Estimation of the Sobol indices in a linear functional multidimensional model

J.C. Fort, T. Klein, A. Lagnoux and B. Laurent*

* Institut de Mathématiques de Toulouse
TOULOUSE - FRANCE

7th International Conference on Sensitivity Analysis of Model Output
Nice - July 1-4, 2013

This work has been partially supported by the French National Research Agency through COSINUS program: Costa-Brava project

INSTITUT de MATHEMATIQUES de TOULOUSE

Let $\mathbb{H}$ a separable Hilbert space endowed with the scalar product $<,>$. Usually $\mathbb{H} = L^2$.
We consider the following linear model

$$Y = \mu + \sum_{k=1}^{p} < \beta^k, X^k > + \varepsilon \tag{1}$$

- X^k are centered stochastic processes $\in \mathbb{H}$ st $\mathbb{E}(\|X^k\|^4) < \infty$;
- β^k are elements of $\mathbb{H}$;
- ε is a centered noise independent of the X^k's st $\mathbb{E}(\|\varepsilon\|^4) < \infty$.

Remark : such a model can arise for example when one wants to define a metamodel to replace an expensive black-box.

Our goal is to quantify the influence of X^k on Y, for $k = 1 \dots p$.

We use as the suggested by Hoeffding decomposition the Sobol index

$$S^{(k)} := \frac{\mathrm{Var}(\mathbb{E}(Y|X^k))}{\mathrm{Var}(Y)}, \qquad k = 1 \dots p.$$

Our goal is to quantify the influence of X^k on Y, for $k = 1 \dots p$.

We use as the suggested by Hoeffding decomposition the Sobol index

$$S^{(k)} := \frac{\mathrm{Var}(\mathbb{E}(Y|X^k))}{\mathrm{Var}(Y)}, \qquad k = 1 \dots p.$$

The model : Let us restrict to $p = 1$ and consider

$$Y = \mu + < \beta, X > + \varepsilon \tag{2}$$

In this setting, the quantity to estimate

$$S = \frac{\mathrm{Var}(\mathbb{E}(Y|X))}{\mathrm{Var}(Y)}$$

is of less interest, but the computations then easily extend to the generic model.

Outline of the talk

Outline of the talk

A first estimation of $\mathrm{Var}(\mathbb{E}(Y|X))$

Precisions on the framework

The observations consist in n i.i.d. copies (X_i, Y_i) of (X, Y).

Since $\mathrm{Var}(Y)$ is naturally estimated by the empirical variance based on $(Y_1, \ldots, Y_n)$

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2,$$

the main purpose is to estimate the quantity $\mathrm{Var}(\mathbb{E}(Y|X))$.

A first estimation of $\mathrm{Var}(\mathbb{E}(Y|X))$

Our approach is based on the so-called Karhunen-Loève decomposition of the processes X and β :

$$X = \sum_{j=1}^{\infty} \sqrt{\lambda_j}\xi_j\varphi_j \quad \text{and} \quad \beta = \sum_{j=1}^{\infty} \gamma_j\varphi_j$$

with ξ_j centered and uncorrelated random variables. Then

$$< X, \varphi_j >= \sqrt{\lambda_j}\xi_j.$$

A first estimation of $\mathrm{Var}(\mathbb{E}(Y|X))$

Notice that

$$\begin{aligned}
\mathbb{E}(YX) &= \mathbb{E}(< X, \beta > X) = \mathbb{E}\left[\left(\sum_{l=1}^{\infty} \sqrt{\lambda_l}\gamma_l\xi_l\right)\left(\sum_{l=1}^{\infty} \sqrt{\lambda_l}\xi_l\varphi_l\right)\right] \\
&= \mathbb{E}\left[\left(\sum_{l=1}^{\infty} \lambda_l\gamma_l\xi_l^2\varphi_l\right)\right] = \sum_{l=1}^{\infty} \lambda_l\gamma_l\varphi_l
\end{aligned}$$

A first estimation of $\mathrm{Var}(\mathbb{E}(Y|X))$

Notice that

$$\begin{aligned}\mathbb{E}(YX) &= \mathbb{E}(< X, \beta > X) = \mathbb{E}\left[\left(\sum_{l=1}^{\infty}\sqrt{\lambda_l}\gamma_l\xi_l\right)\left(\sum_{l=1}^{\infty}\sqrt{\lambda_l}\xi_l\varphi_l\right)\right] \\ &= \mathbb{E}\left[\left(\sum_{l=1}^{\infty}\lambda_l\gamma_l\xi_l^2\varphi_l\right)\right] = \sum_{l=1}^{\infty}\lambda_l\gamma_l\varphi_l\end{aligned}$$

