Estimation de Densité en Dimension Elevée et Classification de Courbes

Laurent Rouvière

Equipe de Probabilités et Statistique (Université Montpellier II)

18 novembre 2005

Plan de l'exposé

Le contexte

- Soient $X_1, \ldots, X_n$ (i.i.d.) issus d'une densité f (inconnue) sur $\mathbb{R}^d$.
- Problème : Trouver une fonction $f_n(x) = f_n(x; X_1, \ldots, X_n)$ qui soit "proche" de f.
- Critère L_1

$$\|f - f_n\|_1 = \int |f - f_n|$$

  → calculable;
  → interprétable "graphiquement";
  → interprétable en terme de probabilités.

Scheffé

$$\int |f - f_n| = 2 \sup_{B \in \mathcal{B}} \left| \int_B f - \int_B f_n \right| .$$

Les histogrammes modifiés

- Soit ℓ un entier et soit $h = 1/\ell$;
- Soit g une densité connue associée à la loi de probabilité ν_g;
- Soit $P = \{A_1, \ldots, A_\ell\}$ une partition de $\mathbb{R}^d$ telle que $\nu_g(A_i) = h$;

$$
\begin{aligned}
f_n(x) &= \left[(1-a_n)\frac{\mu_n(A(x))}{h} + a_n\right] g(x) \\
&= \frac{n\mu_n(A(x)) + 1}{nh+1}\, g(x)
\end{aligned}
$$

où $A(x) = A_i$ si $x \in A_i$.

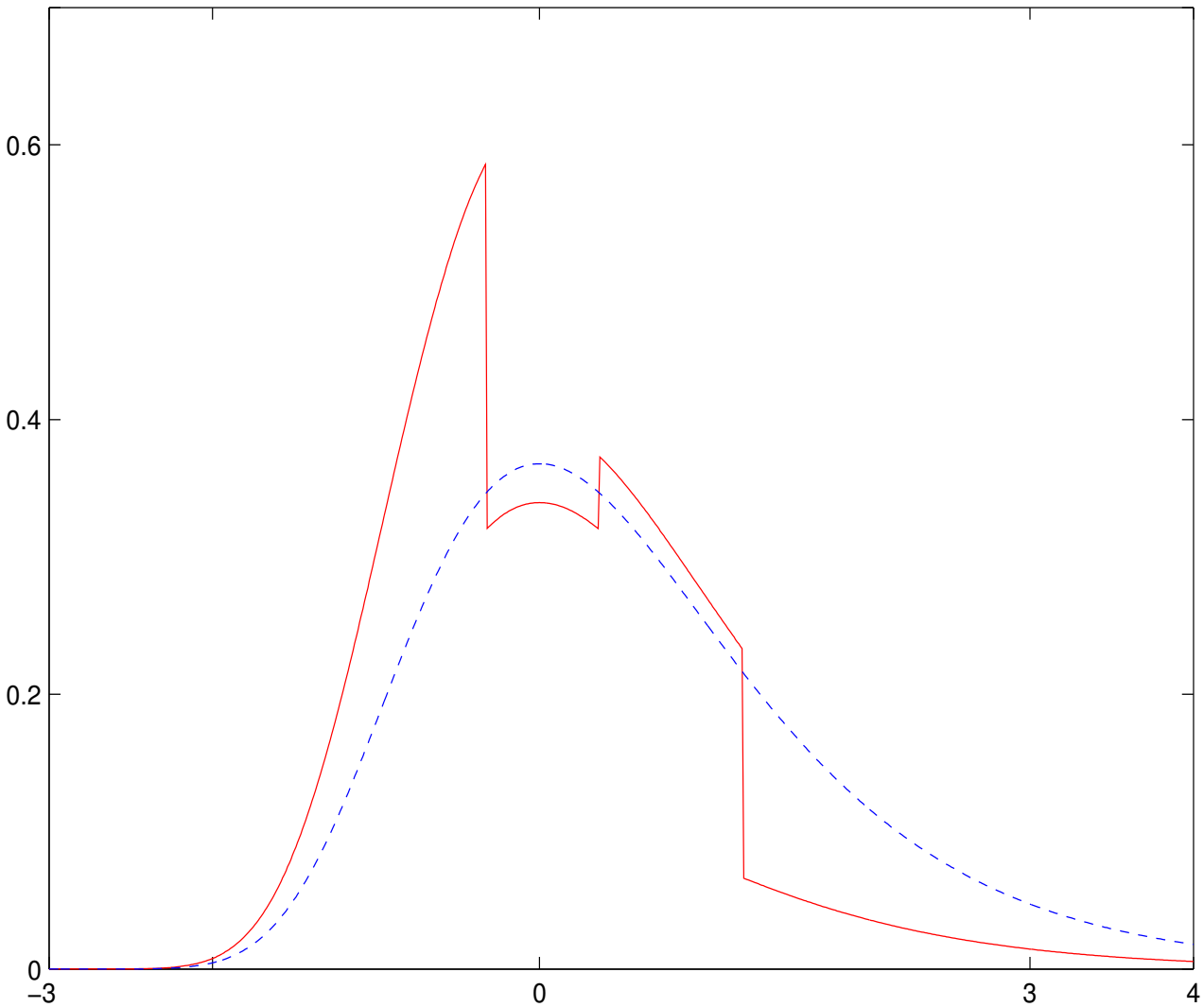
Les histogrammes modifiés

- Soit ℓ un entier et soit $h = 1/\ell$;
- Soit g une densité connue associée à la loi de probabilité ν_g;
- Soit $P = \{A_1, \ldots, A_\ell\}$ une partition de $\mathbb{R}^d$ telle que $\nu_g(A_i) = h$;

$$
\begin{aligned}
f_n(x) &= \left[(1-a_n)\frac{\mu_n\big(A(x)\big)}{h} + a_n\right] g(x) \\
&= \frac{n\mu_n\big(A(x)\big)+1}{nh+1}\, g(x)
\end{aligned}
$$

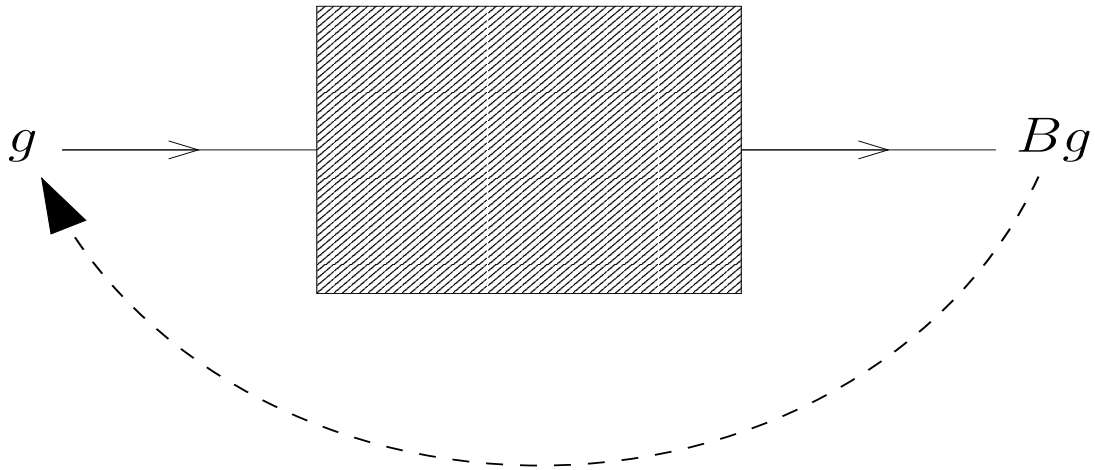
où $A(x) = A_i$ si $x \in A_i$.

Un exemple



Un sytème dynamique

Fixons l'échantillon $X_1, \ldots, X_n$ et le nombre de classes ℓ.



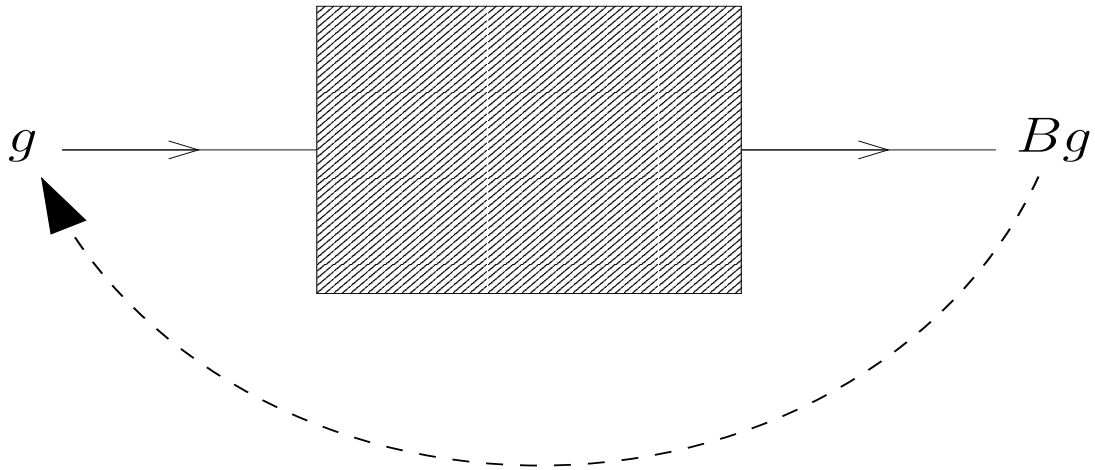
$\longrightarrow$ suite ou trajectoire d'estimateurs de f : $\{B^p g\}_{p \geq 0}$.

Théorème (Berlinet et Biau, 2004)

Si ℓ divise n alors la suite de densités $\{B_\ell^p g\}_{p \geq 0}$ est presque sûrement stationnaire.

Un sytème dynamique

Fixons l'échantillon $X_1, \ldots, X_n$ et le nombre de classes ℓ.



