Sélection de modèles et sélection d'estimateurs pour l'Apprentissage statistique (Cours Peccot)

Deuxième cours: Calibration de pénalités et pénalités minimales

SYLVAIN ARLOT
CNRS – ÉQUIPE SIERRA

Table des matières

Date: 17 Janvier 2011.

1. Problème de calibration de pénalités

Presque toutes les pénalités proposées pour la sélection d'estimateurs ou la sélection de modèles dépendent d'au moins un paramètre (par exemple, un facteur multiplicatif) qu'il faut calibrer précisément en pratique si l'on veut une erreur de prédiction (quasi) minimale.

1.1. **Exemples.** Le problème de la calibration se pose notamment pour la pénalité C_p, introduite au premier cours, qui dépend de la variance σ^2 du bruit, a priori inconnue. Plus généralement, on peut distinguer trois catégories de problèmes (non-exclusives) selon les résultats théoriques disponibles :

(1) La pénalité optimale[1] est connue *à constante multiplicative près*, la valeur optimale de cette constante n'étant pas précisée par la théorie existante. C'est par exemple le cas en détection de ruptures [14, 27], en estimation de densité avec des modèles de mélange Gaussiens [37] ou pour les pénalités de Rademacher locales en classification [10, 25]. Dans les meilleurs cas, on dispose d'un encadrement de la valeur «optimale» de ce paramètre. Ainsi, les pénalités de Rademacher globales en classification [24] diffèrent d'un facteur deux entre théorie et pratique [32] ; voir aussi plus récemment le cas de l'estimation de l'intensité de processus de Poisson [39], où l'on a un paramètre de seuillage optimal connu à un facteur 12 près.

(2) La pénalité optimale est connue théoriquement, mais fait intervenir des *quantités inconnues en pratique*, par exemple σ^2 pour C_p ou C_L [33].

(3) La pénalité optimale est entièrement connue et calculable en pratique, mais seulement *asymptotiquement*, c'est-à-dire que sa forme n'est a priori exacte que lorsque la taille d'échantillon tend vers l'infini. C'est notamment le cas de AIC [2] ou BIC [45]. En particulier, le facteur multiplicatif optimal a de grandes chances de varier avec n, d'une manière totalement inconnue.

Dans les trois cas, en vue d'une utilisation pratique, on a besoin d'une méthode pour calibrer une pénalité à l'aide des données uniquement. Ce cours est consacré à ce problème, et plus particulièrement à une approche assez générale pour le résoudre dans le cas (classique) où la constante à calibrer est un facteur multiplicatif apparaissant devant une pénalité. Pour commencer, étudions plus en profondeur le problème de l'estimation de σ^2 posé par la pénalité C_p.

1.2. **Estimation de σ^2 et moindres carrés en régression.** On se place dans le cadre de la régression homoscédastique sur un design fixe, avec la perte quadratique et des estimateurs des moindres carrés de la forme $\widehat{F}_m = A_m Y$, où A_m est une matrice de projection orthogonale.

[1] ou quasi-optimale, dans la mesure où aucune inégalité-oracle avec constante $1 + o(1)$ n'existe dans les exemples correspondant à ce cas

Commençons par recalculer le risque empirique et son espérance :

$$\begin{aligned}\frac{1}{n}\left\|Y-\widehat{F}_m\right\|^2 &= \frac{1}{n}\left\|F+\varepsilon-A_mF-A_m\varepsilon\right\|^2\\ &= \frac{1}{n}\left\|(I_n-A_m)F\right\|^2+\frac{1}{n}\left\|(I_n-A_m)\varepsilon\right\|^2\\ &\quad+\frac{2}{n}\left\langle (I_n-A_m)F,\,(I_n-A_m)\varepsilon\right\rangle\end{aligned}$$

et donc

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\left\|Y-\widehat{F}_m\right\|^2\right] &= \frac{1}{n}\left\|(I_n-A_m)F\right\|^2+\frac{\sigma^2\operatorname{tr}\left((I_n-A_m)^\top(I_n-A_m)\right)}{n}\\ &= \frac{1}{n}\left\|(I_n-A_m)F\right\|^2+\frac{\sigma^2(n-D_m)}{n}\quad . \end{aligned}\tag{1}$$

Pour tout $m\in\mathcal{M}_n$, on a ainsi l'estimateur naturel de σ^2

$$\widehat{\sigma}^2_m := \frac{1}{n-D_m}\left\|Y-\widehat{F}_m\right\|^2 \tag{2}$$

qui consiste à utiliser les résidus au sein du modèle S_m pour évaluer le niveau de bruit. D'après (1), on a

$$\mathbb{E}\left[\widehat{\sigma}^2_m\right]=\sigma^2+\frac{1}{n-D_m}\left\|(I_n-A_m)F\right\|^2\quad ,$$

et donc le biais de $\widehat{\sigma}^2_m$ comme estimateur de σ^2 dépend de l'erreur d'approximation $n^{-1}\left\|(I_n-A_m)F\right\|^2$.

Plusieurs stratégies sont possibles pour utiliser un estimateur de la forme (2) dans la pénalité C_p de Mallows [33]

$$\operatorname{pen}_{\mathrm{Cp}}(m)=\frac{2\sigma^2D_m}{n}\quad , \tag{3}$$

voir notamment [23] à ce sujet.

Première solution. Fixons un modèle $m_0\in\mathcal{M}_n$ et utilisons $\widehat{\sigma}^2_{m_0}$ comme estimateur de σ^2 dans (3) : on en déduit le critère de choix de modèle

$$\begin{aligned}\operatorname{crit}(m) &= \frac{1}{n}\left\|Y-\widehat{F}_m\right\|^2+\frac{2\widehat{\sigma}^2_{m_0}D_m}{n}\\ &= \frac{1}{n}\left\|Y-\widehat{F}_m\right\|^2+\frac{2D_m}{n(n-D_{m_0})}\left\|Y-\widehat{F}_{m_0}\right\|^2\quad .\end{aligned}$$

L'inconvénient de cette solution est qu'elle repose fortement sur le fait que l'on connaisse un modèle m_0 pour lequel l'erreur d'approximation est effectivement négligeable devant $(1-D_{m_0}/n)\sigma^2$, sous peine de sur-estimer fortement σ^2 et donc de surpénaliser d'un facteur totalement inconnu. De plus, choisir convenablement m_0 n'est pas simple : en le prenant très grand (de dimension proche de n), on est quasi certain de réduire fortement l'erreur d'approximation, mais en contrepartie on se fonde sur des résidus de petite dimension $n-D_{m_0}$, d'où une variance de $\widehat{\sigma}^2_{m_0}$ assez forte. À l'inverse, prendre

m_0 de dimension plus petite rend $\widehat{\sigma}^2_{m_0}$ moins variable, mais plus certainement biaisé. Un bon compromis peut être de choisir D_{m_0} de l'ordre de $n/2$, à condition de connaître avec certitude un modèle de cette dimension qui a une petite erreur d'approximation. Si l'on pense au cas de la détection de ruptures [27, 7], ce n'est pas forcément si simple. Par exemple, en supposant que $x_1 < \cdots < x_n$ et en prenant pour m_0 le modèle de dimension $n/2$ associé à l'ensemble des fonctions constantes sur $[x_{2i-1}, x_{2i}]$ pour $i = 1 \dots n/2$, on obtient un biais égal à

$$\frac{\|F - F_{m_0}\|^2}{n - D_{m_0}} = \frac{1}{n} \sum_{i=1}^{n/2} \left(\eta(x_{(2i)}) - \eta(x_{2i-1}) \right)^2 \quad .$$

Deuxième solution. Utilisons $\widehat{\sigma}^2_m$ comme estimateur de σ^2 dans (3) *pour chaque modèle m séparément*. Ainsi, on choisit $\widehat{m}$ qui minimise le critère

$$\mathrm{crit}(m) = \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{2\widehat{\sigma}^2_m D_m}{n} = \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 \left(1 + \frac{2D_m}{n - D_m} \right) \quad ,$$

aussi appelé critère FPE (Final Prediction Error, [1]). Avec cette approche, C_p peut être vue comme une *pénalité multiplicative*, au sein d'un critère de la forme

$$\mathrm{crit}(m) = \left\| Y - \widehat{F}_m \right\|^2 \left(1 + \frac{\mathrm{pen}_{\mathrm{mult}}(m)}{n - D_m} \right)$$

qui est proportionnel au risque empirique, voir notamment [9] pour une étude non-asymptotique de pénalités de ce type.

