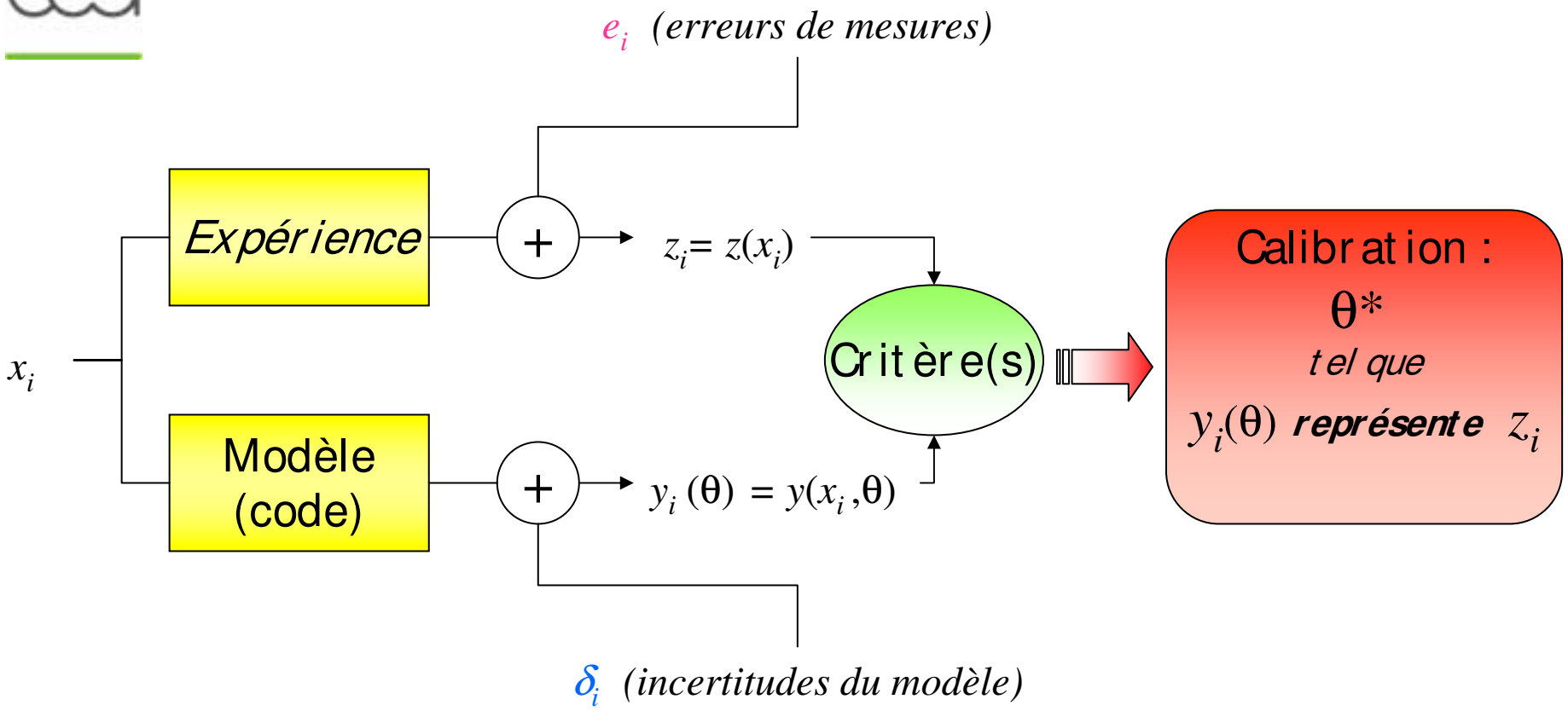
Méthodologie de calibration d'un code de calcul en contexte d'incertitude

V. FEUILLARD, N. DEVICTOR, R. PHAN-TAN-LUU
P. DEHEUVELS

CADARACHE

INTRODUCTION : LA CALIBRATION

Calibration d'un code de calcul



INTRODUCTION

- Objectif :

> *« Trouver le ou les meilleurs jeux de paramètres θ de façon à ce que la réponse du modèle (du code) représente la réalité (réponse expérimentale) »*

Terme *« meilleur »* : défini à partir de critères du vecteur des paramètres θ.

- Difficultés rencontrées :
  - Base de données expérimentale qui n'a pas été construite dans l'objectif de calibrer le code de calcul ou le modèle
  - Situations mal posées → nombre de coefficients à calibrer > nombre d'expériences disponibles

⇒ *PROPOSER UNE METHODOLOGIE*

INTRODUCTION : Méthodologie

4 Étapes

- Étape 1 Collecte d'informations, bien définir le problème
- Étape 2 Analyser la base de données initiale des $\{x_i\}_{i=1..N}$, et des $\{\theta_i\}_{i=1..K}$
- Étape 3 Appliquer la démarche de calibration
- Étape 4 Valider les jeux de paramètres θ

ETAPE 1 Collecte d'information

EN COOPERATION AVEC LES SPECIALISTES DU CODE

- **Bases de données initiale**
  - Entrées x_i ; Réponses expérimentales z_i ; réponses du modèle y_i
  - Domaines de variation et/ou domaines d'intérêt
  - Incertitudes

- **Le vecteur de paramètres θ (avis d'expert)**
  - $\theta \in \Theta$
  - Information a priori
  - Contraintes

- **Fonction de code**
  - Connaissance
  - Linéaire ou non (par rapport à θ , x_i)
  - Degré de continuité, dérivées disponibles

ETAPE 1 Différents contextes

Différentes données possibles :

D1 $= \{ X = (x_1, \ldots, x_n); Y_{obs} = (y_{obs}(x_1, \theta^*), \ldots, y_{obs}(x_n, \theta^*)) \}$

D = D1 ∪ D2

D2 $= \{ \Theta = (\theta_1, \ldots, \ldots \theta_N)$;

$Y(\theta_1) = (f(x_1, \theta_1), \ldots, f(x_n, \theta_1))$;

$Y(\theta_N) = (f(x_1, \theta_N), \ldots, f(x_n, \theta_N)) \}$

Techniques de régression multiple, approche bayésienne, …

Cadre étudié

D = D1 ∪ D2

D2 $= \{ Y(\theta^*) = (f(x_1, \theta^*), \ldots, f(x_n, \theta^*))$;

f fonction de code « connue » }

on connaît f analytiquement, ou ses dérivées

Techniques de régression linéaire ou non linéaire, …

ETAPE 2 Etude de la base de données

- Objectif :

Etude de la qualité de l'information disponible

Cadre étudié : base de données préexistante, absence d'information sur la fonction de code

⇒ Evaluer la qualité de répartition uniforme des données X et Θ dans l'espace

- 2 Approches :
  - Approche « *déterministe* »
  - *Approche « probabiliste » → non décrite aujourd'hui*

ETAPE 2 Approche « déterministe »

Différents types de critères

- Critères d'espacements des points de la BDD

Objectif : *vérifier la régularité des espacement des points*

(Considération de distances entre les points de la BDDE)

- Critères de recouvrement de l'espace

Objectif : *vérifier que les points recouvrent tout l'espace, pas de sous-espace « vide »*

(Considération de distances entre les points de la BDDE et points de l'espace)

- Critères de « bonne » répartition uniforme

Objectif : *vérifier que les points recouvrent « bien » tout l'espace*

(Considération de volumes, comparaison entre nombre de points dans un intervalle et volume de cet intervalle)

ETAPE 2 Critères d'espacements

- Considération de distances (critères *dmin* , *m*)

dmin : *distance minimale entre deux points* $dmin_p = \min_{x_i \neq x_{i'}} \left\| x_i - x_{i'} \right\|_p$

$\boldsymbol{m_{p,\alpha}}$: *« moyenne » des distances entre points* $m_{p,\alpha} = \left[\frac{2}{n(n-1)} \sum_{x_i \neq x_{i'}} \left(\frac{d^{1/p}}{\left\| x_i - x_{i'} \right\|_p} \right)^{\alpha} \right]^{1/\alpha}$

- Considération des espacements (critères γ, λ)

$\boldsymbol{\gamma}$: *rapport du maximum et minimum des espacements entre les points* $\gamma = \max_{i=1..n}\{\gamma_i\} / \min_{i=1..n}\{\gamma_i\}$ $\gamma_i = \min_{x_{i'} : x_{i'} \neq x_i} \left\{ \left\| x_i - x_{i'} \right\| \right\}$

$\boldsymbol{\lambda}$: *« variabilité » des espacements* $\lambda = \frac{1}{\overline{\gamma}} \left(\frac{1}{n} \sum_{i=1}^{n} (\gamma_i - \overline{\gamma})^2 \right)^{1/2}$

Valeurs de référence : $\gamma = 1$, $\lambda = 0$

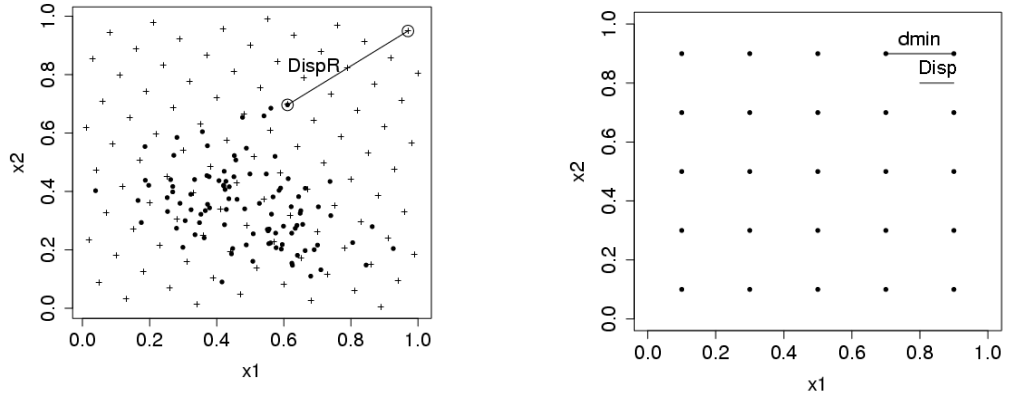
ETAPE 2 Critères de recouvrement de l'espace

- Considération de distances entre points de la BDD et points de l'espace (critères *Dispersion*, h , μ , χ) :

Dispersion : *rayon de la plus grande boule « vide »* $Disp_p(BDDE) = \sup_{x \in \chi} \min_{x_i \in BDDE} \|x - x_i\|_p$

approximation à l'aide de suites dont le recouvrement de l'espace est acceptable,

suites à discrépance faible, $DispR_p(BDDE, SDF) = \max_{x \in SDF} \min_{x_i \in BDDE} \|x_i - x\|_p$



Valeur de référence : $Disp_\infty = 2.dmin_\infty = 1/(2.n^{1/d})$

ETAPE 2 Critères de recouvrement de l'espace

Utilisation des cellules de Voronoi

La cellule de Voronoi associée au point $x \in BDDE$ est l'ensemble des points de l'espace tels que leur distance est inférieure à n'importe quel autre point de la BDDE,

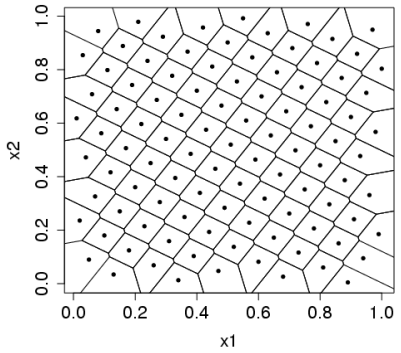
$$V(x) = \left\{ z \in [0,1]^d : \|z - x\| \leq \|z - x'\| , \forall x' \in BDDE \right\}$$

h *maximum des rayons des cellules,* $h = \max_{i=1..n} \{h_i\}$ *où* $h_i = \max_{z \in V(x_i)} \|x_i - z\|$

μ *rapport du maximum et minimum des rayons des cellules,* $\mu = \max_{i=1..n} \{h_i\} / \min_{i=1..n} \{h_i\}$

χ *comparaison entre rayon des cellules et espacement,* $\chi = \max_{i=1..n} \{\chi_i\}$ *où* $\chi_i = 2h_i / \gamma_i$

Valeurs de référence $\mu = 1$, $\chi = 1$



ETAPE 2 Critères de « bonne » répartition uniforme

• Considération de volumes (critères ν, τ, D, *Discrépances*)

Utilisation des cellules de Voronoi

ν *rapport des volumes des cellules,* $\nu = \max_{i=1..n} |V_i| / \min_{i=1..n} |V_i|$ *où* $|V_i|$ *est le volume de* V_i

On pose : $M_i = \frac{1}{V_i} \int_{V_i} (x - \bar{x}_i)(x - \bar{x}_i)^t \, dx$; $T_i = trace(M_i)$; $D_i = \det(M_i - \bar{M}_i)$

τ $\tau = \max_{i=1..n} |T_i - \bar{T}|$

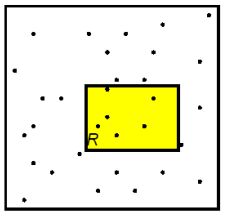
D $D = \max_{i=1..n} \{D_i\}$

Valeurs de référence $\upsilon = 1$, $\tau = 0$ $D = 0$

ETAPE 2 Critères de « bonne » répartition uniforme

Discrépance

« Différence (norme) entre nombre de points compris dans un intervalle et le volume de cet intervalle. »

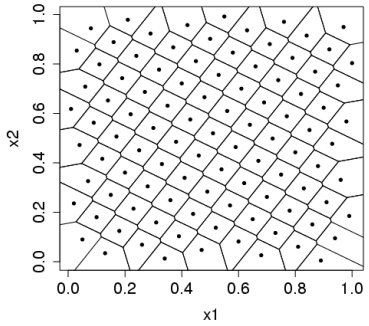


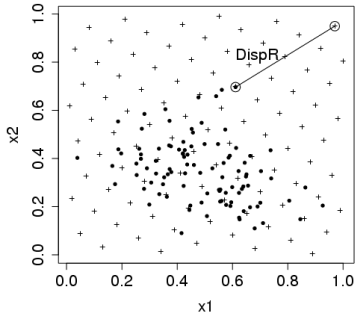
Différentes définitions, selon la norme choisie, selon les intervalles considérés.

