Generalized Sobol indices for dependent variables

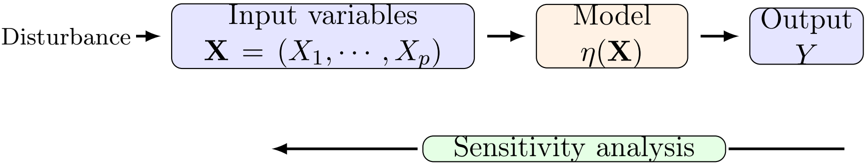
Gaelle Chastaing
Fabrice Gamboa
Clémentine Prieur

July 1st, 2013

Contents

Contents

Mathematical framework



- $\mathbf{X}$ is a random vector of distribution $P_{\mathbf{X}}$, not necessarily independent,
- $\eta \in L^2_{\mathbb{R}}(\mathbb{R}^p)$ i.e., for all $h_1, h_2 \in L^2_{\mathbb{R}}(\mathbb{R}^p)$,

$$\langle h_1, h_2 \rangle = \int_{\mathbb{R}^p} h_1(\mathbf{x}) h_2(\mathbf{x}) dP_{\mathbf{X}}(\mathbf{x}) = \mathbb{E}[h_1(\mathbf{X}) h_2(\mathbf{X})]$$

- Y is in value in $\mathbb{R}$, and $V(Y) \neq 0$.

Use of functional decomposition

- When **X independent**,
  - the Hoeffding decomposition gives

$$Y = \eta(\mathbf{X}) = \eta_\emptyset + \sum_{i=1}^{p} \eta_i(X_i) + \cdots + \eta_{1,\cdots,p}(\mathbf{X}) \qquad (1)$$

(1) **exists and is unique** iff

$$\mathbb{E}[\eta_u(\mathbf{X}_u)\eta_v(\mathbf{X}_v)] = 0 \quad \forall\, u, v \subseteq \{1, \cdots, p\}, \quad u \neq v$$

with $\mathbf{X}_u = (X_{u_1}, \cdots, X_{u_t})$ if $u = (u_1, \cdots, u_t) \subseteq \{1, \cdots, p\}$

The Sobol index

- When **X independent**,
  - Then,

$$V(Y) = \sum_i V[\eta_i(X_i)] + \sum_{i<j} V[\eta_{ij}(X_i, X_j)] + \cdots + V[\eta_{1,\ldots,p}(\mathbf{X})]$$

  - The Sobol index of a group of variables $\mathbf{X}_u$ is

$$S_u = \frac{V(\eta_u)}{V(Y)}$$

  - Moreover

$$\begin{array}{lcl} \eta_\emptyset & = & \mathbb{E}(Y) \\ \eta_i & = & \mathbb{E}(Y|X_i) - \eta_\emptyset \\ \eta_{ij} & = & \mathbb{E}(Y|X_i, X_j) - \eta_i - \eta_j - \eta_\emptyset \\ \ldots & & \end{array}$$

$$S_u = \frac{V[\mathbb{E}(Y|\mathbf{X}_u)] - \sum_{v \subset u} (-1)^{|u|-|v|} V[\mathbb{E}(Y|\mathbf{X}_v)]}{V(Y)}$$

Functional decomposition

- When **X no more independent**,
  - The usual approach [XG08, LRY+10, Can12] is

$$Y \simeq \sum_{u \subset \{1,\cdots,p\}} \eta_u(\mathbf{X}_u) + \varepsilon, \quad \max|u| = d \sim 2,3$$

  - Use of surrogate models to approximate each components (splines, polynomial chaos,. . .) ;
  - ANOVA decomposition

$$V(Y) \simeq \sum_{|u| \leq d} \left[V(\eta_u(\mathbf{X}_u)) + \sum_{v \neq u} \mathrm{Cov}(\eta_u(\mathbf{X}_u), \eta_v(\mathbf{X}_v)) \right]$$

  - Sensitivity index

$$S_u = \frac{V(\eta_u(\mathbf{X}_u)) + \sum_{v \neq u} \mathrm{Cov}(\eta_u(\mathbf{X}_u), \eta_v(\mathbf{X}_v))}{V(Y)}$$

Degree of implication of meta models ?

