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, 31/01/2011

Plan

1. Lundi 10 : Apprentissage statistique et sélection d'estimateurs
2. Lundi 17 : Calibration de pénalités et pénalités minimales
3. Lundi 24 : Rééchantillonnage et pénalisation
4. Aujourd'hui : Validation croisée et pénalités reliées

Plan du cours

Plan

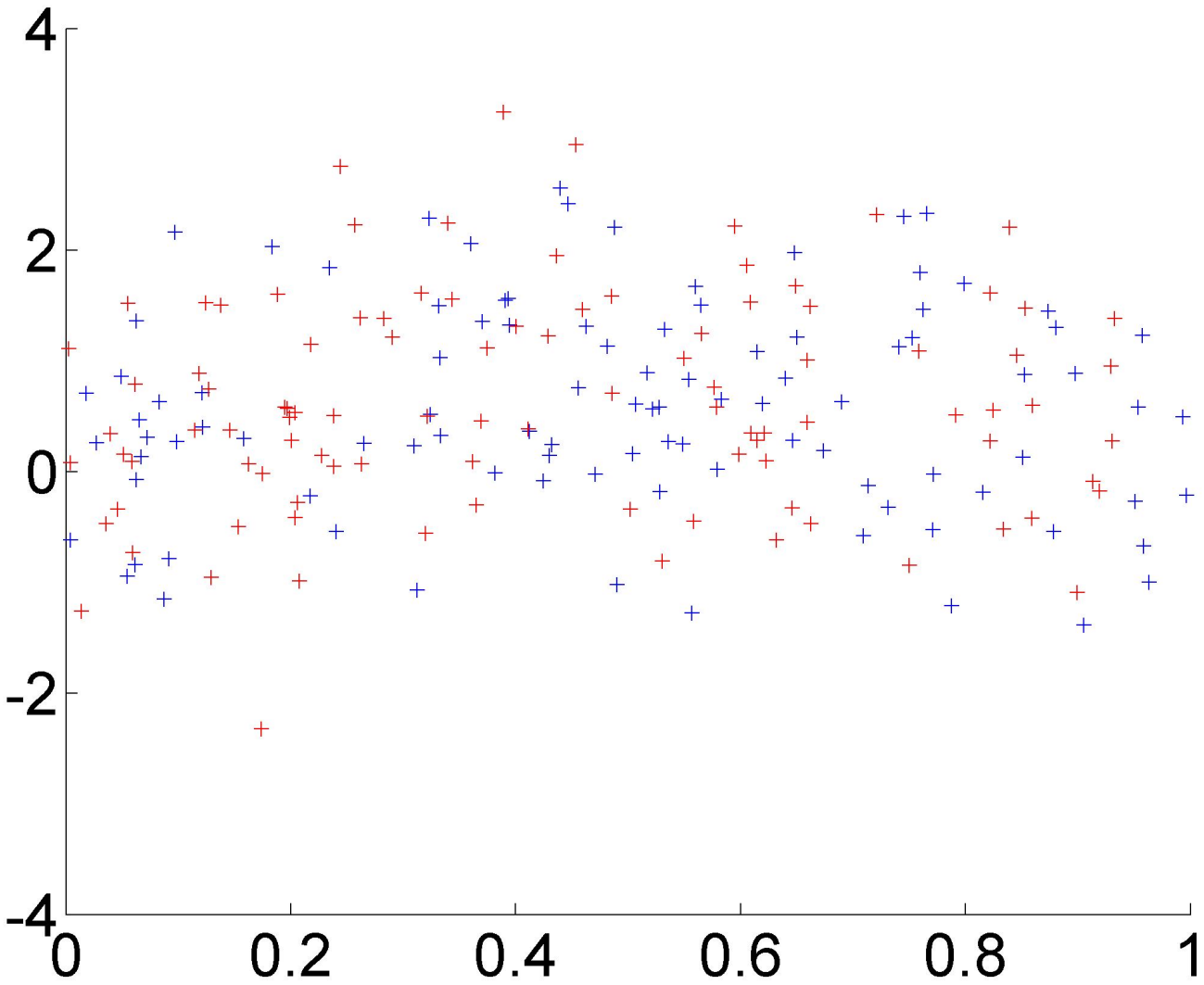
Rappels

- Données : $D_n = (\xi_1, \ldots, \xi_n) \in \Xi^n$, $D_n \sim P^{\otimes n}$
- Perte relative

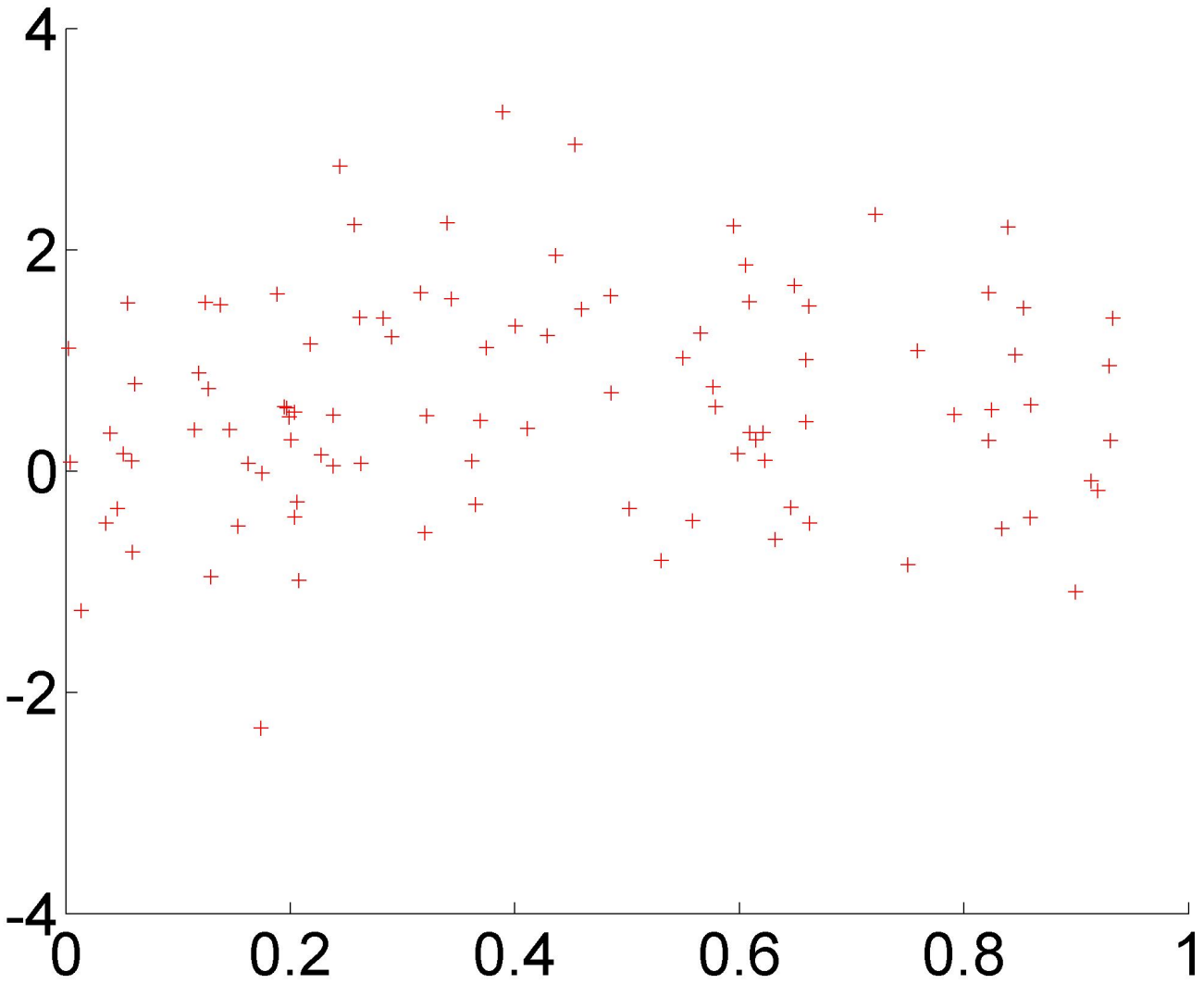
$$\ell\left(s^\star, t\right) = P\gamma(t) - P\gamma(s^\star)$$

- Algorithmes statistiques : $\forall m \in \mathcal{M}_n$, $\mathcal{A}_m : \bigcup_{n\in\mathbb{N}} \Xi^n \mapsto \mathbb{S}$
  $\mathcal{A}_m(D_n) = \widehat{s}_m(D_n) \in \mathbb{S}$ est un estimateur de $s^\star$
- Objectif d'estimation/prédiction : trouver $\widehat{m}(D_n) \in \mathcal{M}$ tel que $\ell\left(s^\star, \widehat{s}_{\widehat{m}(D_n)}(D_n)\right)$ est minimale

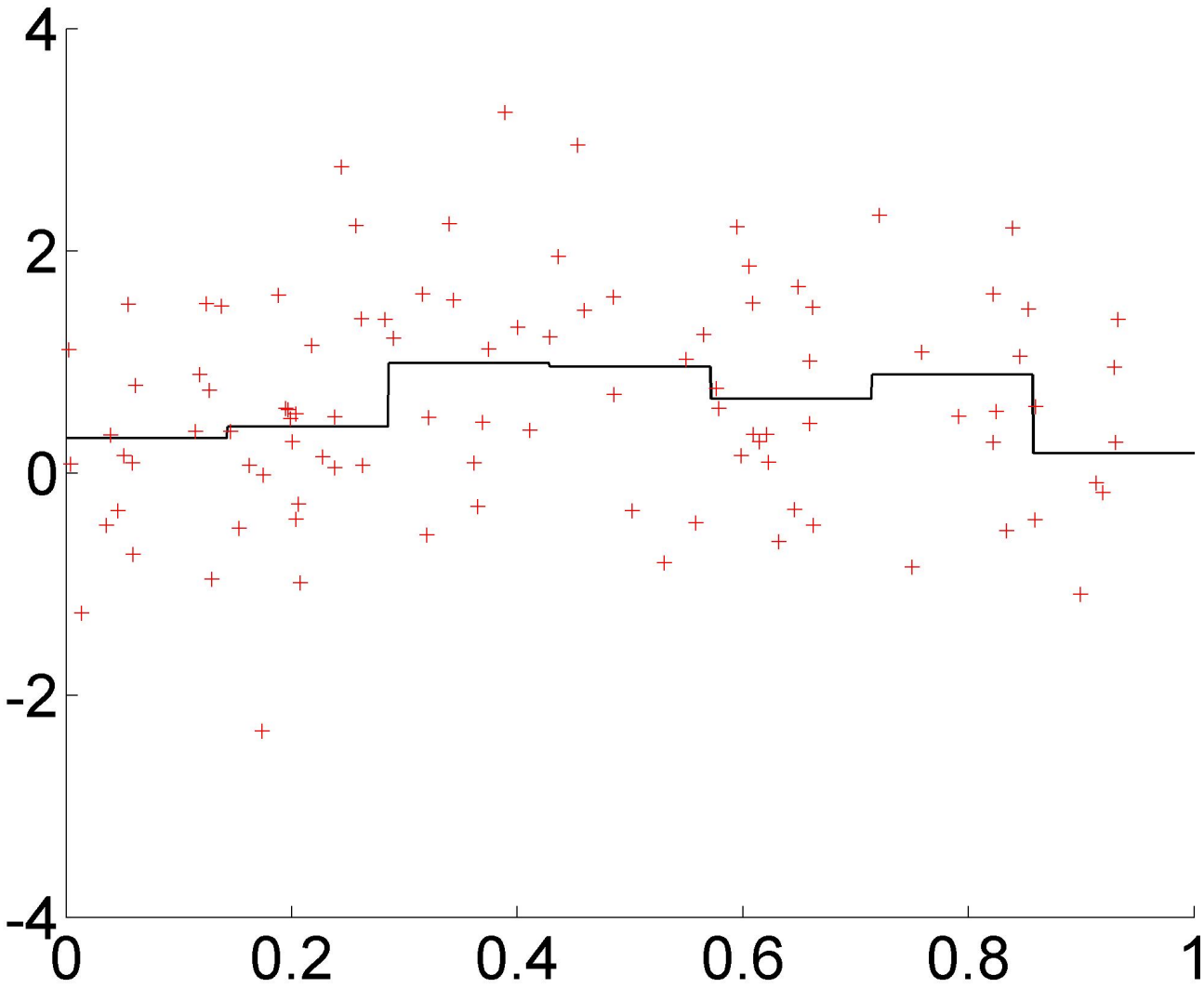
Validation simple (hold-out)



