Validation croisée

Sylvain Arlot (collaborations avec Alain Celisse, Matthieu Lerasle, Nelo Magalhães)

Laboratoire de Mathématiques d'Orsay, Université Paris-Sud

JES 2016, Fréjus
6 Octobre 2016

Plan

1. Problèmes
2. Définition
3. Estimation du risque
4. Sélection d'estimateurs
5. Conclusion

Rappel : problème de prévision

- Données : $D_n = (X_i, Y_i)_{1 \leqslant i \leqslant n}$
  $X_i \in \mathcal{X}$: variable explicative
  $Y_i \in \mathcal{Y}$: variable d'intérêt
  Hypothèse : $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n), \ldots$ i.i.d. $\sim P$

- Prédicteur : $f : \mathcal{X} \to \mathcal{Y}$
  ($\mathcal{F}$: ensemble des prédicteurs)
  Nouvelle observation $X_{n+1} \Rightarrow f(X_{n+1})$ « prévoit » Y_{n+1}

- Mesure de qualité : fonction de coût $c : \mathcal{Y} \times \mathcal{Y} \to [0, +\infty[$
  Risque (erreur de prévision) : $\mathcal{R}_P(f) = \mathbb{E}\Big[c\big(f(X), Y\big)\Big]$

- En résumé : avec D_n uniquement, on cherche un prédicteur $f \in \mathcal{F}$ tel que $\mathcal{R}_P(f)$ est minimal.

Deux problèmes

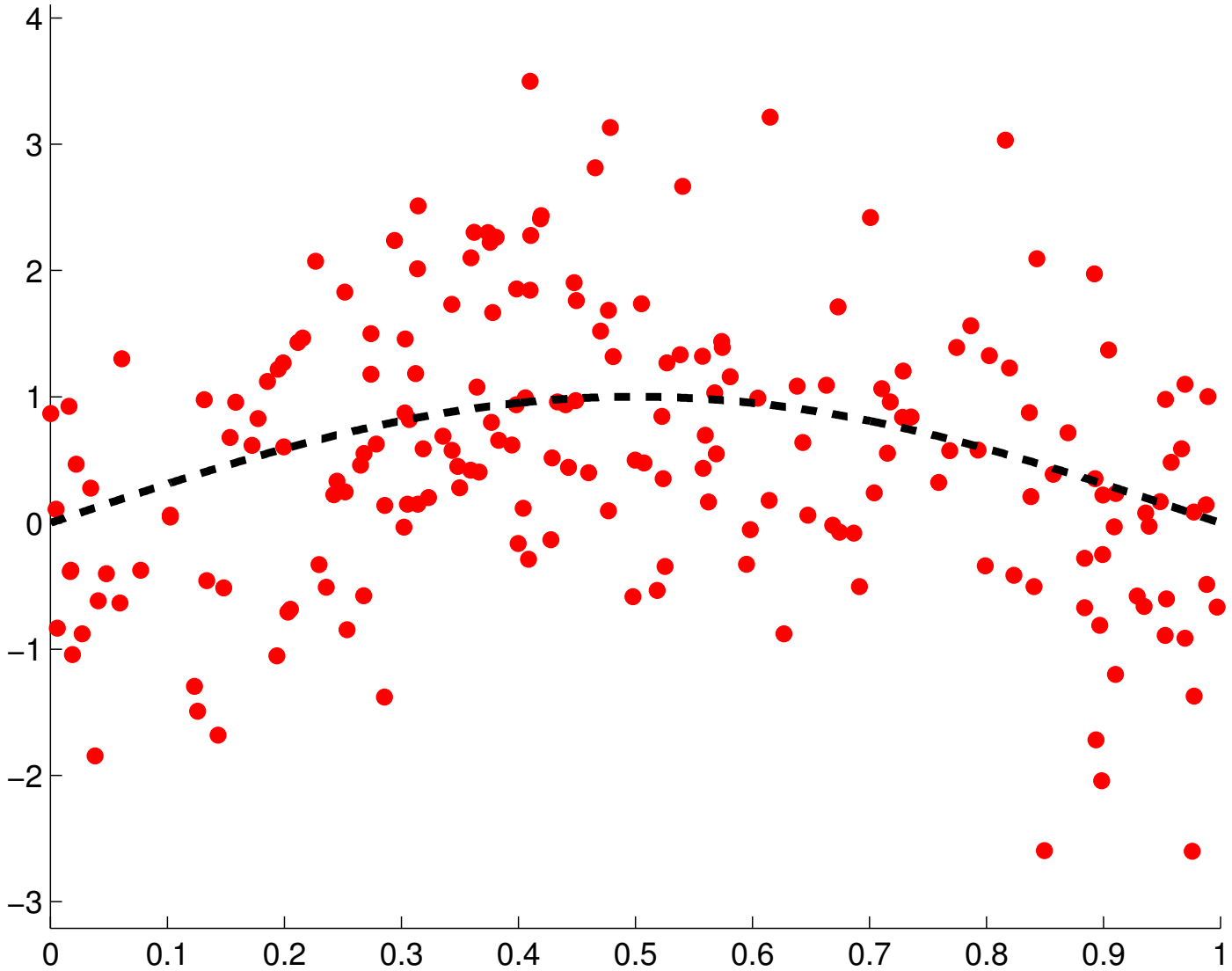
- Règle d'apprentissage $\widehat{f}$

$\Rightarrow$ estimation de son risque $\mathcal{R}_P\left(\widehat{f}(D_n)\right)$?

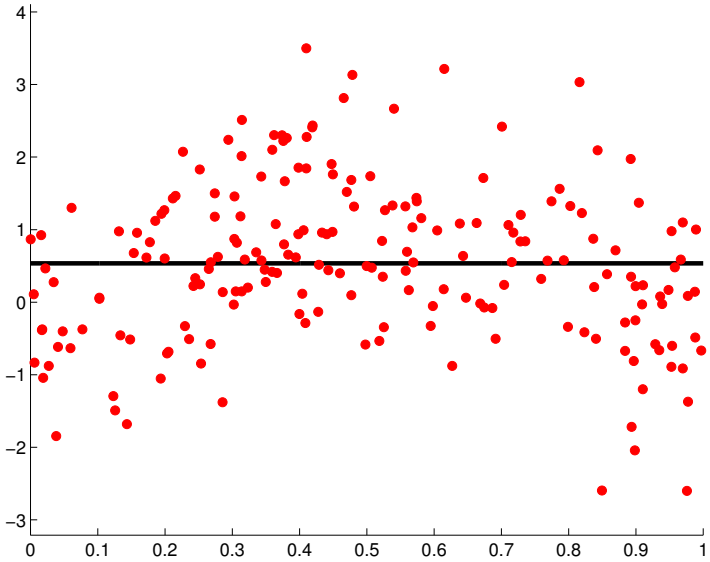
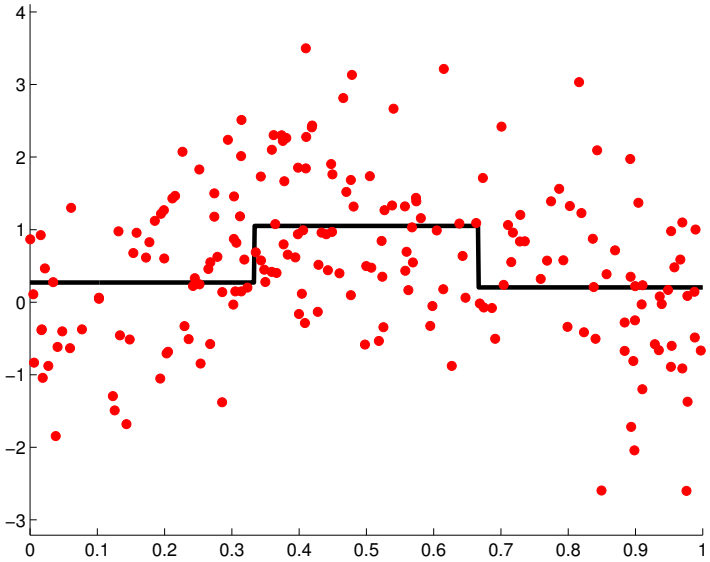
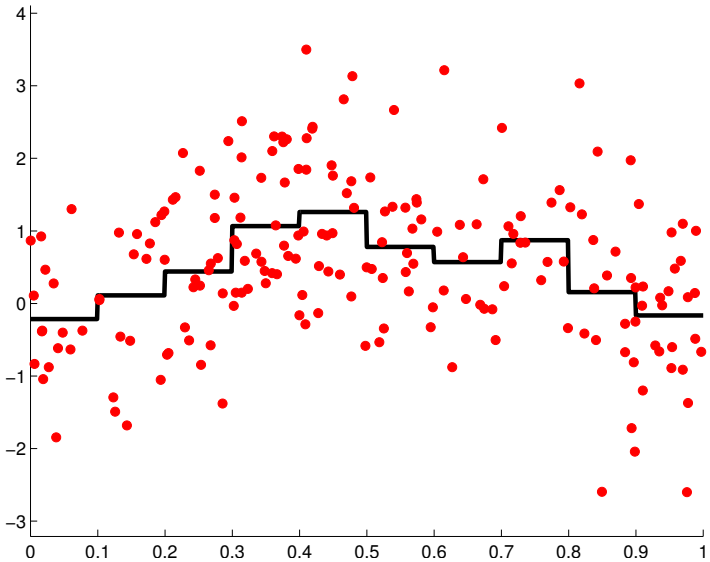
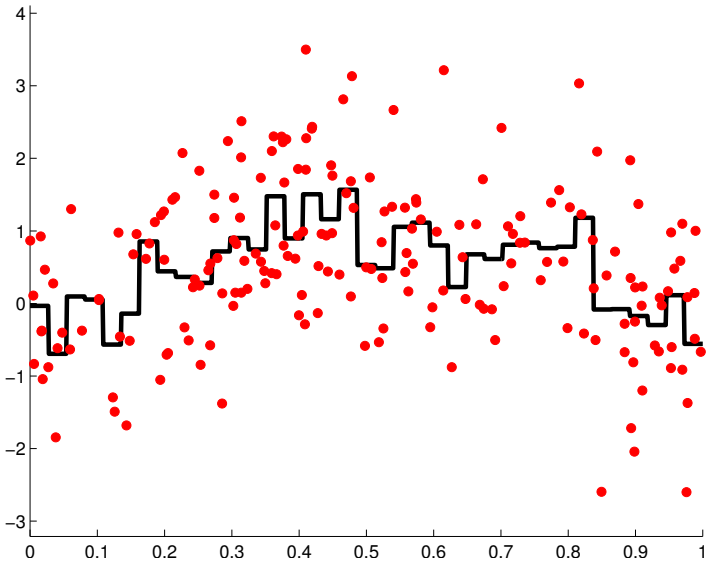
- Famille de règles d'apprentissage $(\widehat{f}_m)_{m \in \mathcal{M}}$

$\Rightarrow$ sélection d'un estimateur $\widehat{f}_{\widehat{m}(D_n)}(D_n)$?

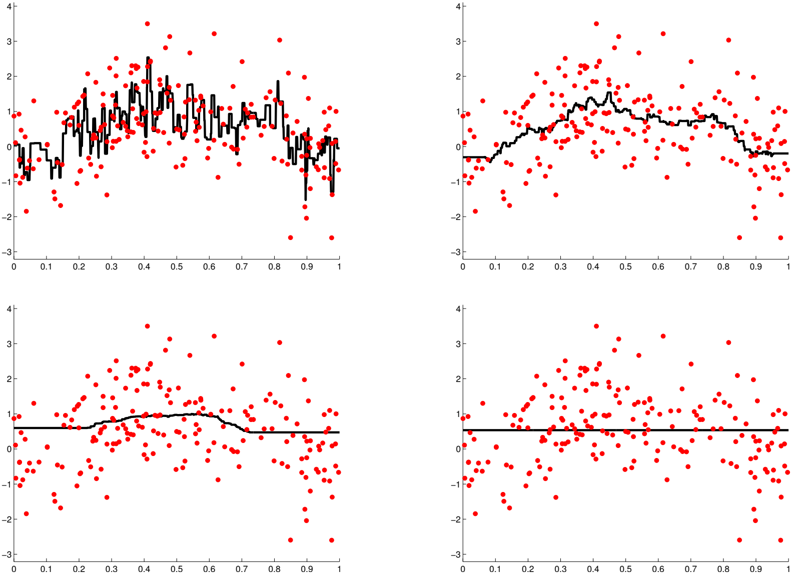
Exemple : régression



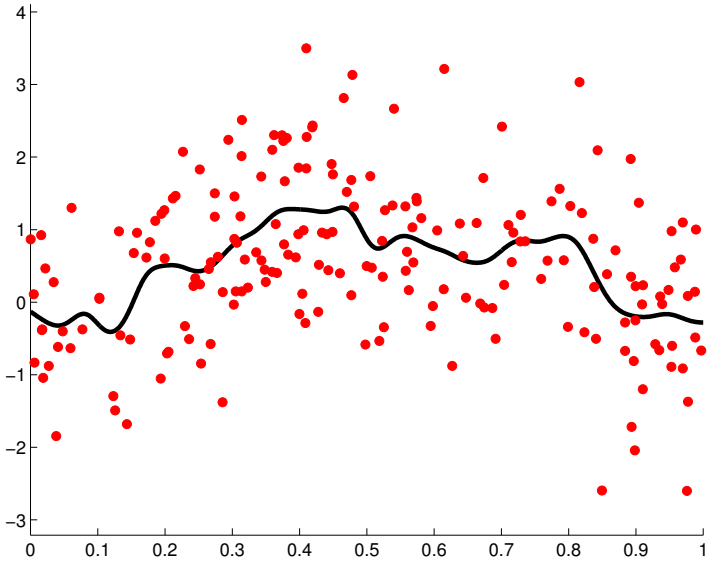
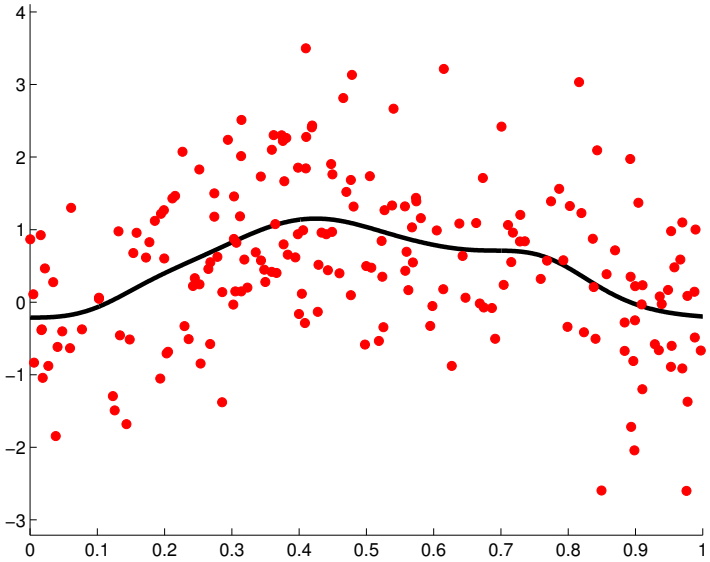
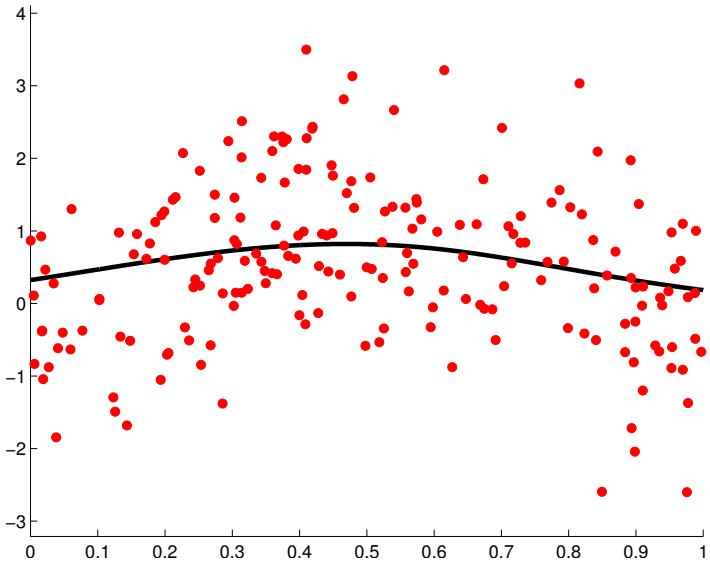
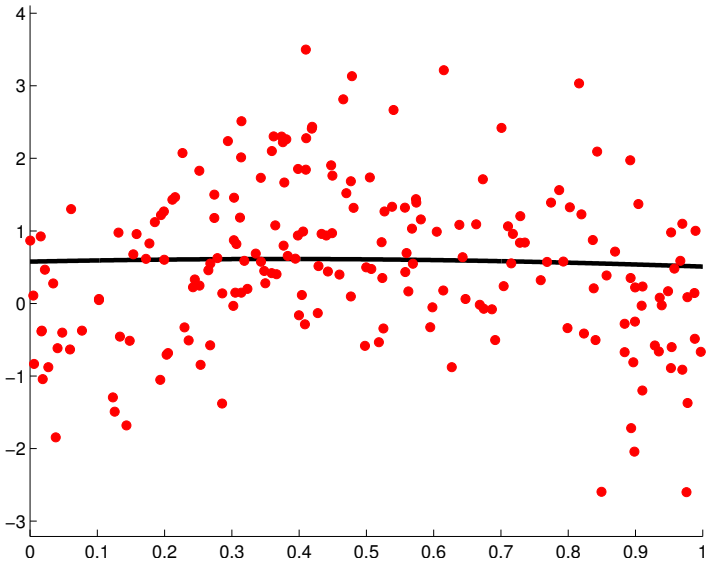
Sélection d'estimateurs (régression) : partitions cubiques

