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, 24/01/2011

Plan

1. Lundi 10 : Apprentissage statistique et sélection d'estimateurs
2. Lundi 17 : Calibration de pénalités et pénalités minimales
3. Aujourd'hui : Rééchantillonnage et pénalisation
4. Lundi 31, 15h30–17h30 : Validation croisée et pénalités reliées

Plan du cours

Plan

Cadre de la régression hétéroscédastique

- Plan d'expérience aléatoire : $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.

$$Y_i = \eta(X_i) + \varepsilon_i$$

$$\mathbb{E}\left[\varepsilon_i \mid X_i\right] = 0 \quad \text{et} \quad \mathbb{E}\left[\varepsilon_i^2 \mid X_i\right] = \sigma^2(X_i)$$

Cadre de la régression hétéroscédastique

- Plan d'expérience aléatoire : $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.

$$Y_i = \eta(X_i) + \varepsilon_i$$

$$\mathbb{E}\left[\varepsilon_i \mid X_i\right] = 0 \quad \text{et} \quad \mathbb{E}\left[\varepsilon_i^2 \mid X_i\right] = \sigma^2(X_i)$$

- Perte quadratique :

$$P\gamma(t) = \mathbb{E}_{(X,Y)\sim P}\left[\gamma(t; (X, Y))\right] = \mathbb{E}_{(X,Y)\sim P}\left[\left(t(X) - Y\right)^2\right]$$

- Perte relative : $\eta = s^\star$ et

$$\ell\left(s^\star, t\right) = P\gamma(t) - P\gamma(s^\star) = \mathbb{E}_{(X,Y)\sim P}\left[\left(s^\star(X) - t(X)\right)^2\right]$$

Régressogrammes

Pour toute partition finie m de $\mathcal{X}$

$$S_m := \left\{ \sum_{\lambda \in m} \alpha_\lambda \mathbb{1}_\lambda \text{ t.q. } \alpha \in \mathbb{R}^m \right\}$$

$\Rightarrow$ estimateur des moindres carrés sur S_m (régressogramme) :

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

Régressogrammes

Pour toute partition finie m de $\mathcal{X}$

$$S_m := \left\{ \sum_{\lambda \in m} \alpha_\lambda \mathbb{1}_\lambda \text{ t.q. } \alpha \in \mathbb{R}^m \right\}$$

$\Rightarrow$ estimateur des moindres carrés sur S_m (régressogramme) :

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

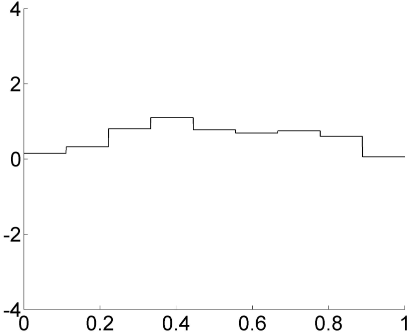
Si pour tout $\lambda \in m$

$$\widehat{p}_\lambda = \widehat{p}_\lambda(D_n) = \frac{1}{n} \operatorname{Card} \{ i \text{ t.q. } X_i \in \lambda \} > 0$$

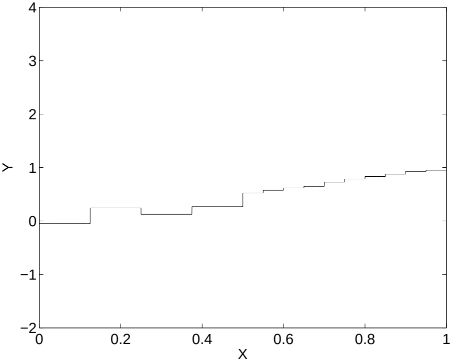
$$\widehat{s}_m = \sum_{\lambda \in m} \widehat{\beta}_\lambda \mathbb{1}_\lambda \qquad \widehat{\beta}_\lambda := \frac{1}{n \widehat{p}_\lambda} \sum_{i \text{ t.q. } X_i \in \lambda} Y_i$$

Régressogrammes : exemples ($\mathcal{X} = [0, 1]$)

$\mathcal{M}_n^{(\mathrm{reg})}$ (partitions régulières)



$\mathcal{M}_n^{(\mathrm{reg},1/2)}$ (partitions régulières sur $[0, 1/2]$ et sur $[1/2, 1]$)



Régressogrammes : biais, pénalité idéale

Régressogrammes : biais, pénalité idéale

$$s_m^\star = \sum_{\lambda \in m} \beta_\lambda \mathbb{1}_\lambda \qquad \beta_\lambda := \mathbb{E}_{(X,Y)\sim P}\left[Y \mid X \in \lambda \right]$$

$$\ell\left(s^\star, s_m^\star\right) = \sum_{\lambda \in m} p_\lambda \left(\sigma_\lambda^{(d)}\right)^2 \qquad \left(\sigma_\lambda^{(d)}\right)^2 := \mathbb{E}\left[\left(\beta_\lambda - s^\star(X)\right)^2 \,\middle|\, X \in \lambda \right]$$

Régressogrammes : biais, pénalité idéale

$$s_m^\star = \sum_{\lambda \in m} \beta_\lambda \mathbb{1}_\lambda \qquad \beta_\lambda := \mathbb{E}_{(X,Y)\sim P}\left[Y \mid X \in \lambda \right]$$

$$\ell\left(s^\star, s_m^\star\right) = \sum_{\lambda \in m} p_\lambda \left(\sigma_\lambda^{(d)}\right)^2 \qquad \left(\sigma_\lambda^{(d)}\right)^2 := \mathbb{E}\left[\left(\beta_\lambda - s^\star(X)\right)^2 \middle| X \in \lambda \right]$$

$$\mathrm{pen}_{\mathrm{id}}(m) = p_1(m) + p_2(m) - \delta(m)$$

$$p_1(m) = P\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(s_m^\star\right)\right) = \sum_{\lambda \in m} p_\lambda \left(\widehat{\beta}_\lambda - \beta_\lambda\right)^2$$

$$p_2(m) = P_n\left(\gamma\left(s_m^\star\right) - \gamma\left(\widehat{s}_m\right)\right) = \sum_{\lambda \in m} \widehat{p}_\lambda \left(\widehat{\beta}_\lambda - \beta_\lambda\right)^2$$

$$\delta(m) = (P_n - P)\gamma\left(s_m^\star\right)$$

Régressogrammes : espérances conditionnelles

$$\mathcal{P}_m := \left(\mathbb{1}_{X_i \in \lambda} \right)_{1 \leq i \leq n, \lambda \in m}$$

$$\mathbb{E}\left[p_1(m) \mid \mathcal{P}_m \right] = \frac{1}{n} \sum_{\lambda \in m} \frac{p_\lambda}{\widehat{p}_\lambda} \sigma_\lambda^2$$

$$\mathbb{E}\left[p_2(m) \mid \mathcal{P}_m \right] = \frac{1}{n} \sum_{\lambda \in m} \sigma_\lambda^2$$

$$\sigma_\lambda^2 := \mathbb{E}_{(X,Y) \sim P}\left[\left(Y - \beta_\lambda \right)^2 \middle| X \in \lambda \right] = \left(\sigma_\lambda^{(d)} \right)^2 + \left(\sigma_\lambda^{(a)} \right)^2$$

$$\left(\sigma_\lambda^{(a)} \right)^2 := \mathbb{E}_{(X,Y) \sim P}\left[\left(\sigma(X) \right)^2 \middle| X \in \lambda \right]$$

Régressogrammes : espérances

$$\mathbb{E}\left[p_1(m) \right] = \frac{1}{n} \sum_{\lambda \in m} \sigma_\lambda^2 \left(1 + \delta_{n,p_\lambda} \right)$$

$$\mathbb{E}\left[p_2(m) \right] = \frac{1}{n} \sum_{\lambda \in m} \sigma_\lambda^2$$

$$\delta_{n,p_\lambda} := \mathbb{E}\left[\frac{p_\lambda}{\widehat{p}_\lambda} \,\middle|\, \widehat{p}_\lambda > 0 \right] - 1$$

$$-\exp(-np) \leq \delta_{n,p} \leq \min\left\{ 1 + \frac{\kappa_1}{(np)^{1/4}} \; , \; \kappa_2 \right\}$$

