Incertitudes en cinétique

François WAHL (Francois.wahl@ifp.fr)
IFP BP3 69390 VERNAISON

Formalisation

Procédés habituellement représentés par un modèle non-linéaire

Modèle $\boxed{\mathbf{y} = f(\mathbf{c};\boldsymbol{\theta}_0)}$

Formalisation

Procédés habituellement représentés par un modèle non-linéaire

Modèle $\boxed{\mathbf{y} = f(\mathbf{c};\boldsymbol{\theta}_0)}$

$\mathbf{c}$ = entrées = compositions des réactifs + conditions opératoires

Formalisation

Procédés habituellement représentés par un modèle non-linéaire

Modèle $\boxed{\mathbf{y} = f(\mathbf{c};\boldsymbol{\theta}_0)}$

$\mathbf{c}$	=	entrées	=	compositions des réactifs + conditions opératoires
$\mathbf{y}$	=	sorties	=	compositions de l'effluent $\boldsymbol{\theta}_0$

Formalisation

Procédés habituellement représentés par un modèle non-linéaire

Modèle $\boxed{\mathbf{y} = f(\mathbf{c}; \boldsymbol{\theta}_0)}$

$\mathbf{c}$ = entrées = compositions des réactifs + conditions opératoires

$\mathbf{y}$ = sorties = compositions de l'effluent $\boldsymbol{\theta}_0$

$\boldsymbol{\theta}_0$ = vrais paramètres du modèle (inconnus) = k, α

Formalisation

Procédés habituellement représentés par un modèle non-linéaire

Modèle $\boxed{\mathbf{y} = f(\mathbf{c};\boldsymbol{\theta}_0)}$

$\mathbf{c}$	=	entrées	=	compositions des réactifs + conditions opératoires
$\mathbf{y}$	=	sorties	=	compositions de l'effluent $\boldsymbol{\theta}_0$
$\boldsymbol{\theta}_0$	=	vrais paramètres du modèle (inconnus) = k, α		
f	=	fonction liant $\mathbf{c}$, $\boldsymbol{\theta}_0$ et $\mathbf{y}$ (solution d'un système d'EDO, résolu numériquement)		

Désulfuration

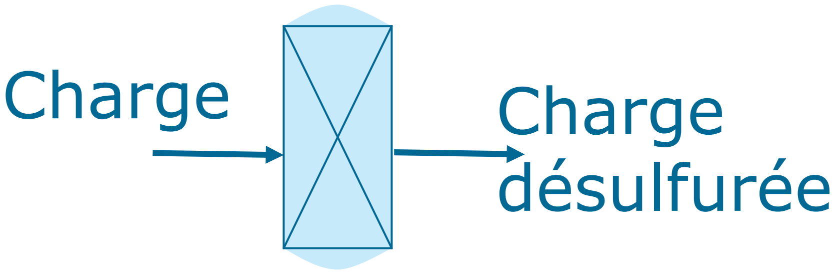


Schéma réactionnel

S $\longrightarrow$ ...

Modèle $\boxed{y = f(\mathbf{c}, \boldsymbol{\theta})}$

$$\frac{dS}{dt} = -kS^{\alpha}$$

$\mathbf{c}$ = [S_0 , t]

$\boldsymbol{\theta}$ = (k, α)

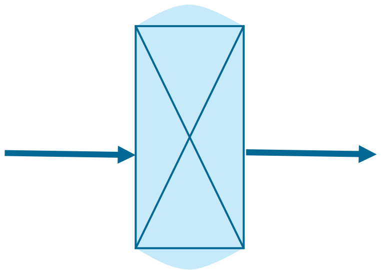
y = S_{Effluent}

f = fonction liant **c**, **θ** et y

$$S_{\text{Effluent}} = (S_0^{1-\alpha} - (1-\alpha)kt)^{1/1-\alpha}$$

Isomérisation des C4

nC4
iC4



nC4
iC4
C3+C5

Conditions opératoires
(P, T, ...)

Schéma réactionnel

$$nC4 \underset{}{\overset{k_1,E_1}{\rightleftarrows}} iC4$$

$$2iC4 \xrightarrow{k_2,E_2} C3+C5$$

$$nC4+iC4 \xrightarrow{k_3,E_3} C3+C5$$

b_{iC4}, b_{C3+C5}

Modèle $\boxed{\mathbf{y} = f(\mathbf{c}, \boldsymbol{\theta})}$

$\mathbf{c}$ = [(nC4, iC4)$_{CHARGE}$ (P,T,...)] — variables contrôlées

$\boldsymbol{\theta}$ = $(k_1, k_2, k_3, E_1, E_2, E_3, b_{iC4}, b_{C3+C5})$ — paramètres à déterminer(catalyseur)

$\mathbf{y}$ = (nC4, iC4, C3+C5)$_{RECETTE}$ — réponses

f = fonction liant **x**, θ et **y** — programme informatique

Formalisation

Procédés habituellement représentés par un modèle non-linéaire

Modèle $\boxed{\mathbf{y} = f(\mathbf{c}, \boldsymbol{\theta}_0)}$

$\mathbf{c}$ = entrées = compositions des réactifs + conditions opératoires

$\mathbf{y}$ = sorties = compositions de l'effluent $\boldsymbol{\theta}_0$

$\boldsymbol{\theta}_0$ = vrais paramètres du modèle (**inconnus**) = k, α

f = fonction liant $\mathbf{c}$, $\boldsymbol{\theta}_0$ et $\mathbf{y}$ (solution d'un système d'EDO, résolu numériquement)

Problématique

Mesures expérimentales $(c_i, y_{obs,i})_{i=1,n}$

Présence de bruit $y_{obs,i} \neq y_{vrai,i}$ et $\mathbf{y}_{obs}$ réalisation de Y aléatoire

Modèle statistique $\boxed{Y = f(\mathbf{c}, \boldsymbol{\theta}_0) + \varepsilon}$

Problématique

Mesures expérimentales $(c_i, y_{obs,i})_{i=1,n}$

Présence de bruit $y_{obs,i} \neq y_{vrai,i}$ et $\mathbf{y}_{obs}$ réalisation de Y aléatoire

Modèle statistique $\boxed{Y = f(\mathbf{c}, \boldsymbol{\theta}_0) + \varepsilon}$

θ_0 estimé par $\boxed{\boldsymbol{\theta}_{obs} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} w_i (y_{obs,i} - f(\mathbf{c}_i, \boldsymbol{\theta}))^2}$

$\Rightarrow$ $\boldsymbol{\theta}_{obs}$ réalisation de $\widehat{\boldsymbol{\theta}}$ aléatoire

Problématique

Mesures expérimentales $(c_i, y_{obs,i})_{i=1,n}$

Présence de bruit $y_{obs,i} \neq y_{vrai,i}$ et $\mathbf{y}_{obs}$ réalisation de Y aléatoire

Modèle statistique $\boxed{Y = f(\mathbf{c}, \boldsymbol{\theta}_0) + \varepsilon}$