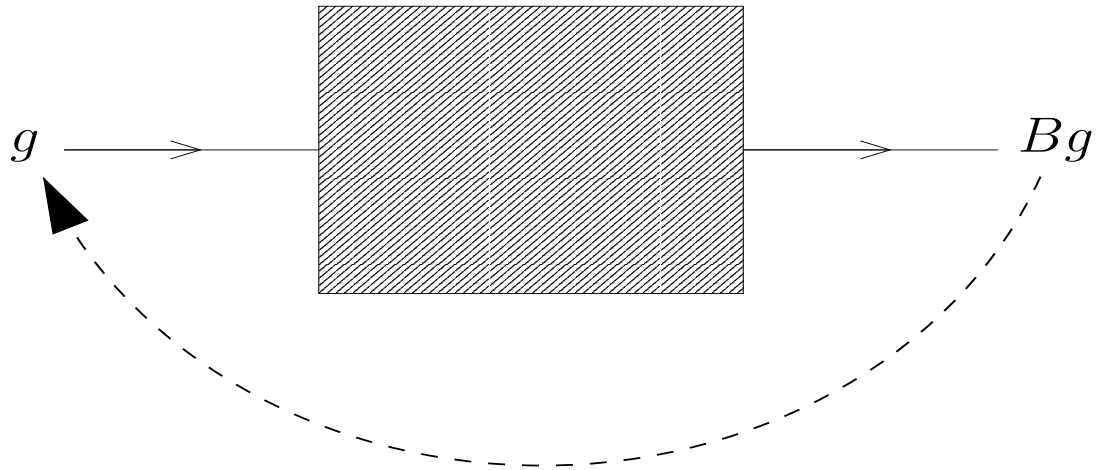
$\longrightarrow$ suite ou trajectoire d'estimateurs de f : $\{B^p g\}_{p \geq 0}$.

Théorème (Berlinet et Biau, 2004)

Si ℓ divise n alors la suite de densités $\{B_\ell^p g\}_{p \geq 0}$ est presque sûrement stationnaire.

Un sytème dynamique

Fixons l'échantillon $X_1, \ldots, X_n$ et le nombre de classes ℓ.



$\longrightarrow$ suite ou trajectoire d'estimateurs de f : $\{B^p g\}_{p \geq 0}$.

Théorème (Berlinet et Biau, 2004)

Si ℓ divise n alors la suite de densités $\{B_\ell^p g\}_{p \geq 0}$ est presque sûrement stationnaire.

Une application

- La densité de référence est un estimateur à noyau : $g = g_{n,h}$.
- La fenêtre est sélectionnée de deux manières différentes :
  1. Méthode Plug-in L_2.
  2. Méthode du double noyau.

Objectif

- Choisir un estimateur f_n dans la famille d'estimateurs "itérés";
- Comparer l'estimateur sélectionné à l'estimateur à noyau initial :

$$\|f - f_n\|_1 < \|f - g_{n,h}\|_1 \ ?$$

Une application

- La densité de référence est un estimateur à noyau : $g = g_{n,h}$.
- La fenêtre est sélectionnée de deux manières différentes :
  1. Méthode Plug-in L_2.
  2. Méthode du double noyau.

Objectif

- Choisir un estimateur f_n dans la famille d'estimateurs "itérés";
- Comparer l'estimateur sélectionné à l'estimateur à noyau initial :

$$\|f - f_n\|_1 < \|f - g_{n,h}\|_1 \ ?$$

Une application

- La densité de référence est un estimateur à noyau : $g = g_{n,h}$.
- La fenêtre est sélectionnée de deux manières différentes :
  1. Méthode Plug-in L_2.
  2. Méthode du double noyau.

Objectif

- Choisir un estimateur f_n dans la famille d'estimateurs "itérés";
- Comparer l'estimateur sélectionné à l'estimateur à noyau initial :

$$\|f - f_n\|_1 < \|f - g_{n,h}\|_1 \ ?$$

Les densités tests

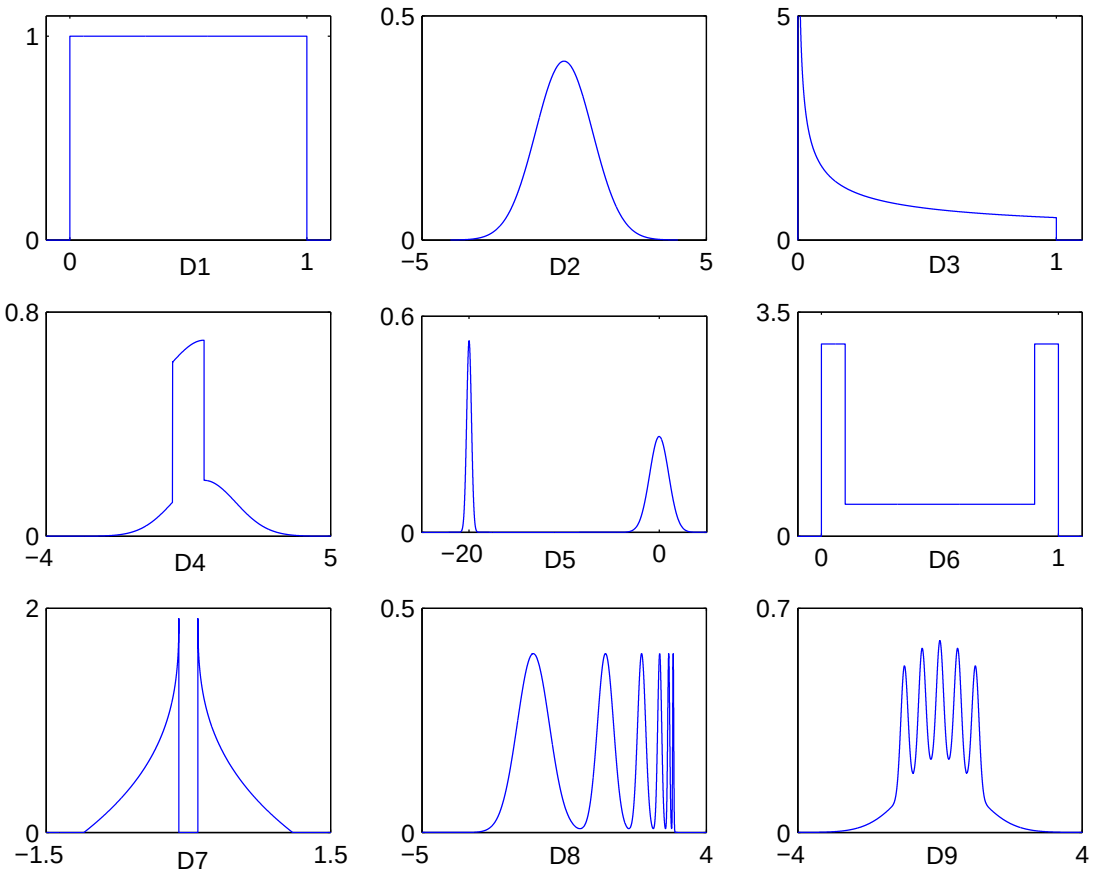
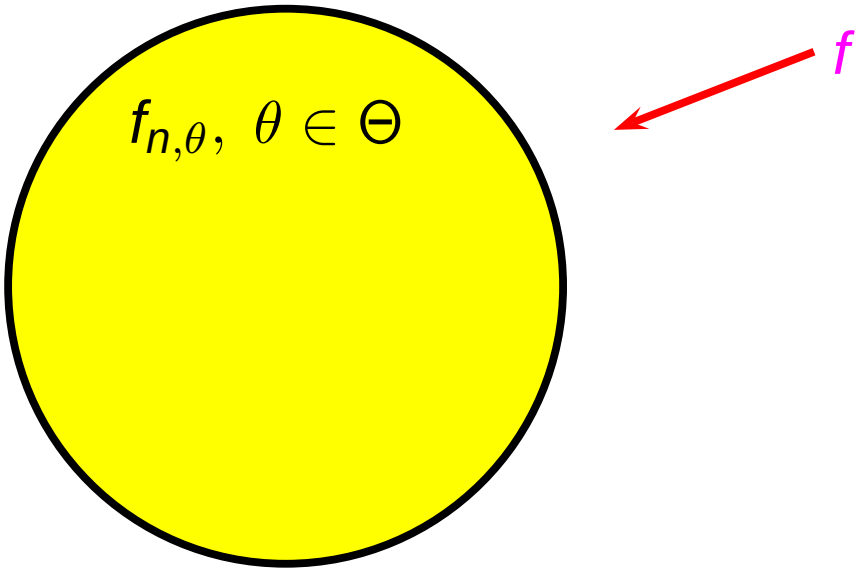


Figure: Les 9 densités tests.

Résultats

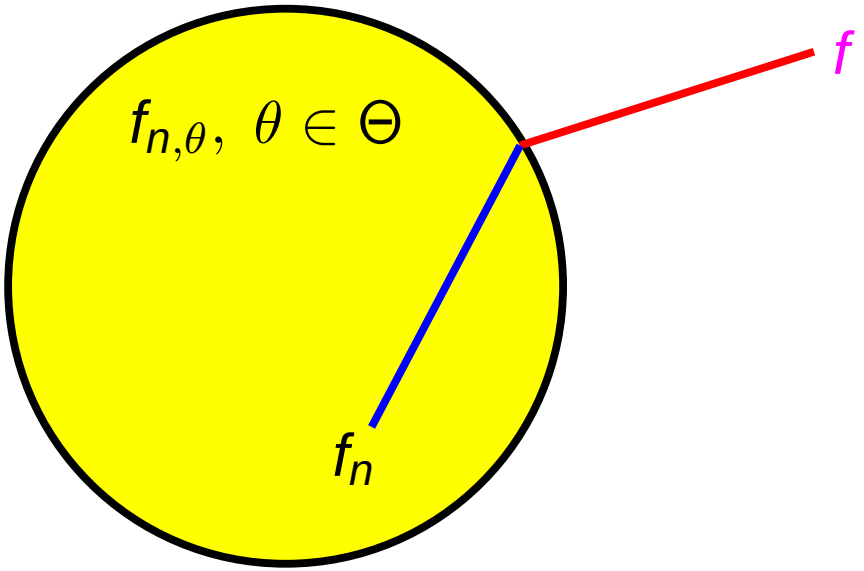
f	$L_1(g_{n,h_{pi}})$	$L_1(f_n)$	**WIN**	$L_1(g_{n,h_{dn}})$	$L_1(f_n)$	**WIN**
D1	0.18	0.20	●	0.21	0.21	●
D2	0.07	0.08	●	0.17	0.14	●
D3	0.37	0.30	●	0.34	0.28	●
D4	0.22	0.20	●	0.22	0.19	●
D5	1.20	0.35	●	0.46	0.32	●
D6	0.75	0.30	●	0.36	0.31	●
D7	0.31	0.27	●	0.40	0.25	●
D8	0.57	0.37	●	0.28	0.27	●
D9	0.34	0.27	●	0.46	0.26	●

Choisir une densité



Etant donné un échantillon aléatoire $X_1, \ldots, X_n$ issu de f, trouver le meilleur $f_{n,\theta}$, $\theta \in \Theta$.