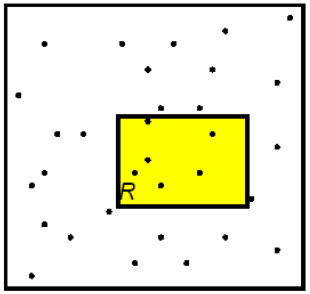
Choix de la discrépance L^2 et de ses variantes (*Hickernell, 1998*)

- ***discrépance L^2*** (*intervalles ancrés à l'origine*)
- ***discrépance L^2 modifiée*** (*intervalles ancrés à l'origine, autre norme*)
- ***discrépance L^2 centrée*** (*intervalles ayant pour borne l'un des sommets du cube unité*)
- ***discrépance L^2 symétrique*** (*intervalles ayant pour borne l'un des sommets « pair» du cube unité)*

ETAPE 2 Critères déterministes







- Critères de régularité des espacements
- Critères de recouvrement de l'espace
- Critères de « bonne » répartition uniforme

CRITERE	VALEUR DE REFERENCE
$dmin_\infty$	$Disp_\infty/2$
γ	1
$m_{2,1}$	$\downarrow$
λ	0
$DispHa_\infty$	$2\,dmin_\infty$
$DispHam_\infty$	$2\,dmin_\infty$
$DispFa_\infty$	$2\,dmin_\infty$
$DispRes_\infty$	$2\,dmin_\infty$
h	$\downarrow$
μ	1
$\mathcal{X}$	$\downarrow$
ν	1
τ	0
D	0
$DiscL2$	$\downarrow$
$RDiscL2$	1
$DiscL2M$	$\downarrow$
$RDiscL2M$	1
$DiscL2C$	$\downarrow$
$RDiscL2S$	1
$DiscL2S$	$\downarrow$
$RDiscL2S$	1

ETAPE 2 Utilisation méthodologie

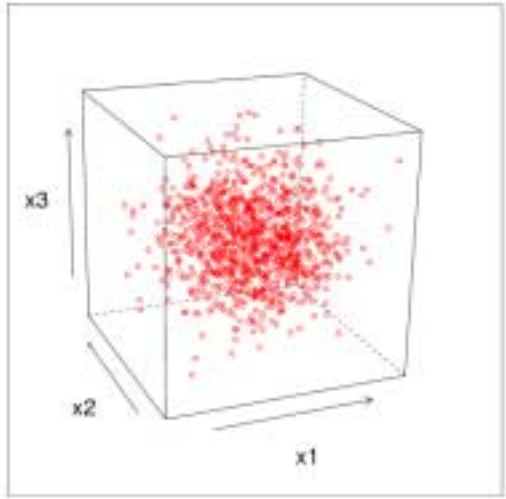
Méthodologie :

3 étapes :

- **Analyse de la base de données initiale**
- **Sélection de points de la base de données**
- **Spécification de points supplémentaires** (si possible et si nécessaire)

Application à un exemple à 3 dimensions

(400 points)



ETAPE 2 Sélection de points

Sélection de points à partir de la base de données initiale

Algorithme (A1)

i. Une distance $dmin$ est fixée. Soit BDD_0 la base de données initiale. Le point x^* le plus proche (au sens de la norme euclidienne) du « centre » du domaine est sélectionné.

   On pose : $BDD_1=\{x^*\}$ et $BDD_2=BDD_0 \setminus \{x^*\}$.

ii. Soient $\{x_1,...x_k\}$ les points inclus dans la boule de centre x^* et de rayon $dmin$.

   On pose : $BDD_2=BDD_2\setminus\{x_1,...x_k\}$.

iii. Soit $x' \in BDD_2$ le point le plus proche de BDD_1 (réalisant le minimum des distances euclidiennes entre les points de BDD_2 et BDD_1).

   On pose : $BDD_1=BDD_1 \cup\{x'\}$ *et* $BDD_2 = BDD_2\setminus\{x'\}$.

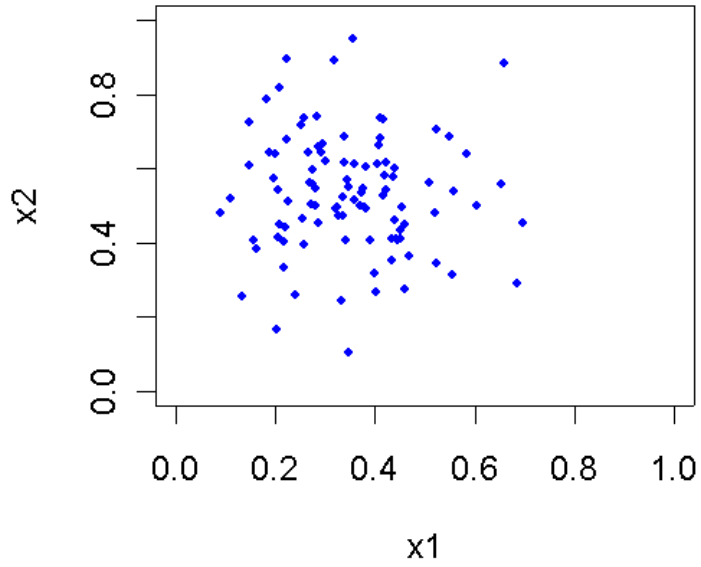
iv. Soient $\{x_{j1},...,x_{jl}\}$ les points inclus dans la boule de centre *x'* et de rayon $dmin$.

   On pose : $BDD_2 = BDD_2\setminus\{ x_{j1},...,x_{jl} \}$

v. Itération des étapes 3 et 4 tant que $BDD_2 \neq \varnothing$.

ETAPE 2 Sélection de points

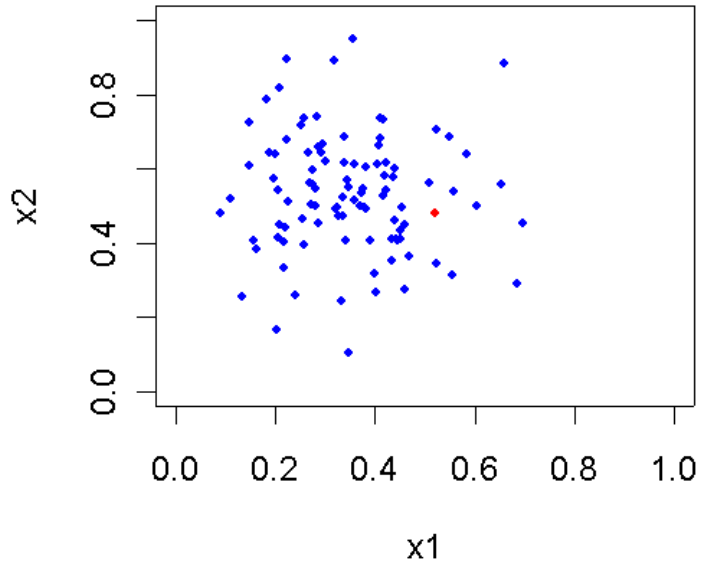
BDD_0



ETAPE 2 Sélection de points

étape *i*

Une distance $dmin$ est fixée. Soit BDD_0 la base de données initiale. Le point x^* le plus proche (au sens de la norme euclidienne) du « centre » du domaine est sélectionné. On pose : $BDD_1=\{x^*\}$ et $BDD_2=BDD_0 \setminus \{x^*\}$.

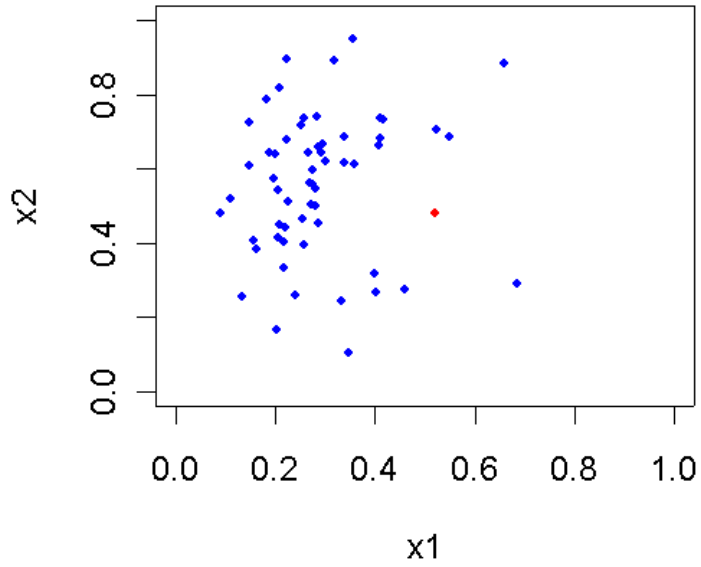


ETAPE 2 Sélection de points

étape *ii*

Soient $\{x_1,...x_k\}$ les points inclus dans la boule de centre x^* et de rayon $dmin$.

On pose : $BDD_2=BDD_2\backslash\{x_1,...x_k\}$.

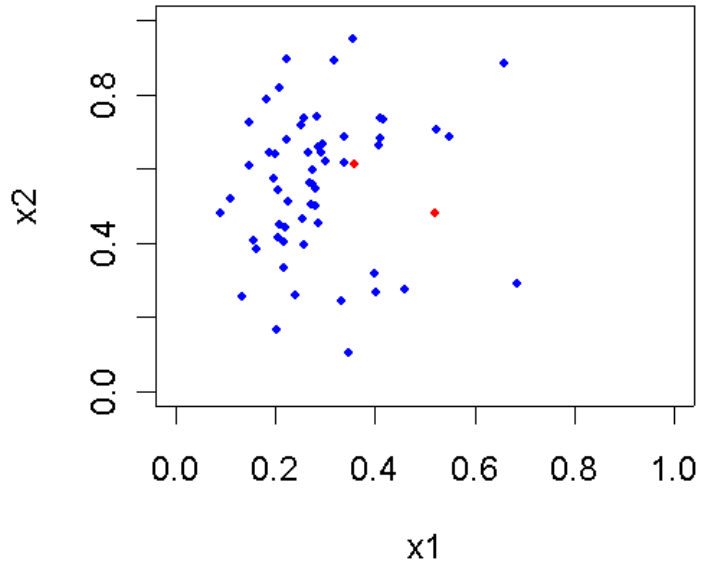


ETAPE 2 Sélection de points

étape *iii*

Soit $x' \in BDD_2$ le point le plus proche de BDD_1 (réalisant le minimum des distances euclidiennes entre les points de BDD_2 et BDD_1).

On pose : $BDD_1 = BDD_1 \cup \{x'\}$ *et* $BDD_2 = BDD_2 \setminus \{x'\}$.

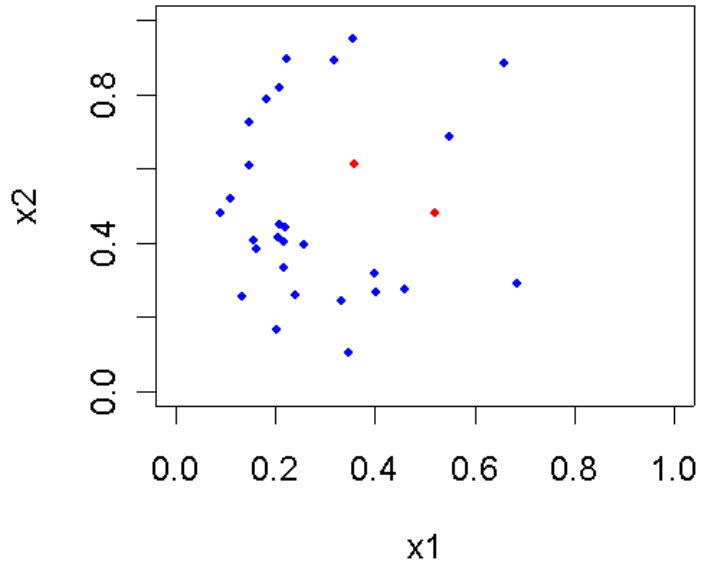


ETAPE 2 Sélection de points

étape *iv*

Soient $\{x_{j1},...,x_{jl}\}$ les points inclus dans la boule de centre *x'* et de rayon $dmin$.

On pose : $BDD_2 = BDD_2 \setminus \{ x_{j1},...,x_{jl} \}$

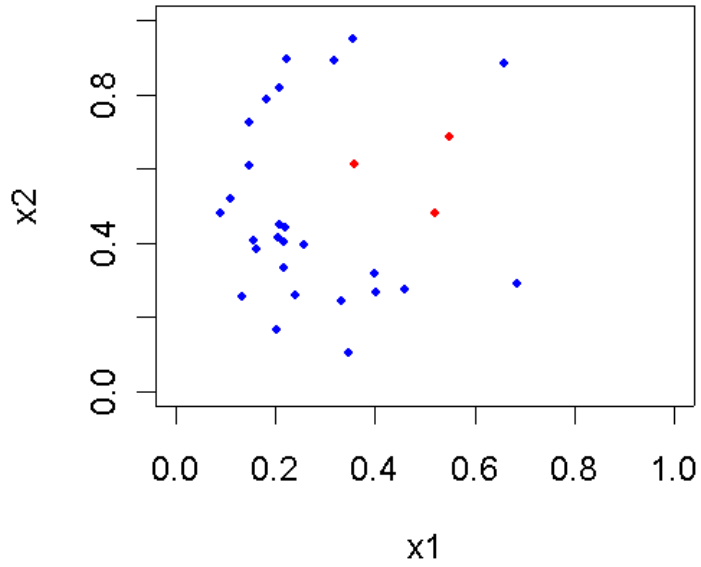


ETAPE 2 Sélection de points

Itération étape *iii*

Soit $x' \in BDD_2$ le point le plus proche de BDD_1 (réalisant le minimum des distances euclidiennes entre les points de BDD_2 et BDD_1).

On pose : $BDD_1 = BDD_1 \cup \{x'\}$ *et* $BDD_2 = BDD_2 \setminus \{x'\}$.

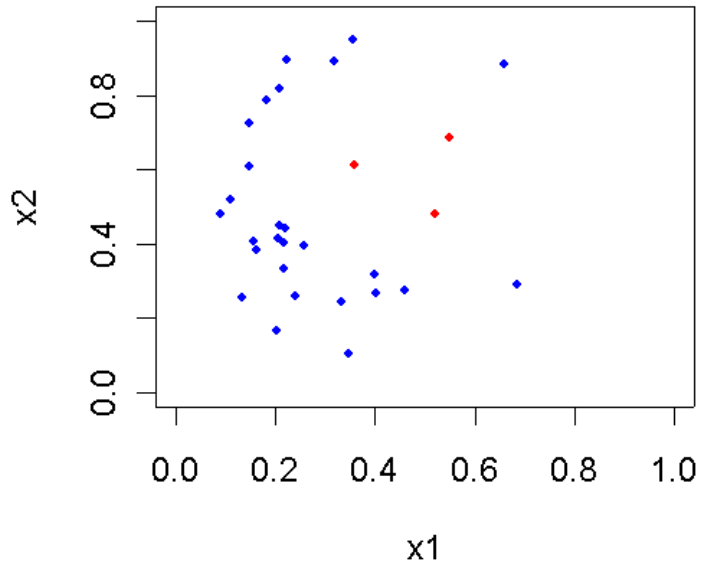


ETAPE 2 Sélection de points

Itération étape *iv*

Soient $\{x_{j1},...,x_{jl}\}$ les points inclus dans la boule de centre *x'* et de rayon $dmin$.

On pose : $BDD_2 = BDD_2 \backslash \{ x_{j1},...,x_{jl} \}$

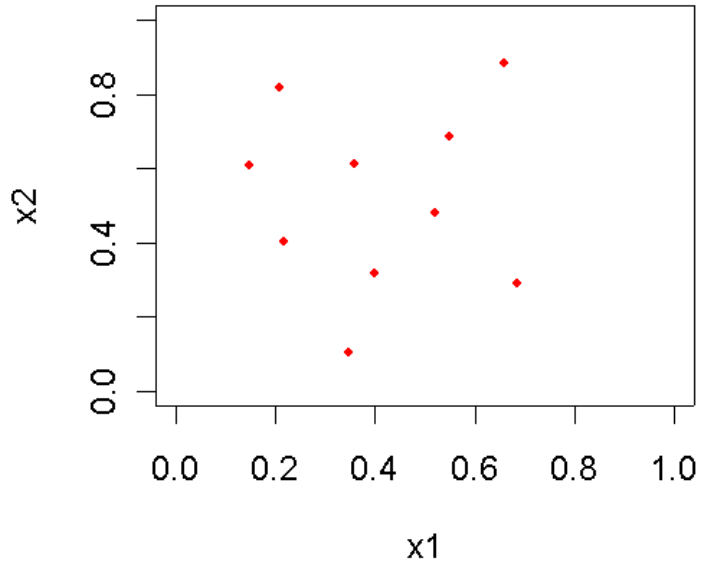


ETAPE 2 Sélection de points

Itération étape *iii*

Soit $x' \in BDD_2$ le point le plus proche de BDD_1 (réalisant le minimum des distances euclidiennes entre les points de BDD_2 et BDD_1).