Régressogrammes : risque, espérance de la pénalité idéale

$$\mathbb{E}\left[\ell\left(s^{\star}, \widehat{s}_m\right)\right] = \sum_{\lambda \in m} p_{\lambda} \left(\sigma_{\lambda}^{(d)}\right)^2 + \frac{1}{n} \sum_{\lambda \in m} \left(1 + \delta_{n, p_{\lambda}}\right) \sigma_{\lambda}^2$$

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{1}{n} \sum_{\lambda \in m} \left(2 + \delta_{n, p_{\lambda}}\right) \sigma_{\lambda}^2$$

Plan

Limitations de pen = pen(D_m)

$$Y = s^\star(X) + \varepsilon \quad \text{avec} \quad X \sim \mathcal{U}([0,1])$$

$$\mathbb{E}\left[\varepsilon^2 \mid X\right] = \sigma(X) \quad \text{et} \quad \int_0^{1/2} \left(\sigma(x)\right)^2 dx \neq \int_{1/2}^{1} \left(\sigma(x)\right)^2 dx$$

$$m \in \mathcal{M}_n^{(\mathrm{reg},1/2)} : \begin{array}{l} D_{m,1} \text{ morceaux sur } \left[0, \frac{1}{2}\right] \\ D_{m,2} \text{ morceaux sur } \left[\frac{1}{2}, 1\right] \end{array}$$

Limitations de pen $=$ pen(D_m)

$$Y = s^\star(X) + \varepsilon \quad \text{avec} \quad X \sim \mathcal{U}([0,1])$$

$$\mathbb{E}\left[\varepsilon^2 \mid X\right] = \sigma(X) \quad \text{et} \quad \int_0^{1/2} \left(\sigma(x)\right)^2 dx \neq \int_{1/2}^{1} \left(\sigma(x)\right)^2 dx$$

$$m \in \mathcal{M}_n^{(\mathrm{reg},1/2)} : \begin{array}{l} D_{m,1} \text{ morceaux sur } \left[0, \frac{1}{2}\right] \\ D_{m,2} \text{ morceaux sur } \left[\frac{1}{2}, 1\right] \end{array}$$

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] \approx \frac{4}{n}\left[D_{m,1} \int_0^{1/2} \left(\sigma(x)\right)^2 dx + D_{m,2} \int_{1/2}^{1} \left(\sigma(x)\right)^2 dx \right]$$

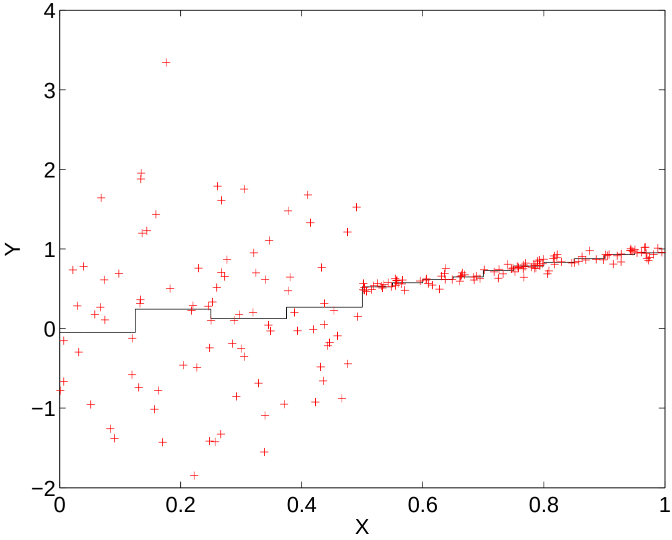
Limitations de pen = pen(D_m) : un exemple

$$Y = s^\star(X) + \varepsilon \quad \text{avec} \quad X \sim \mathcal{U}([0,1])$$
$$\mathcal{L}(\varepsilon \mid X) = \mathcal{N}\left(0, \sigma(X)^2\right)$$

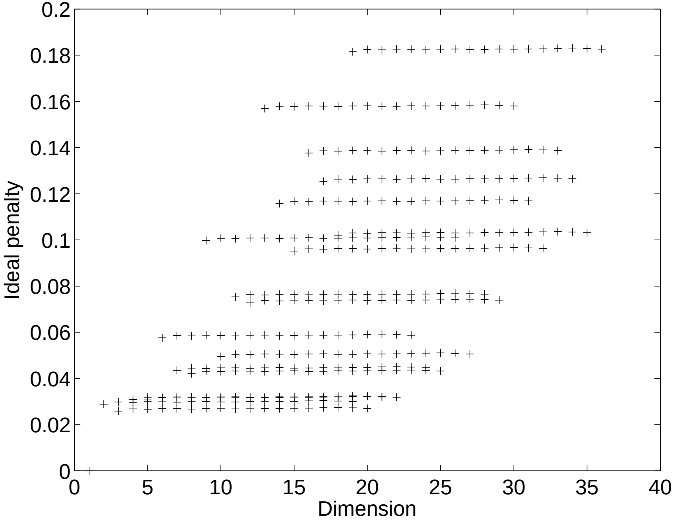
$$s^\star(X) = X \qquad \sigma(X) = \mathbb{1}_{X \leq \frac{1}{2}} + \frac{1}{20}\mathbb{1}_{X > 1/2}$$

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] \approx \frac{2}{n}\left[D_{m,1} + \frac{D_{m,2}}{400}\right]$$

Exemple : données et oracle ($n = 200$)



Exemple : $\mathrm{pen}_{\mathrm{id}}(m)$ en fonction de D_m



Pénalités fonction de la dimension

Lemme

Pour tout $D \in \mathcal{D}_n = \{ D_m \text{ t.q. } m \in \mathcal{M}_n \}$

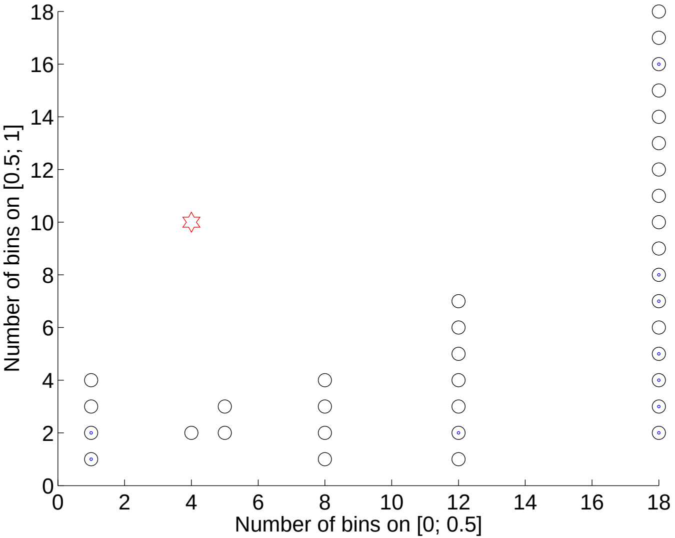
$$\mathcal{M}_{\mathrm{dim}}(D) := \mathrm{argmin}_{m \in \mathcal{M}_n \,\mathrm{t.q.}\, D_m = D} \{ P_n \gamma (\widehat{s}_m) \}$$

$$\mathcal{M}_{\mathrm{dim}} := \bigcup_{D \in \mathcal{D}_n} \mathcal{M}_{\mathrm{dim}}(D)$$

Alors, $\forall F : \mathcal{M}_n \mapsto \mathbb{R} \ \forall (X_i, Y_i)_{1 \leq i \leq n}$