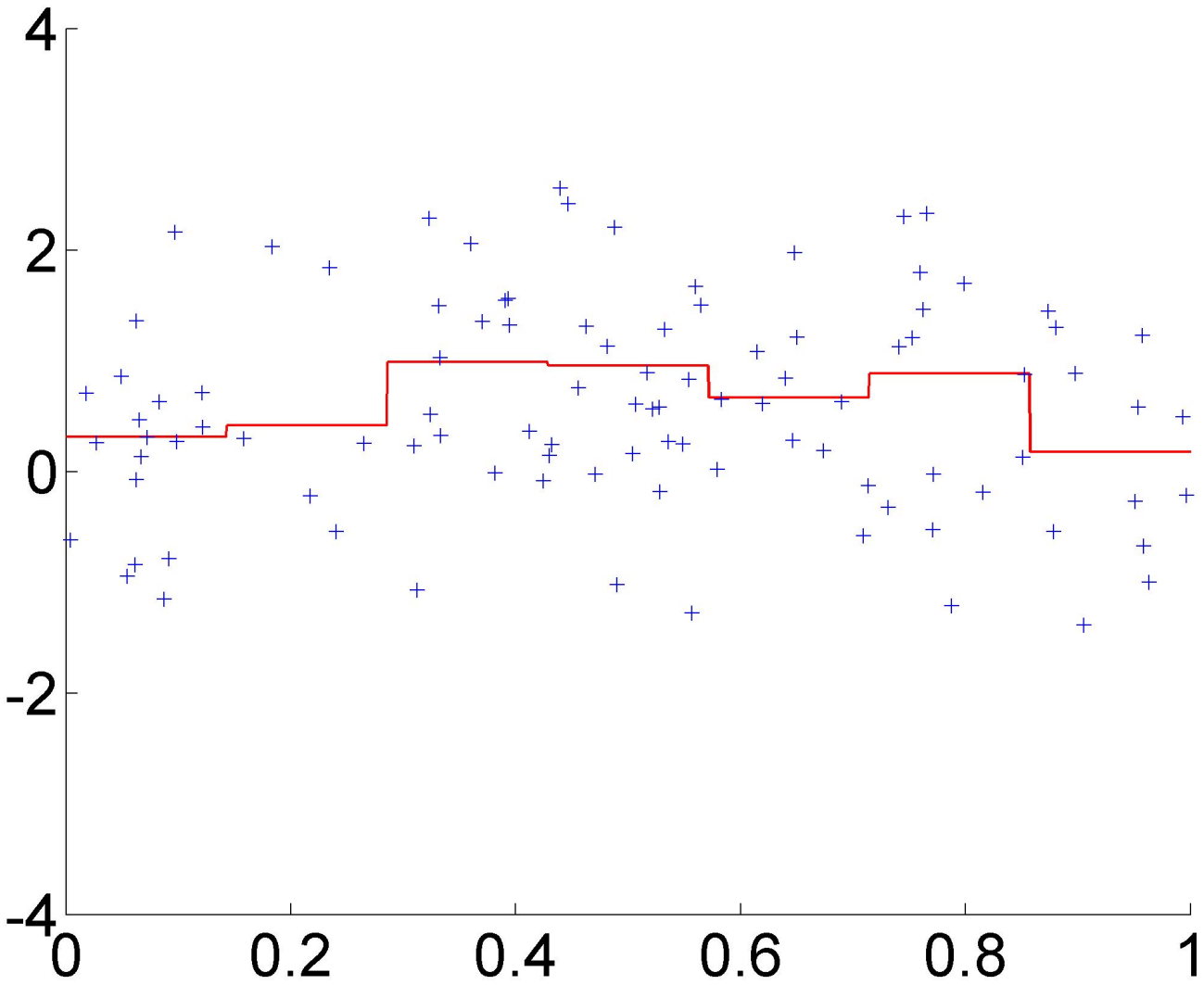
Validation : l'échantillon d'entraînement



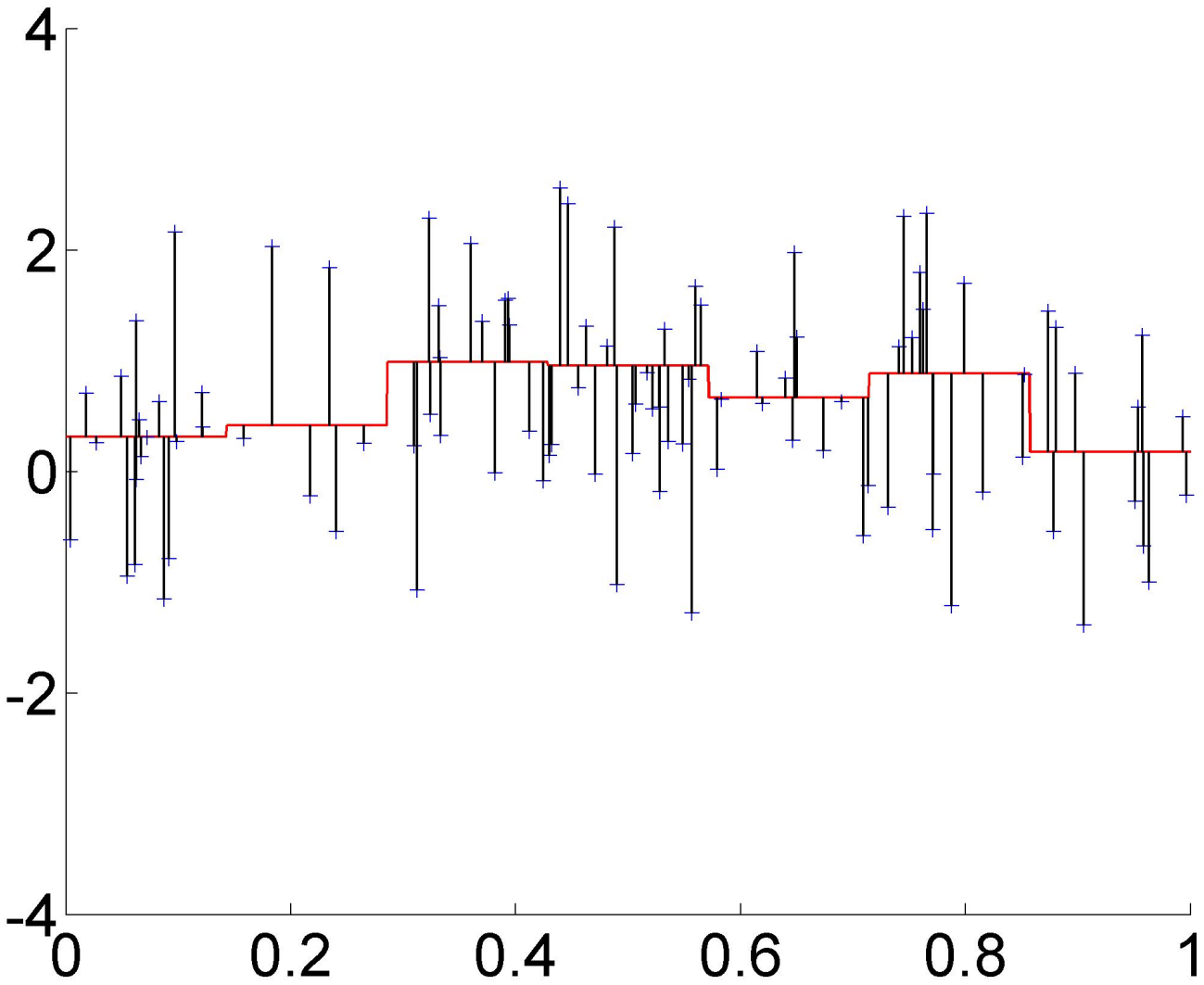
Validation : l'échantillon d'entraînement



Validation : l'échantillon de validation



Validation : l'échantillon de validation



Principe de la validation croisée : hold-out

$$\underbrace{\xi_1 \ , \ \cdots \ , \ \xi_{n_e}}_{\text{Entraînement } (I^{(e)})} \quad , \quad \underbrace{\xi_{n_e+1} \ , \ \cdots \ , \ \xi_n}_{\text{Validation } (I^{(v)})}$$

$$\widehat{s}_m^{(e)} := \mathcal{A}_m \left(D_n^{(e)} \right) \quad \text{où} \quad D_n^{(e)} := \left(\xi_i \right)_{i \in I^{(e)}}$$

$$P_n^{(v)} = \frac{1}{n_v} \sum_{i \in I^{(v)}} \delta_{\xi_i} \qquad n_v := n - n_e$$

$$\Rightarrow \widehat{\mathcal{R}}^{\text{val}} \left(\mathcal{A}_m ; D_n ; I^{(e)} \right) = P_n^{(v)} \gamma \left(\widehat{s}_m^{(e)} \right) = \frac{1}{n_v} \sum_{i \in I^{(v)}} \gamma \left(\mathcal{A}_m \left(D_n^{(e)} \right) ; \xi_i \right)$$

Définition générale de la validation croisée

- $B \geq 1$ ensembles d'entraînement :

$$I_1^{(e)}, \ldots, I_B^{(e)} \subset \{1, \ldots, n\}$$

- Estimateur par validation croisée du risque de $\mathcal{A}_m$:

$$\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\mathcal{A}_m; D_n; \left(I_j^{(e)}\right)_{1 \leq j \leq B}\right) := \frac{1}{B} \sum_{j=1}^{B} \widehat{\mathcal{R}}^{\mathrm{val}}\left(\mathcal{A}_m; D_n; I_j^{(e)}\right)$$

- Algorithme choisi :

$$\widehat{m} \in \mathrm{argmin}_{m \in \mathcal{M}_n} \left\{ \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\mathcal{A}_m; D_n; \left(I_j^{(e)}\right)_{1 \leq j \leq B}\right) \right\}$$

- En général, $\forall j$, $\mathrm{Card}(I_j^{(e)}) = n_e$

Exemples : découpages exhaustifs

- Leave-one-out (LOO), ou delete-one CV, ou validation croisée ordinaire :

$$n_e = n - 1 \qquad B = n$$

(Stone, 1974 ; Allen, 1974 ; Geisser, 1975)

- Leave-p-out (LPO), ou delete-p CV :

$$n_e = n - p \qquad B = \binom{n}{p}$$

Exemples : découpages non-exhaustifs

- Validation croisée "V-fold" (VFCV, Geisser, 1975) : $\mathcal{B} = (B_j)_{1 \leq j \leq V}$ partition de $\{1, \ldots, n\}$

$$\widehat{\mathcal{R}}^{\mathrm{vf}}\left(\mathcal{A}_m; D_n; \mathcal{B}\right) = \frac{1}{V} \sum_{j=1}^{V} \widehat{\mathcal{R}}^{\mathrm{val}}\left(\mathcal{A}_m; D_n; B_j^c\right)$$

- Apprentissage-test répété (RLT, Breiman et al, 1984) : $I_1^{(e)}, \ldots, I_B^{(e)} \subset \{1, \ldots, n\}$ de cardinal n_e, aléatoires et tous différents
- Validation croisée Monte-Carlo (MCCV, Picard et Cook, 1984) : idem avec $I_1^{(e)}, \ldots, I_B^{(e)}$ i.i.d. uniformes parmi les sous-ensembles de taille n_e

Méthodes reliées

- Validation croisée généralisée (GCV) : version invariante par rotation du LOO pour la régression linéaire, plus proche de C_p et C_L que de la validation croisée (Efron, 1986, 2004)
- Approximation analytique du leave-p-out (Shao, 1993)
- Leave-one-out bootstrap (Efron, 1983) :
  version stabilisée du leave-one-out
  correction heuristique du biais $\Rightarrow$ .632 bootstrap
  $\Rightarrow$ .632+ bootstrap (Efron et Tibshirani, 1997)

Biais de l'estimateur validation croisée

- Cible : $P\gamma\left(\mathcal{A}_m(D_n)\right)$
- Biais : si $\forall j$, $\mathrm{Card}(I_j^{(e)}) = n_e$

Biais de l'estimateur validation croisée

- Cible : $P\gamma\left(\mathcal{A}_m(D_n)\right)$
- Biais : si $\forall j$, $\mathrm{Card}(I_j^{(e)}) = n_e$

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\mathcal{A}_m; D_n; \left(I_j^{(e)}\right)_{1\leq j\leq B}\right)\right] = \mathbb{E}\left[P\gamma\left(\mathcal{A}_m\left(D_{n_e}\right)\right)\right]$$

$$\Rightarrow \text{biais } \mathbb{E}\left[P\gamma\left(\mathcal{A}_m\left(D_{n_e}\right)\right)\right] - \mathbb{E}\left[P\gamma\left(\mathcal{A}_m\left(D_n\right)\right)\right]$$

- Algorithme intelligent (Devroye, Györfi & Lugosi, 1996) :
  $n \mapsto \mathbb{E}\left[P\gamma\left(\mathcal{A}_m\left(D_n\right)\right)\right]$ décroissante
  $\Rightarrow$ le biais est positif, minimal pour $n_e = n - 1$
- Exemple : régressogramme :

$$\mathbb{E}\left[P\gamma(\widehat{s}_m(D_n))\right] \approx P\gamma(s_m^\star) + \frac{1}{n}\sum_{\lambda\in m}\sigma_\lambda^2$$

Correction du biais

- Validation croisée V-fold corrigée (Burman, 1989, 1990) :

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

  $+$ idem pour apprentissage-test répété.

