Contributions à la théorie statistique de l'apprentissage: sélection d'estimateurs et détection de ruptures

Sylvain Arlot

[1]Cnrs

[2]École Normale Supérieure (Paris), DI/ENS, Équipe Sierra

Soutenance d'habilitation à diriger des recherches, 3 décembre 2014

Parcours

- Thèse et monitorat à l'Université Paris-Sud (2004 à 2008)
  - Séjour à Berkeley (fév.-mars 2008, P. Bartlett)

Parcours

- Thèse et monitorat à l'Université Paris-Sud (2004 à 2008)
  - Séjour à Berkeley (fév.-mars 2008, P. Bartlett)
- Chargé de recherches CNRS, affecté au Département d'Informatique de l'École normale supérieure
  - Équipe Willow (2008 à 2010)
  - Équipe Sierra (2011 à aujourd'hui)

Pour une théorie utile en pratique

- Comprendre pourquoi certaines procédures sont performantes

Pour une théorie utile en pratique

- Comprendre pourquoi certaines procédures sont performantes
- Pourquoi certaines procédures fonctionnent-elles mieux ?
  Compromis entre complexité algorithmique et performance statistique ?

Pour une théorie utile en pratique

- Comprendre pourquoi certaines procédures sont performantes
- Pourquoi certaines procédures fonctionnent-elles mieux ?
  Compromis entre complexité algorithmique et performance statistique ?
- Corriger les défauts de méthodes couramment utilisées

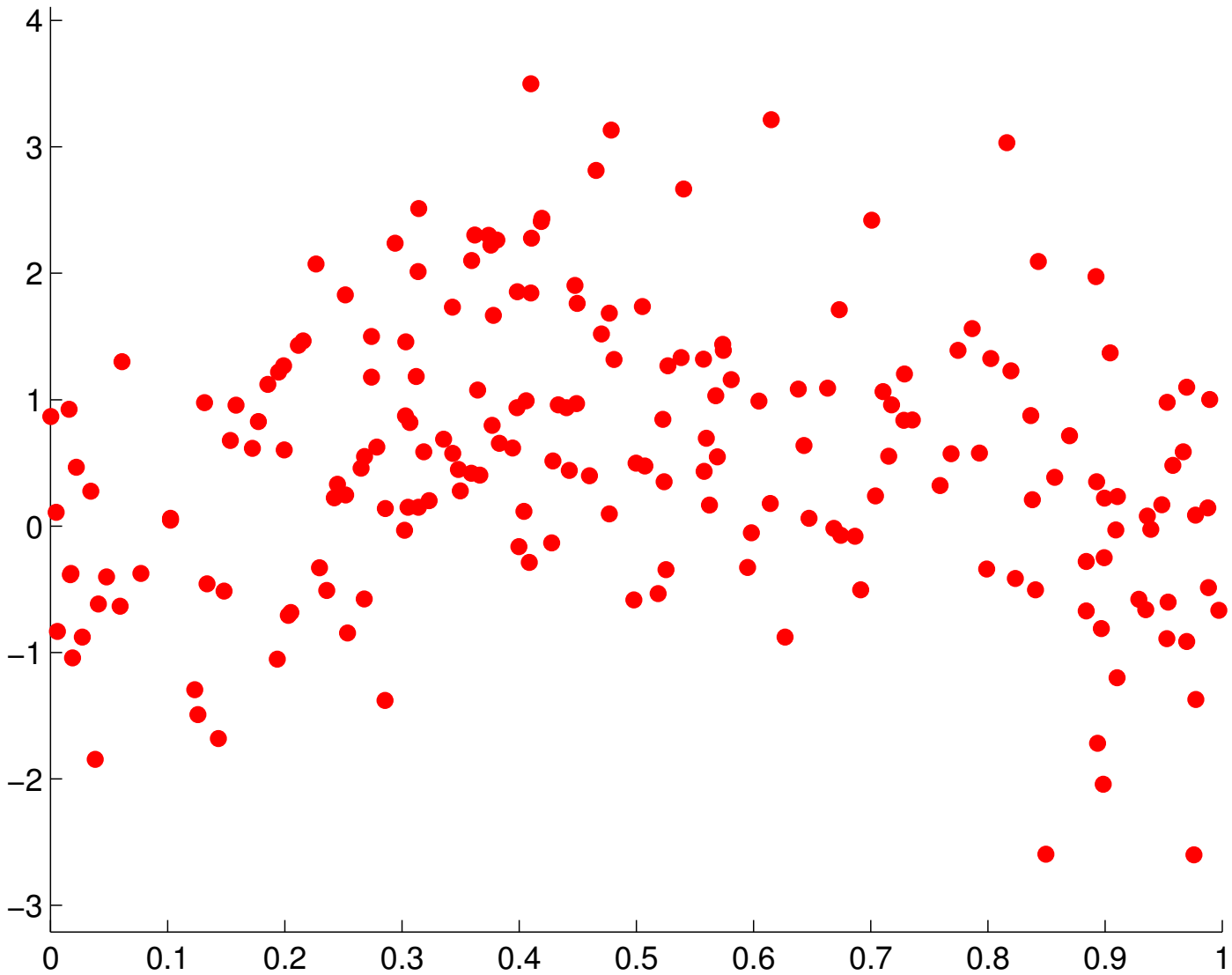
Pour une théorie utile en pratique

- Comprendre pourquoi certaines procédures sont performantes
- Pourquoi certaines procédures fonctionnent-elles mieux ?
  Compromis entre complexité algorithmique et performance statistique ?
- Corriger les défauts de méthodes couramment utilisées
- Proposer de nouvelles méthodes sur des bases théoriques
  e.g., heuristique de pente (Birgé & Massart, 2001)

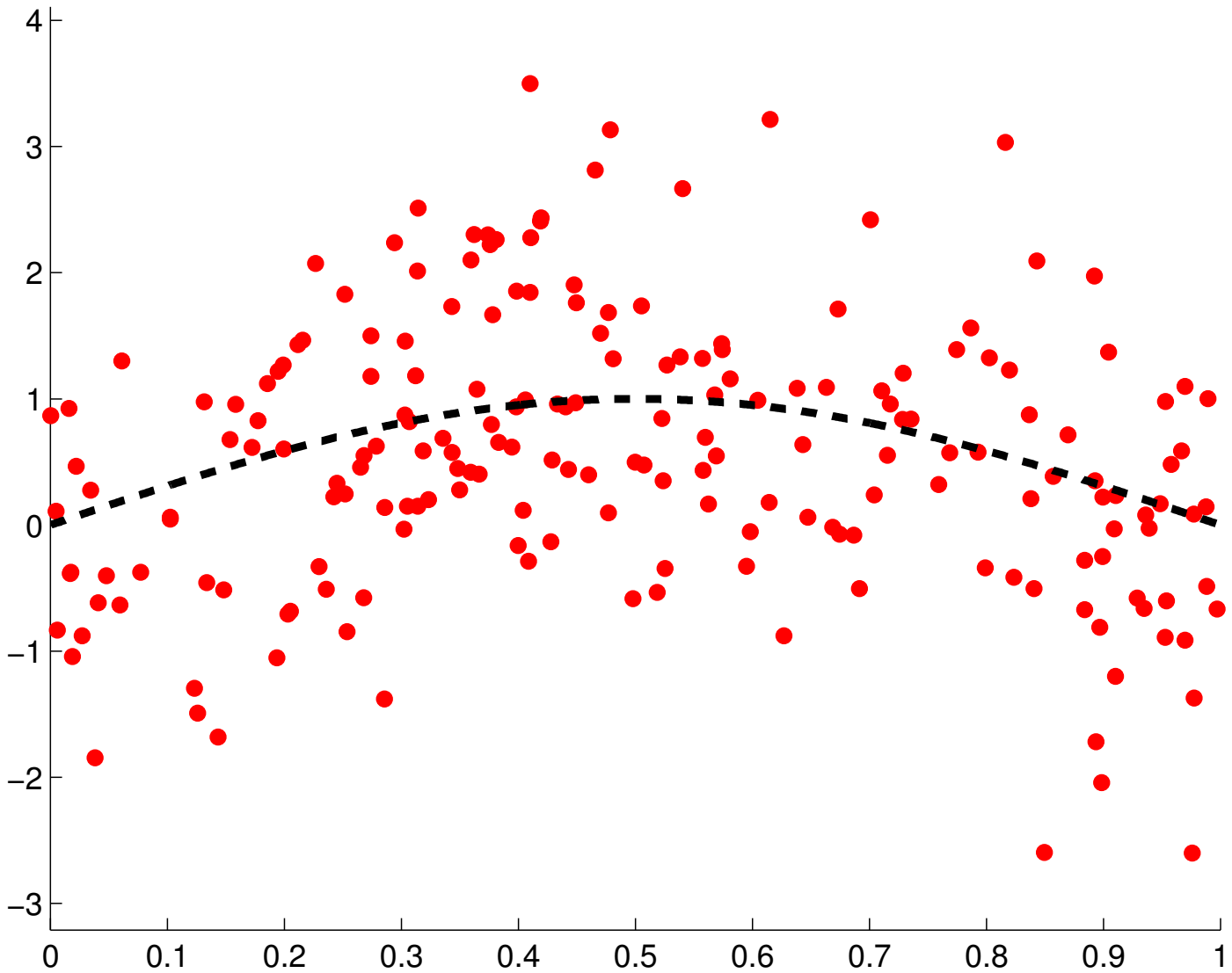
Plan

1. Sélection d'estimateurs
2. Validation croisée
3. Pénalités minimales
4. Détection de ruptures
5. Conclusion

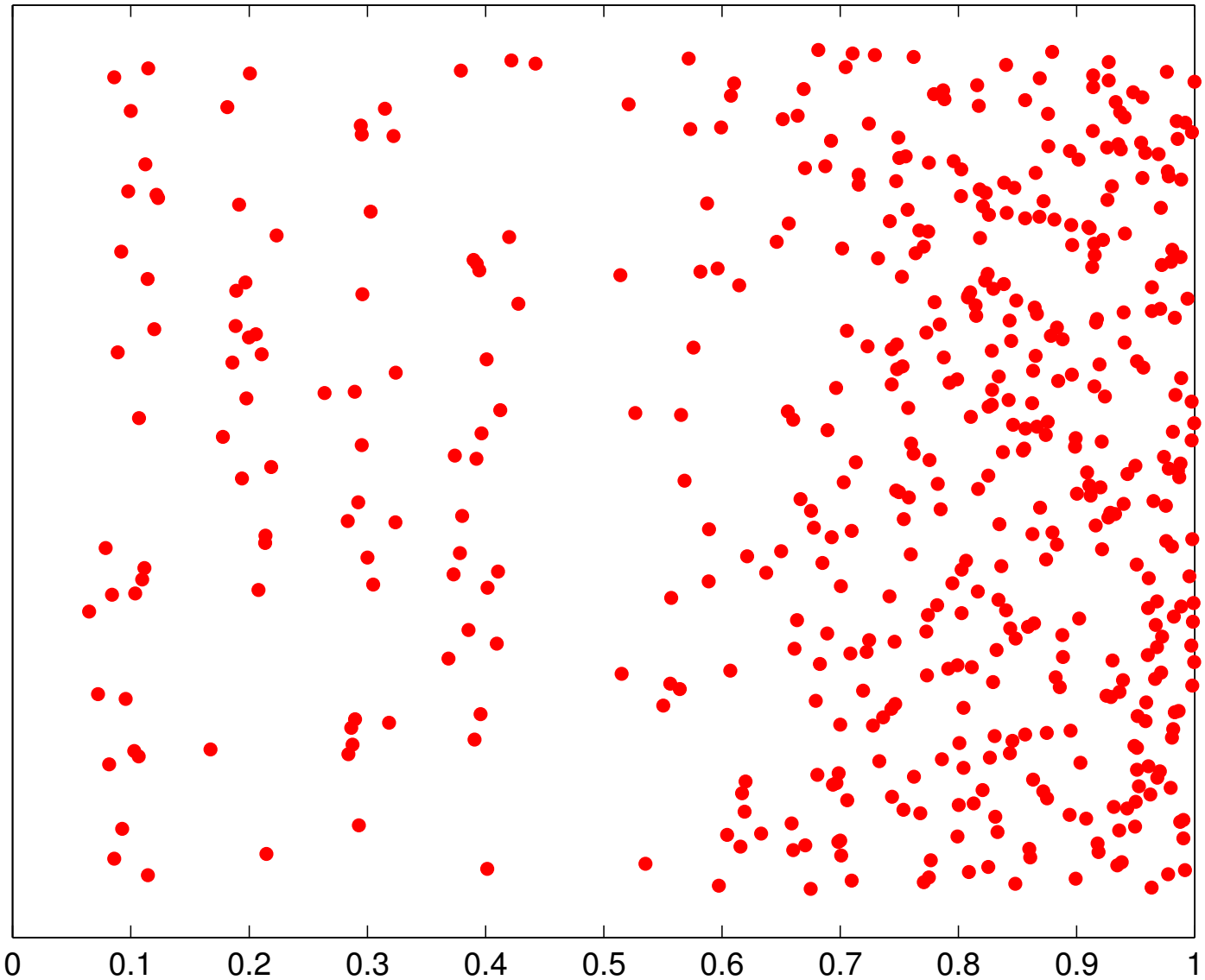
Régression : données $(X_1, Y_1), \ldots, (X_n, Y_n)$



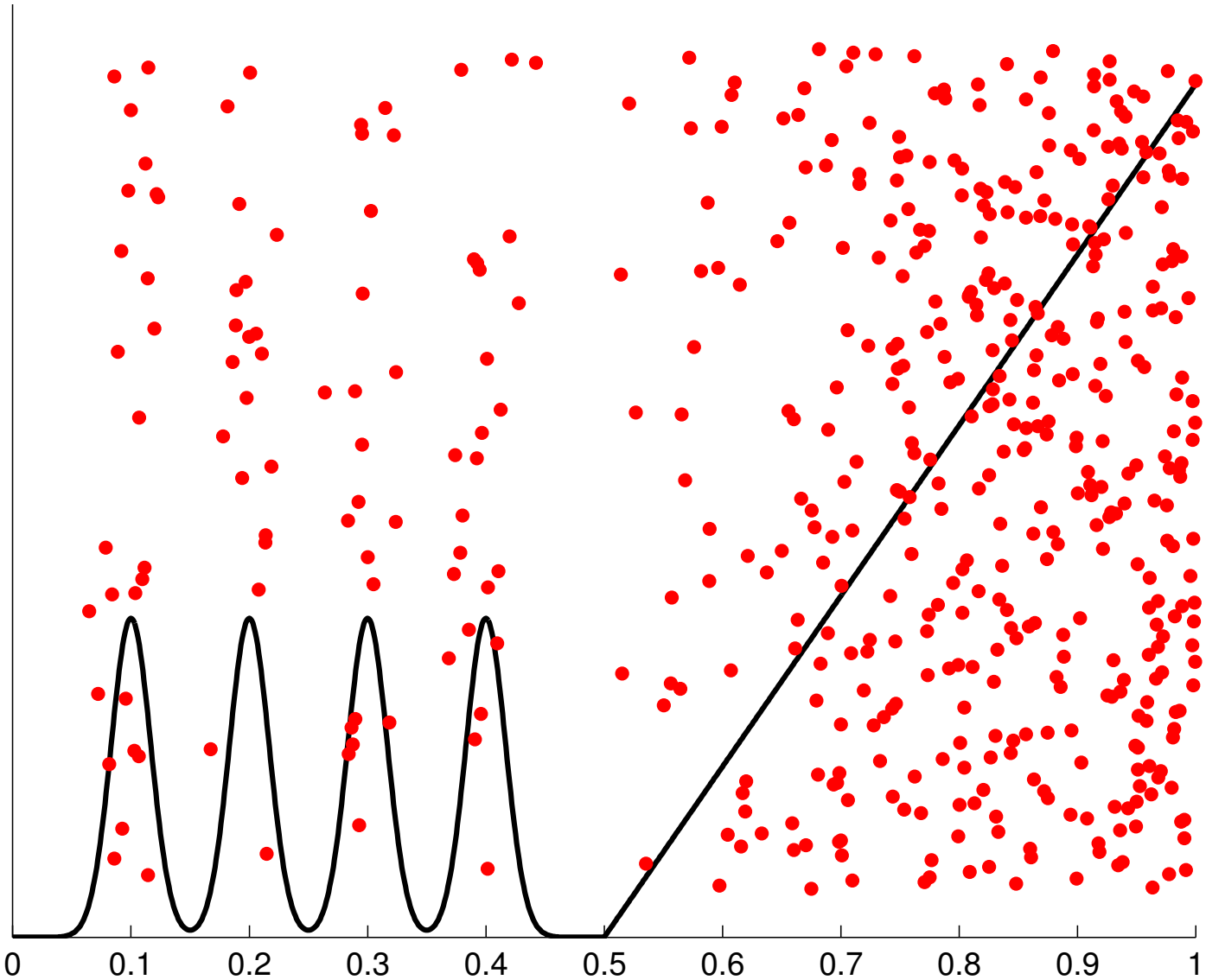
But : reconstruire le signal



Estimation de densité : données $\xi_1, \ldots, \xi_n$



But : estimer la densité $s^\star$ des observations ξ_i



Cadre général

- Données $\xi_1, \ldots, \xi_n \in \Xi$ i.i.d. de loi P
  prédiction : $\xi_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$

Cadre général

- Données $\xi_1, \ldots, \xi_n \in \Xi$ i.i.d. de loi P
  prédiction : $\xi_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$
- But : estimer une caractéristique $s^\star \in \mathbb{S}$ de P
  densité, fonction de régression, meilleur prédicteur, etc.

Cadre général

- Données $\xi_1, \ldots, \xi_n \in \Xi$ i.i.d. de loi P
  prédiction : $\xi_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$
- But : estimer une caractéristique $s^\star \in \mathbb{S}$ de P
  densité, fonction de régression, meilleur prédicteur, etc.
- Fonction de contraste $\gamma : \mathbb{S} \times \Xi \to \mathbb{R}$ telle que

$$s^\star \in \operatorname*{argmin}_{t \in \mathbb{S}} \big\{ P\gamma(t) \big\} \qquad \text{avec} \qquad P\gamma(t) := \mathbb{E}_{\xi \sim P} \big[\gamma(t; \xi) \big]$$

Cadre général

- Données $\xi_1, \ldots, \xi_n \in \Xi$ i.i.d. de loi P
  prédiction : $\xi_i = (X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$
- But : estimer une caractéristique $s^\star \in \mathbb{S}$ de P
  densité, fonction de régression, meilleur prédicteur, etc.
- Fonction de contraste $\gamma : \mathbb{S} \times \Xi \to \mathbb{R}$ telle que

$$s^\star \in \operatorname*{argmin}_{t \in \mathbb{S}} \{ P\gamma(t) \} \qquad \text{avec} \qquad P\gamma(t) := \mathbb{E}_{\xi \sim P} \left[\gamma(t; \xi) \right]$$

- Perte relative

$$\ell(s^\star, t) := P\gamma(t) - P\gamma(s^\star) \geq 0$$

Exemples

- Prédiction : $\xi_i = (X_i, Y_i)$
  $X_{n+1} \quad \rightsquigarrow \quad \ll$ prédire $\gg \; Y_{n+1}$ avec $t(X_{n+1})$?
  $\gamma\big(t; (x, y)\big)$ mesure la $\ll$ distance $\gg$ entre $t(x)$ et y

Exemples

- Prédiction : $\xi_i = (X_i, Y_i)$
  $X_{n+1} \quad \rightsquigarrow \quad \ll$ prédire $\gg Y_{n+1}$ avec $t(X_{n+1})$?
  $\gamma\big(t;(x,y)\big)$ mesure la $\ll$ distance $\gg$ entre $t(x)$ et y

- Régression ($\mathcal{Y} = \mathbb{R}$), moindres carrés :

$$\gamma\big(t;(x,y)\big) = \big(t(x) - y\big)^2 \qquad s^\star(X) = \mathbb{E}[Y|X]$$

Exemples

- Prédiction : $\xi_i = (X_i, Y_i)$
  $X_{n+1} \quad \rightsquigarrow \quad \ll$ prédire $\gg$ Y_{n+1} avec $t(X_{n+1})$?
  $\gamma\big(t;(x,y)\big)$ mesure la $\ll$ distance $\gg$ entre $t(x)$ et y

- Régression ($\mathcal{Y} = \mathbb{R}$), moindres carrés :

$$\gamma\big(t;(x,y)\big) = \big(t(x) - y\big)^2 \qquad s^\star(X) = \mathbb{E}[Y|X]$$

- Classification binaire ($\mathcal{Y} = \{0,1\}$), perte 0–1 :

$$\gamma\big(t;(x,y)\big) = \mathbb{1}_{t(x)\neq y}$$

Exemples

- Prédiction : $\xi_i = (X_i, Y_i)$
  $X_{n+1} \quad \rightsquigarrow \quad \ll$ prédire $\gg$ Y_{n+1} avec $t(X_{n+1})$?
  $\gamma\big(t;(x,y)\big)$ mesure la $\ll$ distance $\gg$ entre $t(x)$ et y

- Régression ($\mathcal{Y} = \mathbb{R}$), moindres carrés :

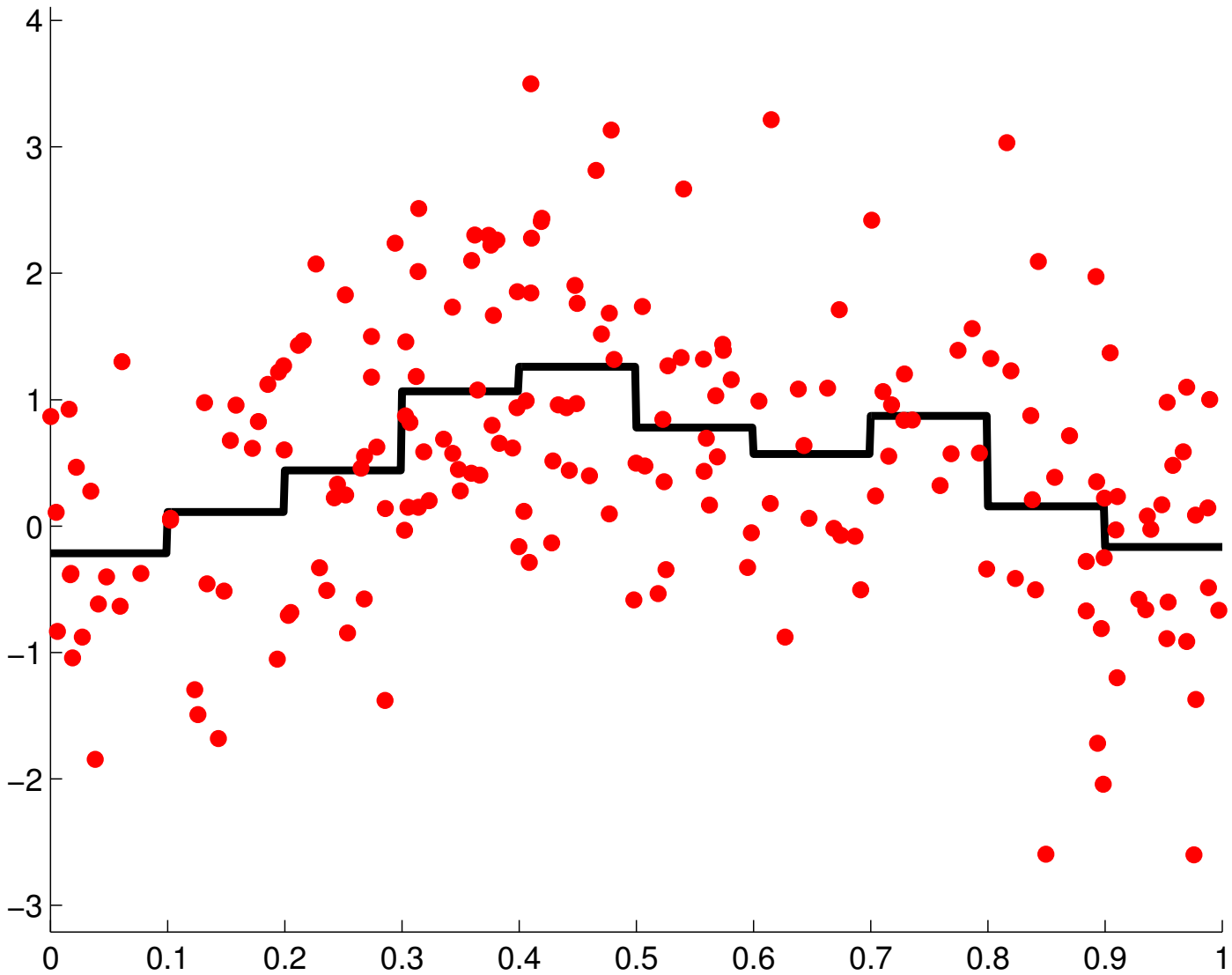
$$\gamma\big(t;(x,y)\big) = \big(t(x) - y\big)^2 \qquad s^\star(X) = \mathbb{E}[Y|X]$$

- Classification binaire ($\mathcal{Y} = \{0,1\}$), perte 0–1 :

