Partie 2 : plans orientés modèle

12-13 décembre 2017

Introduction

- Un plan d'expérience est généralement associé à un modèle (régression, krigeage)
- Principe : choisir les expériences pour maximiser la “qualité” du modèle
- Le modèle est choisi a priori
- Le plan va dépendre du modèle choisi

Cours de régression : données subies

Ici : on choisit nos données !

Plans optimaux pour la régression linéaire

Retour sur la régression (1/2)

Notations

- Base de fonctions : $f_1, \ldots, f_p$
- Plan d'expériences : $\mathbf{x}_1, \ldots, \mathbf{x}_n$
- $\mathbf{F} = \begin{bmatrix} f_1(\mathbf{x}_1) & \ldots & f_p(\mathbf{x}_1) \\ & \vdots & \\ f_1(\mathbf{x}_n) & \ldots & f_p(\mathbf{x}_n) \end{bmatrix} = \begin{bmatrix} \mathbf{f}(\mathbf{x}_1) \\ \vdots \\ \mathbf{f}(\mathbf{x}_n) \end{bmatrix} = \mathbf{f}(\mathbf{X})$

Hypothèse

$Y = \mathbf{f}(\mathbf{x}^*)\beta + \epsilon$ avec ϵ gaussien i.i.d $N(0, \sigma^2)$

Prédicteur en $\mathbf{x}^*$

$\hat{y} = \mathbf{f}(\mathbf{x}^*)\hat{\beta}$

La qualité de l'apprentissage des coefficients est liée à celle de la prédiction

Retour sur la régression (2/2)

Calcul de $\hat{\beta}$

$$\hat{\beta} = \left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{Y}$$

Propriétés de $\hat{\beta}$

- *Best linear unbiased estimate* : $\mathbb{E}(\hat{\beta}) = \beta$
- Covariance : $\left(cov(\hat{\beta}_i, \hat{\beta}_j)\right)_{i,j} = \sigma^2\left(\mathbf{F}^T\mathbf{F}\right)^{-1}$

Variance de prédiction

$$var(\hat{y}) = \sigma^2\mathbf{f}(\mathbf{x}^*)\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\mathbf{f}(\mathbf{x}^*)^T = \sigma^2 d(\mathbf{x}^*)$$

Critère d'optimalité : meilleure connaissance des coefficients

Optimisation d'un plan

Quantité critique : $\mathbf{F}^T\mathbf{F}$

- Pas de dépendance dans les observations
- σ^2 se factorise : pas de dépendance dans le niveau de bruit

Définition du problème d'optimisation

Choisir $\mathbf{x}_1, \ldots, \mathbf{x}_n$ tel que $\left(\mathbf{F}^T\mathbf{F}\right)^{-1}$ ait de "bonnes" propriétés

Exercice : optimisation d'un plan à un point

Modèle :

- Un seul degré de liberté
- $y(x) = \beta_0 x$

Expérience :

- un seul point x^1
- $x^1 \in [0, 1]$

Où placer l'observation de manière optimale ?

- pour minimiser l'erreur sur β_0
- pour minimiser la variance de prédiction

Critères d'optimalité basés sur $\mathbf{F}^T\mathbf{F}$

D-optimalité

- $\min \det \left(\mathbf{F}^T\mathbf{F}\right)^{-1} = \max \det \left(\mathbf{F}^T\mathbf{F}\right)$
- Minimiser le volume de l'ellipsoïde de confiance

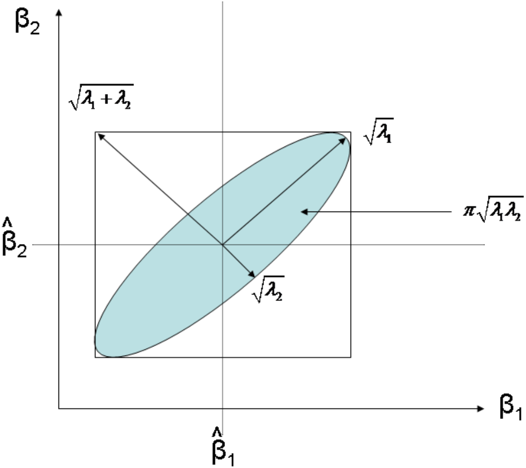
A-optimalité

- $\min tr \left[\left(\mathbf{F}^T\mathbf{F}\right)^{-1}\right]$
- Minimiser la somme des variances des coefficients