On pose : $BDD_1 = BDD_1 \cup \{x'\}$ *et* $BDD_2 = BDD_2 \setminus \{x'\}$.



$BDD_2 \neq \varnothing$. $\Rightarrow$ fin de l'itération

ETAPE 2 Sélection de points

Sélection de points à partir de la base de données initiale

Algorithme (A1)

i. Une distance $dmin$ est fixée. Soit BDD_0 la base de données initiale. Le point x^* le plus proche (au sens de la norme euclidienne) du « centre » du domaine est sélectionné.

   On pose : $BDD_1=\{x^*\}$ et $BDD_2=BDD_0 \setminus \{x^*\}$.

ii. Soient $\{x_1,...x_k\}$ les points inclus dans la boule de centre x^* et de rayon $dmin$.

   On pose : $BDD_2=BDD_2\setminus\{x_1,...x_k\}$.

iii. Soit $x' \in BDD_2$ le point le plus proche de BDD_1 (réalisant le minimum des distances euclidiennes entre les points de BDD_2 et BDD_1).

   On pose : $BDD_1=BDD_1 \cup\{x'\}$ *et* $BDD_2 = BDD_2\setminus\{x'\}$.

iv. Soient $\{x_{j1},...,x_{jl}\}$ les points inclus dans la boule de centre *x'* et de rayon $dmin$.

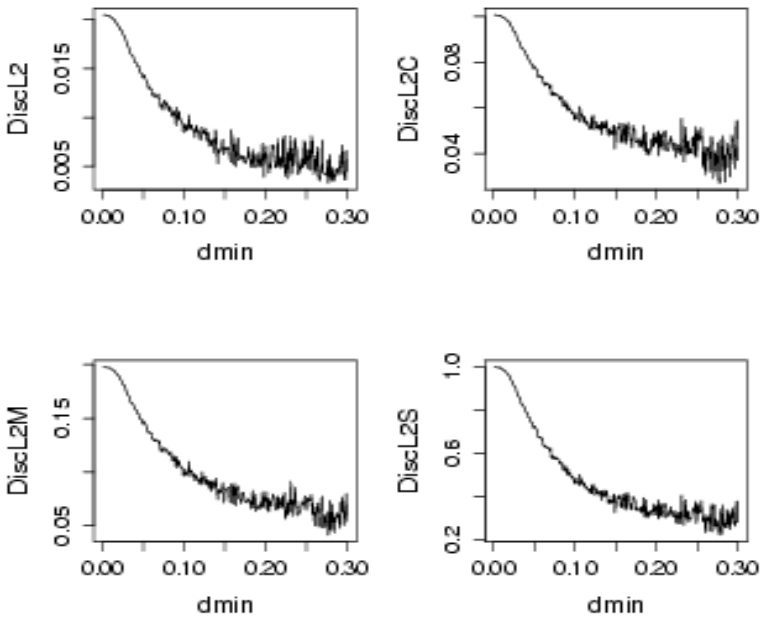
   On pose : $BDD_2 = BDD_2\setminus\{ x_{j1},...,x_{jl} \}$

v. Itération des étapes 3 et 4 tant que $BDD_2 \neq \varnothing$.

→ Espacements réguliers

ETAPE 2 Sélection de points

- Applications de l'algorithme (A1) pour différentes *dmin*



Application de l'algorithme pour :

$0.1 \leq dmin \leq 0.3$ (par pas de 0.001).

- $dmin = 0.277$ réalise le minimum des discrépances, la base de données sélectionnée comporte 16 points
- $dmin = 0.113$ réalise le minimum des discrépances pour $0.1 \leq dmin \leq 0.3$, la base de données sélectionnée comporte 99 points

→ Choix de $dmin = 0.113$

→ avis d'ingénieur ↔ compromis

→ Discrépance la plus faible « possible »

ETAPE 2 Sélection de points, remarques

- Critères d'espacements satisfaisants (choix de l'utilisateur)

- Diminution de la discrépance par rapport à la base de données initiale

- Si la base de données initiale ne recouvre pas l'espace de façon acceptable (par ex : des trous), il en sera de même avec une base sélectionnée à partir de celle-ci.

- Difficulté du choix du nombre de points et de la distance minimale (pas de relation connue en dimension >3)

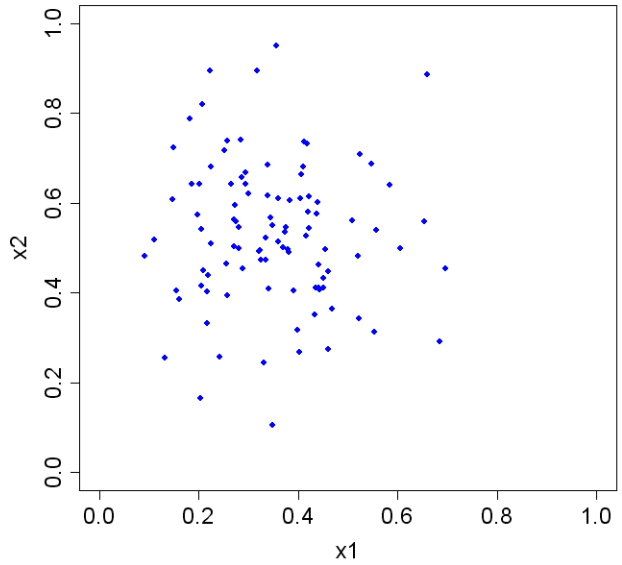
ETAPE 2 Sélection de points

Algorithme A3

i. Choix d'une suite à discrépance faible de n points

ii. Sélection des m points de la base initiale les plus proches de la suite à discrépance faible choisie

iii. On itère les étapes i. et ii. tant que $m < n$



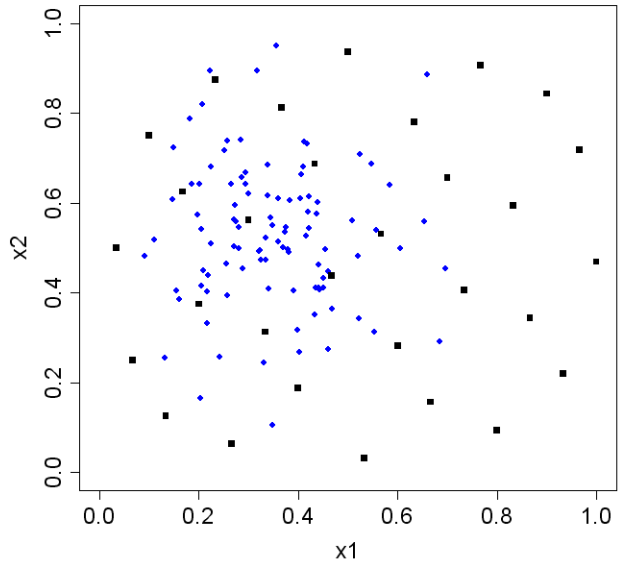
ETAPE 2 Sélection de points

Algorithme A3

i. Choix d'une suite à discrépance faible de n points

ii. Sélection des m points de la base initiale les plus proches de la suite à discrépance faible choisie

iii. On itère les étapes i. et ii. tant que $m < n$



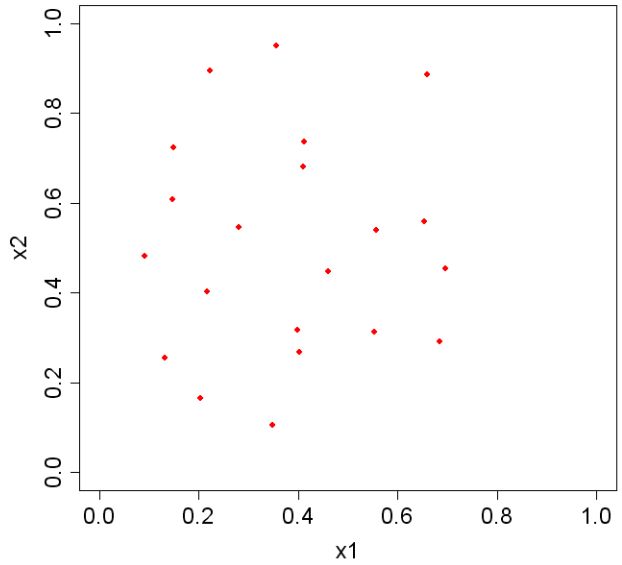
ETAPE 2 Sélection de points

Algorithme A3

i. Choix d'une suite à discrépance faible de n points

ii. Sélection des m points de la base initiale les plus proches de la suite à discrépance faible choisie

iii. On itère les étapes i. et ii. tant que $m < n$



ETAPE 2 Sélection de points

Algorithme A3

i. Choix d'une suite à discrépance faible de n points

ii. Sélection des m points de la base initiale les plus proches de la suite à discrépance faible choisie

iii. On itère les étapes i. et ii. tant que $m < n$

Théorème : *(Rafjlowicz E., Schwabe, 2005)*

Pour certains types de fonction f (continue et à variation bornée), pour une suite d'estimation fn de f de la forme :

$$\hat{f}_n(x) = \sum_{k=1}^{N} \hat{\beta}_{kn} v_k(x) \quad \begin{cases} \text{où } \dfrac{1}{n}\sum_{i=1}^{n} f(x_i) v_i(x) \\ v_1, \ldots, v_k \text{ suite de fonction orthnormale dans } L^2([0,1]^d) \end{cases}$$

Si l'on considère comme points d'expérimentation les points de suites à discrépance faible de Halton ou Hammersley, alors :

$$IMSE(f_n, f) = E\left(\int_{[0,1]^d} (f_n(x) - f(x))^2 dx \right) \longrightarrow 0$$

si $f \in W^{\mu}([0,1]^d)$, avec $\mu > d/2$, et f_n obtenue par régression trigonométrique,

alors : $IMSE(f_n, f) = O(n^{-2\mu/(2\mu+d)})$

ETAPE 2 Spécification de points

- Si la base initiale ou la base sélectionnée ne sont pas jugées de qualité suffisante (ex : dispersion de la base de données initiale trop importante ⇒ une discrépance élevée pour la base sélectionnée)

- Et s'il y a possibilité de réaliser d'autres expériences

→ Application de l'algorithme (A2)

i. On constitue l'ensemble E={$x_1,..,x_n,s_1,...s_n$} où $x_1,..,x_n$ est la base de données sans redondance, et $s_1,...s_n$ est une suite à discrépance faible.

ii. Pour chaque point x_i, on supprime les points de $s_1,...s_n$ dont la distance est inférieure à ε.

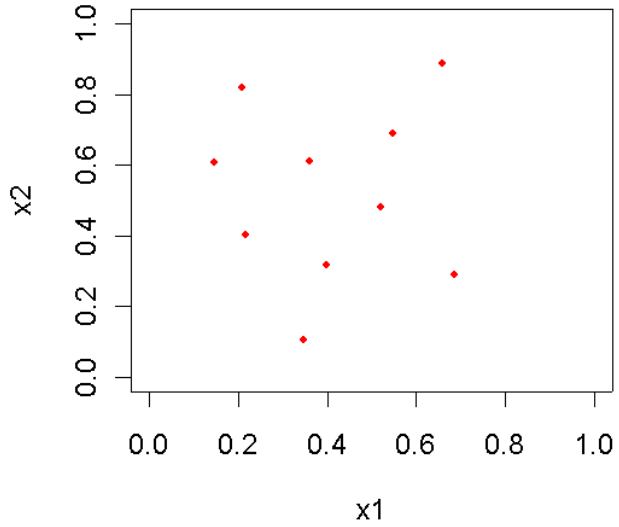
iii. On applique l'algorithme (A1) aux points de la suite $s_1,...s_n$ n'ayant pas été supprimés.

On obtient 380 points.

(Ce nombre dépend de la distance *dmin*, en ajustant cette distance, il est possible d'obtenir une base de données ayant le nombre de points souhaité)

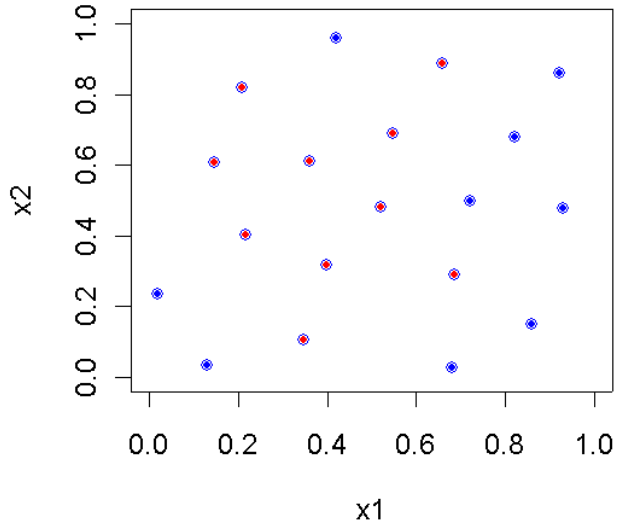
ETAPE 2 Spécification de points

BDD sélectionnée



ETAPE 2 Spécification de points

BDD spécifiée (à l'aide d'un réseau)



4 Étapes

- Étape 1 Collecte d'informations, bien définir le problème
- Étape 2 Analyser la base de données initiale $\{x_i\}_{i=1..N}$ et des $\{\theta_i\}_{i=1..K}$
- Étape 3 Appliquer la démarche de calibration
- Étape 4 Valider les jeux de données θ

ETAPE 3 La calibration

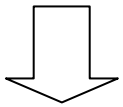
Contexte

$D1 = \{ X = (x_1, \ldots, x_n); Y_{obs} = (y_{obs}(x_1, \theta^*), \ldots, y_{obs}(x_n, \theta^*)) \}$

$D2 = \{ \Theta = (\theta_1, \ldots, \ldots \theta_N) ; Y(\theta_1) = (f(x_1, \theta_1), \ldots, f(x_n, \theta_1)) ;$

$Y(\theta_N) = (f(x_1, \theta_N), \ldots, f(x_n, \theta_N)) \}$

$D = D1 \cup D2$



- Technique par régression multiple
  (Hypothèse de linéarité en les paramètres)
- Technique GLUE
  (approche bayésienne, approximation de la loi a posteriori du paramètre par un loi discrète)
- Technique de Kennedy et O'Hagan
  (approche bayésienne avec approximation de la fonction de code par un processus gaussien)

Etape 3 Technique par régressions multiples

- Méthode

$$\Upsilon = \begin{pmatrix} y(x_1,\theta_1) & y(x_2,\theta_1) & \cdots & y(x_n,\theta_1) \\ y(x_1,\theta_2) & y(x_2,\theta_2) & \cdots & y(x_n,\theta_2) \\ \cdots & \cdots & \cdots & \cdots \\ y(x_1,\theta_k) & y(x_2,\theta_k) & \cdots & y(x_n,\theta_k) \end{pmatrix} \qquad \Theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{pmatrix}$$

- On effectue une régression de $\Upsilon(x_1,.)$ sur Θ : $\Upsilon(x_1,.) = \alpha_1 + \Theta.\beta_1 + \varepsilon_1$

- On effectue une régression de $\Upsilon(x_2,.)$ sur Θ : $\Upsilon(x_2,.) = \alpha_2 + \Theta.\beta_2 + \varepsilon_2$

Finalement

$$\Rightarrow y(x_1,\theta_1) \approx \alpha_1 + \langle \theta_1, \beta_1 \rangle$$

$$\Rightarrow y(x_2,\theta_1) \approx \alpha_1 + \langle \theta_1, \beta_2 \rangle$$

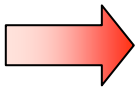
$$\begin{pmatrix} y(x_1,\theta_1) \\ \vdots \\ y(x_n,\theta_1) \end{pmatrix} = \begin{pmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_n \end{pmatrix} + \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_n \end{pmatrix} .\theta_1 + \varepsilon \Rightarrow \hat{\theta} = \left(\hat{B}\, R^{-1} \hat{B}^t\right)^{-1} \hat{B} R^{-1} (y - \hat{\alpha})$$

