

# Traitement des incertitudes en simulation numérique

## Cours 2 : Planification et analyse d'expériences numériques

**Bertrand Iooss**

**Module INSA Toulouse/GMM 5**  
**Planification, risque et incertitudes**

**28 novembre 2012**



# Module Traitement des incertitudes en simulation numérique

Depuis une trentaine d'années, l'industrie a développé des processus et des codes de calcul parfois très lourds pour modéliser des phénomènes complexes !

**La plupart des ingénieurs sont amenés à manipuler ces codes & processus**

1) Il est nécessaire d'**optimiser leur utilisation pour prendre des décisions !**

=> *Analyse de sensibilité, planification d'expérience, développement de modèles réduits*

2) La **validation de leurs résultats** est un problème crucial lorsqu'ils sont utilisés dans des cycles industriels (conception, sûreté, prévision, etc.)

=> *Gestion des incertitudes, calculs fiabilistes*

## 3 cours de 3h15 pour INSA GMM 5 & Master Pro 2 UPS

1.Cours 1 : Introduction, modélisation et propagation d'incertitudes

2.Cours 2 : Planification et analyse d'expériences numériques

3.Cours 3 : Modélisation d'expériences numériques, krigeage

## 3 séances de TP pour INSA GMM 5

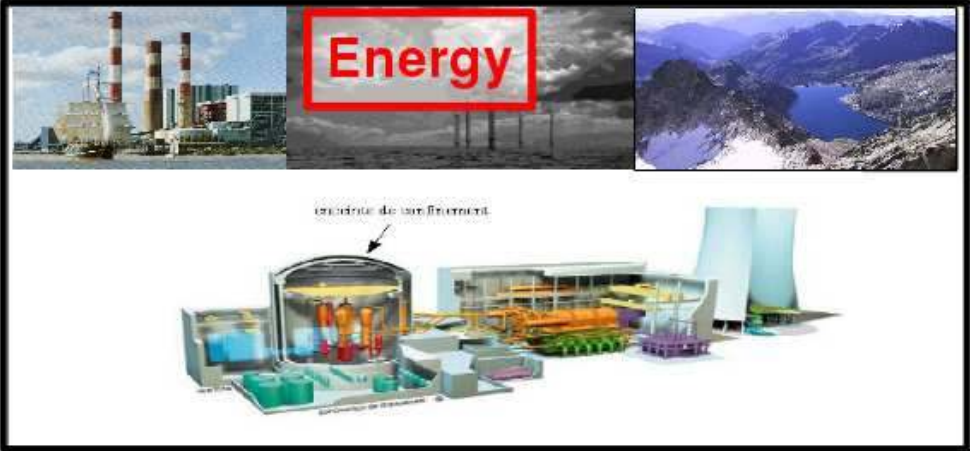
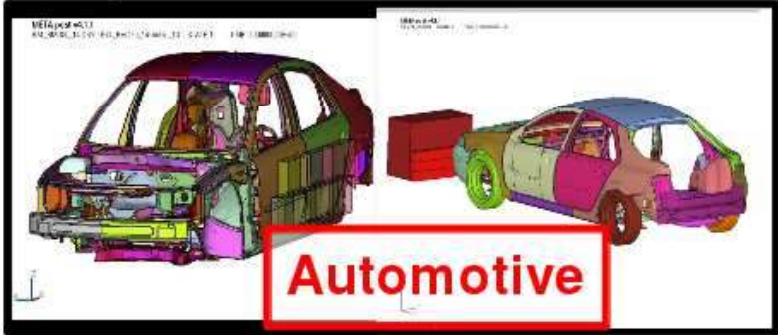
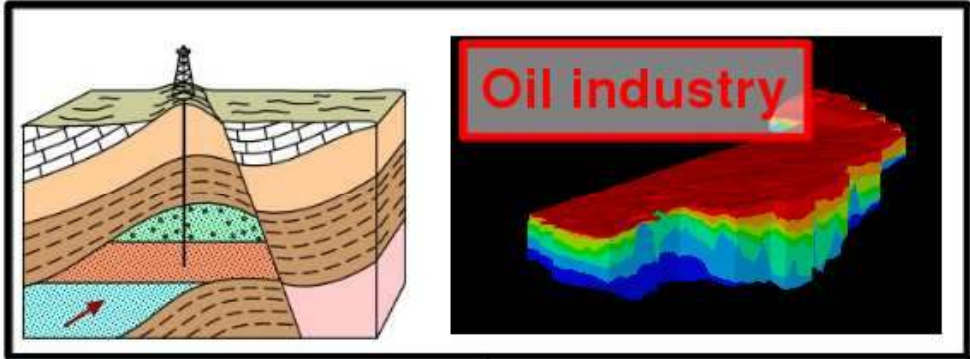
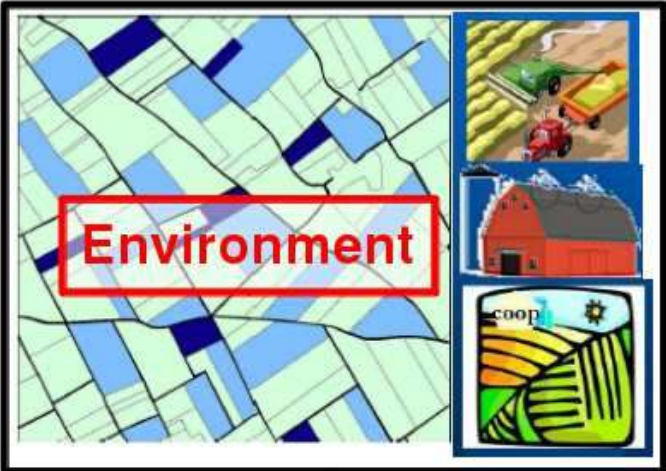
1.TP 1 : Exercices en R

2.TP 2 : Exercices en R

3.TP 3 : Exercices en R

Une note sera délivrée via des compte-rendus réalisés à l'issue des TPs

# Une problématique multi-sectorielle



# Plan du cours 2

1. **Introduction**
2. Planification d'expériences numériques
3. Méthodes d'analyse de sensibilité

# Incertitudes en simulation numérique : les enjeux

## ► Modélisation :

- Explorer au mieux différentes combinaisons des entrées
- Identifier les données influentes pour prioriser la R&D
- Améliorer le modèle

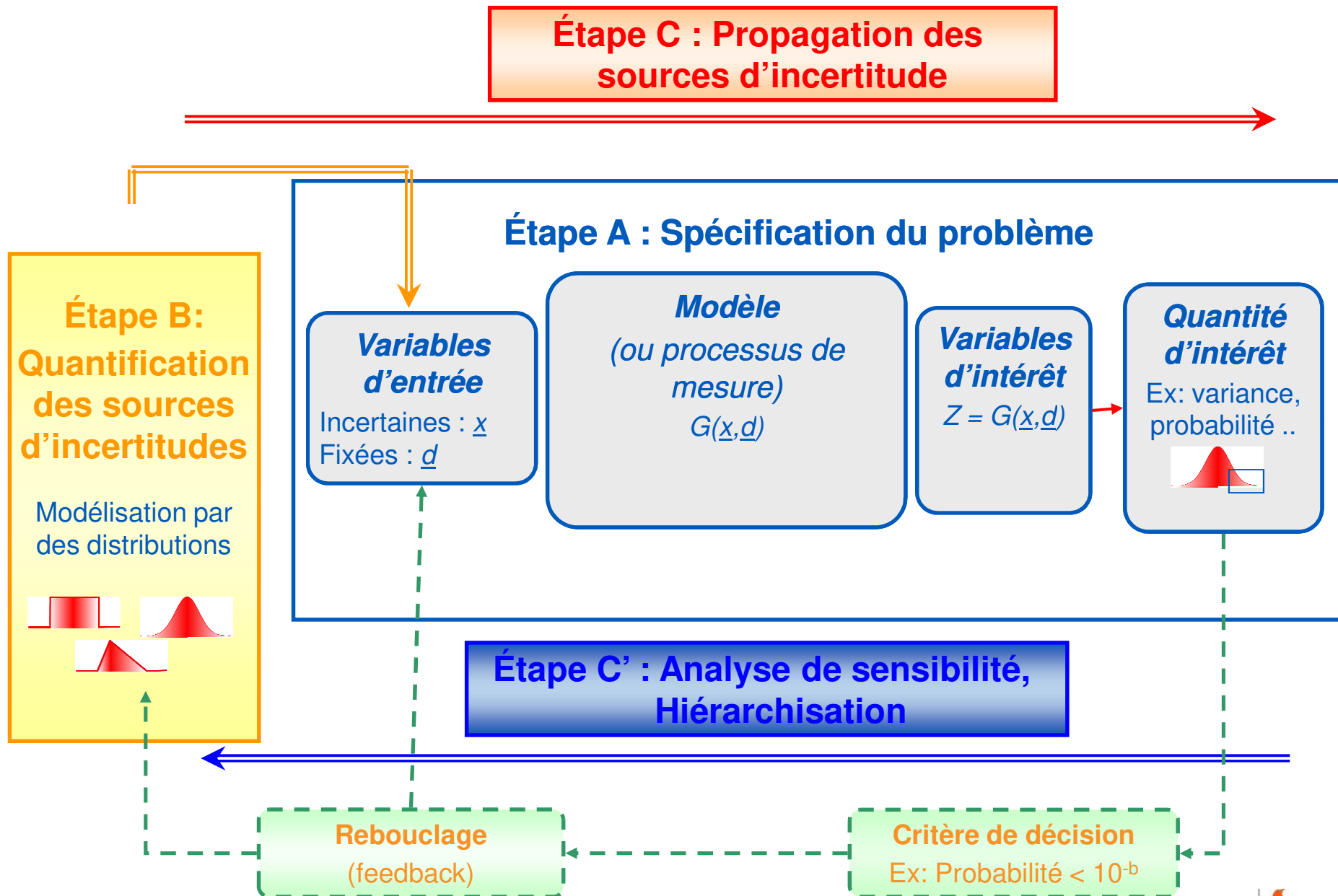
## ► Validation :

- Réduire l'incertitude de prédiction
- Calibrer les paramètres du modèle

## ► Utilisation :

- Études de sûreté : calculer un risque de défaillance (Fiabilité des structures - événements rares), calculer des **marges** (par rapport à une réglementation)
- Conception : optimiser les performances d'un système

# Approches quantitatives : schéma générique introductif



# Rappels sur la propagation d'incertitudes

- ▶ **Enjeu** : Arbitrer entre précision de l'estimateur et coût des calculs
- ▶ Si possible, **Monte Carlo** est à privilégier : indépendant de la dimension des entrées, estimation non biaisée, fournit un intervalle de confiance sur l'estimation  
Mais : coût important en nombre d'évaluations du modèle
- ▶ Si le code de calcul est trop coûteux en CPU, il existe des méthodes alternatives :
  - Méthodes **quasi-Monte Carlo** (cf. cours 2) - Mais : fléau de la dimension
  - Méthodes approchées :
    - Cumul quadratique (développement de Taylor) - Mais : hypothèse linéaires
    - Méthodes FORM/SORM : estimation rapide de  $p_f$  . Cette première estimation peut être utilisée pour construire un tirage d'importance
  - Utilisation d'un modèle de substitution du code de calcul (cf. cours 3) ayant un coût pratiquement nul (**métamodèle**)
    - Attention : un nouveau terme d'erreur apparaît
    - Le calage du métamodèle demande aussi un certain nombre d'appels au vrai modèle G

# Plan du cours 2

1. Introduction
- 2. Planification d'expériences numériques**
3. Méthodes d'analyse de sensibilité

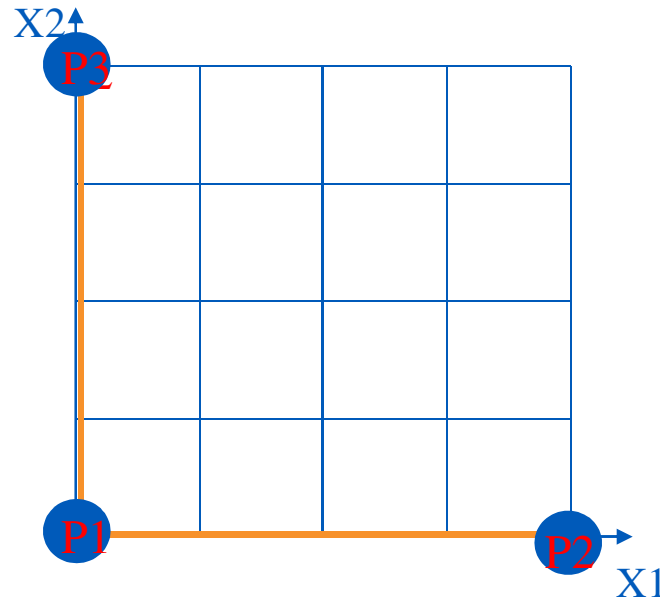


# Objectifs des plans exploratoires

- ▶ Explorer le comportement des réponses d'un code à l'aide d'un nombre limité de calculs
- ▶ Propager les incertitudes (calcul des moments des réponses)
- ▶ Fournir des plans plus efficaces que les plans aléatoires purs pour estimer des indices de sensibilité
- ▶ Fournir un plan initial pour construire un métamodèle

# Mauvais plan exploratoire : le plan « One-At-a-Time » (OAT)

Part de l'idée très répandue que pour analyser les causes d'un phénomène, il faut faire des expériences en **ne bougeant qu'un seul facteur à la fois**



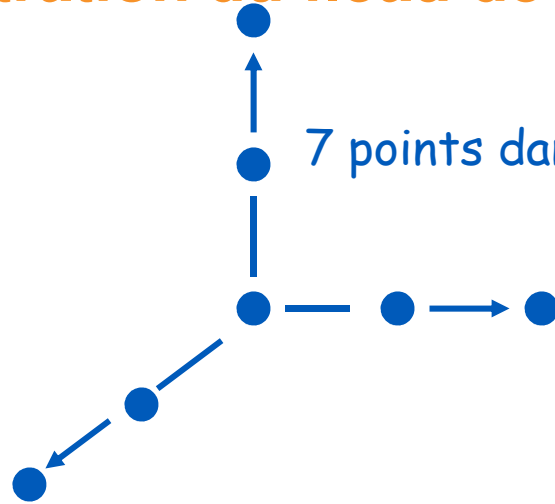
On voit aisément que : OAT apporte des informations, potentiellement fausses

## L'exploration est pauvre :

- ne détecte pas les non monotonies, discontinuités, interactions
- laisse de grandes zones inexplorées dans l'espace des paramètres d'entrée (fléau de la dimension)

Remarque : OAT est un plan de résolution III mais aliases très mal définis

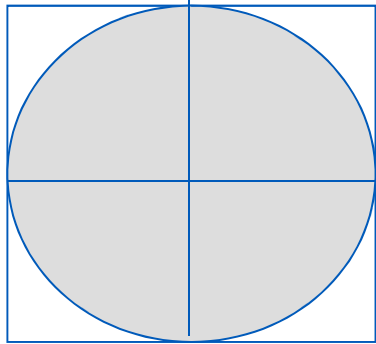
# Illustration du fléau de la dimension



7 points dans un espace 3D

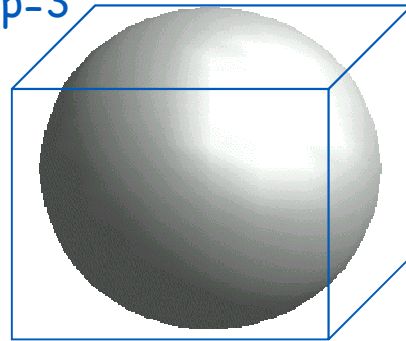
Faible recouvrement de l'espace

p=2



Surf. cercle / Surf. carré ~ 3/4

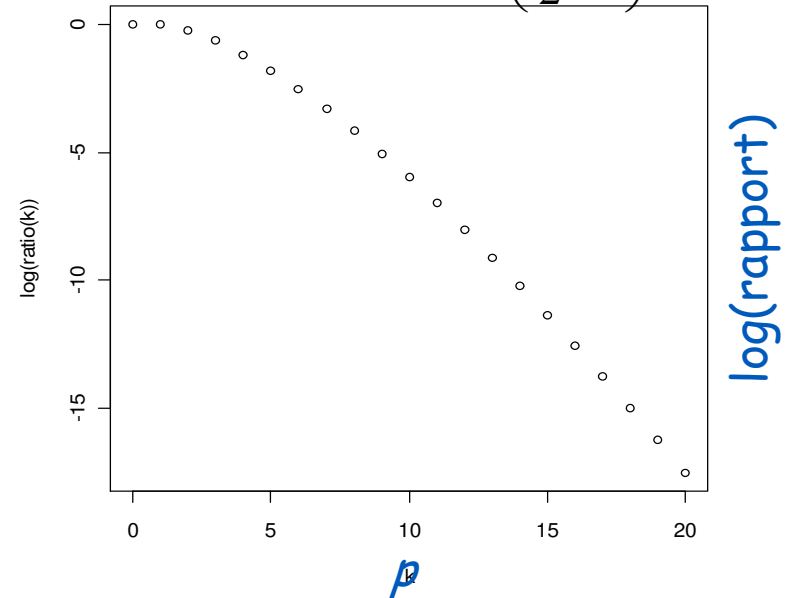
p=3



Vol. sphere / Vol. cube ~ 1/2

p=10 → Rapport ~ 0.0025

$$\text{vol. sphere } (r = 0.5) = \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2} + 1\right)} \left(\frac{1}{2}\right)^p$$



volume de l'hypercube → volume de l'hypersphère (incluse et tangente)

# Exploration « optimale » d'un domaine hypercubique

Placer des points dans le domaine des entrées  $\mathbf{X} \in \mathbb{R}^p$  dans le but de « maximiser » la quantité d'information sur la sortie du modèle  $Y = G(\mathbf{X})$

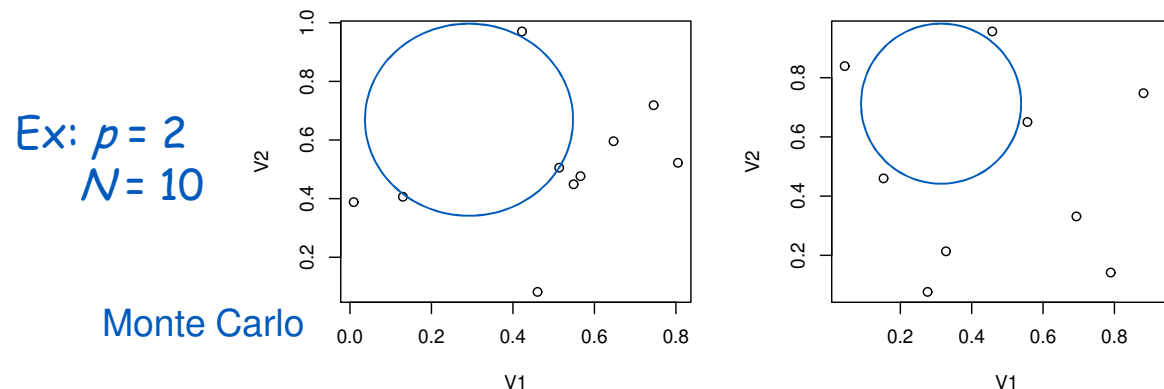
La précision (et donc le coût) de l'exploration dépend de  $p$   
(contrairement à la prop. d'incert.)

Grille régulière à  $n$  niveaux  $\rightarrow N = n^p$  simulations



Pour minimiser  $N$ , on a besoin d'échantillons assurant une bonne couverture de l'espace des entrées

Un échantillon purement aléatoire (Monte Carlo) ne le permet pas



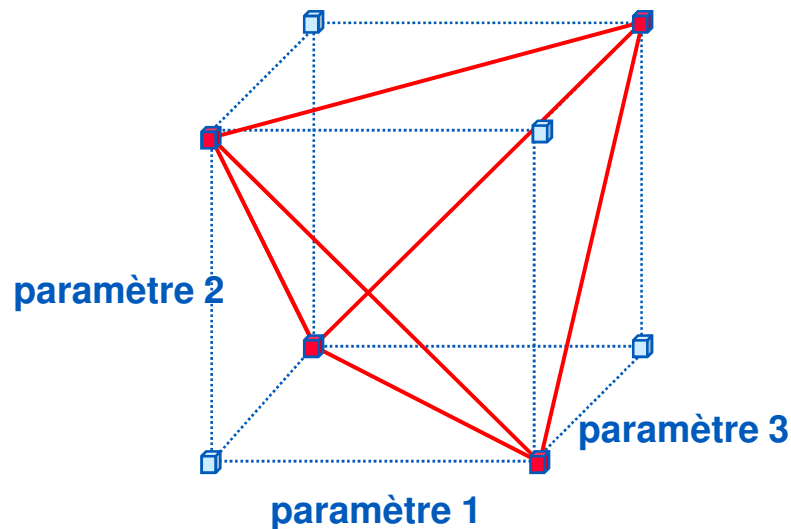
# Plans d'expériences

Experimental Design can be defined as the strategy for setting up experiments [ *performing simulations* ] in such a manner that the information required is obtained as efficiently and precisely as possible (Lewis & Phan-Tan-Luu, 2000)

## Plans classiques (facteurs discrétisés suivant des niveaux)

Exemples :

- Plan factoriel complet  $2^3$
- Plan factoriel fractionnaire  $2^{3-1}$



## Plans classiques (facteurs continus)

Plans optimaux consistant à minimiser une variance ou le déterminant d'une matrice de covariance par rapport à un modèle spécifié (linéaire, polynôme d'ordre deux, ...)

## Plans pour simulations numériques

Spécificités

- expériences déterministes (erreur=0),
- grand nombre de facteurs,
- larges domaines de variation,
- variables d'intérêt multiples,
- modèles fortement non linéaires, ...

