Automatic estimation of first-order Sobol' indices using the replication procedure

Lluís Antoni Jiménez Rugama, Laurent Gilquin,
Élise Arnaud, Fred J. Hickernell,
Hervé Monod, Clémentine Prieur

Illinois Institute of Technology
Email: ljimene1@hawk.iit.edu

Wednesday 5^{th} July, 2017

Outline

- Sobol' Indices—What are they?
- Quasi-Monte Carlo Methods
- Replication Procedure

ANOVA

For $f \in L^2\left([0,1]^d\right)$, and $1{:}d = \{1,\dots,d\}$,

$$f(\boldsymbol{x}) = \sum_{u \subseteq 1:d} f_u(\boldsymbol{x})\,, \qquad f_\varnothing = \mu\,,$$

where,

$$f_u(\boldsymbol{x}) = \int_{[0,1]^{d-|u|}} f(\boldsymbol{x}) d\boldsymbol{x}_{-u} - \sum_{v \subset u} f_v(\boldsymbol{x})\,.$$

- $|u|$ the cardinality of u.
- $-u := u^c = 1{:}d \setminus u$.

e.g. $$\underbrace{\mathrm{e}^{x_1}\cos(x_2)}_{f(x)} = \underbrace{(\mathrm{e}-1)\sin(1)}_{f_\varnothing} + \underbrace{(\mathrm{e}^{x_1}-\mathrm{e}+1)\sin(1)}_{f_{\{1\}}} + \underbrace{(\mathrm{e}-1)(\cos(x)-\sin(1))}_{f_{\{2\}}} + \underbrace{(\mathrm{e}^{x_1}-\mathrm{e}+1)(\cos(x)-\sin(1))}_{f_{\{1,2\}}}$$

Variance decomposition

Under the previous definitions,

$$\sigma_{\varnothing}^2 = 0\,, \qquad \sigma_u^2 = \int_{[0,1]^d} f_u(\boldsymbol{x})^2 d\boldsymbol{x}\,, \qquad \sigma^2 = \int_{[0,1]^d} \left(f(\boldsymbol{x}) - \mu\right)^2 d\boldsymbol{x}\,.$$

The ANOVA identity is,

$$\sigma^2 = \sum_{u \subseteq 1:d} \sigma_u^2\,.$$

Sobol' indices

Sobol' (1990) and (2001) introduced the *global sensitivity* indices which measure the variance explained by any dimension subset $u \subseteq 1{:}d$:

$$\underline{\tau}_u^2 = \sum_{\substack{v \subseteq u \\ v \subseteq 1:d}} \sigma_v^2\,, \quad \text{and} \quad \overline{\tau}_u^2 = \sum_{\substack{v \cap u \neq \varnothing \\ v \subseteq 1:d}} \sigma_v^2\,.$$

We have the following properties,

- $\underline{\tau}_u^2 \leqslant \overline{\tau}_u^2$.
- $\underline{\tau}_u^2 + \overline{\tau}_{-u}^2 = \sigma^2$.

Normalized closed first-order Sobol' indices

From now on, we consider the normalized Sobol' indices and $|u| = 1$,

$$S_u = \frac{\tau_u^2}{\sigma^2} = 1 - \frac{\displaystyle\int_{[0,1]^{2d-1}} f(\boldsymbol{x})\big(f(\boldsymbol{x}) - f(\boldsymbol{x}_u : \boldsymbol{x}'_{-u})\big) d\boldsymbol{x} d\boldsymbol{x}'_{-u}}{\displaystyle\int_{[0,1]^d} f(\boldsymbol{x})^2 d\boldsymbol{x} - \left(\int_{[0,1]^d} f(\boldsymbol{x}) d\boldsymbol{x}\right)^2},$$

satisfying $0 \leqslant S_u \leqslant 1$. More specifically, S_u is composed by,

$$S_u = 1 - \frac{\mu_1}{\mu_2 - \mu_3^2}, \quad \text{where} \begin{cases} \mu_1 \text{ is a } 2d-1 \text{ dim. integral.} \\ \mu_2 \text{ is a } d \text{ dim. integral.} \\ \mu_3 \text{ is a } d \text{ dim. integral.} \end{cases}$$

Error bounds for S_u require more care than error bounds for the μ_j.

