QUELQUES ASPECTS DE LA PLANIFICATION D'EXPÉRIENCES POUR MODÈLES NON PARAMÉTRIQUES

Luc Pronzato

Laboratoire I3S,
CNRS/Univ. Nice Sophia Antipolis, France

Plan

- I) Planification pour l'estimation : exemple 1

Plan

- I) Planification pour l'estimation : exemple 1
- II) Estimation : exemple 2

Plan

- I) Planification pour l'estimation : exemple 1
- II) Estimation : exemple 2
- III) Discrimination : exemple 3

Plan

- I) Planification pour l'estimation : exemple 1
- II) Estimation : exemple 2
- III) Discrimination : exemple 3
- IV) Optimisation et planification d'expériences

Plan

- I) Planification pour l'estimation : exemple 1
- II) Estimation : exemple 2
- III) Discrimination : exemple 3
- IV) Optimisation et planification d'expériences
- V) Modèles non paramétriques

Plan

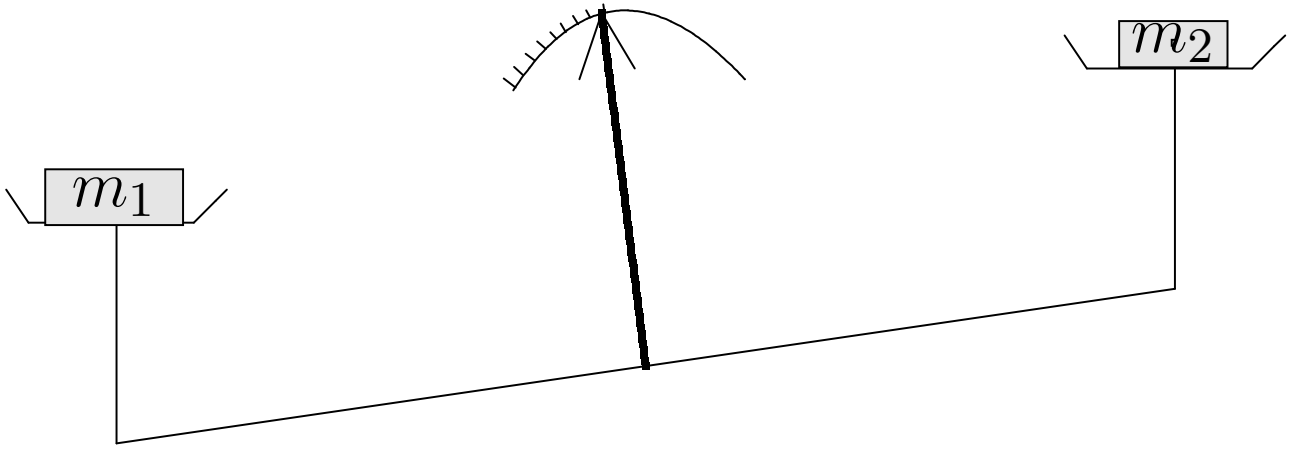
- I) Planification pour l'estimation : exemple 1
- II) Estimation : exemple 2
- III) Discrimination : exemple 3
- IV) Optimisation et planification d'expériences
- V) Modèles non paramétriques
- VI) Planification en non paramétrique

Plan

- I) Planification pour l'estimation : exemple 1
- II) Estimation : exemple 2
- III) Discrimination : exemple 3
- IV) Optimisation et planification d'expériences
- V) Modèles non paramétriques
- VI) Planification en non paramétrique
- VII) Retour sur l'optimisation

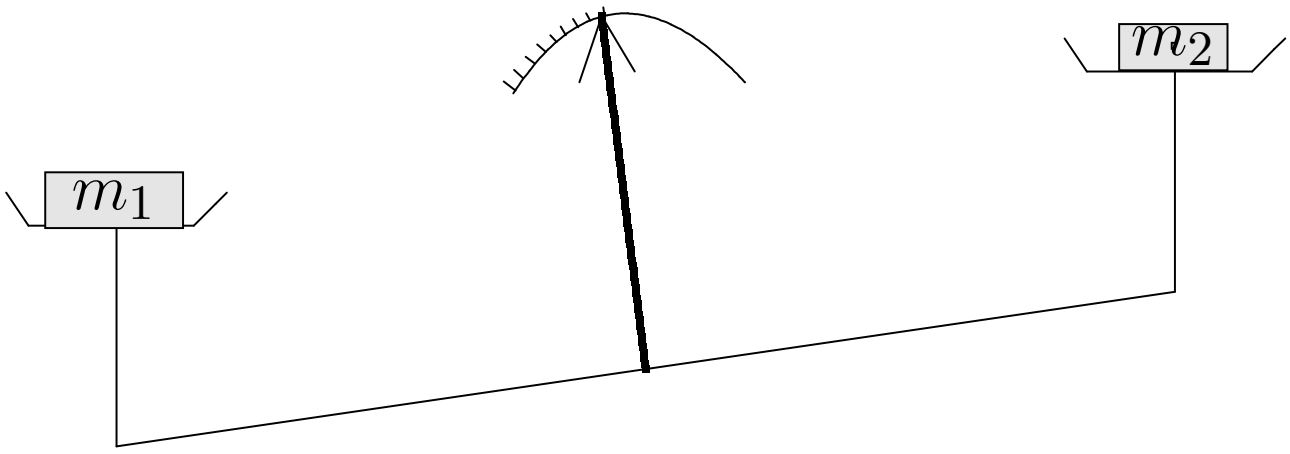
I) Exemple 1 : pesée de 8 objets

$$y = m_1 - m_2 + \epsilon$$



I) Exemple 1 : pesée de 8 objets

$$y = m_1 - m_2 + \epsilon$$



Les objets ont des masses m_i, $i = 1, \ldots, 8$
les erreurs ε_i sont i.i.d. $\mathcal{N}(0, \sigma^2)$

Méthode a: on pèse les 8 objets successivement

$\longrightarrow y(i) = m_i + \varepsilon_i$, $i = 1, \ldots, 8$

$\longrightarrow$ masses estimées : $\hat{m}_i = y(i) \sim \mathcal{N}(m_i, \sigma^2)$

On répète 8 fois, en moyennant les résultats :

$\hat{\hat{m}}_i = y(i) \sim \mathcal{N}(m_i, \sigma^2/8)$ (avec 64 observations. . .)

Méthode a: on pèse les 8 objets successivement

$\longrightarrow y(i) = m_i + \varepsilon_i$, $i = 1, \ldots, 8$

$\longrightarrow$ masses estimées : $\hat{m}_i = y(i) \sim \mathcal{N}(m_i, \sigma^2)$

On répète 8 fois, en moyennant les résultats :

$\hat{\hat{m}}_i = y(i) \sim \mathcal{N}(m_i, \sigma^2/8)$ (avec 64 observations...)

Méthode b: plus sophistiquée...

$$
\begin{aligned}
y(1) &= m_1 + m_2 + m_3 + m_4 + m_5 + m_6 + m_7 + m_8 + \varepsilon_1 \\
y(2) &= m_1 + m_2 + m_3 - m_4 - m_5 - m_6 - m_7 + m_8 + \varepsilon_2 \\
y(3) &= m_1 - m_2 - m_3 + m_4 + m_5 - m_6 - m_7 + m_8 + \varepsilon_3 \\
y(4) &= m_1 - m_2 - m_3 - m_4 - m_5 + m_6 + m_7 + m_8 + \varepsilon_4 \\
y(5) &= -m_1 + m_2 - m_3 + m_4 - m_5 + m_6 - m_7 + m_8 + \varepsilon_5 \\
y(6) &= -m_1 + m_2 - m_3 - m_4 + m_5 - m_6 + m_7 + m_8 + \varepsilon_6 \\
y(7) &= -m_1 - m_2 + m_3 + m_4 - m_5 - m_6 + m_7 + m_8 + \varepsilon_7 \\
y(8) &= -m_1 - m_2 + m_3 - m_4 + m_5 + m_6 - m_7 + m_8 + \varepsilon_8
\end{aligned}
$$

$$
\begin{aligned}
y(1) &= m_1 + m_2 + m_3 + m_4 + m_5 + m_6 + m_7 + m_8 + \varepsilon_1 \\
y(2) &= m_1 + m_2 + m_3 - m_4 - m_5 - m_6 - m_7 + m_8 + \varepsilon_2 \\
y(3) &= m_1 - m_2 - m_3 + m_4 + m_5 - m_6 - m_7 + m_8 + \varepsilon_3 \\
y(4) &= m_1 - m_2 - m_3 - m_4 - m_5 + m_6 + m_7 + m_8 + \varepsilon_4 \\
y(5) &= -m_1 + m_2 - m_3 + m_4 - m_5 + m_6 - m_7 + m_8 + \varepsilon_5 \\
y(6) &= -m_1 + m_2 - m_3 - m_4 + m_5 - m_6 + m_7 + m_8 + \varepsilon_6 \\
y(7) &= -m_1 - m_2 + m_3 + m_4 - m_5 - m_6 + m_7 + m_8 + \varepsilon_7 \\
y(8) &= -m_1 - m_2 + m_3 - m_4 + m_5 + m_6 - m_7 + m_8 + \varepsilon_8
\end{aligned}
$$

$$
\begin{aligned}
\rightarrow \hat{m}_1 &= \frac{y(1) + y(2) + y(3) + y(4) - y(5) - y(6) - y(7) - y(8)}{8} \\
&= m_1 + \frac{\varepsilon_1 + \varepsilon_2 + \varepsilon_3 + \varepsilon_4 - \varepsilon_5 - \varepsilon_6 - \varepsilon_7 - \varepsilon_8}{8}
\end{aligned}
$$

etc.

et $\hat{m}_i \sim \mathcal{N}(m_i, \sigma^2/8)$ avec 8 observations seulement !
(contre 64 avec la méthode a)

et $\hat{m}_i \sim \mathcal{N}(m_i, \sigma^2/8)$ avec 8 observations seulement !
(contre 64 avec la méthode a)
Ici, sélection d'un bon plan = problème de combinatoire :
On a $y(k) = \sum_{i=1}^{8} \mathbf{u}_{ki} m_i + \varepsilon_k = \mathbf{u}_k^\top \mathbf{m} + \varepsilon_k$
par exemple pour $y(2)$ dans la méthode b :
$\mathbf{u}_2 = [1\ 1\ 1\ -1\ -1\ -1\ -1\ 1]^\top$

et $\hat{m}_i \sim \mathcal{N}(m_i, \sigma^2/8)$ avec 8 observations seulement !
(contre 64 avec la méthode a)
Ici, sélection d'un bon plan = problème de combinatoire :
On a $y(k) = \sum_{i=1}^{8} \mathbf{u}_{ki} m_i + \varepsilon_k = \mathbf{u}_k^\top \mathbf{m} + \varepsilon_k$
par exemple pour $y(2)$ dans la méthode b :
$\mathbf{u}_2 = [1\ 1\ 1\ -1\ -1\ -1\ -1\ 1]^\top$

$$\text{estimateur des MC : } \hat{\mathbf{m}} = \arg\min_{\mathbf{m}} \sum_{k=1}^{N} [y(k) - \mathbf{u}_k^\top \mathbf{m}]^2$$

$$= \left(\sum_{k=1}^{N} \mathbf{u}_k \mathbf{u}_k^\top \right)^{-1} \sum_{k=1}^{N} y(k) \mathbf{u}_k$$

$\rightarrow$ Choisir des $\mathbf{u}_k$ tels que $\mathbf{M}_N = \sum_{k=1}^{N} \mathbf{u}_k \mathbf{u}_k^\top$ est non singulière

$\mathbb{E}\{\hat{\mathbf{m}}\} = \mathbf{m}$ (sans biais)

$\mathbb{E}\{\hat{\mathbf{m}}\} = \mathbf{m}$ (sans biais)

$\mathbb{E}\{(\hat{\mathbf{m}} - \mathbf{m})(\hat{\mathbf{m}} - \mathbf{m})^\top\} = \sigma^2 \mathbf{M}_N^{-1}$

$\rightarrow$ minimiser une fonction scalaire de $\mathbf{M}_N^{-1}$

$\mathbb{E}\{\hat{\mathbf{m}}\} = \mathbf{m}$ (sans biais)

$\mathbb{E}\{(\hat{\mathbf{m}} - \mathbf{m})(\hat{\mathbf{m}} - \mathbf{m})^\top\} = \sigma^2 \mathbf{M}_N^{-1}$

$\rightarrow$ minimiser une fonction scalaire de $\mathbf{M}_N^{-1}$

Problème de combinatoire puique $\mathbf{u}_{ki} \in \{-1, 0, 1\}$

[Fisher, 1925 . . .]

$\mathbb{E}\{\hat{\mathbf{m}}\} = \mathbf{m}$ (sans biais)

$\mathbb{E}\{(\hat{\mathbf{m}} - \mathbf{m})(\hat{\mathbf{m}} - \mathbf{m})^\top\} = \sigma^2 \mathbf{M}_N^{-1}$

$\rightarrow$ minimiser une fonction scalaire de $\mathbf{M}_N^{-1}$

Problème de combinatoire puique $\mathbf{u}_{ki} \in \{-1, 0, 1\}$
[Fisher, 1925 . . .]