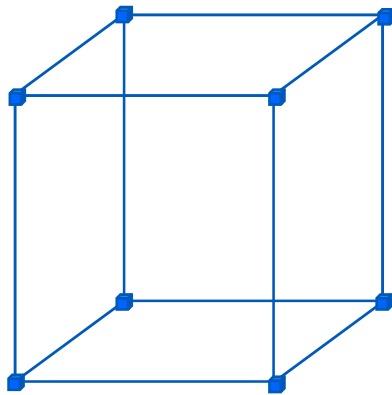
➡ space filling designs

Biblio : Fisher (1917), Box et Wilson (1954), Taguchi (1960), Mitchell (1958), ...

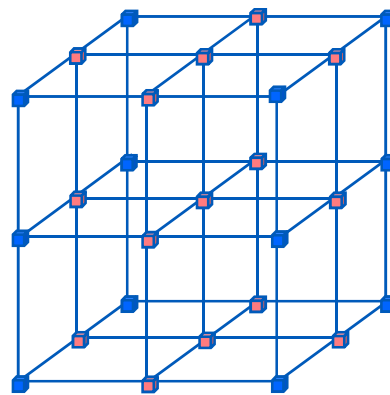
Biblio : Kleijnen (1970), McKay (1979), Morris (1995), Sacks (1989), ...

# Quelques plans d'expériences « classiques »

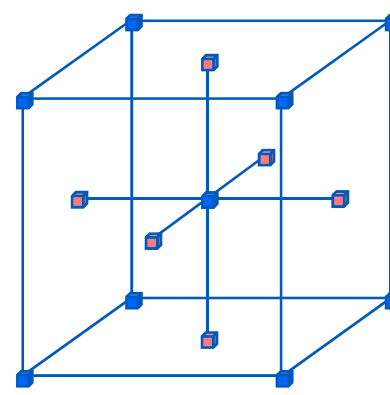
FACTORIEL 2<sup>p</sup>



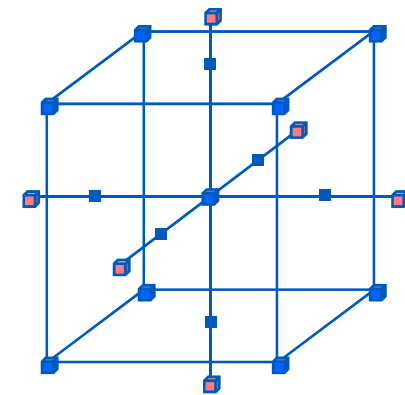
FACTORIEL 3<sup>p</sup>



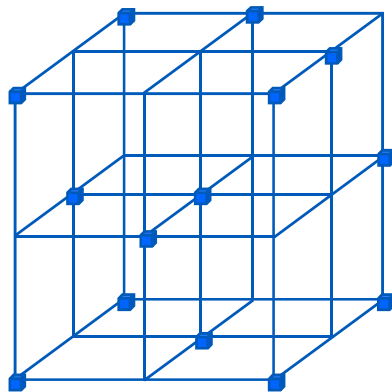
COMPOSITE FCC



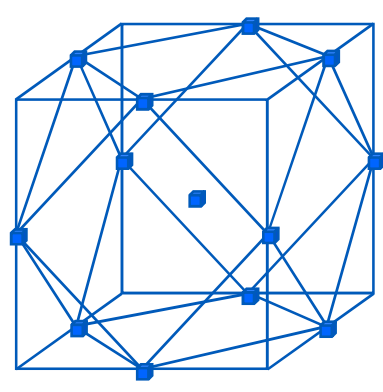
COMPOSITE CCD



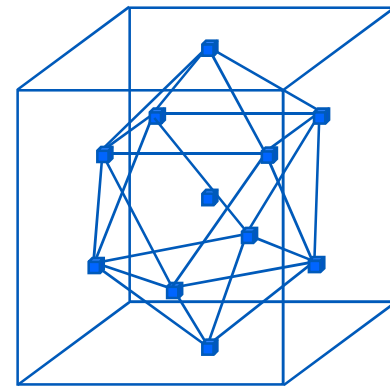
## Recueil de plans standard (tabulés)



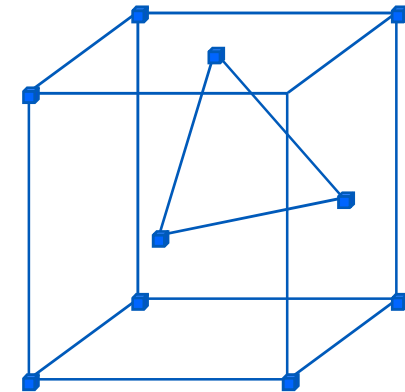
HOKE D6



BOX-BEHNKEN



HYBRID



RECHTSHAFFNER

[ Corre, 2005 ]

# Exploration du domaine : propriétés attendues

- ▶ Répartir « régulièrement » un certain nombre de points  $N$  dans l'espace  $p$ -dimensionnel des entrées ( $\chi$ ), pour construire le plan  $\Xi_N = \left( x_j^{(i)} \right)_{i=1 \dots N, j=1 \dots p}$

$N \sim 50$  à  $1000$  et  $p \sim 5$  à  $50$

- ▶ S'assurer de la robustesse de cette répartition vis-à-vis de la réduction de dimension

Règle d'or à connaître : la plupart du temps, ce sont les effets d'ordre faible qui sont influents

On va donc chercher des plans d'expériences qui répondent à ces objectifs

Question préliminaire :

**Comment définit-on les « bonnes » répartitions ?**  
**=> différents critères possibles**

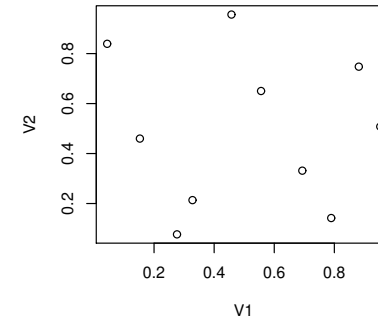
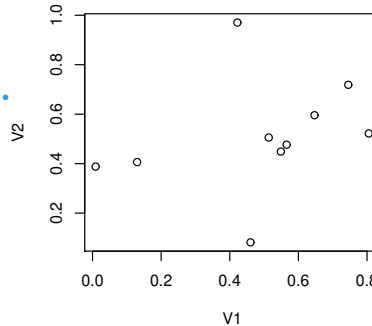
# Comment échantillonner un espace de grande dimension ?

Warning: un échantillon aléatoire pur remplit mal l'espace (surtout si  $p$  est élevé)

1. Plans « space filling » sont de bons candidats pour bien remplir l'espace

Ex:  $p = 2, N = 10$

Echant.  
Monte  
Carlo



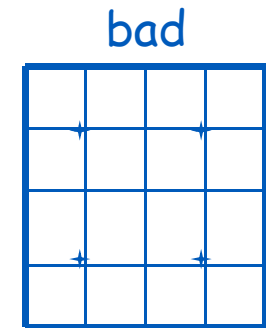
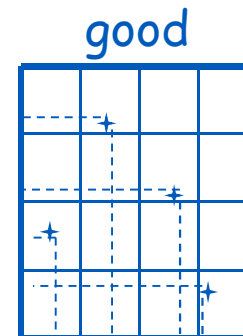
Space  
Filling  
Design

Ces plans sont basés :

- soit sur un critère de distances entre les points du plan : minimax, maximin, ...
- soit sur un critère de répartition uniforme des points (discrédance)

2. Propriété de *projections uniformes sur les marges*  
obtenue via un plan **Hypercube Latin (LHS)**  
chaque entrée est bien échantillonnée.

Ex :  $p = 2, N = 4$



3. Plans LHS optimisés pour avoir les propriétés 1 et 2



## Critères géométriques de remplissage (1/2)

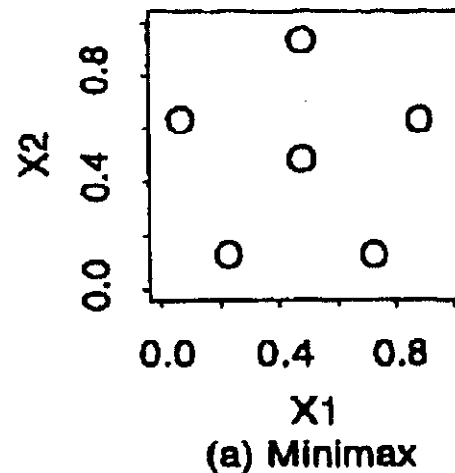
► Minimax design  $D_{MI}$  : Minimise la distance maximale entre un point du domaine et un point du plan

$$\min_D \max_x d(x, D) = \max_x d(x, D_{MI})$$

[ Johnson et al. 1990 ]  
[ Koehler & Owen 1996 ]

$$\text{where } d(x, D) = \min_{x^{(0)} \in D} d(x, x^{(0)})$$

Aucun point du domaine  $[0,1]^p$  n'est trop loin d'un point du plan  $D_{MI}$

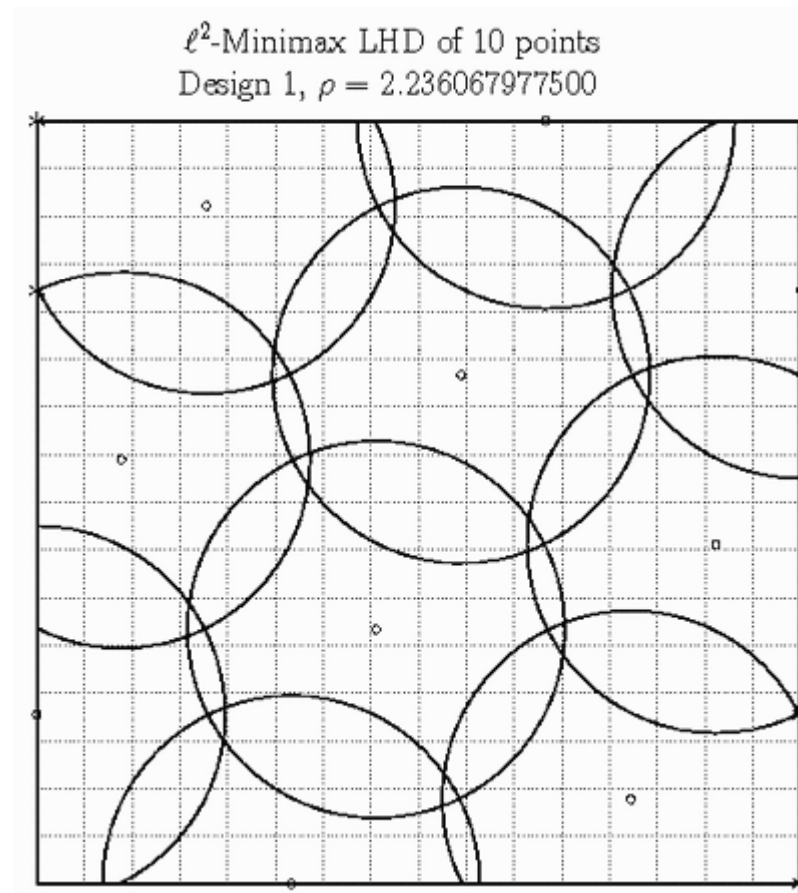


=> L'un des meilleurs plans mais trop coûteux à construire pour  $p > 3$

# Plans minimax

►  $p = 1$  ;  $X_i = (2i-1)/(2N)$  ;  $\phi_{mM} = 1 / 2N$

►  $p > 1$  : recouvrement de sphères



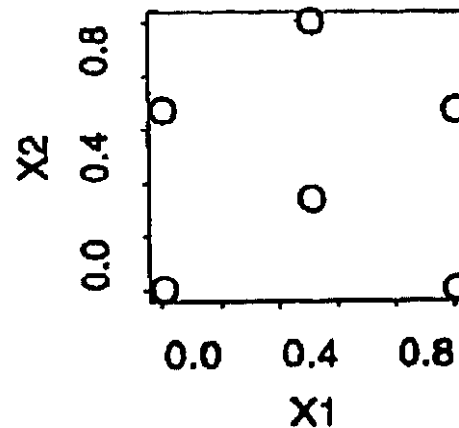
[ [www.spacefillingdesigns.nl](http://www.spacefillingdesigns.nl) ]

## Critères géométriques de remplissage (2/2)

- **Distance mindist** :  $\phi(\Xi^N) = \min_{x^{(1)}, x^{(2)} \in \Xi^N} d(x^{(1)}, x^{(2)})$  (norme  $L_2$  usuellement)

➔ **Maximin design**  $\Xi_{Mm}^N$  : maximise la distance minimale entre les points du plan

$$\max_{\Xi^N} \min_{x^{(1)}, x^{(2)} \in \Xi^N} d(x^{(1)}, x^{(2)}) = \min_{x^{(1)}, x^{(2)} \in \Xi_{Mm}^N} d(x^{(1)}, x^{(2)})$$

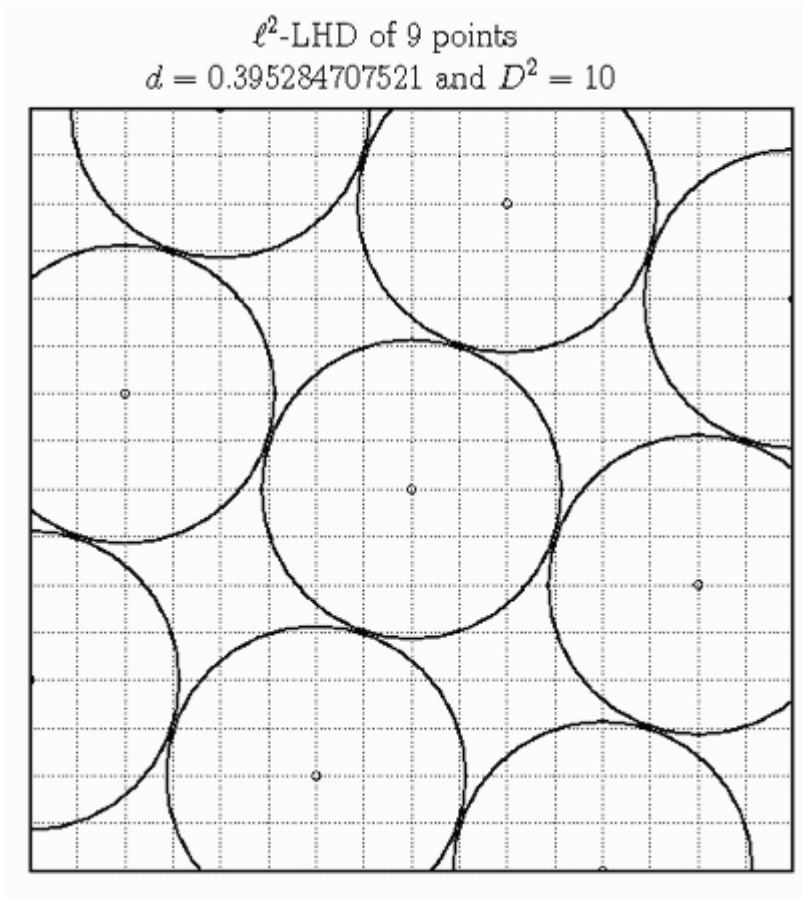


- Moyenne des distances de chaque point à son plus proche voisin,
- Mesure de recouvrement = coefficient de variation de ces distances,
- ...

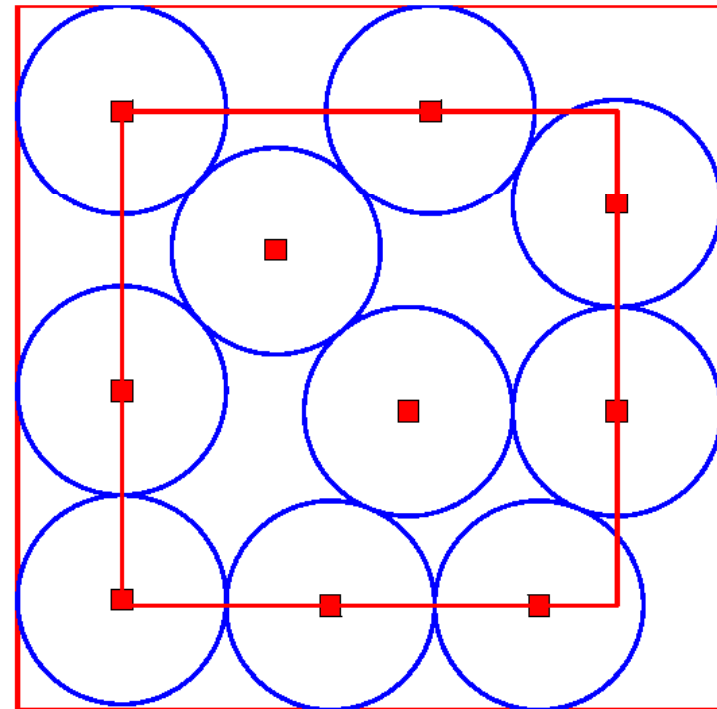
# Plans maximin

▶  $p = 1$  ;  $X_i = (i-1)/(N-1)$  ;  $\phi_{mM} = 1 / (N-1)$

▶  $p > 1$  : empilage de sphères



[ [www.spacefillingdesigns.nl](http://www.spacefillingdesigns.nl) ]



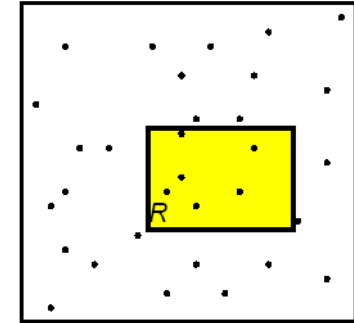
[ [www.packomania.com](http://www.packomania.com) ]

# Un critère statistique : la discrédance

Mesure la **déviatiion maximale entre la répartition des points de l'échantillon et une répartition uniforme** (~ statistique de Kolmogorov-Smirnov)

Interprétation géométrique :

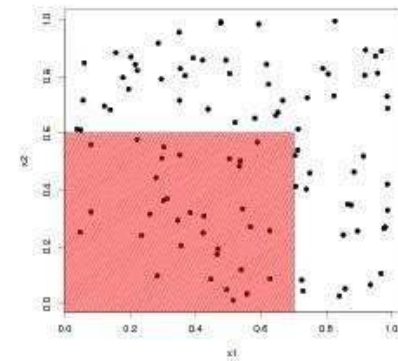
Comparaison entre le volume des intervalles du domaine et le nombre de points contenus dans ces intervalles



$$Q(\mathbf{t}) \subset \mathcal{X} = [0,1[^p, Q(\mathbf{t}) = [0, t_1[ \times [0, t_2[ \times \dots \times [0, t_p[$$

$$\text{discrédance}(\text{plan}) = \sup_{Q(\mathbf{t}) \in [0,1[^p} \left| \frac{N_{Q(\mathbf{t})}}{N} - \prod_{i=1}^p t_i \right|$$

*Plus faible est la discrédance, plus la répartition uniforme des points dans l'espace est bonne*



La discrédance intervient dans la majoration de l'erreur d'intégration d'une fonction

## Lien avec le problème de l'intégration d'une fonction

$$I = \int_{[0,1]^p} G(x) dx$$

$$\text{Monte Carlo : } I_N^{\text{MC}} = \frac{1}{N} \sum_{i=1}^N G(x^{(i)})$$

avec  $(x^{(i)})_{i=1\dots N}$  une séquence aléatoire de points dans  $[0,1]^p$

$$\mathbb{E}(I_N^{\text{MC}}) = I ; \text{Var}(I_N^{\text{MC}}) = \frac{\text{Var}(G)}{N} \Rightarrow \varepsilon = O\left(\frac{1}{\sqrt{N}}\right)$$

Propriété générale (Koksma-Hlawka inequality) :  $\varepsilon \leq V(G) \times \text{disc}(D)$

Avec une suite à faible discrédance  $D$  (séquence quasi-Monte Carlo) :

$$\varepsilon = O\left(\frac{(\ln N)^p}{N}\right)$$

Une suite est uniformément répartie sur  $[0,1]^p$  si  $\lim_{n \rightarrow \infty} \text{Disc}(D_n) = 0$

Choix bien connus : suite de Sobol', Halton, Faure, ...

