Sélection de modèles et sélection d'estimateurs pour l'Apprentissage statistique

Sylvain Arlot

[1]Cnrs

[2]École Normale Supérieure (Paris), Liens, Équipe Sierra

Cours Peccot, Collège de France, 10/01/2011

Plan

1. Aujourd'hui : Apprentissage statistique et sélection d'estimateurs
2. Lundi 17, 14h–16h : Calibration de pénalités et pénalités minimales
3. Lundi 24, 14h–16h : Rééchantillonnage et pénalisation
4. Lundi 31, 15h30–17h30 : Validation croisée et pénalités reliées

Plan du cours

Plan

Cadre général

- Données : $\xi_1, \ldots, \xi_n \in \Xi$ i.i.d. de loi P
- Objectif : estimer une caractéristique $s^\star \in \mathbb{S}$ de la loi P
- Mesure de qualité : fonction de perte

$$\forall t \in \mathbb{S}, \quad \mathcal{L}_P(t) = \mathbb{E}_{\xi \sim P}[\gamma(t;\xi)] = P\gamma(t) \ ,$$

minimale en $t = s^\star$. La fonction $\gamma : \mathbb{S} \times \Xi \mapsto [0, +\infty)$ est appelée contraste.

- Perte relative

$$\ell(s^\star, t) = P\gamma(t) - P\gamma(s^\star) \ .$$

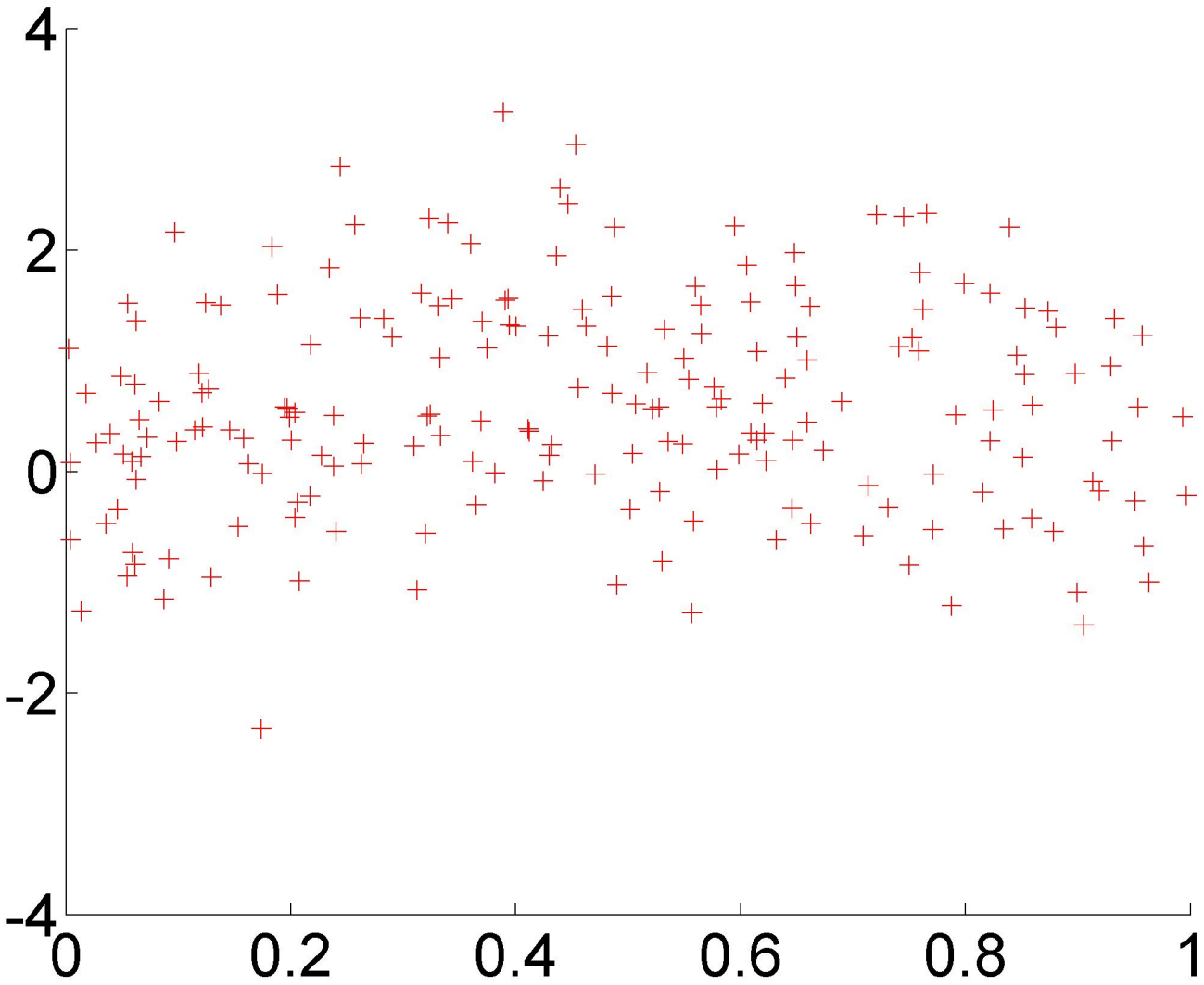
Exemple : prédiction (ou prévision)

- Données : $(X_1, Y_1), \ldots, (X_n, Y_n) \in \Xi = \mathcal{X} \times \mathcal{Y}$
- Objectif : prédire Y sachant X lorsque $(X, Y) = \xi \sim P$
- $s^\star(X)$ est le "meilleur prédicteur" de Y sachant X, i.e., $s^\star$ minimise la fonction de perte

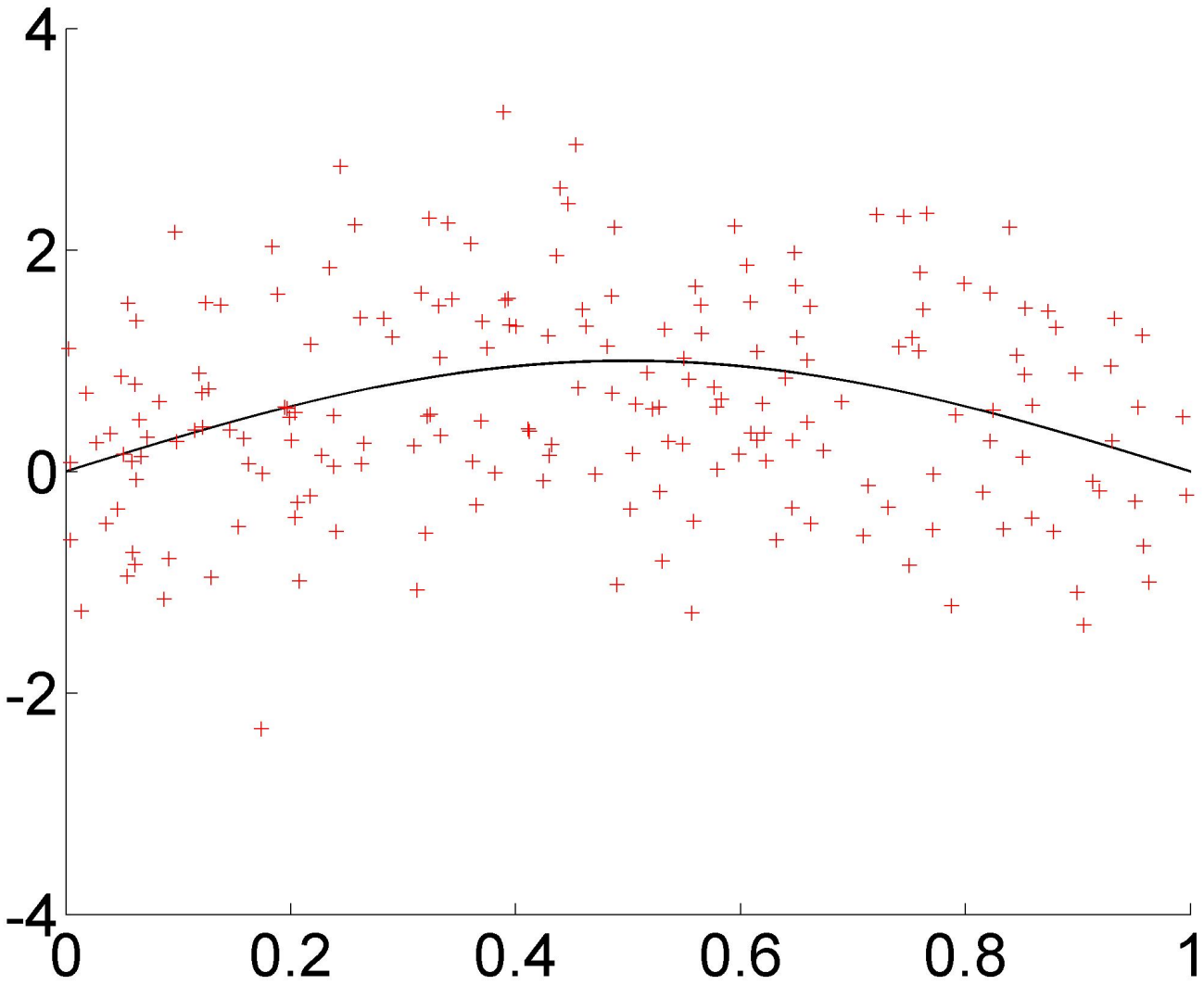
$$P\gamma(t) \quad \text{avec} \quad \gamma(t;(x,y)) = d(t(x), y)$$

mesurant une "distance" entre y et la valeur prédite $t(x)$.

Exemple : régression : données $(X_1, Y_1), \ldots, (X_n, Y_n)$



But : reconstruire le signal



Exemple : régression

- prédiction avec $\mathcal{Y} = \mathbb{R}$
- Données : $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.

$$Y_i = \eta(X_i) + \varepsilon_i \quad \text{avec} \quad \mathbb{E}\left[\varepsilon_i \mid X_i\right] = 0$$

Exemple : régression

- prédiction avec $\mathcal{Y} = \mathbb{R}$
- Données : $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.

$$Y_i = \eta(X_i) + \varepsilon_i \quad \text{avec} \quad \mathbb{E}\left[\varepsilon_i \mid X_i\right] = 0$$

- contraste de moindres carrés : $\gamma(t;(x,y)) = (t(x) - y)^2$

$$\Rightarrow \quad s^\star = \eta \quad \text{et} \quad \ell\left(s^\star, t\right) = \|t - \eta\|_2^2 = \mathbb{E}\left[\left(t(X) - \eta(X)\right)^2\right]$$

Exemple : régression sur un plan d'expérience déterministe

- $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ déterministe ("design fixe")

$$Y = F + \varepsilon \in \mathbb{R}^n \quad \text{avec} \quad F = (\eta(x_1), \ldots, \eta(x_n)) \in \mathbb{R}^n$$

et $\varepsilon_1, \ldots, \varepsilon_n$ indépendantes centrées.

Exemple : régression sur un plan d'expérience déterministe

- $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ déterministe ("design fixe")

$$Y = F + \varepsilon \in \mathbb{R}^n \quad \text{avec} \quad F = (\eta(x_1), \ldots, \eta(x_n)) \in \mathbb{R}^n$$

  et $\varepsilon_1, \ldots, \varepsilon_n$ indépendantes centrées.
- Cas homoscédastique : $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d.

Exemple : régression sur un plan d'expérience déterministe

- $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ déterministe ("design fixe")

$$Y = F + \varepsilon \in \mathbb{R}^n \quad \text{avec} \quad F = (\eta(x_1), \ldots, \eta(x_n)) \in \mathbb{R}^n$$

et $\varepsilon_1, \ldots, \varepsilon_n$ indépendantes centrées.

- Cas homoscédastique : $\varepsilon_1, \ldots, \varepsilon_n$ i.i.d.

- Perte quadratique de $t \in \mathbb{S} = \mathbb{R}^n$:

$$\mathcal{L}_P(t) = \mathbb{E}_Y\left[\frac{1}{n}\|Y - t\|^2\right] = \mathbb{E}_Y\left[\frac{1}{n}\sum_{i=1}^n (Y_i - t_i)^2\right]$$

$$\Rightarrow \quad s^\star = F \quad \text{et} \quad \ell(s^\star, t) = \frac{1}{n}\|F - t\|^2 = \frac{1}{n}\sum_{i=1}^n (\eta(x_i) - t_i)^2$$

Exemple : régression : design fixe vs. design aléatoire

	Design aléatoire	Design fixe
D_n	$(X_i, Y_i)_{1 \leq i \leq n}$ i.i.d. $\sim P$	$Y = F + \varepsilon \in \mathbb{R}^n$
	$(X_{n+1}, Y_{n+1}) \sim P$	$X_{n+1} \sim \mathcal{U}(x_1, \ldots, x_n)$
$\mathbb{S}$	$t : \mathcal{X} \to \mathbb{R}$	$t \in \mathbb{R}^n$
$P\gamma(t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(Y - t(X))^2 \right]$	$E_Y\left[\frac{1}{n} \|Y - t\|^2 \right]$
$s^\star$	$\eta : x \to \mathbb{E}[Y \mid X = x]$	$F = (\eta(x_1), \ldots, \eta(x_n))$
$\ell(s^\star, t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(t(X) - \eta(X))^2 \right]$	$\frac{1}{n} \|F - t\|^2$

avec $\forall x \in \mathbb{R}^n$, $\|x\|^2 = \sum_{i=1}^{n} x_i^2$

Exemple : régression : design fixe vs. design aléatoire

	Design aléatoire	Design fixe
D_n	$(X_i, Y_i)_{1 \leq i \leq n}$ i.i.d. $\sim P$	$Y = F + \varepsilon \in \mathbb{R}^n$
	$(X_{n+1}, Y_{n+1}) \sim P$	$X_{n+1} \sim \mathcal{U}(x_1, \ldots, x_n)$
$\mathbb{S}$	$t : \mathcal{X} \to \mathbb{R}$	$t \in \mathbb{R}^n$
$P\gamma(t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(Y - t(X))^2\right]$	$E_Y\left[\frac{1}{n}\|Y - t\|^2\right]$
$s^\star$	$\eta : x \to \mathbb{E}[Y \mid X = x]$	$F = (\eta(x_1), \ldots, \eta(x_n))$
$\ell(s^\star, t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(t(X) - \eta(X))^2\right]$	$\frac{1}{n}\|F - t\|^2$

