Statistique en grande dimension

Cristina Butucea

PhD Université Paris 6, Actuaire ISUP
Professeur à l'ENSAE

ASPROM - 12 Octobre, 2016

Overview

Thèmes principaux et liens avec les entreprises

Thèmes d'études et recherches

- Modélisation non et semi-paramétrique, en grande dimension
- Estimation de fonctions et tests non paramétriques
- Apprentissage
- Sélection de variables - réduction de la dimension !

Le monde industriel :

- Formation par apprentissage
- Thèses CIFRE (EDF, Safran)
- Cours de formation continue, collaborations, etc.

Étude statistique

Les étapes typiques de la 'chaine' statistique :

- data mining : recherche de phénomènes, liens, liens de cause à effets ; statistiques descriptives ;
- modélisation : un phénomène identifié est 'chiffré', p.ex. un modèle de régression explique la durée de survie d'un patient par son état initial et la posologie de ses traitements
- prévision : peut être descriptive (p.ex. classification ordinale de la difficulté de se garer) ou bien inférentielle (p.ex. intervalle à 95 % de la consommation d'électricité à tel moment de la semaine)

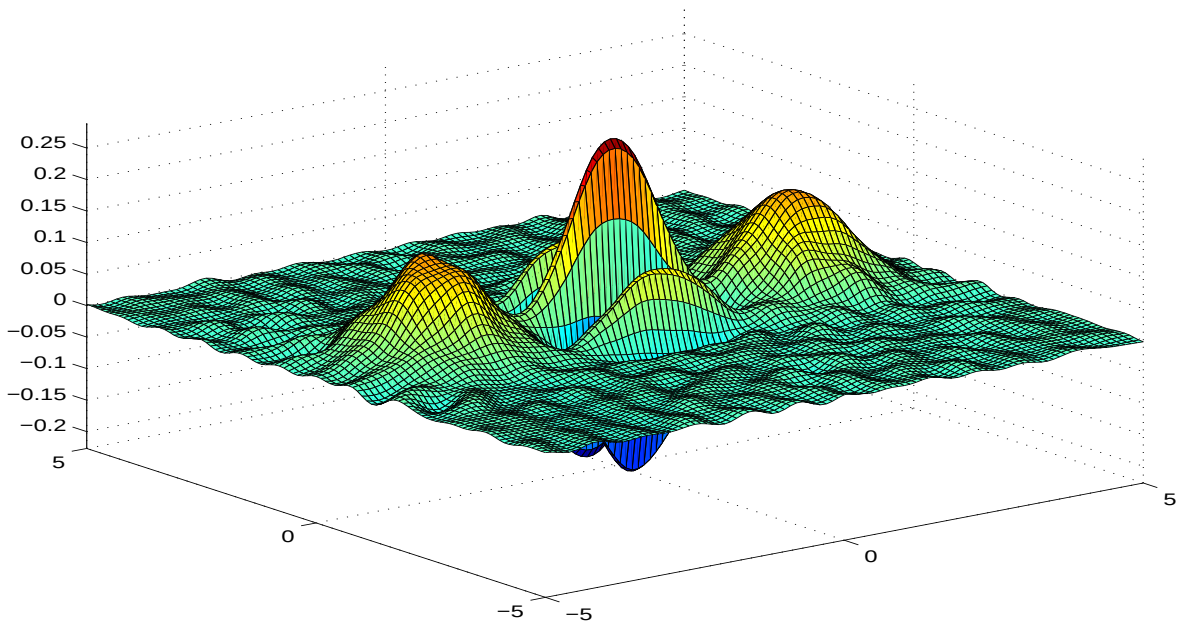
Techniques de lissages

Besoin de reconstituer une fonction, sans lui donner une forme prédéfinie : - choix d'une méthode de lissage,
donc **non paramétrique !**

- méthode des plus proches voisins - choix de la distance qui définit 'voisins'
- méthodes à noyau - choix de la 'fenêtre'
- méthodes de Fourier ou d'ondelettes - choix du nombre de coefficients à estimer

