On the replication of Sobol' sequences to estimate main effects of model inputs

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

University Grenoble Alpes, Illinois Institute of Technology, INRA Jouy-en-Josas

1. Problem Statement

Denote by $\boldsymbol{x} \in [0,1)^s$ a point with components $x_1, \ldots, x_s$. We set $\mathcal{D} = \{1, \ldots, s\}$. Let u be a subset of $\mathcal{D}$, $-u$ its complement and $|u|$ its cardinality. $\boldsymbol{x}_u$ represents a point in $[0,1)^u$ with components $x_j, j \in u$. Given two points $\boldsymbol{x}$ and $\boldsymbol{x}'$, we define the following hybrid point:

$$\boldsymbol{z}_u = (\boldsymbol{x}_u : \boldsymbol{x}'_{-u}), \text{ where: } \begin{cases} z_j = x_j, & \text{if } j \in u \\ z_j = x'_j, & \text{if } j \in -u \end{cases}$$

Let f be the symbol representing the numerical model considered. We assume $f \in \mathcal{L}^2[0,1]^s$. Denote by μ and σ^2 the mean and variance of f. The uncertainty on $\boldsymbol{x}$ is modeled by random variables such that $\boldsymbol{x} \overset{iid}{\sim} \mathcal{U}[0,1)^s$. Recall the Hoeffding [1] decomposition of f:

$$f(\boldsymbol{x}) = \mu + \sum_{\substack{u \subseteq \{1,\ldots,s\} \\ u \neq \varnothing}} f_u(\boldsymbol{x}_u), \tag{1}$$

where: μ is a constant and $\forall u \subseteq \mathcal{D}, u \neq \varnothing, \forall j \in u : \int_{[0,1]} f_u(\boldsymbol{x}_u) dx_j = 0$. Due to orthogonality, by taking the variance of each side in Equation 1 we obtain the following decomposition of the variance:

$$\sigma^2 = \sum_{\substack{u \subseteq \{1,\ldots,s\} \\ u \neq \varnothing}} \sigma^2_v.$$

The problem of interest is the estimation of the quantities:

$$\tau^2_u = \sum_{v \subseteq u} \sigma^2_v, \qquad u \in \mathcal{D}.$$

These quantities correspond to the partial sensitivity Sobol' indices [2] often found written in the literature in their normalized form: τ^2_u/σ^2. When, $|u| = 1$, τ^2_u corresponds to the first-order Sobol' index that evaluates the main effect of

the variable $\boldsymbol{x}_u$. Higher order indices evaluate higher order interactions between variables. In [3], Owen proposed to use the following expression for τ_u^2:

$$\tau_u^2 = \int_{[0,1)^{2s-1}} (f(\boldsymbol{x}) - \mu)(f(\boldsymbol{x}_u : \boldsymbol{x}'_{-u}) - \mu) d\boldsymbol{x} d\boldsymbol{x}'_{-u}, \qquad (2)$$

taking μ as introduced in [4, 5]:

$$\mu = \frac{1}{2} \int_{[0,1)^{2s-1}} f(\boldsymbol{x}) + f(\boldsymbol{x}_u : \boldsymbol{x}'_{-u}) d\boldsymbol{x} d\boldsymbol{x}'_{-u}.$$

The natural way to estimate these two quantities is for $(\boldsymbol{x}_i, \boldsymbol{x}'_i) \overset{iid}{\sim} [0,1)^{2s}$ via:

$$\widehat{\tau_u^2} = \frac{1}{n} \sum_{i=1}^{n} (f(\boldsymbol{x}_i) - \widehat{\mu})(f(\boldsymbol{x}_{i,u} : \boldsymbol{x}'_{i,-u}) - \widehat{\mu}), \qquad (3)$$

$$\widehat{\mu} = \frac{1}{2n} \sum_{i=1}^{n} f(\boldsymbol{x}_i) + f(\boldsymbol{x}_{i,u} : \boldsymbol{x}'_{i,-u}).$$

We consider here a quasi-Monte Carlo sampling strategy to evaluate these two estimators. We focus on the estimation of all first-order Sobol' indices τ_u^2, $u \in \mathcal{D}$. The classical estimation procedure introduced by Sobol' requires to evaluate the quantity $f(\boldsymbol{x}_u : \boldsymbol{x}'_{-u})$ n times for each u while the quantity $f(\boldsymbol{x})$ is only evaluated n times once for all u. The total cost sums up to $n(d+1)$ evaluations of f. This cost can rapidly become prohibitive for models whose evaluation is time consuming.