E-optimalité

- Minimiser la valeur propre maximale de $(\mathbf{F}^T\mathbf{F})^{-1}$:
- $\min \max_{1\leq i\leq n} \frac{1}{\lambda_i}$ (λ_i v.p. de $\mathbf{F}^T\mathbf{F}$)

Critères d'optimalité : interprétation graphique



Critères d'optimalité basés sur la variance de prédiction

G-optimalité

maximum de la variance de prédiction :

$$\min \max_{\mathbf{x} \in \mathbb{X}} \mathbf{f}(\mathbf{x}) \left(\mathbf{F}^T \mathbf{F}\right)^{-1} \mathbf{f}(\mathbf{x})^T$$

Nécessite *a priori* une boucle d'optimisation imbriquée

I-optimalité

intégrale de la variance de prédiction :

$$\min \int_{\mathbb{X}} \mathbf{f}(\mathbf{x}) \left(\mathbf{F}^T \mathbf{F}\right)^{-1} \mathbf{f}(\mathbf{x})^T d\mathbf{x}$$

Optimisation en pratique

Problème très complexe

- Nombre de variables : $n \times d$
- Problème très multimodal (invariances...)

En pratique : algorithmes d'échange

- On détermine le plus mauvais point du plan
- On cherche l'endroit du domaine le plus critique (par exemple, là où la variance est maximale)
- On supprime le mauvais point et on ajoute une observation au point critique

Pour aller plus loin : plans optimaux continus

- On peut faire $n >> p$ expériences
- Vaut-il mieux faire n expériences distinctes ou des répétitions ?

Nouvelle définition d'un plan d'expériences

Pour un plan ξ à n points :

$$\xi = \left\{ \begin{array}{ccc} \mathbf{x}_1, & \dots & , \mathbf{x}_n \\ k_1, & \dots & , k_n \end{array} \right\} \qquad \text{avec : } \sum k_i = N$$

Agrégation des répétitions

- Soit 2 observations $y^1(x_1)$ et $y^2(x_1)$, variance d'erreur σ^2
- On définit : $\bar{y}_1 = \frac{1}{2}(y^1 + y^2)$, variance d'erreur $\sigma^2/2$
- Pas d'information perdue pour la régression !

Le modèle de régression généralisée

On pose : $\mathbf{\Gamma} = \sigma^2 diag\left[k_1^{-1}, ..., k_n^{-1}\right]$
Moindres carrés généralisés :

$$\beta^* = \left(\mathbf{F}^T \mathbf{\Gamma}^{-1} \mathbf{F}\right)^{-1} \mathbf{F}^T \mathbf{\Gamma}^{-1} \mathbf{Y}$$

Passage au continu On pose $\omega_i = k_i/N$. Si $N >> n$, alors les ω_i sont presque continus. Plan continu normalisé :

$$\xi = \left\{ \begin{array}{ccc} \mathbf{x}_1, & \dots & , \mathbf{x}_n \\ \omega_1, & \dots & , \omega_n \end{array} \right\} \qquad \text{avec : } \sum \omega_i = 1$$

Plans continus normalisés

Intérêt de cette définition

- On peut comparer des plans à différents nombres de points
- Le plan est défini par une mesure de probabilité (discrète)
- On a : $M(\xi) = \mathbf{F}^T\Gamma^{-1}\mathbf{F} = \int_D \mathbf{f}(\mathbf{x})^T\mathbf{f}(\mathbf{x})d\xi(\mathbf{x})$

Résultats fondamentaux

- Il existe une mesure D-optimale fini de support de taille $n_0 = \frac{p(p+1)}{2} + 1$
- Théorème d'équivalence généralisé (TEG) : les plans D-optimaux et G-optimaux sont identiques
- Les valeurs optimales de D et G sont connues

Théorème d'équivalence de Kiefer et Wolfowitz (1960)

Les trois conditions sont équivalentes :