## Discrédances $L_2$

Plusieurs définitions, dépendant de la norme et des intervalles considérés

$$D^*(\Xi^N) = \sup_{\mathbf{t} \in [0,1]^p} \left| \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\mathbf{x}^{(i)} \in Q(\mathbf{t})} - \text{Volume}(Q(\mathbf{t})) \right|$$

Choix permettant de faciliter les calculs : discrédance  $L^2$

[ Hickernell 1998 ]

$$\text{Discrédance } L^2 \text{ à l'origine : } D_2^*(\Xi^N) = \left[ \int_{[0,1]^p} \left[ \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\mathbf{x}^{(i)} \in Q(\mathbf{t})} - \text{Volume}(Q(\mathbf{t})) \right]^2 d\mathbf{t} \right]^{1/2}$$

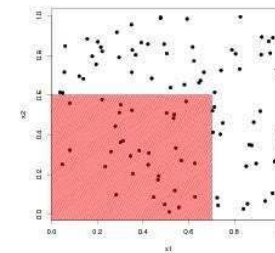
Une propriété manquante : prise en compte de l'uniformité des projections des points sur des sous-espaces de  $[0,1]^p$

=> **Discrédance  $L_2$  modifiée**

$$D_2(\Xi^N) = \left[ \sum_{u \neq \emptyset} \int_{C^u} \left[ \frac{1}{N} \sum_{i=1}^N \mathbf{1}_{\mathbf{x}_u^{(i)} \in Q_u(\mathbf{t})} - \text{Volume}(Q_u(\mathbf{t})) \right]^2 d\mathbf{t} \right]$$

avec  $u \subset \{1, \dots, p\}$

et  $Q_u(\mathbf{t})$  la projection de  $Q(\mathbf{t})$  sur  $C^u$  (cube unité de coordonnées dans  $u$ )

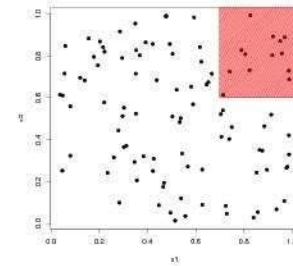


# Calculs de la discrédance en pratique

- Discrédance  $L_2$  centrée (intervalles ancrés en un sommet de l'hypercube)

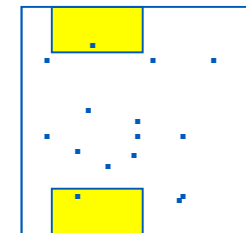
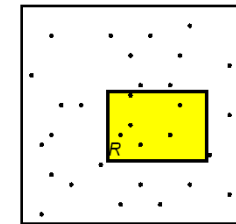
Formule analytique :

$$C^2(\mathbb{E}^N) = \left(\frac{13}{12}\right)^p - \frac{2}{N} \sum_{i=1}^N \prod_{k=1}^p \left(1 + \frac{1}{2} \left|x_k^{(i)} - \frac{1}{2}\right| - \frac{1}{2} \left|x_k^{(i)} - \frac{1}{2}\right|^2\right) + \frac{1}{N^2} \sum_{i,j=1}^N \prod_{k=1}^p \left(1 + \frac{1}{2} \left|x_k^{(i)} - \frac{1}{2}\right| + \frac{1}{2} \left|x_k^{(j)} - \frac{1}{2}\right| - \frac{1}{2} \left|x_k^{(i)} - x_k^{(j)}\right|\right)$$



- Discrédance  $L_2$  wrap-around (supprime les effets de bord en enveloppant le cube unité)

$$W^2(\mathbb{E}^N) = \left(\frac{4}{3}\right)^p + \frac{1}{N^2} \sum_{i,j=1}^N \prod_{k=1}^p \left[\frac{3}{2} - \left|x_k^{(i)} - x_k^{(j)}\right| \left(1 - \left|x_k^{(i)} - x_k^{(j)}\right|\right)\right]$$





# Construction d'une suite à discrédance faible

Séquence 1D de Van der Corput

Écriture d'un nombre  $i$  en base  $b$  :

$$\text{Base } b : i = (\dots a_4 a_3 a_2 a_1 a_0)_b$$

$$\text{Système décimal : } i = \sum_{j=0}^m a_j b^j, 0 \leq a_j \leq b-1$$

Pour obtenir un nombre à l'intérieur de l'intervalle unité :

$$y = (0.a_0 a_1 a_2 a_3 \dots)_b$$

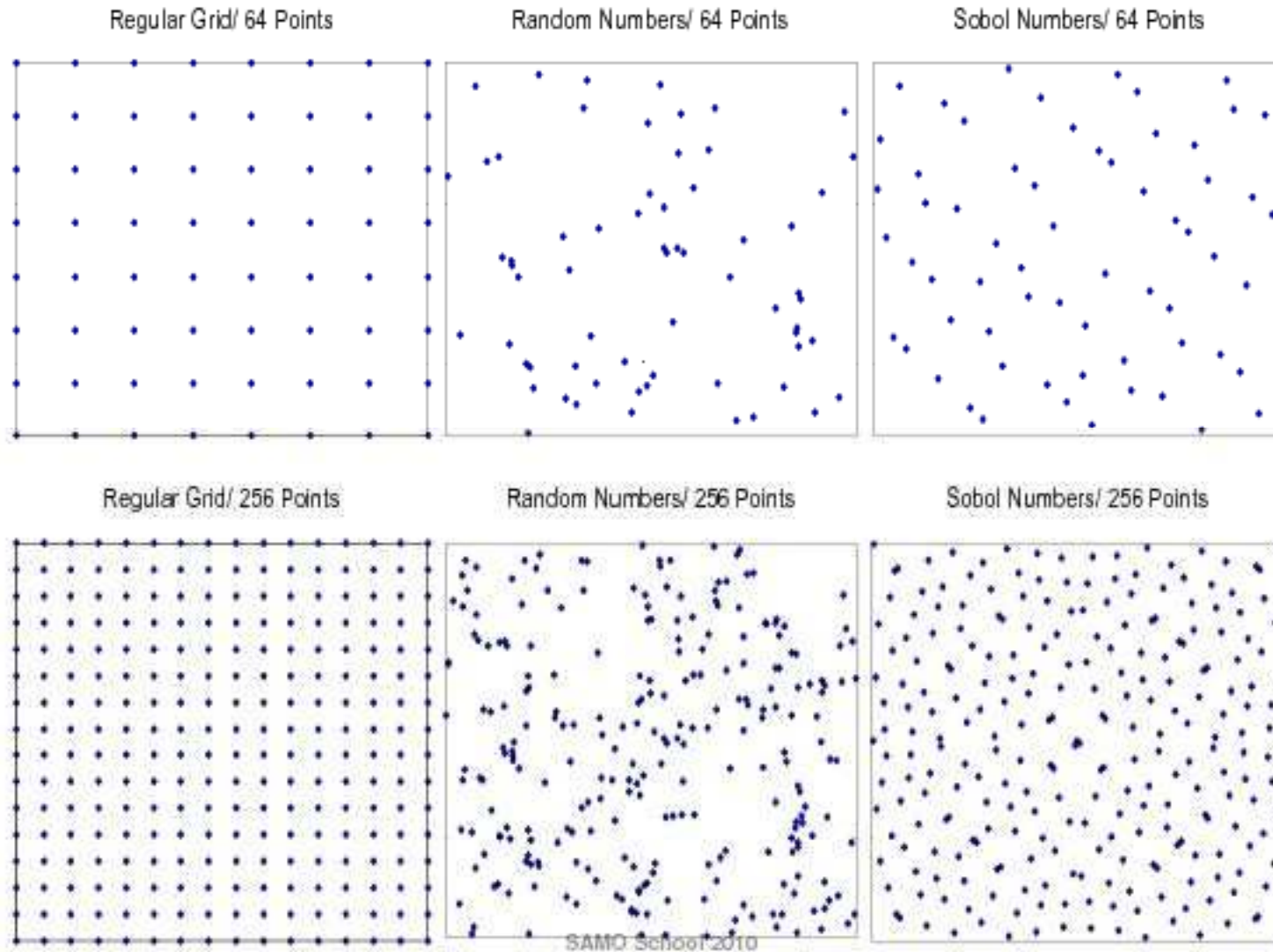
$$\text{Système décimal : } h(i, b) = \sum_{j=0}^m a_j b^{-j-1}$$

**Suite de Halton (dimension  $p$ )** : pour chaque dimension, prendre une base (nombre premier) différente

$$\{h(i, 2), h(i, 3), \dots, h(i, b_p)\}$$

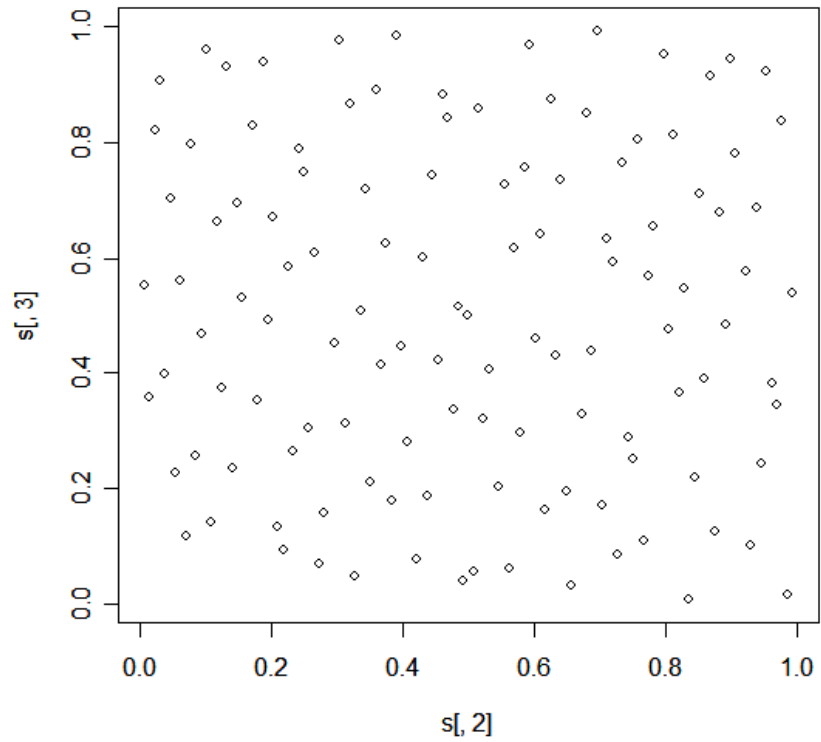
# Exemples en 2D : Suite de Sobol vs. échantillon aléatoire vs. grille régulière

[ Kucherenko, 2010 ]

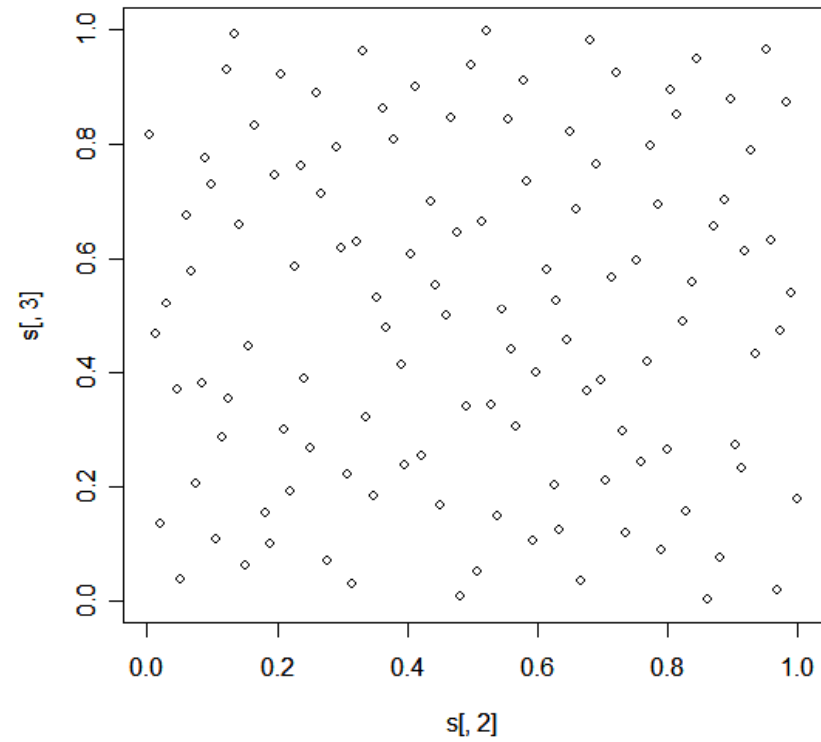


# Exemple - N = 150 - Dimension = 8

Sobol

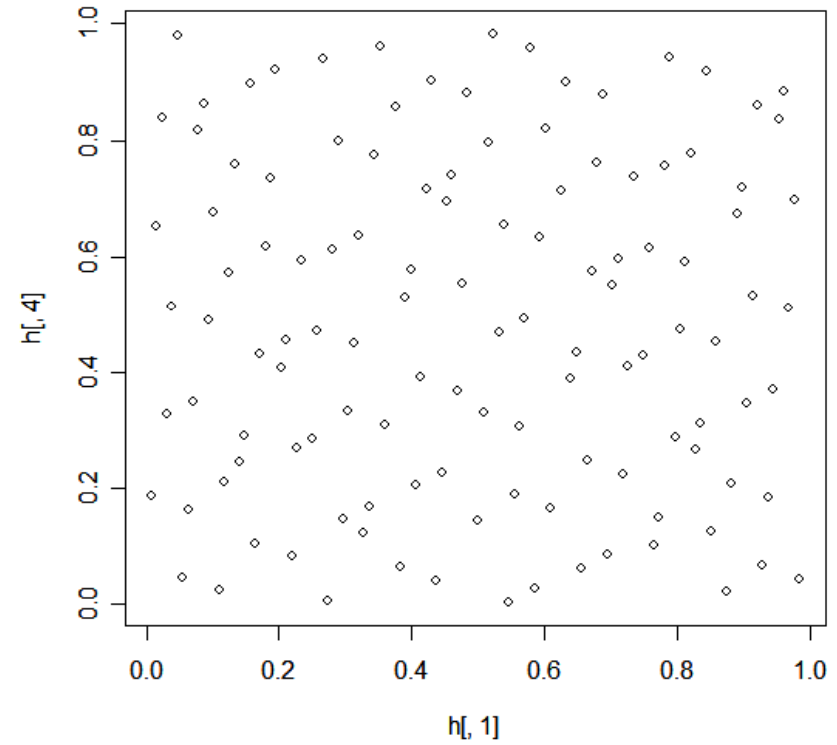
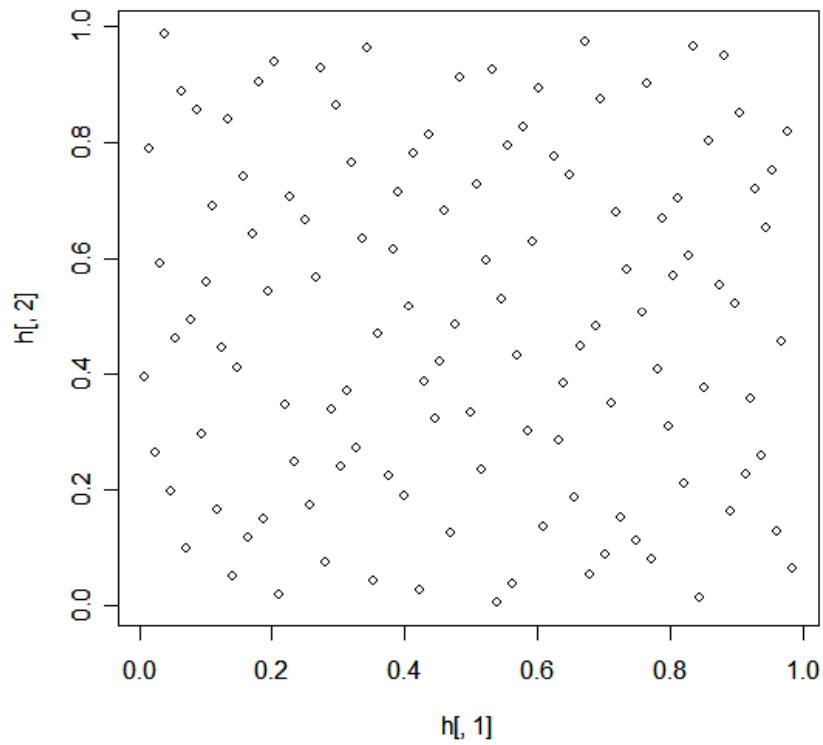


Sobol scrambling Owen



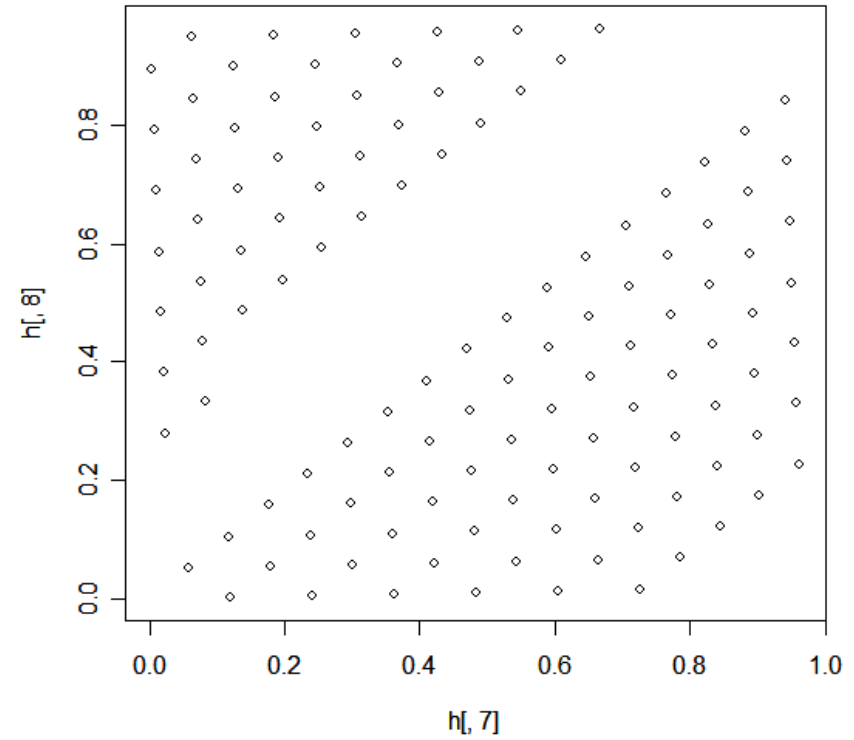
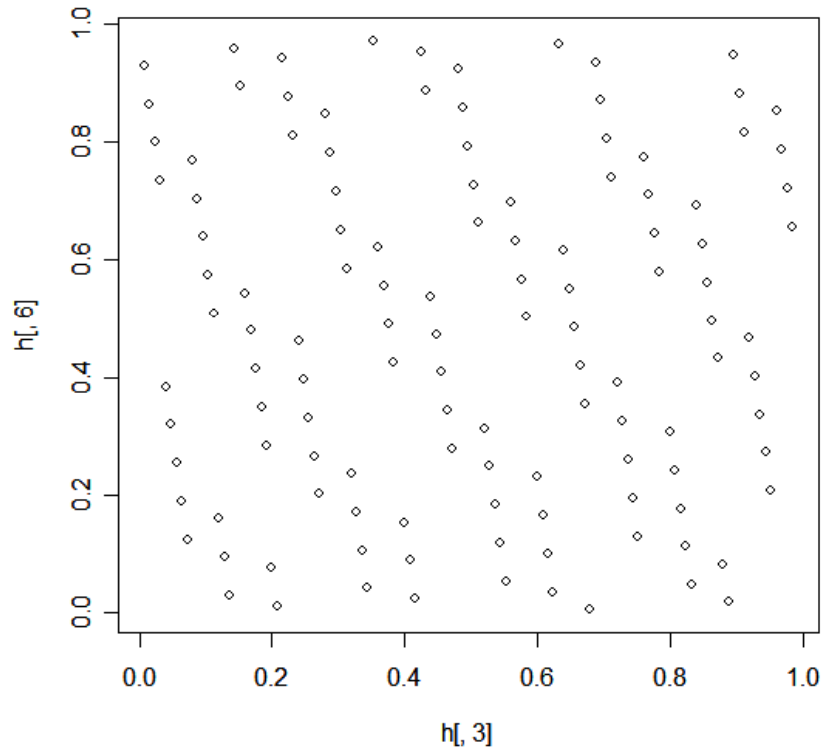
# Exemple - N = 150 - Dimension = 8

Halton



# Pathologies sur les projections 2D

Halton



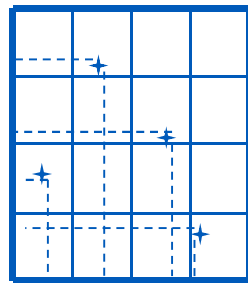
## Propriété importante : robustesse en sous-projections

Le code de calcul  $G(\mathbf{X})$  a souvent de faibles dimensions effectives :

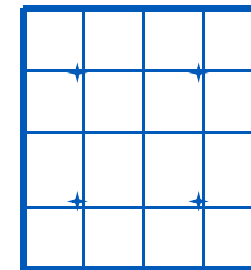
- au sens de la troncature (nombre de variables influentes  $\ll p$ )
- au sens de la superposition (+ **gd ordre d'interaction influente  $\ll p$** )

Il est donc nécessaire que le SFD conserve les propriétés space-filling sur les sous-espace de faibles dimensions (par importance : en dimension  $p'=1$ , puis  $p'=2, \dots$ )

- $p' = 1$  – Le plan LHS nous assure de bonnes projections 1D



good



bad

- $p' \geq 2$  - Les discrécances  $L^2$  modifiées (centrée, wrap-around, ...) prennent en compte les sous-projections dans leurs définitions

*Par contre, les critères de distances vus précédemment n'assurent pas de robustesse en sous-projections*

# Les Hypercubes Latins (LHS)

[ McKay et al. 1979 ]

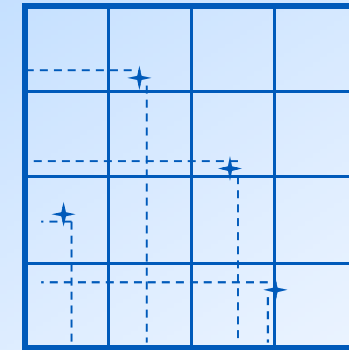
Souvent, seules quelques variables sont influentes



