Un compromis optimal entre explorations et repetitions en analyse de sensibilité

Gildas Mazo

MaIAGE, INRA

Réunion annuelle du réseau ModStatSAP, Paris, Mars 2019

Contexte

L'analyse de sensibilité permet de pointer les facteurs importants d'un modèle

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

qui représente un phénomène d'intérêt.

Si $X_j = x_j$ était fixée à sa vraie valeur, Var Y serait réduit de S_j%.

La quantité S_j est appelé **l'indice de Sobol** de X_j, et vaut

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

Problème

Comment faire une analyse de sensibilité quand le **modèle est stochastique**, c'est à dire

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

où Z est un aléa intrinsèque ?

Que signifie l'analyse de sensibilité dans ce contexte ?

Plan

Indices de Sobol pour modèles stochastiques

Definition

L'indice de Sobol de première espèce est défini comme

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

Definition

L'indice de Sobol de deuxième espèce est défini comme

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

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

Consrtuction des estimateurs

Soit $X = (X_1, \ldots, X_p) \sim P$. Les estimateurs sont construits à partir d'un échantillon Monte-Carlo tiré comme suit :

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

Ensuite, on remplace les espérances par les moyennes empiriques.
Le coût en calcul est de $T = mn(p + 1)$ runs.

Propriétés

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

Theorem

Soit $m \to \infty$ et $n \to \infty$. Alors

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

et

$$\sqrt{n}(\hat{S}''_{j;n,m} - [S''_j + O(1/m)]) \to N(0, \sigma''^2).$$

Corollary

Si $\sqrt{n}/m \to 0$, alors

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

Le nombre optimal de répétitions

Soit $T = mn(p+1)$ le budget de calcul.

Definition and Proposition

Le nombre de répétitions optimal $m^\dagger$ est défini comme l'argument qui minimise

$$\frac{4(p-1)\frac{1}{T}\overbrace{(\zeta_1 m + \zeta_2 + \zeta_3 \frac{1}{m})}^{v(m)}}{\min_{j<j'}(|D_j - D_{j'}|^2)} \geq \mathrm{E} \sum_{j=1}^{p} |\hat{R}_j - R_j|,$$

Supposons pour simplifier que $\sqrt{\zeta_3/\zeta_1}$ est entier.

Theorem

Le nombre optimal de répétitions optimal $m^\dagger$ est donné par $m^\dagger = \sqrt{\zeta_3/\zeta_1}$.

Une procedure qui exploite le rapport bruit/signal du modèle

0. Choisir deux entiers (K, m_0) tels que $m_0 n_0 (p+1) = K < T$.
1. Faire une expérience Monte-Carlo $\mathcal{E}(K, m_0)$ pour avoir une estimation $\hat{m}^{\dagger}_{K,m_0}$ de $m^{\dagger}$. Si $K = 0$, prendre m_0.
2. Faire une expérience Monte-Carlo $\mathcal{E}(T-K, \hat{m}^{\dagger}_{K,m_0})$ pour estimer les indices de sensibilité.

Une propriété oracle

Soit E_{K,m_0} l'excès de variance induit par la méconnaissance de $m^\dagger$, càd

$$E_{K,m_0} = \frac{\frac{1}{T-K} v(\hat{m}^\dagger_{K,m_0}) - \frac{1}{T} v(m^\dagger)}{\frac{1}{T} v(m^\dagger)}$$

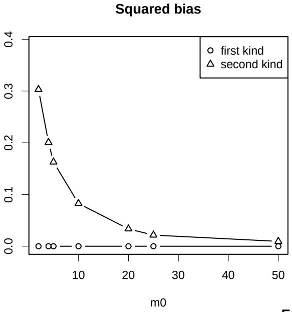
Soit $T \to \infty$.

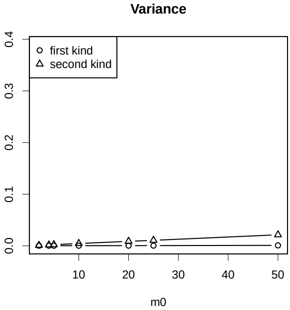
Theorem

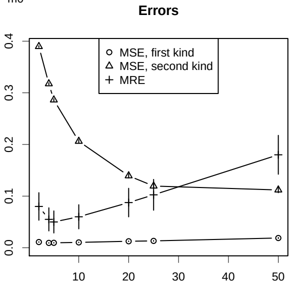
Supposons que $K_T / T \to 0$. Alors $\hat{E}_{K,m_0} \xrightarrow{P} 0$. De plus,