ETAPE 3 Technique par régressions multiples

Modélisation

$$\mathbf{Y} = \alpha + \Theta B + E$$

$$Y_{obs} = \alpha + \theta * B + E'$$



- Hypothèse de linéarité en θ

($\Rightarrow$ utilisation des critères de plan d'expérience)

- Hypothèse d'erreurs gaussiennes

- Méthode utilisée lorsque Y correspond à des expériences
  - Y sont des expériences « coûteuses » dont les paramètres (θ_1 , …, θ_K) sont connus
  - Y_{obs} expériences moins « coûteuses » dont le paramètre θ^* est inconnu $\rightarrow$ *« calibration »*

ETAPE 3 Technique « GLUE »

• Méthode

Loi a priori $\pi(\theta) \rightarrow$ approximation de la loi a posteriori du paramètre θ par une loi discrète (θ_{VA},$L(\theta_{VA}|Y)$)

- Générer N vecteurs θ_i à l'aide de la loi a priori $\pi(\theta)$
- Calcul des probabilités $p(\theta_i)$, $p(Y|\theta_i)$
- Approximation de la loi a posteriori de θ_i :

$$L(\theta_i|Y) = \frac{p(y|\theta_i).p(\theta_i)}{\sum_{i'=1}^{n} p(Y|\theta_{i'})\,p(\theta_{i'})}$$

- Estimation de θ, par exemple, à l'aide de la moyenne $\bar{\theta} = \sum_{i=1}^{n} L(\theta_i|Y).\theta_i$

On utilise d'autres fonctions dites de vraisemblance :

$$L(\theta_i|Y) \propto \left(\frac{1}{\sigma_i^{\,2}}\right)^N \quad avec \quad \sigma_i^{\,2} = \frac{1}{n}\sum_{j=1}^{n}\left(y_{obs}(x_j,\theta) - y_{code}(x_j,\theta_i)\right)^2$$

ETAPE 3 Technique « GLUE »

- Remarque

- On peut ignorer l'approche bayésienne et considérer :

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{j=1}^{n} \left(y_{obs}(x_j, \theta) - y_{code}(x_j, \theta_i) \right)^2 \right\}$$

- On peut aussi considérer différents critères en vue de réaliser une optimisation multi-objectif

Perspective

Approche GLUE avec plusieurs « critères » → optimisation multi-objectifs

ETAPE 3 Technique bayésienne (Kennedy, O'Hagan)

Modélisation $z(x_i) = y'(x_i, \theta_{VA}) + \delta_i + e_i$

- $e_i \sim N(0,\lambda)$ où λ paramètre à estimer
- $y'(.,.) \sim N(m1(.,.), V1[(.,.)(.,.)])$ où $m1(x,\theta) = h1(x,\theta)^T.\beta 1$ et $V1[(,)(,)]$ une fonction de variance stationnaire de paramètres représentés par le vecteur $\psi 1$
- $\delta(.) \sim N(m2(.), V2(.))$ où $m2 = h2(x)^T.\beta 2$ et $V2(x)$ une fonction de variance stationnaire de paramètres représentés par le vecteur $\psi 2$

Notations

$$\beta^t = (\beta_1^t, \beta_2^t)^t \; ; \quad \psi = (\psi_1^t, \psi_2^t)^t \; ; \quad \phi^t = (\lambda, \psi^t) \; ; \quad d^t = (y^t, z^t)$$

ETAPE 3 Technique bayésienne (Kennedy, O'Hagan)

• Méthode

Calcul de la loi de « $d|\ (\theta, \beta, \phi)$ » (d vecteur de toutes les réponses)

- $\beta \sim$ loi uniforme
- $p(\theta, \beta, \phi) \propto p(\theta).p(\phi)$ (indépendance de θ et ϕ)
- $p((\theta, \beta, \phi)|d) \propto p(\theta, \beta, \phi)\ p(\mathrm{d}|\ (\theta, \beta, \phi))$ (formules de Bayes)

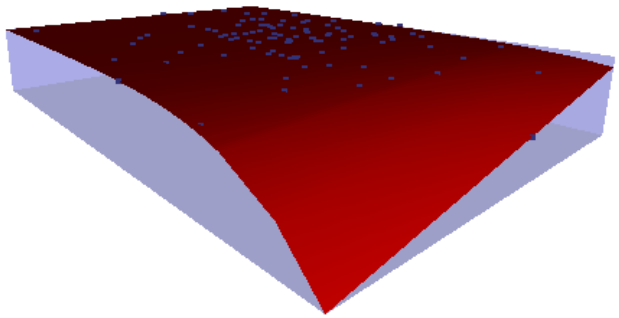
  $\propto p(\theta).p(\phi)\ p(\mathrm{d}|\ (\theta, \beta, \phi))$
- Intégration / $\beta \rightarrow p((\theta, \phi)|d)$
- Approximation : $p(\theta|(\phi = \phi_{est}, d)) \propto p((\theta, \phi est)|d)$ (intégration difficile / ϕ)

• Estimation des paramètres ϕ

- (ψ_1) à partir des $\{y_i = y(x_i, \theta_i)\}_{i=1..N}$ (à l'aide de toutes les simulations)
- (λ, ψ_2) à partir des $\{d_i^T = (z(x_i)^T, y(x_i, \theta)^T)\}_{i=1..N}$ (à partir de toutes les simulations et réponses expérimentales)

ETAPE 3 Application

- Application



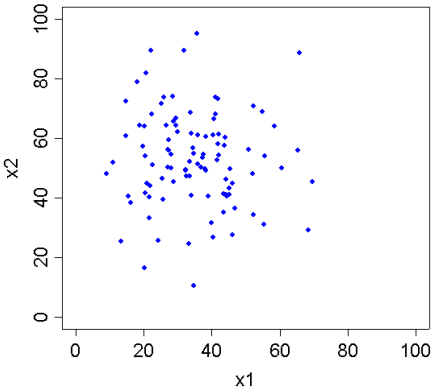
Modèle $y(x,\theta)=\theta_1-\theta_2\, x_1 \exp(-\,\theta_3 x_2)$

- On fixe $\boldsymbol{\theta = (1, 0.5, 0.1)}$
- On considère les domaines de variation

$$X=[0,100]\times[0,100]\ ;\ \Theta=[0.96,2.9]\times[0.2,1.2]\times[0.07;0.2]$$

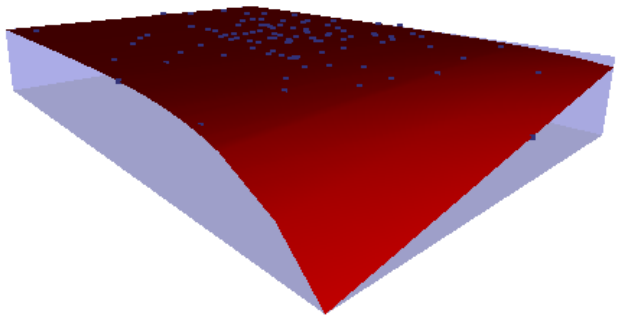
- On considère différentes bases de données initiales pour X :

X_{ini} 100 points



ETAPE 3 Application

- Application



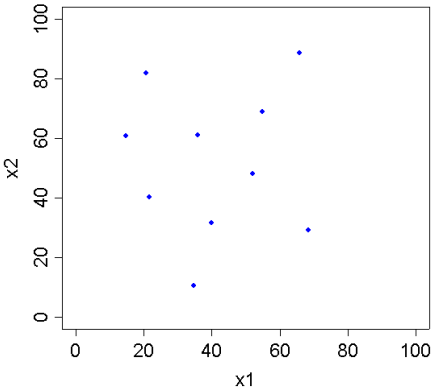
Modèle $y(x,\theta)=\theta_1-\theta_2\, x_1 \exp(-\,\theta_3 x_2)$

- On fixe $\boldsymbol{\theta = (1, 0.5, 0.1)}$
- On considère les domaines de variation

$$X=[0,100]\times[0,100]\ ;\ \Theta=[0.96,2.9]\times[0.2,1.2]\times[0.07;0.2]$$

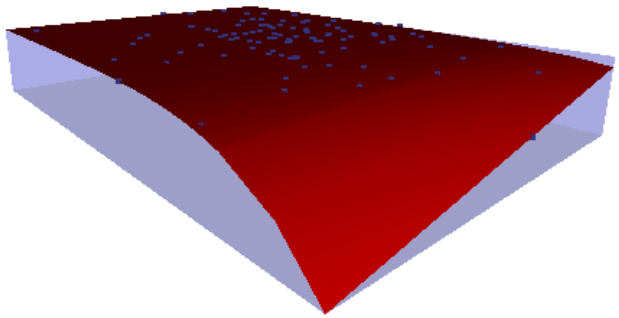
- On considère différentes bases de données initiales pour X :

X_{sel} sélectionnée par l'algorithme A1 (10 points)



ETAPE 3 Application

- Application



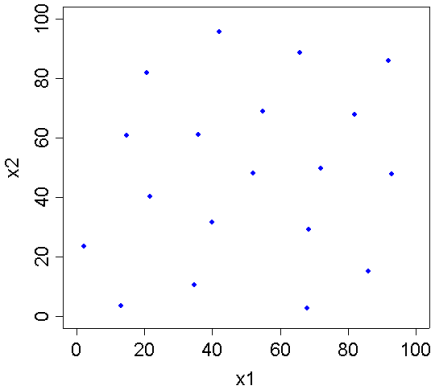
Modèle $y(x,\theta)=\theta_1-\theta_2\, x_1 \exp(-\,\theta_3 x_2)$

- On fixe $\boldsymbol{\theta = (1, 0.5, 0.1)}$
- On considère les domaines de variation

$$X=[0,100]\times[0,100]\ ;\ \Theta=[0.96,2.9]\times[0.2,1.2]\times[0.07;0.2]$$

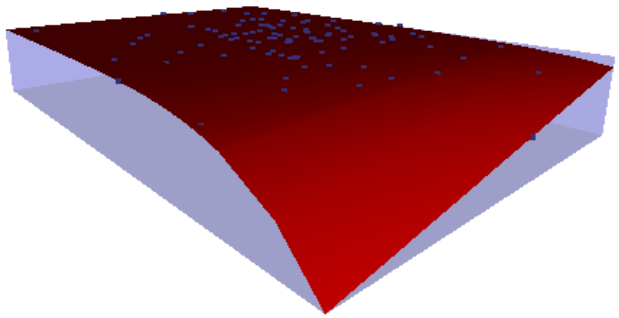
- On considère différentes bases de données initiales pour X :

X_{sp} spécifiée par A3 (19 points)



ETAPE 3 Application

- Application

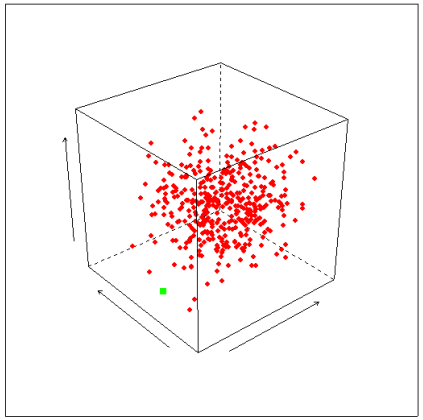


Modèle $y(x,\theta)=\theta_1-\theta_2\, x_1 \exp(-\,\theta_3 x_2)$

- On fixe $\boldsymbol{\theta = (1, 0.5, 0.1)}$
- On considère les domaines de variation

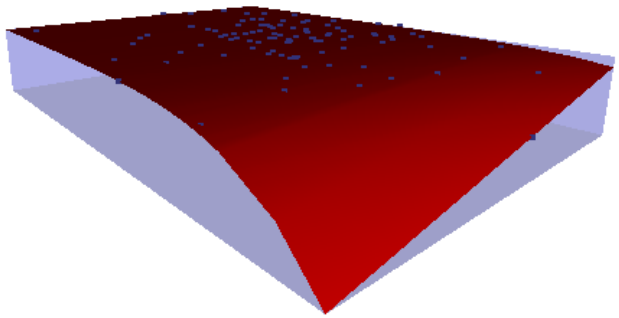
  $X=[0,100]\times[0,100]$; $\Theta=[0.96,2.9]\times[0.2,1.2]\times[0.07;0.2]$
- On considère différents ensembles de paramètres pour Θ

Θ_{ini} 400 points



ETAPE 3 Application

- Application

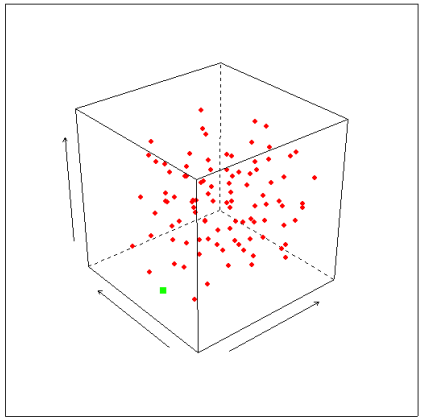


Modèle $y(x,\theta)=\theta_1-\theta_2\, x_1 \exp(-\,\theta_3 x_2)$

- On fixe $\boldsymbol{\theta = (1, 0.5, 0.1)}$
- On considère les domaines de variation

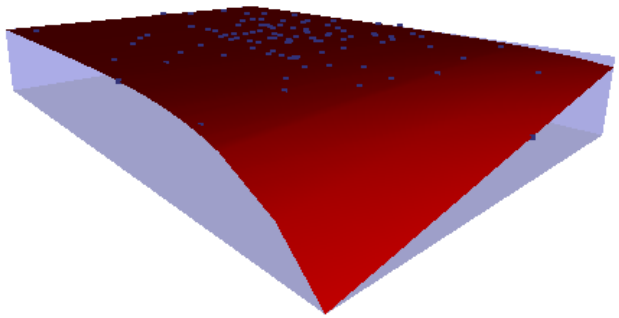
  $X=[0,100]\times[0,100]$; $\Theta=[0.96,2.9]\times[0.2,1.2]\times[0.07;0.2]$
- On considère différents ensembles de paramètres pour Θ

Θ_{sel} sélectionnée par l'algorithme A1 (99 points)



ETAPE 3 Application

- Application



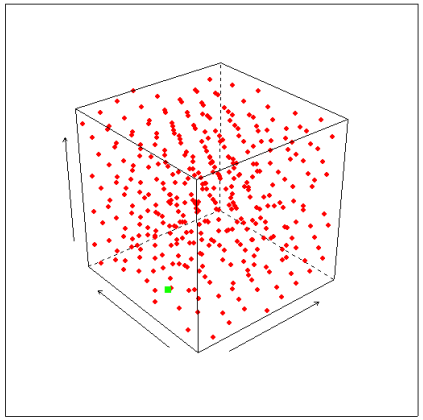
Modèle $y(x,\theta)=\theta_1-\theta_2\, x_1 \exp(-\,\theta_3 x_2)$

- On fixe $\boldsymbol{\theta = (1, 0.5, 0.1)}$
- On considère les domaines de variation

$$X=[0,100]\times[0,100]\ ;\ \Theta=[0.96,2.9]\times[0.2,1.2]\times[0.07;0.2]$$

- On considère différents ensembles de paramètres pour Θ

Θ_{sp} spécifiée par A3 (380 points)



ETAPE 3 Application : Régressions multiples

Estimation du paramètre :

$$\hat{\theta} = \left(\hat{B}\,\hat{B}^t\right)^{-1}\hat{B}\left(y_{obs} - \hat{\alpha}\right) \qquad Er(\hat{\theta}) = \frac{1}{3}\sum_{j=1}^{3}\left(\frac{\hat{\theta}_j}{\theta_j} - 1\right)^2$$