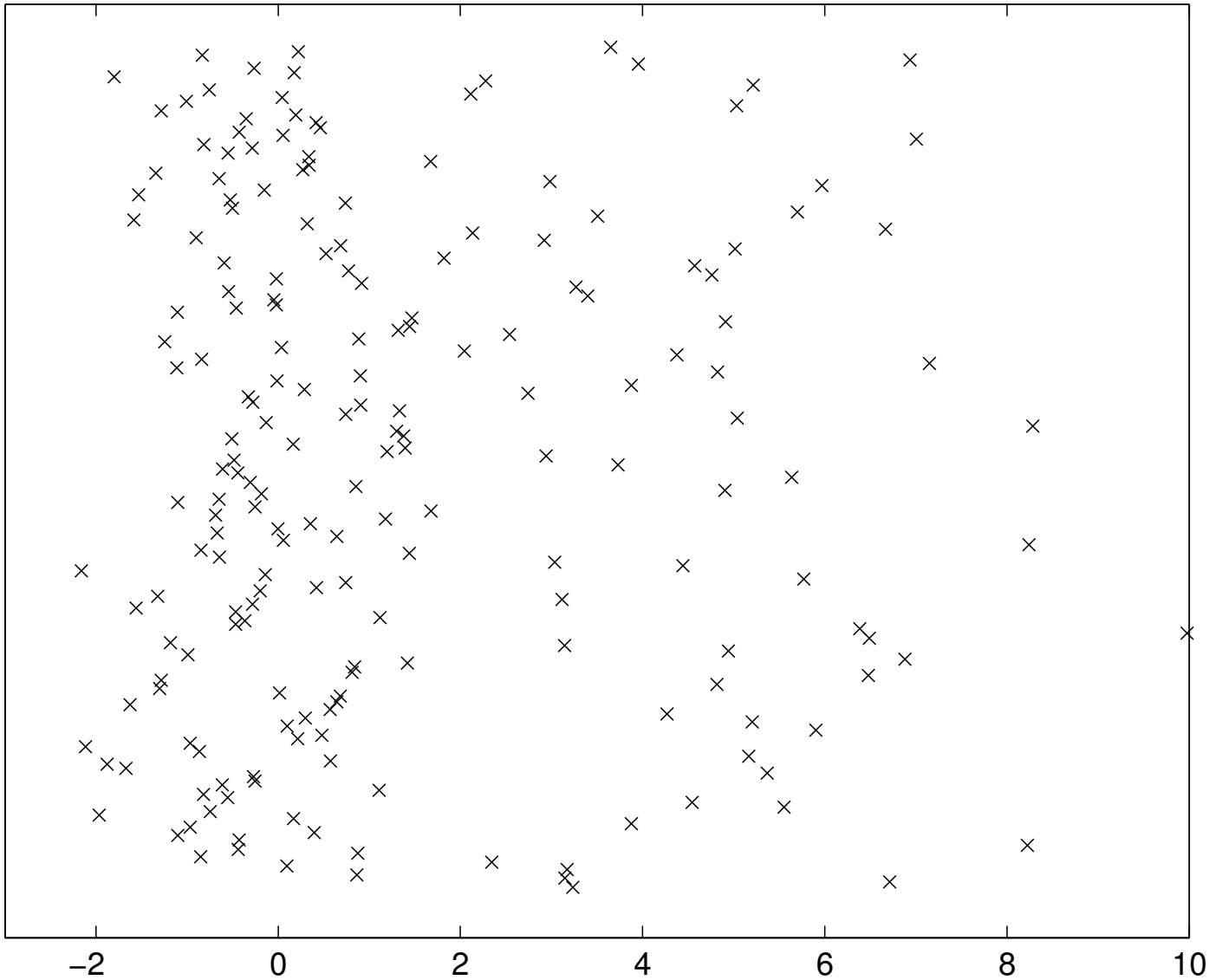
avec $\quad \forall x \in \mathbb{R}^n, \qquad \|x\|^2 = \sum_{i=1}^{n} x_i^2$

Exemple : régression : design fixe vs. design aléatoire

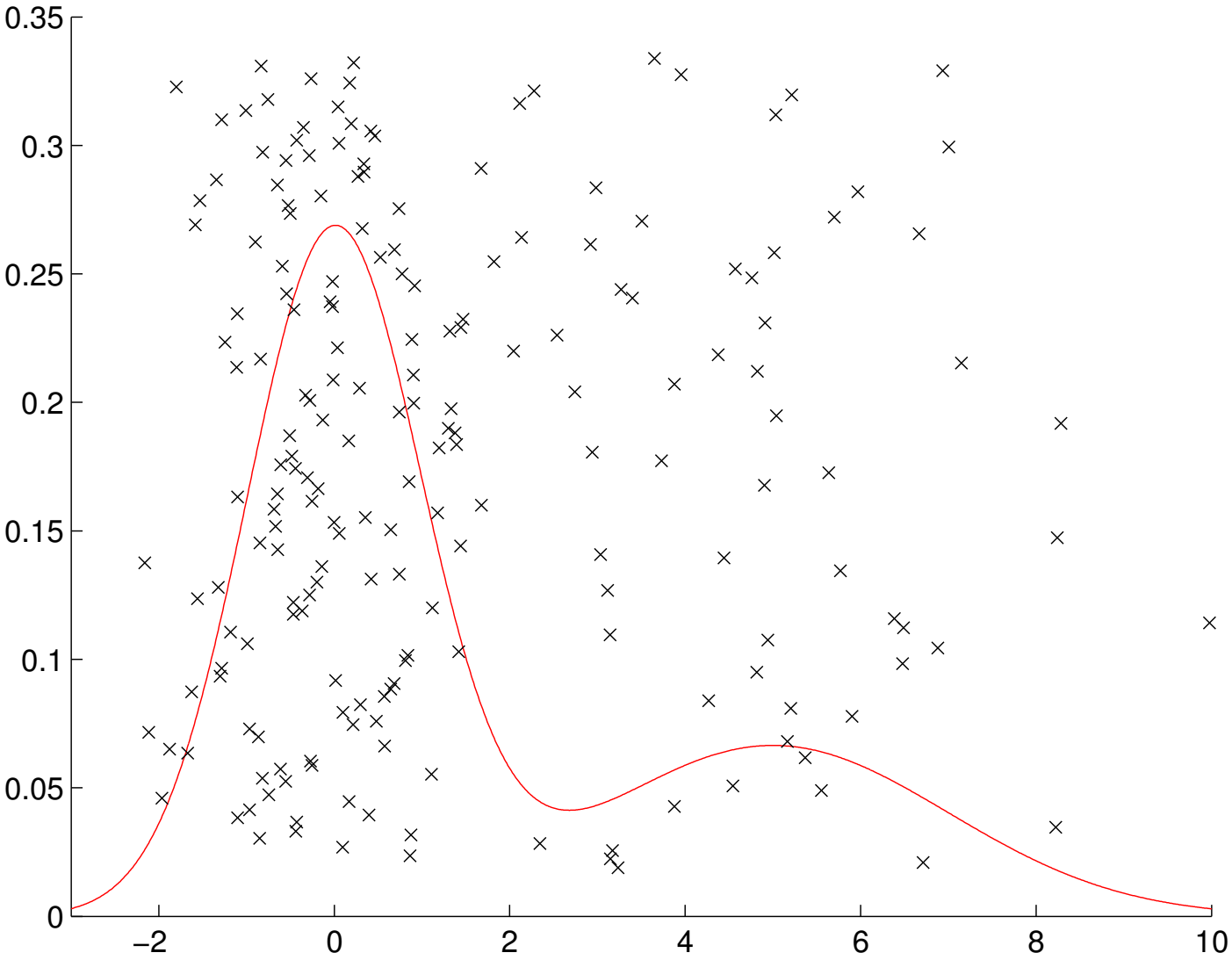
	Design aléatoire	Design fixe
D_n	$(X_i, Y_i)_{1 \leq i \leq n}$ i.i.d. $\sim P$	$Y = F + \varepsilon \in \mathbb{R}^n$
	$(X_{n+1}, Y_{n+1}) \sim P$	$X_{n+1} \sim \mathcal{U}(x_1, \ldots, x_n)$
$\mathbb{S}$	$t : \mathcal{X} \to \mathbb{R}$	$t \in \mathbb{R}^n$
$P\gamma(t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(Y - t(X))^2\right]$	$E_Y\left[\frac{1}{n}\|Y - t\|^2\right]$
$s^\star$	$\eta : x \to \mathbb{E}[Y \mid X = x]$	$F = (\eta(x_1), \ldots, \eta(x_n))$
$\ell(s^\star, t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(t(X) - \eta(X))^2\right]$	$\frac{1}{n}\|F - t\|^2$

avec $\quad \forall x \in \mathbb{R}^n, \qquad \|x\|^2 = \sum_{i=1}^{n} x_i^2$

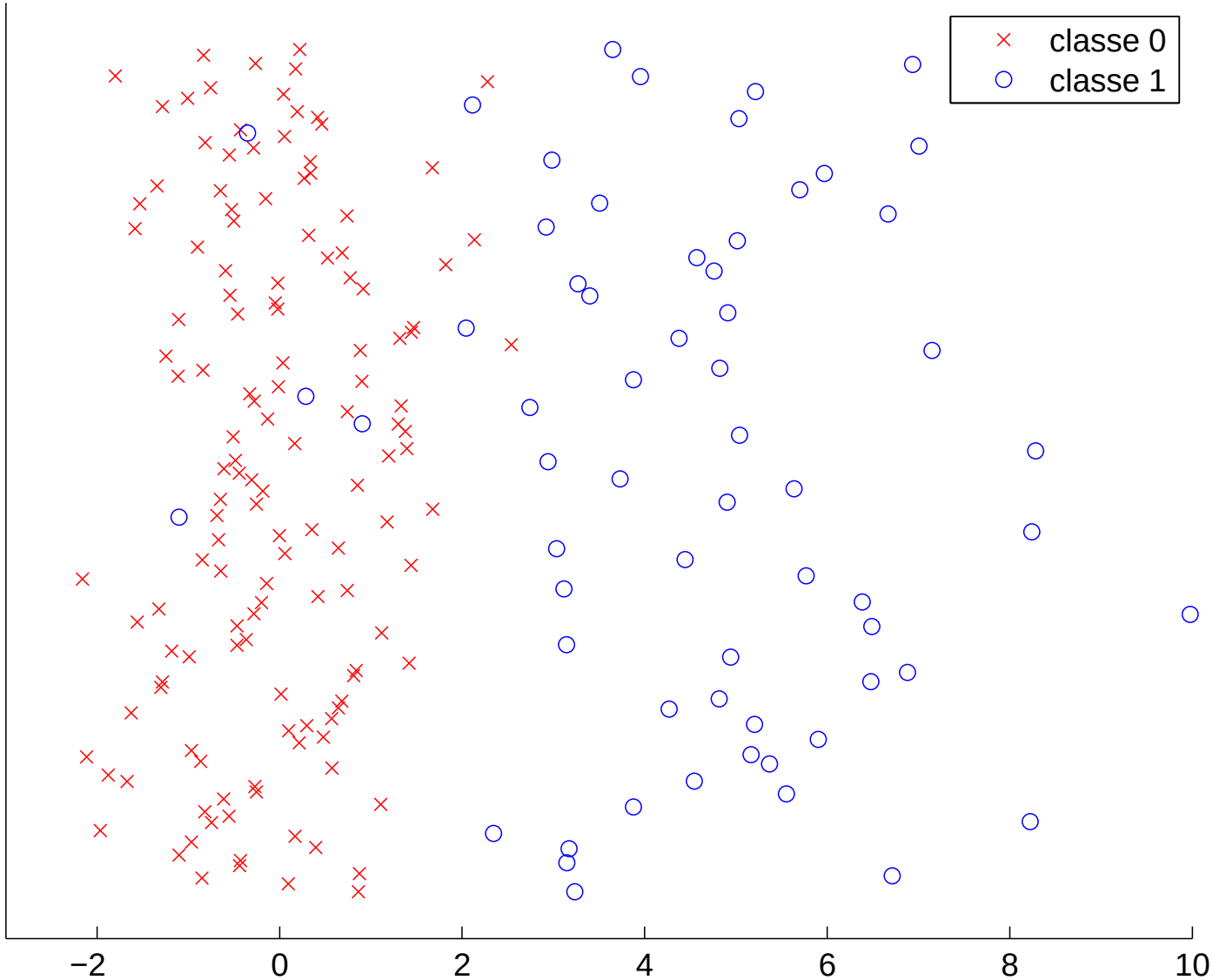
Exemple : estimation de densité ($\Xi = \mathbb{R}$) : données



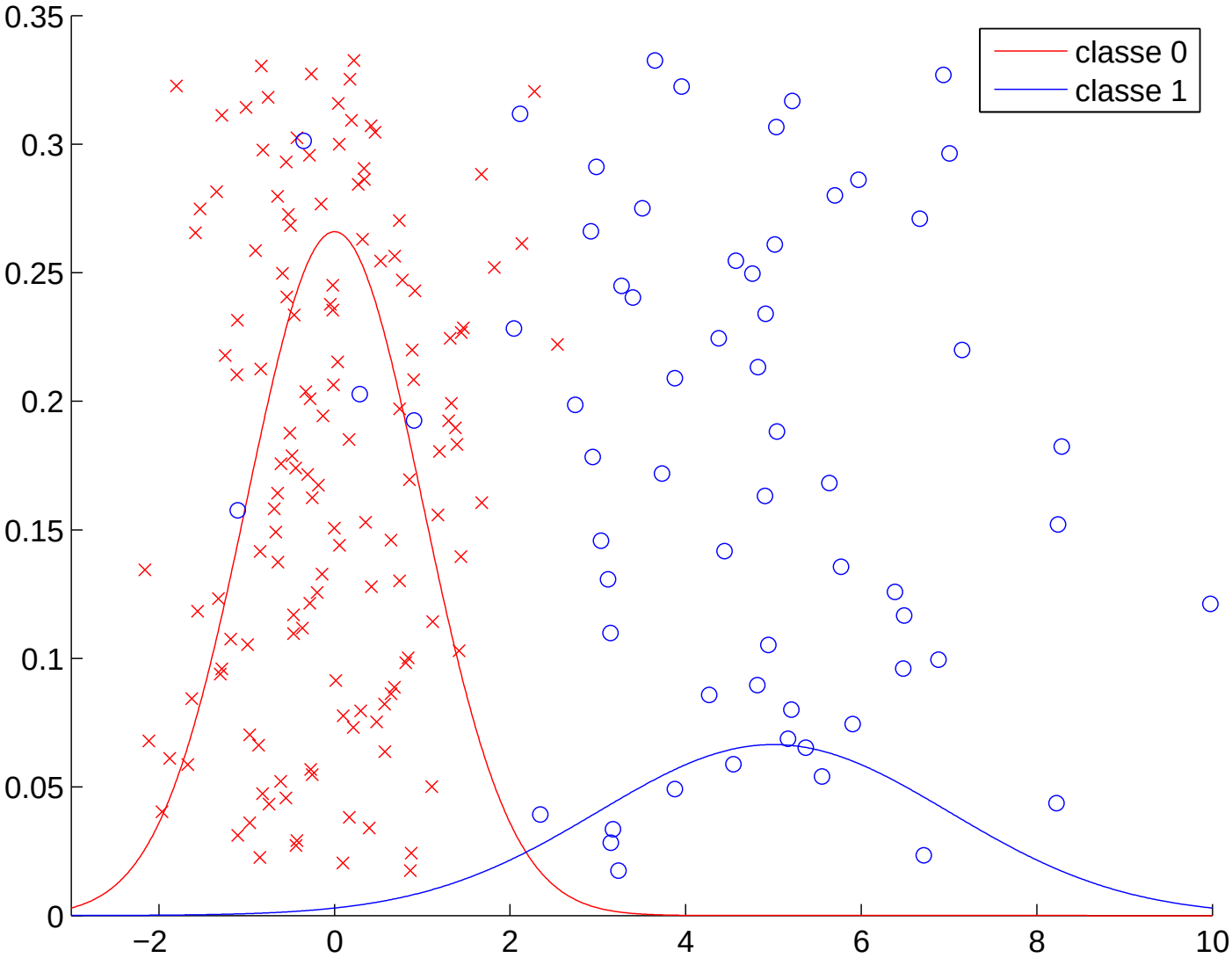
Exemple : estimation de densité ($\Xi = \mathbb{R}$) : données et cible