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

Illustrations. Modèle linéaire, bruit élevé

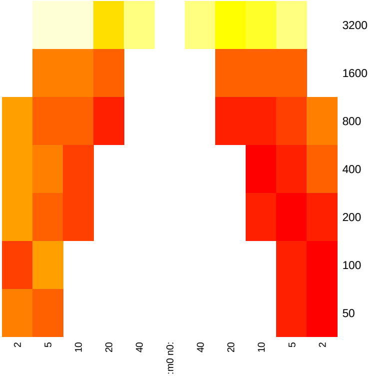




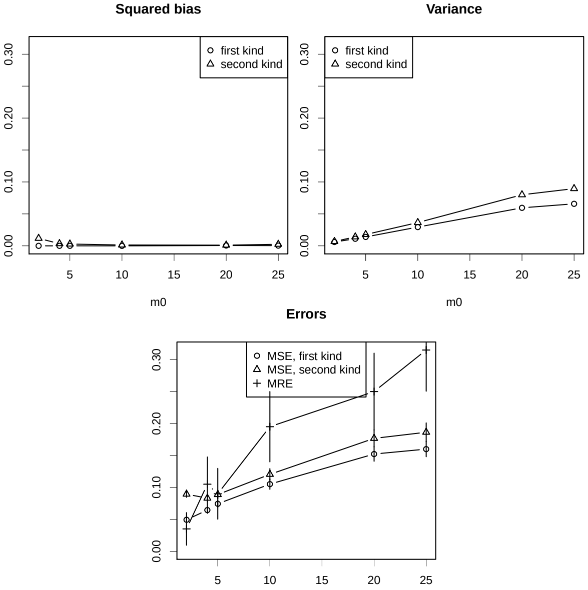


	2	5	10	20	40	:m0 n0 :	40	20	10	5	2
3200	NA	0.25	0.27	0.18	0.24	NA	0.24	0.20	0.23	0.25	NA
1600	NA	0.13	0.14	0.10	NA	NA	NA	0.12	0.11	0.11	NA
800	0.15	0.10	0.11	0.07	NA	NA	NA	0.08	0.07	0.10	0.13
400	0.14	0.12	0.10	NA	NA	NA	NA	NA	0.05	0.06	0.10
200	0.16	0.11	0.10	NA	NA	NA	NA	NA	0.07	0.04	0.08
100	0.10	0.14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

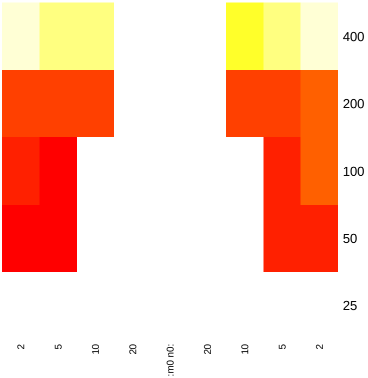
Illustrations. Modèle linéaire, bruit faible



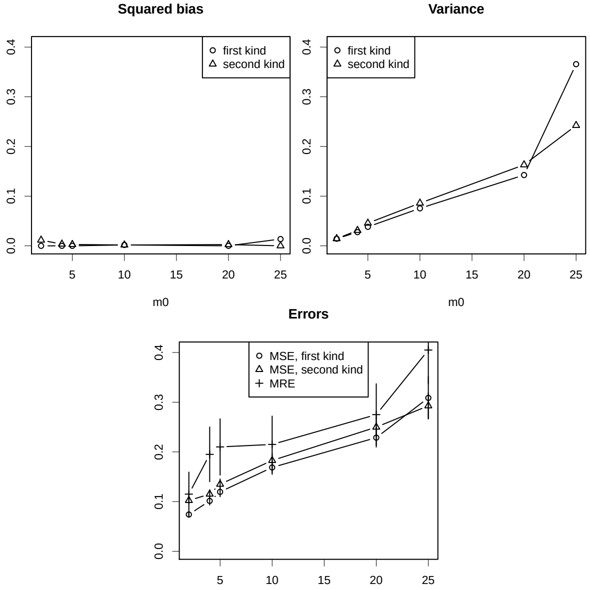
Calibration des paramètres de tunning

	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

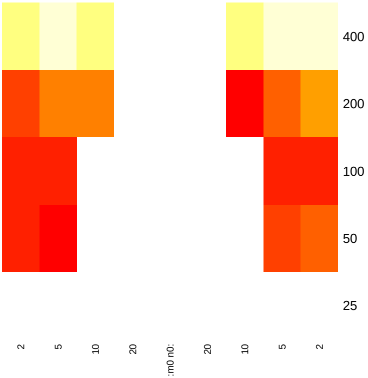
Illustrations. Modèle d'Ishigami randomisé



Calibration des paramètres de tunning

	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

Travail accompli :

- On a formalisé l'analyse de sensibilité pour modèles stochastiques
- On a dégagé les différences statistiques entre les approches considérées
- On a proposé un critère d'optimalité calculable

Perspectives :

- Prise en compte la nature discrète du nombre de répétition optimal
- Extensions