Plus généralement, quand les variables (*facteurs, entrées*) u_k sont des réels, le plan optimal pour l'estimation est obtenu par l'optimisation d'une fonction scalaire de la matrice de covariance (asymptotique) de l'estimateur

Exemple 2 : phamacocinétique

[D'Argenio, 1981], modèle à 2 compartiments
Un produit x est injecté par perfusion ($\rightarrow$ entrée $u(t)$), $x_C(t)$ (produit dans le sang) passe dans un autre tissu $\rightarrow$ $x_P(t)$
Eq. différentielles linéaires :

$$\begin{cases} \frac{dx_C(t)}{dt} &= (-K_{EL} - K_{CP})x_C(t) + K_{PC}x_P(t) + u(t) \\ \frac{dx_P(t)}{dt} &= K_{CP}x_C(t) - K_{PC}x_P(t) \end{cases}$$

On observe la concentration de x dans le sang :
$y(t) = x_C(t)/V + \varepsilon(t)$,
les $(\epsilon(t_i))$ sont i.i.d. $\mathcal{N}(0, \sigma^2)$, $\sigma = 0.2\mu$g/ml

Exemple 2 : phamacocinétique

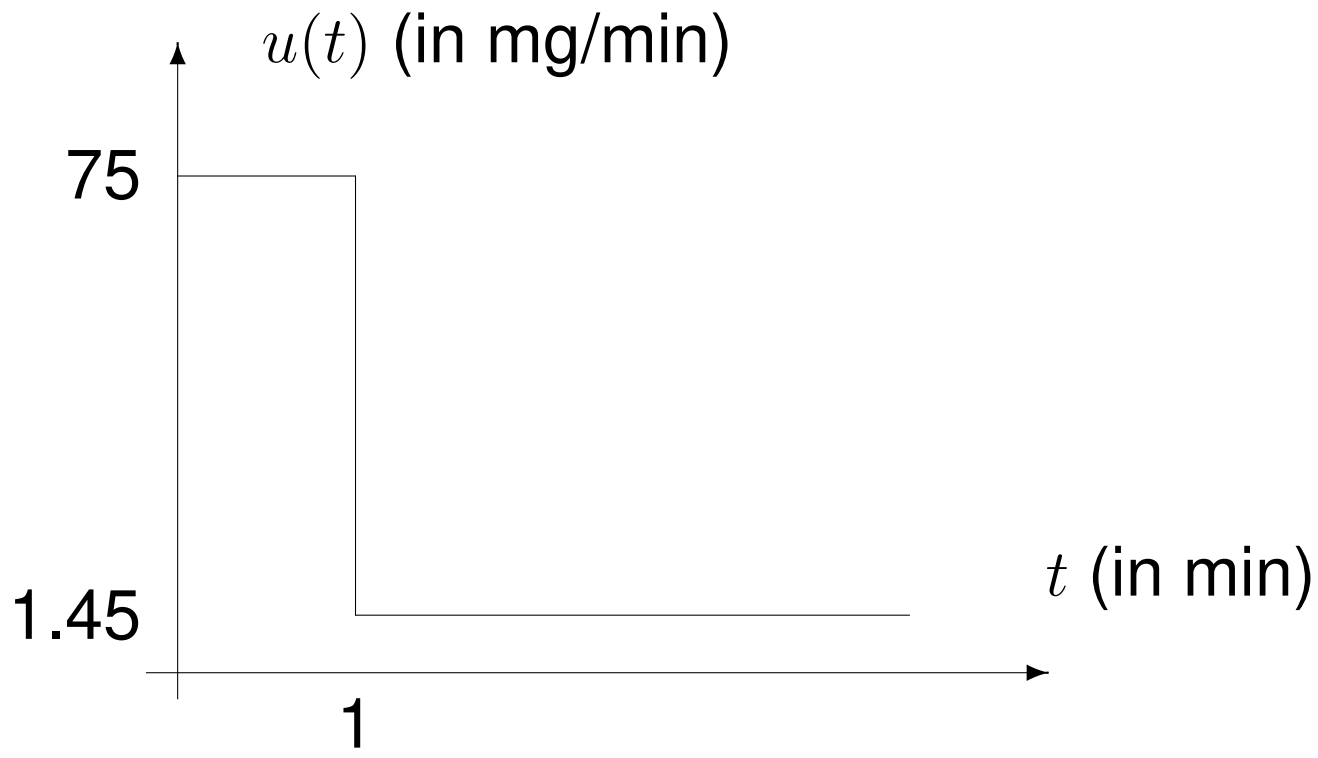
[D'Argenio, 1981], modèle à 2 compartiments
Un produit x est injecté par perfusion ($\rightarrow$ entrée $u(t)$), $x_C(t)$ (produit dans le sang) passe dans un autre tissu $\rightarrow$ $x_P(t)$
Eq. différentielles linéaires :

$$\begin{cases} \frac{dx_C(t)}{dt} &= (-K_{EL} - K_{CP})x_C(t) + K_{PC}x_P(t) + u(t) \\ \frac{dx_P(t)}{dt} &= K_{CP}x_C(t) - K_{PC}x_P(t) \end{cases}$$

On observe la concentration de x dans le sang :
$y(t) = x_C(t)/V + \varepsilon(t)$,
les $(\epsilon(t_i))$ sont i.i.d. $\mathcal{N}(0, \sigma^2)$, $\sigma = 0.2\mu$g/ml

4 paramètres inconnus : $\theta = (K_{CP}, K_{PC}, K_{EL}, V)$

Le profil de l'entrée $u(t)$ est imposé :



$\longrightarrow$ **expérience simulée** avec des valeurs de paramètres

$$\bar{\theta} = (0.066\,\text{min}^{-1},\ 0.038\,\text{min}^{-1},\ 0.0242\,\text{min}^{-1},\ 30\,\text{l})$$

Variables expér. = instants de mesure t_i, $1 \leq t_i \leq 720$ min

Variables expér. = instants de mesure t_i, $1 \leq t_i \leq 720$ min

- Protocole «conventionnel» :

$$\mathbf{t} = (5, 10, 30, 60, 120, 180, 360, 720) \text{ (en min)}$$

Variables expér. = instants de mesure t_i, $1 \leq t_i \leq 720$ min

- Protocole «conventionnel» :

$$\mathbf{t} = (5, 10, 30, 60, 120, 180, 360, 720) \text{ (en min)}$$

- Protocole «optimal» (pour $\bar{\theta}$) :

$$\mathbf{t} = (1, 1, 10, 10, 74, 74, 720, 720) \text{ (en min)}$$

(suppose possible d'obtenir des observations indépendantes à un même instant)

Variables expér. = instants de mesure t_i, $1 \leq t_i \leq 720$ min

- Protocole «conventionnel» :

$$\mathbf{t} = (5, 10, 30, 60, 120, 180, 360, 720) \text{ (en min)}$$

- Protocole «optimal» (pour $\bar{\theta}$) :

$$\mathbf{t} = (1, 1, 10, 10, 74, 74, 720, 720) \text{ (en min)}$$

(suppose possible d'obtenir des observations indépendantes à un même instant)

$\longrightarrow$ 400 simulations

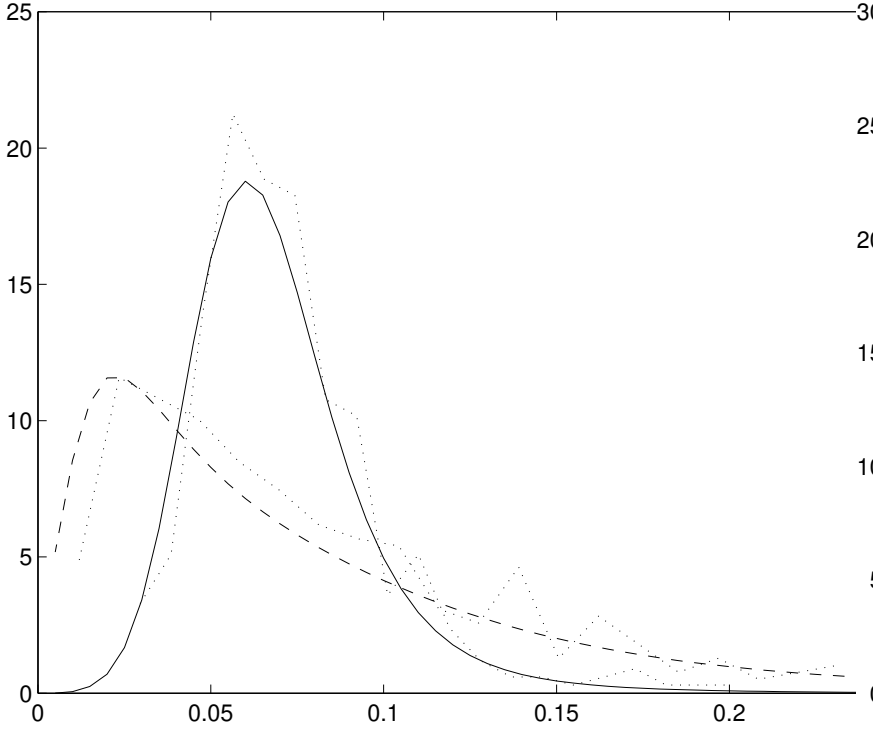
$\longrightarrow$ 400 fois 8 observations, pour chaque protocole

$\longrightarrow$ 400 paramètres estimés (MC) pour chaque protocole. . .

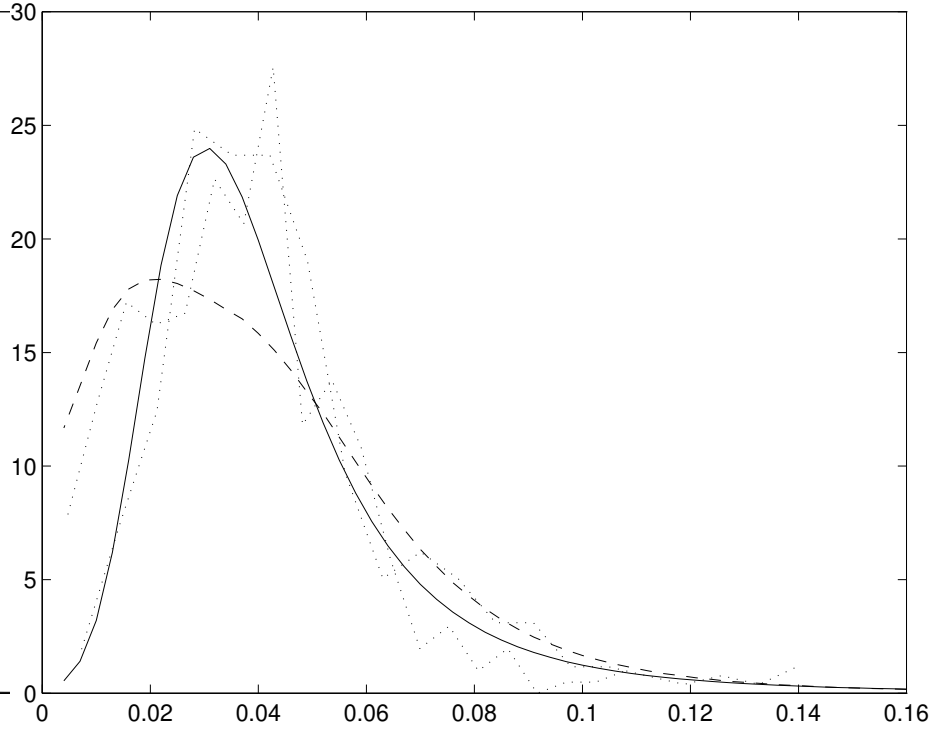
histogrammes des $\hat{\theta}_i$

(approximation des marginales [Pázman & Pronzato 1996])

Marginales et histogrammes pour les 2 protocoles (optimal —, conventionnel - - -)