Choisir une densité



Etant donné un échantillon aléatoire $X_1, \ldots, X_n$ issu de f, trouver le meilleur $f_{n,\theta}$, $\theta \in \Theta$.

$$\mathbf{E}\left\{ \int |f_n - f| \right\} \leq C\big(1 + \Sigma_1(n)\big) \inf_{\theta \in \Theta} \mathbf{E}\left\{ \int |f_{n,\theta} - f| \right\} + \Sigma_2(n).$$

La méthode combinatoire (1)

- Contexte général : $\mathcal{F}$=classe d'estimateurs $f_{n,\theta}$, $\theta \in \Theta$.
- Exemples : fenêtre de l'estimateur à noyau, pas de l'histogramme...
- "Data splitting" : Soit $m < n$,

$$\underbrace{X_1, \ldots, X_{n-m}}_{\downarrow} \qquad \underbrace{X_{n-m+1}, \ldots, X_n}_{\downarrow}$$

construction de l'estimateur : $f_{n-m,\theta}$

échantillon de validation

La méthode combinatoire (2)

- Critère de sélection pour une densité $f_{n-m,\theta} \in \mathcal{F}$:

$$\Delta_\theta = \sup_{A \in \mathcal{A}_\Theta} \left| \int_A f_{n-m,\theta} - \mu_m(A) \right| .$$

- Classe de Yatracos :

$$\mathcal{A}_\Theta = \left\{ \{x : f_{n-m,\theta}(x) > f_{n-m,\theta'}(x)\} : (\theta, \theta') \in \Theta^2 \right\} .$$

- L'estimateur de la distance minimum :

$$\Delta_{\theta^*} < \inf_{\theta \in \Theta} \Delta_\theta + \frac{1}{n} .$$

La méthode combinatoire (2)

- Critère de sélection pour une densité $f_{n-m,\theta} \in \mathcal{F}$:

$$\Delta_\theta = \sup_{A \in \mathcal{A}_\Theta} \left| \int_A f_{n-m,\theta} - \mu_m(A) \right| .$$

- Classe de Yatracos :

$$\mathcal{A}_\Theta = \left\{ \{x : f_{n-m,\theta}(x) > f_{n-m,\theta'}(x)\} : (\theta, \theta') \in \Theta^2 \right\} .$$

- L'estimateur de la distance minimum :

$$\Delta_{\theta^*} < \inf_{\theta \in \Theta} \Delta_\theta + \frac{1}{n} .$$

La méthode combinatoire (2)

- Critère de sélection pour une densité $f_{n-m,\theta} \in \mathcal{F}$:

$$\Delta_\theta = \sup_{A \in \mathcal{A}_\Theta} \left| \int_A f_{n-m,\theta} - \mu_m(A) \right| .$$

- Classe de Yatracos :

$$\mathcal{A}_\Theta = \left\{ \{x : f_{n-m,\theta}(x) > f_{n-m,\theta'}(x)\} : (\theta, \theta') \in \Theta^2 \right\} .$$

- L'estimateur de la distance minimum :

$$\Delta_{\theta^*} < \inf_{\theta \in \Theta} \Delta_\theta + \frac{1}{n} .$$

Théorème (Devroye et Lugosi, 2001)

Soit f_n l'estimateur de la distance minimum. On a :

$$\mathbf{E}\left\{\int |f_n - f|\right\} \leq 3 \inf_{\theta\in\Theta} \mathbf{E}\left\{\int |f_{n-m,\theta} - f|\right\} + 4\mathbf{E}\Delta + \frac{3}{n},$$

où

$$\Delta = \sup_{A\in\mathcal{A}_\Theta} \left| \int_A f - \mu_m(A) \right|.$$

Corollaire

$$\mathbf{E}\left\{\int |f_n - f|\right\} \leq 3 \inf_{\theta\in\Theta} \mathbf{E}\left\{\int |f_{n-m,\theta} - f|\right\} + 8\,\mathbf{E}\left\{\sqrt{\frac{\log 2\, \mathbf{S}_{\mathcal{A}_\Theta}(m)}{m}}\right\} + \frac{3}{n}.$$

où

$$\mathbf{S}_{\mathcal{A}_\Theta}(m) = \max_{x_1,\ldots,x_m\in\mathbb{R}^d} Card\left\{\{x_1,\ldots,x_m\} \cap A : A \in \mathcal{A}_\Theta\right\}.$$

Théorème (Devroye et Lugosi, 2001)

Soit f_n l'estimateur de la distance minimum. On a :

$$\mathbf{E}\left\{\int |f_n - f|\right\} \leq 3 \inf_{\theta\in\Theta} \mathbf{E}\left\{\int |f_{n-m,\theta} - f|\right\} + 4\mathbf{E}\Delta + \frac{3}{n},$$

où

$$\Delta = \sup_{A\in\mathcal{A}_\Theta} \left| \int_A f - \mu_m(A) \right|.$$

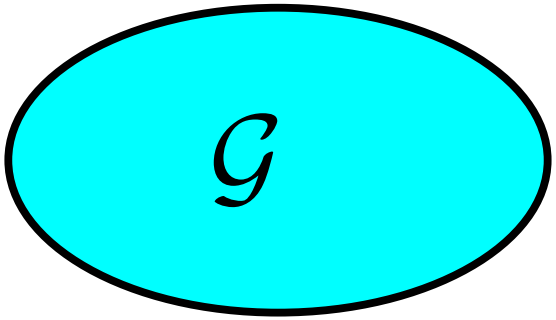
Corollaire

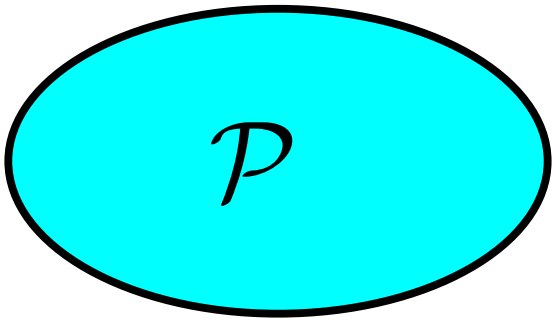
$$\mathbf{E}\left\{\int |f_n - f|\right\} \leq 3 \inf_{\theta\in\Theta} \mathbf{E}\left\{\int |f_{n-m,\theta} - f|\right\} + 8\,\mathbf{E}\left\{\sqrt{\frac{\log 2\,\mathbf{S}_{\mathcal{A}_\Theta}(m)}{m}}\right\} + \frac{3}{n}.$$

où

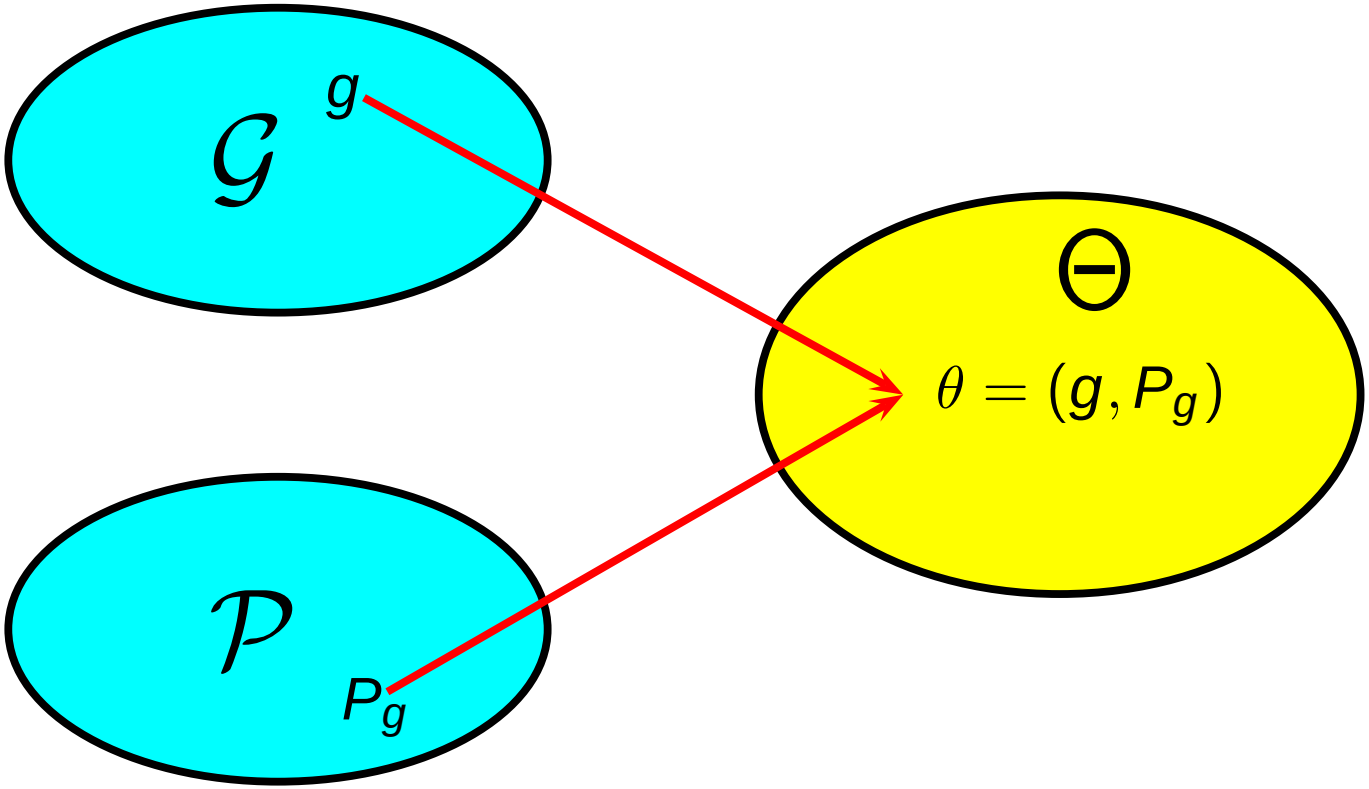
$$\mathbf{S}_{\mathcal{A}_\Theta}(m) = \max_{x_1,\ldots,x_m\in\mathbb{R}^d} Card\left\{\{x_1,\ldots,x_m\} \cap A : A \in \mathcal{A}_\Theta\right\}.$$

