Comparaison de procédures de validation croisée (« V-fold »)

Sylvain Arlot (collaboration avec Matthieu Lerasle)

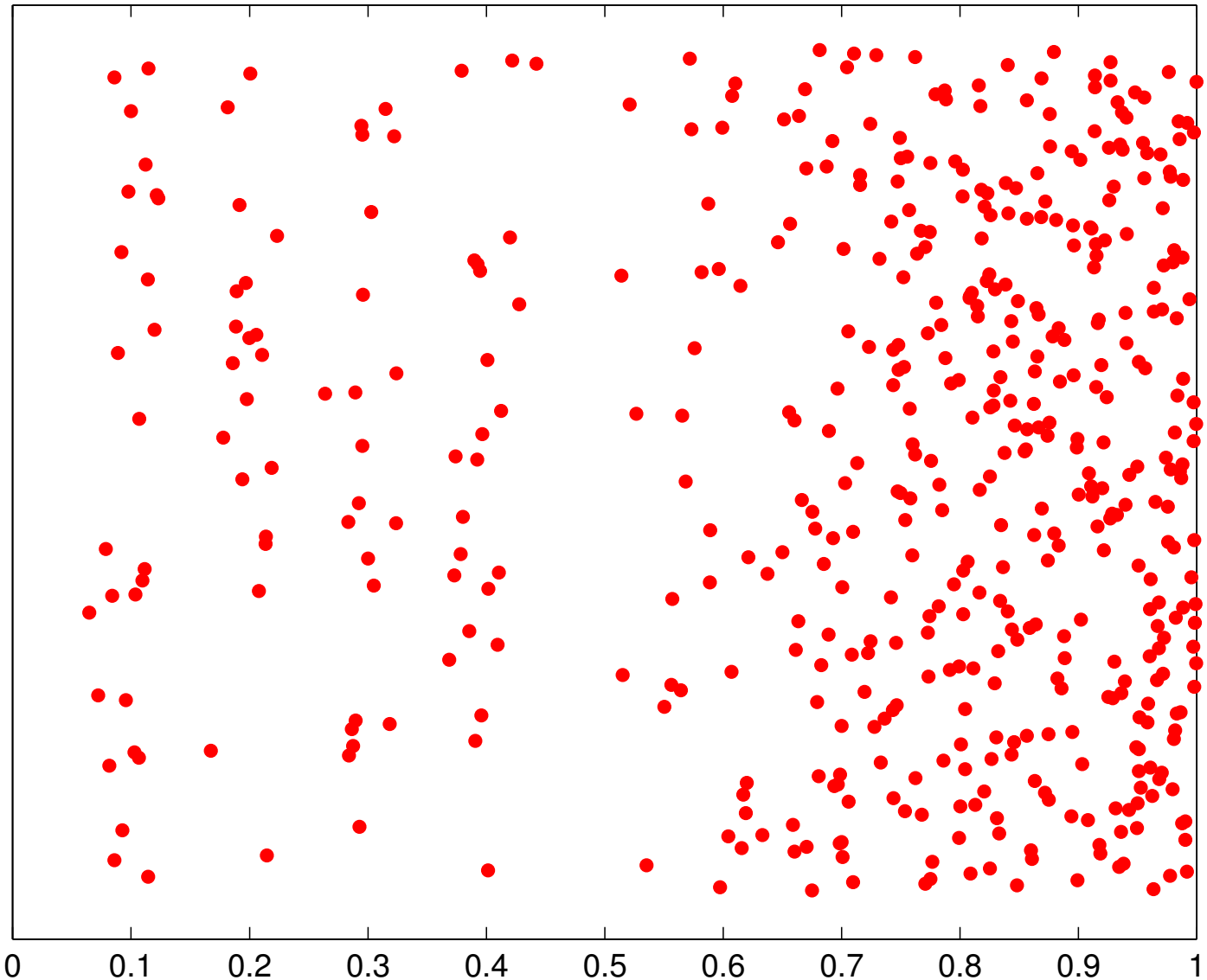
[1]Cnrs

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

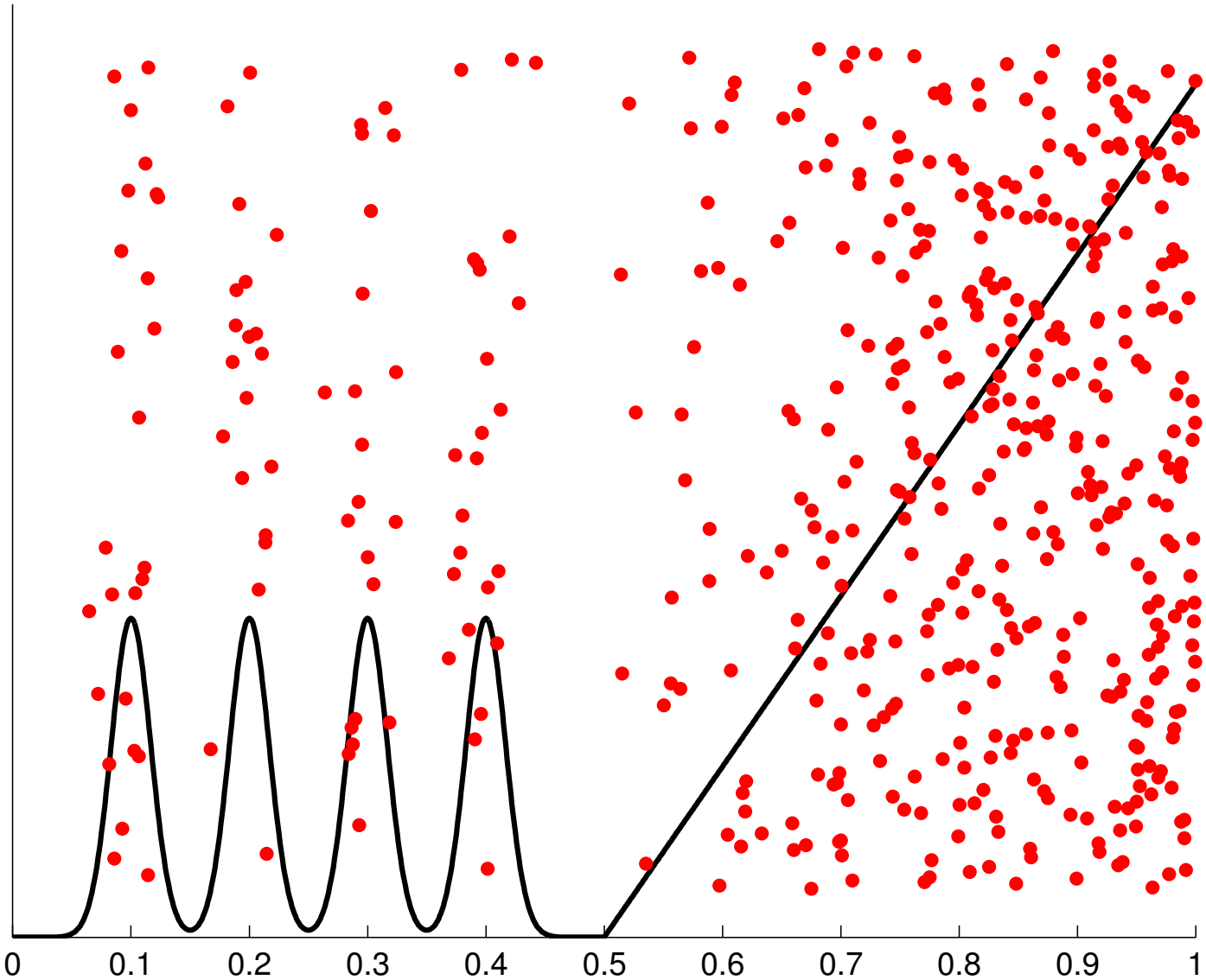
Nice, 20 février 2015

Plan

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



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



Problème : estimation de densité

- Données D_n : $\xi_1, \ldots, \xi_n \in \Xi$ (i.i.d. $\sim P$, densité $s^\star$ par rapport à μ)
- Contraste des moindres carrés $\gamma(t, \xi) = \|t\|^2_{L^2(\mu)} - 2t(\xi)$
- Objectif : apprendre $t \in \mathbb{S} = \{\text{fonctions mesurables } \Xi \to \mathbb{R}\}$ t.q. $\mathbb{E}_{\xi \sim P}\big[\gamma(t;\xi)\big] =: P\gamma(t)$ est minimale.

Problème : estimation de densité

- Données D_n : $\xi_1, \ldots, \xi_n \in \Xi$ (i.i.d. $\sim P$, densité $s^\star$ par rapport à μ)
- Contraste des moindres carrés $\gamma(t, \xi) = \|t\|^2_{L^2(\mu)} - 2t(\xi)$
- Objectif : apprendre $t \in \mathbb{S} = \{\text{fonctions mesurables } \Xi \to \mathbb{R}\}$ t.q. $\mathbb{E}_{\xi \sim P}\big[\gamma(t;\xi)\big] =: P\gamma(t)$ est minimale.

$$P\gamma(t) = \int t^2 \, \mathrm{d}\mu - 2\int t s^\star \, \mathrm{d}\mu = \int (t - s^\star)^2 \, \mathrm{d}\mu - \|s^\star\|^2_{L^2(\mu)}$$

$\Rightarrow$ densité $s^\star \in \operatorname{argmin}_{t \in \mathbb{S}} P\gamma(t)$ et la perte relative vaut

$$\ell(s^\star, t) := P\gamma(t) - P\gamma(s^\star) = \|t - s^\star\|^2_{L^2(\mu)} \, .$$

Problème : estimation de densité

- Données D_n : $\xi_1, \ldots, \xi_n \in \Xi$ (i.i.d. $\sim P$, densité $s^\star$ par rapport à μ)
- Contraste des moindres carrés $\gamma(t, \xi) = \|t\|^2_{L^2(\mu)} - 2t(\xi)$
- Objectif : apprendre $t \in \mathbb{S} = \{\text{fonctions mesurables } \Xi \to \mathbb{R}\}$ t.q. $\mathbb{E}_{\xi \sim P}\big[\gamma(t;\xi)\big] =: P\gamma(t)$ est minimale.

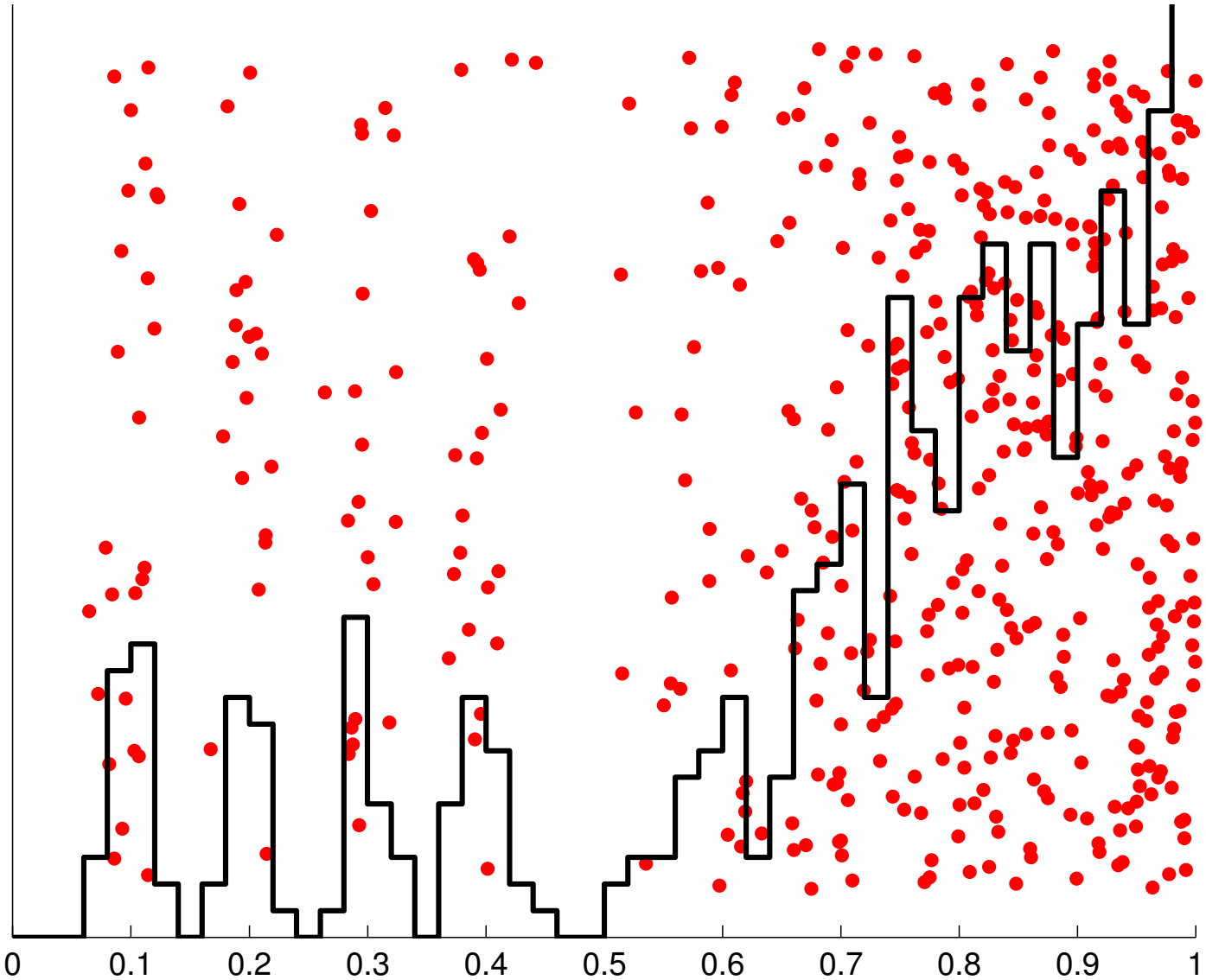
$$P\gamma(t) = \int t^2 \,\mathrm{d}\mu - 2\int t s^\star \,\mathrm{d}\mu = \int (t - s^\star)^2 \,\mathrm{d}\mu - \|s^\star\|^2_{L^2(\mu)}$$

$\Rightarrow$ densité $s^\star \in \operatorname{argmin}_{t \in \mathbb{S}} P\gamma(t)$ et la perte relative vaut