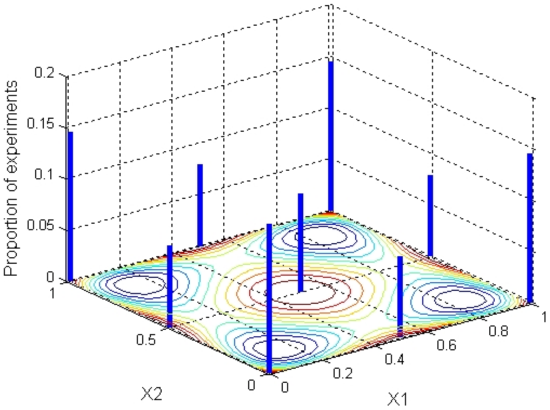
- Un plan est D-optimal
- Un plan est G-optimal
- $\max_{\mathbf{x}^*} \mathbf{f}(\mathbf{x}^*) \left(\mathbf{F}^T \mathbf{\Gamma}^{-1} \mathbf{F}\right)^{-1} \mathbf{f}(\mathbf{x}^*)^T = p$

Conséquences

- Maximiser le déterminant minimise la variance de prédiction maximale.
- On connait la plus petite valeur atteignable !

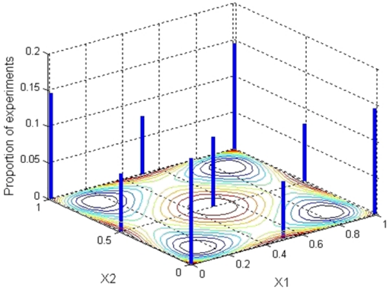
Un plan continu D-optimal (1/2)

- 14.6 % aux coins
- 8.0 % aux milieux des arêtes
- 9.6 % au centre

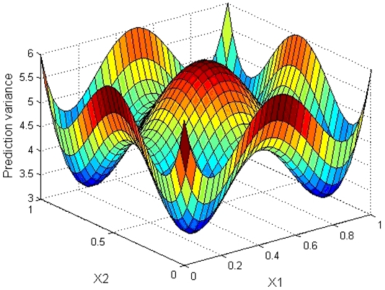


Un plan continu D-optimal (2/2)

Variance de prédiction du modèle :



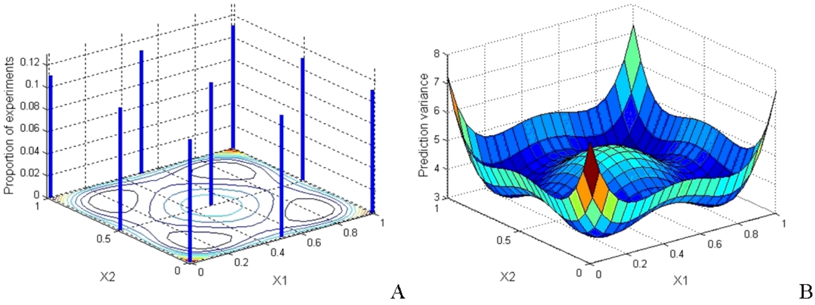
A



B

Le plan factoriel uniforme

Le plan n'est pas G-optimal :



Plans optimaux pour le krigeage

Notation et propriétés

Hypothèse fondamentale

$$Y(\mathbf{x}) \sim \mathcal{PG}\left(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right)$$

On note :

- $m(\mathbf{x}^*) = \mathbb{E}(Y(\mathbf{x}^*)|\mathbf{Y}_n)$: moyenne du krigeage
- $s^2(\mathbf{x}^*) = var(Y(\mathbf{x}^*)|\mathbf{Y}_n)$: variance de prédiction

Propriétés d'interpolation :

$$m(\mathbf{x}_i) = y_i$$

$$s^2(\mathbf{x}_i) = 0$$

Krigeage avec tendance (krigeage universel)

Tendance : $\mu(\mathbf{x}) = \sum_{k=1}^{p} \beta_k f_k(\mathbf{x})$

Estimation de β

Moindres carrés généralisés :

$$\beta^* = \left(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^T\mathbf{K}^{-1}\mathbf{Y}_n$$