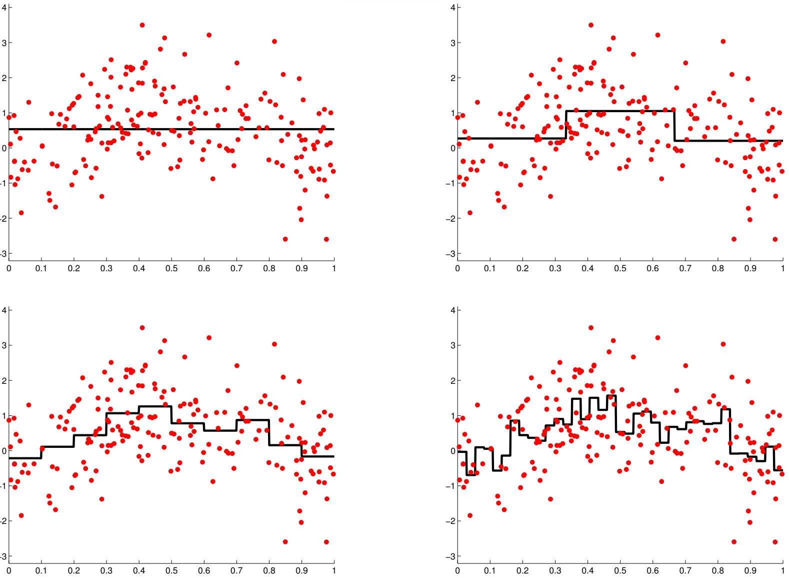
$$\gamma\big(t;(x,y)\big) = \mathbb{1}_{t(x)\neq y}$$

- Estimation de densité (mesure de référence μ) :
  moindres carrés : $\gamma(t;\xi) = \|t\|^2_{L^2(\mu)} - 2t(\xi)$
  log-vraisemblance : $\gamma(t;\xi) = -\log\big(t(\xi)\big)$

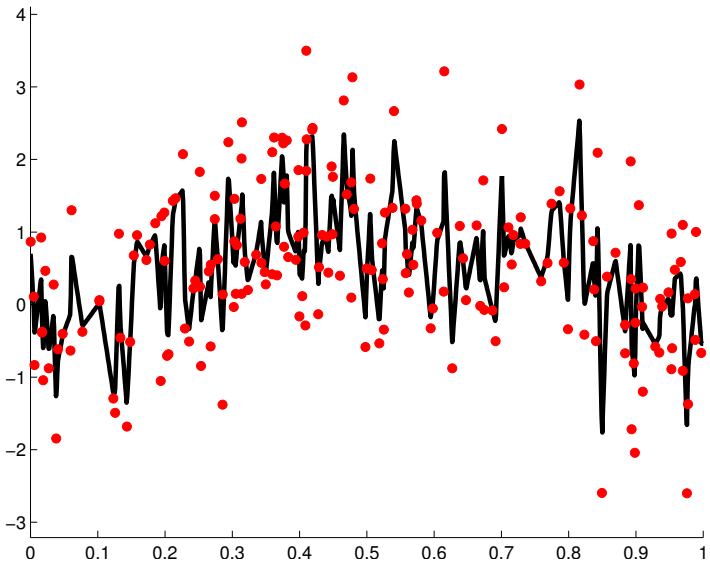
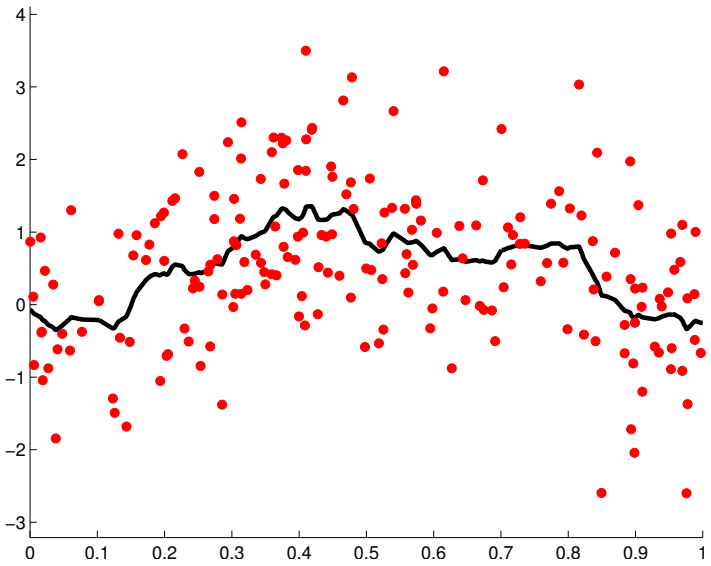
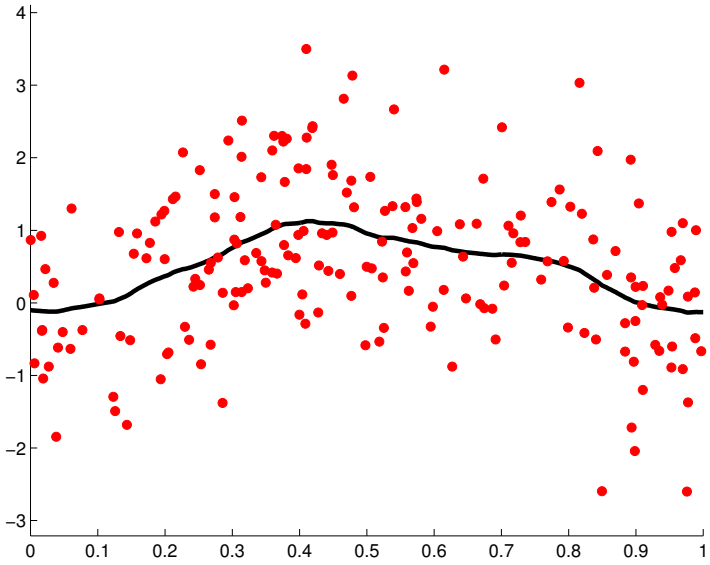
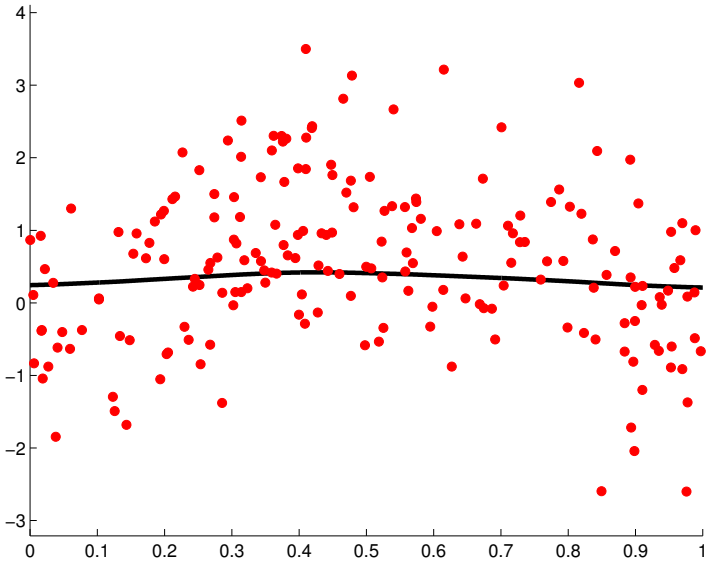
Estimateurs : un régressogramme



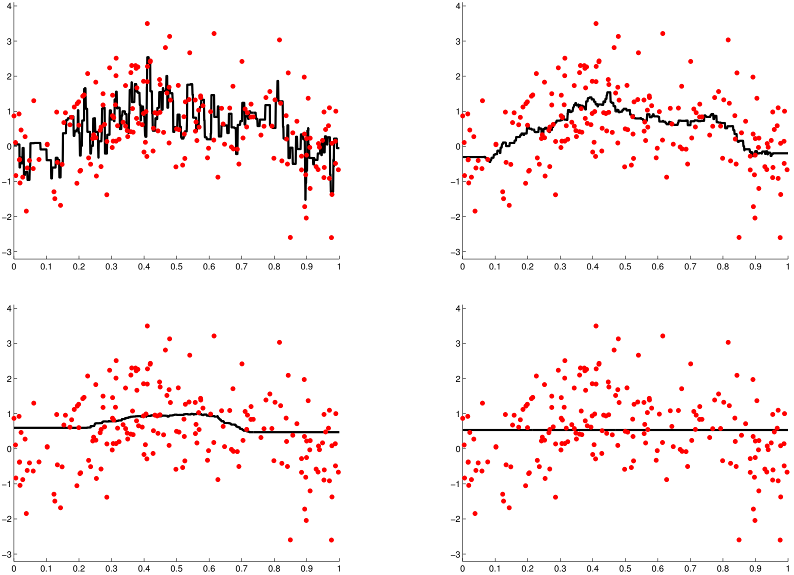
Sélection d'estimateurs : régressogrammes réguliers



Sélection d'estimateurs : régression ridge à noyau

Sélection d'estimateurs : k plus proches voisins



Sélection d'estimateurs

- Estimateur/Algorithme d'apprentissage : $\widehat{s} : D_n \mapsto \widehat{s}(D_n) \in \mathbb{S}$

Sélection d'estimateurs

- Estimateur/Algorithme d'apprentissage : $\widehat{s} : D_n \mapsto \widehat{s}(D_n) \in \mathbb{S}$
- Exemple : estimateur des moindres carrés sur un modèle $S_m \subset \mathbb{S}$

$$\widehat{s}_m \in \operatorname*{argmin}_{t \in S_m} \left\{ P_n \gamma(t) \right\} \quad \text{où} \quad P_n \gamma(t) := \frac{1}{n} \sum_{\xi \in D_n} \gamma(t; \xi)$$

Exemple de modèle : histogrammes

Sélection d'estimateurs

- Estimateur/Algorithme d'apprentissage : $\widehat{s} : D_n \mapsto \widehat{s}(D_n) \in \mathbb{S}$
- Exemple : estimateur des moindres carrés sur un modèle $S_m \subset \mathbb{S}$

$$\widehat{s}_m \in \operatorname*{argmin}_{t \in S_m} \left\{ P_n \gamma(t) \right\} \quad \text{où} \quad P_n \gamma(t) := \frac{1}{n} \sum_{\xi \in D_n} \gamma(t; \xi)$$

  Exemple de modèle : histogrammes

- Famille d'estimateurs $(\widehat{s}_m)_{m \in \mathcal{M}} \Rightarrow$ choisir $\widehat{m} = \widehat{m}(D_n)$?
  e.g., famille de modèles $(S_m)_{m \in \mathcal{M}} \Rightarrow$ famille d'estimateurs des moindres carrés

Sélection d'estimateurs

- Estimateur/Algorithme d'apprentissage : $\widehat{s} : D_n \mapsto \widehat{s}(D_n) \in \mathbb{S}$
- Exemple : estimateur des moindres carrés sur un modèle $S_m \subset \mathbb{S}$

$$\widehat{s}_m \in \operatorname*{argmin}_{t \in S_m} \left\{ P_n \gamma(t) \right\} \quad \text{où} \quad P_n \gamma(t) := \frac{1}{n} \sum_{\xi \in D_n} \gamma(t; \xi)$$

  Exemple de modèle : histogrammes

- Famille d'estimateurs $(\widehat{s}_m)_{m \in \mathcal{M}} \Rightarrow$ choisir $\widehat{m} = \widehat{m}(D_n)$?
  e.g., famille de modèles $(S_m)_{m \in \mathcal{M}} \Rightarrow$ famille d'estimateurs des moindres carrés
- Objectif : minimiser le risque, *i.e.*,
  Inégalité oracle (en espérance ou avec grande probabilité) :

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

Compromis biais-variance

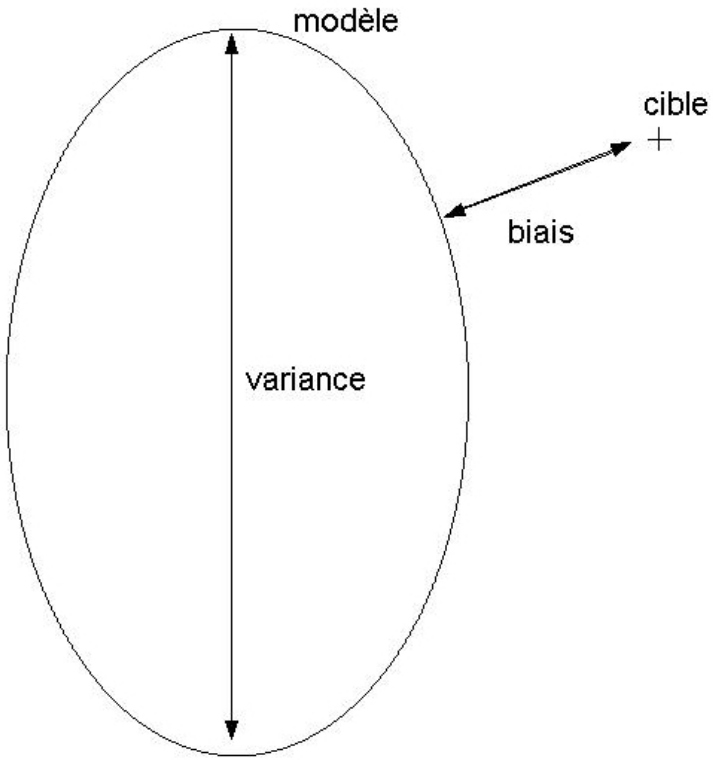
$$\mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right] = \text{Biais} + \text{Variance}$$

Biais ou Erreur d'approximation

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

Variance ou Erreur d'estimation

Régression, moindres carrés : $\dfrac{\sigma^2 \dim(S_m)}{n}$



Compromis biais-variance

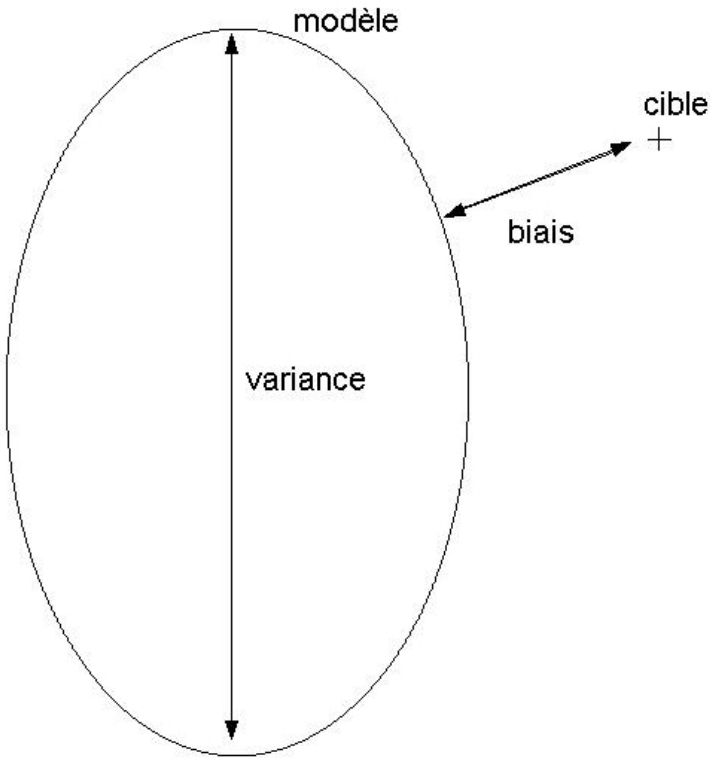
$$\mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right] = \text{Biais} + \text{Variance}$$

Biais ou Erreur d'approximation

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

Variance ou Erreur d'estimation

Régression, moindres carrés : $\dfrac{\sigma^2 \dim(S_m)}{n}$



Compromis biais-variance

$\Leftrightarrow$ éviter le sur-apprentissage et le sous-apprentissage

Un problème général

- Objectif :

$$\text{minimiser } \mathcal{R}(m) \text{ sur } m \in \mathcal{M}$$

Un problème général

- Objectif :

$$\text{minimiser } \mathcal{R}(m) \text{ sur } m \in \mathcal{M}$$

- Méthode :

$$\text{minimiser } \mathcal{C}(m) \text{ sur } m \in \mathcal{M} \quad \Rightarrow \quad \widehat{m}_{\mathcal{C}}$$

Un problème général

- Objectif :

$$\text{minimiser } \mathcal{R}(m) \text{ sur } m \in \mathcal{M}$$

- Méthode :

$$\text{minimiser } \mathcal{C}(m) \text{ sur } m \in \mathcal{M} \quad \Rightarrow \quad \widehat{m}_{\mathcal{C}}$$

- Exemples :
  - sélection d'estimateurs :

$$\mathcal{R}(m) = \ell\left(s^\star, \widehat{s}_m\right)$$

Un problème général

- Objectif :

$$\text{minimiser } \mathcal{R}(m) \text{ sur } m \in \mathcal{M}$$

- Méthode :

$$\text{minimiser } \mathcal{C}(m) \text{ sur } m \in \mathcal{M} \quad \Rightarrow \quad \widehat{m}_{\mathcal{C}}$$

- Exemples :
  - sélection d'estimateurs :

$$\mathcal{R}(m) = \ell\left(s^\star, \widehat{s}_m\right)$$

  - estimateur par minimum de contraste sur S :

$$\mathcal{M} = S \subset \mathbb{S} \qquad \mathcal{R}(t) = \ell\left(s^\star, t\right) \qquad \mathcal{C}(t) = P_n\gamma(t)$$

Un problème général

- Objectif :

$$\text{minimiser } \mathcal{R}(m) \text{ sur } m \in \mathcal{M}$$

- Méthode :

$$\text{minimiser } \mathcal{C}(m) \text{ sur } m \in \mathcal{M} \quad \Rightarrow \quad \widehat{m}_{\mathcal{C}}$$

- Exemples :
  - sélection d'estimateurs :

$$\mathcal{R}(m) = \ell\left(s^\star, \widehat{s}_m\right)$$

  - estimateur par minimum de contraste sur S :

$$\mathcal{M} = S \subset \mathbb{S} \qquad \mathcal{R}(t) = \ell\left(s^\star, t\right) \qquad \mathcal{C}(t) = P_n\gamma(t)$$

  - relaxations (convexes) en optimisation

Analyse du problème : un lemme

Lemme

Si

$$\forall m \in \mathcal{M}, \quad -B(m) \leq \mathcal{C}(m) - \mathcal{R}(m) \leq A(m)$$

alors,

$$\forall \widehat{m}_{\mathcal{C}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{C}(m) \right\}, \quad \mathcal{R}(\widehat{m}_{\mathcal{C}}) - B(\widehat{m}_{\mathcal{C}}) \leq \inf_{m \in \mathcal{M}} \left\{ \mathcal{R}(m) + A(m) \right\} . \tag{1}$$

Analyse du problème : un lemme

Lemme

Si

$$\forall m \in \mathcal{M}, \quad -B(m) \leq \mathcal{C}(m) - \mathcal{R}(m) \leq A(m)$$

alors,

$$\begin{aligned} &\forall \widehat{m}_{\mathcal{C}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{C}(m) \right\}, \\ &\mathcal{R}(\widehat{m}_{\mathcal{C}}) - B(\widehat{m}_{\mathcal{C}}) \leq \inf_{m \in \mathcal{M}} \left\{ \mathcal{R}(m) + A(m) \right\} . \end{aligned} \tag{1}$$

En fait, (1) *a lieu dès que*

$$\begin{aligned} \forall m, m' \in \mathcal{M}, \quad &\left(\mathcal{C}(m) - \mathcal{R}(m) \right) - \left(\mathcal{C}(m') - \mathcal{R}(m') \right) \\ &\leq A(m) + B(m') . \end{aligned}$$

Optimalité au premier ordre

- Principe d'estimation sans biais du risque (Mallows, Akaike, 1973) : choisir $\mathcal{C}$ tel que

$$\forall m \in \mathcal{M}, \qquad \mathbb{E}\big[\mathcal{C}(m)\big] = \mathbb{E}\big[\mathcal{R}(m)\big] \ .$$

Optimalité au premier ordre

- Principe d'estimation sans biais du risque (Mallows, Akaike, 1973) : choisir $\mathcal{C}$ tel que

$$\forall m \in \mathcal{M}, \qquad \mathbb{E}\big[\mathcal{C}(m)\big] = \mathbb{E}\big[\mathcal{R}(m)\big] \ .$$

- Sous réserve d'inégalités de concentration (uniformes sur $m \in \mathcal{M}$), avec grande probabilité,

$$\forall m \in \mathcal{M}, \quad -\delta_n \mathcal{R}(m) \leq \mathcal{C}(m) - \mathcal{R}(m) \leq \delta_n \mathcal{R}(m)$$

avec $\delta_n \in]0,1[$.

Optimalité au premier ordre

- Principe d'estimation sans biais du risque (Mallows, Akaike, 1973) : choisir $\mathcal{C}$ tel que

$$\forall m \in \mathcal{M}, \qquad \mathbb{E}\big[\mathcal{C}(m)\big] = \mathbb{E}\big[\mathcal{R}(m)\big] \ .$$

- Sous réserve d'inégalités de concentration (uniformes sur $m \in \mathcal{M}$), avec grande probabilité,

$$\forall m \in \mathcal{M}, \quad -\delta_n \mathcal{R}(m) \leq \mathcal{C}(m) - \mathcal{R}(m) \leq \delta_n \mathcal{R}(m)$$

avec $\delta_n \in]0,1[$.

$\Rightarrow$ d'après le lemme,

$$\forall \widehat{m}_{\mathcal{C}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \big\{\mathcal{C}(m)\big\}, \quad \mathcal{R}(\widehat{m}_{\mathcal{C}}) \leq \frac{1+\delta_n}{1-\delta_n} \inf_{m \in \mathcal{M}} \big\{\mathcal{R}(m)\big\} \ .$$

Optimalité au premier ordre

- Principe d'estimation sans biais du risque (Mallows, Akaike, 1973) : choisir $\mathcal{C}$ tel que

$$\forall m \in \mathcal{M}, \qquad \mathbb{E}\big[\mathcal{C}(m)\big] = \mathbb{E}\big[\mathcal{R}(m)\big] \ .$$

- Sous réserve d'inégalités de concentration (uniformes sur $m \in \mathcal{M}$), avec grande probabilité,

$$\forall m \in \mathcal{M}, \quad -\delta_n \mathcal{R}(m) \leq \mathcal{C}(m) - \mathcal{R}(m) \leq \delta_n \mathcal{R}(m)$$

avec $\delta_n \in \,]0,1[$.

$\Rightarrow$ d'après le lemme,

$$\forall \widehat{m}_{\mathcal{C}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \big\{\mathcal{C}(m)\big\}, \quad \mathcal{R}(\widehat{m}_{\mathcal{C}}) \leq \frac{1+\delta_n}{1-\delta_n} \inf_{m \in \mathcal{M}} \big\{\mathcal{R}(m)\big\} \ .$$

- Optimal au premier ordre si $\delta_n \to 0$.

Une autre utilisation du lemme

- Choisir $\mathcal{C}$ qui est une borne supérieure sur $\mathcal{R}$, uniformément sur $m \in \mathcal{M}$, c'est-à-dire tel que

$$\forall m \in \mathcal{M}, \qquad \mathcal{C}(m) \geq \mathcal{R}(m) \ .$$

Une autre utilisation du lemme

- Choisir $\mathcal{C}$ qui est une borne supérieure sur $\mathcal{R}$, uniformément sur $m \in \mathcal{M}$, c'est-à-dire tel que

$$\forall m \in \mathcal{M}, \qquad \mathcal{C}(m) \geq \mathcal{R}(m) \ .$$

$\Rightarrow$ d'après le lemme,

$$\forall \widehat{m}_{\mathcal{C}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{C}(m) \right\}, \quad \mathcal{R}(\widehat{m}_{\mathcal{C}}) \leq \inf_{m \in \mathcal{M}} \left\{ \mathcal{C}(m) \right\} \ .$$

Une autre utilisation du lemme

- Choisir $\mathcal{C}$ qui est une borne supérieure sur $\mathcal{R}$, uniformément sur $m \in \mathcal{M}$, c'est-à-dire tel que

$$\forall m \in \mathcal{M}, \qquad \mathcal{C}(m) \geq \mathcal{R}(m) \ .$$

$\Rightarrow$ d'après le lemme,

$$\forall \widehat{m}_{\mathcal{C}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{C}(m) \right\}, \quad \mathcal{R}(\widehat{m}_{\mathcal{C}}) \leq \inf_{m \in \mathcal{M}} \left\{ \mathcal{C}(m) \right\} \ .$$

- Exemples :
  - grandes collections d'estimateurs (sélection de variables, détection de ruptures, etc.)

Une autre utilisation du lemme

- Choisir $\mathcal{C}$ qui est une borne supérieure sur $\mathcal{R}$, uniformément sur $m \in \mathcal{M}$, c'est-à-dire tel que

$$\forall m \in \mathcal{M}, \qquad \mathcal{C}(m) \geq \mathcal{R}(m) \ .$$

$\Rightarrow$ d'après le lemme,

$$\forall \widehat{m}_{\mathcal{C}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{C}(m) \right\}, \quad \mathcal{R}(\widehat{m}_{\mathcal{C}}) \leq \inf_{m \in \mathcal{M}} \left\{ \mathcal{C}(m) \right\} \ .$$

- Exemples :
  - grandes collections d'estimateurs (sélection de variables, détection de ruptures, etc.)
  - minimisation du risque structurel (Vapnik)

Une autre utilisation du lemme

- Choisir $\mathcal{C}$ qui est une borne supérieure sur $\mathcal{R}$, uniformément sur $m \in \mathcal{M}$, c'est-à-dire tel que

$$\forall m \in \mathcal{M}, \qquad \mathcal{C}(m) \geq \mathcal{R}(m) \ .$$

$\Rightarrow$ d'après le lemme,

$$\forall \widehat{m}_{\mathcal{C}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{C}(m) \right\}, \quad \mathcal{R}(\widehat{m}_{\mathcal{C}}) \leq \inf_{m \in \mathcal{M}} \left\{ \mathcal{C}(m) \right\} \ .$$

- Exemples :
  - grandes collections d'estimateurs (sélection de variables, détection de ruptures, etc.)
  - minimisation du risque structurel (Vapnik)
  - relaxations en optimisation

Analyse au second ordre ?