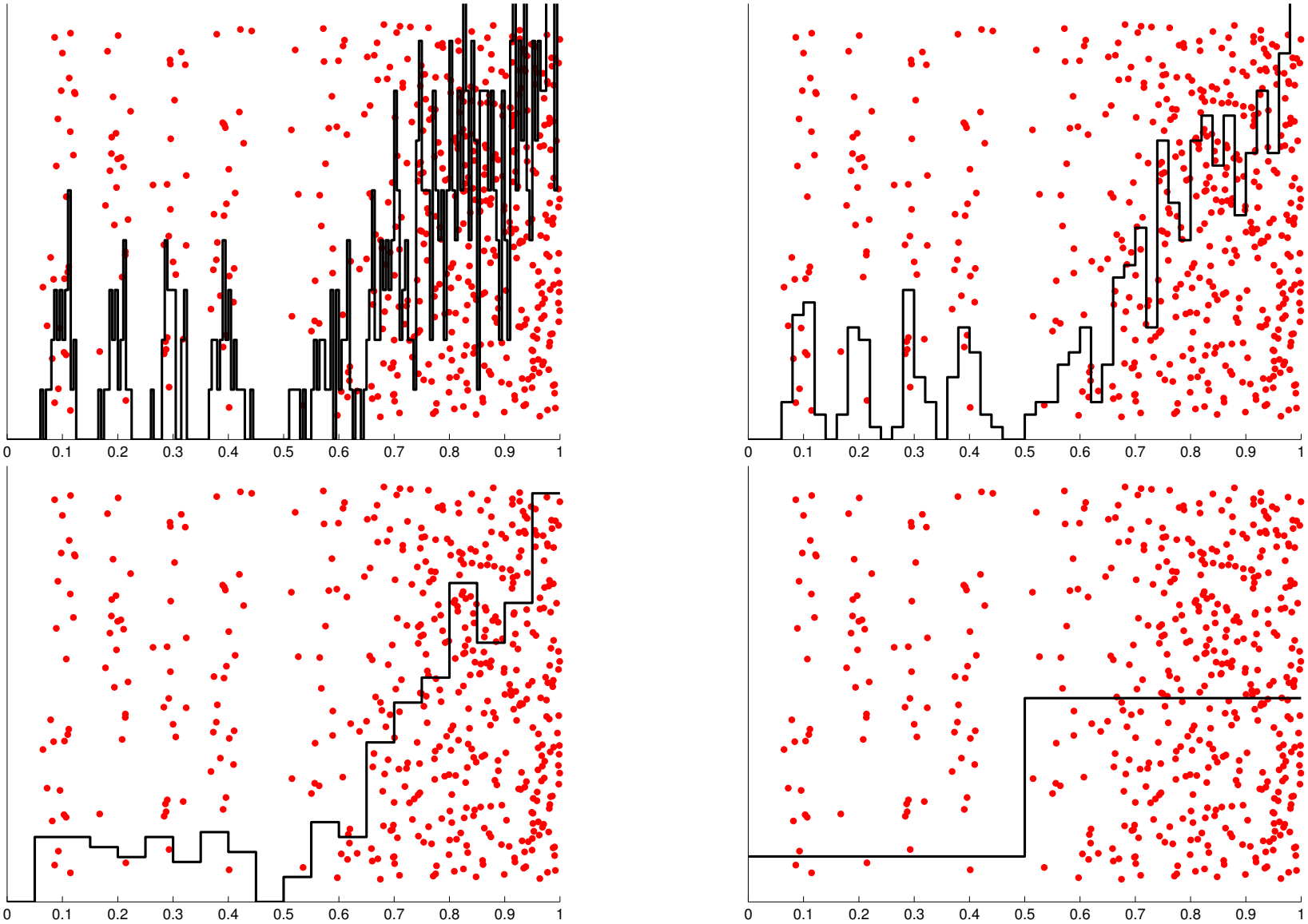
$$\ell(s^\star, t) := P\gamma(t) - P\gamma(s^\star) = \|t - s^\star\|^2_{L^2(\mu)} \ .$$

- Cas particulier d'un cadre général incluant aussi :
  - prédiction (régression/classification)
  - contraste log-vraisemblance en densité

Un estimateur par histogramme



Sélection de modèles : histogrammes réguliers



Sélection de modèles

- Estimateur des moindres carrés sur un modèle $S_m \subset \mathbb{S}$

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

Exemples de modèles : histogrammes, base tronquée (Fourier, ondelettes, etc.).

Sélection de modèles

- Estimateur des moindres carrés sur un modèle $S_m \subset \mathbb{S}$

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

  Exemples de modèles : histogrammes, base tronquée (Fourier, ondelettes, etc.).

- Collection de modèles $(\widehat{s}_m)_{m \in \mathcal{M}} \Rightarrow$ choisir $\widehat{m} = \widehat{m}(D_n)$?

Sélection de modèles

- Estimateur des moindres carrés sur un modèle $S_m \subset \mathbb{S}$

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

  Exemples de modèles : histogrammes, base tronquée (Fourier, ondelettes, etc.).

- Collection de modèles $(\widehat{s}_m)_{m \in \mathcal{M}} \Rightarrow$ choisir $\widehat{m} = \widehat{m}(D_n)$?

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

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

Compromis biais-variance

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

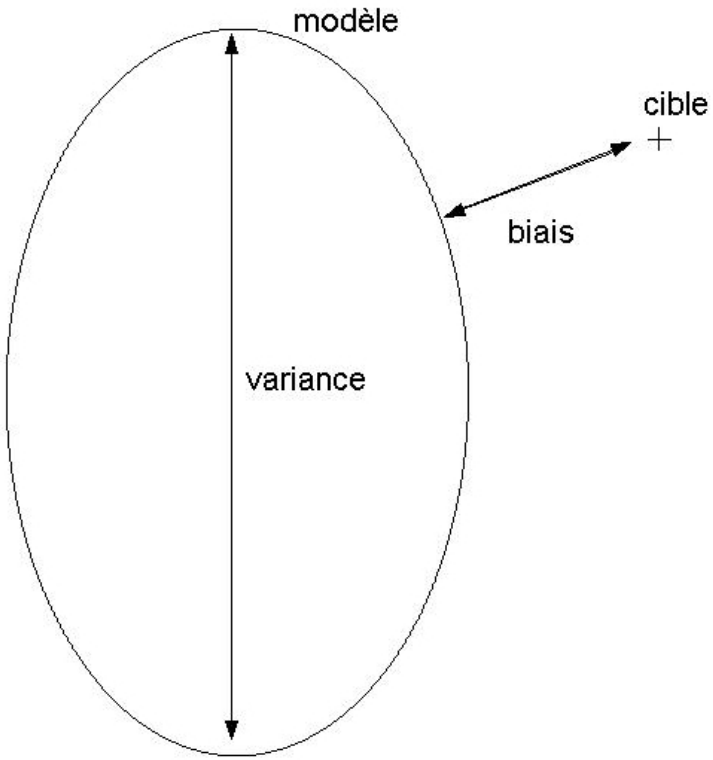
Biais ou Erreur d'approximation

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

Variance ou Erreur d'estimation

histogrammes réguliers sur $\mathbb{R}$ de pas d_m^{-1} :

$$\frac{d_m - \|s_m^\star\|^2_{L^2(\mu)}}{n} \approx \frac{d_m}{n}$$



Compromis biais-variance

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

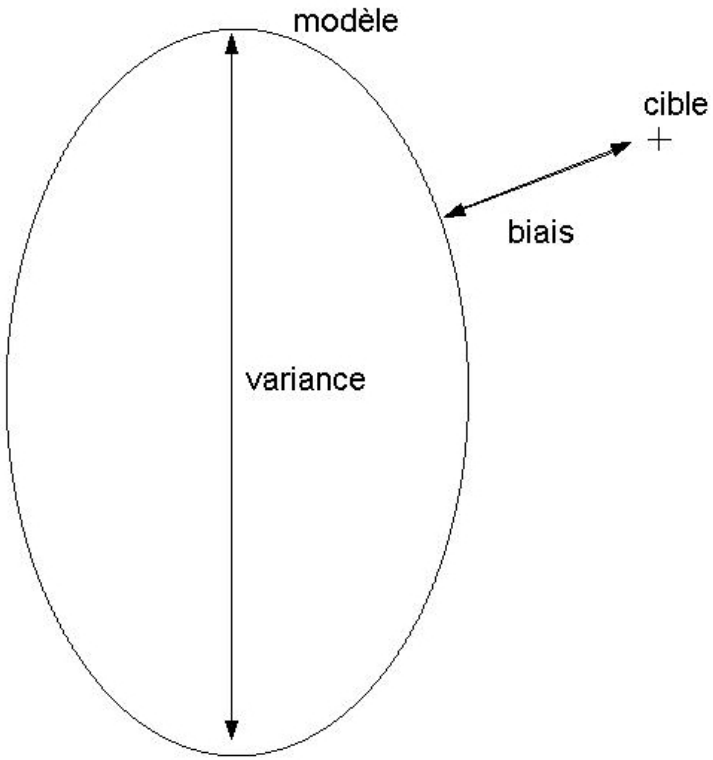
Biais ou Erreur d'approximation

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

Variance ou Erreur d'estimation

histogrammes réguliers sur $\mathbb{R}$ de pas d_m^{-1} :

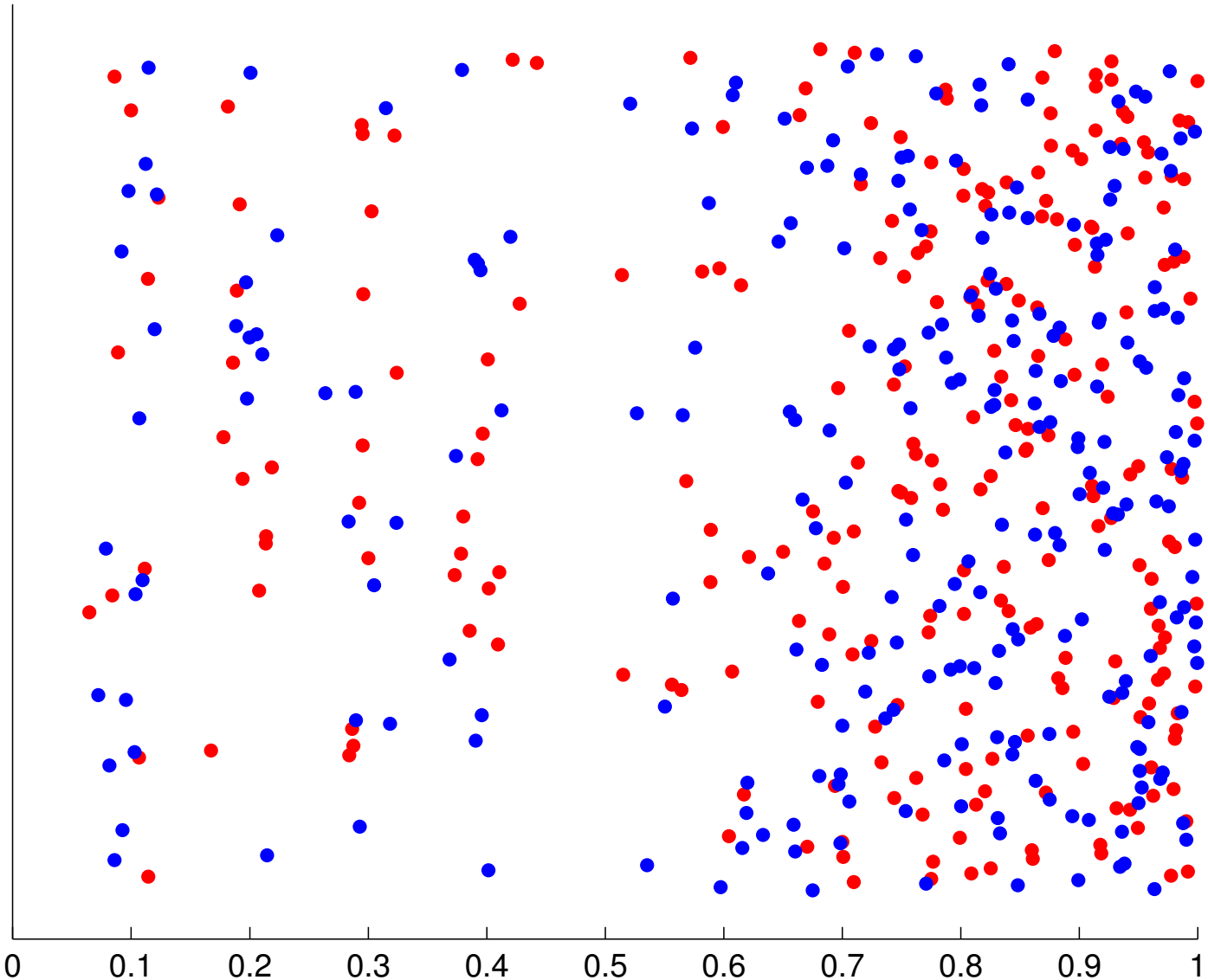
$$\frac{d_m - \|s_m^\star\|^2_{L^2(\mu)}}{n} \approx \frac{d_m}{n}$$