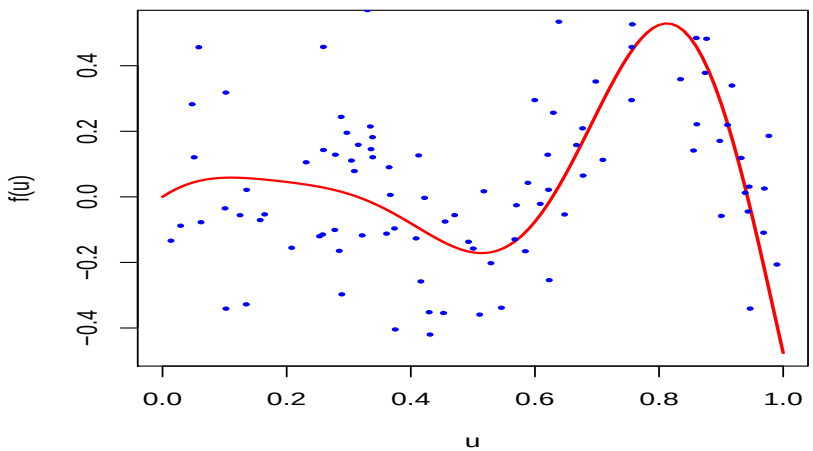
Estimation de l'état Chat de Schrödinger

Butucea, Guta, Artiles *Ann. Statistics, 2007*



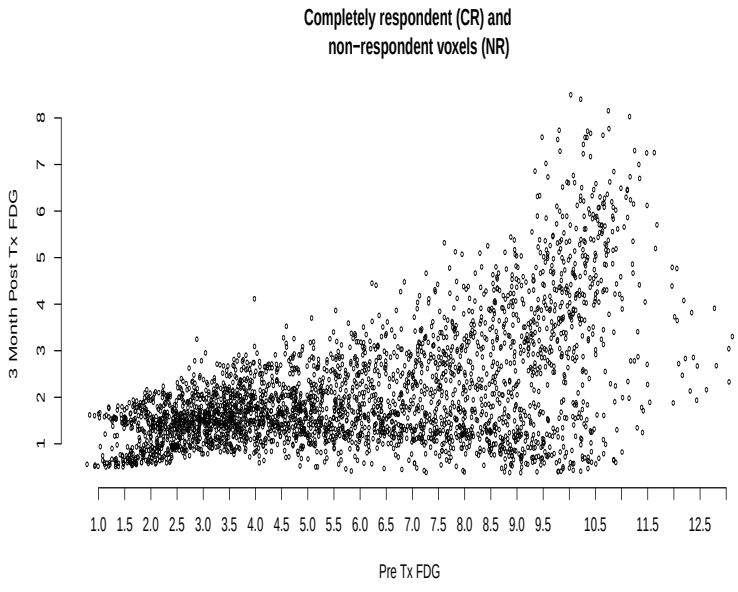
Régression non paramétrique

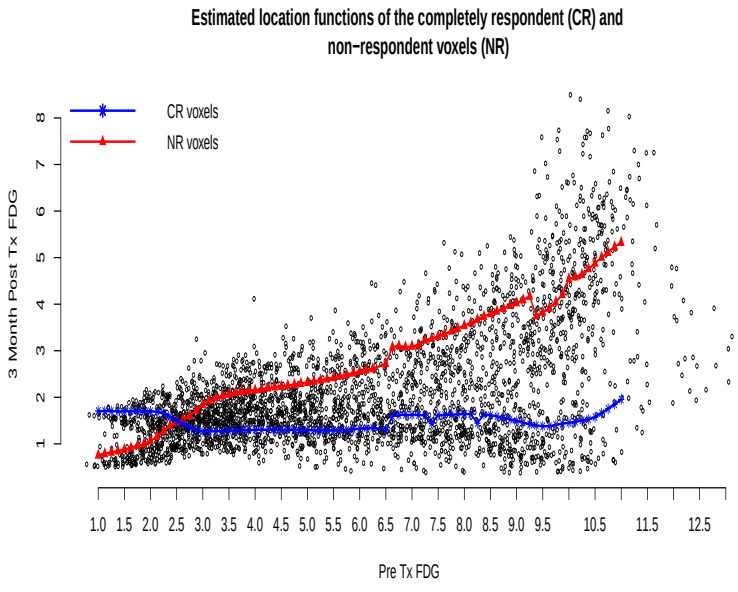
Observations $Y_i = f(X_i) + \xi_i$, i de 1 à $n = 100$.



