An optimal balance between explorations and repetitions in sensitivity analysis

Gildas Mazo (ongoing work)

INRA

MASCOT-NUM Annual Conference, 18-20 March 2019

Context

Sensitivity analysis permits to exhibit the important factors of a model

$$Y = f(X_1, \dots, X_p),$$

which represents a phenomenon of interest.

If $X_j = x_j$ was fixed, Var Y would be reduced by S_j%, where S_j, called **the Sobol index** of X_j, is given by

$$S_j = \frac{\operatorname{Var} \mathrm{E}(Y|X_j)}{\operatorname{Var} Y}.$$

Problem

How to do a sensitivity analysis when **the model is stochastic**, that is,

$$Y = f(X, Z),$$

where Z represents some randomness intrinsic to the model?

What does sensitivity analysis mean in this context?

Plan

1. Definition of the indices
2. Construction of the estimators and first properties
3. The optimal number of repetitions and the noise-to-signal ratio
4. Construction of an oracle procedure
5. Numerical illustrations

Sobol indices for stochastic models

Definition

The Sobol index of first kind is defined as

$$S_j' = \frac{\operatorname{Var} \mathrm{E}(f(X,Z)|X_j)}{\operatorname{Var} f(X,Z)}.$$

Definition

The Sobol index of second kind is defined as

$$S_j'' = \frac{\operatorname{Var} \mathrm{E}([\mathrm{E}\, f(X,Z)|X]|X_j)}{\operatorname{Var}[\mathrm{E}\, f(X,Z)|X]}.$$

Differences

Example

$Y = aX_1 + cX_2 Z$, where X_1, X_2, Z are standard normal and a, c real coefficients.

	$j=1$	$j=2$
S'_j	$\frac{a^2}{a^2+c^2}$	0
S''_j	1	0

The Sobol operator

Let $P \sim X \ni \mathcal{X} \longrightarrow \mathcal{Y}$.

Definition

The Sobol operator on (f, P) is defined as

$$S_j(f, P) = \frac{D_{\{j\}}(f, P) - (\mu(f, P))^2}{D_{\{1,\dots,p\}}(f, P) - (\mu(f, P))^2},$$

where

$$D_A(f, P) = \int \int f(x) f(x'_{-A})\, P(\mathrm{d}x') P(\mathrm{d}x) \quad \text{and} \quad \mu(f, P) = \int f \, \mathrm{d}P,$$

with the notation x'_{-A} standing for the vector whose jth component is x'_j if $j \notin A$ and x_j otherwise, $A \subset \{1, \dots, p\}$.
The quantity $D_{\{j\}}(f, P)$ is called the **discriminator**.

Similarities

Let $\mathcal{Z} \ni Z \sim Q$, $\mathcal{X} \ni X \sim P$ and $f : \mathcal{X} \times \mathcal{Z} \longrightarrow \mathcal{Y}$.

Remark

We have $S'_j = S_j(f, P \times Q)$ and $S''_j = S_j(Qf, P)$, where $Qf : \mathcal{X} \longrightarrow \mathcal{Y}$ is the operator defined as

$$Qf(x) = \int f(x, z)\, Q(\mathrm{d}z).$$

Proposition

We have the identities

$$D_j(f, P \times Q) = D_j(Qf, P), \quad \mu(f, P \times Q) = \mu(Qf, P).$$

Corollary

We have $S'_i < S'_j$ if and only if $S''_i < S''_j$

Construction of the estimators

Let $X = (X_1, \ldots, X_p) \sim P$. The estimators are based on the following Monte-Carlo sampling scheme:

for $i = 1$ to n **do**
  draw two independent copies $X^{(i)}, \widetilde{X}^{(i)}$ from P
  for $A \in \{\{1\}, \ldots, \{p\}, \{1, \ldots, p\}\}$ **do**
    for $k = 1$ to m **do**
      run the computer model at $\widetilde{X}_{-A}$ to get an output $Y_A^{(i,k)}$
    end for
  end for
end for

We need $T = mn(p + 1)$ runs of the model.

Then, substitute expectations for empirical averages. For instance,

$$\widehat{D}_{j;n,m} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{k=1}^{m} Y_{\{1,\dots,p\}}^{(i,k)}\frac{1}{m}\sum_{k'=1}^{m} Y_{\{j\}}^{(i,k')}.$$

Do the same to get $\widehat{\mu}_{n,m}$, $D_{\{1,\dots,p\}}(f, P \times Q)$ and $D_{\{1,\dots,p\}}(Qf, P)$. Finally

$$S'_{j;n,m} = \frac{\widehat{D}_{j;n,m} - \widehat{\mu}^2_{n,m}}{\widehat{D}_{n,m}(f, P \times Q) - \widehat{\mu}^2_{n,m}}, \quad S''_{j;n,m} = \frac{\widehat{D}_{j;n,m} - \widehat{\mu}^2_{n,m}}{\widehat{D}_{n,m}(Qf, P) - \widehat{\mu}^2_{n,m}}.$$

We assume $\mathrm{E}\, f(X,Z)^8 < \infty$.