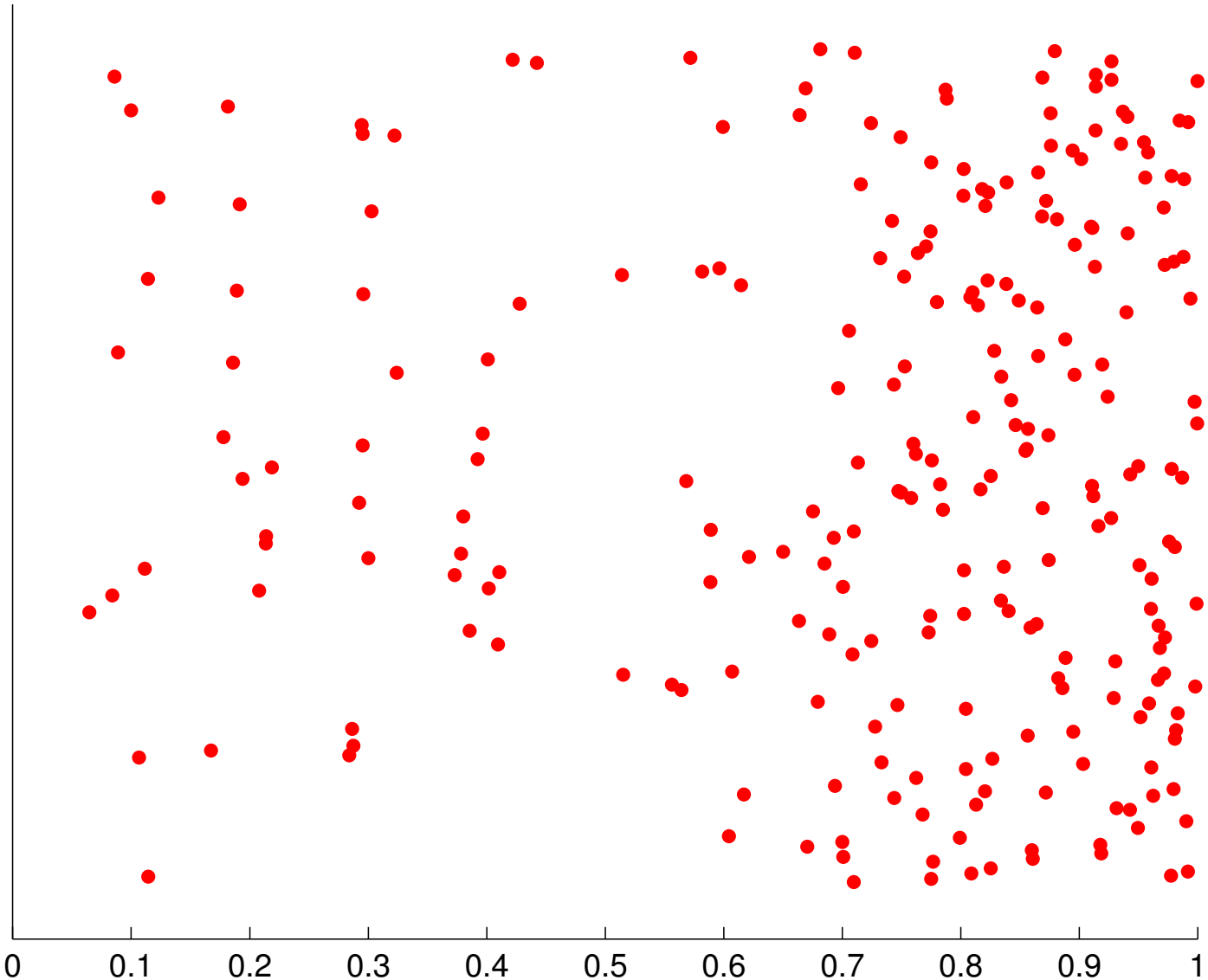
Compromis biais-variance

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

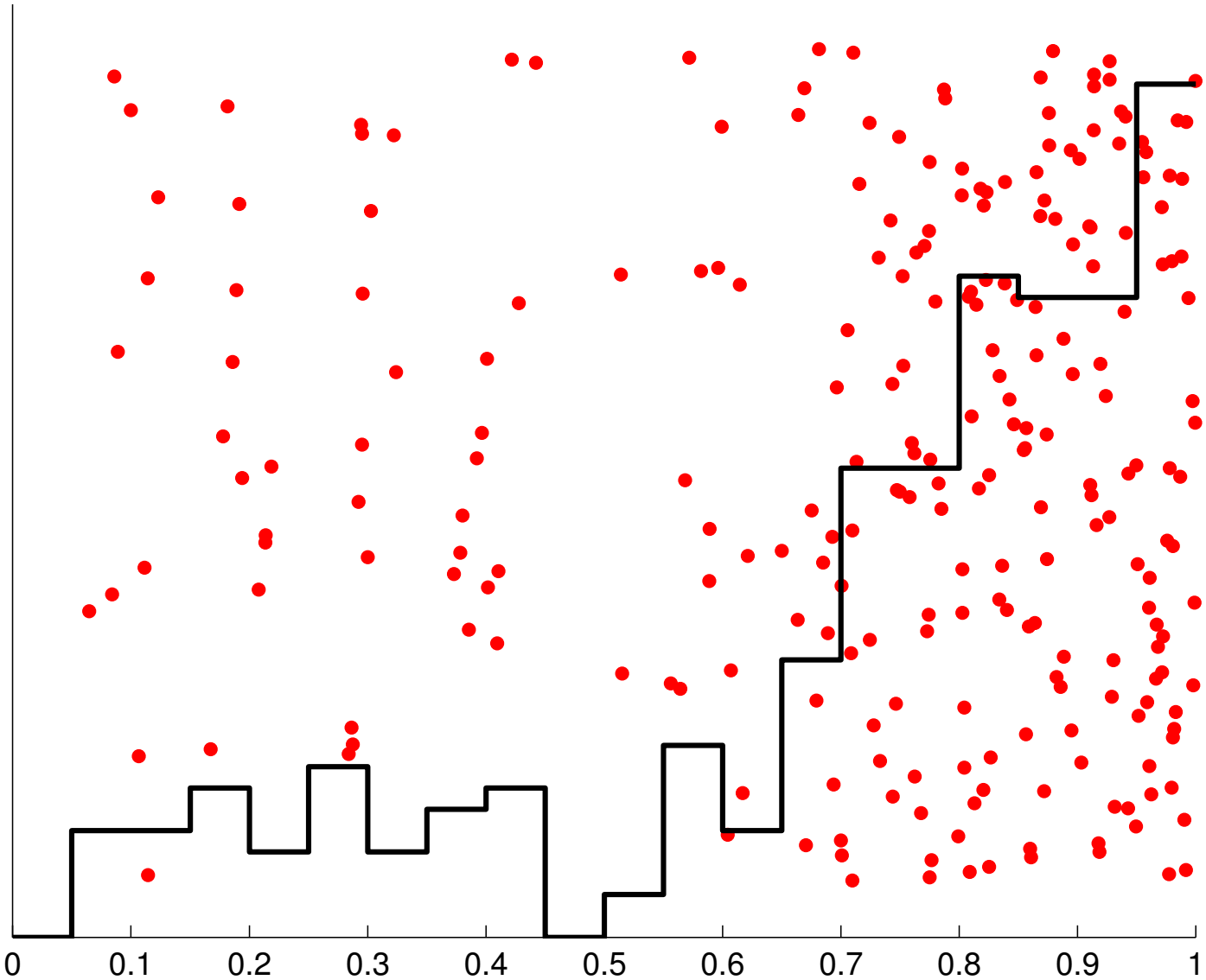
Principe de la validation simple



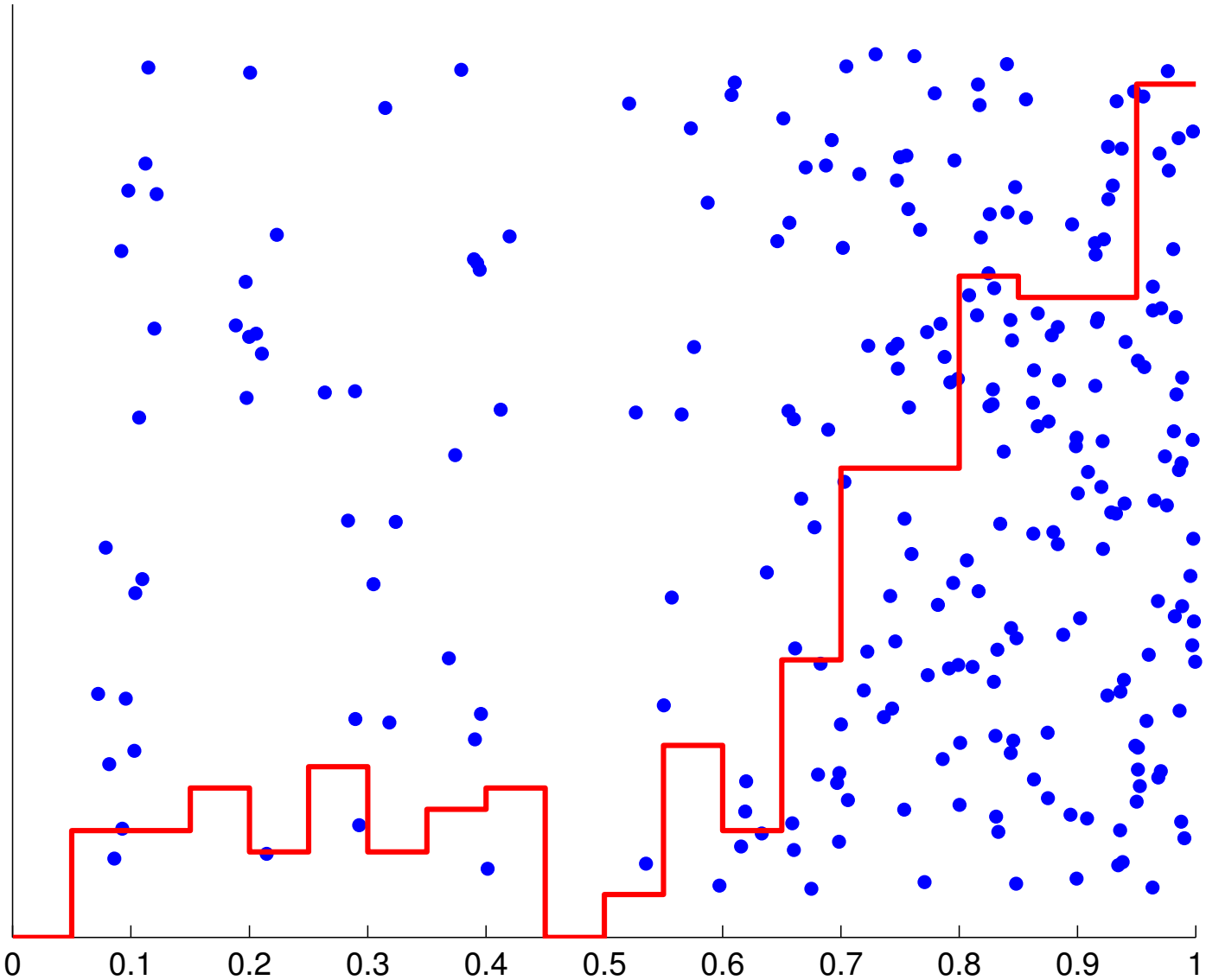
Principe de la validation : échantillon d'entraînement



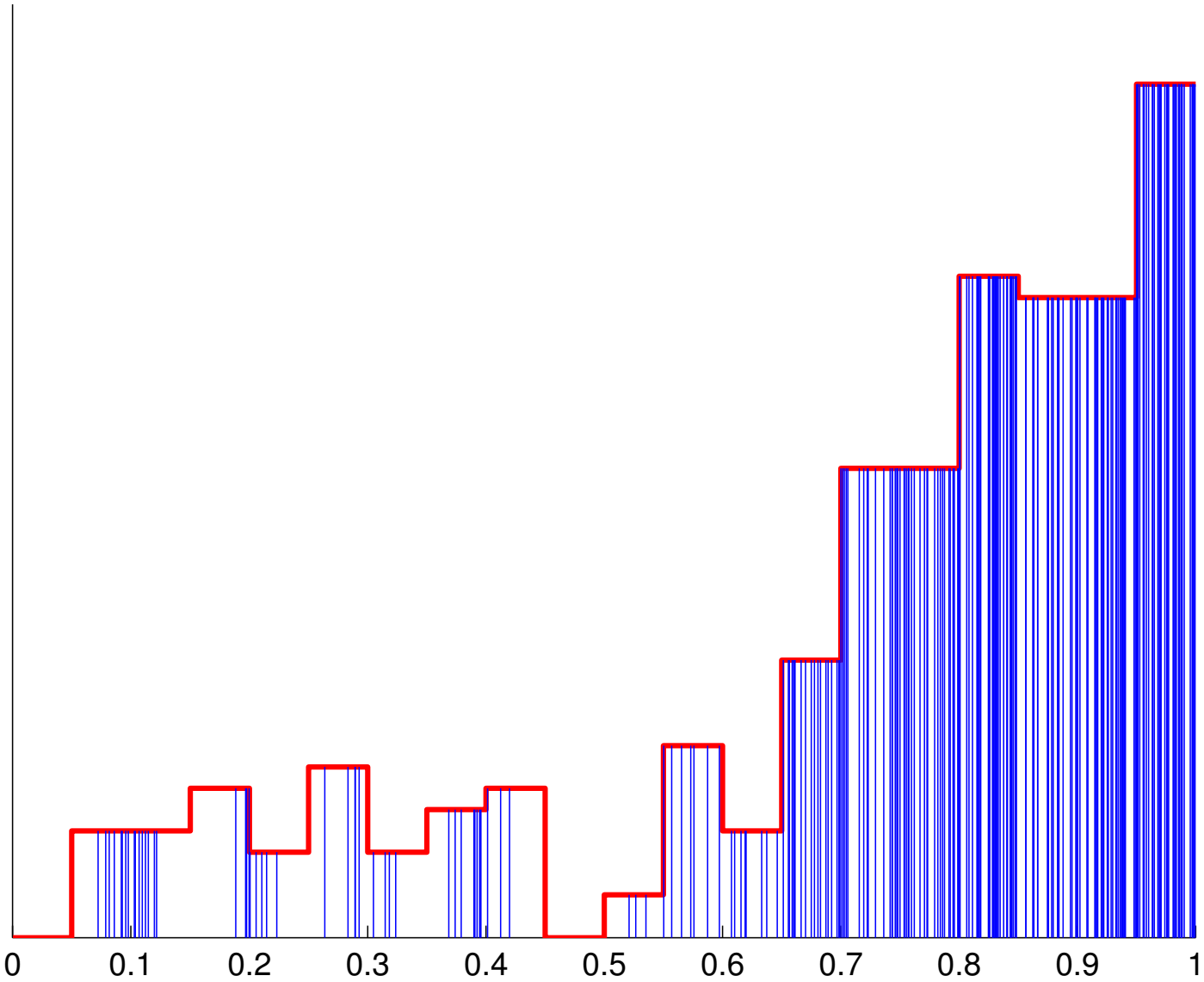
Principe de la validation : échantillon d'entraînement