$\boldsymbol{\theta}_0$ estimé par $\boxed{\boldsymbol{\theta}_{obs} = \underset{\boldsymbol{\theta} \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} w_i (y_{obs,i} - f(\mathbf{c}_i, \boldsymbol{\theta}))^2}$

$\Rightarrow$ $\boldsymbol{\theta}_{obs}$ réalisation de $\widehat{\boldsymbol{\theta}}$ aléatoire

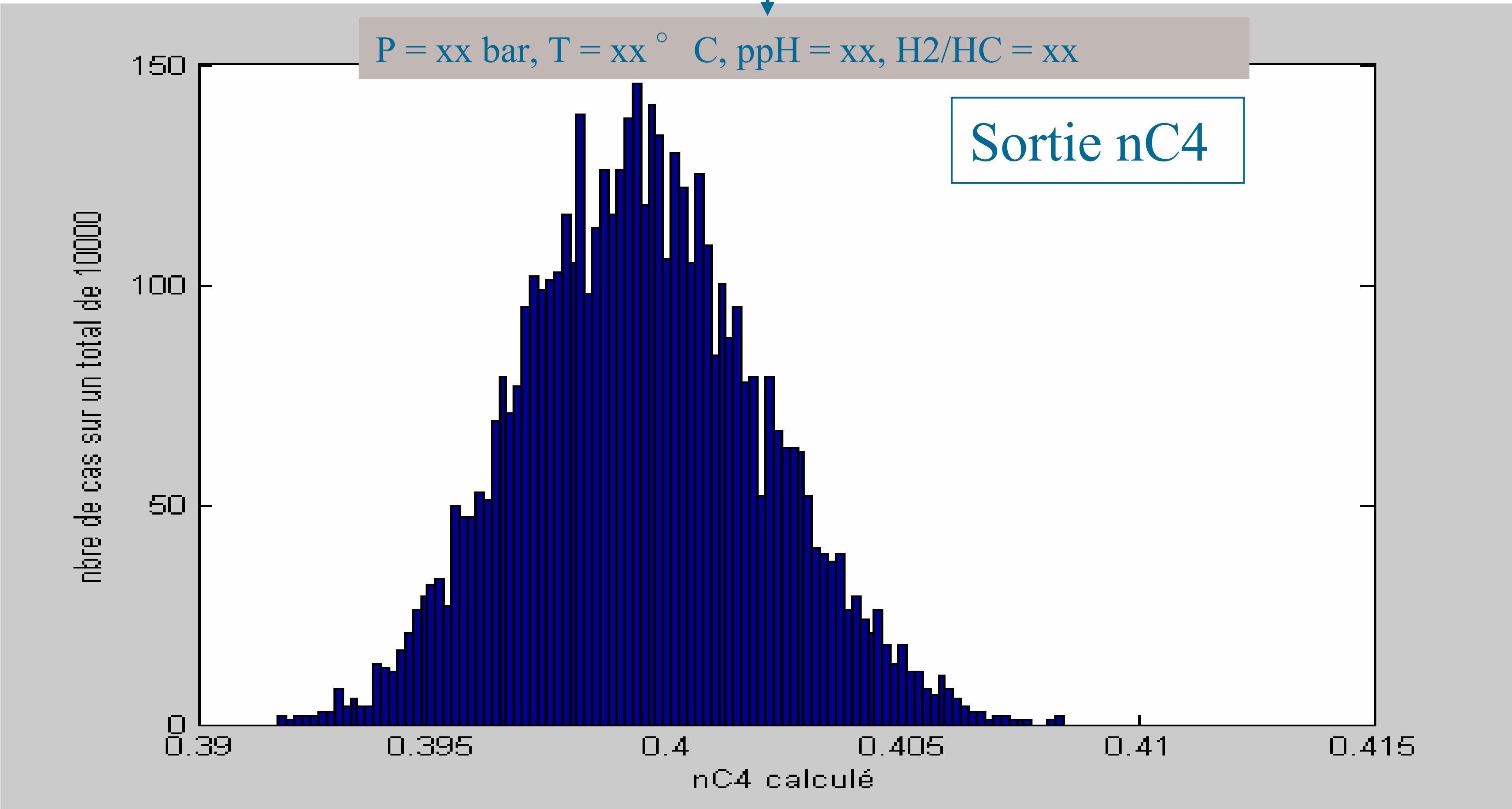
En prédiction $\boxed{\widehat{y} = f(\mathbf{c}, \hat{\boldsymbol{\theta}})}$ $\hat{\boldsymbol{\theta}}$ aléatoire $\Rightarrow$ $\widehat{y}$ aléatoire

Objectif : quantifier l'incertitude sur $\widehat{y}$

Comment : identifier la loi suivie par $\hat{\boldsymbol{\theta}}$

Propagation des Incertitudes : isomérisation des C4

Distribution de $\hat{\theta}$ → $\hat{y} = f(\mathbf{c}, \hat{\theta})$



Innovation
Energie
Environnement

Analyse de sensibilité

Modèle $\boxed{y = f(\mathbf{x})}$

Quelles sont les entrées (de $\mathbf{x}$)

dont les variations influencent le plus, le moins, la variation de y ?

Analyse de sensibilité

Modèle $\boxed{y = f(\mathbf{x})}$

Quelles sont les entrées (de $\mathbf{x}$)

dont les variations influencent le plus, le moins, la variation de y ?

Indices de sensibilté au 1er ordre : $$\boxed{S_i = \frac{\mathrm{var}(E(Y/X_i))}{\mathrm{var}(Y)}}$$

Interprétation

- $E(Y/X_i)$ la fonction de Xi qui approche le mieux Y
- $\mathrm{var}(E(Y/X_i))$ variations de la sortie comme fonction de Xi seul
- $\mathrm{var}(Y)$ normalisation

Analyse de sensibilité

Modèle $\boxed{y = f(\mathbf{x})}$

Quelles sont les entrées (de $\mathbf{x}$)

dont les variations influencent le plus, le moins, la variation de y ?

Indices de sensibilté au 1er ordre : $$\boxed{S_i = \frac{\text{var}(E(Y/X_i))}{\text{var}(Y)}}$$

<u>**Interprétation**</u>

- $E(Y/X_i)$ la fonction de Xi qui approche le mieux Y
- $\text{var}(E(Y/X_i))$ variations de la sortie comme fonction de Xi seul
- $\text{var}(Y)$ normalisation