Le modèle



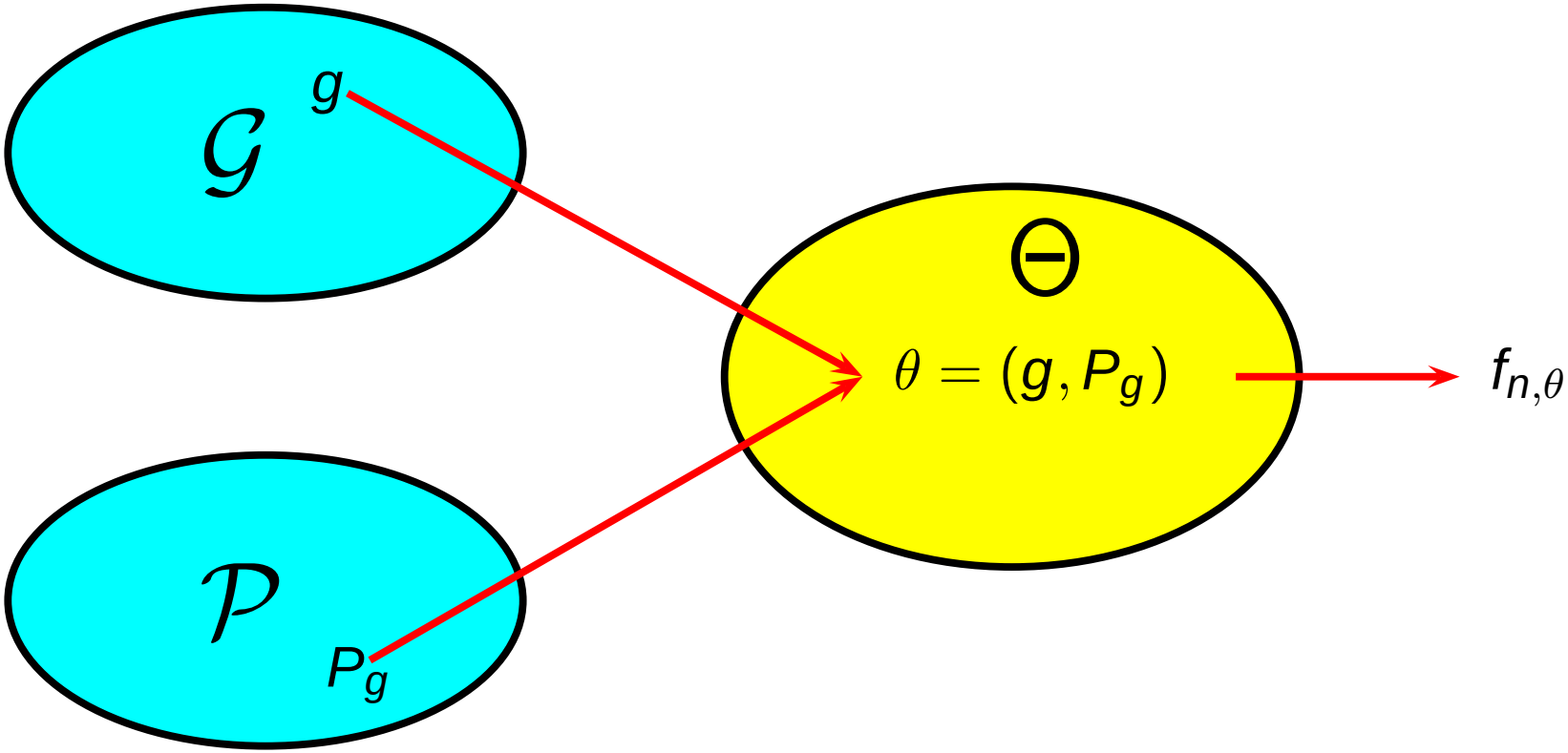


Le modèle



$$\Theta = \left\{(g, P_g), g \in \mathcal{G},\, P_g = \{A_1, \ldots, A_\ell\} \in \mathcal{P}, \ell \leq r,\, \nu_g(A_i) = 1/\ell\right\}.$$

Le modèle



$$\Theta = \left\{(g, P_g), g \in \mathcal{G},\, P_g = \{A_1, \ldots, A_\ell\} \in \mathcal{P}, \ell \leq r,\, \nu_g(A_i) = 1/\ell\right\}.$$

Résultats

$$\mathcal{A}_\Theta = \left\{ \{x : f_{n-m,\theta}(x) > f_{n-m,\theta'}(x)\} : (\theta, \theta') \in \Theta^2 \right\}.$$

Théorème

$$\mathbf{S}_{\mathcal{A}_\Theta}(m) \leq \mathbf{S}_{\mathcal{D}}(m) \left[\mathbf{S}_{\mathcal{P}}(m(n-m)) \right]^{4r},$$

où

$$\mathcal{D} = \left\{ \{(x,z) \in \mathbb{R}^d \times \mathbb{R}_+^* : \alpha z g(x) - g'(x) > 0\} : \alpha \in \mathbb{R}_+^*, (g,g') \in \mathcal{G}^2 \right\}.$$

et $\mathbf{S}_{\mathcal{P}}(j)$ est le coefficient de pulvérisation associé à la classe des ensembles de $\mathcal{P}$.

Résultats

$$\mathcal{A}_\Theta = \left\{ \{x : f_{n-m,\theta}(x) > f_{n-m,\theta'}(x)\} : (\theta, \theta') \in \Theta^2 \right\}.$$

Théorème

$$\mathbf{S}_{\mathcal{A}_\Theta}(m) \leq \mathbf{S}_{\mathcal{D}}(m) \left[\mathbf{S}_{\mathcal{P}}\left(m(n-m)\right) \right]^{4r},$$

où

$$\mathcal{D} = \left\{ \{(x,z) \in \mathbb{R}^d \times \mathbb{R}_+^* : \alpha z g(x) - g'(x) > 0\} : \alpha \in \mathbb{R}_+^*, (g, g') \in \mathcal{G}^2 \right\}.$$

et $\mathbf{S}_{\mathcal{P}}(j)$ *est le coefficient de pulvérisation associé à la classe des ensembles de* $\mathcal{P}$*.*

Résultats

Corollaire (1)

$$\mathbf{E}\left\{ \int |f_n - f| \right\} \leq 3 \inf_{\theta \in \Theta} \mathbf{E}\left\{ \int |f_{n-m,\theta} - f| \right\} + 8\sqrt{\frac{\log 2 + \log \mathbf{S}_{\mathcal{D}}(m) + 4r \log \mathbf{S}_{\mathcal{P}}\left(m(n-m)\right)}{m}} + \frac{3}{n}.$$

Exemple

$\mathcal{G}$: famille exponentielle en dimension 1.

$$\begin{cases} \mathbf{S}_{\mathcal{P}}(j) = \frac{j(j+1)}{2} + 1 \\ \mathbf{S}_{\mathcal{D}}(j) \leq (j+1)^{k+2} \end{cases}$$

Résultats

Corollaire (1)

$$\mathbf{E}\left\{ \int |f_n - f| \right\} \leq 3 \inf_{\theta \in \Theta} \mathbf{E}\left\{ \int |f_{n-m,\theta} - f| \right\} + 8\sqrt{\frac{\log 2 + \log \mathbf{S}_{\mathcal{D}}(m) + 4r \log \mathbf{S}_{\mathcal{P}}\left(m(n-m)\right)}{m}} + \frac{3}{n}.$$

Exemple

$\mathcal{G}$: famille exponentielle en dimension 1.

$$\begin{cases} \mathbf{S}_{\mathcal{P}}(j) = \frac{j(j+1)}{2} + 1 \\ \mathbf{S}_{\mathcal{D}}(j) \leq (j+1)^{k+2} \end{cases}$$

Résultats

Corollaire

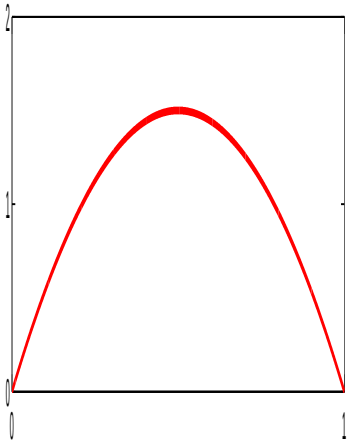
Si $\mathbf{S}_{\mathcal{D}}(j)$ *et* $\mathbf{S}_{\mathcal{P}}(j)$ *sont polynomiaux en* j*. Alors les choix*

$$m = \frac{n}{\log n} \qquad et \qquad r = n^a, \quad 0 < a \leq 1/2,$$

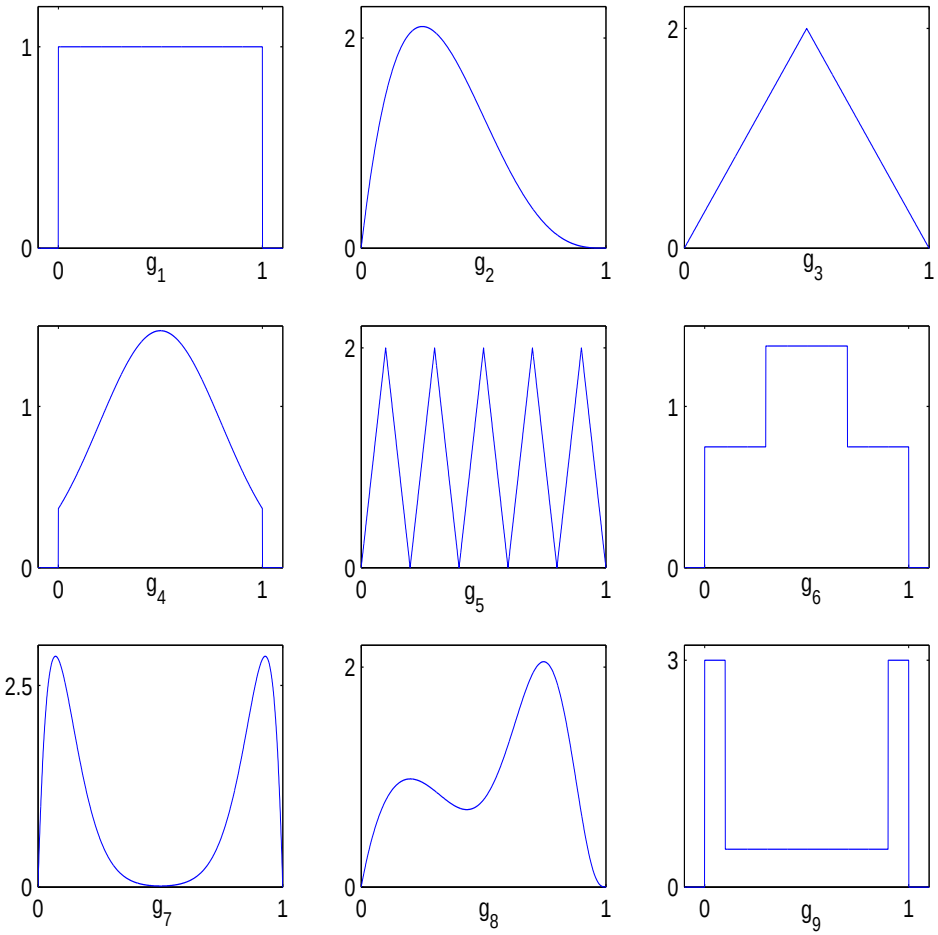
donnent

$$\mathbf{E}\left\{ \int |f_n - f| \right\} \leq 3 \inf_{\theta \in \Theta} \mathbf{E}\left\{ \int |f_{n-m,\theta} - f| \right\} + O\left(\frac{\log n}{n^{(1-a)/2}} \right).$$

