Design and analysis of computer experiments

Bertrand Iooss

29/11/2012
Uncertainty management - The generic methodology

Step A: Problem specification

- Input data
  - Variables: \( x \)
  - Parameters: \( \theta \)
  - Fixed: \( d \)
- Numerical model
  - \( G(x, \theta, d) \)
- Real system
  - \( G^R(x, d) \)

Step B: Quantification of uncertainty sources

- Probabilistic modelling (joint distribution)
- Direct methods, statistics, experts

Step B': Calibration / Validation

- Observed variables
  - \( Z_{obs}(x, d) = Z^R + \epsilon(x) \)

Step C: Propagation of uncertainty sources

- Variables of interest
  - \( Z = G(x, \theta, d) \)
  - \( Z^R = G^R(x, d) \)
  - \( Z = G(x, \theta, d) + b(x, d) \)

Step C': Sensitivity analysis

- Feedback process
- Decision criterion
  - Ex. Probability < 10^-a
Main objectives of sensitivity analysis

• Reduction of the uncertainty of the model outputs by prioritization of the sources
  • Variables to be fixed in order to obtain the largest reduction (or a fixed reduction) of the output uncertainty
    A purely mathematical variable ordering
  • Most influent variables in a given output domain
    if reducibles, then R&D prioritization
    else, modification of the system

• Simplification of a model
  • determination of the non-influent variables, that can be fixed without consequences on the output uncertainty
  • building a simplified model, a metamodel
Outline

1. Design of numerical experiments – Space filling designs

2. Analysis of numerical experiments
Typical engineering practice: One-At-a-Time (OAT) design

Main remarks:

OAT brings some information, but potentially wrong

Exploration is poor: Non monotonicity? Discontinuity? Interaction?

Leave large unexplored zones of the domain (curse of dimensionality)
Model exploration goal

**GOAL**: explore as best as possible the behaviour of the code
Put some points in the whole input space in order to « maximize » the amount of information on the model output

Contrary to an uncertainty propagation step, it depends on $p$

Regular mesh with $n$ levels  \[ N = n^p \] simulations

Ex: $p = 2$, $n = 3$
\[ N = 9 \]

$N = 59049$

To minimize $N$, needs to have some techniques ensuring good « coverage » of the input space

Simple random sampling (Monte Carlo) does not ensure this

Ex: $p = 2$, $N = 10$

Monte Carlo

Optimized design
Objectives

When the objectives is to discover what happens inside the model and when no model computations have been realized, we want to respect the two following constraints:

• To spread the points over the input space in order to capture non linearities of the model output,

• To ensure that this input space coverage is robust with respect to dimension reduction.

Therefore, we look some design which insures the « best coverage » of the input space.

Main question:
• How to define this « best »?
Exploration in physical experimentation

Design of experiments develops strategies to define experiments in order to obtain the required information as efficiently as possible.

**Designs for real experiments**

Estimate parameters of linear regression with a minimal number of points

*Examples:*

- Full factorial design $2^3$
- Fractional factorial design $2^{3-1}$

**Designs for numerical experiments**

*Characteristics*

- Deterministic experiments (no error),
- Large number of input variables,
- Large range of input variation domain,
- Multiple output variables,
- Strong interactions between inputs,
- High non linearity in the model

- Space filling designs (uniform coverage in the input space)
Space filling designs

Sparsity of the space of the input variables in high dimension

The learning design choice is made in order to have an optimal coverage of the input domain

The space filling designs are good candidates.

Example: Sobol sequence

Two possible criteria:
1. Distance criteria between the points: minimax, maximin, ...
2. Uniformity criteria of the design (discrepancy measures)
**Geometrical criteria (1/2)**

- **Minimax design** $D_{MI}$: Minimize the maximal distance between one point of the domain and one point of the design

\[
\min_D \max_x d(x, D) = \max_x d(x, D_{MI})
\]

where \(d(x, D) = \min_{x^{(0)} \in D} d(x, x^{(0)})\)

All points in \([0,1]^p\) are not too far from a design point

\[
\Rightarrow \text{One of the best design, but too expensive to find } D_{MI}
\]

[Johnson et al. 1990]  
[Koehler & Owen 1996]
Minimax design

- $p = 1 \ ; \ X_i = (2i-1)/(2N) \ ; \ \phi_{mm} = 1 / 2N$

- $p > 1$ : sphere recovering
- **Mindist distance**: \( \phi(\Xi^N) = \min_{x^{(1)}, x^{(2)} \in \Xi^N} d(x^{(1)}, x^{(2)}) \) (\( L_2 \) norm for example)