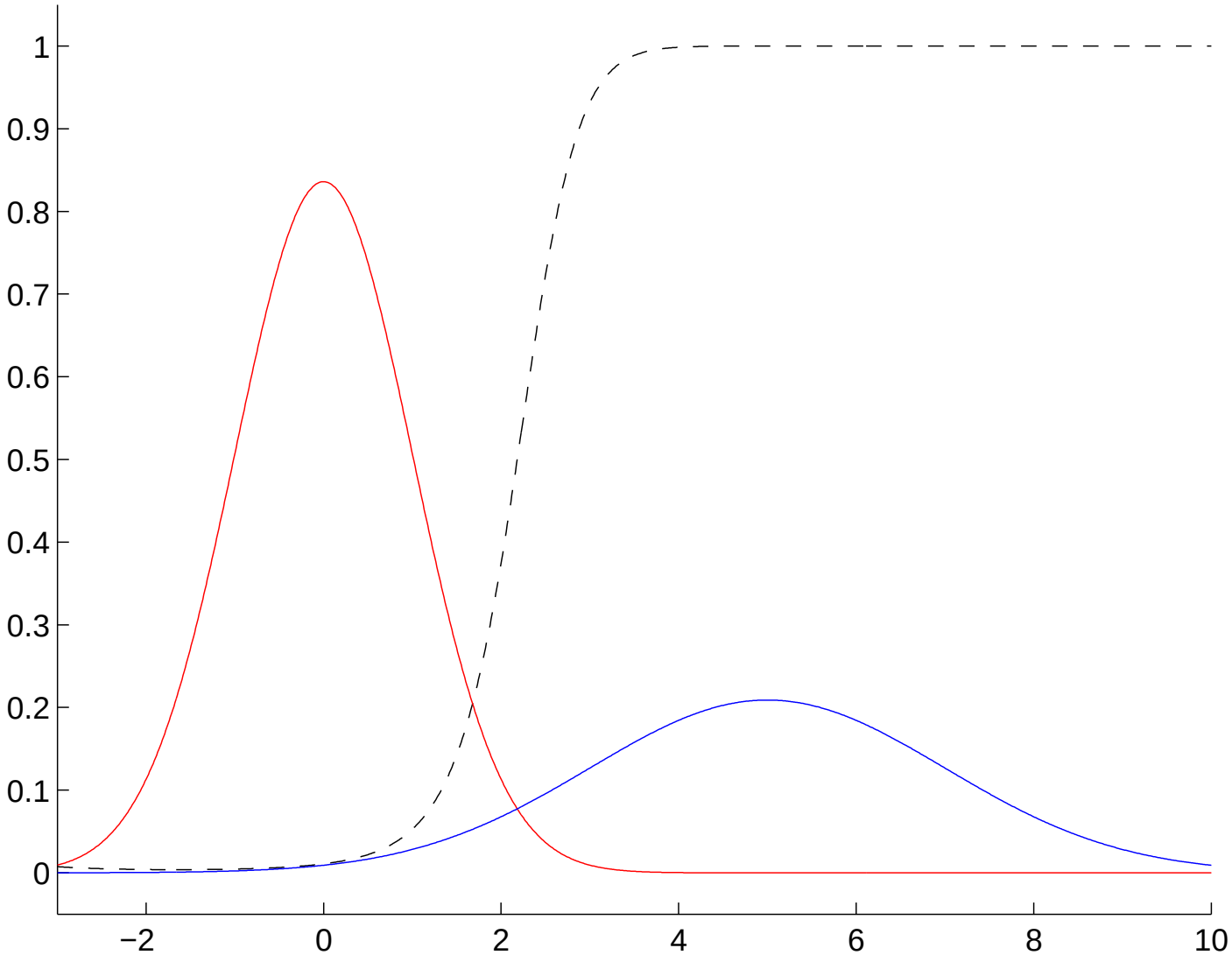
Exemple : classification (prédiction, $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{0, 1\}$)



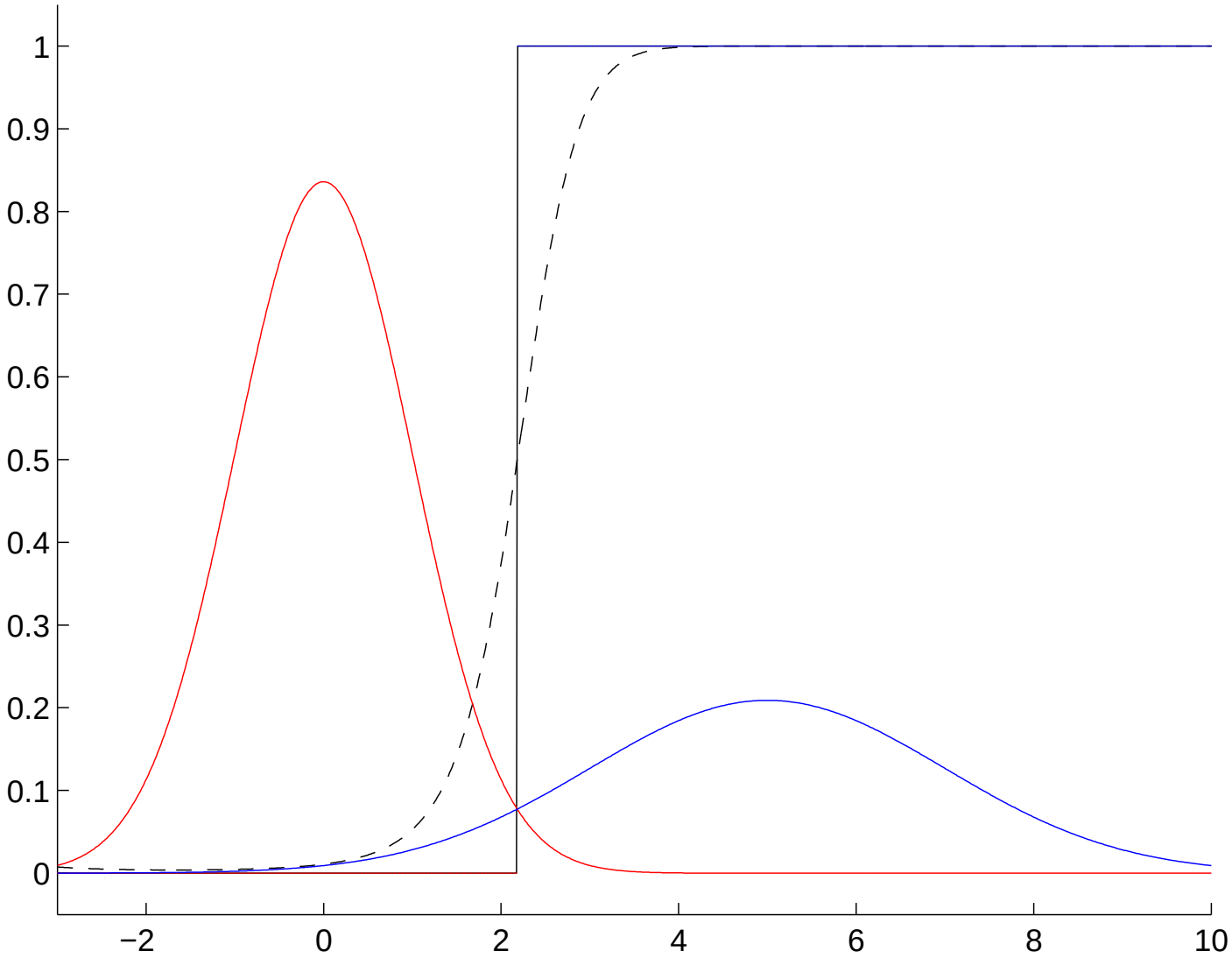
Exemple : classification (prédiction, $\mathcal{X} = \mathbb{R}$ et $\mathcal{Y} = \{0, 1\}$)



Exemple : classification (prédiction, $\mathcal{X} = \mathbb{R}$ et $\mathcal{Y} = \{0, 1\}$)



Exemple : classification (prédiction, $\mathcal{X} = \mathbb{R}$ et $\mathcal{Y} = \{0, 1\}$)



Plan

Qu'est-ce qu'un estimateur ?

- Algorithme statistique ou Règle d'apprentissage :

$$\mathcal{A} : \bigcup_{n\in\mathbb{N}} \Xi^n \mapsto \mathbb{S}$$

$$\text{échantillon } D_n = (\xi_1, \ldots, \xi_n) \mapsto \mathcal{A}(D_n)$$

- $\mathcal{A}(D_n) = \widehat{s}^{\mathcal{A}}(D_n) = \widehat{s}(D_n) \in \mathbb{S}$ est un estimateur de $s^\star$

Qu'est-ce qu'un estimateur ?

- Algorithme statistique ou Règle d'apprentissage :

$$\mathcal{A} : \bigcup_{n\in\mathbb{N}} \Xi^n \mapsto \mathbb{S}$$

$$\text{échantillon } D_n = (\xi_1, \ldots, \xi_n) \mapsto \mathcal{A}(D_n)$$

- $\mathcal{A}(D_n) = \widehat{s}^{\mathcal{A}}(D_n) = \widehat{s}(D_n) \in \mathbb{S}$ est un estimateur de $s^\star$

- Remarque : $P\gamma\left(\widehat{s}^{\mathcal{A}}(D_n)\right)$ et $\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right)$ sont des quantités aléatoires

Qu'est-ce qu'un estimateur ?

- Algorithme statistique ou Règle d'apprentissage :

$$\mathcal{A} : \bigcup_{n\in\mathbb{N}} \Xi^n \mapsto \mathbb{S}$$

$$\text{échantillon } D_n = (\xi_1, \ldots, \xi_n) \mapsto \mathcal{A}(D_n)$$

- $\mathcal{A}(D_n) = \widehat{s}^{\mathcal{A}}(D_n) = \widehat{s}(D_n) \in \mathbb{S}$ est un estimateur de $s^\star$

- Remarque : $P\gamma\left(\widehat{s}^{\mathcal{A}}(D_n)\right)$ et $\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right)$ sont des quantités aléatoires

- Risque de $\widehat{s}^{\mathcal{A}}$:

$$\mathbb{E}_{D_n\sim P^{\otimes n}}\left[P\gamma\left(\widehat{s}^{\mathcal{A}}(D_n)\right)\right] = \mathcal{R}(\mathcal{A}, n)$$

- Excès de risque de $\widehat{s}^{\mathcal{A}}$:

$$\mathbb{E}_{D_n\sim P^{\otimes n}}\left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right)\right] = \mathcal{R}(\mathcal{A}, n) - P\gamma\left(s^\star\right)$$

Consistance, universalité, vitesse d'apprentissage

- Consistance (à P fixée) : $\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \to 0$ quand $n \to +\infty$

Consistance, universalité, vitesse d'apprentissage

- Consistance (à P fixée) : $\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \to 0$ quand $n \to +\infty$
- Consistance universelle :
  $\sup_P \left\{ \overline{\lim}_{n\to\infty} \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = 0$

Consistance, universalité, vitesse d'apprentissage

- Consistance (à P fixée) : $\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \to 0$ quand $n \to +\infty$
- Consistance universelle :
  $\sup_P \left\{ \overline{\lim}_{n\to\infty} \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = 0$
- Consistance universelle uniforme :
  $\overline{\lim}_{n\to\infty} \sup_P \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = 0$ (vitesse d'apprentissage uniforme sur toutes les lois).

Consistance, universalité, vitesse d'apprentissage

- Consistance (à P fixée) : $\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \to 0$ quand $n \to +\infty$
- Consistance universelle :
  $\sup_P \left\{ \overline{\lim}_{n\to\infty} \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = 0$
- Consistance universelle uniforme :
  $\overline{\lim}_{n\to\infty} \sup_P \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = 0$ (vitesse d'apprentissage uniforme sur toutes les lois).
- "No Free Lunch" (cf. Devroye, Györfi & Lugosi, 1996) :
  En classification binaire avec $\mathcal{X}$ infini, $\forall \mathcal{A}$, $\forall n \geq 1$,

$$\sup_P \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = \frac{1}{2}$$

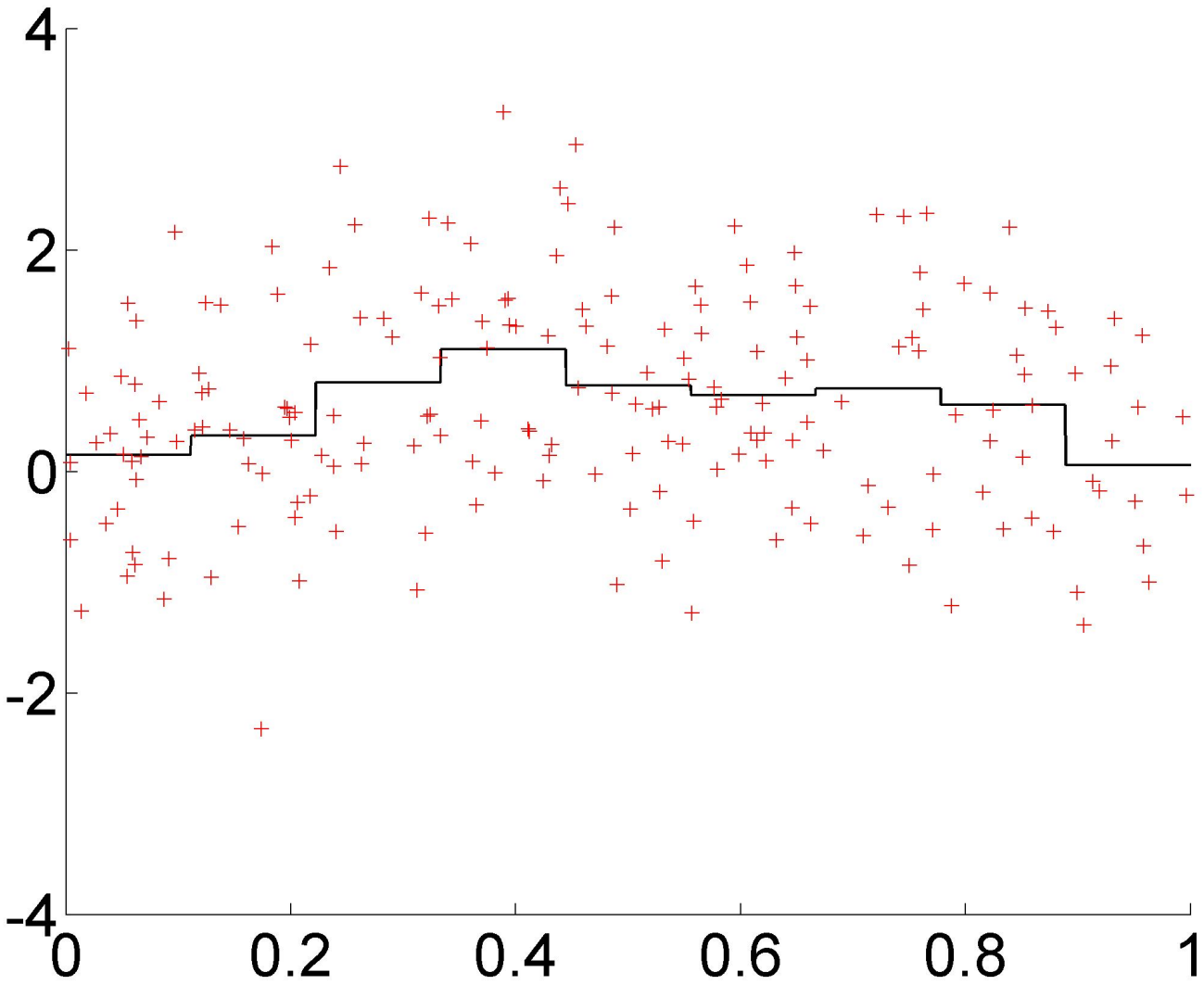
Consistance, universalité, vitesse d'apprentissage

- Consistance (à P fixée) : $\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \to 0$ quand $n \to +\infty$
- Consistance universelle :
  $\sup_P \left\{ \overline{\lim}_{n\to\infty} \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = 0$
- Consistance universelle uniforme :
  $\overline{\lim}_{n\to\infty} \sup_P \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = 0$ (vitesse d'apprentissage uniforme sur toutes les lois).
- "No Free Lunch" (cf. Devroye, Györfi & Lugosi, 1996) :
  En classification binaire avec $\mathcal{X}$ infini, $\forall \mathcal{A}$, $\forall n \geq 1$,

$$\sup_P \left\{ \mathbb{E}_{D_n \sim P^{\otimes n}} \left[\ell\left(s^\star, \widehat{s}^{\mathcal{A}}(D_n)\right) \right] \right\} = \frac{1}{2}$$

$\Rightarrow$ pour avoir une vitesse d'apprentissage uniforme, il faut faire des hypothèses sur P

Estimateur des moindres carrés : régressogramme



Estimateur des moindres carrés

- Cadre : Régression, contraste des moindres carrés
  $\gamma(t;(x,y)) = (t(x)-y)^2$
- Approche naturelle : minimiser un estimateur de
  $P\gamma(t) = \mathbb{E}\left[(t(X)-Y)^2\right]$

Estimateur des moindres carrés

- Cadre : Régression, contraste des moindres carrés
  $\gamma(t;(x,y)) = (t(x)-y)^2$
- Approche naturelle : minimiser un estimateur de
  $P\gamma(t) = \mathbb{E}\left[\left(t(X)-Y\right)^2\right]$
- Critère des moindres carrés :

$$P_n\gamma(t) = \frac{1}{n}\sum_{i=1}^{n}\left(t(X_i)-Y_i\right)^2 \qquad \text{avec} \quad P_n = \frac{1}{n}\sum_{i=1}^{n}\delta_{\xi_i}$$

$$\forall t \in \mathbb{S}, \quad \mathbb{E}\left[P_n\gamma(t)\right] = P\gamma(t)$$