Principe de la validation : échantillon de validation



Principe de la validation : échantillon de validation



Validation croisée

$$\underbrace{\xi_1, \ldots, \xi_{n_e}}$$

Entraînement $D_n^{(e)} \Rightarrow \widehat{s}_m^{(e)} = \widehat{s}_m\left(D_n^{(e)}\right)$

$$\underbrace{\xi_{n_e+1}, \ldots, \xi_n}$$

Validation $D_n^{(v)} \Rightarrow$ évaluer le risque

Validation croisée

$\underbrace{\xi_1, \ldots, \xi_{n_e}}$ $\qquad\qquad$ $\underbrace{\xi_{n_e+1}, \ldots, \xi_n}$

Entraînement $D_n^{(e)} \Rightarrow \widehat{s}_m^{(e)} = \widehat{s}_m\left(D_n^{(e)}\right)$ $\qquad$ Validation $D_n^{(v)} \Rightarrow$ évaluer le risque

- Estimateur « Hold-out » du risque :

$$P_n^{(v)} \gamma\left(\widehat{s}_m^{(e)}\right) = \frac{1}{n_v} \sum_{\xi \in D_n^{(v)}} \gamma\left(\widehat{s}_m^{(e)}; \xi\right) \qquad n_v = |D_n^{(v)}| = n - n_e$$

Validation croisée

$\underbrace{\xi_1, \ldots, \xi_{n_e}}$ $\qquad$ $\underbrace{\xi_{n_e+1}, \ldots, \xi_n}$

Entraînement $D_n^{(e)} \Rightarrow \widehat{s}_m^{(e)} = \widehat{s}_m\left(D_n^{(e)}\right)$ $\qquad$ Validation $D_n^{(v)} \Rightarrow$ évaluer le risque

- Estimateur « Hold-out » du risque :

$$P_n^{(v)} \gamma\left(\widehat{s}_m^{(e)}\right) = \frac{1}{n_v} \sum_{\xi \in D_n^{(v)}} \gamma\left(\widehat{s}_m^{(e)}; \xi\right) \qquad n_v = |D_n^{(v)}| = n - n_e$$

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

$$\widehat{\mathcal{R}}^{\mathrm{vc}}\left(\widehat{s}_m; D_n; (I_j^{(e)})_{1 \leqslant j \leqslant B}\right) = \frac{1}{B} \sum_{j=1}^{B} P_n^{(v,j)} \gamma\left(\widehat{s}_m^{(e,j)}\right) \qquad D_n^{(e,j)} = (\xi_i)_{i \in I_j^{(e)}}$$

Validation croisée

$\underbrace{\xi_1, \ldots, \xi_{n_e}}$ $\qquad$ $\underbrace{\xi_{n_e+1}, \ldots, \xi_n}$

Entraînement $D_n^{(e)} \Rightarrow \widehat{s}_m^{(e)} = \widehat{s}_m\left(D_n^{(e)}\right)$ $\qquad$ Validation $D_n^{(v)} \Rightarrow$ évaluer le risque

- Estimateur « Hold-out » du risque :

$$P_n^{(v)} \gamma\left(\widehat{s}_m^{(e)}\right) = \frac{1}{n_v} \sum_{\xi \in D_n^{(v)}} \gamma\left(\widehat{s}_m^{(e)}; \xi\right) \qquad n_v = |D_n^{(v)}| = n - n_e$$

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

$$\widehat{\mathcal{R}}^{\mathrm{VC}}\left(\widehat{s}_m; D_n; (I_j^{(e)})_{1 \leqslant j \leqslant B}\right) = \frac{1}{B} \sum_{j=1}^{B} P_n^{(v,j)} \gamma\left(\widehat{s}_m^{(e,j)}\right) \qquad D_n^{(e,j)} = (\xi_i)_{i \in I_j^{(e)}}$$

- Sélection de modèles :

$$\widehat{m} \in \operatorname*{argmin}_{m \in \mathcal{M}} \left\{ \widehat{\mathcal{R}}^{\mathrm{VC}}\left(\widehat{s}_m; D_n\right) \right\} .$$

Validation croisée « V-fold »

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

$\underbrace{(\xi_i)_{i \in B_1}}_{\text{validation}}$, $\underbrace{(\xi_i)_{i \in B_2}, \ldots, (\xi_i)_{i \in B_{V-1}}, (\xi_i)_{i \in B_V}}_{\text{entraînement}}$ $\Rightarrow$ $P_n^{(1)} \gamma\left(\widehat{s}_m^{(-1)}\right)$

Validation croisée « V-fold »

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

$$\underbrace{(\xi_i)_{i \in B_1}}_{\text{validation}} , \underbrace{(\xi_i)_{i \in B_2}, \ldots, (\xi_i)_{i \in B_{V-1}}, (\xi_i)_{i \in B_V}}_{\text{entraînement}} \quad \Rightarrow \quad P_n^{(1)} \gamma\left(\widehat{s}_m^{(-1)}\right)$$

$$\underbrace{(\xi_i)_{i \in B_1}}_{\text{entraîn.}} , \underbrace{(\xi_i)_{i \in B_2}}_{\text{validation}} , \underbrace{\ldots, (\xi_i)_{i \in B_{V-1}}, (\xi_i)_{i \in B_V}}_{\text{entraînement}} \quad \Rightarrow \quad P_n^{(2)} \gamma\left(\widehat{s}_m^{(-2)}\right)$$

Validation croisée « V-fold »

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

$\underbrace{(\xi_i)_{i \in B_1}}_{\text{validation}}$, $\underbrace{(\xi_i)_{i \in B_2}, \ldots, (\xi_i)_{i \in B_{V-1}}, (\xi_i)_{i \in B_V}}_{\text{entraînement}}$ $\Rightarrow$ $P_n^{(1)} \gamma\big(\widehat{s}_m^{(-1)}\big)$

$\underbrace{(\xi_i)_{i \in B_1}}_{\text{entraîn.}}$, $\underbrace{(\xi_i)_{i \in B_2}}_{\text{validation}}$, $\underbrace{\ldots, (\xi_i)_{i \in B_{V-1}}, (\xi_i)_{i \in B_V}}_{\text{entraînement}}$ $\Rightarrow$ $P_n^{(2)} \gamma\big(\widehat{s}_m^{(-2)}\big)$

$\vdots$

$\underbrace{(\xi_i)_{i \in B_1}, (\xi_i)_{i \in B_2}, \ldots, (\xi_i)_{i \in B_{V-1}}}_{\text{entraînement}}$, $\underbrace{(\xi_i)_{i \in B_V}}_{\text{validation}}$ $\Rightarrow$ $P_n^{(V)} \gamma\big(\widehat{s}_m^{(-V)}\big)$

Validation croisée « V-fold »

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

$$\underbrace{(\xi_i)_{i \in B_1}}_{\text{validation}} , \underbrace{(\xi_i)_{i \in B_2}, \ldots, (\xi_i)_{i \in B_{V-1}}, (\xi_i)_{i \in B_V}}_{\text{entraînement}} \quad \Rightarrow \quad P_n^{(1)} \gamma\left(\widehat{s}_m^{(-1)}\right)$$

$$\underbrace{(\xi_i)_{i \in B_1}}_{\text{entraîn.}} , \underbrace{(\xi_i)_{i \in B_2}}_{\text{validation}} , \underbrace{\ldots, (\xi_i)_{i \in B_{V-1}}, (\xi_i)_{i \in B_V}}_{\text{entraînement}} \quad \Rightarrow \quad P_n^{(2)} \gamma\left(\widehat{s}_m^{(-2)}\right)$$

$$\vdots$$

$$\underbrace{(\xi_i)_{i \in B_1}, (\xi_i)_{i \in B_2}, \ldots, (\xi_i)_{i \in B_{V-1}}}_{\text{entraînement}}, \underbrace{(\xi_i)_{i \in B_V}}_{\text{validation}} \quad \Rightarrow \quad P_n^{(V)} \gamma\left(\widehat{s}_m^{(-V)}\right)$$

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

Analyse au premier ordre : lemme

- Objectif :

$$\text{minimiser } \mathcal{R}(m) = P\big(\widehat{s}_m(D_n)\big) \text{ sur } m \in \mathcal{M}$$

- Méthode :

$$\text{minimiser } \mathcal{C}(m) = \widehat{\mathcal{R}}^{\mathrm{vf}}\left(\widehat{s}_m; D_n; \mathcal{B}\right) \text{ sur } m \in \mathcal{M}$$

$$\Rightarrow \qquad \widehat{m}_{\mathcal{C}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \big\{\mathcal{C}(m)\big\}$$

Analyse au premier ordre : lemme

- Objectif :

$$\text{minimiser } \mathcal{R}(m) = P\big(\widehat{s}_m(D_n)\big) \text{ sur } m \in \mathcal{M}$$

- Méthode :

$$\text{minimiser } \mathcal{C}(m) = \widehat{\mathcal{R}}^{\text{vf}}\left(\widehat{s}_m; D_n; \mathcal{B}\right) \text{ sur } m \in \mathcal{M}$$
$$\Rightarrow \qquad \widehat{m}_{\mathcal{C}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \big\{\mathcal{C}(m)\big\}$$

Lemme

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

alors, $\quad \forall \widehat{m}_{\mathcal{C}} \in \operatorname*{argmin}_{m \in \mathcal{M}} \big\{\mathcal{C}(m)\big\},$

$$\mathcal{R}(\widehat{m}_{\mathcal{C}}) - B(\widehat{m}_{\mathcal{C}}) \leqslant \inf_{m \in \mathcal{M}} \big\{\mathcal{R}(m) + A(m)\big\} \ .$$

Optimalité au premier ordre

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

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

Optimalité au premier ordre

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

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

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

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

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

Optimalité au premier ordre

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

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

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

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

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

$\Rightarrow$ d'après le lemme,

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

Optimalité au premier ordre

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

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

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

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

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

$\Rightarrow$ d'après le lemme,

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

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

Au premier ordre : biais de la validation croisée

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

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

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

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

Au premier ordre : biais de la validation croisée

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

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

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

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

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

Au premier ordre : biais de la validation croisée

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

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

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

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

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

$\Rightarrow$ sous-optimalité de la validation croisée « V-fold » à V fixé (A. 2008, régressogrammes ; valable plus largement)

Correction du biais et pénalisation « V-fold »

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

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

Correction du biais et pénalisation « V-fold »

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

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

$$= P_n\gamma\left(\widehat{s}_m\right) + \underbrace{\mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n; \mathcal{B})}_{\text{pénalité } V\text{-fold (A. 2008)}}$$

Correction du biais et pénalisation « V-fold »

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

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

$$= P_n\gamma\left(\widehat{s}_m\right) + \underbrace{\mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n; \mathcal{B})}_{\text{pénalité } V\text{-fold (A. 2008)}}$$

- Estimation de densité, moindres carrés (A. & Lerasle, 2014) :

$$\widehat{\mathcal{R}}^{\mathrm{vf}}(\widehat{s}_m; D_n; \mathcal{B}) = P_n\gamma\big(\widehat{s}_m(D_n)\big) + \underbrace{\left(1 + \frac{1}{2(V-1)}\right)}_{\text{surpénalisation}} \mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n; \mathcal{B})$$

$$\widehat{\mathcal{R}}^{\ell\mathrm{po}}(\widehat{s}_m; D_n; \mathcal{B}) = P_n\gamma\big(\widehat{s}_m(D_n)\big) + \overbrace{\left(1 + \frac{1}{2\left(\frac{n}{p}-1\right)}\right)}^{\text{surpénalisation}} \mathrm{pen}_{\mathrm{VF}}(\widehat{s}_m; D_n; \mathcal{B}_{\mathrm{loo}})$$

Inégalités oracle optimales pour la pénalisation « V-fold »

Théorème

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

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

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

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

Inégalités oracle optimales pour la pénalisation « V-fold »

Théorème

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

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

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

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

Valable sous des hypothèses raisonnablement faibles pour :

- Les régressogrammes en régression hétéroscédastique (A. 2008, 2009)
- L'estimation de densité par moindres carrés (A. & Lerasle, 2014 ; Celisse, 2014)

Inégalités oracle optimales pour la pénalisation « V-fold »

Théorème

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

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

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

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

Valable sous des hypothèses raisonnablement faibles pour :

- Les régressogrammes en régression hétéroscédastique (A. 2008, 2009)
- L'estimation de densité par moindres carrés (A. & Lerasle, 2014 ; Celisse, 2014)

\+ Inégalité-oracle sous-optimale pour la validation croisée « V-fold » (constante $1 + \frac{1}{V-1}$).

Analyse au second ordre ?

- Comment comparer $\widehat{m}_{\mathcal{C}_1}$ et $\widehat{m}_{\mathcal{C}_2}$ lorsque

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

Analyse au second ordre ?

- Comment comparer $\widehat{m}_{\mathcal{C}_1}$ et $\widehat{m}_{\mathcal{C}_2}$ lorsque

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

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

Analyse au second ordre ?

- Comment comparer $\widehat{m}_{\mathcal{C}_1}$ et $\widehat{m}_{\mathcal{C}_2}$ lorsque

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

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

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

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

Analyse au second ordre ?

- Comment comparer $\widehat{m}_{\mathcal{C}_1}$ et $\widehat{m}_{\mathcal{C}_2}$ lorsque

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

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

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

  mais $\mathrm{var}\big(\mathcal{C}(m) + Z\big)$ dépend de Z...
- Ce qui compte :

$$\forall m, m' \in \mathcal{M}, \qquad \mathrm{sign}\big(\mathcal{C}(m) - \mathcal{C}(m')\big) = \mathrm{sign}\big(\mathcal{R}(m) - \mathcal{R}(m')\big)$$

Analyse au second ordre ?