Contents

Generalized functional decomposition

Our objective :

- Propose a decomposition for the initial model η ;
- Ensure the existence and uniqueness of such decomposition

Context

- $Y = \eta(\mathbf{X}), \quad \mathbf{X} \sim P_{\mathbf{X}}$
- $p_{\mathbf{X}}$ the density function of $\mathbf{X}$
- $p_{\mathbf{X}_u}$ et $p_{\mathbf{X}_{-u}}$ marginal densities of $\mathbf{X}_u$ and $\mathbf{X}_{-u} = \mathbf{X} \setminus \mathbf{X}_u$

Generalized functional decomposition

Condition

$$\boxed{\exists\ 0 < M \leq 1,\ \forall\ u \subseteq \{1,..,p\},\quad p_X \geq M \cdot p_{\mathbf{X}_u} p_{\mathbf{X}_{-u}}} \qquad \text{(C.1)}$$

Theorem

Under (C.1), we have

$$\eta(\mathbf{X}) = \eta_\emptyset + \sum_i \eta_i(X_i) + \sum_{i<j} \eta_{ij}(X_i, X_j) + \cdots + \eta_{1,\cdots,p}(\mathbf{X})$$

Moreover, this decomposition is unique iff $\eta_u \in H_u^0$, *with*

$$H_u^0 = \left\{h_u(\mathbf{X}_u),\ \mathbb{E}[h_u(\mathbf{X}_u)h_v(\mathbf{X}_v)] = 0, \forall v \subset u, h_v \in H_v^0\right\}$$

Illustration with $p = 3$ inputs

Functional decomposition

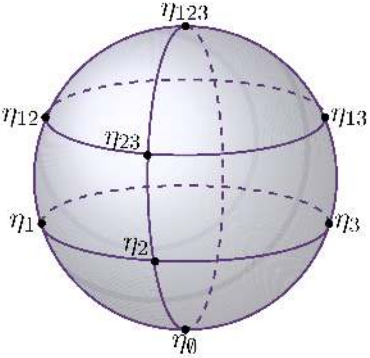
$$\eta(\mathbf{X}) = \eta_\emptyset + \eta_1(X_1) + \eta_2(X_2) + \eta_3(X_3) + \eta_{12}(X_1, X_2) + \eta_{13}(X_1, X_3) + \eta_{23}(X_2, X_3) + \eta_{123}(\mathbf{X})$$

Orthogonality relation representation for

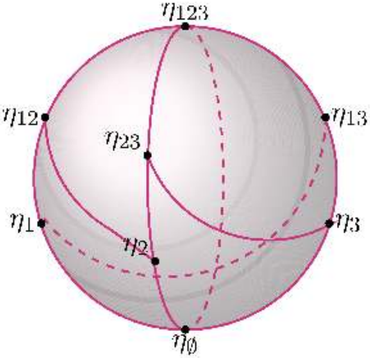
- $P_{\mathbf{X}} = P_{X_1} \otimes \cdots \otimes P_{X_p}$ (Hoeffding decomposition)
- For general $P_{\mathbf{X}}$ (Generalized decomposition)

Illustration

Hoeffding decomposition



Generalized decomposition



Contents

Generalized Sobol sensitivity indices

We have

- $Y = \eta(\mathbf{X}) = \sum_u \eta_u(\mathbf{X}_u)$,
- $\mathbb{E}[\eta_u(\mathbf{X}_u)\eta_v(\mathbf{X}_v)] = 0$ whether $u \subset v$ or $v \subset u$

As a consequence,

$$V(Y) = \sum_u [V(\eta_u) + \sum_{u \cap v \neq u,v} \text{Cov}(\eta_u, \eta_v)]$$