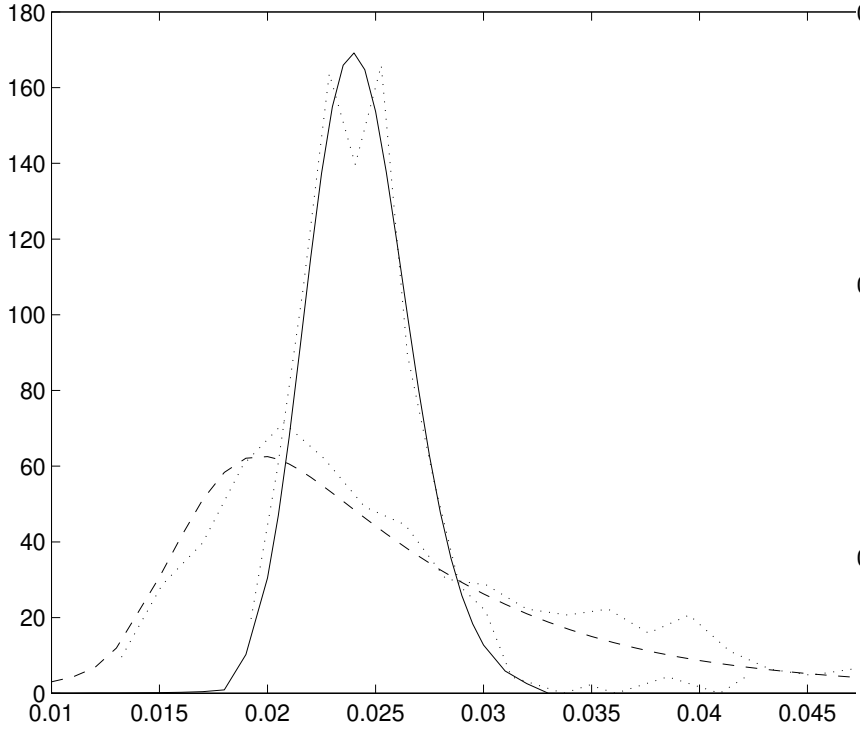


$\hat{K}_{CP}$ $[\bar{K}_{CP} = 0.066]$

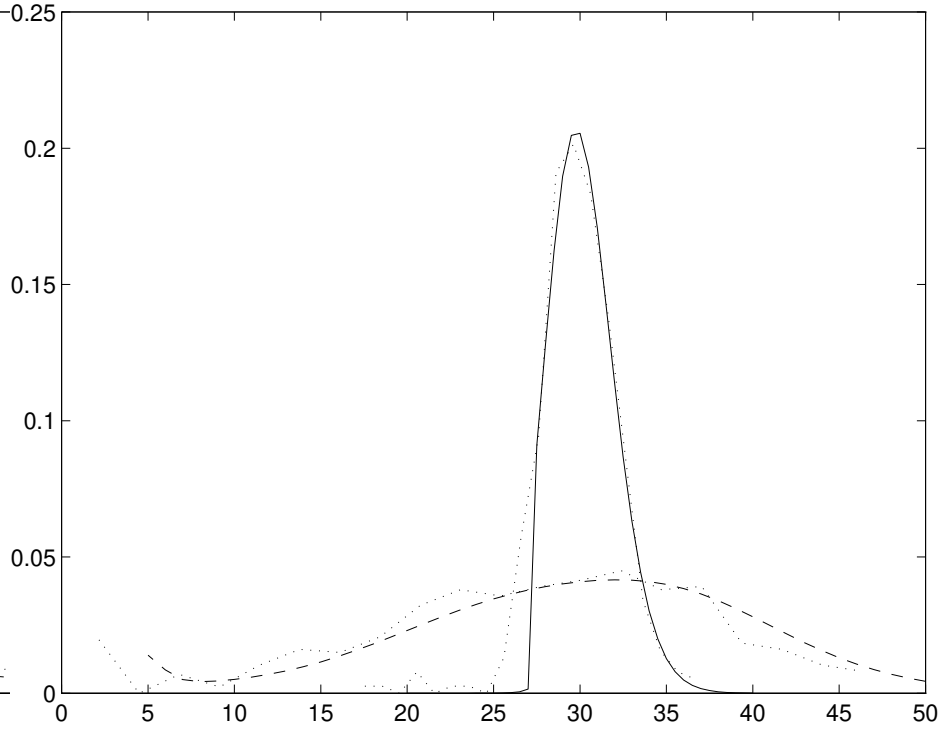


$\hat{K}_{PC}$ $[\bar{K}_{PC} = 0.038]$

Marginales et histogrammes pour les 2 protocoles (optimal —, conventionnel - - -)

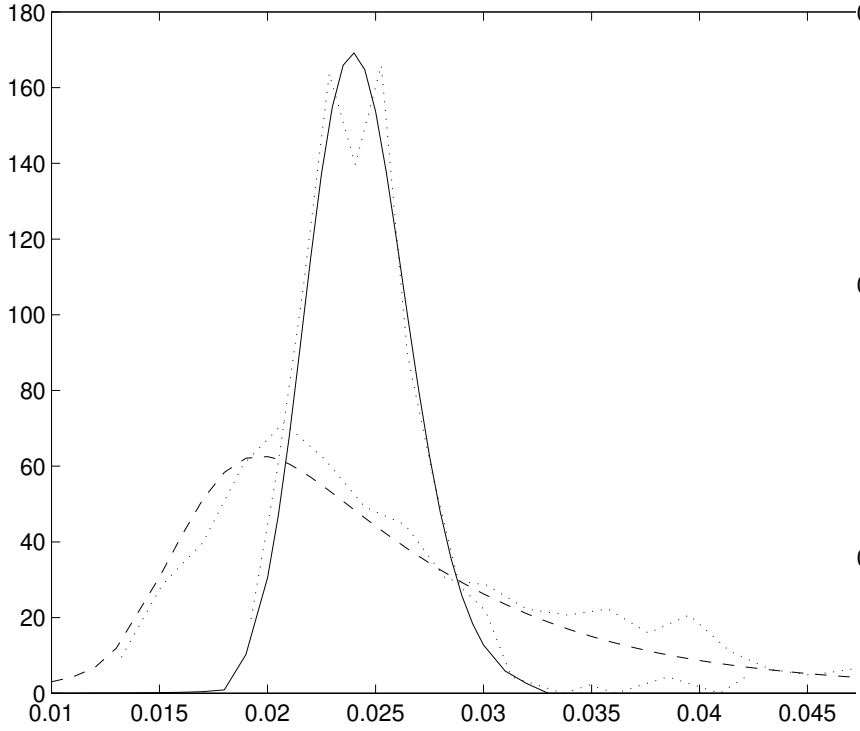


$\hat{K}_{EL}$ $[\bar{K}_{EL} = 0.0242]$

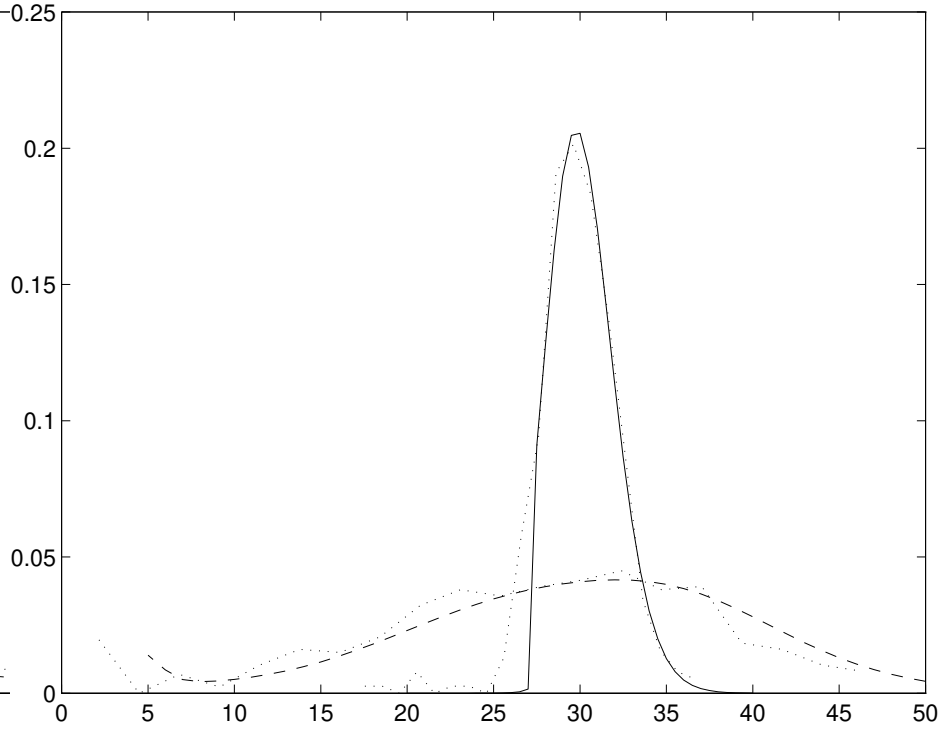


$\hat{V}$ $[\bar{V} = 30]$

Marginales et histogrammes pour les 2 protocoles (optimal —, conventionnel - - -)



$\hat{K}_{EL}$ $[\bar{K}_{EL} = 0.0242]$



$\hat{V}$ $[\bar{V} = 30]$

protocole «optimal» $\longrightarrow$ estimation plus précise

Exemple 3 : discrimination

[Box & Hill, 1967]: discrimination entre structures de modèles

Exemple 3 : discrimination

[Box & Hill, 1967]: discrimination entre structures de modèles

Réaction chimique $A \rightarrow B$

2 facteurs : $\mathbf{u} = (\text{temps } t, \text{ température } T)$

réaction du 1er, 2ème, 3ème ou 4ème ordre ?

Exemple 3 : discrimination

[Box & Hill, 1967]: discrimination entre structures de modèles

Réaction chimique $A \rightarrow B$

2 facteurs : $\mathbf{u} = (\text{temps } t, \text{ température } T)$

réaction du 1er, 2ème, 3ème ou 4ème ordre ?

$\rightarrow$ 4 structures de modèles sont candidates :

$$\begin{aligned}
\eta^{(1)}(\theta_1, \mathbf{u}) &= \exp[-\theta_{11} t \exp(-\theta_{12}/T)]\,, \\
\eta^{(2)}(\theta_2, \mathbf{u}) &= \frac{1}{1 + \theta_{21} t \exp(-\theta_{22}/T)}\,, \\
\eta^{(3)}(\theta_3, \mathbf{u}) &= \frac{1}{[1 + 2\theta_{31} t \exp(-\theta_{32}/T)]^{1/2}}\,, \\
\eta^{(4)}(\theta_4, \mathbf{u}) &= \frac{1}{[1 + 3\theta_{41} t \exp(-\theta_{42}/T)]^{1/3}}\,.
\end{aligned}$$

Expérience simulée

Observations: 2ème structure, $y(\mathbf{u}_j) = \eta^{(2)}(\bar{\theta}_2, \mathbf{u}_j) + \varepsilon_j$,
avec $\bar{\theta}_2 = (400,\, 5000)^\top$ la vraie valeur (inconnue) des paramètres du modèle 2, (ϵ_j) i.i.d. $\mathcal{N}(0, \sigma^2)$, $\sigma = 0.05$

Domaine expérimental admissible : $0 \leq t \leq 150$,
$450 \leq T \leq 600$

Expérience simulée

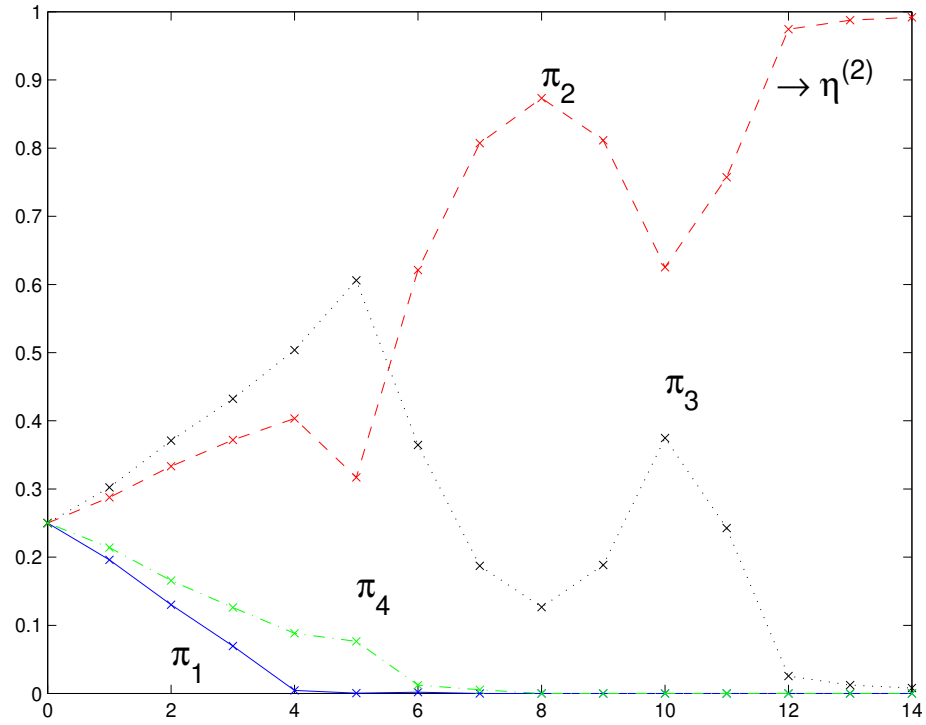
Observations: 2ème structure, $y(\mathbf{u}_j) = \eta^{(2)}(\bar{\theta}_2, \mathbf{u}_j) + \varepsilon_j$,
avec $\bar{\theta}_2 = (400,\, 5000)^\top$ la vraie valeur (inconnue) des paramètres du modèle 2, (ϵ_j) i.i.d. $\mathcal{N}(0, \sigma^2)$, $\sigma = 0.05$

Domaine expérimental admissible : $0 \leq t \leq 150$,
$450 \leq T \leq 600$

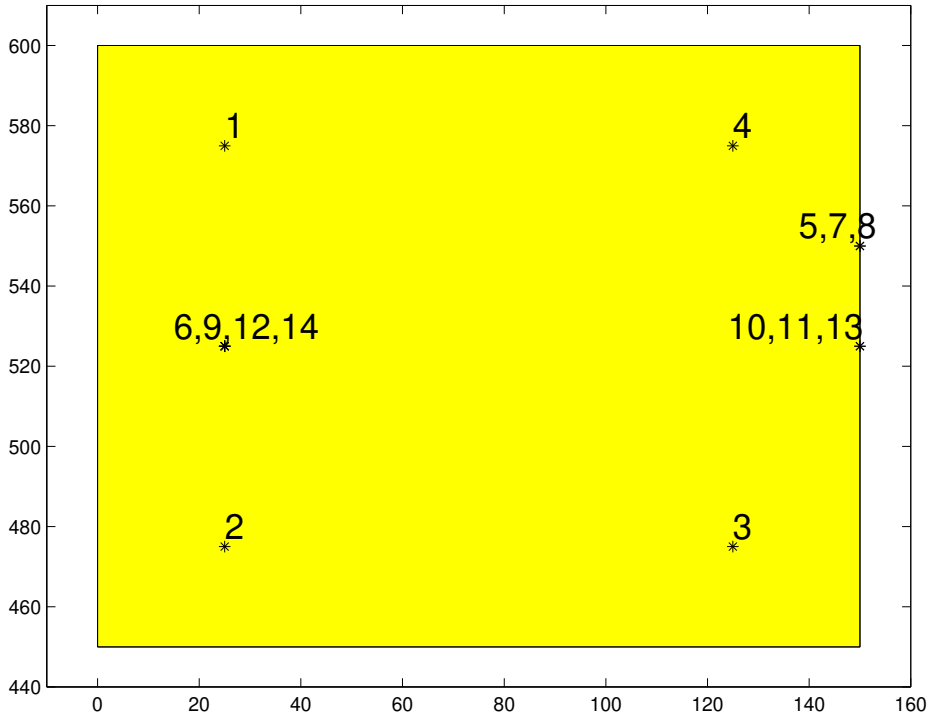
Plan séquentiel : après observation de $y(\mathbf{u}_j)$, $j = 1, \ldots, k$,

