A tradeoff between explorations and repetitions in global sensitivity analysis for stochastic computer models

Gildas Mazo

INRA

UQSay#05, CentraleSupelec Paris-Saclay, October 31, 2019

Sensitivity analysis

Sensitivity analysis permits to exhibit the important factors of a model

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

which represents a phenomenon of interest.

If $X_j = x_j$ were fixed, $\operatorname{Var} Y$ would be reduced by at least $S_j\%$, where S_j, called **the Sobol index** of X_j, is given by

$$S_j = \frac{\operatorname{Var} \mathrm{E}(Y|X_j)}{\operatorname{Var} Y} = \frac{\mathrm{E}(\mathrm{E}(Y|X_j))^2 - (\mathrm{E}\, Y)^2}{\mathrm{E}\, Y^2 - (\mathrm{E}\, Y)^2}.$$

The Sobol decomposition

Sobol (1993) proved that any multidimensional function f can be decomposed uniquely as

$$\begin{aligned} f(x_1, \ldots, x_p) = & f_0 + f_1(x_1) + \cdots + f_p(x_p) \\ & + f_{1,2}(x_1, x_2) + \cdots + f_{p-1,p}(x_{p-1}, x_p) \\ & \vdots \\ & + f_{1,\ldots,p}(x_1, \ldots, x_p). \end{aligned}$$

where

$$\int f_{i_1,\ldots,i_k}(x_{i_1}, \ldots, x_{i_k}) \mathrm{d}x_{i_l} = 0, \quad 1 \leq l \leq k.$$

He gave a statistical interpretation.

Statistical interpretation

Let $X_1, \ldots, X_p$ be independent random variables. Then

$$f(X_1, \ldots, X_p) = f_0 + f_1(X_1) + \cdots + f_{12\ldots n}(X_1, \ldots, X_n),$$

where

$$\mathrm{E}(f_{i_1\ldots i_k}(X_{i_1}, \ldots, X_{i_k}) | X_{i_1}, \ldots, X_{i_{l-1}}, X_{i_{l+1}}, \ldots, X_{i_k}) = 0, \quad 1 \leq l \leq k.$$

In particular,

- the expectation of each of the terms is zero
- the covariance of each pair of terms is zero

and hence

$$\mathrm{Var}\, f(X_1, \ldots, X_p) = \mathrm{Var}\, f_1(X_1) + \cdots + \mathrm{Var}\, f_p(X_p) + \ldots$$

Sobol's proof gives $f_j(X_j) = \mathrm{E}(f(X_1, \ldots, X_p) | X_j) - f_0$.

Monte-Carlo sampling

Let $X = (X_1, \ldots, X_p) \sim P$.

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**
    run the computer model at $\widetilde{X}^{(i)}_{-A}$ to get an output $Y^{(i)}_A$
  end for
end for

Notation: $\widetilde{X}^{(i)}_{\{j\}}$ is $\widetilde{X}^{(i)}$ but the jth component; $\widetilde{X}^{(i)}_{\{1,\ldots,p\}} = X^{(i)}$.

We need $n(p+1)$ runs of the model.

Estimators

The estimator of S_j is given by

$$\widehat{S}_{j;n} = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_0^{(i)} Y_j^{(i)} - \left(\frac{1}{n}\sum_{i=1}^{n} Y_0^{(i)}\right)^2}{\frac{1}{n}\sum_{i=1}^{n} Y_0^{(i)2} - \left(\frac{1}{n}\sum_{i=1}^{n} Y_0^{(i)}\right)^2}$$

We have $\sqrt{n}(\widehat{S}_{j;n} - S_j) \xrightarrow{d} N(0, \sigma^2)$ for some $\sigma > 0$.

Problem

How well can we estimate Sobol-inspired sensitivity indices in **stochastic models**, that is, models of the form

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

where $X = (X_1, \ldots, X_p)$ is the inputs vector and Z represents some randomness intrinsic to the model?

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

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

Definitions do not matter as far as only ranking is of concern

We expand the indices as

$$S'_j = \frac{\overbrace{\mathrm{E}\,\mathrm{E}[f(X,Z)|X]\,\mathrm{E}[f(\widetilde{X}_{-j},Z)|\widetilde{X}_{-j}]}^{D_j} - (\mathrm{E}\,\mathrm{E}[f(X,Z)|X])^2}{\mathrm{E}\,\mathrm{E}[f(X,Z)^2|X] - (\mathrm{E}\,\mathrm{E}[f(X,Z)|X])^2}$$

and

$$S''_j = \frac{\overbrace{\mathrm{E}\,\mathrm{E}[f(X,Z)|X]\,\mathrm{E}[f(\widetilde{X}_{-j},Z)|\widetilde{X}_{-j}]}^{D_j} - (\mathrm{E}\,\mathrm{E}[f(X,Z)|X])^2}{\mathrm{E}\,\mathrm{E}[f(X,Z)|X]^2 - (\mathrm{E}\,\mathrm{E}[f(X,Z)|X])^2},$$

Notice that $S'_i < S'_j$ if and only if $S''_i < S''_j$.