Maximin design \( \Xi^N_{Mm} \):

maximize minimal distance between two points of the design

\[
\max_{\Xi^N} \min_{x^{(1)}, x^{(2)} \in \Xi^N} d(x^{(1)}, x^{(2)}) = \min_{x^{(1)}, x^{(2)} \in \Xi^N_{Mm}} d(x^{(1)}, x^{(2)})
\]
Maximin design

- \( p = 1 \); \( X_i = (i-1)/(N-1) \); \( \phi_{mM} = 1 / (N-1) \)

- \( p > 1 \): sphere packing

\( L^2 \)-LHD of 9 points

\( d = 0.395284707521 \) and \( D^2 = 10 \)

[ www.spacefillingdesigns.nl ]
[ www.packomania.com ]
Space filling measure of a design: the discrepancy

Measure of the maximal deviation between the distribution of the sample’s points to an uniform distribution

⇒ Measure of deviation from the uniformity

Geometrical interpretation:
Comparison between the volume of intervals and the number points within these intervals

\[ Q(t) \in [0,1]^p, Q(t) = [0,t_1] \times [0,t_2] \times \ldots \times [0,t_p] \]

\[
\text{disc}(D) = \sup_{Q(t) \in [0,1]^p} \left| \frac{N_{Q(t)}}{N} - \prod_{i=1}^{p} t_i \right|
\]

Lower the discrepancy is, the more the points of the design D fill the all space
Link with the integration problem

\[ I = \int_{[0,1]^p} f(x) \, dx \]

Monte Carlo: \[ I_N^{MC} = \frac{1}{N} \sum_{i=1}^{N} f(x^{(i)}) \]

with \( (x^{(i)})_{i=1...N} \) a sequence of random points in \([0,1]^p\)

\[ \mathbb{E}(I_N^{MC}) = I ; \quad \text{Var}(I_N^{MC}) = \frac{\text{Var}(N)}{N} \Rightarrow \varepsilon = O\left(\frac{1}{\sqrt{N}}\right) \]

General property (Koksma-Hlawka inequality): \( \varepsilon \leq V(f) \times \text{disc}(D) \)

With a low discrepancy sequence \( D \) (quasi Monte Carlo sequence):

Well-known choice: Sobol’ sequence

\[ \varepsilon = O\left(\frac{(\ln N)^p}{N}\right) \]
**L₂ discrepancy**

Several definitions, depending on considered norms and intervals

\[
D^*_2(\Xi^N) = \sup_{t \in [0,1]^p} \left| \frac{1}{N} \sum_{i=1}^{N} x^{(i)} \in Q(t) - \text{Volume}(Q(t)) \right|
\]

Choice allowing computations: \(L_2\) discrepancy

\[
L_2\text{ discrepancy at origin}: \quad D^*_2(\Xi^N) = \left[ \int_{[0,1]^p} \left[ \frac{1}{N} \sum_{i=1}^{N} x^{(i)} \in Q(t) - \text{Volume}(Q(t)) \right]^2 \ dt \right]^{1/2}
\]

Missing property: taking into account uniformity of the point projections

On lower-dimensional subspaces of \([0,1]^p\)

\(\Rightarrow\) Modified \(L_2\) discrepancies

\[
D_2(\Xi^N) = \left[ \sum_{u \neq \emptyset} \int_{C^u} \left[ \frac{1}{N} \sum_{i=1}^{N} x^{(i)} \in Q_u(t) - \text{Volume}(Q_u(t)) \right]^2 \ dt \right]
\]

with \(u \subset \{1, \ldots, p\}\)

and \(Q_u(t) = \text{projection of } Q(t) \text{ on } C^u\) (unit cube of coordinates in \(u\))
Discrepancy computation in practice

- **Modified \( L_2 \)-discrepancy** (intervals with minimal boundary 0)

- **Centered \( L_2 \)-discrepancy** (intervals with boundary one vertex of the unit cube)

\[
\operatorname{disc}_2(D) = \left( \frac{13}{12} \right)^p - \frac{2}{N} \sum_{i=1}^{N} \prod_{k=1}^{p} \left( 1 + \frac{1}{2} \left| x_k^{(i)} - \frac{1}{2} \right| - \frac{1}{2} \left| x_k^{(i)} - \frac{1}{2} \right|^2 \right)
\]

\[
+ \frac{1}{N^2} \sum_{i,j=1}^{N} \prod_{k=1}^{p} \left( 1 + \frac{1}{2} \left| x_k^{(i)} - \frac{1}{2} \right| + \frac{1}{2} \left| x_k^{(j)} - \frac{1}{2} \right| - \frac{1}{2} \left| x_k^{(i)} - x_k^{(j)} \right| \right)
\]