Estimateur des moindres carrés

- Cadre : Régression, contraste des moindres carrés
  $\gamma(t;(x,y)) = (t(x)-y)^2$
- Approche naturelle : minimiser un estimateur de
  $P\gamma(t) = \mathbb{E}\left[(t(X)-Y)^2 \right]$
- Critère des moindres carrés :

$$P_n\gamma(t) = \frac{1}{n}\sum_{i=1}^{n} (t(X_i)-Y_i)^2 \qquad \text{avec} \quad P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i}$$

$$\forall t \in \mathbb{S}, \quad \mathbb{E}\left[P_n\gamma(t)\right] = P\gamma(t)$$

- Modèle : $S \subset \mathbb{S} \Rightarrow$ Estimateur des moindres carrés sur S :

$$\widehat{s}_S \in \arg\min_{t\in S} \left\{ P_n\gamma(t) \right\} = \arg\min_{t\in S} \left\{ \frac{1}{n}\sum_{i=1}^{n} (t(X_i)-Y_i)^2 \right\}$$

Exemples de modèles (régression)

- histogrammes sur une partition Λ de $\mathcal{X}$
  $\Rightarrow$ l'estimateur des moindres carrés (régressogramme) s'écrit

$$\widehat{s}_m = \sum_{\lambda \in \Lambda} \widehat{\beta}_\lambda \mathbb{1}_\lambda \qquad \widehat{\beta}_\lambda = \frac{1}{\text{Card}\left\{ X_i \in \lambda \right\}} \sum_{X_i \in \lambda} Y_i \ .$$

- décomposition sur un sous-ensemble d'une base orthogonale de $L^2(\mu)$ (Fourier, ondelettes, etc.)
- sélection de variables : $X_i = \left(X_i^{(1)}, \ldots, X_i^{(p)} \right) \in \mathbb{R}^p$ contient p variables pouvant expliquer (linéairement) Y

$$\forall m \subset \{1, \ldots, p\} \ , \quad S_m = \left\{ t : x \in \mathcal{X} \mapsto \sum_{j \in m} \beta_j x^{(j)} \text{ t.q. } \beta \in \mathbb{R}^m \right\}$$

Régression : design fixe vs. design aléatoire

	Design aléatoire	Design fixe
D_n	$(X_i, Y_i)_{1 \leq i \leq n}$ i.i.d. $\sim P$	$Y = F + \varepsilon \in \mathbb{R}^n$
	$(X_{n+1}, Y_{n+1}) \sim P$	$X_{n+1} \sim \mathcal{U}(x_1, \ldots, x_n)$
$\mathbb{S}$	$t : \mathcal{X} \to \mathbb{R}$	$t \in \mathbb{R}^n$
$P\gamma(t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(Y - t(X))^2 \right]$	$E_Y\left[\frac{1}{n} \|Y - t\|^2 \right]$
$s^\star$	$\eta : x \to \mathbb{E}[Y \mid X = x]$	$F = (\eta(x_1), \ldots, \eta(x_n))$
$\ell(s^\star, t)$	$\mathbb{E}_{(X,Y)\sim P}\left[(t(X) - \eta(X))^2 \right]$	$\frac{1}{n} \|F - t\|^2$
	$P_n\gamma(t) = \frac{1}{n}\sum_{i=1}^n (Y_i - t(X_i))^2$	$\frac{1}{n} \|Y - t\|^2$

avec $\forall x \in \mathbb{R}^n$, $\|x\|^2 = \sum_{i=1}^n x_i^2$

Estimateurs par minimum de contraste

- Risque empirique (ou contraste empirique)

$$P_n\gamma(t) = \frac{1}{n}\sum_{i=1}^{n} \gamma\left(t; \xi_i\right)$$

- $\forall t \in \mathbb{S}$, $\mathbb{E}\left[P_n\gamma(t)\right] = P\gamma(t)$
- Estimateur par minimum de contraste (minimisation du risque empirique) sur un modèle $S \subset \mathbb{S}$:

$$\widehat{s}_S \in \arg\min_{t \in S} P_n\gamma(t) \quad \text{avec} \quad P_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{\xi_i}$$

- Exemple supplémentaire : maximum de vraisemblance en estimation de densité : $\gamma(t;\xi) = -\ln(t(\xi))$

Estimateur régularisé : Régression ridge à noyau

- Idée : contrôler la norme de l'estimateur dans un espace fonctionnel $\mathcal{F}$

Estimateur régularisé : Régression ridge à noyau

- Idée : contrôler la norme de l'estimateur dans un espace fonctionnel $\mathcal{F}$
- $\mathcal{F} \subset \mathbb{S}$ est l'espace de Hilbert à noyau reproduisant (RKHS) associé à un noyau défini positif $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \| f \|_{\mathcal{F}}^2 \right\}$$

Estimateur régularisé : Régression ridge à noyau

- Idée : contrôler la norme de l'estimateur dans un espace fonctionnel $\mathcal{F}$
- $\mathcal{F} \subset \mathbb{S}$ est l'espace de Hilbert à noyau reproduisant (RKHS) associé à un noyau défini positif $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \| f \|_{\mathcal{F}}^2 \right\}$$

- Théorème du représentant $\Rightarrow \widehat{f} = \sum_{i=1}^{n} \widehat{\alpha}_i k(X_i, \cdot)$
- Design fixe : $(\widehat{f}(x_i))_{1 \leq i \leq n} = \widehat{F} = K(K + n\lambda I_n)^{-1} Y$

Estimateur régularisé : Régression ridge à noyau

- Idée : contrôler la norme de l'estimateur dans un espace fonctionnel $\mathcal{F}$
- $\mathcal{F} \subset \mathbb{S}$ est l'espace de Hilbert à noyau reproduisant (RKHS) associé à un noyau défini positif $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$

$$\widehat{f} \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda \| f \|_{\mathcal{F}}^2 \right\}$$

- Théorème du représentant $\Rightarrow \widehat{f} = \sum_{i=1}^{n} \widehat{\alpha}_i k(X_i, \cdot)$
- Design fixe : $(\widehat{f}(x_i))_{1 \leq i \leq n} = \widehat{F} = K(K + n\lambda I_n)^{-1} Y$
- Un exemple d'estimateur linéaire $\widehat{F} = AY$
  Autres exemples : moindres carrés, k-plus proches voisins (en régression), Nadaraya-Watson, etc.

D'autres estimateurs régularisés

- Machine à vecteurs de support (SVM) en classification :

$$\arg\min_{f\in\mathcal{F}} \left\{ P_n \gamma_{\text{hinge}}(f) + \lambda \|f\|_{\mathcal{F}}^2 \right\}$$

- Lasso (Tibshirani 1996) : régression, $\mathcal{X} = \mathbb{R}^p$

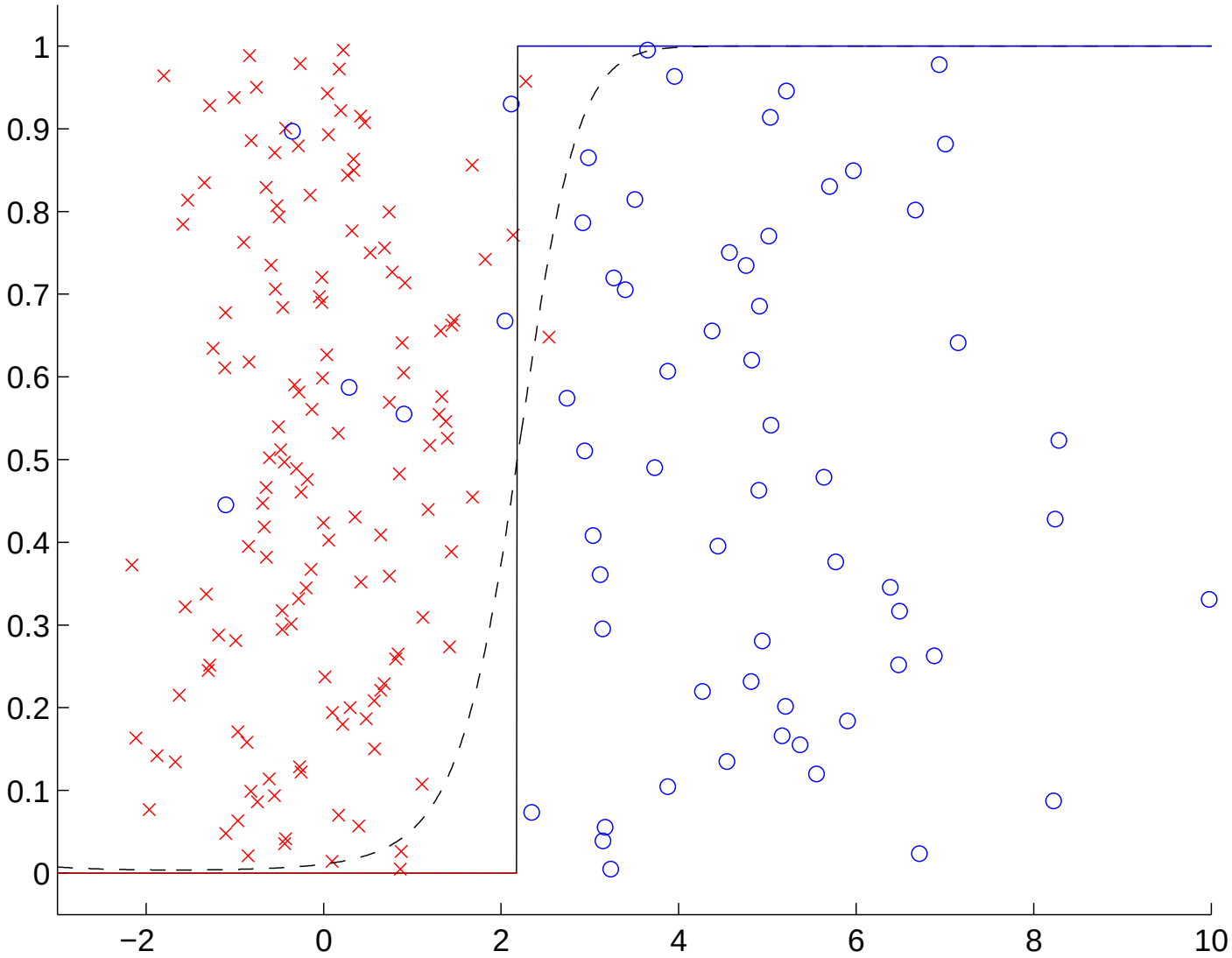
$$\arg\min_{w\in\mathbb{R}^p} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(Y_i - w^\top X_i \right)^2 + \lambda \|w\|_1 \right\}$$

- Lasso structuré

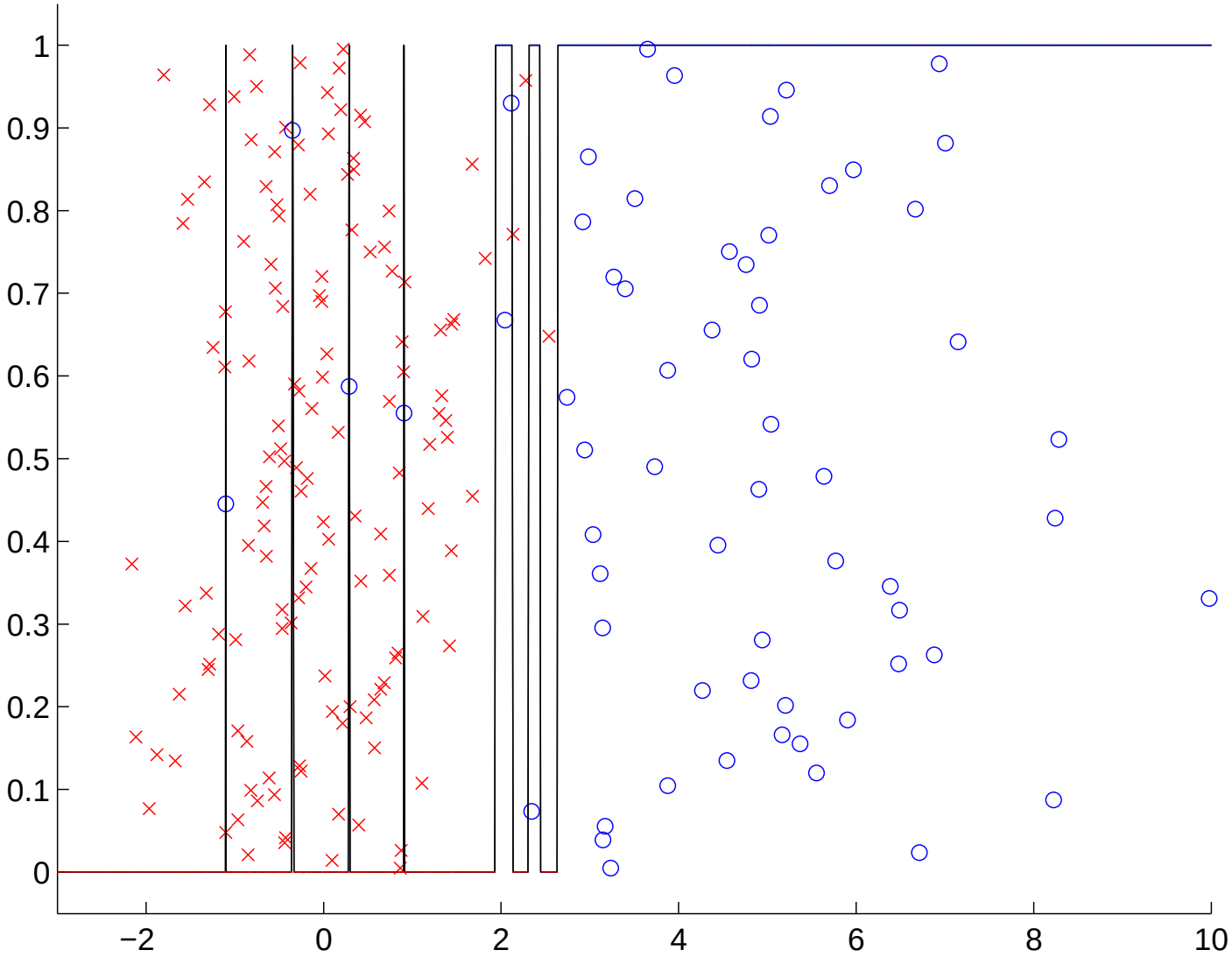
$$\arg\min_{w\in\mathbb{R}^p} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(Y_i - w^\top X_i \right)^2 + \lambda \Omega(w) \right\}$$

par exemple, group Lasso (Yuan & Lin 2006) :
$\Omega(w) = \sum_{g\in\mathcal{G}} \|w_g\|_2$.