Techniques usuelles : SOBOL, FAST

Limitations : Nombre de runs, Indépendance des entrées

Entrées non indépendantes

Méthode de MacKay : nombre de runs très important

Méthode Oakley et O'Hagan : Développements analytiques +

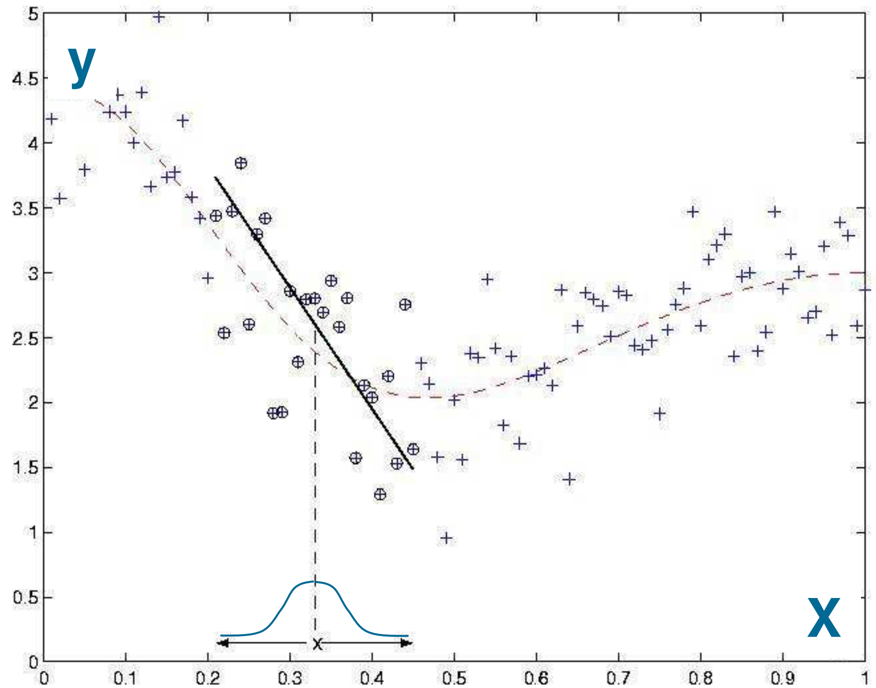
Intégrales multiples (2d-1)

Entrées non indépendantes

Méthode de MacKay : nombre de runs très important

Méthode Oakley et O'Hagan : Développements analytiques + Intégrales multiples (2d-1)

Méthode par polynômes locaux : Sébastien Da Veiga (2007)



Polynômes locaux

Modèle de régression classique

$$\boxed{Y_i = m(X_i) + \sigma(X_i)\varepsilon_i}$$

$$\boxed{m(x) = E(Y / X = x)} \quad \text{et} \quad \boxed{\sigma^2(x) = \mathrm{var}(Y / X = x)}$$

Les ε_i sont centrées réduites indépendantes des X_i

Polynômes locaux

Modèle de régression classique

$$\boxed{Y_i = m(X_i) + \sigma(X_i)\varepsilon_i}$$

$$\boxed{m(x) = E(Y / X = x)} \quad \text{et} \quad \boxed{\sigma^2(x) = \mathrm{var}(Y / X = x)}$$

Les ε_i sont centrées réduites indépendantes des X_i

polynômes d'ordre 1 (ou 2) en pratique

$$\boxed{\beta_0, \beta_1 = \underset{\beta}{\mathrm{argmin}} \sum_{i=1}^{n} \left(y_{obs,i} - (\beta_1 (x_{obs,i} - x) + \beta_0)\right)^2 K\left(\frac{x_{obs,i} - x}{h}\right)}$$

Polynômes locaux

Modèle de régression classique

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i$$

$$m(x) = E(Y / X = x)$$ et $$\sigma^2(x) = \mathrm{var}(Y / X = x)$$

Les ε_i sont centrées réduites indépendantes des X_i

polynômes d'ordre 1 (ou 2) en pratique

$$\beta_0, \beta_1 = \underset{\beta}{\mathrm{argmin}} \sum_{i=1}^{n} \left(y_{obs,i} - (\beta_1(x_{obs,i} - x) + \beta_0) \right)^2 K\left(\frac{x_{obs,i} - x}{h}\right)$$

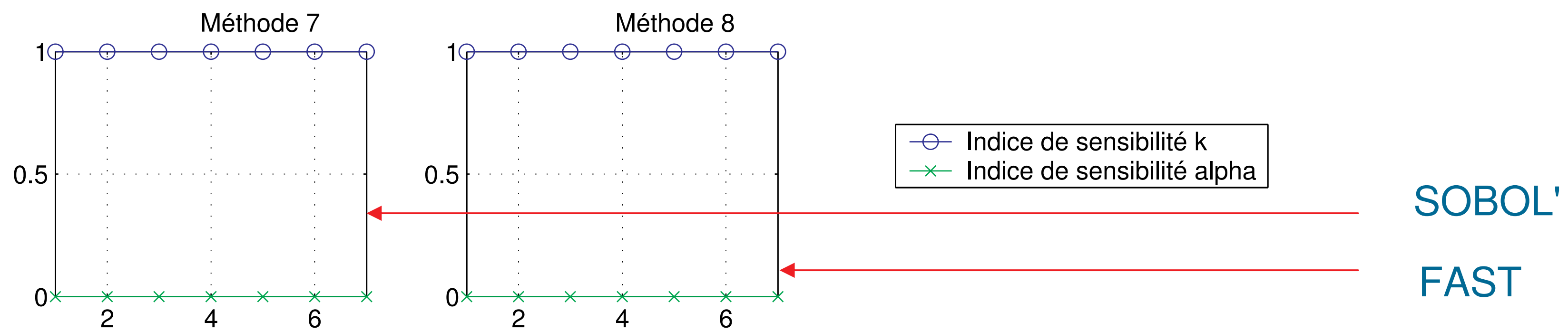
Estimation de Var(E(Y/X))=Var(m(X))

$$\widehat{T}_1 = \frac{1}{n' - 1} \sum_{j=1}^{n'} (\hat{m}(X_j) - \widehat{\overline{m}})^2$$ où $$\widehat{\overline{m}} = \frac{1}{n'} \sum_{j=1}^{n'} \hat{m}(X_j)$$