Caution

Let $Y = aX_1 + cX_2Z$, 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

Monte-Carlo sampling

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 $\widetilde{X}^{(i)}_{-A}$ to get an output $Y_A^{(i,k)}$
    end for
  end for
end for

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

Estimators

The estimators of the indices of the first and of the second kind are given by

$$\widehat{S}'_{j;n,m} = \frac{\overbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{k=1}^{m}Y_0^{(i,k)}\frac{1}{m}\sum_{k'=1}^{m}Y_j^{(i,k')}}^{\widehat{D}_{j;n,m}} - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{k=1}^{m}Y_0^{(i,k)}\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{k=1}^{m}Y_0^{(i,k)2} - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{k=1}^{m}Y_0^{(i,k)}\right)^2}$$

and

$$\widehat{S}''_{j;n,m} = \frac{\overbrace{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{k=1}^{m}Y_0^{(i,k)}\frac{1}{m}\sum_{k'=1}^{m}Y_j^{(i,k')}}^{\widehat{D}_{j;n,m}} - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{k=1}^{m}Y_0^{(i,k)}\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{m}\sum_{k=1}^{m}Y_0^{(i,k)}\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{m}\sum_{k=1}^{m}Y_0^{(i,k)}\right)^2}.$$

Asymptotic properties

We write $\mathbf{S}' = (S'_1, \ldots, S'_p)$ and likewise for $\mathbf{S}''$, $\widehat{\mathbf{S}}'_{n,m}$ and $\widehat{\mathbf{S}}''_{n,m}$.

Theorem

Assume that for all x and all z, $f(x,z) \le F(x)$ for some F with $\mathrm{E}\, F(X)^8 < \infty$. Let $n \to \infty$. Then

$$\sqrt{n}\left(\begin{array}{c}\widehat{\mathbf{S}}'_{n,m} - \mathbf{S}' \\ \widehat{\mathbf{S}}''_{n,m} - \mathbf{S}''\left[1 - \frac{\mathrm{E}\,\mathrm{Var}[f(X,Z)|X]}{\mathrm{E}\,\mathrm{Var}[f(X,Z)|X] + m\,\mathrm{Var}\,\mathrm{E}[f(X,Z)|X]}\right]\end{array}\right) \xrightarrow{d} N(0, \Xi_m),$$

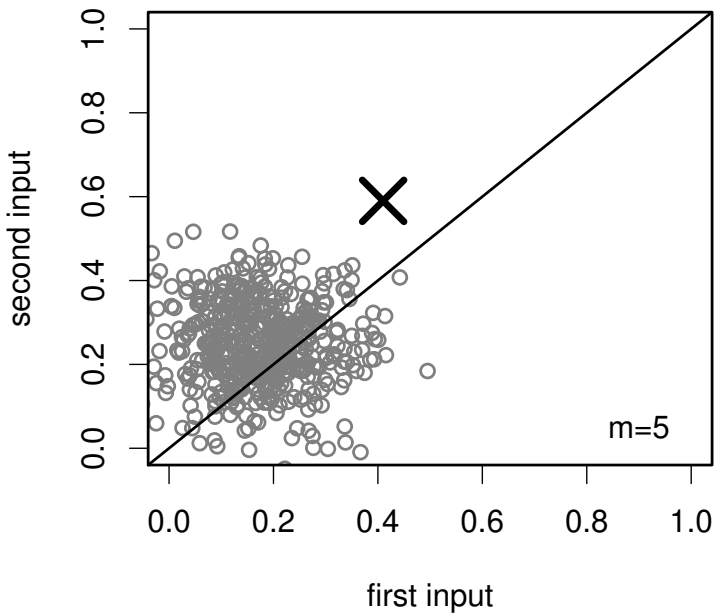
for some nonnegative matrix Ξ_m of size $2p \times 2p$. If $m = m_n \to \infty$ as $n \to \infty$, then, elementwise, $\Xi_m \to \Xi$ for some Ξ. Moreover, the convergence in distribution above still holds with Ξ in place of Ξ_m.