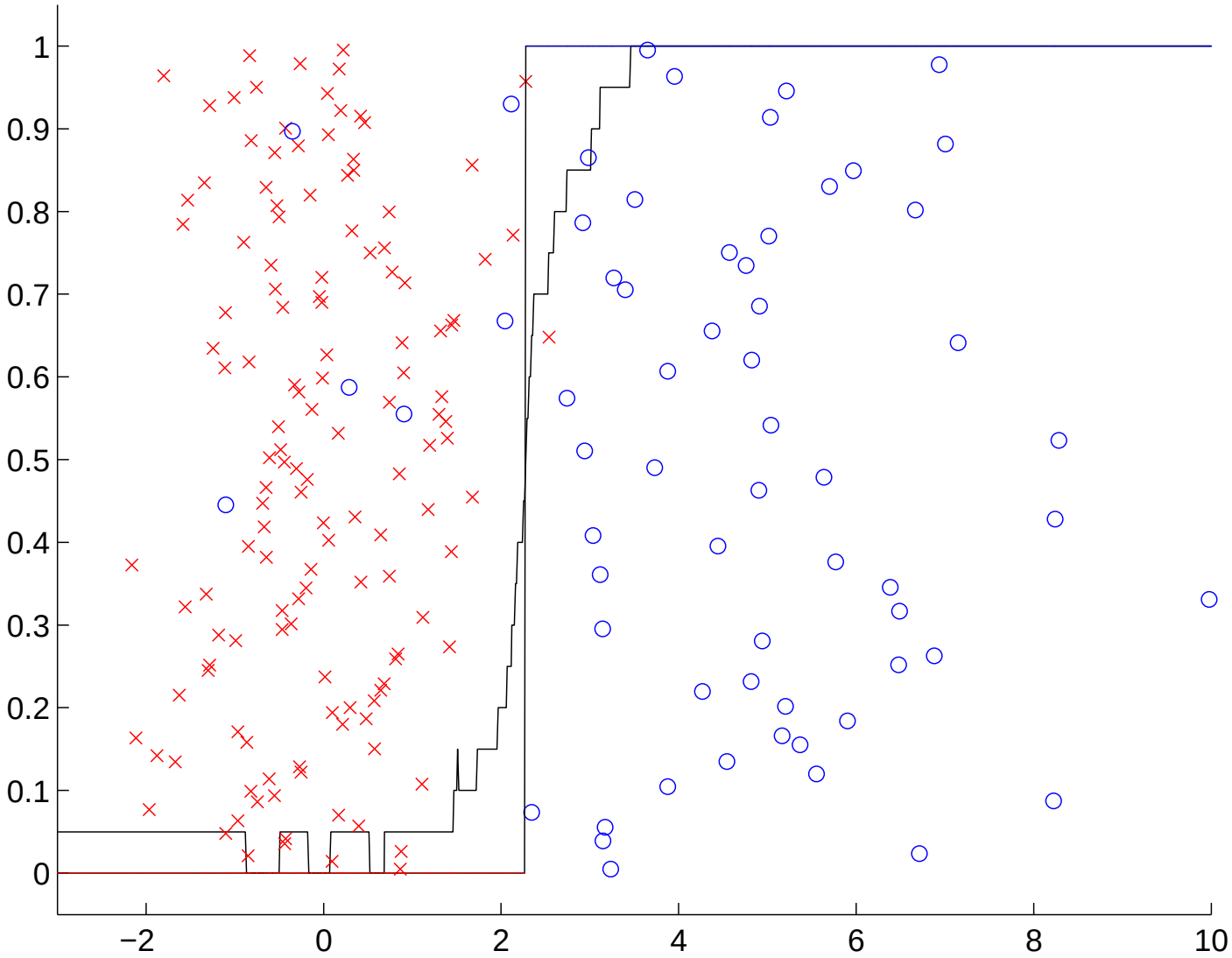
Classification ($\mathcal{X} = \mathbb{R}$)



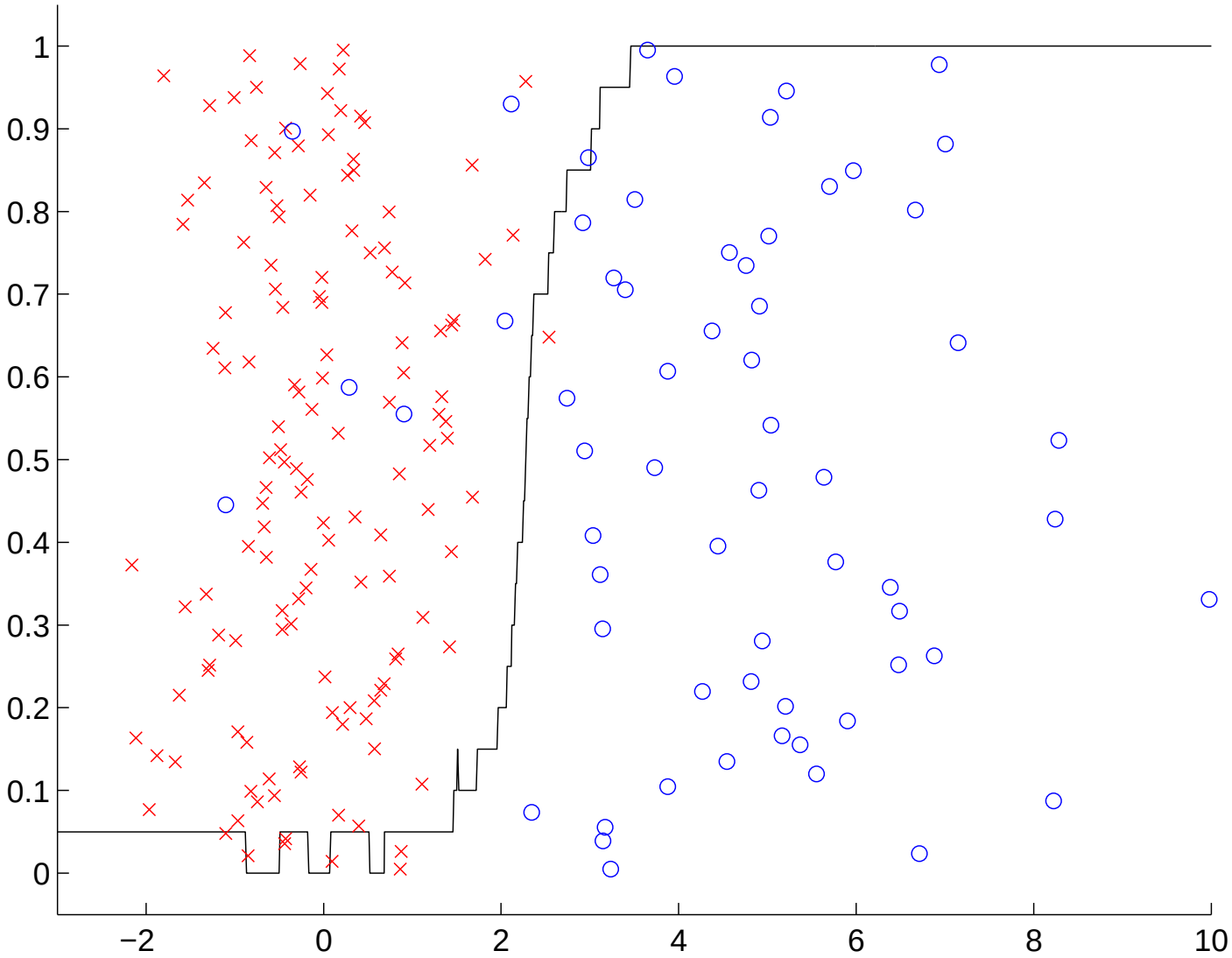
Règle du plus proche voisin



Règle des k plus proches voisins ($k = 20$)

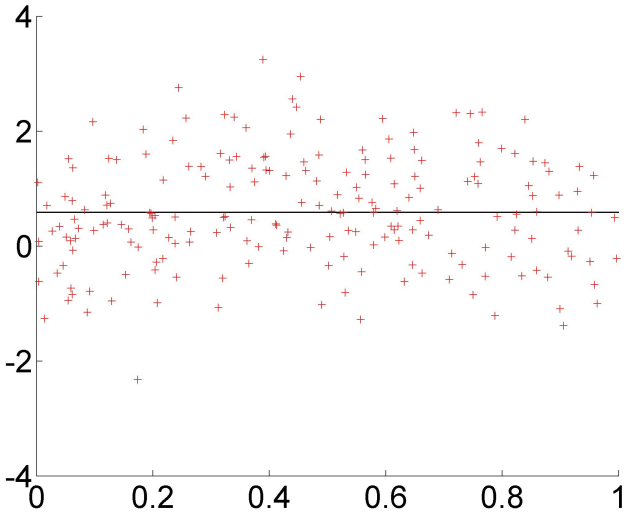


Règle des 20 plus proches voisins : pour la régression

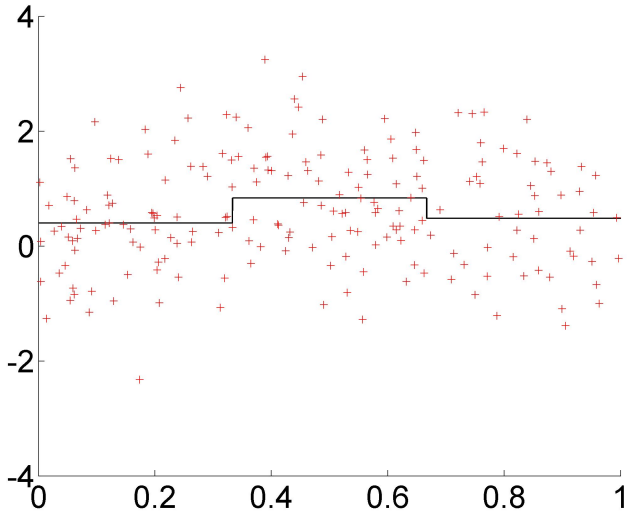


Plan

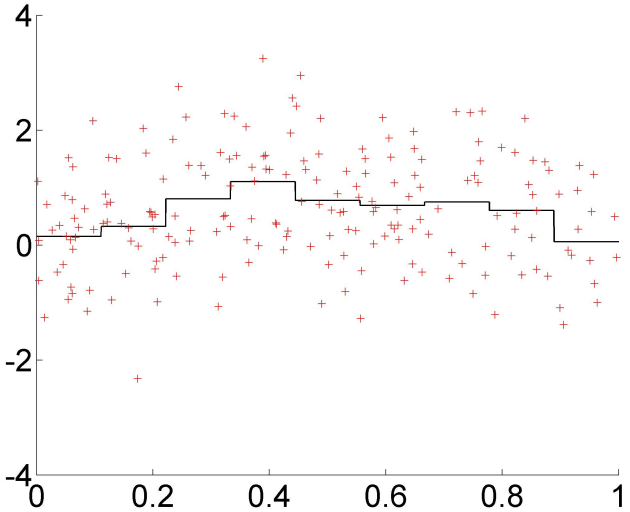
Comment choisir la dimension D ?



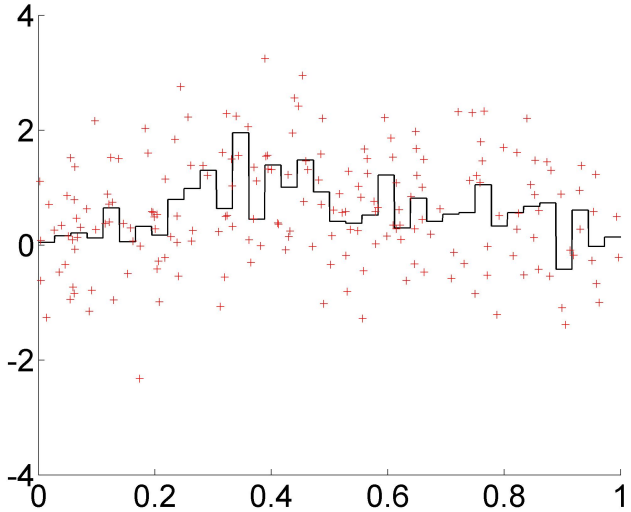
$D = 1$



$D = 3$

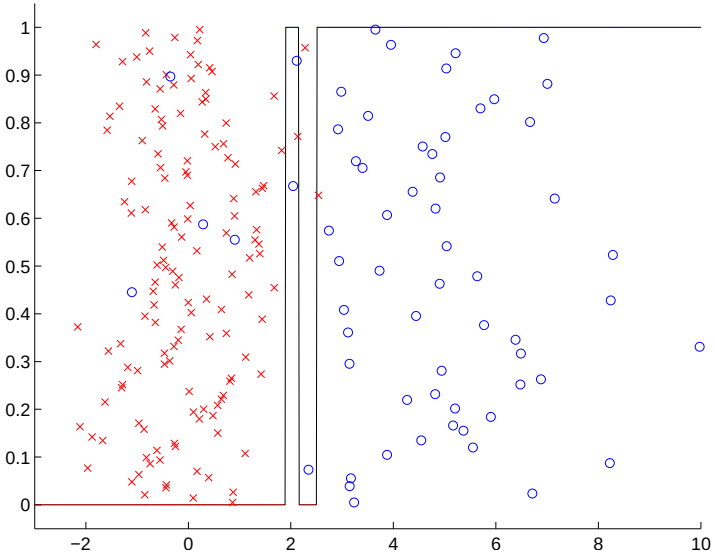


$D = 9$

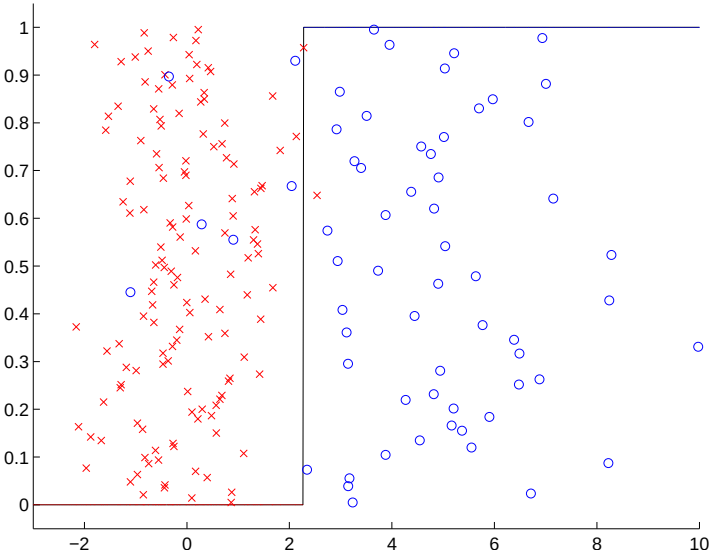


$D = 36$

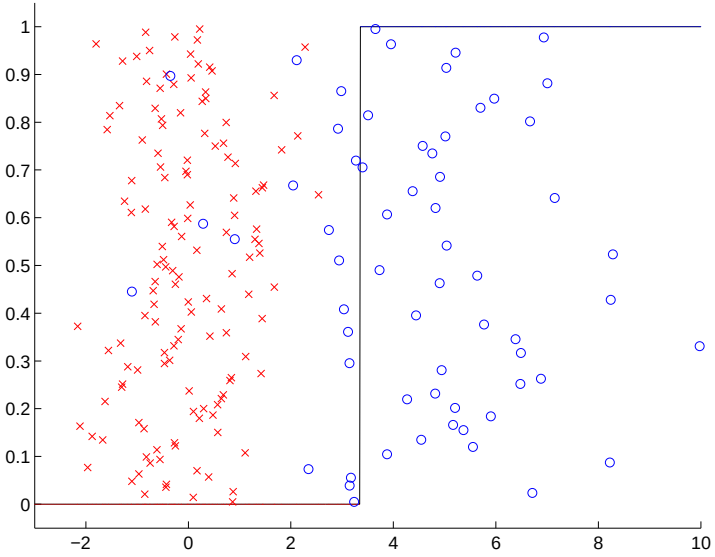
Comment choisir le nombre k de voisins ?



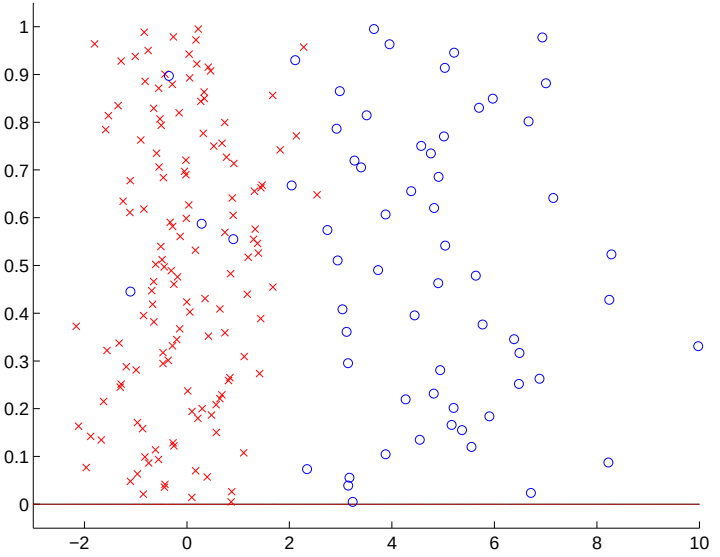
$k = 3$



$k = 20$



$k = 100$



$k = 200$

Problème de la sélection d'estimateur

- Famille de règles d'apprentissage candidates : $(\mathcal{A}_m)_{m \in \mathcal{M}}$
- Problème : comment choisir parmi les estimateurs $(\mathcal{A}_m(D_n))_{m \in \mathcal{M}} = (\widehat{s}_m(D_n))_{m \in \mathcal{M}}$?