En comparaison de la solution précédente, l'avantage ici est que l'on n'a plus besoin de connaître un modèle m_0 ayant une petite erreur d'approximation. En revanche, comme on minimise un critère proportionnel au risque empirique, il est absolument nécessaire d'exclure d'une manière ou d'une autre les très gros modèles, pour lesquels le risque empirique $\left\| Y - \widehat{F}_m \right\|^2$ est quasi nul. Si on utilise la pénalité C_p, $\mathrm{pen}_{\mathrm{mult}}(m) = 2D_m$, il faut exclure les «gros modèles» a priori. Sinon, il faut utiliser une pénalité plus élaborée [9] dont la taille explose lorsque D_m se rapproche de n.

Validation croisée généralisée. Comme remarqué dans [23], une procédure de choix d'estimateurs très proche (malgré son nom) est la validation croisée généralisée (GCV), introduite par Craven et Wahba [22] ; voir notamment [30, 31, 20] pour des résultats théoriques concernant son optimalité comme procédure de choix d'estimateurs. Pour choisir entre des estimateurs linéaires de la forme $\widehat{F}_m = A_m Y$, GCV consiste à minimiser le critère

$$\mathrm{crit}_{\mathrm{GCV}}(m) := \frac{1}{n} \frac{\left\| Y - \widehat{F}_m \right\|^2}{\left(1 - n^{-1} \operatorname{tr}(A_m) \right)^2} \quad ,$$

soit $\mathrm{crit}_{\mathrm{GCV}}(m) = n^{-1} \|Y - \widehat{F}_m\|^2 (1 - n^{-1} D_m)^{-2}$ dans le cas où A_m est une matrice de projection sur un espace vectoriel de dimension D_m. Lorsque

$D_m \ll n$, on a donc

$$\mathrm{crit}_{\mathrm{GCV}}(m) \approx \frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2 \frac{n + D_m}{n - D_m} = \frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2 \left(1 + \frac{2D_m}{n - D_m}\right)$$

ce qui montre que GCV est apparentée à C_p avec σ^2 estimé par $\widehat{\sigma}^2_m$ pour tout m. En particulier, GCV en possède les défauts mentionnés plus haut.

Quelle alternative ? Idéalement, on voudrait pouvoir estimer précisément σ^2 par une quantité indépendante du modèle considéré et qui ne repose pas sur le choix d'un modèle m_0 particulier. De plus, comme notre objectif est d'utiliser cet estimateur de σ^2 au sein de la pénalité (3), plutôt que d'avoir une garantie sur l'estimation de σ^2 (par exemple une majoration de $\mathbb{E}[(\sigma^2 - \widehat{\sigma}^2)^2]$), on aimerait avoir une garantie sur la performance de la procédure de sélection de modèles associée.

L'objectif du reste de ce cours est de proposer une méthode d'estimation de σ^2, et plus généralement de calibration de pénalités, qui possède ces propriétés.

2. Heuristique de pente en régression homoscédastique

Cette section reprouve des résultats pour l'essentiel obtenus initialement par Birgé et Massart [15, 16], en utilisant une présentation plus proche de résultats postérieurs [8, 5]. On considère le problème de sélection de modèles en régression homoscédastique avec la perte quadratique.

2.1. Mise en évidence heuristique d'une pénalité minimale.

Birgé et Massart [15, 16] ont proposé les premiers d'étudier s'il existe une *pénalité minimale* afin d'obtenir un algorithme de calibration de pénalité. L'idée est la suivante : l'inégalité-oracle du Théorème 1 du premier cours concerne la pénalité

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

lorsque $K > 1$. Est-il possible de prouver qu'à l'inverse, lorsque K est trop petite, par exemple lorsque $K \in]0, 1[$, que le modèle sélectionné est mauvais ? Autrement dit, quel est le *niveau minimal de pénalisation* pour éviter le sur-apprentissage, et avoir une inégalité-oracle ?

Formellement, définissons pour tout $C > 0$ la procédure de sélection de modèles par pénalisation

$$\widehat{m}(C) \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2 + C\frac{D_m}{n} \right\} \quad .$$

On voudrait mettre en évidence $C_{\min} > 0$ telle que :

- Si $C > C_{\min}$, $\widehat{m}(C)$ satisfait une inégalité-oracle (avec constante multiplicative dépendant de C mais bornée quand $n \to \infty$).
- Si $C \in]0, C_{\min}[$, le risque de $\widehat{m}(C)$ explose (son rapport au risque de l'oracle n'est pas borné) et l'estimateur $F_{\widehat{m}(C)}$ sur-apprend clairement.

À des inégalités de concentration près, il est raisonnable de penser que $\widehat{m}(C)$ se comporte à peu près comme

$$\begin{aligned} m^\star(C) &\in \arg\min_{m\in\mathcal{M}_n} \left\{ \mathbb{E}\left[\frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + \frac{C D_m}{n} \right] \right\} \\ &= \arg\min_{m\in\mathcal{M}_n} \left\{ \frac{1}{n} \left\| F - F_m \right\|^2 + (C-\sigma^2)\frac{D_m}{n} \right\} \end{aligned} \tag{4}$$

Au vu de (4), on peut distinguer deux cas :

- Si $C < \sigma^2$, alors $m^\star(C)$ minimise un critère strictement décroissant avec D_m (l'erreur d'approximation étant à peu près décroissante en principe, et le terme $(C-\sigma^2)D_m n^{-1}$ étant strictement décroissant). Par conséquent, $D_{m^\star(C)}$ est proche de la dimension maximale $D_{\max} := \max_{m\in\mathcal{M}_n} D_m$ des modèles.
- Si $C > \sigma^2$, alors le critère minimisé par $m^\star(C)$ est strictement croissant au voisinage des plus gros modèles (à cause du terme $(C-\sigma^2)D_m n^{-1}$, l'erreur d'approximation étant quasi-constante en comparaison). Par conséquent, $D_{m^\star(C)}$ est beaucoup plus petite que $D_{\max}$. De plus, le Théorème 1 du premier cours s'applique à $\widehat{m}(C)$.

