Analyse de sensibilité

G. Perrin

guillaume.perrin@univ-eiffel.fr

Année 2022-2023

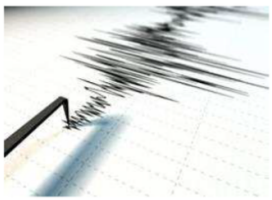
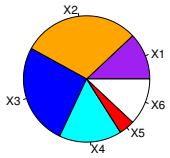
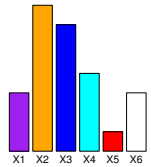





Schéma général

Etape C'
Hiérarchisation des paramètres



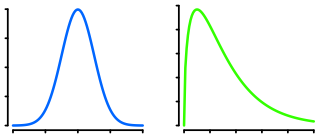


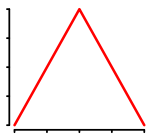
Etape B
Quantification des source d'incertitudes

Etape A
Spécification du problème

Etape C
Propagation des incertitudes

Modélisation des entrées par des distributions

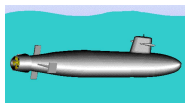




Système réel

Variables d'entrées incertaines $X = (X_1, ..., X_d)$

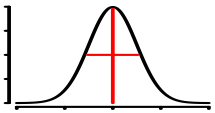
Modèle numérique g



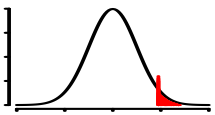
Variables d'intérêt $Y = g(X)$

Quantité d'intérêt :

Moments



Proba défaillance



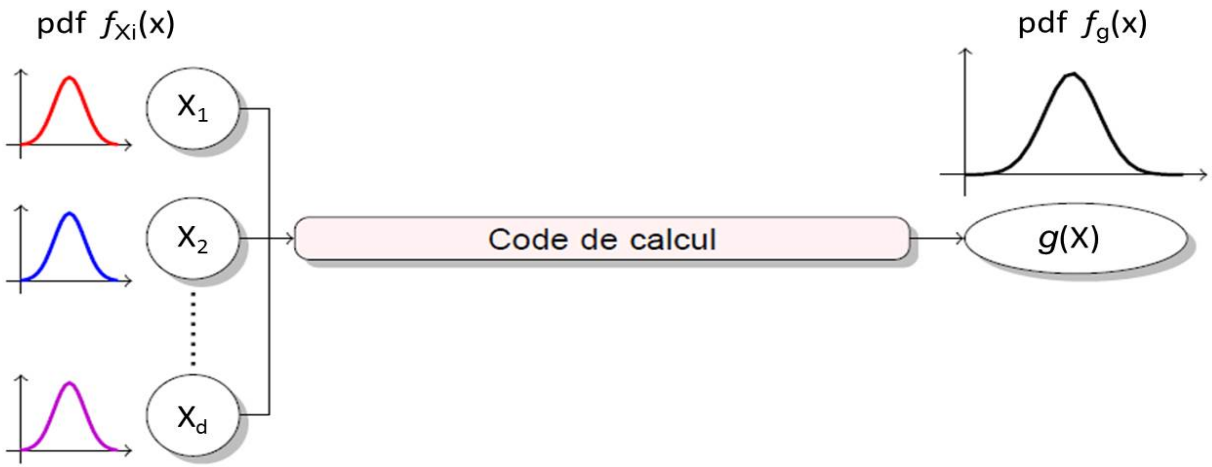
Notations et hypothèses

Notations :

- Paramètre d'entrée : $(X_1, \ldots, X_d) \in \mathbb{R}^d$.
- Modèle de calcul : $g(.)$ (considéré comme coûteux).
- Variable de sortie d'intéret $Y = g(X_1, \ldots, X_d)$.

Hypotheses (valables pour tout le cours) :

- Paramètres d'entrée indépendants.
- Quantité d'intérêt : variabilité de la sortie Y, supposée scalaire.

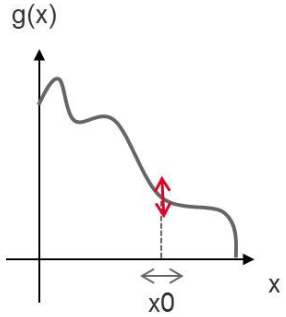


Généralités

Objectif de l'analyse de sensibilité

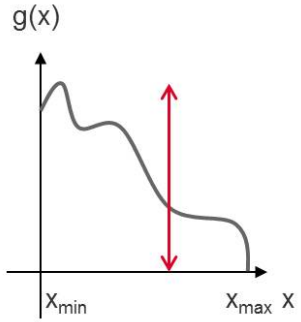
Déterminer les paramètres d'entrée contribuant le plus à la variabilité de la sortie/réponse du modèle

AS locale

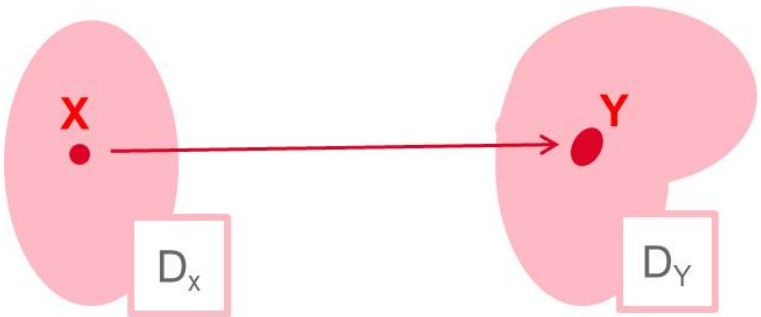


Variation de $g(X)$ autour de X^0

AS globale



Variation globale de $g(X)$ quand X varie dans son domaine incertain

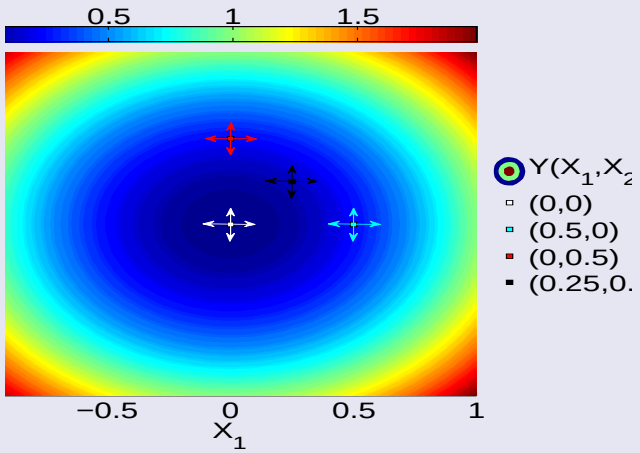


Généralités

Illustration "intuitive" de la distinction entre analyses locale et globale.

On définit : $Y = X_1^2 + X_2^2$, $X_1, X_2 \sim \mathcal{U}(-1, 1)$, X_1 et X_2 indépendants.

Analyse "locale" $\leftrightarrow$ "point-dépendant"



- En se basant sur les dérivées partielles en un point quelconque, les résultats de l'analyse peuvent varier du tout au tout.
- Par une approche par différences finies, les résultats peuvent dépendre du pas de différentiation.

Analyse "globale" $\leftrightarrow$ "distribution-dépendant"

$$\mathrm{var}(Y) = \mathrm{var}(X_1^2) + \mathrm{var}(X_2^2) = \mathrm{var}(Y)/2 + \mathrm{var}(Y)/2.$$

$\Rightarrow$ les paramètres X_1 et X_2 sont **globalement** aussi influents l'un que l'autre sur la dispersion de Y.

Analyse de sensibilité locale

Cumul quadratique :

- Développement de Taylor à l'ordre 1 :

$$g(X) \approx g(X^0) + \sum_{i=1}^{d} \left(\frac{\partial g(X)}{\partial X_i}_{|X=X^0} \right) (X_i - X_i^0)$$

- Retour sur le cumul quadratique. Poids de X_i sur la sortie $Y = g(X)$:

$$\eta_i^2 = \frac{1}{\text{var}(Y)} \left(\frac{\partial g(X)}{\partial X_i}_{|X=X^0} \right)^2 \sigma_{X_i}^2$$

+ Tres simple à mettre en oeuvre (différence finies, différenciation automatique, ...)
− Approche locale. Le passage à l'approche globale ne peut se faire que sous une hypothese de linéarité et d'additivité du modèle

  Ne permet pas de détecter les effets conjoints de plusieurs paramètres.

Deux objectifs et deux types d'ASG

- Analyse qualitative :

  Identifier les paramètres d'entrées non influents sur les variations de la sortie du modèle

  - Réduction de la dimension des paramètres d'entrée incertains : détermination des paramètres que l'on pourra fixer.
  - Construire un modèle simplifié, un métamodèle.

- Analyse quantitative :

  Réduction de l'incertitude de la sortie d'un modèle par hiérarchisation des sources d'incertitude

  - Détermination des paramètres permettant d'obtenir la plus forte réduction (ou une réduction donnée) de l'incertitude de la sortie.
  - Détermination des paramètres les plus influents dans un domaine de variation de la sortie.

  $\Longrightarrow$ Orientation R&D : amélioration connaissance des paramètres d'entrée (campagne de mesures, ...)

Deux objectifs et deux types d'ASG

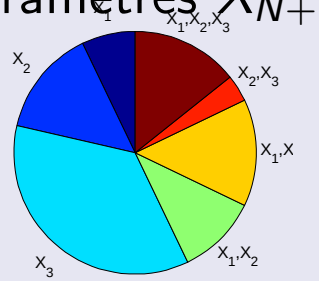
- Soit $Y = g(X_1, \ldots, X_d)$ une quantité d'intérêt dépendant d'un tres grande nombre de paramètres $X_1, \ldots, X_d$, $d \gg 1$.
- Explorer la variabilité de Y revient à explorer $\mathbb{R}^d$...
- Pour rendre cette analyse réaliste, on procede en deux temps :

Etape 1 : analyse qualitative

- Identification de paramètres peu influents $X_1, \ldots, X_N$, $N < d$,
- Réduction de la variabilité d'entrée :
  $Y \approx g(\mathbb{E}(X_1), \ldots, \mathbb{E}(X_N), X_{N+1}, \ldots, X_d)$.

Etape 2 : analyse quantitative

- Décomposition quantitative de la variabilité de Y en fonction des actions individuelles et combinées des paramètres $X_{N+1}, \ldots, X_d$.



Démarche d'une analyse de sensibilité globale

Stratégie de l'analyse de sensibilité globale

Faire varier l'ensemble des paramètre d'entrées du modèle dans son domaine incertain (défini à l'étape B) et analyser les variations des sorties du modèle.