Mélange de régressions

Butucea, Ngueyep Tzoumpe, Vandekerkhove, *Bernoulli, à paraitre, 2016*





Quand ne pas s'en servir

- trop peu de données !
  Un estimateur non paramétrique avec trop peu de données, ne lissera pas correctement

- trop de données !
  Les fonctions estimées ne doivent pas dépendre de trop de variables - *fléau de la dimension*

d	1	2	3	5	10
n	200	1 200	7 000	240 000	168 000 000

Table : Nombre d'observations nécessaires pour un risque MSE du même ordre, risque d'estimation d'une densité de probabilité aux propriétés comparables, quand la dimension d des observations augmente

Big data

Trop d'individus : n est trop grand, alors
réduction du nombre d'individus par échantillonnage

+ Théorie de sondages
- Explorer la totalité de la base

Trop de variables explicatives : d est trop grand, alors
réduction de la dimension

+ Choix étendu de méthodes
+ Possibilité d'appliquer les théories non paramétrique bien quantifiées
- Rarement une quantification a posteriori de la qualité du modèle, globalement sur la base de données toute entière.

Réduction de la dimension

Le nombre de variables d peut dépasser n dans certains cas ! Même une régression linéaire n'aura pas de solution.

On fait **hypothèse de structure**, suivie d'une **hypothèse de parcimonie** *sparsity* pour réduire la complexité du modèle ; par exemple, une structure additive

$$f(x_1, ..., x_d) = f_1(x_1) + ... + f_{j_1}(x_{j_1}) + ... + f_{j_s}(x_{j_s}) + ... + f_d(x_d),$$

de plus, il est souvent raisonable de supposer que toutes les variables n'interviennent pas de manière significative et d'en garder s parmi les d

$$f(x_1, ..., x_d) = f_{j_1}(x_1) + f_{j_2}(x_2) + ... + f_{j_s}(x_{j_s}).$$

Parcimonie

Si le modèle est parcimonieux, c-à-d si on peut garder s variables parmi les d :

$$f(x_1, ..., x_d) = f_{j_1}(x_1) + f_{j_2}(x_2) + ... + f_{j_s}(x_{j_s})$$

et si s est plus petit que le nombre d'individus n,
il existe plusieurs méthodes pour, à la fois :

1. identifier ou sélectionner les bonnes variables à garder
2. estimer le modèle

Nouveau : la sélection de variables !

Méthodes pour sélectionner les variables

- Minimisation d'un risque pénalisé
  -effectue simultanément la sélection et l'estimation du modèle
  -s'assure d'un bonne qualité de l'estimateur
  -limite le nombre de variables rajoutées dans le modèle

  Choix de la pénalité !
  Historiquement : C_p de Mallows, critère d'Akaike (AIC), BIC
  Plus récemment : Lasso, sélection de modèles, slope

  Bons modèles, bonnes propriétés, packages existants
  Contrôle moins bon sur les variables sélectionnées

Méthodes pour sélectionner les variables

- Seuillage des variables
  -l'estimation du modèle se fait dans un deuxième temps
  -implémentation très facile et efficace

Choix du seuil dit de bruit, au dessus duquel la variable est dite significative.

Butucea, Stepanova, 2015 ;
Butucea, Tsybakov, Stepanova, 2015

Sélection par seuillage

Si s connu :
On définit 'l'énergie' d'un signal $\int f_j^2$
et en estimant cet énergie (pas tout le signal)
on établit le seuil qui sépare le bruit d'un signal significatif.
Remarquer le lien avec le problème de test.

$$\begin{cases} H_0 : & f_j = 0 \\ H_1 : & f_j \neq 0 \text{ ou bien } \int f_j^2 \geq \text{seuil.} \end{cases}$$

Si s inconnu :
On compare les sélecteurs obtenus pour différentes valeurs de s et on en choisit un, en n'utilisant que les données d'origine.