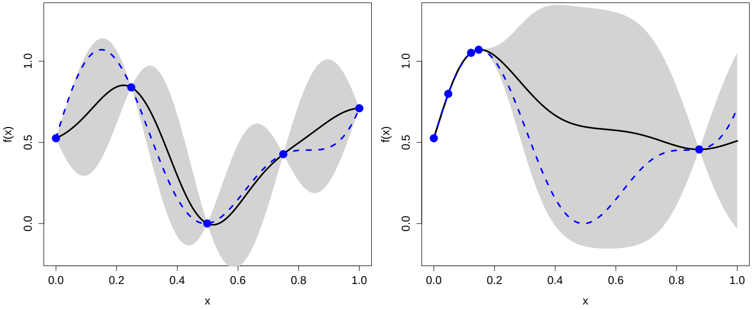
Equations :

$$\begin{aligned}
m_{UK}(\mathbf{x}^*) &= \mathbf{f}(\mathbf{x}^*)\beta^* + \mathbf{k}(\mathbf{x}^*)^T\mathbf{K}^{-1}\left(\mathbf{Y}_n - \mathbf{F}\beta^*\right) \\
s^2_{UK}(\mathbf{x}^*) &= \sigma^2 - \mathbf{k}(\mathbf{x}^*)^T\mathbf{K}^{-1}\mathbf{k}(\mathbf{x}^*) \\
&+ \left(\mathbf{f}(\mathbf{x}^*) - \mathbf{F}^T\mathbf{K}^{-1}\mathbf{k}(\mathbf{x}^*)\right)^T\left(\mathbf{F}^T\mathbf{K}^{-1}\mathbf{F}\right)^{-1}\left(\mathbf{f}(\mathbf{x}^*) - \mathbf{F}^T\mathbf{K}^{-1}\mathbf{k}(\mathbf{x}^*)\right)
\end{aligned}$$

NB : la variance ne dépend pas de la valeur des observations

Effet de la répartition des points

Les deux modèles ont la même covariance et le même nombre de points :



Plans optimaux pour l'apprentissage des paramètres ?

- TEG : les même plans optimisent l'apprentissage des paramètres et la prédiction !
- Pas d'équivalent pour le krigeage...

Plans optimaux pour la tendance (krigeage universel)

Equivalent à la régression avec résidus corrélés ⇒ D-optimalité
Problème : dépend de la covariance $\mathbf{K}$.

D-optimalité... pour la vraisemblance ?

$$l = \log\det \mathbf{K} + \mathbf{Y}^T \mathbf{K}^{-1} \mathbf{Y}$$

Type de plans : plus grande variété des interdistances possibles

Critères d'optimalité basés sur la variance de prédiction

Hypothèse forte : paramètres de covariance connus !

G-optimalité

maximum de la variance de prédiction :

$$\min \max_{\mathbf{x} \in \mathbb{X}} s_{UK}^2(\mathbf{x})$$

I-optimalité

intégrale de la variance de prédiction :

$$\min \int_{\mathbb{X}} s_{UK}^2(\mathbf{x}) d\mathbf{x}$$

Difficultés en pratique

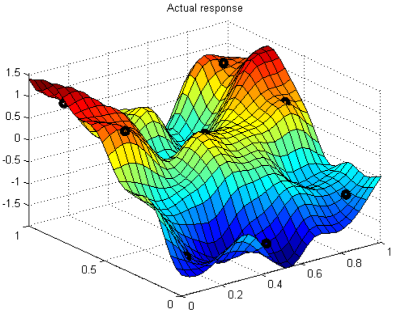
Problème insoluble...

- G-optimalité : nécessite une boucle d'optimisation interne
- I-optimalité : Nécessite un calcul d'intégrale numérique
- Deux critères très coûteux à évaluer !
- Problème global d'optimisation insoluble : grande dimension, multimodal, ...