Ce raisonnement en espérance suggère donc que dans ce cadre, une pénalité minimale existe effectivement, et que $C_{\min} = \sigma^2$. Un tel résultat, outre son intérêt purement théorique, est particulièrement intéressant pour le problème de calibration de pénalité. En effet, si la perte relative $\|F - \widehat{F}_{\widehat{m}(C)}\|^2$ n'est pas observable (ni comparable à l'oracle), la dimension $D_{\widehat{m}(C)}$ l'est. Or, le raisonnement précédent suggère que cette dimension change très rapidement au voisinage de $C_{\min}$, alors qu'elle varie lentement partout ailleurs.

Cette remarque a conduit Birgé et Massart [15, 16] à proposer l'algorithme suivant.

2.2. **Algorithme de calibration de pénalités.**

(1) pour tout $C > 0$, calculer

$$\widehat{m}(C) \in \arg\min_{m\in\mathcal{M}_n} \left\{ \frac{1}{n} \left\| Y - \widehat{F}_m \right\|^2 + C\frac{D_m}{n} \right\}$$

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

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

Remarquons que le premier point de cet algorithme n'est pas coûteux algorithmiquement lorsqu'on a déjà calculé le risque empirique pour tout m. En effet, la trajectoire $(\widehat{m}(C))_{C>0}$ est constante par morceaux, avec au plus $\mathrm{Card}(\mathcal{M}_n)$ morceaux, et peut être calculée efficacement en utilisant un algorithme détaillé par [26, 8].

2.3. **Garanties théoriques.** Tout d'abord, nous avons besoin d'inégalités de concentration. En supposant le bruit Gaussien, la Proposition 3 du premier cours permet de garantir que le raisonnement effectué plus haut est,

pour l'essentiel, correct. Par ailleurs, nous avons implicitement supposé que l'erreur d'approximation ne cause plus de variations majeures du critère empirique pénalisé dès lors que D_m est assez grand. Il faut donc faire une hypothèse sur l'erreur d'approximation, en réalité bien faible car il est suffisant de disposer de deux modèles (un de dimension «modérée» et un de grande dimension) pour lesquels l'erreur d'approximation est majorée.

On obtient le résultat suivant.

Théorème 1. *On se place dans le cadre de la régression à design fixe, avec la perte des moindres carrés. On suppose donnée une famille d'estimateurs des moindres carrés $(\widehat{F}_m)_{m\in\mathcal{M}_n}$ associée à une famille de modèles $(S_m)_{m\in\mathcal{M}_n}$ qui sont des s.e.v. de dimension finie de $\mathbb{R}^n$. On suppose que $\mathrm{Card}(\mathcal{M}_n) \leq C_{\mathcal{M}} n^{\alpha}$ pour des constantes $C_{\mathcal{M}}, \alpha > 0$.*

On suppose que les données sont de la forme $Y = F + \varepsilon$ avec $F \in \mathbb{R}^n$ inconnu et $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ (cadre Gaussien homoscédastique). On note A_m la matrice de projection orthogonale sur S_m, de telle sorte que $\widehat{F}_m = A_m Y$ et $\mathrm{tr}(A_m) = \dim(S_m) = D_m$.

On suppose qu'il existe $m_1, m_2 \in \mathcal{M}_n$ tels que

$$D_{m_1} \geq \frac{n}{2} \quad D_{m_2} \leq \sqrt{n} \quad \text{et} \quad \forall i \in \{1,2\}\ , \quad \frac{1}{n}\left\|(I - A_{m_i})F\right\|^2 \leq \sigma^2 \sqrt{\frac{\ln(n)}{n}}\ .$$

Pour tout $C > 0$, on définit

$$\widehat{m}(C) \in \arg\min_{m\in\mathcal{M}_n} \left\{ \frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2 + C\frac{D_m}{n} \right\}\ .$$

Alors, il existe $n_0(\alpha)$ tel que si $n \geq n_0(\alpha)$, avec probabilité au moins $1 - 4C_{\mathcal{M}} n^{-2}$,

$$\forall C < \left(1 - 42\sqrt{\frac{(\alpha+2)\ln(n)}{n}}\right)\sigma^2\,, \quad D_{\widehat{m}(C)} \geq \frac{n}{3} \tag{5}$$

$$\forall C > \left(1 + 8\frac{\sqrt{(\alpha+2)\ln(n)}}{n^{1/4}}\right)\sigma^2\,, \quad D_{\widehat{m}(C)} \leq n^{3/4} \tag{6}$$

et dans le premier cas, $\left\|F - \widehat{F}_{\widehat{m}(C)}\right\|^2 \geq \ln(n)\inf_{m\in\mathcal{M}_n}\left\{\left\|F - \widehat{F}_m\right\|^2\right\}$.

Une première version, centrée sur le cas $C < \sigma^2$, du Théorème 1 a été obtenue par Birgé et Massart [15, 16], le cas $s^\star = 0$ ayant été préalablement traité par [14, Proposition 8]. La formulation exacte ci-dessus est un cas particulier obtenu ensuite dans [5, 6] dans le cas plus général des estimateurs linéaires.

Preuve du Théorème 1. On découpe cette preuve en 5 étapes principales :

(1) Pour chaque $x \geq 0$ et $m \in \mathcal{M}_n$, on concentre $\langle A_m\varepsilon, \varepsilon\rangle$ autour de $\sigma^2 \operatorname{tr}(A_m)$ et $\langle (A_m - I_n)F, \varepsilon\rangle$ autour de zéro. Ceci définit un événement $\Omega_{m,x}$ de probabilité $1 - 4e^{-x}$, sur lequel

$$\left|\langle A_m\varepsilon, \varepsilon\rangle - \sigma^2 D_m\right| \leq 2\sqrt{xD_m}\sigma^2 + 2x\sigma^2 \tag{7}$$

$$|\langle (A_m - I_n)F, \varepsilon\rangle| \leq \sqrt{2x}\,\|(A_m - I_n)F\|\,\sigma \ . \tag{8}$$

(2) Soit $\Omega_x = \bigcap_{m\in\mathcal{M}} \Omega_{m,x}$, pour lequel une borne d'union donne $\mathbb{P}(\Omega_x) \geq 1 - 4\operatorname{Card}(\mathcal{M}_n)e^{-x}$. On va prendre $x = (\alpha+2)\ln(n)$, et c'est pourquoi $\mathbb{P}(\Omega_x) \geq 1 - 4C_{\mathcal{M}}n^{-2}$.

(3) Pour tout $C > 0$, $\widehat{m}(C)$ minimise

$$\begin{aligned}
\operatorname{crit}_C(m) &= \frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2 + C\frac{D_m}{n} \\
&= \frac{1}{n}\Big[\,\|(I_n - A_m)F\|^2 + \|(I_n - A_m)\varepsilon\|^2 \\
&\qquad +2\,\langle (I_n - A_m)F, (I_n - A_m)\varepsilon\rangle + CD_m\,\Big] \\
&= \frac{1}{n}\Big[\,\|(I_n - A_m)F\|^2 + \|\varepsilon\|^2 - \langle \varepsilon, A_m\varepsilon\rangle \\
&\qquad +2\,\langle (I_n - A_m)F, \varepsilon\rangle + CD_m\,\Big] \\
&= \frac{1}{n}\left[\operatorname{crit}'_C(m) + \|\varepsilon\|^2\right]
\end{aligned}$$

$$\text{avec}\quad \operatorname{crit}'_C(m) = \|(I_n - A_m)F\|^2 + CD_m - \langle \varepsilon, A_m\varepsilon\rangle + 2\,\langle (I_n - A_m)F, \varepsilon\rangle \ .$$