- Résultat asymptotique : biais $= \mathcal{O}(n^{-2})$ (Burman, 1989)

Variabilité de l'estimateur validation croisée

$$\mathrm{var}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\mathcal{A}_m; D_n; \left(I_j^{(e)}\right)_{1\leq j\leq B}\right)\right]$$

Sources de variabilité :

Variabilité de l'estimateur validation croisée

$$\mathrm{var}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\mathcal{A}_m; D_n; \left(I_j^{(e)}\right)_{1\leq j\leq B}\right)\right]$$

Sources de variabilité :

- (n_e, n_v) : cas du hold-out (Nadeau & Bengio, 2003)

$$\begin{aligned}
&\mathrm{var}\left[\widehat{\mathcal{R}}^{\mathrm{val}}\left(\mathcal{A}_m; D_n; I^{(e)}\right)\right] \\
&= \mathbb{E}\left[\mathrm{var}\left(P_n^{(v)}\gamma\left(\mathcal{A}_m(D_n^{(e)})\right) \,\middle|\, D_n^{(e)}\right)\right] + \mathrm{var}\left[P\gamma\left(\mathcal{A}_m(D_{n_e})\right)\right] \\
&= \frac{1}{n_v}\mathbb{E}\left[\mathrm{var}\left(\gamma\left(\widehat{s},\xi\right) \,|\, \widehat{s} = \mathcal{A}_m(D_n^{(e)})\right)\right] + \mathrm{var}\left[P\gamma\left(\mathcal{A}_m(D_{n_e})\right)\right]
\end{aligned}$$

Variabilité de l'estimateur validation croisée

$$\operatorname{var}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\mathcal{A}_m; D_n; \left(I_j^{(e)}\right)_{1\leq j\leq B}\right)\right]$$

Sources de variabilité :

- (n_e, n_v) : cas du hold-out (Nadeau & Bengio, 2003)

$$\begin{aligned}
&\operatorname{var}\left[\widehat{\mathcal{R}}^{\mathrm{val}}\left(\mathcal{A}_m; D_n; I^{(e)}\right)\right] \\
&= \mathbb{E}\left[\operatorname{var}\left(P_n^{(v)}\gamma\left(\mathcal{A}_m(D_n^{(e)})\right) \,\middle|\, D_n^{(e)}\right)\right] + \operatorname{var}\left[P\gamma\left(\mathcal{A}_m(D_{n_e})\right)\right] \\
&= \frac{1}{n_v}\mathbb{E}\left[\operatorname{var}\left(\gamma\left(\widehat{s},\xi\right) \mid \widehat{s} = \mathcal{A}_m(D_n^{(e)})\right)\right] + \operatorname{var}\left[P\gamma\left(\mathcal{A}_m(D_{n_e})\right)\right]
\end{aligned}$$

- Stabilité de l'algorithme $\mathcal{A}_m$ (Bousquet & Elisseff, 2002)
- Nombre de découpages B
- Difficulté : B, n_e, n_v liés pour VFCV et LPO

Résultats sur la variabilité

- Régression linéaire, moindres carrés, cas particulier (Burman, 1989) :

$$\frac{2\sigma^2}{n} + \frac{4\sigma^4}{n^2}\left[4 + \frac{4}{V-1} + \frac{2}{(V-1)^2} + \frac{1}{(V-1)^3}\right] + o\left(n^{-2}\right)$$

Résultats sur la variabilité

- Régression linéaire, moindres carrés, cas particulier (Burman, 1989) :

$$\frac{2\sigma^2}{n} + \frac{4\sigma^4}{n^2}\left[4 + \frac{4}{V-1} + \frac{2}{(V-1)^2} + \frac{1}{(V-1)^3}\right] + o\left(n^{-2}\right)$$

- Quantification explicite en régression (LPO) et estimation de densité (VFCV, LPO) : Celisse (2008)
- LOO très variable lorsque $\mathcal{A}_m$ est instable (e.g., k-NN ou CART), beaucoup moins lorsque $\mathcal{A}_m$ est stable (e.g., estimateur des moindres carrés ; cf. Molinaro et al, 2005)
- Difficulté d'estimer la variabilité de la validation croisée : pas d'estimateur universellement non biaisé (RLT, Nadeau et Bengio, 2003 ; VFCV, Bengio et Grandvalet, 2004), plusieurs estimateurs proposés (ibid. ; Markatou et al, 2005 ; Celisse et Robin, 2008)

Plan

Lien entre estimation du risque et choix d'algorithme

- Principe d'estimation non biaisée du risque
  $\Rightarrow$ la quantité importante (asymptotiquement) est le biais
- Quel est le meilleur critère ?
  En principe, le meilleur $\widehat{m}$ est celui qui minimise le meilleur estimateur du risque.
- Situation parfois plus compliquée (Breiman et Spector, 1992) :
  - Seuls les m "proches" de l'oracle $m^\star$ comptent
  - Surpénalisation parfois nécessaire (beaucoup de modèles et/ou petit rapport signal-sur-bruit)

Rappel : un lemme clé

Lemme

Sur l'événement Ω où pour tout $m, m' \in \mathcal{M}_n$,

$$\begin{aligned}&\left(\mathrm{crit}(m) - P\gamma\left(\widehat{s}_m(D_n)\right)\right) - \left(\mathrm{crit}(m') - P\gamma\left(\widehat{s}_{m'}(D_n)\right)\right)\\ &\qquad \leq A(m) + B(m')\end{aligned}$$

$$\text{on a} \quad \forall \widehat{m} \in \mathrm{argmin}_{m \in \mathcal{M}_n} \left\{\mathrm{crit}(m)\right\}$$

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

Validation croisée pour l'estimation : rôle de n_e

Cadre de la régression linéaire (Shao, 1997) représentatif du comportement général de la validation croisée :

- Si $n_e \sim n$, optimalité asymptotique (CV $\sim C_p$)
- Si $n_e \sim \kappa n$, $\kappa \in]0,1[$, CV $\sim$ GIC$_{1+\kappa^{-1}}$ (i.e., surpénalise d'un facteur $(1+\kappa^{-1})/2 \Rightarrow$ asymptotiquement sous-optimal)

$\Rightarrow$ valable pour LPO (Shao, 1997), RLT (si $B \gg n^2$, Zhang, 1993)

Sous-optimalité de la validation croisée "V-fold"

- $Y = X + \sigma\varepsilon$ avec ε borné et $\sigma > 0$
- $\mathcal{M} = \mathcal{M}_n^{(\mathrm{reg})}$ (histogrammes réguliers sur $\mathcal{X} = [0,1]$)
- $\widehat{m}$ obtenu par validation croisée "V-fold" avec V fixe quand n grandit

Théorème (A. 2008)

Avec probabilité $1 - Ln^{-2}$,

$$\ell\left(s^\star, \widehat{s}_{\widehat{m}}\right) \geq \left(1 + \kappa(V)\right) \inf_{m \in \mathcal{M}_n} \left\{\ell\left(s^\star, \widehat{s}_m\right)\right\}$$

où $\kappa(V) > 0$

Inégalités oracle pour la validation croisée

- Si $n_v \to \infty$ suffisamment vite, il est "facile" de prouver que le hold-out fait au moins aussi bien que

$$\text{argmin}_{m \in \mathcal{M}_n} \left\{ P\gamma\left(\mathcal{A}_m(D_{n_e})\right) \right\}$$

- van der Laan, Dudoit et van der Vaart (2006) : même propriété pour LPO, VFCV et MCCV dans un cadre assez général

Inégalités oracle pour la validation croisée

- Si $n_v \to \infty$ suffisamment vite, il est "facile" de prouver que le hold-out fait au moins aussi bien que

$$\text{argmin}_{m \in \mathcal{M}_n} \left\{ P\gamma\left(\mathcal{A}_m(D_{n_e})\right) \right\}$$

- van der Laan, Dudoit et van der Vaart (2006) : même propriété pour LPO, VFCV et MCCV dans un cadre assez général
- Régressogrammes : VFCV sous-optimale, mais s'adapte aux variations du niveau de bruit (à une constante $C(V) > 1$ près)
- LPO en régression et estimation de densité avec $p/n \in [a, b]$, $0 < a < b < 1$: Celisse (2008)

Inégalités oracle pour la validation croisée

- Si $n_v \to \infty$ suffisamment vite, il est "facile" de prouver que le hold-out fait au moins aussi bien que

$$\text{argmin}_{m \in \mathcal{M}_n} \left\{ P\gamma\left(\mathcal{A}_m(D_{n_e})\right) \right\}$$

- van der Laan, Dudoit et van der Vaart (2006) : même propriété pour LPO, VFCV et MCCV dans un cadre assez général
- Régressogrammes : VFCV sous-optimale, mais s'adapte aux variations du niveau de bruit (à une constante $C(V) > 1$ près)
- LPO en régression et estimation de densité avec $p/n \in [a, b]$, $0 < a < b < 1$: Celisse (2008)

- Question ouverte : comparaison théorique entre méthodes tenant compte de B (et donc de la variabilité de la validation croisée)

Validation croisée pour l'identification : problème

- Famille d'algorithmes $(\mathcal{A}_m)_{m \in \mathcal{M}}$
- Objectif : sélectionner celui qui se comportera le mieux sur un nouvel échantillon de taille $n' \to \infty$

$$m_0 \in \lim_{n' \to \infty} \operatorname{argmin}_{m \in \mathcal{M}} \left\{ \mathbb{E} \left[P\gamma \left(\mathcal{A}_m(D'_{n'}) \right) \right] \right\}$$

- Consistance :

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

- Exemples :
  - identification du vrai modèle en sélection de modèles
  - algorithme paramétrique ou non-paramétrique ?
  - $\widehat{k}$-ppv ou SVM ?
  - ...

Validation croisée avec vote (Yang, 2006)

Deux algorithmes $\mathcal{A}_1$ et $\mathcal{A}_2$

- Pour $m = 1, 2$

$$\left(\widehat{\mathcal{R}}^{\mathrm{val}} \left(\mathcal{A}_m ; D_n ; I_j^{(e)} \right) \right)_{1 \leq j \leq B}$$

$\Rightarrow$ vote majoritaire

$$\mathcal{V}_1(D_n) = \mathrm{Card} \left\{ j \text{ t.q. } \widehat{\mathcal{R}}^{\mathrm{val}} \left(\mathcal{A}_1 ; D_n ; I_j^{(e)} \right) < \widehat{\mathcal{R}}^{\mathrm{val}} \left(\mathcal{A}_2 ; D_n ; I_j^{(e)} \right) \right\}$$

$$\widehat{m} = \begin{cases} 1 & \text{si} \quad \mathcal{V}_1(D_n) > n/2 \\ 2 & \text{sinon} \end{cases}$$

- Validation croisée classique : moyenne sur les découpages puis comparaison

Validation croisée pour l'identification : régression

- "Paradoxe de la validation croisée" (Yang, 2007)
- $r_{n,m}$: asymptotique de $\mathbb{E}\|\mathcal{A}_m(D_n) - s^\star\|_2$
- But : retrouver $\operatorname{argmin}_{m\in\mathcal{M}} r_{n,m}$
- Hypothèse : différence d'au moins une constante $C > 1$ entre $r_{n,1}$ et $r_{n,2}$

Validation croisée pour l'identification : régression

- "Paradoxe de la validation croisée" (Yang, 2007)
- $r_{n,m}$: asymptotique de $\mathbb{E}\|\mathcal{A}_m(D_n) - s^\star\|_2$
- But : retrouver $\operatorname{argmin}_{m\in\mathcal{M}} r_{n,m}$
- Hypothèse : différence d'au moins une constante $C > 1$ entre $r_{n,1}$ et $r_{n,2}$
- VFCV, RLT, LPO (avec vote) sont consistantes si

$$n_v, n_e \to \infty \quad \text{et} \quad \sqrt{n_v} \max_{m\in\mathcal{M}} r_{n_e,m} \to \infty$$

sous des conditions sur $(\|\mathcal{A}_m(D_n) - s^\star\|_p)_{p=2,4,\infty}$

Validation croisée pour l'identification : régression

- Paramétrique vs. paramétrique ($r_{n,m} \propto n^{-1/2}$)
  $\Rightarrow$ la condition devient $n_v \gg n_e \to \infty$
- Non-paramétrique vs. (non-)paramétrique
  ($\max_{m \in \mathcal{M}} r_{n,m} \gg n^{-1/2}$)
  $\Rightarrow$ il suffit d'avoir $n_e / n_v = \mathcal{O}(1)$, et on peut avoir $n_e \sim n$ (mais pas trop proche)
- Intuition :
  - risques estimés avec une précision $\propto n_v^{-1/2}$
  - différence des risques de l'ordre de $\max_{m \in \mathcal{M}} r_{n_e,m}$
    $\Rightarrow$ plus facile de distinguer des algorithmes avec n_e petit car l'écart entre risques est plus grand (discutable en pratique)

Validation croisée en pratique : temps de calcul

- Implémentation naïve : complexité proportionnelle à B
  $\Rightarrow$ LPO inutilisable, LOO parfois
  $\Rightarrow$ VFCV, RLT et MCCV souvent préférables

Validation croisée en pratique : temps de calcul

- Implémentation naïve : complexité proportionnelle à B
  $\Rightarrow$ LPO inutilisable, LOO parfois
  $\Rightarrow$ VFCV, RLT et MCCV souvent préférables
- Formules closes pour le LPO en estimation de densité (moindres carrés) et en régression (estimateurs par projection, par noyau) : Celisse et Robin (2008), Celisse (2008)
  $\Rightarrow$ utilisable par exemple en détection de ruptures (avec programmation dynamique)
- Validation croisée généralisée : généralise une formule pour le LOO en régression linéaire

Validation croisée en pratique : temps de calcul

- Implémentation naïve : complexité proportionnelle à B
  $\Rightarrow$ LPO inutilisable, LOO parfois
  $\Rightarrow$ VFCV, RLT et MCCV souvent préférables
- Formules closes pour le LPO en estimation de densité (moindres carrés) et en régression (estimateurs par projection, par noyau) : Celisse et Robin (2008), Celisse (2008)
  $\Rightarrow$ utilisable par exemple en détection de ruptures (avec programmation dynamique)
- Validation croisée généralisée : généralise une formule pour le LOO en régression linéaire
- En l'absence de formules closes, algorithmes intelligents pour le LOO (analyse discriminante linéaire, Ripley, 1996 ; k-ppv, Daudin et Mary-Huard, 2008) : utilise les résultats obtenus pour les découpages précédents pour éviter de refaire une partie des calculs

Choix d'une méthode de validation croisée

Compromis entre biais, variabilité et temps de calcul :

- Biais : d'autant plus grand que n_e éloigné de n (sauf pour les méthodes avec correction du biais)
  SNR grand : le biais doit être minimal
  SNR petit : un peu de biais peut être préférable ($\Rightarrow n_e = \kappa n$ pour un certain $\kappa \in \,]0,1[\,$)
- Variabilité : en général, décroît avec B et avec n_v, mais cela dépend de la nature des algorithmes considérés (stabilité)
- Temps de calcul : proportionnel à B, sauf cas particuliers

VFCV : B et n_e reliés à $V \Rightarrow$ situation complexe ($V = 10$ n'est pas toujours un bon choix)