- Comment comparer $\mathcal{C}_1$ et $\mathcal{C}_2$ tels que

$$\forall m \in \mathcal{M}, \qquad \mathbb{E}\big[\mathcal{C}_1(m)\big] = \mathbb{E}\big[\mathcal{C}_2(m)\big] \ ?$$

Analyse au second ordre ?

- Comment comparer $\mathcal{C}_1$ et $\mathcal{C}_2$ tels que

$$\forall m \in \mathcal{M}, \qquad \mathbb{E}\left[\mathcal{C}_1(m)\right] = \mathbb{E}\left[\mathcal{C}_2(m)\right] \ ?$$

- Tenir compte de la variance $\mathrm{var}\left(\mathcal{C}_i(m)\right)$?

Analyse au second ordre ?

- Comment comparer $\mathcal{C}_1$ et $\mathcal{C}_2$ tels que

$$\forall m \in \mathcal{M}, \qquad \mathbb{E}\big[\mathcal{C}_1(m)\big] = \mathbb{E}\big[\mathcal{C}_2(m)\big] \ ?$$

- Tenir compte de la variance $\mathrm{var}\big(\mathcal{C}_i(m)\big)$?
- Variance de quelle quantité ?

  Pour toute variable Z, $\widehat{m}_{\mathcal{C}} \in \mathrm{argmin}_{m \in \mathcal{M}} \big\{\mathcal{C}(m) + Z\big\}$

  mais $\mathrm{var}\big(\mathcal{C}(m) + Z\big)$ dépend de Z...

Analyse au second ordre ?

- Comment comparer $\mathcal{C}_1$ et $\mathcal{C}_2$ tels que

$$\forall m \in \mathcal{M}, \qquad \mathbb{E}\left[\mathcal{C}_1(m)\right] = \mathbb{E}\left[\mathcal{C}_2(m)\right] \ ?$$

- Tenir compte de la variance $\mathrm{var}\left(\mathcal{C}_i(m)\right)$?

- Variance de quelle quantité ?

  Pour toute variable Z, $\widehat{m}_{\mathcal{C}} \in \mathrm{argmin}_{m \in \mathcal{M}} \left\{\mathcal{C}(m) + Z\right\}$
  mais $\mathrm{var}\left(\mathcal{C}(m) + Z\right)$ dépend de Z...

$\Rightarrow$ variance des incréments

$$\mathrm{var}\left(\mathcal{C}(m) - \mathcal{C}(m')\right) \ .$$

Plan

1 Sélection d'estimateurs

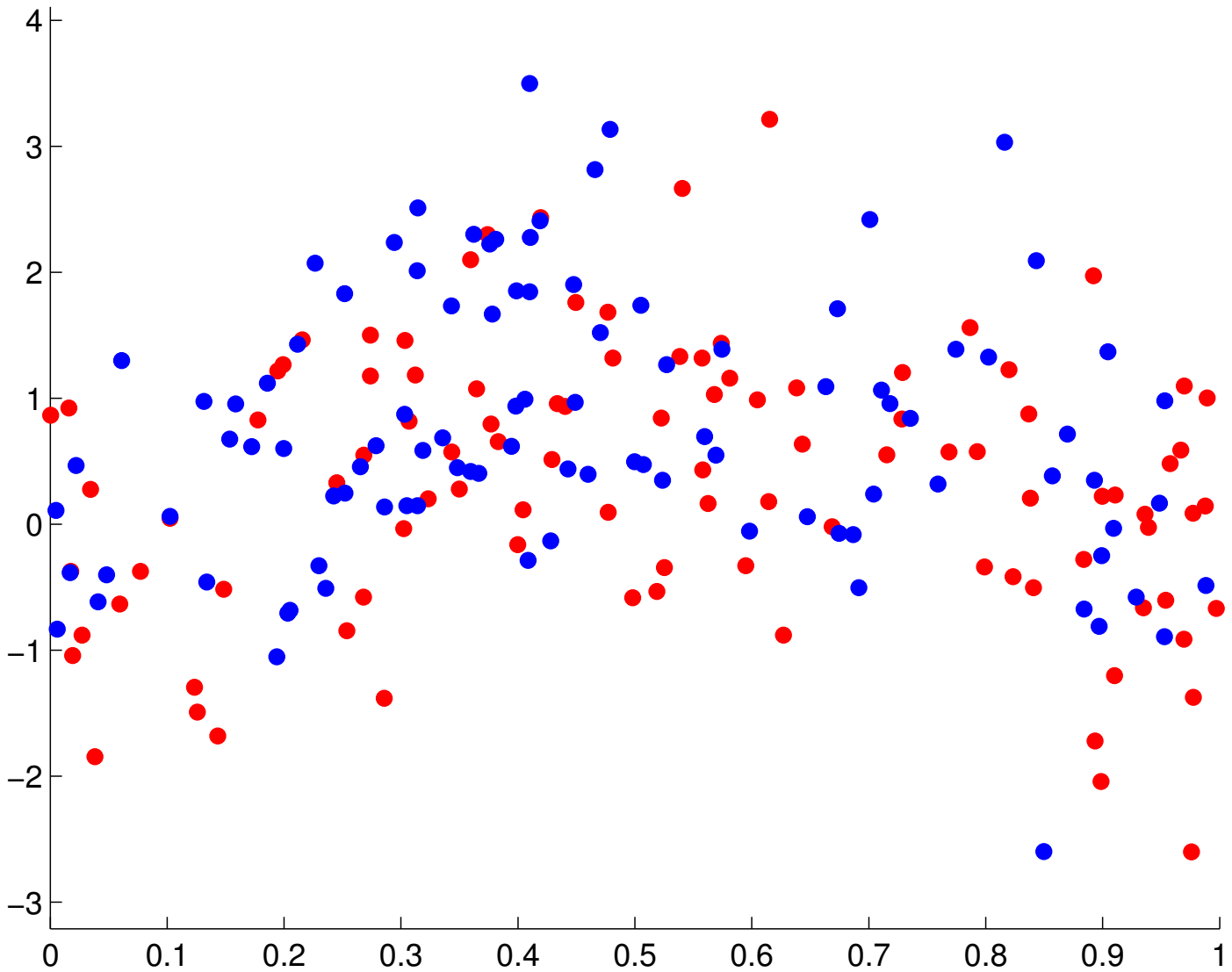
2 Validation croisée

3 Pénalités minimales

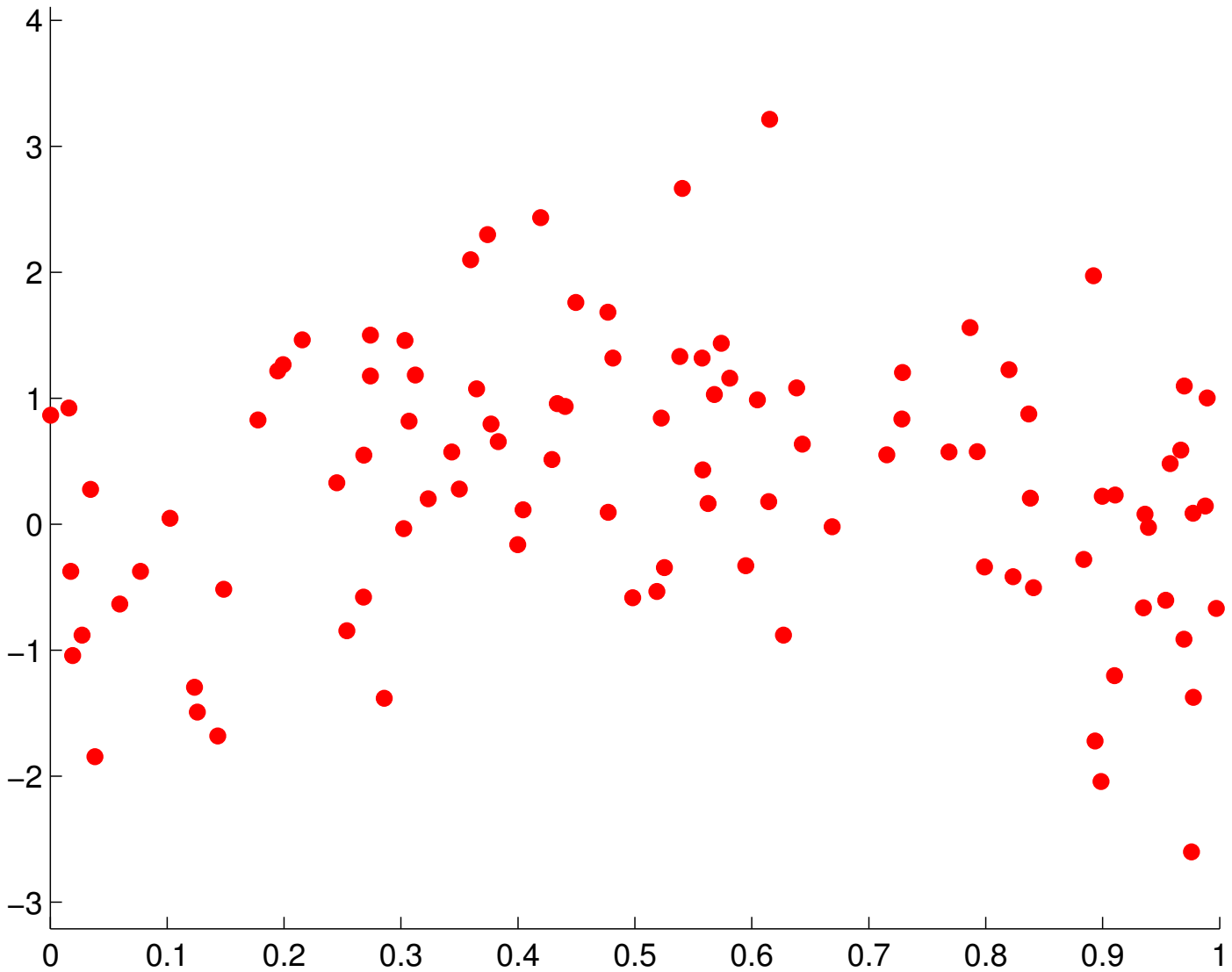
4 Détection de ruptures

5 Conclusion

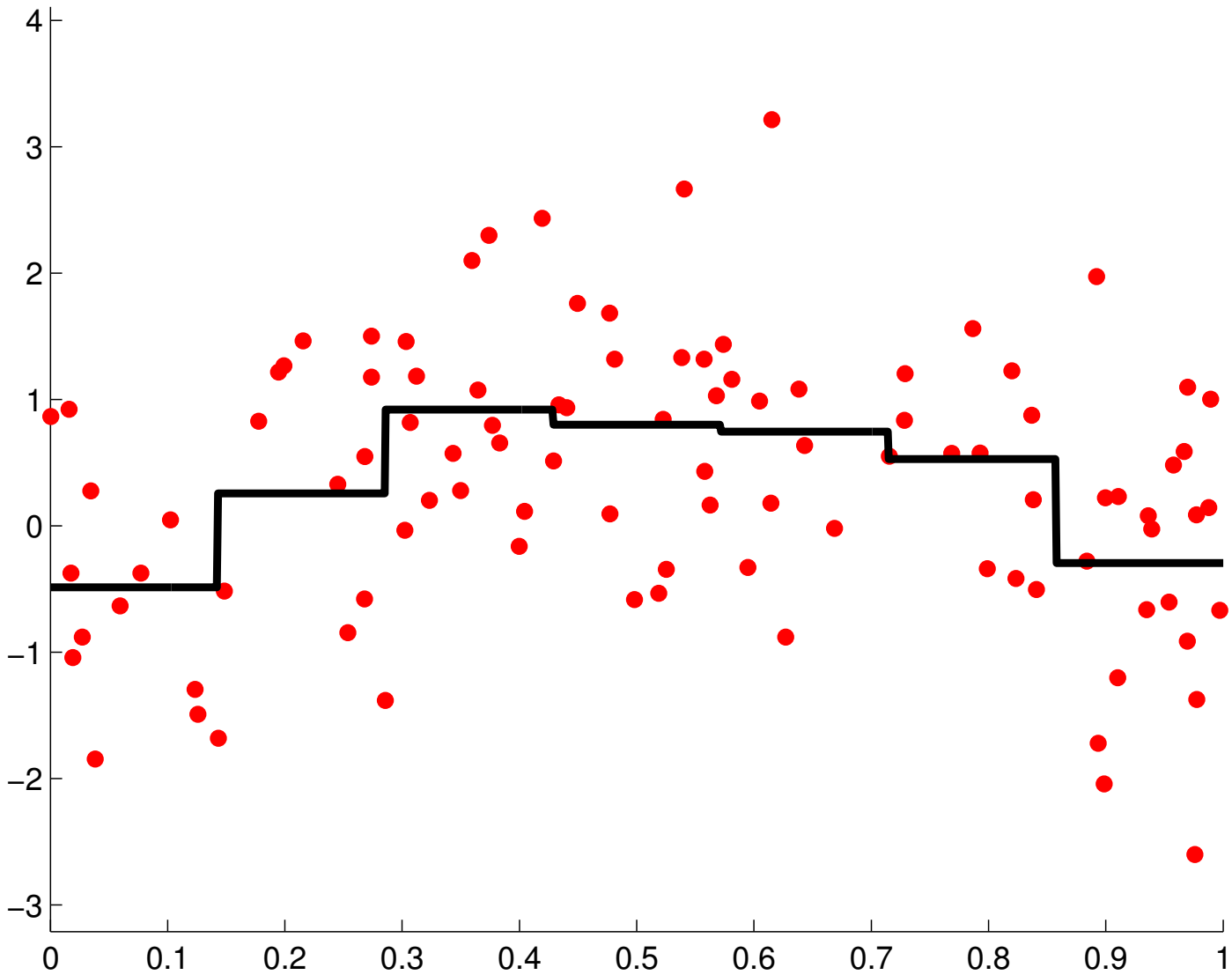
Principe de la validation simple



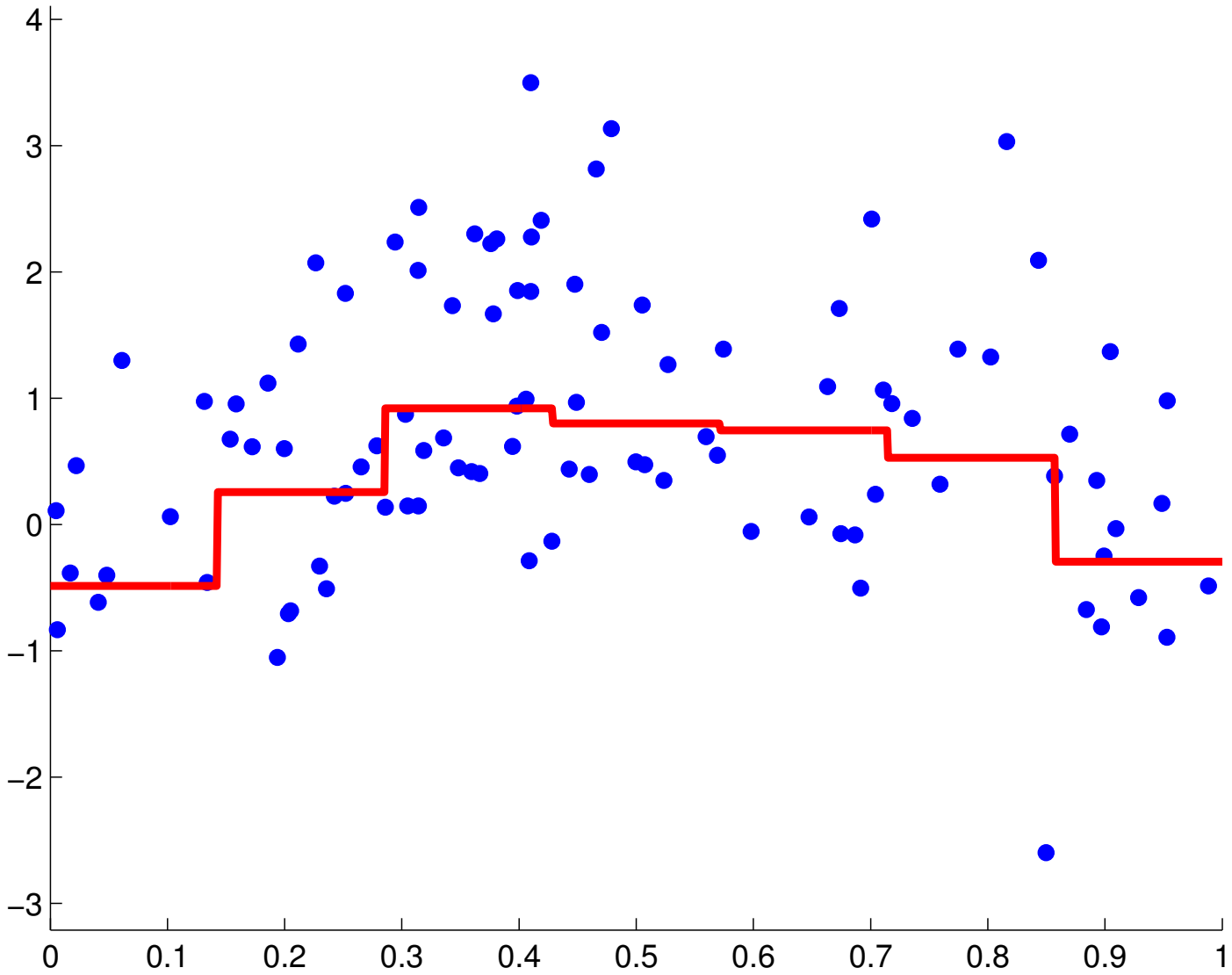
Principe de la validation : échantillon d'entraînement



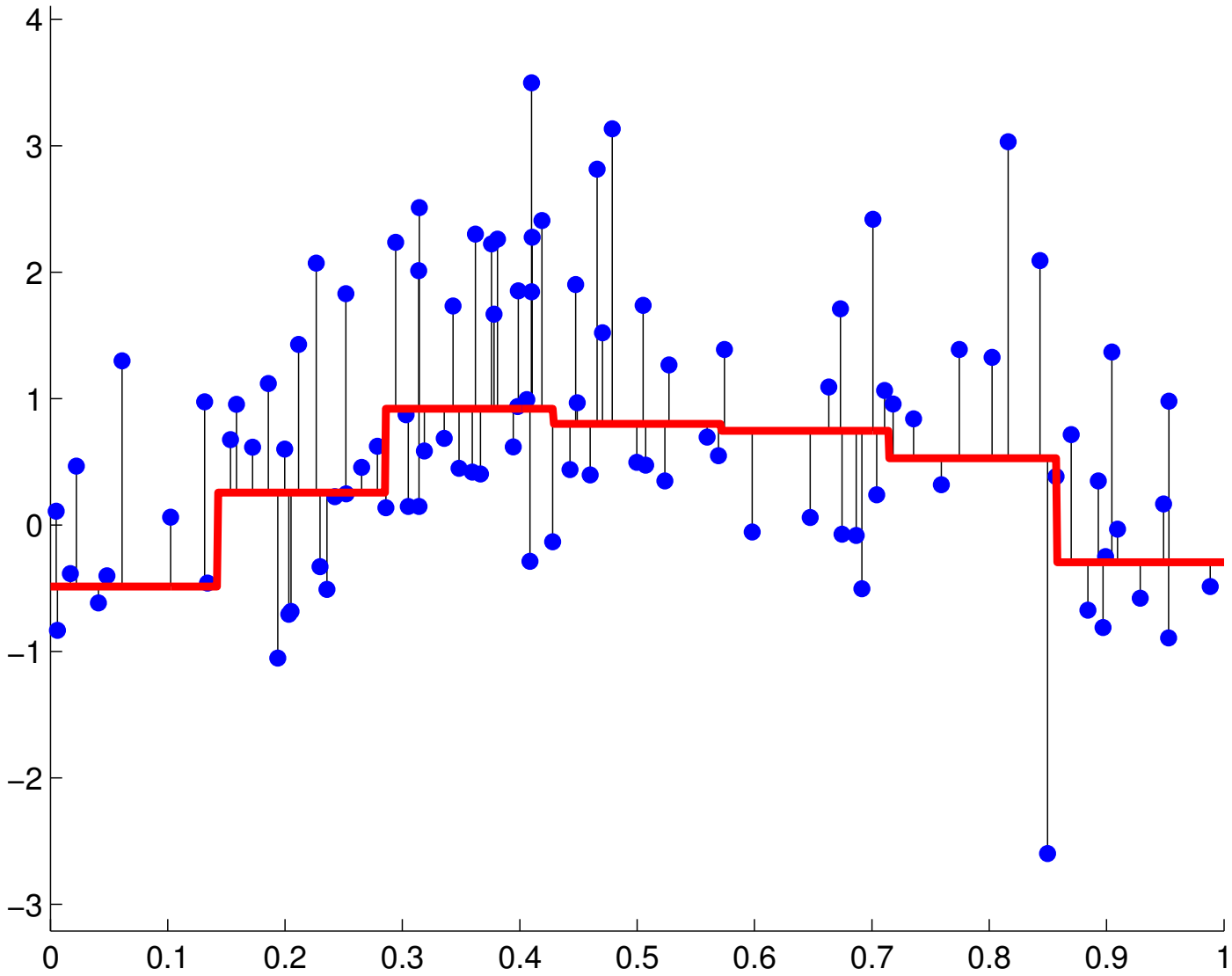
Principe de la validation : échantillon d'entraînement



Principe de la validation : échantillon de validation



Principe de la validation : échantillon de validation



Validation croisée « V-fold »

$$\underbrace{\xi_1, \ldots, \xi_q}_{\text{Entraînement}}, \underbrace{\xi_{q+1}, \ldots, \xi_n}_{\text{Validation}}$$

Validation croisée « V-fold » :

$\mathcal{B} = (B_j)_{1 \leq j \leq V}$ partition de $\{1, \ldots, n\}$

$$\Rightarrow \qquad \widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{s}_m; D_n; \mathcal{B}\right) = \frac{1}{V} \sum_{j=1}^{V} P_n^j \gamma\left(\widehat{s}_m^{(-j)}\right)$$

$$\widehat{m} \in \arg\min_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{s}_m\right) \right\}$$

A. & Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

Au premier ordre : biais de la validation croisée

- Hypothèse : $\mathrm{Card}(B_j) = n/V$ pour tout j.
- Calcul d'espérances :

$$\mathbb{E}\left[P\gamma\left(\widehat{s}_m(D_n)\right) \right] \approx \alpha(m) + \frac{\beta(m)}{n}$$

$$\Rightarrow \quad \mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vf}}(\widehat{s}_m; D_n; \mathcal{B}) \right] = \mathbb{E}\left[P_n^{(j)}\gamma\left(\widehat{s}_m^{(-j)}\right) \right] = \mathbb{E}\left[P\gamma\left(\widehat{s}_m^{(-j)}\right) \right]$$

$$\approx \alpha(m) + \frac{V}{V-1}\frac{\beta(m)}{n}$$

Au premier ordre : biais de la validation croisée

- Hypothèse : $\mathrm{Card}(B_j) = n/V$ pour tout j.
- Calcul d'espérances :

$$\mathbb{E}\left[P\gamma\left(\widehat{s}_m(D_n)\right) \right] \approx \alpha(m) + \frac{\beta(m)}{n}$$

$$\Rightarrow \quad \mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vf}}(\widehat{s}_m; D_n; \mathcal{B}) \right] = \mathbb{E}\left[P_n^{(j)}\gamma\left(\widehat{s}_m^{(-j)}\right) \right] = \mathbb{E}\left[P\gamma\left(\widehat{s}_m^{(-j)}\right) \right]$$

$$\approx \alpha(m) + \frac{V}{V-1}\frac{\beta(m)}{n}$$

$\Rightarrow$ biais, décroissant avec V, tend vers zéro quand $V \to +\infty$

Au premier ordre : biais de la validation croisée

- Hypothèse : $\mathrm{Card}(B_j) = n/V$ pour tout j.
- Calcul d'espérances :

$$\mathbb{E}\left[P\gamma\left(\widehat{s}_m(D_n)\right) \right] \approx \alpha(m) + \frac{\beta(m)}{n}$$

$$\Rightarrow \quad \mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vf}}(\widehat{s}_m; D_n; \mathcal{B}) \right] = \mathbb{E}\left[P_n^{(j)}\gamma\left(\widehat{s}_m^{(-j)}\right) \right] = \mathbb{E}\left[P\gamma\left(\widehat{s}_m^{(-j)}\right) \right]$$

$$\approx \alpha(m) + \frac{V}{V-1}\frac{\beta(m)}{n}$$

$\Rightarrow$ biais, décroissant avec V, tend vers zéro quand $V \to +\infty$

$\Rightarrow$ sous-optimalité de la validation croisée « V-fold » à V fixé (prouvé pour les régressogrammes, valable plus largement)

A. V-fold cross-validation improved : V-fold penalization, 2008. arXiv:0802.0566v2

Correction du biais et pénalisation « V-fold »

- Validation croisée « V-fold » corrigée (Burman, 1989) :

$$\widehat{\mathcal{R}}^{\mathrm{vf,corr}}(\widehat{s}_m; D_n; \mathcal{B}) := \widehat{\mathcal{R}}^{\mathrm{vf}}(\widehat{s}_m; D_n; \mathcal{B}) + P_n\gamma(\widehat{s}_m) - \frac{1}{V}\sum_{j=1}^{V} P_n\gamma\left(\widehat{s}_m^{(-j)}\right)$$

Correction du biais et pénalisation « V-fold »