- Comment comparer $\widehat{m}_{\mathcal{C}_1}$ et $\widehat{m}_{\mathcal{C}_2}$ lorsque

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

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

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

  mais $\mathrm{var}\big(\mathcal{C}(m) + Z\big)$ dépend de Z...
- Ce qui compte :

$$\forall m, m' \in \mathcal{M}, \qquad \mathrm{sign}\big(\mathcal{C}(m) - \mathcal{C}(m')\big) = \mathrm{sign}\big(\mathcal{R}(m) - \mathcal{R}(m')\big)$$

$\Rightarrow$ variance des incréments

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

Variance et sélection de modèles : heuristique

- $\forall m \notin \operatorname{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right]$, on veut minimiser

$\mathbb{P}(\widehat{m}_{\mathcal{C}} = m)$

Variance et sélection de modèles : heuristique

- $\forall m \notin \operatorname{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right]$, on veut minimiser

$$\mathbb{P}(\widehat{m}_{\mathcal{C}} = m) = \mathbb{P}\big(\forall m' \in \mathcal{M}, \mathcal{C}(m) - \mathcal{C}(m') < 0\big)$$
$$\leqslant \min_{m' \neq m} \mathbb{P}\big(\mathcal{C}(m) - \mathcal{C}(m') < 0\big)$$

Variance et sélection de modèles : heuristique

- $\forall m \notin \operatorname{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right]$, on veut minimiser

$$\begin{aligned}
\mathbb{P}(\widehat{m}_{\mathcal{C}} = m) &= \mathbb{P}\big(\forall m' \in \mathcal{M}, \mathcal{C}(m) - \mathcal{C}(m') < 0\big) \\
&\leqslant \min_{m' \neq m} \mathbb{P}\big(\mathcal{C}(m) - \mathcal{C}(m') < 0\big) \\
&\approx \min_{m' \neq m} \mathbb{P}\left(\mathbb{E}\left[\mathcal{C}(m) - \mathcal{C}(m')\right] - \mathcal{N}\sqrt{\operatorname{var}\big(\mathcal{C}(m) - \mathcal{C}(m')\big)} < 0\right)
\end{aligned}$$

Variance et sélection de modèles : heuristique

- $\forall m \notin \operatorname{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right]$, on veut minimiser

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

Variance et sélection de modèles : heuristique

- $\forall m \notin \operatorname{argmin}_{m \in \mathcal{M}} \mathbb{E}\left[\ell\left(s^\star, \widehat{s}_m\right)\right]$, on veut minimiser

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

- Hypothèses : $\forall m \in \mathcal{M}$, $\mathbb{E}[\mathcal{C}_1(m)] = \mathbb{E}[\mathcal{C}_2(m)]$ et $\forall i$, $\operatorname{argmin}_{m \in \mathcal{M}} \mathbb{E}[\mathcal{C}_i(m)] = \operatorname{argmin}_{m \in \mathcal{M}} \mathbb{E}[\mathcal{R}(m)]$

  $\Rightarrow$ on veut minimiser $\operatorname{var}\left(\mathcal{C}_i(m) - \mathcal{C}_i(m')\right)$

Variance et sélection de modèles (densité, moindres carrés)

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

Théorème (A. & Lerasle, 2014)

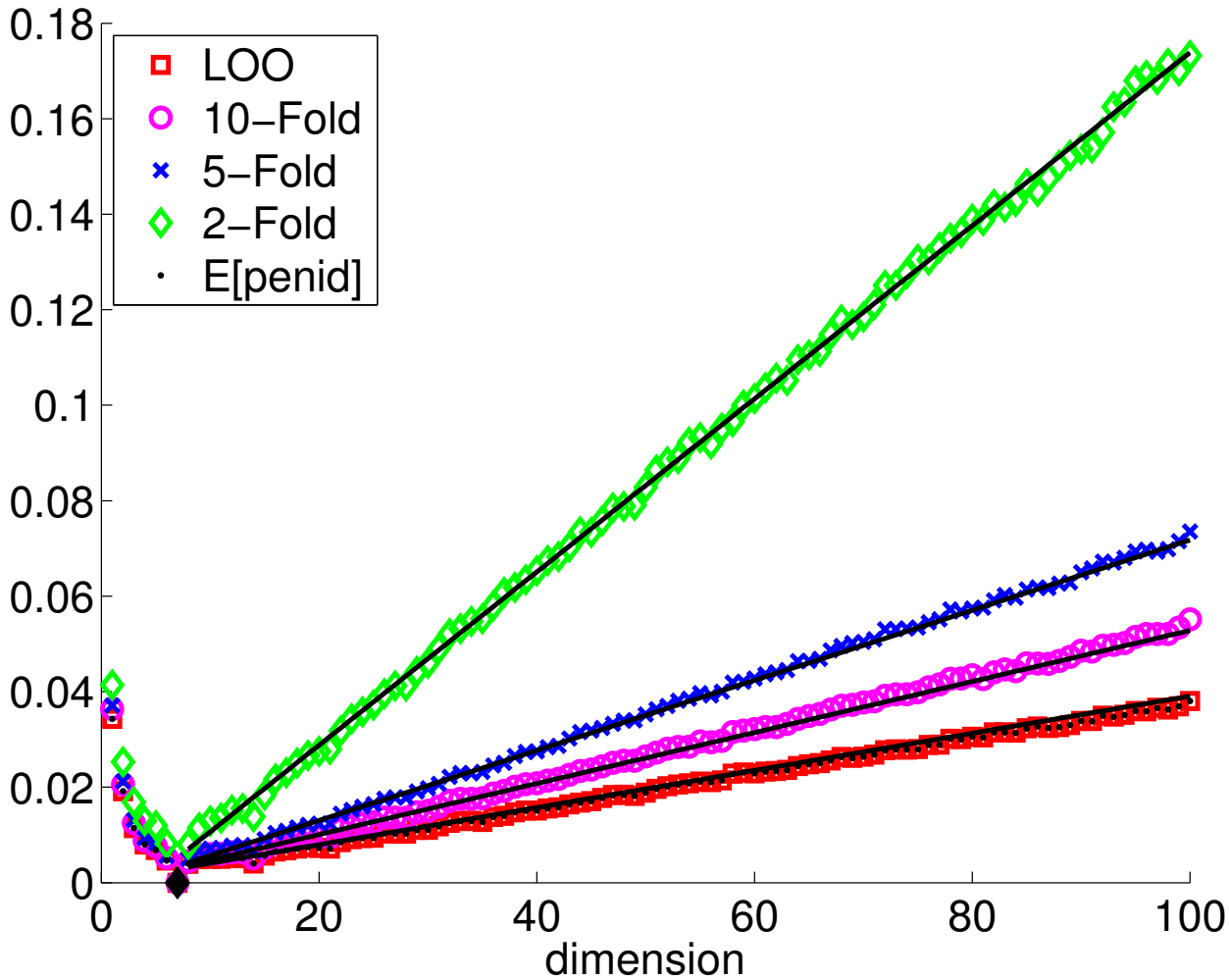
$$\mathrm{var}\big(\Delta(m, m', V)\big) = 4\left(1 + \frac{2}{n} + \frac{1}{n^2}\right)\frac{\mathrm{var}_P(s^\star_m - s^\star_{m'})}{n} + 2\left(1 + \frac{4}{V-1} - \frac{1}{n}\right)\underbrace{\frac{B(m, m')}{n^2}}_{\geqslant 0}$$

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

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

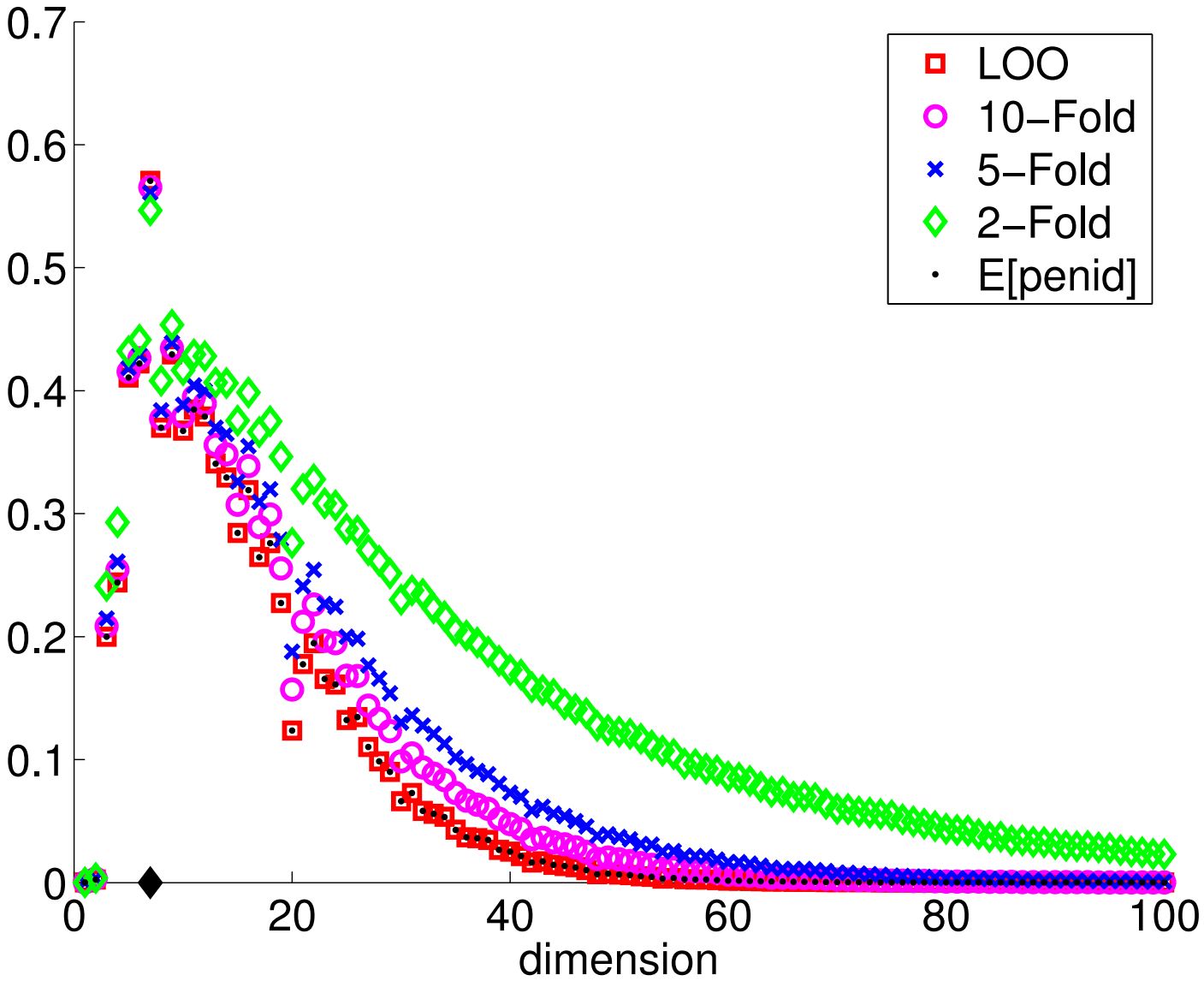
Les deux termes sont du même ordre si $\left\|s^\star_m - s^\star_{m'}\right\| \approx d_m/n$.

Variance de $\widehat{\mathcal{R}}^{\mathrm{vf,corr}}(\widehat{s}_m) - \widehat{\mathcal{R}}^{\mathrm{vf,corr}}(\widehat{s}_{m^\star})$ vs. (d_m, V)

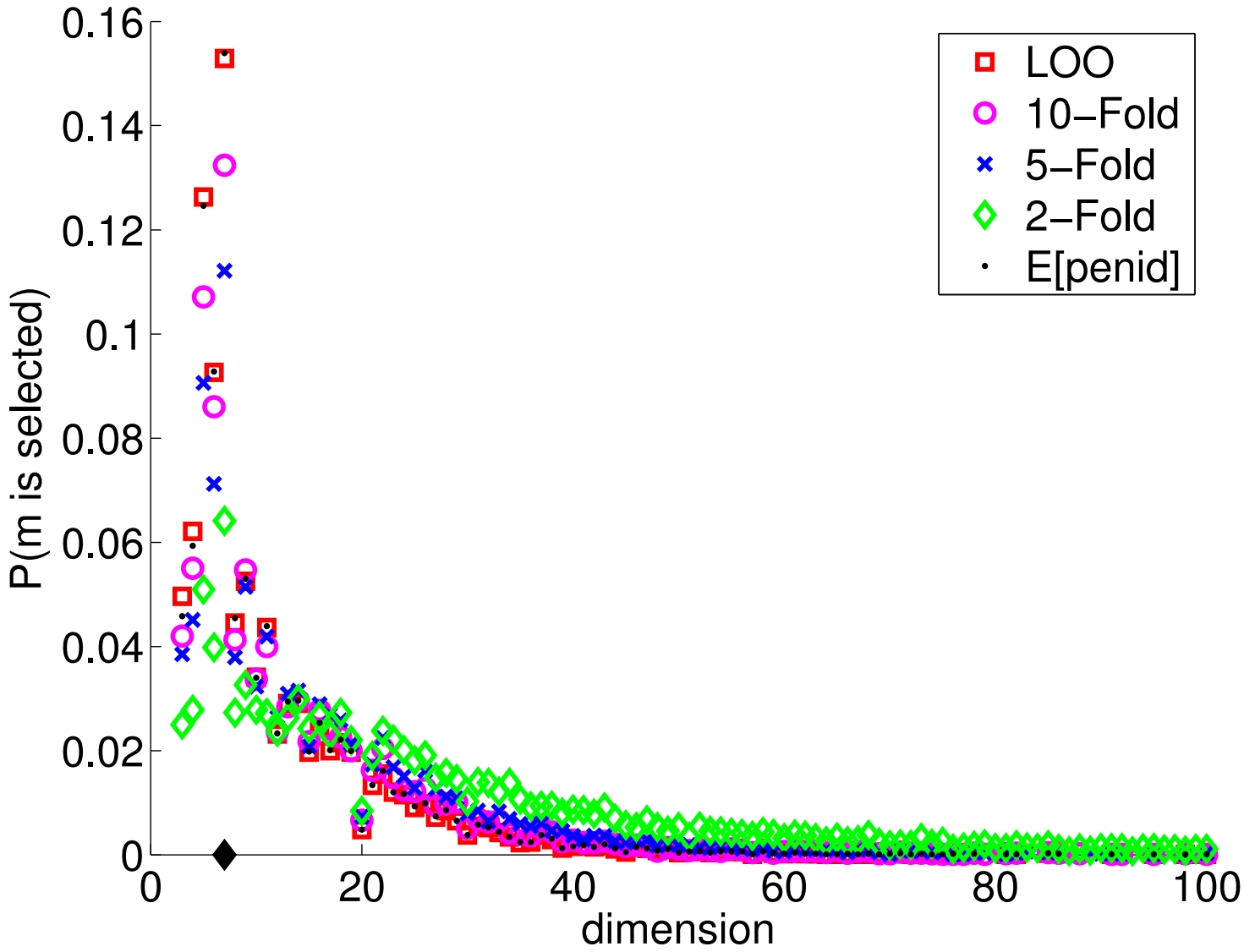


$$\mathrm{var}(\Delta(m, m', V)) \approx n^{-2}\left[29\left(1 + \tfrac{0.8}{V-1}\right) + 3.7\left(1 + \tfrac{3.8}{V-1}\right)(d_m - d_{m^\star})\right]$$