- **Symetric \( L_2 \)-discrepancy** (intervals with boundary one « even » vertex of the unit cube)
Sobol’s sequence vs. Random sample vs. regular grid

[From: Kucherenko, 2010]
Example - N = 150 - Dimension = 8

Sobol

Sobol scrambling Owen

Venez’ course 2012 – Design and analysis - B. Iooss
Example - N = 150 - Dimension = 8

Halton
Pathologies on 2D projections

Halton
Important property: robustness in terms of subprojections

Most of the times, the function $f(X)$ has low effective dimensions:
- in the truncation sense ($p_1 = \text{number of influent inputs}$) $\Rightarrow p_1 \ll p$
- in the superposition sense ($p_2 = \text{higher order of influent interaction}$) $\Rightarrow p_2 \ll p$

Then, we need SFD which keeps their space-filling properties in low-dimensional subspaces (by importance: in dimensions $p' = 1$, then $p' = 2$, ...)

- $p' = 1 \Rightarrow \text{LHS ensures good 1D projection properties}$

- $p' \geq 2$
  In their definition, the modified $L^2$-discrepancy criteria take into account subprojections

In contrary design points distance criteria are not robust at all
Latin Hypercube Sample (LHS)

Most often, only a small number of variables are influent

**Property:** Uniform projections on margins

**Principle:** $p$ variables, $N$ points $\Rightarrow LHS(p,N)$

Divide each dimension in $N$ intervals
Take one point in each stratum

Each level is taken only one time by each variable
$\Rightarrow$ Each column of the design is a permutation of $\{1,2,\ldots,N\}$

Exemple : $p=2$, $N=4$
Algorithm of LHS($p,N$) – Stein method

ran = matrix(runif(N*p), nrow=N, ncol=p)  # tirage de N x p valeurs selon loi U[0,1]
x = matrix(0, nrow=N, ncol=p)  # construction de la matrice x

for (i in 1:p) {
    idx = sample(1:N)  # vecteur de permutations des entiers {1,2,...,N}
    P = (idx-ran[,i]) / N  # vecteur de probabilités
    x[,i] <- quantile_selon_la_loi (P)  
}

Example:  $p=2$, $N=10$, $X_1 \sim U[0,1]$, $X_2 \sim N(0,1)$

(a) Simple Random Sampling

(b) Latin Hypercube Sampling
Optimisation of LHS => Space-filling LHS

Simple method: produce a large number (for ex 1000) of different LHS. Then, choose the best with respect to a criterion $\phi(.)$ (« space filling »)

Example : LHS(2,16)

Maximin criterion

But: the number of LHS is huge: $(N!)^p$

Methods via optimization algo (ex: minimisation of $\phi(.)$ via simulated annealing):

1. Initialisation of a design $\Xi$ (LHS initial) and a temperature $T$

2. While $T > 0$ :
   1. Produce a neighbor $\Xi_{\text{new}}$ of $\Xi$ (permutation of 2 components in a column)
   2. replace $\Xi$ by $\Xi_{\text{new}}$ with proba $\min\left(\exp\left(-\frac{\phi(\Xi_{\text{new}}) - \phi(\Xi)}{T}\right), 1\right)$
   3. decrease $T$

3. Stop criterion $\Rightarrow \Xi$ is the optimal solution
Examples of optimized LHS
Joining the two properties (space filling and LHS)

Example: $p = 2 - N = 16$

Maximin LHS

Low wrap-around discrepancy LHS

For comparison: Sobol sequence
Summary on the design of numerical experiments

**Goal:** Sample a high dimensional space in an « optimal » manner (obtain the maximum of information on the behaviour of the output $Z / X \in \mathbb{R}^p$)

**Problem:** a pure random sample (Monte Carlo) badly fills the space

1. « Space filling » designs are good candidates:

- Based on a distance criterion between points (minimax, maximin, ...)