- estimer $\hat{\theta}_i^k$ (MC) pour $i = 1, 2, 3, 4$
- calculer la probabilité *a posteriori* $\pi_i(k)$ que le modèle i soit correct pour $i = 1, 2, 3, 4$

Initialisation : $\pi_i(0) = 1/4$, $i = 1, \ldots, 4$ et $\mathbf{u}_1, \ldots, \mathbf{u}_4$ donnés



Probabilités $\pi_i(k)$



Choix des facteurs $\mathbf{u}_k$

Université Nice SOPHIA ANTIPOLIS

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

Une méthode séquentielle simple pour choisir entre deux structures $\eta^{(1)}(\theta_1, \mathbf{u})$ et $\eta^{(2)}(\theta_2, \mathbf{u})$ [Atkinson & Fedorov 1975]

- Après l'observation de $y(\mathbf{u}_1), \ldots, y(\mathbf{u}_k)$ estimer $\hat{\theta}_1^k$ et $\hat{\theta}_2^k$ pour les deux modèles
- placer le nouveau point $\mathbf{u}_{k+1}$ là où $[\eta^{(1)}(\hat{\theta}_1^k, \mathbf{u}) - \eta^{(2)}(\hat{\theta}_2^k, \mathbf{u})]^2$ est maximum
- $k \longrightarrow k+1$, répéter

Une méthode séquentielle simple pour choisir entre deux structures $\eta^{(1)}(\theta_1, \mathbf{u})$ et $\eta^{(2)}(\theta_2, \mathbf{u})$ [Atkinson & Fedorov 1975]

- Après l'observation de $y(\mathbf{u}_1), \ldots, y(\mathbf{u}_k)$ estimer $\hat{\theta}_1^k$ et $\hat{\theta}_2^k$ pour les deux modèles
- placer le nouveau point $\mathbf{u}_{k+1}$ là où $[\eta^{(1)}(\hat{\theta}_1^k, \mathbf{u}) - \eta^{(2)}(\hat{\theta}_2^k, \mathbf{u})]^2$ est maximum
- $k \longrightarrow k+1$, répéter

Si plus de 2 modèles : estimer $\hat{\theta}_i^k$ pour tous, placer le prochain point en utilisant les deux modèles reproduisant le mieux les données (voir [Atkinson & Cox 1974; Hill 1978] pour une synthèse)

4) Planification et optimisation

Observations $y(i) = \eta(\bar{\theta}, \mathbf{u}_i) + \epsilon_i$, erreurs $(\epsilon_i)_i$ i.i.d. $\mathcal{N}(0, \sigma^2)$

Objectif : maximiser $\mathbb{E}\{y\} = \eta(\bar{\theta}, \mathbf{u})$

déterminer $\mathbf{u}^* = \mathbf{u}^*(\bar{\theta}) = \arg\max_{\mathbf{u}} \eta(\bar{\theta}, \mathbf{u})$

4) Planification et optimisation

Observations $y(i) = \eta(\bar{\theta}, \mathbf{u}_i) + \epsilon_i$, erreurs $(\epsilon_i)_i$ i.i.d. $\mathcal{N}(0, \sigma^2)$

Objectif : maximiser $\mathbb{E}\{y\} = \eta(\bar{\theta}, \mathbf{u})$

$$\text{déterminer } \mathbf{u}^* = \mathbf{u}^*(\bar{\theta}) = \arg\max_{\mathbf{u}} \eta(\bar{\theta}, \mathbf{u})$$

$\rightarrow$ estimer $\hat{\theta} = \hat{\theta}[\mathbf{y}]$, puis utiliser $\mathbf{u}^*(\hat{\theta})$

Quel critère pour choisir l'expérience ?

4) Planification et optimisation

Observations $y(i) = \eta(\bar{\theta}, \mathbf{u}_i) + \epsilon_i$, erreurs $(\epsilon_i)_i$ i.i.d. $\mathcal{N}(0, \sigma^2)$

Objectif : maximiser $\mathbb{E}\{y\} = \eta(\bar{\theta}, \mathbf{u})$

$$\text{déterminer } \mathbf{u}^* = \mathbf{u}^*(\bar{\theta}) = \arg\max_{\mathbf{u}} \eta(\bar{\theta}, \mathbf{u})$$

$\rightarrow$ estimer $\hat{\theta} = \hat{\theta}[\mathbf{y}]$, puis utiliser $\mathbf{u}^*(\hat{\theta})$

Quel critère pour choisir l'expérience ?

- a) précision sur $\hat{\theta}$?

4) Planification et optimisation

Observations $y(i) = \eta(\bar{\theta}, \mathbf{u}_i) + \epsilon_i$, erreurs $(\epsilon_i)_i$ i.i.d. $\mathcal{N}(0, \sigma^2)$

Objectif : maximiser $\mathbb{E}\{y\} = \eta(\bar{\theta}, \mathbf{u})$

$$\text{déterminer } \mathbf{u}^* = \mathbf{u}^*(\bar{\theta}) = \arg\max_{\mathbf{u}} \eta(\bar{\theta}, \mathbf{u})$$

$\rightarrow$ estimer $\hat{\theta} = \hat{\theta}[\mathbf{y}]$, puis utiliser $\mathbf{u}^*(\hat{\theta})$

Quel critère pour choisir l'expérience ?

- a) précision sur $\hat{\theta}$?
- ou b) précision sur $\mathbf{u}^*(\hat{\theta})$?

4) Planification et optimisation

Observations $y(i) = \eta(\bar{\theta}, \mathbf{u}_i) + \epsilon_i$, erreurs $(\epsilon_i)_i$ i.i.d. $\mathcal{N}(0, \sigma^2)$

Objectif : maximiser $\mathbb{E}\{y\} = \eta(\bar{\theta}, \mathbf{u})$

$$\text{déterminer } \mathbf{u}^* = \mathbf{u}^*(\bar{\theta}) = \arg\max_{\mathbf{u}} \eta(\bar{\theta}, \mathbf{u})$$

$\rightarrow$ estimer $\hat{\theta} = \hat{\theta}[\mathbf{y}]$, puis utiliser $\mathbf{u}^*(\hat{\theta})$

Quel critère pour choisir l'expérience ?

- a) précision sur $\hat{\theta}$?
- ou b) précision sur $\mathbf{u}^*(\hat{\theta})$?
- ou c) coût $C(\hat{\theta}|\bar{\theta})$

par ex., $C(\hat{\theta}|\bar{\theta}) = \eta[\bar{\theta}, \mathbf{u}^*(\bar{\theta})] - \eta[\bar{\theta}, \mathbf{u}^*(\hat{\theta})] \geq 0$

$\rightarrow$ minimiser le risque bayésien $R = \mathbb{E}\{C(\hat{\theta}[\mathbf{y}]|\bar{\theta})\}$

(espérance sur $\mathbf{y}$ et $\bar{\theta} \sim$ loi *a priori* $\pi(\cdot)$)

- a,b,c = planification «standard» : en 2 étapes

1) choisir des $\mathbf{u}_i$ pour l'estimation

2) estimer $\hat{\theta}$ et construire $\mathbf{u}^*(\hat{\theta})$

Mais chaque $\eta(\bar{\theta}, \mathbf{u}_i)$ est loin de l'optimum $\mathbf{u}^*$!
(choisis pour estimer, pas pour optimiser)

• a,b,c = planification «standard» : en 2 étapes

1) choisir des $\mathbf{u}_i$ pour l'estimation

2) estimer $\hat{\theta}$ et construire $\mathbf{u}^*(\hat{\theta})$

Mais chaque $\eta(\bar{\theta}, \mathbf{u}_i)$ est loin de l'optimum $\mathbf{u}^*$!
(choisis pour estimer, pas pour optimiser)

• On veut parfois les deux à la fois : $\eta(\bar{\theta}, \mathbf{u}_i)$ aussi grand que possible pour tous les i

$\rightarrow$ choisir $\mathbf{u}_i$ proche de $\mathbf{u}^*(\bar{\theta})$ . . . qui est inconnu !

$\rightarrow$ plan séquentiel

• a,b,c = planification «standard» : en 2 étapes
1) choisir des $\mathbf{u}_i$ pour l'estimation
2) estimer $\hat{\theta}$ et construire $\mathbf{u}^*(\hat{\theta})$

Mais chaque $\eta(\bar{\theta}, \mathbf{u}_i)$ est loin de l'optimum $\mathbf{u}^*$!
(choisis pour estimer, pas pour optimiser)

• On veut parfois les deux à la fois : $\eta(\bar{\theta}, \mathbf{u}_i)$ aussi grand que possible pour tous les i
$\rightarrow$ choisir $\mathbf{u}_i$ proche de $\mathbf{u}^*(\bar{\theta})$ . . . qui est inconnu !

$\rightarrow$ plan séquentiel

$$\cdots \rightarrow \mathbf{u}_i \rightarrow \text{observer } y(i) \rightarrow \text{estimer } \hat{\theta}^i = \hat{\theta}(\mathbf{y}_1^i) \rightarrow \mathbf{u}_{i+1} \rightarrow \cdots$$

retour d'information$\leftrightarrow$ problème de contrôle (système dynamique)

Chaque $\mathbf{u}_i$ a deux objectifs :

1) aider à estimer $\bar{\theta}$

2) essayer de maximiser $\eta(\bar{\theta}, \mathbf{u})$ (doit être près de $\mathbf{u}^*(\bar{\theta})$)

$\longrightarrow$ on parle de commande duale

Chaque $\mathbf{u}_i$ a deux objectifs :

1) aider à estimer $\bar{\theta}$

2) essayer de maximiser $\eta(\bar{\theta}, \mathbf{u})$ (doit être près de $\mathbf{u}^*(\bar{\theta})$)

$\longrightarrow$ on parle de commande duale

On peut s'intéresser aux propriétés théoriques asymptotiques (nb. de pas $N \to \infty$)

Chaque $\mathbf{u}_i$ a deux objectifs :

1) aider à estimer $\bar{\theta}$

2) essayer de maximiser $\eta(\bar{\theta}, \mathbf{u})$ (doit être près de $\mathbf{u}^*(\bar{\theta})$)

$\longrightarrow$ on parle de commande duale

On peut s'intéresser aux propriétés théoriques asymptotiques (nb. de pas $N \to \infty$)

ou rechercher une stratégie performante pour N fini

les deux sont DIFFICILES !

Optimisation «sans modèle» (on se rapproche du non-paramétrique...)
[Kiefer & Wolfowitz 1952]

on observe $y(\mathbf{u}_k) = f(\mathbf{u}_k) + \varepsilon_k$,

stratégie $[\mathbf{u}_{k+1}]_i = [\mathbf{u}_k]_i + \frac{\gamma_k}{c_k}\left[y(\mathbf{u}_k + c_k[\mathbf{e}]_i) - y(\mathbf{u}_k)\right]$, $i = 1, \ldots, d$

avec $\gamma_k, c_k > 0$, décroisssants, $\sum_k \gamma_k = \infty$, $\sum_k \gamma_k c_k < \infty$, $\sum_k \gamma_k^2 / c_k^2 < \infty$

Optimisation «sans modèle» (on se rapproche du non-paramétrique...)
[Kiefer & Wolfowitz 1952]

on observe $y(\mathbf{u}_k) = f(\mathbf{u}_k) + \varepsilon_k$,

stratégie $[\mathbf{u}_{k+1}]_i = [\mathbf{u}_k]_i + \frac{\gamma_k}{c_k}\left[y(\mathbf{u}_k + c_k[\mathbf{e}]_i) - y(\mathbf{u}_k)\right]$, $i = 1, \ldots, d$

avec $\gamma_k, c_k > 0$, décroisssants, $\sum_k \gamma_k = \infty$, $\sum_k \gamma_k c_k < \infty$, $\sum_k \gamma_k^2 / c_k^2 < \infty$

Converge vers un maximum local, mais lentement

(il faut beaucoup d'observations y_i...)

Optimisation «sans modèle» (on se rapproche du non-paramétrique...)
[Kiefer & Wolfowitz 1952]

on observe $y(\mathbf{u}_k) = f(\mathbf{u}_k) + \varepsilon_k$,

stratégie $[\mathbf{u}_{k+1}]_i = [\mathbf{u}_k]_i + \frac{\gamma_k}{c_k}\left[y(\mathbf{u}_k + c_k[\mathbf{e}]_i) - y(\mathbf{u}_k)\right]$, $i = 1, \ldots, d$

avec $\gamma_k, c_k > 0$, décroisssants, $\sum_k \gamma_k = \infty$, $\sum_k \gamma_k c_k < \infty$, $\sum_k \gamma_k^2 / c_k^2 < \infty$

Converge vers un maximum local, mais lentement
(il faut beaucoup d'observations y_i...)