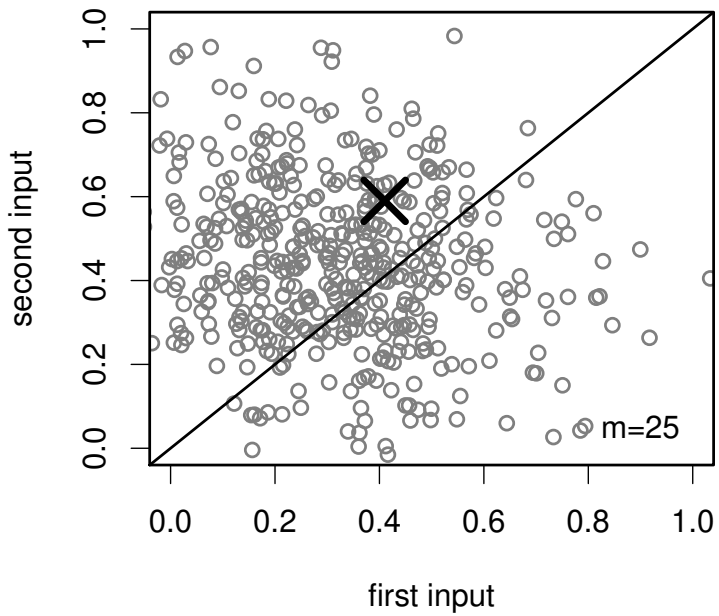
Corollary

If $\sqrt{n}/m \to 0$, then $\sqrt{n}([\widehat{\mathbf{S}}'_{n,m} - \mathbf{S}']^\top, [\widehat{\mathbf{S}}''_{n,m} - \mathbf{S}'']^\top) \to N(0, \Xi)$.

Finite sample sizes

The couple (m, n) controls the bias-variance tradeoff.





What is best?

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

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

The continuous number of repetitions

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

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

and is called the **continuous number of repetitions**.

Relation between the optimal number of repetitions and the continuous number of repetitions

Only some couples (m, n) are possible. For instance with $T = (p+1)mn = (2+1) \times 100 = 300$:

m	1	2	4	5	10	20	25	50	100
n	100	50	25	20	10	5	4	2	1

Let $m^\dagger$ be the optimal number of repetitions. We have

(i) $m^\dagger = 1$ if $m^* \leq 1$

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

(iii) If $1 < m^* < T/(p+1)$ then

$$m^\dagger = \begin{cases} \ulcorner m \urcorner^* & \text{if } \llcorner m \lrcorner^* \ulcorner m \urcorner^* \leq m^{*2} \\ \llcorner m \lrcorner^* & \text{if } \llcorner m \lrcorner^* \ulcorner m \urcorner^* \geq m^{*2}. \end{cases}$$

Estimation of the continuous number of repetitions

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

$$\sqrt{n}\left(\widehat{m}^*_{n,m} - \left[m^* + \frac{C + o(1)}{m}\right]\right) \to N(0, \sigma^2),$$

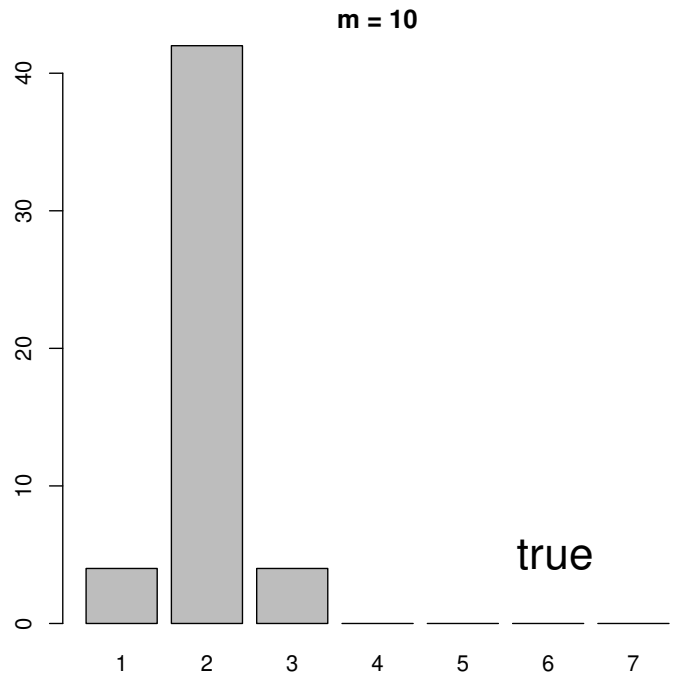
for some constant C and variance σ^2.