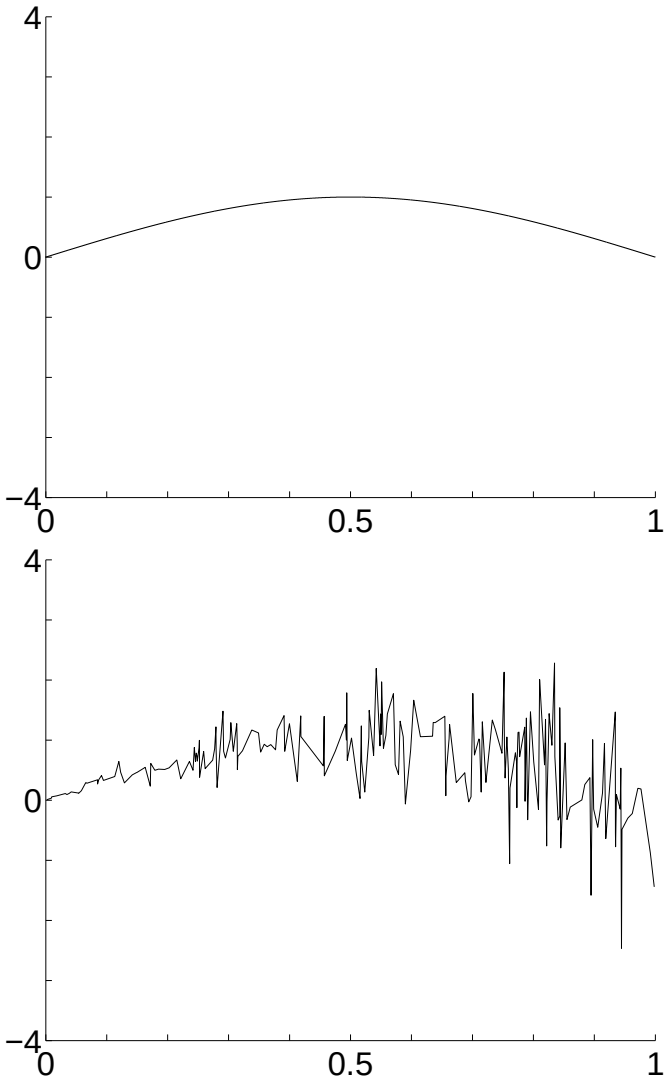
Choix des échantillons d'entraînement

- Recommandation usuelle : tenir compte de la structure des données, par exemple :
  - répartition des X_i dans l'espace en régression
  - répartition des Y_i dans les classes en classification
  - ...

  mais pas de résultat théorique très clair (simulations de Breiman et Spector, 1992 : différence non-significative).

- Dépendance entre les $I_j^{(e)}$?
  Intuitivement, mieux vaut donner à toutes les données des rôles similaires dans les tâches d'entraînement et de validation $\Rightarrow$ VFCV
  Mais pas de résultat clair comparant VFCV (forte dépendance), RLT (faible dépendance) et MCCV (indépendance).

VFCV : Simulations : sin, $n = 200$, $\sigma(x) = x$, 2 pas



Modèles : $\mathcal{M}_n = \mathcal{M}_n^{(\mathrm{reg},1/2)}$

$$\frac{\mathbb{E}\left[\ell\left(s^\star, \widehat{s}_{\widehat{m}}\right)\right]}{\mathbb{E}\left[\inf_{m\in\mathcal{M}_n}\left\{\ell\left(s^\star, \widehat{s}_m\right)\right\}\right]}$$

calculé avec $N = 1000$ échantillons

Mallows	3.69 ± 0.07
2-fold	2.54 ± 0.05
5-fold	2.58 ± 0.06
10-fold	2.60 ± 0.06
20-fold	2.58 ± 0.06
leave-one-out	2.59 ± 0.06

Universalité de la validation croisée ?

- Heuristique quasi universelle (données i.i.d., pas d'autre hypothèse explicite)

- Mais $D_n \mapsto \mathcal{A}_{\widehat{m}(D_n)}$ reste un algorithme d'apprentissage
  $\Rightarrow$ concerné par les "No Free Lunch Theorems"

- Hypothèses implicites de la validation croisée :
  - on estime bien l'erreur de généralisation à partir d'un nombre fini n_v de points
  - le comportement d'un algorithme avec n_e points est représentatif de son comportement avec n points

  $+$ les hypothèses de l'estimation sans biais du risque

Séries temporelles et données dépendantes

- non valable a priori (données non i.i.d.)
- Processus de Markov stationnaire $\Rightarrow$ CV fonctionne toujours (Burman et Nolan, 1992)
- Corrélations positives $\Rightarrow$ risque de sur-apprentissage (Hart et Wehrly, 1986 ; Opsomer et. al, 2001)

Séries temporelles et données dépendantes

- non valable a priori (données non i.i.d.)
- Processus de Markov stationnaire $\Rightarrow$ CV fonctionne toujours (Burman et Nolan, 1992)
- Corrélations positives $\Rightarrow$ risque de sur-apprentissage (Hart et Wehrly, 1986 ; Opsomer et. al, 2001)
- Solution : si dépendances à courte distance, choisir $I^{(e)}$ et $I^{(v)}$ tels que

$$\min_{i \in I^{(e)}, j \in I^{(v)}} |i - j| \geq h > 0$$

$\Rightarrow$ CV modifiée (Chu et Marron, 1991), "h-block CV" (avec correction du biais, Burman et al. 1994), etc.

Grand nombre de modèles

- Sélection de modèles en régression, nombre exponentiel de modèles par dimension $\Rightarrow$ pénalité minimale de l'ordre de $\ln(n)D_m/n$ (Birgé et Massart, 2007)

  $\Rightarrow$ la validation croisée sur-apprend (sauf peut-être si $n_e \ll n$)

Grand nombre de modèles

- Sélection de modèles en régression, nombre exponentiel de modèles par dimension $\Rightarrow$ pénalité minimale de l'ordre de $\ln(n)D_m/n$ (Birgé et Massart, 2007)

  $\Rightarrow$ la validation croisée sur-apprend (sauf peut-être si $n_e \ll n$)

- Wegkamp (2003) : hold-out pénalisé

- A. et Celisse (2009) : regroupement des modèles par dimension, application en détection de ruptures

Plan

Détection de ruptures et sélection de modèles

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

- But : détecter les ruptures dans la moyenne η 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)

- Ici : pas d'hypothèse sur la variance $\sigma(t_i)^2$

Approche classique (Lebarbier, 2005 ; Lavielle, 2005)

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

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

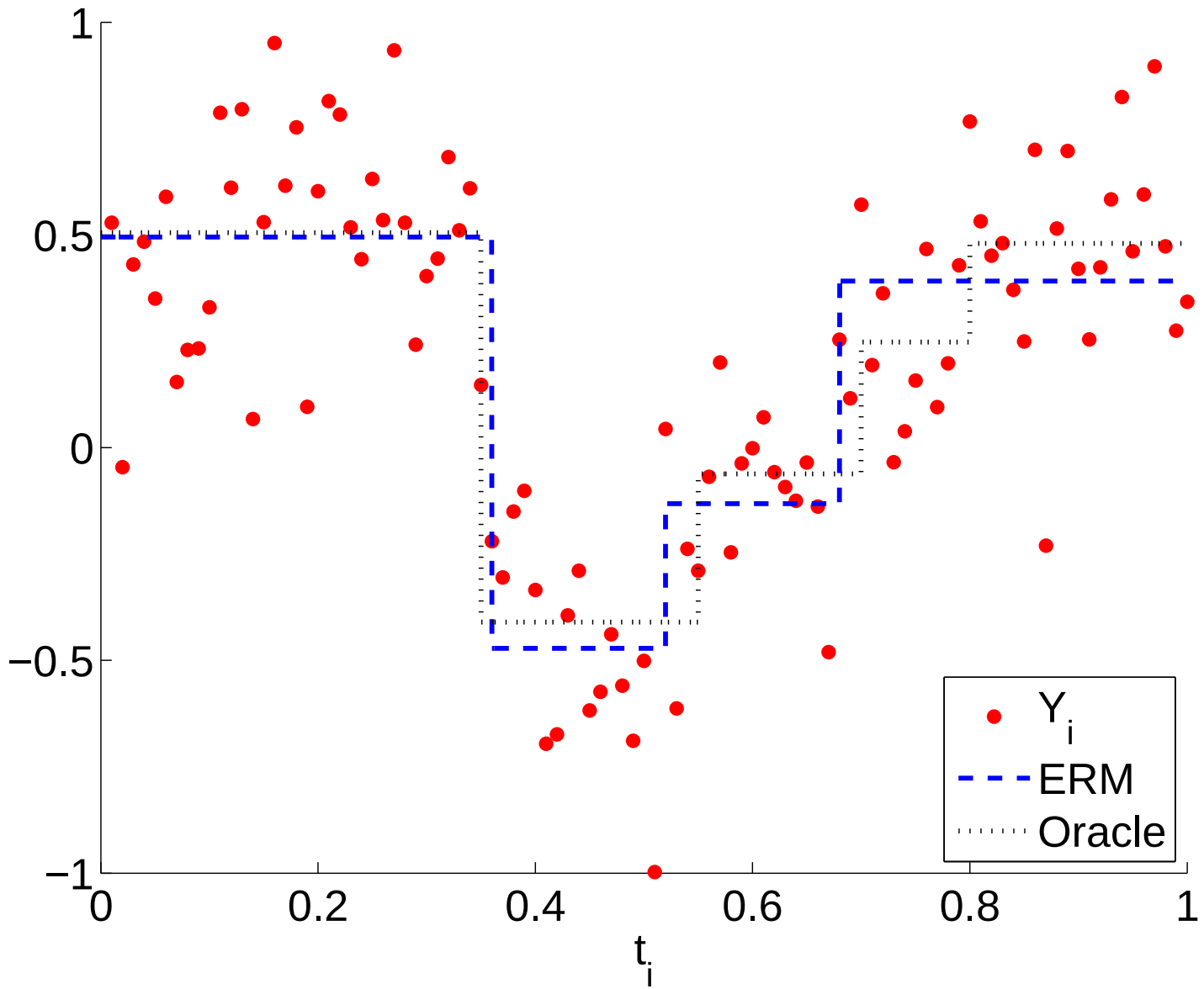
- Revient à aggréger les modèles de même dimension :

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

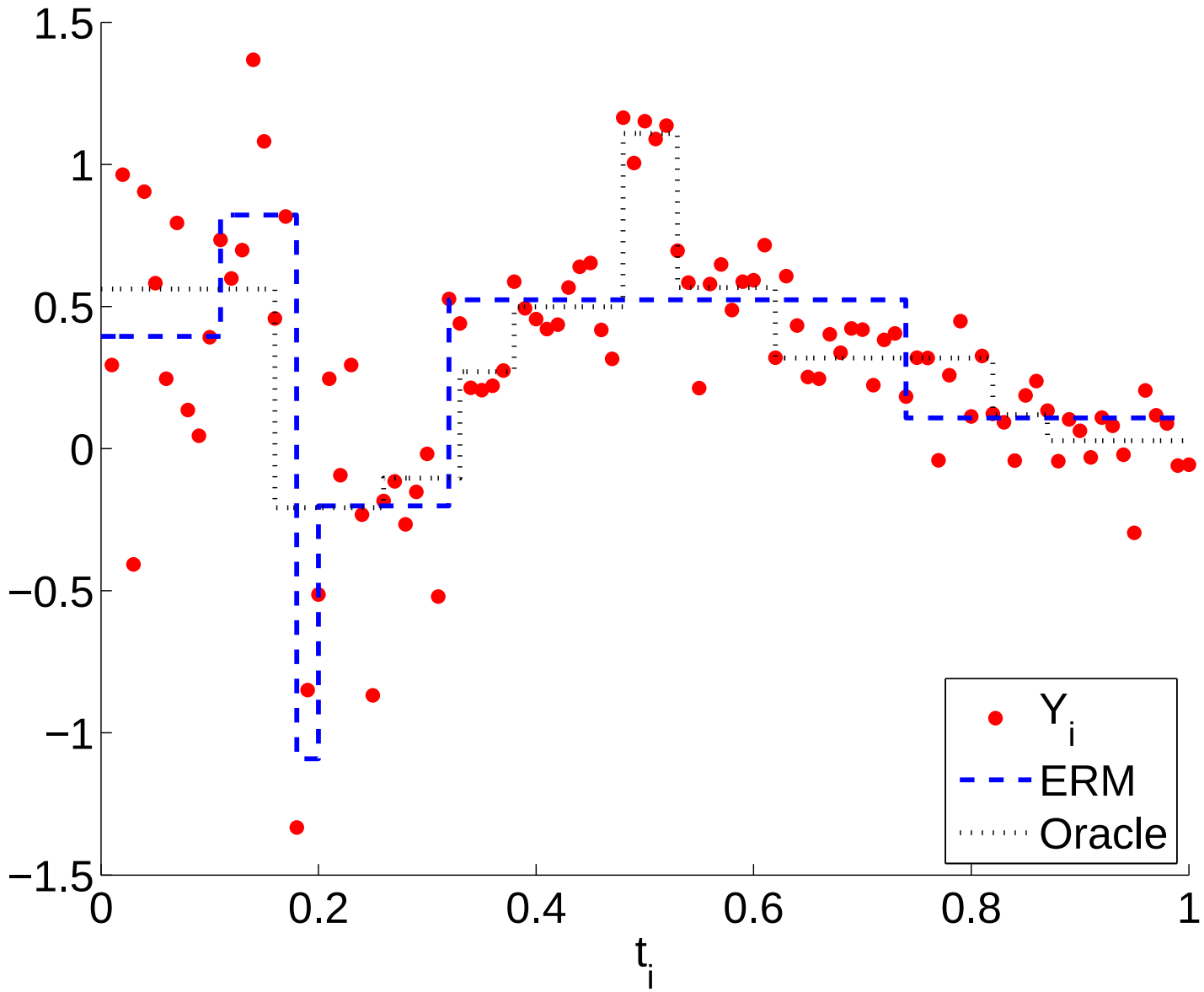
$$\widehat{s}_D \in \operatorname{argmin}_{t \in \widetilde{S}_D} \left\{ P_n \gamma\left(t\right) \right\} \qquad \text{programmation dynamique}$$