Par conséquent, $\widehat{m}(C)$ minimise $\operatorname{crit}'_C(m)$. De plus, en utilisant (7) et (8), sur Ω_x, pour tout $m \in \mathcal{M}_n$ et $C > 0$,

$$\begin{aligned}
\operatorname{crit}'_C(m) &\leq (C - \sigma^2)D_m + \|(I_n - A_m)F\|^2 + \left[2\sqrt{xD_m} + 2x\right]\sigma^2 \\
&\qquad + \sqrt{2x}\,\|(A_m - I_n)F\|\,\sigma \\
&\leq (C - \sigma^2)D_m + \|(I_n - A_m)F\|^2 + \sqrt{2(\alpha+2)\ln(n)}\,\|(A_m - I_n)F\|\,\sigma \\
&\qquad + \left[2\sqrt{(\alpha+2)\ln(n)n} + 2(\alpha+2)\ln(n)\right]\sigma^2 \\
&\leq (C - \sigma^2)D_m + \|(I_n - A_m)F\|^2 + 3\sqrt{(\alpha+2)\ln(n)n}\sigma^2 \qquad (9) \\
&\qquad + \sqrt{2(\alpha+2)\ln(n)}\,\|(A_m - I_n)F\|\,\sigma \\
\operatorname{crit}'_C(m) &\geq (C - \sigma^2)D_m - \left[2\sqrt{xD_m} + 2x\right]\sigma^2 \\
&\qquad + \|(I_n - A_m)F\|^2 - \sqrt{2x}\,\|(A_m - I_n)F\|\,\sigma \\
&= (C - \sigma^2)D_m - \left[2\sqrt{xD_m} + 2x\right]\sigma^2 \\
&\qquad + \left(\|(I_n - A_m)F\| - \sqrt{\frac{x\sigma^2}{2}}\right)^2 - \frac{x\sigma^2}{2} \\
&\geq (C - \sigma^2)D_m - \left[2\sqrt{xn} + \frac{5x}{2}\right]\sigma^2
\end{aligned}$$

$$= (C-\sigma^2)D_m - \left[2\sqrt{(\alpha+2)\ln(n)n} + \frac{5(\alpha+2)\ln(n)}{2}\right]\sigma^2$$
$$\geq (C-\sigma^2)D_m - 3\sqrt{(\alpha+2)\ln(n)n}\sigma^2 \tag{10}$$

en supposant que $n \geq n_0(\alpha)$.

(4) On suppose que $C < \sigma^2$, et l'on cherche à prouver que $D_{\widehat{m}(C)} \geq n/3$ dès que C est suffisamment éloignée de σ^2. Comme $\widehat{m}(C)$ minimise crit'_C, il suffit par exemple de montrer que

$$\inf_{m\in\mathcal{M}_n \text{t.q.} D_m < n/3} \left\{ \mathrm{crit}'_C(m) \right\} > \mathrm{crit}'_C(m_1) \ .$$

Commençons par majorer $\mathrm{crit}'_C(m_1)$: sur Ω_x, d'après (9) et en utilisant les hypothèses sur m_1, pour tout $C \in \left]0, \sigma^2\right[$,

$$\begin{aligned}\mathrm{crit}'_C(m_1) &\leq (C-\sigma^2)D_{m_1} + \|(I_n - A_{m_1})F\|^2 + 3\sqrt{(\alpha+2)\ln(n)n}\sigma^2 \\ &\quad + \sqrt{2(\alpha+2)\ln(n)}\,\|(A_m - I_n)F\|\,\sigma \\ &\leq (C-\sigma^2)\frac{n}{2} + \sqrt{\ln(n)n}\sigma^2 + 3\sqrt{(\alpha+2)\ln(n)n}\sigma^2 \\ &\quad + \sqrt{2(\alpha+2)\ln(n)\sqrt{\ln(n)n}}\sigma^2 \\ &\leq (C-\sigma^2)\frac{n}{2} + 4\sqrt{(\alpha+2)\ln(n)n}\sigma^2 \end{aligned} \tag{11}$$

en supposant que $n \geq n_0(\alpha)$ (quitte à augmenter la valeur de $n_0(\alpha)$).

Par ailleurs, pour tout $m \in \mathcal{M}_n$ tel que $D_m < n/3$, d'après (9), sur Ω_x, pour tout $C \in \left]0, \sigma^2\right[$ et $n \geq n_0(\alpha)$,

$$\begin{aligned}\mathrm{crit}'_C(m) &\geq (C-\sigma^2)D_m - 3\sqrt{(\alpha+2)\ln(n)n}\sigma^2 \\ &\geq (C-\sigma^2)\frac{n}{3} - 3\sqrt{(\alpha+2)\ln(n)n}\sigma^2 \ .\end{aligned} \tag{12}$$

La borne inférieure de (12) est strictement supérieure à la borne supérieure de (11) si et seulement si

$$(C-\sigma^2)\frac{n}{3} - 3\sqrt{(\alpha+2)\ln(n)n}\sigma^2 > (C-\sigma^2)\frac{n}{2} + 4\sqrt{(\alpha+2)\ln(n)n}\sigma^2$$

que l'on peut réécrire

$$(\sigma^2 - C)\frac{n}{6} > 7\sqrt{(\alpha+2)\ln(n)n}\sigma^2 \ ,$$

d'où l'on déduit (5).

La minoration de $\left\|\widehat{F}_{\widehat{m}(C)} - F\right\|^2$ en découle directement, voir [5].

(5) On suppose que $C > \sigma^2$, et l'on cherche à prouver que $D_{\widehat{m}(C)} \leq n^{3/4}$ dès que C est suffisamment éloignée de σ^2. Comme $\widehat{m}(C)$ minimise crit'_C, il suffit par exemple de montrer que

$$\inf_{m\in\mathcal{M}_n \text{t.q.} D_m > n^{3/4}} \left\{ \mathrm{crit}'_C(m) \right\} > \mathrm{crit}'_C(m_2) \ .$$

Commençons par majorer $\mathrm{crit}'_C(m_2)$, de la même manière que l'on a obtenu (11) : sur Ω_x, d'après (10) et en utilisant les hypothèses sur

m_2, pour tout $C > \sigma^2$,

$$\mathrm{crit}'_C(m_2) \leq (C-\sigma^2)\sqrt{n} + 4\sqrt{(\alpha+2)\ln(n)n}\sigma^2 \tag{13}$$

en supposant que $n \geq n_0(\alpha)$.