- $\Rightarrow$ une méthode d'exploration de l'espace des paramètres d'entrée (plan d'expériences)
- $\Rightarrow$ une méthode stat. d'analyse des sorties du modèle aux points du PE

- jeu de n réalisations de X $\begin{pmatrix} X_1^{(1)} & \dots & X_d^{(1)} \\ \vdots & \vdots & \vdots \\ X_1^{(n)} & \dots & X_d^{(n)} \end{pmatrix}$
- Simulations correspondantes : $g(X^{(1)}), \dots, g(X^{(n)})$
- Méthode statistique reposant sur des hypotheses du comportement du modèle g.

Criteres de choix d'une méthode d'ASG

- Contraintes :
  - Le temps de calcul du modèle $\rightarrow$ limite du nombre de simulations réalisables.
  - Le nombre et le type de paramètres d'entrée (continus, discrets)
- Régularité du modèle :
  - Propriété (connue ou supposée) du modèle : linéarité, monotonie, continuité,...
  - Plus le modèle est irrégulier et plus il faudra explorer finement l'espace des paramètres pour capturer ces irrégularités.
- Objectif de l'AS Précision :
  - Analyse de sensibilité quantitative ou qualitative.
  - Niveau de précision des résultats de l'AS.

Choix d'une méthode d'ASG

Pour un objectif fixé, le choix d'une méthode doit être le meilleur compromis entre le nombre de paramètres d'entrée, le nombre de simulations possibles et la connaissance *a priori* que l'on a du modèle.

Deux grandes familles de méthodes d'ASG

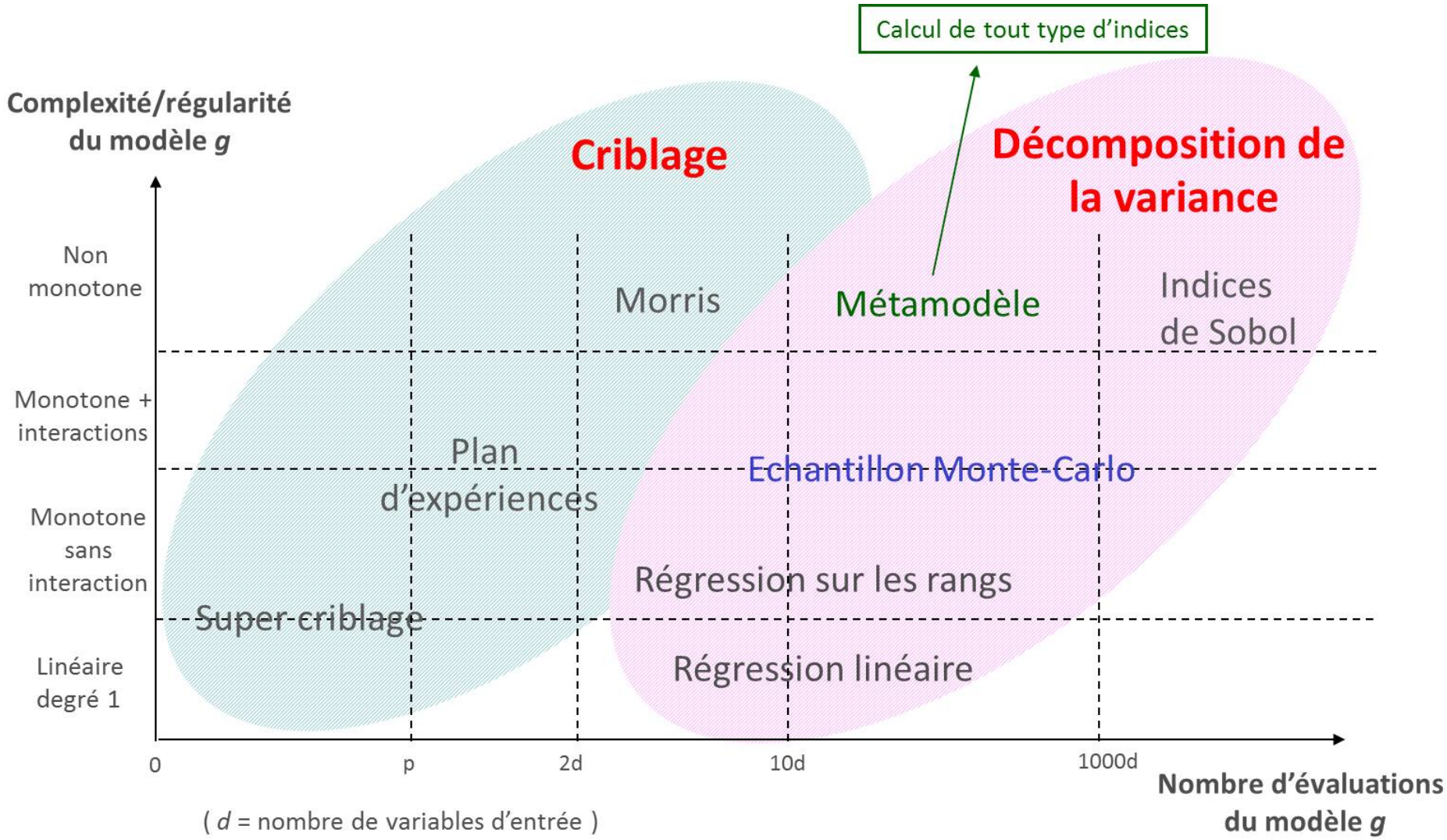
1. Méthodes de criblage (screening)
   - Plan d'expériences classiques
   - Criblage à tres grande dimension
   - Méthode de Morris

   $\longrightarrow$ $n \approx d/2$ à $10d$

2. Méthodes quantitatives basées sur la décomposition de la variance
   - Techniques de regression linéaire (ou monotone)
   - Méthode de Sobol

   $\longrightarrow$ $n \approx 2d$ à $10^4 d$

Deux grandes familles de méthodes d'ASG (schéma issu de [looss 2009])



Deux grandes familles de méthodes d'ASG

1. Méthodes de criblage (screening)
   - Plan d'expériences classiques
   - Criblage à tres grande dimension
   - Méthode de Morris

   $\longrightarrow$ $n \approx d/2$ à $10d$

2. Méthodes quantitatives basées sur la décomposition de la variance
   - Techniques de regression linéaire (ou monotone)
   - Méthode de Sobol

   $\longrightarrow$ $n \approx 2d$ à $10^4 d$

Analyse de sensibilité qualitative

- Nous voulons effectuer l'analyse de sensibilité qualitative de

$$g : X \mapsto g(X) \in \mathbb{R}$$

$$X = (X_1, X_2, \ldots, X_d) \in \mathbb{R}^d.$$

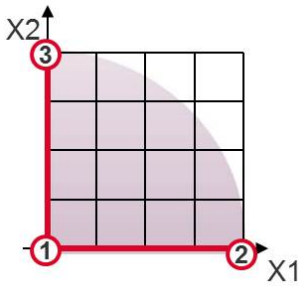
- Objectif (rappel)

  Identifier les paramètres d'entrées non influents sur les variations de la sortie du modèle

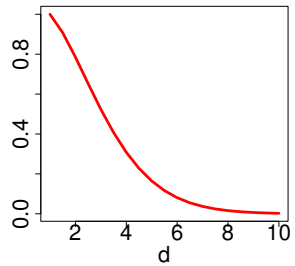
- Modélisation des paramètres d'entrée (étape A) :
  - Pour chaque paramètre d'entrée X_i, on définit précisement un domaine d'étude par un intervalle de variation $[x_{\min}^{(i)}, x_{\max}^{(i)}]$
  - Le domaine d'étude est symbolisé par un hypercube.
  - Les variations sur les paramètres d'entrées impliquent une variation sur la sortie.

Du bon et mauvais usage du plan OAT (One-AT-a-Time)

- Peu coûteux : $n = d + 1$
- Part de l'idée tres répandue que pour analyser les causes d'un phénomene, il faut faire des expériences en ne bougeant qu'un seul paramètre à la fois.
- OAT apporte des informations, potentiellement fausses ne détecte pas les non monotonies, discontinuités, interactions
- Laisse de grandes zones inexplorées dans l'espace des X :



Volume sphere/Volume cube :



Hypothese implicite sur le modèle

Le modèle est linéaire sans interactions pour analyse quantitative.
Le modèle est monotone sans interactions pour criblage.

Plan d'expériences classiques (1/4)

Plan factoriels complets pour modèle linéaire

- On se limite à deux niveaux pour chaque paramètres : $x_{\min} \rightarrow -1$ et $x_{\max} \rightarrow +1$
- Les expériences sont faites aux sommets du domaine de X (hypercube).
- Le plan complet balaye l'intégralité des 2^d expériences à réaliser.

Exp.	X_1	X_2	X_3
1	-1	-1	-1
2	$+1$	-1	-1
3	-1	$+1$	-1
4	$+1$	$+1$	-1
5	-1	-1	$+1$
6	$+1$	-1	$+1$
7	-1	$+1$	$+1$
8	$+1$	$+1$	$+1$

Plan d'expériences classiques (2/4)

Plans factoriels fractionnaires à deux niveaux

- On souhaite réduire le nombre d'essais d'un plan complet.
- Il est possible de diviser le nombre d'essais par une puissance de 2.
  On parle de plan $2^{d-p} \leftrightarrow d$ paramètres mais 2^{d-p} expériences.
- Règle de construction : certains paramètres sont alors nécessairement confondus (ou aliasés) avec des interactions. Exemple de plan 2^{5-2} :

Exp.	X_1	X_2	X_3	$X_4 = X_1X_2$	$X_5 = X_1X_3$
1	-1	-1	-1	$+1$	$+1$
2	$+1$	-1	-1	-1	-1
3	-1	$+1$	-1	-1	$+1$
4	$+1$	$+1$	-1	$+1$	-1
5	-1	-1	$+1$	$+1$	-1
6	$+1$	-1	$+1$	-1	$+1$
7	-1	$+1$	$+1$	-1	-1
8	$+1$	$+1$	$+1$	$+1$	$+1$

Plan d'expériences classiques (3/4)

Plans de résolution III pour criblage :

Hypothese sur le modèle

Le modèle est monotone sans interactions

- Peu coûteux : $n \geq d + 1$ avec $n = 0 \equiv 4$
  $d = 20 \rightarrow n = 24$, $d = 80 \rightarrow n = 84$
- Modèle statistique pour le criblage (additif) : $\widehat{g}(X) = \beta_0 + \sum_{i=1}^{d} \beta_i X_i$
- Interprétation rapide : les coefficients du modèle linéaire β_i représentent les effets principaux des paramètres.
- Plan optimisé pour l'analyse de sensibilité.
- Aucun effet principal n'est confondu avec un autre effet principal $\rightarrow$ Estimation sans biais des effets principaux si hypothese vérifiée.
- Si modèle linéaire sans interactions : analyse quantitative possible des effets.