- Validation croisée « V-fold » corrigée (Burman, 1989) :

$$\widehat{\mathcal{R}}^{\mathrm{vf,corr}}(\widehat{s}_m; D_n; \mathcal{B}) := \widehat{\mathcal{R}}^{\mathrm{vf}}(\widehat{s}_m; D_n; \mathcal{B}) + P_n\gamma(\widehat{s}_m) - \frac{1}{V}\sum_{j=1}^{V} P_n\gamma\left(\widehat{s}_m^{(-j)}\right)$$

- Heuristique de rééchantillonnage (Efron, 1983), sous-échantillonnage « V-fold » et pénalisation
  $\Rightarrow$ pénalité « V-fold »

$$\mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n; \mathcal{B}) := \frac{V-1}{V}\sum_{j=1}^{V}\left(P_n - P_n^{(-j)}\right)\gamma\left(\widehat{s}_m^{(-j)}\right)$$

A. V-fold cross-validation improved : V-fold penalization, 2008. arXiv:0802.0566v2

Correction du biais et pénalisation « V-fold »

- Validation croisée « V-fold » corrigée (Burman, 1989) :

$$\widehat{\mathcal{R}}^{\mathrm{vf,corr}}(\widehat{s}_m; D_n; \mathcal{B}) := \widehat{\mathcal{R}}^{\mathrm{vf}}(\widehat{s}_m; D_n; \mathcal{B}) + P_n\gamma(\widehat{s}_m) - \frac{1}{V}\sum_{j=1}^{V} P_n\gamma\left(\widehat{s}_m^{(-j)}\right)$$

- Heuristique de rééchantillonnage (Efron, 1983), sous-échantillonnage « V-fold » et pénalisation
  $\Rightarrow$ pénalité « V-fold »

$$\mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n; \mathcal{B}) := \frac{V-1}{V}\sum_{j=1}^{V}\left(P_n - P_n^{(-j)}\right)\gamma\left(\widehat{s}_m^{(-j)}\right)$$

A. V-fold cross-validation improved : V-fold penalization, 2008. arXiv:0802.0566v2

- $\widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{s}_m; D_n; \mathcal{B}\right) = P_n\gamma\left(\widehat{s}_m(D_n)\right) + \mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n; \mathcal{B})$

Correction du biais et pénalisation « V-fold »

- Validation croisée « V-fold » corrigée (Burman, 1989) :

$$\widehat{\mathcal{R}}^{\mathrm{vf,corr}}(\widehat{s}_m; D_n; \mathcal{B}) := \widehat{\mathcal{R}}^{\mathrm{vf}}(\widehat{s}_m; D_n; \mathcal{B}) + P_n\gamma(\widehat{s}_m) - \frac{1}{V}\sum_{j=1}^{V} P_n\gamma\left(\widehat{s}_m^{(-j)}\right)$$

- Heuristique de rééchantillonnage (Efron, 1983), sous-échantillonnage « V-fold » et pénalisation
  $\Rightarrow$ pénalité « V-fold »

$$\mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n; \mathcal{B}) := \frac{V-1}{V}\sum_{j=1}^{V}\left(P_n - P_n^{(-j)}\right)\gamma\left(\widehat{s}_m^{(-j)}\right)$$

A. V-fold cross-validation improved : V-fold penalization, 2008. arXiv:0802.0566v2

- $\widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{s}_m; D_n; \mathcal{B}\right) = P_n\gamma\left(\widehat{s}_m(D_n)\right) + \mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n; \mathcal{B})$
- Sans biais si $\mathbb{E}\left[(P - P_n)\left(\widehat{s}_m(D_n)\right)\right] = \gamma(m)/n$

Inégalités oracle optimales pour la pénalisation « V-fold »

Théorème

Avec probabilité $1 - n^{-2}$, $\forall \delta > 0$,

$$\forall \widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ P_n \gamma\left(\widehat{s}_m(D_n)\right) + \mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n; \mathcal{B}) \right\},$$

$$\ell\left(s^\star, \widehat{s}_{\widehat{m}}\right) \leq \left(1+\delta\right) \inf_{m \in \mathcal{M}} \left\{ \ell\left(s^\star, \widehat{s}_m\right) \right\} + \frac{L\left[\log\left(\mathrm{Card}(\mathcal{M})\right) \vee \log(n)\right]^\alpha}{\delta^\beta \, n}$$

$\Rightarrow$ Optimal au premier ordre si $\mathrm{Card}(\mathcal{M}) \leq a n^b$

Inégalités oracle optimales pour la pénalisation « V-fold »

Théorème

Avec probabilité $1 - n^{-2}$, $\forall \delta > 0$,

$$\forall \widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ P_n \gamma\left(\widehat{s}_m(D_n)\right) + \mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n; \mathcal{B}) \right\},$$

$$\ell\left(s^\star, \widehat{s}_{\widehat{m}}\right) \leq \left(1+\delta\right) \inf_{m \in \mathcal{M}} \left\{ \ell\left(s^\star, \widehat{s}_m\right) \right\} + \frac{L\left[\log\left(\mathrm{Card}(\mathcal{M})\right) \vee \log(n)\right]^\alpha}{\delta^\beta \, n}$$

$\Rightarrow$ Optimal au premier ordre si $\mathrm{Card}(\mathcal{M}) \leq a n^b$

Valable sous des hypothèses raisonnablement faibles pour :

- Les régressogrammes en régression hétéroscédastique

A. V-fold cross-validation improved : V-fold penalization, 2008. arXiv:0802.0566v2

A. Model selection by resampling penalization. *Electronic Journal of Statistics*, 3:557–624, 2009.

Inégalités oracle optimales pour la pénalisation « V-fold »

Théorème

Avec probabilité $1 - n^{-2}$, $\forall \delta > 0$,

$$\forall \widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ P_n \gamma\left(\widehat{s}_m(D_n)\right) + \mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n; \mathcal{B}) \right\},$$

$$\ell\left(s^\star, \widehat{s}_{\widehat{m}}\right) \leq \left(1+\delta\right) \inf_{m \in \mathcal{M}} \left\{ \ell\left(s^\star, \widehat{s}_m\right) \right\} + \frac{L\left[\log\left(\mathrm{Card}(\mathcal{M})\right) \vee \log(n)\right]^\alpha}{\delta^\beta \, n}$$

$\Rightarrow$ Optimal au premier ordre si $\mathrm{Card}(\mathcal{M}) \leq a n^b$

Valable sous des hypothèses raisonnablement faibles pour :

- Les régressogrammes en régression hétéroscédastique

A. V-fold cross-validation improved : V-fold penalization, 2008. arXiv:0802.0566v2

A. Model selection by resampling penalization. *Electronic Journal of Statistics*, 3:557–624, 2009.

- L'estimation de densité par moindres carrés

A. & Lerasle. Why $V = 5$ is enough in V-fold cross-validation, 2014. arXiv:1210.5830v2.

A. & Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79, 2010.

Variance et sélection de modèles : heuristique

- Heuristique : pour tout $m \notin \operatorname{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right]$, on veut minimiser $\mathbb{P}(\widehat{m}_{\mathcal{C}} = m)$

Variance et sélection de modèles : heuristique

- Heuristique : pour tout $m \notin \operatorname{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right]$, on veut minimiser $\mathbb{P}(\widehat{m}_{\mathcal{C}} = m)$

$$= \mathbb{P}\left(\forall m' \in \mathcal{M}, \mathcal{C}(m) - \mathcal{C}(m') < 0\right)$$

$$\leq \min_{m' \neq m} \mathbb{P}\left(\mathcal{C}(m) - \mathcal{C}(m') < 0\right)$$

Variance et sélection de modèles : heuristique

- Heuristique : pour tout $m \notin \operatorname{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right]$, on veut minimiser $\mathbb{P}(\widehat{m}_{\mathcal{C}} = m)$

$$= \mathbb{P}\left(\forall m' \in \mathcal{M}, \mathcal{C}(m) - \mathcal{C}(m') < 0\right)$$

$$\leq \min_{m' \neq m} \mathbb{P}\left(\mathcal{C}(m) - \mathcal{C}(m') < 0\right)$$

$$\approx \min_{m' \neq m} \mathbb{P}\left(\mathbb{E}\left[\mathcal{C}(m) - \mathcal{C}(m')\right] - \mathcal{N}\sqrt{\operatorname{var}\left(\mathcal{C}(m) - \mathcal{C}(m')\right)} < 0\right)$$

Variance et sélection de modèles : heuristique

- Heuristique : pour tout $m \notin \operatorname{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right]$, on veut minimiser $\mathbb{P}(\widehat{m}_{\mathcal{C}} = m)$

$$= \mathbb{P}\left(\forall m' \in \mathcal{M}, \mathcal{C}(m) - \mathcal{C}(m') < 0\right)$$

$$\leq \min_{m' \neq m} \mathbb{P}\left(\mathcal{C}(m) - \mathcal{C}(m') < 0\right)$$

$$\approx \min_{m' \neq m} \mathbb{P}\left(\mathbb{E}\left[\mathcal{C}(m) - \mathcal{C}(m')\right] - \mathcal{N}\sqrt{\operatorname{var}\left(\mathcal{C}(m) - \mathcal{C}(m')\right)} < 0\right)$$

$$= \overline{\Phi}\left(\max_{m' \neq m} \frac{\mathbb{E}\left[\mathcal{C}(m) - \mathcal{C}(m')\right]}{\sqrt{\operatorname{var}\left(\mathcal{C}(m) - \mathcal{C}(m')\right)}}\right) \qquad \text{où} \quad \Phi(t) = \mathbb{P}(\mathcal{N} > t)$$

Variance et sélection de modèles : heuristique

- Heuristique : pour tout $m \notin \operatorname{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right]$, on veut minimiser $\mathbb{P}(\widehat{m}_{\mathcal{C}} = m)$

$$
\begin{aligned}
&= \mathbb{P}\left(\forall m' \in \mathcal{M}, \mathcal{C}(m) - \mathcal{C}(m') < 0\right) \\
&\leq \min_{m' \neq m} \mathbb{P}\left(\mathcal{C}(m) - \mathcal{C}(m') < 0\right) \\
&\approx \min_{m' \neq m} \mathbb{P}\left(\mathbb{E}\left[\mathcal{C}(m) - \mathcal{C}(m')\right] - \mathcal{N}\sqrt{\operatorname{var}\left(\mathcal{C}(m) - \mathcal{C}(m')\right)} < 0\right) \\
&= \overline{\Phi}\left(\max_{m' \neq m} \frac{\mathbb{E}\left[\mathcal{C}(m) - \mathcal{C}(m')\right]}{\sqrt{\operatorname{var}\left(\mathcal{C}(m) - \mathcal{C}(m')\right)}}\right) \qquad \text{où} \quad \Phi(t) = \mathbb{P}(\mathcal{N} > t)
\end{aligned}
$$

- Biais constant parmi $\left(\widehat{\mathcal{R}}^{\mathrm{vf,corr}}\right)_{2 \leq V \leq n} \Rightarrow$ comparer

$$
\operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{s}_m; D_n; \mathcal{B}\right) - \widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{s}_{m'}; D_n; \mathcal{B}\right)\right)
$$

Variance et sélection de modèles (densité, moindres carrés)

$$\Delta(m, m', V) = \widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{s}_m\right) - \widehat{\mathcal{R}}^{\mathrm{vf,corr}}\left(\widehat{s}_{m'}\right)$$

Théorème

$$\mathrm{var}\left(\Delta(m, m', V)\right) = 4\left(1 + \frac{2}{n} + \frac{1}{n^2}\right)\frac{\mathrm{var}_P\left(s^\star_m - s^\star_{m'}\right)}{n} + 2\left(1 + \frac{4}{V-1} - \frac{1}{n}\right)\underbrace{\frac{B(m, m')}{n^2}}_{\geq 0}$$

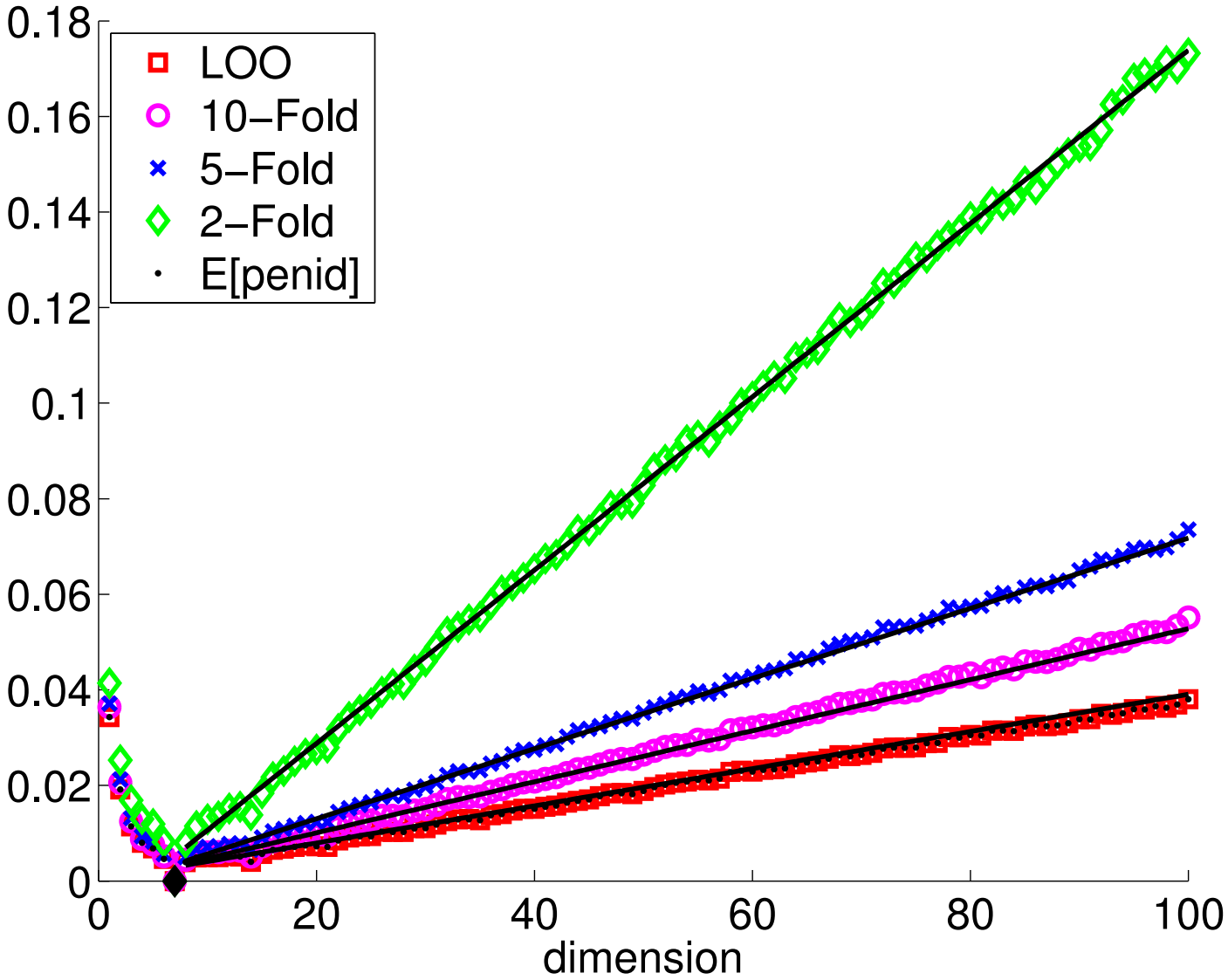
Si de plus $S_m \subset S_{m'}$ sont deux modèles d'histogrammes réguliers de pas d_m^{-1}, $d_{m'}^{-1}$, alors

$$B(m, m') \propto \left\|s^\star_m - s^\star_{m'}\right\| d_m \ .$$

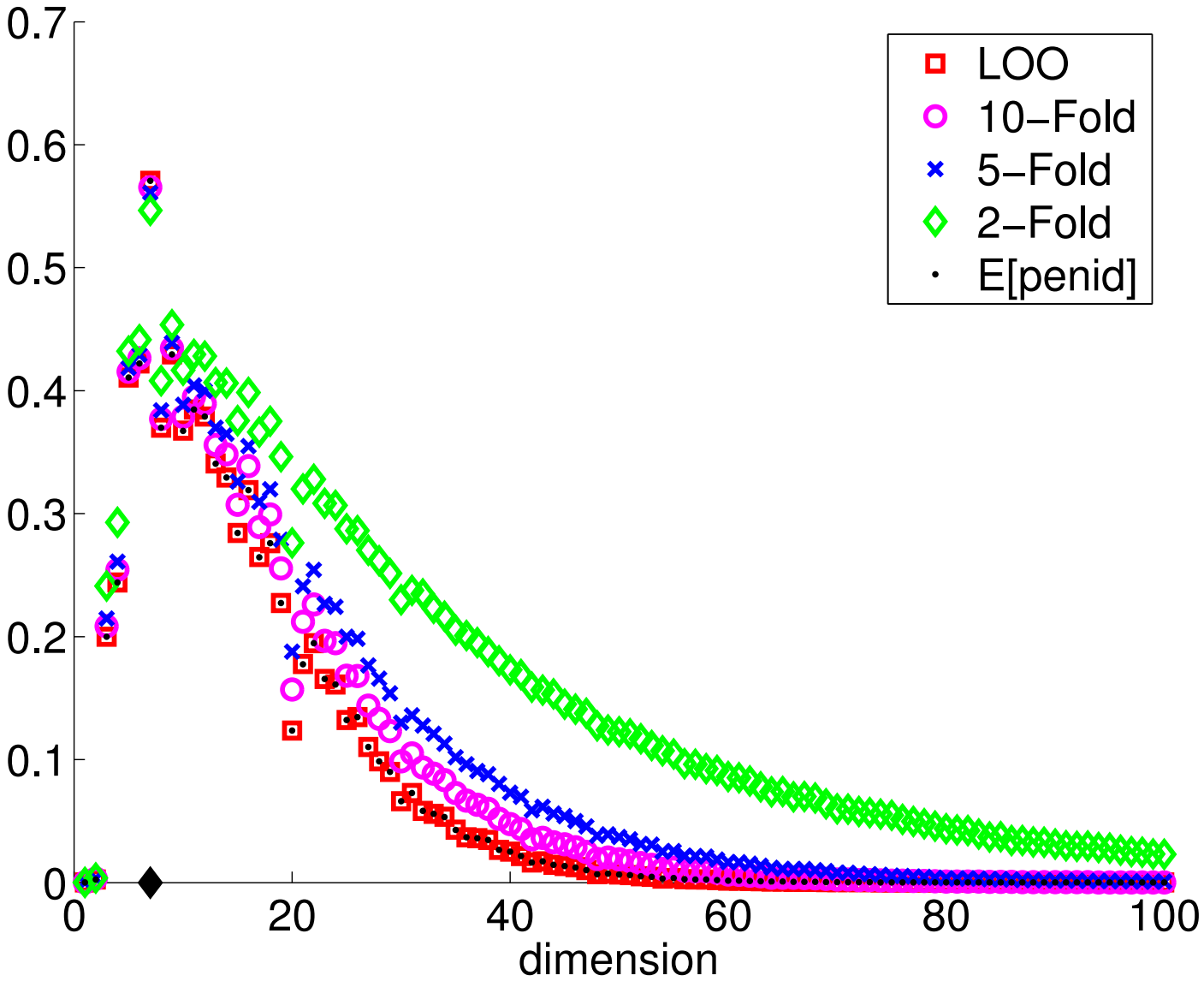
A. & Lerasle. Why $V = 5$ is enough in V-fold cross-validation, 2014. arXiv:1210.5830v2.

Les deux termes sont du même ordre si $\left\|s^\star_m - s^\star_{m'}\right\| \approx d_m/n$.

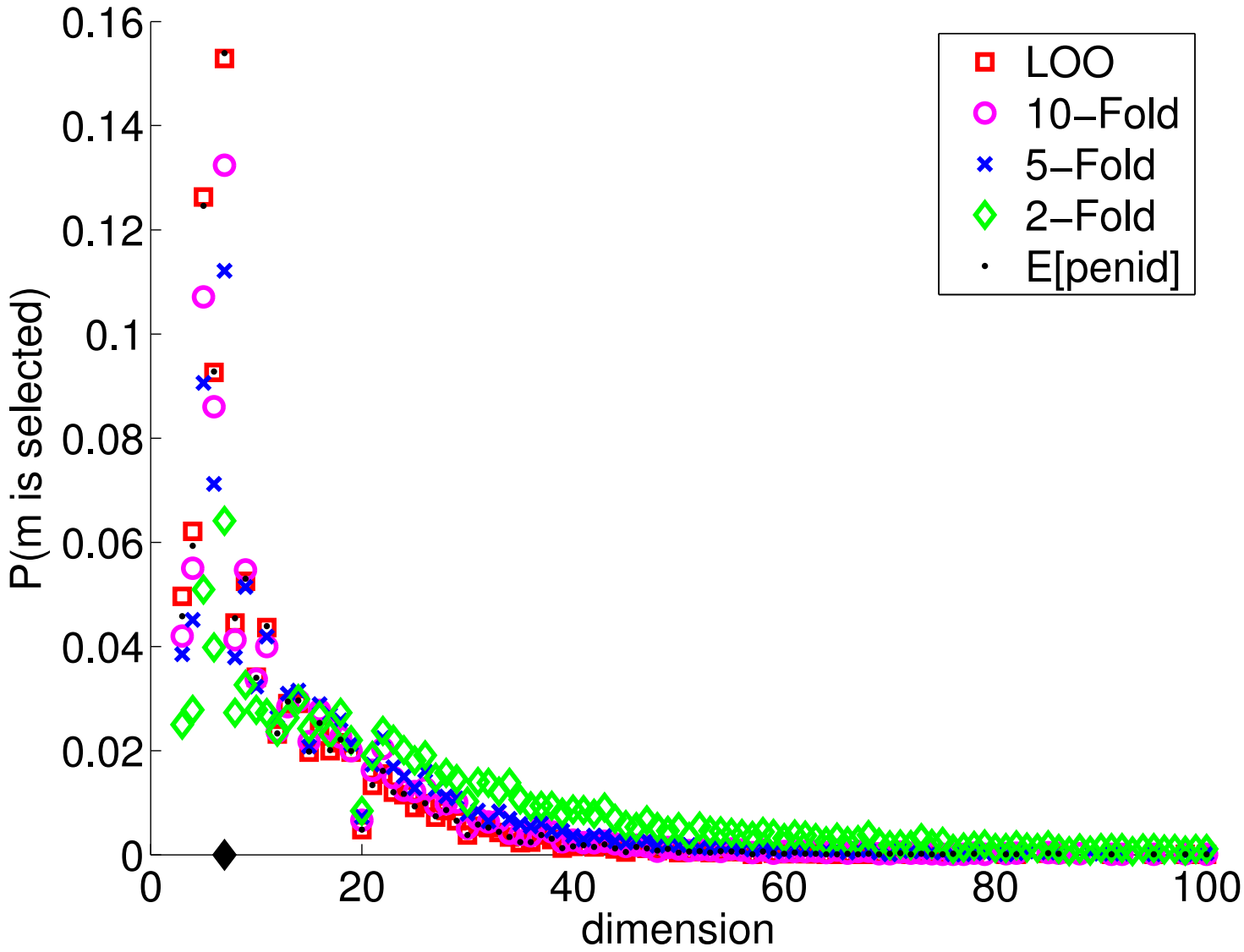
Variance de $\mathcal{C}(m) - \mathcal{C}(m^\star)$ en fonction de d_m et V



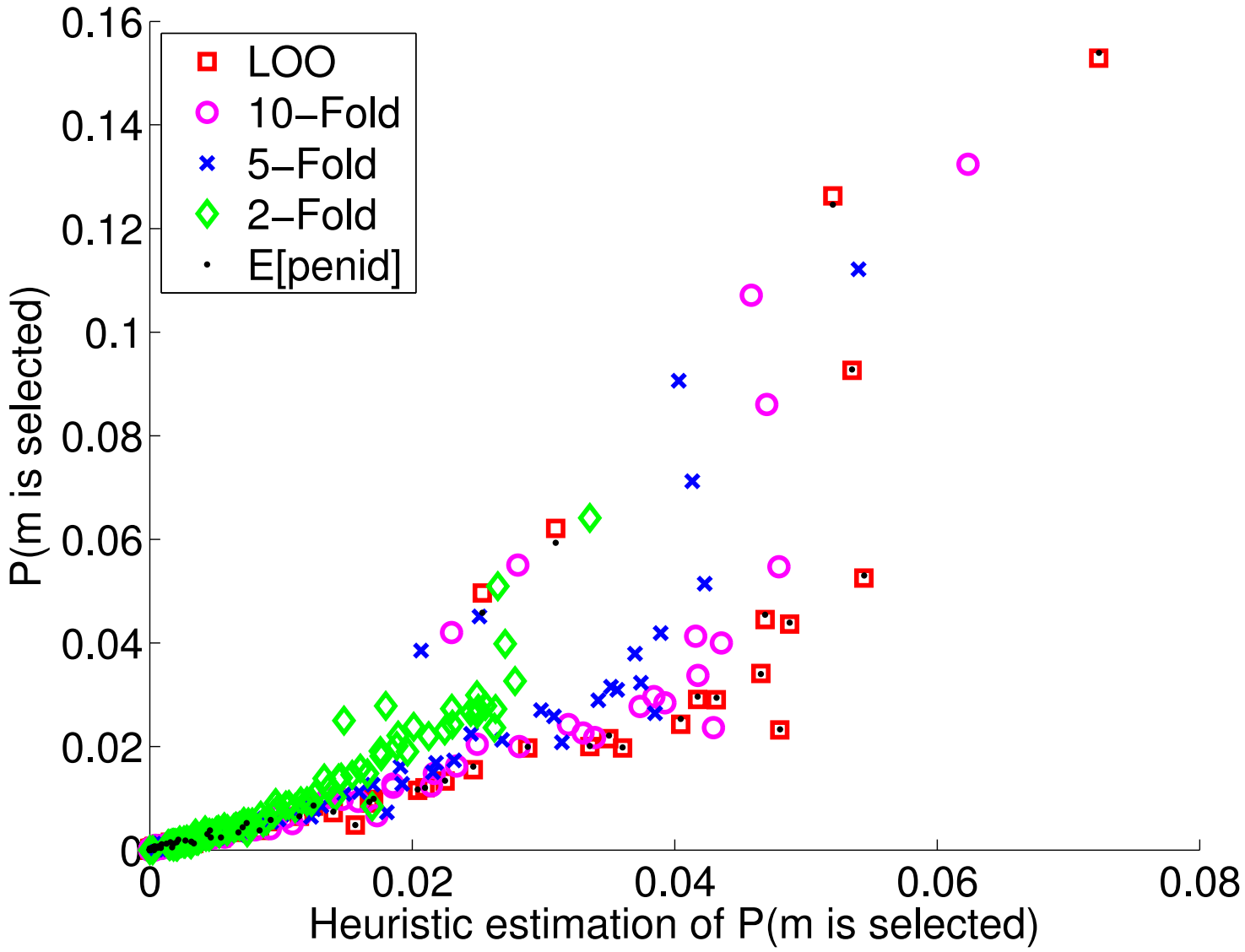
$\overline{\Phi}(\max_{m' \neq m} \mathbb{E}[\mathcal{C}(m) - \mathcal{C}(m')] / \sqrt{\mathrm{var}(\mathcal{C}(m) - \mathcal{C}(m'))})$



Évaluation de l'heuristique : $\mathbb{P}(\widehat{m}_{\mathcal{C}} = m)$, densité



Évaluation de l'heuristique, densité (2)



Conclusion : choix de V pour le « V-fold »

- Biais : fonction décroissante de V
  peut être supprimé (pénalisation)
- Variance : fonction décroissante de V
  proche de son minimum pour $V = 5$ ou 10

$\Rightarrow$ meilleure performance pour les plus grands V
(pas nécessairement vrai en général en apprentissage)

Conclusion : choix de V pour le « V-fold »

- Biais : fonction décroissante de V
  peut être supprimé (pénalisation)
- Variance : fonction décroissante de V
  proche de son minimum pour $V = 5$ ou 10

$\Rightarrow$ meilleure performance pour les plus grands V
(pas nécessairement vrai en général en apprentissage)

- Complexité algorithmique : $\mathcal{O}(V)$ en général

Plan

1 Sélection d'estimateurs

2 Validation croisée

3 Pénalités minimales

4 Détection de ruptures

5 Conclusion

Motivation (1) : calibration de pénalités

Pénalisation :

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ P_n \gamma(\widehat{s}_m) + \text{pen}(m) \right\}$$

$\Rightarrow$ pénalité idéale $\text{pen}_{\text{id}}(m) = (P - P_n)\gamma(\widehat{s}_m)$.

