- mse for $S_{17}, \ldots, S_{24}$
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),
- …
References

Some references I


Thanks for your attention