Bonnes pratiques

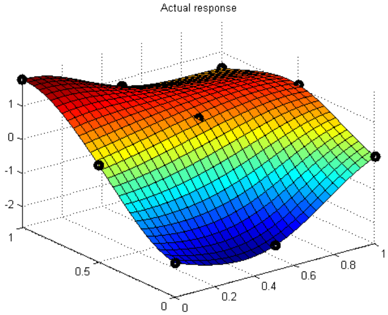
- Optimisation pour une famille de plans (factoriels, LHS)
- **Construction séquentielle**

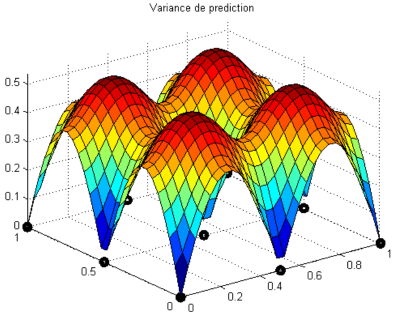
Exemple 2D : vraie fonction



Prédiction avec un plan factoriel à 3 niveaux

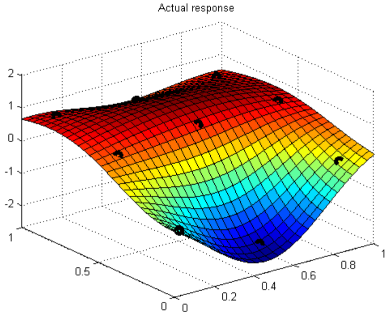
- IMSE $= 0.3647$
- maxMSE $= 0.51$

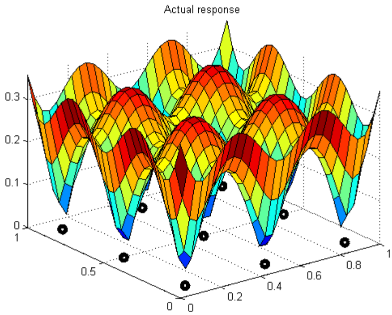




Prédiction avec un plan factoriel de côté 80%

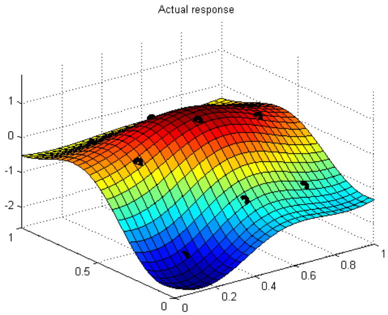
- IMSE = 0.2322
- maxMSE = 0.33

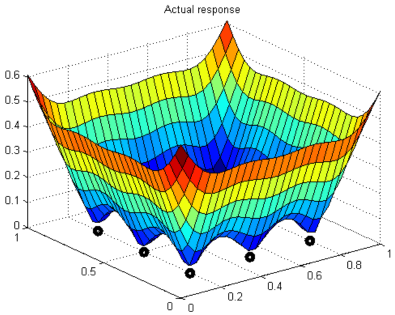




Prédiction avec un plan factoriel de côté 60%

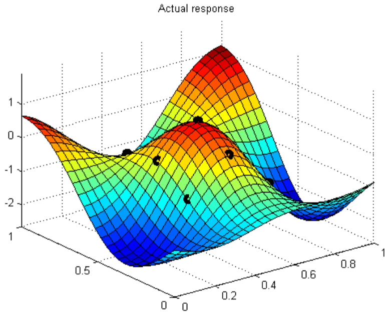
- IMSE = 0.2279
- maxMSE = 0.6

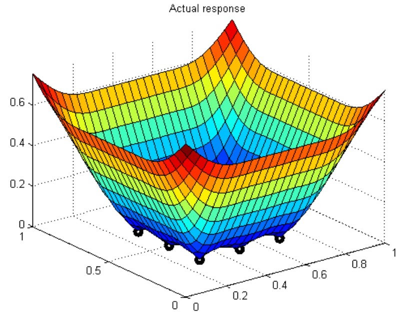




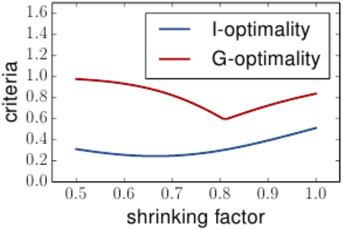
Prédiction avec un plan factoriel de côté 40%

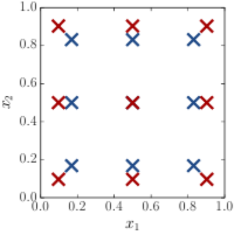
- IMSE $= 0.3005$
- maxMSE $= 0.75$





Contraction optimale selon les 2 critères





$\Rightarrow$ Les solutions optimales sont différentes !