- Based on a criterion of uniform distribution of the points (*discrepancy*)

2. Property of uniform projections on margins can be obtained via the *Latin hypercube designs* (LHS)

3. It is possible to couple 1 and 2
Outline

1. Design of numerical experiments

2. Analysis of numerical experiments - Sensitivity analysis
Sensitivity analysis notions

- **Sensitivity**, for example \( \partial Y / \partial X_i \)

  Donne une idée de la manière dont peut répondre la réponse en fonction de variations potențielles des facteurs

- **Contribution** = sensitivity x importance, for example \( \frac{\partial Y}{\partial X_i} \sigma(X_i) \)

  Permet de déterminer le poids d’une variable d’entrée (ou groupe de variables) sur l’incertitude de la variable d’intérêt (la sortie)
Main objectives of sensitivity analysis

• Reduction of the uncertainty of the model outputs by prioritization of the sources
  • Variables to be fixed in order to obtain the largest reduction (or a fixed reduction) of the output uncertainty
    A purely mathematical variable ordering
  • Most influent variables in a given output domain
    → if reducibles, then R&D prioritization
    → else, modification of the system

• Simplification of a model
  • determination of the non-influential variables, that can be fixed without consequences on the output uncertainty
  • building a simplified model, a metamodel
Overall classification of sensitivity analysis methods
(quantity of interest = variability of the output)

Three types of answers:

1. **Screening**:
   - classical design of experiments,
   - numerical design of experiments (*Morris*, sequential bifurcation)

2. **Quantitative measures of global influence**:
   - correlation/regression on values/ranks
   - statistical tests,
   - functional variance decomposition (Sobol),
   - other measures: entropy, distribution distances

3. **Deep exploration of sensitivities**
   - smoothing techniques (param./non parametric)
   - metamodels
Screening with $n < p$ (supersaturated designs)

Many inputs ($p >> 10$) and cpu time costly computer code

**Objective:** less computations than number of inputs

**Hypotheses:**
- Number of influent inputs $\ll$ total number of inputs
- Monotony of the model, no interaction between inputs
- Knowledge of the direction of the output variation / each input

**Example:** method of sequential bifurcations

2 runs

$X_1 \quad X_2 \quad X_3 \quad X_4 \ldots \quad X_p$  

$y^+ \quad y^-$
Screening with \( n < p \) (supersaturated designs)

Many inputs \((p \gg 10)\) and cpu time costly computer code

**Objective:** less computations than number of inputs

**Hypotheses:**
- Number of influent inputs \(< \) total number of inputs
- Monotony of the model, no interaction between inputs
- Knowledge of the direction of the output variation / each input

**Example: method of sequential bifurcations**
Screening with \( n < p \) (supersaturated designs)

Many inputs \((p >> 10)\) and cpu time costly computer code

**Objective:** less computations than number of inputs

**Hypotheses:**
- Number of influential inputs \(<\) total number of inputs
- Monotony of the model, no interaction between inputs
- Knowledge of the direction of the output variation / each input

**Example:** method of sequential bifurcations

Number of runs: \( n \sim p / 2 \)
Screening without hypothesis on function: Morris’ method

- Discretization of input space
- Needs p+1 experiments
- OAT (One-at-A-Time)

\[ d_{X_1} = \frac{f(P_3) - f(P_2)}{\Delta} \]

\[ d_{X_2} = \frac{f(P_2) - f(P_1)}{\Delta} \]
Morris’ method

- OAT design is repeated $R$ times (total: $n = R \times (p+1)$ experiments)
- It gives an $R$-sample for each elementary effect

\[
\begin{align*}
\{d^i_{X1}\}_{i=1\ldots R} \\
\{d^i_{X2}\}_{i=1\ldots R}
\end{align*}
\]

- Sensitivity measures:
  \[
  \mu^*_i = \mathbb{E}(|d_{X_i}|) \\
  \sigma_i = \sigma(d_{X_i})
  \]
Morris: sensitivity measures

- $\mu_i^* = \mathbb{E}(|dX_i|)$ is a measure of the sensitivity:
  - Important value $\rightarrow$ important effects (in mean)
  - $\rightarrow$ sensitive model to input variations

- $\sigma_i = \sigma(dX_i)$ is a measure of the interactions and of the non linear effects:
  - Important value $\rightarrow$ different effects in the R-sample
  - $\rightarrow$ effects which depend on the value:
    - of the input $X_i$ $\Rightarrow$ non linear effect
    - or of the other inputs $\Rightarrow$ interaction
  (the distinction between the two cases is impossible)
Morris : example

20 factors
210 simulations
→ Graph (μ*, σ)