Motivation (1) : calibration de pénalités

Pénalisation :

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ P_n \gamma(\widehat{s}_m) + \operatorname{pen}(m) \right\}$$

$\Rightarrow$ pénalité idéale $\operatorname{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma(\widehat{s}_m)$.

- C_p et C_L (Mallows, 1973) :

$$\frac{2\sigma^2 D_m}{n} \qquad \frac{2\sigma^2 \operatorname{tr}(A_m)}{n}$$

Motivation (1) : calibration de pénalités

Pénalisation :

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ P_n \gamma(\widehat{s}_m) + \operatorname{pen}(m) \right\}$$

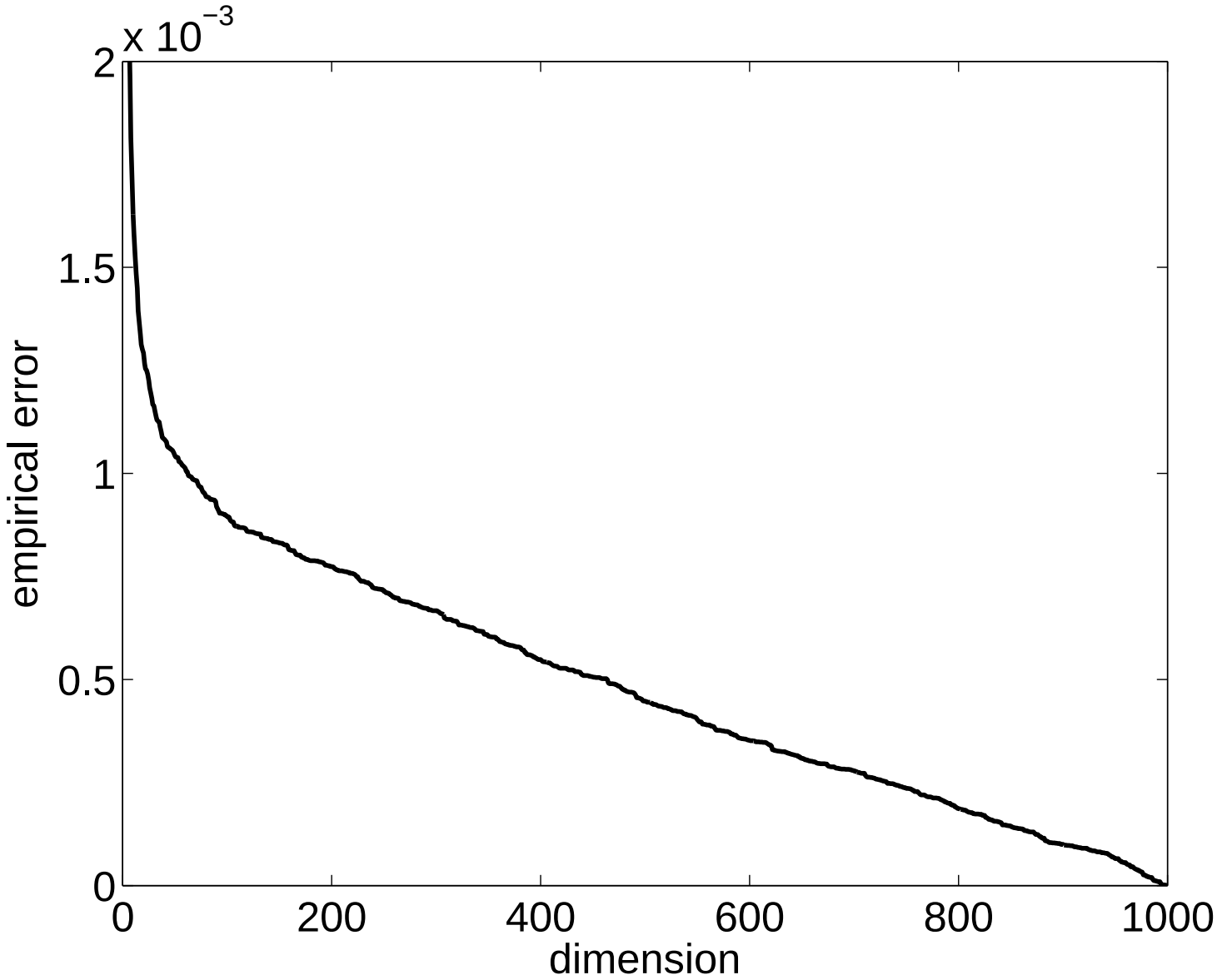
$\Rightarrow$ pénalité idéale $\operatorname{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma(\widehat{s}_m)$.

- C_p et C_L (Mallows, 1973) :

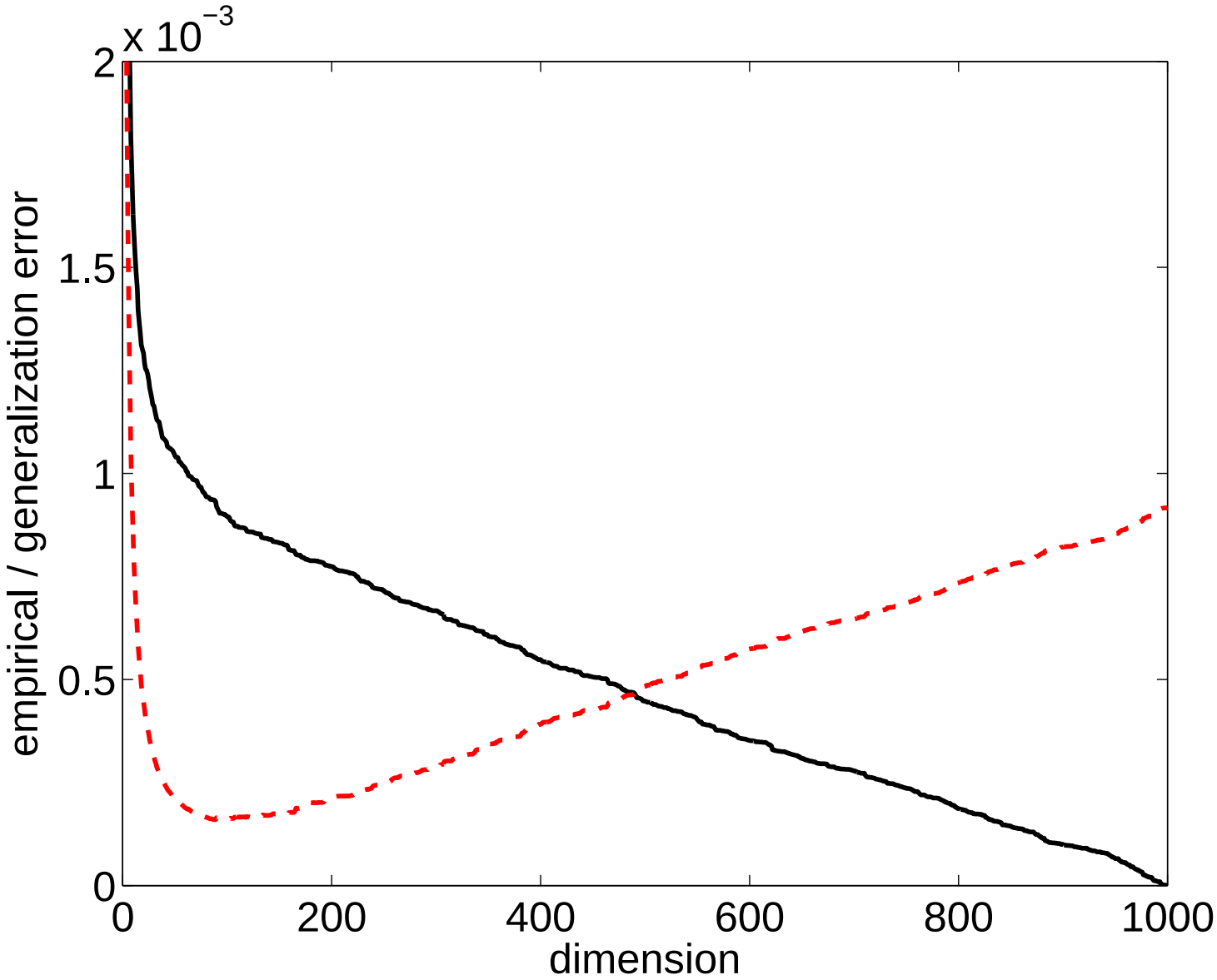
$$\frac{2\sigma^2 D_m}{n} \qquad \frac{2\sigma^2 \operatorname{tr}(A_m)}{n}$$

- Pénalités justifiées asymptotiquement : AIC, BIC
- Pénalités par rééchantillonnage
- Nombreuses pénalités faisant intervenir une ou plusieurs constante dont la valeur optimale est inconnue (complexités de Rademacher, détection de ruptures, etc.)

Motivation (2) : « L-curve » et heuristique de coude



Motivation (2) : « L-curve » et heuristique de coude



Motivation (3) : niveau minimal de pénalisation

Régression, moindres carrés :

- Pénalité optimale (C_p, Mallows, 1973) :

$$\frac{2\sigma^2 D_m}{n}$$

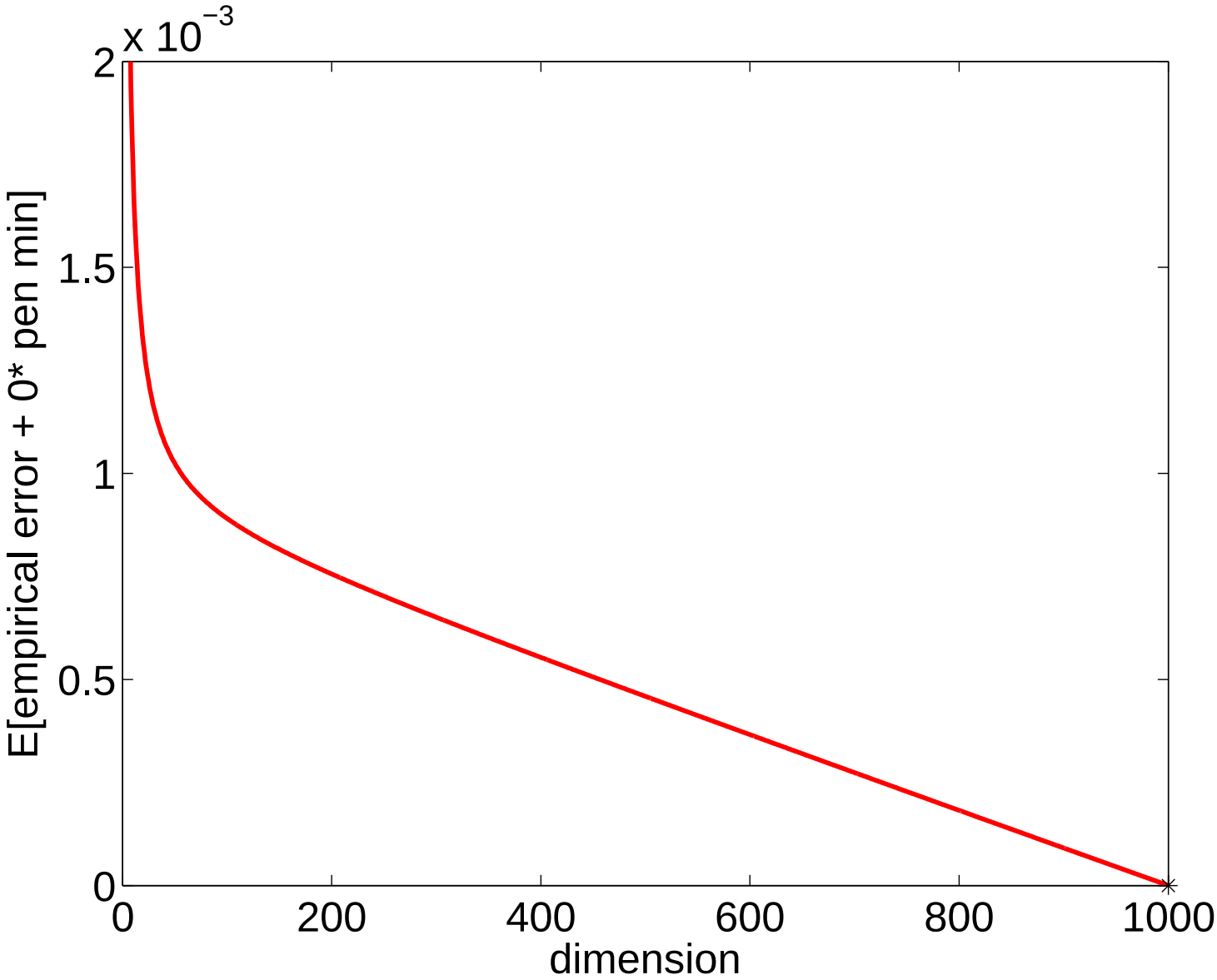
- Inégalité oracle démontrée pour

$$\frac{K\sigma^2 D_m}{n}$$

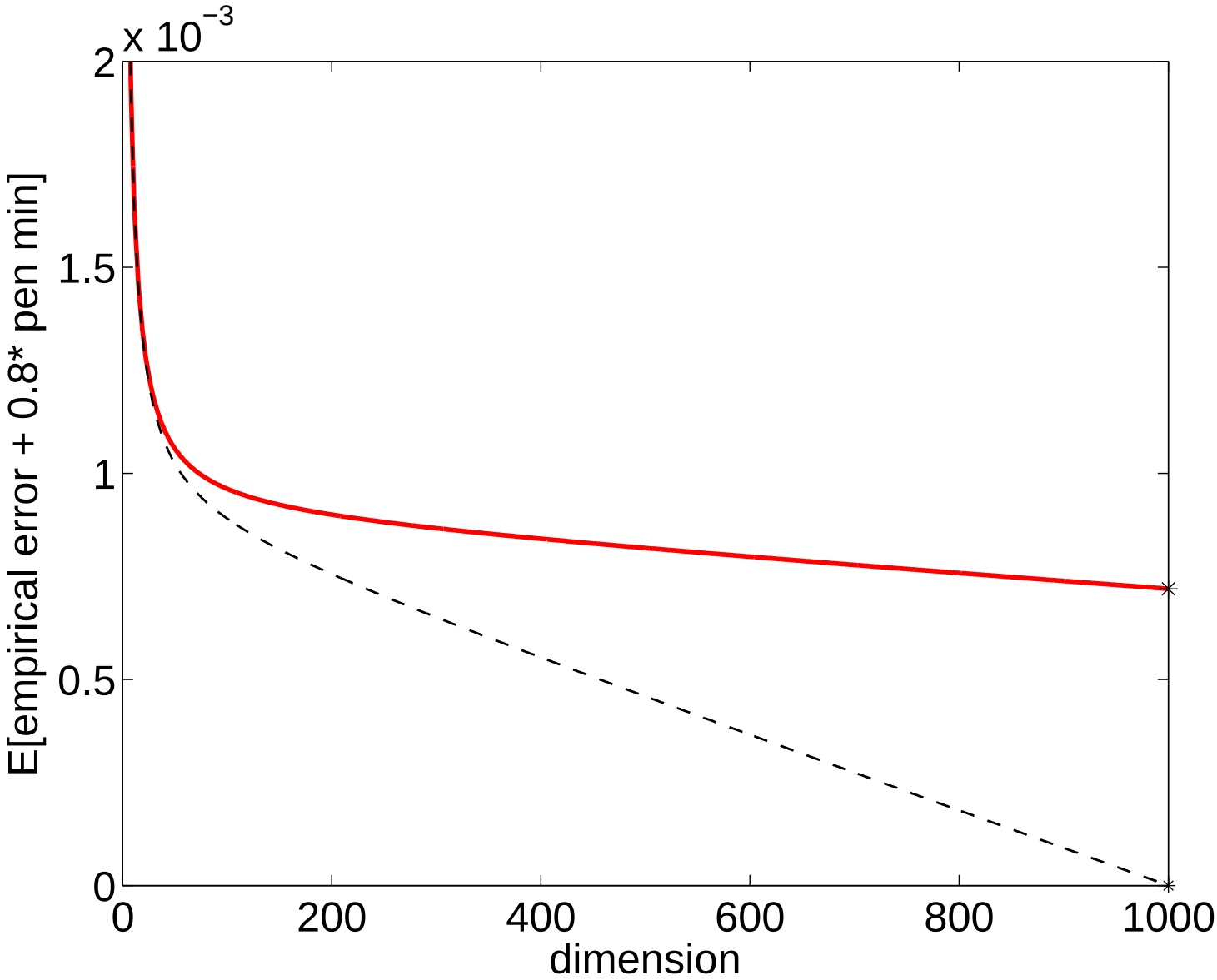
pour tout $K > 1$ (Birgé & Massart, 2001)

$\Rightarrow$ Que se passe-t-il pour $K \leq 1$?

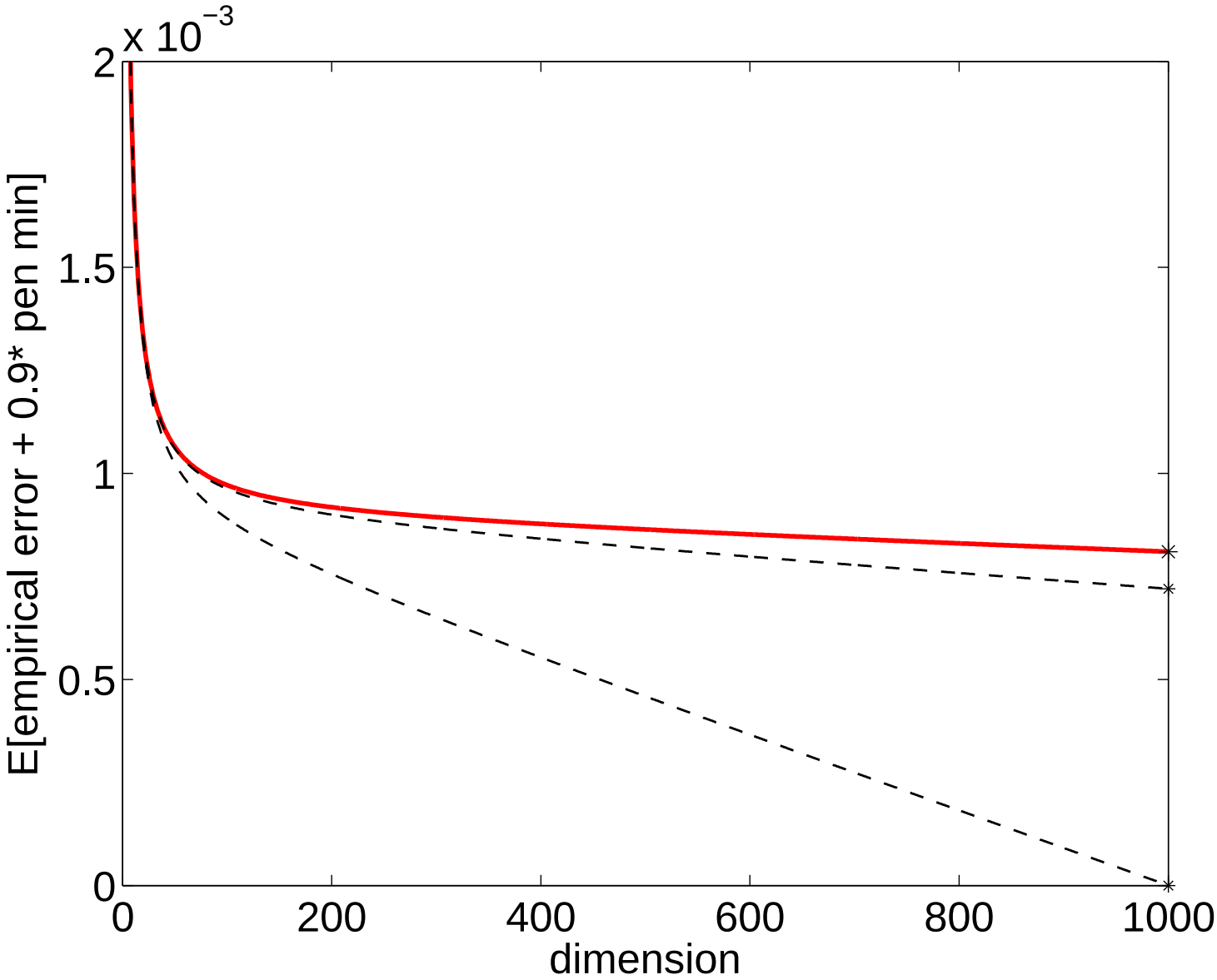
$\mathbb{E}[$ Risque empirique $] + 0 \times \sigma^2 D_m n^{-1}$ (moindres carrés)