Plan d'expériences classiques (4/4)

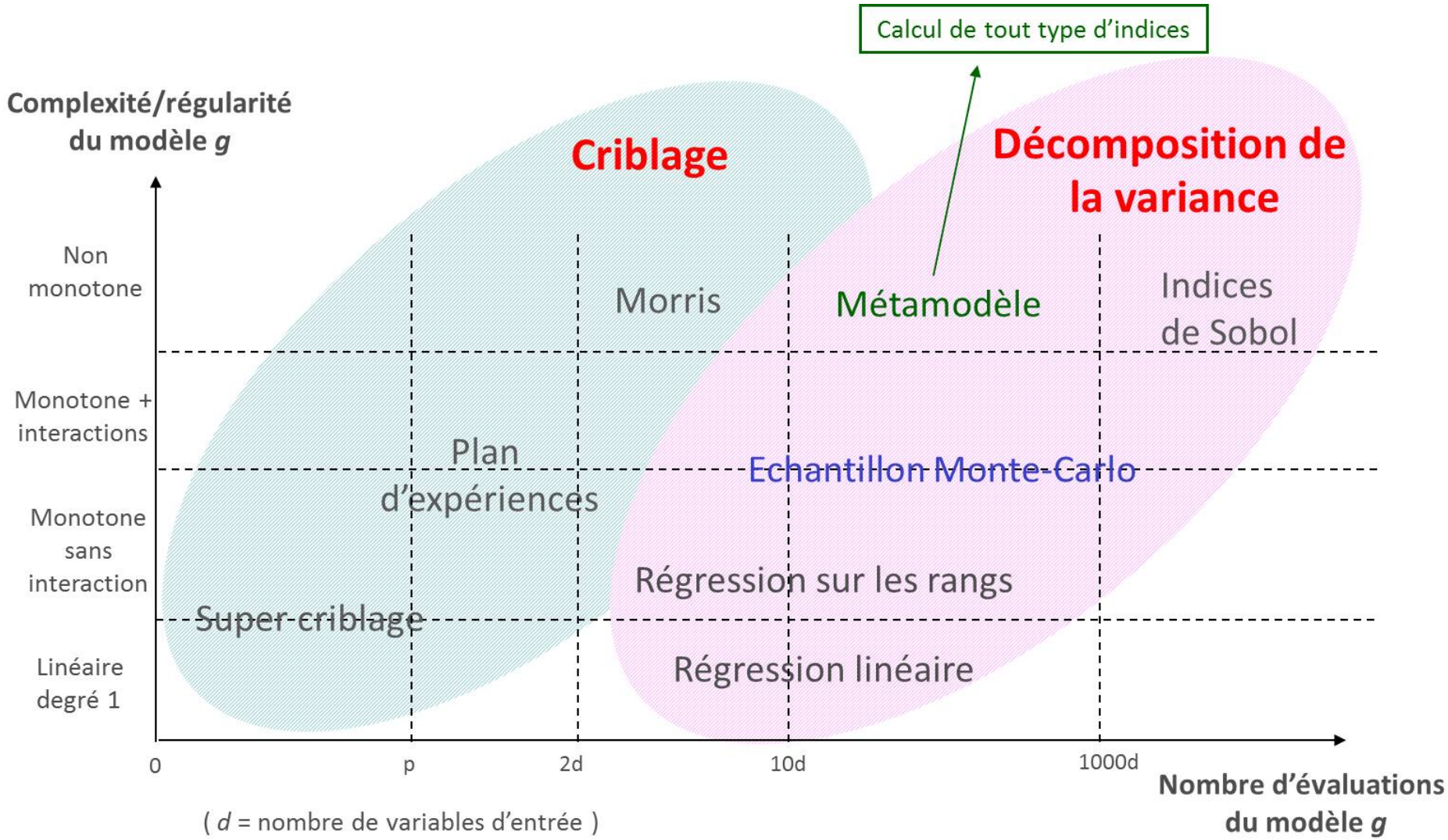
Plans de résolution IV pour criblage :

Hypothese sur le modèle

Le modèle est monotone et les interactions > 2 sont nulles

- Peu coûteux : $2\times$ coût plan résolution III
  $d = 30 \rightarrow n = 64$, $d = 60 \rightarrow n = 136$
- Construction : un plan de résolution III et son opposé.
- Même modèle statistique et même interprétation qu'un plan de résolution III.
- Plan optimisé pour l'analyse de sensibilité.
- Aucun effet principal n'est confondu avec un autre effet principal ou une interaction d'ordre 2 $\rightarrow$ Estimation sans biais des effets principaux même s'il existe des interactions d'ordre 2.
- Possibilité de détecter certaines interactions (hypotheses supplémentaires)

Deux grandes familles de méthodes d'ASG (schéma issu de [looss 2009])



Criblage avec $n < d$ (1/2)

Criblage par groupe :

Hypothese sur le modèle

Le modèle est monotone sans interactions.
Le sens de variation de la sortie pour chaque paramètre est connu.

- Hypothese supplémentaire : le nombre de paramètres influents est petit par rapport au nombre de paramètres d'entrée ($\sim 10\%$)
- Idée : regrouper les d paramètres en G groupes et traiter chaque groupe comme un paramètre individuel (plan de rés. III par ex.)
- Les groupes *inactifs* sont éliminés.
- Les paramètres des groupes *actifs* peuvent être traités soit individuellement soit de nouveau par groupe.
- Le regroupement dépend de la connaissance que l'on a *a priori* sur les paramètres Pas d'effets opposés dans les groupes
- Inconvénient : méthode ne s'appliquant qu'à une sortie à la fois (séquentielle)

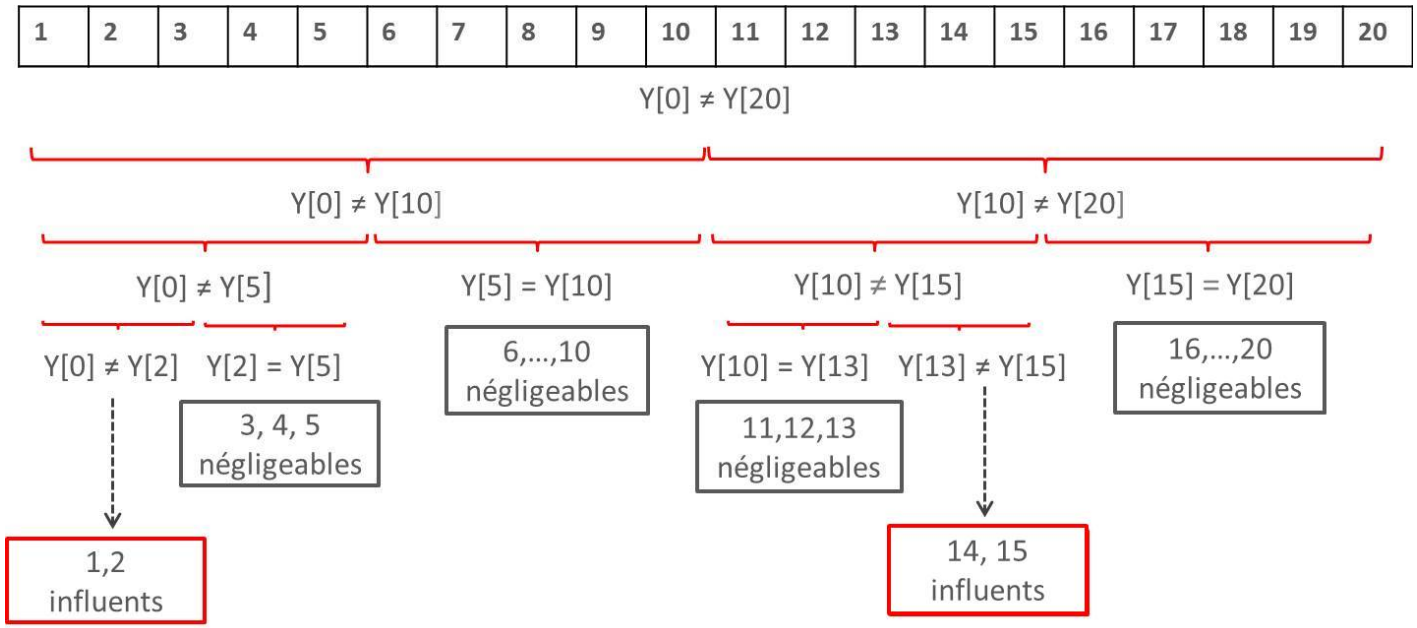
Criblage avec $n < d$ (2/2)

Bifurcations séquentielles :