Sensitivity index for the group $\mathbf{X}_u$ is

$$\boxed{S_u = \frac{V(\eta_u(\mathbf{X}_u)) + \sum_{u \cap v \neq u,v} \text{Cov}(\eta_u(\mathbf{X}_u), \eta_v(\mathbf{X}_v))}{V(Y)}}$$

and

$$\sum_u S_u = 1$$

Contents

Procedure

Remind that $Y = \sum_u \eta_u(\mathbf{X}_u)$ with

$$\eta_u \in H_u^0 := \{h_u(\mathbf{X}_u),\ \mathbb{E}[h_u(\mathbf{X}_u)h_v(\mathbf{X}_v)] = 0, \forall v \subset u\}$$
$$H_u^0 \perp H_v^0, \quad \forall\, v \subset u$$

Idea :

1. Construct $H_u^{0,L} = \text{Span}\left\{\phi_1^u, \cdots, \phi_{L_u}^u\right\} \simeq H_u^0, \forall\, u$;
2. Estimate components $(\eta_u)_u$ by the LSE

$$(\eta_u)_u \in \operatorname*{Arg\,min}_{\tilde{\eta}_u \in H_u^0} \mathbb{E}[(Y - \sum_{u \in \{1,\cdots,p\}} \tilde{\eta}_u(\mathbf{X}_u))^2];$$

3. Estimate empirically the generalized sensitivity indices

$$S_u = \frac{V(\eta_u(\mathbf{X}_u)) + \sum_{u \cap v \neq u,v} \text{Cov}(\eta_u(\mathbf{X}_u), \eta_v(\mathbf{X}_v))}{V(Y)}.$$

Step 1 : Construction of the approximation spaces

$$H_u^0 = \left\{ h_u(\mathbf{X}_u),\ \mathbb{E}[h_u(\mathbf{X}_u) h_v(\mathbf{X}_v)] = 0, \forall v \subset u, h_v \in H_v^0 \right\}, \forall\ u$$

Definition of the finite approximation spaces $H_u^{0,L} \simeq H_u^0$, $\forall\ u$ by

- use of truncated orthonormal basis $(\phi_{l_i}^i)_{l_i=0}^L$ of $L^2(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_{X_i})$, $\forall\ i = 1, \cdots, p$;
- inductive procedure on $|u|$;
- exploitation of extended basis to reconstitute the hierarchical orthogonality.

Procedure

1. Initialization : for any $i \in \{1, \cdots, p\}$, construct

$$H_i^{0,L} \simeq H_i^0 := \{h_i(X_i),\ \mathbb{E}[h_i(X_i)] = 0\}$$

Proceed as

- choose a truncated orthonormal basis $(\phi_{l_i}^i)_{l_i=0}^L$ of $L^2(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_{X_i})$ with $\phi_0^i = 1$. Then,

$$\mathbb{E}(\phi_{l_i}^i) = 0, \quad \forall\ l_i = 1, \cdots, L$$

Set

$$H_i^{0,L} = \text{Span}\left\{\phi_1^i, \cdots, \phi_L^i\right\}$$

Finally,

$$h_i(X_i) = \sum_{l_i=1}^{L_i} \beta_{l_i}^i \phi_{l_i}^i(X_i),\ \text{and}\ \mathbb{E}[h_i(X_i)] = 0$$

Procedure(2)

❷ For any $i \neq j \in \{1, \cdots, p\}$, construct

$$H_{ij}^{0,L} \simeq H_{ij}^{0} := \{h_{ij}(X_i, X_j), \mathbb{E}[h_{ij}h_i] = \mathbb{E}[h_{ij}h_j] = \mathbb{E}[h_{ij}] = 0\}$$

Proceed recursively as

- set

$$\phi_{l_{ij}}^{ij}(X_i, X_j) = \phi_{l_i}^{i} \otimes \phi_{l_j}^{j}(X_i, X_j) + \sum_{k=1}^{L} \lambda_k^i \phi_k^i(X_i) + \sum_{k=1}^{L} \lambda_k^j \phi_k^j(X_j) + C$$