	X_{ini}	X_{sel}	X_{spec}
Θ_{ini}	(1.017, 0.364 ,0.0716)	(1.072, 0.359 ,0.069)	(0.994, 0.482 ,0.0898)
Er	0.05164	0.06027	0.00392
Θ_{sel}	(0.999,0.463,0.0903)	(1.052, 0.377, 0.073)	**(1.050, 0.483, 0.0916)**
Er	0.00492	0.04537	**0.00357**
Θ_{spec}	(1.018,0.595, 0.124)	(0.930, 0.630, 0.130)	(1.06, 0.502 ,0.113)
Er	0.031341	0.05417	0.00684

En général (100 cas tests)

Pour les différentes bases X :

- des résultats « comparables » pour X_{ini} et X_{sel}
- de bien meilleurs résultats pour X_{spec}

Pour les différents ensembles Θ :

- des résultats « comparables » pour Θ_{ini} et Θ_{sel}
- de meilleurs résultats pour Θ_{spec}

ETAPE 3 Application Méthode GLUE

Estimation du paramètre

$$L(\theta_i|Y) \propto \left(\frac{1}{\sigma_i^{\,2}}\right)^N \quad avec \quad \sigma_i^{\,2} = \frac{1}{n}\sum_{j=1}^{n}\left(y_{obs}(x_j,\theta) - y_{code}(x_j,\theta_i)\right)^2 \qquad \bar{\theta} = \sum_{i=1}^{n} L(\theta_i|Y).\theta_i$$

	X_{ini}	X_{sel}	X_{spec}
Θ_{ini}	(1.030,0.613,0.112)	(1.257,0.610,0.115)	(1.353,0.517,0.102)
Er	0.02213	0.04565	0.04205
Θ_{sel}	(1.047,0.605,0.110)	(1.200,0.598,0.112)	(1.302,0.497,0.100)
Er	0.01877	0.03094	0.03041
Θ_{spec}	**(1.036,0.598,0.108)**	(1.213,0.619,0.110)	(1.299,0.493,0.0946)
Er	**0.01537**	0.03734	0.03084

Pour les différentes bases X :

- de meilleurs résultats pour X_{ini}
- des résultats comparables pour X_{sel} et X_{spec}

Pour les différents ensembles Θ :

- des résultats « comparables » pour Θ_{sel} et Θ_{spec}

ETAPE 3 Application Remarques

- La méthode par régressions multiples
  - Elle permet en général d'obtenir des résultats plus précis sur les paramètres dont la relation avec la réponse est linéaire. (Le cadre théorique permet l'utilisation de plan d'expérience)

- La méthode GLUE
  - Elle permet d'obtenir une estimation du paramètre recherché en pondérant les paramètres initiaux ⇒ la base de données initiale des paramètres a une influence sur l'estimation, il est donc important de bien recouvrir tout le domaine (absence de techniques de plan d'expérience pour cette méthode)

- Application en cours de la méthode de Kennedy et O'Hagan. Utilisation de la library BACCO de R.
  - Application délicate à mettre en oeuvre
  - Ne marche que pour un nombre restreint de données
  - Le paramètre estimé peut ne pas avoir de « sens physique »

4 Étapes

- Étape 1 Collecte d'informations, bien définir le problème
- Étape 2 Analyser la base de données initiale $\{x_i\}_{i=1..N}$ et des $\{\theta_i\}_{i=1..K}$
- Étape 3 Appliquer la démarche de calibration
- Étape 4 Valider les jeux de données θ

ETAPE 4 Valider le jeux de paramètres θ

- Cohérence des résultats

- A minima retrouver les résultats des expériences aux incertitudes près.

On se focalisera sur des données sélectionnées pour l'estimation, et on vérifiera sur les données restantes la qualité de la prédiction.

CONCLUSION PERSPECTIVE

- Combiner Régressions multiples et technique bayésienne
  - Détecter les linéarités en certains paramètres → les estimer par la méthode de régressions multiples
  - Estimer les paramètres « restants » par une méthode plus élaborée, (technique bayésienne de Kennedy et O'Hagan)
- Combiner régression trigonométrique, GLUE
  - Remplacer la fonction de code par une régression trigonométrique en θ
  - Approximation de la loi a posteriori de par θ une loi discrète

Références Analyse de la BDD (approche « déterministe »)

• Rafajlowicz E. and Schwabe R., Halton and Hammersley sequences in multivariate nonparametric regression, *Statistics and probability letters*, 2005

• Hickernell F.J., A generalized discrepancy and quadrature error bound, *Mathematics of computation*, 1998

• Burkardt J. and Gunzburger, Uniformity measures for point samples in *hypercube, www.csit.fsu.edu/~burkardt2004:/ptmeas.pdf*, 2004

Références Régressions multiples

- Sundberg R., Multivariate Calibration – Direct and Indirect methodology, *Scandinavian Journal of Statistics*, vol 26, 161-207, 1999

- Brown, Philip J., Inverse prediction, Report UKC/IMS/00/18 University of Kent at Canterbury

Références GLUE

• Beven K.J. and A. Binley, The future of distributed models: model calibration and uncertainty prediction, *Hydrol. Process*., 6, 279-298, 1992.

• Ratto M., Tarantola S., Saltelli A., Sensivity analysis in model calibration : GSA-GLUE approach, *Computer Physics Communications*

Références Kennedy et O'Hagan

• Kennedy M., and O'Hagan A., Bayesian calibration of computer models*, J.R. Statis. Soc, 2001*)

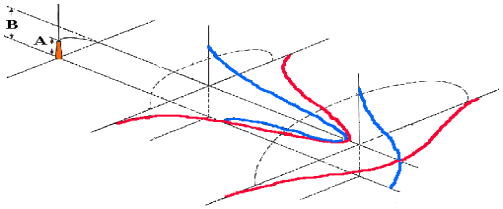
• Kennedy M., and O'Hagan A., Supplementary details on Bayesian calibration of computer models, *Internal report, University of Sheffield* http://www.shef.ac.uk/~st1a0/ps/CALSUP.PS, 2001

• Hankin R.K.S. Introducing Bacco, an R bundle for Bayesian analysis of computer code output, *Journal of Statistical Software*, 2005

Application envisageable

Application : Accident d'usine chimique

- *x* comporte deux composantes correspondant à des coordonnées Nord-Est.
- θ correspond au « terme source » et la vitesse de déposition des substances.
- $y = y(x, \theta)$ est le « Gaussian plume model »
- *z* : la mesure du dépôt de substances en certains points.



(Kennedy M., O'Hagan A., *Bayesian calibration of computer models, J.R. Statis. Soc, 2001*)

Application envisageable

Application à la qualification du modèle de frottement isotherme dans le RJH (Dumas M., Gaudier F.) :

x :

G : la vitesses massique axiale

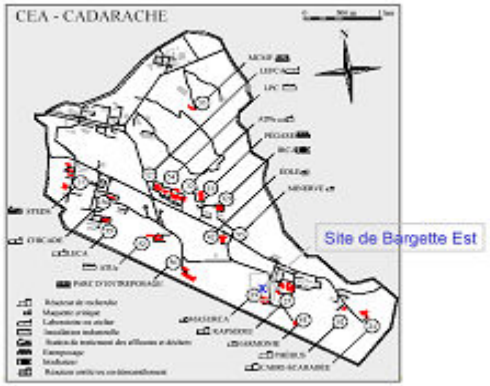
H : la hauteur frottante

D_h : le diamètre hydraulique

ΔP^{Mes} : la perte de charge expérimentale

ρ_l : la masse volumique liquide

ρ_c : la masse volumique liquide dans la canalisation



Réponse « observée » : $z(\underbrace{G, H, \rho_l, \rho_c, D_h, \Delta P^{Mes}}_{x}) = \frac{2\rho_l D_h}{G^2 H} \times \left(\frac{\Delta P^{Mes}}{H} - (\rho_l - \rho_c) g \right)$

Réponse à calibrer : $y(\underbrace{G, H, \rho_l, \rho_c, D_h, \Delta P^{Mes}}_{x}, \underbrace{\theta_1, \theta_2}_{\theta}) = \theta_1 \left(\frac{G D_h}{\mu} \right)^{-\theta_2}$

Application envisageable

Références :

Sensivity analysis in model calibration : GSA-GLUE approach, M. Ratto, S. Tarantola, A. Saltelli, Computer Physics Communications ,

Évoution d'une réaction irréversible isotherme

L'équation considérée est : $y_B = 1 + (y_B{}^0 - 1)\exp\left(-k_\infty \exp\left(-E/RT\right)t\right)$

Le vecteur des paramètres est : $\theta = \left(y_B{}^0, k_\infty, E\right)^t$

Pour simuler les observations $z(ti)$ des erreurs normales ont été rajoutées à la solution analytique $y(t)$ (avec un θ fixé étant la vraie valeur que l'on cherche à estimer)

La fonction de « vraisemblance » considérée est : $L(\theta_i|Y) \propto \left(\frac{1}{\sigma_i{}^2}\right)^N$

$$\sigma_i{}^2 = \frac{1}{N_{obs}} \sum_{j=1}^{N_{obs}} \left(\hat{Y}(t_j) - Y(t_j)\right)^2$$

Critères

Exemple :

pour θ, connaître la loi de probabilité de

$$e_{l^k}(x_1,...,x_n,\theta_j)=\frac{1}{N}\sum_{i=1}^{N}\left|y_{obs}(x_i,\theta^*)-y(x_i,\theta_j)\right|^k \quad \text{pour } k=1 \text{ ou k}=2$$

- optimisation multi-objectif des fonctions (critères de validation)
  - « Performance » $E(\theta)=E\{e_{l^k}(x_1,...,x_n,\theta)\}$
  - « robustesse » $V(\theta)=Var\{e_{l^k}(x_1,...,x_n,\theta)\}$
  - « fiabilité »

    $p(\theta)=P\{e_{l^k}(x_1,...,x_n,\theta)>\varepsilon\}\leq\alpha$ pour ε et α donnés

ETAPE 2 Approche probabilistes

Utilisation de tests d'uniformité

Hypothèse H0 :

la suite $\{x_i\}_{i=1..N}$ *est une réalisation de v.a. i.i.d. uniformes dans* $[0,1]^d$

(après avoir éventuellement normalisé la BDDE).

- Tests périodiques (« Serial Tests ») :

Partition du pavé unité en $h^d = k$ cellules cubiques de même volume : $A_1,\ldots,A_k$.

On note N_j le nombre de points de la suite $\{x_i\}_{i=1..n}$ dans la cellule A_j

considération de statistiques de la forme :

$$Y = \sum_{j=1}^{k} \{f_{n,k}(N_j)\}$$

- Test de Cramer-Von Mises multivarié :

$$\int_{[0,1]^d} (F_n(x) - x_1 \ldots x_d)^2 \, dx_1 \ldots dx_d$$

ETAPE 2 Tests périodiques (Serial Tests)

Exemple de fonctions $f_{n,k}(x)$:

Y^2	$f_{n,k}(x)$	Nom
D_δ	$2x[(x/m)^\delta - 1] / [\delta(1-\delta)]$ avec $\delta > -1$	Puissance de divergence (divergence power)
X^2	$(x-m)^2 / m$	Pearson
G^2	$2x \ln(x/m)$	Log-vraisemblance (loglikelihood)
$-H$	$(x/n) \log_2(x/m)$	Negative entropy
N_b	$1_{[x=b]}$	Nombre de cellules ayant b points
W_b	$1_{[x \geq b]}$	Nombre de cellules ayant au moins b points
N_0	$1_{[x=0]}$	Nombre de cellules vides
C	$(x-1)1_{[x>b]}$	Nombre de collisions

$m = n/k$

- $m > 1$ « *dense serial tests* »
- $m < 1$ « *sparse serial tests* »

Théorème : *Si* $k \to \infty$, $n \to \infty$ *et* $n/k \to \lambda$ *avec* $0 < \lambda < \infty$ *alors*

$$D_\delta^N = \frac{D_\delta - k\mu}{\sigma_N} \to N(0,1) \qquad \text{où } \sigma_N = Var[D_\delta]$$

ETAPE 2 Tests périodiques (Serial Tests)

PROPRIETES

- Moyenne et variance connues
- Théorèmes

Théorème 1 :

pour $\delta >$ -1, *k fixé et* $n \to \infty$ *(« dense case ») alors*

$$D_\delta^C = \frac{D_\delta - k\mu + (k-1)\sigma_C}{\sigma_C} \to \mathcal{X}^2(k-1) \qquad \text{où } \sigma_C^2 = Var[D_\delta]/(2(k-1))$$

Théorème 2 :

Si $k \to \infty$, $n \to \infty$ *et* $n/k \to \lambda$ *avec* $0<\lambda<\infty$ *alors*

$$D_\delta^N = \frac{D_\delta - k\mu}{\sigma_N} \to N(0,1) \qquad \text{où } \sigma_N = Var[D_\delta]$$

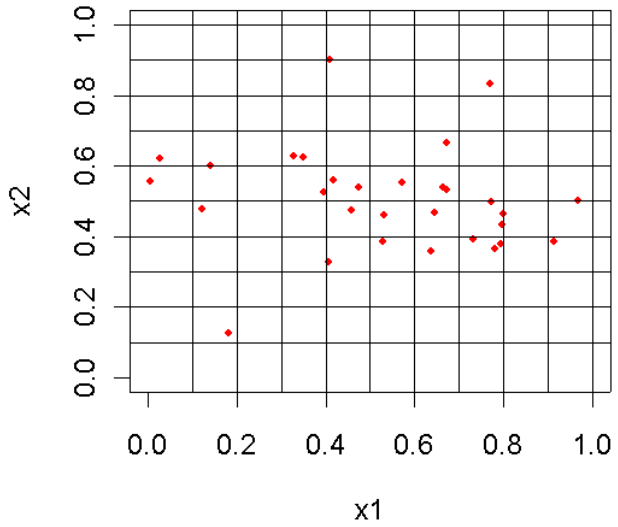
ETAPE 2 Tests périodiques (Serial Tests)

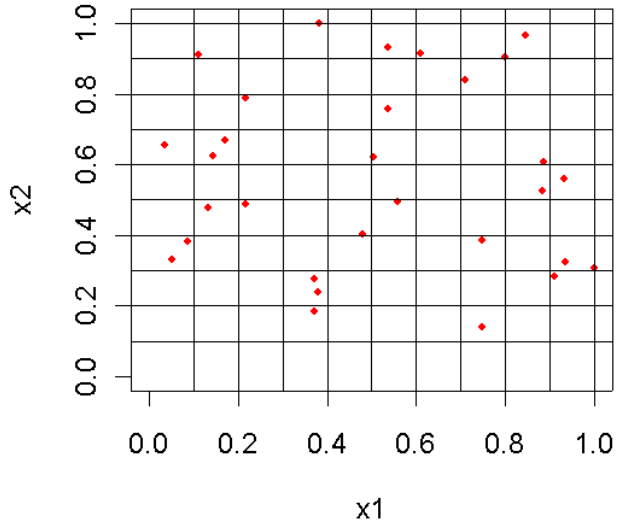
Exemple :

• Application du Théorème 2 :

Région de confiance à 70% : $R_c = \{ \mid D_\delta \mid \leq 1.04 \}$

$\Rightarrow$ Région de rejet : $R_r = \{ \mid D_\delta \mid > 1.04 \}$





$D^{10}{}_{0.5} = 2.15 \rightarrow$ on rejette

$D^{10}{}_{0.5} = -0.11 \rightarrow$ on accepte

ETAPE 2 Test de Cramer Von-Mises

- Statistiques de Cramer Von-Mises :

$$\int_{[0,1]^d} \alpha_n(x_1,...,x_d)^2\,dx_1...dx_d \quad \text{où} \quad \alpha_n(x_1,...,x_d) = \sqrt{n}\left(F_n(x_1,...,x_d)) - x_1...x_d\right)$$

Loi tabulée ⇒ tests.