Theorem

Let $n \to \infty$. The following statements are true.

(i) *If m is fixed, then*

$$\sqrt{n}(\widehat{S}'_{j;n,m} - S'_j) \xrightarrow{d} N(0, \sigma'^2_{j;m}), \quad \textit{and}$$

$$\sqrt{n}(\widehat{S}''_{j;n,m} - \left[S''_j + O\left(\frac{\mathrm{E}\,\mathrm{Var}^{\mathbf{X}} f}{m}\right)\right]) \xrightarrow{d} N(0, \sigma''^2_{j;m}).$$

(ii) *If $m \to \infty$, then the conlusions in (i) hold with $\lim_{m\to\infty} \sigma'_{j;m}$ and $\lim_{m\to\infty} \sigma''_{j;m}$ in place of $\sigma'_{j;m}$ and $\sigma''_{j;m}$.*

Corollary

If $\sqrt{n}/m \to 0$, then $\sqrt{n}(\widehat{S}''_{j;n,m} - S''_j) \to N(0, \sigma''^2)$.

The optimal number of repetitions

Let $T = mn(p+1)$ be the computing budget.

Definition and Proposition

The optimal number of repetitions $m^\dagger$ is defined as the argument that minimises

$$\frac{\overbrace{4(p-1)\sum_{j=1}^{p} \operatorname{Var} \widehat{D}_{j;n,m}}^{v(m)/T}}{\min_{j<j'}(|D_j - D_{j'}|^2)} \geq \underbrace{\mathrm{E}\sum_{j=1}^{p} |\widehat{R}_{j;n,m} - R_j|}_{MRE},$$

The noise-to-signal ratio

Lemma

For some constants ζ_1,ζ_2 and ζ_3, we have

$$\sum_{j=1}^{p} \operatorname{Var} \widehat{D}_{j;n,m} = \frac{1}{T}\left(\zeta_1 m + \zeta_2 + \frac{\zeta_3}{m}\right).$$

The minimum over all real m is attained at

$$m^* \equiv \sqrt{\frac{\zeta_3}{\zeta_1}} = \sqrt{\frac{\sum_{j=1}^{p} \operatorname{E} \operatorname{Var}^{\mathbf{X}} f(X,Z) f(\widetilde{X}_{-j}, Z_j)}{\sum_{j=1}^{p} \operatorname{Var} \operatorname{E}^{\mathbf{X}} f(X,Z) f(\widetilde{X}_{-j}, Z_j)}},$$

and is called the **noise-to-signal ratio**.

Relation between the optimal number of repetitions and the noise-to-signal ratio

To simplify this talk, suppose that the noise-to-signal ratio is an integer compatible with the budget equation.

Corollary

The optimal number of repetitions $m^\dagger$ is given according to the following three cases.

(i) *If $m^* \leq 1$ then $m^\dagger = 1$.*

(ii) *If $m^* \geq T/(p+1)$ then $m^\dagger = T/(p+1)$.*

(iii) *If $1 < m^* < T/(p+1)$ then $m^\dagger = m^*$.*

Estimation of the noise-to-signal ratio

By substituting expectations for empirical averages, we can construct an estimator $\widehat{m}^*_{n,m}$ based on the same Monte-Carlo experiment as before.

Theorem

Let $n \to \infty$ and $m \to \infty$. Then $\sqrt{n}(\widehat{m}^*_{n,m} - [m^* + O(1/m)]) \to N(0, \sigma^2)$, where $\sigma > 0$.

Corollary

Let $\sqrt{n}/m \to 0$. Then $\sqrt{n}(\widehat{m}^*_{n,m} - m^*) \to N(0, \sigma^2)$.

Estimation of sensitivity indices by exploiting the noise-to-signal ratio

Consider the two-step procedure given below.

0. Choose two integers (K, m_0) such that $m_0 n_0(p+1) = K < T$.
1. Do a Monte-Carlo experiment $\mathcal{E}(K, m_0)$ to get an estimate $\widehat{m}^{\dagger}_{K,m_0}$ of $m^{\dagger}$. If $K = 0$, take m_0.
2. Do a Monte-Carlo experiment $\mathcal{E}(T-K, \widehat{m}^{\dagger}_{K,m_0})$ to estimate the sensitivity indices.