- compute the coefficients $(C, \lambda_1^i, \cdots, \lambda_L^i, \lambda_1^j, \cdots, \lambda_L^j)$ by solving

$$\begin{cases} \mathbb{E}[\phi_{l_{ij}}^{ij} \cdot \phi_k^i] = 0, & \forall\, k \in \{1, \cdots, L\} \\ \mathbb{E}[\phi_{l_{ij}}^{ij} \cdot \phi_k^j] = 0, & \forall\, k \in \{1, \cdots, L\} \\ \mathbb{E}[\phi_{l_{ij}}^{ij}] = 0 \end{cases} \quad (2)$$

Set $H_{ij}^{0,L} = \text{Span}\left\{\phi_1^{ij}, \cdots \phi_{L_{ij}}^{ij}\right\}$, $L_{ij} = L \times L$.

❸ The same idea for any $|u| = k$

Construction in practice

Suppose that

- we get a i.i.d. n-sample $(y^l, \mathbf{x}^l)_{l=1,\cdots,n}$
- define the empirical expected value $\mathbb{E}_{\mathrm{n}}(\cdot)$ as

$$\mathbb{E}_{\mathrm{n}}(g) = \frac{1}{n}\sum_{l=1}^{n} g(\mathbf{x}^l)$$

- $n \gg L$

The same procedure can be applied by replacing $\mathbb{E}(\cdot)$ by $\mathbb{E}_{\mathrm{n}}(\cdot)$.
Finally,

$$H_{u,\mathrm{n}}^{0,L} = \mathrm{Span}\left\{\hat{\phi}_{1,\mathrm{n}}^{u}, \cdots, \hat{\phi}_{L_u,\mathrm{n}}^{u}\right\}, \quad \forall\ u$$

Step 2 : Least-squared estimation

- We get

$$\eta_u(\mathbf{X}_u) \simeq \sum_{l_u=1}^{L_u} \beta_{l_u}^u \hat{\phi}_{l_u,\mathrm{n}}^u(\mathbf{X}_u), \quad \forall\, u$$

- Estimate components (η_u) from the minimization

$$\min_{\beta_{l_u}^u} \frac{1}{n} \sum_{s=1}^{n} [y^l - \sum_u \sum_{l_u=1}^{L_u} \beta_{l_u}^u \hat{\phi}_{l_u,\mathrm{n}}^u(\mathbf{x}_u^l)]^2$$

- The total number of components $(\beta_{l_u}^u)$ is

$$pL + \binom{p}{2} L^2 + \binom{p}{3} L^3 + \cdots$$

Unfeasible scheme for large p !

Solution

- Order of ANOVA $d \ll p$, i.e.

$$Y \simeq \sum_{|u| \leq d} \eta_u(\mathbf{X}_u)$$

- Use of a penalization to select a small number of $(\hat{\phi}^u_{l_u,\mathrm{n}})$,

$$\min_{\beta^u_{l_u}} \frac{1}{n} \sum_{s=1}^{n} [y^l - \sum_{|u| \leq d} \sum_{l_u=1}^{L_u} \beta^u_{l_u} \hat{\phi}^u_{l_u,\mathrm{n}}(\mathbf{x}^l_u)]^2 + \lambda J(\beta^1_1, \cdots, \beta^u_{l_u}, \cdots)$$

  - $J(\beta) = \|\beta\|_0 := \sum_j \mathbb{1}(\beta_j \neq 0)$ by greedy algorithm ;
    - Discrete process ; sparse selection from a dictionary ;
  - $J(\beta) = \|\beta\|_1 := \sum_j |\beta_j|$ by LARS algorithm.
    - Convex relaxation of the ℓ_0 penalty ; build up the regression function by successive steps ;