Simulations



Densité à estimer



Densités de référence

Résultats

	$n = 200, m = 50, r = 16$		
g	$\int \lvert f_{n,g} - f \rvert$	$\int \lvert f_{n,\theta_g^\star} - f \rvert$	$\hat{\ell}_n$
g_1	0.21	0.15	9.68
g_2	0.33	0.30	12.92
g_3	0.17	0.11	7.28
g_4	0.18	0.10	8.28
g_5	0.43	0.40	14.28
g_6	0.23	0.19	10.84
g_7	0.82	0.81	15.64
g_8	0.22	0.17	9.04
g_9	0.24	0.17	10.92
g_{1-9}	**0.22**	**0.10**	**8.28**

Résultats

	$n = 200, m = 50, r = 16$		
g	$\int \lvert f_{n,g} - f \rvert$	$\int \lvert f_{n,\theta_g^\star} - f \rvert$	$\hat{\ell}_n$
g_1	0.21	0.15	9.68
g_2	**0.33**	0.30	12.92
g_3	0.17	0.11	7.28
g_4	0.18	0.10	8.28
g_5	**0.43**	0.40	14.28
g_6	**0.23**	0.19	10.84
g_7	**0.82**	0.81	15.64
g_8	**0.22**	0.17	9.04
g_9	**0.24**	0.17	10.92
g_{1-9}	**0.22**	**0.10**	**8.28**

Estimateurs à noyau

Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de densité f

$$f_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)$$

Estimateurs à noyau variable

Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de densité f

$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(x, X_i)} K\left(\frac{x - X_i}{h(x, X_i)} \right)$$

où $h : \mathbb{R}^d \times \mathbb{R}^d \to (0, \infty)$.

Estimateurs à noyau variable

Soit $X_1, \ldots, X_n$ un échantillon i.i.d. de densité f

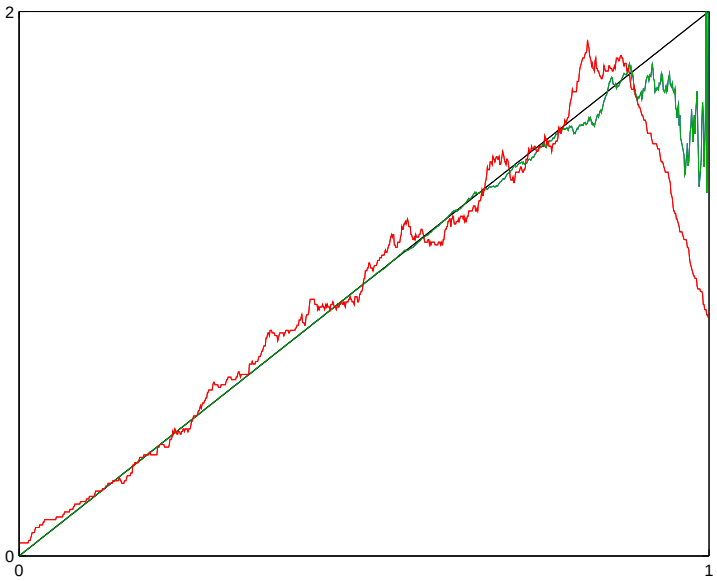
$$f_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h(x, X_i)} K\left(\frac{x - X_i}{h(x, X_i)} \right)$$

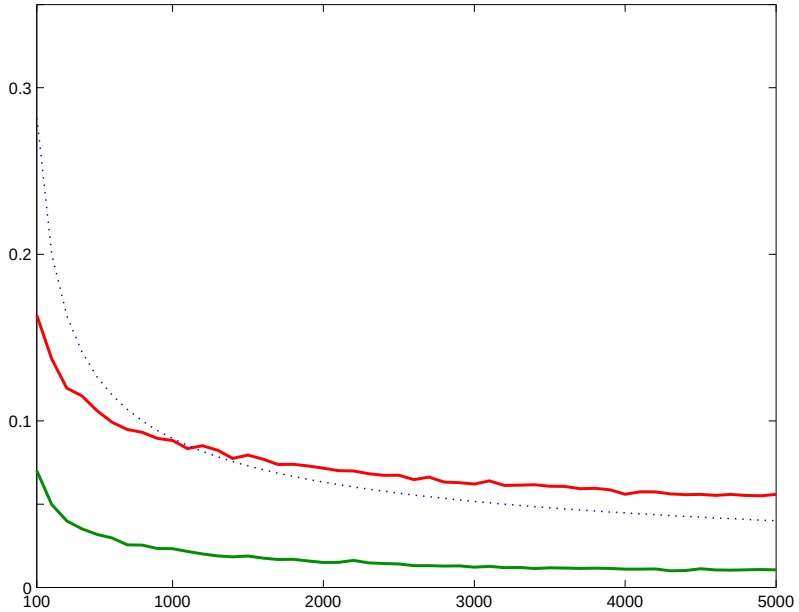
où $h : \mathbb{R}^d \times \mathbb{R}^d \to (0, \infty)$.

Bonnes propriétés :

- En grande dimension;
- En terme de vitesses de convergence (réduction du biais);
- A distance finie pour certaines classes de densités.

Exemple





— f densité à estimer;

— f_{n,h_0} tel que $h_0 = \text{argmin}_{h>0} \int |f_{n,h} - f|$;

— $f_{n,h_0(x)}$ pour une certaine fonction $h_0(x)$.

— La borne $\sqrt{8/n}$;

— $\min_h L_1(f_{n,h})$;

— $L_1(f_{n,h_0(x)})$.

Le modèle

On considère des fenêtres de la forme

$$h(x, X_i, \theta) = \phi(x, X_i, \lambda)$$

avec

- $\lambda \in \mathbb{R}^p$
- $\lambda \to \phi(x, X_i, \lambda)$ polynomiale de degré $\leq \ell$.

Le modèle

On considère des fenêtres de la forme

$$h(x, X_i, \theta) = \sum_{j_1=1}^{r_1} \phi(x, X_i, \lambda_{j_1}) \mathbf{1}_{B_{j_1}^1}(x)$$

avec

- $\lambda_{j_1 j_2} \in \mathbb{R}^p$
- $\lambda_{j_1 j_2} \to \phi(x, X_i, \lambda_{j_1 j_2})$ polynomiale de degré $\leq \ell$.
- $P_1 = \{B_1^1, \ldots, B_{r_1}^1\} \in \mathcal{P}_1$

Le modèle

On considère des fenêtres de la forme

$$h(x, X_i, \theta) = \sum_{(j_1, j_2) \in J} \phi(x, X_i, \lambda_{j_1 j_2}) \mathbf{1}_{B_{j_1}^1 \times B_{j_2}^2}(x, X_i)$$

avec

- $\lambda_{j_1 j_2} \in \mathbb{R}^p$
- $\lambda_{j_1 j_2} \to \phi(x, X_i, \lambda_{j_1 j_2})$ polynomiale de degrés $\leq \ell$.
- $P_1 = \{B_1^1, \ldots, B_{r_1}^1\} \in \mathcal{P}_1$
- $P_2 = \{B_1^2, \ldots, B_{r_2}^2\} \in \mathcal{P}_2$
- $J = \{1, \ldots, r_1\} \times \{1, \ldots, r_2\}$

Le modèle

On considère des fenêtres de la forme

$$h(x, X_i, \theta) = \sum_{(j_1, j_2) \in J} \phi(x, X_i, \lambda_{j_1 j_2}) \mathbf{1}_{B^1_{j_1} \times B^2_{j_2}}(x, X_i)$$

avec

- $\lambda_{j_1 j_2} \in \mathbb{R}^p$
- $\lambda_{j_1 j_2} \to \phi(x, X_i, \lambda_{j_1 j_2})$ polynomiale de degré $\leq \ell$.
- $P_1 = \{B^1_1, \ldots, B^1_{r_1}\} \in \mathcal{P}_1$
- $P_2 = \{B^2_1, \ldots, B^2_{r_2}\} \in \mathcal{P}_2$
- $J = \{1, \ldots, r_1\} \times \{1, \ldots, r_2\}$

La méthode combinatoire

On utilise la méthode combinatoire pour sélectionner θ dans :

$$\Theta = \left\{ (P_1, P_2, \underline{\lambda}) : P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2, \underline{\lambda} \in \mathbb{R}^{r_1 r_2 p} \right\}.$$

Théorème

$$\mathbf{S}_{\mathcal{A}_\Theta}(m) \leq \psi\big(n, m, \ell, \mathbf{S}_{\mathcal{P}}(n)\big)^{r_1 r_2 p},$$

où ψ est polynomial en ses arguments.

Corollaire

En particulier, le choix $m = n/\log n$ donne

$$\mathbf{E}\left\{ \int |f_n - f| \right\} \leq 5\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right) \inf_{\theta \in \Theta} \mathbf{E}\left\{ \int |f_{n,\theta} - f| \right\} + O\left(\frac{\log n}{\sqrt{n}}\right).$$

La méthode combinatoire

On utilise la méthode combinatoire pour sélectionner θ dans :

$$\Theta = \left\{ (P_1, P_2, \underline{\lambda}) : P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2, \underline{\lambda} \in \mathbb{R}^{r_1 r_2 p} \right\}.$$

Théorème

$$\mathbf{S}_{\mathcal{A}_\Theta}(m) \leq \psi\big(n, m, \ell, \mathbf{S}_{\mathcal{P}}(n)\big)^{r_1 r_2 p},$$

où ψ est polynomial en ses arguments.