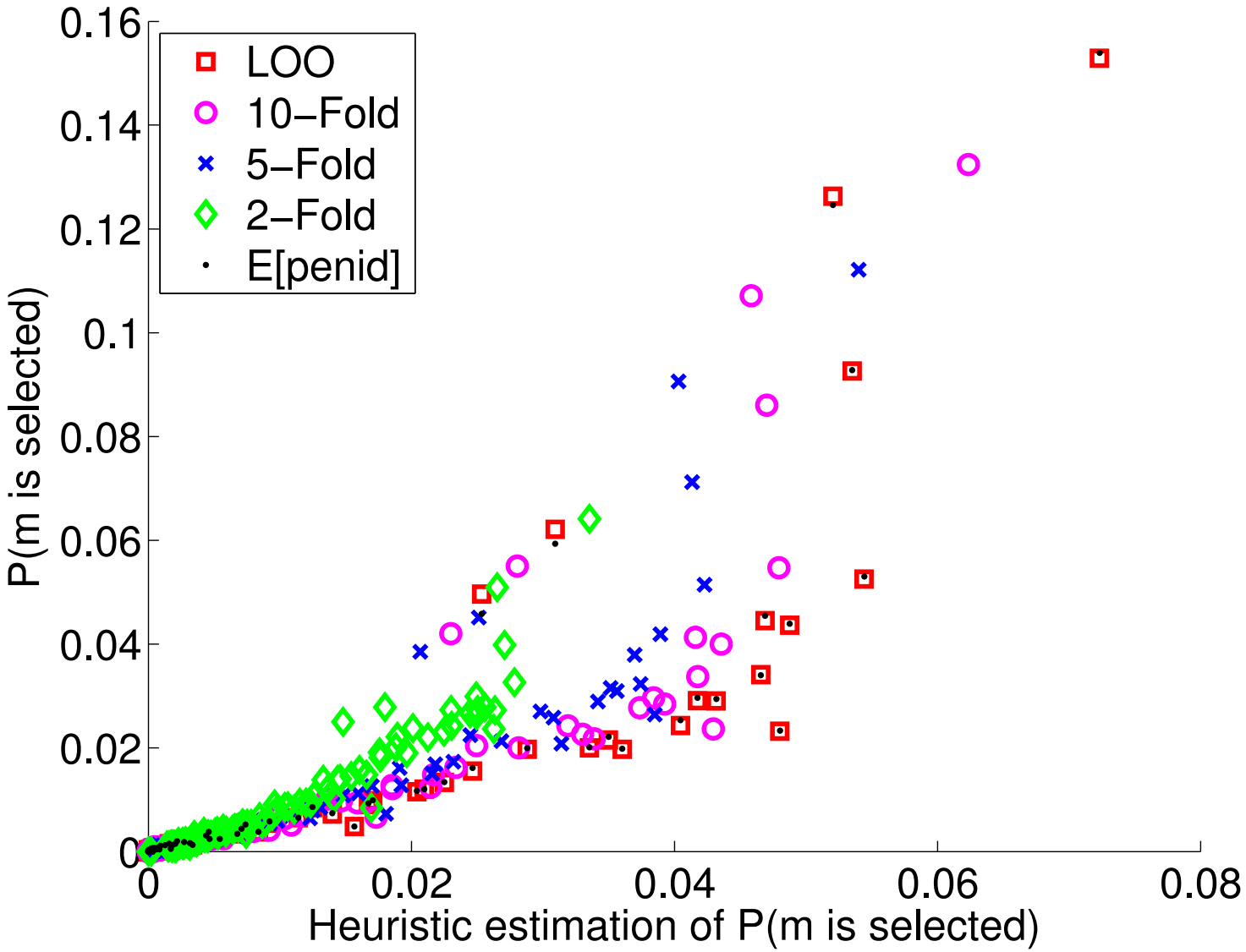
$\overline{\Phi}(\max_{m' \neq m} \mathbb{E}[\mathcal{C}(m) - \mathcal{C}(m')] / \sqrt{\mathrm{var}(\mathcal{C}(m) - \mathcal{C}(m'))})$



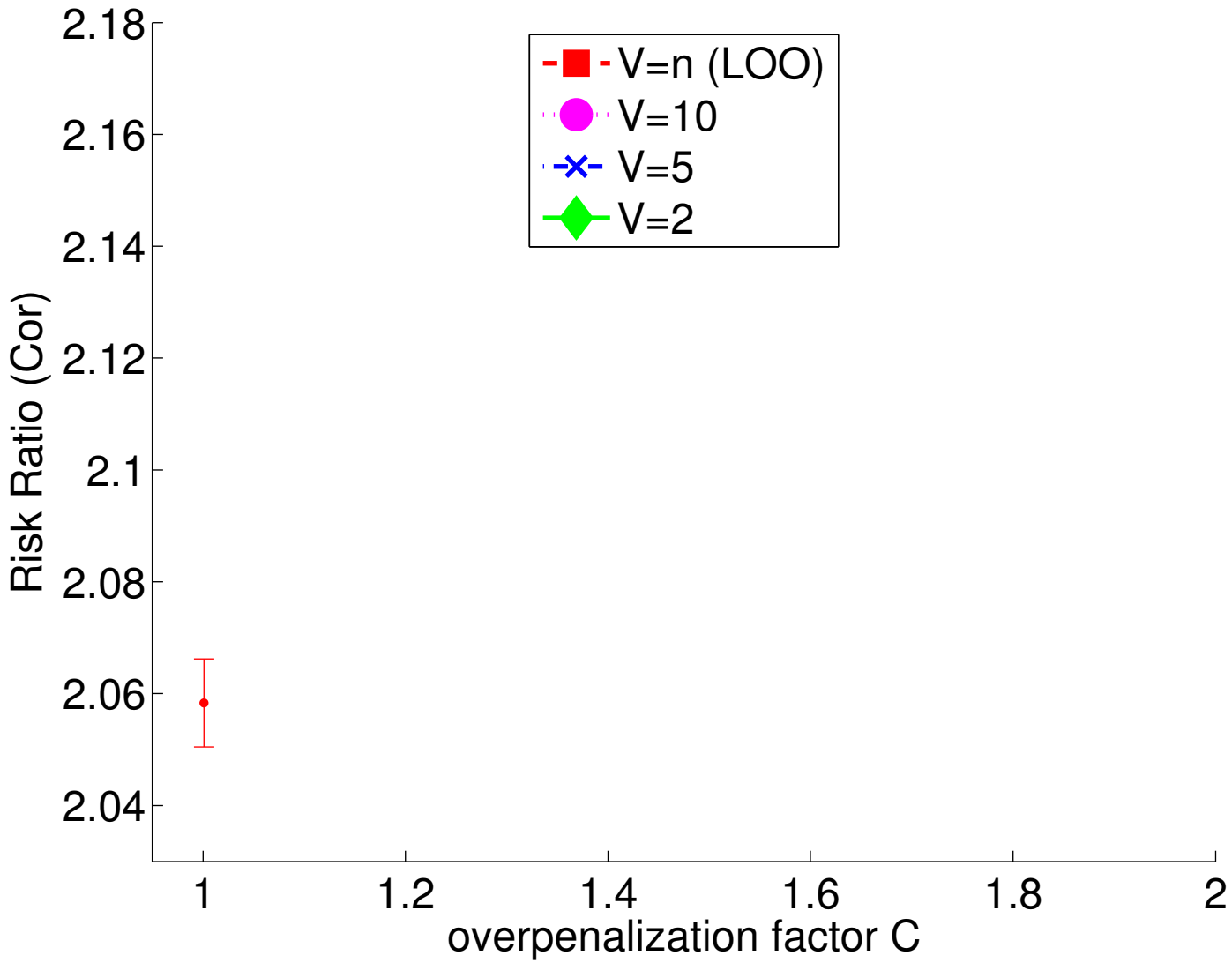
Évaluation de l'heuristique : $\mathbb{P}(\widehat{m}_{\mathcal{C}} = m)$



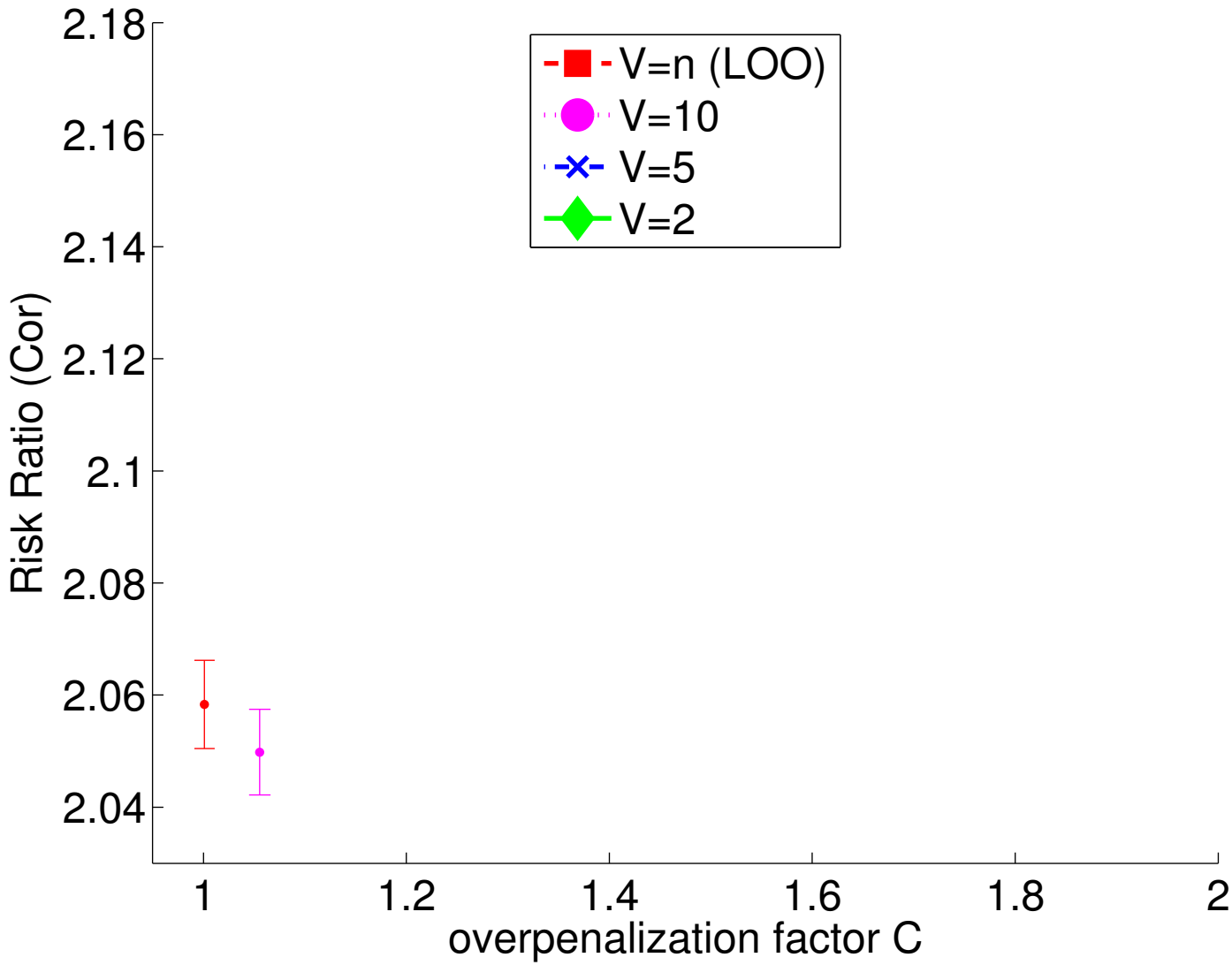
Évaluation de l'heuristique (2)