As a consequence, $\gamma_j = \frac{1}{\lambda_j} < \mathbb{E}(YX), \varphi_j >$ that is naturally estimated by

$$\widehat{\gamma}_j = \frac{1}{\lambda_j}\frac{1}{n}\sum_{i=1}^{n} < X_i, \varphi_j > Y_i.$$

A first estimation of $\mathrm{Var}(\mathbb{E}(Y|X))$

- First, we have

$$\widehat{\gamma}_j = \frac{1}{\lambda_j}\frac{1}{n}\sum_{i=1}^{n} < X_i, \varphi_j > Y_i.$$

A first estimation of $\mathrm{Var}(\mathbb{E}(Y|X))$

- First, we have

$$\widehat{\gamma}_j = \frac{1}{\lambda_j}\frac{1}{n}\sum_{i=1}^{n} < X_i, \varphi_j > Y_i.$$

- Second, expansion in the KL basis gives

$$\mathrm{Var}(\mathbb{E}(Y|X)) = \mathbb{E}(< \beta, X >^2) = \sum_{j=1}^{\infty} \lambda_j \gamma_j^2.$$

A natural estimation of $\mathrm{Var}(\mathbb{E}(Y|X))$ is then

$$\widehat{E}_m^1 = \sum_{l=1}^{m} \frac{1}{\lambda_l}\frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} Y_i < X_i, \varphi_l > Y_j < X_j, \varphi_l > .$$

A second estimation of $\mathrm{Var}(\mathbb{E}(Y|X))$

- We consider another design of experiment : let ε' be a copy of ε, independent of X and ε and

$$\begin{cases} Y & = \mu + < X, \beta > + \varepsilon \\ Y^X & = \mu + < X, \beta > + \varepsilon' \end{cases}$$

- Now the observations consist in
  (1) n-sample of (X, Y) : (X_i, Y_i), $1 \leq i \leq n$.
  (2) n-sample of (X, Y^X) : (X_i, Y_i^X), $1 \leq i \leq n$.

A second estimation of $\mathrm{Var}(\mathbb{E}(Y|X))$

- We consider another design of experiment : let ε' be a copy of ε, independent of X and ε and

$$\begin{cases} Y & = \mu + < X, \beta > + \varepsilon \\ Y^X & = \mu + < X, \beta > + \varepsilon' \end{cases}$$

- Now the observations consist in
  (1) n-sample of (X, Y) : (X_i, Y_i), $1 \leq i \leq n$.
  (2) n-sample of (X, Y^X) : (X_i, Y_i^X), $1 \leq i \leq n$.
- $\mathrm{Var}(Y)$ is naturally estimated by the empirical variance based on $(Y_1, \ldots, Y_n)$ and $(Y_1^X, \ldots, Y_n^X)$

$$\frac{1}{2n} \sum_{i=1}^{n} \left[(Y_i)^2 + \left(Y_i^X\right)^2 \right] - \left(\frac{1}{2n} \sum_{i=1}^{n} \left[Y_i + Y_i^X \right] \right)^2 .$$

A second estimation of $\mathrm{Var}(\mathbb{E}(Y|X))$

- It remains to estimate $\mathrm{Var}(\mathbb{E}(Y|X))$ that can be rewritten as

$$\mathrm{Var}(\mathbb{E}(Y|X)) = \mathrm{Cov}(Y, Y^X).$$

- A natural estimation of $\mathrm{Var}(\mathbb{E}(Y|X))$ is then :

$$\widehat{E}^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^X - \left(\frac{1}{2n}\sum_{i=1}^{n}\left[Y_i + Y_i^X\right]\right)^2.$$

Straighforwardly $\widehat{E}^1_m$ is biased and

$$\mathbf{B}_m = \mathbb{E}(\widehat{E}^1_m) - \mathrm{Var}(\mathbb{E}(Y|X)) = \sum_{l=m+1}^{\infty} \lambda_l \gamma_l^2$$

whereas $\widehat{E}^2$ is unbiased.

Straighforwardly $\widehat{E}_m^1$ is biased and

$$\mathbf{B}_m = \mathbb{E}(\widehat{E}_m^1) - \mathrm{Var}(\mathbb{E}(Y|X)) = \sum_{l=m+1}^{\infty} \lambda_l \gamma_l^2$$

whereas $\widehat{E}^2$ is unbiased.