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

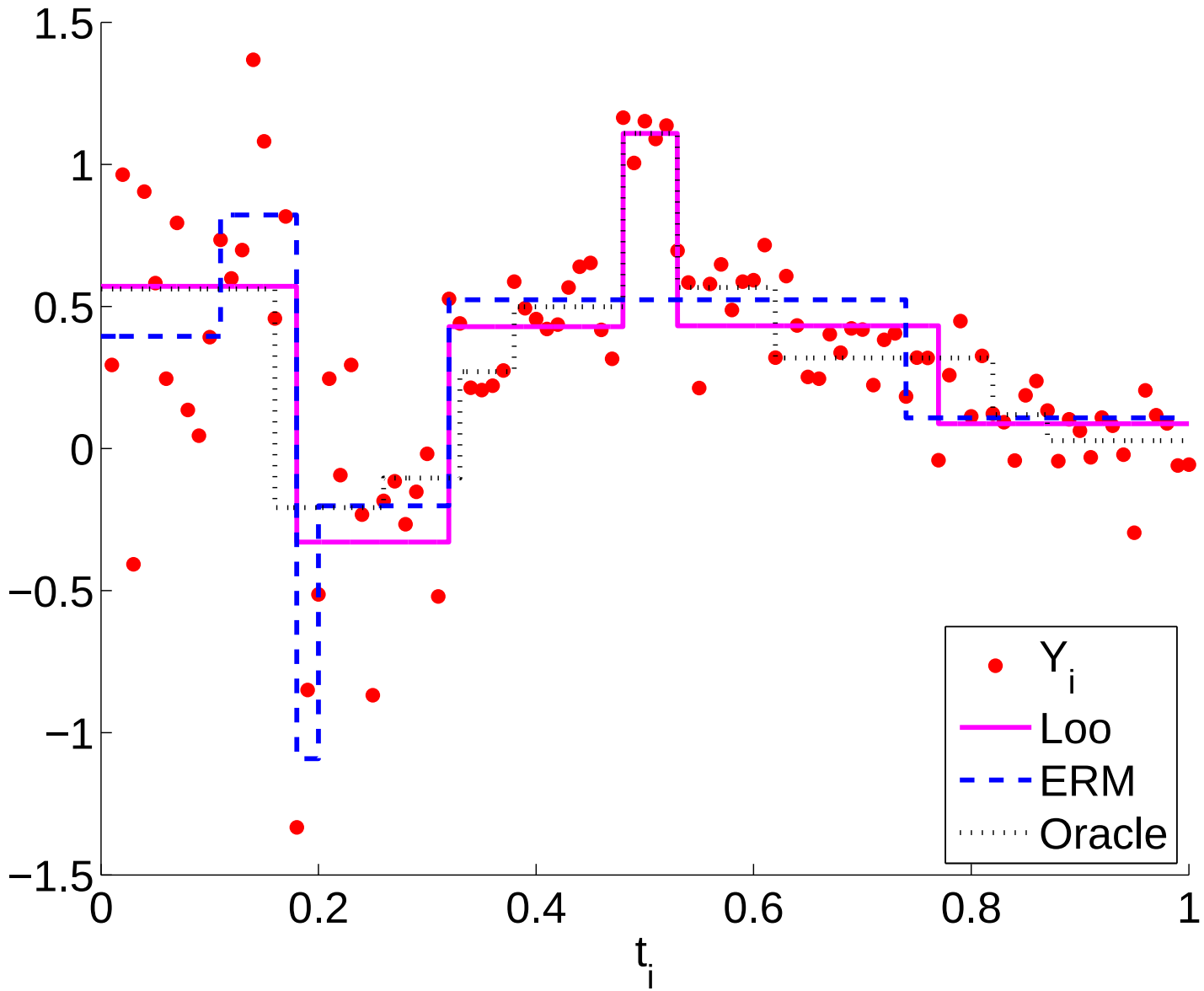
$D = 4$, homoscédastique ; $n = 100$, $\sigma = 0.25$



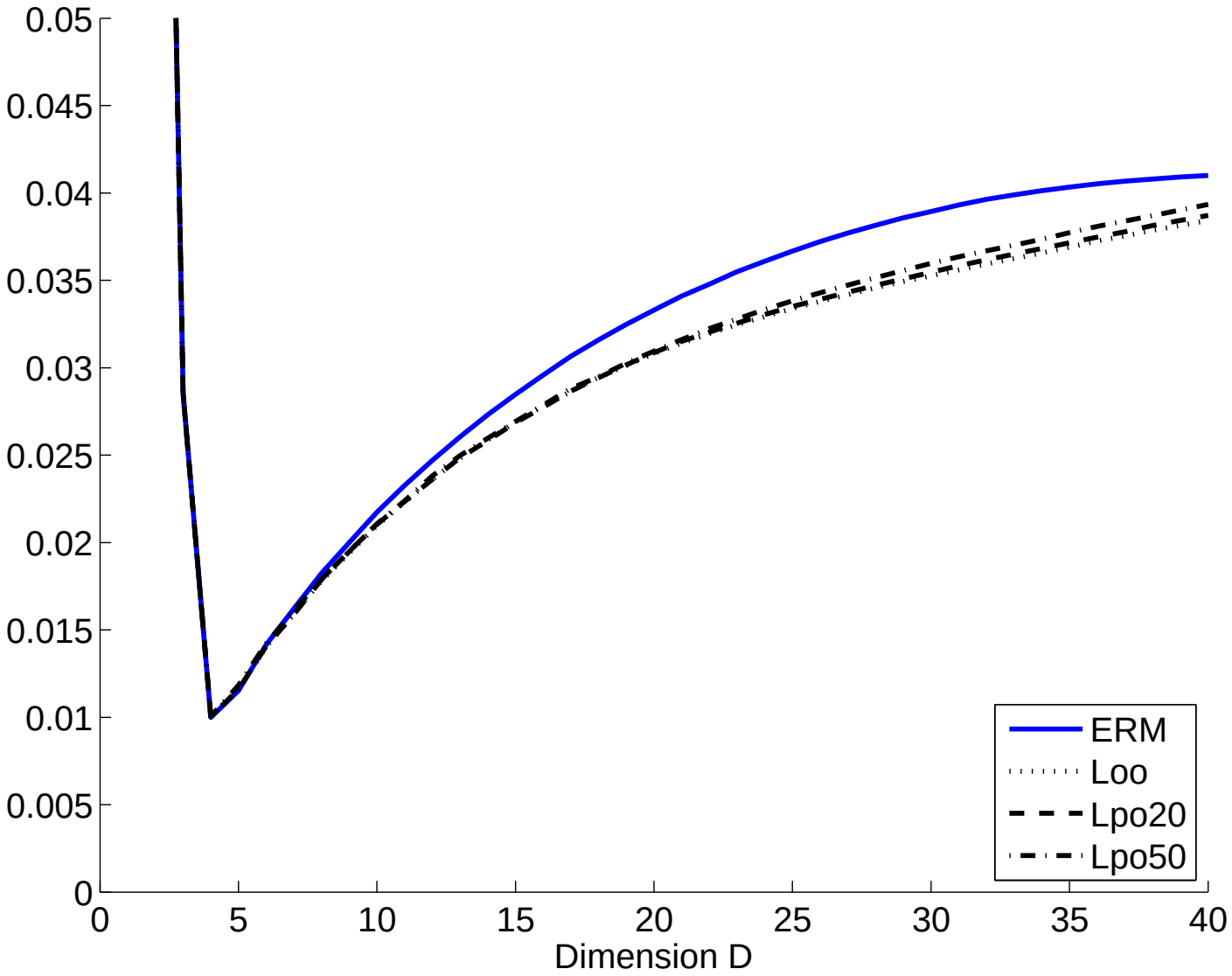
$D = 6$, hétéroscédastique ; $n = 100$, $\|\sigma\| = 0.30$



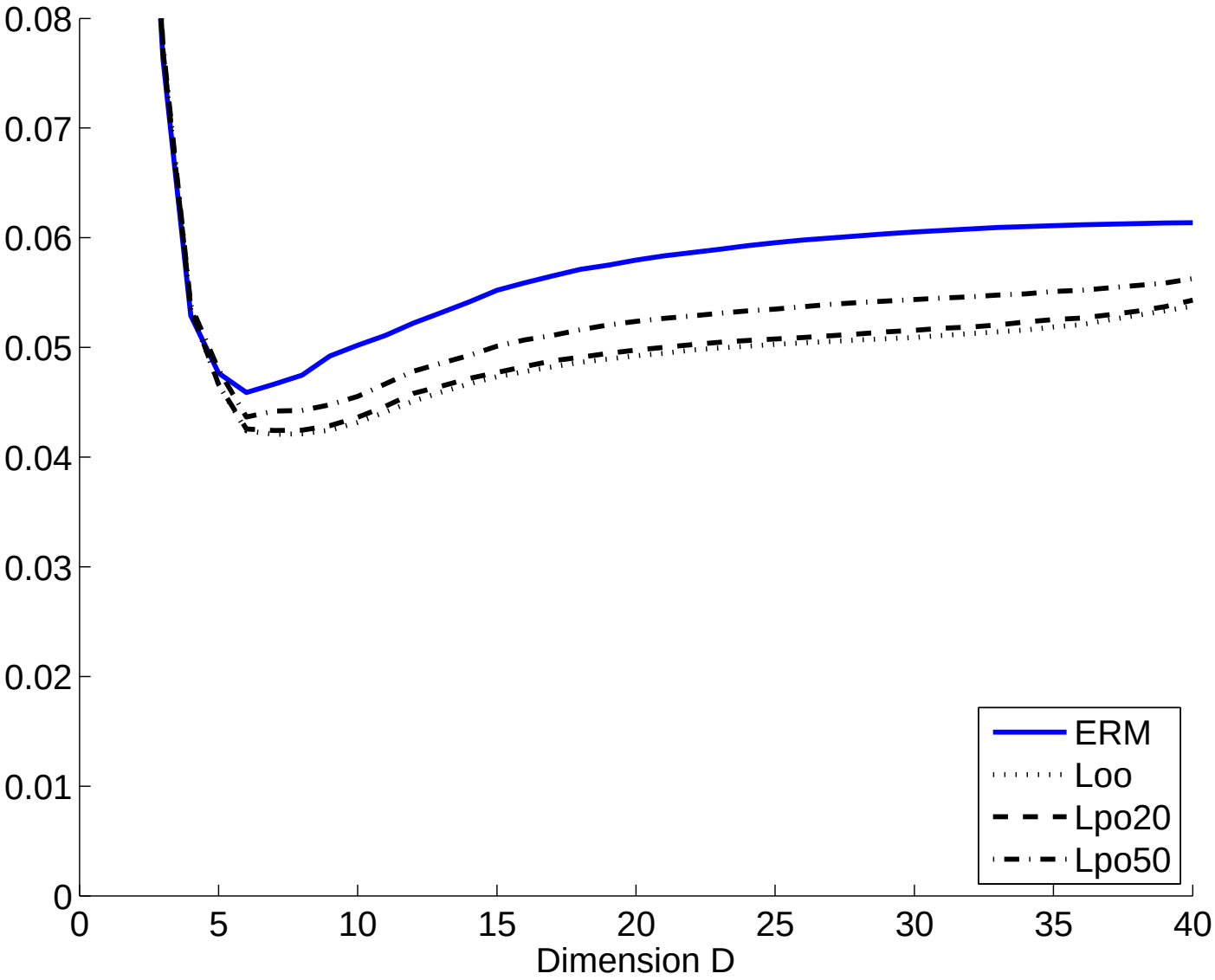
$D = 6$, hétéroscédastique ; $n = 100$, $\|\sigma\| = 0.30$