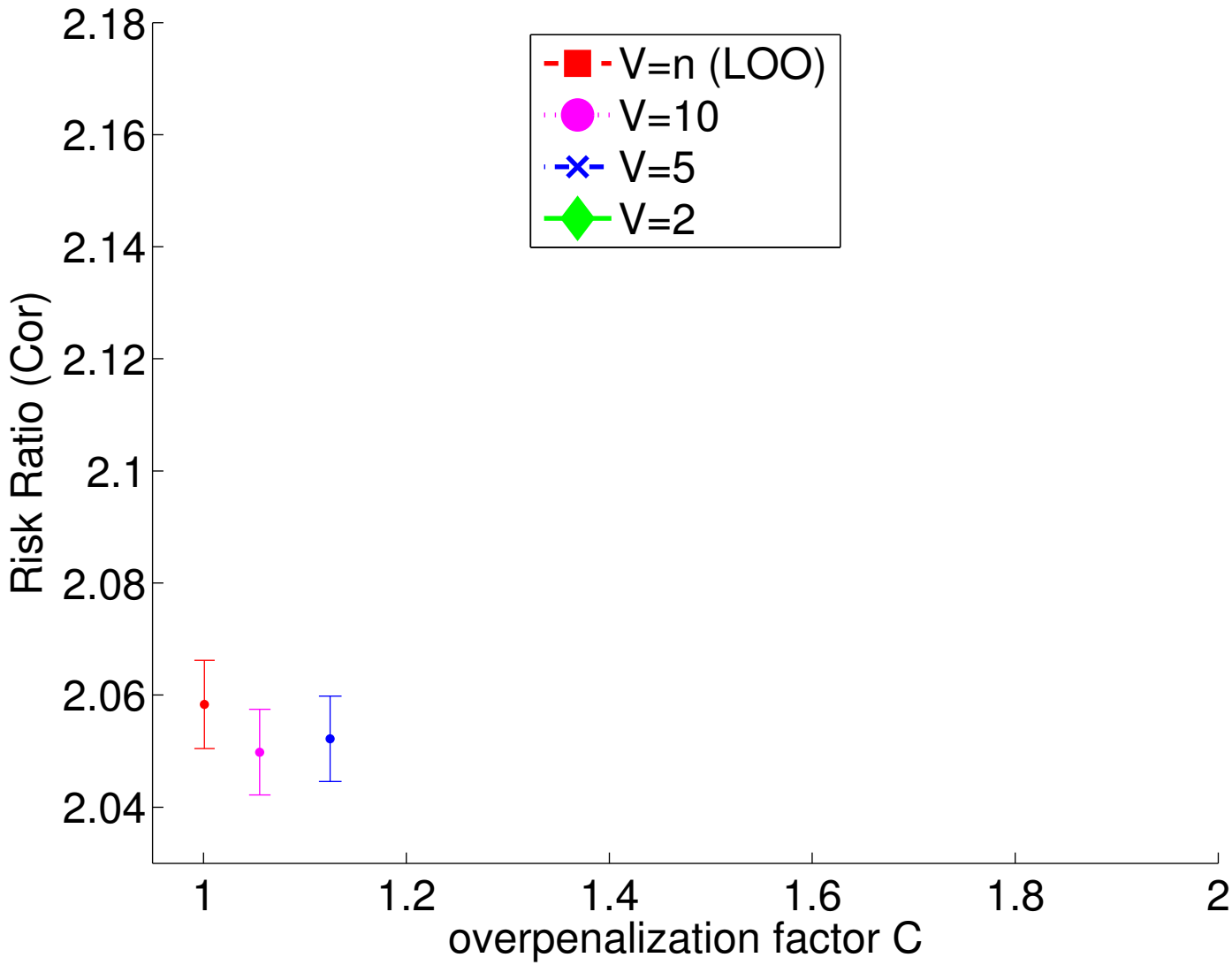
Simulations : validation croisée « V-fold »



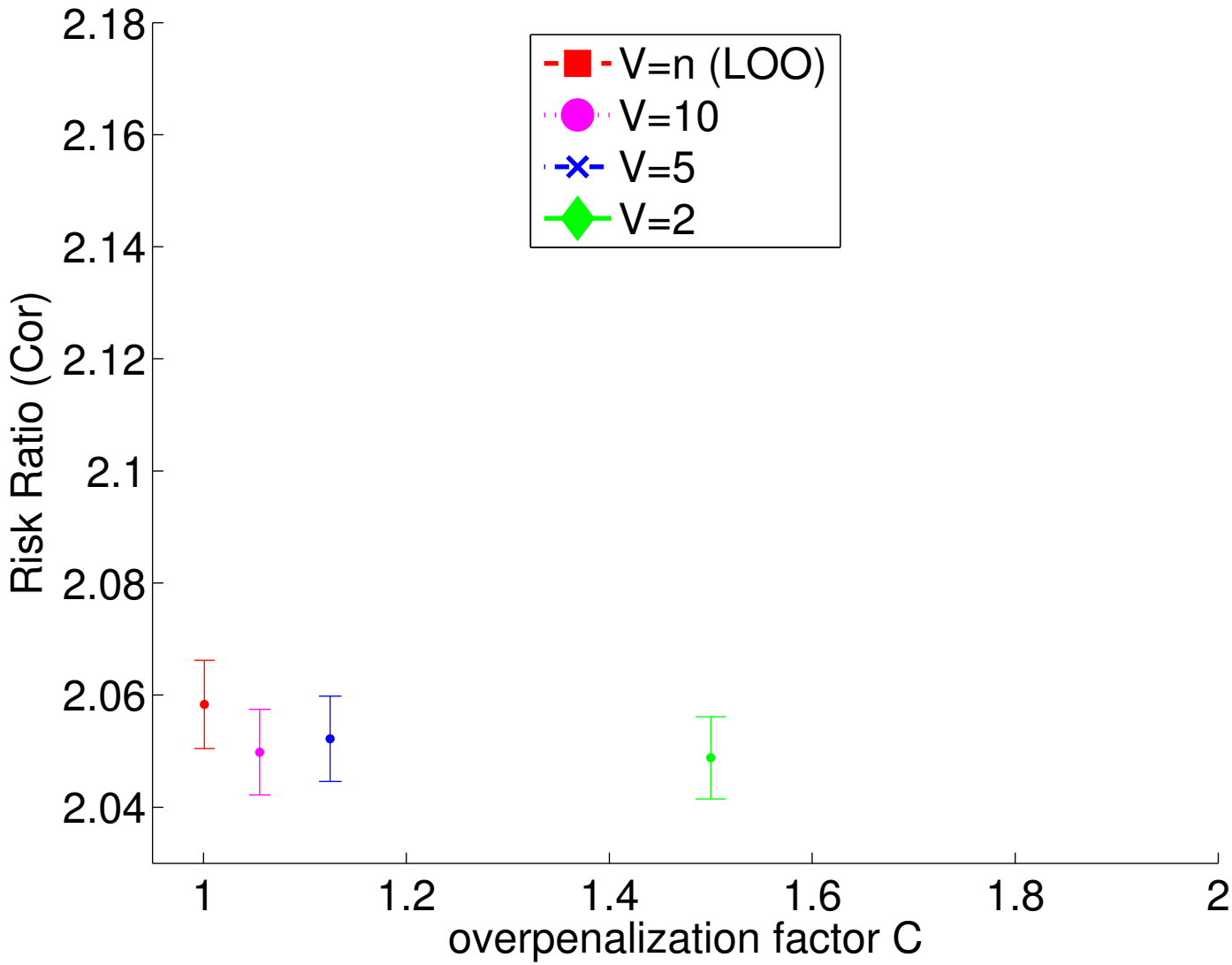
Simulations : validation croisée « V-fold »



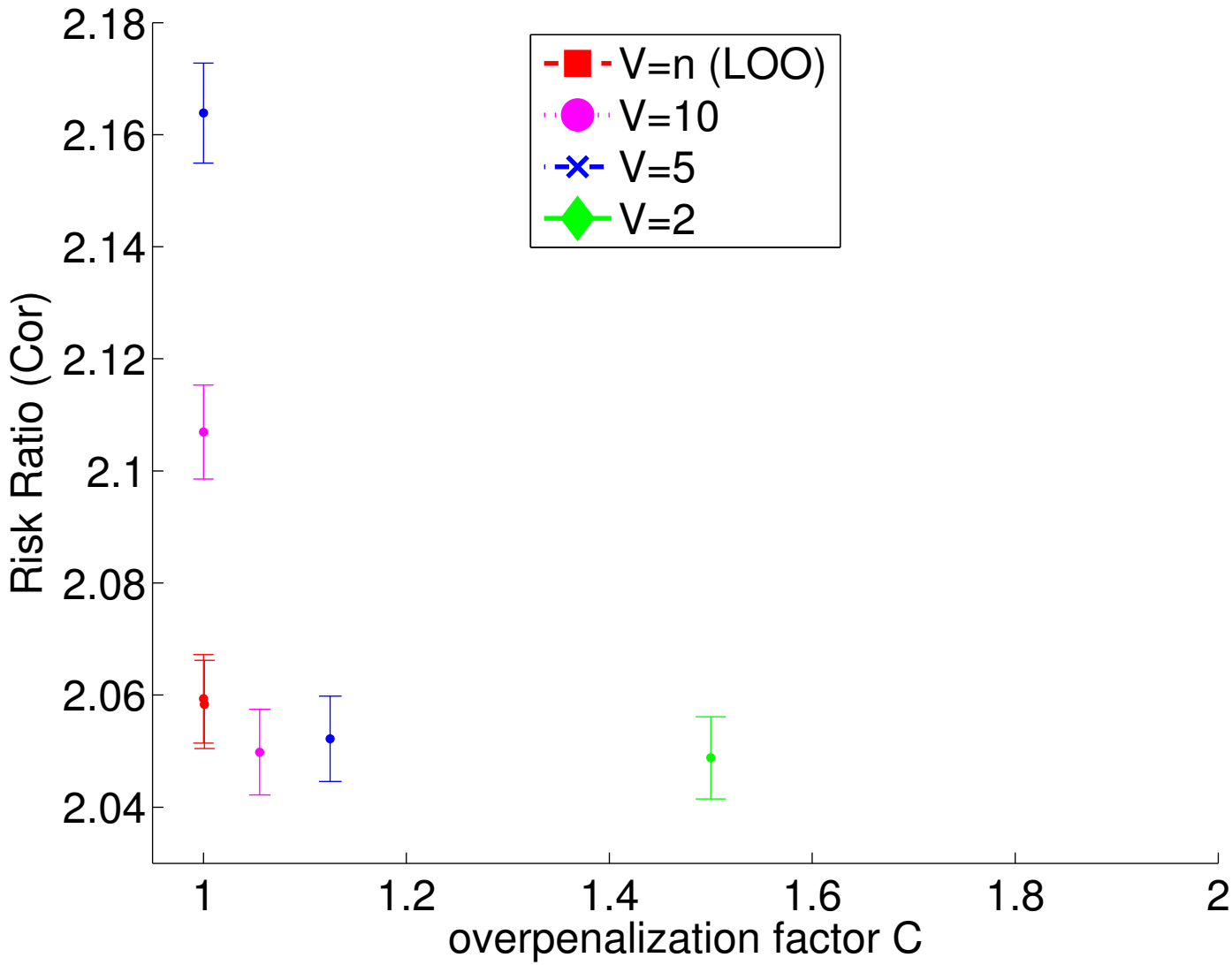
Simulations : validation croisée « V-fold »



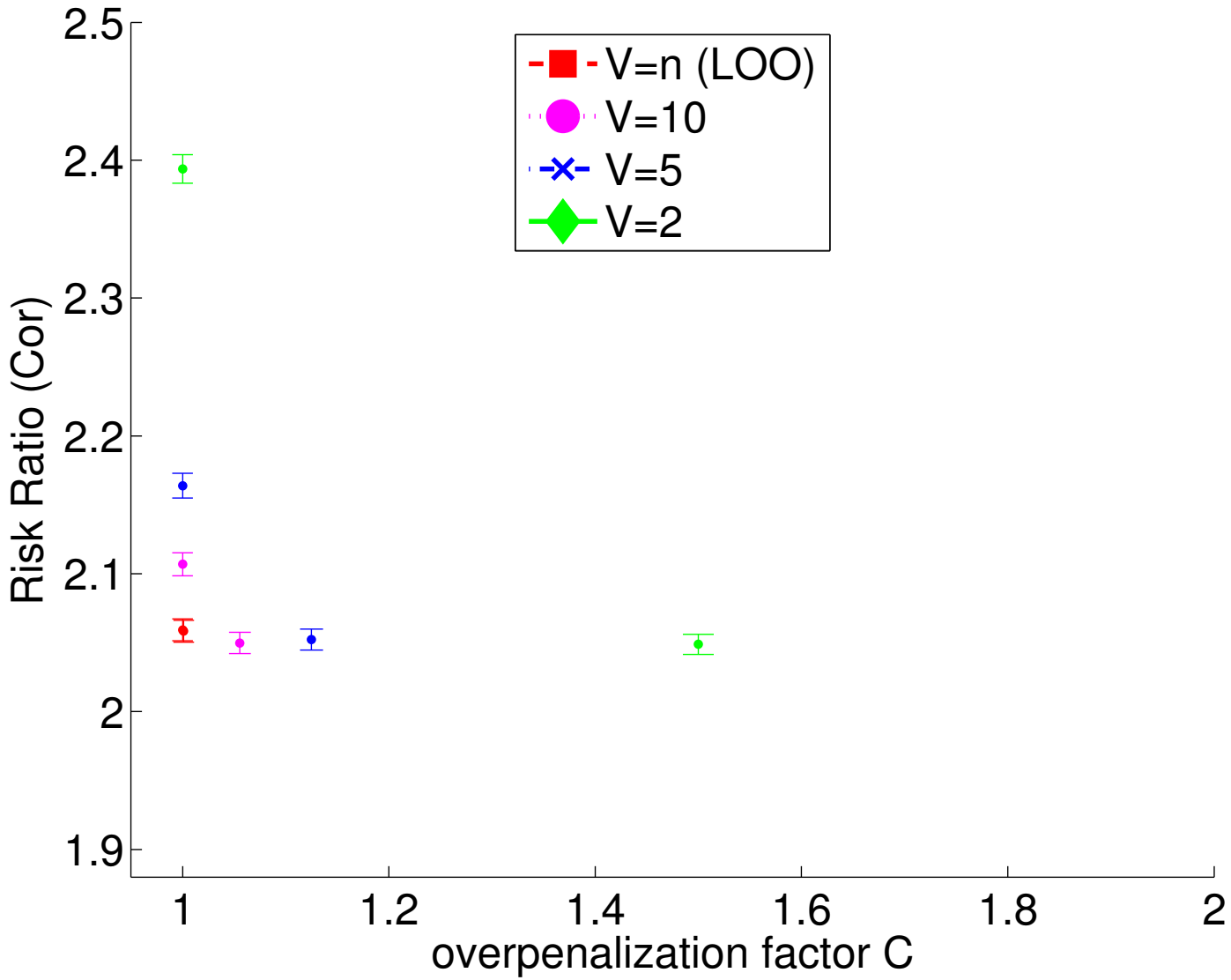
Simulations : validation croisée « V-fold »