Distinction between 3 groups:
1. Negligible effects
2. Linear effects
3. Non linear effects and/or with interactions

Cas test : non monotonic function of Morris
Example: fuel irradiation computation in HTR

Computer code ATLAS (CEA): simulation of the HTR fuel (fuel particles) behaviour under irradiation

Noyau de matière fissile
Carbone pyrolytique poreux
Carbone pyrolytique dense
Carbure de Silicium

Contamination sources: failure of particles

Reliability studies are needed

The failure of a particle can be caused by the failure of the external thick layers (IPyC, SiC, OPyC)

Output variables are representative of failure phenomena: maximal orthoradial strains in external layers

Number of particles inside a reactor: $10^9$ to $10^{10}$
3 uncertainty types for the inputs

• 10 parameters of fuel particle manufacturing process (thickness, ...)
  Specifications → truncated Gaussian distributions

• 5 parameters of irradiation (temperature, ...)
  Interval [min,max] → uniform distributions

• 28 behaviour laws (functions of temperature, flux, ...)
  Expert judgment → multiplicative constants (~ U[0.95,1.05])

Example: Law of PyC densification
**Results of Morris**

\[ p = 43 \text{ inputs}, \ 20 \text{ repetitions}, \ n = 860 \text{ runs}, \ \text{unitary cost} \sim 1 \text{ mn} \Rightarrow \text{total}=14h \]

Large sensitivities to these inputs (thickness, irradiation temperature)
Small interaction effects

Influence of creep and densification laws of PyC

**Conclusion:**

**Morris method provides qualitative information about output variations due to potential variations of inputs**

Useful in order to identify potential influential inputs
Overall classification of sensitivity analysis methods
(quantity of interest = variability of the output)

Three types of answers:

1. **Screening**:
   - classical design of experiments,
   - numerical design of experiments (Morris, sequential bifurcation)

2. **Quantitative measures of global influence**:
   - correlation/regression on values/ranks
   - statistical tests,
   - functional variance decomposition (Sobol),
   - other measures: entropy, distribution distances

3. **Deep exploration of sensitivities**
   - smoothing techniques (param./non parametric)
   - metamodels
Sensitivity analysis for one scalar output

Sample \((X \in \mathbb{R}^p, Y(X) \in \mathbb{R})\) of size \(N > p\)

Preliminary step: graphical visualization (for ex: scatterplots)

**Remark:** it can be a Monte Carlo sample, a quasi-Monte Carlo sample or any other designs
Graphical representation: scatterplots

Measure the linear character of the cloud

\[ N \text{ runs} \]

Graphs Output with respect to each input

\[ \rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} \]

\[ \hat{\rho} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y}) \]

Example:
\[ N = 300 \]

Venez' course 2012 – Design and analysis - B. Iooss
Flood model - Scatterplots – Output S

$Q = \text{river flowrate} \sim \text{Gumbel on } [500,3000]$  
$Ks = \text{friction coefficient} \sim \text{normal on } [15,50]$  
$Zv = \text{downstream river bed height} \sim \text{triangular on } [49,51]$  
$Hd = \text{dyke height} \sim \text{triangular on } [7,9]$  
$Cb = \text{bank height} \sim \text{triangular on } [55,56]$  

\[ S = Z_v + H - H_d - C_b \text{ avec } H = \left( \frac{Q}{BK_s} \sqrt{\frac{Z_m - Z_v}{L}} \right)^{0.6} \]

Monte Carlo sample - $N = 100$
Flood model - Scatterplots – Output Cp

\[ C_p = \mathbb{I}_{S>0} + \left\{ 0.2 + 0.8 \left[ 1 - \exp \left( -\frac{1000}{S^4} \right) \right] \right\} \mathbb{I}_{S\leq 0} + \frac{1}{20} \left( H_d \mathbb{I}_{H_d>8} + 8 \mathbb{I}_{8\leq H_d} \right), \]

Monte Carlo sample - \( N = 100 \)

**Major drawback:** only first order relations between inputs are analyzed and not their interactions (=> needs of other data analysis tools)
Sensitivity analysis for one scalar output

Sample \(( \mathbf{X} \in \mathbb{R}^p, Y(\mathbf{X}) \in \mathbb{R} )\) of size \(N > p\)

Preliminary step: graphical visualization (for ex: scatterplots)

Quantitative sensitivity analysis methodology

[Saltelli et al. 00, Helton et al. 06 ]

? Linear relation ?

\((R^2)\)

Oui

Non

? Monotonic relation ?

