Variance-based methods

S. KUCHERENKO

Imperial College London (UK)

s.kucherenko@imperial.ac.uk

Outline

High-Dimensional Model Representation (HDMR)

ANOVA decomposition

Sobol' Sensitivity Indices (SI)

Improved formulas for Sobol' Main Effect SI

High-Dimensional Model Representation (HDMR)

Consider a model *f(x)*, *x* is a vector of input independent variables, *f(x)* is integrable. Decomposition of *f(x)* is called HDMR:

$$f(\mathbf{x}) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_{i}\sum_{j>i} f_{ij}(x_i, x_j) + \ldots + f_{1,2,\ldots,n}(x_1, x_2, \ldots, x_n)$$

For simplisity we assume $\mathbf{x} \in H^n = [0,1]^n$

An example in 3 dimensions:

$$\begin{aligned} f(\mathbf{x}) = {} & f_0 + f_1(x_1) + f_2(x_2) + f_3(x_3) \\ & + f_{12}(x_1, x_2) + f_{13}(x_1, x_3) + f_{23}(x_2, x_3) \\ & + f_{123}(x_1, x_2, x_3) \end{aligned}$$

There are infinite ways to build such an expansion →

High-Dimensional Model Representation (HDMR)

An example in 2 dimensions: $f(x_1, x_2) = 4x_1^2 + 3x_2$

a) $f_0 = 0;$ $f_1(x_1) = 4x_1^2$ $f_2(x_2) = 3x_2$ $f_{12}(x_1, x_2) = 0$

b) $f_0 = 5;$

$$f_1(x_1) = 4x_1^2 - 2x_1$$

$$f_2(x_2) = 3x_2 - \sqrt{x_2}$$

$$f_{12}(x_1, x_2) = 2x_1 + \sqrt{x_2} - 5$$

ANOVA decomposition

$$y = f(\mathbf{x}) = f_0 + \sum_{i=1}^{n} f_i\left(x_i\right) + \sum_{i}\sum_{j>i} f_{ij}\left(x_i, x_j\right) + \ldots + f_{1,2,\ldots,n}\left(x_1, x_2, \ldots, x_n\right)$$

Impose a constraint: each term in the HDMR should be

$$\int f_{i_1 i_2 \ldots i_s}(x_{i_1}, x_{i_2}, \ldots, x_{i_s})dx_j = 0 \quad \forall j = i_1, i_2, \ldots, i_s$$

ANOVA decomposition

THEN the HDMR has the following properties:

1) $$f_0 = \int_{H^n} f(\mathbf{x})\, d\mathbf{x}$$

2) Any pair of terms in the HDMR is orthogonal:

$$\int f_{i_1,\ldots,i_s} f_{j_1,\ldots,j_l}\, d\mathbf{x} = 0 \qquad \text{for } (i_1,\ldots,i_s) \neq (j_1,\ldots,j_l)$$

3) HDMR is unique → ANOVA decomposition (Sobol' 1993)

ANOVA decomposition and Sobol' Sensitivity Indices

$$Y = f(X)$$
$$X = (X_1, X_2, ..., X_n) \in H^n$$
$$0 \le X_i \le 1$$

$f(x)$ is L2 integrable

ANOVA decomposition is unique:

$$Y = f(X) = f_0 + \sum_{i=1}^{n} f_i(X_i) + \sum_i \sum_{j>i} f_{ij}\left(X_i, X_j\right) + ... + f_{1,2,...,k}\left(X_1, X_2, ..., X_n\right),$$

$$\int_0^1 f_{i_1 \ldots i_s}(X_{i_1}, ,..., X_{i_s}) dX_{i_p} = 0, \ \forall p, \ 1 \le p \le s, \rightarrow \int_0^1 f_{i_1 \ldots i_s} f_{i_1 \ldots i_l} dX_{i_p} dX_{i_l} = 0, \ \forall i_p \ne i_l$$

Let's square each side and integrate over dx :

$$\int_{H^n} (f(X) - f_0)^2 dx = \int_{H^n} \Big(\sum_{i=1}^{n} f_i(X_i) + \sum_i \sum_{j>i} f_{ij}\left(X_i, X_j\right) + ... + f_{1,2,...,k}\left(X_1, X_2, ..., X_n\right)\Big)^2 dx$$