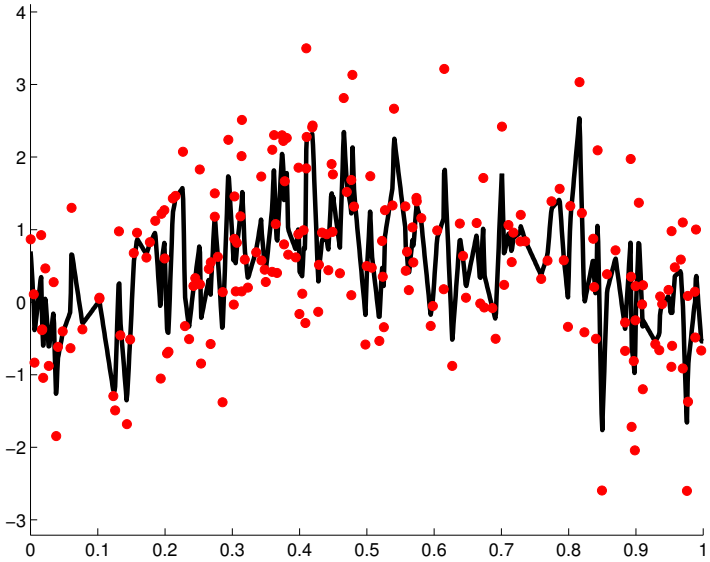
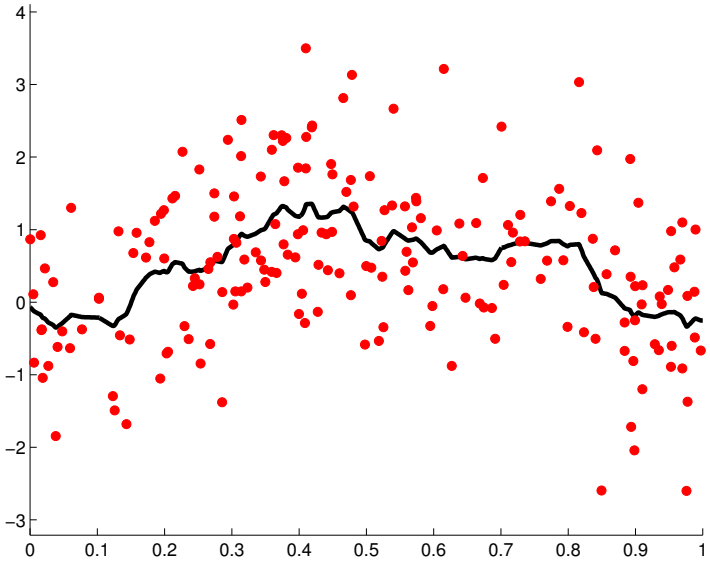
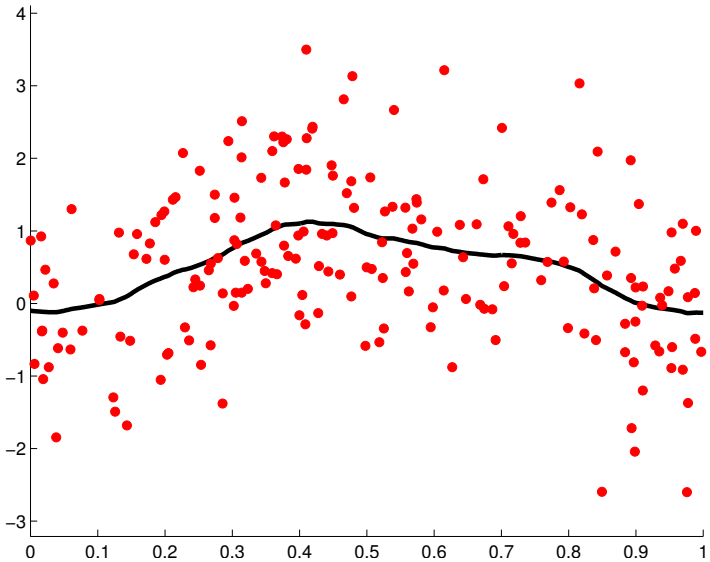
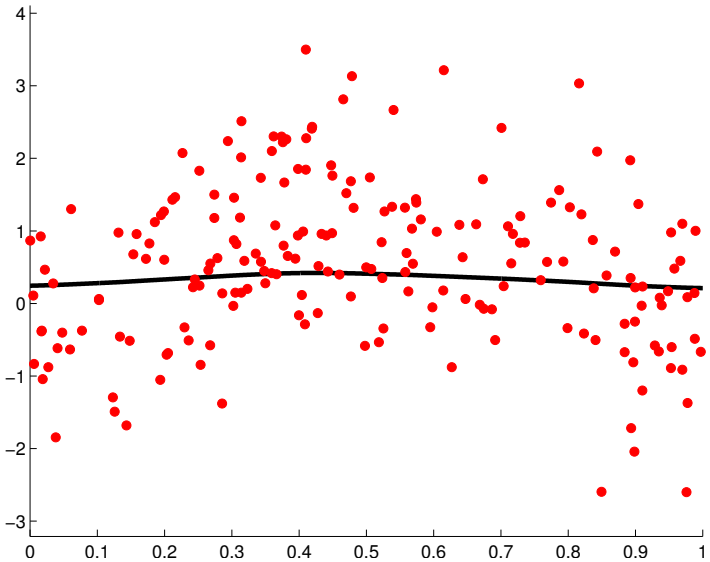
Sélection d’estimateurs (régression) : k plus proches voisins



Sélection d'estimateurs (régression) : Nadaraya-Watson

Sélection d'estimateurs (régression) : ridge à noyau

Sélection d'estimateurs

- Estimateur/Règle d'apprentissage : $\widehat{f} : D_n \mapsto \widehat{f}(D_n) \in \mathcal{F}$
- Exemple : estimateur des moindres carrés sur $S_m \subset \mathcal{F}$:

$$\widehat{f}_m \in \operatorname*{argmin}_{f \in S_m} \left\{ \widehat{\mathcal{R}}_n(f) \right\} \quad \text{où} \quad \widehat{\mathcal{R}}_n(f) := \frac{1}{n} \sum_{i=1}^{n} c(f(X_i), Y_i)$$

Exemples de modèles S_m : histogrammes, e. v.$\{\varphi_1, \ldots, \varphi_D\}$

Sélection d'estimateurs

- Estimateur/Règle d'apprentissage : $\widehat{f} : D_n \mapsto \widehat{f}(D_n) \in \mathcal{F}$
- Exemple : estimateur des moindres carrés sur $S_m \subset \mathcal{F}$:

$$\widehat{f}_m \in \operatorname*{argmin}_{f \in S_m} \left\{ \widehat{\mathcal{R}}_n(f) \right\} \quad \text{où} \quad \widehat{\mathcal{R}}_n(f) := \frac{1}{n} \sum_{i=1}^{n} c(f(X_i), Y_i)$$

Exemples de modèles S_m : histogrammes, e. v.$\{\varphi_1, \ldots, \varphi_D\}$

- Famille d'estimateurs $(\widehat{f}_m)_{m \in \mathcal{M}} \Rightarrow$ choisir $\widehat{m} = \widehat{m}(D_n)$?

Sélection d'estimateurs

- Estimateur/Règle d'apprentissage : $\widehat{f} : D_n \mapsto \widehat{f}(D_n) \in \mathcal{F}$
- Exemple : estimateur des moindres carrés sur $S_m \subset \mathcal{F}$:

$$\widehat{f}_m \in \operatorname*{argmin}_{f \in S_m} \left\{ \widehat{\mathcal{R}}_n(f) \right\} \quad \text{où} \quad \widehat{\mathcal{R}}_n(f) := \frac{1}{n} \sum_{i=1}^{n} c(f(X_i), Y_i)$$

Exemples de modèles S_m : histogrammes, e. v. $\{\varphi_1, \ldots, \varphi_D\}$

- Famille d'estimateurs $(\widehat{f}_m)_{m \in \mathcal{M}} \Rightarrow$ choisir $\widehat{m} = \widehat{m}(D_n)$?
- Exemples :
  - choix de modèles
  - « calibration » d'hyperparamètres (choix de k ou d'une distance pour k-ppv, choix du paramètre de régularisation, choix d'un noyau, etc.)
  - choix entre des méthodes de natures différentes
    ex. : k-ppv ou splines de lissage ?

Sélection d'estimateurs : deux objectifs

- Estimation : minimiser le risque de l'estimateur final, i.e.,
  Inégalité oracle (en espérance ou avec grande probabilité) :

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

Sélection d'estimateurs : deux objectifs

- Estimation : minimiser le risque de l'estimateur final, i.e., Inégalité oracle (en espérance ou avec grande probabilité) :

$$\ell(f^\star, \widehat{f}_{\widehat{m}}) \leqslant C \inf_{m \in \mathcal{M}} \{\ell(f^\star, \widehat{f}_m)\} + R_n$$

- Identification : choisir le « meilleur » estimateur/modèle asymptotiquement, en supposant qu'il est bien défini, i.e., Consistance en sélection :

$$\mathbb{P}(\widehat{m}(D_n) = m^\star) \xrightarrow[n \to \infty]{} 1 \,.$$

Équivalent à l'estimation dans le cadre paramétrique.

Sélection d'estimateurs : deux objectifs

- Estimation : minimiser le risque de l'estimateur final, i.e., Inégalité oracle (en espérance ou avec grande probabilité) :

$$\ell(f^\star, \widehat{f}_{\widehat{m}}) \leqslant C \inf_{m \in \mathcal{M}} \{\ell(f^\star, \widehat{f}_m)\} + R_n$$

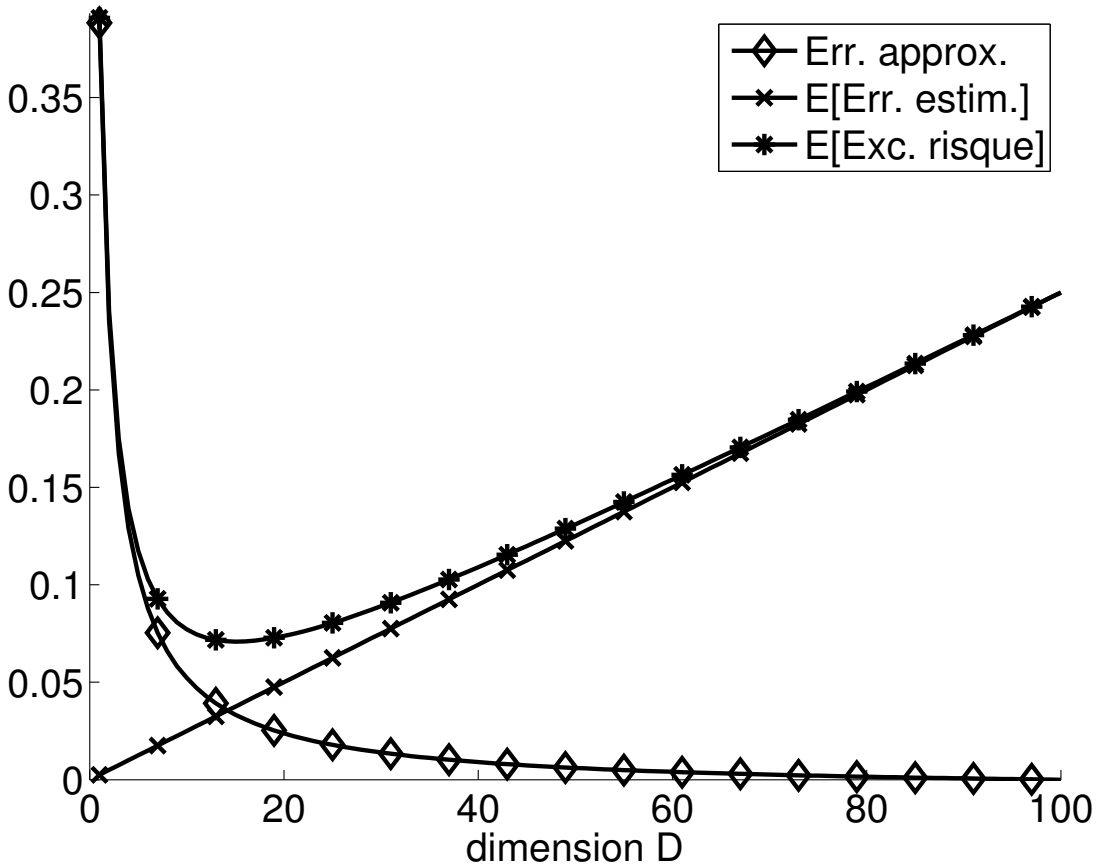
- Identification : choisir le « meilleur » estimateur/modèle asymptotiquement, en supposant qu'il est bien défini, i.e., Consistance en sélection :

$$\mathbb{P}(\widehat{m}(D_n) = m^\star) \xrightarrow[n \to \infty]{} 1\,.$$

Équivalent à l'estimation dans le cadre paramétrique.

- Double objectif avec une seule procédure (dilemme AIC-BIC) ? Non en général (Yang, 2005). Parfois possible.