Entrées non indépendantes sur le modèle HDS

$$\frac{dS}{dt} = -kS^{\alpha}$$

$$S_{\text{Effluent}} = (S_0^{1-\alpha} - (1-\alpha)kt)^{1/(1-\alpha)}$$

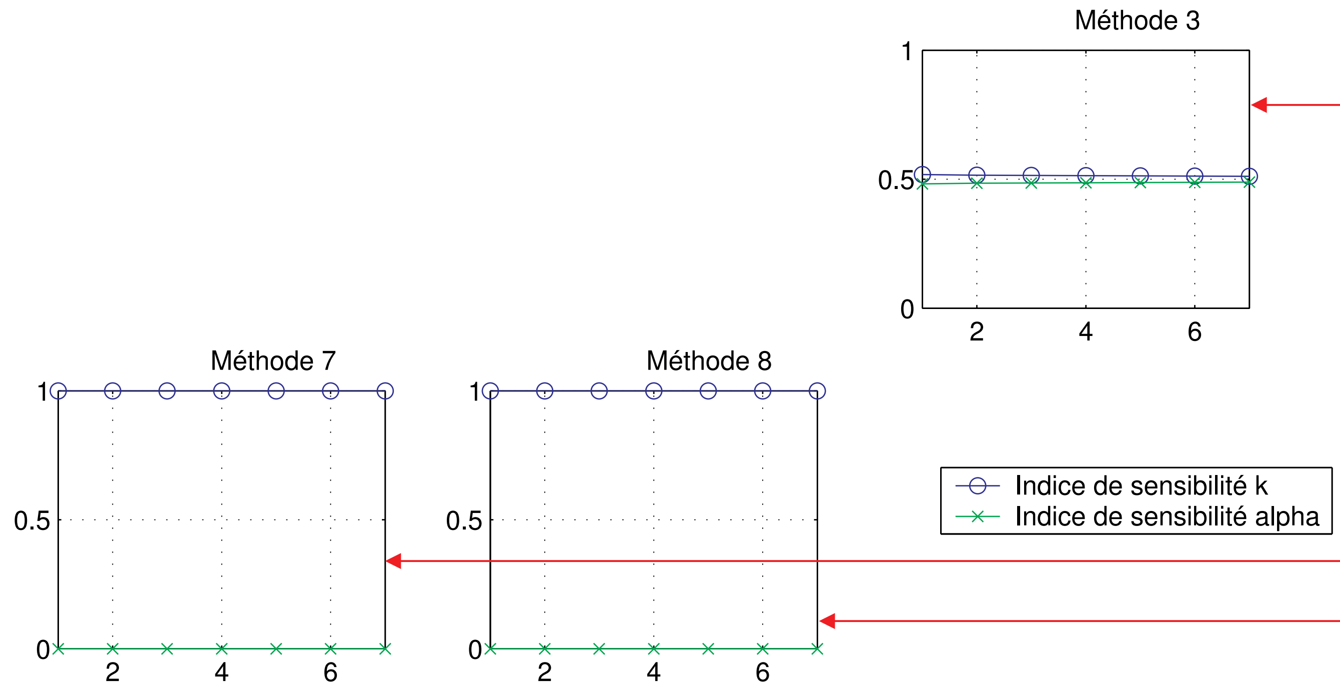


les entrées sont supposées indépendantes

Entrées non indépendantes sur le modèle HDS

$$\frac{dS}{dt} = -kS^{\alpha}$$

$$S_{\text{Effluent}} = \left(S_0^{1-\alpha} - (1-\alpha)kt\right)^{1/(1-\alpha)}$$



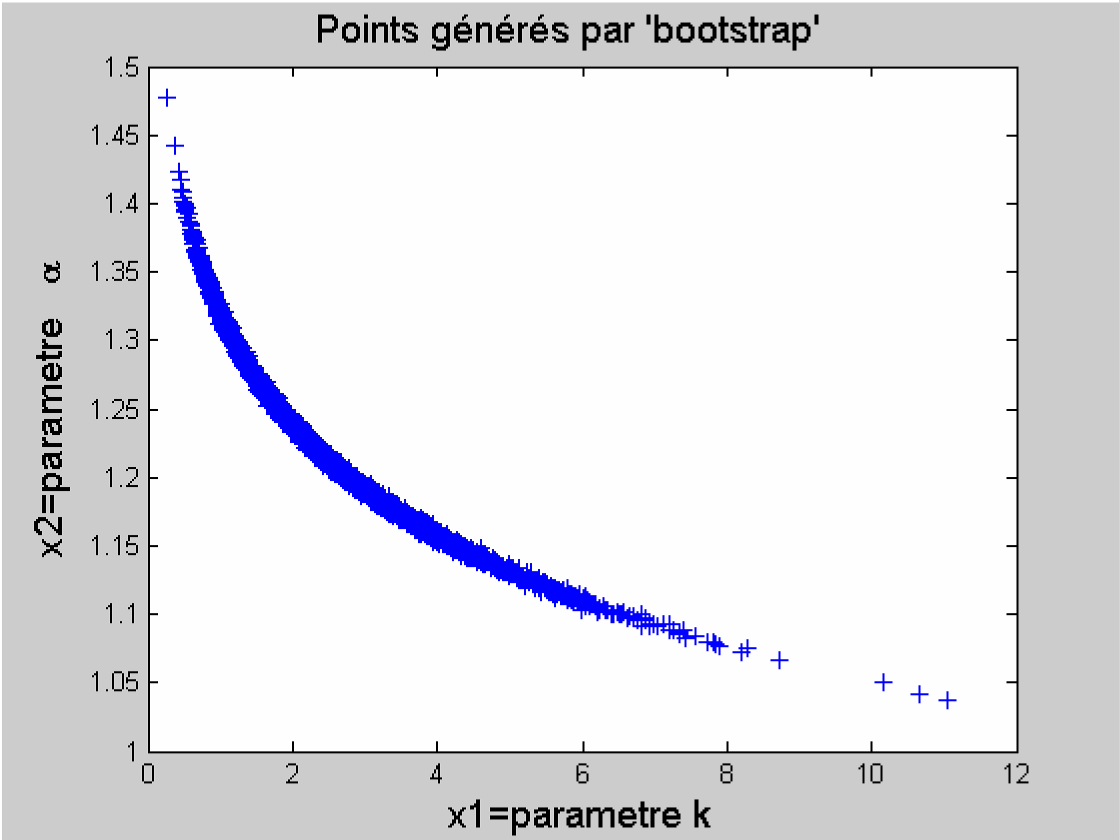
Polynomes locaux

SOBOL'

FAST

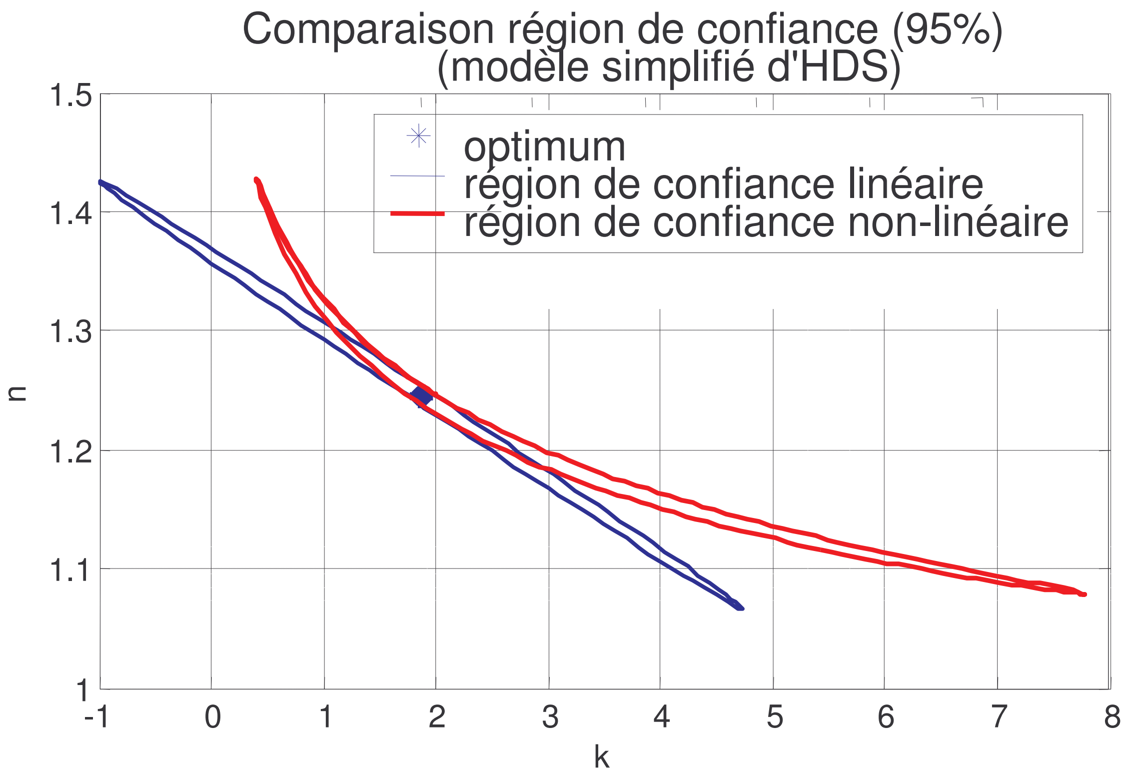
les entrées sont supposées indépendantes