Total variance:

$$D = \int_{H^n} (f(X) - f_0)^2 dx$$

ANOVA decomposition and Sobol' Sensitivity Indices

Due to the orthogonality of the terms:

$$\int_0^1 f_{i_1 \ldots i_s} f_{i_1 \ldots i_l} dX_{i_p} dX_{i_l} = 0, \quad \forall i_p \neq i_l$$

we obtain variance decomposition :

$$D = \sum_i D_i + \sum_{i,j} D_{ij} + \ldots + D_{1,2,\ldots,n}$$

Partial variances:

$$D_{ij} = \iint f_{ij}^2(x_i, x_j) dx_i dx_j - \left[\iint f_{ij}(x_i, x_j) dx_i dx_j \right]^2 = \iint f_{ij}^2(x_i, x_j) dx_i dx_j$$

Sobol' SI:

$$1 = \sum_{i=1}^{n} S_i + \sum_{i<j} S_{ij} + \sum_{i<j<l} S_{ijl} + \ldots + S_{1,2,\ldots,n}$$

Sobol' Sensitivity Indices (SI)

- *Definition:* $$\boxed{S_{i_1 \ldots i_s} = D_{i_1 \ldots i_s} / D}$$

$$D_{i_1 \ldots i_s} = \int_0^1 f_{i_1 \ldots i_s}^2 \left(x_{i_1}, \ldots, x_{is} \right) dx_{i_1}, \ldots, x_{is}$$ *- partial variances*

$$D = \int_0^1 \left(f\left(x\right) - f_0 \right)^2 dx$$ *- total variance*

- *Sensitivity indices for subsets of variables:* $x = \left(y, z \right)$

$$D_y = \sum_{s=1}^{m} \sum_{\left(i_1 \langle \ldots \langle i_s \right) \in K} D_{i_1, \ldots, i_s}$$

The total variance (Homma&Saltelli 96) : $D_y^{tot} = D_y + D_{yz} = D - D_z$

Corresponding Sobol' sensitivity indices (SI):

$$S_y = D_y / D, \qquad S_y^{tot} = D_y^{tot} / D.$$

How to use Sobol' Sensitivity Indices?

$$\boxed{0 \le S_y \le S_y^{tot} \le 1}$$

- $S_y^{tot} - S_y$ *accounts for all interactions between y and z, x=(y,z).*
- *The important indices in practice are* S_i *and* S_i^{tot}

  $S_i^{tot} = 0 \rightarrow f(x)$ does not depend on x_i ;

  $S_i = 1 \rightarrow f(x)$ depends only on x_i ;

  $S_i = S_i^{tot}$ *corresponds to the absence of interactions between* x_i and other variables

  If $\sum_{s=1}^{n} S_i = 1$, then function has an additive structure: $f(x) = f_0 + \sum_i f_i(x_i)$

- *Fixing unessential variables*

  If $S_z^{tot} << 1 \rightarrow f(x)$ does not depend on z so it can be fixed

  $f(x) \approx f(y, z_0) \rightarrow$ complexity reduction, from k to $k - k_z$ variables

ANOVA decomposition. Finding component functions

$$f(x) = f_0 + \sum_{i=1}^{n} f_i(x_i) + \sum_i \sum_{j>i} f_{ij}(x_i, x_j) + \ldots + f_{1,2,\ldots,n}(x_1, x_2, \ldots, x_n),$$

$$\int_0^1 f_{i_1 \cdots i_s}(x_{i_1}, , \ldots, x_{i_s}) dx_{i_k} = 0, \ \forall k, \ 1 \le k \le s$$

$$\int_0^1 f_v(x_v) f_u(x_u) dx = 0, \ \forall v \ne u.$$

$$f_0 = \int_{H^n} f(x) dx,$$

$$f_i(x_i) = \int_{H^n} f(x) \prod_{j \ne i}^{n} dx_j - f_0,$$

$$f_{ij}(x_i, x_j) = \int_{H^n} f(x) \prod_{k \ne i, j}^{n} dx_k - f_i(x_i) - f_j(x_j) - f_0, \ldots$$