Pourquoi plan d'expériences ? A chaque pas, plan avec $d+1$ points:

$U_k = \mathbf{u}_k$ et $\mathbf{u}_k + c_k[\mathbf{e}]_i$, $i = 1, \ldots, d$

$\simeq$ estimation du gradient $\nabla_u f$ par différences finies

$\simeq$ approximation linéaire (en $\mathbf{u}$) de f

$\simeq$ estimation du gradient $\nabla_u f$ par différences finies
$\simeq$ approximation linéaire (en $\mathbf{u}$) de f

Méthode de la surface de réponse : [Box & Wilson 1951]
utilise des modèles linéaires et/ou quadratiques en $\mathbf{u}$

$\simeq$ estimation du gradient $\nabla_u f$ par différences finies
$\simeq$ approximation linéaire (en $\mathbf{u}$) de f

Méthode de la surface de réponse : [Box & Wilson 1951]
utilise des modèles linéaires et/ou quadratiques en $\mathbf{u}$

Remarques TRES IMPORTANTES en paramérique :

• Pour un modèle paramétrique donné, un plan optimal conduit à répéter des observations toujours aux mêmes points (conséquence du Th. de Caratheodory)

$\simeq$ estimation du gradient $\nabla_u f$ par différences finies
$\simeq$ approximation linéaire (en $\mathbf{u}$) de f

Méthode de la surface de réponse : [Box & Wilson 1951]
utilise des modèles linéaires et/ou quadratiques en $\mathbf{u}$

Remarques TRES IMPORTANTES en paramérique :

- Pour un modèle paramétrique donné, un plan optimal conduit à répéter des observations toujours aux mêmes points (conséquence du Th. de Caratheodory)
- Ce sont les points qui apportent le plus d'information sur le modèle (sur ses paramètres)

$\simeq$ estimation du gradient $\nabla_u f$ par différences finies
$\simeq$ approximation linéaire (en $\mathbf{u}$) de f

Méthode de la surface de réponse : [Box & Wilson 1951] utilise des modèles linéaires et/ou quadratiques en $\mathbf{u}$

Remarques TRES IMPORTANTES en paramérique :

- Pour un modèle paramétrique donné, un plan optimal conduit à répéter des observations toujours aux mêmes points (conséquence du Th. de Caratheodory)
- Ce sont les points qui apportent le plus d'information sur le modèle (sur ses paramètres)
- Ce sont les points où notre prédiction du comportement du modèle est la plus incertaine (conséquence du Th. d'équivalence de Kiefer-Wolfowitz [1960])

Relie l'optimalité dans l'espace des paramètres θ à l'optimalité dans l'espace des réponses y

quand $N \to \infty$

$$\begin{cases} N\mathbf{var}[\hat{\theta}^N] & \to \mathbf{M}_F^{-1}(\xi, \bar{\theta}) \\ N\mathbf{var}[\eta(\hat{\theta}^N, u)] & \to \frac{\partial \eta(\theta,u)}{\partial \theta^\top} \mathbf{M}_F^{-1}(\xi, \bar{\theta}) \frac{\partial \eta(\theta,u)}{\partial \theta} \end{cases}$$

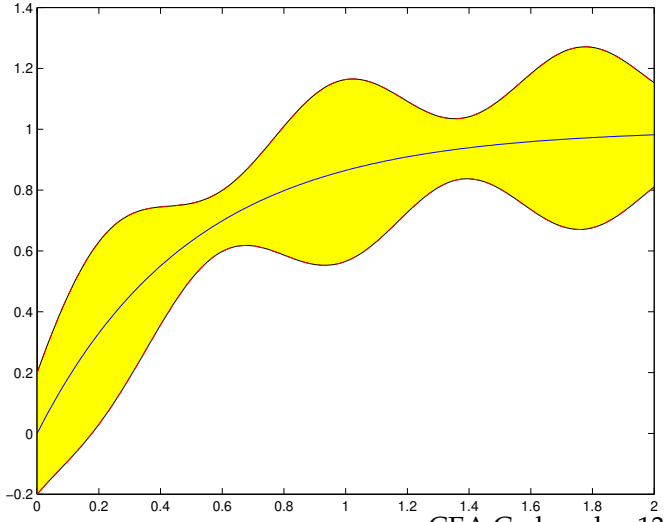
Relie l'optimalité dans l'espace des paramètres θ à l'optimalité dans l'espace des réponses y

$$\text{quand } N \to \infty \left\{ \begin{array}{rcl} N\mathbf{var}[\hat{\theta}^N] & \to & \mathbf{M}_F^{-1}(\xi, \bar{\theta}) \\ N\mathbf{var}[\eta(\hat{\theta}^N, u)] & \to & \frac{\partial \eta(\theta,u)}{\partial \theta^\top} \mathbf{M}_F^{-1}(\xi, \bar{\theta}) \frac{\partial \eta(\theta,u)}{\partial \theta} \end{array} \right.$$

Doptimalité $\Leftrightarrow$ G-optimality

ξ_D minimise $\det \mathbf{M}_F^{-1}(\xi, \bar{\theta})$ et minimise le maximum de la variance de la prédiction sur le domaine expérimental $\mathcal{U}$

$\eta(\hat{\theta}, u)$, $\eta(\hat{\theta}, u) \pm 2\sigma$



5) Modèles non-paramétriques

Plan d'expériences : apprentissage actif
A partir de données «d'apprentissage»
$\mathcal{D} = \{[u_1, y(u_1)], \ldots, [u_N, y(u_N)]\}$ on souhaite prédire $y(u)$ pour un nouveau u par un modèle non-paramétrique (beaucoup de paramètres, mais sans intérêt) :
réseau de neurones, machine à vecteurs de support (SVM), estimateur de Nadaraya-Watson, splines, krigeage (Kriging) (= processus gaussien), etc.

5) Modèles non-paramétriques

Plan d'expériences : apprentissage actif
A partir de données «d'apprentissage»
$\mathcal{D} = \{[u_1, y(u_1)], \ldots, [u_N, y(u_N)]\}$ on souhaite prédire $y(u)$ pour un nouveau u par un modèle non-paramétrique (beaucoup de paramètres, mais sans intérêt) :
réseau de neurones, machine à vecteurs de support (SVM), estimateur de Nadaraya-Watson, splines, krigeage (Kriging) (= processus gaussien), etc.

$\hat{y}_{\mathcal{D}}(u)$ est la prédiction en u
$\mathbf{y} = [y(u_1), \ldots, y(u_N)]^\top$ est le vecteurs des observations disponibles

1) Réseau de neurones, une couche cachée (RBF)

$\hat{y}_{\mathcal{D}}(u) = \sum_{k=1}^{N} a_k K(u - u_k)$, par ex. $K(z) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{z^2}{2h^2}\right)$

$\rightarrow$ les *nœuds* coincident avec les facteurs u_k

Soit $\mathbf{a} = (a_1, \ldots, a_N)^\top$

$\hat{\mathbf{a}}$ minimise $\gamma \sum_{k=1}^{N} [y(u_k) - \hat{y}_{\mathcal{D}}(u_k)]^2 + \mathbf{a}^\top \mathbf{G} \mathbf{a}$

$\rightarrow (\mathbf{G} + \mathbf{I}/\gamma)\hat{\mathbf{a}} = \mathbf{y}$ avec $\mathbf{G}_{i,j} = K(u_i - u_j)$

1) Réseau de neurones, une couche cachée (RBF)

$\hat{y}_{\mathcal{D}}(u) = \sum_{k=1}^{N} a_k K(u - u_k)$, par ex. $K(z) = \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{z^2}{2h^2}\right)$

$\rightarrow$ les *nœuds* coincident avec les facteurs u_k

Soit $\mathbf{a} = (a_1, \ldots, a_N)^\top$

$\hat{\mathbf{a}}$ minimise $\gamma \sum_{k=1}^{N} [y(u_k) - \hat{y}_{\mathcal{D}}(u_k)]^2 + \mathbf{a}^\top \mathbf{G} \mathbf{a}$

$\rightarrow (\mathbf{G} + \mathbf{I}/\gamma)\hat{\mathbf{a}} = \mathbf{y}$ avec $\mathbf{G}_{i,j} = K(u_i - u_j)$

Version paramétrique : M ***nœuds*** N_i **fixés,** $i = 1, \ldots, M$ répartis dans $\mathcal{U}$

- prédire par $\hat{y}_{\mathcal{D}}(u) = \sum_{i=1}^{M} a_i K(u - N_i)$
- estimer $\mathbf{a}$ par MC : $\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \sum_{k=1}^{N} [y(u_k) - \hat{y}_{\mathcal{D}}(u_k)]^2$

(modèle linéaire comme dans l'exemple 1)

2) Nadaraya [1964] & Watson [1964]: lissage local

$$\hat{y}_{\mathcal{D}}(u) = \frac{1}{G(u)} \sum_{i=1}^{N} K(u - u_i) y(u_i)$$

avec $G(u) = \sum_{i=1}^{N} K(u - u_i)$ et $K(\cdot)$ un noyau symétrique unimodal (par ex. la densité de la loi normale)

3) SVM (voir par ex. [Suykens *et al.*, 2002]

prédiction $\hat{y}_{\mathcal{D}}(u) = \phi^{\top}(u)\theta$, $\dim(\theta) = p$ très grande
(éventuellement $\infty \rightarrow$ modèle linéaire de dim. infinie)

3) SVM (voir par ex. [Suykens *et al.*, 2002]

prédiction $\hat{y}_{\mathcal{D}}(u) = \phi^{\top}(u)\theta$, $\dim(\theta) = p$ très grande (éventuellement $\infty \rightarrow$ modèle linéaire de dim. infinie)

Classique : modèle non-paramétrique $\Rightarrow$ beaucoup de paramètres !

3) SVM (voir par ex. [Suykens *et al.*, 2002]

prédiction $\hat{y}_{\mathcal{D}}(u) = \phi^\top(u)\theta$, $\dim(\theta) = p$ très grande (éventuellement $\infty \rightarrow$ modèle linéaire de dim. infinie)

Classique : modèle non-paramétrique $\Rightarrow$ beaucoup de paramètres !

Pour $f_\theta(u) = \sum_{i=1}^{\infty} \theta_i \phi_i(u)$ on définit

$\langle f_\theta(\cdot), f_{\theta'}(\cdot) \rangle$ = produit scalaire entre $f_\theta(\cdot)$ et $f_{\theta'}(\cdot)$

$= \sum_{i=1}^{\infty} \theta_i \theta'_i$

$\rightarrow$ RKHS (Reproducing Kernel Hilbert Space), avec des noyaux $K(x,z) = \sum_{i=1}^{\infty} \phi_i(x)\phi_i(z)$

et

$$\begin{aligned} \langle f_\theta(\cdot), K(\cdot, z) \rangle &= f_\theta(z) \\ \langle K(x,\cdot), K(\cdot, z) \rangle &= K(x,z) \end{aligned}$$

Choisir $\hat{\theta}$ qui minimise $\gamma \sum_{k=1}^{N} L[y_k - \theta^\top \phi(u_k)] + \frac{1}{2}\|\theta\|^2$

$\hat{\theta}$ satisfait $-\sum_{k=1}^{N} \underbrace{\gamma L'[y_k - \theta^\top \phi(u_k)]}_{=a_k} \phi_i(u_k) + \hat{\theta}_i = 0$, pour tout i

$\rightarrow \hat{\theta}_i = \sum_{k=1}^{N} a_k \phi_i(u_k)$, pour tout i

Choisir $\hat{\theta}$ qui minimise $\gamma \sum_{k=1}^{N} L[y_k - \theta^\top \phi(u_k)] + \frac{1}{2}\|\theta\|^2$

$\hat{\theta}$ satisfait $-\sum_{k=1}^{N} \underbrace{\gamma L'[y_k - \theta^\top \phi(u_k)]}_{=a_k} \phi_i(u_k) + \hat{\theta}_i = 0$, pour tout i

$\rightarrow \hat{\theta}_i = \sum_{k=1}^{N} a_k \phi_i(u_k)$, pour tout i

Prédiction en u : (*Th. de représentation*)

$$\hat{y}_{\mathcal{D}}(u) = \hat{\theta}^\top \phi(u) = \sum_{i=1}^{\infty} \hat{\theta}_i \phi_i(u) = \sum_{k=1}^{N} a_k \underbrace{\sum_{i=1}^{\infty} \phi_i(u_k)\phi_i(u)}_{=K(u_k,u)}$$

avec les a_k qui satisfont

$$a_k = \gamma L' \left[y_k - \sum_{i=1}^{N} a_i K(u_i, u_k) \right]$$

Choisir $\hat{\theta}$ qui minimise $\gamma \sum_{k=1}^{N} L[y_k - \theta^\top \phi(u_k)] + \frac{1}{2}\|\theta\|^2$

$\hat{\theta}$ satisfait $-\sum_{k=1}^{N} \underbrace{\gamma L'[y_k - \theta^\top \phi(u_k)]}_{=a_k} \phi_i(u_k) + \hat{\theta}_i = 0$, pour tout i

$\rightarrow \hat{\theta}_i = \sum_{k=1}^{N} a_k \phi_i(u_k)$, pour tout i

Prédiction en u : (*Th. de représentation*)

$$\hat{y}_{\mathcal{D}}(u) = \hat{\theta}^\top \phi(u) = \sum_{i=1}^{\infty} \hat{\theta}_i \phi_i(u) = \sum_{k=1}^{N} a_k \underbrace{\sum_{i=1}^{\infty} \phi_i(u_k)\phi_i(u)}_{=K(u_k,u)}$$

avec les a_k qui satisfont

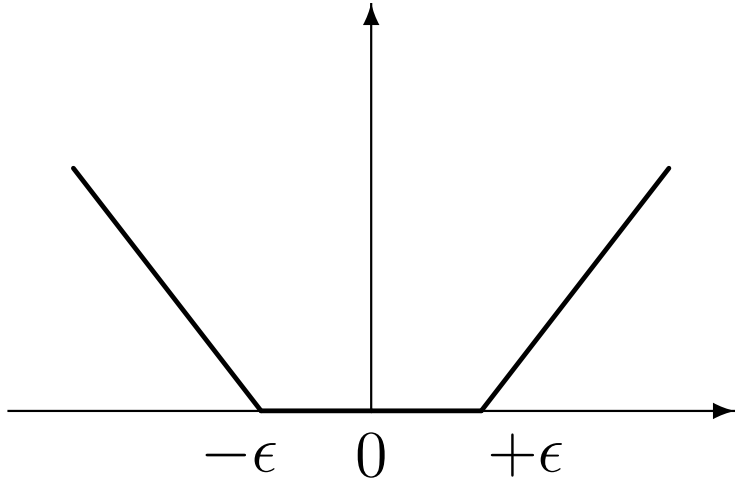
$$a_k = \gamma L' \left[y_k - \sum_{i=1}^{N} a_i K(u_i, u_k) \right]$$

Soit $\mathbf{a} = (a_1, \ldots, a_N)^\top$

Différentes fonctions de coût $L \rightarrow$ différentes méthodes pour choisir $\mathbf{a}$

- LS-SVM [Suykens *et al.*, 2002]: $L(u) = u^2/2$ et $\mathbf{a}$ satisfait $[\Omega + \mathbf{I}/\gamma]\mathbf{a} = \mathbf{y}$, avec $\Omega_{i,j} = K(u_i, u_j)$

- LS-SVM [Suykens *et al.*, 2002]: $L(u) = u^2/2$ et $\mathbf{a}$ satisfait $[\Omega + \mathbf{I}/\gamma]\mathbf{a} = \mathbf{y}$, avec $\Omega_{i,j} = K(u_i, u_j)$
- Coût SVM classique : $L(x) = \epsilon + \max\{|x| - \epsilon, 0\}$ (non différentiable)



$\mathbf{a}$ obtenu par résolution d'un problème convexe

$\Rightarrow$ de nombreux a_k valent zéro dans $\hat{y}_{\mathcal{D}}(u) = \sum_{k=1}^{N} a_k K(u_k, u)$
ceux $\neq 0 \rightarrow$ les «vecteurs de support» u_k

On peut travailler directement avec $K(u, z)$
Condition sur K ?