An oracle property

Let $\widehat{E}_{K,m_0}$ be the excess of variance incurred by our ignorance:

$$\widehat{E}_{K,m_0} = \frac{\frac{1}{T-K} v(\widehat{m}^*_{K,m_0}) - \frac{1}{T} v(m^*)}{\frac{1}{T} v(m^*)}.$$

Let $T \to \infty$.

Theorem

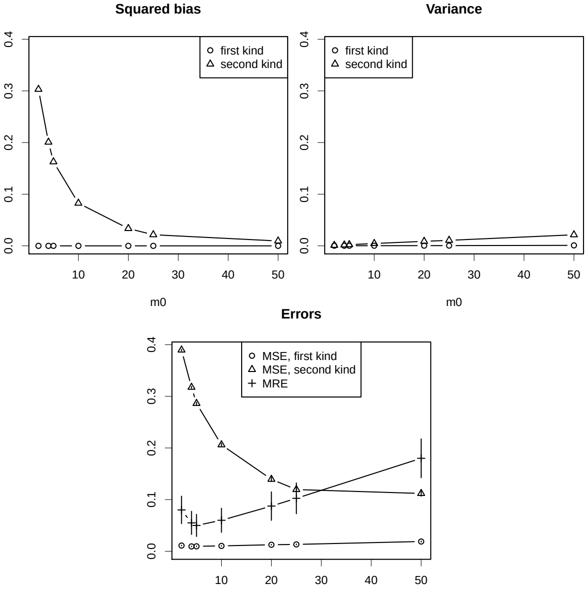
Assume $K_T/T \to c/(c+1)$, $c \in [0, \infty)$. Then $\widehat{E}_{K,m_0} \xrightarrow{P} c$. Moreover, we have

$$T^{2/5} \widehat{E}_{T^{3/5}, T^{1/5}} = O_P(1).$$

Proposition

The rates above are optimal among all powers of T.

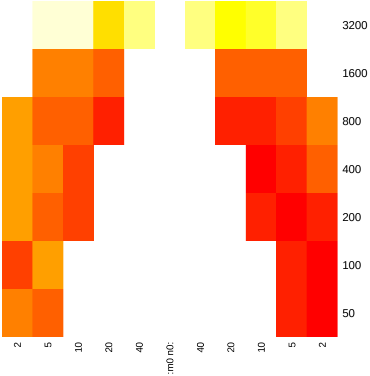
Illustrations. Linear model, high noise



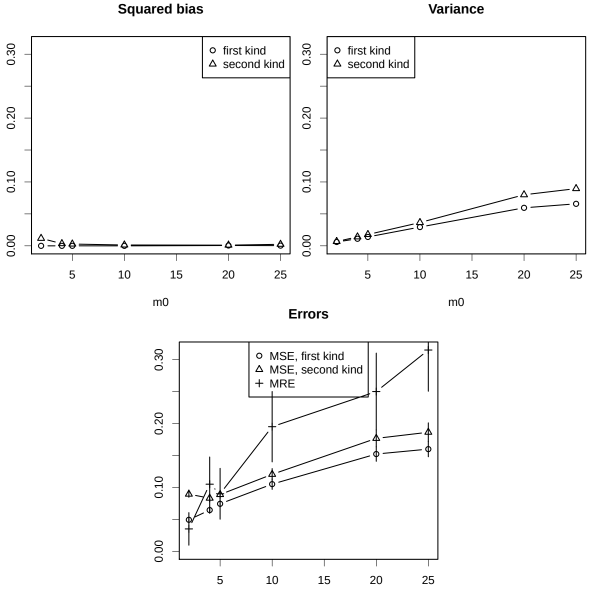
Calibration of the tunning parameters

	2	5	10	20	40	:m0 n0:	40	20	10	5	2
3200	NA	0.25	0.27	0.18	0.24	NA	0.24	0.20	0.23	0.25	NA
1600	NA	0.13	0.14	0.10	NA	NA	NA	0.12	0.11	0.11	NA
800	0.15	0.10	0.11	0.07	NA	NA	NA	0.08	0.07	0.10	0.13
400	0.14	0.12	0.10	NA	NA	NA	NA	NA	0.05	0.06	0.10
200	0.16	0.11	0.10	NA	NA	NA	NA	NA	0.07	0.04	0.08
100	0.10	0.14	NA	NA	NA	NA	NA	NA	NA	0.07	0.06
50	0.12	0.12	NA	NA	NA	NA	NA	NA	NA	0.08	0.06

Table: MRE for various calibrations: $K/(p+1) = 50, 100, \dots$ and $m_0 = 2, 5, \dots$ The greatest values depend on K and hence the values for n_0 have been given instead. For instance, for $K/(p+1) = 200 = m_0 n_0$, the available MREs are for $m_0 = 2, 5, 10, 20, 40, 100$.