2^n integral evaluations

ANOVA decomposition. Test case

$$f(x_1, x_2) = f_0 + f_1(x_1) + f_2(x_2) + f_{12}(x_1, x_2),$$

$$f(x_1, x_2) = x_1 x_2 \in H^2 \to f_0 = \frac{1}{4},$$

$$f_1(x_1) = \int_{H^n} f(x)\,dx_2 - f_0 = \frac{1}{2}x_1 - \frac{1}{4},$$

$$f_2(x_2) = \int_{H^n} f(x)\,dx_1 - f_0 = \frac{1}{2}x_2 - \frac{1}{4},$$

$$f_{12}(x_1, x_2) = x_1 x_2 - \frac{1}{2}x_1 - \frac{1}{2}x_2 + \frac{1}{4}.$$

$$S_1 = \frac{\int_{H^2} f_1^2(x_1)\,dx_1}{D} = \frac{3}{7},$$

$$S_2 = S_1 = \frac{3}{7},\ S_{12} = \frac{1}{7}.$$

$$S_1^{tot} = S_1 + S_{12} = \frac{4}{7},\ S_2^{tot} = S_2 + S_{12} = \frac{4}{7}$$

Evaluation of Sobol' Sensitivity Indices

Straightforward use of ANOVA decomposition requires

2^n integral evaluations – not practical !

There are efficient formulas for evaluation of Sobol' SI (Sobol' 2001):

$$S_y = \frac{1}{D}\left[\int_0^1 f(y,z')^2\, dydzdz' - f_0^2\right]$$

$$S_y^{tot} = \frac{1}{2D}\int_0^1 [f(y,z) - f(y',z)]^2\, dydzdz'$$

$$D = \int_0^1 f^2(y,z)\, dydz - f_0^2$$

Evaluation is reduced to high-dimensional integration by MC/QMC methods.

Definition of Sobol' SI for subsets

Consider two sets of variables : x=(y,z)

ANOVA decomposition:

$$f(y,z) = f_0 + g_1(y) + g_2(z) + g_{12}(y,z) \qquad (1)$$

$$\int g_1(y)dy = \int g_2(z)dz = \int g_{12}(y,z)dy = \int g_{12}(y,z)dz = 0 \qquad (2)$$

We square and integrate (1) and because of (2)

$$D = D_y + D_z + D_{yz}$$

Define

$$D_y^{tot} = D_y + D_{yz}$$

$$D_z^{tot} = D_z + D_{yz}$$

$$S_y = \frac{D_y}{D}, S_y^{tot} = \frac{D_y^{tot}}{D}$$

Formulas for the main effect Sobol' SI

$$S_y = \frac{1}{D}[\int_0^1 f(y,z')^2 dydz' - f_0^2] =$$

$$= \frac{1}{D}[\int_0^1 f(y,z) f(y,z') dydzdz' - f_0^2].$$

That is

$$D_y = [\int_0^1 f(y,z) f(y,z') dydzdz' - f_0^2]$$

We need to prove that

$$\int_0^1 f(y,z) f(y,z') dydzdz' = f_0^2 + D_y :$$

Recall that $D_y = \int g_1(y)^2 dy$

Derivation of formulas for the main effect Sobol' SI

$$D_y = [\int_0^1 f(y,z) f(y,z')dydzdz' - f_0^2]$$

$$\int_0^1 f(y,z) f(y,z')dydzdz' = \int dy \int f(y,z)dz \int f(y,z')dz'$$

$$= \int dy \left[\int f(y,z)dz \right]^2 = \int dy \left[\int (f_0 + g_1(y) + \cancel{g_2(z)} + \cancel{g_{12}(y,z)})dz \right]^2$$

$$= \int dy \left[\int (f_0 + g_1(y))dz \right]^2 = \int dy [f_0 + g_1(y)]^2$$

$$= f_0^2 + 2 f_0 \int \cancel{g_1(y)dy} + \int g_1(y)^2 dy = f_0^2 + D_y$$

Similarly

$$D_z = [\int_0^1 f(y,z) f(y',z)dydy'dz - f_0^2]$$

Derivation of formula for the total effect Sobol' SI

Jansen's formula (1994), Sobol (2001):

$$S_y^{tot} = \frac{1}{2D}\int_0^1 [f(y,z) - f(y',z)]^2 dydzdy',$$

$$D_y^{tot} = \frac{1}{2}\int_0^1 [f(y,z) - f(y',z)]^2 dydzdy'$$

$$= \frac{1}{2}\int_0^1 [f(y,z)]^2 dydz + \frac{1}{2}\int_0^1 [f(y',z)]^2 dy'dz - \int_0^1 f(y,z)f(y',z)dydy'dz$$

$$= \int_0^1 [f(y,z)]^2 dydz - (D_z + f_0^2) = D - D_z = (D_y + D_z + D_{yz}) - D_z = D_y + D_{yz}$$