An improvement to reduce this cost lies in the use of replicated designs. The notion of replicated designs was first introduced by McKay [6] through the introduction of replicated Latin Hypercubes samples. Here, we define replicated designs in a wider framework:

Definition 1 (Replicated designs). Let $\mathcal{P} = \{\boldsymbol{x}_i\}_{i=0}^{n-1}$ and $\mathcal{P}' = \{\boldsymbol{x}'_i\}_{i=0}^{n-1}$ be two point sets, $\boldsymbol{x}_i, \boldsymbol{x}'_i \in [0,1)^s$. Let $\mathcal{P}^u = \{\boldsymbol{x}_{i,u}\}_{i=0}^{n-1}$ (resp. $\mathcal{P}'^u$), $u \subsetneq \mathcal{D}$, denote the subset of dimensions of $\mathcal{P}$ (resp. $\mathcal{P}'$) indexed by u. We say that $\mathcal{P}$ and $\mathcal{P}'$ are two replicated designs of order r if:
$\forall u \subsetneq \mathcal{D}$ such that $|u| = r$, $\mathcal{P}^u$ and $\mathcal{P}'^u$ are the same point set in $[0,1)^r$.

The replication procedure introduced and studied in [7, 8] allows to estimate all first-order Sobol' indices with two replicated designs of order 1. In this procedure, the quantity $f(\boldsymbol{x}_u : \boldsymbol{x}'_{-u})$ is only evaluated n times once for all u, resulting in a total of $2n$ evaluations of f.

Our objective here is to combine the use of Sobol' sequences [9] with the replication procedure to iteratively estimate all quantities τ_u^2, $u \in \mathcal{D}$, $|u| = 1$. This approach reduces to construct two replicated Sobol' sequences iteratively refined with new point sets. In this fashion, two approaches are considered: one multiplicative and one additive. These two approaches are detailed in Section 3. In the next section, we provide some background on digital nets and sequences.

2

2. Digital Nets Background

In this section we introduce a particular type of quasi-Monte Carlo points to estimate the quantities in (2) for $|u| = 1$, according to the replication method.

The concepts of (t, m, s)-net and (t, s)-sequence were first introduced by Niederreiter in [10]. ~~Digital nets are accurately defined as (t, m, s)-nets in $[0, 1)^s$ whose quality is measured by the parameter t, called t-value.~~

Definition 2 $((t, m, s) - net)$**.** Let $\mathcal{A}$ be the set of all elementary intervals $A \in [0, 1)^s$ where $A = \prod_{j=1}^{s} [\alpha_j b^{-\gamma_j}, (\alpha_j + 1) b^{-\gamma_j})$, with integers $s \geqslant 1$, $b \geqslant 2$, $\gamma_j \geqslant 0$, and $b^{\gamma_j} > \alpha_j \geqslant 0$. For $m \geqslant t \geqslant 0$, the point set $\mathcal{P} \in [0, 1)^s$ with b^m points is a $(t, m, s) - net$ in base b if every A with volume b^{t-m} contains b^t points of $\mathcal{P}$.

A (t, m, s)-net is defined such that all elementary intervals of volume at least b^{t-m} will enclose a proportional number of points. The most evenly spread nets are $(0, m, s)$-nets, since each elementary interval of the smallest volume possible, b^{-m}, contains exactly one point. Note that if $t' > t$, (t, m, s)-nets are always (t', m, s)-nets.

Definition 3 $((t, s)$-sequence$)$**.** For integers $s \geqslant 1$, $b \geqslant 2$, and $t \geqslant 0$, the sequence $\{\boldsymbol{x}_i\}_{i \in \mathbb{N}_0}$ is a (t, s)-sequence in base b, if for every set $\mathcal{P}_{\ell, m} = \{\boldsymbol{x}_i\}_{i=\ell b^m}^{(\ell+1)b^m - 1}$ with $\ell \geqslant 0$ and $m \geqslant t$, $\mathcal{P}_{\ell, m}$ is a (t, m, s)-net in base b.