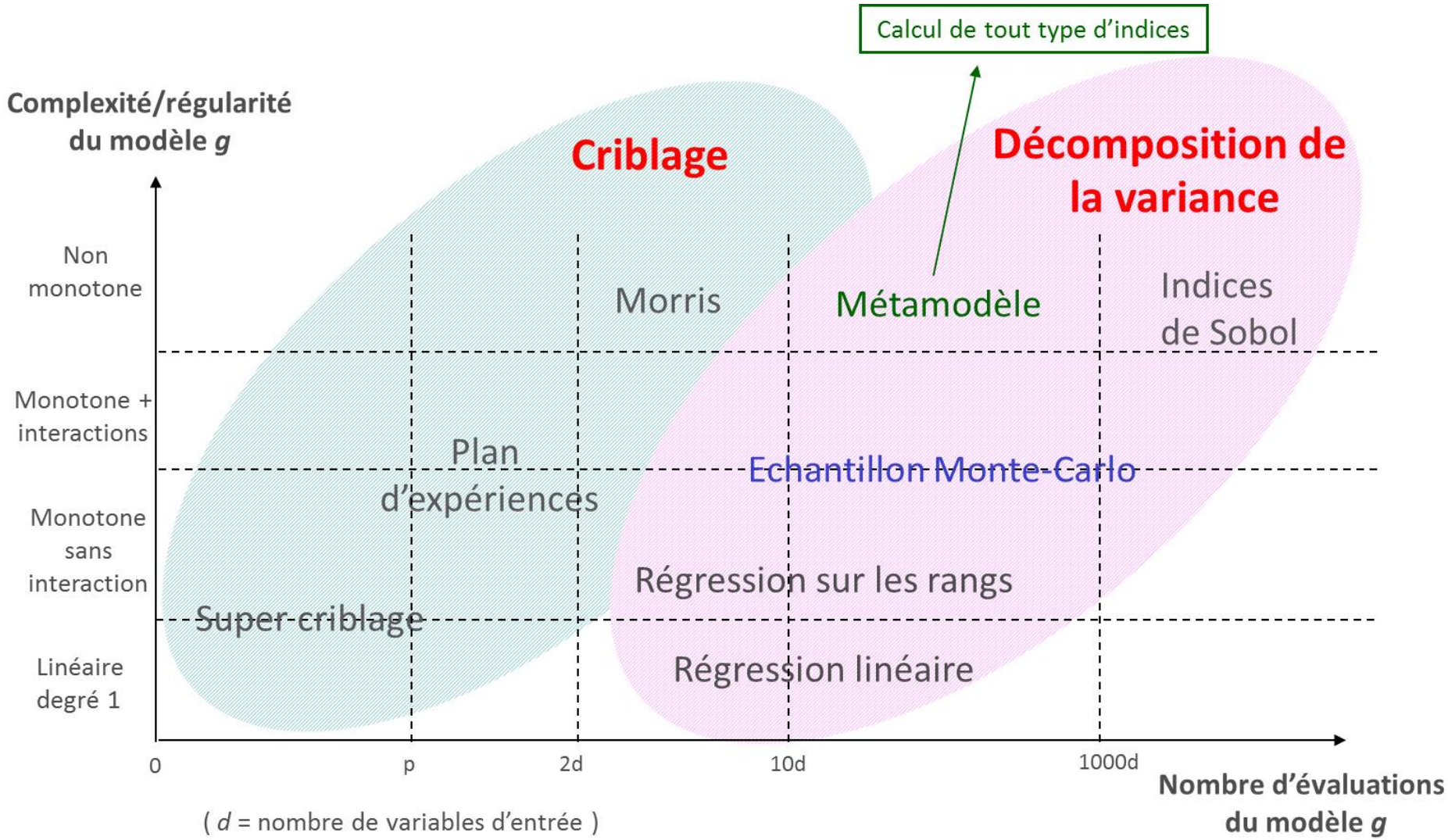
Hypothese sur le modèle

Le modèle est monotone sans interactions.
Le sens de variation de la sortie pour chaque paramètre est connu.

- Méthode dichotomique ($\approx$ Méthode de criblage avec 2 groupes)
  $\mathrm{Y}[j] : \{g(X) : (X_1, \ldots, X_j)$ en $(+)$ et $(X_{j+1}, \ldots, X_d)$ en $(\text{-})\}$



Deux grandes familles de méthodes d'ASG (schéma issu de [looss 2009])



Méthode de Morris (1/3)

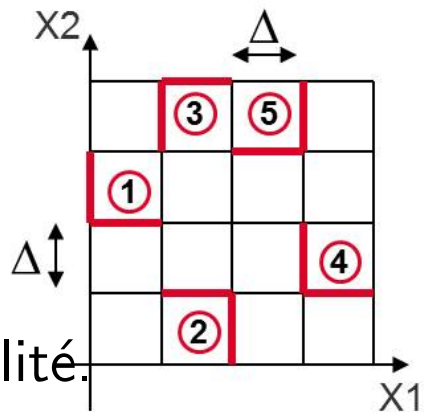
Hypothese sur le modèle

Aucune... Mais mieux vaut une certaine régularité.

- Basée sur une discrétisation du domaine des paramètres d'entrée
- Répétitions de R plans OAT. Coût : $R(d+1)$

  $\rightarrow$ R effets élémentaires pour chaque X_i :

  $$d_i = \pm \frac{g(\dots, X_i \pm \Delta, \dots) - (g(\dots, X_i, \dots)}{\Delta}$$



- La moyenne des $\mathbb{E}(|d_i|)$ est une mesure de la sensibilité.

  Valeur importante $\rightarrow$ effets importants (en moyenne). Modèle sensible aux variations de l'entrée.

- La var. des $\sigma(|d_i|)$ est une mesure des interactions non linéaires

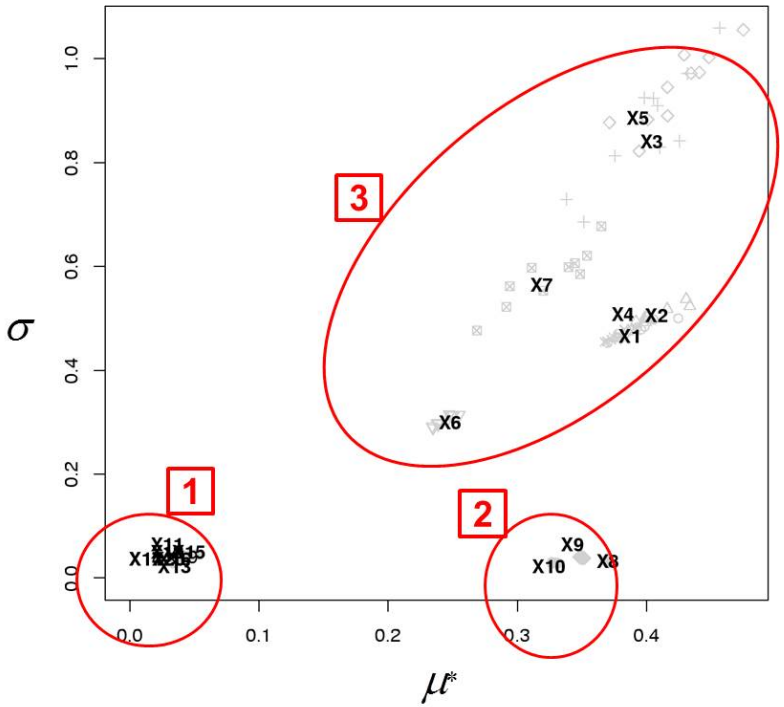
  Valeur importante $\rightarrow$ effets différents les uns des autres. Effet non linéaire ou interactions.

Méthode de Morris (2/3)

Criblage plus informatif :

Permet de classer les paramètres d'entrée en 3 catégories :

1. les variables ayant des effets négligeables ;
2. les variables ayant des effets linéaires *et* sans interactions ;
3. les variables ayant des effets non linéaires *et/ou* des interactions.



Graphe (μ^*, σ)

Fonction non monotone de Morris.
20 paramètres, 210 simulations.

Méthode de Morris (3/3)

Choix non néligeable de variables internes à la méthode :

- Δ le pas de discrétisation de l'espace

  $\rightarrow$ détermine le caractere global ou local des effets estimés. Δ petit : variations locales.

- R le nombre de répétitions.

  $\rightarrow$ Plus Δ sera petit, plus R devra être grand pour explorer correctement le domaine des paramètres d'entrée.

- En pratique, c'est le nombre de simulations n qui est fixé.
  $\rightarrow$ nombre de répétitions R possibles : $\lfloor n/(d+1) \rfloor$

  R petit : effets globaux

  R grand : effets plus locaux

- Usage courant : $R \in [5, 10]$

  (d'où la certaine régularité attendue du modèle...)

Deux grandes familles de méthodes d'ASG

1. Méthodes de criblage (screening)
   - Plan d'expériences classiques
   - Criblage à tres grande dimension
   - Méthode de Morris

   $\longrightarrow$ $n \approx d/2$ à $10d$

2. Méthodes quantitatives basées sur la décomposition de la variance
   - Techniques de regression linéaire (ou monotone)
   - Méthode de Sobol

   $\longrightarrow$ $n \approx 2d$ à $10^4 d$

Analyse de sensibilité quantitative

- Nous voulons effectuer l'analyse de sensibilité quantitative de

$$g : X \mapsto g(X) \in \mathbb{R}$$

$$X = (X_1, X_2, \ldots, X_d) \in \mathbb{R}^d.$$

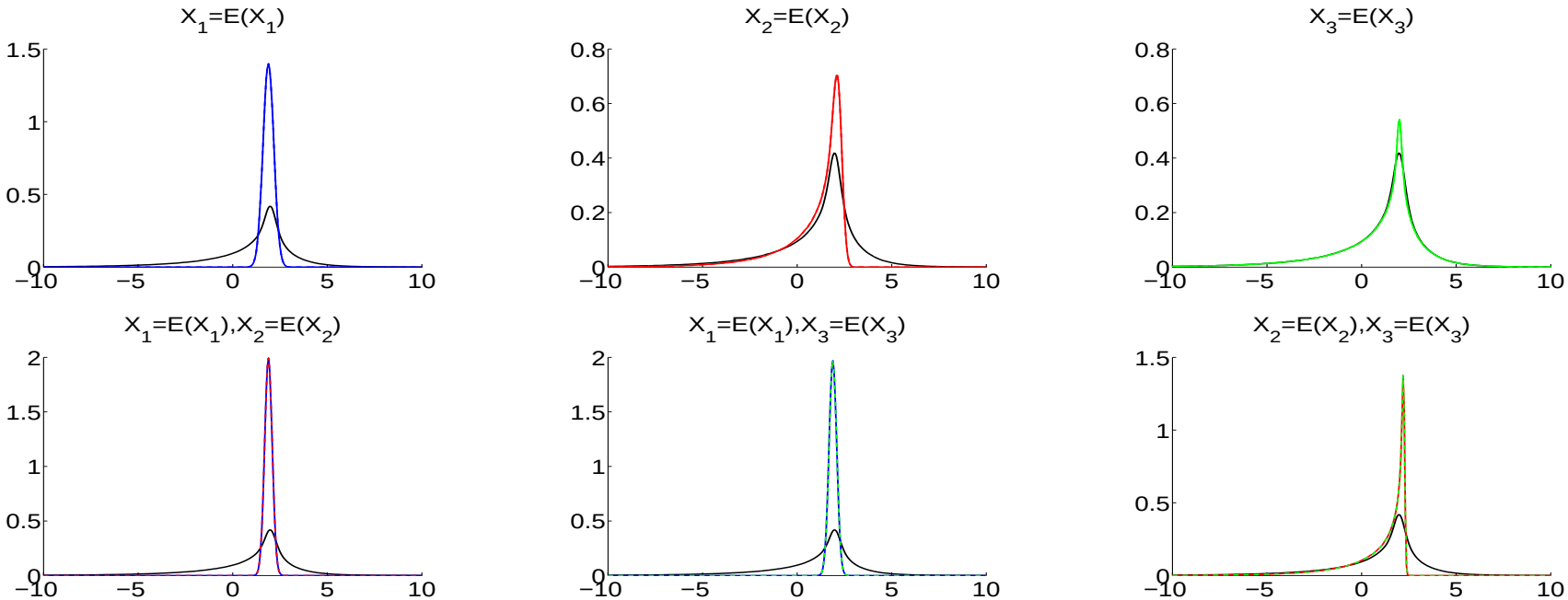
- Objectif (rappel)

  Quantifier la part de chaque variable X_i ou groupe de variable $(X_{i_1}, X_{i_2}, \ldots, X_{i_u})$ sur la "variabilité" de la sortie.

- Modélisation des paramètres d'entrée (étape A) :
  - les paramètres d'entrées X sont modélisées par des v.a.
  - Les variations sur les paramètres d'entrées impliquent une variation sur la sortie.