Recall

$$D_z = [\int_0^1 f(y,z)f(y',z)dydy'dz - f_0^2],$$

$$\int_0^1 [f(y,z)]^2 dydz - f_0^2 = D$$

Evaluation of Sobol' Main Effect SI with small values

Original Sobol' formula:

$$\underline{x} = (y, z), \quad x' = (y', z')$$

using values $f(y, z), \quad f(y, z'), \quad f(y', z)$

$$S_y = \frac{1}{D} \int_0^1 f(y, z) f(y, z') \, dydzdz' - f_0^2$$

for small indices $S_y << 1$

$$\int_0^1 f(y, z) f(y, z') \, dydzdz' \approx f_0^2$$

$\rightarrow$ loss of accuracy

Improved formula for Sobol' Main Effect SI

Notice that $f_0^2 = \int_0^1 f(y,z)dydz \int_0^1 f(y',z')\ dy'dz'$

using values $f(y,z),\ \ f(y',z),\ \ f(y',z')$

$$S_y = \frac{1}{D}\int_0^1 f(y,z)f(y',z)\ dydy'dz - \ f_0^2 \rightarrow$$

$$S_y = \frac{1}{D}[\int_0^1 f(y',z')[f(y',z) - f(y,z)]\ dydy'dzdz'$$

--gives much more accurate results (Kucherenko, Mauntz, 2002)

Additional advantage (Saltelli 2002):

Requires $N(n+2)$ model evalution rather than

$N(2n+1)$ for original Sobol' formulas.

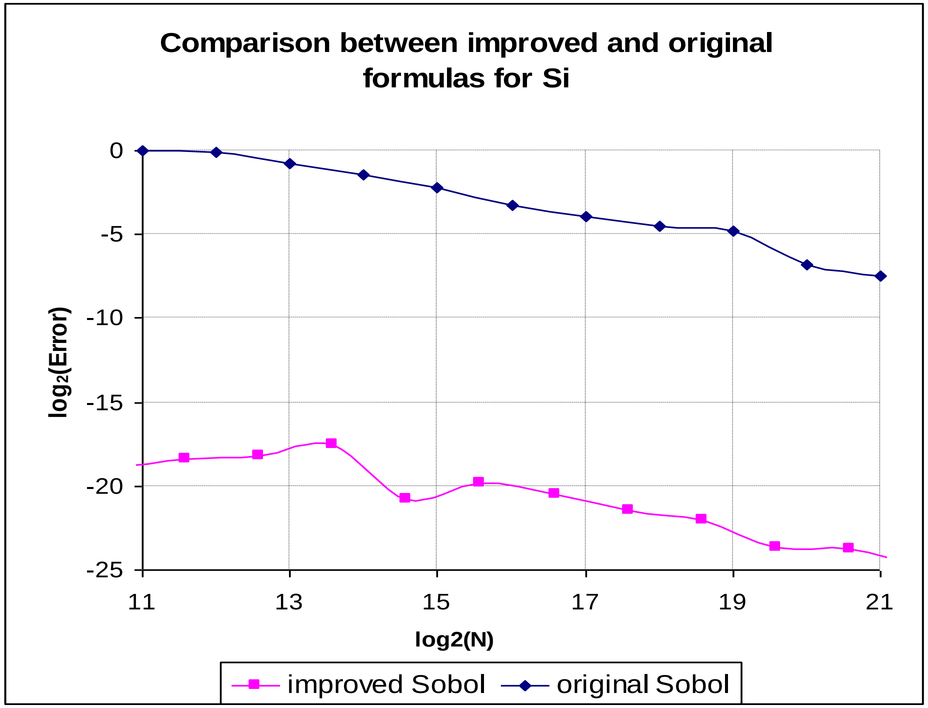
Further improvements: Sobol' and Mishetskaya 2007, A. Owen 2012