\((R^2*)\)

Sensitivity indices

Linear regression between \(X_i\) and \(Y\)

Regression coefficients

Regression on ranks

Sobol' indices

Venez' course 2012 – Design and analysis - B. Iooss
Sensitivity indices in case of linear inputs/output relation

Independent input variables $X = (X_1, \ldots, X_p)$

Sample: $n$ realizations of $(X, Y)$

$$Y = \beta_0 + \sum_{i=1}^p \beta_i X_i$$

- SRC index: $SRC(X_i) = \beta_i \sqrt{\frac{\text{Var}(X_i)}{\text{Var}(Y)}}$

Sign of $\beta_i$ gives the direction of variation of $Y$ in fct of $X_i$

- SRC is similar to the linear correlation coefficient (Pearson)

- Validity of the linear model via
  The residuals diagnostics and $R^2$:
  $$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

- We have $R^2 = \sum_{i=1}^p SRC^2(X_i)$ => nice interpretation of SRC
Flood model - Output S

Monte Carlo sample - $N = 100$

<table>
<thead>
<tr>
<th>Q</th>
<th>Ks</th>
<th>Zv</th>
<th>Hd</th>
<th>Cb</th>
</tr>
</thead>
<tbody>
<tr>
<td>31%</td>
<td>15%</td>
<td>18%</td>
<td>29%</td>
<td>6%</td>
</tr>
</tbody>
</table>

Sensitivity indices ($SRC^2$)

The model is linear ($R^2 = 0.99$)

$SRC$ coefficients are sufficient for the quantitative sensitivity analysis
Sensitivity analysis for one scalar output

Sample \(( X \in \mathbb{R}^p, Y(X) \in \mathbb{R} )\) of size \(N > p\)

Preliminary step: graphical visualization (for ex: scatterplots)

Quantitative sensitivity analysis methodology

\[ \text{Saltelli et al. 00, Helton et al. 06} \]

- ? Linear relation ?
  - Oui
  - Non
    - Linear regression between \(X_i\) and \(Y\)
    - Regression coefficients

- ? Monotonic relation ?
  - Oui
  - Non
    - Regression on ranks

Sensitivity indices

\[ S_i = \frac{\text{Var}[E(Y|X_i)]}{\text{Var}(Y)} \]
Functional decomposition

\[ y = f(x) = f_0 + \sum_{i=1}^{p} f_i(x_i) + \sum_{i} \sum_{j>i} f_{ij}(x_i, x_j) + \ldots + f_{1,2,\ldots,p}(x_1, x_2, \ldots, x_p) \]

with \( f(x) \in L^2(x) \ \ x \in [0;1]^p \)

Infinity of possible decompositions

BUT, unicity of decomposition if: \( \int f_{i_1\ldots i_s}(x_{i_1}, \ldots, x_{i_s})dx_j = 0 \ \forall j = i_1, \ldots, i_s \)

Properties ( \( x_i \sim U[0,1] \) for \( i=1,\ldots,p \), the \( x_i \)s are independent)

\[ f_0 = \int f(x) \, dx = E(y) \]
\[ f_i(x_i) = \int f(x) \, dx_{-i} - f_0 = E(y | x_i) - f_0 \]
\[ f_{ij}(x_i, x_j) = E(y | x_i, x_j) - E(y | x_i) - E(y | x_j) + f_0 \]

Example: \( f(x_1, x_2) = x_1 + x_2 ; x_1 \sim U[0;1] ; x_2 \sim U[0;1] \)

\[ f_0 = 1 ; f_1(x_1) = x_1 - \frac{1}{2} ; f_2(x_2) = x_2 - \frac{1}{2} ; f_{12}(x_1, x_2) = 0 \]
Another example

\[ f(x_1, x_2) = 4x_1^2 + 3x_2 \]

\[ x_1, x_2 \in U[-1/2; 1/2] \]

\[ f_0 = 0 \]

\[ f_1(x_1) = 4x_1 \]

\[ f_2(x_2) = 3x_2 \]

\[ f_{12}(x_1, x_2) = 0 \]

\[ f_0 = E(y) = \int_{-1/2}^{1/2} \int_{-1/2}^{1/2} (4x_1^2 + 3x_2) \, dx_1 \, dx_2 = \frac{1}{3} \]

\[ f_1(x_1) = E(y \mid x_1) - f_0 = \int_{-1/2}^{1/2} (4x_1^2 + 3x_2) \, dx_2 = 4x_1^2 - \frac{1}{3} \]