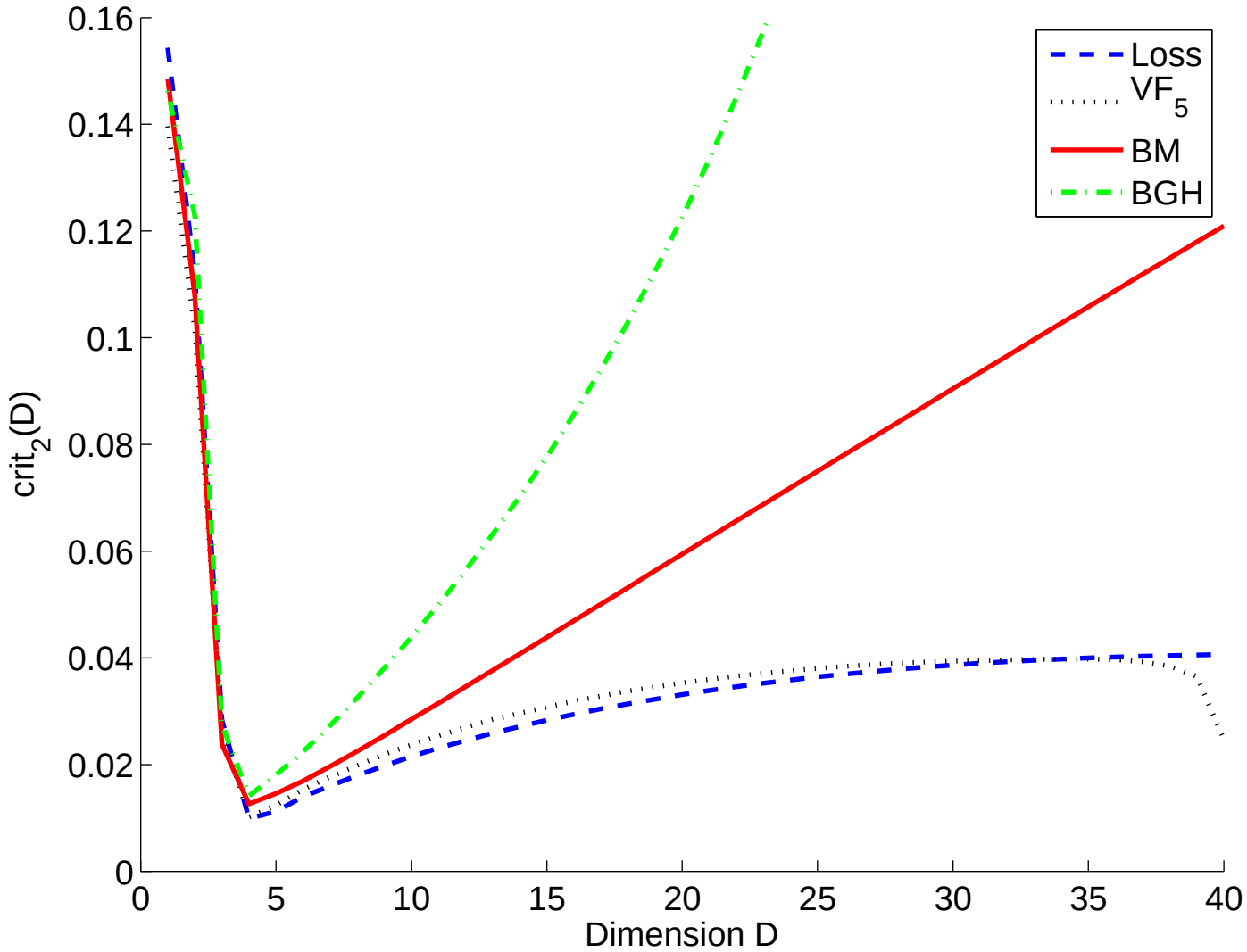
Homoscédastique : perte en fonction de D



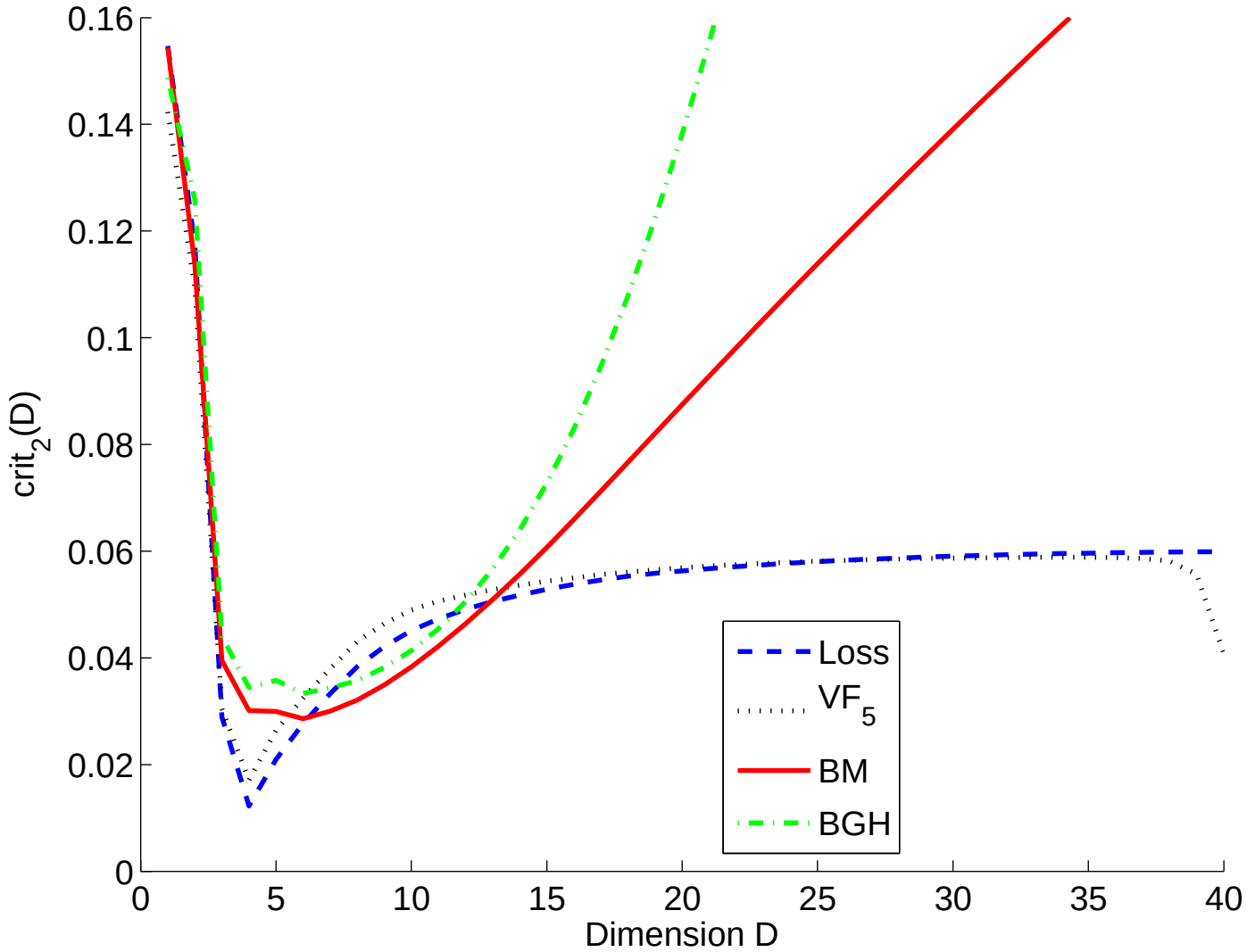
Hétéroscédastique : perte en fonction de D



Homoscédastique : perte estimée en fonction de D



Hétéroscédastique : perte estimée en fonction de D



Détection de ruptures en 2 étapes (A. & Celisse, 2010)

1. $\forall D \in \{1, \ldots, D_{\max}\}$, choisir

$$\widehat{m}(D) \in \operatorname{argmin}_{m \in \mathcal{M}_n,\, D_m = D} \{ \operatorname{crit}_1(m; (t_i, Y_i)_i) \}$$

Exemples pour crit_1 : risque empirique, ou estimateurs leave-p-out ou V-fold du risque (programmation dynamique)

Détection de ruptures en 2 étapes (A. & Celisse, 2010)

1. $\forall D \in \{1, \ldots, D_{\text{max}}\}$, choisir

$$\widehat{m}(D) \in \text{argmin}_{m \in \mathcal{M}_n,\, D_m = D} \{ \text{crit}_1(m; (t_i, Y_i)_i) \}$$

Exemples pour crit_1 : risque empirique, ou estimateurs leave-p-out ou V-fold du risque (programmation dynamique)

2. Sélectionner

$$\widehat{D} \in \text{argmin}_{D \in \{1, \ldots, D_{\text{max}}\}} \{ \text{crit}_2(D; (t_i, Y_i)_i; \text{crit}_1(\cdot)) \}$$

Exemples pour crit_2 : critère empirique pénalisé, estimateur V-fold du risque

Méthodes concurrentes

- $[\mathrm{Emp}, \mathrm{BM}]$: suppose $\sigma(\cdot) \equiv \sigma$

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

Méthodes concurrentes

- [Emp, BM] : suppose $\sigma(\cdot) \equiv \sigma$

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

- BGH (Baraud, Giraud & Huet 2009) : pénalité multiplicative, $\sigma(\cdot) \equiv \sigma$

$$\operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) \left[1 + \frac{\mathrm{pen}_{\mathrm{BGH}}(m)}{n - D_m} \right] \right\}$$

Méthodes concurrentes

- [Emp, BM] : suppose $\sigma(\cdot) \equiv \sigma$

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

- BGH (Baraud, Giraud & Huet 2009) : pénalité multiplicative, $\sigma(\cdot) \equiv \sigma$

$$\operatorname{argmin}_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) \left[1 + \frac{\mathrm{pen}_{\mathrm{BGH}}(m)}{n - D_m} \right] \right\}$$

- ZS (Zhang & Siegmund, 2007) : BIC modifié, $\sigma(\cdot) \equiv \sigma$

Méthodes concurrentes

- [Emp, BM] : suppose $\sigma(\cdot) \equiv \sigma$

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

- BGH (Baraud, Giraud & Huet 2009) : pénalité multiplicative, $\sigma(\cdot) \equiv \sigma$

$$\text{argmin}_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) \left[1 + \frac{\text{pen}_{\text{BGH}}(m)}{n - D_m} \right] \right\}$$

- ZS (Zhang & Siegmund, 2007) : BIC modifié, $\sigma(\cdot) \equiv \sigma$
- PML (Picard et al., 2005) : maximum de vraisemblance pénalisé, cherche les ruptures de (η, σ)

$$\text{argmin}_{m \in \mathcal{M}_n} \left\{ \sum_{\lambda \in m} n \widehat{p}_\lambda \log \left(\frac{1}{n \widehat{p}_\lambda} \sum_{t_i \in \lambda} \left(Y_i - \widehat{s}_m(t_i) \right)^2 \right) + \widehat{C}'' D_m \right\}$$

Simulations : comparaison à l'oracle (risque quadratique)

$$\frac{\mathbb{E}\left[\ell\left(s^\star, \widehat{s}_{\widehat{m}}\right)\right]}{\mathbb{E}\left[\inf_{m \in \mathcal{M}_n}\left\{\ell\left(s^\star, \widehat{s}_m\right)\right\}\right]}$$

$N = 10\,000$ échantillons

$\mathcal{L}(\varepsilon)$ $\sigma(\cdot)$ η	Gaussien homosc. s_2	Gaussien hétérosc. s_2	Gaussien hétérosc. s_3
$[\mathrm{Loo}, \mathrm{VF}_5]$	4.02 ± 0.02	4.95 ± 0.05	5.59 ± 0.02
$[\mathrm{Emp}, \mathrm{VF}_5]$	3.99 ± 0.02	5.62 ± 0.05	6.13 ± 0.02
$[\mathrm{Emp}, \mathrm{BM}]$	3.58 ± 0.02	9.25 ± 0.06	6.24 ± 0.02
BGH	3.52 ± 0.02	10.13 ± 0.07	6.31 ± 0.02
ZS	3.62 ± 0.02	6.50 ± 0.05	6.61 ± 0.02
PML	4.34 ± 0.02	2.73 ± 0.03	4.99 ± 0.03

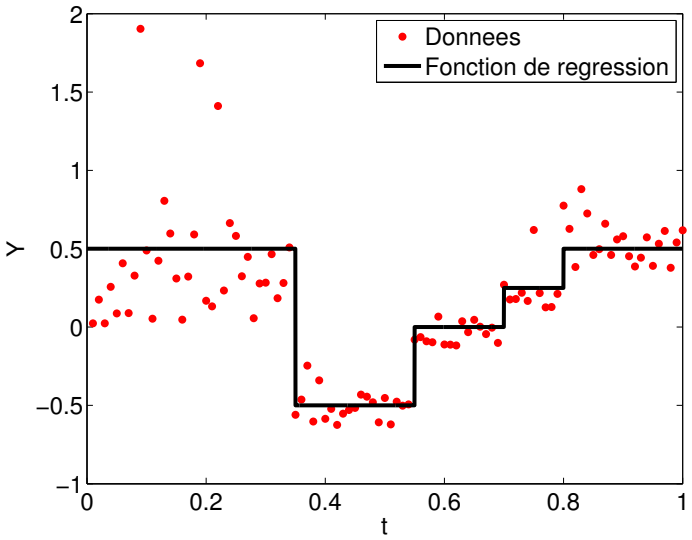
Simulations : comparaison à l'oracle (risque quadratique)

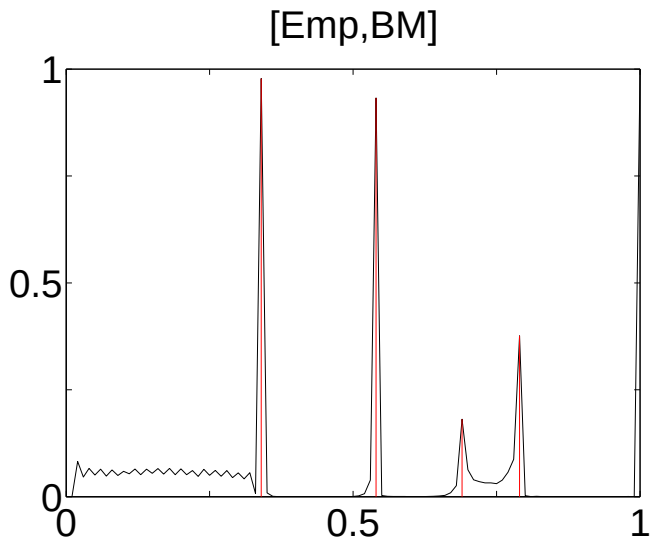
$$\frac{\mathbb{E}\left[\ell\left(s^\star, \widehat{s}_{\widehat{m}}\right)\right]}{\mathbb{E}\left[\inf_{m \in \mathcal{M}_n}\left\{\ell\left(s^\star, \widehat{s}_m\right)\right\}\right]}$$