Plans adaptatifs

Plans adaptatifs

Principe

- Remplacer un problème d'optimisation de dimension $n \times d$
- ... par n problèmes à d dimensions

Problème d'optimisation complet (C = maxMSE ou IMSE) :

$$\min_{\mathbf{x}_1,\ldots,\mathbf{x}_n} C(\mathbf{x}_1, \ldots, \mathbf{x}_n)$$

Problème d'optimisation adaptatif (C = maxMSE ou IMSE) :

Pour i allant de 1 (ou $1 \leq k \leq n$) à n :

$$\min_{\mathbf{x}_i} C(\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}, \mathbf{x}_i)$$

Planification adaptative en cherchant le maximum de variance

On choisit $C = s^2(\mathbf{x})$.

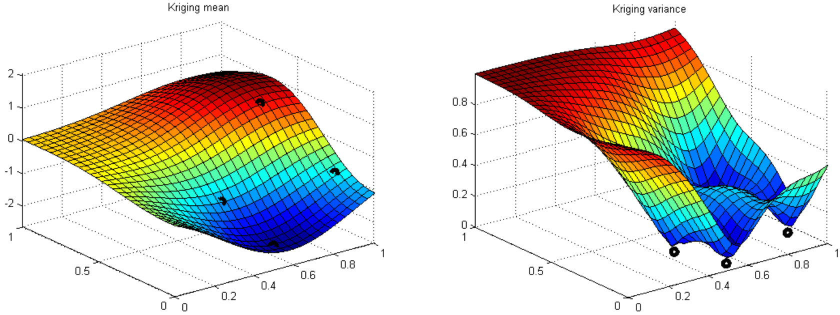
Algorithme

1. on construit un plan initial (taille k)
   Pour i allant de $k+1$ à n :
2. On cherche le point où la variance est maximum :
   $\mathbf{x}* = \arg\max_D s^2(\mathbf{x})$
3. On ajoute une nouvelle observation en ce point : $\mathbf{x}_i = \mathbf{x}*$
4. On répète les opérations 2 et 3