Problème de la sélection d'estimateur

- Famille de règles d'apprentissage candidates : $(\mathcal{A}_m)_{m \in \mathcal{M}}$
- Problème : comment choisir parmi les estimateurs $(\mathcal{A}_m(D_n))_{m \in \mathcal{M}} = (\widehat{s}_m(D_n))_{m \in \mathcal{M}}$?

- Exemples :
  - choix de modèles
  - calibration (choix de k ou d'une distance pour k-ppv, choix du paramètre de régularisation, choix d'un noyau, etc.)
  - choix entre des méthodes de nature différente, e.g., k-ppv et SVM

Objectif d'estimation ou de prédiction

- Objectif principal : trouver $\widehat{m}$ tel que $\ell\left(s^\star, \widehat{s}_{\widehat{m}(D_n)}(D_n)\right)$ est minimale
- Oracle : $m^\star \in \arg\min_{m \in \mathcal{M}_n} \left\{ \ell\left(s^\star, \widehat{s}_m(D_n)\right) \right\}$

Objectif d'estimation ou de prédiction

- Objectif principal : trouver $\widehat{m}$ tel que $\ell\left(s^\star, \widehat{s}_{\widehat{m}(D_n)}(D_n)\right)$ est minimale
- Oracle : $m^\star \in \arg\min_{m \in \mathcal{M}_n} \left\{ \ell\left(s^\star, \widehat{s}_m(D_n)\right) \right\}$
- Inégalité oracle (en espérance ou avec grande probabilité) :

$$\ell\left(s^\star, \widehat{s}_{\widehat{m}}\right) \leq C \inf_{m \in \mathcal{M}_n} \left\{ \ell\left(s^\star, \widehat{s}_m(D_n)\right) \right\} + R_n$$

- Non-asymptotique : tous les paramètres peuvent varier avec n, à commencer par la collection d'estimateurs $\mathcal{M} = \mathcal{M}_n$

Objectif d'estimation ou de prédiction

- Objectif principal : trouver $\widehat{m}$ tel que $\ell\left(s^\star, \widehat{s}_{\widehat{m}(D_n)}(D_n)\right)$ est minimale
- Oracle : $m^\star \in \arg\min_{m \in \mathcal{M}_n} \left\{ \ell\left(s^\star, \widehat{s}_m(D_n)\right) \right\}$
- Inégalité oracle (en espérance ou avec grande probabilité) :

$$\ell\left(s^\star, \widehat{s}_{\widehat{m}}\right) \leq C \inf_{m \in \mathcal{M}_n} \left\{ \ell\left(s^\star, \widehat{s}_m(D_n)\right) \right\} + R_n$$

- Non-asymptotique : tous les paramètres peuvent varier avec n, à commencer par la collection d'estimateurs $\mathcal{M} = \mathcal{M}_n$
- Adaptation (e.g., au sens minimax) à la régularité de $s^\star$, aux variations du bruit $\mathbb{E}\left[\varepsilon^2 \,\middle|\, X\right]$, etc. (pourvu que $(\mathcal{A}_m)_{m \in \mathcal{M}_n}$ est bien choisie)

Objectif d'identification

- Hypothèse supplémentaire (cas de la sélection de modèles) : $s^\star \in S_{m_0}$ pour un certain $m_0 \in \mathcal{M}_n$
- Objectif supplémentaire : sélectionner $\widehat{m} = m_0$ avec une probabilité maximale
- Consistance :

$$\mathbb{P}\left(\widehat{m} = m_0\right) \xrightarrow[n\to\infty]{} 1 \ .$$

Objectif d'identification

- Hypothèse supplémentaire (cas de la sélection de modèles) : $s^\star \in S_{m_0}$ pour un certain $m_0 \in \mathcal{M}_n$
- Objectif supplémentaire : sélectionner $\widehat{m} = m_0$ avec une probabilité maximale
- Consistance :

$$\mathbb{P}\left(\widehat{m} = m_0\right) \xrightarrow[n\to\infty]{} 1 \quad .$$

- Estimation *et* identification ?
  Objectifs incompatibles en général (Yang, 2005)
  Parfois possible de concilier les deux (e.g., Yang, 2005 ; van Erven et al., 2008)

Sélection de modèles : biais et variance

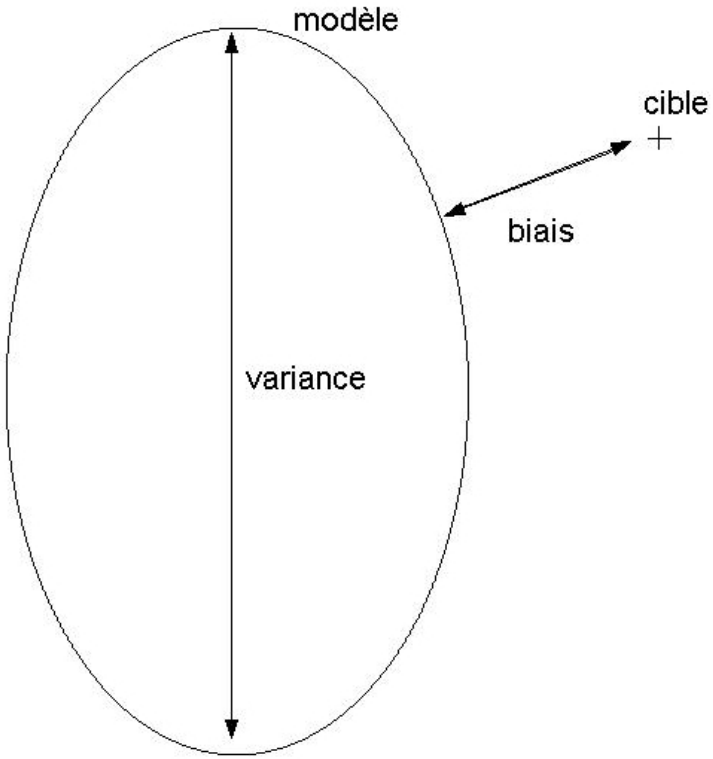
$$\mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m(D_n)\right)\right] = \text{Biais} + \text{Variance}$$

Biais ou Erreur d'approximation

$$\ell\left(s^\star, s_m^\star\right) := \inf_{t \in S_m} \left\{\ell\left(s^\star, t\right)\right\}$$

Variance ou Erreur d'estimation

$$\mathbb{E}\left[P\gamma\left(\widehat{s}_m(D_n)\right)\right] - P\gamma\left(s_m^\star\right)$$



Sélection de modèles : biais et variance

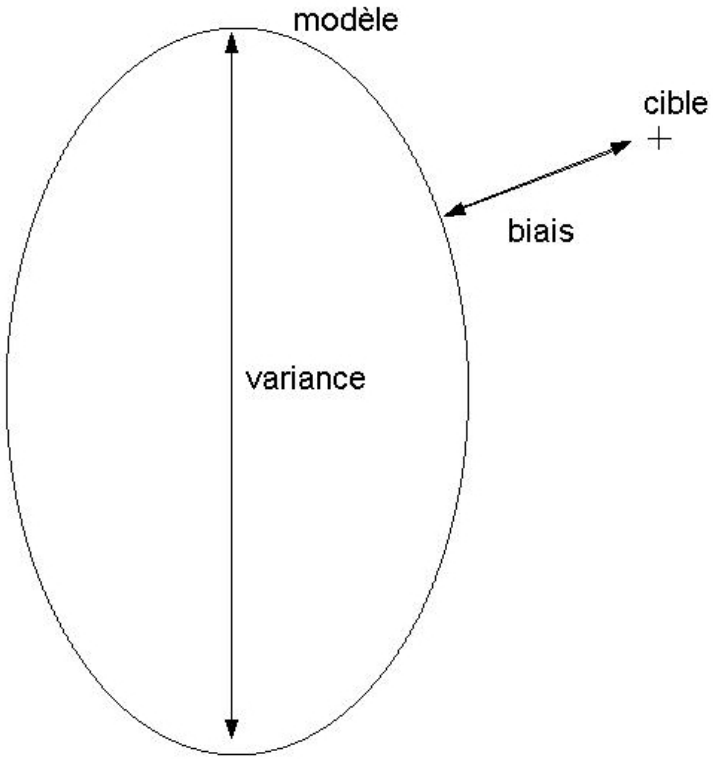
$$\mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m(D_n)\right)\right] = \text{Biais} + \text{Variance}$$

Biais ou Erreur d'approximation

$$\ell\left(s^\star, s_m^\star\right) := \inf_{t \in S_m} \left\{\ell\left(s^\star, t\right)\right\}$$

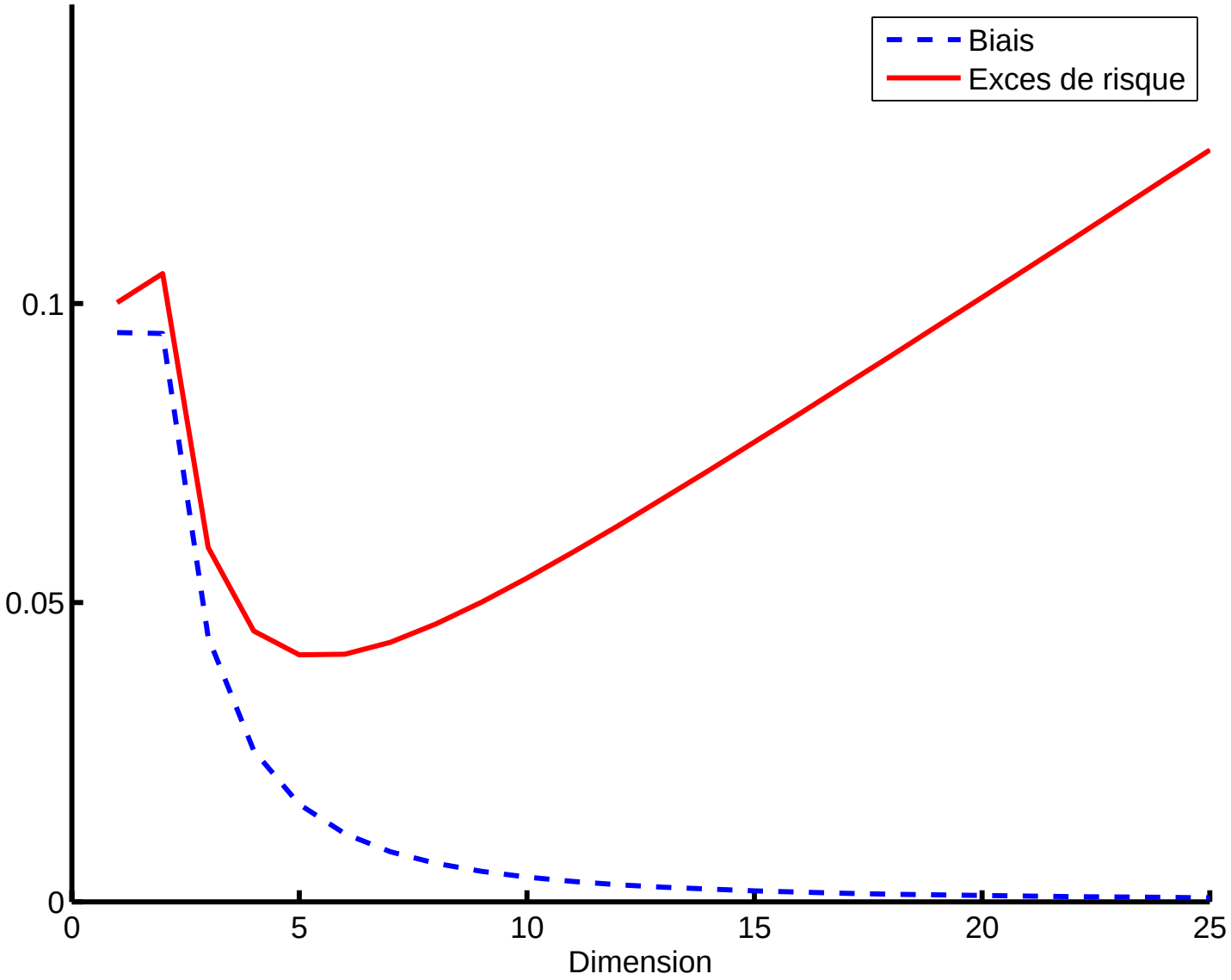
Variance ou Erreur d'estimation

$$\mathbb{E}\left[P\gamma\left(\widehat{s}_m(D_n)\right)\right] - P\gamma\left(s_m^\star\right)$$



Compromis biais-variance
$\Rightarrow$ éviter le sur-apprentissage et le sous-apprentissage

Compromis biais-variance



Exemple : régression homoscédastique sur un design fixe

$$Y = F + \varepsilon \quad \text{avec} \quad \mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$$

$$\widehat{F}_m = A_m Y \quad \text{avec} \quad A_m = A_m^\top = A_m^2 \quad \text{et} \quad \mathrm{tr}(A_m) = \dim(S_m)$$

$\Rightarrow$ Décomposition biais-variance du risque

Exemple : régression homoscédastique sur un design fixe

$$Y = F + \varepsilon \quad \text{avec} \quad \mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$$

$$\widehat{F}_m = A_m Y \quad \text{avec} \quad A_m = A_m^\top = A_m^2 \quad \text{et} \quad \operatorname{tr}(A_m) = \dim(S_m)$$

$\Rightarrow$ Décomposition biais-variance du risque

$$F_m = \arg\min_{t \in S_m} \left\{ \|t - F\|^2 \right\} = A_m F$$

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\|(A_m - I)F\|^2 + \frac{\sigma^2 \dim(S_m)}{n}$$

$$= \text{Biais} + \text{Variance}$$

Principe de l'estimation sans biais du risque

$$\widehat{m} \in \arg \min_{m \in \mathcal{M}_n} \{ \mathrm{crit}(m) \}$$

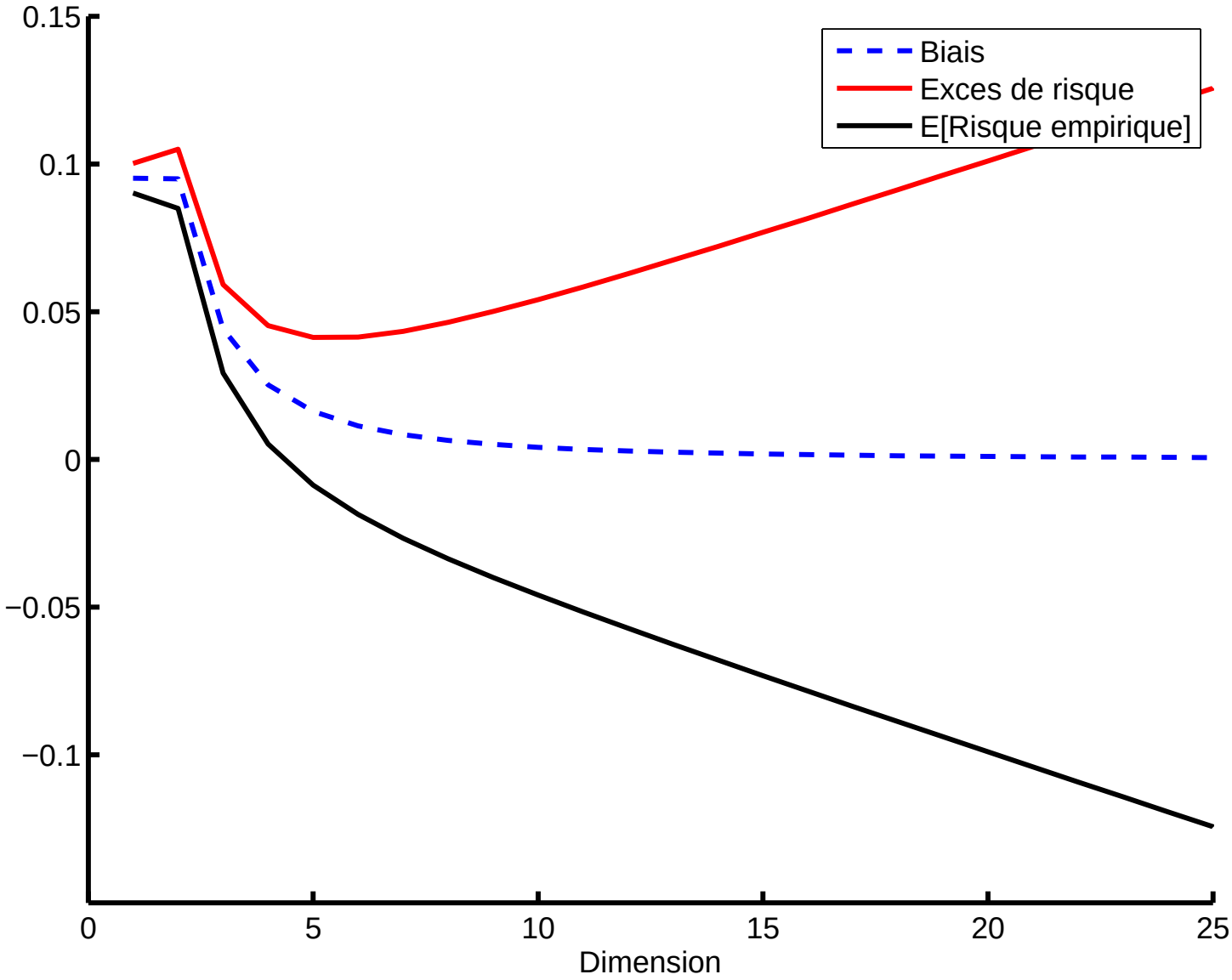
$$\mathrm{crit}_{\mathrm{id}}(m) = \ell \left(s^\star, \widehat{s}_m(D_n) \right)$$

Heuristique :

$$\mathrm{crit}(m) \approx \mathbb{E} \left[\ell \left(s^\star, \widehat{s}_m(D_n) \right) \right]$$

$\Rightarrow$ valide si $\mathrm{Card}(\mathcal{M}_n)$ n'est pas trop grand
(+ inégalités de concentration)

Pourquoi faut-il pénaliser le risque empirique ?