**Propriété :** Projections uniforme sur les marginales

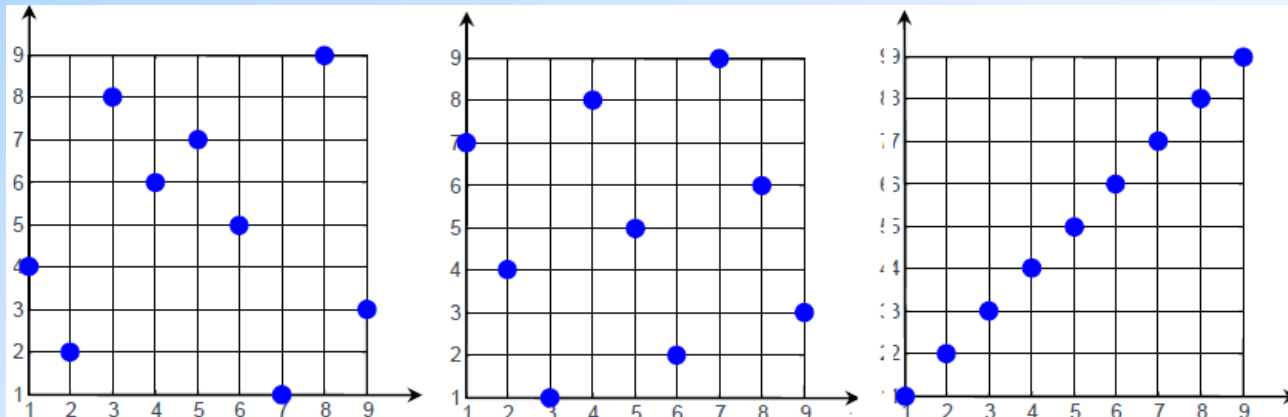
**Principe :**  $p$  variables,  $N$  points  $\Rightarrow LHS(p,N)$

- On divise chaque dimension en  $N$  intervalles
- Tirage aléatoire d'un point dans chaque strate :



Exemple :  $p=2, N=4$

Chacun des niveaux est pris une fois et une seule par chaque facteur  
 $\Rightarrow$  chacune des colonnes du plan est donc une permutation de  $\{ 1,2,\dots,N \}$



Choix du LHS par optimisation de différents critères

- Remplissage
- Indépendance
- Uniformité

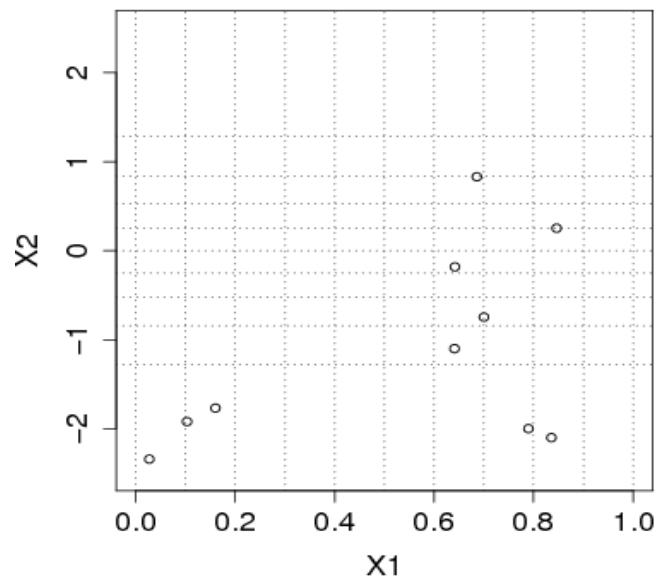
# Algorithme LHS( $p, N$ ) – Méthode de Stein

```
ran = matrix(runif(N*p), nrow=N, ncol=p) #tirage de N x p valeurs selon loi U[0,1]
x = matrix(0, nrow=N, ncol=p)          # construction de la matrice x

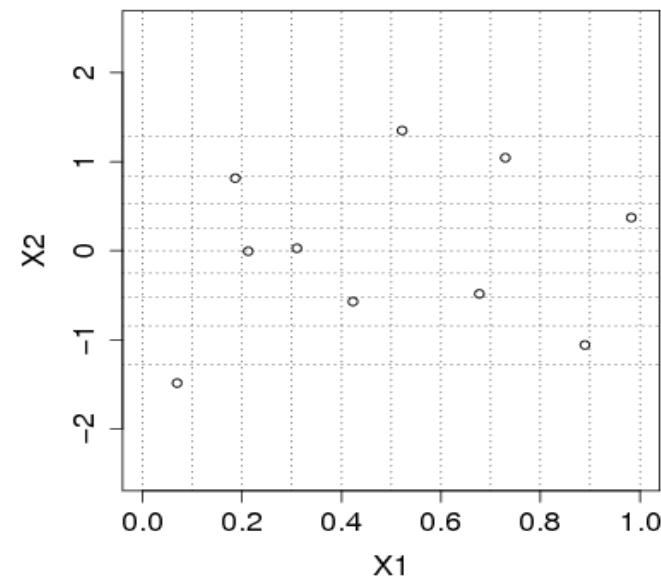
for (i in 1:p) {
  idx = sample(1:N) #vecteur de permutations des entiers {1,2,...,N}
  P = (idx-ran[,i]) / N    # vecteur de probabilités
  x[,i] <- quantile_selon_la_loi (P)  }
```

Exemple :  $p=2$ ,  $N=10$ ,  $X_1 \sim U[0,1]$ ,  $X_2 \sim N(0,1)$

(a) Simple Random Sampling



(b) Latin Hypercube Sampling





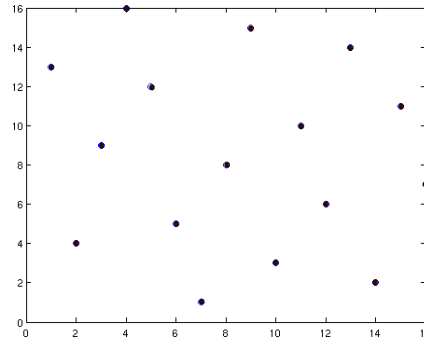
# Optimisation de LHS => Space-filling LHS

[ Park 1993;  
Morris & Mitchell 1995 ]

Méthode simple : générer un grand nombre (par ex. 1000) de LHS différents.  
Puis, choisir le meilleur au sens d'un critère  $\phi(.)$  (« space filling »)

Exemple : LHS(2,16)

critère maximin



MAIS : le nombre de LHS possibles est énorme :  $(N!)^P$

Méthode par algos d'optimisation (ex : minimisation de  $\phi(.)$  par recuit simulé) :

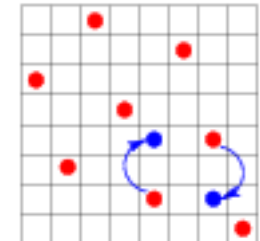
1. Initialisation d'un plan  $\Xi$  (LHS initial) et d'une température  $T$

2. Tant que  $T > 0$  :

1. générer un voisin  $\Xi_{new}$  de  $\Xi$  (voisin = permutation de 2 composantes dans une colonne)

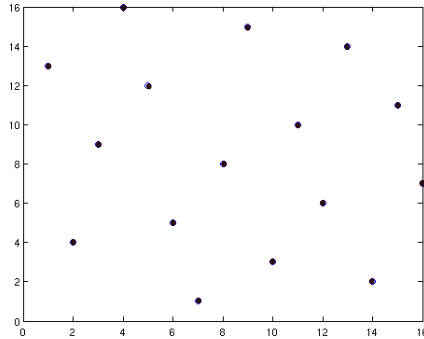
2. remplacer  $\Xi$  par  $\Xi_{new}$  avec la probabilité  $\min\left(\exp\left[-\frac{\phi(\Xi_{new}) - \phi(\Xi)}{T}\right], 1\right)$

3. faire décroître  $T$



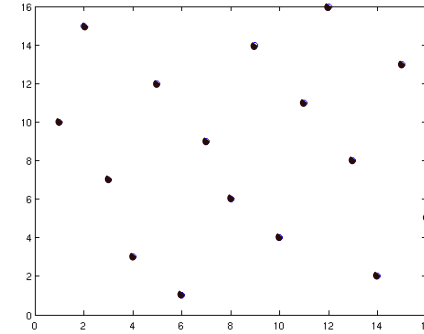
3. Critère d'arrêt =>  $\Xi$  contient la solution optimale générée par le recuit simulé

# Exemples



Maximin LHS

Exemple : LHS(2,16)



LHS à faible discrédance

Sur des tests numériques ( $N=100$ ), on voit qu'à partir de la dimension 10, le LHS maximin se comporte comme un LHS standard (dimension 40 pour un LHS à discrédance centrée faible)

**Cela confirme la pertinence de la discrédance  $L^2$  centrée en terme de sous-projection**

En pratique, par exemple, on peut lancer plusieurs ( $\sim 10$ ) optimisations suivant un critère de discrédance, et on compare les résultats suivant un critère de distance afin de choisir le meilleur plan

# Récapitulatif sur la planification d'expériences numériques

**Enjeu** : Échantillonner un espace de grande dimension de manière « optimale »  
(obtenir le plus d'informations possible sur le comportement de la sortie  $Z / \mathbf{X} \in \mathbb{R}^p$ )

Problème : un échantillon aléatoire pur (Monte Carlo) remplit mal l'espace

1. Plans « space filling » sont de bons candidats pour bien remplir l'espace :

- Basés sur un critère de distances entre les points du plan (minimax, maximin, ...),  
*justification théorique pour la construction du métamodèle de krigage*
- Basés sur un critère de répartition uniforme des points (discrépance) ;  
*justification théorique lorsque l'on calcule la moyenne de la fonction  $f(\mathbf{X})$*

2. Propriété de projections uniformes sur les marges peut être obtenue via les **plans hypercubes latins** (LHS)

3. Il est possible de coupler les 2 propriétés en construisant des LHS optimisés

# Synthèse : propriétés des Space Filling Designs

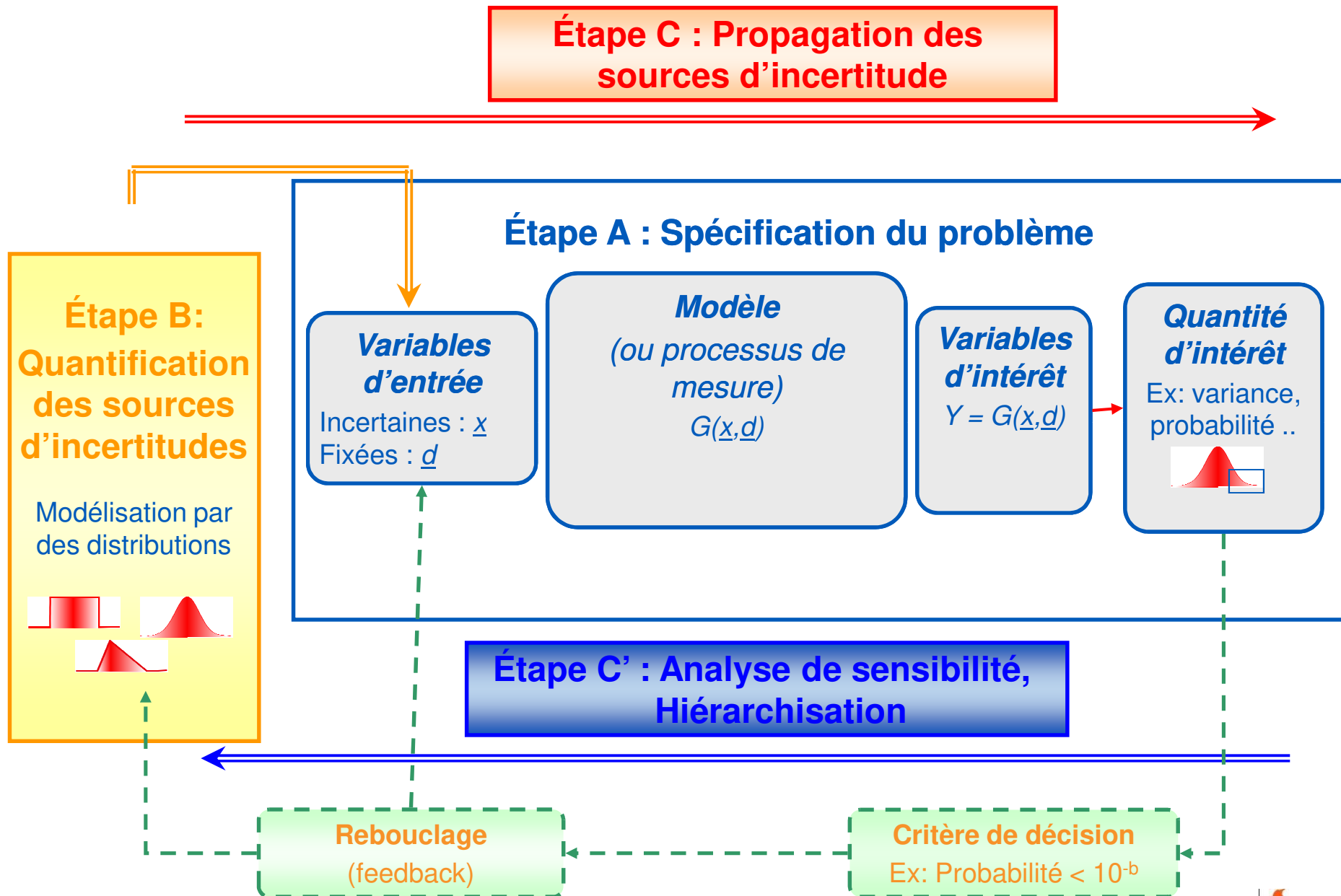
	Orthogonalité des colonnes	Motifs, alignement	Plan séquentiel	Réduction dimension
Monte Carlo	Non	Non	Oui	Oui
Faure, Halton, Sobol	Oui	Oui, en dim. élevée	Oui	Oui, mais pathologies
LHS à discréc centrée faible	Non	?	Non	Oui
LHS maximin	Non	Oui	Non	Non

- A part le Monte Carlo, tous ces plans sont dits « à bon remplissage de l'espace »
- les suites à faible discrécance remplissent le cube unité de manière extrêmement régulière (et parfois trop => techniques de scrambling)

# Plan du cours 2

1. Introduction
2. Planification d'expériences numériques
- 3. Méthodes d'analyse de sensibilité**

# Approches quantitatives : schéma générique introductif



# Deux notions pour l'analyse de sensibilité

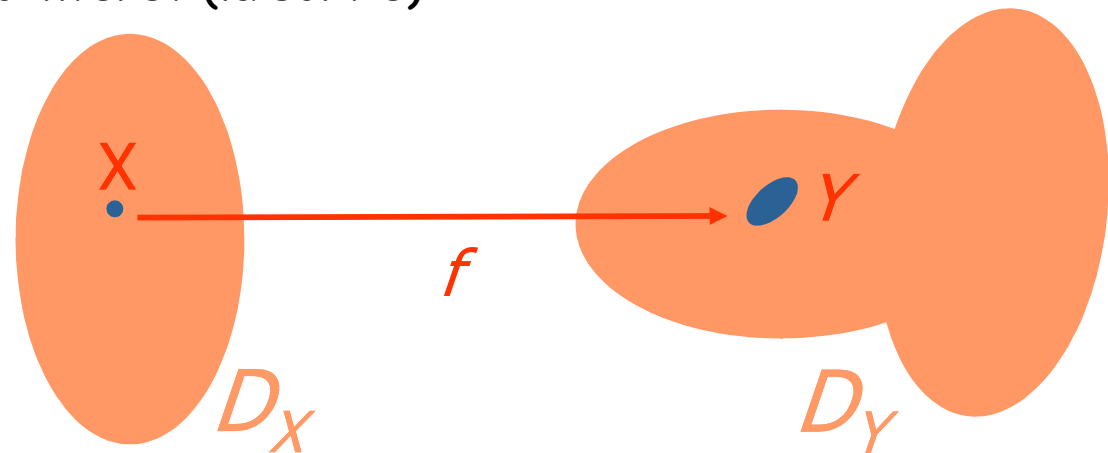
◆ **sensibilité**, par exemple  $\partial Y / \partial X_i$

Donne une idée de la manière dont peut répondre la réponse en fonction de **variations potentielles** des facteurs

◆ « **contribution** » = **sensibilité** x **importance**, par exemple  $\frac{\partial Y}{\partial X_i} \sigma(X_i)$

Permet de déterminer le **poids** d'une variable d'entrée (ou groupe de variables) sur l'incertitude de la variable d'intérêt (la sortie)

**Distinction  
local vs. global**



*C'est l'impact vis-à-vis de la quantité d'intérêt qui est étudié :*

- *variabilité globale (variance, entropie, ...)*
- *quantile, probabilité de dépassement, ...*

# Méthodes locales

- **Cumul quadratique**

$$Y(\mathbf{X}) = Y(\mathbf{X}^0) + \sum_{i=1}^p \left( \frac{\partial Y}{\partial X_i} \right)_{\mathbf{X}^0} (X_i - X_i^0)$$

- ▶ Contribution de  $X_i$  sur la sortie  $Y$ : 
$$C(X_i) = \frac{(\partial Y / \partial X_i)^2 \sigma_{X_i}^2}{\sum_{i=1}^p (\partial Y / \partial X_i)^2 \sigma_{X_i}^2}$$

Calcul des dérivées par différences finies, dérivées exactes, différentiation automatique de codes (fortran, C)

- **Autres indices de sensibilité locaux**

- ▶ Indices issus des méthodes FORM/SORM mesurent les sensibilités par rapport au dépassement d'un seuil (autour du point de conception)



# Deux grands objectifs de l'analyse de sensibilité globale

## ► Réduction de l'incertitude de la sortie d'un modèle par hiérarch. des sources

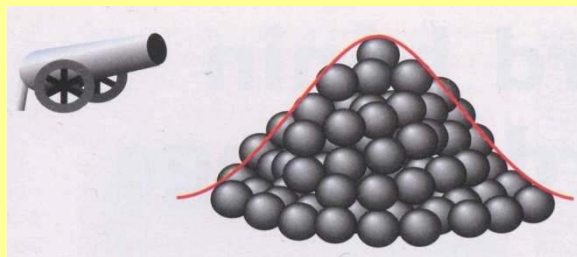
- Variables à fixer pour obtenir la **plus forte réduction** (ou une réduction donnée) **de l'incertitude de la sortie**
  - Variables les plus influentes dans un domaine de valeurs de la sortie
- si réductibles, priorité de R&D

## ► Simplifier un modèle

- **détermination des variables non influentes**, que l'on pourra fixer
- construire un modèle simplifié, un métamodèle - cf. Loïc

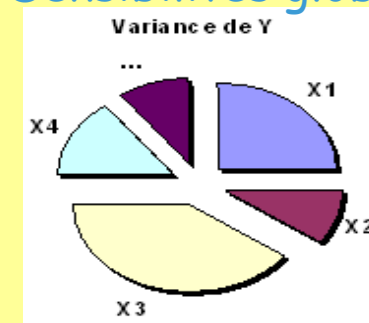
### Approche statistique (échantillonnage Monte Carlo)

#### Propagation



coût élevé pour probas faibles

#### Sensibilités globales



coût  $\propto p$

# Rappels sur l'analyse de sensibilité

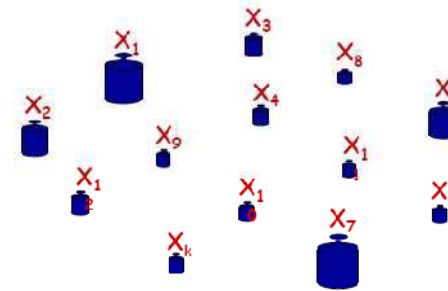
**Enjeu** : décomposer la variabilité globale de la sortie  $Z = G(\mathbf{X})$  (due aux incertitudes sur les entrées  $\mathbf{X}$ ) en part de variabilité due à chaque entrée  $X_i, i=1, \dots, p$

Problème : comme pour la planification, le coût en nombre  $N$  d'évaluations de  $G(\cdot)$  dépend de  $p$

## 1. Le criblage (screening) :

- plans d'expériences classiques,
- plans d'expériences numériques

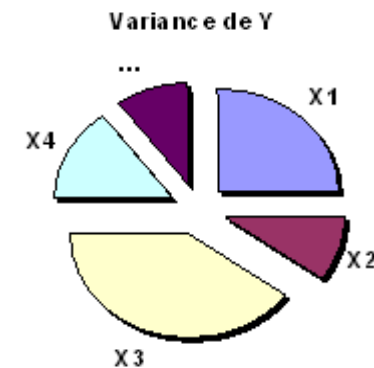
$$N \sim p/2 \text{ à } 10 p$$



## 2. Les mesures d'influence globale :

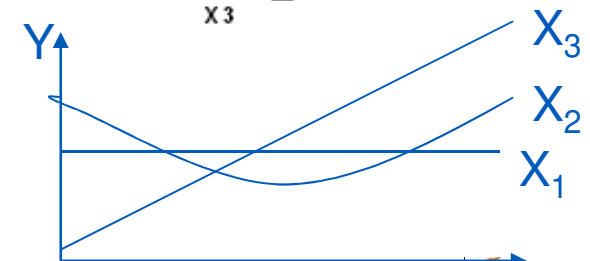
- corrélation/régression sur les valeurs/rangs,
- décomposition de la variance fonctionnelle (Sobol),

$$N \sim 2p \text{ à } 1e4 p$$



## 3. Exploration fine des sensibilités - $N \sim 10p$ à $100 p$

- Méthodes de lissage (param./non param.)
- Métamodèles



# Criblage avec $N < p$ (plans supersaturés)

Beaucoup d'entrées ( $p \gg 10$ ) et code coûteux

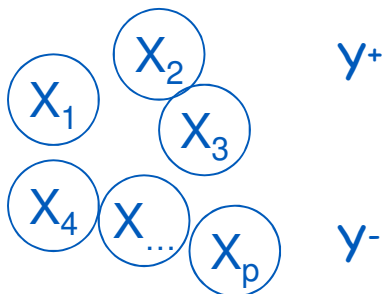
Contrainte : réaliser moins de calculs qu'il n'y a d'entrées

Hypothèses :

- Nombre d'entrées influentes  $\ll p$
- Monotonie du modèle, pas d'interaction entre entrées
- Connaissance du sens de variation de la sortie / chaque entrée
- Possibilité de planification séquentielle