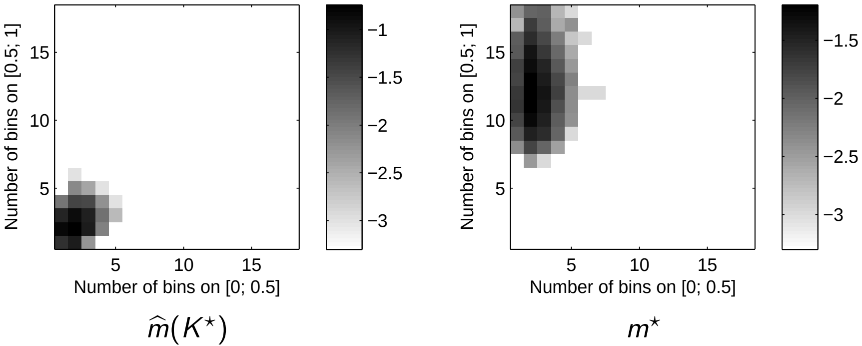
$$\mathrm{argmin}_{m \in \mathcal{M}_n} \{ P_n \gamma (\widehat{s}_m) + F(D_m) \} \subset \mathcal{M}_{\mathrm{dim}}$$

Modèles sélectionnables avec pen(D_m)



Limitations de pen = pen(D_m) : $\widehat{m}(D^\star) \neq m^\star$

Densités de $(D_{\widehat{m}(D^\star),1}, D_{\widehat{m}(D^\star),2})$ et $(D_{m^\star,1}, D_{m^\star,2})$ sur $N = 1000$ échantillons



Vers une preuve : concentration de $\mathrm{pen}_{\mathrm{id}}$

Hypothèses : $\|Y\|_\infty \leq A < \infty$ et $\sigma(\cdot) \geq \sigma_{\min} > 0$

- Concentration de p_1 et p_2 :
  si $\min_{\lambda \in m} \{ n p_\lambda \} \geq \Diamond \ln(n)$, avec probabilité au moins $1 - L n^{-\gamma}$, pour $i = 1, 2$

$$|p_i(m) - \mathbb{E}[p_i(m)]| \leq \frac{L_{A,\sigma_{\min},\gamma} (\ln(n))^2}{\sqrt{D_m}} \mathbb{E}[p_2(m)]$$

- Inégalité de Bernstein : avec probabilité $1 - 2e^{-x}$,

$$\forall \theta \in]0,1] \quad , \quad |(P_n - P)(\gamma(s_m^\star) - \gamma(s^\star))| \leq \theta \ell(s^\star, s_m^\star) + \frac{6A^2 x}{\theta n}$$

Heuristique de preuve : espérances

$$\mathbb{E}\left[p_1(m)\right] \approx \mathbb{E}\left[p_2(m)\right] \approx \frac{\beta_1 D_{m,1}}{n} + \frac{\beta_2 D_{m,1}}{n}$$

$$\beta_1 = 2 \int_0^{1/2} \sigma^2 \qquad \beta_2 = 2 \int_{1/2}^{1} \sigma^2$$

$$\ell\left(s^\star, s_m^\star\right) \approx \frac{\alpha_1}{D_{m,1}^2} + \frac{\alpha_2}{D_{m,2}^2}$$

$$\alpha_1 = \frac{1}{48} \int_0^{1/2} \left(s^{\star\prime}\right)^2 \qquad \alpha_2 = \frac{1}{48} \int_{1/2}^{1} \left(s^{\star\prime}\right)^2$$

Heuristique de preuve : espérances

$$\beta_1 = 2\int_0^{1/2} \sigma^2 \qquad \beta_2 = 2\int_{1/2}^1 \sigma^2$$

$$\alpha_1 = \frac{1}{48}\int_0^{1/2} \left(s^{\star\prime}\right)^2 \qquad \alpha_2 = \frac{1}{48}\int_{1/2}^1 \left(s^{\star\prime}\right)^2$$

$$P_n\gamma\left(\widehat{s}_m\right) - P\gamma\left(s^\star\right) \approx \frac{\alpha_1}{D_{m,1}^2} + \frac{\alpha_2}{D_{m,2}^2} - \frac{\beta_1 D_{m,1}}{n} - \frac{\beta_2 D_{m,1}}{n}$$

$$\ell\left(s^\star, \widehat{s}_m\right) \approx \frac{\alpha_1}{D_{m,1}^2} + \frac{\alpha_2}{D_{m,2}^2} + \frac{\beta_1 D_{m,1}}{n} + \frac{\beta_2 D_{m,1}}{n}$$

Heuristique de preuve : espérances

$$\beta_1 = 2 \int_0^{1/2} \sigma^2 \quad > \quad \beta_2 = 2 \int_{1/2}^{1} \sigma^2$$

$$\alpha_1 = \frac{1}{48} \int_0^{1/2} \left(s^{\star\prime} \right)^2 \qquad \alpha_2 = \frac{1}{48} \int_{1/2}^{1} \left(s^{\star\prime} \right)^2$$

$$P_n\gamma\left(\widehat{s}_m\right) - P\gamma\left(s^\star\right) \approx \frac{\alpha_1}{D_{m,1}^2} + \frac{\alpha_2}{D_{m,2}^2} - \frac{\beta_1 D_{m,1}}{n} - \frac{\beta_2 D_{m,1}}{n}$$

$$\ell\left(s^\star, \widehat{s}_m\right) \approx \frac{\alpha_1}{D_{m,1}^2} + \frac{\alpha_2}{D_{m,2}^2} + \frac{\beta_1 D_{m,1}}{n} + \frac{\beta_2 D_{m,1}}{n}$$

$$m^\star \approx \left(\left(\frac{2\alpha_1 n}{\beta_1} \right)^{1/3} , \left(\frac{2\alpha_2 n}{\beta_2} \right)^{1/3} \right)$$

Limitations de pen = pen(D_m) : théorie

$$Y = s^\star(X) + \varepsilon \quad \text{avec} \quad X \sim \mathcal{U}([0,1]) \ , \qquad \mathbb{E}\left[\varepsilon^2 \,\middle|\, X \right] = \sigma(X)$$

$$\text{et} \quad \sigma_a^2 = \int_0^{1/2} (\sigma(x))^2 \, dx \neq \int_{1/2}^{1} (\sigma(x))^2 \, dx = \sigma_b^2$$

Théorème (A. 2008)

Si $\mathcal{M} = \mathcal{M}_n^{(\mathrm{reg},1/2)}$, sous des hypothèses "raisonnables" sur $(s^\star, \varepsilon, \sigma)$, il existe $\eta(\sigma_a^2/\sigma_b^2) > 0$ telle qu'avec probabilité au moins $1 - C(\lVert \varepsilon \rVert_\infty, \sigma_a^2, \sigma_b^2, \lVert s^{\star\prime} \rVert_\infty, \lVert s^{\star\prime\prime} \rVert_\infty) n^{-2}$

$$\forall F \ , \ \forall \widehat{m}_F \in \arg\min_{m \in \mathcal{M}_n} \left\{ P_n \gamma\left(\widehat{s}_m\right) + F(D_m) \right\} \ ,$$

$$\ell\left(s^\star, \widehat{s}_{\widehat{m}_F}\right) \geq \left(1 + \eta\left(\frac{\sigma_a^2}{\sigma_b^2}\right)\right) \inf_{m \in \mathcal{M}_n} \left\{ \ell\left(s^\star, \widehat{s}_m\right) \right\}$$

Quelles motivations pour estimer la forme de la pénalité ?

- $\mathrm{pen}(D) = F(D) \Rightarrow$ perte d'un facteur $(1 + \eta) > 1$

Quelles motivations pour estimer la forme de la pénalité ?

- $\mathrm{pen}(D) = F(D) \Rightarrow$ perte d'un facteur $(1 + \eta) > 1$
- $\mathrm{pen}(m) = 2\mathbb{E}\left[\sigma(X)^2\right] D_m/n \Rightarrow$ explosion du risque possible

Quelles motivations pour estimer la forme de la pénalité ?