On peut travailler directement avec $K(u, z)$
Condition sur K ?

Mercer : $\int K(u, z)g(u)g(z)\,du\,dz \geq 0$ pour tout $g(\cdot)$ dans $\mathcal{L}_2$
(pour définir un produit salaire)
$\rightarrow$ considérer $K(u, z)$ comme la covariance d'un processus dans l'espace des u

On peut travailler directement avec $K(u, z)$
Condition sur K ?

Mercer : $\int K(u, z)g(u)g(z)\,du\,dz \geq 0$ pour tout $g(\cdot)$ dans $\mathcal{L}_2$
(pour définir un produit salaire)
$\rightarrow$ considérer $K(u, z)$ comme la covariance d'un processus dans l'espace des u

- Espace hilbertien engendré par une fonction de covariance
- Processus aléatoire à trajectoires dans un RKHS

4) Processus gaussien & krigeage (voir par ex. [Stein 1999])

Modèle $\boxed{y(u_k) = \bar{\theta}_0 + P(u_k, \omega) + \varepsilon(u_k)}$ avec

4) Processus gaussien & krigeage (voir par ex. [Stein 1999])

Modèle $\boxed{y(u_k) = \bar{\theta}_0 + P(u_k, \omega) + \varepsilon(u_k)}$ avec

- $P(u, \omega)$ un processus stationnaire du second-ordre,

de covariance $\mathbb{E}\{P(u,\omega)P(z,\omega)\} = K(u,z) = \sigma_P^2 C(u-z)$

4) Processus gaussien & krigeage (voir par ex. [Stein 1999])

Modèle $\boxed{y(u_k) = \bar{\theta}_0 + P(u_k, \omega) + \varepsilon(u_k)}$ avec

- $P(u, \omega)$ un processus stationnaire du second-ordre,

de covariance $\mathbb{E}\{P(u,\omega)P(z,\omega)\} = K(u,z) = \sigma_P^2 C(u-z)$

- (ε_k) des erreurs i.i.d., de moyenne nulle et de variance σ^2

4) Processus gaussien & krigeage (voir par ex. [Stein 1999])

Modèle $\boxed{y(u_k) = \bar{\theta}_0 + P(u_k, \omega) + \varepsilon(u_k)}$ avec

- $P(u, \omega)$ un processus stationnaire du second-ordre,

de covariance $\mathbb{E}\{P(u,\omega)P(z,\omega)\} = K(u,z) = \sigma_P^2 C(u-z)$

- (ε_k) des erreurs i.i.d., de moyenne nulle et de variance σ^2

On définit la matrice $\mathbf{C}_P$ $(N \times N)$ par $[\mathbf{C}_P]_{i,j} = C(u_i - u_j)$

Krigeage : meilleur prédicteur linéaire sans biais en u:

$\hat{y}_{\mathcal{D}}(u) = \mathbf{v}^\top(u)\mathbf{y}$

$\mathbf{v}(u)$ minimise $\mathbb{E}\{(\mathbf{v}^\top\mathbf{y} - [\bar{\theta}_0 + P(u,\omega)])^2\}$ sous la contrainte

$\mathbb{E}\{\mathbf{v}^\top\mathbf{y}\} = \bar{\theta}_0 \sum_{i=1}^N v_i = \mathbb{E}\{y(u)\} = \bar{\theta}_0 \Rightarrow \sum_{i=1}^N v_i = 1$

Solution explicite !

Krigeage : meilleur prédicteur linéaire sans biais en u:

$\hat{y}_{\mathcal{D}}(u) = \mathbf{v}^\top(u)\mathbf{y}$

$\mathbf{v}(u)$ minimise $\mathbb{E}\{(\mathbf{v}^\top\mathbf{y} - [\bar{\theta}_0 + P(u,\omega)])^2\}$ sous la contrainte

$\mathbb{E}\{\mathbf{v}^\top\mathbf{y}\} = \bar{\theta}_0 \sum_{i=1}^N v_i = \mathbb{E}\{y(u)\} = \bar{\theta}_0 \Rightarrow \sum_{i=1}^N v_i = 1$

Solution explicite !

Posons $\mathbf{C}_y = \sigma^2\mathbf{I}_N + \sigma_P^2\mathbf{C}_P$, $\mathbf{1} = (\underbrace{1,\ldots,1}_{N \text{ termes}})^\top$

$\mathbf{c}(u) = \sigma_P^2[C(u-u_1),\ldots,C(u-u_N)]^\top$

$\hat{\theta}_0 = \dfrac{\mathbf{1}^\top\mathbf{C}_y^{-1}\mathbf{y}}{\mathbf{1}^\top\mathbf{C}_y^{-1}\mathbf{1}}$ (estimateur des MC pondérés de θ_0)

alors, $\boxed{\hat{y}_{\mathcal{D}}(u) = \mathbf{v}^\top(u)\mathbf{y} = \hat{\theta}^0 + \mathbf{c}^\top(u)\mathbf{C}_y^{-1}(\mathbf{y} - \hat{\theta}^0\mathbf{1})}$

$(= \hat{\theta}^0 + \sum_{k=1}^N a_k K(u,u_k))$

De plus, erreur quadratique moyenne (EQM) de la prédiction $\hat{y}_{\mathcal{D}}(u)$ en u :

$$\rho^2_{\mathcal{D}}(u) = \sigma^2_P - \left[\begin{array}{cc} \mathbf{c}^\top(u) & 1 \end{array}\right] \left[\begin{array}{cc} \mathbf{C}_y & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{array}\right]^{-1} \left[\begin{array}{c} \mathbf{c}(u) \\ 1 \end{array}\right]$$

et, si $\sigma^2 = 0$ (pas d'erreur de mesure ε_k)

$\hat{y}_{\mathcal{D}}(u_i) = y(u_i)$ et $\rho^2_{\mathcal{D}}(u_i) = 0$ pour tout i

$\Rightarrow \hat{y}_{\mathcal{D}}(u)$ est un interpolateur parfait !

De plus, erreur quadratique moyenne (EQM) de la prédiction $\hat{y}_{\mathcal{D}}(u)$ en u :

$$\rho^2_{\mathcal{D}}(u) = \sigma^2_P - \begin{bmatrix} \mathbf{c}^\top(u) & 1 \end{bmatrix} \begin{bmatrix} \mathbf{C}_y & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{c}(u) \\ 1 \end{bmatrix}$$

et, si $\sigma^2 = 0$ (pas d'erreur de mesure ε_k)

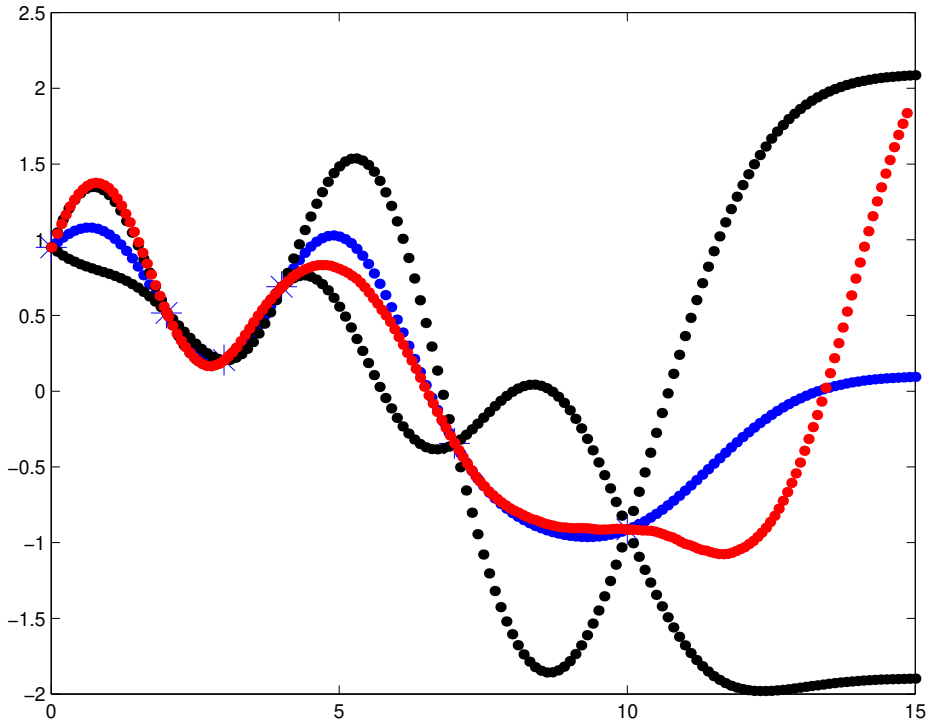
$\hat{y}_{\mathcal{D}}(u_i) = y(u_i)$ et $\rho^2_{\mathcal{D}}(u_i) = 0$ pour tout i

$\Rightarrow \hat{y}_{\mathcal{D}}(u)$ est un interpolateur parfait !

Erreurs de modèle = processus aléatoire

$\longrightarrow$ inférence statistique même pour un système purement déterministe

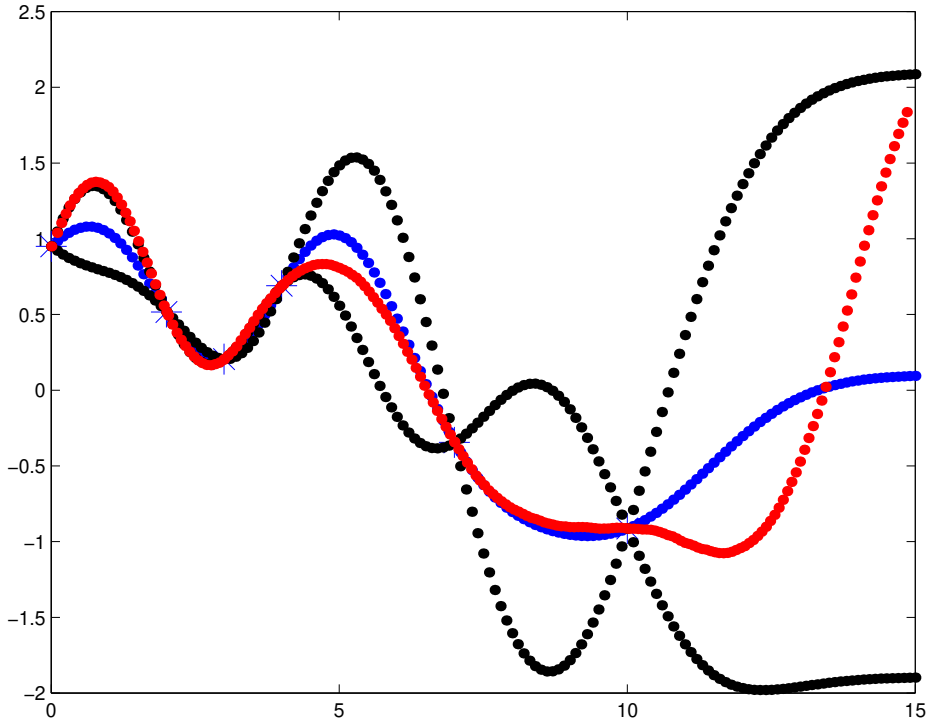
aussi quand expérience = simulation (ex. éléments finis) — computer experiments — [Sacks *et al.*, 1989]



processus $y(u)$

prédiction $\hat{y}_{\mathcal{D}}(u)$

et $\hat{y}_{\mathcal{D}}(u) \pm 2\rho_{\mathcal{D}}(u)$



processus $y(u)$
prédiction $\hat{y}_{\mathcal{D}}(u)$
et $\hat{y}_{\mathcal{D}}(u) \pm 2\rho_{\mathcal{D}}(u)$

Si le processus $P(u, \omega)$ et les erreurs ε_k appartiennent à une famille paramétrée $\longrightarrow$ estimer les paramètres

Par ex., processus gaussien avec $C(z) = C(\beta, z)$ + erreurs normales ε_k

$\longrightarrow$ estimer β, σ_P^2 et σ^2 par Maximum de Vraisemblance

Extensions :

1) Remplacer la moyenne constante θ_0 par un modèle linéaire $\mathbf{r}^\top(u)\theta$

• meilleur prédicteur linéaire sans biais $\longrightarrow$ krigeage universel

• processus généralisé, covariances généralisées, fonction aléatoires intrinsèques (permet d'étendre la classe des fonctions de covariance que l'on peut utiliser, équivalence avec les splines [Vasquez, 2005])

Extensions :

1) Remplacer la moyenne constante θ_0 par un modèle linéaire $\mathbf{r}^\top(u)\theta$

- meilleur prédicteur linéaire sans biais $\rightarrow$ krigeage universel
- processus généralisé, covariances généralisées, fonction aléatoires intrinsèques (permet d'étendre la classe des fonctions de covariance que l'on peut utiliser, équivalence avec les splines [Vasquez, 2005])

2) Mettre une loi *a priori* sur θ (krigeage bayesien)

Extensions :

1) Remplacer la moyenne constante θ_0 par un modèle linéaire $\mathbf{r}^\top(u)\theta$

• meilleur prédicteur linéaire sans biais $\rightarrow$ krigeage universel

• processus généralisé, covariances généralisées, fonction aléatoires intrinsèques (permet d'étendre la classe des fonctions de covariance que l'on peut utiliser, équivalence avec les splines [Vasquez, 2005])

2) Mettre une loi *a priori* sur θ (krigeage bayesien)

3) Prédire la dérivée de $y(u)$ (ou prédire $y(u)$ à partir d'observations de la dérivée aux points u_i) [Vasquez, 2005]

6) Planification en non paramétrique

où placer les u_k ?

