Minimizing the Number of Function Evaluations to Estimate Sobol' Indices Using Quasi-Monte Carlo

Lluís Antoni Jiménez Rugama
Joint work with: Fred J. Hickernell (IIT),
Clémentine Prieur (Univ. Grenoble), Elise Arnaud (Univ. Grenoble),
Hervé Monod (INRA), and Laurent Gilquin (Univ. Grenoble)

Room 120, Bldg E1, Department of Applied Mathematics
Illinois Institute of Technology, Chicago, 60616 IL
Email: ljimene1@hawk.iit.edu

Thursday 26th May, 2016

Outline

- Introduction
  - ANOVA
  - Sobol' Indices
- Quasi-Monte Carlo Methods
- Replicated Method

Outline

- Introduction
  - ANOVA—The ANalysis Of VAriance decomposition.
  - Sobol' Indices
- Quasi-Monte Carlo Methods
- Replicated Method

ANOVA

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

$$f(\boldsymbol{x}) = \sum_{u \subseteq \mathcal{D}} f_u(\boldsymbol{x}) \,, \qquad f_\varnothing = \mu \,,$$

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.
- $-u := u^c = \mathcal{D} \backslash u$.

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,

$$\sigma^2 = \sum_{u \subseteq \mathcal{D}} \sigma_u^2\,.$$

Outline

- Introduction
  - ANOVA
  - Sobol' Indices—Measuring the importance of each input.
- Quasi-Monte Carlo Methods
- Replicated Method

Sobol' Indices

Sobol' introduced the *global sensitivity* indices which measure the variance explained by any dimension subset $u \in \mathcal{D}$:

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

We have the following properties,

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

Sobol' Indices - Probabilistic Framework

For $\boldsymbol{X} \sim U[0,1]^d$, Sobol' indices can also be presented in the following form,

$$\underline{\tau}_u^2 = \mathrm{Var}\left[\mathbb{E}\left(f(\boldsymbol{X})|\boldsymbol{X}_u\right)\right] = \mathrm{Var}\left(f(\boldsymbol{X})\right) - \mathbb{E}\left[\mathrm{Var}\left(f(\boldsymbol{X})|\boldsymbol{X}_u\right)\right],$$

$$\overline{\tau}_u^2 = \mathrm{Var}\left(f(\boldsymbol{X})\right) - \mathrm{Var}\left[\mathbb{E}\left(f(\boldsymbol{X})|\boldsymbol{X}_{-u}\right)\right] = \mathbb{E}\left[\mathrm{Var}\left(f(\boldsymbol{X})|\boldsymbol{X}_{-u}\right)\right].$$

The Normalized Sobol' Indices

One may also use the normalized definition of the Sobol' indices,

$$S_u = \frac{\underline{\tau}_u^2}{\sigma^2} = \frac{\mathrm{Var}\left[\mathbb{E}\left(f(\boldsymbol{X})|\boldsymbol{X}_u\right)\right]}{\mathrm{Var}\left(f(\boldsymbol{X})\right)} = 1 - \frac{\mathbb{E}\left[\mathrm{Var}\left(f(\boldsymbol{X})|\boldsymbol{X}_u\right)\right]}{\mathrm{Var}\left(f(\boldsymbol{X})\right)},$$

$$S_u^{\text{tot}} = \frac{\overline{\tau}_u^2}{\sigma^2} = 1 - \frac{\mathrm{Var}\left[\mathbb{E}\left(f(\boldsymbol{X})|\boldsymbol{X}_{-u}\right)\right]}{\mathrm{Var}\left(f(\boldsymbol{X})\right)} = \frac{\mathbb{E}\left[\mathrm{Var}\left(f(\boldsymbol{X})|\boldsymbol{X}_{-u}\right)\right]}{\mathrm{Var}\left(f(\boldsymbol{X})\right)}.$$

satisfying $0 \leqslant S_u \leqslant S_u^{\text{tot}} \leqslant 1$.