Exemple: méthode des bifurcations séquentielles

2 calculs



# Criblage avec $N < p$ (plans supersaturés)

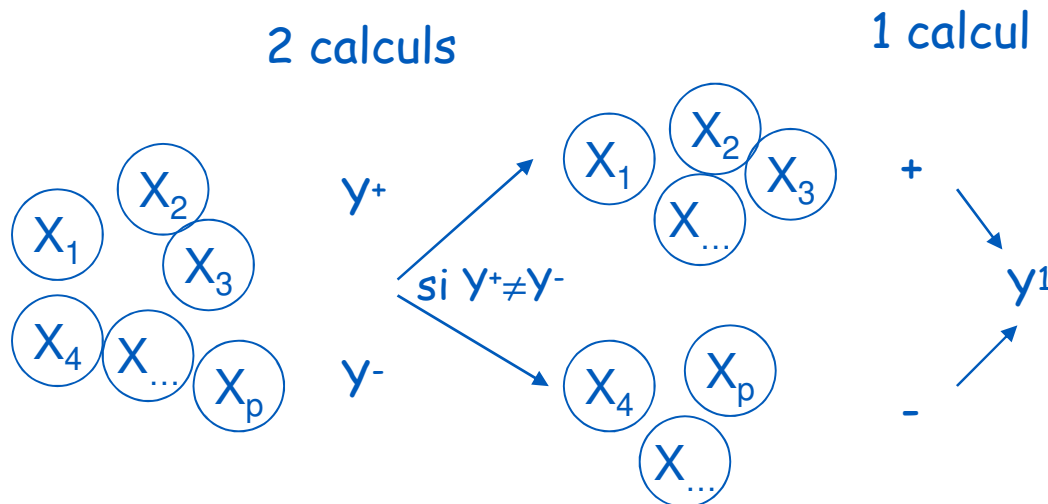
Beaucoup d'entrées ( $p \gg 10$ ) et code coûteux

Contrainte : réaliser moins de calculs qu'il n'y a d'entrées

Hypothèses :

- Nombre d'entrées influentes  $\ll p$
- Monotonie du modèle, pas d'interaction entre entrées
- Connaissance du sens de variation de la sortie / chaque entrée
- Possibilité de planification séquentielle

Exemple: méthode des bifurcations séquentielles



# Criblage avec $N < p$ (plans supersaturés)

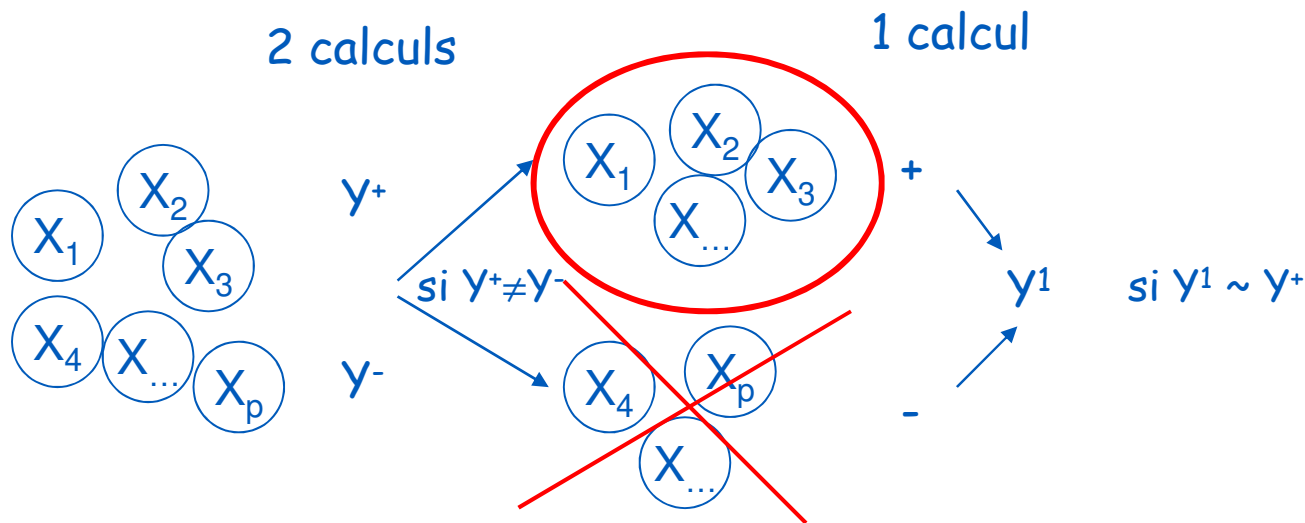
Beaucoup d'entrées ( $p \gg 10$ ) et code coûteux

Contrainte : réaliser moins de calculs qu'il n'y a d'entrées

Hypothèses :

- Nombre d'entrées influentes  $\ll p$
- Monotonie du modèle, pas d'interaction entre entrées
- Connaissance du sens de variation de la sortie / chaque entrée
- Possibilité de planification séquentielle

Exemple: méthode des bifurcations séquentielles



# Criblage avec $N < p$ (plans supersaturés)

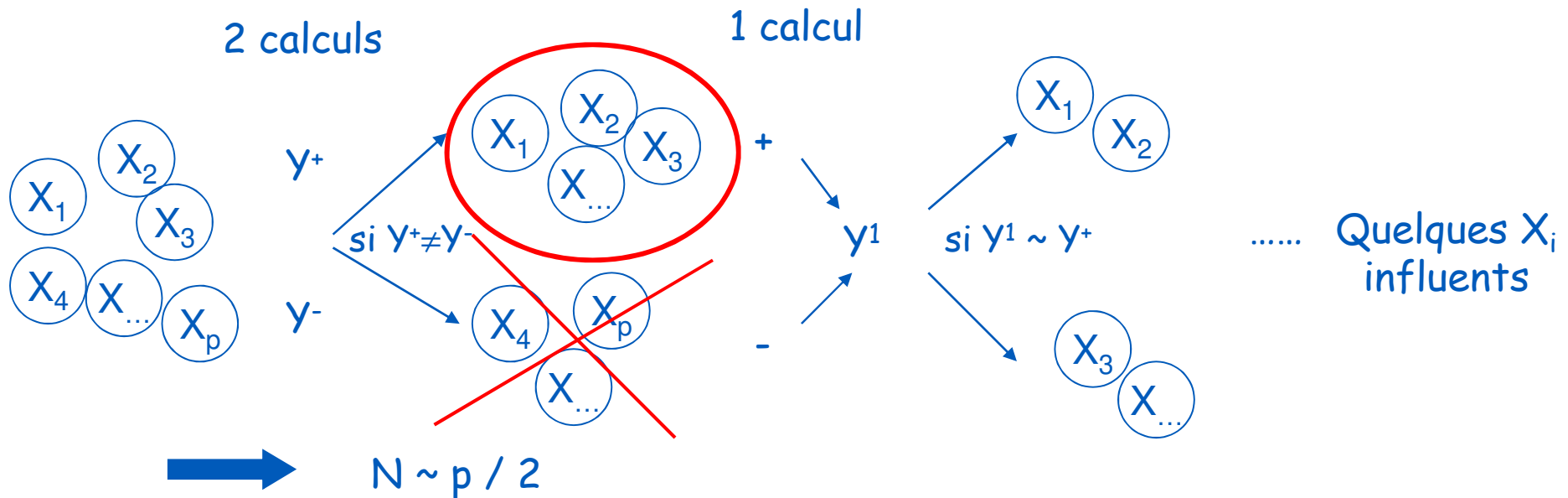
Beaucoup d'entrées ( $p \gg 10$ ) et code coûteux

Contrainte : réaliser moins de calculs qu'il n'y a d'entrées

Hypothèses :

- Nombre d'entrées influentes  $\ll p$
- Monotonie du modèle, pas d'interaction entre entrées
- Connaissance du sens de variation de la sortie / chaque entrée
- Possibilité de planification séquentielle

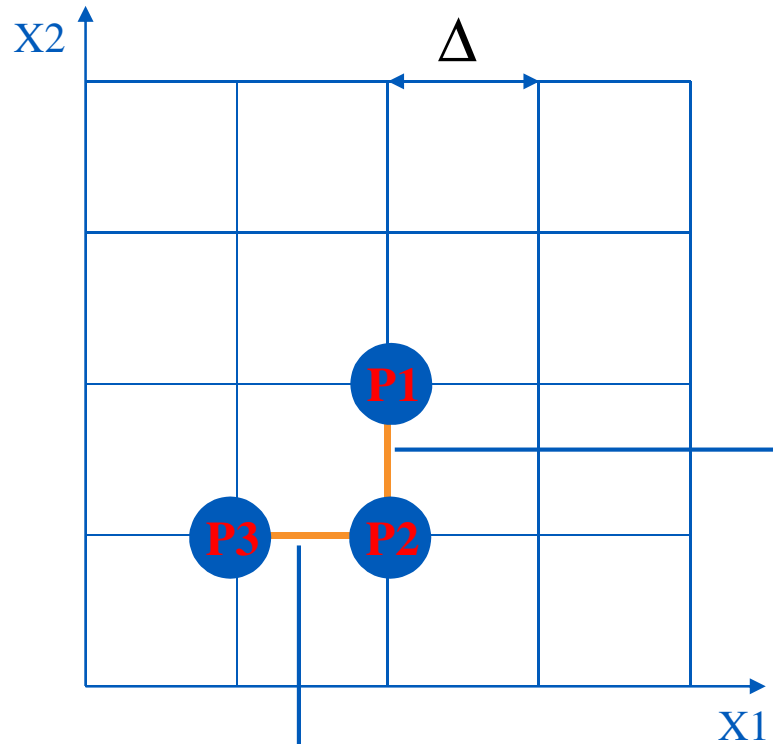
Exemple: méthode des bifurcations séquentielles



# Un criblage plus informatif : la méthode de Morris

- Méthode de « **screening** » (criblage)
  - un modèle comportant beaucoup de variables d'entrée est difficile à explorer ...  
... mais souvent, seulement quelques entrées sont influentes
  - objectif qualitatif : identifier **rapidement** ces entrées
- La méthode de [*Morris 91*] permet de classer les entrées en trois groupes selon leurs **effets** :
  1. effets négligeables
  2. effets linéaires et sans interaction
  3. effets non linéaires et/ou avec interactions
- Pas d'hypothèses sur le modèle...  
... mais mieux vaut une certaine régularité...

# Méthode de Morris : plan d'expériences



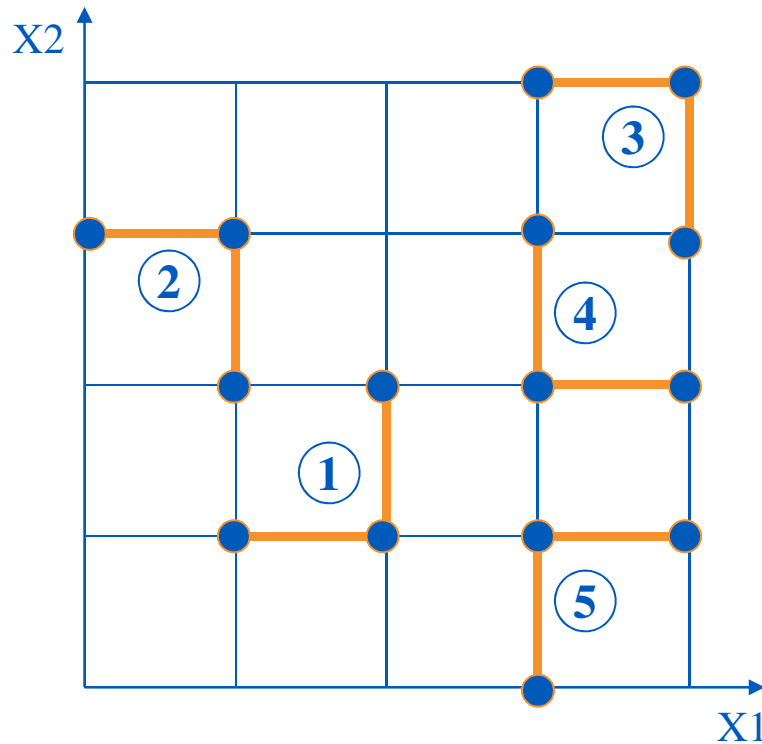
$$d_{X_1} = \frac{f(P_3) - f(P_2)}{\Delta}$$

$$d_{X_2} = \frac{f(P_2) - f(P_1)}{\Delta}$$

- Discrétisation de l'espace
- Nécessite p+1 expériences
- OAT (One-at-A-Time)
- Permet de calculer un effet élémentaire pour chaque entrée



## Méthode de Morris : plan d'expériences (suite)



- Le plan d'expériences est répété  $R$  fois (au total :  $N = R \cdot (p+1)$  expériences)

- Ceci donne  $R$ -échantillons pour chaque effet élémentaire

$$\{d_{X1}^i\}_{i=1\dots R}$$

$$\{d_{X2}^i\}_{i=1\dots R}$$

- Mesures de sensibilité

Moyenne des effets

$$\mu_i^* = E(|d_{X_i}|)$$

Dispersion des effets

$$\sigma_i = \sigma(d_{X_i})$$

## Méthode de Morris : mesures de sensibilité

- $\mu_i^* = E(|d_{X_i}|)$  est une mesure de la **sensibilité** :

valeur importante → effets importants (en moyenne)  
→ modèle sensible aux variations de l'entrée

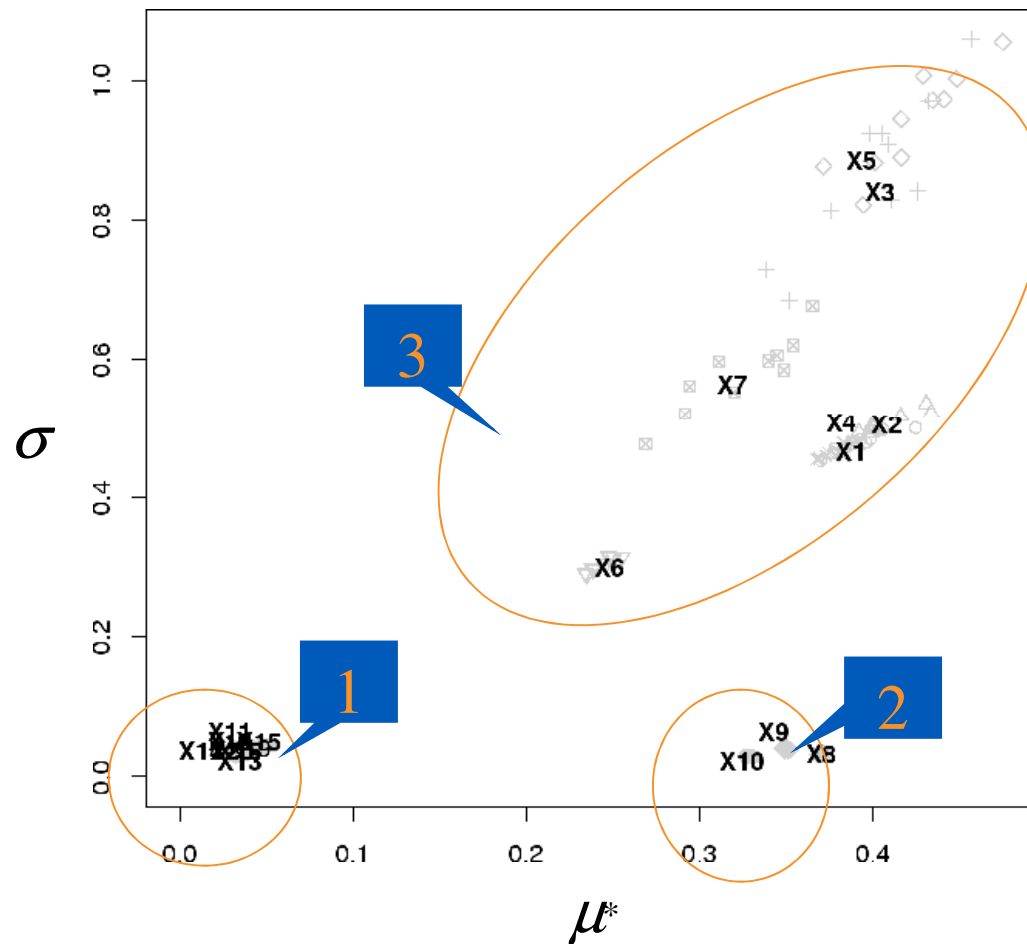
- $\sigma_i = \sigma(d_{X_i})$  est une mesure des **interactions**  
et des **effets non linéaires** :

valeur importante → effets différents les uns des autres  
→ effets qui dépendent de la valeur :

- soit de l'entrée elle-même : effet non linéaire
- soit des autres entrées : interaction

(la méthode ne permet pas de distinguer les 2 cas)

# Méthode de Morris : exemple



20 facteurs  
210 simulations  
→ Graphe ( $\mu^*$ ,  $\sigma$ )

On distingue les 3 groupes:

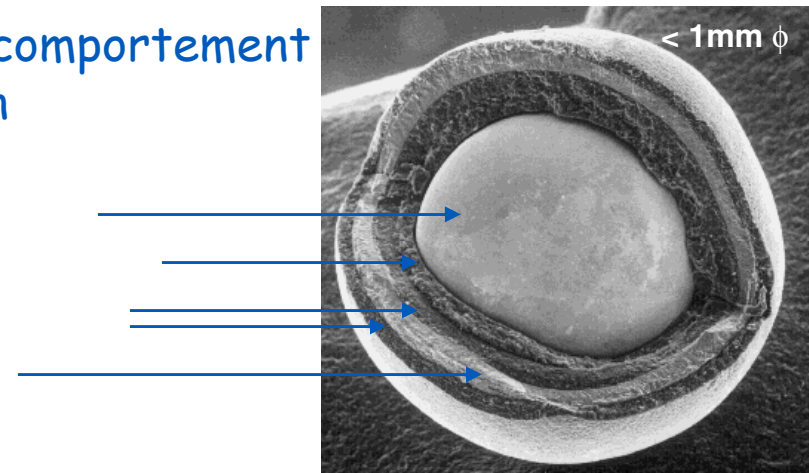
1. Effets négligeables
2. Effets linéaires
3. Effets non linéaires et/ou avec interactions

Cas test : fonction non monotone de Morris (source Saltelli)

## Exemple : code combustible HTR

Code de calcul ATLAS (CEA) : simulation du comportement du combustible à particules sous irradiation

Noyau de matière fissile  
Carbone pyrolytique poreux  
Carbone pyrolytique dense  
Carbure de Silicium



Nombre de particules dans un réacteur : de  $10^9$  à  $10^{10}$  !