- $\mathrm{pen}(D) = F(D) \Rightarrow$ perte d'un facteur $(1 + \eta) > 1$
- $\mathrm{pen}(m) = 2\mathbb{E}\left[\sigma(X)^2\right] D_m / n \Rightarrow$ explosion du risque possible
- $\mathrm{pen}(m) = 2 \left\lVert \sigma \right\rVert_\infty^2 D_m / n \Rightarrow$ inégalité-oracle avec constante $\mathcal{O}(\max \sigma^2 / \min \sigma^2)$

Quelles motivations pour estimer la forme de la pénalité ?

- $\mathrm{pen}(D) = F(D) \Rightarrow$ perte d'un facteur $(1 + \eta) > 1$
- $\mathrm{pen}(m) = 2\mathbb{E}\left[\sigma(X)^2\right] D_m/n \Rightarrow$ explosion du risque possible
- $\mathrm{pen}(m) = 2 \left\|\sigma\right\|_\infty^2 D_m/n \Rightarrow$ inégalité-oracle avec constante $\mathcal{O}(\max \sigma^2 / \min \sigma^2)$

$\Rightarrow$ nécessaire d'estimer $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$ pour avoir un oracle avec constante $(1 + o(1))$ et éviter le risque de sur-apprentissage

Plan

Heuristique de rééchantillonnage (bootstrap, Efron 1979)

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

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

Heuristique de rééchantillonnage (bootstrap, Efron 1979)

$$
\begin{tikzcd}
\text{Monde réel :} \arrow[d, squiggly] & P \arrow[r, "\text{échantillonnage}"] & P_n \arrow[r, Rightarrow] & \widehat{s}_m \\
\text{Monde bootstrap :} & P_n & &
\end{tikzcd}
$$

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

Heuristique de rééchantillonnage (bootstrap, Efron 1979)

$$\begin{array}{lcl} \text{Monde réel :} & & P \xrightarrow{\text{échantillonnage}} P_n \Longrightarrow \widehat{s}_m \\ \Big\downarrow & & \\ \text{Monde bootstrap :} & & P_n \xrightarrow{\text{rééchantillonnage}} P_n^W \Longrightarrow \widehat{s}_m^W \end{array}$$

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

Rééchantillonnage à poids échangeables

$$P_n^W := \frac{1}{n} \sum_{i=1}^{n} W_i \delta_{\xi_i}$$

- Bootstrap :

$$W \sim \mathcal{M}\left(n; \frac{1}{n}, \ldots, \frac{1}{n} \right)$$

Rééchantillonnage à poids échangeables

$$P_n^W := \frac{1}{n} \sum_{i=1}^{n} W_i \delta_{\xi_i}$$

- Efron(m) ou Bootstrap "m out of n" :

$$\frac{m}{n} W \sim \mathcal{M}\left(m; \frac{1}{n}, \ldots, \frac{1}{n} \right)$$

Rééchantillonnage à poids échangeables

$$P_n^W := \frac{1}{n} \sum_{i=1}^n W_i \delta_{\xi_i} \quad \text{ou} \quad \frac{1}{\sum_k W_k} \sum_{i=1}^n W_i \delta_{\xi_i} = \frac{1}{n} \sum_{i=1}^n \frac{W_i}{\overline{W}} \delta_{\xi_i}$$

- Efron(m) ou Bootstrap "m out of n" :

$$\frac{m}{n} W \sim \mathcal{M}\left(m; \frac{1}{n}, \ldots, \frac{1}{n} \right)$$

- Sous-échantillonnage :
  - Random-hold out(q) ou validation aléatoire, $q \in \{ 1, \ldots, n-1 \}$:

$$W_i = \frac{n}{q} \mathbb{1}_{i \in I} \quad \text{avec} \quad I \sim \mathcal{U}\left(\mathfrak{P}_q(\{ 1, \ldots, n \}) \right)$$

  - Rademacher(p) ou Bernoulli :

$$pW_1, \ldots, pW_n \text{ i.i.d. } \sim \mathcal{B}(p)$$

Justification théorique asymptotique

Théorème (van der Vaart & Wellner, 1996)

Soit $(W_{n,1}, \ldots, W_{n,n}) \in \mathbb{R}^n$ un vecteur aléatoire positif, échangeable, indépendant de $\xi_{1\ldots n}$, borné et tel que

$$n^{-1} \sum_{i=1}^{n} \left(W_{n,i} - \overline{W}_n \right)^2 \xrightarrow{(p)} c^2 > 0 \enspace .$$

Alors, lorsque n tend vers l'infini,

$$\sup_{h \in BL_1} \left| \mathbb{E}_W \left[h \left(\sqrt{n} \left(P_n^W - \overline{W}_n P_n \right) \right) \right] - \mathbb{E} \left[h \left(c \mathbb{G} \right) \right] \right| \xrightarrow{(p)} 0$$

où $\mathbb{G}$ est le processus Gaussien limite de $\sqrt{n}(P_n - P)$, de moyenne nulle et de fonction de covariance $\operatorname{cov}(f, g) = P(fg) - P(f)P(g)$.

Utilisations classiques du rééchantillonnage

- estimation de la variance, du risque quadratique
- estimation et/ou correction du biais
- intervalles de confiance, p-valeurs
- estimation de l'erreur de prédiction, sélection de modèles
- stabilisation (bagging, forêts aléatoires)
- ...

Estimateur de la variance par rééchantillonnage

Cadre :

$$\xi_1, \ldots, \xi_n \text{ i.i.d. } \sim P \qquad \mathbb{E}[\xi_i] = \mu \quad \mathbb{E}\left[(\xi_i - \mu)^2\right] = \sigma^2$$

$$\sigma^2 = n\mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\xi_i - \mu\right)^2\right] = n\mathbb{E}\left[\left(\mathbb{E}_{\xi\sim P_n}\xi - \mathbb{E}_{\xi\sim P}\xi\right)^2\right]$$
$$= n\mathbb{E}\left[F(P, P_n)\right]$$

$\Rightarrow$ estimateur par rééchantillonnage

$$\widehat{\sigma^2_W} = n\mathbb{E}_W\left[F(P_n, P_n^W)\right]$$

Estimateur de la variance par rééchantillonnage

$$\widehat{\sigma^2_W} = n\mathbb{E}_W \left[F(P_n, P_n^W) \right]$$

$$\widehat{\sigma^2_W} = \frac{R_V^{(W)}}{n} \left[\sum_{i=1}^{n} (\xi_i - \mu)^2 - \frac{1}{n-1} \sum_{i \neq j} (\xi_i - \mu)(\xi_j - \mu) \right]$$

$$R_V^{(W)} := \mathbb{E}_W \left[\left(\frac{nW_i}{\sum_{k=1}^{n} W_k} - 1 \right)^2 \right]$$

Comparaison avec l'estimateur classique

Estimateur classique de la variance :

$$\widehat{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\xi_i - \frac{1}{n} \sum_{k=1}^{n} \xi_k \right)^2$$

Comparaison avec l'estimateur classique

Estimateur classique de la variance :

$$\widehat{\sigma^2} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\xi_i - \frac{1}{n} \sum_{k=1}^{n} \xi_k \right)^2$$

$$\widehat{\sigma^2_W} = R_V^{(W)} \widehat{\sigma^2}$$

$$\Rightarrow \qquad \mathbb{E}\left[\widehat{\sigma^2_W} \right] = R_V^{(W)} \sigma^2$$

Rééchantillonnage et structure

- Propriétés de $F(P, P_n) = \left(\mathbb{E}_{\xi \sim P_n} \xi - \mathbb{E}_{\xi \sim P} \xi \right)^2$:
  - échangeabilité
  - invariance par translation
  - homogénéité
  - polynôme en les ξ_i et en $\mathbb{E}_{\xi \sim P} \xi$