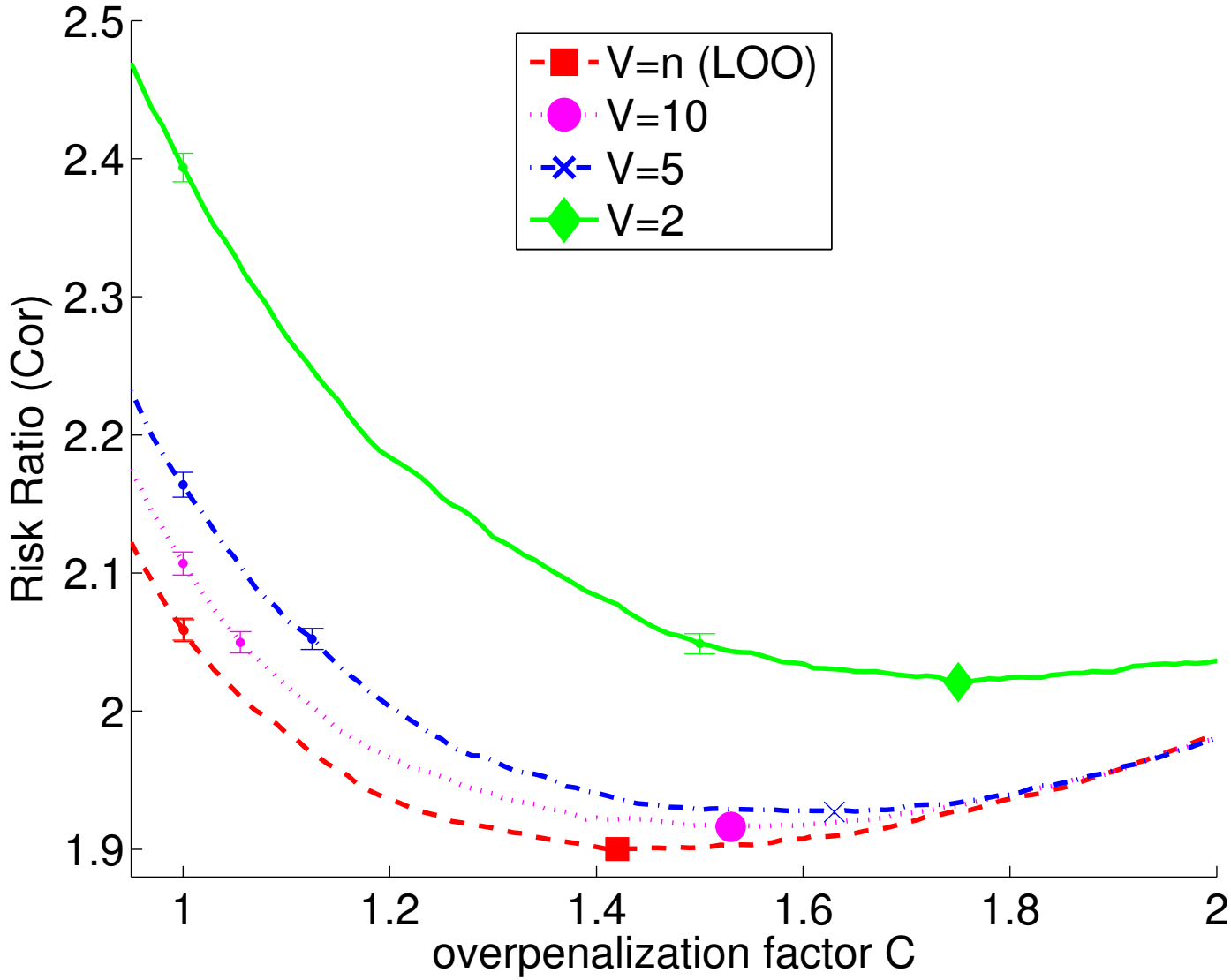
Simulations : pénalisation « V-fold »



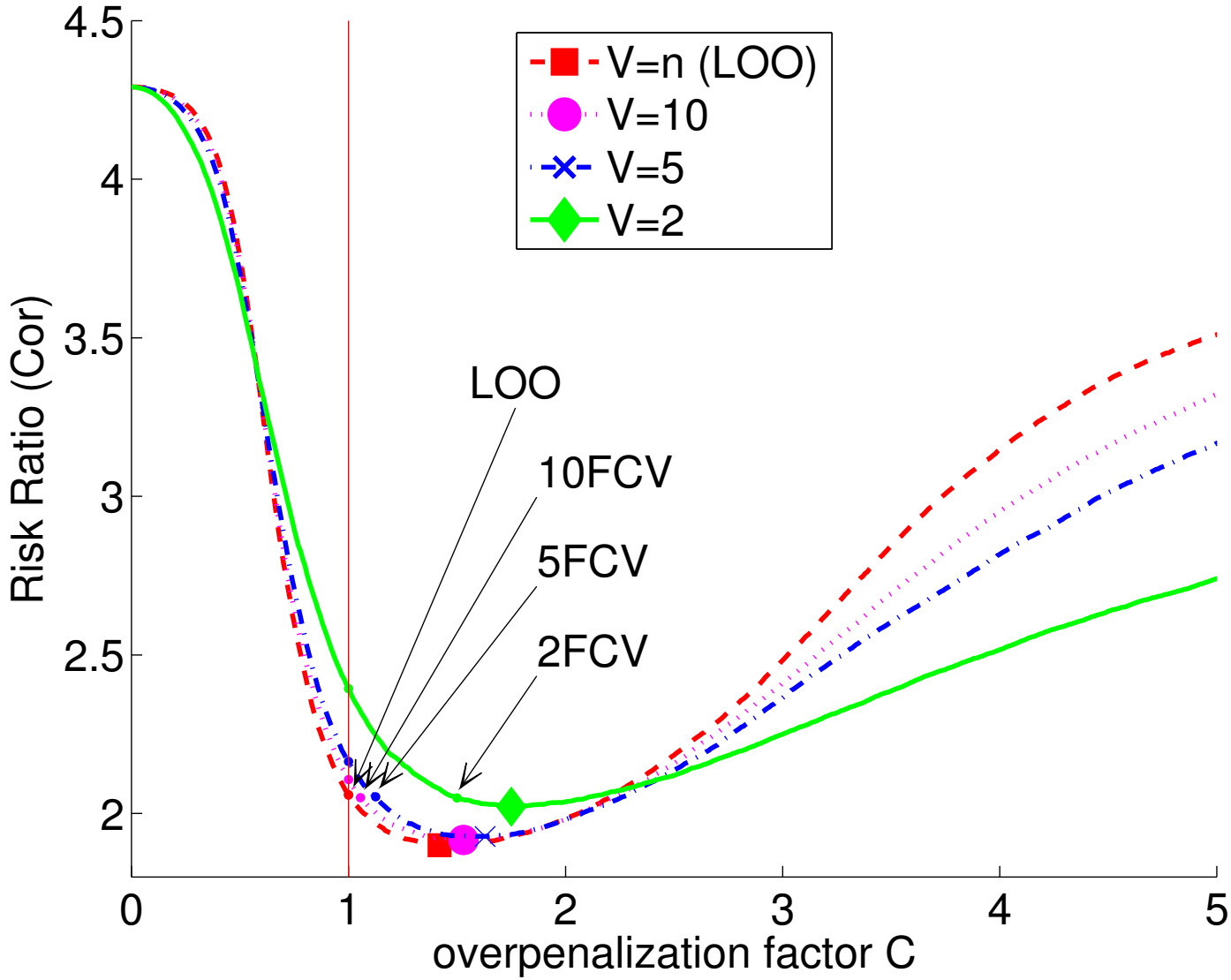
Simulations : pénalisation « V-fold »



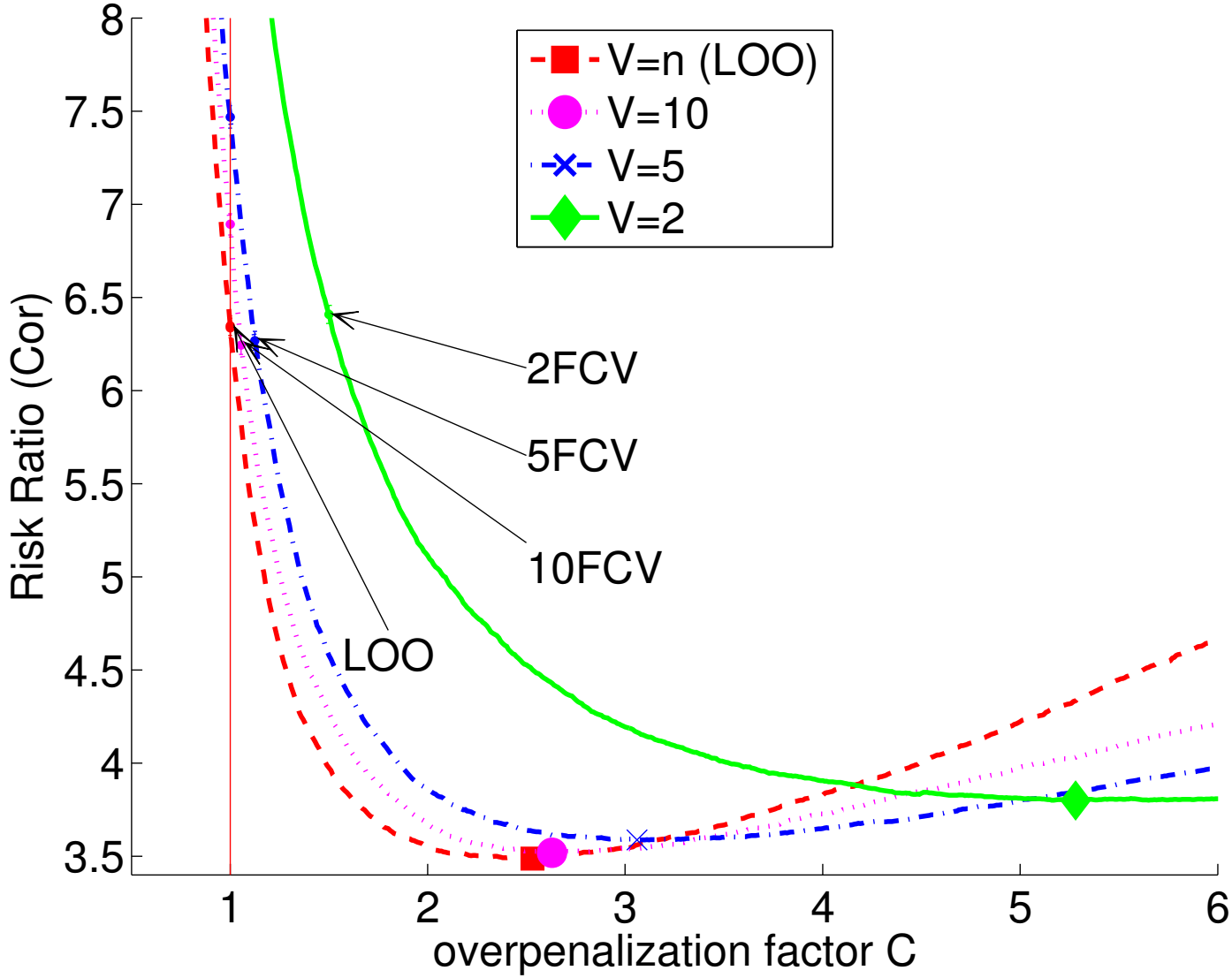
Simulations : surpénalisation



Simulations : conclusion



Simulations : cadre « pratiquement paramétrique »



« V-fold » et sélection de modèles : conclusion

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

« V-fold » et sélection de modèles : conclusion

- Temps de calcul : $\mathcal{O}(V)$ en général
- Validation croisée « V-fold » :
  - Biais : diminue avec V / peut être supprimé
  - Variance : diminue avec V / quasi minimal pour $V \in [5, 10]$

  $\Rightarrow$ performace optimale lorsque V est maximal, quasi optimale pour $V = 10$...

« V-fold » et sélection de modèles : conclusion

- Temps de calcul : $\mathcal{O}(V)$ en général
- Validation croisée « V-fold » :
  - Biais : diminue avec V / peut être supprimé
  - Variance : diminue avec V / quasi minimal pour $V \in [5, 10]$

  $\Rightarrow$ performace optimale lorsque V est maximal, quasi optimale pour $V = 10$...
  ... si le facteur de surpénalisation optimal $C^\star \approx 1$ (nombreux cas possibles).

« V-fold » et sélection de modèles : conclusion

- Temps de calcul : $\mathcal{O}(V)$ en général
- Validation croisée « V-fold » :
  - Biais : diminue avec V / peut être supprimé
  - Variance : diminue avec V / quasi minimal pour $V \in [5, 10]$

  $\Rightarrow$ performace optimale lorsque V est maximal, quasi optimale pour $V = 10$...
  ... si le facteur de surpénalisation optimal $C^\star \approx 1$ (nombreux cas possibles).
- Pénalisation « V-fold » :
  - Découplage entre biais et variance $\Rightarrow$ plus simple à comprendre et utiliser.
  - Biais : choisi directement à travers C, sans contrainte.
  - Variance : diminue avec V / quasi minimale pour $V \in [5, 10]$.

Généralité de ces résultats

- Valable en régression par moindres carrés et en estimation de densité par moindres carrés ou noyaux (travail en cours avec M. Lerasle et N. Magalhães).

Généralité de ces résultats

- Valable en régression par moindres carrés et en estimation de densité par moindres carrés ou noyaux (travail en cours avec M. Lerasle et N. Magalhães).
- Correction du biais / pénalisation « V-fold » : valable lorsque

$$\mathbb{E}\Big[(P - P_n)\gamma\big(\widehat{s}_m\big)\Big] \approx \frac{\gamma(m)}{n} \quad .$$

Sinon : « V-fold » répété ou VC Monte-Carlo avec n_e bien choisi.

Généralité de ces résultats

- Valable en régression par moindres carrés et en estimation de densité par moindres carrés ou noyaux (travail en cours avec M. Lerasle et N. Magalhães).
- Correction du biais / pénalisation « V-fold » : valable lorsque

$$\mathbb{E}\left[(P - P_n)\gamma\left(\widehat{s}_m\right)\right] \approx \frac{\gamma(m)}{n} \quad .$$

  Sinon : « V-fold » répété ou VC Monte-Carlo avec n_e bien choisi.
- Variance : d'autres comportements sont possibles (expérimentalement).

Généralité de ces résultats

- Valable en régression par moindres carrés et en estimation de densité par moindres carrés ou noyaux (travail en cours avec M. Lerasle et N. Magalhães).
- Correction du biais / pénalisation « V-fold » : valable lorsque

$$\mathbb{E}\Big[(P - P_n)\gamma\big(\widehat{s}_m\big)\Big] \approx \frac{\gamma(m)}{n} \quad .$$

  Sinon : « V-fold » répété ou VC Monte-Carlo avec n_e bien choisi.
- Variance : d'autres comportements sont possibles (expérimentalement).
- Tout peut se tester sur des données synthétiques : tracer

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