Un sélecteur

$\hat{\eta} = \hat{\eta}(X_1, \ldots, X_n)$ est un estimateur à valeurs binaires 0 ou 1 :

$$\hat{\eta} = (\hat{\eta}_1, \ldots, \hat{\eta}_d), \quad \hat{\eta}_j \in \{0, 1\}.$$

Pour chaque coordonnées $\hat{\eta}_j$ vaut 1 si l'énergie estimée dépasse le seuil,
vaut 0, sinon.

Perte de Hamming d'un sélecteur compte le nombre de variables mal classées :

$$|\hat{\eta} - \eta| := \sum_{j=1}^{d} |\hat{\eta}_j - \eta_j|.$$

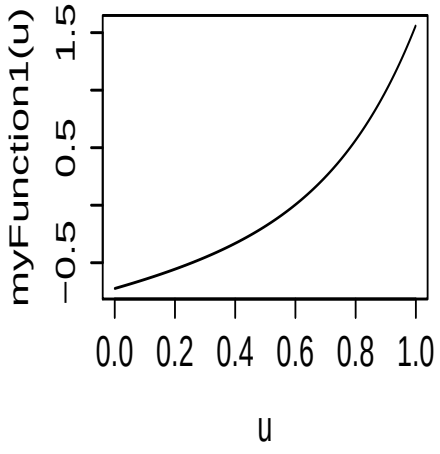
Exemple numérique

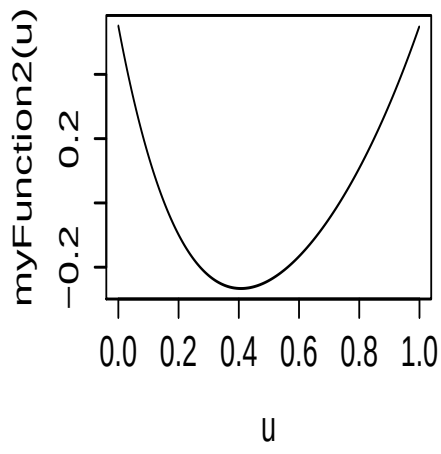
Supposons que 5 variables sont actives (non-nulles), définies sur $[0,1]$ par

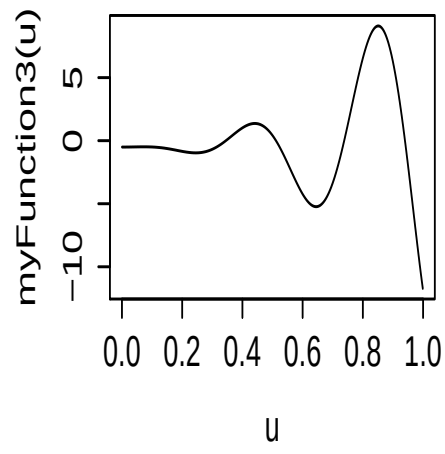
$$\begin{aligned} f_1(x) &= x^2 \cdot (2^{x-1} - (x-0.5)^2) \cdot e^x - 0.4741, \\ f_2(x) &= x^2 \cdot (2^{x-1} - (x-1)^5) - 0.4494, \\ f_3(x) &= 15x^2 2^{x-1} \cos(15x) - 0.0338, \\ f_4(x) &= x + 1/2, \\ f_5(x) &= 5 \cdot (x-0.7)^3 + 0.29, \end{aligned}$$

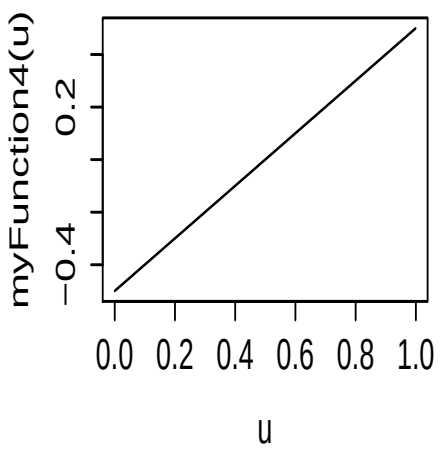
où $\int_0^1 f_j(x)dx = 0$ pour tous les $j = 1, \ldots, 5$.
La dimension d vaut 50, 100, 500, $1\,000$, $5\,000$, $10\,000$, $50\,000$.
On observe l'équivalent de $n = 10\,000$ observations.