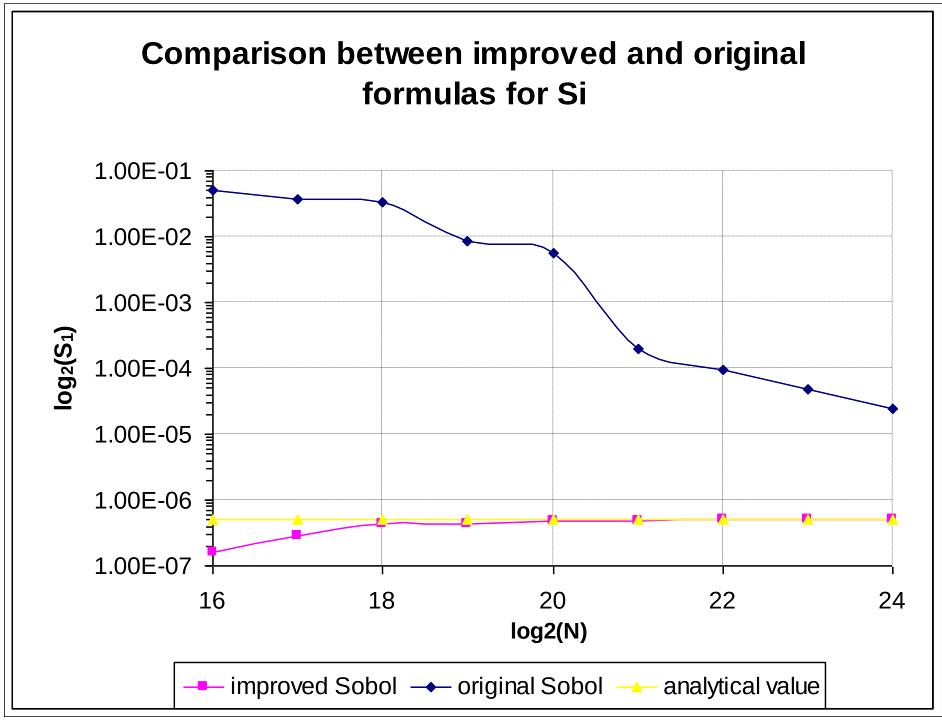
Saltelli 2002 - computation of second order indices at no extra costs.

Improved formula for Sobol' Main Effect SI

$$Test: f(x) = \sum_{i=1}^{n} i x_i\,,\quad S_i = S^T = \frac{6}{n(n+1)(2n+1)}$$

$$n = 180,\quad S_1 = 5.1\,10^{-7}$$





Improved formula have much higher convergence rate than the original Sobol' formula.

Evaluation of Sobol' Sensitivity Indices. Monte Carlo estimates

Main effect SI:

$$S_y = \frac{\frac{1}{N}\sum_{j=1}^{N} f\left(x_j^{'}\right)\left(f\left(y_j^{'}, z_j\right) - f\left(x_j\right)\right)}{D}$$

Total order effect SI:

$$S_y^T = \frac{\frac{1}{N}\sum_{j=1}^{N}\left(f\left(x_j\right) - f\left(y_j^{'}, z_j\right)\right)^2}{2D}$$

Each MC trial reqires three function values for $f(x), f(x'), f(y^{'}, z)$

The total number of function evaluations for a set $(S_i, S_i^T),\ i = 1, \ldots, n$

is equal to $N_F = N(n+2)$.

How to sample ?

Evaluation of Sobol' Sensitivity Indices. MC and QMC Sampling

To sample x and x' (they are vector points in H^n):

A. For Monte Carlo sample 2n random numbers

$$\xi_j = (\gamma_1^j, \gamma_2^j, ..., \gamma_n^j), \xi_j^{'} = (\gamma_{n+1}^j, \gamma_{n+2}^j, ..., \gamma_{2n}^j), \; j = 1, 2, ..., N$$

B. For Quasi Monte Carlo sample one

2n-dimensional quasi random number

$Q_j = (q_1^j, q_2^j, ..., q_{2n}^j)$ and split it into two points

$$\xi_j = (q_1^j, q_2^j, ..., q_n^j), \xi_j^{'} = (q_{n+1}^j, q_{n+2}^j, ..., q_{2n}^j), \; j = 1, 2, ..., N$$

How to use Sobol' sequence generators. MATLAB version

Specification (www.broda.co.uk) :

Successive calls to the function

SobolSeq(i,n)

generates an n- dimensional vector containing the Cartesian coordinates of the i-th point of the Sobol' sequence in the n- dimensional unit cube $[0,1]^n$.