- Utilisation de processus définis sur les marges

$$\int_{[0,1]^d} \alpha_{n,0}(x_1,...,x_d)^2\,dx_1...dx_d \quad \text{où} \quad \alpha_{n,0}(x_1,...,x_d) = \alpha_n(x) - \sum_{1\le j\le d} x_j \alpha_n(x_1,...,x_{j-1},1,x_{j+1},...,x_d)$$

$$+ \sum_{1\le j<l\le d} x_j \alpha_n(x_1,...,x_{j-1},1,x_{j+1},...,x_{l-1},1,x_{l+1},...,x_d)$$

$$+...+(-1)^d x_1...x_d \alpha_n(1,...,1)$$

On a : $\alpha_{n,0}(x) \xrightarrow{n} B(x)$ avec $E\left(B(x)B(x')\right) = \prod_{j=1}^{d}\left\{\min(x_j, x_j') - x_j x_k'\right\}$

Tabulation de la loi possible à l'aide d'un développement de Kahunen-Loeve (Deheuvels, 2005) ⇒ tests

ETAPE 2 Processus défini sur les marges

Soit $\{B_0(x), \in x[0,1]^d\}$ *un processus gaussien centré de fonction de variance :*

$$E(\mathbf{B}_0(x)\mathbf{B}_0(x')) = \prod_{j=1}^{d} \left\{ \min(x^j, x^{j'}) - x^j x^{j'} \right\}$$

Théorème 5 *Le développement de Karhunen-Loeve de* B_0 *est donné par :*

$$B_0(x) = \sum_{k_1 \geq 1} \cdots \sum_{k_d \geq 1} \sqrt{\lambda_{k_1,\ldots,k_d}} e_{k_1,\ldots,k_d}(t) Y_{k_1,\ldots,k_d} \qquad (1.7)$$

où :

$$\lambda_{k_1,\ldots,k_d} = \prod_{j=1}^{d} \left\{ \frac{2 \times \frac{1}{2}}{z_{1/2,k_j}} \right\}^2$$

$$e_{k_1,\ldots,k_d}(x) = \prod_{j=1}^{d} \left[x_j^{\frac{1}{2\times 1/2} - \frac{1}{2}} \left\{ \frac{J_{1/2}(z_{1/2,k}.x_j^{\frac{1}{2\times 1/2}})}{\sqrt{1/2} J_{1/2-1}(z_{1/2,k})} \right\} \right]$$

et $\{Y_{k_1,\ldots,k_d}, k_1 \geq 1, \ldots, k_d \geq 1\}$, *une suite de variable aléatoires indépendantes et identiquement distribuées de loi normale centré réduite* $N(0,1)$.
$J_{1/2}(x_j)$ *est la fonction de Bessel :*

$$J_{1/2}(x_j) = \frac{x^{1/2}}{2} \sum_{k=0}^{\infty} \frac{\frac{-1}{4} x_j^{2k}}{\Gamma(\frac{1}{2} + k + 1) k!}$$

et $\{z_{1/2,k}, k \geq 1\}$ *la suite croissante des* 0 *de* $J_{1/2}$.

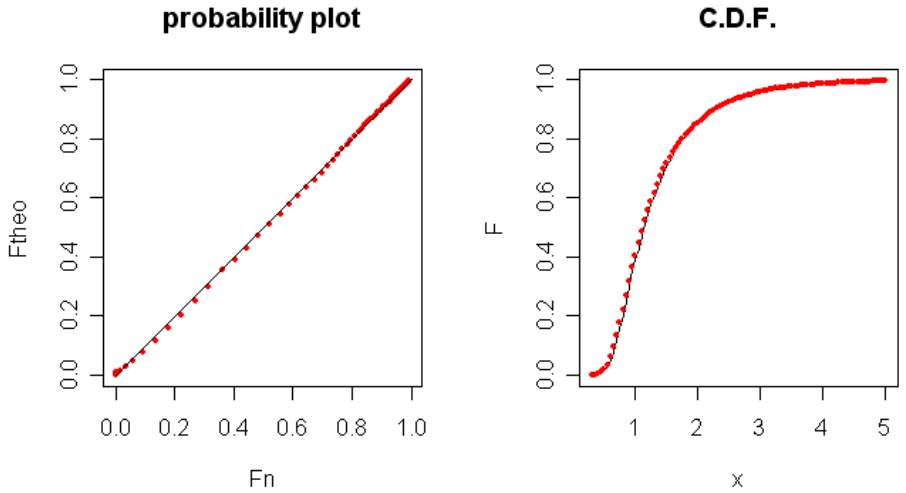
$$\Delta_n(x^1,x^2) = \sqrt(n)(U_n(x^1,x^2) - x^1 U_n(1,x^2) - x^2 U_n(x_1,1) + x^1 x^2)$$

La loi limite de ce processus est celle d'un processus gaussien centré et de fonction de covariance :

$$E[\mathbf{B}_{1,2}(x^1,x^2)\mathbf{B}_{1,2}(x^{1'},x^{2'})] = (\min(x^1,x^{1'}) - x^1 x^{1'})(\min(x^2,x^{2'}) - x^1 x^{2'})$$

Sous l'hypothèse $H0$:

$$\int_I \Delta_n(x^1,x^2)dx^1dx^2 = \sum_{k=1}^{\infty}\left(\prod_{j=1}^{2}\left\{\frac{2\times\frac{1}{2}}{z_{1/2,k_j}}\right\}^2\right)Y^2_{k_1,k_2}$$



comparaison de la loi obtenue par … (à l'aide de la fonction caractéristique) et par le développement de K.L. tronqué

ETAPE 2 Notion de discrépance

On note :

f_u **la projection d'une fonction** f **sur** $[0,1]^u$ **avec** $u \subset \{1,..,d\}$ **et** $\|f\|_2 = \sum_{u \subset \{1,..,d\}} \|f_u\|_{L^2([0,1]^u)}$

Dicrépances de type L2

La discrépance L2: $DiscL^2 = \left\| \frac{\#(J(x))}{n} - \lambda(J(x)) \right\|_{L^2([0,1]^d)}$

$E(DiscL^2) = \frac{1}{n}\left(\left(\frac{1}{2}\right)^d - \left(\frac{1}{3}\right)^d \right)$

La discrépance L2 modifiée: $DiscL^2M = \left\| \frac{\#(J(x))}{n} - \lambda(J(x)) \right\|_2$

$E(DiscL^2M) = \frac{1}{n}\left[\left(\frac{4}{3}+\frac{1}{6}\right)^d - \left(\frac{4}{3}\right)^d \right]$

La discrépance L2 centrée: $DiscL^2M = \left\| \frac{\#(J(x,a))}{n} - \lambda(J(x)) \right\|_2$

(a le plus proche sommet de x)

$E(DiscL^2C) = \frac{1}{n}\left[\left(\frac{13}{12}+\frac{1}{6}\right)^d - \left(\frac{13}{12}\right)^d \right]$

La discrépance L2 symétrique: $DiscL^2M = \left\| \#\frac{(J(x,a_p))}{n} - \lambda(J(x)) \right\|_2$

(a le plus proche sommet « pair » de x)

$E(DiscL^2S) = \frac{1}{n}\left[\left(\frac{4}{3}+\frac{2}{6}\right)^d - \left(\frac{4}{3}\right)^d \right]$

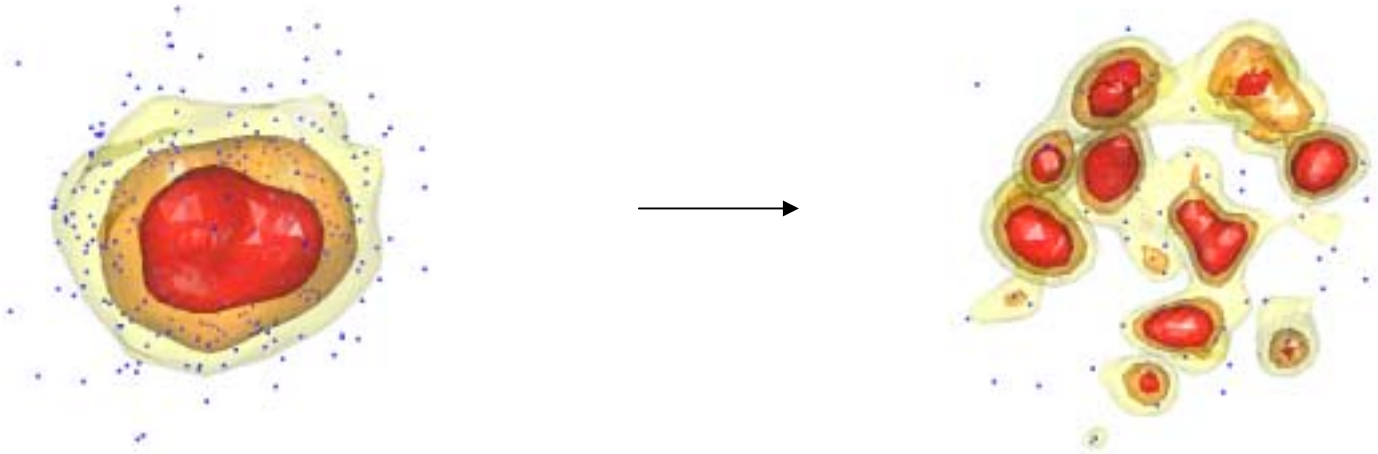
« Valeurs de références Supérieures »

ETAPE 2 Sélection de points

Algorithme A3

i. Estimation de la densité des points de la base de données initiale.

ii. Tirage des points parmi la BDDI avec une probabilité inversement proportionnelle à la valeur de la densité en ces points



densité de la base initiale

densité de la base sélectionnée

ETAPE 2 Remarques

- Contrôle du nombre de points sélectionnés

- Valeur de la discrépance obtenue comparable à celle des bases obtenues par les Algorithmes A1 et A2

- Hypothèse d'uniformité rejetée si la base initiale ne recouvre pas de façon « acceptable » le pavé unité (ex : des trous)

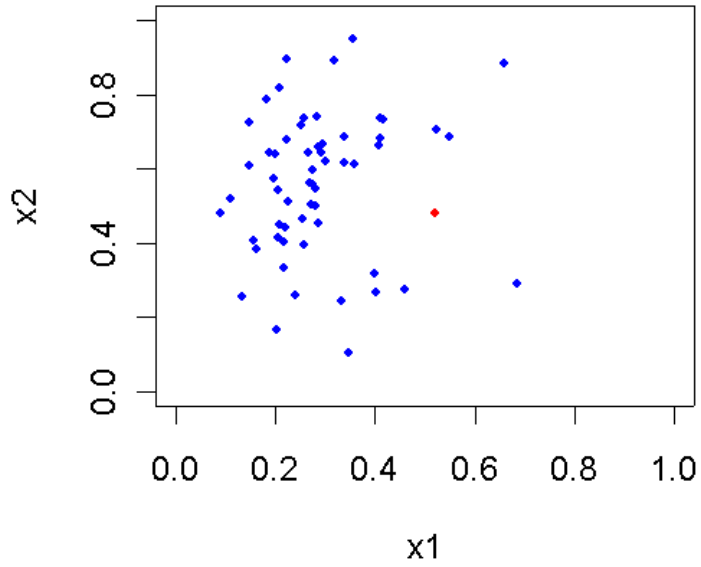
- Méthode limitée car l'estimation de la densité devient délicate en dimension>6 (nombre de points nécessaire très important)

Méthodologie : Etape 2

étape *ii*

Soient $\{x_1,...x_k\}$ les points inclus dans la boule de centre x^* et de rayon $dmin$.

On pose : $BDD_2=BDD_2\backslash\{x_1,...x_k\}$.

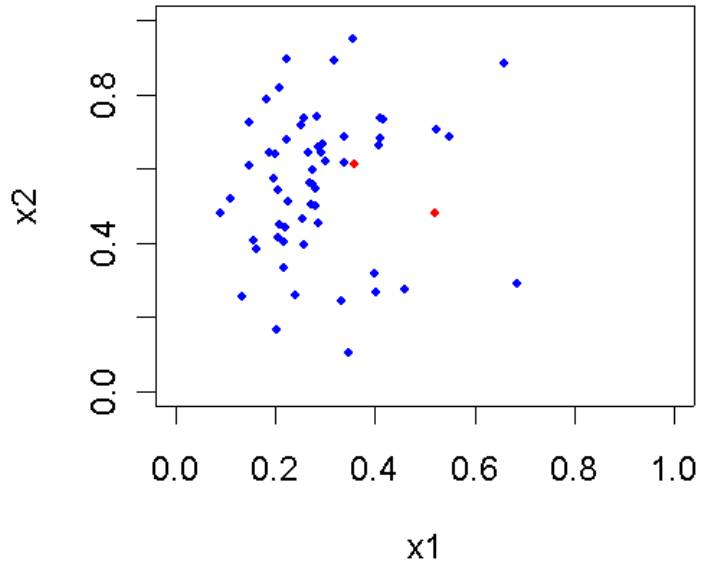


Méthodologie : Etape 2

étape *iii*

Soit $x' \in BDD_2$ le point le plus proche de BDD_1 (réalisant le minimum des distances euclidiennes entre les points de BDD_2 et BDD_1).

On pose : $BDD_1 = BDD_1 \cup \{x'\}$ *et* $BDD_2 = BDD_2 \setminus \{x'\}$.

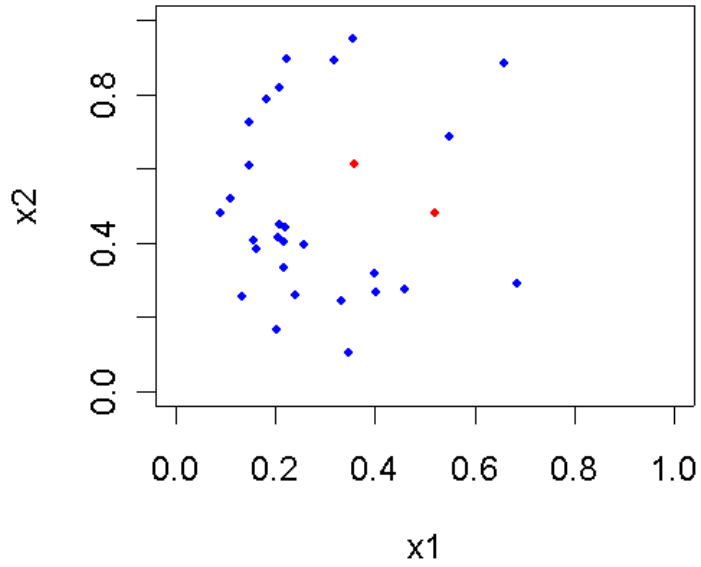


Méthodologie : Etape 2

étape *iv*

Soient $\{x_{j1},...,x_{jl}\}$ les points inclus dans la boule de centre *x'* et de rayon $dmin$.