Corollaire

En particulier, le choix $m = n/\log n$ donne

$$\mathbf{E}\left\{ \int |f_n - f| \right\} \leq 5\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right) \inf_{\theta \in \Theta} \mathbf{E}\left\{ \int |f_{n,\theta} - f| \right\} + O\left(\frac{\log n}{\sqrt{n}}\right).$$

La méthode combinatoire

On utilise la méthode combinatoire pour sélectionner θ dans :

$$\Theta = \left\{ (P_1, P_2, \underline{\lambda}) : P_1 \in \mathcal{P}_1, P_2 \in \mathcal{P}_2, \underline{\lambda} \in \mathbb{R}^{r_1 r_2 p} \right\}.$$

Théorème

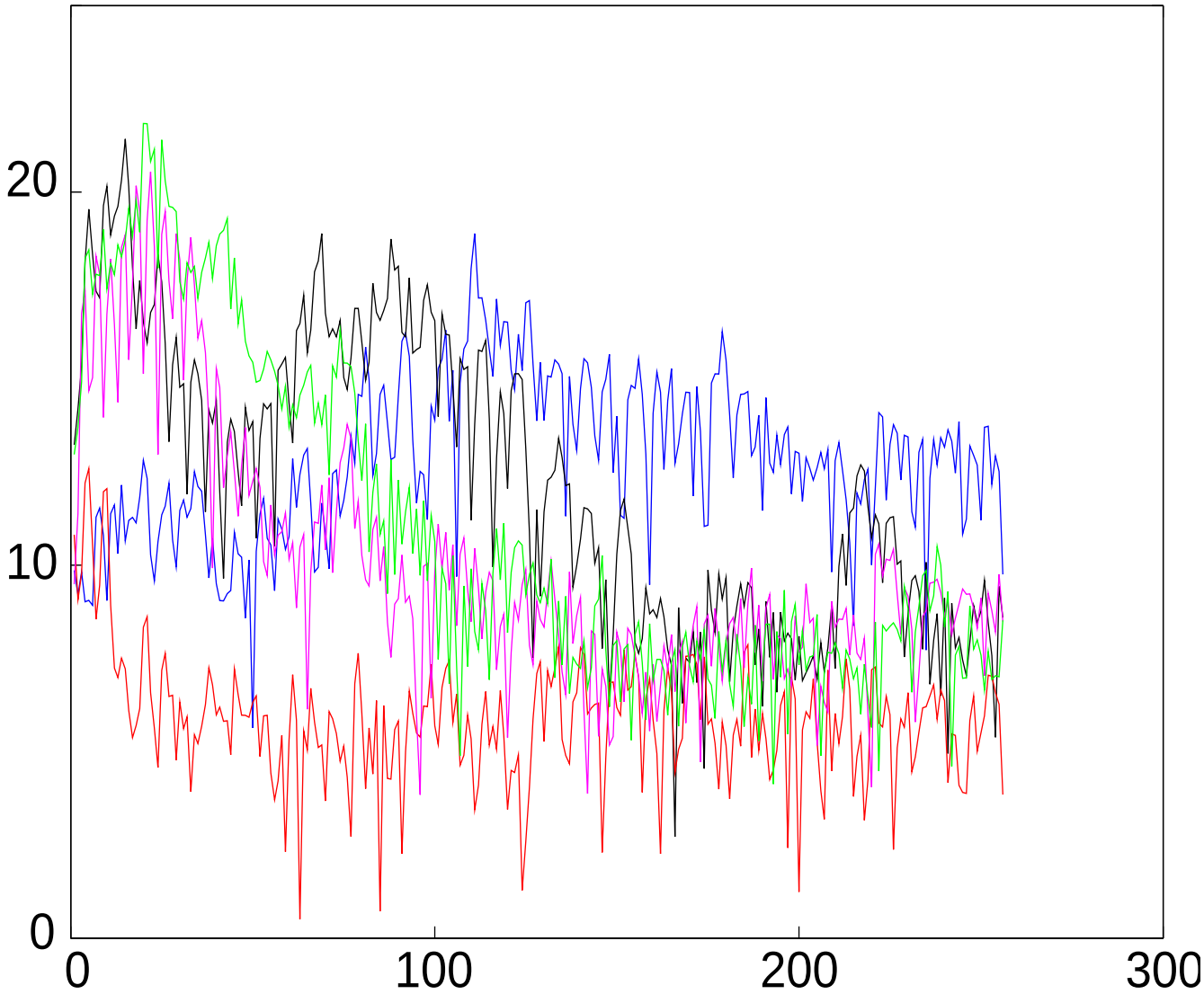
$$\mathbf{S}_{\mathcal{A}_\Theta}(m) \leq \psi\big(n, m, \ell, \mathbf{S}_{\mathcal{P}}(n)\big)^{r_1 r_2 p},$$

où ψ est polynomial en ses arguments.

Corollaire

En particulier, le choix $m = n/\log n$ donne

$$\mathbf{E}\left\{ \int |f_n - f| \right\} \leq 5\left(1 + O\left(\frac{1}{\sqrt{\log n}}\right)\right) \inf_{\theta \in \Theta} \mathbf{E}\left\{ \int |f_{n,\theta} - f| \right\} + O\left(\frac{\log n}{\sqrt{n}}\right).$$



“sh”, “iy”, “dcl”, “ao”, “aa”.

Le modèle mathématique

- Soit $(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$ un échantillon de copies indépendantes du couple $(X, Y) \in \mathcal{F} \times \{0, 1\}$. Le problème de la classification consiste à prédire le label inconnu Y d'une nouvelle observation X.
- Le statisticien crée une règle de classification

$$g : \mathcal{F} \to \{0, 1\}$$

qui représente sa prédiction concernant le label de X.

- On définit la probabilité d'erreur pour une règle par

$$L(g) = \mathbf{P}\{g(X) \neq Y\}.$$

Le modèle mathématique

- Soit $(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$ un échantillon de copies indépendantes du couple $(X, Y) \in \mathcal{F} \times \{0, 1\}$. Le problème de la classification consiste à prédire le label inconnu Y d'une nouvelle observation X.
- Le statisticien crée une règle de classification

$$g : \mathcal{F} \to \{0, 1\}$$

qui représente sa prédiction concernant le label de X.

- On définit la probabilité d'erreur pour une règle par

$$L(g) = \mathbf{P}\{g(X) \neq Y\}.$$

Le modèle mathématique

- Soit $(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$ un échantillon de copies indépendantes du couple $(X, Y) \in \mathcal{F} \times \{0, 1\}$. Le problème de la classification consiste à prédire le label inconnu Y d'une nouvelle observation X.
- Le statisticien crée une règle de classification

$$g : \mathcal{F} \to \{0, 1\}$$

qui représente sa prédiction concernant le label de X.

- On définit la probabilité d'erreur pour une règle par

$$L(g) = \mathbf{P}\{g(X) \neq Y\}.$$

Classification

- La règle de Bayes

$$g^*(x) = \begin{cases} 0 & \text{si } \mathbf{P}\{Y = 0 | X = x\} \geq \mathbf{P}\{Y = 1 | X = x\} \\ 1 & \text{sinon,} \end{cases}$$

est optimale au sens où pour toutes règles $g : \mathcal{F} \to \{0, 1\}$,

$$\mathbf{P}\{g^*(X) \neq Y\} \leq \mathbf{P}\{g(X) \neq Y\}.$$

- Le problème est alors de construire une règle raisonnable $\hat{g}$ à partir des observations $(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$ telle que

$$\lim_{n \to \infty} \mathbf{E} L(\hat{g}) = L^*.$$

- En dimension finie : arbres, plus proches voisins, noyau...

Classification

- La règle de Bayes

$$g^*(x) = \begin{cases} 0 & \text{si } \mathbf{P}\{Y = 0 | X = x\} \geq \mathbf{P}\{Y = 1 | X = x\} \\ 1 & \text{sinon,} \end{cases}$$

est optimale au sens où pour toutes règles $g : \mathcal{F} \to \{0, 1\}$,

$$\mathbf{P}\{g^*(X) \neq Y\} \leq \mathbf{P}\{g(X) \neq Y\}.$$

- Le problème est alors de construire une règle raisonnable $\hat{g}$ à partir des observations $(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$ telle que

$$\lim_{n \to \infty} \mathbf{E}L(\hat{g}) = L^\star .$$

- En dimension finie : arbres, plus proches voisins, noyau...

Classification

- La règle de Bayes

$$g^*(x) = \begin{cases} 0 & \text{si } \mathbf{P}\{Y = 0 | X = x\} \geq \mathbf{P}\{Y = 1 | X = x\} \\ 1 & \text{sinon,} \end{cases}$$