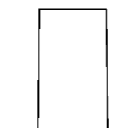
Sources de contamination : rupture de particules

→ Etudes de fiabilité [Cannamela 07]

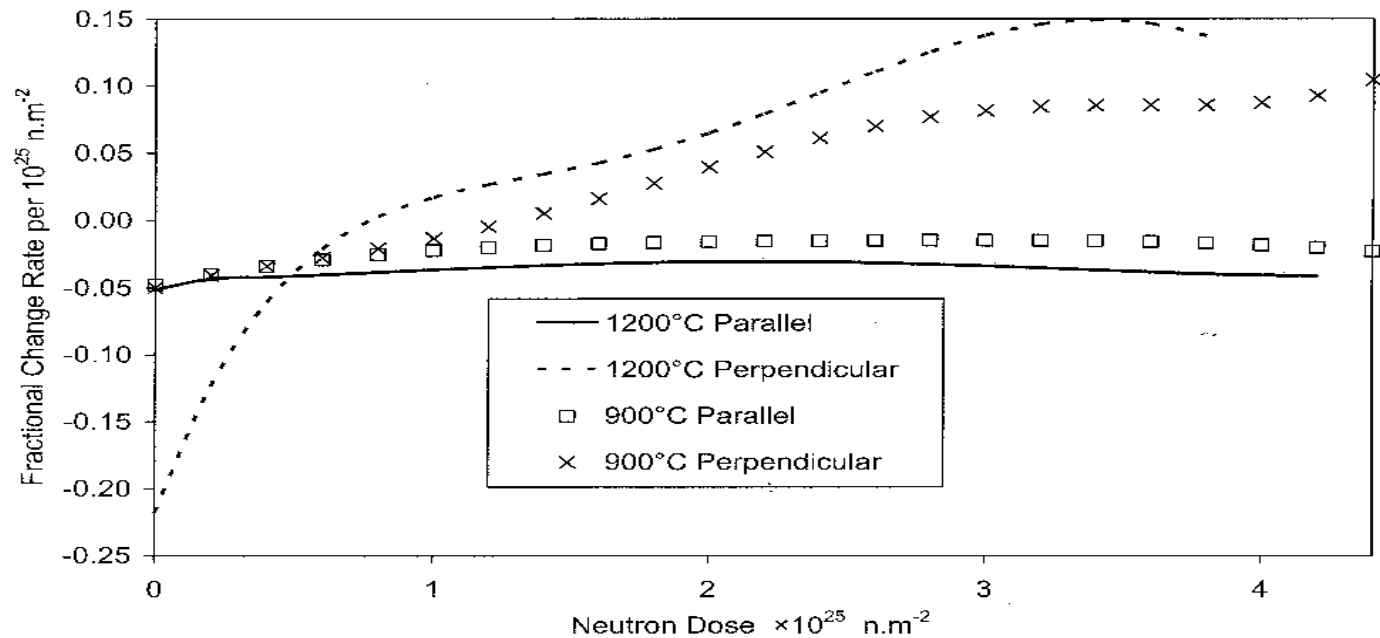
La rupture d'une particule peut être provoquée par la rupture des couches denses externes (IPyC, SiC, OPyC)

Les réponses sont choisies pour être **représentatives** du phénomène de rupture : contraintes orthoradiales maximales dans les couches externes

### 3 sources d'incertitudes en entrée

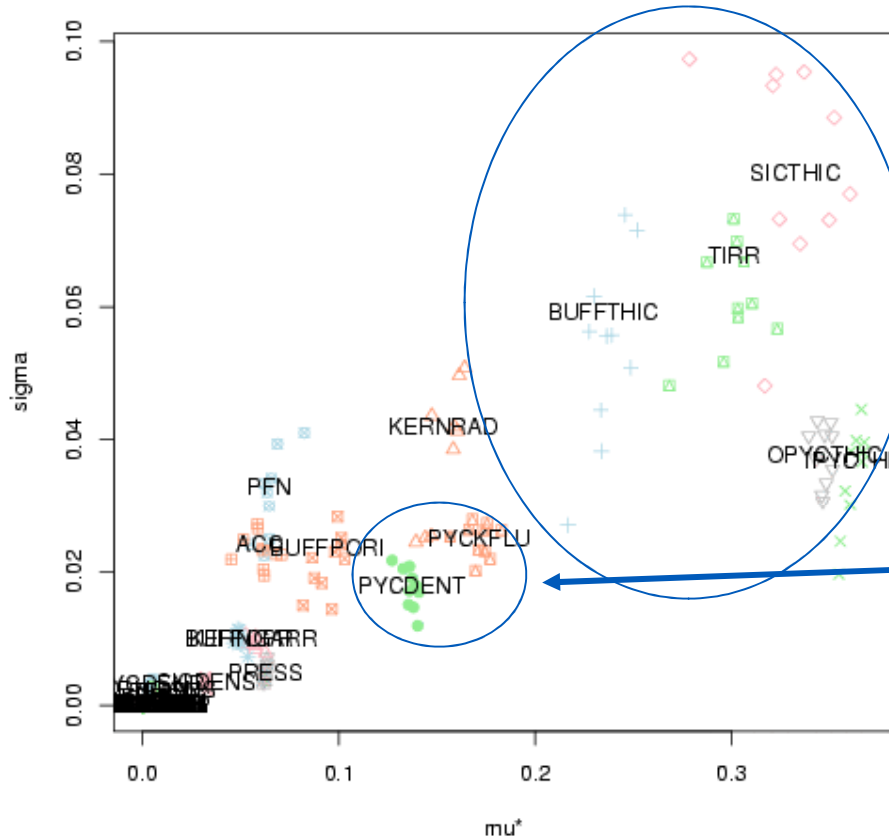
- 10 paramètres de fabrication des particules (épaisseurs, ...)  
Spécifications de fabrication → lois normales tronquées 
- 5 paramètres d'irradiation (température, ...)  
Intervalle [min,max] → lois uniformes 
- 28 lois de comportement (fonctions des température, flux, ...)  
Avis d'expert → constantes multiplicatives (de loi U[0.95,1.05])

*Exemple :  
loi de densification  
du PyC*



# Résultats de Morris

$p = 43$  entrées, 20 répétitions,  $n = 860$  calculs, coût unitaire  $\sim 1$  mn  $\rightarrow$  14h



Grande sensibilité à ces entrées (épaisseurs, température d'irradiation)  
Interactions faibles

Les lois sur le fluage et la densification du PyC sont les lois auxquelles le code est le plus sensible

## Conclusion :

La méthode de Morris donne une idée de la manière dont peut répondre la sortie en fonction de **variations potentielles** des entrées...

**→ Utile pour identifier les entrées potentiellement influentes**

# Rappels sur l'analyse de sensibilité

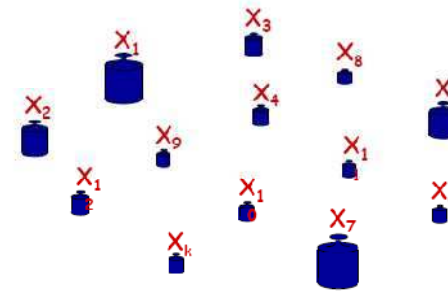
**Enjeu** : décomposer la variabilité globale de la sortie  $Z = G(\mathbf{X})$  (due aux incertitudes sur les entrées  $\mathbf{X}$ ) en part de variabilité due à chaque entrée  $X_i, i=1, \dots, p$

Problème : comme pour la planification, le coût en nombre  $N$  d'évaluations de  $G(\cdot)$  dépend de  $p$

## 1. Le criblage (screening) :

- plans d'expériences classiques,
- plans d'expériences numériques

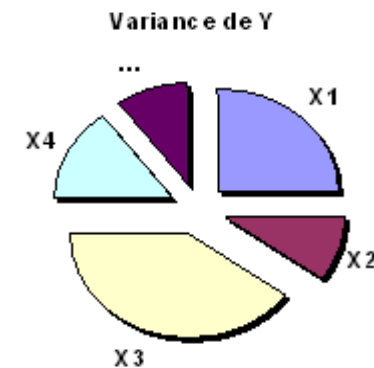
$$N \sim p/2 \text{ à } 10 p$$



## 2. Les mesures d'influence globale :

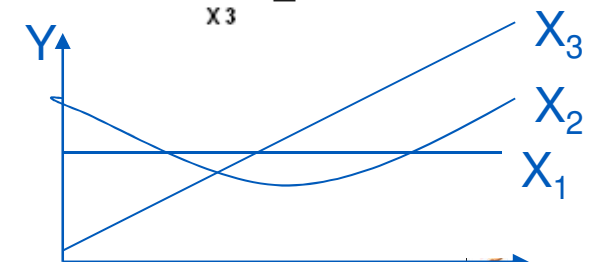
- corrélation/régression sur les valeurs/rangs,
- décomposition de la variance fonctionnelle (Sobol),

$$N \sim 2p \text{ à } 1e4 p$$



## 3. Exploration fine des sensibilités - $N \sim 10p$ à $100 p$

- Méthodes de lissage (param./non param.)
- Métamodèles



## Analyse de sensibilité pour 1 sortie scalaire

Échantillon  $(\mathbf{X}, Y(\mathbf{X}))$  de taille  $N > p$ , de préférence de taille  $N \gg p$   
Étape préliminaire : visualisation graphique (par ex : scatterplots)



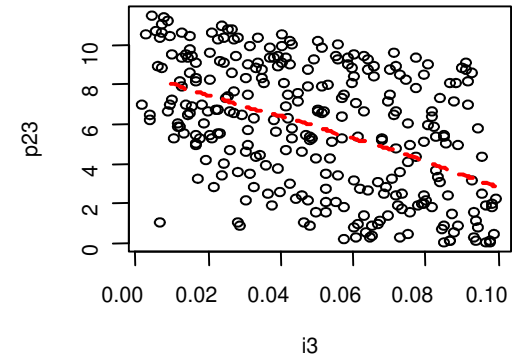
# Représentation graphique : scatterplots

Mesure le caractère linéaire du nuage de points

$N$  calculs

Grappe Sortie / chaque entrée

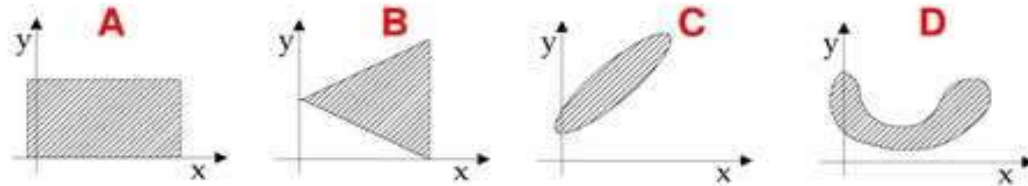
Exemple :  
 $N = 300$



$$\rho = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

$$\hat{\rho} = \frac{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

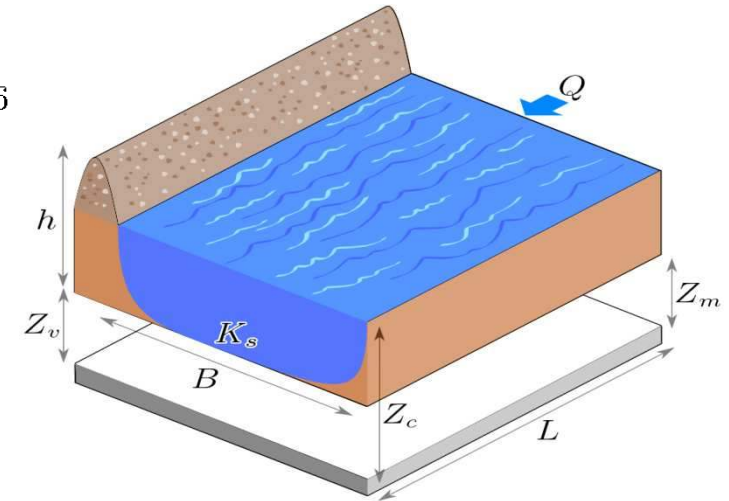
▪ Nuage de points : exemples



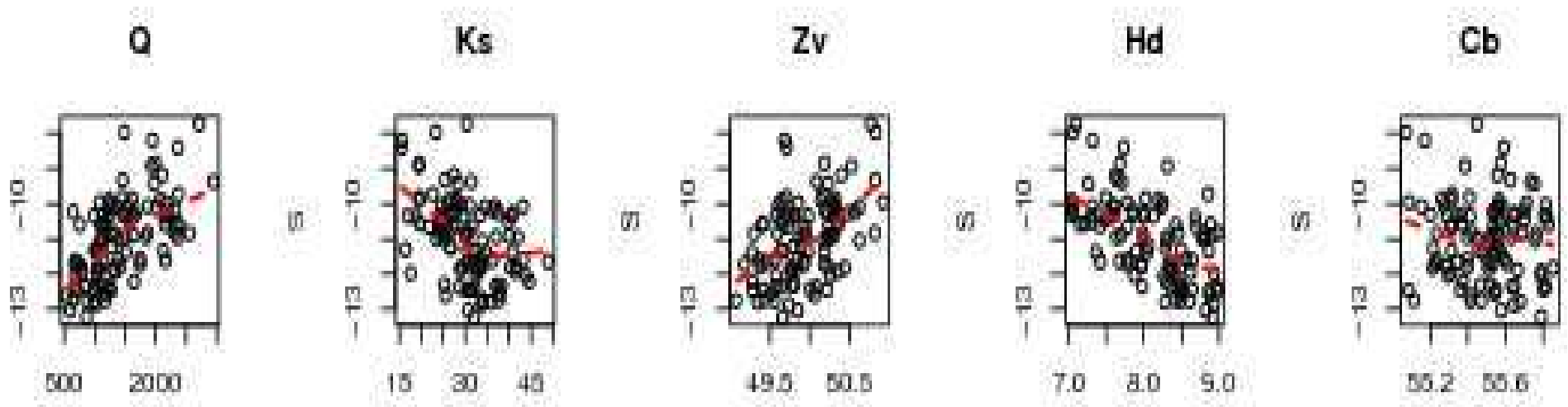
- 1 : corrélation non linéaire
- 2 : absence de liaison en moyenne mais pas en dispersion
- 3 : corrélation linéaire
- 4 : absence de liaison

# Modèle de crues - Scatterplots – Sortie S

$$S = Z_v + H - H_d - C_b \text{ avec } H = \left( \frac{Q}{BK_s \sqrt{\frac{Z_m - Z_v}{L}}} \right)^{0.6}$$



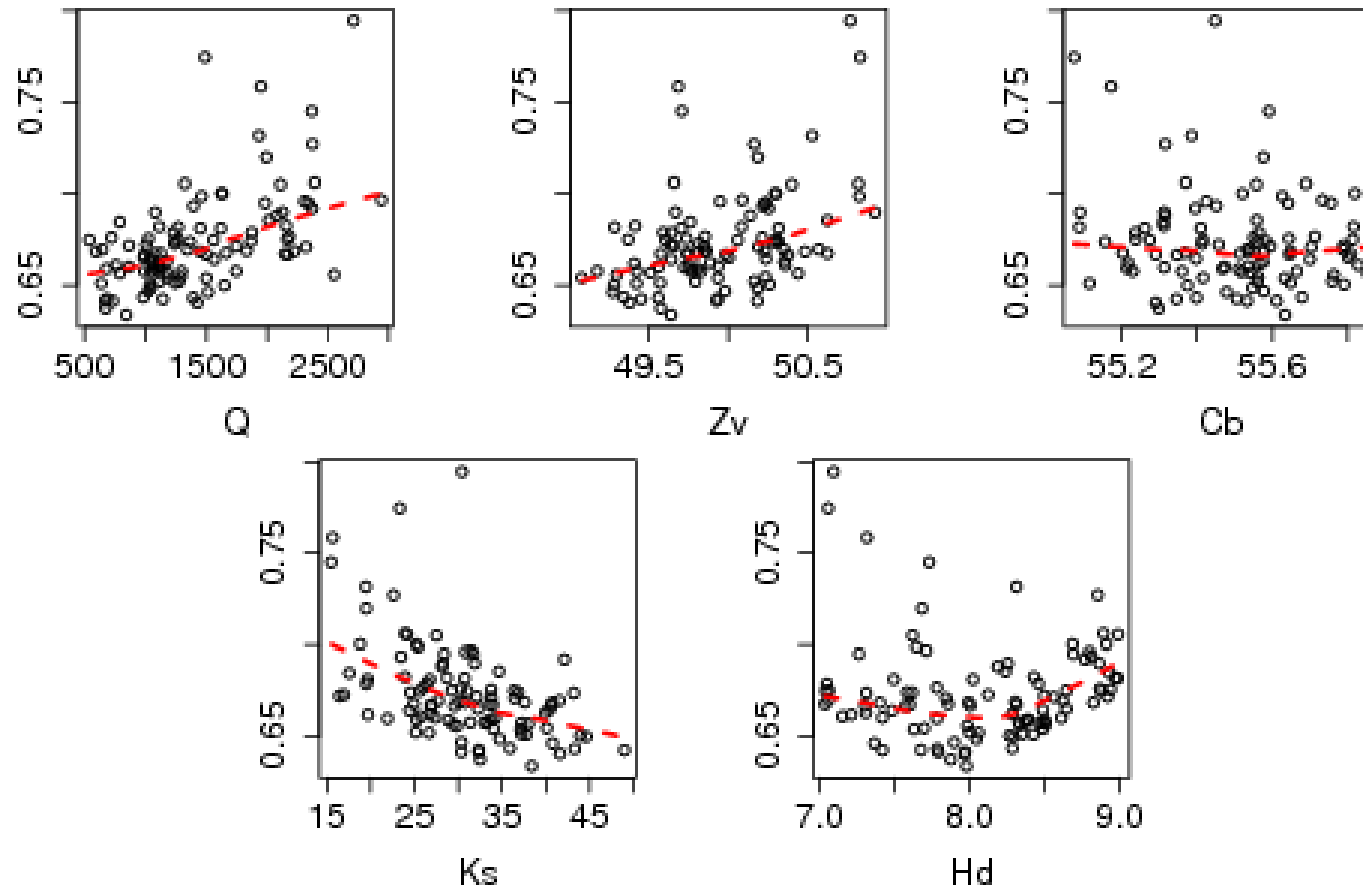
Echantillon Monte Carlo -  $N = 100$



## Modèle de crues - Scatterplots – Sortie Cp

$$C_p = \mathbb{1}_{S>0} + \left\{ 0.2 + 0.8 \left[ 1 - \exp \left( -\frac{1000}{S^4} \right) \right] \right\} \mathbb{1}_{S \leq 0} \\ + \frac{1}{20} (H_d \mathbb{1}_{H_d > 8} + 8 \mathbb{1}_{8 \leq H_d}) ,$$

Monte Carlo -  $N = 100$



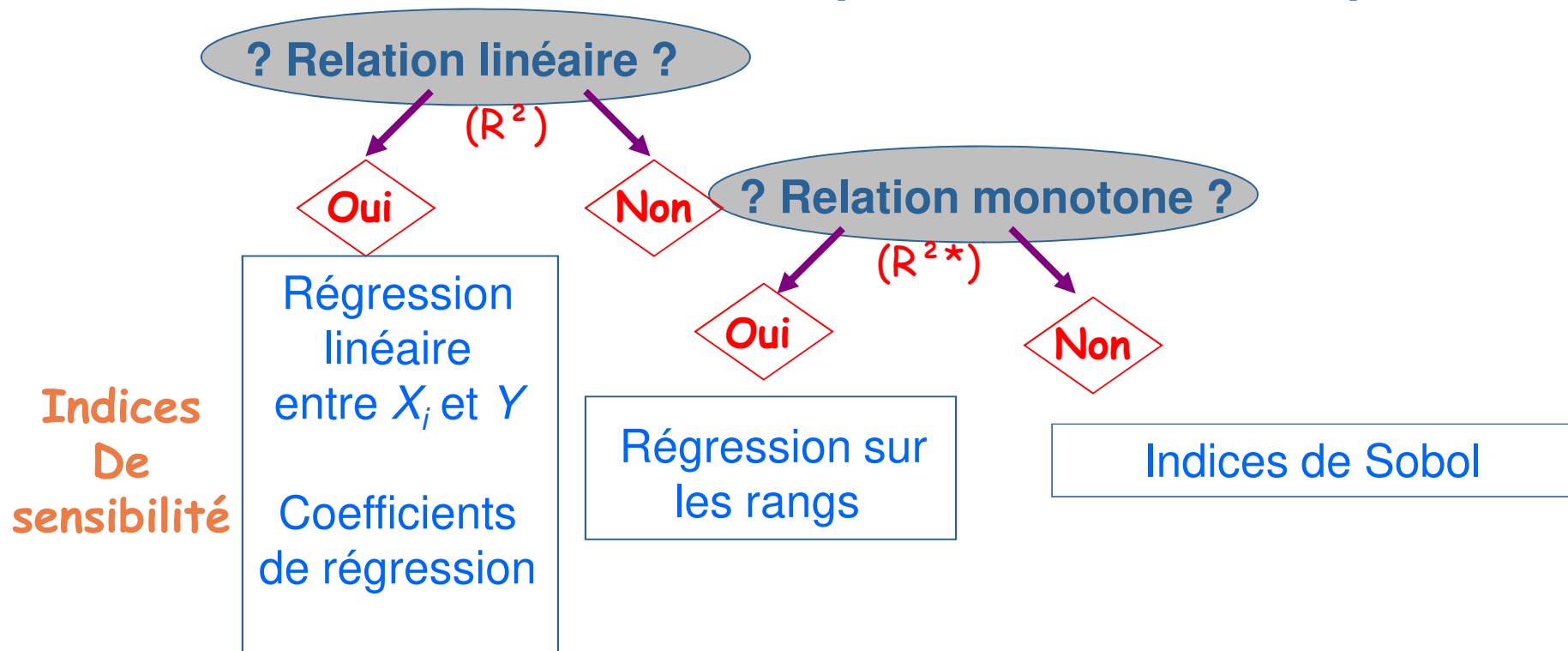
**Limite majeure : analyse seulement les relations du premier ordre et pas les effets des interactions entre les entrées**

# Analyse de sensibilité pour 1 sortie scalaire

Échantillon  $(X, Y(X))$  de taille  $N > p$ , de préférence de taille  $N \gg p$   
Étape préliminaire : visualisation graphique (par ex : scatterplots)

## Méthodologie d'analyse de sensibilité quantitative

[Saltelli et al. 00, Helton et al. 06]



# Indices de sensibilité dans le cas d'une relation entrées/sortie linéaire

Variables d'entrées  $\mathbf{X} = (X_1, \dots, X_p)$  indépendantes

Echantillon :  $N$  réalisations de  $(\mathbf{X}, Y)$

$$Y = \beta_0 + \sum_{i=1}^p \beta_i X_i$$

► L'indice SRC :  $\text{SRC}(X_i) := \beta_i \sqrt{\frac{\text{Var}(X_i)}{\text{Var}(Y)}}$

Le signe de  $\beta_i$  donne le sens de variation de  $Y / X_i$

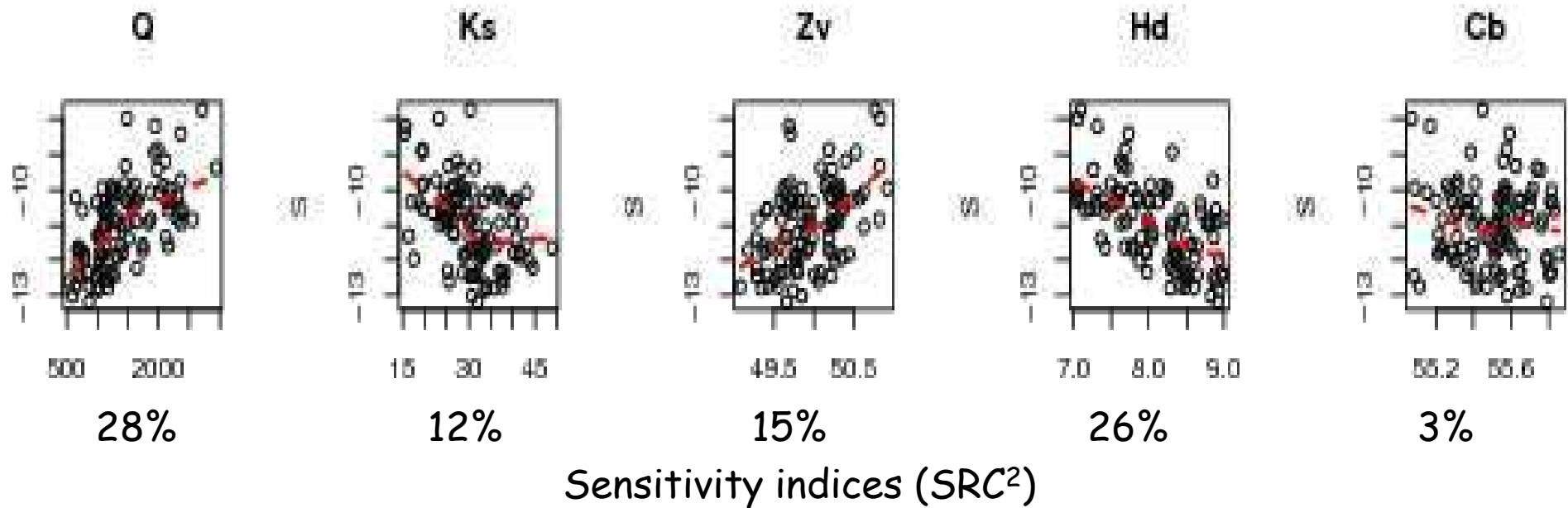
► Similaire au **coefficient de corrélation linéaire (Pearson)**