Input parameters:

i - index of a point (i=[0,2**31-1]),

n - dimension of the Sobol' sequence;

Syntax:

r = SobolSeq(i,n)

How to use Sobol' sequence generators

To sample n independent inputs $x = (x_1, x_n, ..., x_n)$ in H^n

A. Monte Carlo: sample n random numbers $(\gamma_1^j, \gamma_2^j, ..., \gamma_n^j) = \xi_j, \ j = 1, 2, ..., N$

B. Quasi Monte Carlo: sample one n-dimensional quasi random vector

$\xi_j = (q_1^j, q_2^j, ..., q_n^j);$

to sample another vector - increase index $j \to j+1$.

Sets $\{q_k^j\}, \{q_p^j\}, j = 1, 2, ..., N, k \neq p$ (different dimensions) are independent;

Vectors $\xi_j = (q_1^j, q_2^j, ..., q_n^j), \xi_{j+1} = (q_1^{j+1}, q_2^{j+1}, ..., q_n^{j+1}),.. \ \ j = 1, 2, ..., N$ are dependent

INDEX	x1	X2	X3
1	0.5	0.5	0.5
2	0.25	0.75	0.25
3	0.75	0.25	0.75
4	0.125	0.625	0.875
5	0.625	0.125	0.375
6	0.375	0.375	0.625
7	0.875	0.875	0.125
8	0.0625	0.9375	0.6875

Different formulas for the main effect index

	D_y	Monte Carlo estimator
Sobol'	$\int f(x) f(y,z') dx dz' - f_0^2$	$\frac{1}{N}\sum_{k=1}^{N} f(y,z) f(y,z') - \left[\frac{1}{N}\sum_{k=1}^{N} f(y,z)\right]^2$
Kucherenko 2002	$\int f(x)[f(y,z') - f(x')] dx dx'$	$\frac{1}{N}\sum_{k=1}^{N} f(y,z)[f(y,z') - f(y',z')]$
Owen 2012	$\int [f(x) - f(y'',z)][f(y,z') - f(x')] dx dx' dx''$	$\frac{1}{N}\sum_{k=1}^{N} [f(y,z) - f(y'',z)][f(y,z') - f(y',z')]$
Sobol-Myshetzskay (Oracle) 2007	$\int [f(x) - \mu][f(y,z') - f(x')] dx dx' dx''$	$\frac{1}{N}\sum_{k=1}^{N} [f(y,z) - \mu][f(y,z') - f(y',z')]$

Comparison of computational costs

Method	Sobol'	S-K	Owen	Oracle
Number of function evaluations N_{CPU}	$N(2n+1)$	$N(n+2)$	$N(2n+2)$	$N(n+2)$

The root mean square error (RMSE) is determined using K independent runs

$$\boxed{\varepsilon_i\,(N) = \left[\frac{1}{K}\sum_{k=1}^{K}(S_i^{(n),k} - S_i^{(a)})^2\right]^{1/2}}$$