est optimale au sens où pour toutes règles $g : \mathcal{F} \to \{0, 1\}$,

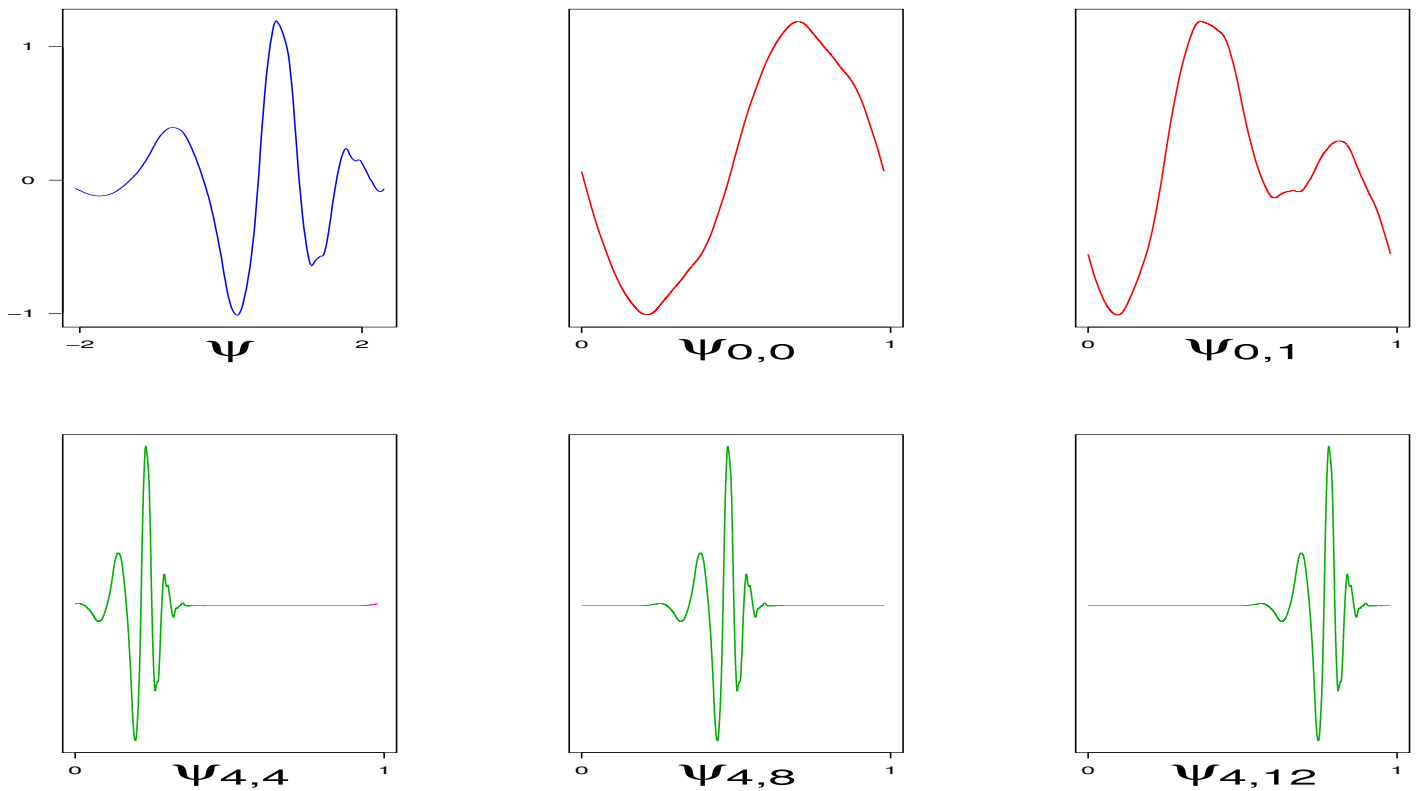
$$\mathbf{P}\{g^*(X) \neq Y\} \leq \mathbf{P}\{g(X) \neq Y\}.$$

- Le problème est alors de construire une règle raisonnable $\hat{g}$ à partir des observations $(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$ telle que

$$\lim_{n \to \infty} \mathbf{E}L(\hat{g}) = L^\star .$$

- En dimension finie : arbres, plus proches voisins, noyau...

Ondelettes de Daubechies



— : ondelette mère
— : bases de W_0
— : éléments de la base de W_4.

Réduction de la dimension

- $V_0 \subset V_1 \subset V_2 \subset \ldots \subset V_J \subset \ldots \subset L_2([0,1])$;
- Les observations X_i sont approchées par

$$X_i(t) \approx \sum_{j=1}^{2^J} X_{ij}\psi_j(t)\,.$$

- échantillon d'apprentissage $(X_1, Y_1), \ldots, (X_n, Y_n)$;
- échantillon test $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$.

Réduction de la dimension

- $V_0 \subset V_1 \subset V_2 \subset \ldots \subset V_J \subset \ldots \subset L_2([0,1])$;
- Les observations X_i sont approchées par

$$X_i(t) \approx \sum_{j=1}^{2^J} X_{ij} \psi_j(t) \, .$$

- échantillon d'apprentissage $(X_1, Y_1), \ldots, (X_n, Y_n)$;
- échantillon test $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$.

Réduction de la dimension

- $V_0 \subset V_1 \subset V_2 \subset \ldots \subset V_J \subset \ldots \subset L_2([0,1])$;
- Les observations X_i sont approchées par

$$X_i(t) \approx \sum_{j=1}^{2^J} X_{ij}\psi_j(t)\,.$$

- échantillon d'apprentissage $(X_1, Y_1), \ldots, (X_n, Y_n)$;
- échantillon test $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$.

Un seuillage global

- Ordonnons les 2^J fonctions $\{\psi_1, \ldots, \psi_{2^J}\}$ en $\{\psi_{j_1}, \ldots, \psi_{j_{2^J}}\}$ suivant

$$\sum_{i=1}^{n} {X_{ij_1}}^2 \geq \sum_{i=1}^{n} {X_{ij_2}}^2 \geq \ldots \geq \sum_{i=1}^{n} {X_{ij_{2^J}}}^2 .$$

- Les coefficients d'ondelettes sont rangés à l'aide un seuillage global suivant la moyenne des carrés des coefficients de l'échantillon d'apprentissage.

- On notera

$$\mathbf{X}_i^{(d)} = (X_{ij_1}, \ldots, X_{ij_d}) .$$

Un seuillage global

- Ordonnons les 2^J fonctions $\{\psi_1, \ldots, \psi_{2^J}\}$ en $\{\psi_{j_1}, \ldots, \psi_{j_{2^J}}\}$ suivant

$$\sum_{i=1}^{n} X_{ij_1}{}^2 \geq \sum_{i=1}^{n} X_{ij_2}{}^2 \geq \ldots \geq \sum_{i=1}^{n} X_{ij_{2^J}}{}^2 .$$

- Les coefficients d'ondelettes sont rangés à l'aide un seuillage global suivant la moyenne des carrés des coefficients de l'échantillon d'apprentissage.

- On notera

$$\mathbf{x}_i^{(d)} = (X_{ij_1}, \ldots, X_{ij_d}) .$$

Un seuillage global

- Ordonnons les 2^J fonctions $\{\psi_1, \ldots, \psi_{2^J}\}$ en $\{\psi_{j_1}, \ldots, \psi_{j_{2^J}}\}$ suivant

$$\sum_{i=1}^{n} {X_{ij_1}}^2 \geq \sum_{i=1}^{n} {X_{ij_2}}^2 \geq \ldots \geq \sum_{i=1}^{n} {X_{ij_{2^J}}}^2 .$$

- Les coefficients d'ondelettes sont rangés à l'aide un seuillage global suivant la moyenne des carrés des coefficients de l'échantillon d'apprentissage.

- On notera

$$\mathbf{x}_i^{(d)} = (X_{ij_1}, \ldots, X_{ij_d}) .$$

La procédure

- Pour chaque $d = 1, \ldots, 2^J$, soit $D_n^{(d)}$ une classe de règles $g^{(d)} : \mathbb{R}^d \times (\mathbb{R}^d \times \{0,1\})^n \to \{0,1\}$.
- Soit $\mathbb{S}_{C_n^{(d)}}(m)$ le coefficient de pulvérisation associé à cette classe, et soit $\mathbb{S}_{C_n}^{(J)}(m)$ le coefficient de pulvérisation de toutes les règles $\{g^{(d)} : d = 1, \ldots, 2^J\}$.
- On sélectionne d et $g^{(d)}$ par minimisation de la probabilité d'erreur empirique basée sur l'échantillon de validation :

$$\left(\hat{d}, \hat{g}^{(\hat{d})}\right) = \underset{1 \leq d \leq 2^J,\, g \in D_n^{(d)}}{\operatorname{argmin}} \left[\frac{1}{m} \sum_{i=n+1}^{n+m} \mathbf{1}_{[g^{(d)}(\mathbf{X}_i^{(d)}) \neq Y_i]} \right].$$

La procédure

- Pour chaque $d = 1, \ldots, 2^J$, soit $D_n^{(d)}$ une classe de règles $g^{(d)} : \mathbb{R}^d \times (\mathbb{R}^d \times \{0,1\})^n \to \{0,1\}$.
- Soit $\mathbb{S}_{C_n^{(d)}}(m)$ le coefficient de pulvérisation associé à cette classe, et soit $\mathbb{S}_{C_n}^{(J)}(m)$ le coefficient de pulvérisation de toutes les règles $\{g^{(d)} : d = 1, \ldots, 2^J\}$.

- On sélectionne d et $g^{(d)}$ par minimisation de la probabilité d'erreur empirique basée sur l'échantillon de validation :

$$\left(\hat{d}, \hat{g}^{(\hat{d})}\right) = \underset{1 \leq d \leq 2^J,\, g \in D_n^{(d)}}{\operatorname{argmin}} \left[\frac{1}{m} \sum_{i=n+1}^{n+m} \mathbf{1}_{[g^{(d)}(\mathbf{X}_i^{(d)}) \neq Y_i]} \right].$$

Théorème

$$\mathbf{E}\left\{L_{n+m}(\hat{g})\right\} - L^* \leq L^*_{2^J} - L^* + \mathbf{E}\left\{ \inf_{d=1,\ldots,2^J, g^{(d)} \in D_n^{(d)}} L_n(g^{(d)}) \right\} - L^*_{2^J}$$

$$+ 2\mathbf{E}\left\{ \sqrt{\frac{8 \log\left(4\mathbb{S}^{(J)}_{\mathbf{C}_n}(2m)\right)}{m}} + \frac{1}{\sqrt{(m/2)\log\left(4\mathbb{S}^{(J)}_{\mathbf{C}_n}(2m)\right)}} \right\}.$$

Corollaire

Si chaque ensemble $D_n^{(J)}$ contient une règle convergente. Si,de plus,

$$\lim_{n\to\infty} m = \infty, \qquad et \qquad \lim_{n\to\infty} \mathbf{E}\left\{ \frac{\log \mathbb{S}^{(J)}_{\mathbf{C}_n}(2m)}{m} \right\} = 0,$$

alors

$$\lim_{J\to\infty} \lim_{n\to\infty} \mathbf{E}\left\{L_{n+m}(\hat{g})\right\} = L^*.$$

Théorème

$$\mathbf{E}\left\{L_{n+m}(\hat{g})\right\} - L^* \leq L^*_{2^J} - L^* + \mathbf{E}\left\{\inf_{d=1,\ldots,2^J, g^{(d)} \in D_n^{(d)}} L_n(g^{(d)})\right\} - L^*_{2^J}$$

$$+ 2\mathbf{E}\left\{\sqrt{\frac{8\log\left(4\mathbb{S}^{(J)}_{\mathbf{C}_n}(2m)\right)}{m}} + \frac{1}{\sqrt{(m/2)\log\left(4\mathbb{S}^{(J)}_{\mathbf{C}_n}(2m)\right)}}\right\}.$$

Corollaire

Si chaque ensemble $D_n^{(J)}$ contient une règle convergente. Si,de plus,

$$\lim_{n\to\infty} m = \infty, \qquad et \qquad \lim_{n\to\infty} \mathbf{E}\left\{\frac{\log \mathbb{S}^{(J)}_{\mathbf{C}_n}(2m)}{m}\right\} = 0,$$

alors

$$\lim_{J\to\infty} \lim_{n\to\infty} \mathbf{E}\left\{L_{n+m}(\hat{g})\right\} = L^*.$$