Exemple

Plan de départ : 4 points

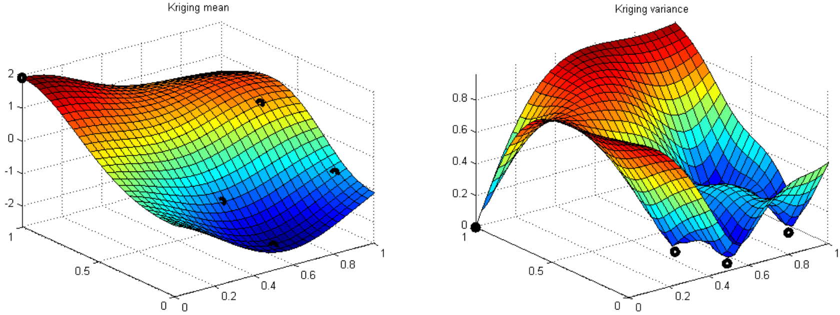
- IMSE = 0.5985
- maxMSE = 0.9991



Exemple

5 points

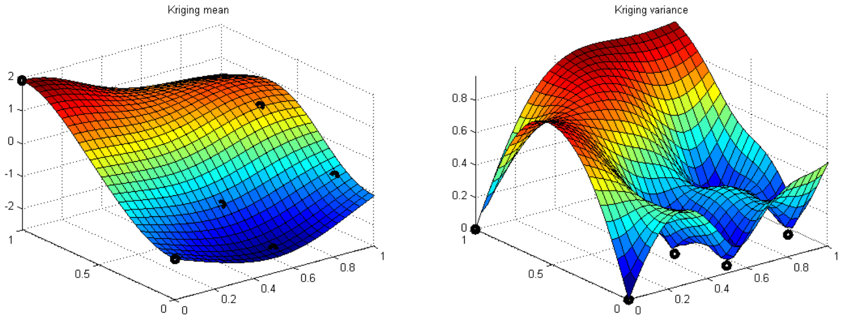
- IMSE = 0.5462
- maxMSE = 0.9665



Exemple

6 points

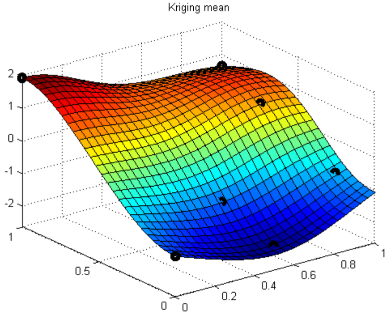
- IMSE = 0.5011
- maxMSE = 0.9466

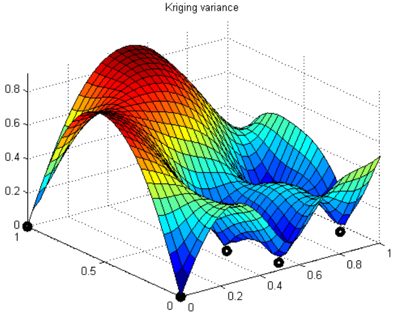


Exemple

7 points

- IMSE = 0.4619
- maxMSE = 0.9035

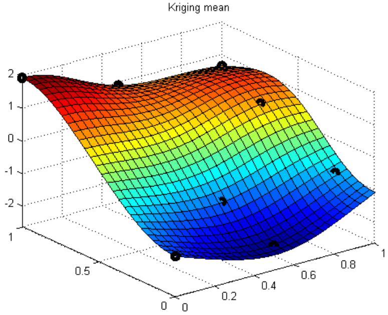


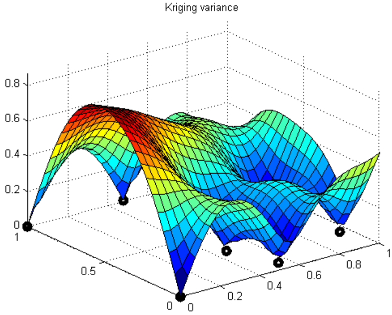


Exemple

8 points

- IMSE = 0.3885
- maxMSE = 0.8632

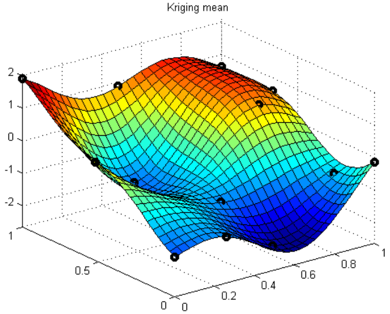


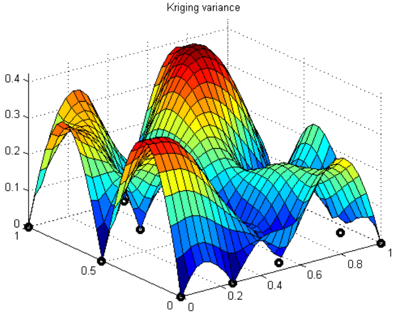


Exemple

14 points

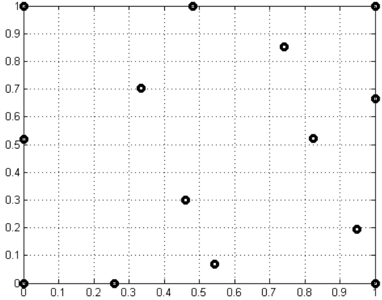
- IMSE = 0.2009
- maxMSE = 0.4226



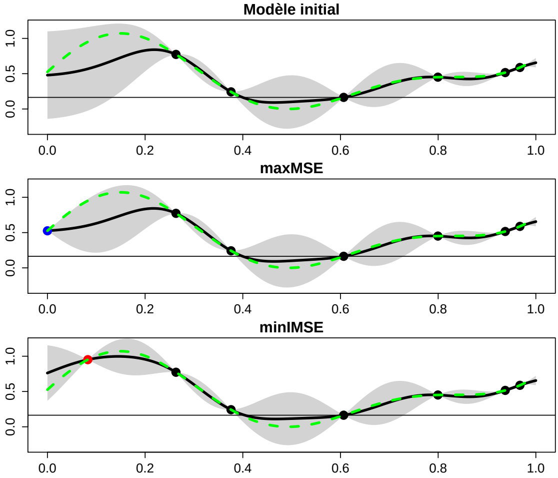


Exemple : plan final

- Bon remplissage d'espace
- Tendance à échantillonner sur les bords



Ajouter un point où la variance est maximale $\neq$ ajouter un point pour minimiser la variance maximale !