Par ailleurs, pour tout $m \in \mathcal{M}_n$ tel que $D_m > n^{3/4}$, d'après (9), sur Ω_x, pour tout $C > \sigma^2$ et $n \geq n_0(\alpha)$,

$$\mathrm{crit}'_C(m) \geq (C-\sigma^2)n^{3/4} - 3\sqrt{(\alpha+2)\ln(n)n}\sigma^2 \ . \tag{14}$$

La borne inférieure de (14) est strictement supérieure à la borne supérieure de (13) si et seulement si

$$(C-\sigma^2)(n^{3/4}-\sqrt{n}) > 7\sqrt{(\alpha+2)\ln(n)n}\sigma^2$$

qui est vérifiée dès lors que n est assez grand et

$$(C-\sigma^2) > \frac{8\sqrt{(\alpha+2)\ln(n)}}{n^{1/4}}\sigma^2 \ ,$$

d'où l'on déduit (6). Remarquons qu'en remplaçant $n^{3/4}$ dans l'énoncé par $b_n \gg \sqrt{n}$, on obtiendrait de la même manière que sur Ω_x, pour tout $n \geq n_0(\alpha)$ (et $b_n \geq 8\sqrt{n}/7$),

$$\forall C > \left(1 + \frac{8\sqrt{(\alpha+2)\ln(n)}}{b_n n^{-1/2}}\right)\sigma^2 \ , \qquad D_{\widehat{m}(C)} \leq b_n \ .$$

□

Pour justifier pleinement l'algorithme de calibration de pénalité de la Section 2.2, il reste à prouver que l'on a une inégalité-oracle avec constante $1+o(1)$ lorsqu'on utilise une pénalité de la forme

$$\mathrm{pen}(m) = \frac{2\widehat{C}D_m}{n}$$

avec $\widehat{C}$ aléatoire (dépendant des données), dont on sait seulement qu'avec grande probabilité,

$$(2-\eta_-)\sigma^2 = 2\left(1 - 42\sqrt{\frac{(\alpha+2)\ln(n)}{n}}\right)\sigma^2$$

$$\leq 2\widehat{C} \leq 2\left(1 + 8\frac{\sqrt{(\alpha+2)\ln(n)}}{n^{1/4}}\right)\sigma^2 = (2+\eta_+)\sigma^2$$

L'idée clé ici est de remarquer que l'événement défini au Théorème 1 du premier cours est le même pour toutes les valeurs de K (et le même que l'événement défini au Théorème 1, si l'on prend $x = (\alpha+2)\ln(n)$). Par conséquent, $\widehat{m}(2\widehat{C})$ vérifie l'inégalité-oracle du Théorème 1 du premier cours en prenant un sup sur $K \in \left[(2-\eta_-)\sigma^2, (2+\eta_+)\sigma^2\right]$ pour le membre de droite. Comme $C(K)$ est une fonction de K décroissante sur $]1,2]$ et croissante sur $[2,+\infty[$, on a montré qu'avec probabilité au moins $1 - 4C_{\mathcal{M}}n^{-2}$,

si $n \geq n_0(\alpha)$,

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}(2\widehat{C})} - F\right\|^2 \leq \left(\max\left\{1+\eta_+\,,\,(1-\eta_-)^{-1}\right\}+\delta\right)\inf_{m\in\mathcal{M}_n}\left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\} + \frac{\max\left\{C(2-\eta_-)\,,\,C(2+\eta_+)\right\}x\sigma^2}{\delta n}\ .$$

Quitte à augmenter à nouveau $n_0(\alpha)$, on en déduit l'inégalité-oracle : avec probabilité au moins $1 - 4C_{\mathcal{M}}n^{-2}$, pour tout $n \geq n_0(\alpha)$, pour tout $\delta > 0$,

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}(2\widehat{C})} - F\right\|^2 \leq \left(1 + 16\frac{\sqrt{(\alpha+2)\ln(n)}}{n^{1/4}} + \delta\right)\inf_{m\in\mathcal{M}_n}\left\{\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right\} + \frac{Lx\sigma^2}{\delta n}\ ,$$

où $L > 0$ est une constante numérique.

3. L'heuristique de pente

Les résultats de la section précédente ont naturellement conduit à une heuristique plus générale, appelée *heuristique de pente*, et à un algorithme de calibration de pénalités associé. Cette heuristique et cet algorithme ont été initialement formulés dans la prépublication [15] et dans l'article qui a suivi [16]. On les trouve également exposés dans [34], [17, Section 2] et [35, Section 8.5.2]. Précisons que la terminologie «pente» correspond au fait que dans le cadre de la Section 2, le risque empirique a un comportement linéaire en fonction de D_m pour les plus gros modèles, et que la *pente* du risque empirique valant $-\sigma^2/n$, elle permet d'estimer σ^2.

3.1. Heuristique de pente : première formulation.

En utilisant les notations introduites au premier cours, on peut résumer l'heuristique de pente ainsi :

(1) il existe une *pénalité minimale* $\mathrm{pen}_{\min}$ telle que si

$$\widehat{m}_{\min}(C) \in \arg\min_{m\in\mathcal{M}_n}\left\{P_n\gamma\left(\widehat{s}_m\right) + C\,\mathrm{pen}_{\min}(m)\right\}$$

le rapport entre l'excès de risque de $\widehat{m}_{\min}(C)$ et celui de l'oracle est borné pour $C > 1$ et non-borné pour $C < 1$ (quand $n \to \infty$).

(2) soit $\mathcal{C}_m$ une mesure de *complexité* du modèle S_m (par exemple, sa dimension), supposée observable. Alors, la pénalité minimale est détectable au sens où $\mathcal{C}_{\widehat{m}_{\min}(C)}$ «saute» au voisinage de $C = 1$.

(3) la pénalité optimale (au sens de l'objectif de prédiction) $\mathrm{pen}_{\mathrm{opt}}$ est reliée à la pénalité minimale par la relation

$$\mathrm{pen}_{\mathrm{opt}}(m) \approx 2\,\mathrm{pen}_{\min}(m)\ .$$

3.2. Algorithme correspondant.

On en déduit l'algorithme de calibration de pénalités, supposant connue $\mathrm{pen}_0(m) \propto \mathrm{pen}_{\min}(m)$ ainsi qu'une mesure de complexité $\mathcal{C}_m$ pour chaque $m \in \mathcal{M}_n$:

(1) pour tout $C > 0$, calculer

$$\widehat{m}(C) \in \arg\min_{m \in \mathcal{M}_n} \{ P_n \gamma(\widehat{s}_m) + C \operatorname{pen}_0(m) \} \quad ,$$

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

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

Notons que la proposition d'utiliser une mesure de complexité $\mathcal{C}_m$ quelconque (et pas nécessairement la dimension D_m) a été proposée initialement par [28].

3.3. **Quelle pénalité minimale ?** Un candidat naturel pour construire une pénalité minimale a été proposé par [17, Section 2], [35, Section 8.5.2] puis [8], il s'agit de l'*excès de risque empirique* au sein du modèle S_m, défini pour tout $m \in \mathcal{M}_n$ par

$$\widehat{v}_m = p_2(m) := P_n(\gamma(s_m^\star) - \gamma(\widehat{s}_m)) \quad .$$

Nous utiliserons dans la suite la notation p_2, qui est utilisée notamment par [8].

En effet, lorsque $\operatorname{pen}(m) = p_2(m)$ (ou son espérance), le critère pénalisé

$$\operatorname{crit}(m) = P_n\gamma(\widehat{s}_m) + p_2(m) = P_n\gamma(s_m^\star)$$

a pour espérance $P\gamma(s_m^\star)$, qui est (approximativement) décroissante car elle est (à constante additive près) égale à l'erreur d'approximation $\ell(s^\star, s_m^\star)$. Dès lors que $\mathbb{E}[p_2(m)]$ croît suffisamment vite avec la complexité de S_m, on peut donc escompter que pénaliser d'un facteur $(1+\delta)\mathbb{E}[p_2(m)]$ avec $\delta > 0$ suffit à sélectionner un modèle de complexité «raisonnable», tandis que pénaliser d'un facteur $(1-\delta)\mathbb{E}[p_2(m)]$ avec $\delta > 0$ condamne à sélectionner un modèle de complexité quasi-maximale dans la famille $\mathcal{M}_n$.