$\Rightarrow$ $\mathbb{E}_W [F(P_n, P_n^W)]$ possède les mêmes propriétés

Rééchantillonnage et structure

- Propriétés de $F(P, P_n) = (\mathbb{E}_{\xi \sim P_n} \xi - \mathbb{E}_{\xi \sim P} \xi)^2$:
  - échangeabilité
  - invariance par translation
  - homogénéité
  - polynôme en les ξ_i et en $\mathbb{E}_{\xi \sim P} \xi$

$\Rightarrow$ $\mathbb{E}_W[F(P_n, P_n^W)]$ possède les mêmes propriétés

$$\Rightarrow \qquad \mathbb{E}_W \left[F(P_n, P_n^W) \right] \propto \widehat{\sigma^2}$$

Rééchantillonnage et concentration

Sur-concentration de l'estimateur par rééchantillonnage :

$$\operatorname{var}\left(n \left(\frac{1}{n} \sum_{i=1}^{n} \xi_i - \mu \right)^2 \right) = 2\sigma^4 + \frac{\mathbb{E}\left[(\xi_1 - \mu)^4 \right] - 3\sigma^4}{n}$$

$$\operatorname{var}\left(\frac{1}{R_V^{(W)}} \widehat{\sigma_W^2} \right) = \frac{1}{n} \left(\mathbb{E}\left[(\xi_1 - \mu)^4 \right] - \sigma^4 \right) + \frac{2}{n(n-1)} \sigma^4$$

Calcul de la constante multiplicative

$$R_V^{(W)} := \mathbb{E}_W\left[\left(\frac{nW_i}{\sum_{k=1}^n W_k} - 1\right)^2\right]$$

Calcul de la constante multiplicative

$$R_V^{(W)} := \mathbb{E}_W \left[\left(\frac{nW_i}{\sum_{k=1}^n W_k} - 1 \right)^2 \right]$$

Efron(m) : $R_V^{(W)} = \frac{n-1}{m}$

Rademacher(p) : $R_V^{(W)} = \frac{1 + \delta_{n,p}}{p} - 1 \approx \frac{1}{p} - 1$

Random hold-out(q) : $R_V^{(W)} = \frac{n}{q} - 1$

Leave-one-out $=$ Rho$(n-1)$: $R_V^{(W)} = \frac{1}{n-1}$

Estimation par rééchantillonnage de $\mathrm{pen}_{\mathrm{id}}(m)$

- Pénalité idéale :

$$(P - P_n)(\gamma(\widehat{s}_m)) = F(P, P_n)$$

- Estimateur de $\mathbb{E}\left[F(P, P_n)\right]$ par rééchantillonnage :

$$\mathrm{pen}(m) = C_W \mathbb{E}\left[(P_n - P_n^W)(\gamma(\widehat{s}_m^W)) \,\middle|\, (X_i, Y_i)_{1 \leq i \leq n} \right]$$

- bootstrap (Efron, 1983 ; Shibata, 1997), bootstrap "m out of n" pour l'identification (Shao, 1996), poids échangeables généraux (A. 2009)

- Constante C_W : pourquoi ? comment l'estimer ?

Pénalités de Rademacher

- Pénalités globales :

$$\mathrm{pen}_{\mathrm{id}}(m) \leq \mathrm{pen}_{\mathrm{id}}^{\mathrm{glo}}(m) = \sup_{t \in S_m} (P - P_n)\gamma(t)$$

Pénalités de Rademacher

- Pénalités globales :

$$\mathrm{pen}_{\mathrm{id}}(m) \leq \mathrm{pen}_{\mathrm{id}}^{\mathrm{glo}}(m) = \sup_{t \in S_m} (P - P_n)\gamma(t)$$

- Pénalités de Rademacher globales en classification (Koltchinskii, Panchenko 2001 ; Bartlett, Boucheron, Lugosi 2002), poids échangeables (Fromont 2004)

$$\mathbb{E}\left[\sup_{t \in S_m} \left\{ \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \gamma\left(t; \xi_i\right) \right\} \,\middle|\, P_n \right]$$
$$\text{avec} \quad \varepsilon_1, \ldots, \varepsilon_n \text{ i.i.d. } \sim \mathcal{U}(\{-1, +1\})$$

- complexités de Rademacher locales (Bartlett, Bousquet and Mendelson, 2004 ; Koltchinskii, 2006)

Plan

Rappels

$$\widehat{s}_m = \sum_{\lambda \in m} \widehat{\beta}_\lambda \mathbb{1}_\lambda \quad \text{avec} \quad \widehat{\beta}_\lambda := \frac{1}{n \widehat{p}_\lambda} \sum_{i \text{ t.q. } X_i \in \lambda} Y_i$$

$$\widehat{p}_\lambda = \widehat{p}_\lambda(D_n) = \frac{1}{n} \operatorname{Card} \{ i \text{ t.q. } X_i \in \lambda \}$$

$$\mathrm{pen}_{\mathrm{id}}(m) = p_1(m) + p_2(m) - \delta(m)$$

$$p_1(m) = P\left(\gamma\left(\widehat{s}_m\right) - \gamma\left(s_m^\star\right)\right) = \sum_{\lambda \in m} \left[p_\lambda \left(\widehat{\beta}_\lambda - \beta_\lambda \right)^2 \right]$$

$$p_2(m) = P_n\left(\gamma\left(s_m^\star\right) - \gamma\left(\widehat{s}_m\right)\right) = \sum_{\lambda \in m} \left[\widehat{p}_\lambda \left(\widehat{\beta}_\lambda - \beta_\lambda \right)^2 \right]$$

$$\delta(m) = (P_n - P)\gamma\left(s_m^\star\right)$$

Pénalité par rééchantillonnage

$$\mathrm{pen}_W(m) = \frac{C_W}{n} \sum_{\lambda \in m} \frac{R_{1,W} + R_{2,W}}{n\widehat{p}_\lambda - 1} \left(S_{\lambda,2} - \frac{1}{n\widehat{p}_\lambda} S_{\lambda,1}^2 \right) \mathbb{1}_{n\widehat{p}_\lambda \geq 2}$$

$$\text{avec} \quad S_{\lambda,1} := \sum_{X_i \in \lambda} (Y_i - \beta_\lambda) \qquad S_{\lambda,2} := \sum_{X_i \in \lambda} (Y_i - \beta_\lambda)^2$$

$$R_{1,W}(n, \widehat{p}_\lambda) := \mathbb{E}\left[\frac{(W_1 - \widehat{W}_\lambda)^2}{\widehat{W}_\lambda^2} \,\middle|\, X_1 \in \lambda \ , \ \widehat{W}_\lambda > 0 \right]$$

$$\text{et} \quad R_{2,W}(n, \widehat{p}_\lambda) := \mathbb{E}\left[\frac{(W_1 - \widehat{W}_\lambda)^2}{\widehat{W}_\lambda} \,\middle|\, X_1 \in \lambda \right]$$

Valeurs pour R_1 et R_2 : exemples

$$R_{1,W}(n,\widehat{p}_\lambda) \sim R_{2,W}(n,\widehat{p}_\lambda) \text{ lorsque } n\widehat{p}_\lambda \to \infty$$

$$C_{W,\infty}(n) := \lim_{n\widehat{p}_\lambda \to \infty} \frac{1}{R_{2,W}(n,\widehat{p}_\lambda)}$$

Efron(m) : $R_{2,W}(n,\widehat{p}_\lambda) = \frac{n}{m}\left(1 - \frac{1}{n\widehat{p}_\lambda}\right)$ $\quad C_{W,\infty} = \frac{m}{n}$

Rademacher(p) : $R_{2,W}(n,\widehat{p}_\lambda) = \frac{1}{p} - 1$ $\quad C_{W,\infty} = \frac{p}{1-p}$

Random hold-out(q) : $R_{2,W}(n,\widehat{p}_\lambda) = \frac{n}{q} - 1$ $\quad C_{W,\infty} = \frac{q}{n-q}$

Leave-one-out : $R_{2,W}(n,\widehat{p}_\lambda) = \frac{1}{n-1}$ $\quad C_{W,\infty} = n-1$