~~Digital sequences are usually computed with $\infty \times \infty$-matrices $C_1, \ldots, C_s$ over $\mathbb{F}_b$, often referred in the literature as generating matrices. However, to construct the first b^m points, one only needs the knowledge of the blocks $C_1^{\infty \times m}, \ldots, C_s^{\infty \times m}$.~~ This is due to its construction, described as follows. For each $i = 0, \ldots, b^m - 1$, the point $\boldsymbol{x}_i = (x_{i,1}, \ldots, x_{i,s})^\mathsf{T}$ of the sequence is obtained dimension-wise:

$$(x_{i,j,1}, x_{i,j,2}, \ldots)^\mathsf{T} = C_j^{\infty \times m} \boldsymbol{i}, \qquad j = 1, \ldots, s, \tag{4}$$

where $x_{i,j} = \sum_{k \geqslant 1} x_{i,j,k} b^{-k}$ and $\boldsymbol{i} = (i_0, \ldots, i_{m-1})^\mathsf{T}$ is the b-ary decomposition of i.

In order to relate digital nets and sequences to the definition of replicated design, we introduce the following concept.

Definition 4. Two digital sequences $\{\boldsymbol{x}_i\}_{i \in \mathbb{N}_0}$ and $\{\boldsymbol{x}'_i\}_{i \in \mathbb{N}_0}$ are *digitally replicated* of order r if for all $m \geqslant 0$, $\{\boldsymbol{x}_i\}_{i=0}^{b^m-1}$ and $\{\boldsymbol{x}'_i\}_{i=0}^{b^m-1}$ are two replicated designs of order r.

Sobol' sequences are digital sequences attractive for their easy implementation and optimized properties [11]. In section 3, we present two approaches to construct replicated point sets based on Sobol' sequences. The multiplicative approach uses two Sobol' sequences that are replicated by construction, while the additive approach takes an initial number of points of two Sobol' sequences and applies digital shifts and scramblings to extend the point set. In this case, at step k the size of each replicated Sobol' sequence is $k \times 2^m$ where 2^m is the size of the initial Sobol' sequence.

3. Iterative constructions of replicated point sets

Sobol' sequences are particularly interesting for having ~~unitriangular~~ generating matrices over $\mathbb{F}_2$. This property leads to the result below,

Lemma 1. *All Sobol' sequences are digitally replicated of order 1.*

Proof. Consider any two s-dimensional Sobol' sequences generated by $C_1, \ldots, C_s$ and $C'_1, \ldots, C'_s$. Since these matrices are upper triangular, to compute the first 2^m points one only needs the knowledge of the $m \times m$ blocks. Because any two matrices $C_j^{m \times m}$ and $C_j'^{m \times m}$ are square and full rank, the products $C_j^{m \times m} \boldsymbol{i}$ and $C_j'^{m \times m} \boldsymbol{i}$ are one-to one and onto, for all $i = 0, \ldots, 2^m - 1$. Therefore, they generate the same point sets. $\square$

Two different matrices $C^{m \times m}$ and $C'^{m \times m}$ produce the same set of 2^m values. However, they will appear in different order. It is possible to generate them in the same order using the corresponding upper triangular matrix $U^{m \times m}$, such that $C^{m \times m} = C'^{m \times m} U^{m \times m}$. This matrix can be thought as permutation of the input binary vectors $\boldsymbol{i}$.

3.1. Multiplicative approach

Denote by $\mathcal{P}$ and $\mathcal{P}'$ the two designs iteratively refined with new points sets using the multiplicative approach. The goal of this approach is to ~~generate a 2s-dimensional Sobol' sequence~~ with good properties to estimate all first order Sobol' indices. Design $\mathcal{P}$ is formed with the first s dimensions of the sequence and design $\mathcal{P}'$ with the s reminding ones. As shown in Lemma 1, for any $m \geqslant 0$ and 2^m points, each dimension of the sequence will have exactly the same values, although probably with different order. Thus, designs $\mathcal{P}$ and $\mathcal{P}'$ are nested replicated designs of order 1 and both are s-dimensional Sobol' sequences.
~~The only constraint we have is that for each $u = 1, \ldots, s$, to estimate the quantity τ_u^2 the 2s-dimensional Sobol' sequence must ensure that $x_{i,u} = x_{i,u+s}$ for $i \in \mathbb{N}_0$. Therefore, we will need s different 2s-dimensional sequences.~~ These sequences can be constructed using a single 2s-dimensional sequence and considering the corresponding upper triangular matrices U_u.