Enjeux du problème (rappel)



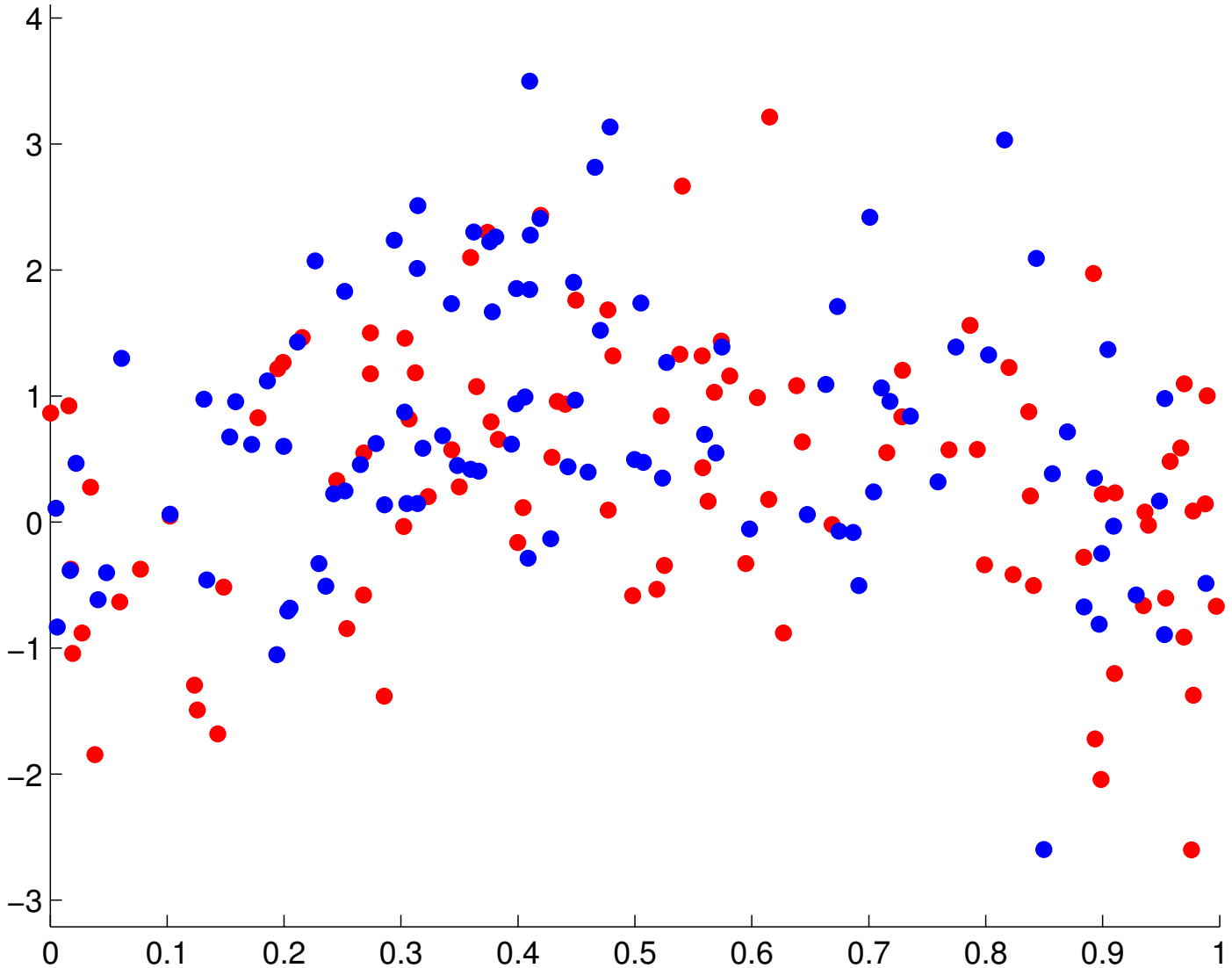
Sous-apprentissage

Sur-apprentissage

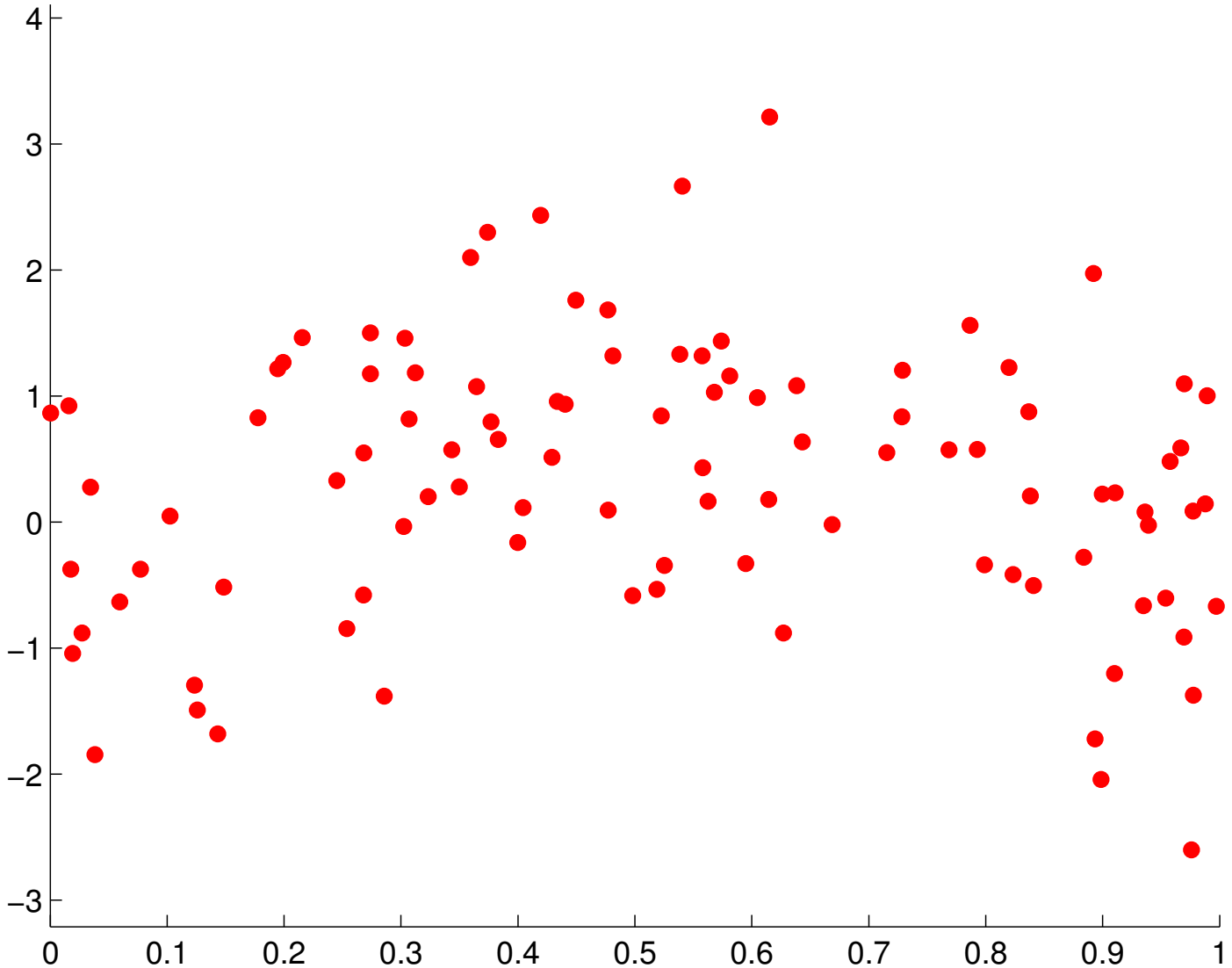
Plan

1. Problèmes
2. Définition
3. Estimation du risque
4. Sélection d'estimateurs
5. Conclusion

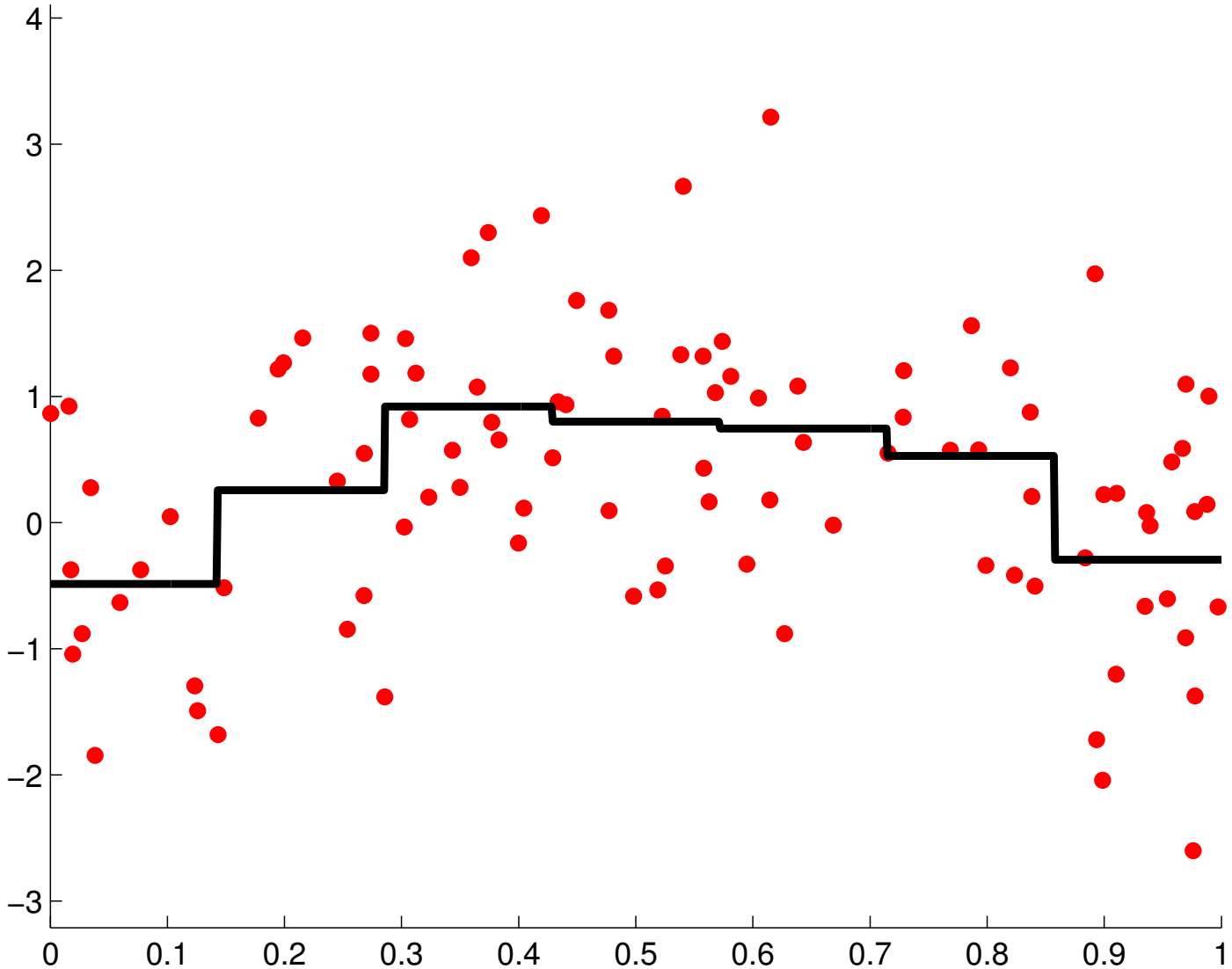
Principe de la validation simple



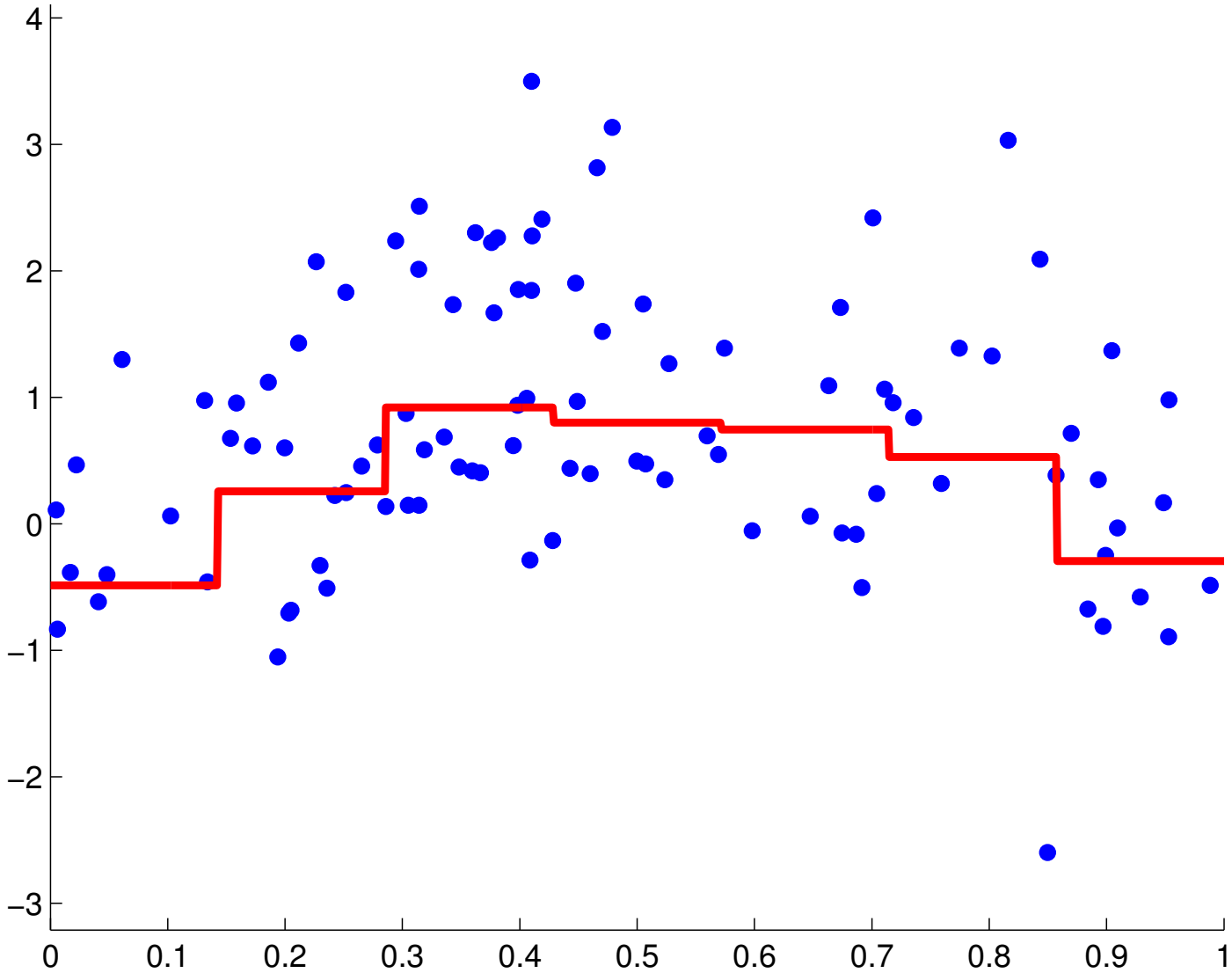
Principe de la validation : échantillon d'entraînement