Improving the estimator

Suppose that $\boldsymbol{\mu} := (\mu_1, \mu_2, \mu_3) \in [\widehat{\boldsymbol{\mu}} - \textbf{err}, \widehat{\boldsymbol{\mu}} + \textbf{err}]$ for

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=0}^{n-1} g_j(\boldsymbol{x}_i)$$

and some data-based $\widehat{\boldsymbol{\mu}}$ and **err**. A natural way to estimate S_u is

$$\widehat{S}_u = 1 - \frac{\widehat{\mu}_1}{\widehat{\mu}_2 - \widehat{\mu}_3^2},$$

Nonetheless,

$$\widetilde{S}_u = 1 - \frac{1}{2}\left(\max_{\boldsymbol{\mu} \in [\widehat{\boldsymbol{\mu}} - \textbf{err}, \widehat{\boldsymbol{\mu}} + \textbf{err}]} \frac{\mu_1}{\mu_2 - \mu_3^2} + \min_{\boldsymbol{\mu} \in [\widehat{\boldsymbol{\mu}} - \textbf{err}, \widehat{\boldsymbol{\mu}} + \textbf{err}]} \frac{\mu_1}{\mu_2 - \mu_3^2} \right),$$

guarantees the tightest absolute error bound.

Why?

For instance, if $\boldsymbol{\mu} \in [1,1] \times [1,3] \times [0,0]$, then

$$\frac{1}{3} \leqslant S_u \leqslant 1,$$
$$\left|S_u - \widehat{S}_u\right| = \left|S_u - \frac{1}{2}\right| \leqslant \frac{1}{2},$$
$$\left|S_u - \widetilde{S}_u\right| = \left|S_u - \frac{2}{3}\right| \leqslant \frac{1}{3},$$

In this case, $1/3$ is the smallest error bound possible.

A deeper study is provided in (Hickernell et al., 2017+) and (Jiménez Rugama and Gilquin, 2017).

Outline

- Sobol' Indices
- Quasi-Monte Carlo Methods—How can we estimate S_u efficiently?
- Replication Procedure

Adaptive quasi-Monte Carlo cubature

To estimate S_u using $\widetilde{S}_u$ we need to approximate μ_1, μ_2, and μ_3 such that $\mu_j \in [\widehat{\mu}_j - \mathsf{err}_j, \widehat{\mu}_j + \mathsf{err}_j]$.

Assuming that integrands defining μ_1, μ_2, and μ_3 satisfy some specific conditions on the decay of the Fourier coefficients, in (Hickernell and Jiménez Rugama, 2016) and (Jiménez Rugama and Hickernell, 2016) we provided two adaptive quasi-Monte Carlo cubatures that compute $\widehat{\mu}_{j,n}$ and $\mathsf{err}_{j,n}$ such that

$$\left|\mu_j - \widehat{\mu}_{j,n}\right| \leqslant \mathsf{err}_{j,n} \tag{1}$$

Given the error tolerance ε, the number of points n is increased until

$$\max_{\boldsymbol{\mu}\in[\widehat{\boldsymbol{\mu}}_{j,n}-\mathbf{err}_{j,n},\widehat{\boldsymbol{\mu}}_{j,n}+\mathbf{err}_{j,n}]} \left|S_u(\boldsymbol{\mu}) - \widetilde{S}_u\right| \leqslant \varepsilon$$

Estimating μ_1, μ_2, and μ_3 automatically

Given ε and $\boldsymbol{x} \mapsto f(\boldsymbol{x})$, we want $\widehat{\mu}$ such that

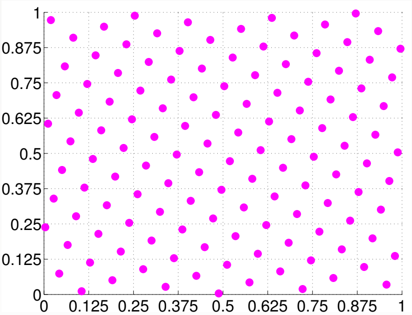
$$\left| \int_{[0,1]^d} f(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - \widehat{\mu}\left(\boldsymbol{x} \mapsto f(\boldsymbol{x}), \varepsilon\right) \right| \leqslant \varepsilon,$$