The Normalized Sobol' Indices

One may also use the normalized definition of the Sobol' indices,

$$S_u = \frac{\underline{\tau}_u^2}{\sigma^2} = \frac{\mathrm{Var}\left[\mathbb{E}\left(f(\boldsymbol{X})|\boldsymbol{X}_u\right)\right]}{\mathrm{Var}\left(f(\boldsymbol{X})\right)} = 1 - \frac{\mathbb{E}\left[\mathrm{Var}\left(f(\boldsymbol{X})|\boldsymbol{X}_u\right)\right]}{\mathrm{Var}\left(f(\boldsymbol{X})\right)},$$

$$S_u^{\text{tot}} = \frac{\overline{\tau}_u^2}{\sigma^2} = 1 - \frac{\mathrm{Var}\left[\mathbb{E}\left(f(\boldsymbol{X})|\boldsymbol{X}_{-u}\right)\right]}{\mathrm{Var}\left(f(\boldsymbol{X})\right)} = \frac{\mathbb{E}\left[\mathrm{Var}\left(f(\boldsymbol{X})|\boldsymbol{X}_{-u}\right)\right]}{\mathrm{Var}\left(f(\boldsymbol{X})\right)}.$$

satisfying $0 \leqslant S_u \leqslant S_u^{\text{tot}} \leqslant 1$. More specifically, S_u is composed by,

$$S_u = 1 - \frac{I^{(1)}}{I^{(2)} - \left(I^{(3)}\right)^2}, \quad \text{where} \begin{cases} I^{(1)} \text{ is a } 2d - |u| \text{ dim. integral.} \\ I^{(2)} \text{ is a } d \text{ dim. integral.} \\ I^{(3)} \text{ is a } d \text{ dim. integral.} \end{cases}$$

Error bounds for S_u require more care than error bounds for $I^{(k)}$.

Outline

- Introduction
  - ANOVA
  - Sobol' Indices
- Quasi-Monte Carlo Methods—How can we compute high dimensional integrals efficiently?
- Replicated Method

Why Quasi-Monte Carlo?

To estimate S_u we need to approximate $I^{(1)}$, $I^{(2)}$, and $I^{(3)}$. However, in high dimensions we need a suitable technique:

Method	Convergence
Trapezoidal rule:	$\mathcal{O}(n^{-2/d})$
Simpson's rule:	$\mathcal{O}(n^{-4/d})$
IID Monte Carlo:	$\mathcal{O}(n^{-1/2})$
Quasi-Monte Carlo:	$\mathcal{O}(n^{-1+\varepsilon})$

(n: number of data points)

Estimating $I^{(1)}$, $I^{(2)}$, and $I^{(3)}$ automatically

Given ε_a and $\boldsymbol{x} \mapsto f(\boldsymbol{x})$, we want $\hat{I}$ such that

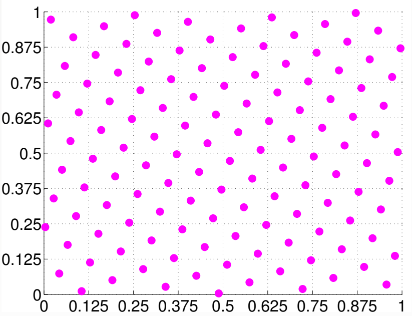
$$\left| \int_{[0,1)^d} f(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} - \hat{I}\left(\boldsymbol{x} \mapsto f(\boldsymbol{x}), \varepsilon_a\right) \right| \leqslant \varepsilon_a,$$

where

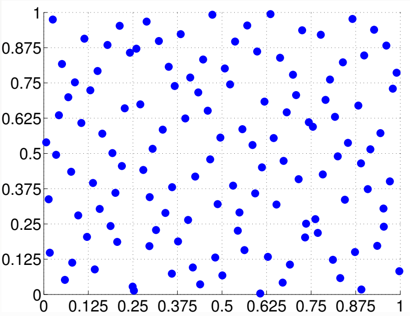
$$\hat{I}\left(\boldsymbol{x} \mapsto f(\boldsymbol{x}), \varepsilon_a\right) = \frac{1}{2^m} \sum_{i=0}^{2^m-1} f(\boldsymbol{z}_i),$$

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

Examples of Sequences



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



Digitally shifted scrambled Sobol' sequence.