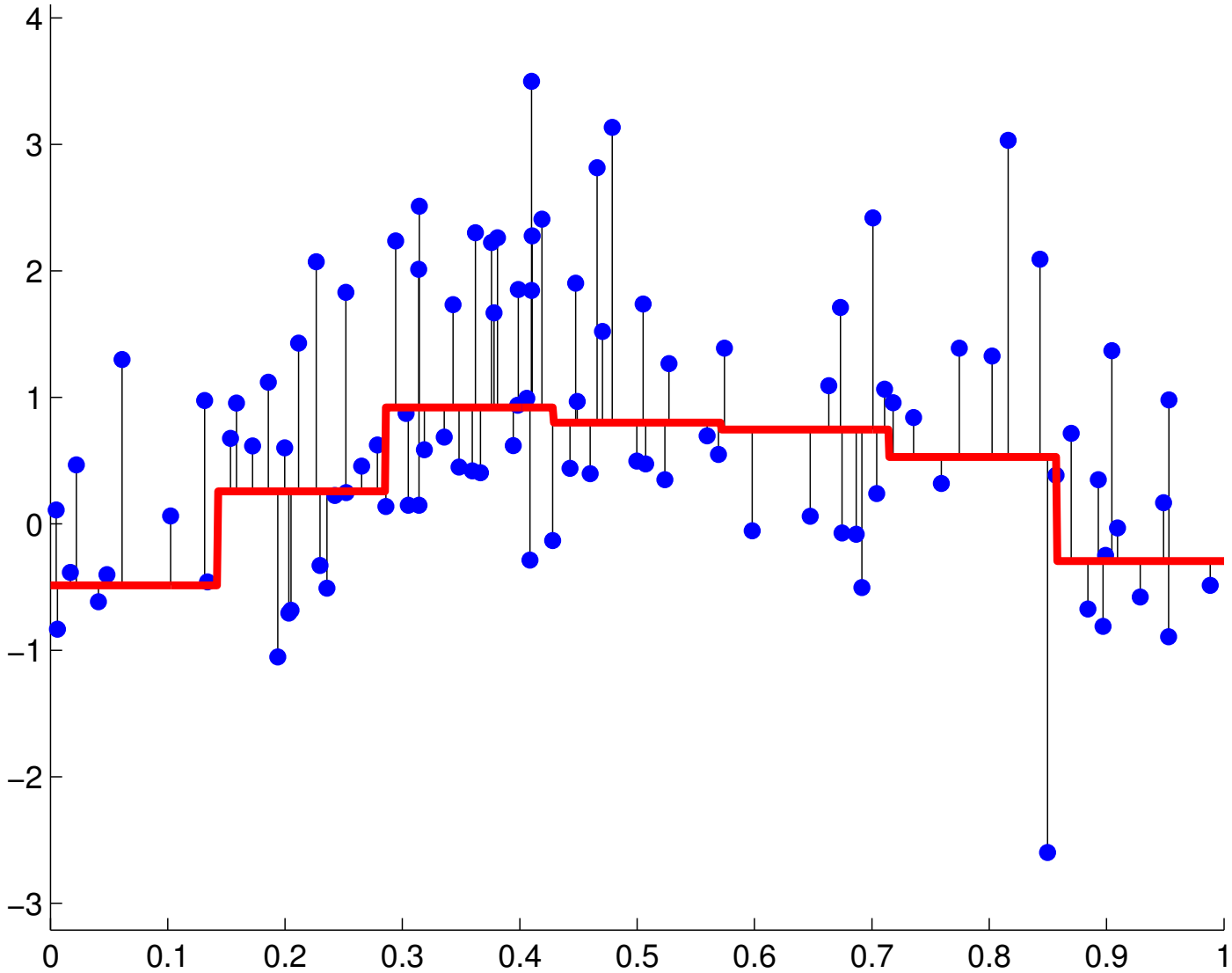
Principe de la validation : échantillon d'entraînement



Principe de la validation : échantillon de validation



Principe de la validation : échantillon de validation



Validation croisée

$\underbrace{(X_1, Y_1), \ldots, (X_{n_e}, Y_{n_e})}$ $\qquad$ $\underbrace{(X_{n_e+1}, Y_{n_e+1}), \ldots, (X_n, Y_n)}$

Entraînement $D_n^E \Rightarrow \widehat{f}_m(D_n^E)$ $\qquad$ Validation $D_n^{E^c} \Rightarrow$ évaluer le risque

Validation croisée

$\underbrace{(X_1, Y_1), \ldots, (X_{n_e}, Y_{n_e})}$ $\qquad$ $\underbrace{(X_{n_e+1}, Y_{n_e+1}), \ldots, (X_n, Y_n)}$

Entraînement $D_n^E \Rightarrow \widehat{f}_m(D_n^E)$ $\qquad$ Validation $D_n^{E^c} \Rightarrow$ évaluer le risque

- estimateur « hold-out » du risque :

$$\widehat{\mathcal{R}}^{\text{val}}(\widehat{f}_m; D_n; E) = \widehat{\mathcal{R}}_n^{E^c}(\widehat{f}_m(D_n^E)) = \frac{1}{\text{Card}(E^c)} \sum_{i \in E^c} c(\widehat{f}_m(D_n^E; X_i), Y_i)$$

Validation croisée

$\underbrace{(X_1, Y_1), \ldots, (X_{n_e}, Y_{n_e})}$ $\qquad$ $\underbrace{(X_{n_e+1}, Y_{n_e+1}), \ldots, (X_n, Y_n)}$

Entraînement $D_n^E \Rightarrow \widehat{f}_m(D_n^E)$ $\qquad$ Validation $D_n^{E^c} \Rightarrow$ évaluer le risque

- estimateur « hold-out » du risque :

$$\widehat{\mathcal{R}}^{\mathrm{val}}(\widehat{f}_m; D_n; E) = \widehat{\mathcal{R}}_n^{E^c}(\widehat{f}_m(D_n^E)) = \frac{1}{\mathrm{Card}(E^c)} \sum_{i \in E^c} c(\widehat{f}_m(D_n^E; X_i), Y_i)$$

- validation croisée : moyenne d'estimateurs « hold-out »

$$\widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{f}_m; D_n; (E_j)_{1 \leqslant j \leqslant V}) = \frac{1}{V} \sum_{j=1}^{V} \widehat{\mathcal{R}}^{\mathrm{val}}(\widehat{f}_m; D_n; E_j)$$

Validation croisée

$\underbrace{(X_1, Y_1), \ldots, (X_{n_e}, Y_{n_e})}$ $\qquad$ $\underbrace{(X_{n_e+1}, Y_{n_e+1}), \ldots, (X_n, Y_n)}$

Entraînement $D_n^E \Rightarrow \widehat{f}_m(D_n^E)$ $\qquad$ Validation $D_n^{E^c} \Rightarrow$ évaluer le risque

- estimateur « hold-out » du risque :

$$\widehat{\mathcal{R}}^{\mathrm{val}}(\widehat{f}_m; D_n; E) = \widehat{\mathcal{R}}_n^{E^c}(\widehat{f}_m(D_n^E)) = \frac{1}{\mathrm{Card}(E^c)} \sum_{i \in E^c} c(\widehat{f}_m(D_n^E; X_i), Y_i)$$

- validation croisée : moyenne d'estimateurs « hold-out »

$$\widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{f}_m; D_n; (E_j)_{1 \leqslant j \leqslant V}) = \frac{1}{V} \sum_{j=1}^{V} \widehat{\mathcal{R}}^{\mathrm{val}}(\widehat{f}_m; D_n; E_j)$$

- sélection d'estimateurs :

$$\widehat{m}^{\mathrm{vc}}(D_n; (E_j)_{1 \leqslant j \leqslant V}) \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{f}_m; D_n; (E_j)_{1 \leqslant j \leqslant V}) \right\}$$

Validation croisée : exemples

- Méthodes exhaustives : tous les sous-ensembles de taille n_e
  $\Rightarrow$ leave-one-out ($n_e = n - 1$)

$$\widehat{\mathcal{R}}^{\text{loo}}(\widehat{f}_m; D_n) = \widehat{\mathcal{R}}^{\text{vc}}\left(\widehat{f}_m; D_n; (\{j\}^c)_{1 \leqslant j \leqslant n}\right) = \frac{1}{n} \sum_{j=1}^{n} c(\widehat{f}_m(D_n^{(-j)}; X_j), Y_j)$$

  $\Rightarrow$ leave-p-out ($n_e = n - p$)

Validation croisée : exemples

- Méthodes exhaustives : tous les sous-ensembles de taille n_e
  $\Rightarrow$ leave-one-out ($n_e = n - 1$)

$$\widehat{\mathcal{R}}^{\mathrm{loo}}(\widehat{f}_m; D_n) = \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m; D_n; (\{j\}^c)_{1 \leqslant j \leqslant n}\right) = \frac{1}{n} \sum_{j=1}^{n} c(\widehat{f}_m(D_n^{(-j)}; X_j), Y_j)$$

  $\Rightarrow$ leave-p-out ($n_e = n - p$)

- Validation croisée « V-fold » : $\mathcal{B} = (B_j)_{1 \leqslant j \leqslant V}$ partition de $\{1, \ldots, n\}$

$$\Rightarrow \widehat{\mathcal{R}}^{\mathrm{vf}}(\widehat{f}_m; D_n; (B_j)_{1 \leqslant j \leqslant V}) = \frac{1}{V} \sum_{j=1}^{V} \widehat{\mathcal{R}}_n^{B_j}\left(\widehat{f}_m\left(D_n^{B_j^c}\right)\right)$$

Validation croisée : exemples

- Méthodes exhaustives : tous les sous-ensembles de taille n_e
  $\Rightarrow$ leave-one-out ($n_e = n - 1$)

$$\widehat{\mathcal{R}}^{\mathrm{loo}}(\widehat{f}_m; D_n) = \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m; D_n; (\{j\}^c)_{1 \leqslant j \leqslant n}\right) = \frac{1}{n} \sum_{j=1}^{n} c(\widehat{f}_m(D_n^{(-j)}; X_j), Y_j)$$

  $\Rightarrow$ leave-p-out ($n_e = n - p$)

- Validation croisée « V-fold » : $\mathcal{B} = (B_j)_{1 \leqslant j \leqslant V}$ partition de $\{1, \ldots, n\}$

$$\Rightarrow \widehat{\mathcal{R}}^{\mathrm{vf}}(\widehat{f}_m; D_n; (B_j)_{1 \leqslant j \leqslant V}) = \frac{1}{V} \sum_{j=1}^{V} \widehat{\mathcal{R}}_n^{B_j}\left(\widehat{f}_m\left(D_n^{B_j^c}\right)\right)$$

- Validation croisée Monte-Carlo / Apprentissage Test Répété :

$E_1, \ldots, E_V$ i.i.d. uniforme

Deux hypothèses

Dans cet exposé :

$$(E_j)_{1 \leqslant j \leqslant V} \text{ est indépendante de } D_n \qquad (\textbf{Ind})$$

$$\text{Card}(E_1) = \text{Card}(E_2) = \cdots = \text{Card}(E_V) = n_e \qquad (\textbf{Reg})$$

Pour la VC « V-fold » : $n_e = \dfrac{n(V-1)}{V}$

Plan

1. Problèmes
2. Définition
3. Estimation du risque
4. Sélection d'estimateurs
5. Conclusion

Biais

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\right)\right] = \frac{1}{V}\sum_{j=1}^{V}\mathbb{E}\left[\widehat{\mathcal{R}}_n^{E_j^c}\left(\widehat{f}_m(D_n^{E_j})\right)\right]$$

Biais

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\right)\right] = \frac{1}{V}\sum_{j=1}^{V}\mathbb{E}\left[\widehat{\mathcal{R}}_n^{E_j^c}\left(\widehat{f}_m(D_n^{E_j})\right)\right]$$

$$= \frac{1}{V}\sum_{j=1}^{V}\mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m(D_n^{E_j})\right)\right] \qquad \textbf{(Ind)}$$

Biais

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\right)\right] = \frac{1}{V}\sum_{j=1}^{V}\mathbb{E}\left[\widehat{\mathcal{R}}_n^{E_j^c}\left(\widehat{f}_m\left(D_n^{E_j}\right)\right)\right]$$

$$= \frac{1}{V}\sum_{j=1}^{V}\mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m\left(D_n^{E_j}\right)\right)\right] \qquad \textbf{(Ind)}$$

$$= \mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m(D_{n_e})\right)\right] \qquad \textbf{(Reg)}$$

Biais

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{VC}}\left(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\right)\right] = \frac{1}{V}\sum_{j=1}^{V}\mathbb{E}\left[\widehat{\mathcal{R}}_n^{E_j^c}\left(\widehat{f}_m(D_n^{E_j})\right)\right]$$

$$= \frac{1}{V}\sum_{j=1}^{V}\mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m(D_n^{E_j})\right)\right] \qquad \textbf{(Ind)}$$

$$= \mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m(D_{n_e})\right)\right] \qquad \textbf{(Reg)}$$

Biais pour l'estimation du risque :

$$\mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m(D_{n_e})\right)\right] - \mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m(D_n)\right)\right]$$

$\Rightarrow$ tout dépend de $n \to \mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m(D_n)\right)\right]$

Biais

$$\begin{aligned}
\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{VC}}\left(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\right)\right] &= \frac{1}{V}\sum_{j=1}^{V}\mathbb{E}\left[\widehat{\mathcal{R}}_n^{E_j^c}\left(\widehat{f}_m(D_n^{E_j})\right)\right] \\
&= \frac{1}{V}\sum_{j=1}^{V}\mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m(D_n^{E_j})\right)\right] \qquad \textbf{(Ind)} \\
&= \mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m(D_{n_e})\right)\right] \qquad \textbf{(Reg)}
\end{aligned}$$

Biais pour l'estimation du risque :

$$\mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m(D_{n_e})\right)\right] - \mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m(D_n)\right)\right]$$

$\Rightarrow$ tout dépend de $n \to \mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m(D_n)\right)\right]$

Attention ! $D_n \to \widehat{f}_m(D_n)$ doit être fixée avant d'avoir vu une seule observation ; sinon, on a un biais encore plus fort.

Biais de la validation croisée : exemple générique

Hypothèse :

$$\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big] = \alpha(m) + \frac{\beta(m)}{n}$$

(e.g., moindres carrés/ridge/k-ppv en régression, moindres carrés/noyaux en estimation de densité).

Biais de la validation croisée : exemple générique

Hypothèse :

$$\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big] = \alpha(m) + \frac{\beta(m)}{n}$$

(e.g., moindres carrés/ridge/k-ppv en régression, moindres carrés/noyaux en estimation de densité).

$$\Rightarrow \qquad \mathbb{E}\Big[\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\Big] = \alpha(m) + \frac{n}{n_e}\frac{\beta(m)}{n}$$

Biais de la validation croisée : exemple générique

Hypothèse :

$$\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big] = \alpha(m) + \frac{\beta(m)}{n}$$

(e.g., moindres carrés/ridge/k-ppv en régression, moindres carrés/noyaux en estimation de densité).

$$\Rightarrow \qquad \mathbb{E}\Big[\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\Big] = \alpha(m) + \frac{n}{n_e}\frac{\beta(m)}{n}$$

$\Rightarrow$ Biais :

- fonction décroissante de n_e,
- minimal pour $n_e = n - 1$,
- négligeable si $n_e \sim n$.

Biais de la validation croisée : exemple générique

Hypothèse :

$$\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big] = \alpha(m) + \frac{\beta(m)}{n}$$

(e.g., moindres carrés/ridge/k-ppv en régression, moindres carrés/noyaux en estimation de densité).

$$\Rightarrow \qquad \mathbb{E}\Big[\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\Big] = \alpha(m) + \frac{n}{n_e}\frac{\beta(m)}{n}$$

$\Rightarrow$ Biais :

- fonction décroissante de n_e,
- minimal pour $n_e = n - 1$,
- négligeable si $n_e \sim n$.

$\Rightarrow$ V-fold : le biais diminue quand V augmente, disparaît quand $V \to +\infty$.