Entrées non indépendantes sur le modèle HDS



Modèle simple d'HDS : développement linéaire

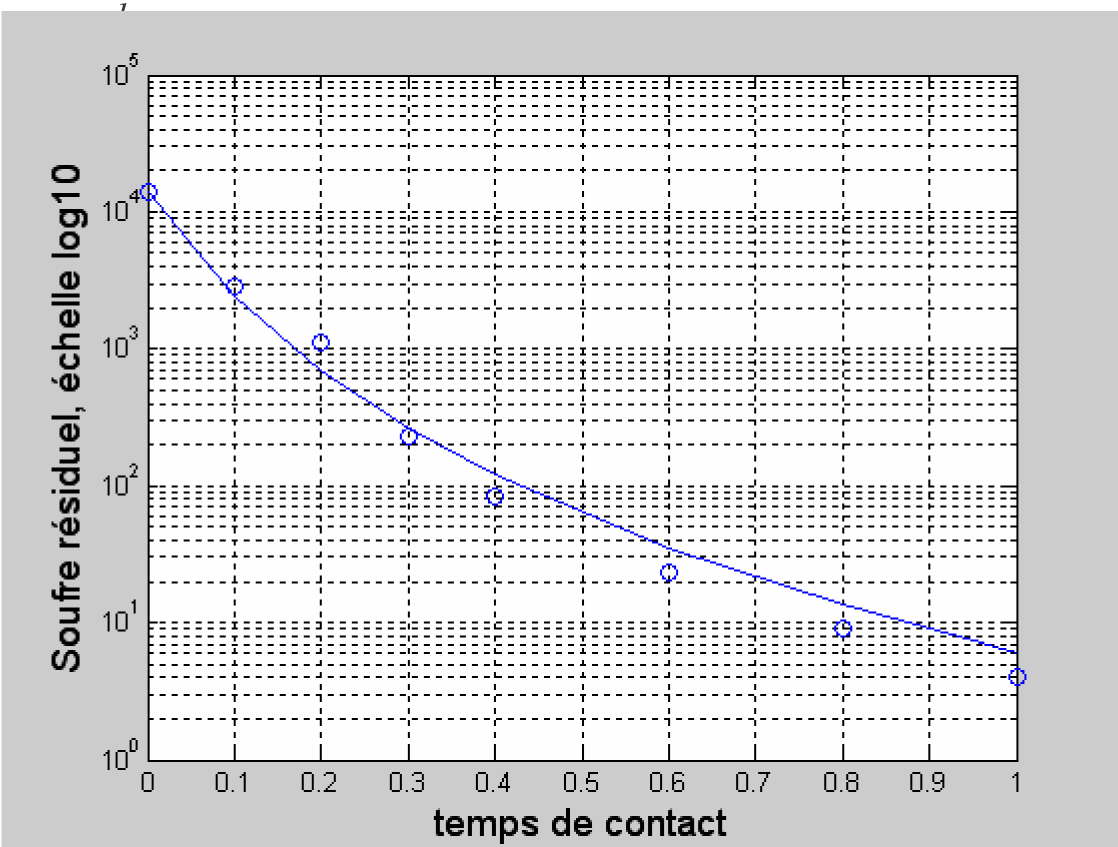
L'approximation linéaire (classique) ne suffit pas



Modèle simple d'HDS : Résultats

t contact	0	0.1	0.2	0.3	0.4	0.6	0.8	1.0
y mesuré	14000	2892	1109	230	84	23	9	4

Fonction coût $\sum_i (y_i^{mesuré} - y_i^{calculé})^2 \longrightarrow \hat{\theta}_{ref} = [k \quad \alpha] = [2.35 \quad 1.23]$



Bootstrap

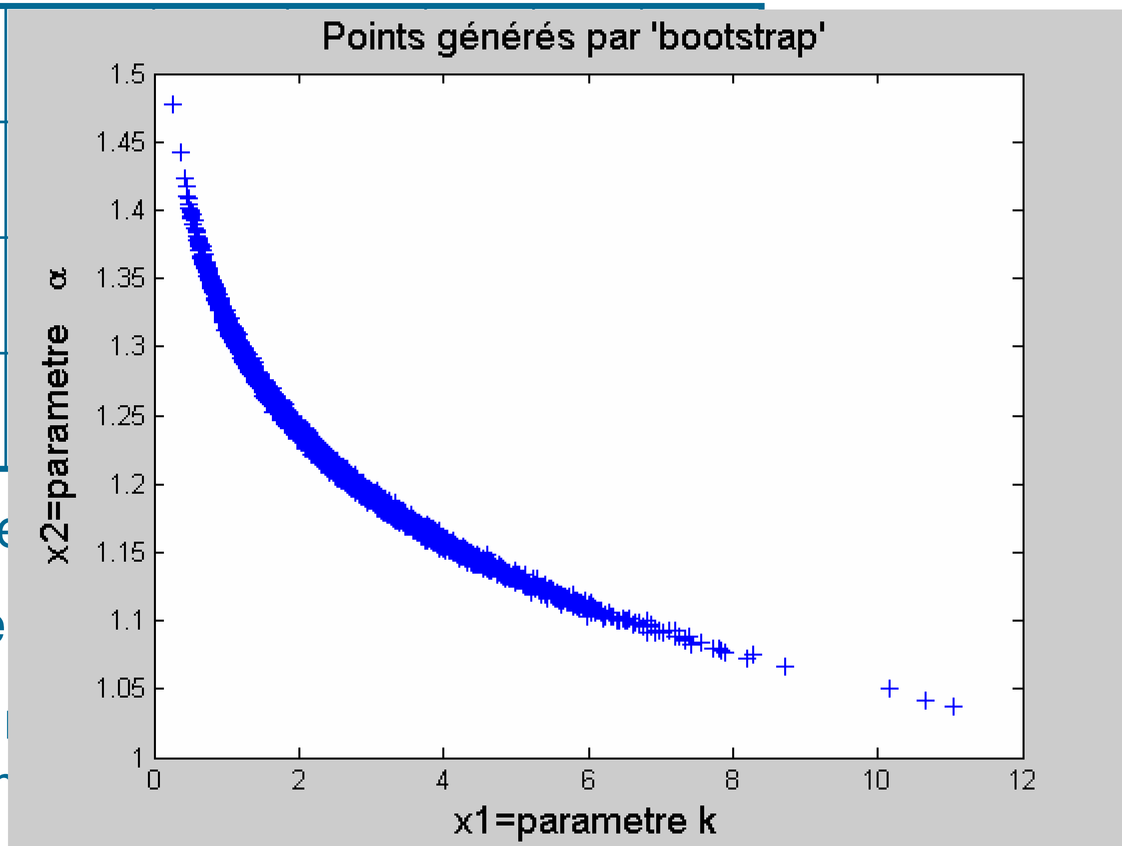
t contact	0	0.1	0.2	0.3	0.4	0.6	0.8	1.0
y mesuré	14000	2892	1109	230	84	23	9	4
y_{ref} calculé	14000	2492	695	265	121	35	13	6
résidus	0	-500	-514	35	37	12	4	2