Définissons également, pour tout $m \in \mathcal{M}_n$,

$$p_1(m) = P(\gamma(\widehat{s}_m) - \gamma(s_m^\star))$$
$$\delta(m) = (P - P_n)\gamma(s_m^\star) \quad .$$

La première quantité, $p_1(m)$, est l'*excès de risque* au sein du modèle S_m. La seconde quantité, $\delta(m)$, est d'espérance nulle car $s_m^\star$ est déterministe. Avec ces notations, on peut écrire la pénalité idéale

$$\operatorname{pen}_{\mathrm{id}}(m) = (P - P_n)\gamma(\widehat{s}_m) = p_1(m) - \delta(m) + p_2(m) \quad .$$

D'après le principe d'estimation sans biais du risque, l'espérance de la pénalité idéale

$$\mathbb{E}[\operatorname{pen}_{\mathrm{id}}(m)] = \mathbb{E}[p_1(m)] + \mathbb{E}[p_2(m)]$$

est une pénalité optimale. Si l'on admet que $p_2(m)$ ou son espérance est effectivement une pénalité minimale, on peut réécrire le troisième point de l'heuristique de pente sous la forme

$$\mathbb{E}[p_1(m)] \approx \mathbb{E}[p_2(m)]$$

ou encore

$$p_1(m) \approx p_2(m) \ ,$$

c'est-à-dire, *l'excès de risque au sein de S_m est proche de l'excès de risque empirique au sein de S_m.*

Dans le cadre de la section précédente, on a montré que

$$p_1(m) = \frac{1}{n}\left\|F - \widehat{F}_m\right\|^2 - \frac{1}{n}\left\|F - A_m F\right\|^2$$
$$p_2(m) = \frac{1}{n}\left\|Y - A_m F\right\|^2 - \frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2$$
$$\mathbb{E}\left[p_1(m)\right] = \frac{D_m\sigma^2}{n} = \mathbb{E}\left[p_2(m)\right]$$

et que $p_1(m)$ comme $p_2(m)$ se concentrent autour de leur espérance (sous l'hypothèse Gaussienne). En s'inspirant de ce premier résultat, on peut identifier les principaux ingrédients d'une validation théorique de l'heuristique de pente dans un cadre général[2] :

- Concentration de l'excès de risque $p_1(m)$ autour de son espérance.
- Concentration de l'excès de risque empirique $p_2(m)$ autour de son espérance.
- Proximité de leurs espérances respectives : $\mathbb{E}\left[p_1(m)\right] \approx \mathbb{E}\left[p_2(m)\right]$.
- Contrôle de $\mathbb{E}\left[p_2(m)\right]$ en fonction de la complexité $\mathcal{C}_m$: $\mathbb{E}\left[p_2(m)\right]$ doit compenser largement la décroissance de l'erreur d'approximation pour les m de grande complexité.
- Contrôle des termes de restes (déviations et espérances).

3.4. **Résultats mathématiques existants.** Des résultats théoriques validant l'heuristique de pente ont été obtenus dans les cadres suivants :

- Estimateurs des moindres carrés en *régression homoscédastique Gaussienne* sur un plan d'expérience fixe [16]. Notons que des résultats partiels (explosion du risque sous la pénalité minimale) y sont aussi prouvés pour une famille de modèles de complexité exponentielle ; un résultat similaire est aussi prouvé lorsque la famille de modèles a une complexité intermédiaire, avec l'hypothèse additionnelle $s^\star = 0$.
- *Régressogrammes* avec la perte des moindres carrés en régression *hétéroscédastique* sur un plan d'expérience aléatoire [8], avec des données *non-Gaussiennes.*
- *Estimation de densité par moindres carrés*, avec des données i.i.d. [36] ou *mélangeantes* [29]. Notons que le cadre de l'estimation de densité par moindres carrés est remarquable car on y a $p_1(m) = p_2(m)$ p.s. Par ailleurs, c'est dans ce cadre qu'a été suggérée initialement l'idée d'utiliser une mesure de complexité $\mathcal{C}_m$ différente de la dimension D_m.
- Estimateurs de moindres carrés en régression hétéroscédastique bornée [43], pour des familles de modèles incluant les *polynômes par morceaux.*

[2]cadre potentiellement plus général que celui de la sélection de modèles, pour peu que l'on soit capables de définir l'équivalent de $s^\star_m$ pour chaque estimateur $\widehat{s}_m$

– Estimation de densité par *maximum de vraisemblance* sur des modèles d'histogrammes [41]. Notons que ce résultat (comme le précédent) est issu d'une approche générale visant à étudier l'heuristique de pente pour des estimateurs par minimum de contraste, en supposant que le contraste γ est «régulier» [40]. En particulier, il s'agit du premier résultat concernant l'heuristique de pente où la perte n'est pas mesurée avec le contraste des moindres carrés.

Voir également l'article de survol [12] sur l'heuristique de pente.

3.5. **Résultats empiriques.** De plus, un nombre encore plus important de travaux expérimentaux donne des arguments en faveur de l'utilisation de l'heuristique de pente dans divers autres cadres, par exemple :

– la détection de ruptures [27],
– les modèles de mélanges Gaussiens [38],
– la classification non-supervisée (choix du nombre de classes) [11],
– la géométrie computationnelle [19],
– la calibration du Lasso [21].

Une liste plus complète est indiquée dans l'article de survol [12] sur l'heuristique de pente.

4. Estimateurs linéaires en régression

Les résultats mentionnés ci-dessus concernent tous des problèmes de sélection de modèles. Il est naturel de se demander dans quelle mesure on peut généraliser l'heuristique de pente à des problèmes de sélection d'estimateurs. Le cadre le plus simple et le plus naturel à considérer est celui de la sélection d'estimateurs linéaires en régression, qui englobe notamment le cadre traité en Section 2. Incidemment, considérer ce cadre amène à reformuler l'heuristique de pente d'une manière plus générale. Les résultats présentés dans cette section sont issus de [5, 6].

On se place dans le cadre de la régression homoscédastique sur un plan d'expérience fixe, où l'on observe $Y \in \mathbb{R}^n$ avec $Y = F + \varepsilon$ et

$$\forall i \in \{1, \dots, n\} \ , \quad \mathbb{E}[\varepsilon_i] = 0 \quad \text{et} \quad \mathbb{E}\left[\varepsilon_i^2\right] = \sigma^2 \ .$$

L'objectif est d'estimer F.

On dit alors qu'un estimateur $\widehat{F}$ de F est un *estimateur linéaire* lorsque

$$\widehat{F} = AY$$

pour une matrice $A \in \mathcal{M}_n(\mathbb{R})$ déterministe. Notons que le plan d'expérience $x_1, \dots, x_n$ étant également déterministe, la matrice A est autorisée à dépendre de $x_1, \dots, x_n$.

4.1. **Exemples.** Un premier exemple est celui que nous avons largement traité jusqu'à présent : si A est une matrice de projection orthogonale sur un sous-espace vectoriel de $\mathbb{R}^n$, alors $\widehat{F} = AY$ est un estimateur des moindres carrés. Mentionnons trois autres exemples principaux.