Analyse de sensibilité quantitative

Exercice : à partir des schémas, indiquer la/les variable(s) les plus influentes sur la variabilité de $Y = g(X_1, X_2, X_3)$. Indiquer la variable à surveiller pour garantir que $Y > 0$.



$$\begin{cases} Y = g(X_1, X_2, X_3) = X_1(X_2 - X_1) + X_3, \\ X_1 \sim \mathcal{N}(0.1, 1), \quad X_2 \sim \mathcal{N}(1, 2), \quad X_3 \sim \mathcal{N}(2, 0.2). \end{cases}$$

Analyse de sensibilité quantitative

- De maniere générale, l'impact élémentaire S_i d'une variable X_i sur la variabilité de la sortie $Y = g(X_1, \ldots, X_d)$ peut s'écrire sous la forme :

$$S_i = \mathbb{E}_{X_i}\left[d\left(f_Y, f_{Y|X_i}\right)\right],$$

  $d\left(f_Y, f_{Y|X_i}\right) \leftrightarrow$ distance **à définir** entre les PDFs de Y et $Y|X_i$.
- On peut proposer par exemple :
  - $d\left(f_Y, f_{Y|X_i}\right) = \sum_{p=1}^{P}(\mathbb{E}\left[Y^p\right] - \mathbb{E}\left[(Y|X_i)^p\right])^2$ (moments statistiques),
  - $d\left(f_Y, f_{Y|X_i}\right) = \int h(f_Y(y), f_{Y|X_i}(y)) f_{Y|X_i}(y) dy$, où h est une fonction à définir (par ex. $h(f_Y(y), f_{Y|X_i}(y)) = \log\left(\frac{f_{Y|X_i}(y)}{f_Y(y)}\right)\ldots$),...
- Pour ce cours d'introduction, on se limitera au cas où $d\left(f_Y, f_{Y|X_i}\right) = (\mathbb{E}\left[Y\right] - \mathbb{E}\left[Y|X_i\right])^2$, ce qui conduira à ce que l'on désigne usuellement par **"indices de Sobol"**, pour se concentrer davantage sur l'impact d'une ou d'un groupe de variables sur la variabilité de la sortie.

Exercice : montrer que $\mathbb{E}_{X_i}\left[(\mathbb{E}\left[Y\right] - \mathbb{E}\left[Y|X_i\right])^2\right] = \mathrm{var}\left[\mathbb{E}\left[Y|X_i\right]\right]$.

Décomposition de la variance

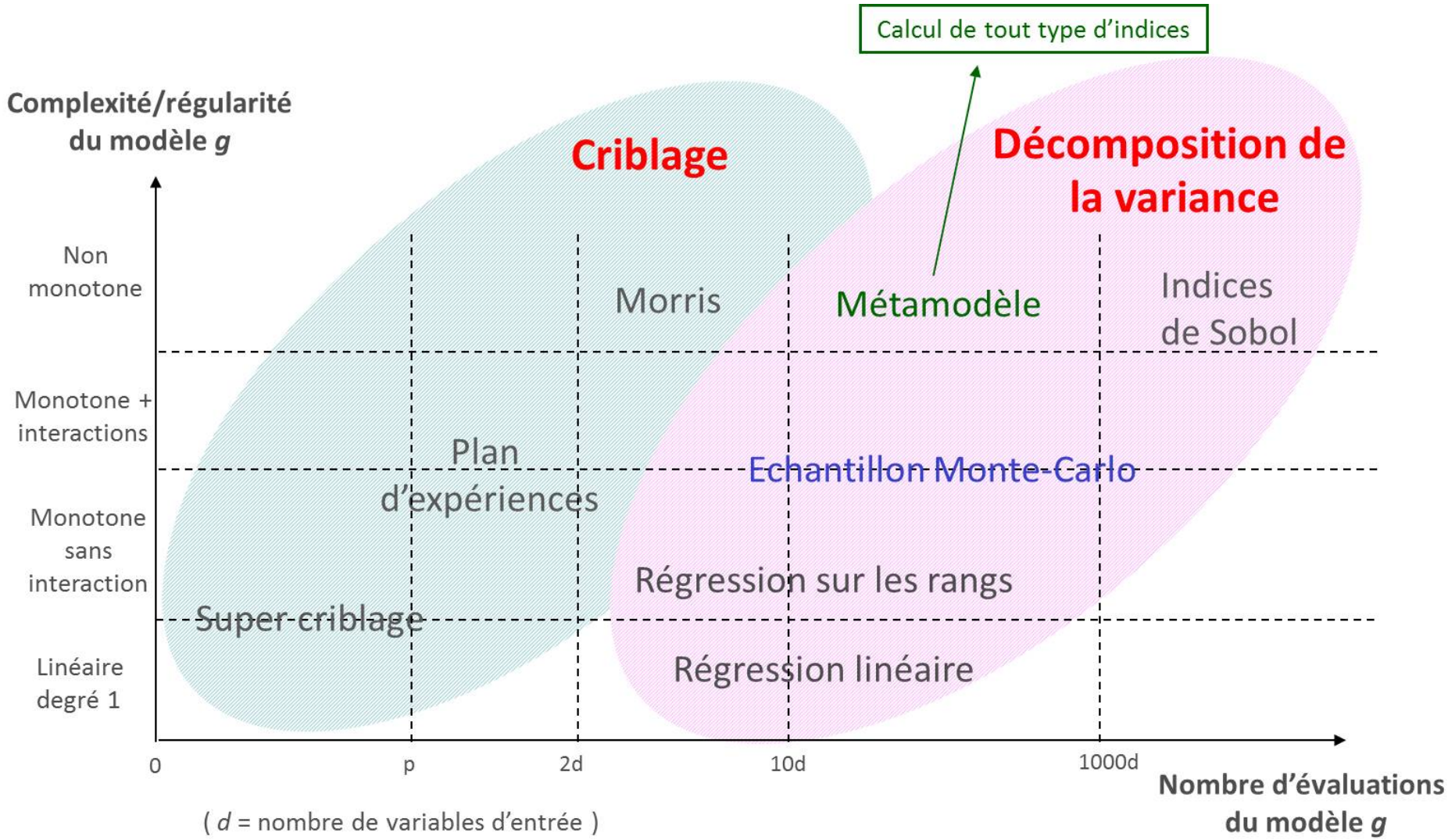
- Effectuer une analyse de sensibilité quantitative revient à **décomposer la variabilité** de $Y = g(X_1, \ldots, X_d)$ (c'est à dire la dispersion de sa distribution) en une somme de **contributions positives** pouvant être attribuées à chaque paramètre ou groupe de paramètres d'entrées, et à en évaluer les **poids** respectifs.
- Cet objectif, généralement trop ambitieux, peut être simplifié, en se limitant à la **décomposition de la variance** de Y :

$$\mathrm{var}(Y) = \sum_i V_i + \sum_{i,j} V_{ij} + \ldots + V_{1,2,\ldots,d}.$$

- Dans un grand nombre de configurations, seuls les termes du premier ordre V_i sont accessibles, en terme de coût de calcul. L'analyse au 1er ordre pourra être jugée suffisante si $\sum_i V_i / \mathrm{var}(Y) \geq 1 - \varepsilon$.
- On parle également de contribution totale T_i d'un paramètre X_i à la variance de Y :

$$T_i = V_i + \sum_j V_{ij} + \ldots + V_{1,2,\ldots,d}.$$

Deux grandes familles de méthodes d'ASG (schéma issu de [looss 2009])



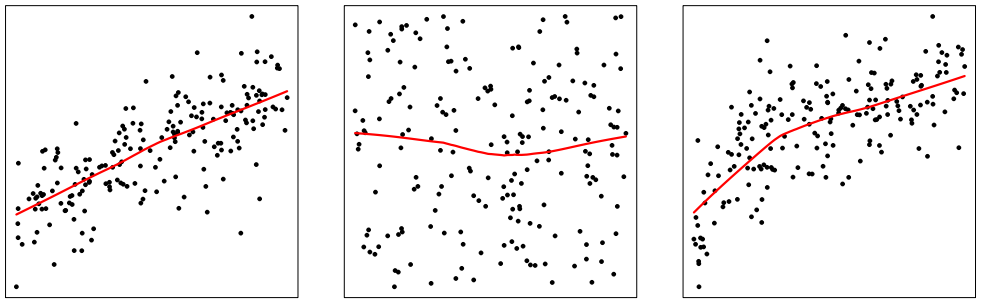
Méthodes d'ASG basées sur la régression (1/4)

On dispose d'un échantillon $(X, g(X))$ de taille $n \gg d$

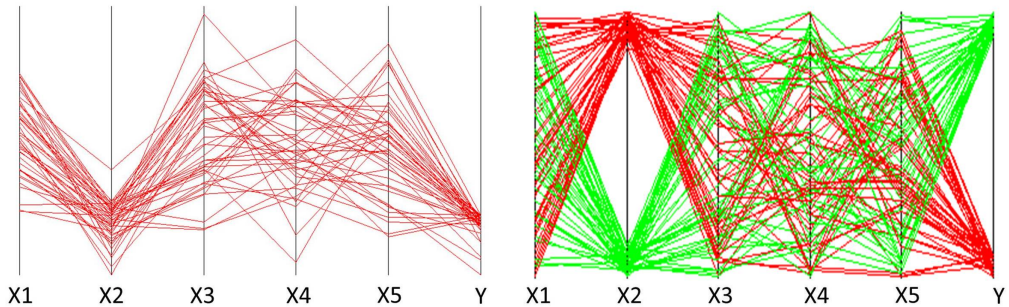
$\rightarrow$ Voir le cours sur la *planification d'expériences*

Etape préliminaire : visualisation graphique

- **Scatterplot :**
  Détection des relations du premier ordre uniquement



- **Cobwebplot :**
  Détection des interactions possible



Méthodes d'ASG basées sur la régression (2/4)

Hypothese sur le modèle

Le modèle est linéaire sans interactions

- Métamodèle linéaire : $\widehat{Y} = \widehat{g}(X) = \beta_0 + \sum_{i=1}^{d} \beta_i X_i$
- On suppose que les paramètres d'entrées sont indépendants :

$$\mathrm{var}(\widehat{Y}) = \sum_{i=1}^{d} \beta_i^2 \mathrm{var}(X_i) = \sum_{i=1}^{d} V_i, \quad V_i = \beta_i^2 \mathrm{var}(X_i), \quad SRC^2(X_i) = \frac{V_i}{\mathrm{var}(\widehat{Y})}.$$

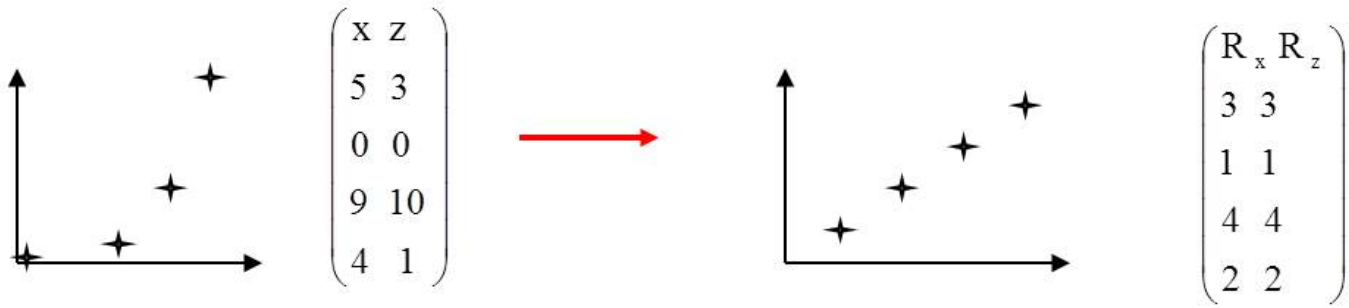
- Remarques :
  - → Le signe de β_i donne le sens de variation de Y pour X_i.
  - → On retrouve les coefficients de corrélation linéaire de Pearson.
  - → Pas d'influence combinée des paramètres sur la variabilité de Y.
  - → L'hypothese de linéarité doit être validée en s'assurant que $1 - \sum_i V_i/\mathrm{var}(Y)$ est faible !