On considère y_{ref} , calculé avec les paramètres $\widehat{\theta}_{ref}$, comme vrai

- Faire un grand nombre de fois (pour b=1,B)

1. rajouter une erreur aléatoire sur chacun des y_{ref} calculés, pour obtenir des nouvelles mesures simulées (en tirant parmi les résidus)
2. redéterminer les paramètres $\widehat{\theta}_b$ à partir des mesures simulées

t contact	0	0.1
y mesuré	14000	2892
y_{ref} calculé	14000	2492
résidus	0	-500



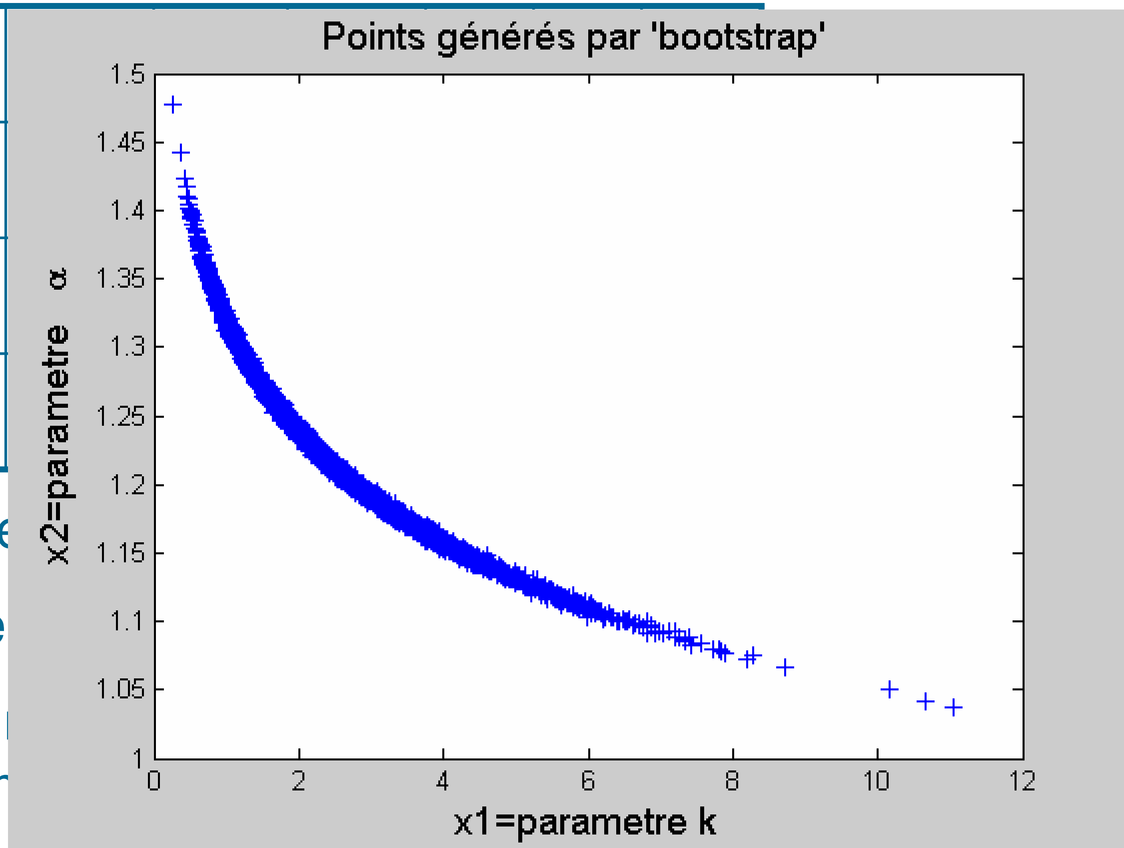
On considère y_{ref} , calculé ave

- Faire un grand nombre de

1. rajouter une erreur aléatoi
des nouvelles mesures sir
2. redéterminer les paramètres θ_b à partir des mesures simulées

Bootstrap

t contact	0	0.1
y mesuré	14000	2892
y_{ref} calculé	14000	2492
résidus	0	-500



On considère y_{ref} , calculé ave

- Faire un grand nombre de

1. rajouter une erreur aléatoi
des nouvelles mesures sir
2. redéterminer les paramètres θ_b à partir des mesures simulées

Avantage	:	Méthode générale, peu d'hypothèses
Inconvénient	:	Coût de calcul (B optimisations)

Bootstrap : coût de calcul

- 1 point 'bootstrap' = 1 jeu de paramètres = 1 optimisation
- fléau de la dimension = nombre de parametres

- Isomérisation des C4
  - 8 paramètres
  - 10000 optimisations = une nuit de calcul
- Isomérisation des C5
  - ~20 paramètres
  - 20000 optimisations = un mois de calcul

Idée

Objectif : quantifier l'incertitude sur $\widehat{y}$

Comment : identifier la loi suivie par $\hat{\boldsymbol{\theta}}$

- Une réalisation $\boldsymbol{\theta}^*{}_b$ demande la résolution d'un problème de minimisation

$$\boldsymbol{\theta}^*{}_b = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} w_i (y^*{}_{b,i} - f(\mathbf{c}_i, \boldsymbol{\theta}))^2$$

Idée

Objectif : quantifier l'incertitude sur $\widehat{y}$

Comment : identifier la loi suivie par $\hat{\boldsymbol{\theta}}$

- Une réalisation $\boldsymbol{\theta}^*{}_b$ demande la résolution d'un problème de minimisation

$$\boxed{\boldsymbol{\theta}^*{}_b = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} w_i (y^*{}_{b,i} - f(\mathbf{c}_i, \boldsymbol{\theta}))^2}$$