Correction du biais

Définition (Burman, 1989) :

$$\widehat{\mathcal{R}}^{\mathrm{vc-cor}}\left(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\right) = \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\right) + \widehat{\mathcal{R}}_n\left(\widehat{f}_m(D_n)\right) - \frac{1}{V}\sum_{j=1}^{V} \widehat{\mathcal{R}}_n\left(\widehat{f}_m(D_n^{E_j})\right)$$

Correction du biais

Définition (Burman, 1989) :

$$\widehat{\mathcal{R}}^{\mathrm{vc-cor}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big) = \widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big) + \widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big) - \frac{1}{V}\sum_{j=1}^{V} \widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n^{E_j})\big)$$

Proposition (3.1)

Hypothèses : (**Ind**) *et* $\exists \gamma(m)$, $\forall n \geqslant 1$,

$$\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big) - \widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big)\Big] = \frac{\gamma(m)}{n}$$

Alors :

$$\mathbb{E}\Big[\widehat{\mathcal{R}}^{\mathrm{vc-cor}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\Big] = \mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big]$$

Variance

Proposition (3.2)

On suppose (**Ind**) *et* (**Reg**)*. Alors :*

$$\begin{aligned}\operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{val}}\left(\widehat{f}_m; D_n; E_0\right)\right) &\geqslant \operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\right)\right) \\ &\geqslant \operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{lpo}}\left(\widehat{f}_m; D_n; n - n_e\right)\right)\end{aligned}$$

Variance

Proposition (3.2)

On suppose (**Ind**) *et* (**Reg**). *Alors* :

$$\begin{aligned}\operatorname{var}\big(\widehat{\mathcal{R}}^{\mathrm{val}}(\widehat{f}_m; D_n; E_0)\big) &\geqslant \operatorname{var}\Big(\widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V})\Big)\\ &\geqslant \operatorname{var}\big(\widehat{\mathcal{R}}^{\mathrm{lpo}}(\widehat{f}_m; D_n; n-n_e)\big)\end{aligned}$$

Proposition (3.3)

On suppose (**Ind**) *et* (**Reg**).
Pour la VC Monte-Carlo (E_j *iid uniformes*), *on a* :

$$\begin{aligned}\operatorname{var}\Big(\widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V})\Big) &= \operatorname{var}\big(\widehat{\mathcal{R}}^{\mathrm{lpo}}(\widehat{f}_m; D_n; n-n_e)\big)\\ &+\frac{1}{V}\Big[\underbrace{\operatorname{var}\big(\widehat{\mathcal{R}}^{\mathrm{val}}(\widehat{f}_m; D_n; E_1)\big) - \operatorname{var}\big(\widehat{\mathcal{R}}^{\mathrm{lpo}}(\widehat{f}_m; D_n; n-n_e)\big)}_{\textit{variance de permutation}}\Big]\end{aligned}$$

Variance : estimation de densité L^2 (A. & Lerasle 2012)

Histogramme régulier de pas $h_m > 0$:

$$\mathrm{var}\left(\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m\right)\right) = \frac{C_1(n, V, n_e)}{n^2}\mathcal{W}_1(h_m, P) + \frac{C_2(n, V, n_e)}{n}\mathcal{W}_2(h_m, P)$$

Variance : estimation de densité L^2 (A. & Lerasle 2012)

Histogramme régulier de pas $h_m > 0$:

$$\mathrm{var}\left(\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m\right)\right) = \frac{C_1(n, V, n_e)}{n^2}\mathcal{W}_1(h_m, P) + \frac{C_2(n, V, n_e)}{n}\mathcal{W}_2(h_m, P)$$

Si $n \to +\infty$, au premier ordre :

	$C_1(n, V, n_e)$	$C_2(n, V, n_e)$
V-fold, $V \to \infty$	$1 + \frac{4}{V}$	1

Variance : estimation de densité L^2 (A. & Lerasle 2012)

Histogramme régulier de pas $h_m > 0$:

$$\mathrm{var}\left(\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m\right)\right) = \frac{C_1(n, V, n_e)}{n^2}\mathcal{W}_1(h_m, P) + \frac{C_2(n, V, n_e)}{n}\mathcal{W}_2(h_m, P)$$

Si $n \to +\infty$, au premier ordre :

	$C_1(n, V, n_e)$	$C_2(n, V, n_e)$
V-fold, $V \to \infty$	$1 + \frac{4}{V}$	1
hold-out, $n_e \sim n\tau$	$\frac{1}{\tau^2} + \frac{2}{\tau(1-\tau)} > 11$	$\frac{1}{1-\tau}$
leave-p-out, $n_e \sim n\tau$	1	1

Variance : estimation de densité L^2 (A. & Lerasle 2012)

Histogramme régulier de pas $h_m > 0$:

$$\mathrm{var}\big(\widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{f}_m)\big) = \frac{C_1(n,V,n_e)}{n^2}\mathcal{W}_1(h_m,P) + \frac{C_2(n,V,n_e)}{n}\mathcal{W}_2(h_m,P)$$

Si $n \to +\infty$, au premier ordre :

	$C_1(n,V,n_e)$	$C_2(n,V,n_e)$
V-fold, $V \to \infty$	$1 + \frac{4}{V}$	1
hold-out, $n_e \sim n\tau$	$\frac{1}{\tau^2} + \frac{2}{\tau(1-\tau)} > 11$	$\frac{1}{1-\tau}$
leave-p-out, $n_e \sim n\tau$	1	1
Monte-Carlo $n_e = \frac{n(V-1)}{V}$ V-fold	> 1 si $V \geqslant 3$	$2 - \frac{1}{V}$

Plan

1. Problèmes
2. Définition
3. Estimation du risque
4. Sélection d'estimateurs
5. Conclusion

Estimation du risque $\neq$ sélection d'estimateurs

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m\right) \right\} \quad \text{vs.} \quad m^\star \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{R}_P\left(\widehat{f}_m(D_n)\right) \right\}$$

- Pour tout Z (déterministe ou aléatoire),

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m\right) + Z \right\}$$

$\Rightarrow$ biais et variance inutiles.

Estimation du risque $\neq$ sélection d'estimateurs

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m \right) \right\} \quad \text{vs.} \quad m^\star \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{R}_P\left(\widehat{f}_m(D_n) \right) \right\}$$

- Pour tout Z (déterministe ou aléatoire),

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m \right) + Z \right\}$$

  $\Rightarrow$ biais et variance inutiles.

- Classement parfait parmi $(\widehat{f}_m)_{m \in \mathcal{M}} \Leftrightarrow \forall m, m' \in \mathcal{M}$,

$$\mathrm{signe}\left(\widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{f}_m) - \widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{f}_{m'}) \right) = \mathrm{signe}\left(\mathcal{R}_P(\widehat{f}_m) - \mathcal{R}_P(\widehat{f}_{m'}) \right)$$

Estimation du risque $\neq$ sélection d'estimateurs

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m \right) \right\} \quad \text{vs.} \quad m^\star \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{R}_P\left(\widehat{f}_m(D_n) \right) \right\}$$

- Pour tout Z (déterministe ou aléatoire),

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m \right) + Z \right\}$$

  $\Rightarrow$ biais et variance inutiles.

- Classement parfait parmi $(\widehat{f}_m)_{m \in \mathcal{M}} \Leftrightarrow \forall m, m' \in \mathcal{M}$,

$$\operatorname{signe}\left(\widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{f}_m) - \widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{f}_{m'})\right) = \operatorname{signe}\left(\mathcal{R}_P(\widehat{f}_m) - \mathcal{R}_P(\widehat{f}_{m'})\right)$$

$\Rightarrow$ $\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m\right) - \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_{m'}\right)\right]$ doit être du bon signe (heuristique d'estimation sans biais du risque : AIC, C_p, leave-one-out...)

Estimation du risque $\neq$ sélection d'estimateurs

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m\right) \right\} \quad \text{vs.} \quad m^\star \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{R}_P\left(\widehat{f}_m(D_n)\right) \right\}$$

- Pour tout Z (déterministe ou aléatoire),

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m\right) + Z \right\}$$

  $\Rightarrow$ biais et variance inutiles.

- Classement parfait parmi $(\widehat{f}_m)_{m \in \mathcal{M}} \Leftrightarrow \forall m, m' \in \mathcal{M}$,

$$\operatorname{signe}\left(\widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{f}_m) - \widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{f}_{m'})\right) = \operatorname{signe}\left(\mathcal{R}_P(\widehat{f}_m) - \mathcal{R}_P(\widehat{f}_{m'})\right)$$

$\Rightarrow$ $\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m\right) - \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_{m'}\right)\right]$ doit être du bon signe (heuristique d'estimation sans biais du risque : AIC, C_p, leave-one-out...)

$\Rightarrow$ $\operatorname{var}\left(\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m\right) - \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_{m'}\right)\right)$ doit être minimal (heuristique détaillée : A. & Lerasle 2012)

Analyse au premier ordre : espérance

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m\right) \right\} \quad \text{vs.} \quad m^\star \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{R}_P\left(\widehat{f}_m(D_n)\right) \right\}$$

- Hypothèse :

$$\mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m(D_n)\right)\right] = \alpha(m) + \frac{\beta(m)}{n}$$

(e.g., moindres carrés/ridge/k-ppv en régression, moindres carrés/noyaux en estimation de densité).

Analyse au premier ordre : espérance

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m\right) \right\} \quad \text{vs.} \quad m^\star \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathcal{R}_P\left(\widehat{f}_m(D_n)\right) \right\}$$

- Hypothèse :

$$\mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m(D_n)\right)\right] = \alpha(m) + \frac{\beta(m)}{n}$$

(e.g., moindres carrés/ridge/k-ppv en régression, moindres carrés/noyaux en estimation de densité).

- Quantités clés :

$$\mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m\right) - \mathcal{R}_P\left(\widehat{f}_{m'}\right)\right] = \alpha(m) - \alpha(m') + \frac{\beta(m) - \beta(m')}{n}$$

$$\mathbb{E}\left[\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m\right) - \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_{m'}\right)\right] = \alpha(m) - \alpha(m') + \frac{n}{n_e}\frac{\beta(m) - \beta(m')}{n}$$

$\Rightarrow$ VC favorise m de complexité $\beta(m)$ plus petite, d'autant plus que n_e diminue.

VC pour l'estimation : grandes lignes ($\mathcal{M}$ « petite »)

- Au premier ordre, le biais détermine la performance de :
  leave-p-out, VC V-fold,
  VC Monte-Carlo si $B \gg n^2$
  ou si n_v assez grand (y compris le hold-out)
- VC se comporte comme

$$\underset{m \in \mathcal{M}}{\operatorname{argmin}} \left\{ \mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m(D_{n_e}) \right) \right] \right\}$$

VC pour l'estimation : grandes lignes ($\mathcal{M}$ « petite »)

- Au premier ordre, le biais détermine la performance de :
  leave-p-out, VC V-fold,
  VC Monte-Carlo si $B \gg n^2$
  ou si n_v assez grand (y compris le hold-out)
- VC se comporte comme

$$\operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \mathbb{E}\left[\mathcal{R}_P\left(\widehat{f}_m(D_{n_e}) \right) \right] \right\}$$

$\Rightarrow$ optimalité au premier ordre si $n_e \sim n$

$\Rightarrow$ sous-optimal sinon
e.g., VC V-fold avec V fixe.

- Résultats théoriques en régression et estimation de densité par moindres carrés, au moins.

Analyse au second ordre (1) : variance

$$\mathrm{var}\left(\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\right) - \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_{m'}; D_n; (E_j)_{1\leqslant j\leqslant V}\right)\right)$$

Analyse au second ordre (1) : variance

$$\mathrm{var}\left(\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\right) - \widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_{m'}; D_n; (E_j)_{1\leqslant j\leqslant V}\right)\right)$$

Estimation de densité L^2, histogrammes réguliers :

$$= \frac{C_1(n,V,n_e)}{n^2}\mathcal{W}_1(h_m, h'_m P) + \frac{C_2(n,V,n_e)}{n}\mathcal{W}_2(h_m, h'_m, P)$$

Différences avec la variance du critère VC seul ?

C_1, C_2 identiques, mais $\dfrac{\mathcal{W}_1}{n^2} \approx \dfrac{\mathcal{W}_2}{n}$ pour les m, m' « qui comptent »

$\Rightarrow$ même classement entre méthodes, quelques différences quantitatives

Analyse au second ordre (2) : pénalisation

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big) + \underbrace{\mathrm{pen}(m; D_n)}_{\text{pénalité}} \right\}$$

Pénalité idéale :

$$\mathrm{pen}_{\mathrm{id}}(m; D_n) := \mathcal{R}_P\big(\widehat{f}_m(D_n)\big) - \widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big)$$

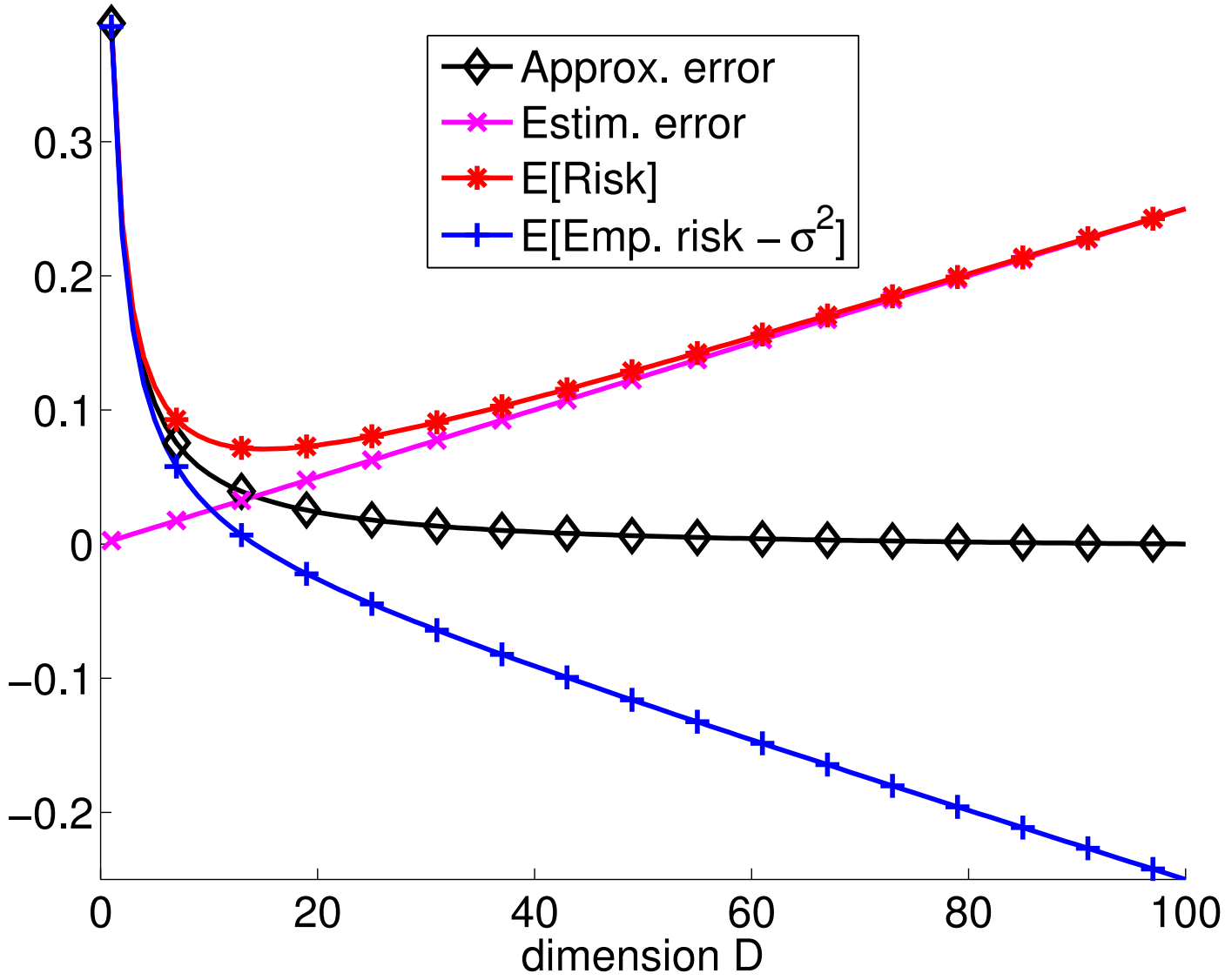
Estimation sans biais du risque :

$$\mathbb{E}\big[\mathrm{pen}(m; D_n)\big] \approx \mathbb{E}\Big[\mathrm{pen}_{\mathrm{id}}(m; D_n)\Big] \qquad \text{(à translation près)}$$