where

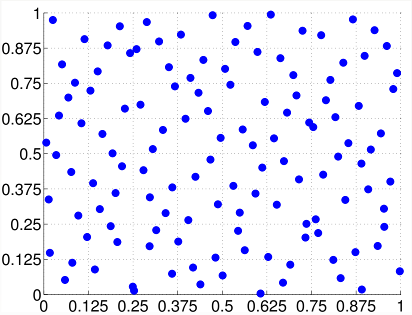
$$\widehat{\mu}\left(\boldsymbol{x} \mapsto f(\boldsymbol{x}), \varepsilon\right) = \frac{1}{2^m} \sum_{i=0}^{2^m-1} f(\boldsymbol{z}_i \oplus \boldsymbol{\Delta}),$$

for some automatic choice of m and $\{\boldsymbol{z}_i\}_{i=0}^{\infty} \in \begin{Bmatrix} \text{Lattice} \\ \text{Digital} \end{Bmatrix}$ sequence.

Examples of sequences



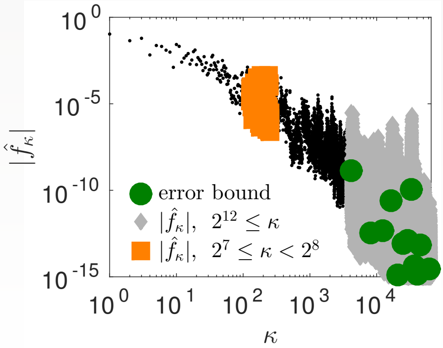
Shifted rank-1 lattice sequence with generating vector $(1, 47)$.



Digitally shifted scrambled Sobol' sequence.

Adaptive algorithm for integrands $f \in \mathcal{C}$

$$\left| \int_{[0,1]^d} f(\boldsymbol{x})\, d\boldsymbol{x} - \frac{1}{2^m} \sum_{i=0}^{2^m-1} f(\boldsymbol{x}_i) \right| \leqslant \overbrace{\sum \bullet}^{\text{Dual net/lat Fourier coef}} \leqslant \mathfrak{C}(r,m) \sum_{\kappa=\lfloor 2^{m-r-1} \rfloor}^{2^{m-r}-1} \left| \tilde{f}_{m,\kappa} \right| \leqslant \varepsilon$$



$$\mathcal{C} = \left\{ \begin{array}{l} \sum \blacksquare \text{ bounds } \sum \blacklozenge \\ \sum \blacklozenge \text{ bounds } \sum \bullet \end{array} \right\}$$

Outline

- Sobol' Indices
- Quasi-Monte Carlo Methods
- Replication Procedure—Reducing the number of function evaluations to compute *first-order* indices.

Number of function evaluations to estimate $\widetilde{S}_1, \ldots, \widetilde{S}_d$

Computing all the indices one by one, if one requires n points for each estimation, the total number of function evaluations is

$$2dn\,,$$

However, if all indices are computed together, some evaluations can be saved. Therefore, the number of function evaluations becomes

$$(1+d)n\,,$$

Finally, under a special set of quasi-Monte Carlo sequences, this number can be decreased to

$$2n\,.$$

Normalized first-order Sobol' indices

Given $\boldsymbol{x}, \boldsymbol{x}' \in [0,1]^d$, we define the following point,

$$(\boldsymbol{x}_u : \boldsymbol{x}'_{-u}) := (x'_1, \ldots, x'_{u-1}, x_u, x'_{u+1}, \ldots, x'_d) \in [0,1]^d.$$

This point is used in the definition of S_u:

$$S_u = 1 - \frac{\int_{[0,1]^{2d-1}} f(\boldsymbol{x})\big(f(\boldsymbol{x}) - f(\boldsymbol{x}_u : \boldsymbol{x}'_{-u})\big) d\boldsymbol{x} d\boldsymbol{x}'_{-u}}{\int_{[0,1]^d} f(\boldsymbol{x})^2 d\boldsymbol{x} - \left(\int_{[0,1]^d} f(\boldsymbol{x}) d\boldsymbol{x}\right)^2}.$$