On pose : $BDD_2 = BDD_2 \setminus \{ x_{j1},...,x_{jl} \}$

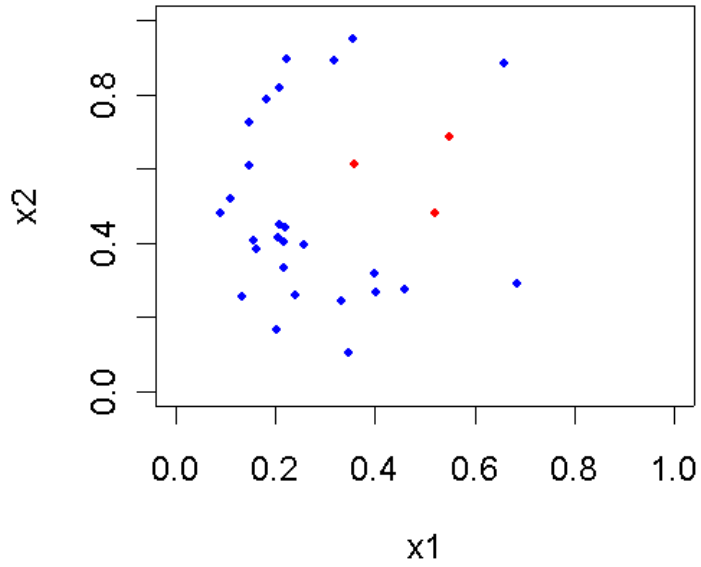


Méthodologie : Etape 2

Itération étape *iii*

Soit $x' \in BDD_2$ le point le plus proche de BDD_1 (réalisant le minimum des distances euclidiennes entre les points de BDD_2 et BDD_1).

On pose : $BDD_1 = BDD_1 \cup \{x'\}$ *et* $BDD_2 = BDD_2 \setminus \{x'\}$.

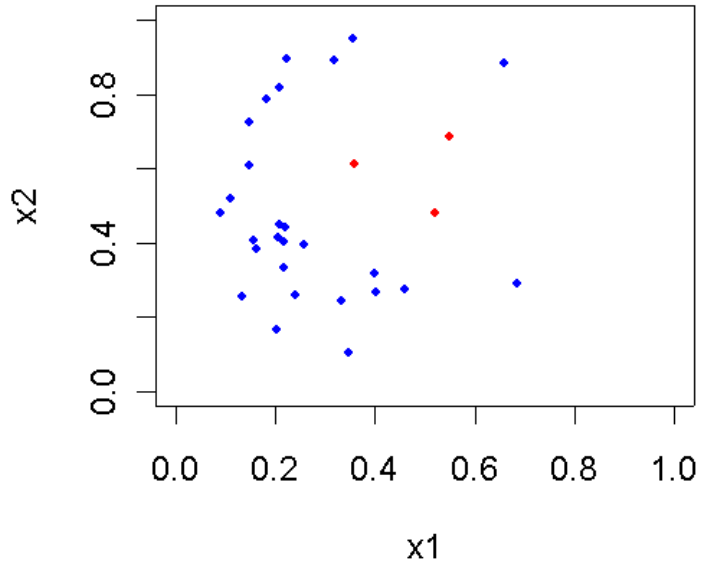


Méthodologie : Etape 2

Itération étape *iv*

Soient $\{x_{j1},...,x_{jl}\}$ les points inclus dans la boule de centre *x'* et de rayon $dmin$.

On pose : $BDD_2 = BDD_2 \setminus \{ x_{j1},...,x_{jl} \}$

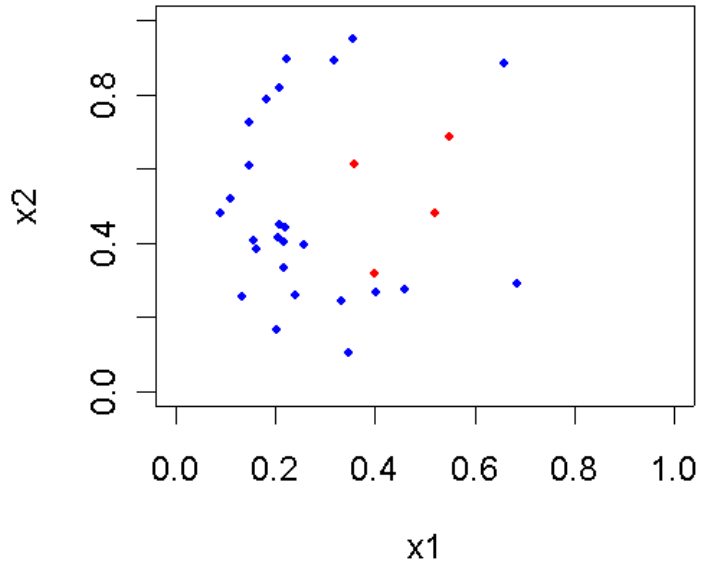


Méthodologie : Etape 2

Itération étape *iii*

Soit $x' \in BDD_2$ le point le plus proche de BDD_1 (réalisant le minimum des distances euclidiennes entre les points de BDD_2 et BDD_1).

On pose : $BDD_1 = BDD_1 \cup \{x'\}$ *et* $BDD_2 = BDD_2 \setminus \{x'\}$.

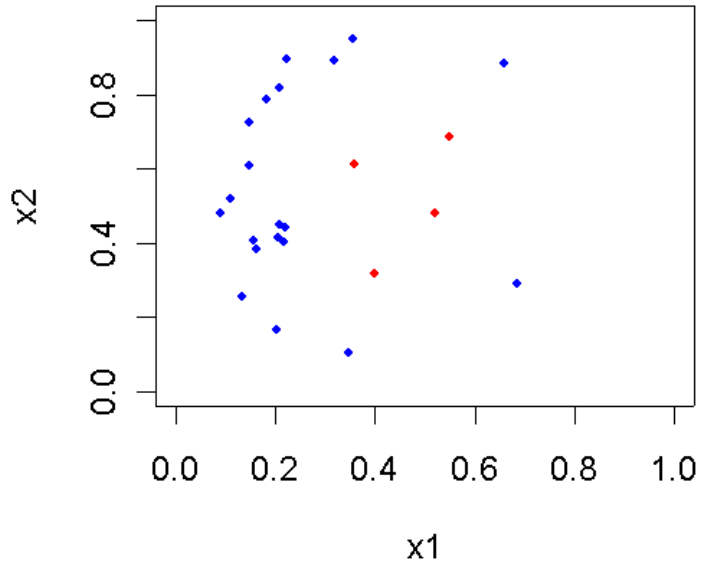


Méthodologie : Etape 2

Itération étape *iv*

Soient $\{x_{j1},...,x_{jl}\}$ les points inclus dans la boule de centre *x'* et de rayon $dmin$.

On pose : $BDD_2 = BDD_2 \backslash \{ x_{j1},...,x_{jl} \}$

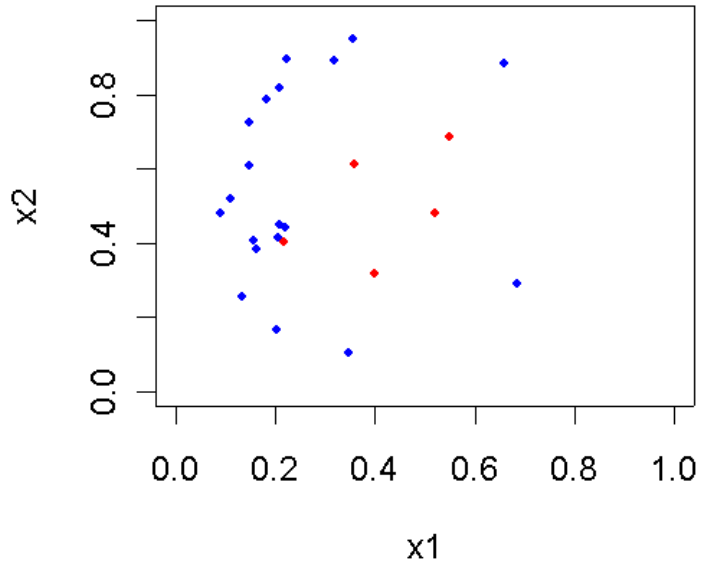


Méthodologie : Etape 2

Itération étape *iii*

Soit $x' \in BDD_2$ le point le plus proche de BDD_1 (réalisant le minimum des distances euclidiennes entre les points de BDD_2 et BDD_1).

On pose : $BDD_1 = BDD_1 \cup \{x'\}$ *et* $BDD_2 = BDD_2 \setminus \{x'\}$.

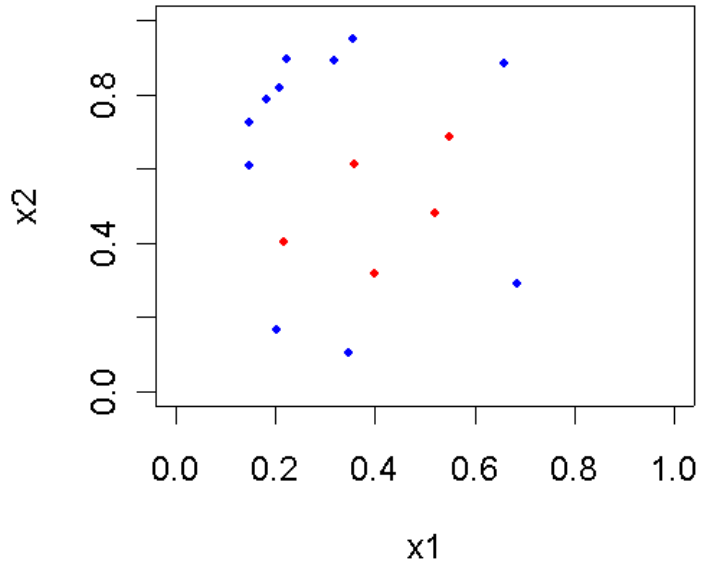


Méthodologie : Etape 2

Itération étape *iv*

Soient $\{x_{j1},...,x_{jl}\}$ les points inclus dans la boule de centre *x'* et de rayon $dmin$.

On pose : $BDD_2 = BDD_2 \setminus \{ x_{j1},...,x_{jl} \}$

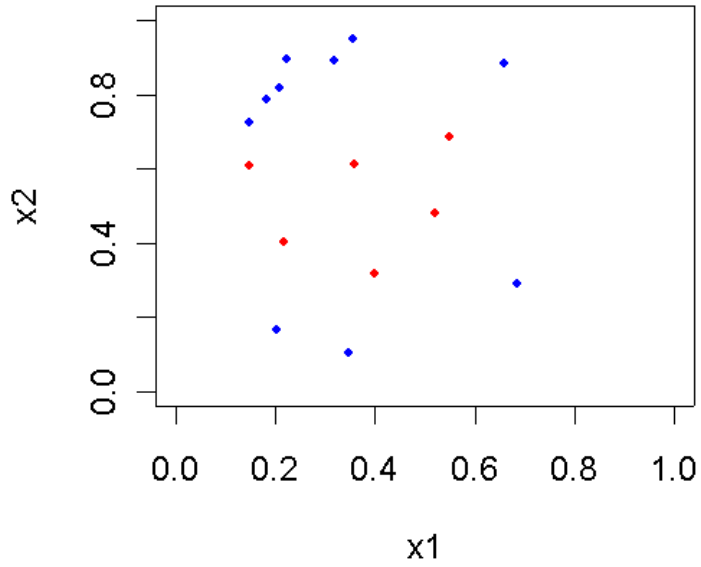


Méthodologie : Etape 2

Itération étape *iii*

Soit $x' \in BDD_2$ le point le plus proche de BDD_1 (réalisant le minimum des distances euclidiennes entre les points de BDD_2 et BDD_1).

On pose : $BDD_1 = BDD_1 \cup \{x'\}$ *et* $BDD_2 = BDD_2 \setminus \{x'\}$.

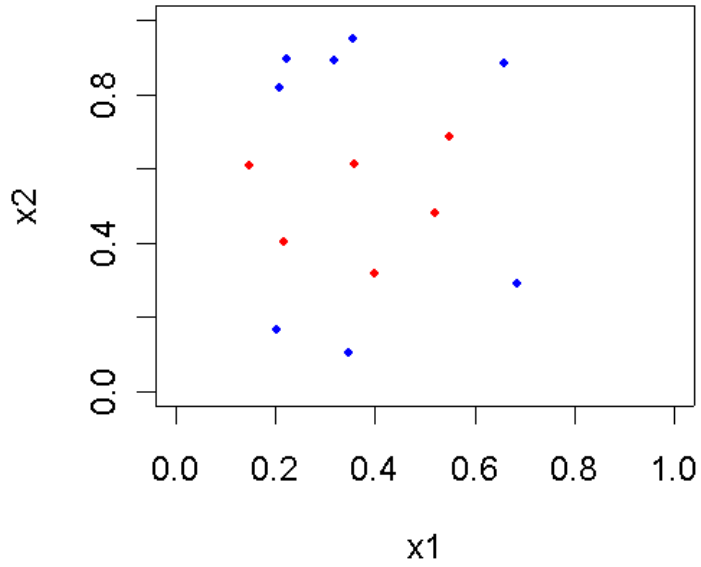


Méthodologie : Etape 2

Itération étape *iv*

Soient $\{x_{j1},...,x_{jl}\}$ les points inclus dans la boule de centre *x'* et de rayon $dmin$.

On pose : $BDD_2 = BDD_2 \setminus \{ x_{j1},...,x_{jl} \}$

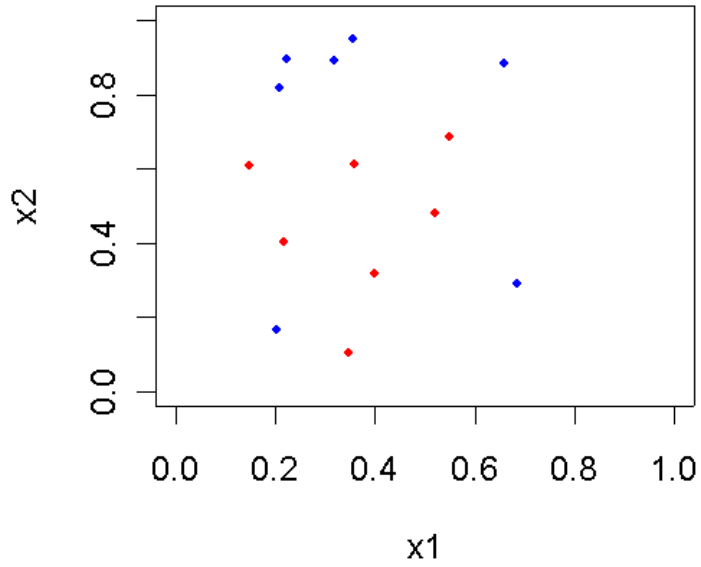


Méthodologie : Etape 2

Itération étape *iii*

Soit $x' \in BDD_2$ le point le plus proche de BDD_1 (réalisant le minimum des distances euclidiennes entre les points de BDD_2 et BDD_1).

On pose : $BDD_1 = BDD_1 \cup \{x'\}$ *et* $BDD_2 = BDD_2 \setminus \{x'\}$.

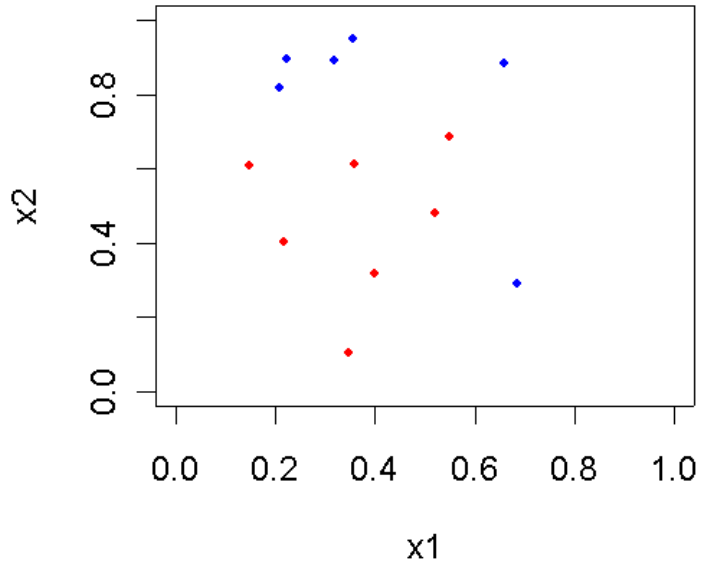


Méthodologie : Etape 2

Itération étape *iv*

Soient $\{x_{j1},...,x_{jl}\}$ les points inclus dans la boule de centre *x'* et de rayon $dmin$.