Step 3 : Estimation of sensitivity indices

- We get

$$\hat{\eta}_u = \sum_{l_u=1}^{L_u} \hat{\beta}_{l_u}^u \hat{\phi}_{l_u,\mathrm{n}}^u, \quad \forall\ u, |u| \leq d$$

- Empirical estimation of sensitivity indices S_u

$$\hat{S}_u = \frac{\widehat{V}(\hat{\eta}_u(\mathbf{X}_u)) + \sum_{u \cap v \neq u,v} \widehat{\mathrm{Cov}}(\hat{\eta}_u(\mathbf{X}_u), \hat{\eta}_v(\mathbf{X}_v))}{\widehat{V}(Y)}.$$

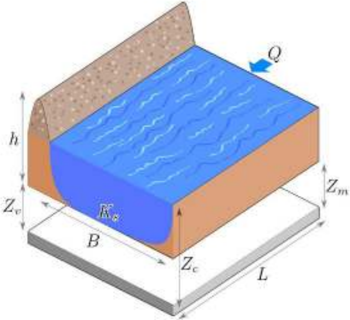
with

$$\widehat{\mathrm{Cov}}(\hat{\eta}_u(\mathbf{X}_u), \hat{\eta}_v(\mathbf{X}_v)) = \frac{1}{n} \sum_{l=1}^{n} \hat{\eta}_u(\mathbf{x}_u^l) \hat{\eta}_v(\mathbf{x}_v^l), \quad \forall\ u, v$$

$$\widehat{V}(Y) = \frac{1}{n} \sum_{l=1}^{n} (y^l - \bar{y})^2, \quad \bar{y} = \frac{1}{n} \sum_{l=1}^{n} y^l$$

Contents

River flood inundation

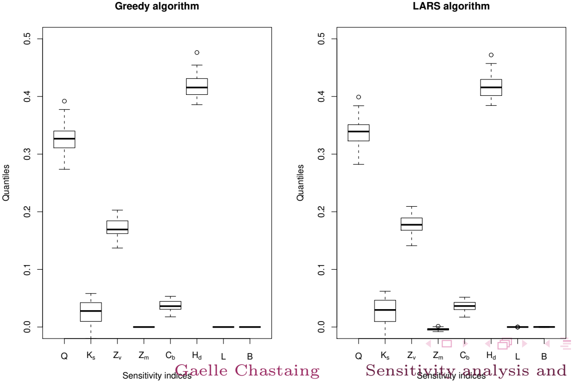


- Aim : prevent from inundation
- The maximal overflow is modelized by eight parameters
- Why randomness and uncertainty ?
  - spatio-temporal variability ;
  - errors of measurements.

River flood inundation

Test with

- $n = 500$ observations, $d = 2$;
- pairwise correlations ;
- Hermite basis of degree 10, i.e. $P = 2880$ parameters



Thank you for your attention !

Y. Caniou.
Analyse de sensibilité globale pour les modèles imbriqués et multiéchelles.
PhD thesis, Université Blaise Pascal - Clermont II, 2012.

G. Chastaing, F. Gamboa, and C. Prieur.
Generalized hoeffding-sobol decomposition for dependent variables -Application to sensitivity analysis.
Electronic Journal of Statistics, 6 :2420–2448, 2012.

G. Chastaing, F. Gamboa, and C. Prieur.
Generalized sobol sensitivity indices for dependent variables : Numerical methods.
Available at `http ://arxiv.org/abs/1303.4372`, 2013.

G. Li, H. Rabitz, P.E. Yelvington, O. Oluwole, F. Bacon, Kolb C.E., and J. Schoendorf.
Global sensitivity analysis with independent and/or correlated inputs.
Journal of Physical Chemistry A, 114 :6022–6032, 2010.

I.M. Sobol.
Sensitivity estimates for nonlinear mathematical models.
Mathematical Modeling and Computational Experiment, 1(4) :407–414, 1993.

C. Xu and G. Gertner.
Uncertainty and sensitivity analysis for models with correlated parameters.
Reliability Engineering & System Safety, 93 :1563–1573, 2008.