Espérances

$$\mathbb{E}\left[\, Y_i - \beta_\lambda \,\middle|\, X_i \in \lambda \,\right] = 0 \quad \text{et} \quad \mathbb{E}\left[\left(Y_i - \beta_\lambda \right)^2 \,\middle|\, X_i \in \lambda \right] = \sigma_\lambda^2$$

$$\mathbb{E}\left[\mathrm{pen}_W(m) \,\middle|\, \mathcal{P}_m\right] = \frac{C_W}{n} \sum_{\lambda \in m} \left(R_{1,W} + R_{2,W} \right) \sigma_\lambda^2 \mathbb{1}_{n\widehat{p}_\lambda \geq 2}$$

$$\mathbb{E}\left[\mathrm{pen}_W(m)\right] = \frac{C_W}{C_{W,\infty}} \frac{1}{n} \sum_{\lambda \in m} \left(2 + \overline{\delta}_{n,p_\lambda}^{(\mathrm{penW})} \right) \sigma_\lambda^2$$

avec $\overline{\delta}_{n,p_\lambda}^{(\mathrm{penW})} \to 0$ quand $np_\lambda \to +\infty$

$\Rightarrow$ adaptation à l'hétéroscédasticité

Concentration

Proposition (A. 2009)

- *Données bornées* : $\|Y_i\|_\infty \leq A < \infty$
- *Bruit minoré* : $\sigma(X_i) \geq \sigma_{\min} > 0$
- $\mathcal{L}(W)$ *parmi* $Efr(n)$, $Rad(1/2)$, $Rho(n/2)$, Loo

Pour tout $A_n \geq 2$, *avec probabilité au moins* $1 - L_1 n^{-\gamma}$,

$$\left| \mathrm{pen}_W(m) - \mathbb{E}\left[\mathrm{pen}_W(m) \mid \mathcal{P}_m \right] \right| \mathbb{1}_{\min_{\lambda \in m} \{ n \widehat{p}_\lambda \} \geq A_n} \leq \frac{C_W}{C_{W,\infty}} \frac{L_2(A/\sigma_{\min}, \gamma) \ln(n)}{\sqrt{A_n D_m}} \mathbb{E}\left[p_2(m) \right]$$

Inégalité-oracle “trajectorielle” non-asymptotique

- $\mathcal{L}(W)$ parmi Efr(n), Rad($1/2$), Rho($n/2$), Loo
- $C_W \approx C_{W,\infty}$

Inégalité-oracle “trajectorielle” non-asymptotique

- $\mathcal{L}(W)$ parmi Efr(n), Rad($1/2$), Rho($n/2$), Loo
- $C_W \approx C_{W,\infty}$
- $\mathrm{Card}(\mathcal{M}_n) \leq C_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$
- Données bornées : $\|Y_i\|_\infty \leq A < \infty$
- Bruit minoré : $\sigma(X_i) \geq \sigma_{\min} > 0$

Inégalité-oracle "trajectorielle" non-asymptotique

- $\mathcal{L}(W)$ parmi Efr(n), Rad($1/2$), Rho($n/2$), Loo
- $C_W \approx C_{W,\infty}$

- $\mathrm{Card}(\mathcal{M}_n) \leq C_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$
- Données bornées : $\|Y_i\|_\infty \leq A < \infty$
- Bruit minoré : $\sigma(X_i) \geq \sigma_{\min} > 0$

- $s^\star \in \mathcal{H}(\alpha, R)$ non-constante
- Présélection de modèles : $\forall m \in \mathcal{M} \quad , \quad \min_{\lambda \in m} n\widehat{p}_\lambda \geq 3$

Inégalité-oracle "trajectorielle" non-asymptotique

- $\mathcal{L}(W)$ parmi $\mathrm{Efr}(n)$, $\mathrm{Rad}(1/2)$, $\mathrm{Rho}(n/2)$, Loo
- $C_W \approx C_{W,\infty}$

- $\mathrm{Card}(\mathcal{M}_n) \leq C_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$
- Données bornées : $\lVert Y_i \rVert_\infty \leq A < \infty$
- Bruit minoré : $\sigma(X_i) \geq \sigma_{\min} > 0$

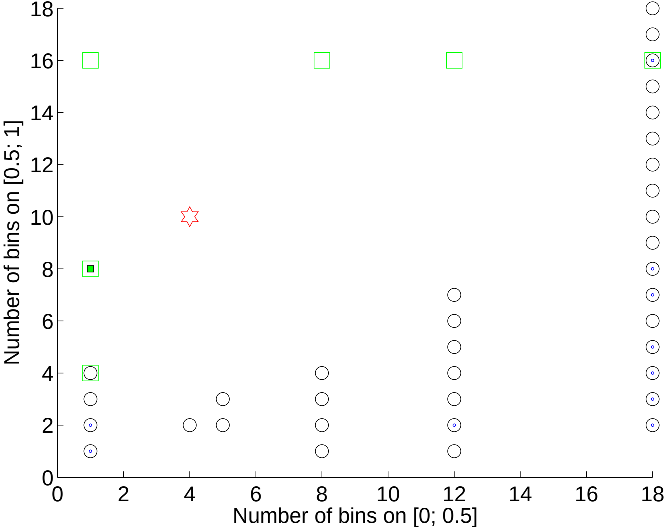
- $s^\star \in \mathcal{H}(\alpha, R)$ non-constante
- Présélection de modèles : $\forall m \in \mathcal{M} \quad , \quad \min_{\lambda \in m} n\widehat{p}_\lambda \geq 3$

Théorème (A. 2009)

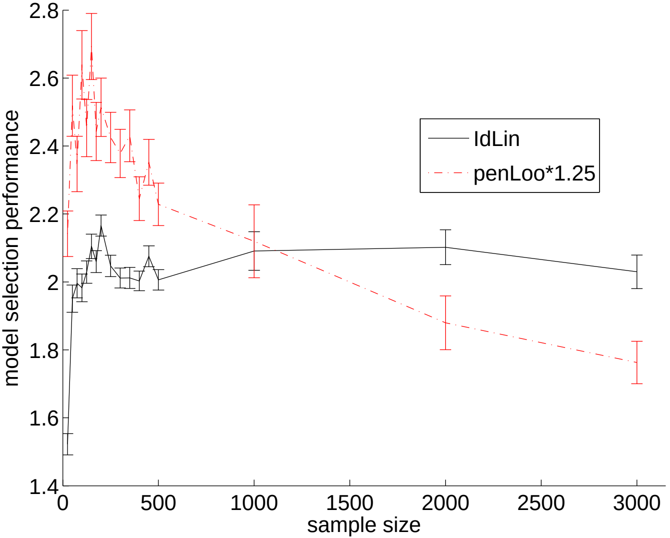
Avec probabilité au moins $1 - \Diamond n^{-2}$,

$$\ell\left(s^\star, \widehat{s}_{\widehat{m}}\right) \leq \left(1 + (\ln(n))^{-1/5}\right) \inf_{m \in \mathcal{M}} \left\{ \ell\left(s^\star, \widehat{s}_m\right) \right\}$$

Modèles sélectionnables : penLoo mieux que pen(D_m)



Simulations : $1.25 \times \mathrm{pen}_{\mathrm{Loo}}(m)$ vs. $K^{\star} D_m$



Adaptation

$$\widetilde{s} := \widehat{s}_{\widehat{m}} \quad \text{avec} \quad \widehat{m} \in \operatorname*{argmin}_{\substack{m \in \mathcal{M}_n^{(\mathrm{reg})} \\ \min_{\lambda \in m}\{ n\widehat{p}_\lambda \} \geq 3}} \left\{ P_n \gamma \left(\widehat{s}_m \right) + \mathrm{pen}_W(m) \right\}$$

Hypothèses :