Replicated designs

For the cubature, we must evaluate

$$
\begin{aligned}
f(\boldsymbol{x}_i) &= f(x_{i,1},\ldots,x_{i,u-1},x_{i,u},x_{i,u+1},\ldots,x_{i,d})\,,\\
f(\boldsymbol{x}_{i,u},\boldsymbol{x}'_{i,-u}) &= f(x'_{i,1},\ldots,x'_{i,u-1},x_{i,u},x'_{i,u+1},\ldots,x'_{i,d})
\end{aligned}
$$

for all u. We show that $f(\boldsymbol{x}_i)$ and $f(\boldsymbol{x}'_i)$ are enough. We focus on well uniformly distributed points $\boldsymbol{x}_i$ and $\boldsymbol{x}'_i$ such that,

$$
\begin{pmatrix}
x_{0,1} & \cdots & x_{0,d}\\
\vdots & \vdots & \vdots\\
x_{i,1} & \cdots & x_{i,d}\\
\vdots & \vdots & \vdots
\end{pmatrix},
\qquad
\begin{pmatrix}
x'_{0,1} & \cdots & x'_{0,d}\\
\vdots & \vdots & \vdots\\
x'_{i,1} & \cdots & x'_{i,d}\\
\vdots & \vdots & \vdots
\end{pmatrix}
=
\begin{pmatrix}
x_{\pi_1(0),1} & \cdots & x_{\pi_d(0),d}\\
\vdots & \ddots & \vdots\\
x_{\pi_1(i),1} & \cdots & x_{\pi_d(i),d}\\
\vdots & & \vdots
\end{pmatrix},
$$

where the permutations π_u reorder the $\boldsymbol{x}_{i,u}$ into the $\boldsymbol{x}'_{i,u}$.

Sobol' points

By construction, Sobol' points have this property when $n = 2^m$. For instance, if $d = 2$ we can use a fourth dimensional Sobol' sequnce $\{\boldsymbol{z}_i\}_{i\in\mathbb{N}_0}$:

$$\begin{pmatrix} z_1 & z_2 \\ x_1 & x_2 \\ 0 & 0 \\ 0.5 & 0.5 \\ 0.25 & 0.75 \\ 0.75 & 0.25 \\ 0.125 & 0.625 \\ 0.625 & 0.125 \\ 0.375 & 0.375 \\ 0.875 & 0.875 \\ \vdots & \vdots \end{pmatrix}, \quad \begin{pmatrix} z_3 & z_4 \\ x'_1 & x'_2 \\ 0 & 0 \\ 0.5 & 0.5 \\ 0.25 & 0.75 \\ 0.75 & 0.25 \\ 0.875 & 0.875 \\ 0.375 & 0.375 \\ 0.625 & 0.125 \\ 0.125 & 0.625 \\ \vdots & \vdots \end{pmatrix}, \quad \begin{pmatrix} \pi_1 & \pi_2 \\ 0 & 0 \\ 1 & 1 \\ 2 & 2 \\ 3 & 3 \\ 7 & 7 \\ 6 & 6 \\ 5 & 5 \\ 4 & 4 \\ \vdots & \vdots \end{pmatrix}.$$

No need to evaluate $f(\boldsymbol{x}_{i,u} : \boldsymbol{x}'_{i,-u})$ for d different u

Given the right order of our points $\boldsymbol{x}'_i$ into $\boldsymbol{x}_i$, i.e. π_u^{-1}:

$$\begin{pmatrix} \boldsymbol{x}'_{\pi_u^{-1}(0)} \\ \vdots \\ \boldsymbol{x}'_{\pi_u^{-1}(n)} \\ \vdots \end{pmatrix} = \begin{pmatrix} x'_{\pi_u^{-1}(0),1} & \cdots & x_{0,u} & \cdots & x'_{\pi_u^{-1}(0),d} \\ \vdots & & \vdots & & \vdots \\ x'_{\pi_u^{-1}(n),1} & \cdots & x_{n,u} & \cdots & x'_{\pi_u^{-1}(n),d} \\ \vdots & & \vdots & & \vdots \end{pmatrix}.$$