On pose : $BDD_2 = BDD_2 \setminus \{ x_{j1},...,x_{jl} \}$

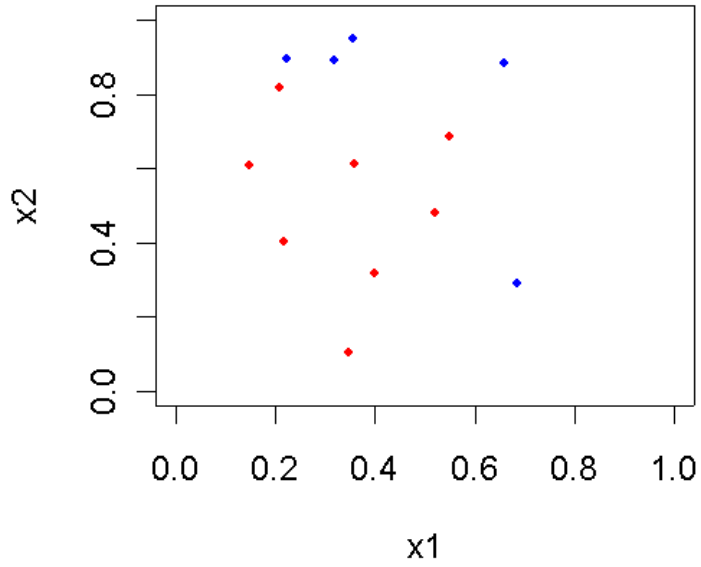


Méthodologie : Etape 2

Itération étape *iii*

Soit $x' \in BDD_2$ le point le plus proche de BDD_1 (réalisant le minimum des distances euclidiennes entre les points de BDD_2 et BDD_1).

On pose : $BDD_1 = BDD_1 \cup \{x'\}$ *et* $BDD_2 = BDD_2 \setminus \{x'\}$.

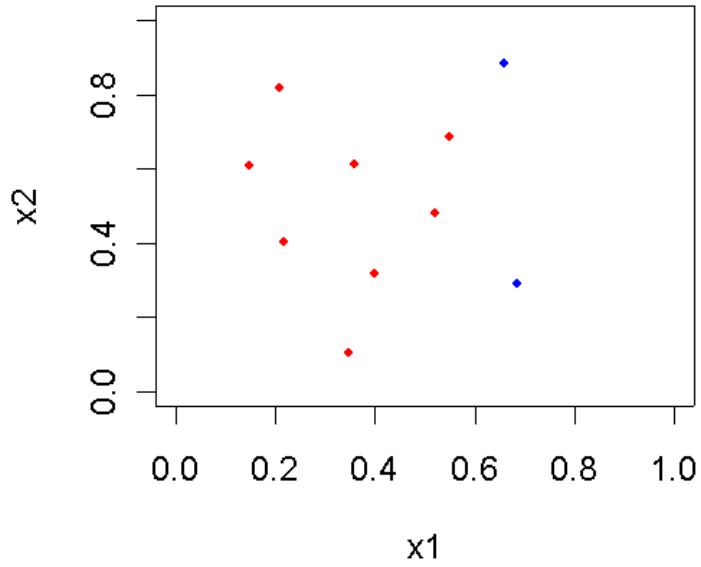


Méthodologie : Etape 2

Itération étape *iv*

Soient $\{x_{j1},...,x_{jl}\}$ les points inclus dans la boule de centre *x'* et de rayon $dmin$.

On pose : $BDD_2 = BDD_2 \backslash \{ x_{j1},...,x_{jl} \}$

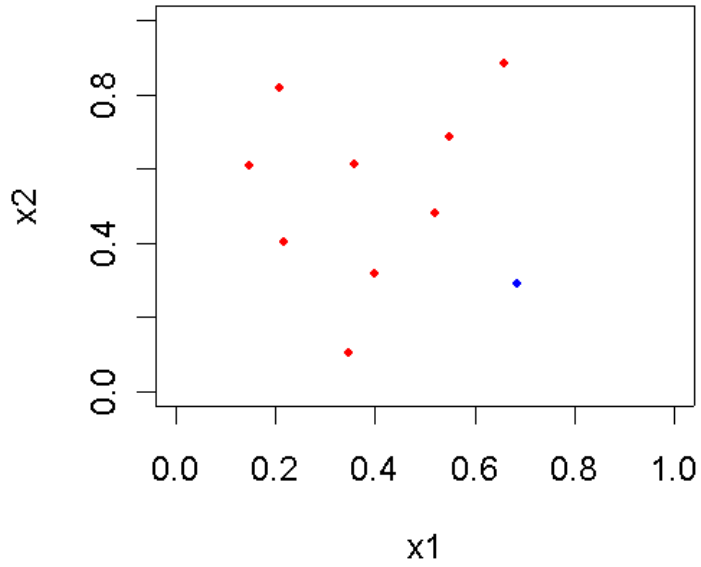


Méthodologie : Etape 2

Itération étape *iii*

Soit $x' \in BDD_2$ le point le plus proche de BDD_1 (réalisant le minimum des distances euclidiennes entre les points de BDD_2 et BDD_1).

On pose : $BDD_1 = BDD_1 \cup \{x'\}$ *et* $BDD_2 = BDD_2 \setminus \{x'\}$.

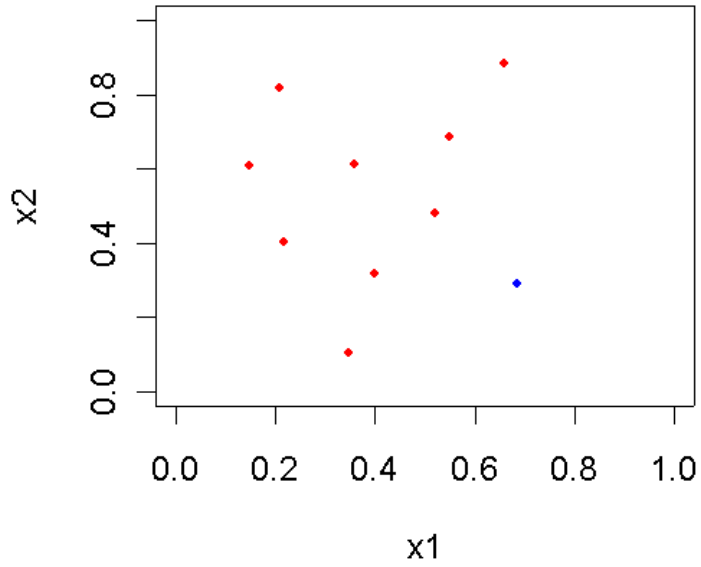


Méthodologie : Etape 2

Itération étape *iv*

Soient $\{x_{j1},...,x_{jl}\}$ les points inclus dans la boule de centre *x'* et de rayon $dmin$.

On pose : $BDD_2 = BDD_2 \backslash \{ x_{j1},...,x_{jl} \}$

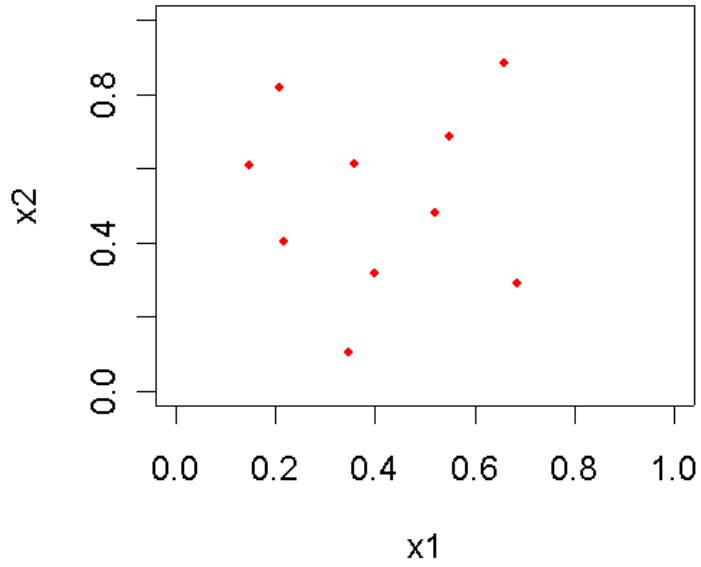


Méthodologie : Etape 2

Itération étape *iii*

Soit $x' \in BDD_2$ le point le plus proche de BDD_1 (réalisant le minimum des distances euclidiennes entre les points de BDD_2 et BDD_1).

On pose : $BDD_1 = BDD_1 \cup \{x'\}$ *et* $BDD_2 = BDD_2 \setminus \{x'\}$.



$BDD_2 \neq \varnothing$. $\Rightarrow$ fin de l'itération

ETAPE 2 Approche « déterministe », remarque

Théorème :

Pour certains types de fonction f (continue et à variation bornée), pour une suite d'estimation f_n de f de la forme :

$$\hat{f}_n(x) = \sum_{k=1}^{N} \hat{\beta}_{kn} v_k(x) \quad \begin{cases} \text{où } \dfrac{1}{n}\sum_{i=1}^{n} f(x_i) v_i(x) \\ v_1, \ldots, v_k \text{ suite de fonction orthnormale dans } L^2([0,1]^d) \end{cases}$$

Si l'on considère comme points d'expérimentation les points de suites à discrépance faible de Halton ou Hammersley, alors :

$$IMSE(f_n, f) = E\left(\int_{[0,1]^d} (f_n(x) - f(x))^2 dx \right) \longrightarrow 0$$

si $f \in W^{\mu}([0,1]^d)$, avec $\mu > d/2$, et f_n obtenue par régression trigonométrique, alors : $IMSE(f_n, f) = O(n^{-2\mu/(2\mu+d)})$

ETAPE 2 Sélection de points, remarques

- Valeur de la discrépance comparable à celle obtenue avec l'algorithme A1

- pas de contrôle de la distance entre points ⇒ critères d'espacements moins satisfaisant que pour l'algorithme A1

- Si la base de données initiale ne recouvre pas l'espace de façon acceptable (par ex : des trous), il en sera de même avec une base sélectionnée à partir de celle-ci (même remarque que pour 'algorithme A1)

Méthodologie : Etape 1

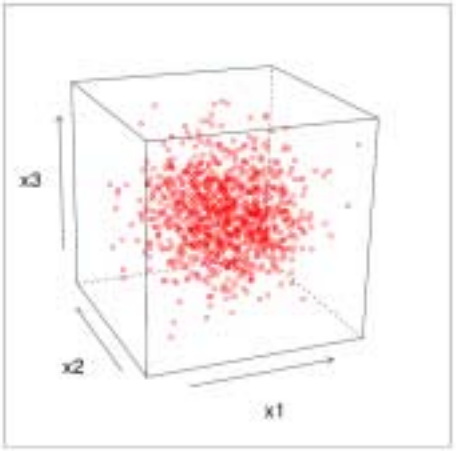
Interprétation des différents critères pour les suites à discrépance faible (400 points)

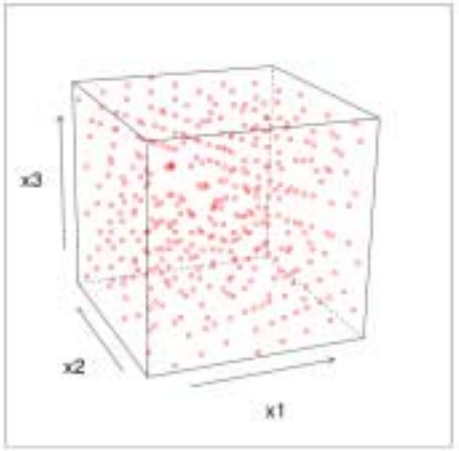
CRITERE	Halton	Hammersley	Faure	Réseau	VALEUR DE REFERENCE
$dmin_2$	0.0476	0.0415	0.0342	0.0931	
$dmin_\infty$	0.0384	0.03	0.0304	0.0718	$Disp_\infty/2$
γ	3.72	3.72	4.86	1.88	1
$m_{2,1}$	3.200	3.192	3.212	3.195	
λ	0.261	0.166	0.236	0.0620	0
$DispHa_\infty(1000)$	0.130	0.141	0.162	0.138	$2.dmin_\infty$
$DispHam_\infty$	0.131	0.122	0.113	0.112	$2.dmin_\infty$
$DispFa_\infty$	0.115	0.106	0.123	0.129	$2.dmin_\infty$
$DispRes_\infty$	0.123	0.107	0.128	0.112	$2.dmin_\infty$
h	0.203	0.169	0.204	0.183	$\downarrow$
μ	2.164	1.737	2.248	1.883	1
$\mathcal{X}$	6.117	6.186	8.422	3.939	$\downarrow$
ν	9.367	2.853	5.990	6.564	1
τ	0.00236	0.00210	0.003301	0.00222	0
D	1.042E-8	1.340E-9	5.144E-9	2.992 E-9	0
$DiscL2$	0.00416	0.00286	0.00391	0.00335	$\downarrow$
$DiscL2M$	0.0109	0.00759	0.00862	0.006752	$\downarrow$
$DiscL2C$	0.00827	0.00652	0.00730	0.00602	$\downarrow$
$DiscL2S$	0.0274	0.0212	0.0300	0.0267	$\downarrow$

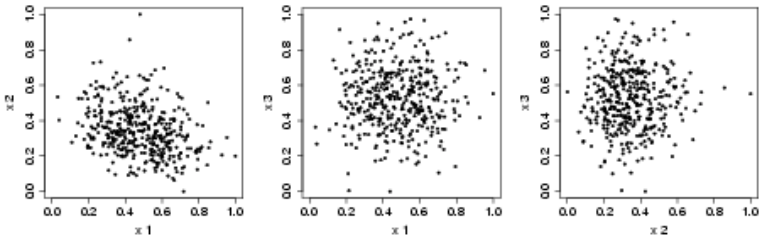
Méthodologie : Etape 3

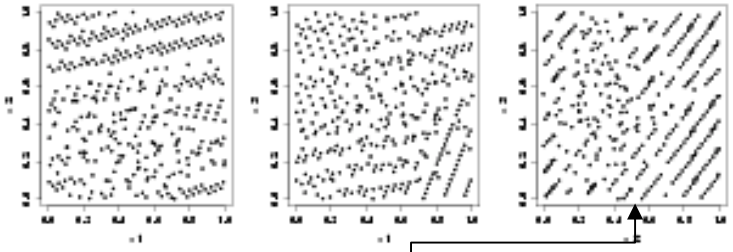
Comparaison visuelle

Moins de points mais mieux « equirépartis »









Effet du choix de la suite

Recommandations

• Choix de la suite à discrépance faible délicat pour la comparaison et le calcul de la dispersion : la propriété de « discrépance faible » est asymptotique.

• Choix de la suite à discrépance faible délicat pour la spécification de nouveaux points : les points spécifiés auront les mêmes propriétés que ceux de la suite à discrépance faible utilisée.

• Nécessité d'une interaction avec les spécialistes de la physique étudiée pour appliquer la méthodologie et interpréter les résultats obtenus : nombre de points minimum à conserver, choix d'une distance caractérisant une redondance d'information.