► Validité du modèle linéaire via les diagnostics de la régression et **le  $R^2$**  :  $R^2 = 1 - \frac{\sum_{i=1}^N (\hat{Y}_i - Y_i)^2}{\sum_{i=1}^N (Y_i - \bar{Y})^2}$

► On a  $R^2 = \sum_{i=1}^p \text{SRC}^2(X_i)$  , ce qui permet d'interpréter aisément les SRC

# Flood model - Output S

Monte Carlo sample -  $N=100$



Le modèle linéaire est valide ( $R^2=0.99$ )

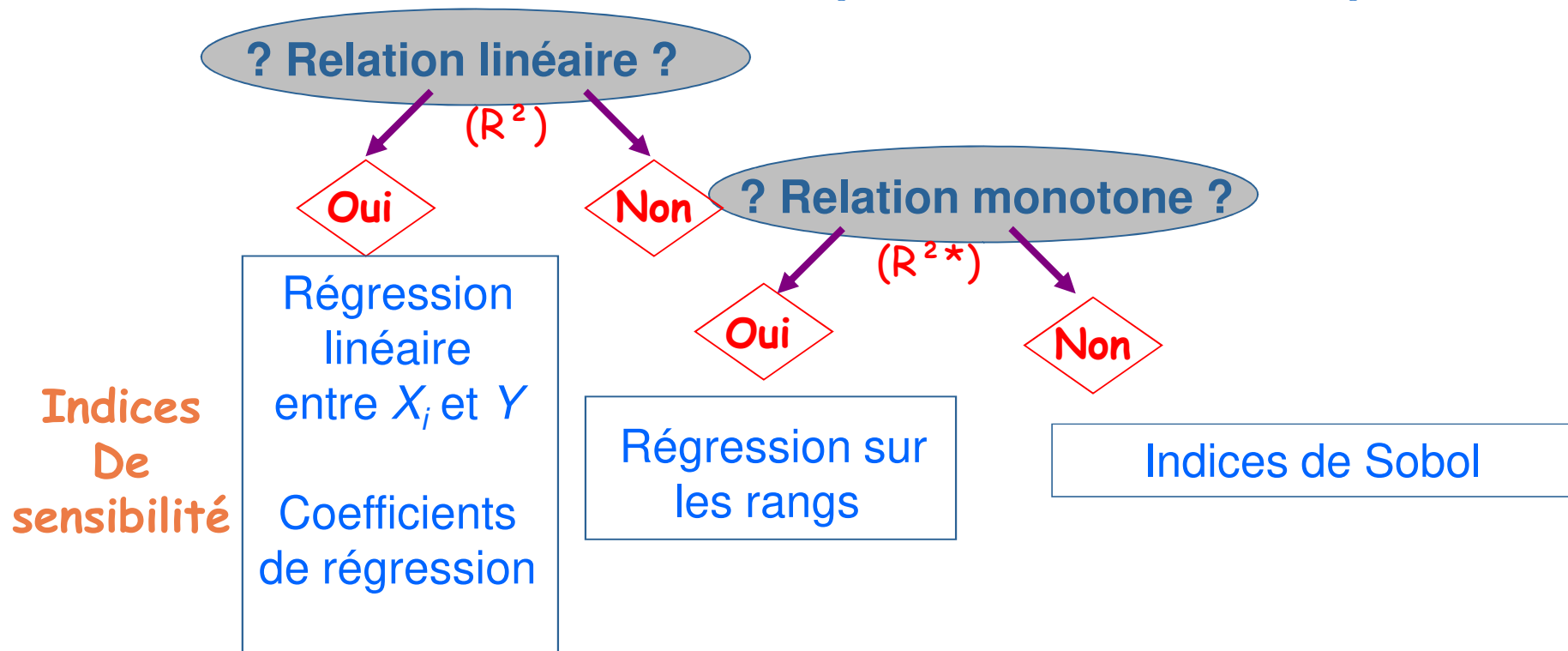
Les coefficients SRC sont suffisants

# Analyse de sensibilité pour 1 sortie scalaire

Échantillon  $(X, Y(X))$  de taille  $N > p$ , de préférence de taille  $N \gg p$   
Étape préliminaire : visualisation graphique (par ex : scatterplots)

## Méthodologie d'analyse de sensibilité quantitative

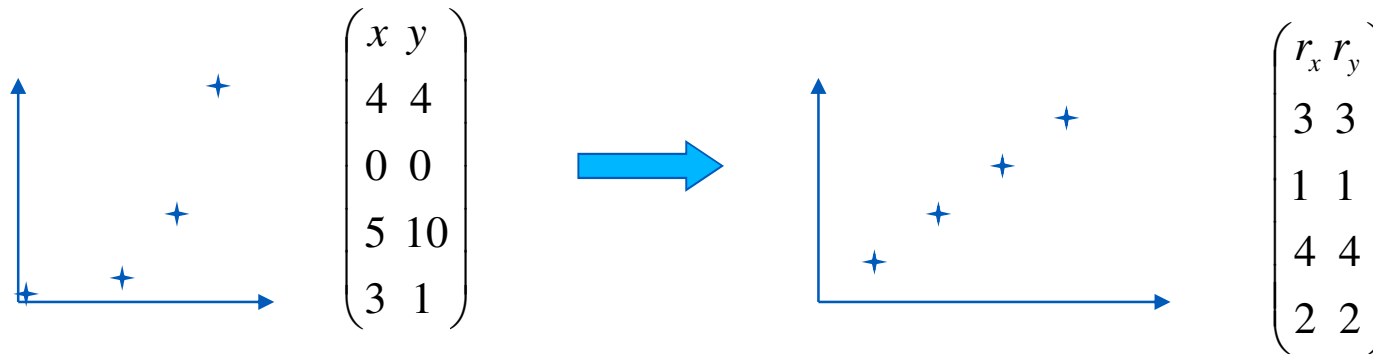
[Saltelli et al. 00, Helton et al. 06]



# Indices de sensibilité dans le cas d'une relation entrées/sortie monotone

Transformation des rangs :

à chaque individu  $X^{(j)}$  ( $j=1, \dots, N$ ), on associe son rang  $R_X^{(j)}$  (le rang varie de 1 à  $N$ )



- Le coefficient de corrélation des rangs (Spearman) = mesure le caractère monotone de la relation entre  $X_i$  et  $Y$

$$\rho_S = \frac{\text{COV}(R_X, R_Y)}{\sigma_{R_X} \sigma_{R_Y}}$$

- Relation monotone = relation linéaire sur les rangs

→ Indices de sensibilité : SRRC

- Validation : diagnostics de la régression,  $R^{2*}$

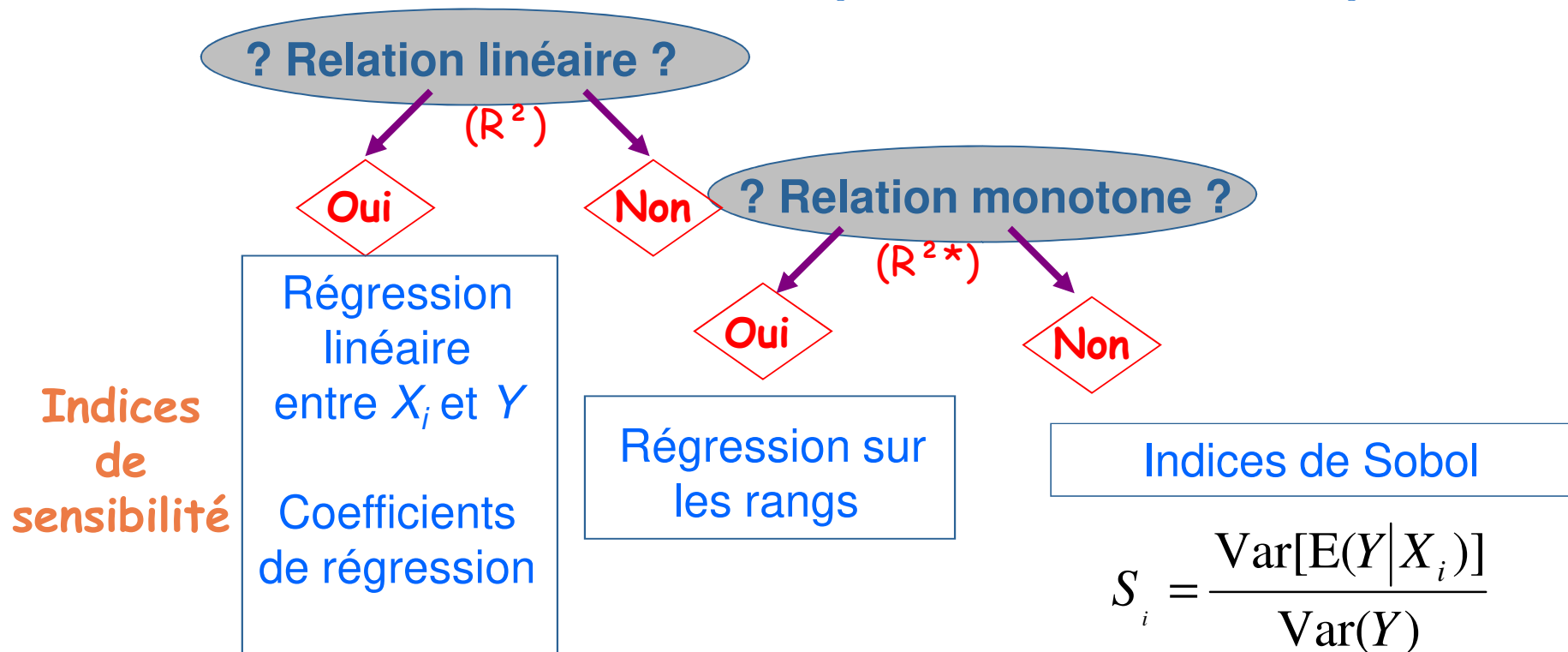


# Analyse de sensibilité pour 1 sortie scalaire

Échantillon  $(X, Y(X))$  de taille  $N > p$ , de préférence de taille  $N \gg p$   
Étape préliminaire : visualisation graphique (par ex : scatterplots)

## Méthodologie d'analyse de sensibilité quantitative

[Saltelli et al. 00, Helton et al. 06]



## Décomposition fonctionnelle

$$y = f(\mathbf{x}) = f_0 + \sum_{i=1}^p f_i(x_i) + \sum_i \sum_{j>i} f_{ij}(x_i, x_j) + \dots + f_{1,2,\dots,p}(x_1, x_2, \dots, x_p)$$

$$\text{avec } f(\mathbf{x}) \in L^2(\mathbf{x}) \quad \mathbf{x} \in [0;1]^p$$

[Hoeffding 1946]

*Il existe une infinité de décompositions possibles*

**MAIS, la décomposition est unique si :**  $\int f_{i_1 \dots i_s}(x_{i_1}, \dots, x_{i_s}) dx_j = 0 \quad \forall j = i_1, \dots, i_s$

Propriétés ( $x_i \sim U[0,1]$  pour  $i=1, \dots, p$ , les  $x_i$  sont indépendants)

$$f_0 = \int f(\mathbf{x}) d\mathbf{x} = E(y)$$

$$f_i(x_i) = \int f(\mathbf{x}) dx_{-i} - f_0 = E(y | x_i) - f_0$$

$$f_{ij}(x_i, x_j) = E(y | x_i, x_j) - E(y | x_i) - E(y | x_j) + f_0$$

Exercice :  $f(x_1, x_2) = x_1 + x_2$  ;  $x_1 \sim U[0;1]$  ;  $x_2 \sim U[0;1]$   
 $f_0 = ?$  ;  $f_1(x_1) = ?$  ;  $f_2(x_2) = ?$  ;  $f_{12}(x_1, x_2) = ?$

## Décomposition fonctionnelle

$$y = f(\mathbf{x}) = f_0 + \sum_{i=1}^p f_i(x_i) + \sum_i \sum_{j>i} f_{ij}(x_i, x_j) + \dots + f_{1,2,\dots,p}(x_1, x_2, \dots, x_p)$$

$$\text{avec } f(\mathbf{x}) \in L^2(\mathbf{x}) \quad \mathbf{x} \in [0;1]^p$$

[Hoeffding 1946]

*Il existe une infinité de décompositions possibles*

**MAIS, la décomposition est unique si :**  $\int f_{i_1 \dots i_s}(x_{i_1}, \dots, x_{i_s}) dx_j = 0 \quad \forall j = i_1, \dots, i_s$

Propriétés ( $x_i \sim U[0,1]$  pour  $i=1, \dots, p$ , les  $x_i$  sont indépendants)

$$f_0 = \int f(\mathbf{x}) d\mathbf{x} = E(y)$$

$$f_i(x_i) = \int f(\mathbf{x}) dx_{-i} - f_0 = E(y | x_i) - f_0$$

$$f_{ij}(x_i, x_j) = E(y | x_i, x_j) - E(y | x_i) - E(y | x_j) + f_0$$

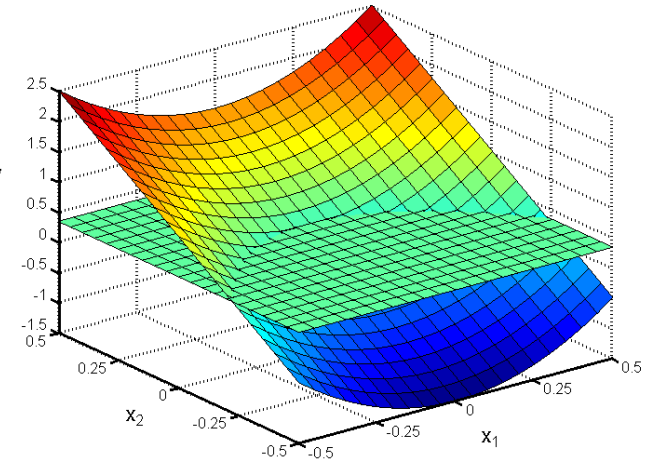
Exercice :  $f(x_1, x_2) = x_1 + x_2$  ;  $x_1 \sim U[0;1]$  ;  $x_2 \sim U[0;1]$  ;  $x_1 \perp x_2$

$$f_0 = 1 ; f_1(x_1) = x_1 - \frac{1}{2} ; f_2(x_2) = x_2 - \frac{1}{2} ; f_{12}(x_1, x_2) = 0$$

## Un autre exemple

$$f(x_1, x_2) = 4x_1^2 + 3x_2$$

$$x_1, x_2 \in U[-1/2; 1/2]$$



~~$$f_0 = 0$$~~

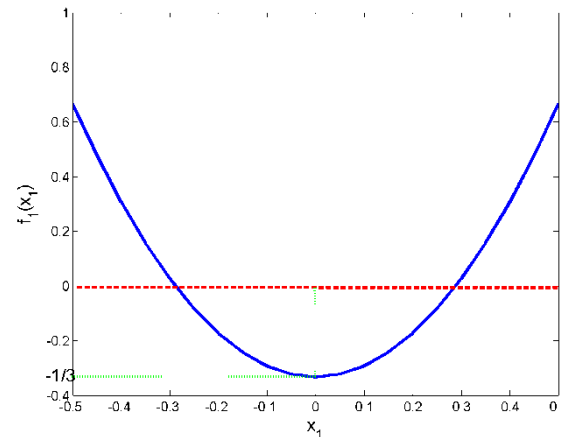
~~$$f_1(x_1) = 4x_1^2$$~~

~~$$f_2(x_2) = 3x_2$$~~

~~$$f_{12}(x_1, x_2) = 0$$~~

$$f_0 = E(y) = \int_{-1/2}^{1/2} \int_{-1/2}^{1/2} (4x_1^2 + 3x_2) dx_1 dx_2 = \frac{1}{3}$$

$$f_1(x_1) = E(y | x_1) - f_0 = \int_{-1/2}^{1/2} (4x_1^2 + 3x_2) dx_2 = 4x_1^2 - \frac{1}{3}$$



$$f_2(x_2) = E(y | x_2) - f_0 = 3x_2$$

$$f_{12}(x_1, x_2) = 0$$

# Indices de sensibilité sans hypothèse sur le modèle

**ANOVA fonctionnelle** [Efron & Stein 81] (hyp. de  $X_i$  indépendants) :

$$\text{Var}(Y) = \sum_{i=1}^p V_i(Y) + \sum_{i < j} V_{ij}(Y) + \dots + V_{12\dots p}(Y)$$

où  $V_i(Y) = \text{Var}[E(Y|X_i)]$  ;  $V_{ij} = \text{Var}[E(Y|X_i X_j)] - V_i - V_j, \dots$

**Définition des indices de Sobol :**

▶ Indice de sensibilité d'ordre 1 :  $S_i = \frac{V_i}{\text{Var}(Y)}$

▶ Indice de sensibilité d'ordre 2 :  $S_{ij} = \frac{V_{ij}}{\text{Var}(Y)}$

▶ ...

Exercice :

$$f(X_1, X_2) = X_1^2 + X_2 ; X_1 \sim U[0,1] ; X_2 \sim U[0,1] ; X_1 \perp X_2$$

$$S_1 = ? \quad ; \quad S_2 = ?$$

# Indices de sensibilité sans hypothèse sur le modèle

**ANOVA fonctionnelle** [Efron & Stein 81] (hyp. de  $X_i$  indépendants) :

$$\text{Var}(Y) = \sum_{i=1}^p V_i(Y) + \sum_{i<j}^p V_{ij}(Y) + \dots + V_{12\dots p}(Y)$$

$$\text{où } V_i(Y) = \text{Var}[E(Y|X_i)] ; V_{ij} = \text{Var}[E(Y|X_i X_j)] - V_i - V_j, \dots$$

**Définition des indices de Sobol :**

► Indice de sensibilité d'ordre 1 :  $S_i = \frac{V_i}{\text{Var}(Y)}$

► Indice de sensibilité d'ordre 2 :  $S_{ij} = \frac{V_{ij}}{\text{Var}(Y)}$

► ...

Exercice :  $f(X_1, X_2) = X_1^2 + X_2$  ;  $X_1 \sim U[0,1]$  ;  $X_2 \sim U[0,1]$  ;  $X_1 \perp X_2$

$$S_1 = \frac{16}{31} \quad ; \quad S_2 = \frac{15}{31}$$

## L'autre exemple

$$y = f(x_1, x_2) = 4x_1^2 + 3x_2 \quad x_1, x_2 \in U[-1/2, 1/2]$$

On a vu :

$$f_0 = E(y) = \frac{1}{3}$$

$$f_1(x_1) = E(y | x_1) - f_0 = 4x_1^2 - \frac{1}{3}$$

$$f_2(x_2) = E(y | x_2) - f_0 = 3x_2$$

$$f_{12}(x_1, x_2) = 0$$

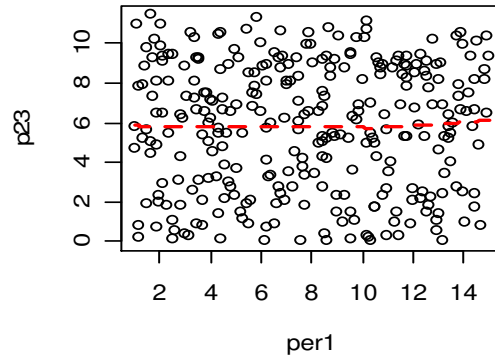
$$S_1 = \frac{\text{Var}[f_1(x_1)]}{V} = \frac{0.08}{0.838} = 0.106$$

$$S_2 = \frac{\text{Var}[f_2(x_2)]}{V} = \frac{0.75}{0.838} = 0.894$$

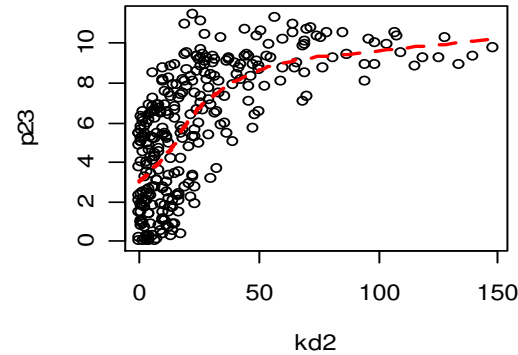
# Interprétation graphique

Les indices de Sobol du premier ordre mesurent la variabilité des espérances conditionnelles (courbes de tendance dans les scatterplots)

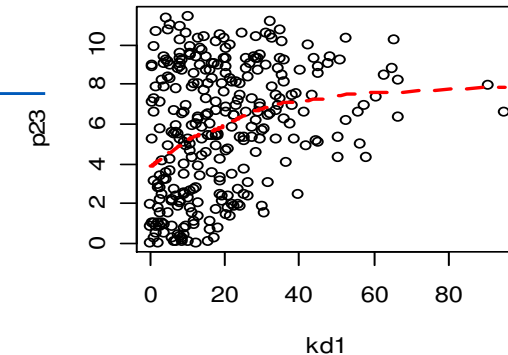
Indice nul



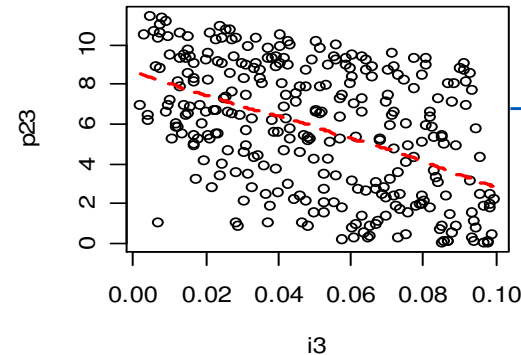
Indice élevé



Indice faible



Indice moyen



La courbe rouge correspond à un lissage, c'est-à-dire à  $E(Y | X_i)$

Le lissage est réalisé ici par polynômes locaux :  $\hat{f}(\mathbf{x}) = \hat{\alpha}(\mathbf{x}) + \mathbf{x} \hat{\beta}(\mathbf{x})$

Autres méthodes envisageables : moyenne mobile, polynômes, splines, modèles additifs, ...