For QMC, the convergence rate is $\varepsilon_{QMC} = \dfrac{O(\ln N)^d}{N}$

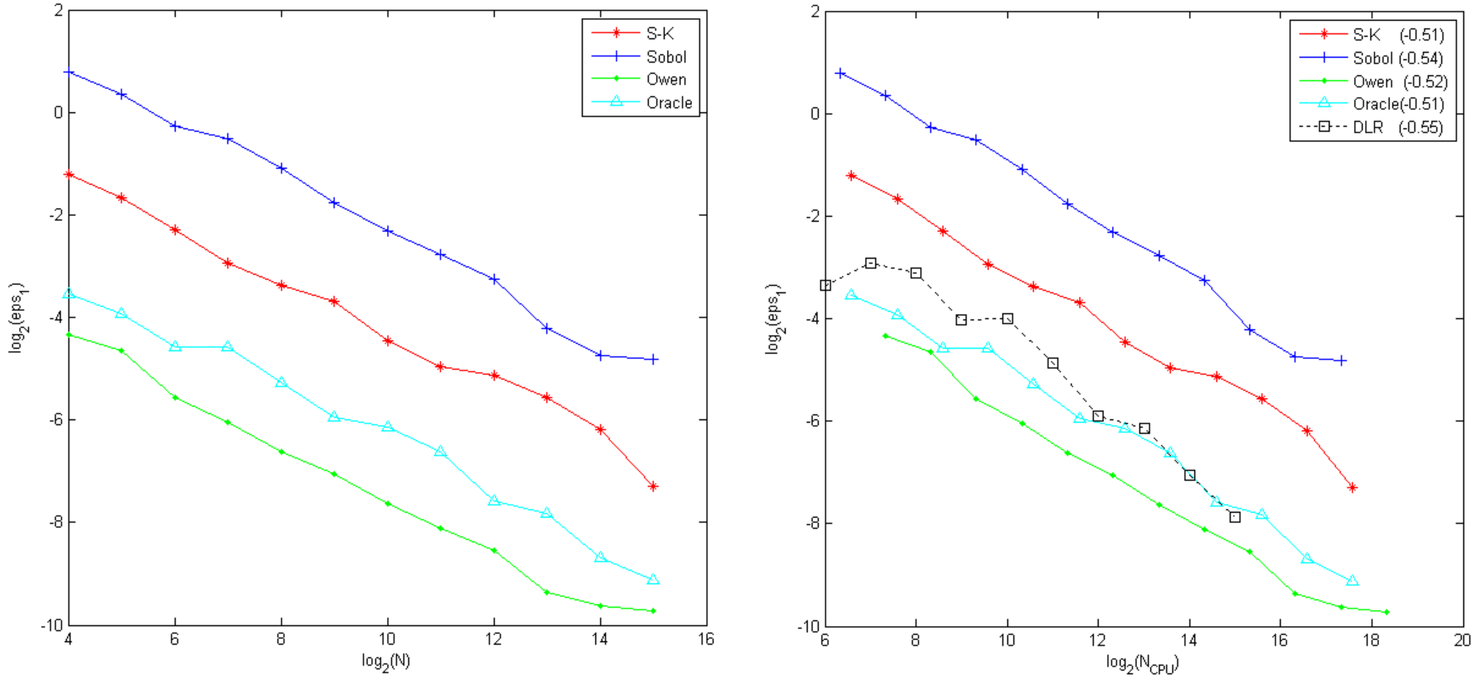
In practice, RMSE is approximated by $\boxed{cN^{-\alpha},\ 0<\alpha<1}$

Convergence significantly improves when using QMC (Sobol' sequences) sampling.

Comparison of different formulas

$$f(x) = a_1 x_1 + a_2 x_2 + \ldots + a_n x_n, \quad x_i \sim N(\mu_i, \sigma_i^2)$$

$$n = 4, \quad \mu = (1, 3, 5, 7), \quad \sigma = (1, 1.5, 2, 2.5), \quad a_i = 1, \ i = 1,2,3,4$$



Log2(RMSE) versus Lof2(N) for i=1, S1= 0.0741

Improved formulas have much higher convergence rate than the original Sobol' formula. Owen and Oracle - outperforming other methods

References

Saltelli A., P. Annoni ,I. Azzini, F. Campolongo, M. Ratto and S. Tarantola (2010) Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index, *Computer Physics Communications* 181, 259–270

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D. Saisana, M., and Tarantola, S., 2008, Global Sensitivity Analysis. The Primer, John Wiley & Sons.

Homma, T. and A. Saltelli (1996). Importance measures in global sensitivity analysis of nonlinear models. Reliability Engineering and System Safety, 52, 1–17.

Sobol' (2001) Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, *Mathematics and Computers in Simulation,* 55, 271–280

Kucherenko S., Feil B., Shah N., Mauntz W. (2011) The identification of model effective dimensions using global sensitivity analysis Reliability Engineering and System Safety 96, 440–449

Kucherenko S, Song S (2017) Different numerical estimators for main effect global sensitivity indices, Reliability Engineering & System Safety, 165, 222–238