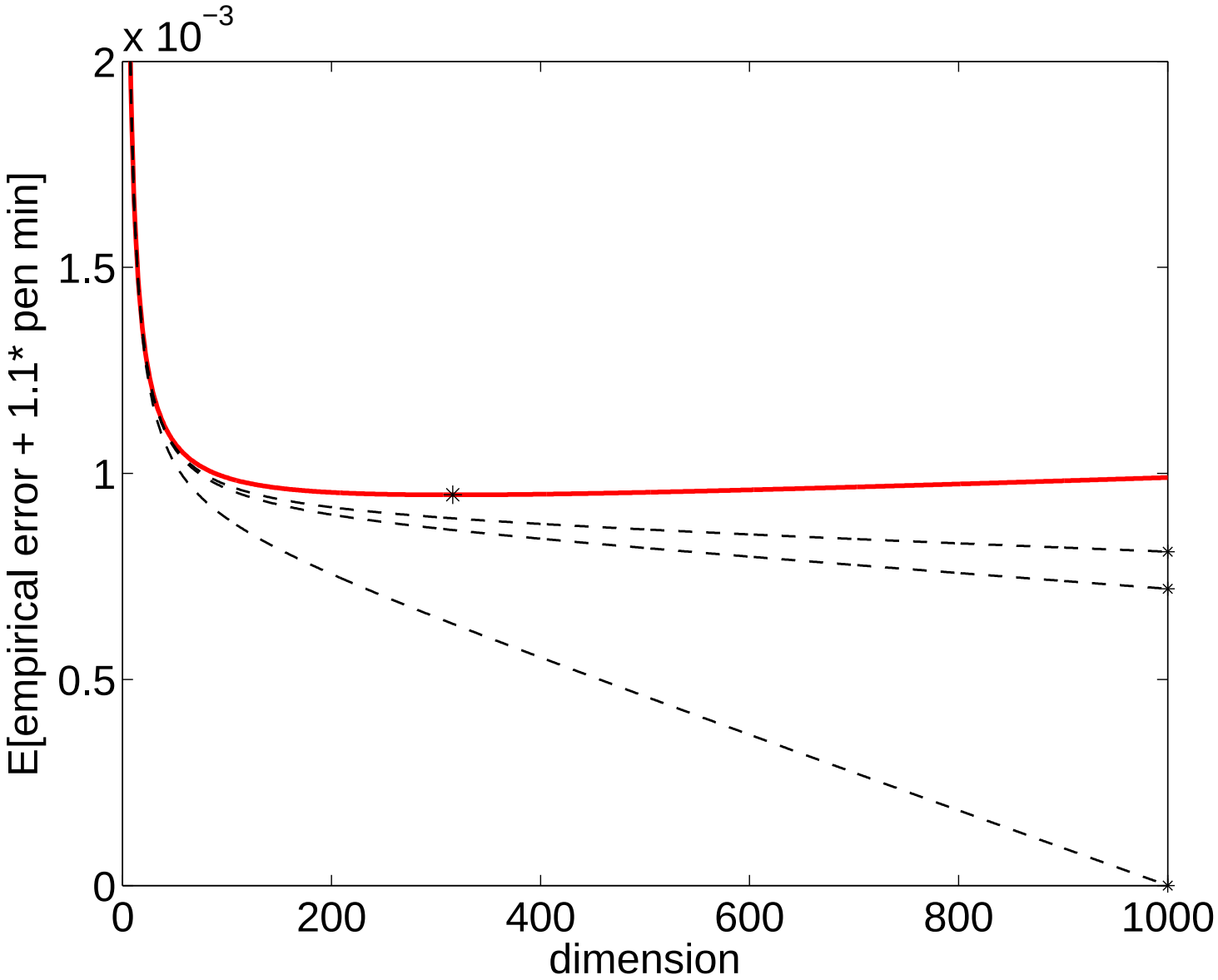
$\mathbb{E}[$ Risque empirique $] + 0.8 \times \sigma^2 D_m n^{-1}$ (moindres carrés)



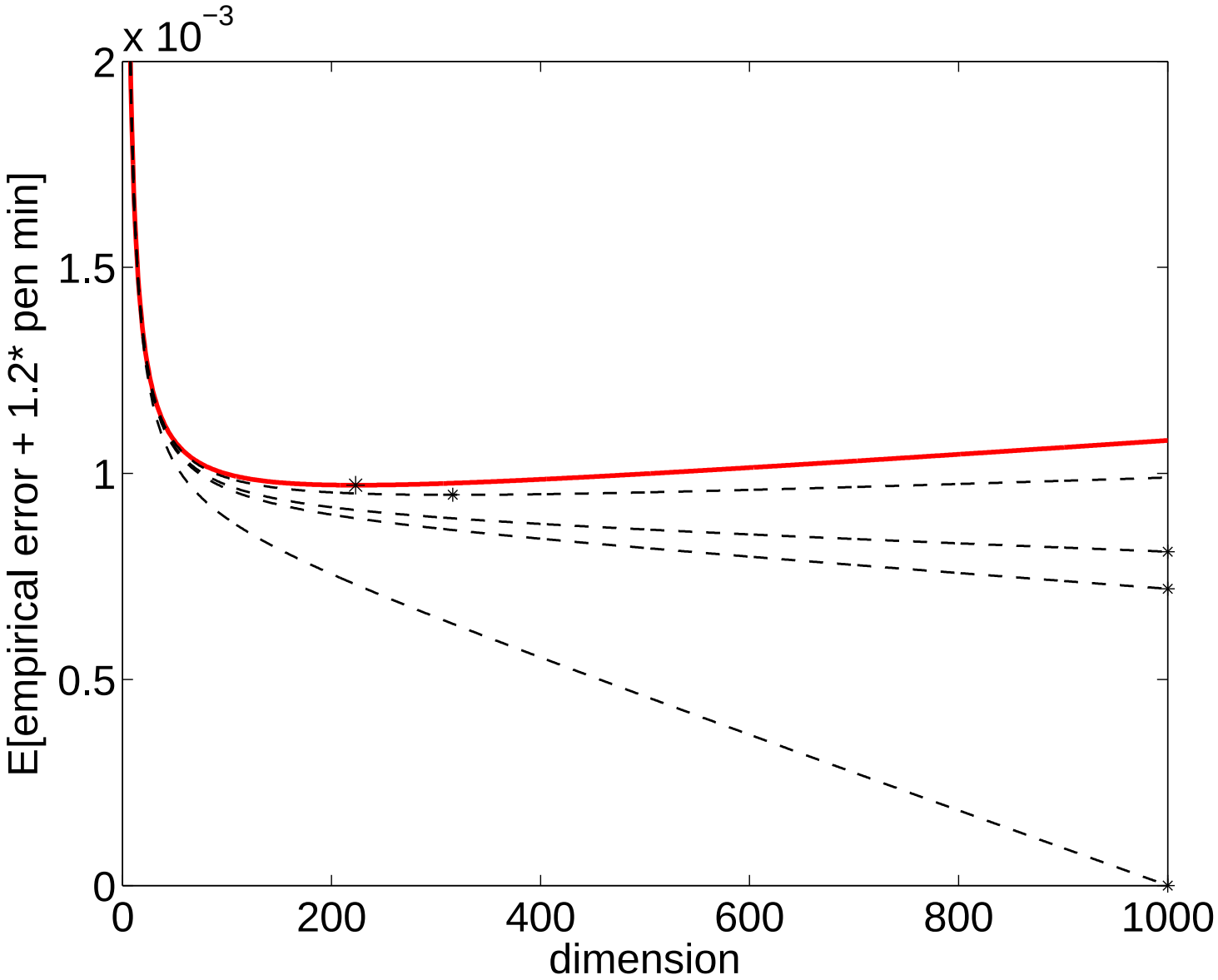
$\mathbb{E}[$ Risque empirique $] + 0.9 \times \sigma^2 D_m n^{-1}$ (moindres carrés)



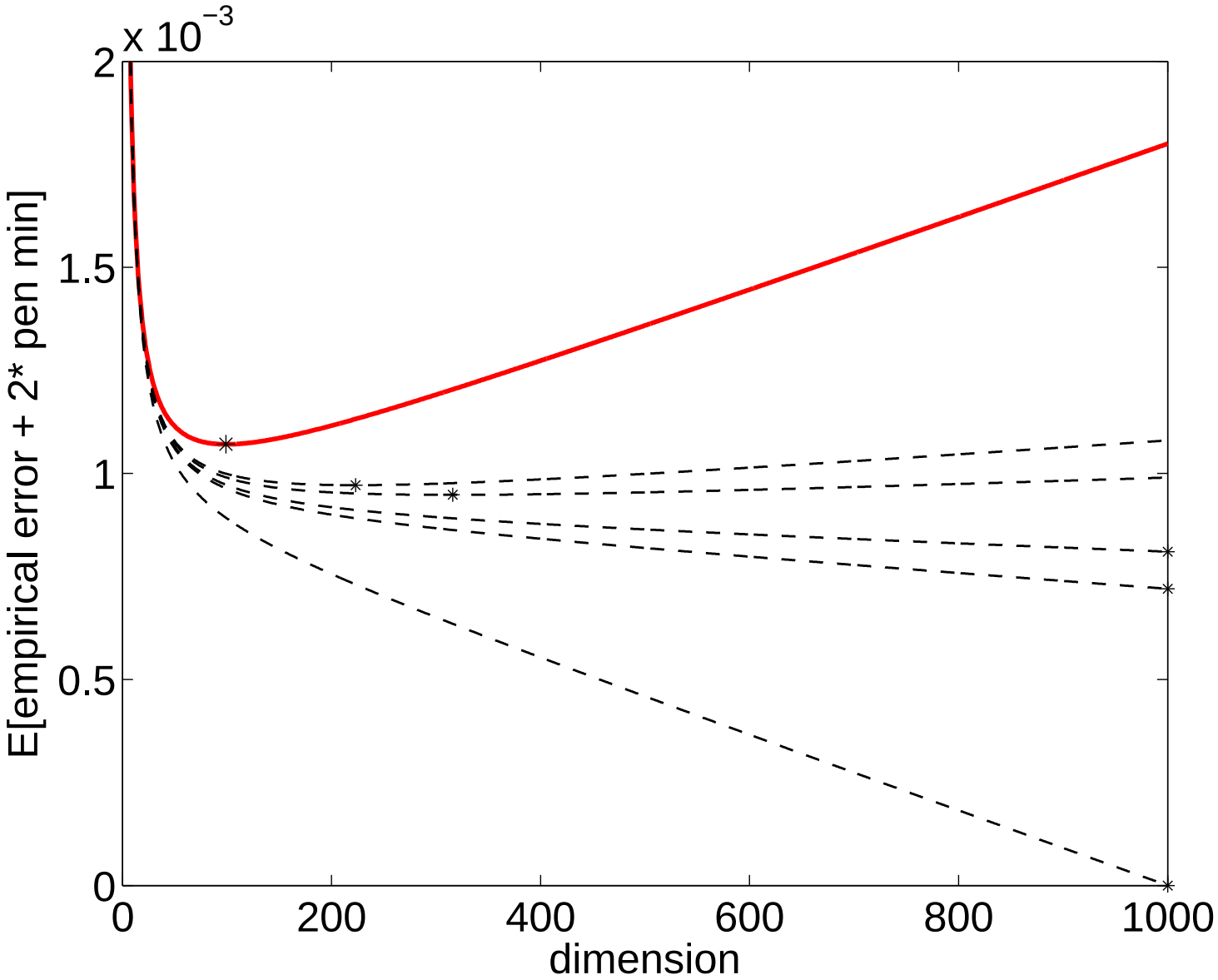
$\mathbb{E}[$ Risque empirique $] + 1.1 \times \sigma^2 D_m n^{-1}$ (moindres carrés)



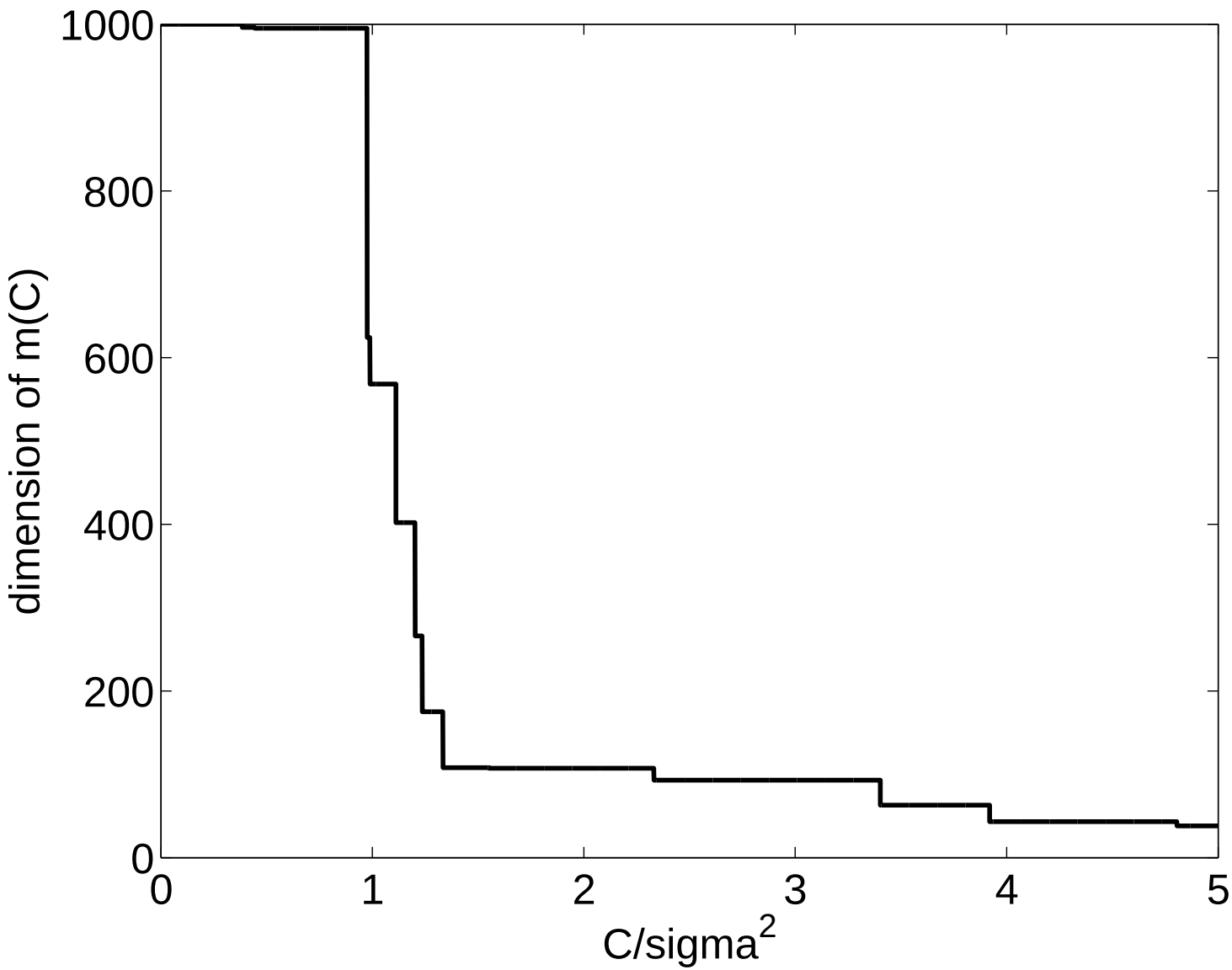
$\mathbb{E}[$ Risque empirique $] + 1.2 \times \sigma^2 D_m n^{-1}$ (moindres carrés)



$\mathbb{E}[$ Risque empirique $] + 2 \times \sigma^2 D_m n^{-1}$ (moindres carrés)



Moindres carrés : Saut de dimension



Moindres carrés : algorithme (Birgé & Massart 2007)

1. pour tout $C > 0$, calculer

$$\widehat{m}(C) \in \underset{m \in \mathcal{M}_n}{\operatorname{argmin}} \left\{ P_n \gamma(\widehat{s}_m) + C \frac{D_m}{n} \right\}$$

2. trouver $\widehat{C}_{\text{min}}$ tel que $D_{\widehat{m}(C)}$ est « très grande » lorsque $C < \widehat{C}_{\text{min}}$ et « raisonnablement petite » lorsque $C > \widehat{C}_{\text{min}}$

3. choisir $\widehat{m} = \widehat{m}\left(2\widehat{C}_{\text{min}}\right)$

Mise en œuvre pratique : package CAPUSHE (Baudry, Maugis & Michel, 2011)

`http://www.math.univ-toulouse.fr/~maugis/CAPUSHE.html`

Théorème (1) : Saut de dimension / Pénalité minimale

Théorème (Birgé & Massart 2007, reformulé)

Hypothèses :

- *plan d'expérience déterministe, bruit Gaussien, variance* σ^2 *constante,* $\exists m_0 \in \mathcal{M}$, $S_{m_0} = \mathbb{S}$,
- $\inf_{m \in \mathcal{M}} \left\{ \mathbb{E}[\ell(s^\star, \widehat{s}_m)] \right\} \leq \sigma^2 \delta_n$, $\delta_n \leq 1/20$.

Alors, $\forall \gamma > 0, n \geq n_0(\gamma)$, *avec prob. au moins* $1 - 4\,\mathrm{Card}(\mathcal{M}) n^{-\gamma}$,

$$\forall C < \left(1 - \eta_n^-\right)\sigma^2 , \quad D_{\widehat{m}(C)} \geq \frac{9n}{10}$$

$$\forall C > \left(1 + \eta_n^+\right)\sigma^2 , \quad D_{\widehat{m}(C)} \leq \frac{n}{10}$$

avec $\eta_n^- = 81\sqrt{\frac{\gamma \log(n)}{n}}$, $\eta_n^+ = \eta_n^- + 20\delta_n$.

A. Minimal penalties and the slope heuristics : a survey, 2014.

Théorème (2) : Inégalité-oracle

Théorème (Birgé & Massart 2007, reformulé)

Sous les hypothèses précédentes, avec probabilité au moins $1 - 4\,\mathrm{Card}(\mathcal{M})n^{-\gamma}$, *pour tout*

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ P_n \gamma(\widehat{s}_m) + 2\widehat{C}_{\min} \frac{D_m}{n} \right\} ,$$

si $n \geq n_0(\gamma)$, *pour tout* $\eta \geq 2\max\{\eta_n^-, \eta_n^+\}$,

$$\ell\left(s^\star, \widehat{s}_{\widehat{m}}\right) \leq (1+3\eta) \inf_{m \in \mathcal{M}} \left\{ \ell(s^\star, \widehat{s}_m) \right\} + \frac{880\sigma^2\gamma\log(n)}{\eta n}$$

A. Minimal penalties and the slope heuristics : a survey, 2014.

Estimateurs linéaires

Définition :

$$\left(\widehat{s}_m(X_i)\right)_{1\leq i\leq n} = A_m \left(Y_i\right)_{1\leq i\leq n}$$

Estimateurs linéaires

Définition :

$$\left(\widehat{s}_m(X_i)\right)_{1\leq i\leq n} = A_m \left(Y_i\right)_{1\leq i\leq n}$$

Exemples :

- moindres carrés

Estimateurs linéaires

Définition :

$$\left(\widehat{s}_m(X_i)\right)_{1\leq i\leq n} = A_m\left(Y_i\right)_{1\leq i\leq n}$$

Exemples :

- moindres carrés
- k plus proches voisins
- Nadaraya-Watson
- splines (Wahba, 1990), régression ridge à noyaux :

$$\widehat{s}_\lambda \in \operatorname*{argmin}_{t\in\mathcal{H}}\left\{ P_n\gamma(t) + \lambda\|t\|_{\mathcal{H}}^2 \right\} \ .$$

Généralisation aux estimateurs linéaires ?

Moindres carrés

$$\mathrm{pen}_{\mathrm{Cp}}(m) = \frac{2\sigma^2 D_m}{n}$$

$$\underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ P_n \gamma(\widehat{s}_m) + C \frac{D_m}{n} \right\}$$

$\Rightarrow D_{\widehat{m}(C)} \ll$ saute $\gg$ en $\widehat{C}_{\min} \approx \sigma^2$

$\Rightarrow$ choix optimal avec $\widehat{m}\left(2\widehat{C}_{\min}\right)$

Généralisation aux estimateurs linéaires ?

Moindres carrés

$$\mathrm{pen}_{\mathrm{Cp}}(m) = \frac{2\sigma^2 D_m}{n}$$

$$\underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ P_n \gamma(\widehat{s}_m) + C \frac{D_m}{n} \right\}$$

$\Rightarrow D_{\widehat{m}(C)} \ll$ saute $\gg$ en $\widehat{C}_{\min} \approx \sigma^2$

$\Rightarrow$ choix optimal avec $\widehat{m}\left(2\widehat{C}_{\min}\right)$

Estimateurs linéaires

$$\mathrm{pen}_{\mathrm{CL}}(m) = \frac{2\sigma^2 \operatorname{tr}(A_m)}{n}$$

Généralisation aux estimateurs linéaires ?

Moindres carrés

$$\mathrm{pen}_{\mathrm{Cp}}(m) = \frac{2\sigma^2 D_m}{n}$$

$$\underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ P_n \gamma(\widehat{s}_m) + C \frac{D_m}{n} \right\}$$

$\Rightarrow D_{\widehat{m}(C)}$ « saute » en $\widehat{C}_{\min} \approx \sigma^2$

$\Rightarrow$ choix optimal avec $\widehat{m}\left(2\widehat{C}_{\min}\right)$

Estimateurs linéaires

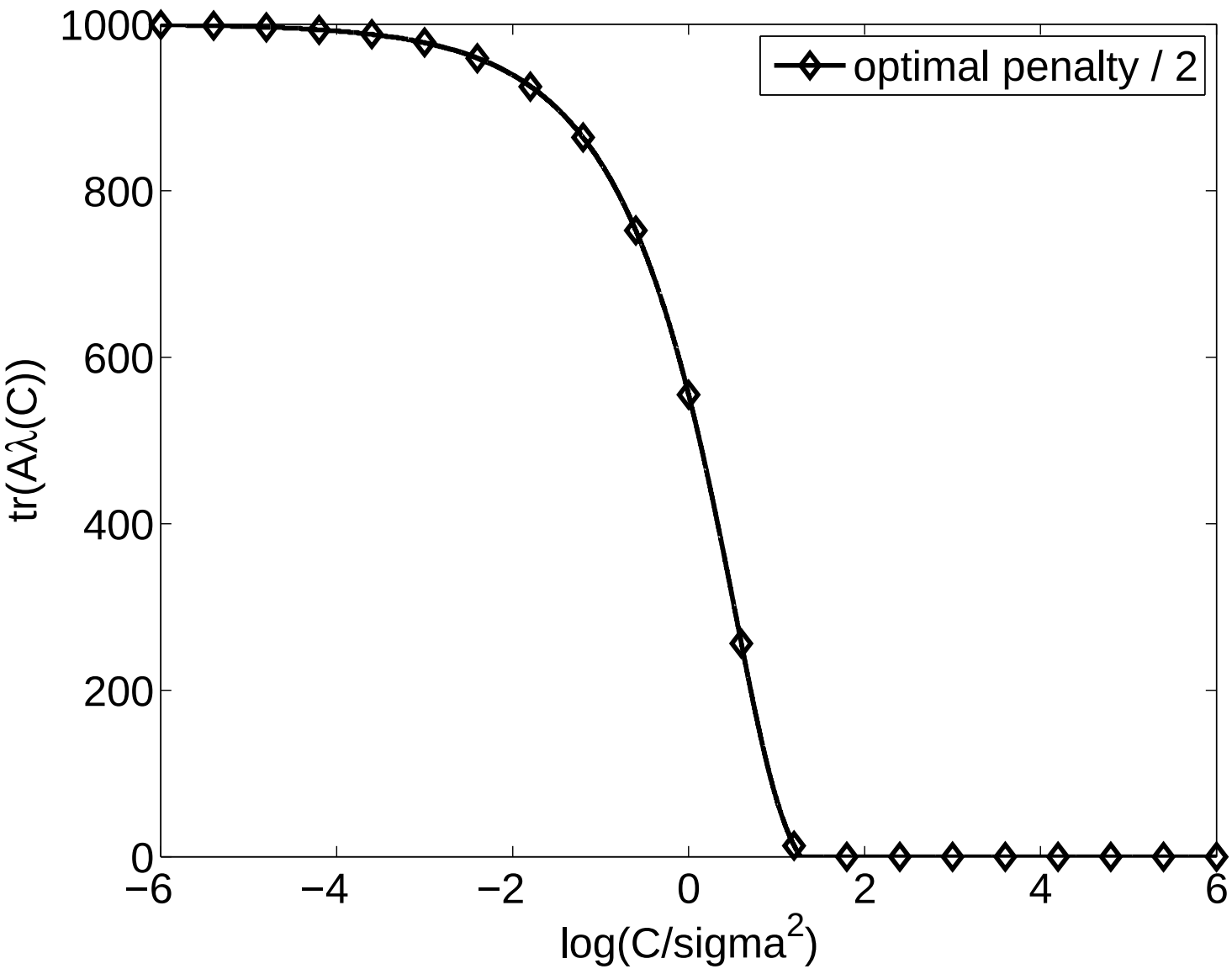
$$\mathrm{pen}_{\mathrm{CL}}(m) = \frac{2\sigma^2 \operatorname{tr}(A_m)}{n}$$

$$\underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ P_n \gamma(\widehat{s}_m) + C \frac{\operatorname{tr}(A_m)}{n} \right\}$$

$\operatorname{tr}\left(A_{\widehat{m}(C)}\right)$ saute-t-il en $\widehat{C}_{\min} \approx \sigma^2$?

choix optimal avec $\widehat{m}\left(2\widehat{C}_{\min}\right)$?