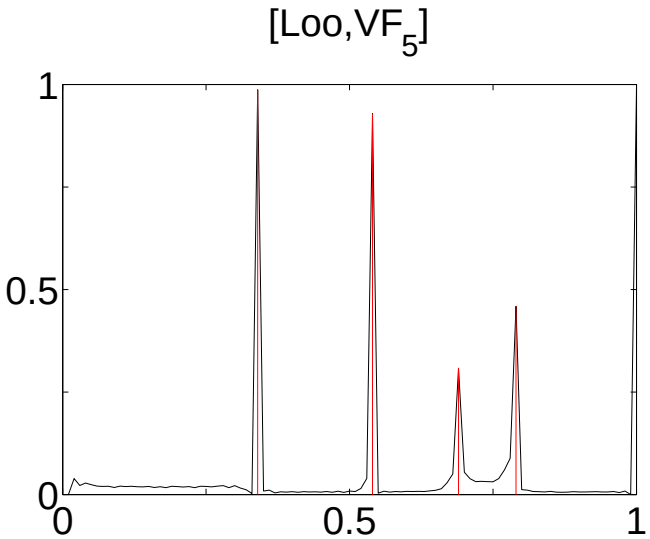
$N = 10\,000$ échantillons

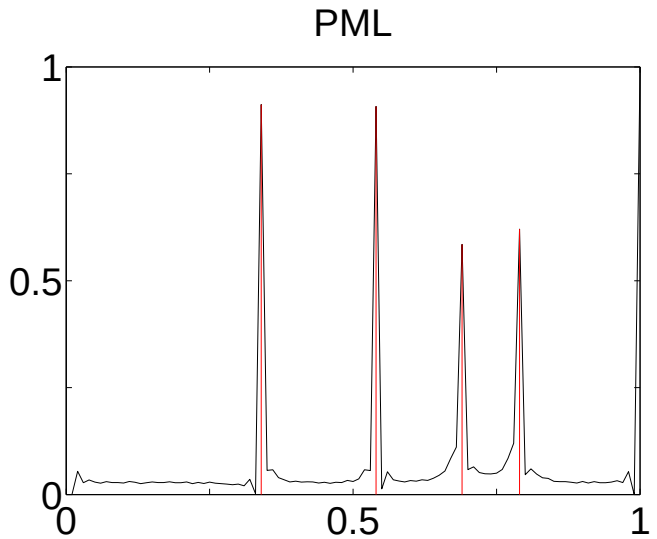
$\mathcal{L}(\varepsilon)$ $\sigma(\cdot)$ η	Gaussien homosc. s_2	Exponentiel hétérosc. s_2	Exponentiel hétérosc. s_3
$[\mathrm{Loo}, \mathrm{VF}_5]$	4.02 ± 0.02	4.47 ± 0.05	5.11 ± 0.03
$[\mathrm{Emp}, \mathrm{VF}_5]$	3.99 ± 0.02	5.98 ± 0.07	6.22 ± 0.04
$[\mathrm{Emp}, \mathrm{BM}]$	3.58 ± 0.02	10.81 ± 0.09	6.45 ± 0.04
BGH	3.52 ± 0.02	11.67 ± 0.09	6.42 ± 0.04
ZS	3.62 ± 0.02	9.34 ± 0.09	6.83 ± 0.04
PML	4.34 ± 0.02	5.04 ± 0.06	5.40 ± 0.03

Simulations : positions des ruptures

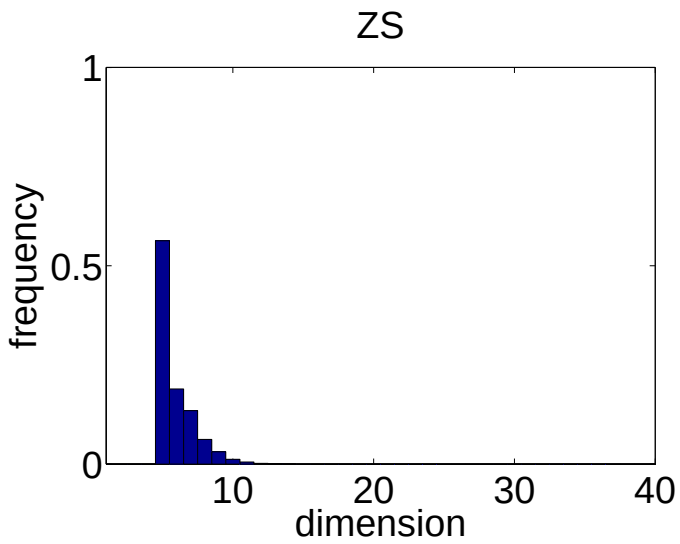


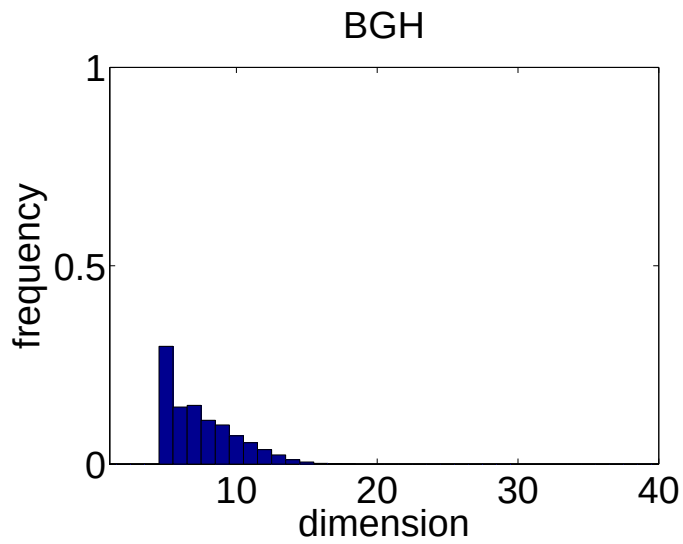


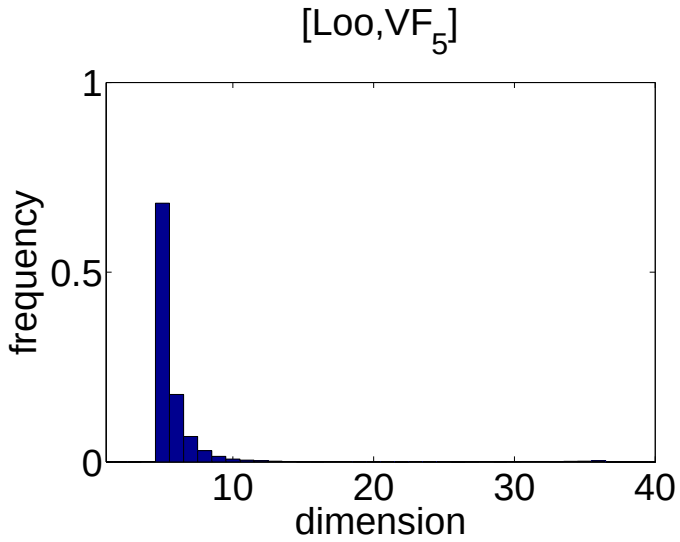


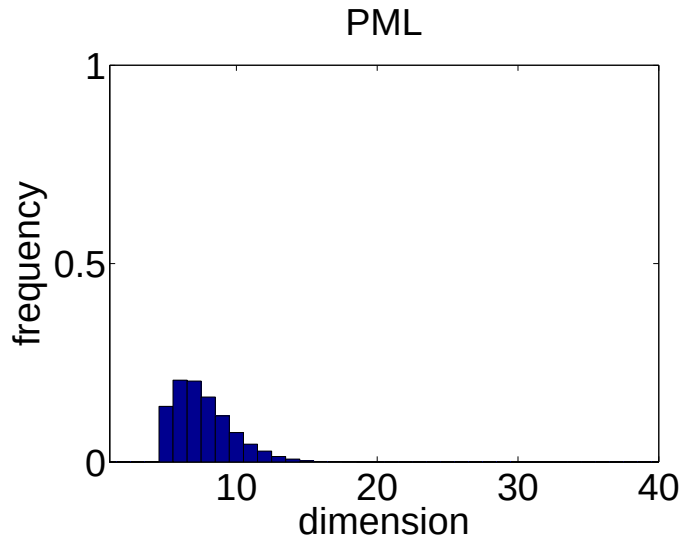


Simulations : dimensions sélectionnées ($D_0 = 5$)









Plan

Heuristique de rééchantillonnage (bootstrap, Efron 1979)

Monde réel : $P \xrightarrow{\text{échantillonnage}} P_n \Longrightarrow \widehat{s}_m$

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

Heuristique de rééchantillonnage (bootstrap, Efron 1979)

Monde réel : $P \xrightarrow{\text{échantillonnage}} P_n \Longrightarrow \widehat{s}_m$

Monde bootstrap : P_n

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

Heuristique de rééchantillonnage (bootstrap, Efron 1979)

Monde réel : $P \xrightarrow{\text{échantillonnage}} P_n \Longrightarrow \widehat{s}_m$

$\Big\downarrow$

Monde bootstrap : $P_n \xrightarrow{\text{rééchantillonnage}} P_n^W \Longrightarrow \widehat{s}_m^W$

$$(P - P_n)\gamma\left(\widehat{s}_m\right) = F(P, P_n) \rightsquigarrow F(P_n, P_n^W) = (P_n - P_n^W)\gamma\left(\widehat{s}_m^W\right)$$

Heuristique de rééchantillonnage (bootstrap, Efron 1979)

Monde réel : $P \xrightarrow{\text{échantillonnage}} P_n \Longrightarrow \widehat{s}_m$

$\Big\downarrow$

Monde bootstrap : $P_n \xrightarrow{\text{sous-échantillonnage}} P_n^W \Longrightarrow \widehat{s}_m^W$

$$(P - P_n)\gamma\left(\widehat{s}_m\right) = F(P, P_n) \rightsquigarrow F(P_n, P_n^W) = (P_n - P_n^W)\gamma\left(\widehat{s}_m^W\right)$$

V-fold : $$P_n^W = \frac{1}{n - \mathrm{Card}(B_J)} \sum_{i \notin B_J} \delta_{(X_i, Y_i)} \quad \text{avec } J \sim \mathcal{U}(1, \ldots, V)$$

Pénalités V-fold (A. 2008)

- Pénalité idéale :

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

- Pénalité V-fold (A., 2008) :

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

$$\widehat{s}_m^{(-B_j)} = \widehat{s}_m \left(D_n^{(-B_j)} \right)$$

- Modèle sélectionné :

$$\widehat{m} \in \mathrm{argmin}_{m \in \mathcal{M}_n} \{ P_n \gamma(\widehat{s}_m) + \mathrm{pen}(m) \}$$

Calcul d'espérances

Hypothèses :

$$\left.\begin{array}{r} \mathcal{B} = (B_j)_{1 \leq j \leq V} \text{ partition de } \{1, \ldots, n\} \\ \text{et} \quad \forall j \in \{1, \ldots, V\} \quad , \qquad \mathrm{Card}(B_j) = \dfrac{n}{V} \end{array}\right\} \qquad \textbf{(RegPart)}$$

$$\forall 1 \leq N \leq n \quad , \qquad \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m; D_N)\right] = \frac{\gamma_m}{N} \qquad \textbf{(Epenid)}$$

Calcul d'espérances

Hypothèses :

$$\left.\begin{array}{r} \mathcal{B} = (B_j)_{1 \leq j \leq V} \text{ partition de } \{1, \ldots, n\} \\ \text{et} \quad \forall j \in \{1, \ldots, V\} \ , \qquad \text{Card}(B_j) = \dfrac{n}{V} \end{array}\right\} \qquad \textbf{(RegPart)}$$

$$\forall 1 \leq N \leq n \ , \qquad \mathbb{E}\left[\text{pen}_{\text{id}}(m; D_N)\right] = \frac{\gamma_m}{N} \qquad \textbf{(Epenid)}$$

Proposition (A. 2011)

$$\mathbb{E}\left[\text{pen}_{\text{VF}}(m; D_n; C; \mathcal{B})\right] = \frac{C}{V-1} \mathbb{E}\left[\text{pen}_{\text{id}}(m; D_n)\right]$$

Concentration : hypothèses supplémentaires

Pour tout $N \in \{1, \dots, n\}$,

$$\mathbb{P}\left(|p_1(m; D_N) - \mathbb{E}\left[p_1(m; D_N) \right]| \leq w_N \mathbb{E}\left[p_1(m; D_N) \right] \right) \geq 1 - q_N \quad (\mathbf{C}p_1)$$

$$\mathbb{P}\left(|p_2(m; D_N) - \mathbb{E}\left[p_2(m; D_N) \right]| \leq w_N \mathbb{E}\left[p_2(m; D_N) \right] \right) \geq 1 - q_N \quad (\mathbf{C}p_2)$$

$$\left.\begin{array}{l} \exists S_m \subset \mathbb{S} \text{ t.q. } s_m^\star \in S_m \ , \ \widehat{s}_m(D_N) \in S_m \quad \text{p.s.} \\ \text{et} \quad \forall t \in S_m \ , \ \forall x \geq 0 \ , \\ \mathbb{P}\left(|\delta(t; D_N) - \delta(s^\star; D_N)| \leq \inf_{\eta \in]0,1]} \left\{ \eta \ell\left(s^\star, t\right) + \dfrac{K_\delta x}{\eta N} \right\} \right) \\ \geq 1 - 2e^{-x} \end{array}\right\} \quad (\mathbf{C}\delta)$$

$$p_1(m; D_N) = P\gamma\left(\widehat{s}_m(D_N)\right) - P\gamma\left(s_m^\star\right)$$

$$p_2(m; D_N) = P_N\gamma\left(s_m^\star\right) - P_N\gamma\left(\widehat{s}_m(D_N)\right)$$

$$\delta(t; D_N) = (P_N - P)\gamma\left(t\right)$$

Concentration : résultat

Proposition (A. 2011)