$\hat{\boldsymbol{\theta}}_b$ est une fonction des résidus $(\varepsilon_i)_{i=1,n}$

$$\boxed{\begin{aligned} y^*{}_{b,i} - f(\mathbf{c}_i, \boldsymbol{\theta}) &= y^*{}_{b,i} - f(\mathbf{c}_i, \boldsymbol{\theta}_{obs}) + f(\mathbf{c}_i, \boldsymbol{\theta}_{obs}) - f(\mathbf{c}_i, \boldsymbol{\theta}) \\ &= \varepsilon_i \qquad\qquad\qquad\quad + f(\mathbf{c}_i, \boldsymbol{\theta}_{obs}) - f(\mathbf{c}_i, \boldsymbol{\theta}) \end{aligned}}$$

Idée : Fonction des résidus

Objectif : quantifier l'incertitude sur $\widehat{y}$

Comment : identifier la loi suivie par $\hat{\boldsymbol{\theta}}$

- Une réalisation $\boldsymbol{\theta}^*{}_b$ demande la résolution d'un problème de minimisation
- Idée : On considère que le bootstrap échantillonne une surface de réponse, qu'on ajuste à partir de quelques points seulement

$$\boxed{\boldsymbol{\theta}^*{}_b = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} w_i (y^*{}_{b,i} - f(\mathbf{c}_i, \boldsymbol{\theta}))^2}$$

$\hat{\boldsymbol{\theta}}_b$ est une fonction des résidus $(\varepsilon_i)_{i=1,n}$

$$\boxed{\begin{aligned} y^*{}_{b,i} - f(\mathbf{c}_i, \boldsymbol{\theta}) &= y^*{}_{b,i} - f(\mathbf{c}_i, \boldsymbol{\theta}_{obs}) + f(\mathbf{c}_i, \boldsymbol{\theta}_{obs}) - f(\mathbf{c}_i, \boldsymbol{\theta}) \\ &= \varepsilon_i \qquad\qquad\qquad\quad + f(\mathbf{c}_i, \boldsymbol{\theta}_{obs}) - f(\mathbf{c}_i, \boldsymbol{\theta}) \end{aligned}}$$

Idée : Fonction des résidus

Objectif : quantifier l'incertitude sur $\widehat{y}$

Comment : identifier la loi suivie par $\hat{\boldsymbol{\theta}}$

• Une réalisation $\boldsymbol{\theta}^*_b$ demande la résolution d'un problème de minimisation

• Idée : On considère que le bootstrap échantillonne une surface de réponse, qu'on ajuste à partir de quelques points seulement

❖ soit par **SVR** ('Support Vector Machine for Regression')

❖ soit par **PCE** ('Polynomial Chaos Expansion')

$$\boxed{\boldsymbol{\theta}^*_b = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} w_i \left(y^*_{b,i} - f(\mathbf{c}_i, \boldsymbol{\theta})\right)^2}$$

$\hat{\boldsymbol{\theta}}_b$ est une fonction des résidus $(\varepsilon_i)_{i=1,n}$

$$\begin{aligned} y^*_{b,i} - f(\mathbf{c}_i, \boldsymbol{\theta}) &= y^*_{b,i} - f(\mathbf{c}_i, \boldsymbol{\theta}_{obs}) + f(\mathbf{c}_i, \boldsymbol{\theta}_{obs}) - f(\mathbf{c}_i, \boldsymbol{\theta}) \\ &= \varepsilon_i \qquad\qquad\qquad\quad + f(\mathbf{c}_i, \boldsymbol{\theta}_{obs}) - f(\mathbf{c}_i, \boldsymbol{\theta}) \end{aligned}$$

Polynômes de chaos

$\hat{\theta}_b$ est une fonction des résidus $(\varepsilon_i)_{i=1,n}$

Chaque composante du vecteur de paramètres $\hat{\theta}_b$
peut s'écrire comme une somme de polynômes orthogonaux

$$\hat{\theta}_b = a_0\Gamma_0 + \sum_i a_i\Gamma_1(\varepsilon_i) + \sum_{i,j} a_{i,j}\Gamma_{1,1}(\varepsilon_i, \varepsilon_j) + \ldots$$

Pour chacune des q composantes de $\hat{\theta}_b$

on est ramené à un système d'équations linéaires pour trouver

(n+1)(n+2)/2 coefficients $\boldsymbol{a_i}$ pour n résidus à l'ordre 2

à l'aide de **b équations = b estimations bootstrap**

Introduction des dérivées / aux résidus

$$\begin{cases} \hat{\theta} = a_0\Gamma_0 + \sum_i a_i\Gamma_1(\varepsilon_i) + \sum_{i,j} a_{i,j}\Gamma_{1,1}(\varepsilon_i,\varepsilon_j) + \ldots \\ \dfrac{\partial\hat{\theta}}{\partial\varepsilon_i} = a_0\Gamma_0 + a_i\dfrac{\partial\Gamma_1}{\partial\varepsilon_i} + \sum_j a_{i,j}\dfrac{\partial\Gamma_{1,1}}{\partial\varepsilon_i} + \ldots \end{cases}$$

$$\begin{cases} \hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}} \sum_{i=1}^{n} w_i(y_i - f(\mathbf{c}_i,\boldsymbol{\theta}))^2 \quad \text{avec} \quad C = \sum_{i=1}^{n} w(y_i - f(\mathbf{c}_i,\boldsymbol{\theta}))^2 \\ \dfrac{\partial\hat{\theta}}{\partial\varepsilon_i} = -\left(\dfrac{\partial^2 C}{\partial\varepsilon_i{}^2}\right)^{-1}\dfrac{\partial^2 C}{\partial\theta.\partial\varepsilon_i} \end{cases}$$

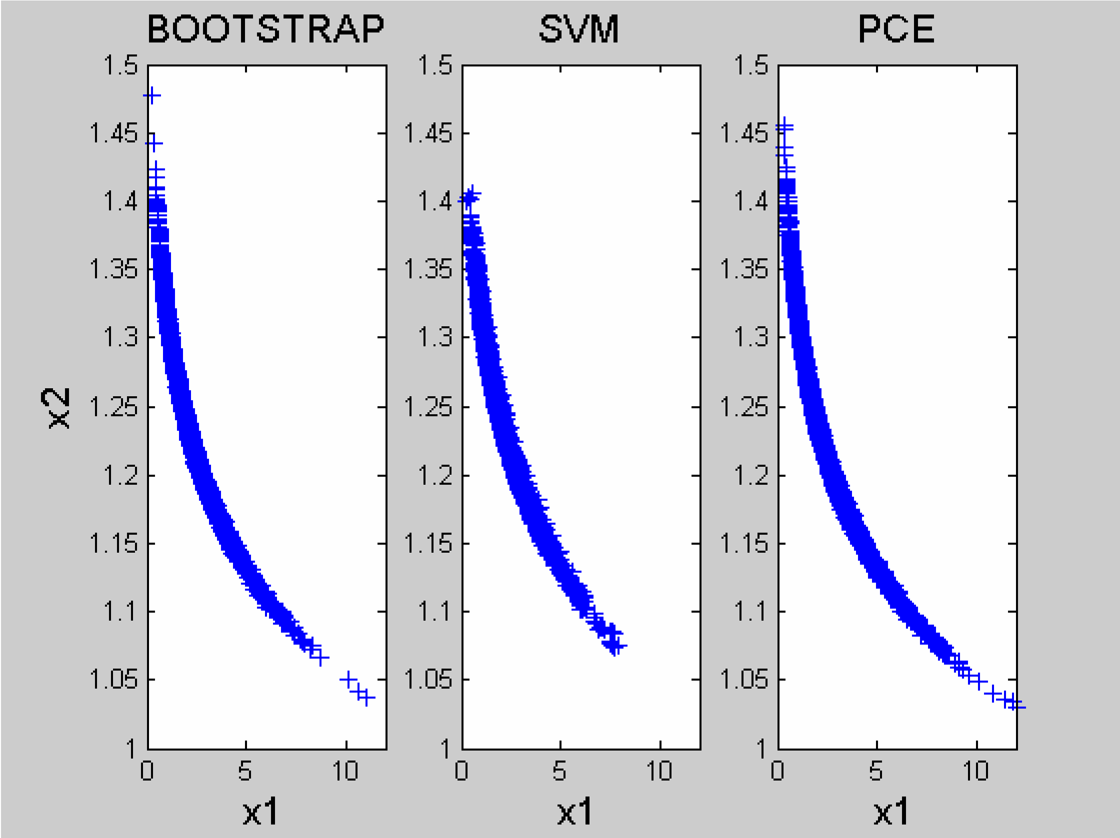
Pour chacune des q composantes de $\hat{\boldsymbol{\theta}}_b$