Pénalisation

- Pénalisation : $\mathrm{crit}(m) = P_n \gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m)$

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m) \right\}$$

Pénalisation

- Pénalisation : $\mathrm{crit}(m) = P_n \gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m)$

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + \mathrm{pen}(m) \right\}$$

- Pénalité idéale :

$$\mathrm{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma\left(\widehat{s}_m\right)$$

- Heuristique de Mallows :

$$\mathrm{pen}(m) \approx \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] \Rightarrow \text{inégalité oracle}$$

Exemple : régression homoscédastique sur un design fixe

Rappel :

$$Y = F + \varepsilon \quad \text{avec} \quad \mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$$

$$\widehat{F}_m = A_m Y \quad \text{avec} \quad A_m = A_m^\top = A_m^2 \quad \text{et} \quad \operatorname{tr}(A_m) = \dim(S_m)$$

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\left\|(A_m - I)F\right\|^2 + \frac{\sigma^2 \dim(S_m)}{n}$$

$\Rightarrow$ Risque empirique ? Pénalité idéale ? Son espérance ?

Exemple : régression homoscédastique sur un design fixe

Rappel :

$$Y = F + \varepsilon \quad \text{avec} \quad \mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2$$

$$\widehat{F}_m = A_m Y \quad \text{avec} \quad A_m = A_m^\top = A_m^2 \quad \text{et} \quad \mathrm{tr}(A_m) = \dim(S_m)$$

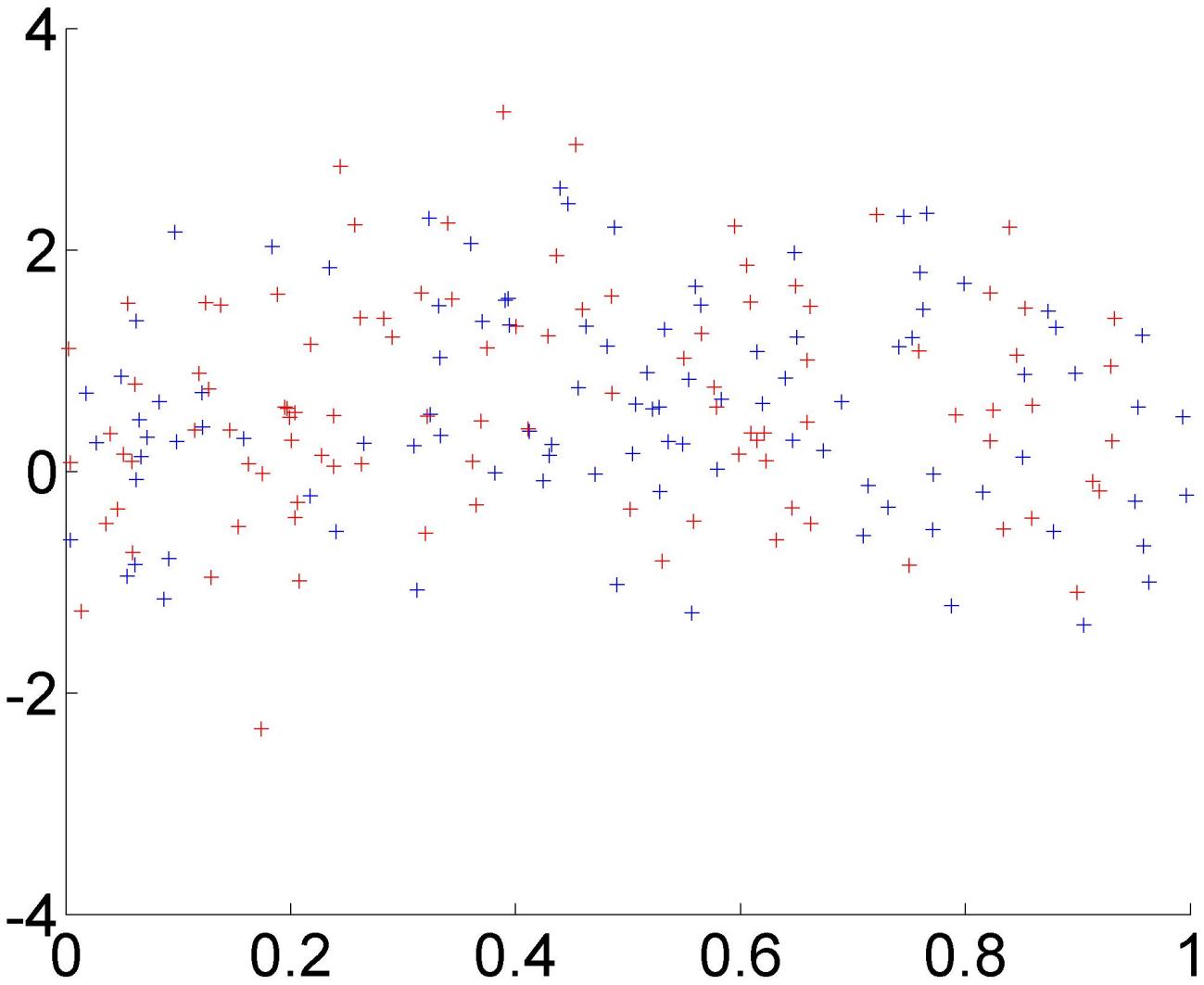
$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\left\|(A_m - I)F\right\|^2 + \frac{\sigma^2 \dim(S_m)}{n}$$

$\Rightarrow$ Risque empirique ? Pénalité idéale ? Son espérance ?

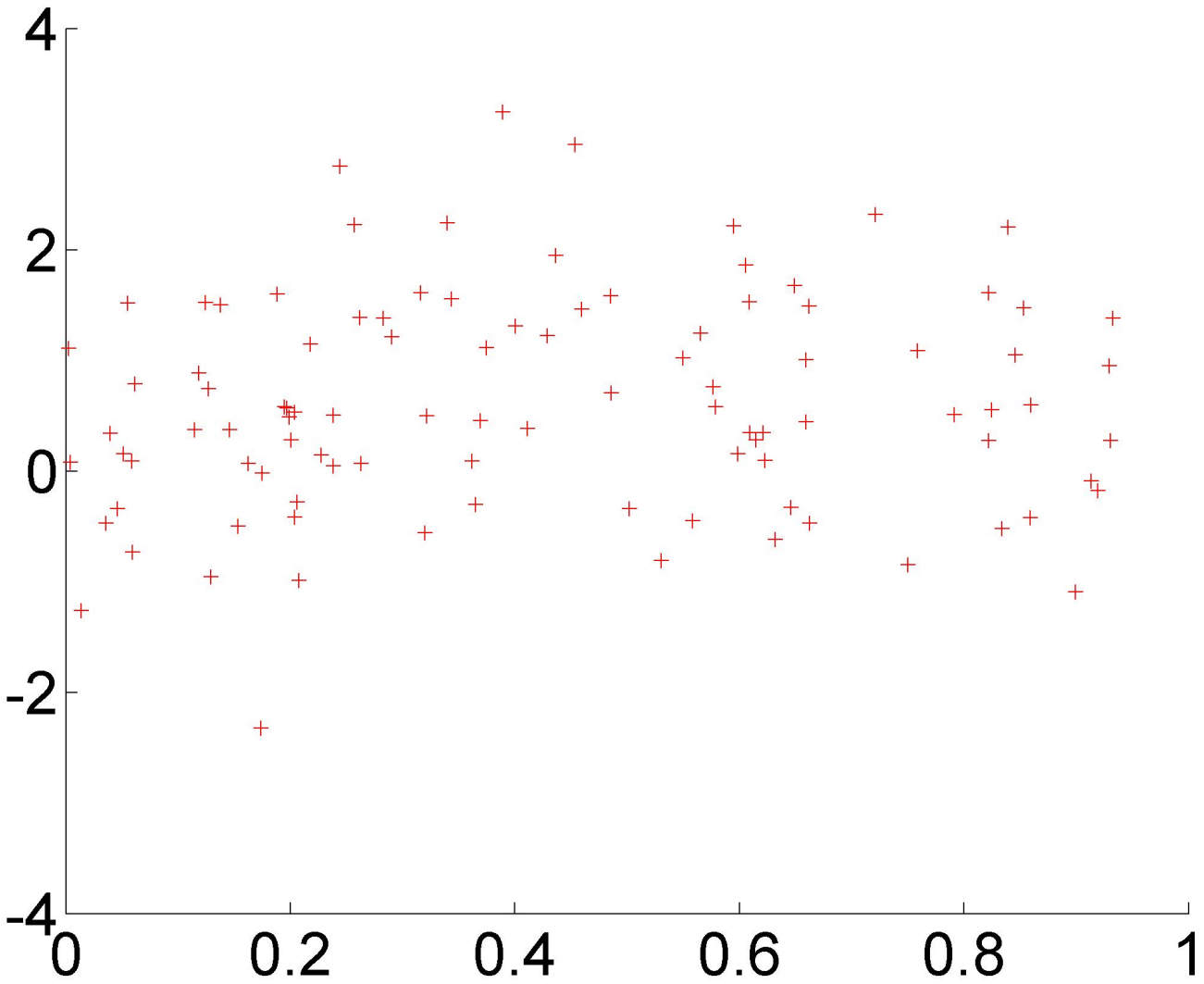
$$\mathrm{pen}_{\mathrm{id}}(m) = \frac{2}{n}\langle A_m\varepsilon, \varepsilon\rangle + \frac{2}{n}\langle (A_m - I_n)F, \varepsilon\rangle$$

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{2\sigma^2 D_m}{n} \quad \Rightarrow \quad C_p \text{ (Mallows, 1973)}$$

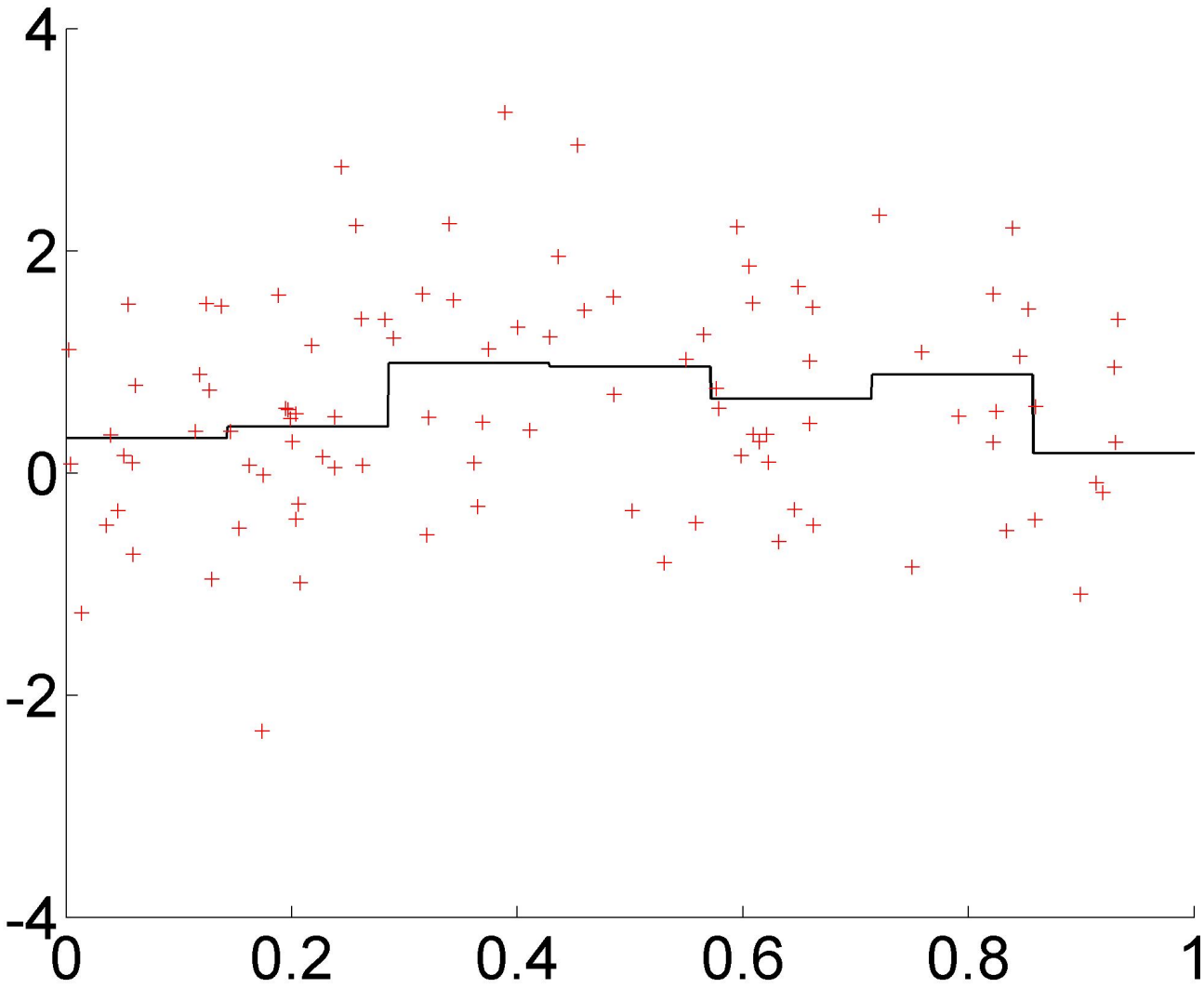
Validation simple (hold-out)