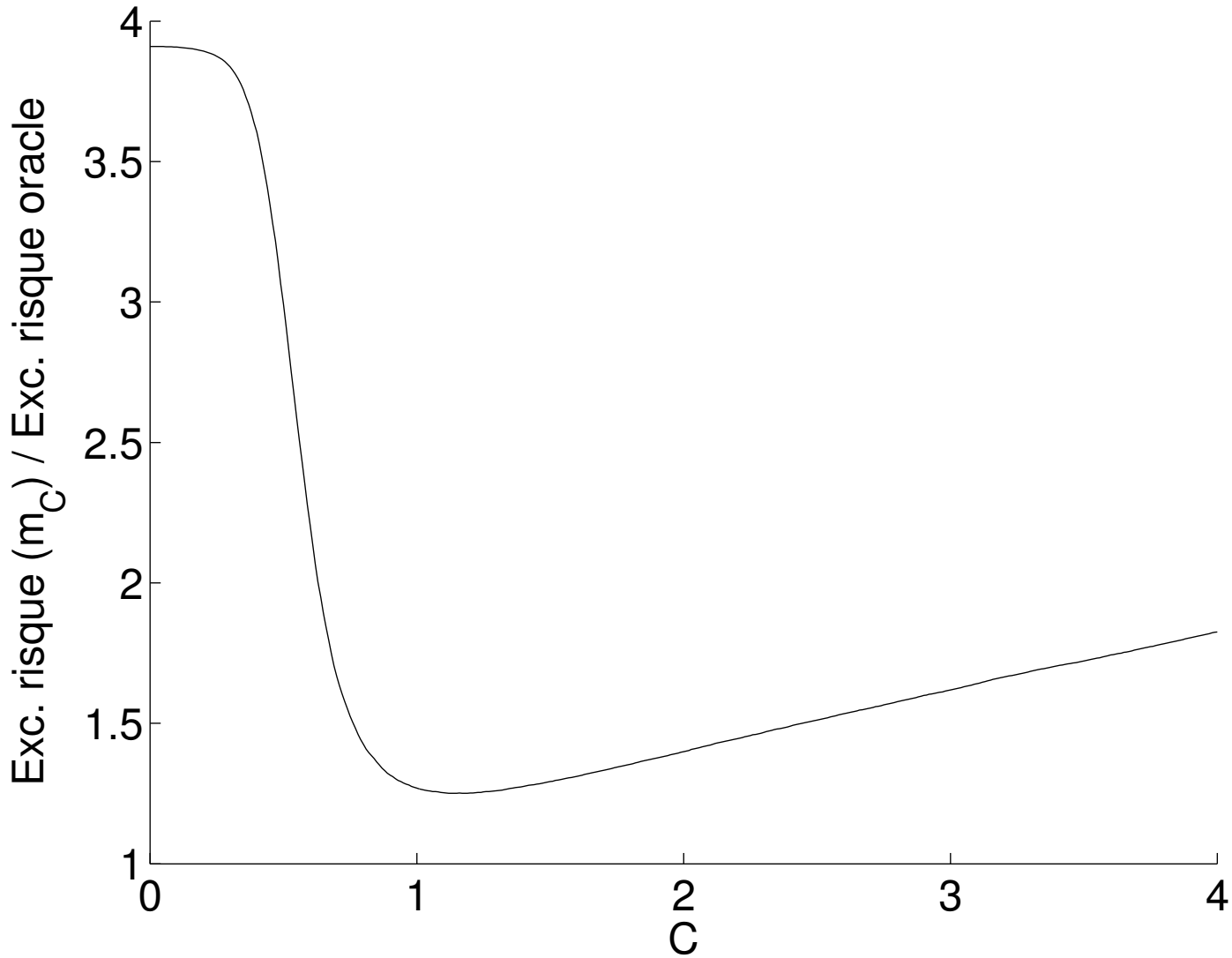
Majoration du risque :

$$\mathrm{pen}(m; D_n) \geqslant \mathrm{pen}_{\mathrm{id}}(m; D_n) \qquad \text{(à translation près)}$$

Pourquoi pénaliser ?



Surpénalisation : pen $= C\,\mathbb{E}[\mathrm{pen}_{\mathrm{id}}] \Rightarrow C$ optimal ?



Surpénalisation et validation croisée

Hypothèses : (**Ind**), (**Reg**),

$$\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big] = \alpha(m) + \frac{\beta(m)}{n} \qquad \text{et} \qquad \mathbb{E}\Big[\widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big)\Big] = \alpha(m) - \frac{\beta(m)}{n}$$

Alors :

$$\mathbb{E}\Big[\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\Big] = \mathbb{E}\Big[\widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big) + \underbrace{\frac{1}{2}\left(1+\frac{n}{n_e}\right)}_{\text{facteur de surpénalisation}} \mathrm{pen}_{\mathrm{id}}(m)\Big]$$

Surpénalisation et validation croisée

Hypothèses : (**Ind**), (**Reg**),

$$\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big] = \alpha(m) + \frac{\beta(m)}{n} \qquad \text{et} \qquad \mathbb{E}\Big[\widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big)\Big] = \alpha(m) - \frac{\beta(m)}{n}$$

Alors :

$$\mathbb{E}\Big[\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\Big] = \mathbb{E}\Big[\widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big) + \underbrace{\frac{1}{2}\left(1+\frac{n}{n_e}\right)}_{\text{facteur de surpénalisation}} \mathrm{pen}_{\mathrm{id}}(m)\Big]$$

$$\mathbb{E}\Big[\widehat{\mathcal{R}}^{\mathrm{vf}}\big(\widehat{f}_m; D_n; (B_j)_{1\leqslant j\leqslant V}\big)\Big] = \mathbb{E}\Big[\widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big) + \overbrace{\left(1+\frac{1}{2(V-1)}\right)} \mathrm{pen}_{\mathrm{id}}(m)\Big]$$

Surpénalisation et validation croisée

Hypothèses : (**Ind**), (**Reg**),

$$\mathbb{E}\Big[\mathcal{R}_P\big(\widehat{f}_m(D_n)\big)\Big] = \alpha(m) + \frac{\beta(m)}{n} \qquad \text{et} \qquad \mathbb{E}\Big[\widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big)\Big] = \alpha(m) - \frac{\beta(m)}{n}$$

Alors :

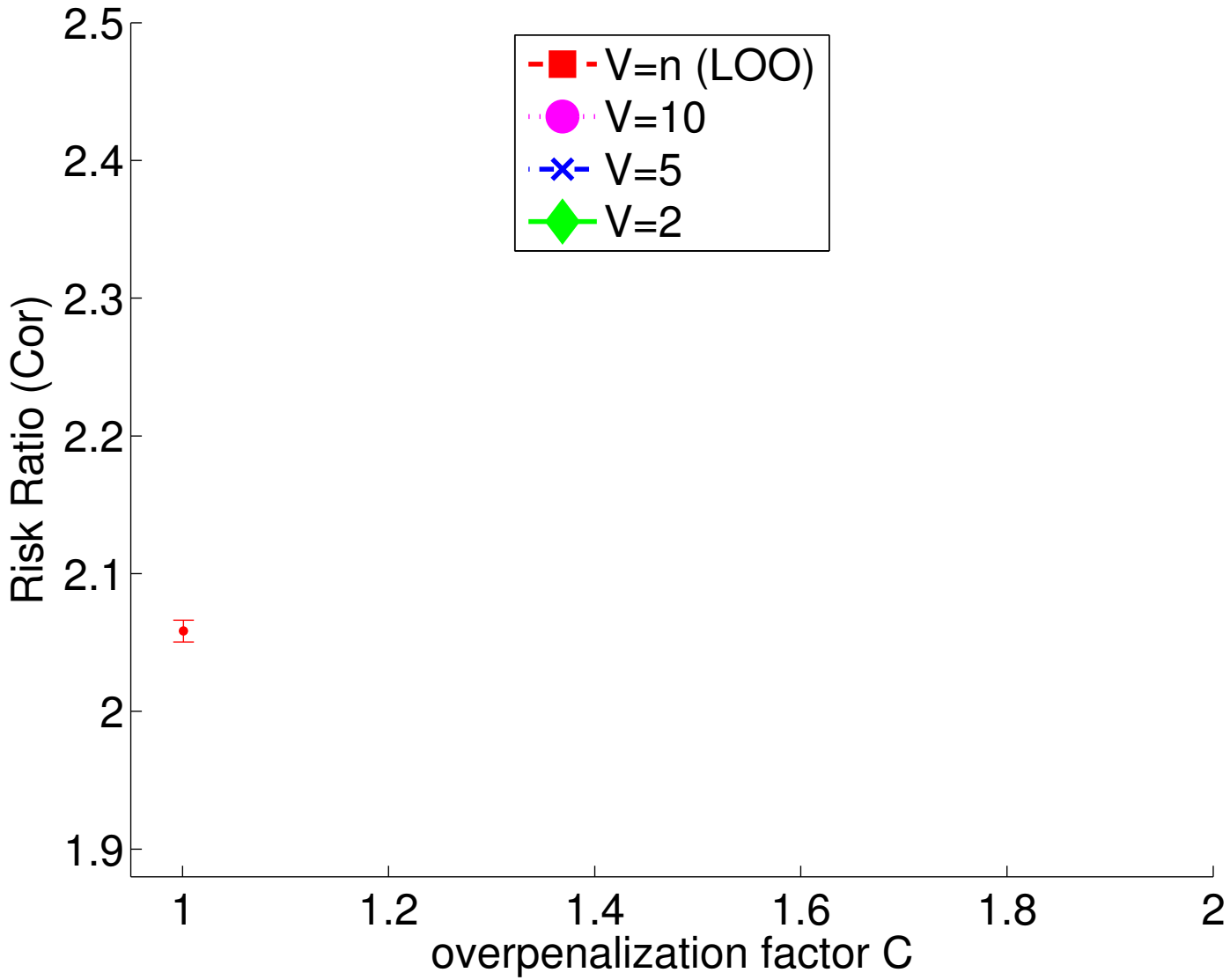
$$\mathbb{E}\Big[\widehat{\mathcal{R}}^{\mathrm{vc}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\Big] = \mathbb{E}\Big[\widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big) + \underbrace{\frac{1}{2}\left(1+\frac{n}{n_e}\right)}_{\text{facteur de surpénalisation}} \mathrm{pen}_{\mathrm{id}}(m)\Big]$$

$$\mathbb{E}\Big[\widehat{\mathcal{R}}^{\mathrm{vf}}\big(\widehat{f}_m; D_n; (B_j)_{1\leqslant j\leqslant V}\big)\Big] = \mathbb{E}\Big[\widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big) + \overbrace{\left(1+\frac{1}{2(V-1)}\right)} \mathrm{pen}_{\mathrm{id}}(m)\Big]$$

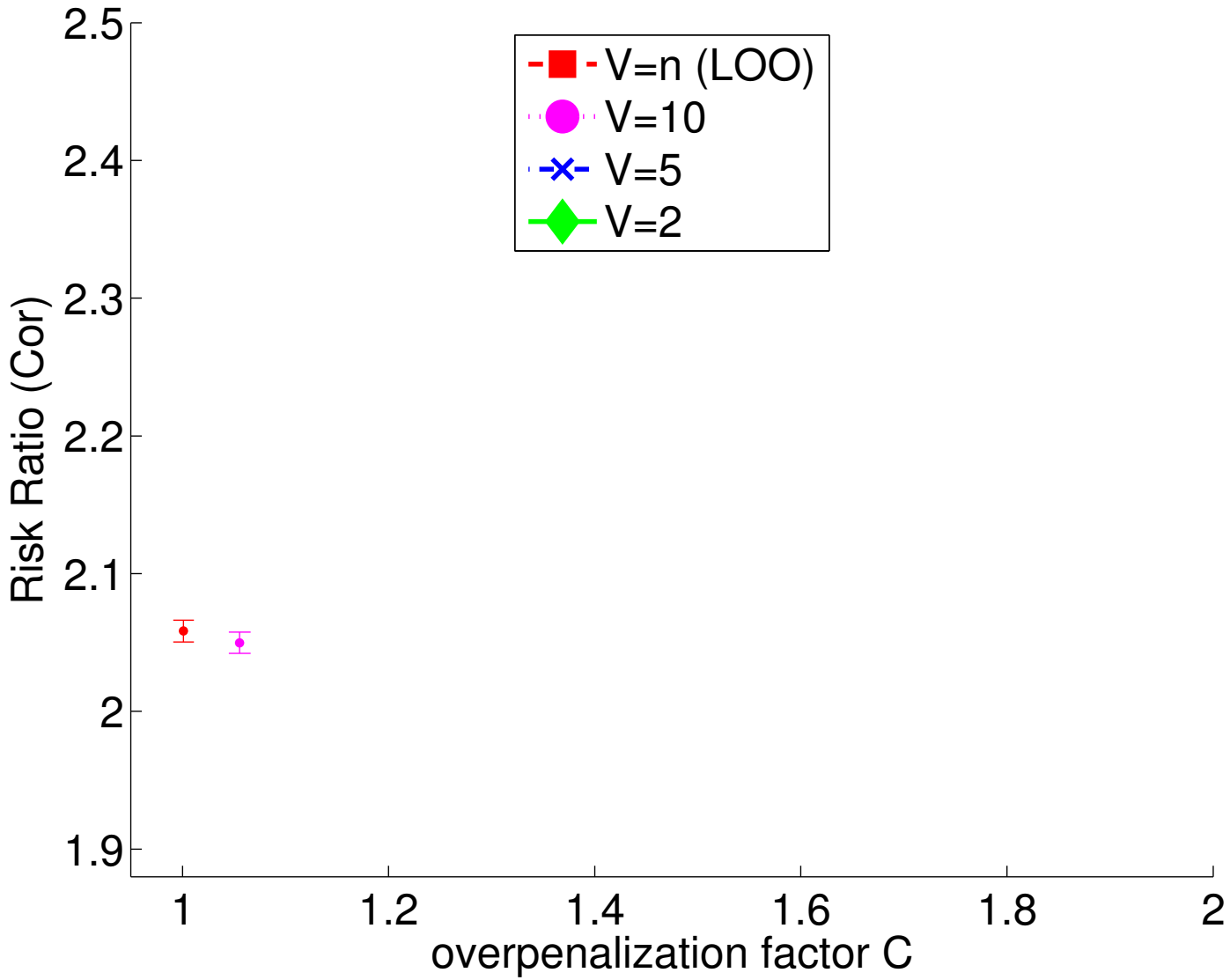
En corrigeant le biais : pas de surpénalisation !

$$\mathbb{E}\Big[\widehat{\mathcal{R}}^{\mathrm{vc-cor}}\big(\widehat{f}_m; D_n; (E_j)_{1\leqslant j\leqslant V}\big)\Big] = \mathbb{E}\Big[\widehat{\mathcal{R}}_n\big(\widehat{f}_m(D_n)\big) + \mathrm{pen}_{\mathrm{id}}(m; D_n)\Big]$$