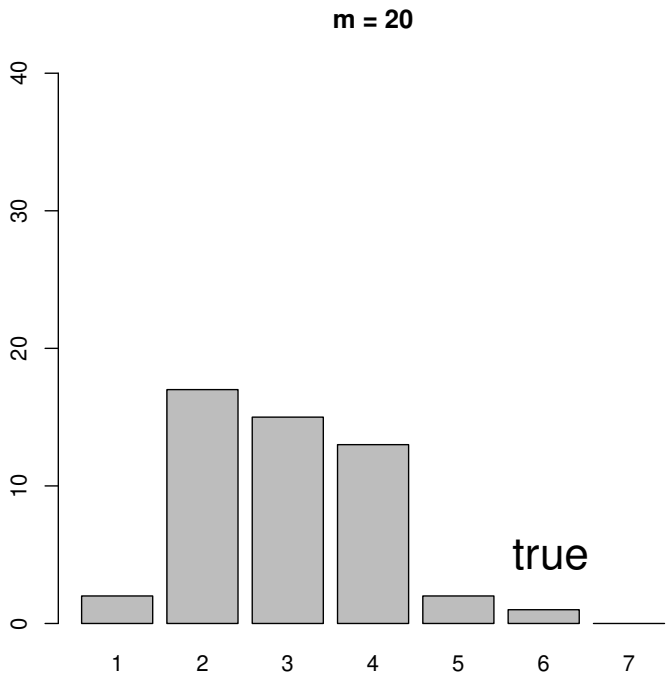
Corollary

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

Finite sample sizes

The couple (m, n) again controls the bias-variance tradeoff.





Here $T = (2 + 1) \times 100$.

Estimation of sensitivity indices by exploiting the continuous number of repetitions

Choose integers (K, m_0, n_0) such that $m_0 n_0 (p+1) = K < T$.

1. Do a Monte-Carlo experiment to get an estimate $\widehat{m}^{\dagger}_{m_0,n_0}$ of $m^{\dagger}$. If $K = 0$, take m_0.
2. With the remaining budget $T - K$, do a Monte-Carlo experiment with number of repetitions $\widehat{m}^{\dagger}_{m_0,n_0}$ to estimate the sensitivity indices.

An oracle property

Let $\widehat{E}_{m_0,n_0}$ be the excess of variance incurred by our ignorance:

$$\widehat{E}_{m_0,n_0} = \frac{\frac{1}{T-K} v(\widehat{m}^*_{m_0,n_0}) - \frac{1}{T} v(m^*)}{\frac{1}{T} v(m^*)}.$$

Theorem

Take $m_0 = T^{1/5}(p+1)^{-1/3}$ and $n_0 = T^{2/5}(p+1)^{-2/3}$. Then there are some constants C and $\sigma > 0$ such that

$$T^{2/5}\widehat{E}_{m_0,n_0} \xrightarrow{d} \frac{v''(m^*)(p+1)^{2/3}(C+\sigma W)^2}{2v(m^*)},$$

where $W \sim N(0,1)$. Moreover, if $|C| > 0$, then the rate $T^{2/5}$ is optimal.

Numerical tests

The models are

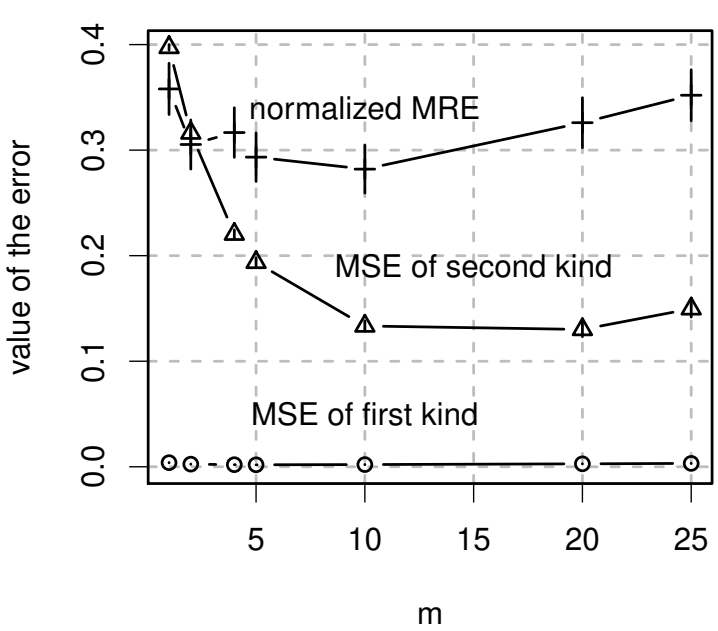
$$Y = X_1 + 1.2X_2 + \sigma Z, \quad X_1, X_2, Z \sim N(0,1), \sigma \in \{4, 0.9\}.$$