- Données bornées : $\| Y_i \|_\infty \leq A < \infty$
- Bruit minoré : $\sigma(X_i) \geq \sigma_{\min} > 0$
- Densité de X minorée : $\forall I \subset \mathcal{X}, \ \mathbb{P}\left(X \in I \right) \geq c_X^{\min} \operatorname{Leb}(I)$
- $s^\star = \eta \in \mathcal{H}(\alpha, R)$ avec $\alpha \in \,]0,1]$:

$$\forall x_1, x_2 \in \mathcal{X} \ , \ \ |s^\star(x_1) - s^\star(x_2)| \leq R \, \| x_1 - x_2 \|_\infty^\alpha$$

Adaptation

$$\widetilde{s} := \widehat{s}_{\widehat{m}} \quad \text{avec} \quad \widehat{m} \in \operatorname*{argmin}_{\substack{m \in \mathcal{M}_n^{(\mathrm{reg})} \\ \min_{\lambda \in m} \{ n \widehat{p}_\lambda \} \geq 3}} \left\{ P_n \gamma \left(\widehat{s}_m \right) + \mathrm{pen}_W(m) \right\}$$

$$\mathbb{E}\left[\ell \left(s^\star, \widetilde{s} \right) \right] \leq K_2 R^{\frac{2d}{2\alpha+d}} n^{\frac{-2\alpha}{2\alpha+d}} \left\| \sigma \right\|_\infty^{\frac{4\alpha}{2\alpha+d}} + \frac{K_3 A^2}{n^2}$$

Adaptation

$$\widetilde{s} := \widehat{s}_{\widehat{m}} \quad \text{avec} \quad \widehat{m} \in \operatorname*{argmin}_{\substack{m \in \mathcal{M}_n^{(\mathrm{reg})} \\ \min_{\lambda \in m}\{ n\widehat{p}_\lambda \} \geq 3}} \left\{ P_n \gamma\left(\widehat{s}_m \right) + \mathrm{pen}_W(m) \right\}$$

$$\mathbb{E}\left[\ell\left(s^\star, \widetilde{s} \right) \right] \leq K_2 R^{\frac{2d}{2\alpha+d}} n^{\frac{-2\alpha}{2\alpha+d}} \left\| \sigma \right\|_\infty^{\frac{4\alpha}{2\alpha+d}} + \frac{K_3 A^2}{n^2}$$

- et si $\sigma(\cdot)$ est K_σ-Lipschitz plus un maximum de J_σ sauts :

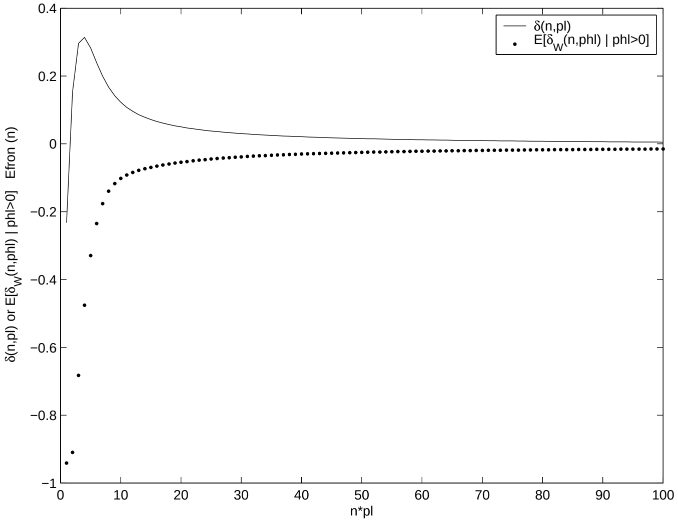
$$\mathbb{E}\left[\ell\left(s^\star, \widetilde{s} \right) \right] \leq K_2 R^{\frac{2d}{2\alpha+d}} n^{\frac{-2\alpha}{2\alpha+d}} \left\| \sigma \right\|_{L^2(\mathrm{Leb})}^{\frac{4\alpha}{2\alpha+d}} + \frac{K_4 A^2}{n^2}$$

Comparaison théorique des poids : rappels

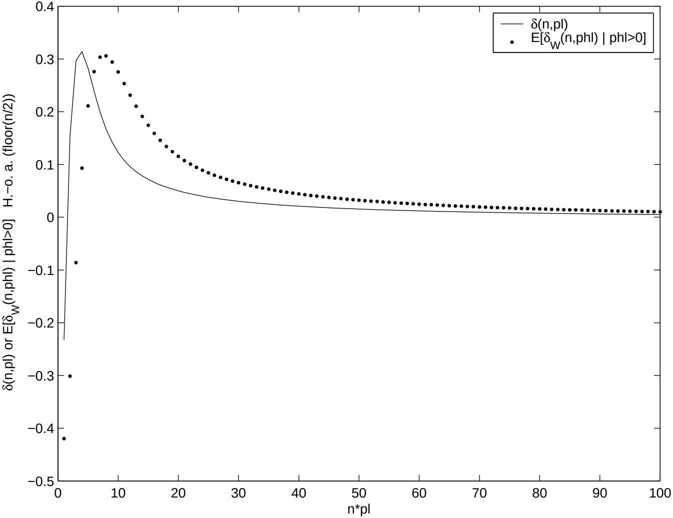
$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{1}{n} \sum_{\lambda \in m} \left(2 + \delta_{n,p_\lambda} \right) \sigma_\lambda^2$$

et $$\mathbb{E}\left[\mathrm{pen}_W(m)\right] = \frac{1}{n} \sum_{\lambda \in m} \left(2 + \overline{\delta}_{n,\widehat{p}_\lambda}^{(\mathrm{penW})} \right) \sigma_\lambda^2$$

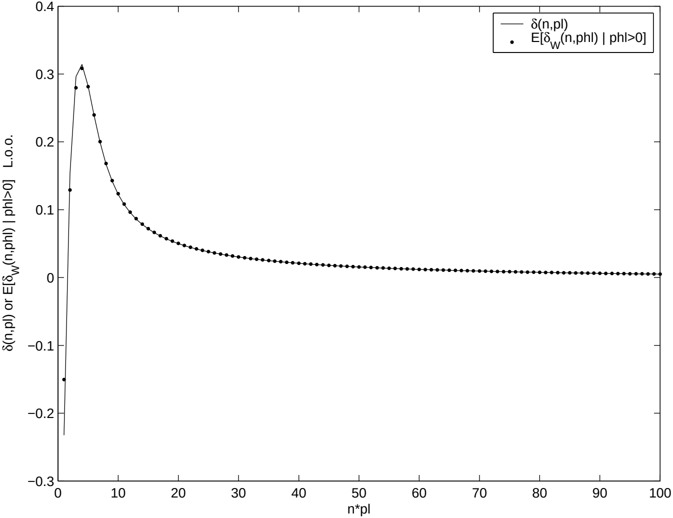
$\overline{\delta}^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}$ vs. δ_{n,p_λ} : Efron(n) ≈ Poisson(1)



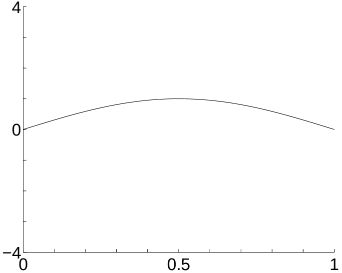
$\overline{\delta}^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}$ vs. δ_{n,p_λ} : $\mathrm{Rho}(n/2) \approx \mathrm{Rad}(1/2)$

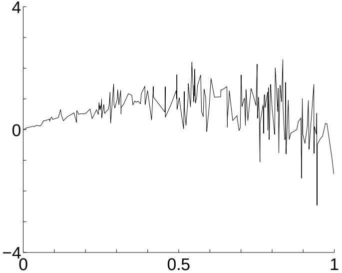


$\overline{\delta}^{(\mathrm{penW})}_{n,\widehat{p}_\lambda}$ vs. δ_{n,p_λ} : Leave-one-out