Régression ridge à noyau. On suppose donné un noyau défini positif $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, et l'on cherche une fonction $f : \mathcal{X} \to \mathbb{R}$ dans l'espace de Hilbert à noyau reproduisant (RKHS) associé $\mathcal{F}$, muni de la norme $\|\cdot\|_{\mathcal{F}}$. Alors, l'estimateur par régression ridge — aussi appelé estimateur par splines régularisés dans le cas des noyaux associés aux splines [46] — est obtenu comme solution du problème de minimisation [44]

$$\min_{f\in\mathcal{F}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(x_i))^2 + \lambda \|f\|_{\mathcal{F}}^2 \right\} .$$

D'après le théorème du représentant, la solution unique de ce problème de minimisation est de la forme $\widehat{f} = f_{\widehat{\alpha}}$ avec $f_\alpha = \sum_{i=1}^n \alpha_i k(\cdot, x_i)$ pour $\alpha \in \mathbb{R}^n$. Notons K la matrice de noyau $n \times n$, définie par $K_{ab} = k(x_a, x_b)$. On a alors $\|f_\alpha\|_{\mathcal{F}}^2 = \alpha^\top K\alpha$ et $(f_\alpha(x_i))_{1\le i\le n} = K\alpha$, et donc

$$\widehat{\alpha} \in \arg\min_{\alpha\in\mathbb{R}^n} \left\{ \frac{1}{n}\|Y - K\alpha\|^2 + \lambda\alpha^\top K\alpha \right\} .$$

S'agissant de la minimisation d'une fonction strictement convexe (si $\lambda > 0$), $\widehat{\alpha}$ est défini de manière unique comme solution de la condition du premier ordre, et l'on en déduit que $\widehat{\alpha} = (K + n\lambda I)^{-1}Y$. Ainsi, $\widehat{F} = (\widehat{f}(x_i))_{1\le i\le n} = A_{\lambda,K}Y$ avec une matrice de régularisation $A_{\lambda,K} = K(K + n\lambda I_n)^{-1}$, paramétrée par le paramètre de régularisation $\lambda > 0$ et la matrice de noyau K.

Remarquons que le cas de la régression ridge en est un exemple : si X est la matrice de design (de taille $n \times p$, dont la i-ème ligne contient les coordonnées de x_i dans une base de $\mathbb{R}^p$), cela correspond à la matrice de noyau $K = XX^\top$.

k plus proches voisins. Soit d une distance sur $\mathcal{X}$. Pour tout $x \in \mathcal{X}$, on peut ainsi définir l'ensemble $\mathcal{E}_k(x)$ des k plus proches voisins de $\mathcal{X}$ parmi $\{x_1, \dots, x_n\}$. La valeur en x_i de l'estimateur des k plus proches voisins vaut alors

$$\widehat{F}_i = \frac{1}{k}\sum_{x_j\in\mathcal{E}_k(x_i)} Y_j = \sum_{1\le j\le n}\left(\frac{1}{k}\mathbb{1}_{x_j\in\mathcal{E}_k(x_i)}Y_j\right)$$

si bien que $\widehat{F} = A(k)Y$ pour la matrice $A(k) \in \mathcal{M}_n(\mathbb{R})$ définie par

$$\forall i, j \in \{1, \dots, n\} , \quad A(k)_{i,j} = \frac{1}{k}\mathbb{1}_{x_j\in\mathcal{E}_k(x_i)} .$$

Estimateurs de Nadaraya-Watson. Soit $K : \mathcal{X} \times \mathcal{X} \mapsto [0, +\infty[$, fonction appelée «noyau» (à ne pas confondre avec les noyaux utilisés en régression ridge). L'estimateur de Nadaraya-Watson est alors défini par $\widehat{F} = A(K)Y$ avec

$$\forall i, j \in \{1, \dots, n\} , \quad A(K)_{i,j} = \frac{K(x_i, x_j)}{\sum_{1\le\ell\le n} K(x_i, x_\ell)} .$$

Typiquement, le noyau est défini par $K(x, y) = g(d(x, y)/h)$ pour une distance d sur $\mathcal{X}$, une largeur de bande $h > 0$ et une fonction $g : [0, +\infty[\mapsto$

$[0,+\infty[$ décroissante. On parle alors de *noyau Gaussien* lorsque $g(t) = \exp(-t^2)$, et de *noyau fenêtre* lorsque $g(t) = \mathbb{1}_{t\in[0,1]}$.

4.2. **Sélection d'estimateurs linéaires.** Soit $(A_m)_{m\in\mathcal{M}_n}$ une famille de matrices (déterministes) et $(\widehat{F}_m)_{m\in\mathcal{M}_n}$ les estimateurs linéaires associés. L'objectif est de résoudre le problème de sélection d'estimateurs ainsi posé. Il est intéressant de noter que les calculs faits dans le cadre des moindres carrés se transposent quasi-intégralement ici, en faisant uniquement attention à ne pas utiliser la relation $A_m = A_m^\top = A_m^\top A_m$ qui caractérise les estimateurs par projection.

Risque. Tout d'abord, la perte relative de $\widehat{F}_m$ s'écrit

$$\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 = \frac{1}{n}\left\|(A_m - I_n)F\right\|^2 + \frac{1}{n}\left\|A_m\varepsilon\right\|^2 + \frac{2}{n}\left\langle A_m\varepsilon,\, (A_m - I_n)F\right\rangle \quad , \tag{15}$$

$$\text{d'où}\quad \mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - F\right\|^2\right] = \frac{1}{n}\left\|(A_m - I_n)F\right\|^2 + \frac{\sigma^2\operatorname{tr}(A_m^\top A_m)}{n} \quad . \tag{16}$$

En particulier, on peut encore définir dans ce cadre une erreur d'approximation $n^{-1}\|(A_m - I)F\|^2$, l'équivalent de $s_m^\star$ étant ici $F_m = A_m F = \mathbb{E}[A_m Y] = A_m\mathbb{E}[Y]$. Insistons sur le fait pour un problème de sélection d'estimateur général, disposer d'une telle notion naturelle de biais est loin d'être évident.

Le deuxième terme apparaissant dans l'expression du risque (16) peut donc être interprété comme une erreur d'estimation, où l'on a remplacé la dimension D_m du modèle par $\operatorname{tr}(A_m^\top A_m)$.

Risque empirique et pénalité idéale. On peut de la même manière calculer le risque empirique de $\widehat{F}_m$:

$$\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 = \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{2}{n}\left\langle A_m\varepsilon,\, \varepsilon\right\rangle + \frac{1}{n}\left\|\varepsilon\right\|^2 - \frac{2}{n}\left\langle (A_m - I_n)\, F,\, \varepsilon\right\rangle \tag{17}$$

$$= \frac{1}{n}\left\|(A_m - I_n)F\right\|^2 + \frac{1}{n}\left\langle \left(A_m^\top A_m - 2A_m\right)\varepsilon,\, \varepsilon\right\rangle + \frac{1}{n}\left\|\varepsilon\right\|^2 + \frac{2}{n}\left\langle \varepsilon,\, \left(A_m^\top - I_n\right)(A_m - I_n)\, F\right\rangle \tag{18}$$

en utilisant (15). On déduit de (18) l'espérance du risque empirique

$$\mathbb{E}\left[\frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2\right] = \frac{1}{n}\left\|(A_m - I_n)F\right\|^2 + \frac{\sigma^2\left(\operatorname{tr}\left(A_m^\top A_m\right) - 2\operatorname{tr}\left(A_m\right)\right)}{n} \quad .$$

Par ailleurs, on peut extraire de (17) la pénalité idéale (en lui ajoutant $n^{-1}\left\|\varepsilon\right\|^2$ qui ne dépend pas de m)

$$\begin{aligned}\operatorname{pen}_{\mathrm{id}}(m) &= \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 - \frac{1}{n}\left\|\widehat{F}_m - Y\right\|^2 + \frac{1}{n}\left\|\varepsilon\right\|^2 \\ &= \frac{2}{n}\left\langle A_m\varepsilon,\, \varepsilon\right\rangle + \frac{2}{n}\left\langle (A_m - I_n)F,\, \varepsilon\right\rangle \end{aligned} \tag{19}$$

puis son espérance

$$\mathbb{E}\left[\mathrm{pen}_{\mathrm{id}}(m)\right] = \frac{2\sigma^2 \operatorname{tr}\left(A_m\right)}{n} \ . \tag{20}$$

La pénalité (20) est aussi appelée C_L de Mallows [33]. Dans la mesure où $\operatorname{tr}(A_m)$ est la quantité qui généralise la dimension d'un modèle, on l'appelle souvent *nombre de degrés de liberté généralisé*.