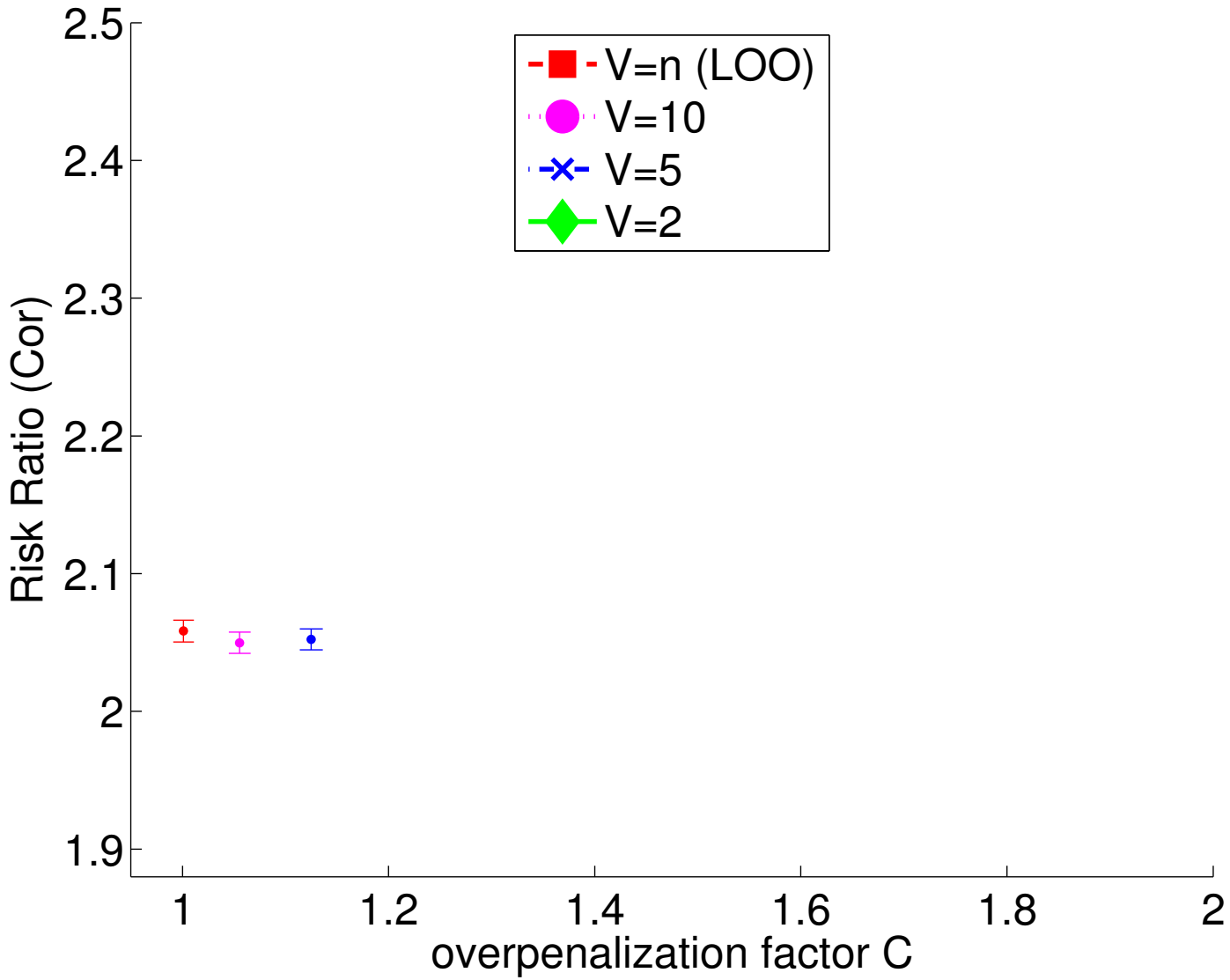
Simulation (estimation de densité L^2) : VC V-fold



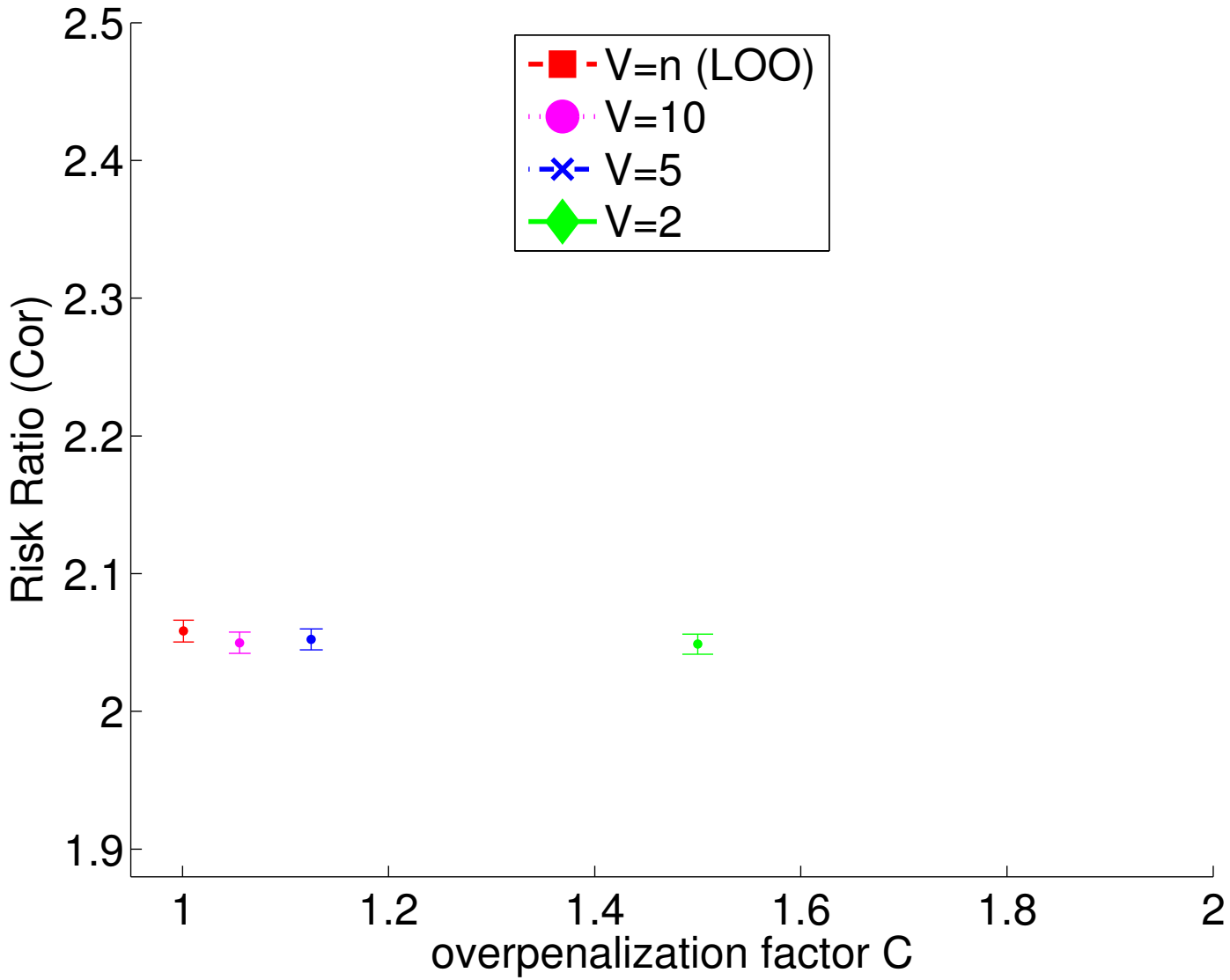
Simulation (estimation de densité L^2) : VC V-fold



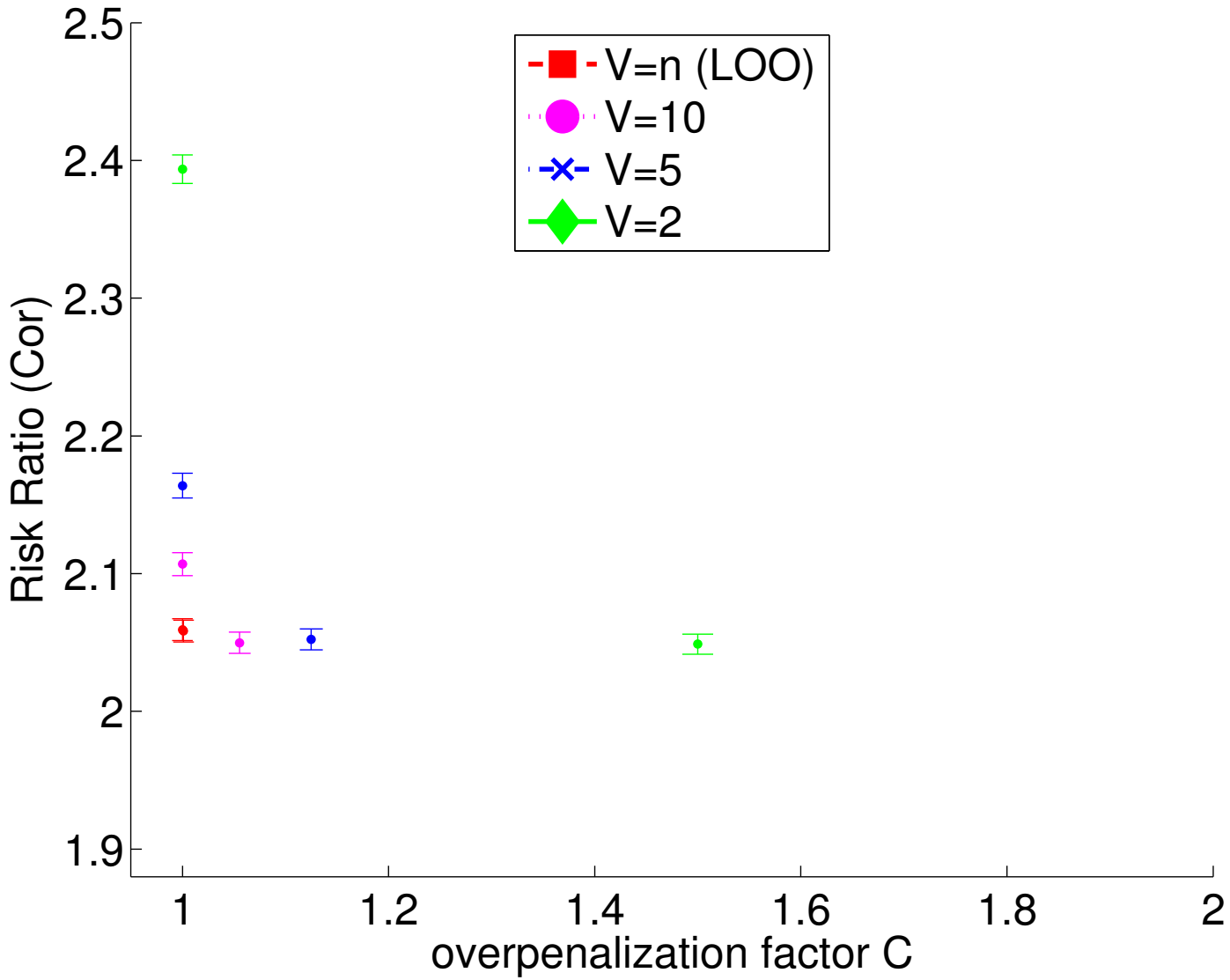
Simulation (estimation de densité L^2) : VC V-fold



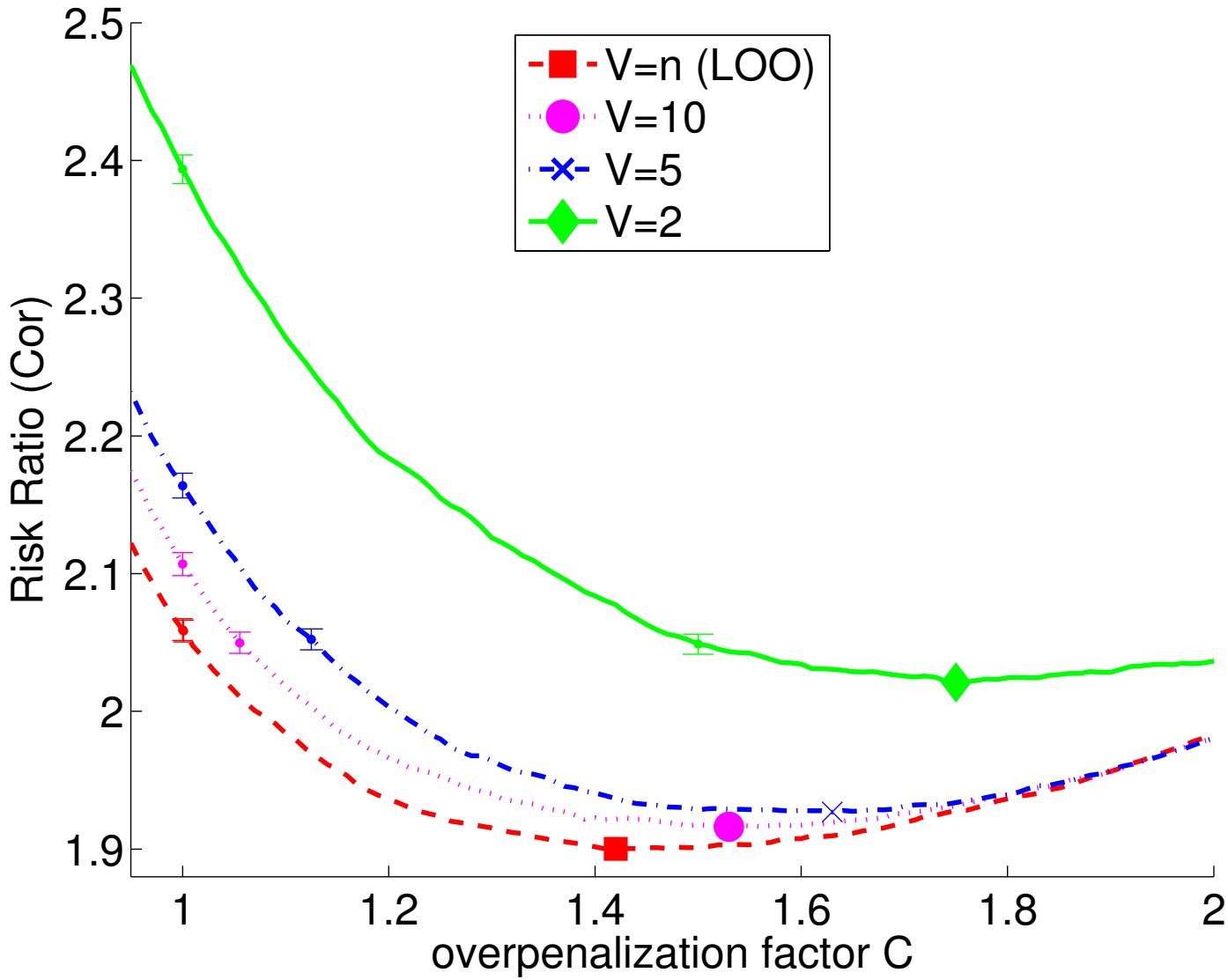
Simulation (estimation de densité L^2) : VC V-fold