Méthodes d'ASG basées sur la régression (3/4)

Hypothese sur le modèle

Le modèle est monotone sans interactions

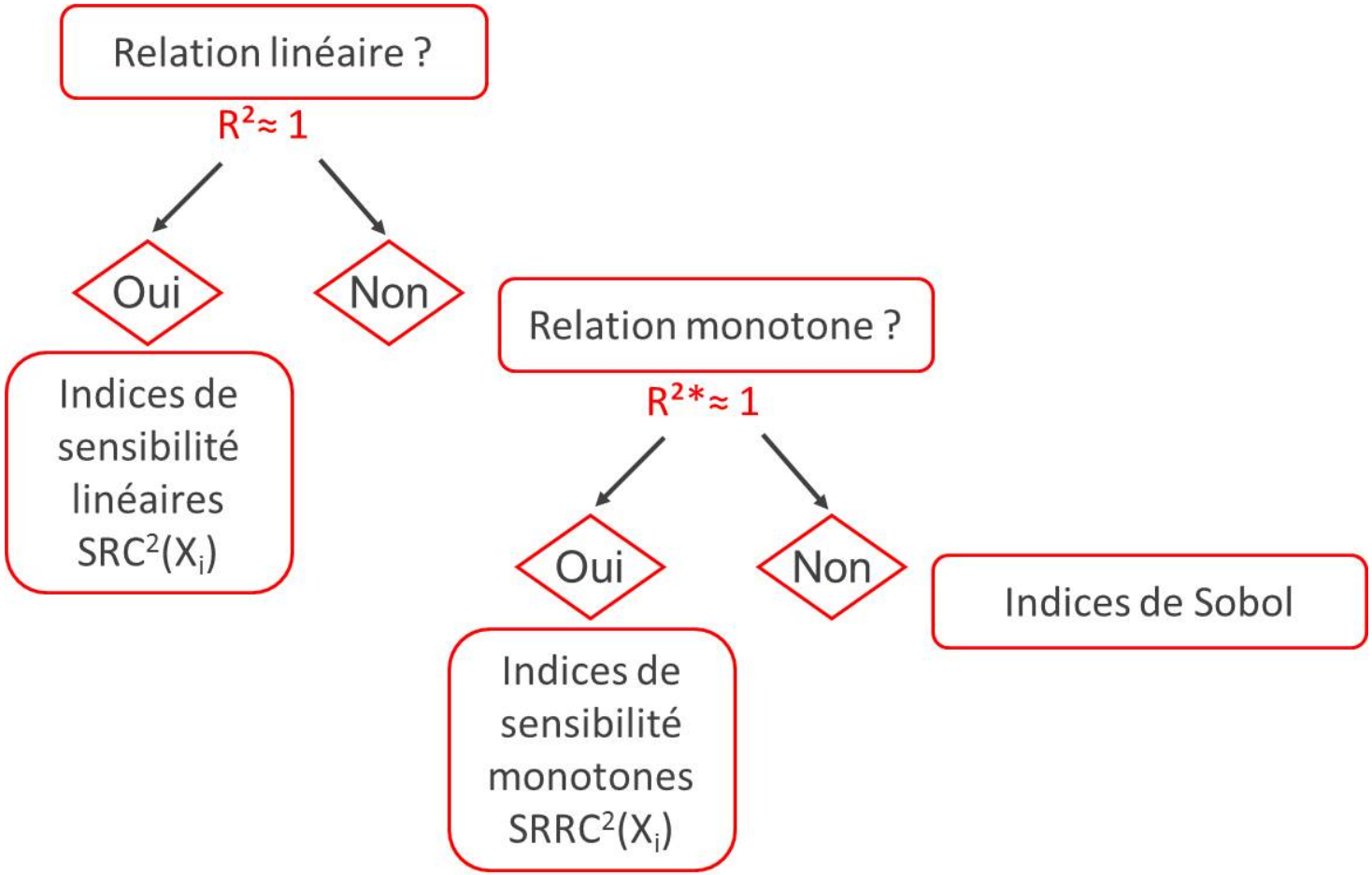
- Transformation des rangs : à chaque couple $(X^{(j)}, Y^{(j)})$, (j=1,. . .,n), on associe les rangs $(R_{X^{(j)}}, R_{Y^{(j)}})$ (les rangs varient de 1 à n).



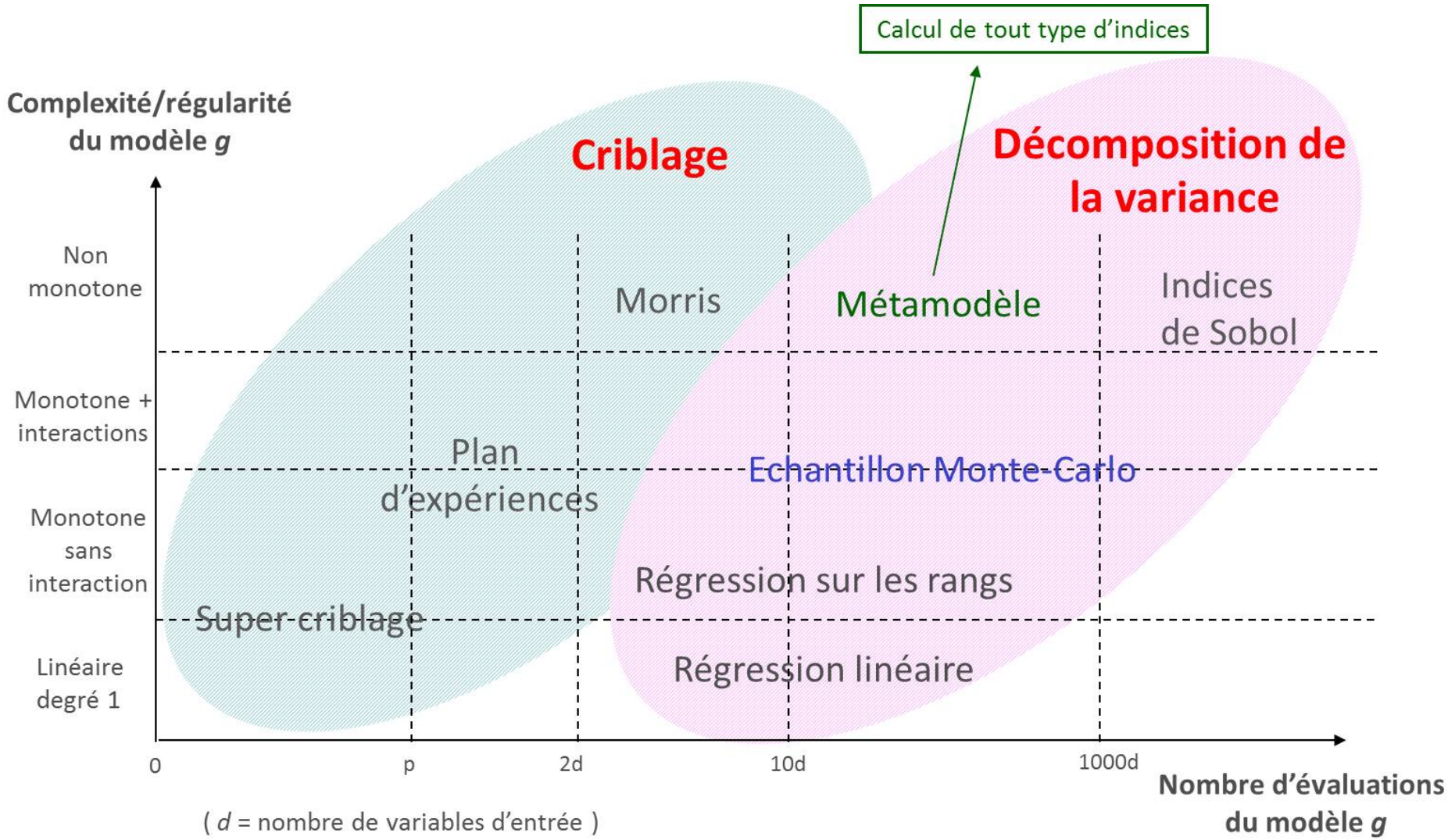
- Similaire au coefficient de corrélation des rangs de Spearman.
- Relation monotone = relation linéaire sur les rangs.
  -1cm→ Régression linéaire à partir de l'échantillon $(R_{X^{(j)}}, R_{Y^{(j)}})$, $j = 1, \ldots, n$.
- Indices de sensibilité : $SRRC^2(X_i)$ (indices plutôt qualitatifs)
- Validation : diagnostics de la régression, R^{2*}, Q_2^*

Méthodes d'ASG basées sur la régression (4/4)

Méthodologie d'analyse de sensibilité quantitative



Deux grandes familles de méthodes d'ASG (schéma issu de [looss 2009])



Indices de Sobol du premier ordre

Mesure de l'impact de la variable X_i sur la variance de $g(X)$.