Given the generators $C_1, \ldots, C_s, C'_1, \ldots, C'_s$, we define the U_u matrices such that $C_u = C'_u U_u$. Then, each sequence generated by $C_1, \ldots, C_s, C'_1 U_u, \ldots, C'_s U_u$ will satisfy $x_{i,u} = x_{i,u+s}$ for all $i \in \mathbb{N}_0$.

Proposition 1. For any $m \geqslant 0$, and t_m depending on m, if the Sobol' (t_m, m, s)-nets are generated by $C_1, \ldots, C_s$, then $C_1 U, \ldots, C_s U$, where U is an upper triangular matrix, generate exactly the same (t_m, m, s)-nets.

Proof. For any m and $2^{m-1} \leqslant i < 2^m$, the image by U, $\boldsymbol{j} = U\boldsymbol{i}$, will also satisfy $2^{m-1} \leqslant j < 2^m$. The reason is because U is upper triangular, therefore $j_{m-1} = i_{m-1}$. Thus, by induction on m, the (t_m, m, s)-nets generated after applying the permutation U contain exactly the same points. $\square$

4

Sobol' generating matrices C are constructed through primitive polynomials and initial directional numbers. How to choose them is deeply discussed in [12]. The t-values of pairwise projections are already optimized in [11]. ~~Thus, we will consider the $2s$-dimensional Sobol' sequence described by them to construct the other s $2s$-dimensional sequences.~~

3.2. Additive Approach

Denote by $\mathcal{P}$ and $\mathcal{P}'$ the two replicated designs iteratively refined with new points sets using the additive approach. These two designs possess a block structure:

$$\mathcal{P} = B_0 \cup B_1 \cup B_2 \ldots$$

$$\mathcal{P}' = B'_0 \cup B'_1 \cup B'_2 \ldots$$

We detail below the steps of the construction of these two designs.

Step 1. We start by fixing m and choosing a set $C_1^m, \ldots, C_s^m$ of generating matrices to construct an initial Sobol' sequence $\{\boldsymbol{x}_i\}_{i=0}^{2^m-1}$ of 2^m points.

Step 2. At step l of the recursive estimation, $l \geq 0$, a new points set $B_l = \{\boldsymbol{x}_i^{(l)}\}_{i=0}^{2^m-1}$ of 2^m points is added to $\mathcal{P}$. B_l is obtained by applying a digital shift to the initial Sobol' sequence. The digital shift operation is applied as follows:

for each $i = 0, \ldots, 2^m - 1$, the points $\boldsymbol{x}_i^{(l)} = (x_{i,1}^{(l)}, \ldots, x_{i,s}^{(l)})^\intercal$ is obtained by:

$$(x_{i,j,1}^{(l)}, \ldots, x_{i,j,m}^{(l)})^\intercal = C_j^m \boldsymbol{i} + \boldsymbol{e}_j, \qquad j = 1, \ldots, s, \qquad (5)$$

where $x_{i,j}^{(l)} = \sum_{k=1}^m x_{i,j,k}^{(l)} 2^{-k}$ and $\boldsymbol{e}_j = (e_{j,0}, \ldots, e_{j,m-1})^\intercal \in ker(C_j^m)$ the kernel of C_j^m.

Step 3. Then, the points set $B'_l = \{\boldsymbol{x'}_i^{(l)}\}_{i=0}^{2^m-1}$ is obtained from B_l using Tezuka's i-binomial scrambling [14]. This scrambling consists in the following:

for each $i = 0, \ldots, 2^m - 1$, the points $\boldsymbol{x'}_i^{(l)} = (x'_{i,1}^{(l)}, \ldots, x'_{i,s}^{(l)})^\intercal$ is obtained by:

$$(x'_{i,j,1}^{(l)}, \ldots, x'_{i,j,m}^{(l)})^\intercal = L_j^m (x_{i,j,1}^{(l)}, \ldots, x_{i,j,m}^{(l)})^\intercal, \qquad (6)$$

5

where $x'^{(l)}_{i,j} = \sum_{k=1}^{m} x'^{(l)}_{i,j,k} 2^{-k}$ and L_j^m is an invertible lower triangular matrix of size $m \times m$ over the Galois field $\mathbb{F}_2$.