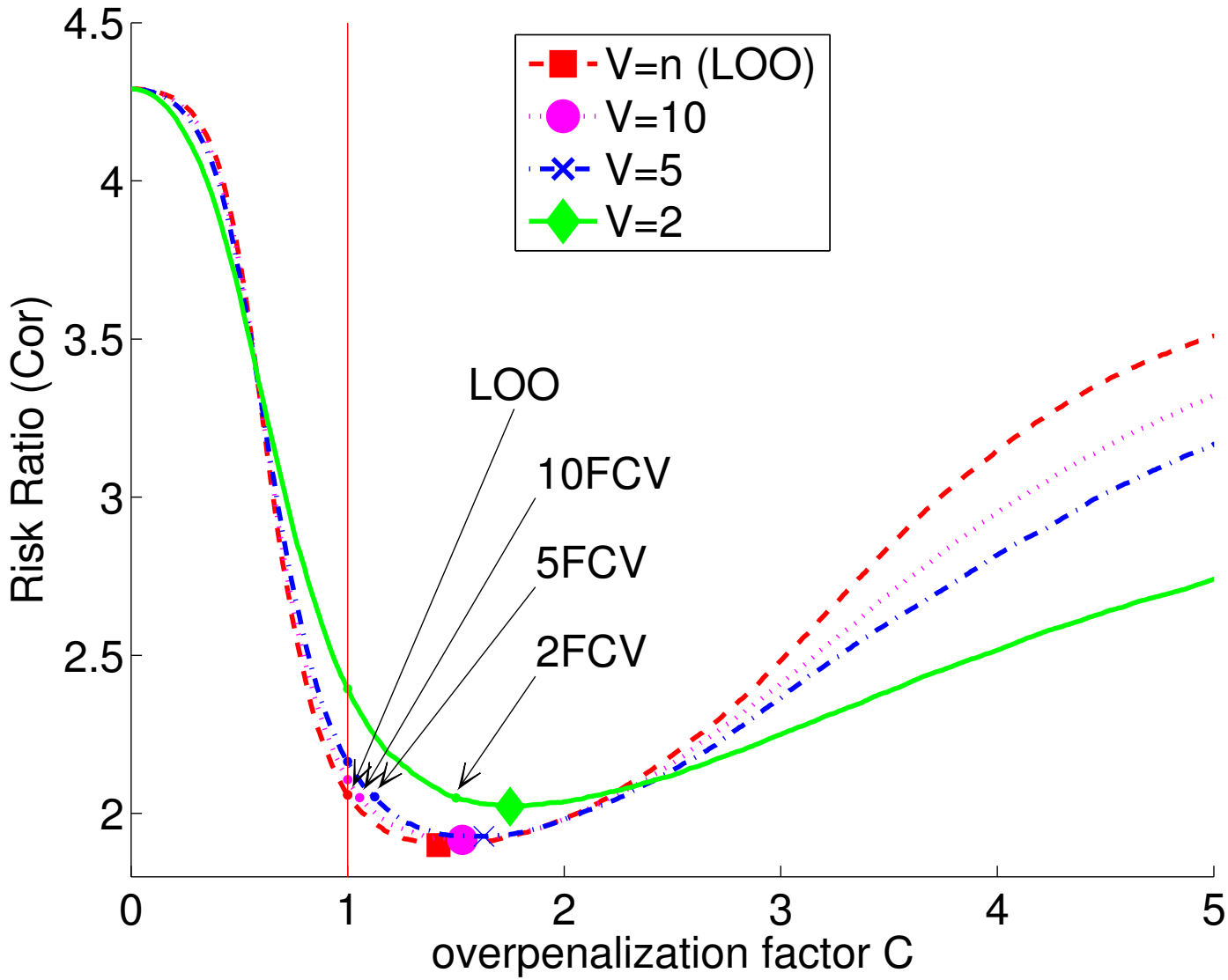
Simulation (estimation de densité L^2) : pénalisation V-fold



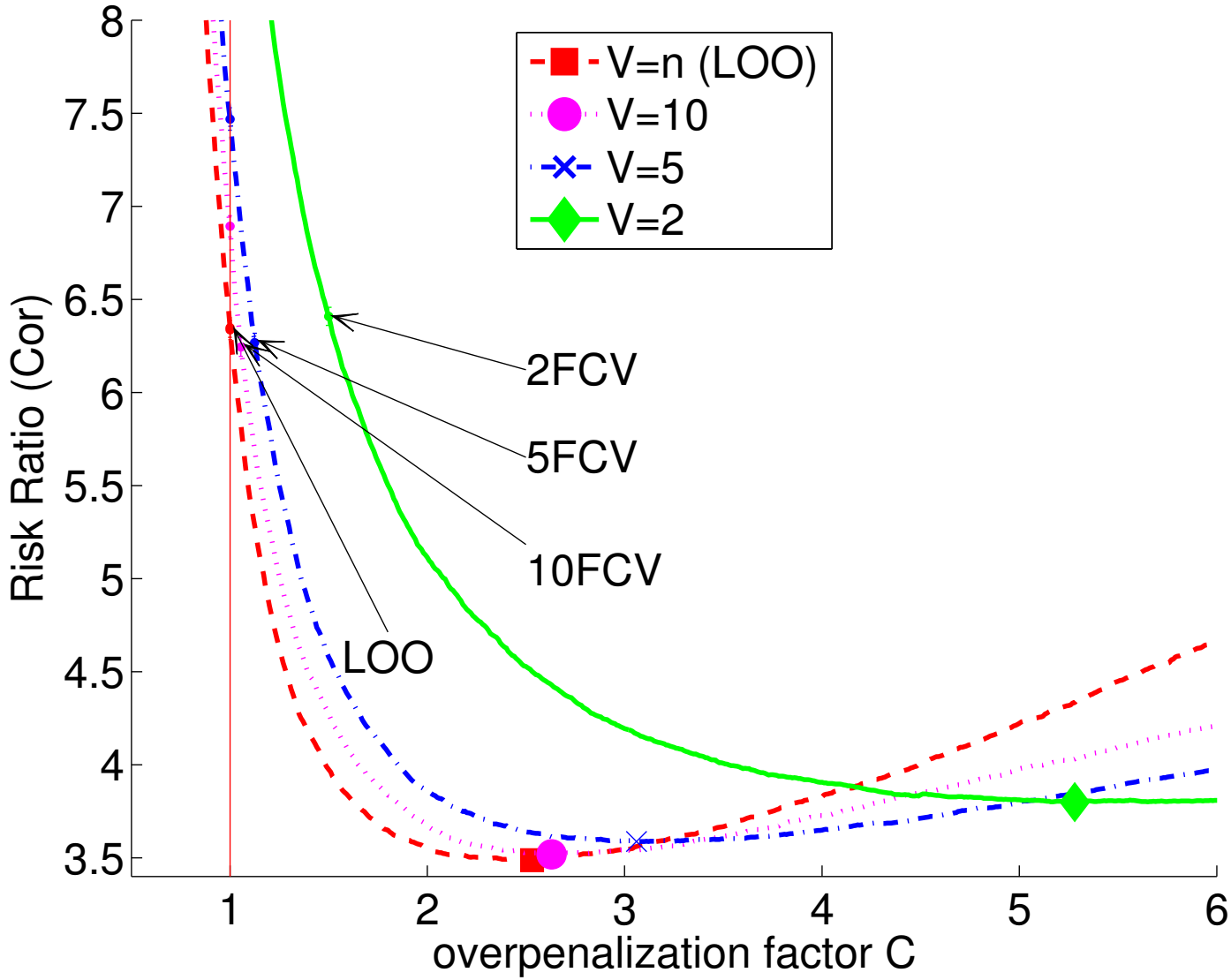
Simulation (estimation de densité L^2) : surpénalisation



Simulation (estimation de densité L^2) : conclusion



Simulation (estimation de densité L^2) : cadre différent



Plan

1. Problèmes
2. Définition
3. Estimation du risque
4. Sélection d'estimateurs
5. Conclusion

Risque de l'estimateur sélectionné ?

$$\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_{\widehat{m}^{\mathrm{vc}}\left(D_n;(E_j)_{1\leqslant j\leqslant V}\right)};D_n;(E_j)_{1\leqslant j\leqslant V}\right)$$

est une estimation biaisée du risque

$$\mathcal{R}_P\left(\widehat{f}_{\widehat{m}^{\mathrm{vc}}\left(D_n;(E_j)_{1\leqslant j\leqslant V}\right)}\right)$$

Risque de l'estimateur sélectionné ?

$$\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{f}_{\widehat{m}^{\mathrm{vc}}\left(D_n;(E_j)_{1\leqslant j\leqslant V}\right)};D_n;(E_j)_{1\leqslant j\leqslant V}\right)$$

est une estimation biaisée du risque

$$\mathcal{R}_P\left(\widehat{f}_{\widehat{m}^{\mathrm{vc}}\left(D_n;(E_j)_{1\leqslant j\leqslant V}\right)}\right)$$

$\Rightarrow$ découpage en trois sous-échantillons :
entraînement / validation / test

Choix d'un type de validation croisée (cas « régulier »)

- Temps de calcul : $\mathcal{O}(V)$ en général

Choix d'un type de validation croisée (cas « régulier »)

- Temps de calcul : $\mathcal{O}(V)$ en général
- Taille n_e de l'échantillon d'entraînement :
  $\Rightarrow$ biais / facteur de surpénalisation (décroissant avec n_e)
  possibilité de débiaiser (mais le souhaite-t-on ?)

Choix d'un type de validation croisée (cas « régulier »)

- Temps de calcul : $\mathcal{O}(V)$ en général
- Taille n_e de l'échantillon d'entraînement :
  $\Rightarrow$ biais / facteur de surpénalisation (décroissant avec n_e)
  possibilité de débiaiser (mais le souhaite-t-on ?)
- Nombre V de découpages à n_e fixé $\Rightarrow$ variance (décroissante), quasi minimale pour V « petit » (en fonction de n_e...)
  $\Rightarrow$ compromis temps de calcul / précision

Choix d'un type de validation croisée (cas « régulier »)

- Temps de calcul : $\mathcal{O}(V)$ en général
- Taille n_e de l'échantillon d'entraînement :
  $\Rightarrow$ biais / facteur de surpénalisation (décroissant avec n_e)
  possibilité de débiaiser (mais le souhaite-t-on ?)
- Nombre V de découpages à n_e fixé $\Rightarrow$ variance (décroissante), quasi minimale pour V « petit » (en fonction de n_e...)
  $\Rightarrow$ compromis temps de calcul / précision
- « V-fold » : biais et variance reliés $\Rightarrow$ $V = 5$ ou 10 ?

Choix d'un type de validation croisée (cas « régulier »)

- Temps de calcul : $\mathcal{O}(V)$ en général
- Taille n_e de l'échantillon d'entraînement :
  $\Rightarrow$ biais / facteur de surpénalisation (décroissant avec n_e)
  possibilité de débiaiser (mais le souhaite-t-on ?)
- Nombre V de découpages à n_e fixé $\Rightarrow$ variance (décroissante), quasi minimale pour V « petit » (en fonction de n_e...)
  $\Rightarrow$ compromis temps de calcul / précision
- « V-fold » : biais et variance reliés $\Rightarrow$ $V = 5$ ou 10 ?
- Découplage biais/variance $\Rightarrow$ choix plus simple de n_e et V :
  - VC Monte-Carlo
  - VC « V-fold » répétée
  - pénalisation « V-fold » : facteur de surpénalisation C choisi librement

Validation croisée ou procédure spécifique ?

- Par exemple : C_p ou validation croisée ?

Validation croisée ou procédure spécifique ?

- Par exemple : C_p ou validation croisée ?

- Si les hypothèses de C_p sont vérifiées :
  estimateurs et risque des moindres carrés, modèles = espaces vectoriels, bruit homoscédastique
  $\Rightarrow$ C_p (la validation croisée paye le prix de sa polyvalence)

Validation croisée ou procédure spécifique ?

- Par exemple : C_p ou validation croisée ?
- Si les hypothèses de C_p sont vérifiées :
  estimateurs et risque des moindres carrés, modèles = espaces vectoriels, bruit homoscédastique
  $\Rightarrow$ C_p (la validation croisée paye le prix de sa polyvalence)
- Sinon (ou si on n'a pas confiance...)
  $\Rightarrow$ validation croisée (plus robuste)

Validation croisée ou procédure spécifique ?

- Par exemple : C_p ou validation croisée ?
- Si les hypothèses de C_p sont vérifiées :
  estimateurs et risque des moindres carrés, modèles = espaces vectoriels, bruit homoscédastique
  $\Rightarrow$ C_p (la validation croisée paye le prix de sa polyvalence)
- Sinon (ou si on n'a pas confiance...)
  $\Rightarrow$ validation croisée (plus robuste)
- Très souvent, aucune méthode spécifique $\Rightarrow$ validation croisée

Généralité de ces résultats ?

- Au moins valable pour les moindres carrés (régression / estimation de densité) et les noyaux en estimation de densité.

Généralité de ces résultats ?

- Au moins valable pour les moindres carrés (régression / estimation de densité) et les noyaux en estimation de densité.
- Correction du biais / pénalisation V-fold : valable si

$$\mathbb{E}\Big[\mathcal{R}_P(\widehat{f}_m(D_n)) - \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n))\Big] = \frac{\gamma(m)}{n}\,.$$

Sinon : utiliser le V-fold répété ou la VC Monte-Carlo avec n_e bien choisi.

Généralité de ces résultats ?

- Au moins valable pour les moindres carrés (régression / estimation de densité) et les noyaux en estimation de densité.
- Correction du biais / pénalisation V-fold : valable si

$$\mathbb{E}\Big[\mathcal{R}_P(\widehat{f}_m(D_n)) - \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n))\Big] = \frac{\gamma(m)}{n}\,.$$

  Sinon : utiliser le V-fold répété ou la VC Monte-Carlo avec n_e bien choisi.
- Variance : d'autres comportements possibles dans d'autres cadres (expériences).

Généralité de ces résultats ?

- Au moins valable pour les moindres carrés (régression / estimation de densité) et les noyaux en estimation de densité.
- Correction du biais / pénalisation V-fold : valable si

$$\mathbb{E}\Big[\mathcal{R}_P(\widehat{f}_m(D_n)) - \widehat{\mathcal{R}}_n(\widehat{f}_m(D_n))\Big] = \frac{\gamma(m)}{n}\,.$$

  Sinon : utiliser le V-fold répété ou la VC Monte-Carlo avec n_e bien choisi.
- Variance : d'autres comportements possibles dans d'autres cadres (expériences).
- Tout peut se vérifier sur des données simulées : tracer

$$n \to \mathbb{E}\Big[\mathcal{R}_P(\widehat{f}_m(D_n))\Big] \qquad \text{et} \qquad m \to \mathrm{var}\Big(\widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{f}_m) - \widehat{\mathcal{R}}^{\mathrm{vc}}(\widehat{f}_{m^\star})\Big)\,.$$

Validation croisée pour l'identification

- Différence principale : valeur de la constante de surpénalisation optimale $C^\star$, souvent $C^\star \to +\infty$ lorsque $n \to +\infty$.

$\Leftrightarrow$ Paradoxe de la validation croisée (Yang, 2006, 2007) : $n_e \ll n$ peut être nécessaire !

- Pourquoi ? n_e plus petit $\Rightarrow$ plus facile de distinguer les deux procédures... si n_e assez grand (régime asymptotique).

- Remarque : objectif d'estimation, cadre paramétrique $\Rightarrow$ phénomène similaire.

Données dépendantes

- $D_n^E, D_n^{E^c}$ pas indépendants $\Rightarrow$ l'heuristique de la VC s'écroule !

$\Rightarrow$ problèmes possibles pour l'estimation du risque (Hart & Wehrly, 1986 ; Opsomer et al., 2001).

- Solution pour une dépendance à courte portée :
  supprimer des données dans chaque sous-échantillon $\Rightarrow$ séparer (temporellement) les échantillons d'entraînement et de validation.

Grande famille d'estimateurs/modèles

- Sélection de modèles/estimateurs parmi une famille « exponentielle » (exclu implicitement dans tous les résultats précédents).
  $\Rightarrow$ l'espérance ne décrit plus le comportement au premier ordre !

- Exemples : sélection de variables avec $p \geqslant n$ variables, détection de ruptures.

- Solution : regrouper les modèles $\Rightarrow$ un estimateur par « dimension » (e.g., minimiseur du risque empirique) fonctionne pour la détection de ruptures (A. & Celisse, 2010).

Questions ?

Et vous, comment utilisez-vous la validation croisée ?

Quel(s) choix pour V, n_e, etc. ?

Quel(s) comportement(s) avez-vous observé en pratique ?