Illustrations

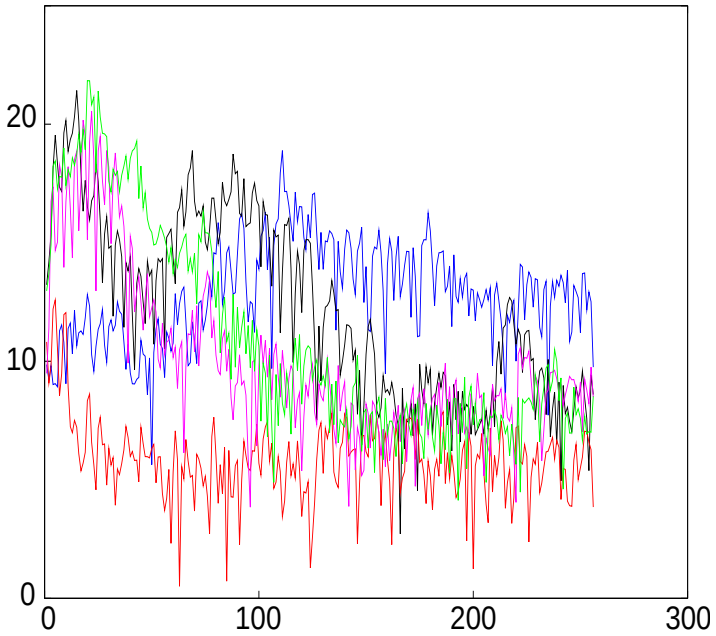
- W-QDA quand $D_n^{(d)}$ représente une “Analyse discriminante quadratique” en dimension d.
- W-NN quand $D_n^{(d)}$ contient toute les règles des kppv.
- W-T quand $D_n^{(d)}$ contient tous les arbres binaires.
- W-BOOST lorsque l’on applique l’algorithme “Adaboost” sur les coefficients sélectionnés;
- F-NN désigne une méthode basée sur les coefficients de Fourier.
- MPLSR désigne la régression PLS multivariée.

Illustrations

- W-QDA quand $D_n^{(d)}$ représente une “Analyse discriminante quadratique” en dimension d.
- W-NN quand $D_n^{(d)}$ contient toute les règles des kppv.
- W-T quand $D_n^{(d)}$ contient tous les arbres binaires.
- W-BOOST lorsque l’on applique l’algorithme “Adaboost” sur les coefficients sélectionnés;

- F-NN désigne une méthode basée sur les coefficients de Fourier.
- MPLSR désigne la régression PLS multivariée.

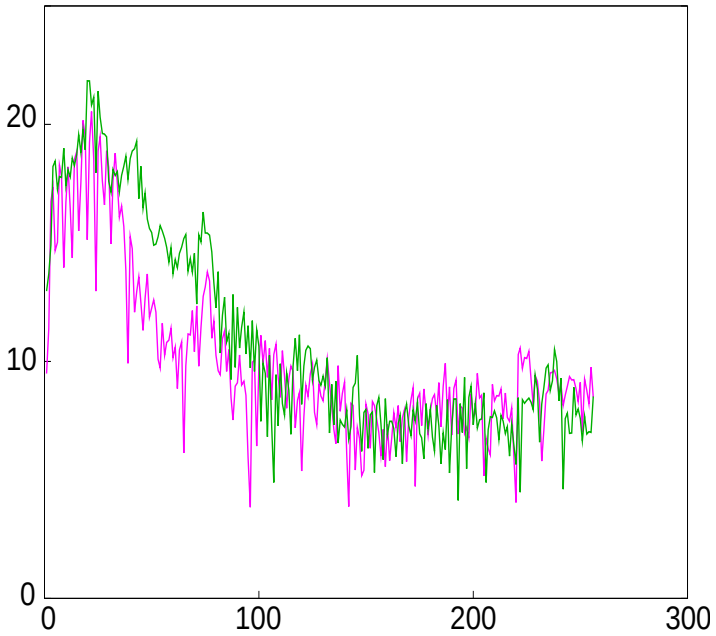
Reconnaissance vocale



“sh”, “iy”, “dcl”, “ao”, “aa”.

4509 observations, n=m=250
50 partitions

Reconnaissance vocale

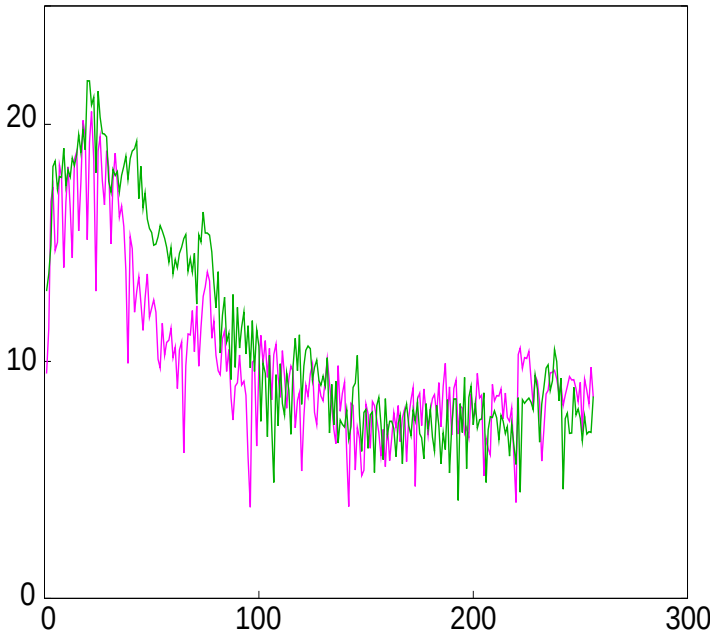


“ao”, “aa”.

1717 observations, n=m=250
50 partitions

Méthodes	**Erreurs**	$\hat{d}$
W-QDA	0.23	19
W-NN	0.21	22
W-T	0.22	9
W-BOOST	0.21	21
F-NN	0.25	42
MPLSR	0.20	5

Reconnaissance vocale

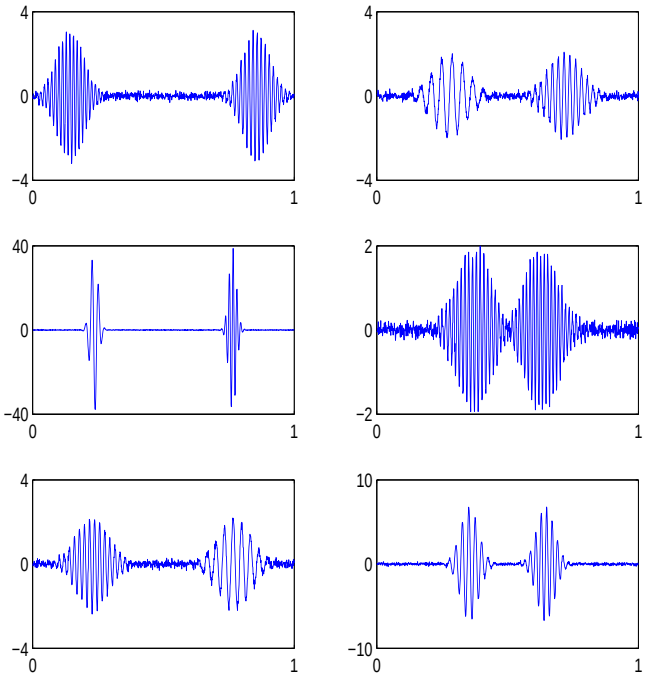


"ao", "aa".

1717 observations, n=m=250
50 partitions

Méthodes	**Erreurs**	$\hat{d}$
W-QDA	0.23	19
W-NN	0.21	22
W-T	0.22	9
W-BOOST	0.21	21
F-NN	0.25	42
MPLSR	0.20	5

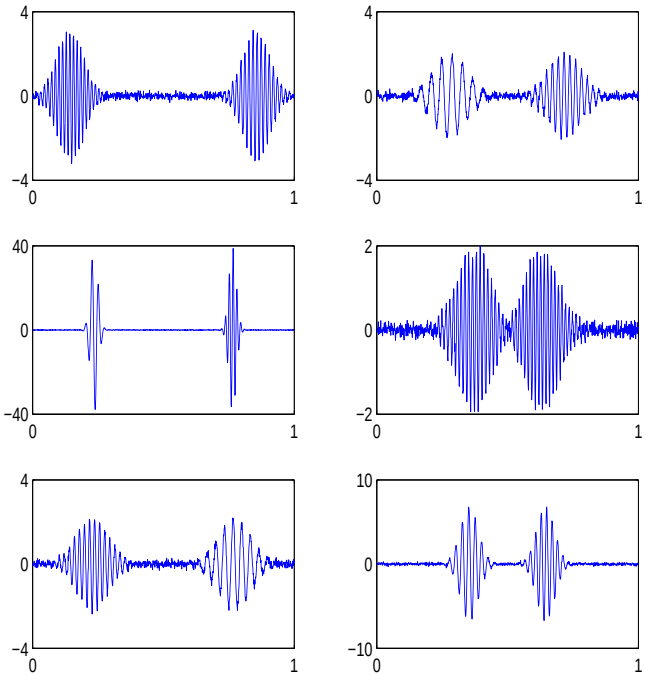
Un exemple simulé



500 observations, $n = m = 50$,
50 répétitions

Méthodes	**Erreurs**	$\hat{d}$
W-QDA	0.08	8
W-NN	0.12	24
W-T	0.13	20
W-BOOST	0.08	43
F-NN	0.35	76
MPLSR	0.44	4

Un exemple simulé



500 observations, $n = m = 50$,
50 répétitions

Méthodes	**Erreurs**	$\hat{d}$
W-QDA	0.08	8
W-NN	0.12	24
W-T	0.13	20
W-BOOST	0.08	43
F-NN	0.35	76
MPLSR	0.44	4

Perspectives

1. Densité
   - Itérer les histogrammes modifiés en “grande dimension”;
   - Mise en oeuvre effective sur machines des méthodes combinatoires;

2. Classification pour des courbes dicrétisées
   - “Design” aléatoire;
   - Obtenir des vitesses;
   - Méthodes à noyau.

Perspectives

1. Densité
   - Itérer les histogrammes modifiés en “grande dimension”;
   - Mise en oeuvre effective sur machines des méthodes combinatoires;

2. Classification pour des courbes dicrétisées
   - “Design” aléatoire;
   - Obtenir des vitesses;
   - Méthodes à noyau.