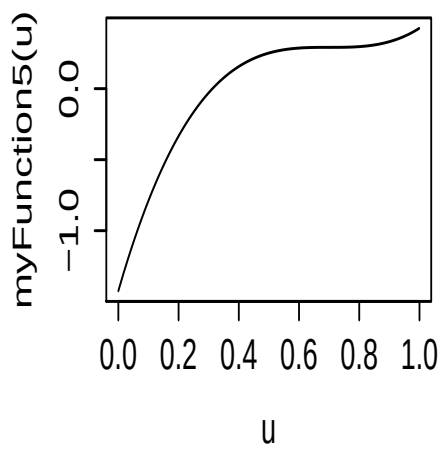
Représentation graphique :











s/d	0.1	0.05	0.01	$5 \cdot 10^{-3}$
$err(\hat{\eta})$	0.1272	0.1420	0.1616	0.1856

s/d	10^{-3}	$5 \cdot 10^{-4}$	10^{-4}
$err(\hat{\eta})$	0.2156	0.2248	0.2340

Table : Erreur estimée quand s/d tend vers zéro, $s = 5$.

La Table 1 indique une erreur qui augmente, mais très lentement, qui reste remarquablement stable quand d augmente, pour de très petites valeurs de s/d.

$s/d = 5/d$	0.1	0.05	0.01	$5 \cdot 10^{-3}$
$\sum_{k=1}^{K} \mathbb{I}(s_{\hat{m}}^{(k)} \leq s)/K$	0.886	0.896	0.918	0.840

$s/d = 5/d$	10^{-3}	$5 \cdot 10^{-4}$	10^{-4}
$\sum_{k=1}^{K} \mathbb{I}(s_{\hat{m}}^{(k)} \leq s)/K$	0.838	0.904	0.866

Table : Estimated probability of underestimating s for almost full selection as s/d tends to zero, $s = 5$.

On modifie $f_{5,l}(x) = l((x-0.7)^3 + 0.058)$, avec l dans le tableau :

l	0.01	0.5	1	2
$\|f_{5,l}\|_2$	0.0009	0.0460	0.0920	0.1841
$err(\widehat{\eta}(s_{\widehat{m}})), d = 1000$	0.3856	0.3684	0.3892	0.3924
$err(\widehat{\eta}(s_{\widehat{m}})), d = 5000$	0.3992	0.3948	0.4092	0.3916

l	3	4	5	6	7
$\|f_{5,l}\|_2$	0.2761	0.3681	0.4601	0.5522	0.6442
$err(\widehat{\eta}(s_{\widehat{m}})), d = 1000$	0.3620	0.1844	0.1848	0.1856	0.1684
$err(\widehat{\eta}(s_{\widehat{m}})), d = 5000$	0.3976	0.2064	0.2044	0.1992	0.1948

Table : Estimated error of the adaptive almost full selector for different choices of l, $s = 5$.

$s/d = 10/d$	0.2	0.1	0.02	0.01
$err(\hat{\eta}(s_{\hat{m}}))$	0.1320	0.1456	0.1562	0.1890

$s/d = 10/d$	$2 \cdot 10^{-3}$	10^{-3}	$2 \cdot 10^{-4}$
$err(\hat{\eta}(s_{\hat{m}}))$	0.2054	0.2254	0.2298

Table : Estimated error for almost full selection as s/d tends to zero, $s = 10$.

l	0.01	0.5	1	2
$err(\widehat{\eta}(s_{\widehat{m}})), d = 1000$	0.2850	0.2690	0.2798	0.2772
$err(\widehat{\eta}(s_{\widehat{m}})), d = 5000$	0.2946	0.2942	0.3038	0.2932

l	3	4	5	6	7
$err(\widehat{\eta}(s_{\widehat{m}})), d = 1000$	0.2626	0.1798	0.1768	0.1850	0.1690
$err(\widehat{\eta}(s_{\widehat{m}})), d = 5000$	0.2960	0.1974	0.2024	0.1946	0.1942

Table : Estimated error of the adaptive almost full selector for different choices of l, $s = 10$.