\[ f_2(x_2) = E(y \mid x_2) - f_0 = 3x_2 \]

\[ f_{12}(x_1, x_2) = 0 \]
Sensitivity indices without model hypotheses

Functional ANOVA [Efron & Stein 81] (hyp. of independent $X_i$ s):

$$\text{Var} (Y) = \sum_{i=1}^{p} V_i (Y) + \sum_{i<j}^{p} V_{ij} (Y) + \cdots + V_{12...p} (Y)$$

where $V_i (Y) = \text{Var}[E(Y|X_i)]$

$V_{ij} = \text{Var}[E(Y|X_i,X_j)] - V_i - V_j , ...$

Sobol indices definition:

• First order sensitivity indices: $S_i = \frac{V_i}{\text{Var}(Y)}$

• Second order sensitivity indices: $S_{ij} = \frac{V_{ij}}{\text{Var}(Y)}$

• ...
Another example

\[ y = f(x_1, x_2) = 4x_1^2 + 3x_2 \quad \text{for} \quad x_1, x_2 \in U[-1/2, 1/2] \]

On a vu :

\[ f_0 = E(y) = \frac{1}{3} \]

\[ f_1(x_1) = E(y \mid x_1) - f_0 = 4x_1^2 - \frac{1}{3} \]

\[ f_2(x_2) = E(y \mid x_2) - f_0 = 3x_2 \]

\[ f_{12}(x_1, x_2) = 0 \]

\[ S_1 = \frac{\text{Var}[f_1(x_1)]}{V} = \frac{0.08}{0.838} = 0.106 \]

\[ S_2 = \frac{\text{Var}[f_2(x_2)]}{V} = \frac{0.75}{0.838} = 0.894 \]
Graphical interpretation

First order Sobol’ indices measure the variability of conditional expectations (mean trend curves in the scatterplots)

Null index

High index

Small index
Sobol’ indices properties

\[
1 = \sum_{i=1}^{p} S_i + \sum_{i} \sum_{j} S_{ij} + \sum_{i} \sum_{j} \sum_{k} S_{ijk} \ldots + S_{1,2,\ldots,k}
\]

\[
\sum_{i} S_i \leq 1 \quad \text{Always}
\]

\[
\sum_{i} S_i = 1 \quad \text{Additive model}
\]

1 - \sum_{i} S_i \quad \text{Measure the degree of interactions between variables}

\textbf{Examples:} p = 4 \text{ gives } 4 \text{ indices } S_i, 6 \text{ indices } S_{ij}, 4 \text{ indices } S_{ijk}, 1 \text{ indice } S_{ijkl}

\textbf{General case:} 2^p - 1 \text{ indices to be estimated}

\textbf{Total sensitivity index:} \quad S_{Ti} = S_i + \sum_{j} S_{ij} + \sum_{j,k} S_{ijk} + \ldots = 1 - S_{~i}

[ Homma & Saltelli 1996 ]
Flood model

Sortie S – Indices 1er ordre

Sortie S – Indices totaux

Sortie Cp – Indices 1er ordre

Sortie Cp – Indices totaux
Sobol indices computation

- Indices for $X_i$ ($1^{st}$ order and total):

$$S_i = \frac{V_i}{\text{Var}(Y)} \text{ and } S_{T_i} = 1 - \frac{V_{\sim i}}{\text{Var}(Y)}$$

- Formulations of the conditional variances:

Let $\mathbf{X} = (X_i, X_{\sim i})$ and $\mathbf{X}'$ an independent copy of $\mathbf{X}$

$$V_i(Y) = \text{Var}[E(Y|X_i)] = \int E^2(Y|X_i)dX_i - \left(\int E(Y|X_i)dX_i\right)^2 = \text{Cov}[f(X_i, X_{\sim i}), f(X_i, X_{\sim i}')]$$

$$V_{\sim i}(Y) = \text{Var}[E(Y|X_{\sim i})] = \text{Cov}[f(X_i, X_{\sim i}), f(X_i', X_{\sim i}')]$$
Direct estimation via Monte Carlo

\[ 2 \text{ i.i.d. samples: } (X_i^{(j)})_{i=1,...,p; j=1,...,n} \text{ and } (X_i^{n(j)})_{i=1,...,p; j=1,...,n} \]