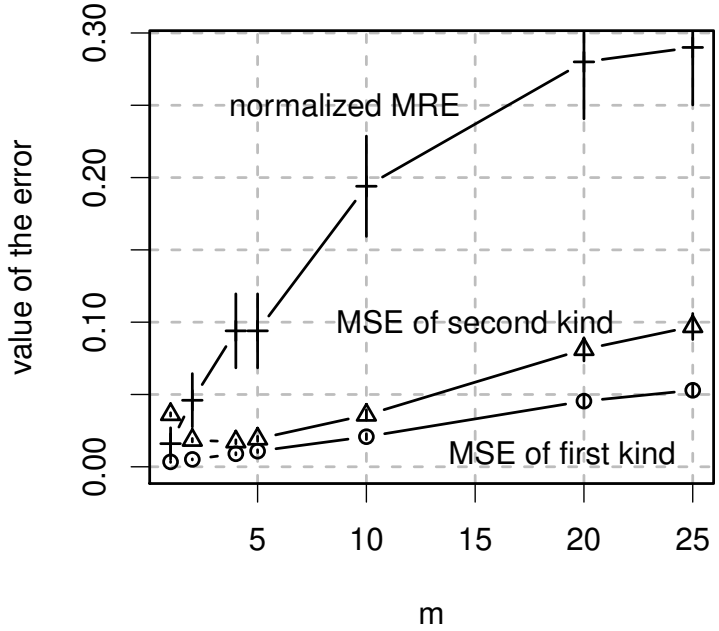
We set $T = (p+1)mn = (2+1) \times 500 = 1500$.