Pas de saut de dimension avec une pénalité $\propto \mathrm{tr}(A_m)$



Pénalité minimale pour les estimateurs linéaires

$$\mathbb{E}\left[\ell(s^\star, \widehat{s}_m)\right] = \text{Err. approx.} + \underbrace{\frac{\mathrm{tr}(A_m^\top A_m)\sigma^2}{n}}_{\text{Err. estim.}}$$

Pénalité minimale pour les estimateurs linéaires

$$\mathbb{E}\left[\ell(s^\star, \widehat{s}_m)\right] = \text{Err. approx.} + \underbrace{\frac{\operatorname{tr}(A_m^\top A_m)\sigma^2}{n}}_{\text{Err. estim.}}$$

$$\mathbb{E}\left[P_n \gamma(\widehat{s}_m)\right] = \sigma^2 + \text{Err. approx.} - \frac{\left(2\operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m)\right)\sigma^2}{n}$$

Pénalité minimale pour les estimateurs linéaires

$$\mathbb{E}\left[\ell(s^\star, \widehat{s}_m)\right] = \text{Err. approx.} + \underbrace{\frac{\operatorname{tr}(A_m^\top A_m)\sigma^2}{n}}_{\text{Err. estim.}}$$

$$\mathbb{E}\left[P_n \gamma(\widehat{s}_m)\right] = \sigma^2 + \text{Err. approx.} - \frac{\left(2\operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m)\right)\sigma^2}{n}$$

$$\Rightarrow \quad \text{pénalité optimale} \quad \frac{\left(2\operatorname{tr}(A_m)\right)\sigma^2}{n}$$

Pénalité minimale pour les estimateurs linéaires

$$\mathbb{E}\left[\ell(s^\star, \widehat{s}_m)\right] = \text{Err. approx.} + \underbrace{\frac{\operatorname{tr}(A_m^\top A_m)\sigma^2}{n}}_{\text{Err. estim.}}$$

$$\mathbb{E}\left[P_n\gamma(\widehat{s}_m)\right] = \sigma^2 + \text{Err. approx.} - \underbrace{\frac{\left(2\operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m)\right)\sigma^2}{n}}_{\text{pénalité minimale}}$$

$$\Rightarrow \quad \text{pénalité optimale} \quad \frac{\left(2\operatorname{tr}(A_m)\right)\sigma^2}{n}$$

Pénalité minimale pour les estimateurs linéaires

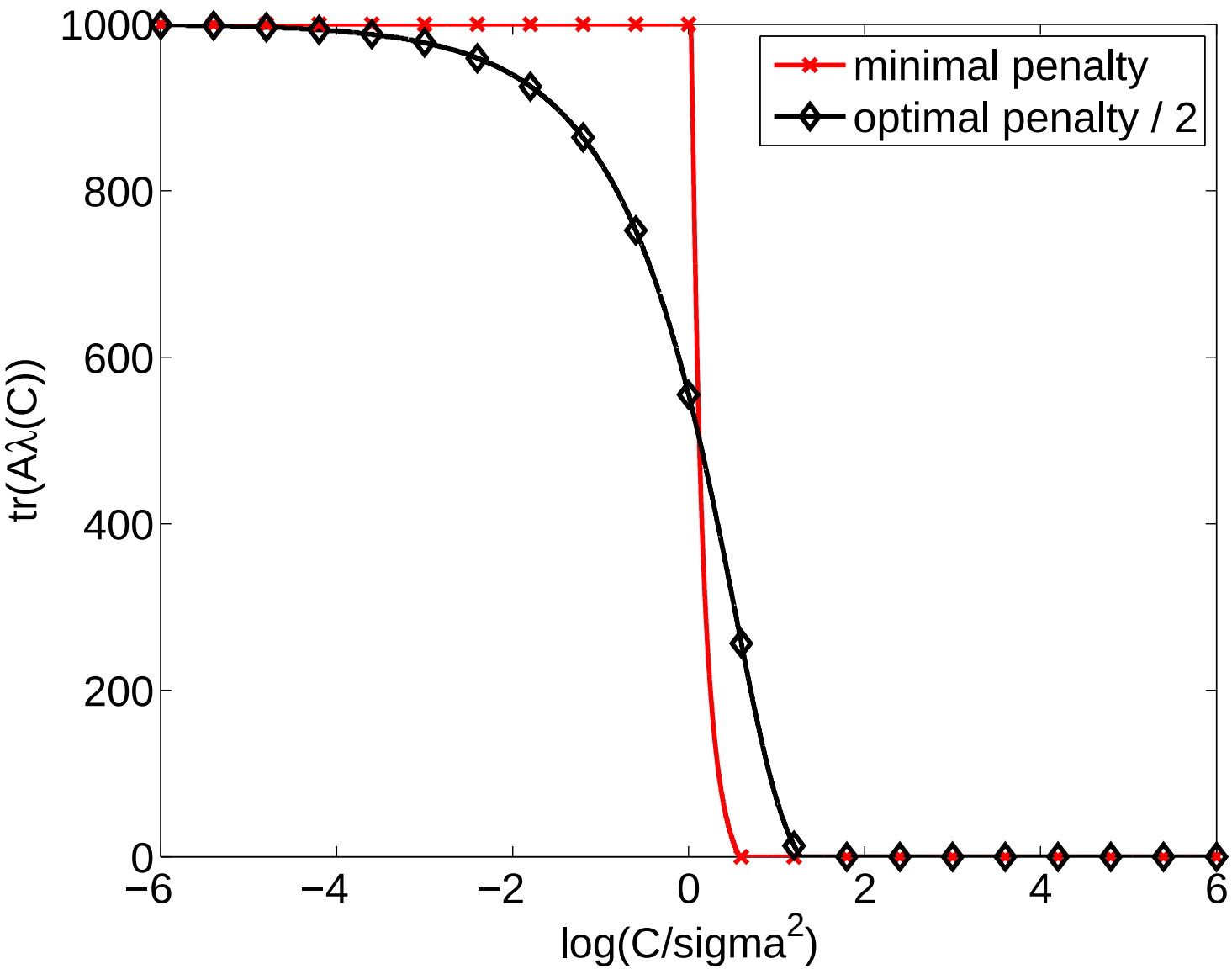
$$\mathbb{E}\left[\ell(s^\star, \widehat{s}_m)\right] = \text{Err. approx.} + \underbrace{\frac{\operatorname{tr}(A_m^\top A_m)\sigma^2}{n}}_{\text{Err. estim.}}$$

$$\mathbb{E}\left[P_n\gamma(\widehat{s}_m)\right] = \sigma^2 + \text{Err. approx.} - \underbrace{\frac{\left(2\operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m)\right)\sigma^2}{n}}_{\text{pénalité minimale}}$$

$$\Rightarrow \quad \text{pénalité optimale} \quad \frac{\left(2\operatorname{tr}(A_m)\right)\sigma^2}{n}$$

$$\widehat{m}(C) \in \operatorname*{argmin}_{m\in\mathcal{M}} \left\{ P_n\gamma(\widehat{s}_m) + C \times \frac{2\operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m)}{n} \right\}$$

Saut de « dimension » (régression ridge)



Algorithme de calibration de pénalités

1. pour tout $C > 0$, calculer

$$\widehat{m}_{\min}(C) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ P_n \gamma(\widehat{s}_m) + \frac{C \left(2 \operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m) \right)}{n} \right\}$$

2. trouver $\widehat{C}_{\min}$ tel que $\operatorname{tr}(A_{\widehat{m}_{\min}(C)})$ est « très grande » lorsque $C < \widehat{C}_{\min}$ et « raisonnablement petite » lorsque $C > \widehat{C}_{\min}$,

3. choisir

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ P_n \gamma(\widehat{s}_m) + \frac{2 \widehat{C}_{\min} \operatorname{tr}(A_m)}{n} \right\}$$

$\Rightarrow$ Inégalité-oracle optimale (au premier ordre) :

A. & Bach. Data-driven calibration of linear estimators with minimal penalties. In *Advances in Neural Information Processing Systems 22*, pages 46–54, 2009.

Pénalités minimales : bilan

- Moindres carrés et k-plus proches voisins :

$$\text{pen}_{\text{opt}} = 2\,\text{pen}_{\text{min}}$$

mais pas pour les autres estimateurs linéaires (kernel ridge, Nadaraya-Watson, etc.) !

Pénalités minimales : bilan

- Moindres carrés et k-plus proches voisins :

$$\text{pen}_{\text{opt}} = 2\,\text{pen}_{\text{min}}$$

  mais pas pour les autres estimateurs linéaires (kernel ridge, Nadaraya-Watson, etc.) !

- Application à l'apprentissage multi-tâches, *via* l'estimation d'une matrice de covariance en régression multivariée

  Solnon, A. & Bach. Multi-task regression using minimal penalties. *Journal of Machine Learning Research*, 13:2773–2812, 2012.

Pénalités minimales : bilan

- Moindres carrés et k-plus proches voisins :

$$\text{pen}_{\text{opt}} = 2\,\text{pen}_{\text{min}}$$

mais pas pour les autres estimateurs linéaires (kernel ridge, Nadaraya-Watson, etc.) !

- Application à l'apprentissage multi-tâches, *via* l'estimation d'une matrice de covariance en régression multivariée

Solnon, A. & Bach. Multi-task regression using minimal penalties. *Journal of Machine Learning Research*, 13:2773–2812, 2012.

- Régression hétéroscédastique, régressogrammes :
$\text{pen}_{\text{opt}} \approx 2\,\text{pen}_{\text{min}}$

A. & Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245–279, 2009.

Pénalités minimales : bilan

- Moindres carrés et k-plus proches voisins :

$$\text{pen}_{\text{opt}} = 2\,\text{pen}_{\text{min}}$$

mais pas pour les autres estimateurs linéaires (kernel ridge, Nadaraya-Watson, etc.) !

- Application à l'apprentissage multi-tâches, *via* l'estimation d'une matrice de covariance en régression multivariée

Solnon, A. & Bach. Multi-task regression using minimal penalties. *Journal of Machine Learning Research*, 13:2773–2812, 2012.

- Régression hétéroscédastique, régressogrammes :
$\text{pen}_{\text{opt}} \approx 2\,\text{pen}_{\text{min}}$

A. & Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10:245–279, 2009.

- Article de survol :

A. Minimal penalties and the slope heuristics : a survey, 2014.

Plan

1 Sélection d'estimateurs

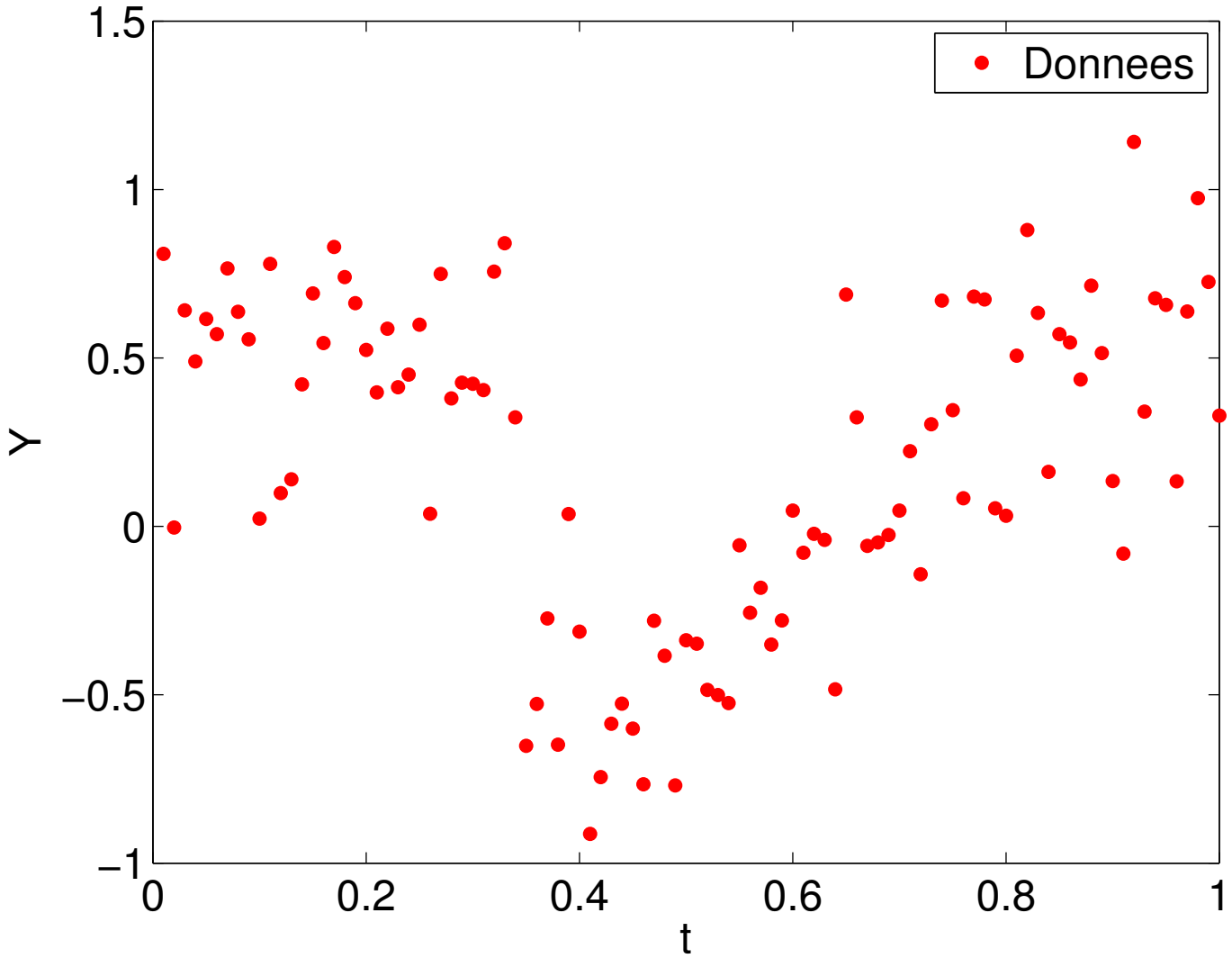
2 Validation croisée

3 Pénalités minimales

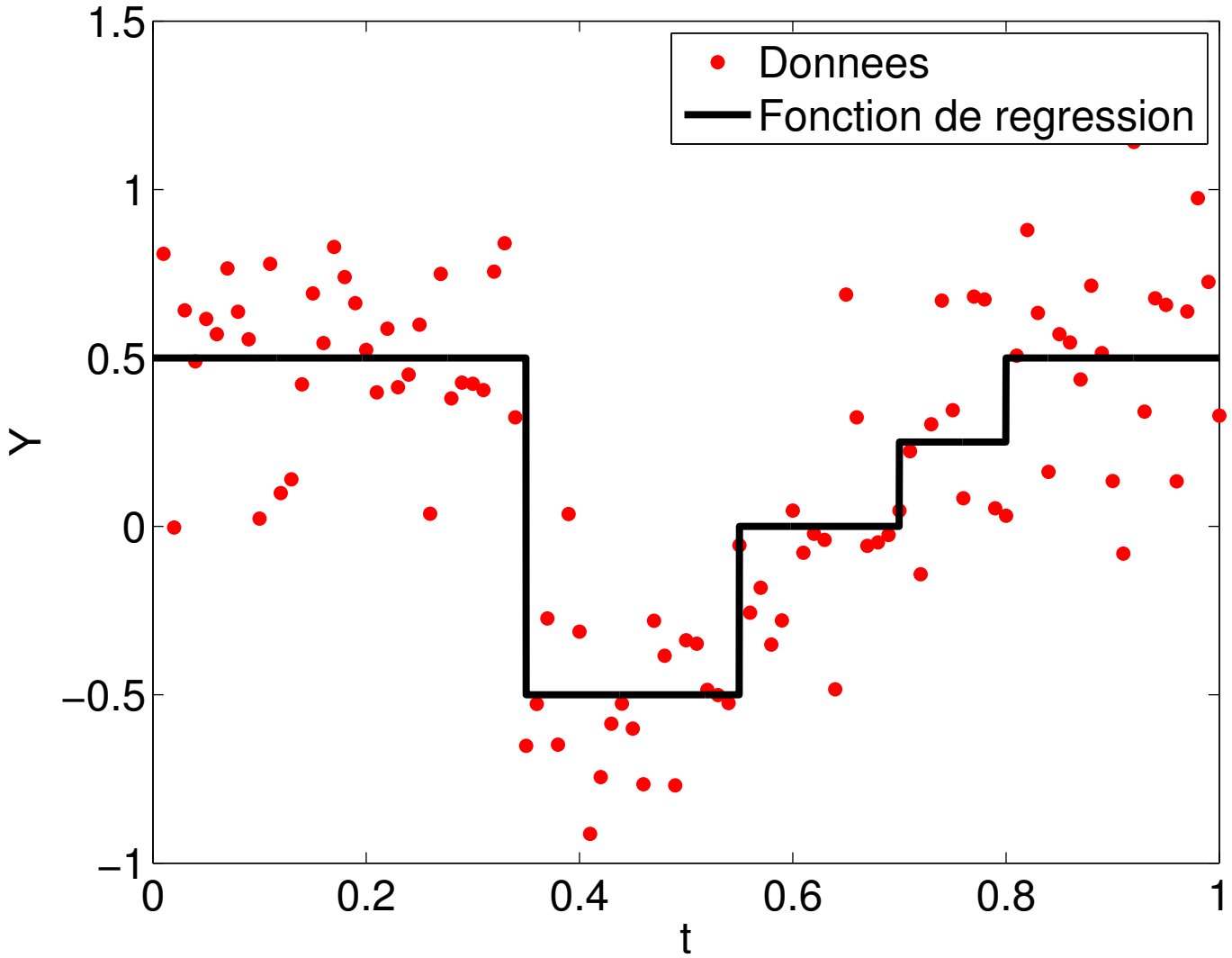
4 Détection de ruptures

5 Conclusion

Détection de ruptures : données



Détection de ruptures : objectif



Détection de ruptures et sélection de modèles

$$Y_i = s^\star(t_i) + \sigma(t_i)\,\varepsilon_i \quad \text{avec} \quad \mathbb{E}\left[\varepsilon_i\right] = 0 \quad \mathbb{E}\left[\varepsilon_i^2\right] = 1$$

- But : détecter les ruptures dans la moyenne $s^\star$ du signal Y

$\Rightarrow$ Sélection de modèles, collection des régressogrammes avec $\mathcal{M}_n = \mathfrak{P}_{\text{interv}}(\{\, t_1, \ldots, t_n \,\})$ (ensemble des partitions en intervalles)

Approche classique (Lebarbier, 2005)

- Pénalité « Birgé-Massart » (suppose $\sigma(t_i) \equiv \sigma$) :

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}_n} \left\{ P_n \gamma(\widehat{s}_m) + \frac{C\sigma^2 D_m}{n} \left(5 + 2 \log\left(\frac{n}{D_m} \right) \right) \right\}$$

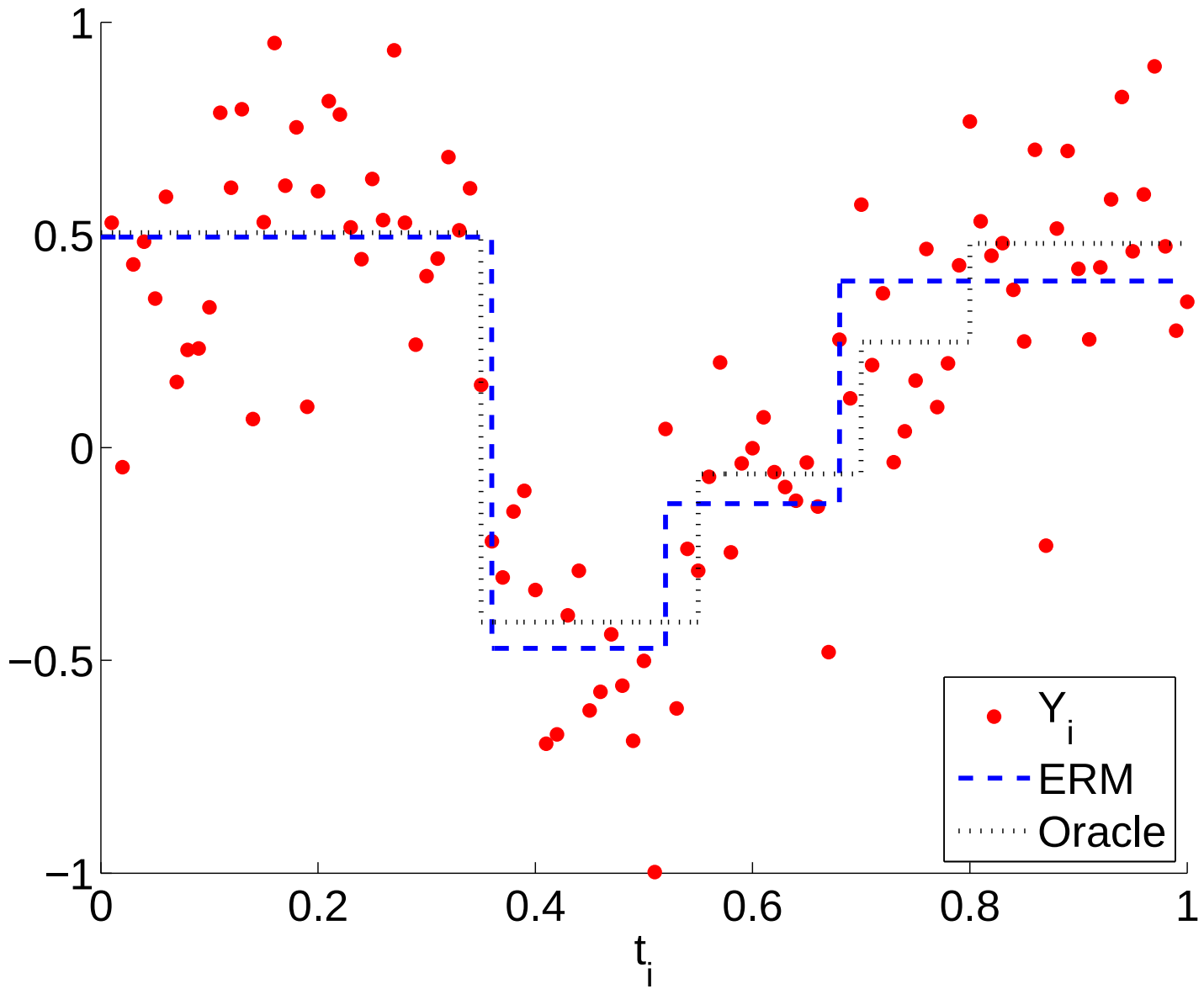
- Revient à aggréger les modèles de même dimension :

$$\widetilde{S}_D := \bigcup_{m \in \mathcal{M}_n \,,\, D_m = D} S_m$$

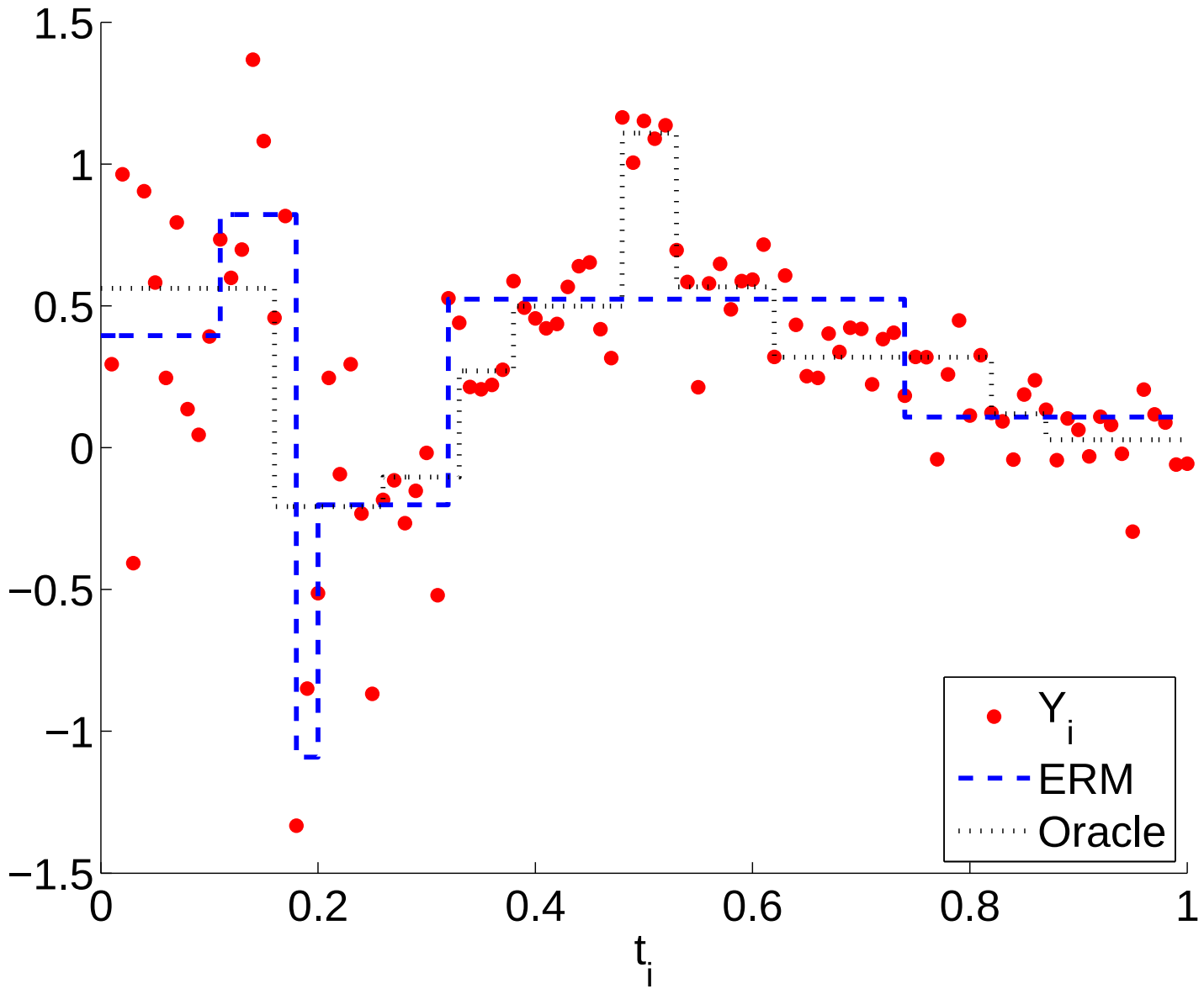
$$\widehat{s}_D \in \operatorname*{argmin}_{t \in \widetilde{S}_D} \left\{ P_n \gamma\left(t\right) \right\} \qquad \text{programmation dynamique}$$

$$\widehat{D} \in \operatorname*{argmin}_{1 \leq D \leq n} \left\{ P_n \gamma\left(\widehat{s}_D\right) + \frac{C\sigma^2 D}{n} \left(5 + 2 \log\left(\frac{n}{D} \right) \right) \right\}$$

$D = 4$, homoscédastique ; $n = 100$, $\sigma = 0{,}25$



$D = 6$, hétéroscédastique ; $n = 100$, $\|\sigma\| = 0{,}30$



Détection de ruptures par validation croisée

- Échec de la minimisation du risque empirique dans le cas hétéroscédastique suggéré par