On suppose : $V \geq 2$, (**RegPart**), (**Epenid**), ($\mathbf{C}p_1$), ($\mathbf{C}p_2$) *et* ($\mathbf{C}\delta$) *avec* $\gamma_m \geq 0$, $K_\delta > 0$ *et* $(w_k), (q_k)$ *décroissantes positives.*

Alors, $\forall C > 0, x \geq 0$, *avec probabilité* $1 - 2V\left(q_{\frac{n(V-1)}{V}} + 2e^{-x}\right)$,

$\forall \eta \in]0,1]$,

$$\begin{aligned}
&\left|\mathrm{pen}_{\mathrm{VF}}(m; D_n; C; \mathcal{B}) - \mathbb{E}\left[\mathrm{pen}_{\mathrm{VF}}(m; D_n; C; \mathcal{B})\right] - \mathcal{Z}\right| \\
&\qquad \leq \frac{4C}{V}\left(\eta + 2w_{\frac{n(V-1)}{V}}\right)\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m; D_n)\right] \\
&\qquad\qquad + \frac{C}{V}\left(2\eta\ell(s^\star, s_m^\star) + \frac{4K_\delta x V}{\eta n}\right)
\end{aligned}$$

où $\mathcal{Z} = \mathcal{Z}(D_n; C; \mathcal{B}) = \frac{C}{V}\sum_{j=1}^{V}\left(\delta\left(s^\star; D_n^{(B_j)}\right) - \delta\left(s^\star; D_n^{(-B_j)}\right)\right)$

Inégalité-oracle pour la pénalisation "V-fold"

Théorème (A. 2008–2011)

Si de plus $w_k \to 0$, $C = V - 1$ *et* $\exists (\kappa_k)_{k \geq 1}$ *décroissante,*

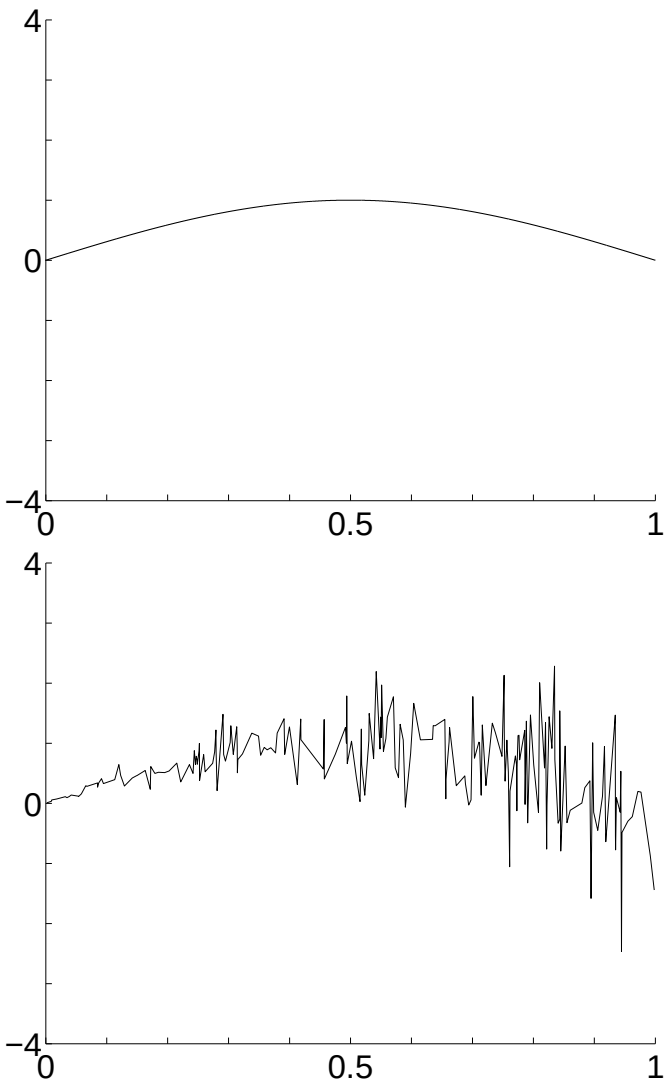
$$\forall N \geq 1 \ , \quad 0 \leq \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m; D_N)\right] \leq \kappa_N \mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\left(D_N\right)\right)\right]$$

Alors, avec probabilité $1 - L_1 V \,\mathrm{Card}(\mathcal{M}_n)(q_{\frac{n(V-1)}{V}} + e^{-x})$, *pour tout* $\eta_k \to 0$,

$$\ell\left(s^\star, \widehat{s}_{\widehat{m}_{\mathrm{pen}_{\mathrm{VF}}}(D_n)}\right) \leq \left[1 + L_2\left(\eta_n + \frac{1}{n} + w_{\frac{n(V-1)}{V}}\right)\right] \times \inf_{m \in \mathcal{M}_n} \left\{\ell\left(s^\star, \widehat{s}_m(D_n)\right)\right\} + \frac{L_3 K_\delta x V}{\eta_n n}$$

Exemple : régressogrammes *sous de bonnes hypothèses* ($\|Y\|_\infty \leq A$, $\sigma(\cdot) \geq \sigma_{\min} > 0$, ...)

Simulations : sin, $n = 200$, $\sigma(x) = x$, $\mathcal{M}_n = \mathcal{M}_n^{(\mathrm{reg},1/2)}$



Mallows	3.69 ± 0.07
2-fold	2.54 ± 0.05
5-fold	2.58 ± 0.06
10-fold	2.60 ± 0.06
20-fold	2.58 ± 0.06
leave-one-out	2.59 ± 0.06
pen 2-f	3.06 ± 0.07
pen 5-f	2.75 ± 0.06
pen 10-f	2.65 ± 0.06
pen Loo	2.59 ± 0.06
Mallows $\times 1.25$	3.17 ± 0.07
pen 2-f $\times 1.25$	2.75 ± 0.06
pen 5-f $\times 1.25$	2.38 ± 0.06
pen 10-f $\times 1.25$	2.28 ± 0.05
pen Loo $\times 1.25$	2.21 ± 0.05

Choix de V : estimation de densité (A. & Lerasle, 2011)

- Estimation de densité par moindres carrés : sous (**RegPart**),

$$\begin{aligned}
&\operatorname{var}\left(\left(\operatorname{pen}_{\mathrm{VF}}(m)-\operatorname{pen}_{\mathrm{id}}(m)\right)-\left(\operatorname{pen}_{\mathrm{VF}}(m')-\operatorname{pen}_{\mathrm{id}}(m')\right)\right)\\
&=\frac{8}{n^2}\left[1+\frac{1}{V-1}\right]F\left(m,m'\right)+\frac{4}{n}\operatorname{var}_P\left(s^\star_m-s^\star_{m'}\right)
\end{aligned}$$

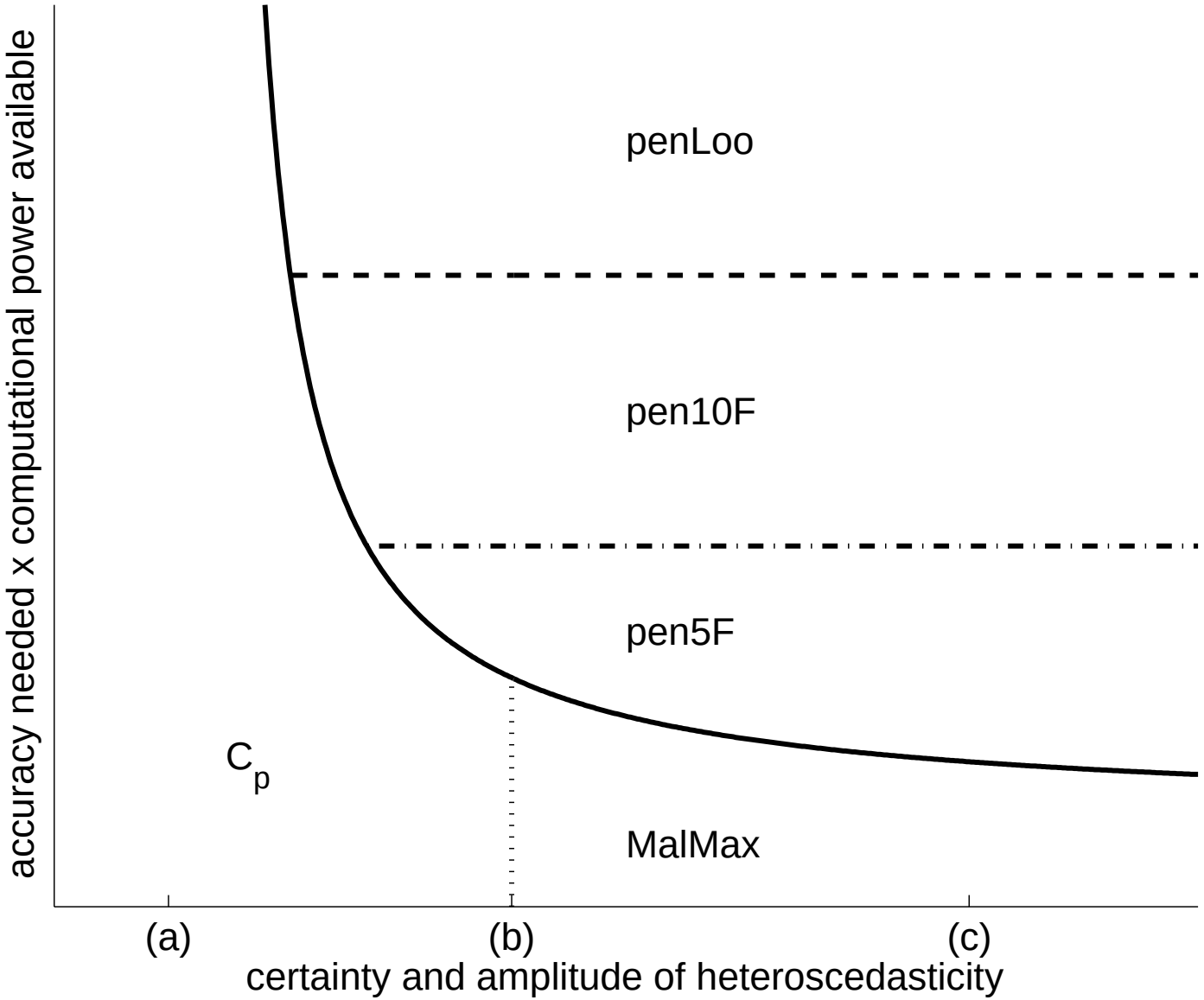
  avec $F\left(m,m'\right)>0$.

- Pour les histogrammes réguliers,

$$F\left(m,m'\right)\leq\left(D_m+D_{m'}\right)\left\|s^\star\right\|^2+2\left\|s^\star\right\|^4$$

Plan

Choix d'une procédure de sélection d'estimateurs



Bilan

- étudier des méthodes utilisées en pratique :
  - heuristiques de "coude" dans la L-curve, heuristique de pente
  - (pénalités par) rééchantillonnage
  - validation croisée

Bilan

- étudier des méthodes utilisées en pratique :
  - heuristiques de "coude" dans la L-curve, heuristique de pente
  - (pénalités par) rééchantillonnage
  - validation croisée
- utiliser la théorie pour proposer de nouvelles méthodes :
  - pénalités minimales pour les estimateurs linéaires
  - pénalités V-fold pour corriger le biais de la VFCV

Bilan

- étudier des méthodes utilisées en pratique :
  - heuristiques de "coude" dans la L-curve, heuristique de pente
  - (pénalités par) rééchantillonnage
  - validation croisée
- utiliser la théorie pour proposer de nouvelles méthodes :
  - pénalités minimales pour les estimateurs linéaires
  - pénalités V-fold pour corriger le biais de la VFCV
- résultats théoriques assez fins pour expliquer des différences observées en pratique :
  - comparaison des poids de rééchantillonnage
  - rôle de V pour les méthodes "V-fold"

Bilan

- étudier des méthodes utilisées en pratique :
  - heuristiques de "coude" dans la L-curve, heuristique de pente
  - (pénalités par) rééchantillonnage
  - validation croisée
- utiliser la théorie pour proposer de nouvelles méthodes :
  - pénalités minimales pour les estimateurs linéaires
  - pénalités V-fold pour corriger le biais de la VFCV
- résultats théoriques assez fins pour expliquer des différences observées en pratique :
  - comparaison des poids de rééchantillonnage
  - rôle de V pour les méthodes "V-fold"
- résultats non-asymptotiques

Problèmes ouverts

- étudier des méthodes utilisées en pratique :
  - validation croisée et pénalités par rééchantillonnage hors des cadres "jouet" (régressogrammes, estimation de densité par moindres carrés) ?
  - pénalités minimales avec un contraste différent des moindres carrés (SVM, Lasso, etc.) ?

Problèmes ouverts

- étudier des méthodes utilisées en pratique :
  - validation croisée et pénalités par rééchantillonnage hors des cadres "jouet" (régressogrammes, estimation de densité par moindres carrés) ?
  - pénalités minimales avec un contraste différent des moindres carrés (SVM, Lasso, etc.) ?
- résultats théoriques assez fins pour expliquer des différences observées en pratique :
  - choix d'un rééchantillonnage / d'une méthode de validation croisée ?
  - explication de la variabilité (non-systématique) du leave-one-out ?

Problèmes ouverts

- étudier des méthodes utilisées en pratique :
  - validation croisée et pénalités par rééchantillonnage hors des cadres "jouet" (régressogrammes, estimation de densité par moindres carrés) ?
  - pénalités minimales avec un contraste différent des moindres carrés (SVM, Lasso, etc.) ?
- résultats théoriques assez fins pour expliquer des différences observées en pratique :
  - choix d'un rééchantillonnage / d'une méthode de validation croisée ?
  - explication de la variabilité (non-systématique) du leave-one-out ?
- résultats non-asymptotiques :
  - phénomène de surpénalisation ?