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

$$f(x) = \sum_{u \subseteq \mathcal{D}} f_u(x), \quad f_{\emptyset} = \mu,$$

where,

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

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

ljimene1@hawk.iit.edu
Variance Decomposition

Under the previous definitions,

$$
\sigma^2_{\emptyset} = 0, \quad \sigma^2_u = \int_{[0,1]^d} f_u(x)^2 dx, \quad \sigma^2 = \int_{[0,1]^d} (f(x) - \mu)^2 dx.
$$

The ANOVA identity is,

$$
\sigma^2 = \sum_{u \subseteq D} \sigma^2_u.
$$
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 $u \in \mathcal{D}$:

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

We have the following properties,

- $\tau_u^2 \leq \overline{\tau}_u^2$.
- $\tau_u^2 + \overline{\tau}_u^2 = \sigma^2$. 

ljimene1@hawk.iit.edu
Sobol’ Indices - Probabilistic Framework

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

$$
\bar{\tau}_u^2 = \text{Var} \left[ \mathbb{E} (f(X)|X_u) \right] = 
\text{Var} (f(X)) - \mathbb{E} \left[ \text{Var} (f(X)|X_u) \right],
$$

$$
\tilde{\tau}_u^2 = \text{Var} (f(X)) - \text{Var} \left[ \mathbb{E} (f(X)|X_{-u}) \right] = 
\mathbb{E} \left[ \text{Var} (f(X)|X_{-u}) \right].
$$
The Normalized Sobol’ Indices

One may also use the normalized definition of the Sobol’ indices,\

\[ S_u = \frac{\tau_u^2}{\sigma^2} = \frac{\text{Var} \left[ \mathbb{E} (f(X) | X_u) \right]}{\text{Var} (f(X))} = 1 - \frac{\mathbb{E} \left[ \text{Var} (f(X) | X_u) \right]}{\text{Var} (f(X))}, \]

\[ S_{u_{\text{tot}}} = \frac{\tau_u^2}{\sigma^2} = 1 - \frac{\text{Var} \left[ \mathbb{E} (f(X) | X_{-u}) \right]}{\text{Var} (f(X))} = \frac{\mathbb{E} \left[ \text{Var} (f(X) | X_{-u}) \right]}{\text{Var} (f(X))}. \]

satisfying \( 0 \leq S_u \leq S_{u_{\text{tot}}} \leq 1. \)
The Normalized Sobol’ Indices

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

\[
S_u = \frac{\tau_u^2}{\sigma^2} = \frac{\text{Var} \left[ \mathbb{E} \left( f(X) \mid X_u \right) \right]}{\text{Var} (f(X))} = 1 - \frac{\mathbb{E} \left[ \text{Var} \left( f(X) \mid X_u \right) \right]}{\text{Var} (f(X))},
\]

\[
S_u^{\text{tot}} = \frac{\tau_u^2}{\sigma^2} = 1 - \frac{\text{Var} \left[ \mathbb{E} \left( f(X) \mid X_{-u} \right) \right]}{\text{Var} (f(X))} = \frac{\mathbb{E} \left[ \text{Var} \left( f(X) \mid X_{-u} \right) \right]}{\text{Var} (f(X))}.
\]

satisfying \( 0 \leq S_u \leq S_u^{\text{tot}} \leq 1 \). More specifically, \( S_u \) is composed by,

\[
S_u = 1 - \frac{I^{(1)}}{I^{(2)} - (I^{(3)})^2},
\]

where \( I^{(1)} \) is a \( 2d - |u| \) dim. integral, \( I^{(2)} \) is a \( d \) dim. integral, and \( I^{(3)} \) is a \( d \) dim. integral.

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:

<table>
<thead>
<tr>
<th>Method</th>
<th>Convergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trapezoidal rule:</td>
<td>$O(n^{-2/d})$</td>
</tr>
<tr>
<td>Simpson’s rule:</td>
<td>$O(n^{-4/d})$</td>
</tr>
<tr>
<td>IID Monte Carlo:</td>
<td>$O(n^{-1/2})$</td>
</tr>
<tr>
<td>Quasi-Monte Carlo:</td>
<td>$O(n^{-1+\varepsilon})$</td>
</tr>
</tbody>
</table>

($n$: number of data points)
**Estimating $I^{(1)}$, $I^{(2)}$, and $I^{(3)}$ automatically**

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

$$\left| \int_{[0,1]^d} f(x) \, dx - \hat{I}(x \mapsto f(x), \varepsilon_a) \right| \leq \varepsilon_a,$$

where

$$\hat{I}(x \mapsto f(x), \varepsilon_a) = \frac{1}{2m} \sum_{i=0}^{2^m-1} f(z_i),$$

for some automatic and adaptive choice of $m$ and $\{z_i\}_{i=0}^\infty \in \{\text{Lattice}, \text{Digital}\}$ sequence.
Examples of Sequences

Shifted rank-1 lattice sequence with generating vector $(1, 47)$. Digitally shifted scrambled Sobol’ sequence.
Outline

- Introduction
  - ANOVA
  - Sobol’ Indices
- Quasi-Monte Carlo Methods
  - Replicated Method—Reducing the number of function evaluations to compute first-order indices.
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 \(x, x' \in [0, 1]^d\), we define the following point,

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

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

\[
S_u = 1 - \frac{\int_{[0,1]^{2d-1}} f(x)(f(x) - f(x_u : x'_{-u})) \, dx \, dx'_{-u}}{\int_{[0,1]^d} f(x)^2 \, dx - \left(\int_{[0,1]^d} f(x) \, dx\right)^2} = H(g, g_u).
\]
Number of Function Evaluations

We will focus on reducing the number of function evaluations, and to estimate \( \sigma_u^2/\sigma^2 \), 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 $u$:

$$g(x, x') = f(x_1, \ldots, x_{u-1}, x_u, x_{u+1}, \ldots, x_d),$$

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

Hence, we can construct our points $x'_i$ as follows,

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

Given the right order of points:

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

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

\[
\begin{pmatrix}
f(\mathbf{x}'_0) \\
\vdots \\
f(\mathbf{x}'_n)
\end{pmatrix}
= \begin{pmatrix}
y_0 \\
\vdots \\
y_n
\end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix}
g_u(\mathbf{x}_0, \mathbf{x}'_0) \\
\vdots \\
g_u(\mathbf{x}_n, \mathbf{x}'_n)
\end{pmatrix}
= \begin{pmatrix}
y_{\pi_u^{-1}(0)} \\
\vdots \\
y_{\pi_u^{-1}(n)}
\end{pmatrix}.
\]
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 II