Some statistical questions :

1. Are $\widehat{E}_m^1$ and $\widehat{E}^2$ "good" estimators for $\mathrm{Var}(\mathbb{E}(Y|X))$?
2. Are they consistent ? If yes, what is the rate of convergence ?
   Answer : Central Limit Theorem (cv in $\sqrt{n}$).
3. Are they asymptotically efficient ?
4. Can we measure their quality at a fixed n ?
   Answer : Berry-Esseen and/or concentration inequalities.
5. Are the estimators and designs of experiment comparable ?

Outline of the talk

Asymptotic properties of $\widehat{E}_m^1$

Consistency : $\widehat{E}_m^1$ and $\widehat{E}^2 \underset{n\to\infty}{\overset{\mathbb{P}}{\to}}$ are consistent.

Asymptotic properties of $\widehat{E}_m^1$

Consistency : $\widehat{E}_m^1$ and $\widehat{E}^2 \underset{n\to\infty}{\overset{\mathbb{P}}{\to}}$ are consistent.

Asymptotic normality

$$\begin{aligned}\widehat{E}_m^1 &= \sum_{l=1}^{m} \frac{1}{\lambda_j} \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} Y_i < X_i, \varphi_l > Y_j < X_j, \varphi_l > \\ &= U_n K + P_n L - \mathbf{B}_m + \mathrm{Var}(\mathbb{E}(Y|X))\end{aligned}$$

$$\begin{aligned}\text{with} \quad U_n K &= \sum_{l=1}^{m} \frac{1}{\lambda_l} \frac{1}{n(n-1)} \sum_{1\leq i\neq j\leq n} Z_{i,l}^c \quad \underbrace{Z_{j,l}^c}_{Y_i < X_i, \varphi_l > -\mathbb{E}(Y_j < X_j, \varphi_l >)} \\ \text{and} \quad P_n L &= \frac{2}{n} \sum_{l=1}^{m} \sum_{i=1}^{n} \gamma_l Z_{i,l}^c .\end{aligned}$$

Asymptotic properties of $\widehat{E}_m^1$

We want to show

$$\mathbf{B}_m^2 = o\left(\frac{1}{n}\right), \quad U_n K = o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right), \quad \sqrt{n} P_n L \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, C(\beta))$$

Asymptotic properties of $\widehat{E}_m^1$

We want to show

$$\mathbf{B}_m^2 = o\left(\frac{1}{n}\right), \quad U_n K = o_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right), \quad \sqrt{n} P_n L \underset{n\to\infty}{\xrightarrow{\mathcal{L}}} \mathcal{N}(0, C(\beta))$$

Assumptions :

- (A1) $\mathbb{E}(\|X\|^4) < +\infty$ and $\mathbb{E}(\varepsilon^4) < +\infty$.
- (A2) $\sup_{l\geq 1} \mathbb{E}(\xi_l^4) < +\infty$.
- (A3) there exist $C > 0$ and $\delta > 1$ such that

$$\forall l \geq 1, \quad \lambda_l \leq C l^{-\delta}.$$

Now let $m = m(n) = \sqrt{n} h(n)$, where $h(n)$ satisfies : $h(n) \to 0$ and $\forall \alpha > 0$, $n^\alpha h(n) \to +\infty$ as $n \to +\infty$.

Theorem (Asymptotic normality)

(i) Since $\widehat{E}^1_m - Var(\mathbb{E}(Y|X)) = U_nK + P_nL - \mathbf{B}_m$ and assuming (A1-3) and $n^{1/2(\delta+2s)} << m << \sqrt{n}$, one gets

$$\begin{cases} \mathbf{B}_m^2 = o\left(\frac{1}{n}\right) \quad \mathbb{E}\left((U_nK)^2\right) = o\left(\frac{1}{n}\right) \\ \sqrt{n}P_nL \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0, 4\mathrm{Var}(Y < X, \beta >)) \end{cases}$$

then $\sqrt{n}(\widehat{E}^1_m - \mathrm{Var}(\mathbb{E}(Y|X))) \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0, 4\mathrm{Var}(Y < X, \beta >))$.