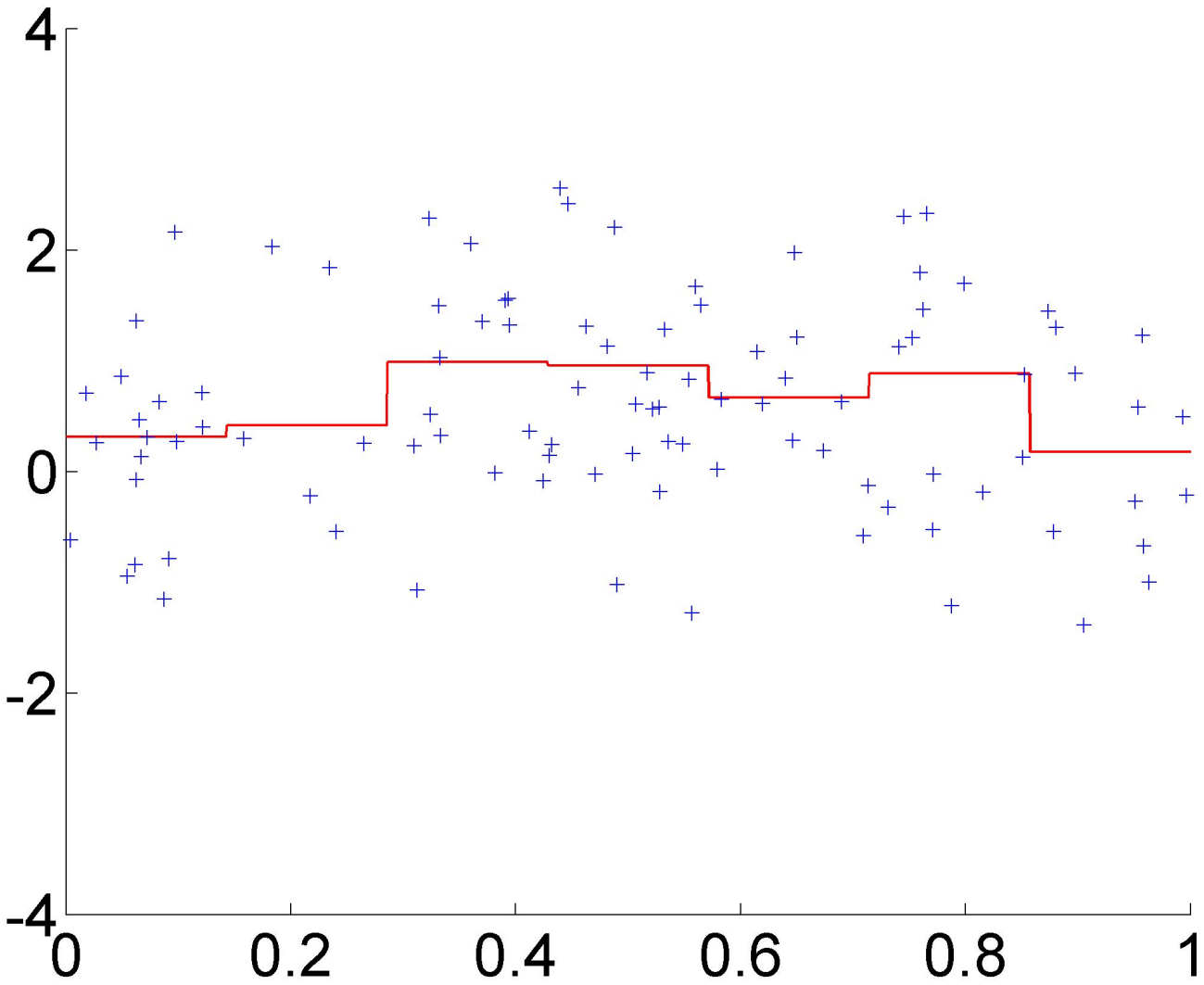
Validation : l'échantillon d'entraînement



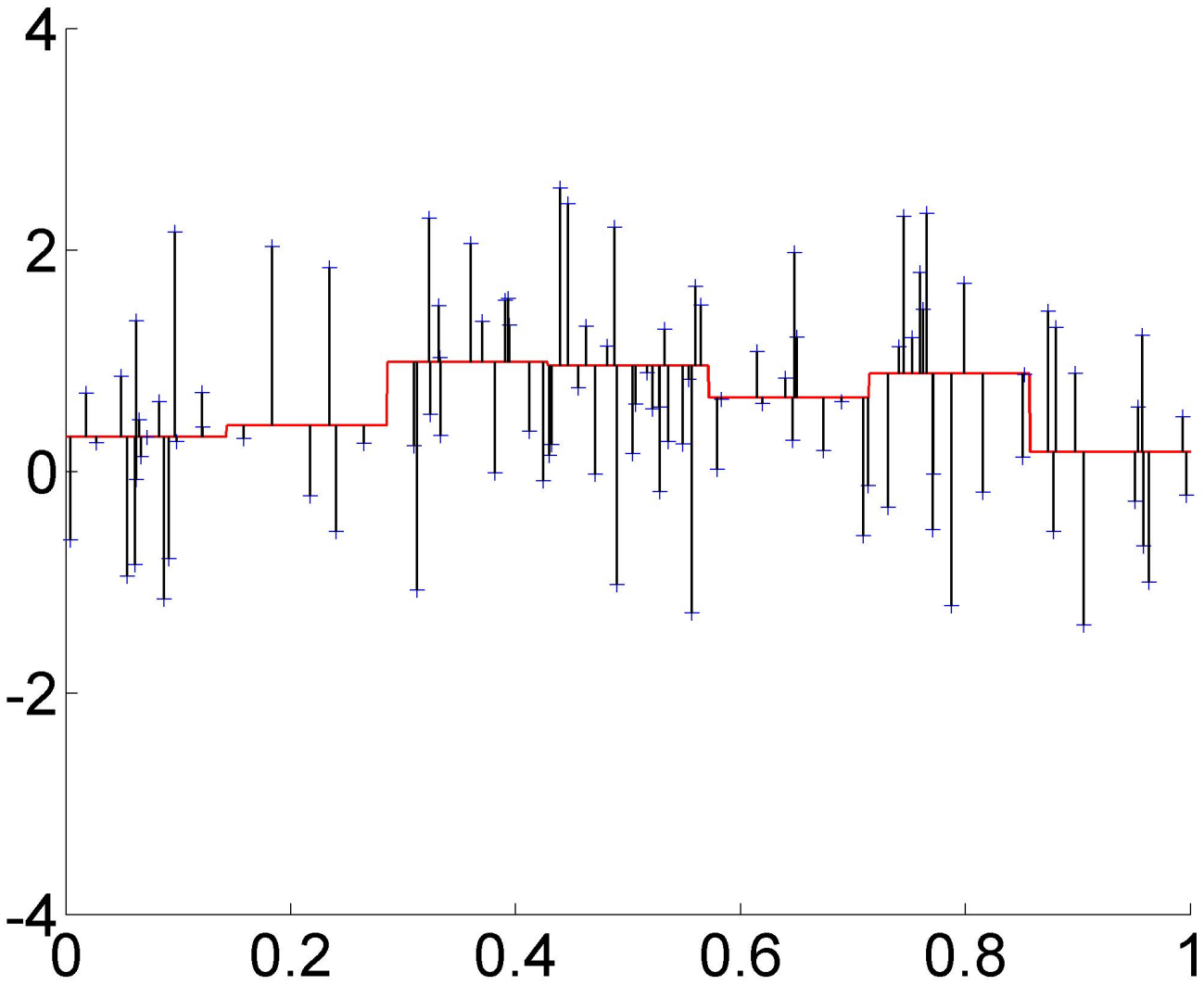
Validation : l'échantillon d'entraînement



Validation : l'échantillon de validation



Validation : l'échantillon de validation



Procédures fondées sur l'estimation sans biais du risque

Idée :

$$\mathbb{E}\left[\mathrm{crit}(m)\right] \approx \mathbb{E}\left[P\gamma\left(\widehat{s}_m\right)\right] \quad \text{soit} \quad \mathbb{E}\left[\mathrm{pen}(m)\right] \approx \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] \ .$$

Exemples :

- FPE (Akaike, 1970), SURE (Stein, 1981)
- certains types de validation croisée (e.g., leave-p-out, $p \ll n$)
- log-vraisemblance : AIC (Akaike, 1973), AICc (Sugiura, 1978 ; Hurvich et Tsai, 1989)
- moindres carrés : C_p, C_L (Mallows, 1973), GCV (Craven et Wahba, 1979)
- pénalités covariance (Efron, 2004)
- pénalité bootstrap (Efron, 1983), par rééchantillonnage (Arlot, 2009)
- ...

Plan

Un lemme clé

Lemme

Soit pen : $\mathcal{M}_n \mapsto \mathbb{R}$ *une pénalité (pouvant dépendre des données). Sur l'événement* Ω *où pour tout* $m, m' \in \mathcal{M}_n$,

$$\begin{aligned}&\left(\mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m, D_n)\right) - \left(\mathrm{pen}(m') - \mathrm{pen}_{\mathrm{id}}(m', D_n)\right) \\ &\qquad \leq A(m) + B(m')\end{aligned}$$

$$\textit{on a} \quad \forall \widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \mathrm{pen}(m) \right\}$$

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 - B(\widehat{m}) \leq \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 + A(m) \right\}$$

Inégalité-oracle pour la régression Gaussienne (1)

Hypothèses :

- Régression à design fixe, contraste des moindres carrés
- Bruit Gaussien homoscédastique : $\varepsilon \sim \mathcal{N}(0, \sigma^2)$
- Famille de modèles de complexité polynômiale : $\mathrm{Card}(\mathcal{M}_n) \leq C n^\alpha$
- Pour tout $m \in \mathcal{M}_n$, $\widehat{F}_m = A_m Y = \Pi_{S_m} Y$ (estimateur des moindres carrés)
- Pénalité

$$\mathrm{pen}(m) = \frac{K \sigma^2 \dim(S_m)}{n} \qquad \text{avec} \quad K > 1$$

Inégalité-oracle pour la régression Gaussienne (2) : rappels

$$-B(m) \leq \mathrm{pen}(m) - \mathrm{pen}_{\mathrm{id}}(m, D_n) \leq A(m)$$

$$\Rightarrow \quad \frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 - B(\widehat{m}) \leq \inf_{m \in \mathcal{M}_n}\left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 + A(m)\right\}$$

$$\mathrm{pen}_{\mathrm{id}}(m, D_n) = \frac{2}{n}\langle A_m\varepsilon, \varepsilon\rangle + \frac{2}{n}\langle (A_m - I_n)F, \varepsilon\rangle$$

Premier terme d'espérance $\frac{2\sigma^2 \dim(S_m)}{n}$, deuxième terme centré.

Inégalité-oracle pour la régression Gaussienne (3)

Deux résultats de concentration Gaussienne (cf. Massart 2007) :

Proposition

Soit ξ un vecteur Gaussien standard de $\mathbb{R}^n$, $\alpha \in \mathbb{R}^n$, $M \in \mathcal{M}_n(\mathbb{R})$. Alors, pour tout $x \geq 0$,

$$\mathbb{P}\left(|\langle \xi, \alpha \rangle| \leq \sqrt{2x}\, \|\alpha\|_2 \right) \geq 1 - 2e^{-x}$$

$$\mathbb{P}\left(|\langle \xi, M\xi \rangle - \mathrm{tr}(M)| \leq 2\sqrt{x\, \mathrm{tr}\left(M^\top M\right) + 2\, \|M\|\, x} \right) \geq 1 - 2e^{-x}$$

Inégalité-oracle pour la régression Gaussienne (4)

Schéma de la preuve :

- Pour tout $m \in \mathcal{M}_n$,
  concentrer $\langle A_m \varepsilon, \varepsilon \rangle$ autour de $\sigma^2 \dim(S_m)$
  et $\langle (A_m - I_n) F, \varepsilon \rangle$ autour de 0

- Appliquer le lemme sur l'intersection de ces $\mathrm{Card}(\mathcal{M}_n)$ événements

- Contrôler les termes de reste

Inégalité-oracle pour la régression Gaussienne (5)

Théorème (Birgé & Massart 2007)

Pour tout $x \geq 0$, *avec probabilité* $1 - 4\,\mathrm{Card}(\mathcal{M}_n)e^{-x}$, *pour tout*

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{K\sigma^2 \dim(S_m)}{n} \right\} ,$$

on a l'inégalité-oracle $\forall \delta > 0$,

$$\frac{1}{n} \left\| \widehat{F}_{\widehat{m}} - F \right\|^2 \leq \left(\frac{1 + (K-2)_+}{1 - (2-K)_+} + \delta \right) \inf_{m \in \mathcal{M}_n} \left\{ \frac{1}{n} \left\| \widehat{F}_m - F \right\|^2 \right\} + \frac{C(K) x \sigma^2}{\delta n}$$

Plan

Probabilités : concentration de la mesure

- Processus empiriques :

$$(P_n - P)\gamma(t) \quad \text{ou} \quad \sup_{t \in S} \{ (P_n - P)\gamma(t) \}$$

- Concentration des termes quadratiques, $\|M\varepsilon\|^2$, statistiques du χ^2 (en les écrivant comme un sup, ou en considérant la concentration de U-statistiques générales)

- Quantités plus complexes, par exemple la "pénalité idéale"

$$(P - P_n)\gamma\left(\widehat{s}_m(D_n)\right)$$

Probabilités

- Contrôles ou calculs d'espérances :

$$\mathbb{E}\left[\sup_{t\in S}\left\{(P_n - P)\gamma(t)\right\}\right]$$

$$\mathbb{E}\left[(P - P_n)\gamma\left(\widehat{s}_m(D_n)\right)\right]$$

- Contrôle du risque vu comme une fonction de n

$$\mathbb{E}\left[P\gamma\left(\widehat{s}_m(D_n)\right)\right]$$

- Processus de rééchantillonnage
- Contrôle de termes de reste (variances, déviations, ...) par rapport aux espérances
- ...

Théorie de l'approximation

- Terme de biais $\ell(s^\star, S_m)$
- Contrôle nécessaire pour obtenir un résultat d'adaptation à partir d'une inégalité-oracle
- En retour, comment bien choisir $(S_m)_{m \in \mathcal{M}_n}$ sachant que $P \in \mathcal{P}$?
- Contrôle de $\ell(s^\star, S_m)$ (majoration et minoration) utile pour contrôler $\dim(S_{\widehat{m}})$ et $\dim(S_{m^\star})$

Optimisation : pour des raisons pratiques

- $\widehat{s}_m(D_n)$ souvent défini comme un arg min

$\Rightarrow$ Calcul de $\widehat{s}_m(D_n)$ pour chaque m (éventuellement approché) ?

$\Rightarrow$ Calcul direct rapide de la famille $\left(\widehat{s}_m(D_n)\right)_{m \in \mathcal{M}_n}$ (chemin de régularisation, e.g. LARS-Lasso) ?

Optimisation : pour des raisons pratiques

- $\widehat{s}_m(D_n)$ souvent défini comme un arg min

$\Rightarrow$ Calcul de $\widehat{s}_m(D_n)$ pour chaque m (éventuellement approché) ?

$\Rightarrow$ Calcul direct rapide de la famille $\left(\widehat{s}_m(D_n)\right)_{m \in \mathcal{M}_n}$ (chemin de régularisation, e.g. LARS-Lasso) ?

- Calcul de $\widehat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \mathrm{crit}(m) \right\}$ sans énumérer tous les $m \in \mathcal{M}_n$? (e.g., programmation dynamique pour la détection de ruptures : Bellman & Dreyfus, 1962 ; Rigaill, 2010)

Optimisation : pour des raisons pratiques

- $\widehat{s}_m(D_n)$ souvent défini comme un arg min

$\Rightarrow$ Calcul de $\widehat{s}_m(D_n)$ pour chaque m (éventuellement approché) ?

$\Rightarrow$ Calcul direct rapide de la famille $\left(\widehat{s}_m(D_n)\right)_{m\in\mathcal{M}_n}$ (chemin de régularisation, e.g. LARS-Lasso) ?

- Calcul de $\widehat{m} \in \arg\min_{m\in\mathcal{M}_n} \left\{ \mathrm{crit}(m) \right\}$ sans énumérer tous les $m \in \mathcal{M}_n$? (e.g., programmation dynamique pour la détection de ruptures : Bellman & Dreyfus, 1962 ; Rigaill, 2010)

- Les procédures les plus intéressantes à étudier sont celles qui peuvent être implémentées efficacement.

Optimisation : pour des raisons théoriques

- $\widehat{s}_m(D_n)$ souvent défini comme un arg min

$\Rightarrow$ les conditions KKT permettent de le caractériser

- Ex : pénalité idéale pour le Lasso (Efron et al. 2004 ; Zou, Hastie & Tibshirani 2007)

- RKHS et méthodes à noyau : théorème du représentant
- ...

Plan

Plan des séances restantes

- Lundi 17, 14h–16h : Calibration de pénalités et pénalités minimales
- Lundi 24, 14h–16h : Rééchantillonnage et pénalisation
- Lundi 31, 15h30–17h30 : Validation croisée et pénalités reliées