6) Planification en non paramétrique

où placer les u_k ?

A) APPROCHE SANS MODÈLE : remplir l'espace

- $u \in \mathcal{U}$ = espace admissible,
- $\mathcal{S} \subset \mathcal{U}$ les sites choisis u_k, $k = 1, \ldots, N$

6) Planification en non paramétrique

où placer les u_k ?

A) APPROCHE SANS MODÈLE : remplir l'espace

- $u \in \mathcal{U}$ = espace admissible,
- $\mathcal{S} \subset \mathcal{U}$ les sites choisis u_k, $k = 1, \ldots, N$

[Johnson *et al.*, 1990]:

- distance **maximin** :

$$\max_{\mathcal{S}} \min_{u \neq u' \in \mathcal{S}^2} d(u, u')$$

$\Rightarrow$ étaler les u_k autant que possible sur $\mathcal{U}$

($\rightarrow$ points sur les bords de $\mathcal{U}$)

6) Planification en non paramétrique

où placer les u_k ?

A) APPROCHE SANS MODÈLE : remplir l'espace

- $u \in \mathcal{U}$ = espace admissible,
- $\mathcal{S} \subset \mathcal{U}$ les sites choisis u_k, $k = 1, \ldots, N$

[Johnson *et al.*, 1990]:

- distance **maximin** :

$$\max_{\mathcal{S}} \min_{u \neq u' \in \mathcal{S}^2} d(u, u')$$

$\Rightarrow$ étaler les u_k autant que possible sur $\mathcal{U}$

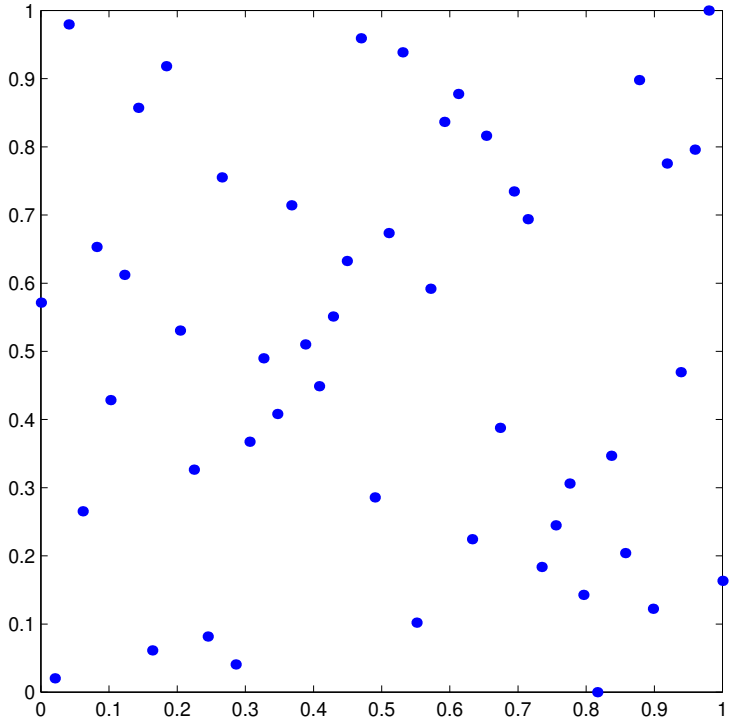
($\rightarrow$ points sur les bords de $\mathcal{U}$)

- distance **minimax** : pour $\mathcal{U}$ un ensemble fini

$$\min_{\mathcal{S}} \max_{z \in \mathcal{U}} \underbrace{\min_{u \in \mathcal{S}} d(z, u)}_{= d(z, \mathcal{S})}$$

Pour assurer de bonnes propriétés de projection (pour chaque coordonnée de u) : travailler dans la classe des **Hypercubes Latins** [Morris & Mitchell 1995]

$\mathcal{U} = [0,1]^d$, les $[u_k]_i$, $k = 1, \ldots, N$, prennent les valeurs $0, 1/(N-1), 2/(N-1), \ldots, 1$, pour chaque $i = 1, \ldots, d$



B) Approche avec modèle :

Il faut une caractérisation de l'incertitude sur $\hat{y}_D(u)$

B) Approche avec modèle :

Il faut une caractérisation de l'incertitude sur $\hat{y}_D(u)$

Soit $\rho_D(u) = EQM(u) = \mathbb{E}\{[\hat{y}_{\mathcal{D}}(u) - y(u)]^2\} - \sigma^2$
($\mathbb{E}$ porte sur $\mathbf{y}$ et $y(u)$)

Une question générale en non paramétrique :

Comment décroît $EQM(u)$ quand N augmente ?

B) APPROCHE AVEC MODÈLE :

Il faut une caractérisation de l'incertitude sur $\hat{y}_D(u)$

Soit $\rho_D(u) = EQM(u) = \mathbb{E}\{[\hat{y}_{\mathcal{D}}(u) - y(u)]^2\} - \sigma^2$
($\mathbb{E}$ porte sur $\mathbf{y}$ et $y(u)$)

Une question générale en non paramétrique :

Comment décroît $EQM(u)$ quand N augmente ?

. . . pas évident ! l'effet est local : on observe en u,
$\Rightarrow$ $\rho_{\mathcal{D}}(z)$ décroît pour z près de u
(krigeage sans erreur de mesure : $\rho_{\mathcal{D}}(u)$ devient 0)

mais peu de changement loin de u. . .

Hypothèse classique :
$\mathcal{D}$ forme une suite de paires i.i.d. $[u_k, y(u_k)]$, voir par ex. [Cucker & Smale 2001; Bartlett 2003] $\longrightarrow$ **pas de planification**

Hypothèse classique :
$\mathcal{D}$ forme une suite de paires i.i.d. $[u_k, y(u_k)]$, voir par ex. [Cucker & Smale 2001; Bartlett 2003] $\rightarrow$ **pas de planification**

Planification difficile, peu de résultats précis...
[Cheng *et al.*, 1998] : répartir les $u_i \in I\!R$ avec une densité de probabilité proportionnelle à $|y''(u)|^{2/9}$

Hypothèse classique :
$\mathcal{D}$ forme une suite de paires i.i.d. $[u_k, y(u_k)]$, voir par ex. [Cucker & Smale 2001; Bartlett 2003] $\rightarrow$ **pas de planification**

Planification difficile, peu de résultats précis...
[Cheng *et al.*, 1998] : répartir les $u_i \in I\!R$ avec une densité de probabilité proportionnelle à $|y''(u)|^{2/9}$

...Pour le krigeage, on peut utiliser $EQM(u) = \rho^2_{\mathcal{D}}(u)$

- $\mathcal{S}$ minimise l'EQM maximale = $\max_{u \in \mathcal{U}} \rho^2_{\mathcal{D}}(u)$

(lié à la distance minimax, [Johnson *et al.*, 1990])

- $\mathcal{S}$ minimise l'EQM maximale = $\max_{u \in \mathcal{U}} \rho^2_{\mathcal{D}}(u)$

(lié à la distance minimax, [Johnson *et al.*, 1990])

ou

- $\mathcal{S}$ minimise l'EQM moyenne = $\int_{\mathcal{U}} \rho^2_{\mathcal{D}}(u)\, \pi(du)$

(π une mesure de probabilité pour u)

• $\mathcal{S}$ minimise l'EQM maximale = $\max_{u \in \mathcal{U}} \rho^2_{\mathcal{D}}(u)$
(lié à la distance minimax, [Johnson *et al.*, 1990])
ou
• $\mathcal{S}$ minimise l'EQM moyenne = $\int_{\mathcal{U}} \rho^2_{\mathcal{D}}(u)\, \pi(du)$
(π une mesure de probabilité pour u)
ou
• **Maximum Entropy Sampling** : [Shewry & Wynn 1987, Wynn 2004]
Soit $\mathrm{ent}_Y = -\int \log \varphi(Y) \varphi(Y) dY$
$Y_{\mathcal{A}} = y(u)$ pour $u \in \mathcal{A}$ (et donc $Y_{\mathcal{S}} = \mathbf{y}$ = vecteur des observations)

• $\mathcal{S}$ minimise l'EQM maximale = $\max_{u \in \mathcal{U}} \rho^2_{\mathcal{D}}(u)$
(lié à la distance minimax, [Johnson *et al.*, 1990])
ou
• $\mathcal{S}$ minimise l'EQM moyenne = $\int_{\mathcal{U}} \rho^2_{\mathcal{D}}(u)\, \pi(du)$
(π une mesure de probabilité pour u)
ou
• **Maximum Entropy Sampling** : [Shewry & Wynn 1987, Wynn 2004]
Soit $\mathrm{ent}_Y = -\int \log \varphi(Y) \varphi(Y) dY$
$Y_{\mathcal{A}} = y(u)$ pour $u \in \mathcal{A}$ (et donc $Y_{\mathcal{S}} = \mathbf{y}$ = vecteur des observations)

Une approche bayesienne minimiserait l'espérance de l'entropie *a posteriori*
$\mathbb{E}\{\mathrm{ent}(Y_{\mathcal{U}\setminus\mathcal{S}} | Y_{\mathcal{S}})\}$ . . . ce qui est difficile

On décompose l'entropie en $\mathrm{ent}(Y_{\mathcal{U}}) = \mathrm{ent}(Y_{\mathcal{S}}) + \mathbb{E}\{\mathrm{ent}(Y_{\mathcal{U}\setminus\mathcal{S}}|Y_{\mathcal{S}})\}$, puisque $\mathrm{ent}(Y_{\mathcal{U}})$ est fixée, choisir $\mathcal{S}$ qui maximise $\mathrm{ent}(Y_{\mathcal{S}})$

On décompose l'entropie en $\mathrm{ent}(Y_{\mathcal{U}}) = \mathrm{ent}(Y_{\mathcal{S}}) + \mathbb{E}\{\mathrm{ent}(Y_{\mathcal{U}\setminus\mathcal{S}}|Y_{\mathcal{S}})\}$, puisque $\mathrm{ent}(Y_{\mathcal{U}})$ est fixée, choisir $\mathcal{S}$ qui maximise $\mathrm{ent}(Y_{\mathcal{S}})$

Pour le krigeage avec processus gaussien et sans erreurs de mesure («computer experiments»)

$\Rightarrow$ maximiser $\det(\mathbf{C}_P)$, avec $[\mathbf{C}_P]_{i,j} = C(u_i - u_j)$

(lié à la distance maximin, [Johnson *et al.*, 1990])

On décompose l'entropie en $\mathrm{ent}(Y_{\mathcal{U}}) = \mathrm{ent}(Y_{\mathcal{S}}) + \mathbb{E}\{\mathrm{ent}(Y_{\mathcal{U}\setminus\mathcal{S}}|Y_{\mathcal{S}})\}$, puisque $\mathrm{ent}(Y_{\mathcal{U}})$ est fixée, choisir $\mathcal{S}$ qui maximise $\mathrm{ent}(Y_{\mathcal{S}})$

Pour le krigeage avec processus gaussien et sans erreurs de mesure («computer experiments»)

$\Rightarrow$ maximiser $\det(\mathbf{C}_P)$, avec $[\mathbf{C}_P]_{i,j} = C(u_i - u_j)$

(lié à la distance maximin, [Johnson *et al.*, 1990])

• Application du krigeage à l'optimisation globale dans la partie VII

Remarque : différence avec modèle paramétrique

$y(u_k) = \eta(\bar{\theta}, u_k) + \varepsilon_k$

$\eta(\theta, u)$ **est fixé (pas modifié quand d'autres** u_k **sont utilisés)**

Même problème : choisir des facteurs $U_1^N = (u_1, \ldots, u_N)$

assurant une «bonne prédiction» $\hat{y}_{\mathcal{D}}(u)$ quand $u \in \mathcal{U} \subset I\!R^d$

Remarque : différence avec modèle paramétrique

$y(u_k) = \eta(\bar{\theta}, u_k) + \varepsilon_k$

$\eta(\theta, u)$ **est fixé (pas modifié quand d'autres** u_k **sont utilisés)**