(ii) Since $\mathbb{E}(Y^4) < \infty$,

$$\sqrt{n}(\widehat{E}^2 - \mathrm{Var}(\mathbb{E}(Y|X))) \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}(0, \mathrm{Var}((Y - \mathbb{E}(Y))(Y^X - \mathbb{E}(Y^X)))).$$

Comments

We may assume that $h(n) = 1/\log(n)$, and hence $m(n) = \sqrt{n}/\log n$, to fill the condition

$$\forall \alpha > 0, \lim_{n \to \infty} n^{\alpha} h(n) = +\infty.$$

The estimator $\widehat{V}_m^X$ converges at the parametric rate $1/\sqrt{n}$, for any β. We could have chosen a smaller value of m leading to the same asymptotic efficiency, but depending on δ.

Asymptotic properties of $\widehat{S}^1_m$ and $\widehat{S}^2$

Using the so-called Delta method, one can extend these properties of the numerators to the estimators of the Sobol index S :

Theorem (Asymptotic Normality)

(i) Under the same assumptions as in the previous theorem, we have

$$\sqrt{n}\left(\widehat{S}^1_m - S\right) \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}\left(0, \frac{\mathrm{Var}(U)}{(\mathrm{Var}(Y))^2}\right)$$

where $U := 2Y < X, \beta > -S(Y - \mathbb{E}(Y))^2$.

(ii) Since $\mathbb{E}(Y^4) < \infty$,

$$\sqrt{n}\left(\widehat{S}^2 - S\right) \underset{n\to\infty}{\overset{\mathcal{L}}{\to}} \mathcal{N}\left(0, \frac{\mathrm{Var}(V)}{(\mathrm{Var}(Y))^2}\right)$$

where $V :=$
$(Y - \mathbb{E}(Y))(Y^X - \mathbb{E}(Y)) - S^X/2\left((Y - \mathbb{E}(Y))^2 + (Y^X - \mathbb{E}(Y))^2\right)$.

Remark

- *For independent inputs, we establish more generally in the product space*
  - *the consistency*
  - *the asymptotic normality*
  - *the asymptotic efficiency*

  of $\widehat{S}^1_m := (\widehat{S}^{(1,1)}_m, \ldots, \widehat{S}^{(1,p)}_m)$ *and* $\widehat{S}^2 := (\widehat{S}^{(2,1)}, \ldots, \widehat{S}^{(2,p)})$ *to the vector of Sobol indices*

  $$S := (S^{(1)}, \ldots, S^{(p)}),$$

  the indices 1 and 2 refer to the first and second estimators.
- *One can also generalize these results to Sobol indices defined for subsets* $I \subset \{1, \ldots, p\}$.

Outline of the talk

Estimators considered

A first estimation of $\mathrm{Var}(\mathbb{E}(Y|X))$

A second estimation of $\mathrm{Var}(\mathbb{E}(Y|X))$

Asymptotic properties of the estimators

Numerical Applications

Conclusion

We consider the model with $p = 2$, $\mu = 0$ and $\varepsilon = 0$:

$$Y =< \beta^1, X^1 > + < \beta^2, X^2 >$$

1. First Model : $\gamma^i = (\gamma_1^i, \gamma_2^i, \gamma_3^i, \ldots)$ for $i = 1, \ldots, 2$

$$\gamma_l^i = l^{\delta_i} \quad \text{for} \quad 1 \leq l \leq L \quad \text{and} \quad \gamma_l^i = 0 \quad \text{for} \quad l > L;$$

   with $i = 1 \ldots 2$ and $\delta_i = (-1/2 - 1/100)$.
2. Second Model : $\gamma^i = (0, \gamma_2^i, \gamma_3^i, \ldots)$ for $i = 1, \ldots, 2$.
3. Third Model : $\gamma^i = (\gamma_3^i, \gamma_4^i, \gamma_5^i, \ldots)$ for $i = 1, \ldots, 2$.

We perform $N_{sim} = 5000$ simulations and we study the influence of the parameter n, where $3n$ observations are used for both methods. We set $L = 500$ and $m = \lfloor \sqrt{3n} / \log(3n) \rfloor$.