Thus, evaluating $f(\boldsymbol{x}'_i)$, one can directly obtain the $f(\boldsymbol{x}_{i,u} : \boldsymbol{x}'_{i,-u})$:

$$\begin{pmatrix} f(\boldsymbol{x}'_0) \\ \vdots \\ f(\boldsymbol{x}'_i) \\ \vdots \end{pmatrix} = \begin{pmatrix} y_0 \\ \vdots \\ y_i \\ \vdots \end{pmatrix} \Longrightarrow \begin{pmatrix} f(\boldsymbol{x}_{0,u} : \boldsymbol{x}'_{0,-u}) \\ \vdots \\ f(\boldsymbol{x}_{i,u} : \boldsymbol{x}'_{i,-u}) \\ \vdots \end{pmatrix} := \begin{pmatrix} y_{\pi_u^{-1}(0)} \\ \vdots \\ y_{\pi_u^{-1}(i)} \\ \vdots \end{pmatrix}$$

Sobol' Indices Example from Bratley et al. (1992)

$$f(\boldsymbol{x}) = -x_1 + x_1x_2 - x_1x_2x_3 + \cdots + x_1x_2x_3x_4x_5x_6$$

$$\varepsilon = 1\mathsf{E}{-3}, \qquad n = 65\ 536$$

u	1	2	3	4	5	6
S_u	0.6529	0.1791	0.0370	0.0133	0.0015	0.0015
$\widetilde{S}_u$	0.6523	0.1796	0.0372	0.0136	0.0015	0.0017
$\widehat{S}_u = S_u(\widehat{\boldsymbol{\mu}}_n)$	0.6396	0.1787	0.0319	0.0124	0.0000	0.0000

Summary

- We can study how each dimension explains the overall variance of a model using Sobol' Indices.
- Our quasi-Monte Carlo automatic cubatures can be adapted to estimate these indices automatically.
- *First-order* Sobol' Indices can be estimated using only $2n$ quasi-Monte Carlo function evaluations.
- Under some conditions, one may also use scrambled Sobol' sequences that keep the replication property.
- The same algorithms can be designed for rank-1 lattices.

References I

Bratley, P., B. L. Fox, and H. Niederreiter. 1992. *Implementation and tests of low-discrepancy sequences* **2**, 195–213.

Cools, R. and D. Nuyens (eds.) 2016. *Monte Carlo and quasi-Monte Carlo methods: MCQMC, Leuven, Belgium, April 2014*, Springer Proceedings in Mathematics and Statistics, vol. 163, Springer-Verlag, Berlin.

Gilquin, L., Ll. A. Jiménez Rugama, E. Arnaud, F. J. Hickernell, H. Mond, and C. Prieur. 2016. *Iterative construction of replicated designs based on Sobol' sequences*, C. R. Math. Acad. Sci. Paris **355**, 10–14.

Hickernell, F. J. and Ll. A. Jiménez Rugama. 2016. *Reliable adaptive cubature using digital sequences*, Monte Carlo and quasi-Monte Carlo methods: MCQMC, Leuven, Belgium, April 2014. arXiv:1410.8615 [math.NA].

Hickernell, F. J., Ll. A. Jiménez Rugama, and D. Li. 2017+. *Adaptive quasi-monte carlo methods*. Under review.

Jiménez Rugama, Ll. A. and L. Gilquin. 2017. *Reliable error estimation for Sobol' indices*, Statistics and Computing. in press.

References II

Jiménez Rugama, Ll. A. and F. J. Hickernell. 2016. *Adaptive multidimensional integration based on rank-1 lattices*, Monte Carlo and quasi-Monte Carlo methods: MCQMC, Leuven, Belgium, April 2014, pp. 407–422. arXiv:1411.1966.

Owen, Art B. 2013. *Variance components and generalized Sobol' indices*, SIAM/ASA Journal on Uncertainty Quantification **1**, no. 1, 19–41.

Saltelli, Andrea. 2002. *Making best use of model evaluations to compute sensitivity indices*, Computer Physics Communications **145**, no. 2, 280 –297.

Sobol', I. M. 1990. *On sensitivity estimation for nonlinear mathematical models*, Matem. Mod. **2**, no. 1, 112–118.

Sobol', I.M. 2001. *Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates*, Mathematics and Computers in Simulation (MATCOM) **55**, no. 1, 271–280.

Tissot, J.-Y. and C. Prieur. 2015. *A randomized orthogonal array-based procedure for the estimation of first- and second-order Sobol' indices*, Journal of Statistical Computation and Simulation **85**, no. 7, 1358–1381.