Même problème : choisir des facteurs $U_1^N = (u_1, \ldots, u_N)$

assurant une «bonne prédiction» $\hat{y}_{\mathcal{D}}(u)$ quand $u \in \mathcal{U} \subset \mathbb{R}^d$

$$EQM(u) = \mathbb{E}\{[\hat{y}_{\mathcal{D}}(u) - y(u)]^2\} - \sigma^2$$

(espérance sur $y(u)$ et $\mathcal{D}$ pour U_1^N fixé)

La variance domine le carré du biais, et

$$EQM(u) \simeq \underbrace{\mathbb{E}\{[\hat{y}_{\mathcal{D}}(u) - \mathbb{E}\{\hat{y}_{\mathcal{D}}(u)\}]^2\}}_{V(u|U_1^N)}$$

Ici $\hat{y}_{\mathcal{D}}(u) = \eta(\hat{\theta}^N, u)$ avec $\hat{\theta}^N$ estimé à partir de $\mathcal{D}$, et

$V(u|U_1^N)$ est relié à $\mathrm{Var}(\hat{\theta}^N)$, décroît en $1/N$, **pour tout** u**!**

Remarque : différence avec modèle paramétrique

$y(u_k) = \eta(\bar{\theta}, u_k) + \varepsilon_k$

$\eta(\theta, u)$ **est fixé (pas modifié quand d'autres** u_k **sont utilisés)**

Même problème : choisir des facteurs $U_1^N = (u_1, \ldots, u_N)$

assurant une «bonne prédiction» $\hat{y}_{\mathcal{D}}(u)$ quand $u \in \mathcal{U} \subset I\!R^d$

$$EQM(u) = \mathbb{E}\{[\hat{y}_{\mathcal{D}}(u) - y(u)]^2\} - \sigma^2$$

(espérance sur $y(u)$ et $\mathcal{D}$ pour U_1^N fixé)

La variance domine le carré du biais, et

$$EQM(u) \simeq \underbrace{\mathbb{E}\{[\hat{y}_{\mathcal{D}}(u) - \mathbb{E}\{\hat{y}_{\mathcal{D}}(u)\}]^2\}}_{V(u|U_1^N)}$$

Ici $\hat{y}_{\mathcal{D}}(u) = \eta(\hat{\theta}^N, u)$ avec $\hat{\theta}^N$ estimé à partir de $\mathcal{D}$, et

$V(u|U_1^N)$ est relié à $\mathrm{Var}(\hat{\theta}^N)$, décroît en $1/N$, **pour tout** u**!**

modèle paramétrique $\Rightarrow$ $\hat{\theta}^N$ a un effet global sur $\hat{y}_{\mathcal{D}}(u)$

7) Optimisation globale et krigeage

On veut maximiser $y(u)$, on observe (sans erreur) en des u_i que l'on choisit

Comment les choisir au mieux ?

Krigeage $\rightarrow EQM$ de la prédiction en u = $\rho^2_{\mathcal{D}}(u)$

7) Optimisation globale et krigeage

On veut maximiser $y(u)$, on observe (sans erreur) en des u_i que l'on choisit

Comment les choisir au mieux ?

Krigeage $\rightarrow EQM$ de la prédiction en u = $\rho^2_{\mathcal{D}}(u)$

Approche bayesienne, après observation de $\mathcal{D}_k = \{[u_1, y(u_1)], \ldots, [u_k, y(u_k)]\}$, $y(u)$ est distribué avec la densité $\varphi(y|\mathcal{D}_k, u)$ de la loi normale $\mathcal{N}(\hat{y}_{\mathcal{D}_k}(u), \rho^2_{\mathcal{D}_k}(u))$

7) Optimisation globale et krigeage

On veut maximiser $y(u)$, on observe (sans erreur) en des u_i que l'on choisit

Comment les choisir au mieux ?

Krigeage $\rightarrow EQM$ de la prédiction en u = $\rho^2_{\mathcal{D}}(u)$

Approche bayesienne, après observation de $\mathcal{D}_k = \{[u_1, y(u_1)], \ldots, [u_k, y(u_k)]\}$, $y(u)$ est distribué avec la densité $\varphi(y|\mathcal{D}_k, u)$ de la loi normale $\mathcal{N}(\hat{y}_{\mathcal{D}_k}(u), \rho^2_{\mathcal{D}_k}(u))$

Stratégie optimale (algorithme) $\rightarrow$ problème de Programmation Dynamique Stochastique...

...que l'on ne sait pas résoudre !

[Mockus *et al.* 1978, Mockus 1989, Schonlau *et al.* 1998, Jones *et al.* 1998] $\longrightarrow$ approximation à un pas, «expected improvement»

u_{k+1} maximizes $EI(u) = \int_{y_k^{\max}}^{\infty} [y - y_k^{\max}] \varphi(y|\mathcal{D}_k, u) dy$

avec $y_k^{\max}$ la valeur maximale des $y(u_i)$ observés

[Mockus *et al.* 1978, Mockus 1989, Schonlau *et al.* 1998, Jones *et al.* 1998] $\rightarrow$ approximation à un pas, «expected improvement»

u_{k+1} maximizes $EI(u) = \int_{y_k^{\max}}^{\infty} [y - y_k^{\max}] \varphi(y|\mathcal{D}_k, u) dy$

avec $y_k^{\max}$ la valeur maximale des $y(u_i)$ observés

- Converge vers l'optimum global de $y(u)$ (on va finir par observer partout)

[Mockus *et al.* 1978, Mockus 1989, Schonlau *et al.* 1998, Jones *et al.* 1998] $\rightarrow$ approximation à un pas, «expected improvement»

u_{k+1} maximizes $EI(u) = \int_{y_k^{\max}}^{\infty} [y - y_k^{\max}] \varphi(y|\mathcal{D}_k, u) dy$

avec $y_k^{\max}$ la valeur maximale des $y(u_i)$ observés

- Converge vers l'optimum global de $y(u)$ (on va finir par observer partout)

- Il faut résoudre un problème d'optimisation globale à chaque itération (mais sans calculer y en de nouveaux points)

[Mockus *et al.* 1978, Mockus 1989, Schonlau *et al.* 1998, Jones *et al.* 1998] $\longrightarrow$ approximation à un pas, «expected improvement»

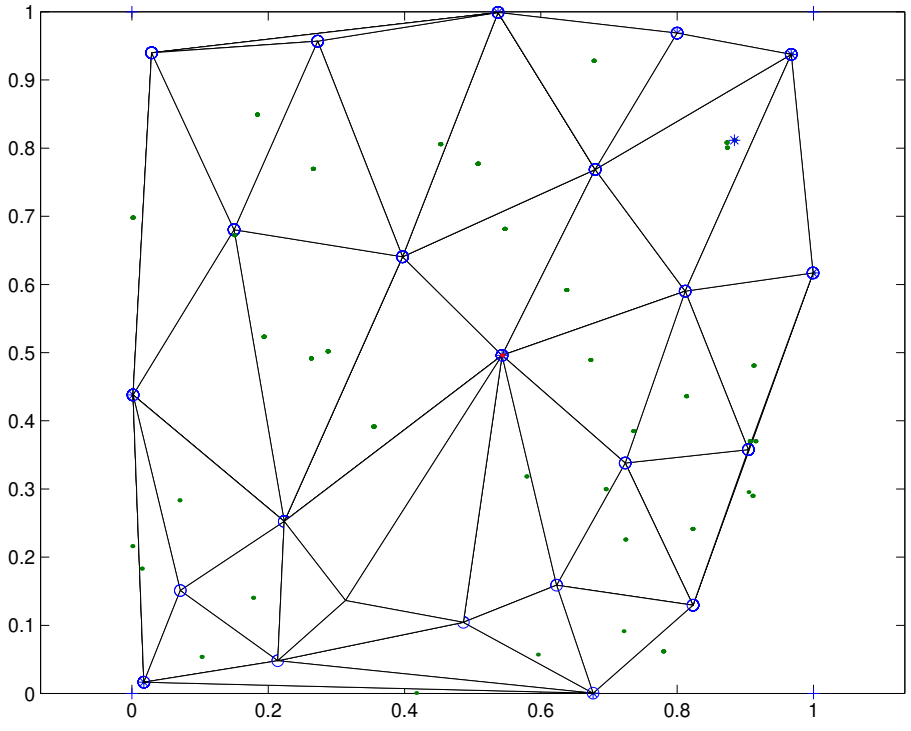
u_{k+1} maximizes $EI(u) = \int_{y_k^{\max}}^{\infty} [y - y_k^{\max}] \varphi(y|\mathcal{D}_k, u) dy$

avec $y_k^{\max}$ la valeur maximale des $y(u_i)$ observés

- Converge vers l'optimum global de $y(u)$ (on va finir par observer partout)

- Il faut résoudre un problème d'optimisation globale à chaque itération (mais sans calculer y en de nouveaux points)

- Les points «intéressants» sont loin de ceux déjà utilisés $\longrightarrow$ initialiser la recherche au centre des simplexes d'une triangulation de Delaunay [Bates & Pronzato 2001]



Possibilité d'utiliser une information sur les dérivées [Leary *et al.*, 2004]

References

- A.C. Atkinson and D.R. Cox. Planning experiments for discriminating between models (with discussion). *Journal of Royal Statistical Society*, B36:321–348, 1974.
- A.C. Atkinson and V.V. Fedorov. The design of experiments for discriminating between two rival models. *Biometrika*, 62(1):57–70, 1975.
- A.C. Atkinson and V.V. Fedorov. Optimal design: experiments for discriminating between several models. *Biometrika*, 62(2):289–303, 1975.
- P.L. Bartlett. Prediction algorithms: complexity, concentration and convexity. In *Prep. 13th IFAC Symposium on System Identification, Rotterdam*, pages 1507–1517, August 2003.
- R. Bates and L. Pronzato. Emulator-based global optimisation using lattices and Delaunay tesselation. In P. Prado and R. Bolado, editors, *Proc. 3rd Int. Symp. on Sensitivity Analysis of Model Output*, pages 189–192, Madrid, June 2001.
- G.E.P. Box and W.J. Hill. Discrimination among mechanistic models. *Technometrics*, 9(1):57–71, 1967.

- G.E.P. Box and K.B. Wilson. On the experimental attainment of optimum conditions (with discussion). *Journal of Royal Statistical Society*, B13(1):1–45, 1951.
- M.-Y. Cheng, P. Hall and M. Titterington. Optimal design for curve estimation by local linear smoothing. *Bernoulli*, 4(1):3–14, 1998.
- F. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of the AMS*, 39(1):1–49, 2001.
- D.Z. D'Argenio. Optimal sampling times for pharmacokinetic experiments. *Journal of Pharmacokinetics and Biopharmaceutics*, 9(6):739–756, 1981.
- R.A. Fisher. *Statistical Methods for Research Workers*. Oliver & Boyd, Edimbourgh, 1925.
- P.D.H. Hill. A review of experimental design procedures for regression model discrimination. *Technometrics*, 20:15–21, 1978.
- M.E. Johnson, L.M. Moore, and D. Ylvisaker. Minimax and maximin distance designs. *Journal of Statistical Planning and Inference*, 26:131–148, 1990.

- D. Jones, M. Schonlau and W.J. Welch. Efficient global optimization of expensive black-box functions. *J. Global Optimization*, 13:455–492, 1998.
- J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Math. Stat.*, 23:462–466, 1952.
- J. Kiefer and J. Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.
- S. Leary, A. Bhaskar and A.J. Keane. A derivative based surrogate model for approximating and optimizing the output of an expensive computer simulation. *J. Global Optimization*, 30:39–58, 2004.
- J. Mockus. *Bayesian Approach to Global Optimization, Theory and Applications*. Kluwer, Dordrecht, 1989.
- J. Mockus, V. Tiesis, and A. Zilinskas. The application of Bayesian methods for seeking the extremum. In L.C.W. Dixon and G.P. Szego, editors, *Towards Global Optimisation 2*, pages 117–129. North Holland, Amsterdam, 1978.
- M.D. Morris and T.J. Mitchell. Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, 43:381–402, 1995.

- E.A. Nadaraya. On estimating regression. *Theory of Probability and its Applications*, 9:141–142, 1964.
- A. Pázman and L. Pronzato. A Dirac function method for densities of nonlinear statistics and for marginal densities in nonlinear regression. *Statistics & Probability Letters*, 26:159–167, 1996.
- J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn. Design and analysis of computer experiments (with discussion). *Statistical Science*, 4(4):409–435, 1989.
- M. Schonlau, W.J. Welch, and D.R. Jones. Global versus local search in constrained optimization of computer models. In *New Developments and Applications in Experimental Design, Lecture Notes — Monograph Series, vol. 34*, pages 11–25. IMS, Hayward, 1998.
- M.C. Shewry and H.P. Wynn. Maximum entropy sampling. *Applied Statistics*, 14:165–170, 1987.
- M.L. Stein. *Interpolation of Spatial Data. Some Theory for Kriging*. Springer, Heidelberg, 1999.
- J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle. *Least Squares Support vector Machines*. World Scientific, New Jersey, 2002.

- E. Vasquez. *Modélisation comportementale de systèmes non-linéaires multivariables par méthodes à noyaux et applications*. Thèse de doctorat, université Paris XI Orsay, mai 2005.
- G.S. Watson. Smooth regression analysis. *Sankhya, Series A*, 26:359–372, 1964.
- H.P. Wynn. Maximum entropy sampling and general equivalence theory. In A. Di Bucchianico, H. Läuter, and H.P. Wynn, editors, *mODa'7 – Advances in Model–Oriented Design and Analysis, Proceedings of the 7th Int. Workshop, Heeze (Netehrlands)*, pages 211–218. Physica Verlag, Heidelberg, June 2004.