First Model : $S = (0.5107, 0.4893)$		
n	RMSE($\hat{S}_m$)	RMSE($\hat{S}_{SPF}$)
10^2	$10^{-2}[7.17, 7.21]$	$10^{-2}[8.95, 9.14]$
10^3	$10^{-2}[2.26, 2.20]$	$10^{-2}[2.79, 2.83]$
Second Model : $S = (0.7535, 0.2465)$		
n	RMSE($\hat{S}_m$)	RMSE($\hat{S}_{SPF}$)
10^2	$10^{-2}[8.07, 5.45]$	$10^{-2}[7.80, 9.90]$
10^3	$10^{-2}[2.52, 1.71]$	$10^{-2}[2.41, 3.13]$
Third Model : $S = (0.8655, 0.1345)$		
n	RMSE($\hat{S}_m$)	RMSE($\hat{S}_{SPF}$)
10^2	$10^{-1}[3.01, 0.48]$	$10^{-2}[7.12, 9.97]$
10^3	$10^{-2}[4.67, 1.28]$	$10^{-2}[2.24, 3.17]$

We consider the model with $p = 4$, $\mu = 0$ and $\varepsilon = 0$:

$$Y = \sum_{k=1}^{4} < \beta^k, X^k >$$

1. First Model : $\gamma^i = (\gamma^i_1, \gamma^i_2, \gamma^i_3, \ldots)$ for $i = 1, \ldots, 4$

$$\gamma^i_l = (l+1)^{\delta_i} \quad \text{for} \quad 1 \leq l \leq L \quad \text{and} \quad \gamma^i_l = 0 \quad \text{for} \quad l > L;$$

   with $i = 1 \ldots 4$ and $\delta_i = (-1/2 - 1/100, -1, -2, 3/2)$.
2. Second Model : $\gamma^i = (0, \gamma^i_2, \gamma^i_3, \ldots)$ for $i = 1, \ldots, 4$.
3. Third Model : $\gamma^i = (\gamma^i_3, \gamma^i_4, \gamma^i_5, \ldots)$ for $i = 1, \ldots, 4$.

We perform $N_{sim} = 5000$ simulations and we study the influence of the parameter n, where $5n$ observations are used for both methods. We set $L = 500$ and $m = \lfloor \sqrt{5n} / \log(5n) \rfloor$.

First Model : $S = (0.5438, 0.2639, 0.0635, 0.1288)$		
n	RMSE($\hat{S}_m$)	RMSE($\hat{S}_{SPF}$)
10^2	$10^{-2}[5.55, 4.29, 2.35, 3.22]$	$10^{-2}[9.92, 9.80, 9.75, 9.63]$
10^3	$10^{-2}[1.82, 1.36, 0.72, 0.99]$	$10^{-2}[3.13, 3.12, 3.11, 3.06]$

Second Model : $S = (0.7080, 0.2085, 0.0200, 0.0635])$		
n	RMSE($\hat{S}_m$)	RMSE($\hat{S}_{SPF}$)
10^2	$10^{-2}[6.35, 3.92, 1.47, 2.31]$	$10^{-1}[1.04, 0.99, 0.99, 0.99]$
10^3	$10^{-2}[1.92, 1.22, 0.41, 0.73]$	$10^{-2}[3.29, 3.15, 3.19, 3.14]$

Third Model : $S = (0.7561, 0.1871, 0.0112, 0.0456)$		
n	RMSE($\hat{S}_m$)	RMSE($\hat{S}_{SPF}$)
10^2	$10^{-2}[6.14, 3.72, 1.22, 2.01]$	$10^{-1}[1.07, 1.00, 1.01, 0.99]$
10^3	$10^{-2}[1.97, 1.17, 0.33, 0.60]$	$10^{-2}[3.36, 3.16, 3.14, 3.13]$

Outline of the talk

1. We construct two different estimators of

$$S := (S^{(1)}, \ldots, S^{(p)}),$$

   based on two different designs of experiment for the functional linear regression.
2. The first one $\widehat{S}_m^1$ is based on the Karhunen-Loève expansion of the covariance operator $\Gamma(f) = \mathbb{E}(< X, f > X)$ and performs better for large values of p.
3. Nevertheless, it is more complex and requires the knowledge of the λ_j and φ_j that can be estimated in a future work.
4. The second is more general and applies whatever the context but is performing as well.

Bibliography

J .C. Fort, T. Klein, A. Lagnoux, B. Laurent. “Estimation of the Sobol indices in a linear functional multidimensional model”, *JSPI*, 2013.

N. Hilgert, A. Mas, N. Verzelen. “Minimax adaptive tests for the functional linear model”, *Annals of Statistics*, in press, Arxiv : 1206.1094.

A. Janon, T. Klein, A. Lagnoux, M. Nodet, C. Prieur. “ Asymptotic normality and efficiency of a Sobol index estimator”, *ESAIM P&S*, 2013.

I.M. Sobol. “Sensitivity estimates for nonlinear mathematical models”, *Math. Mod. Comput. Exp.*, 1993.