IMSE séquentiel

Problème d'optimisation complet :

$$\min C(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \int s_n^2(\mathbf{x}) d\mathbf{x}$$

Problème d'optimisation adaptatif :

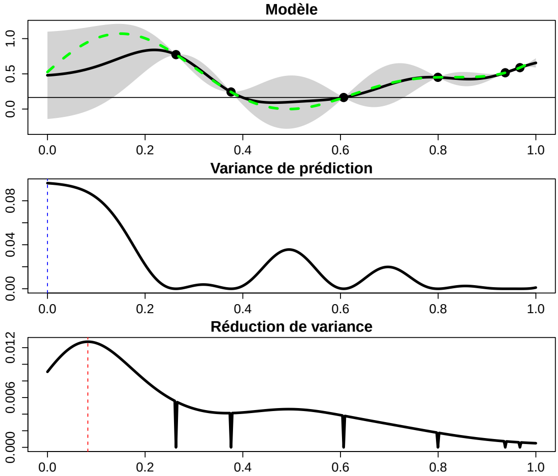
On a déjà trouvé $\mathbf{x}_1, \ldots, \mathbf{x}_{i-1}$. On cherche :

$$\min_{\mathbf{x}_i} \int s_i^2(\mathbf{x}) d\mathbf{x}$$

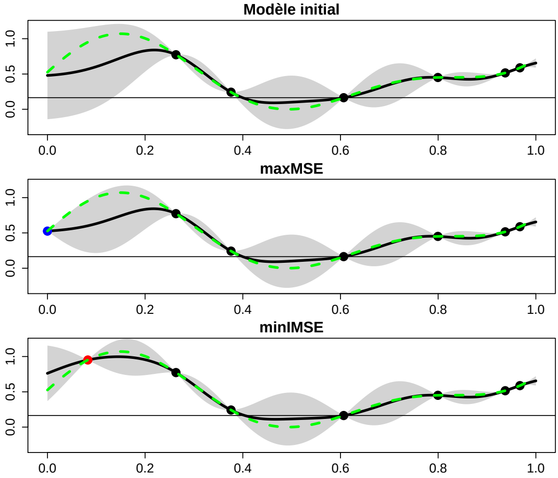
Nécessite de construire un nouveau krigeage pour chaque point candidat !

On a diminué la dimension du problème, mais pas son coût

IMSE séquentiel



maxMSE vs. IMSE séquentiel



Remarques finales

Plans adaptatifs et krigeage

Modèle “souple”, donc bien adapté à la séquentialité.
Intérêt plus discutable en régression

Avantage des plans adaptatifs

Prise en compte la valeur des réponses !

Applications

- Optimisation $\Rightarrow$ Cours de R. Le Riche
- Planification ciblée $\Rightarrow$ Cours de Y. Richet

Plans d'expériences : conclusion générale, perspectives

Retour sur la définition

Objectifs

- Organisation réfléchie des expériences
- Génération contrôlée des données

pour **maximiser l'information obtenue**, soit :

- Mesurer l'influence des variables d'entrée sur la réponse ;
- Permettre l'utilisation de modèles reproduisant la complexité du processus étudié ;
- Maximiser la qualité de l'inférence du modèle

Approches géométriques

Objectif 1 : mesure d'effets simple

Plans OAT, factoriels

Objectif 2 : remplissage d'espace

Différentes mesures de remplissage

- Discrépance
- Distortion
- Quadrature
- Distances :
  - distance minimale entre les points
  - distance maximale entre un point du domaine et un point du plan

Propriétés en projections

Approches orientées modèle

Régression : plans optimaux

Maximisation de l'information de Fisher $\Rightarrow$ matrice de covariance des coefficients de la régression

- D-optimalité : déterminant
- G-optimalité : variance de prédiction maximale

TEG : les plans D- et G-optimaux sont les mêmes !

Krigeage : utilisation de la variance de prédiction

- I-optimalité : variance de prédiction moyenne
- G-optimalité : variance de prédiction maximum

En pratique : remplissage d'espace et construction séquentielle