The budget can be decomposed as

1	2	4	5	10	20	25	50	100	125	250	500
500	250	125	100	50	25	20	10	5	4	2	1

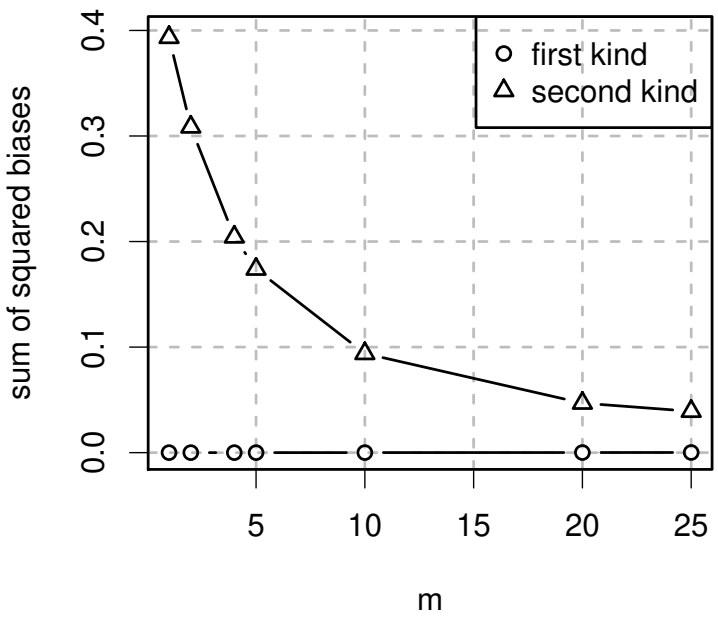
MREs and MSEs



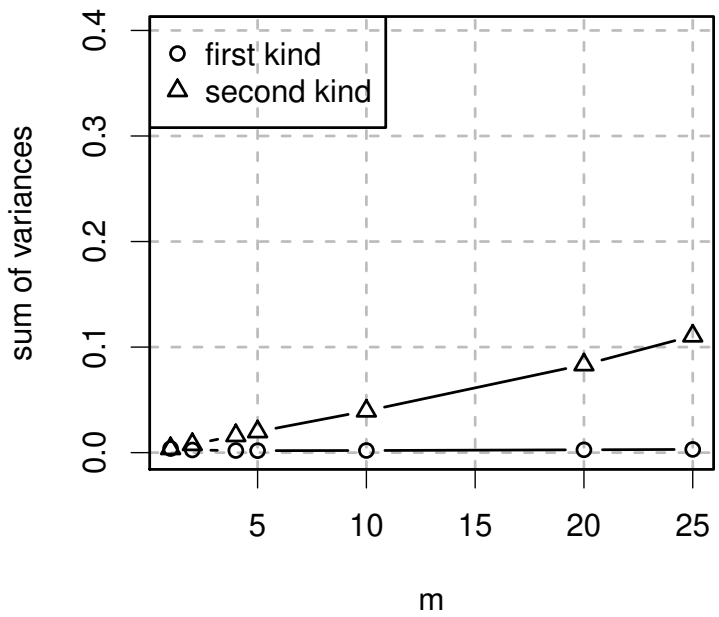


$\sigma = 4$

squared bias

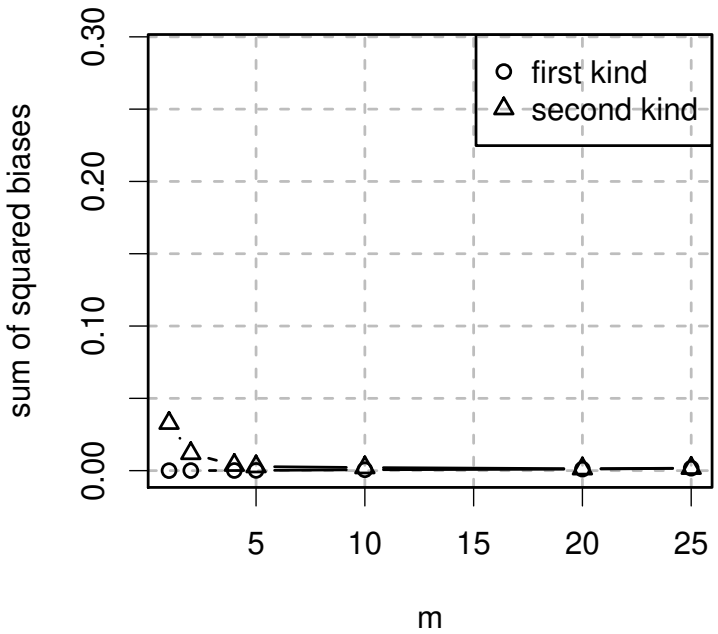


variance

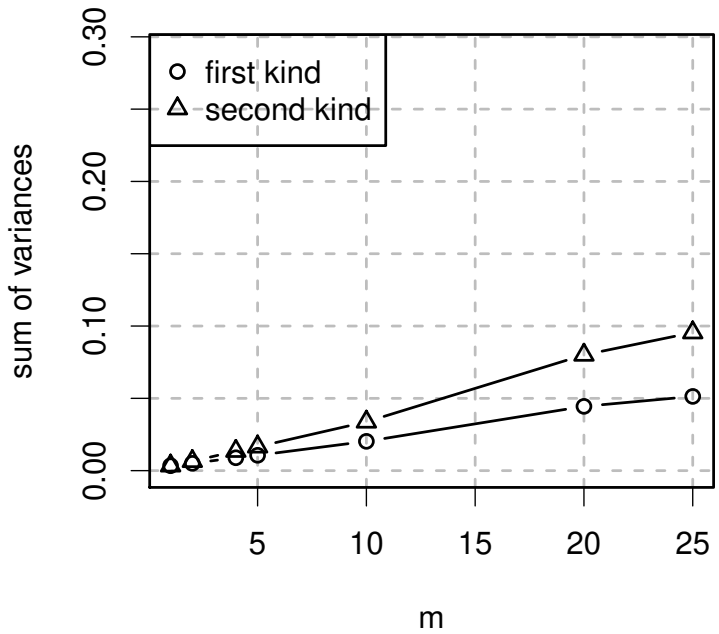


$\sigma = 0.9$

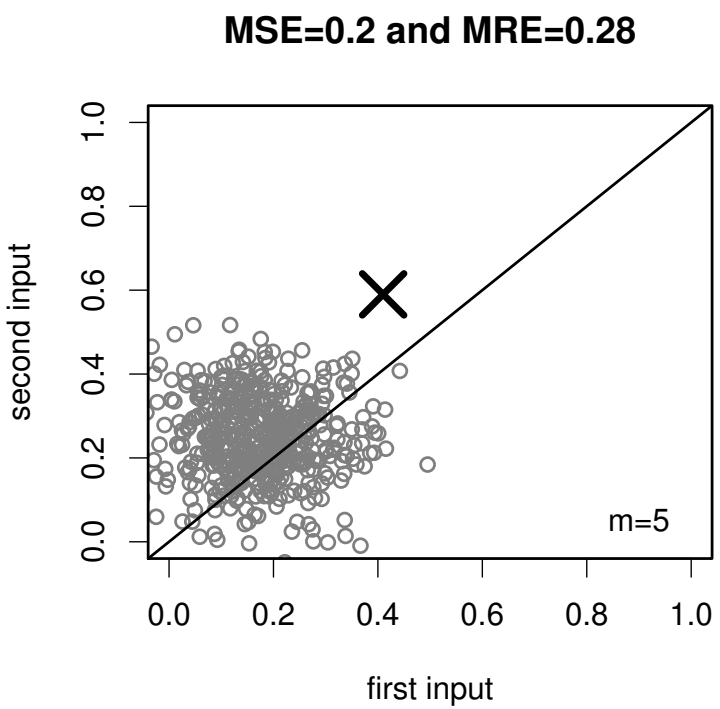
squared bias

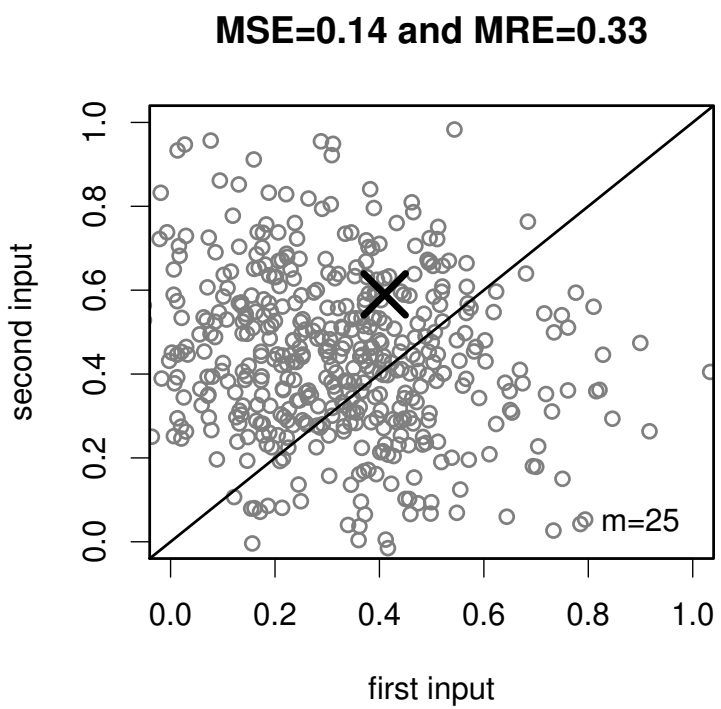


variance



Answer to the quiz





What is the performance of the two-stage procedure?

$\sigma = 4$

	m_0				n_0			
$K/3$	2	5	10	20	20	10	5	2
400	0.43	0.42	0.42	-	-	0.42	0.39	0.40
200	0.38	0.39	0.37	-	-	0.35	0.35	0.34
100	0.36	0.37	-	-	-	-	0.32	0.30
50	0.39	0.33	-	-	-	-	0.33	0.31

$\sigma = 0.9$

	m_0				n_0			
$K/3$	2	5	10	20	20	10	5	2
400	0.18	0.15	0.17	-	-	0.16	0.18	0.20
200	0.05	0.04	0.04	-	-	0.06	0.05	0.07
100	0.02	0.04	-	-	-	-	0.04	0.04
50	0.03	0.02	-	-	-	-	0.02	0.04

Sensitivity analysis of a SIR model

A closed population of size N is followed at each time event, where an infection or a recovery occurs:

$$N = \underbrace{S_i}_{\text{susceptible}} + \underbrace{I_i}_{\text{infectious}} + \underbrace{R_i}_{\text{recovered}}, \quad i = 1, 2, \ldots,$$

The time between two consecutive events i and $i-1$ is an exponential random variable depending on parameters R_0, τ, I_{i-1} and S_{i-1}.

Since I_0 and S_0 are assumed to be known and fixed, **it remains two parameters:** $X_1 = R_0$ **and** $X_2 = \tau$.