Outline

Normalized *First-Order* Sobol' Indices

In this particular case, we consider $|u| = 1$ and want to estimate $S_u = \sigma_u^2/\sigma^2$. For this purpose, given $\boldsymbol{x}, \boldsymbol{x}' \in [0,1]^d$, we define the following point,

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

Thus, one can use the following integral form to build an estimator:

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

Number of Function Evaluations

We will focus on reducing the number of function evaluations, and to estimate σ_u^2/σ^2, only g and g_u are evaluated.

Computing all the indices one by one, if one requires n points for each estimation, the total number of function evaluations of g and g_u are

$$2dn\,,$$

However, if all indices are computed together, g only needs to be evaluated once. Therefore, the number of function evaluations becomes

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

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

$$2n\,.$$

Replicated Designs

Functions g and g_u only share input dimension u:

$$g(\boldsymbol{x}, \boldsymbol{x}') = f(x_1, \ldots, x_{u-1}, x_u, x_{u+1}, \ldots, x_d)\,,$$
$$g_u(\boldsymbol{x}, \boldsymbol{x}') = f(x'_1, \ldots, x'_{u-1}, x_u, x'_{u+1}, \ldots, x'_d)\,.$$

Hence, we can construct our points $\boldsymbol{x}'_i$ as follows,

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

The Right Function Values

Given the right order of points:

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

Therefore, we only need to evaluate $g_u(\boldsymbol{x}, \boldsymbol{x}')$ once:

$$\begin{pmatrix} f(\boldsymbol{x}'_0) \\ \vdots \\ f(\boldsymbol{x}'_n) \\ \vdots \end{pmatrix} = \begin{pmatrix} y_0 \\ \vdots \\ y_n \\ \vdots \end{pmatrix} \Longrightarrow \begin{pmatrix} g_u(\boldsymbol{x}_0, \boldsymbol{x}'_0) \\ \vdots \\ g_u(\boldsymbol{x}_n, \boldsymbol{x}'_n) \\ \vdots \end{pmatrix} = \begin{pmatrix} y_{\pi_u^{-1}(0)} \\ \vdots \\ y_{\pi_u^{-1}(n)} \\ \vdots \end{pmatrix}$$

Conclusions

- 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 (not depending on d).

References I

Hickernell, F. J. and Ll. A. Jiménez Rugama. 2015+. *Reliable adaptive cubature using digital sequences*, Monte Carlo and quasi-Monte Carlo methods 2014. to appear, arXiv:1410.8615 [math.NA].

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

Joe, Stephen and Frances Y. Kuo. 2008. *Constructing sobol sequences with better two-dimensional projections*, SIAM J. Scientific Computing **30**, no. 5, 2635–2654.

M. D. McKay, W. J. Conover, R. J. Beckman. 1979. *A comparison of three methods for selecting values of input variables in the analysis of output from a computer code*, Technometrics **21**, no. 2, 239–245.

Niederreiter, H. 1992. *Random number generation and quasi-Monte Carlo methods*, CBMS-NSF Regional Conference Series in Applied Mathematics, SIAM, Philadelphia.

Owen, A. B. 1998. *Scrambling Sobol' and Niederreiter-Xing points*, J. Complexity **14**, 466–489.

References II

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

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

Sobol', I. M. 1967. *The distribution of points in a cube and the approximate evaluation of integrals*, U.S.S.R. Comput. Math. and Math. Phys. **7**, 86–112.

Tissot, Jean-Yves and Clémentine Prieur. 2014. *A randomized Orthogonal Array-based procedure for the estimation of first- and second-order Sobol' indices*, Journal of Statistical Computation and Simulation, 1–24.