Variance (classical estimator): \[ \hat{V}(Y) = \frac{1}{n} \sum_{k=1}^{n} f(X^{(k)})^2 - \hat{f}_0^2 \] avec \[ \hat{f}_0 = \frac{1}{n} \sum_{k=1}^{n} f(X^{(k)}) \]

Conditional variances estimation:

\[ \hat{V}_i(Y) = \frac{1}{n} \sum_{k=1}^{n} f(X_1^{(k)},...,X_{i-1}^{(k)},X_i^{(k)},X_{i+1}^{(k)},...,X_p^{(k)})f(X_1^{n(k)},...,X_{i-1}^{n(k)},X_i^{n(k)},X_{i+1}^{n(k)},...,X_p^{n(k)}) - \hat{f}_0^2 \]

Indices 1\(^{st}\) order: cost = \( n \, (p + 1) \)

\[ \hat{V}_{-i}(Y) = \frac{1}{n} \sum_{k=1}^{n} f(X_1^{(k)},...,X_{i-1}^{(k)},X_i^{(k)},X_{i+1}^{(k)},...,X_p^{(k)})f(X_1^{n(k)},...,X_{i-1}^{n(k)},X_i^{n(k)},X_{i+1}^{n(k)},...,X_p^{n(k)}) - \hat{f}_0^2 \]

Indices 1\(^{st}\) order + total indices: cost = \( n \, (p + 2) \), by inverting

\[ (X_i^{(j)})_{i=1,...,p; j=1,...,n} \text{ and } (X_i^{n(j)})_{i=1,...,p; j=1,...,n} \text{ in } \hat{V}_{-i}(Y) \]

In practice, \( n \sim 1e4 \) => problem of the cost in terms of required model runs

Other formula (Jansen-Sobol estimator):

\[ \hat{V}_i = \frac{1}{n} \sum_{k=1}^{n} f(X_{k,1}^{(2)},...,X_{k,p}^{(2)}) \left[ f(X_{k,1}^{(1)},...,X_{k,i-1}^{(1)},X_{k,i}^{(2)},X_{k,i+1}^{(1)},...,X_{k,p}^{(1)}) - f(X_{k,1}^{(1)},...,X_{k,p}^{(1)}) \right] \]
The sampling-based approaches

Sample \( (X \in \mathbb{R}^p, Y(X) \in \mathbb{R}) \) of size \( N > p \)

? Linear relation ?

Yes

\( (R^2) \)

Linear regression between \( X \) and \( Y \)

Regression coefficients

Sensitivity indices

? Monotonic relation ?

No

\( (R^2*) \)

Regression on ranks

Sobol indices

? CPU time cost of the model ?

Negligible

Monte Carlo

\( N > 1000 \ p \)

Small

Quasi-MC, FAST, RBD,

\( N > 100 \ p \)

Large

Smoothing Metamodel

\( N > 10 \ p \)
Flood model – Output Cp

From the 100-size Monte Carlo sample, a Gaussian process metamodel is fitted

Predictivity of the Gp metamodel : $Q^2 = 99\%$

$N=1e5$
100 replicates

$N \times (p+2) \times 100 = 7e7$ evaluations

<table>
<thead>
<tr>
<th>Indices (en %)</th>
<th>$Q$</th>
<th>$K_s$</th>
<th>$Z_v$</th>
<th>$H_d$</th>
<th>$C_b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_i$ modèle</td>
<td>35.5</td>
<td>15.9</td>
<td>18.3</td>
<td>12.5</td>
<td>3.8</td>
</tr>
<tr>
<td>$S_i$ métamodèle</td>
<td>38.9</td>
<td>16.8</td>
<td>18.8</td>
<td>13.9</td>
<td>3.7</td>
</tr>
<tr>
<td>$S_{Ti}$ modèle</td>
<td>48.2</td>
<td>25.3</td>
<td>22.9</td>
<td>18.1</td>
<td>3.8</td>
</tr>
<tr>
<td>$S_{Ti}$ métamodèle</td>
<td>45.5</td>
<td>21.0</td>
<td>21.3</td>
<td>16.8</td>
<td>4.3</td>
</tr>
</tbody>
</table>
Classification of sensitivity analysis methods

Calculations of all types of indices (Sobol, distribution-based, ...) + main effects $E(Y | X_i)$

Number of model $f$ evaluations

$p = \text{number of input variables}$

Venez’ course 2012 – Design and analysis - B. Iooss
Bibliography

– Fang et al., *Design and modeling for computer experiments*, Chapman & Hall, 2006


– Koehler & Owen, Computer experiments, 1996