# Une autre interprétation

▶ Question centrale de l'analyse de sensibilité : la sortie  $Y$  est-elle plus ou moins variable lorsque l'on fixe une des entrées  $X_i$  ?

▶ On étudie  $\text{Var}(Y|X_i = x_i)$  et on prend sa moyenne sur les valeurs de  $X_i$  :  $E[\text{Var}(Y|X_i = x_i)]$

▶ Plus  $E[\text{Var}(Y|X_i = x_i)]$  est petit, plus  $X_i$  est influente

▶ **Théorème de la variance totale :**

$$\text{Var}(Y) = E[\text{Var}(Y|X_i)] + \text{Var}[E(Y|X_i)]$$

▶ On en déduit les indices de Sobol :  $S_i = \frac{\text{Var}[E(Y|X_i)]}{\text{Var}(Y)}$

▶ Au passage, si le modèle est linéaire :  $S_i = \text{SRC}^2(X_i)$

## Propriétés des indices de Sobol

$$1 = \sum_{i=1}^p S_i + \sum_i \sum_j S_{ij} + \sum_i \sum_j \sum_k S_{ijk} \dots + S_{1,2,\dots,k}$$

$$\sum_i S_i \leq 1 \quad \text{Toujours}$$

$$\sum_i S_i = 1 \quad \text{Le modèle est additif}$$

$$1 - \sum_i S_i \quad \text{Mesure le degré d'interactions entre variables}$$

Exemples :  $p=4$  donne 4 indices  $S_i$ , 6 indices  $S_{ij}$ , 4 indices  $S_{ijk}$ , 1 indice  $S_{ijkl}$

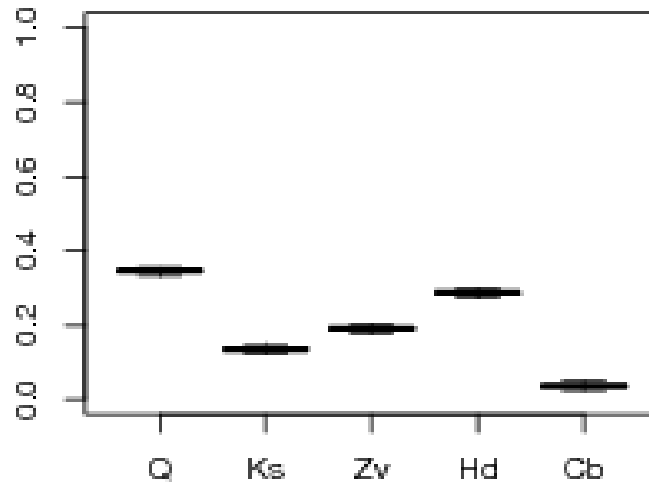
Cas général :  $2^p - 1$  indices à estimer

$$\text{Indice de sensibilité total : } S_{Ti} = S_i + \sum_j S_{ij} + \sum_{j,k} S_{ijk} + \dots = 1 - S_{\sim i}$$

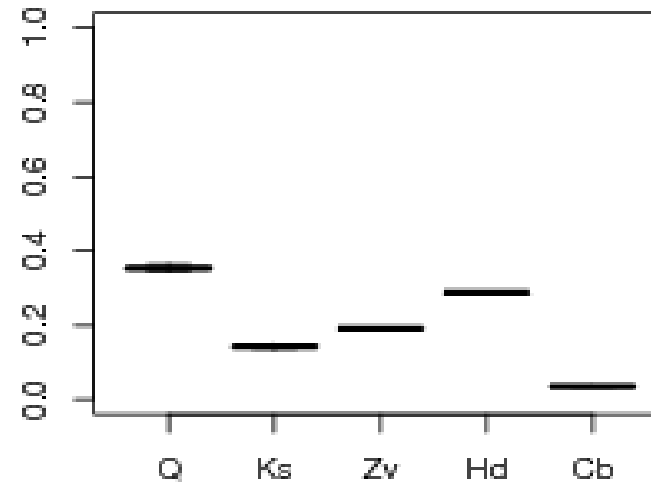
[ Homma & Saltelli 1996 ]

# Modèle de crues

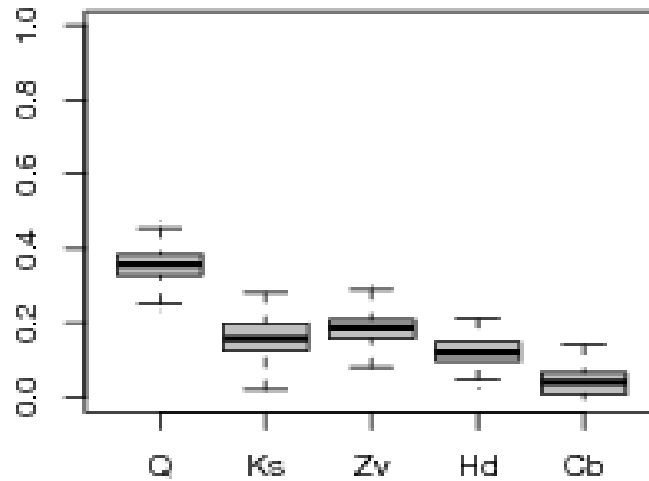
## Sortie S - Indices 1er ordre



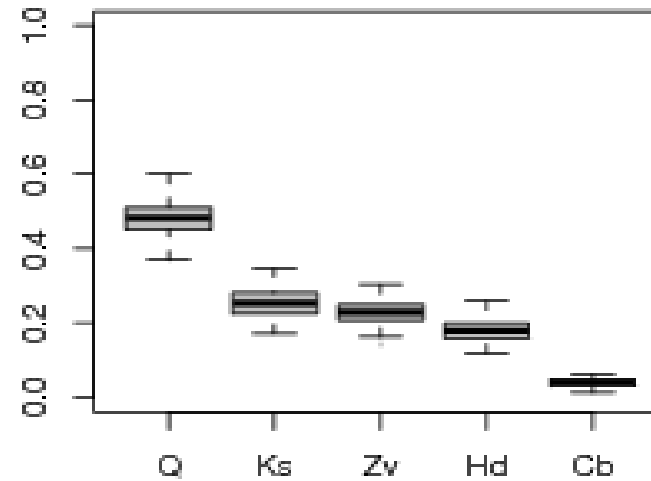
## Sortie S - Indices totaux



## Sortie Cp - Indices 1er ordre



## Sortie Cp - Indices totaux



# Calcul des indices de Sobol

- ◆ Formulations des indices pour  $X_i$  (1<sup>er</sup> ordre et total) :

$$S_i = \frac{V_i}{\text{Var}(Y)} \text{ et } S_{T_i} = 1 - \frac{V_{\sim i}}{\text{Var}(Y)}$$

- ◆ Formulations des variances conditionnelles :

Soient  $\mathbf{X} = (X_i, X_{\sim i})$  et  $\mathbf{X}'$  une copie indépendante de  $\mathbf{X}$

$$V_i(Y) = \text{Var}[E(Y|X_i)] = \int E^2(Y|X_i) dX_i - \left( \int E(Y|X_i) dX_i \right)^2 = \text{Cov}[f(X_i, X_{\sim i}), f(X_i, X'_{\sim i})]$$

$$V_{\sim i}(Y) = \text{Var}[E(Y|X_{\sim i})] = \text{Cov}[f(X_i, X_{\sim i}), f(X'_i, X_{\sim i})]$$

# Estimation des indices de Sobol par Monte Carlo

Soient 2 échantillons i.i.d :  $(X_i^{(j)})_{i=1,\dots,p;j=1,\dots,n}$  et  $(X_i'^{(j)})_{i=1,\dots,p;j=1,\dots,n}$

► Variance (estimateur classique) :  $\hat{V}(Y) = \frac{1}{n} \sum_{k=1}^n f(\mathbf{X}^{(k)})^2 - \hat{f}_0^2$  avec  $\hat{f}_0 = \frac{1}{n} \sum_{k=1}^n f(\mathbf{X}^{(k)})$

► Estimations des variances conditionnelles :

$$\hat{V}_i(Y) = \frac{1}{n} \sum_{k=1}^n f(X_1^{(k)}, \dots, X_{i-1}^{(k)}, X_i^{(k)}, X_{i+1}^{(k)}, \dots, X_p^{(k)}) f(X_1'^{(k)}, \dots, X_{i-1}'^{(k)}, X_i^{(k)}, X_{i+1}'^{(k)}, \dots, X_p'^{(k)}) - f_0^2$$

Indices 1<sup>er</sup> ordre : coût =  $n(p+1)$

$$\hat{V}_{\sim i}(Y) = \frac{1}{n} \sum_{k=1}^n f(X_1^{(k)}, \dots, X_{i-1}^{(k)}, X_i^{(k)}, X_{i+1}^{(k)}, \dots, X_p^{(k)}) f(X_1^{(k)}, \dots, X_{i-1}^{(k)}, X_i'^{(k)}, X_{i+1}^{(k)}, \dots, X_p^{(k)}) - f_0^2$$

Indices 1<sup>er</sup> ordre + indices totaux : coût =  $n(p+2)$

Astuce : on intervertit  $(X_i^{(j)})_{i=1,\dots,p;j=1,\dots,n}$  et  $(X_i'^{(j)})_{i=1,\dots,p;j=1,\dots,n}$  dans  $\hat{V}_{\sim i}(Y)$

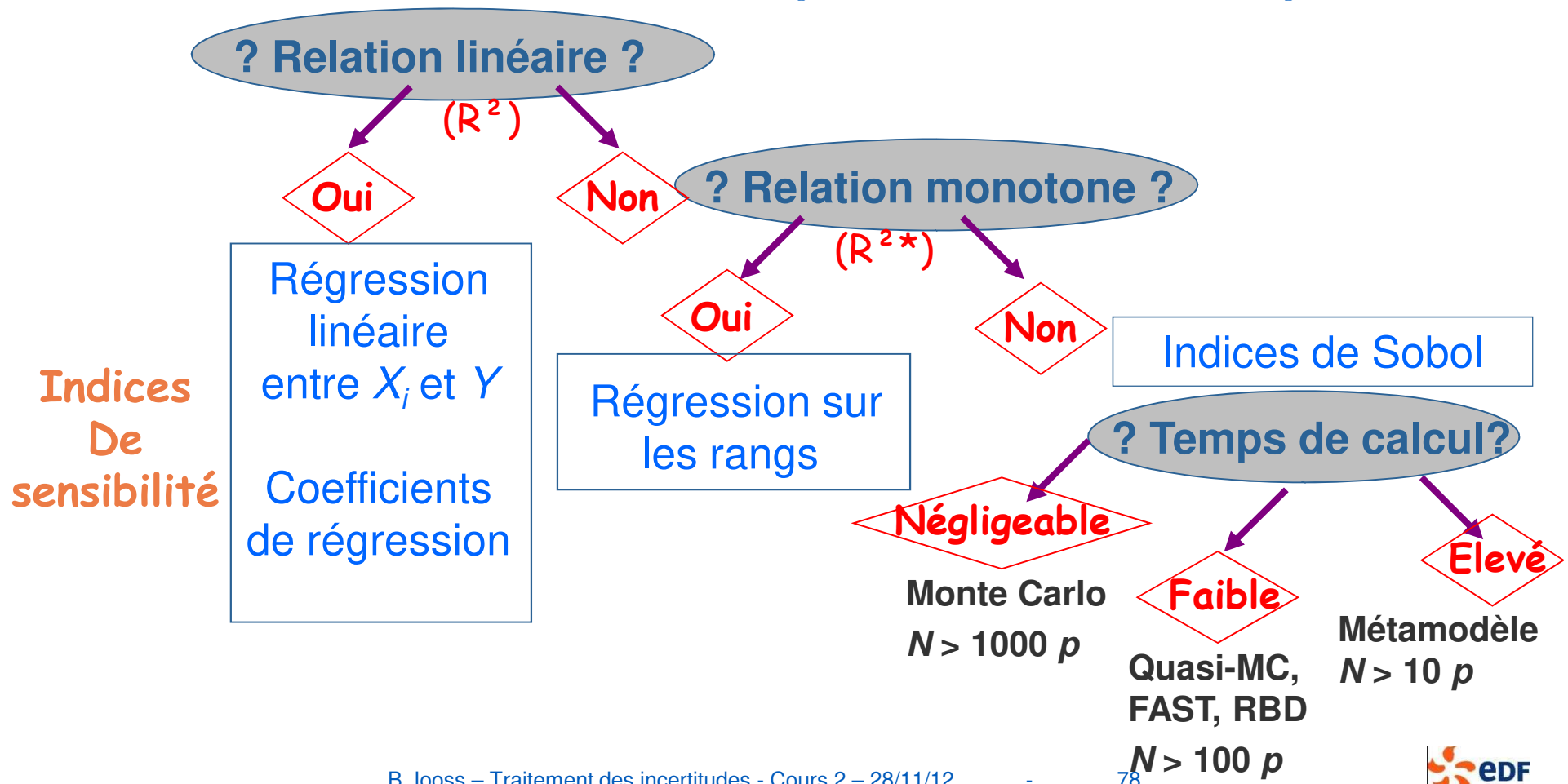
En pratique,  $n \sim 1e4 \Rightarrow$  problème du coût en nombre d'évaluations nécessaires

# Analyse de sensibilité pour 1 sortie scalaire

Échantillon  $(X, Y(X))$  de taille  $N > p$ , de préférence de taille  $N \gg p$   
Étape préliminaire : visualisation graphique (par ex : scatterplots)

## Méthodologie d'analyse de sensibilité quantitative

[Saltelli et al. 00, Helton et al. 06]



# Rappels sur l'analyse de sensibilité

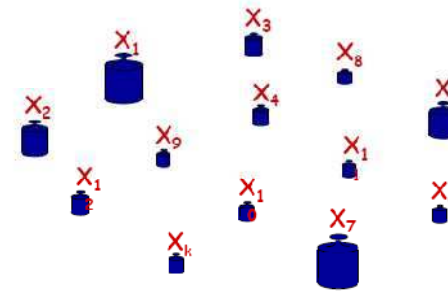
**Enjeu** : décomposer la variabilité globale de la sortie  $Z = G(\mathbf{X})$  (due aux incertitudes sur les entrées  $\mathbf{X}$ ) en part de variabilité due à chaque entrée  $X_i, i=1, \dots, p$

Problème : comme pour la planification, le coût en nombre  $N$  d'évaluations de  $G(\cdot)$  dépend de  $p$

## 1. Le criblage (screening) :

- plans d'expériences classiques,
- plans d'expériences numériques

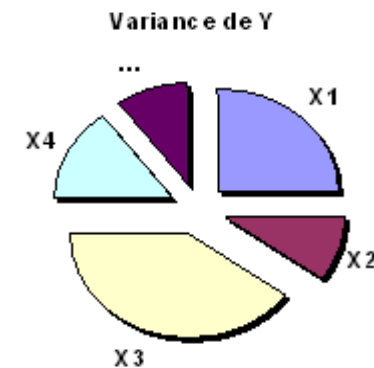
$$N \sim p/2 \text{ à } 10 p$$



## 2. Les mesures d'influence globale :

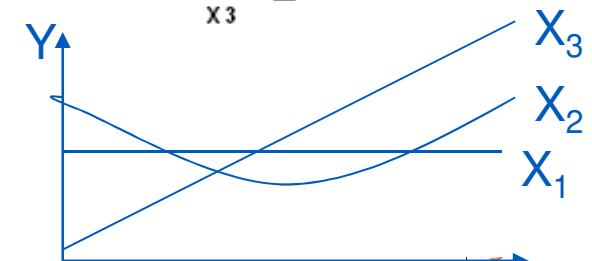
- corrélation/régression sur les valeurs/rangs,
- décomposition de la variance fonctionnelle (Sobol),

$$N \sim 2p \text{ à } 1e4 p$$



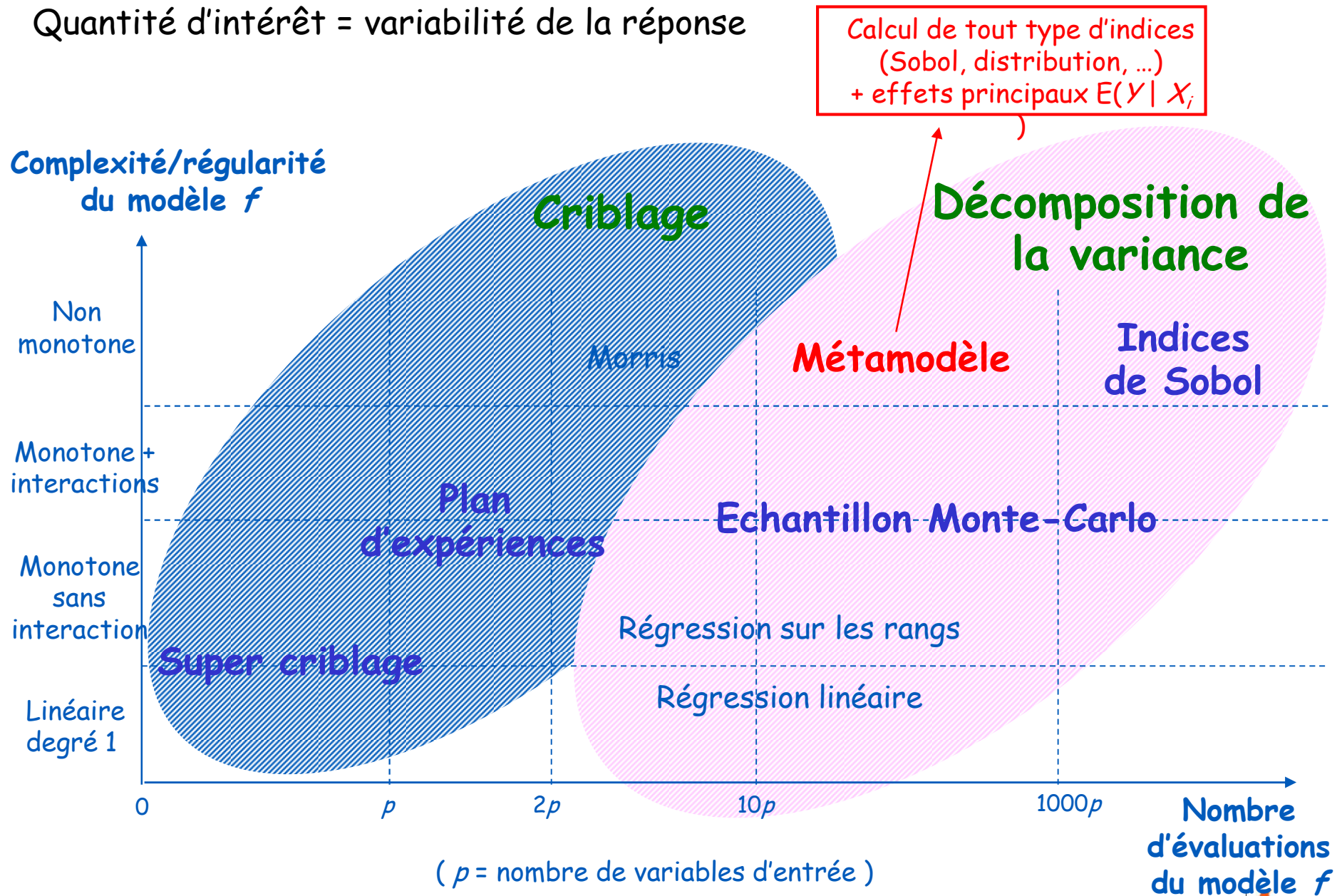
## 3. Exploration fine des sensibilités - $N \sim 10p$ à $100 p$

- Méthodes de lissage (param./non param.)
- **Métamodèles => cours 3**



# Choisir sa méthode d'analyses de sensibilité

Quantité d'intérêt = variabilité de la réponse





# Crédits & Bibliographie

- Formation « Démarche Incertitudes », IMdR-LNE
- Summer School SAMO, Fiesole, Italy, 2010
- Fang et al., *Design and modeling for computer experiments*, Chapman & Hall, 2006
- Kleijnen, *The design and analysis of simulation experiments*, Springer, 2008
- Koehler & Owen, *Computer experiments*, 1996
- Faivre et al., *Analyse de sensibilité et exploration de modèles – Applications aux sciences de la nature et de l'environnement*, Editions Quaé, à paraître
- Saltelli et al., *Sensitivity analysis*, Wiley, 2000

Ce cours est disponible sur : <http://www.gdr-mascotnum.fr/doku.php?id=iooss1#academic>