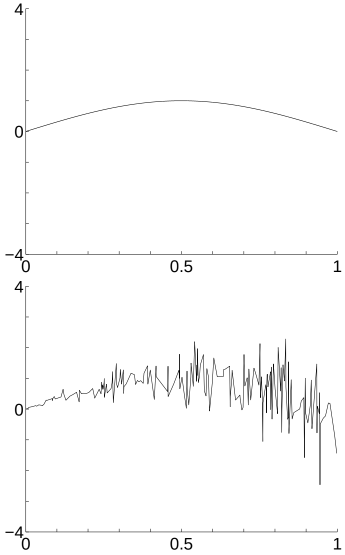
$s^\star(x) = \sin(2\pi x)$ $n = 200$ $\sigma(x) = x$ $\mathcal{M} = \mathcal{M}^{(\mathrm{reg},1/2)}$





Mallows	3.69 ± 0.07
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	2.30 ± 0.05
pen Efr	3.15 ± 0.07
pen Loo	2.59 ± 0.06
pen Rho	2.50 ± 0.06
pen Rad	2.49 ± 0.06

$s^\star(x) = \sin(2\pi x)$ $\quad n = 200$ $\quad \sigma(x) = x$ $\quad \mathcal{M} = \mathcal{M}^{(\mathrm{reg},1/2)}$



Mallows	3.69 ± 0.07
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}]$	2.30 ± 0.05
pen Efr	3.15 ± 0.07
pen Loo	2.59 ± 0.06
pen Rho	2.50 ± 0.06
pen Rad	2.49 ± 0.06
Mallows $\times 1.25$	3.17 ± 0.07
$\mathbb{E}[\mathrm{pen}_{\mathrm{id}}] \times 1.25$	2.03 ± 0.04
pen Efr $\times 1.25$	2.60 ± 0.06
pen Loo $\times 1.25$	2.22 ± 0.05
pen Rho $\times 1.25$	2.14 ± 0.05
pen Rad $\times 1.25$	2.14 ± 0.05

Plan

Estimation de densité par moindres carrés

- μ mesure de référence sur Ξ
- $f = dP/d\mu \in \mathbb{S} = L^2(\mu)$

Estimation de densité par moindres carrés

- μ mesure de référence sur Ξ
- $f = dP/d\mu \in \mathbb{S} = L^2(\mu)$
- $\gamma(t;\xi) = \|t\|^2_{L^2(\mu)} - 2t(\xi)$
  $\Rightarrow P\gamma(t) = \|t\|^2_{L^2(\mu)} - 2\left\langle t,\, f\right\rangle_{L^2(\mu)}$
  $\Rightarrow s^\star = f$ et $\ell\left(s^\star, t\right) = \|t - s^\star\|^2_{L^2(\mu)}$

Estimation de densité par moindres carrés

- μ mesure de référence sur Ξ
- $f = dP/d\mu \in \mathbb{S} = L^2(\mu)$

- $\gamma(t;\xi) = \|t\|^2_{L^2(\mu)} - 2t(\xi)$

  $\Rightarrow P\gamma(t) = \|t\|^2_{L^2(\mu)} - 2\langle t, f\rangle_{L^2(\mu)}$

  $\Rightarrow s^\star = f$ et $\ell(s^\star, t) = \|t - s^\star\|^2_{L^2(\mu)}$

- $(\psi_\lambda)_{\lambda\in m}$ b.o.n. de S_m

  $\Rightarrow s^\star_m = \sum_{\lambda\in m}(P\psi_\lambda)\psi_\lambda$ et $\widehat{s}_m = \sum_{\lambda\in m}(P_n\psi_\lambda)\psi_\lambda$

Estimation de densité par moindres carrés

- μ mesure de référence sur Ξ
- $f = dP/d\mu \in \mathbb{S} = L^2(\mu)$

- $\gamma(t;\xi) = \|t\|^2_{L^2(\mu)} - 2t(\xi)$

  $\Rightarrow P\gamma(t) = \|t\|^2_{L^2(\mu)} - 2\langle t, f\rangle_{L^2(\mu)}$

  $\Rightarrow s^\star = f$ et $\ell(s^\star, t) = \|t - s^\star\|^2_{L^2(\mu)}$

- $(\psi_\lambda)_{\lambda\in m}$ b.o.n. de S_m

  $\Rightarrow s^\star_m = \sum_{\lambda\in m}(P\psi_\lambda)\psi_\lambda$ et $\widehat{s}_m = \sum_{\lambda\in m}(P_n\psi_\lambda)\psi_\lambda$

$$\begin{aligned}\mathrm{pen}_{\mathrm{id}}(m) &= (P - P_n)\gamma(\widehat{s}_m) = 2(P_n - P)(\widehat{s}_m)\\ &= 2\|s^\star_m - \widehat{s}_m\|^2_{L^2(\mu)} + 2(P_n - P)(s^\star_m)\end{aligned}$$

Cas i.i.d. (Lerasle 2009)

$$\mathrm{pen}_{\mathrm{id}}(m) = 2(P_n - P)(\widehat{s}_m)$$

$$\mathrm{pen}_W(m) = C_W \mathbb{E}_W \left[2(P_n^W - \overline{W} P_n)(\widehat{s}_m^W) \right]$$

Cas i.i.d. (Lerasle 2009)

$$\mathrm{pen}_W(m) = C_W \mathbb{E}_W \left[2(P_n^W - \overline{W} P_n)(\widehat{s}_m^W) \right]$$

$\Rightarrow$ $\mathrm{pen}_W(m)$ ne dépend de W que par une constante multiplicative

Cas i.i.d. (Lerasle 2009)

$$\mathrm{pen}_W(m) = C_W \mathbb{E}_W \left[2(P_n^W - \overline{W} P_n)(\widehat{s}_m^W) \right]$$

$\Rightarrow$ $\mathrm{pen}_W(m)$ ne dépend de W que par une constante multiplicative

$\Rightarrow$ $\mathbb{E}\left[\mathrm{pen}_W(m)\right] = C_W \operatorname{var}\left(W_1 - \overline{W}\right) \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$

$+$ concentration de $\mathrm{pen}_W(m)$ autour de son espérance (plus rapide que $\mathrm{pen}_{\mathrm{id}}(m)$)

Cas i.i.d. (Lerasle 2009)

$$\mathrm{pen}_W(m) = C_W \mathbb{E}_W \left[2(P_n^W - \overline{W} P_n)(\widehat{s}_m^W) \right]$$

$\Rightarrow$ $\mathrm{pen}_W(m)$ ne dépend de W que par une constante multiplicative

$\Rightarrow$ $\mathbb{E}\left[\mathrm{pen}_W(m)\right] = C_W \operatorname{var}\left(W_1 - \overline{W}\right) \mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right]$
$+$ concentration de $\mathrm{pen}_W(m)$ autour de son espérance (plus rapide que $\mathrm{pen}_{\mathrm{id}}(m)$)

$\Rightarrow$ inégalité-oracle avec constante $1 + o(1)$ sous de bonnes hypothèses sur P et $\mathcal{M}_n$

Cas dépendant (Lerasle 2010)

- Hypothèses de mélange (β ou τ)

- Découpage des données en blocs $\Rightarrow$ on garde un bloc sur deux
- On rééchantillonne les blocs (qui sont quasi-indépendants entre eux)

$\Rightarrow$ Inégalité-oracle (avec un oracle n'utilisant pas toutes les données de départ)

Plan

Limites

- Temps de calcul

$\Rightarrow$ alternative : poids non-échangeables (e.g., V-fold)

- Résultats non-asymptotiques : possible sans calculs explicites ?

La semaine prochaine

- Lundi 31, 15h30–17h30 : Validation croisée et pénalités reliées

Lundi 31, 13h30–14h40 :
Deux approches nouvelles en sparsité (I) : Les liens de la régularisation L1 avec la sélection de modèles parmi des boules L1
par Caroline Meynet (Université Paris-Sud).
Salle W, ENS, 45 rue d'Ulm
`http://www.math.ens.fr/-Seminaires-?id_seance=178`