  A. Choosing a penalty for model selection in heteroscedastic regression, 2010. arXiv:0812.3141v2.

Détection de ruptures par validation croisée

- Échec de la minimisation du risque empirique dans le cas hétéroscédastique suggéré par

  A. Choosing a penalty for model selection in heteroscedastic regression, 2010. arXiv:0812.3141v2.

- La validation croisée s'adapte à l'hétéroscédasticité :

  A. V-fold cross-validation improved : V-fold penalization, 2008. arXiv:0802.0566v2.

Détection de ruptures par validation croisée

- Échec de la minimisation du risque empirique dans le cas hétéroscédastique suggéré par

  A. Choosing a penalty for model selection in heteroscedastic regression, 2010. arXiv:0812.3141v2.

- La validation croisée s'adapte à l'hétéroscédasticité :

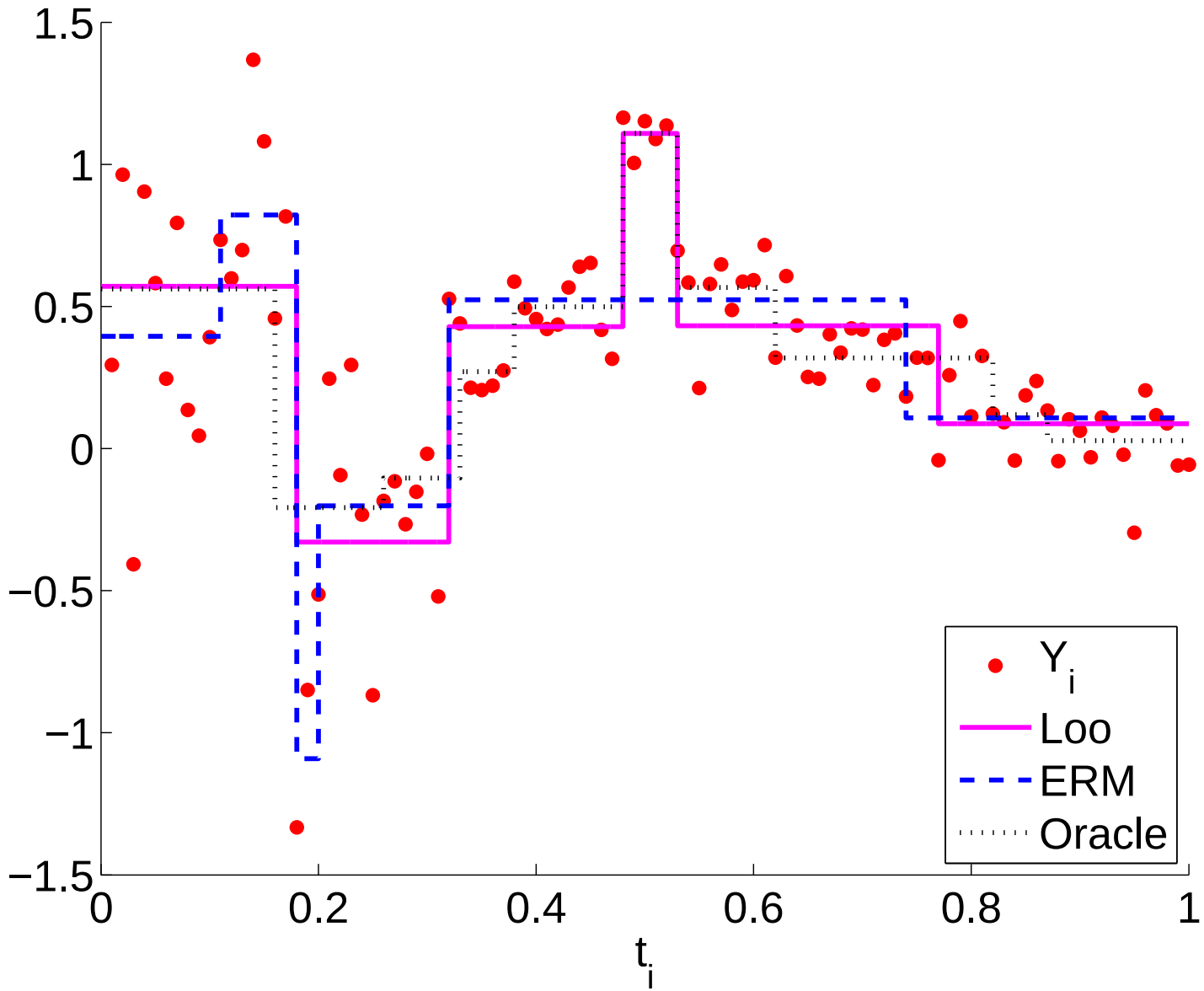
  A. V-fold cross-validation improved : V-fold penalization, 2008. arXiv:0802.0566v2.

$\Rightarrow$ Algorithme :

$$\left(\widehat{s}_m\right)_{D_m=D} \quad \rightsquigarrow \quad \widehat{s}_D^{\mathrm{CV}} = \widehat{s}_{\widehat{m}^{\mathrm{CV}}(D)} \quad \rightsquigarrow \quad \widetilde{s} = \widehat{s}_{\widehat{D}^{\mathrm{CV}}}^{\mathrm{CV}}$$

A. & Celisse. Segmentation of the mean of heteroscedastic data via cross-validation. *Statistics and Computing*, pages 1–20, 2010.

$D = 6$, hétéroscédastique ; $n = 100$, $\|\sigma\| = 0{,}30$



Données de grande dimension ou complexes

$Y_1, \ldots, Y_n \in \mathcal{Y}$ de loi constante par morceaux

- $\mathcal{Y}$ de grande dimension (structure euclidienne inadaptée) ou qui n'est pas un espace vectoriel

Données de grande dimension ou complexes

$$Y_1, \ldots, Y_n \in \mathcal{Y} \qquad \text{de loi constante par morceaux}$$

- $\mathcal{Y}$ de grande dimension (structure euclidienne inadaptée) ou qui n'est pas un espace vectoriel

$\Rightarrow$ Idée : considérer un noyau défini positif $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ et chercher les ruptures dans la série

$$\Phi(Y_1), \ldots, \Phi(Y_n) \in \mathcal{H} \qquad \text{où} \quad \Phi(y) = k(y, \cdot)$$

et $\mathcal{H}$ est le RKHS associé à k.

Données de grande dimension ou complexes

$Y_1, \ldots, Y_n \in \mathcal{Y}$ de loi constante par morceaux

- $\mathcal{Y}$ de grande dimension (structure euclidienne inadaptée) ou qui n'est pas un espace vectoriel

$\Rightarrow$ Idée : considérer un noyau défini positif $k : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ et chercher les ruptures dans la série

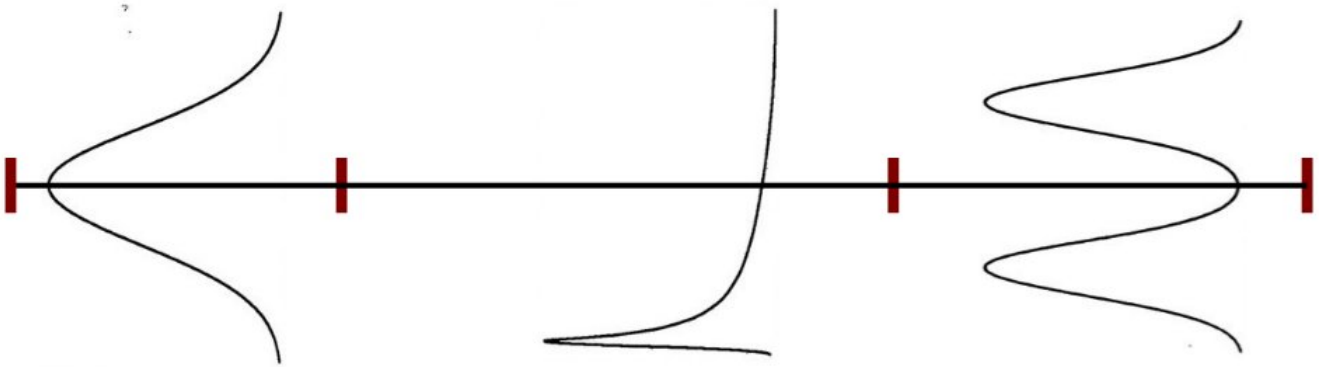
$$\Phi(Y_1), \ldots, \Phi(Y_n) \in \mathcal{H} \qquad \text{où} \quad \Phi(y) = k(y, \cdot)$$

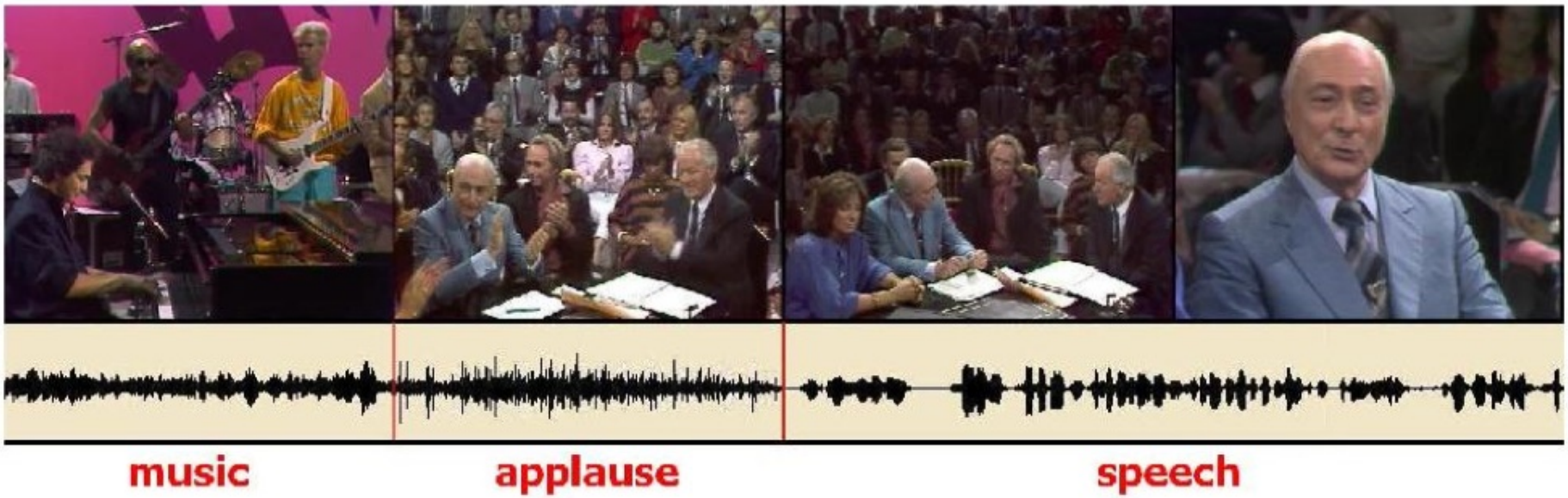
et $\mathcal{H}$ est le RKHS associé à k.

- Pénalité similaire à celle de Lebarbier (2005)
  $\Rightarrow$ inégalité-oracle.

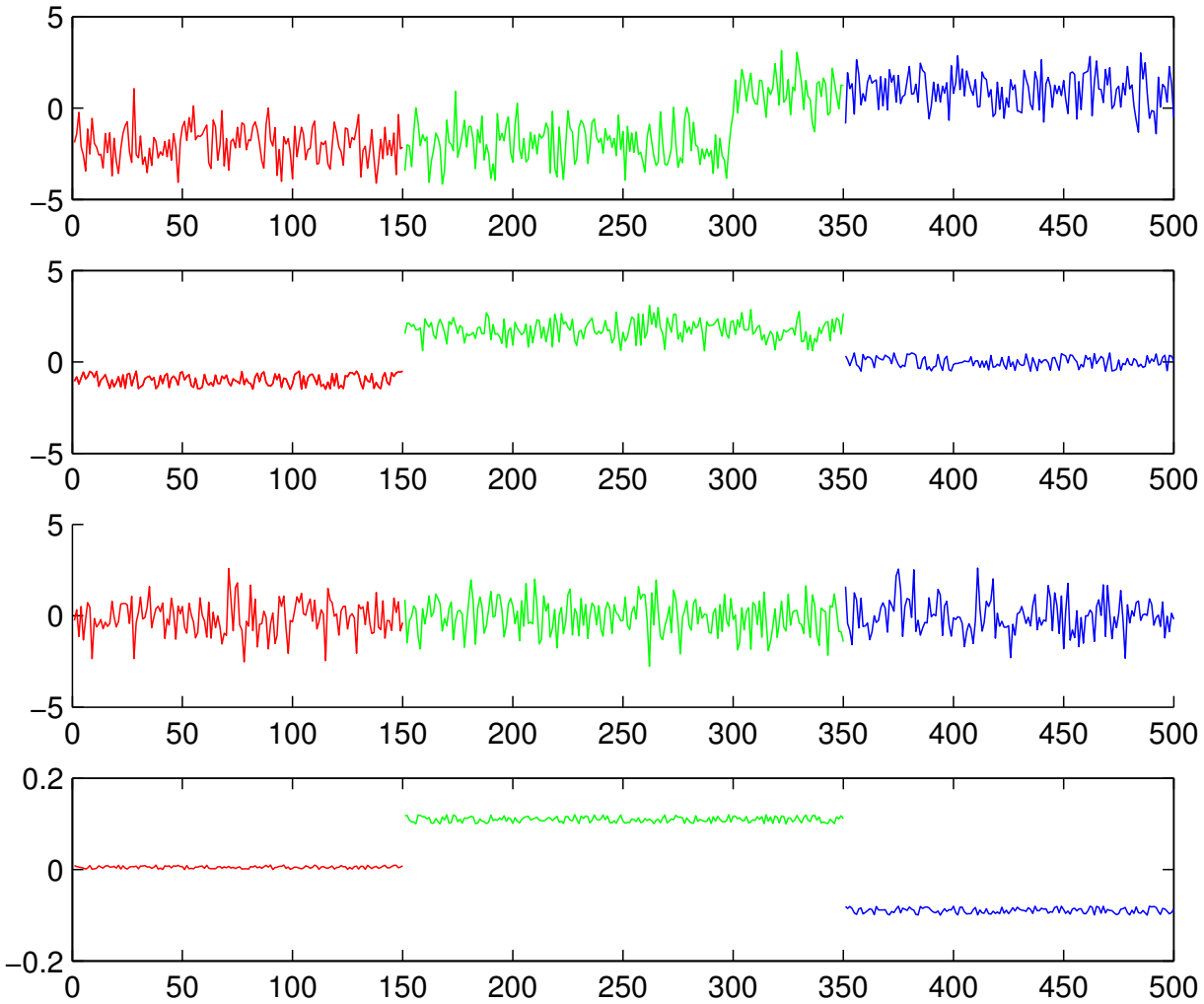
A., Celisse & Harchaoui. Kernel change-point detection, 2012. arXiv:1202.3878v1.

Détection de ruptures à noyaux : applications





Apprentissage de métrique : objectif



Apprentissage de métrique

$$Y = (Y_1, \ldots, Y_n) \in \left(\mathbb{R}^d\right)^n$$

$$\forall B \succeq 0, \qquad \widehat{m}_B(Y) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ P_n \gamma_B(\widehat{s}_m) + \lambda D_m \right\} \quad \Rightarrow \quad \widehat{B} \ ?$$

Apprentissage de métrique

$$Y = (Y_1, \ldots, Y_n) \in \left(\mathbb{R}^d\right)^n$$

$$\forall B \succeq 0, \qquad \widehat{m}_B(Y) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ P_n \gamma_B(\widehat{s}_m) + \lambda D_m \right\} \quad \Rightarrow \quad \widehat{B} \ ?$$

- Hypothèse : on dispose de séries segmentées

$$\left(Y^1, m^1\right), \ldots, \left(Y^N, m^N\right) \ .$$

Apprentissage de métrique

$$Y = (Y_1, \ldots, Y_n) \in \left(\mathbb{R}^d\right)^n$$

$$\forall B \succeq 0, \qquad \widehat{m}_B(Y) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ P_n \gamma_B(\widehat{s}_m) + \lambda D_m \right\} \quad \Rightarrow \quad \widehat{B} \ ?$$

- Hypothèse : on dispose de séries segmentées

$$(Y^1, m^1), \ldots, (Y^N, m^N) \ .$$

$\Rightarrow$ prédiction structurée : idéalement, on voudrait

$$\widehat{B} \in \operatorname*{argmin}_{B \succeq 0} \left\{ \frac{1}{N} \sum_{j=1}^{N} \gamma\left(m^j, \widehat{m}_B\left(Y^j\right)\right) + \underbrace{\Omega(B)}_{\text{régularisation}} \right\}$$

Apprentissage de métrique

$$Y = (Y_1, \ldots, Y_n) \in \left(\mathbb{R}^d\right)^n$$

$$\forall B \succeq 0, \qquad \widehat{m}_B(Y) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ P_n \gamma_B(\widehat{s}_m) + \lambda D_m \right\} \quad \Rightarrow \quad \widehat{B} \ ?$$

- Hypothèse : on dispose de séries segmentées

$$\left(Y^1, m^1\right), \ldots, \left(Y^N, m^N\right) \ .$$

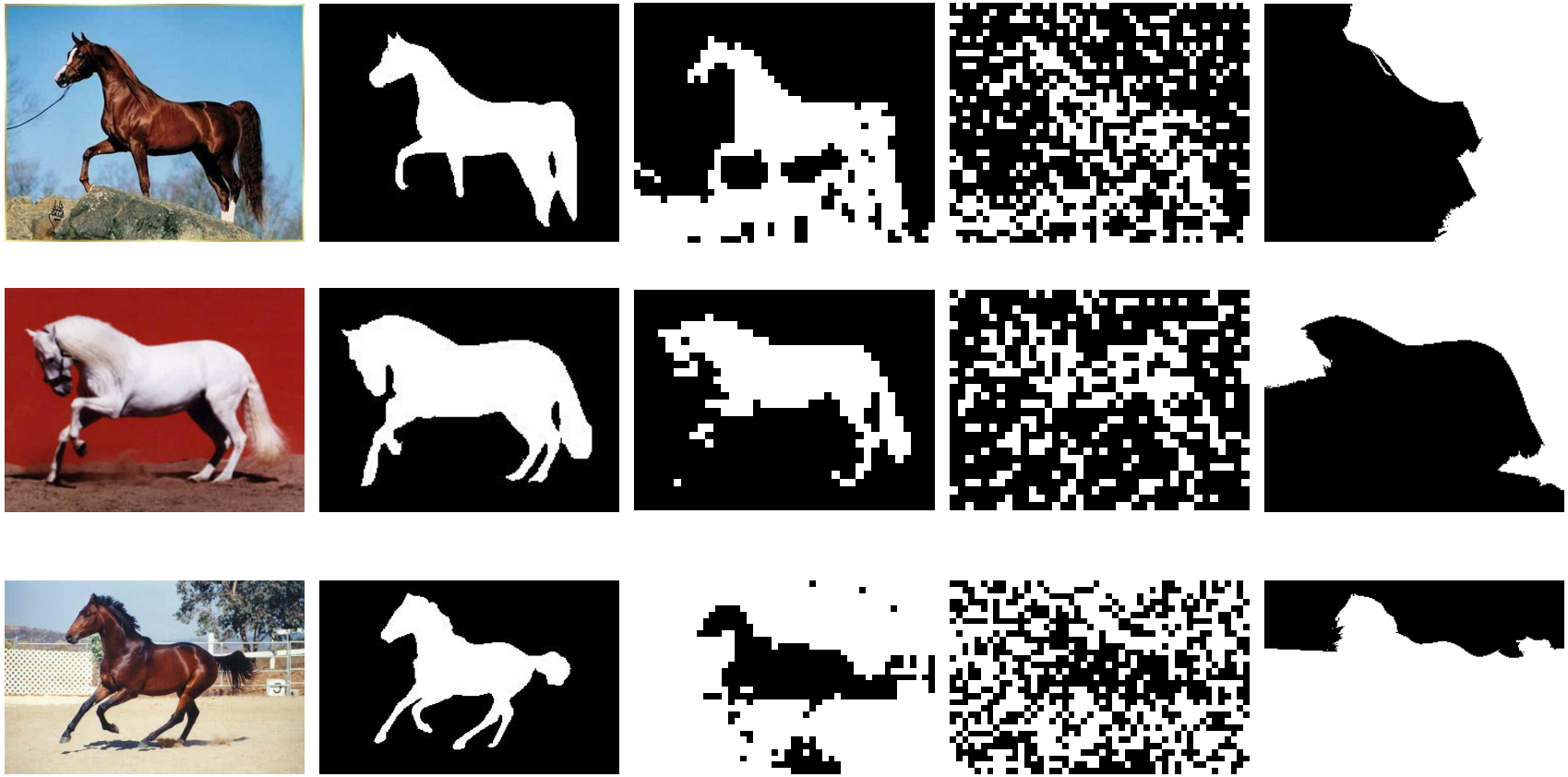
$\Rightarrow$ prédiction structurée : idéalement, on voudrait

$$\widehat{B} \in \operatorname*{argmin}_{B \succeq 0} \left\{ \frac{1}{N} \sum_{j=1}^{N} \gamma\left(m^j, \widehat{m}_B\left(Y^j\right)\right) + \underbrace{\Omega(B)}_{\text{régularisation}} \right\}$$

$\Rightarrow$ relaxation convexe (Tsochantaridis *et al.*, 2005)

Lajugie, A. & Bach. Large-margin metric learning for constrained partitioning problems. In *International Conference on Machine Learning* (*ICML*), volume 32, pages 297–305, 2014.

Apprentissage de métrique : résultats (segmentation 2d)



données notre méthode normalized cuts

cible sans prior

Autres travaux

- Régions de confiance et tests multiples par rééchantillonnage

  A., Blanchard & Roquain. Some nonasymptotic results on resampling in high dimension, I : Confidence regions. II : Multiple tests. *The Annals of Statistics*, 38(1):51–99, 2010.

- Rééchantillonnage et précision d'extrapolation dans une série temporelle, application en astronomie :

  Desmars, A., Arlot, Lainey & Vienne. Estimating the accuracy of satellite ephemerides using the bootstrap method. *Astronomy and Astrophysics*, 499:321–330, 2009.

- Adaptation à la condition de marge en classification :

  A. & Bartlett. Margin adaptive model selection in statistical learning. *Bernoulli*, 17(2):687–713, 2011.

- Erreur d'approximation de forêts purement aléatoires :

  A. & Genuer. Analysis of purely random forests bias, 2014. arXiv:1407.3939v1.

- Apprentissage de métrique pour l'alignement dynamique de séquences :

  Garreau, Lajugie, A. & Bach. Metric learning for temporal sequence alignment. In *Advances in Neural Information Processing Systems 27*, 2014.

Bilan

- Comprendre pourquoi certaines procédures sont performantes
  $\Rightarrow$ heuristique de coude, forêts (purement) aléatoires (Breiman, 2001).

Bilan

- Comprendre pourquoi certaines procédures sont performantes
  $\Rightarrow$ heuristique de coude, forêts (purement) aléatoires (Breiman, 2001).
- Pourquoi certaines procédures fonctionnent-elles mieux ?
  Compromis entre complexité algorithmique et performance statistique ?
  $\Rightarrow$ choix de V pour la validation croisée « V-fold ».

Bilan

- Comprendre pourquoi certaines procédures sont performantes
  $\Rightarrow$ heuristique de coude, forêts (purement) aléatoires (Breiman, 2001).
- Pourquoi certaines procédures fonctionnent-elles mieux ?
  Compromis entre complexité algorithmique et performance statistique ?
  $\Rightarrow$ choix de V pour la validation croisée « V-fold ».
- Corriger les défauts de méthodes couramment utilisées
  $\Rightarrow$ correction du biais de la validation croisée « V-fold ».

Bilan

- Comprendre pourquoi certaines procédures sont performantes
  $\Rightarrow$ heuristique de coude, forêts (purement) aléatoires (Breiman, 2001).
- Pourquoi certaines procédures fonctionnent-elles mieux ? Compromis entre complexité algorithmique et performance statistique ?
  $\Rightarrow$ choix de V pour la validation croisée « V-fold ».
- Corriger les défauts de méthodes couramment utilisées
  $\Rightarrow$ correction du biais de la validation croisée « V-fold ».
- Proposer de nouvelles méthodes sur des bases théoriques
  $\Rightarrow$ généralisation de l'heuristique de pente aux estimateurs linéaires, détection de ruptures par validation croisée ou à noyaux, etc.

Perspectives

- Comparaisons au second ordre
  - Rendre l'heuristique quantitative
  - Application à la surpénalisation ?

Perspectives

- Comparaisons au second ordre
  - Rendre l'heuristique quantitative
  - Application à la surpénalisation ?
- Validation croisée : extension à d'autres cadres
  - Correction du biais
  - Prise en compte de la variance (comportements différents ?)

Perspectives

- Comparaisons au second ordre
  - Rendre l'heuristique quantitative
  - Application à la surpénalisation ?
- Validation croisée : extension à d'autres cadres
  - Correction du biais
  - Prise en compte de la variance (comportements différents ?)
- Pénalités minimales
  - Grandes familles de modèles
  - Surpénalisation automatique ?

Perspectives

- Comparaisons au second ordre
  - Rendre l'heuristique quantitative
  - Application à la surpénalisation ?
- Validation croisée : extension à d'autres cadres
  - Correction du biais
  - Prise en compte de la variance (comportements différents ?)
- Pénalités minimales
  - Grandes familles de modèles
  - Surpénalisation automatique ?
- Détection de ruptures
  - Prise en compte de l'hétéroscédasticité dans le cadre hilbertien
  - Choix du noyau ?