- Si $X_i = x_i$, on peut définir $\delta_i(x_i) = \mathrm{var}(Y) - \mathrm{var}(Y \mid X_i = x_i)$.
  $\Rightarrow$ plus $\delta_i(x_i)$ est grand, et plus X_i a de chance de jouer un rôle important sur la variance de Y.
- $\delta_i(x_i)$ est une fonction de x_i, si bien que pour évaluer de maniere **robuste** le rôle de X_i sans être dépendant de x_i, on peut se concentrer sur la moyenne de δ_i.
  $\Rightarrow$ plus $\mathbb{E}(\delta_i)$ est grand, et plus X_i a de chance de jouer un rôle important sur la variance de Y :
  $$\mathbb{E}(\delta_i) = \mathbb{E}(\mathrm{var}(Y) - \mathrm{var}(Y|X_i)) = V(Y) - \mathbb{E}(\mathrm{var}(Y \mid X_i)).$$

Indices de Sobol du premier ordre

En normalisant par la variance de Y, et en remarquant que $V(Y) = \mathrm{var}(\mathbb{E}(Y \mid X_i)) + \mathbb{E}(\mathrm{var}(Y \mid X_i))$, on nomme S_i les indices de Sobol du premier ordre, définis par : $0 \leq S_i = \frac{\mathrm{var}(\mathbb{E}(Y \mid X_i))}{\mathrm{var}(Y)} \leq 1$.

Indices de Sobol généralisés (1/2)

- La décomposition de Hoeffding-Sobol stipule que toute fonction de carré intégrable peut s'écrire de maniere unique sous la forme :

$$g(X) = g_0 + \sum_{i=1}^{d} g_i(X_i) + \sum_{1\le i<j\le d} g_{i,j}(X_,X_j) + \cdots + g_{1,2,\ldots,d}(X_a, X_2, \ldots, X_d)$$

$$\begin{aligned} g_0 &= \mathbb{E}[g(X)] \\ g_i(X_i) &= \mathbb{E}[g(X)|X_i] - g_0 \\ g_{i,j}(X_i, X_j) &= \mathbb{E}[g(X)|X_i, X_j] - g_i(X_i) - g_j(X_j) - g_0 \\ &\cdots \end{aligned}$$

- On peut alors écrire :

$$\begin{aligned} \mathrm{var}(g(X)) &= \sum_{i=1}^{d} \mathrm{var}(g_i(X_i)) + \sum_{1\le i<j\le d} \mathrm{var}(g_{i,j}(X_i, X_j)) + \cdots + \mathrm{var}(g_{1,2,. \\ &= \sum_{i=1}^{d} V_i + \sum_{1\le i<j\le d} V_{ij} + \cdots + V_{1,\ldots,d}. \end{aligned}$$

Indices de Sobol généralisés (2/2)

En divisant l'expression précédente par la variance de Y, on obtient :

$$1 = \sum_{i=1}^{d} S^i + \sum_{1 \leq i < j \leq d} S_{ij} + \cdots + S_{12\ldots d},$$

$$0 \leq S_i,\ S_{ij}, \ldots, S_{12\ldots d} \leq 1,$$

- les S_i correspondent aux indices de Sobol du premier ordre, qui s'interpretent comme la part de variance de Y expliquée par l'effet individuel de X_i seulement,
- les S_{ij} correspondent aux indices de Sobol du second ordre, qui s'interpretent comme la part de variance de Y expliquée par l'interaction (X_i, X_j) et non expliquée par les effets individuels de X_i et X_j, $\cdots$,
- $S_{12\ldots d}$ correspond à l'indice de Sobol d'ordre d, qui s'interprete comme la part de variance de Y expliquée par l'interaction $(X_1, X_2, \ldots, X_d)$ et non expliquée par tous les autres effets individuels ou combinés de paramètres.

Indices de Sobol totaux

On définit par ailleurs les indices de Sobol totaux associés au paramètre X_i, et nommés T_i, comme la somme des indices faisant intervenir X_i (*i.e.* la **somme de toutes le contributions de** X_i) :

$$T^i = S^i + \sum_{j \neq i} S_{ij} + \sum_{1 \leq j < k \leq d, i \neq j, i \neq k} S_{ijk} + \cdots + S_{12\ldots d}$$

En définissant $X_{(-i)} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d)$, on a ($V = \mathrm{var}(Y)$) :

$$V = \underbrace{\mathrm{var}\left(\mathbb{E}(Y|X_i)\right) + \mathrm{var}\left(Y - \mathbb{E}(Y|X_i) - \mathbb{E}(Y|X_{(-i)})\right)}_{\text{contribution totale de } X_i} + \underbrace{\mathrm{var}\left(\mathbb{E}(Y|X_{(-i)})\right)}_{\text{complémentaire}},$$

En remarquant que $\mathrm{var}(Y) = \mathrm{var}\left(\mathbb{E}(Y|X_{(-i)})\right) + \mathbb{E}\left(\mathrm{var}(Y|X_{(-i)})\right)$:

$$\begin{aligned} T_i &= 1 - \frac{\mathrm{var}\left(\mathbb{E}(Y|X_{(-i)})\right)}{\mathrm{var}(Y)}, \\ &= \frac{\mathbb{E}\left(\mathrm{var}(Y|X_{(-i)})\right)}{\mathrm{var}(Y)}. \end{aligned}$$

Approche pratique concernant les indices de Sobol

1. On commence par calculer $\mathrm{var}(Y)$ (une boucle de calcul), $\mathbb{E}\left(\mathrm{var}(Y|X_{(-i)})\right)$ (deux boucles de calcul) et $\mathrm{var}(\mathbb{E}(Y|X_i))$ (deux boucles de calcul).

2. On en déduit les indices de Sobol totaux et du premier ordre :

$$S_i = \frac{\mathrm{var}(\mathbb{E}(Y|X_i))}{\mathrm{var}(Y)}, \quad T_i = \frac{\mathbb{E}\left(\mathrm{var}(Y|X_{(-i)})\right)}{\mathrm{var}(Y)}.$$

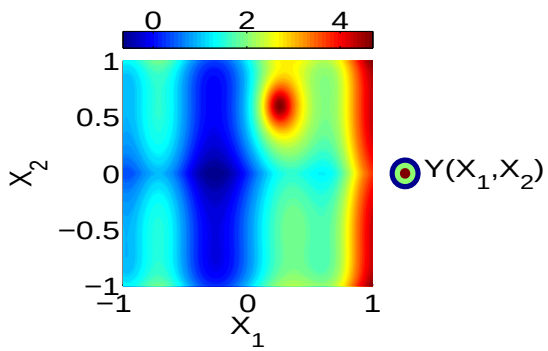
3. Si $S_i \approx T_i$, alors l'action des interactions incluant X_i peut être négligé, et l'on se contente de ce résultat. Si ce n'est pas le cas, on essaye (si le budget de calcul le permet) d'évaluer S_{ij},... puis S_{ijk} si nécessaire, et ainsi de suite...

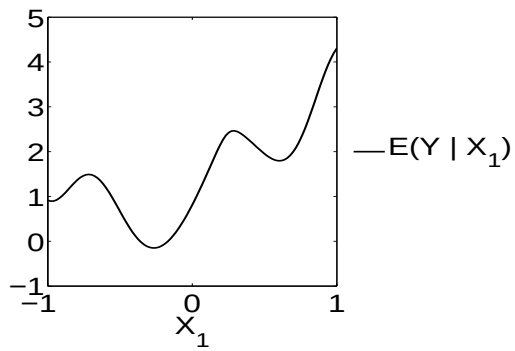
Illustration des indices de Sobol

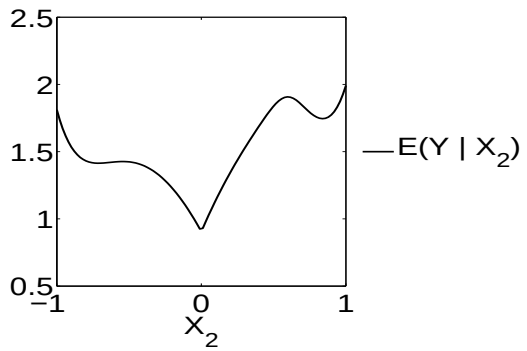
Exemple :

$$Y = g(X_1, X_2) = 0.2\exp(X_1 - 3) + 2.2|X_2| + 1.3X_2^6 - 2X_2^2 - 0.5X_2^4 - 0.5X_1^4 + 2.5X_1^2 + 0.7X_1^3 + \frac{3}{(8X_1-2)^2+(5X_2-3)^2+1} + \sin(5X_1)\cos(3X_1^2),$$

$X_1 \sim \mathcal{U}(-1,1)$, $X_2 \sim \mathcal{U}(-1,1)$.







$$\mathrm{var}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2$$

$$= \frac{1}{4}\int_{-1}^{1}\int_{-1}^{1} Y(x_1,x_2)^2 dx_1 dx_2 - \left(\frac{1}{2}\int_{-1}^{1}\int_{-1}^{1} Y(x_1,x_2) dx_1 dx_2\right)^2 \approx 1.23,$$

$$\mathrm{var}(\mathbb{E}(Y|X_1)) \approx 1.15, \quad \mathrm{var}(\mathbb{E}(Y|X_2)) \approx 0.075.$$

Estimation Monte-Carlo des indices de Sobol

- Soit $(\boldsymbol{X}^{(j)})_{j=1,\dots,n}$ un échantillon Monte-Carlo de $\boldsymbol{X} = (X_1, \dots, X_d)$.
- Estimation du numérateur $\mathrm{var}(\mathbb{E}(g(\boldsymbol{X})|X_i))$.
  - $\mathrm{var}(\mathbb{E}(g(\boldsymbol{X})|X_i)) = \mathbb{E}_{X_i}\left(\mathbb{E}_{X_{(-i)}}(g(\boldsymbol{X})|X_i)^2\right) - \mathbb{E}_{\boldsymbol{X}}\left(g(\boldsymbol{X})\right)^2$.
  - Astuce : soit $\widetilde{\boldsymbol{X}}$ une variable aléatoire de même loi que $\boldsymbol{X}$ telle que $\boldsymbol{X}$ et $\widetilde{\boldsymbol{X}}$ soient indépendantes, on définit la variable aléatoire :