on est ramené à un système d'équations linéaires pour trouver

(n+1)(n+2)/2 coefficients a_i pour n résidus à l'ordre 2

à l'aide de **b(q+1) équations**

Désulfuration : Comparaison des résultats

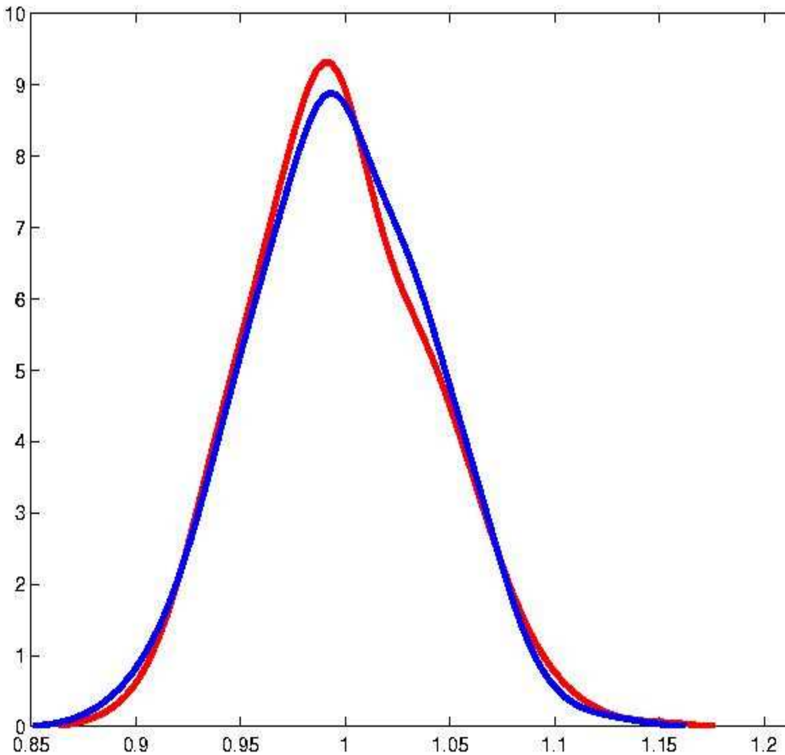


Bootstrap	: **5000** points
SVM	: **500** points, 5000 points générés
PCE (avec dérivées)	: **50** points, 5000 points générés

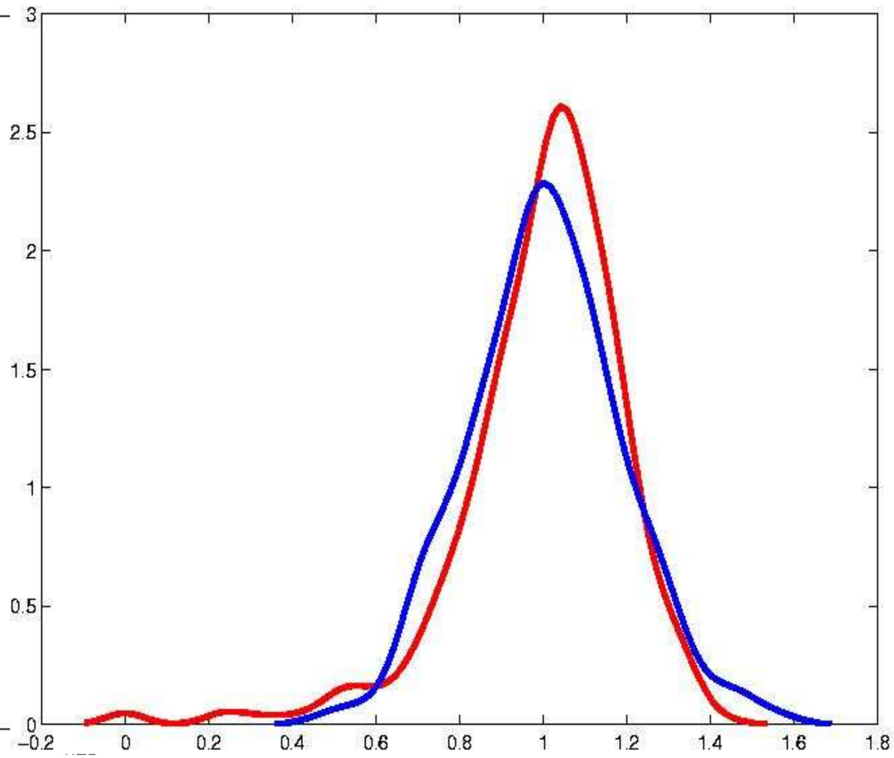
Distribution des paramètres 1 et 7

Rouge = bootstrap 10000, Bleu = SVM 200

= SVM 50 + Deriv (Vasquez & al 2003), (Lazaro & al 2005)



Paramètre 1



Paramètre 7

Conclusions

$$S_i = \frac{\mathrm{var}(E(Y / X_i))}{\mathrm{var}(Y)}$$

- **Indices de sensibilité et Polynômes locaux**
  - **faciles à appréhender**
  - **résultats théoriques de convergence**

$$\boldsymbol{\theta}^*{}_b = \underset{\boldsymbol{\theta}\in\Theta}{\mathrm{argmin}} \sum_{i=1}^{n} w_i (y^*{}_{b,i} - f(\mathbf{c}_i, \boldsymbol{\theta}))^2$$

- **Bootstrap : Pb de temps de calcul**
- **Solutions : Surfaces de réponse**
  - **SVR**
  - **Polynômes de chaos**
- **Surfaces de réponse et dérivées**
  - **Limitations : Calcul des dérivées**
  - **Pb en grande dimension**