3200
1600
800
400
200
100
50
2
5
10
20
40
:m0 n0:
40
20
10
5
2

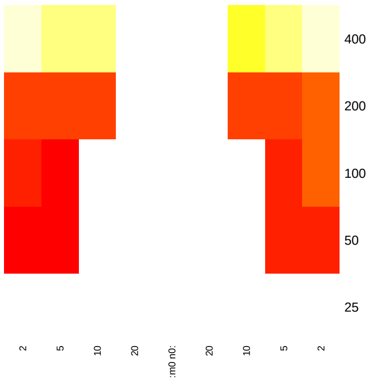
Illustrations. Linear model, low noise



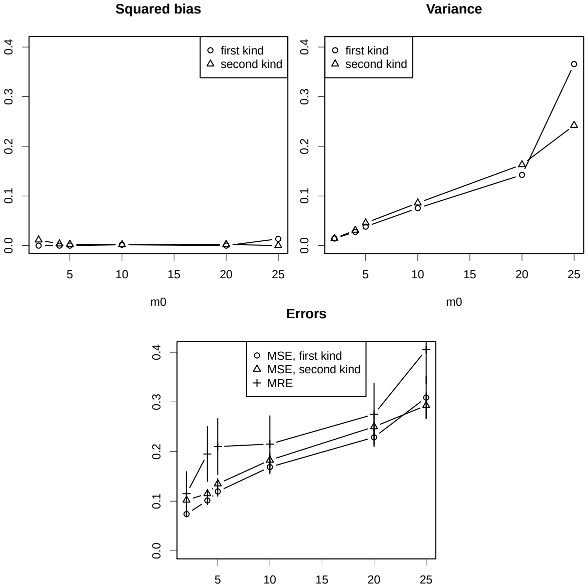
Calibration of the tunning parameters

	2	5	10	20	:m0 n0:	20	10	5	2
400	0.18	0.17	0.17	NA	NA	NA	0.16	0.18	0.19
200	0.05	0.05	0.06	NA	NA	NA	0.05	0.06	0.06
100	0.04	0.02	NA	NA	NA	NA	NA	0.04	0.06
50	0.02	0.01	NA	NA	NA	NA	NA	0.03	0.03
25	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table: MRE for various calibrations: $K/(p+1) = 50, 100, \dots$ and $m_0 = 2, 5, \dots$ The greatest values depend on K and hence the values for n_0 have been given instead. For instance, for $K/(p+1) = 200 = m_0 n_0$, the available MREs are for $m_0 = 2, 5, 10, 20, 40, 100$.


400
200
100
50
25
2
5
10
20
:m0 n0:
20
10
5
2

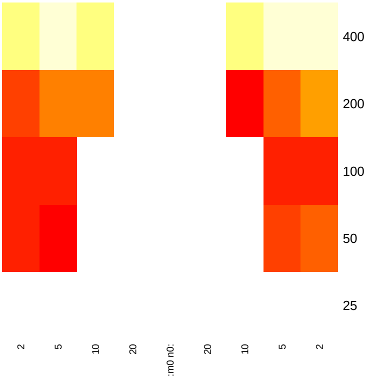
Illustrations. Randomized Ishigami model



Calibration of the tunning parameters

	2	5	10	20	:m0 n0:	20	10	5	2
400	0.22	0.24	0.23	NA	NA	NA	0.23	0.25	0.25
200	0.10	0.12	0.13	NA	NA	NA	0.08	0.11	0.14
100	0.08	0.09	NA	NA	NA	NA	NA	0.08	0.09
50	0.09	0.06	NA	NA	NA	NA	NA	0.10	0.11
25	NA	NA	NA	NA	NA	NA	NA	NA	NA

Table: MRE for various calibrations: $K/(p+1) = 50, 100, \ldots$ and $m_0 = 2, 5, \ldots$ The greatest values depend on K and hence the values for n_0 have been given instead. For instance, for $K/(p+1) = 200 = m_0 n_0$, the available MREs are for $m_0 = 2, 5, 10, 20, 40, 100$.


400
200
100
50
25
2
5
10
20
:m0 n0:
20
10
5
2

Conclusion

We have:

- proposed a formalism of sensitivity analysis for stochastic models;
- emphasized differences between two kinds of approaches;
- proposed a computable optimality criterion.

Future work:

- aggregate sensitivity estimators of steps 1 and 2
- consider higher-order Sobol indices
- extend to non-scalar inputs or outputs