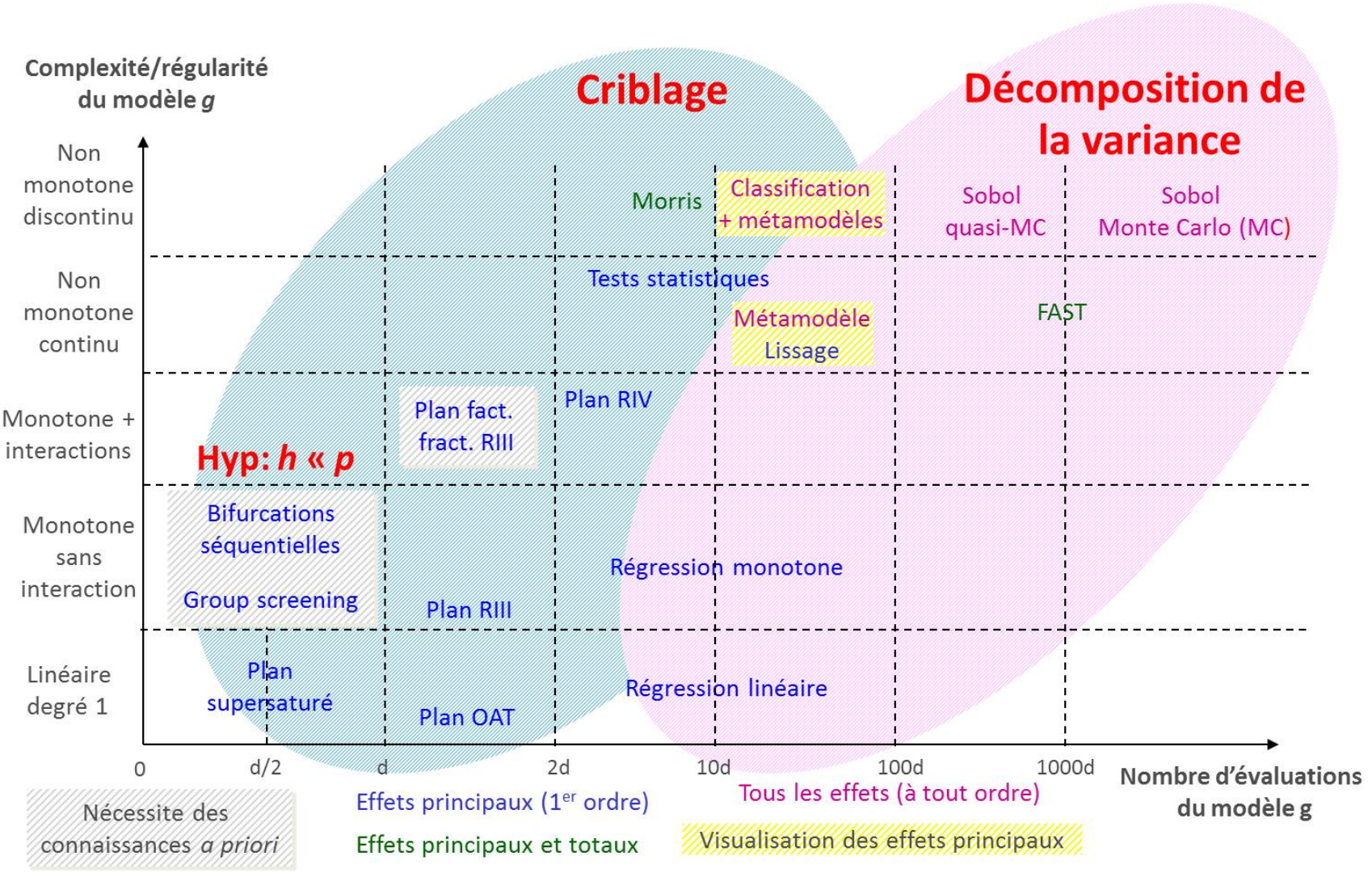
$$\boldsymbol{T} = (\widetilde{X}_1, \widetilde{X}_2, \dots, \widetilde{X}_{i-1}, X_i, \widetilde{X}_{i+1}, \dots, \widetilde{X}_d), \text{ si bien que :}$$

$$\begin{aligned}
\mathbb{E}_{X_i}\left(\mathbb{E}_{X_{(-i)}}(g(\boldsymbol{X})|X_i)^2\right) &= \mathbb{E}_{X_i}\left(\mathbb{E}_{X_{(-i)}}(g(\boldsymbol{X})|X_i)\mathbb{E}_{T_{(-i)}}(g(\boldsymbol{T})|X_i)\right), \\
&= \mathbb{E}_{X_i}\left(\mathbb{E}_{X_{(-i)},T_{(-i)}}((g(\boldsymbol{X})|X_i)(g(\boldsymbol{T})|X_i))\right) \text{ par indépendance,} \\
&= \mathbb{E}(g(\boldsymbol{X})g(\boldsymbol{T})).
\end{aligned}$$

- On a donc l'estimateur Monte-Carlo suivant des indices de Sobol :

$$S_i = \frac{\mathrm{var}(\mathbb{E}(g(\boldsymbol{X})|X_i))}{\mathrm{var}(g(\boldsymbol{X}))} \approx \frac{\frac{1}{n}\sum_{j=1}^n g(\boldsymbol{X}^{(j)})g(\boldsymbol{T}^{(j)}) - \frac{1}{n}\sum_{j=1}^n g(\boldsymbol{X}^{(j)})\frac{1}{n}\sum_{j=1}^n g(\boldsymbol{T}^{(j)})}{\frac{1}{n}\sum_{j=1}^n g(\boldsymbol{X}^{(j)})^2 - \left(\frac{1}{n}\sum_{j=1}^n g(\boldsymbol{X}^{(j)})\right)^2}.$$

Classification des méthodes d'ASG (schéma issu de [Iooss 2009])



Autres approches d'ASG (liste non exhaustive)

Autres approches d'ASG (non abordées dans ce cours) :

- Basées sur des méthodes d'approximation de modèle : méthode de lissage, métamodèle (voir cours sur *les métamodèles*)

  $\longrightarrow$ exploration fine des densités.

- Basées sur des méthodes d'analyse de données : tests statistiques, arbre de classification, ...

  $\longrightarrow$ criblage

- Indices globaux basés sur les dérivées : DGSM (Derivative-based global sensitivity measure)

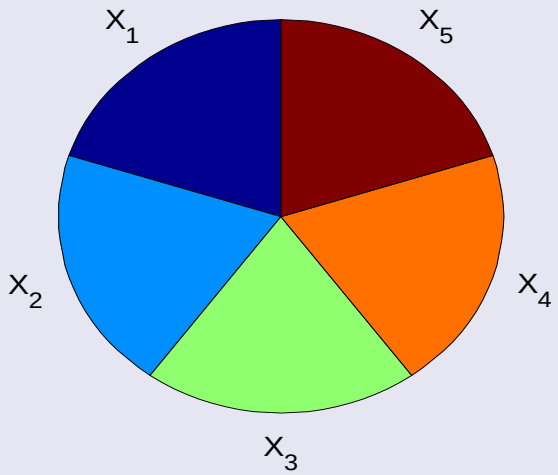
  $\longrightarrow$ criblage

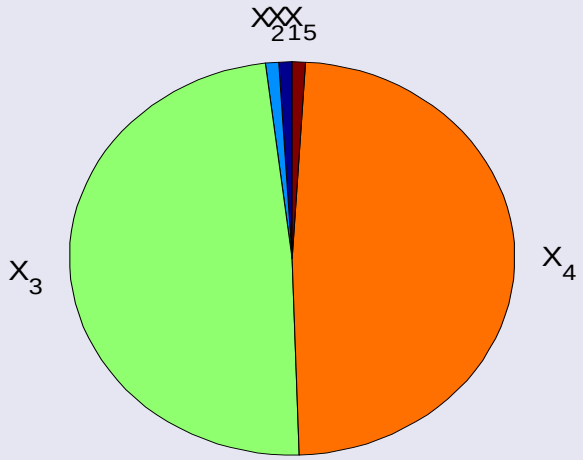
Conclusion

- Utiliser la méthode la plus simple en fonction de :
  - l'objectif de l'étude,
  - temps d'exécution du modèle numérique (et maniabilité)
  - nombre de paramètres d'entrée, type de paramètres,
  - la connaissance *a priori* que l'on a du modèle,
  - les contraintes potentielles (plusieurs sorties, ...)
- Avoir une approche itérative :
  - la validation *a posteriori* de la méthode utilisée permet de savoir s'il est nécessaire d'en utiliser une plus performante
  - exploration de l'espace des paramètres d'entrée de plus en plus finement (si nécessaire)
- Ne pas négliger et sous-estimer les étapes du choix des paramètres et de définition des domaines de variation (et/ou modalités, pdf)
- Les résultats d'une AS ne sont valables que pour les domaines de variation/pdf choisis au départ.

Attention !

Une analyse de sensibilité globale dépend COMPLeTEMENT des distributions des paramètres d'entrée ! ! !





$Y = X_1 + X_2 + X_3 + X_4 + X_5$,
Cas 1 : $X_1, X_2, X_3, X_4, X_5 \sim \mathcal{N}(0, 1)$.
Cas 2 : $X_1, X_2, X_5 \sim \mathcal{N}(0, \sqrt{0.1})$, $X_3, X_4 \sim \mathcal{N}(0, \sqrt{5})$.

Bibliographie

- Livres de références :
  - Saltelli & al., *Sensitivity analysis*, Wiley, 2000.
  - Saltelli & al., *Global Sensitivity analysis - The Primer*, Wiley, 2008.
- Ouvrage pédagogique en français :

  Faivre & al., *Analyse de sensibilité et exploration de modèles - Applications aux sciences de la nature et de l'environnement*, Editions Quaé, 2013.
- Article pédagogique en français : Iooss, Journal de la SFdS, 152(11), 2011.
- Librairies **R** : sensitivity, planor, CompModSA, multisensi

Quelques extensions...

- Utilisation d'autres quantités d'intérêt :
  - Indices distributionnels (distance entre distributions). (E. Borgonovo)
  - Généralisation des indices de Sobol pour différentes quantités d'intérêt. (J.-C. Fort, T. Klein, N. Rachdi)
  - Indices pour probabilités de défaillance. (These de P. Lemaître)
- Sortie fonctionnelle : la sortie résumée par une vecteur (décomposition sur une base de fonctions) puis métamodélisation des coefficients de la base.
  - approche composante par composante (courbe ou carte d'indices de sensibilité) (These d' A. Marrel)
  - indices généralisée (These de M. Lamboni)
- Entrées non indépendantes :
  - Généralisation de la décomposition ANOVA pour des variables corrélées (These de G. Chastaing)
  - ANCOVA décomposition (ANalysis of COVAvariance) (These de

Quizz

- **Pour l'AS d'un modèle à beaucoup de paramètres, vaut-il mieux privilégier**
  ☐ une analyse quantitative ? ☐ une analyse qualitative ? ☐ l'une puis l'autre ? ☐ l'autre puis l'une ?
- **L'utilisation de l'approximation linéaire est selon vous :**
  ☐ dangereuse ☐ utile ☐ inutile ☐ utile si validée
- **Les plans OAT pour des modèles à beaucoup de paramètres couvrent une partie de l'espace**
  ☐ importante ☐ réduite ☐ négligeable ☐ intéressante
- **Pour pouvoir retirer X_i des paramètres variables d'un modèle, il faut que**
  ☐ $S_i \approx 1$ ☐ $S_i \approx 0$ ☐ $T_i \approx 1$ ☐ $T_i \approx 0$
- **Quels sont les points délicats de l'analyse de sensibilité globale ?**
  ☐ distributions-dépendants ☐ lourds numériquement ☐ difficiles à interpréter ☐ hypotheses-dépendants