Sensitivity analysis of a SIR model

Require: R_0, τ, N, S^0, I^0

$i = 0$

while $S^i > 0$ and $I^i > 0$ **do**

$i = i + 1$

draw $T^i \sim \text{Exp}(\text{mean} = \tau/[R_0 S^{i-1} I^{i-1}/N + I^{i-1}])$

draw $u \sim \text{Unif}(0, 1)$

if $u \leq [R_0 S^{i-1} I^{i-1}/N]/[R_0 S^{i-1} I^{i-1}/N + I^{i-1}]$ **then**

$I^i = I^{i-1} + 1$

$S^i = S^{i-1} - 1$

else

$I^i = I^{i-1} - 1$

$R^i = R^{i-1} + 1$

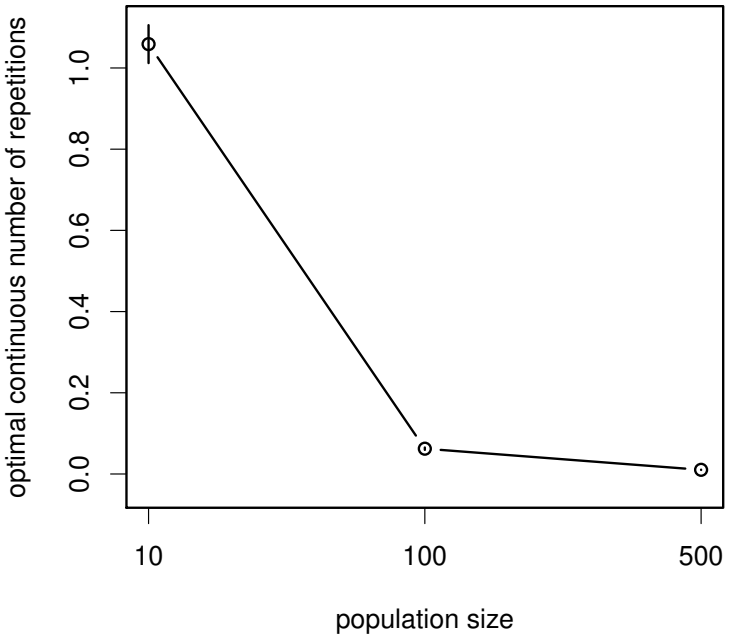
end if

end while

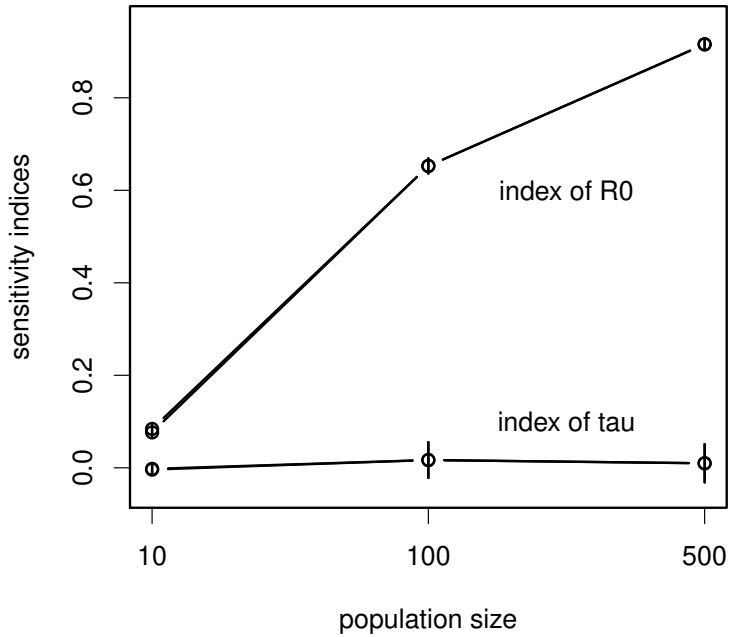
return $S^0 - S^i$

Number of repetitions and sensitivity indices in terms of population size

Number of repetitions



sensitivity indices



Some code

```
> mylinearmodel2 <- function(x){ qexp(pnorm(x[1])) +
+                                1.2 * qunif(pnorm(x[2])) +
+                                .9 * rnorm(1) }
> dosa <- do_SA(total=3000,
+               firststage=list(nb_calls=300,nb_repet=10),
+               model=mylinearmodel2,
+               FoI=1:2)
```

```
$SI
       1stkind   2ndkind
[1,] 0.5151459 0.5151459
[2,] 0.1473089 0.1473089

$optimal_nb_repet
[1] 1

$optimal_nb_explo
[1] 900

$minimizer
[1] 0.1767121
```

Conclusion

- Given a fixed computing budget, the number of repetitions and the number of explorations control the bias-variance tradeoff of the sensitivity indices
- Choosing a good decomposition of the budget may be a delicate matter since optimality depends on the error criterion
- A choice that minimizes a bound of the MRE, an objective criterion to rank the model inputs correctly, exists.
- Further work needs to be done: total sensitivity indices, computation of confidence intervals, combination of the two stages, minimization of the MSE, tests on real-life models, etc.