Step 4. ~~The two blocks B_l and B'_l are then both randomized using the same Owen's scrambling for each pair of j-th column of the two blocks.~~

Step 5. B_l and B'_l are used to compute model outputs and estimate all first-order indices $\underline{\tau}_u^2, u \in \mathcal{D}$.

Steps 2 to 5 are iterated until a stopping criterion is reached. There is a finite number of choices for the vector $\boldsymbol{e}_j$ and matrix L_j^m respectively constructed in Steps 2 and 3. The choice of m must take into account this limitation.

At each step l, B_l and B'_l are two replicated designs of order 1 thus enabling the estimation of all first-order indices. Denote by K the step at which the recursive estimation stopped. The total cost of the additive approach to estimate all first-order indices equals $2 \times K \times 2^m$.

The additive approach is attractive due to the slow growth of the size of $\mathcal{P}$ and $\mathcal{P}'$. At the opposite of the multiplicative approach, the main drawback is that neither $\mathcal{P}$ or $\mathcal{P}'$ possesses the structure of a Sobol' sequence.

References

[1] W. Hoeffding, A class of statistics with asymptotically normal distributions, Annals of Mathematical Statistics 19 (3) (1948) 293–325.

[2] I. M. Sobol', Sensitivity indices for nonlinear mathematical models, Mathematical Modeling and Computational Experiment 1 (1993) 407–414.

[3] A. B. Owen, Better estimation of small sobol' sensitivity indices, ACM Trans. Model. Comput. Simul. 23 (2), doi:10.1145/2457459.2457460.

[4] H. Monod, C. Naud, D. Makowski, Uncertainty and sensitivity analysis for crop models, chap. 3, Elsevier, 55–100, 2006.

[5] A. Janon, T. Klein, A. Lagnoux, M. Nodet, C. Prieur, Asymptotic normality and efficiency of two Sobol' index estimators, ESAIM Probab. Stat. 18 (2014) 342–364, doi:10.1051/ps/2013040.

[6] M. D. McKay, W. J. Conover, R. J. Beckman, A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics 21 (1979) 239–245, doi:10.2307/1271432.

[7] T. A. Mara, O. R. Joseph, Comparison of some efficient methods to evaluate the main effect of computer model factors, Journal of Statistical Computation and Simulation 78 (2) (2008) 167–178, doi: 10.1080/10629360600964454.

[8] J. Y. Tissot, C. Prieur, A randomized orthogonal array-based procedure for the estimation of first- and second-order Sobol' indices, J. Statist. Comput. Simulation 85 (2014) 1358–1381, doi:10.1080/00949655.2014.971799.

[9] I. Sobol', On the distribution of points in a cube and the approximate evaluation of integrals, USSR Computational Mathematics and Mathematical Physics 7 (4) (1967) 86–112, doi:10.1016/0041-5553(67)90144-9.

[10] H. Niederreiter, Random Number Generation and Quasi-Monte Carlo Methods, vol. 63, CBMS-NSF Regional Conference Series in Applied Math., siam,philadelphia, pa edn., 1992.

[11] S. Joe, F. Y. Kuo, Constructing Sobol Sequences with Better Two-Dimensional Projections, SIAM J. Sci. Comput. 30 (5) (2008) 26352654, doi:10.1137/070709359.

[12] S. Joe, F. Y. Kuo, Remark on Algorithm 659: Implementing Sobol's quasirandom sequence generator, ACM Trans. Math. Softw. 29 (1) (2003) 49–57, doi:10.1145/641876.641879.

[13] A. B. Owen, Randomly permuted (t,m,s)-nets and (t,s)-sequences, vol. 106, Springer-Verlag, 351–368, 1995.

[14] S. Tezuka, H. Faure, I-binomial scrambling of digital nets and sequences, J. Complexity 19 (6) (2003) 744–757, doi:10.1016/S0885-064X(03)00035-9.