Pénalité minimale. Grâce au fait que $s_m^\star$ dispose d'une définition naturelle dans ce cadre de sélection d'estimateurs, on peut calculer $p_2(m)$ et son espérance, candidat naturel au titre de pénalité minimale :

$$\begin{aligned} p_2(m) &= \frac{1}{n}\left\|Y - F_m\right\|^2 - \frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2 \\ &= \frac{2}{n}\left\langle \varepsilon,\, A_m\varepsilon\right\rangle - \frac{1}{n}\left\|A_m\varepsilon\right\|^2 - \frac{2}{n}\left\langle \varepsilon,\, A_m^\top(I_n - A_m)F\right\rangle \end{aligned}$$

donc

$$\mathrm{pen}_{\min}(m) = \mathbb{E}\left[p_2(m)\right] = \frac{\sigma^2\left(2\operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m)\right)}{n} \ . \tag{21}$$

Contrairement à ce que l'heuristique de pente préconise, *la pénalité minimale n'est donc pas proportionnelle à la pénalité optimale* pour les estimateurs linéaires. Néanmoins, la pénalité minimale étant connue à la constante σ^2 près, on peut toujours l'utiliser pour estimer σ^2, et *in fine* utiliser cet estimateur de σ^2 au sein de la pénalité C_L (20).

Remarquons qu'il est nécessaire ici de faire l'hypothèse que $\operatorname{tr}(A_m^\top A_m) \leq (2-c)\operatorname{tr}(A_m)$ pour une constante $c > 0$. En effet, la mesure de complexité naturelle pour les estimateurs linéaires est $\mathcal{C}_m = \operatorname{tr}(A_m)$, donc pour observer un phénomène de saut de complexité autour de la pénalité minimale (21), il faut que $\mathrm{pen}_{\min}(m)$ soit une fonction strictement croissante de $\operatorname{tr}(A_m)$. Cette hypothèse n'est pas trop restrictive car dans tous les exemples mentionnés en Section 4.1 on a

$$\operatorname{tr}(A_m^\top A_m) \leq (2-c)\operatorname{tr}(A_m)$$

et donc

$$\frac{\sigma^2 \operatorname{tr}(A_m)}{n} \leq \mathrm{pen}_{\min}(m) \leq \frac{2\sigma^2 \operatorname{tr}(A_m)}{n} \ .$$

4.3. **Algorithme de calibration de pénalités.** On obtient ainsi l'algorithme de calibration de pénalités suivant, proposé initialement dans [5] :

(1) pour tout $C > 0$, calculer

$$\widehat{m}_{\min}(C) \in \arg\min_{m\in\mathcal{M}_n}\left\{\frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2 + \frac{C\left(2\operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m)\right)}{n}\right\} \ ,$$

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

(3) choisir

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}_n} \left\{ \frac{1}{n}\left\|Y-\widehat{F}_m\right\|^2 + \frac{2\widehat{C}_{\min}\operatorname{tr}(A_m)}{n} \right\} \ .$$

4.4. **Comparaison avec le cas des moindres carrés.** Il est intéressant de comparer cet algorithme à celui proposé par Birgé et Massart dans le cas de l'heuristique de pente. Au lieu d'un facteur 2 entre pénalité minimale et optimale, nous avons ici des pénalités minimales et optimales de formes différentes :

$$\operatorname{pen}_{\min}(m) = \frac{\sigma^2\left(2\operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m)\right)}{n}$$
$$\operatorname{pen}_{\mathrm{opt}}(m) = \frac{\sigma^2\left(2\operatorname{tr}(A_m)\right)}{n}$$

et dont le rapport

$$\frac{\operatorname{pen}_{\mathrm{opt}}(m)}{\operatorname{pen}_{\min}(m)} = \frac{2\operatorname{tr}(A_m)}{2\operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m)} \in \ ]1,2]$$

sous l'hypothèse $\operatorname{tr}(A_m^\top A_m) \leq \operatorname{tr}(A_m)$ (qui est vérifiée dans tous les exemples mentionnés précédemment, voir la nouvelle version de [6]).

Dans le cas particulier des estimateurs des moindres carrés, on a $A_m^\top A_m = A_m$ et donc $2\operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m)$, d'où l'heuristique de pente

$$\frac{\operatorname{pen}_{\mathrm{opt}}(m)}{\operatorname{pen}_{\min}(m)} = 2 \ .$$

Le fait que A_m soit une matrice de projection orthogonale est donc pratiquement le seul cas où le rapport entre pénalité minimale et optimale est constant (indépendant de m) et vaut 2.

Plus précisément, dans le cas des estimateurs linéaires en régression homoscédastique, l'heuristique de pente est valide si et seulement si $\operatorname{tr}(A_m^\top A_m) \approx \operatorname{tr}(A_m)$ pour tout $m \in \mathcal{M}_n$. Un fait remarquable est que cette égalité est vérifiée dans le cas des k plus proches voisins.

En effet, soit A une matrice associée à un estimateur des k plus proches voisins. On a donc

$$\forall i,j \in \{1,\dots,n\} \ , \quad A_{i,j} \in \left\{0, \frac{1}{k}\right\}$$
$$\text{et} \quad \forall i \in \{1,\dots,n\} \ , \quad A_{i,i} = \frac{1}{k} \quad \text{et} \quad \sum_{j=1}^{n} A_{i,j} = 1 \ .$$

En particulier,

$$\operatorname{tr}(A) = \frac{n}{k}$$

et

$$\operatorname{tr}(A^\top A) = \sum_{1\leq i,j\leq n} A_{i,j}^2 = \sum_{1\leq i,j\leq n} \frac{A_{i,j}}{k} = \sum_{1\leq i\leq n} \frac{1}{k} = \frac{n}{k} \ ,$$

si bien que

$$\mathrm{tr}(A) = \mathrm{tr}(A^\top A) \ .$$

4.5. **Garanties théoriques.** De la même manière que pour les estimateurs des moindres carrés, on peut justifier rigoureusement l'algorithme de calibration proposé en Section 4.3 pour les estimateurs linéaires. On obtient les résultats suivants, prouvés en détail dans [6]. Tout d'abord, l'algorithme de la Section 4.3 estime effectivement σ^2.

Théorème 2. *On se place dans le cadre de la régression sur un plan d'expérience déterministe avec la perte des moindres carrés, et l'on fait les hypothèses suivantes, où $(A_m)_{m\in\mathcal{M}_n}$ est une famille de matrices $n\times n$ et $(\widehat{F}_m)_{m\in\mathcal{M}_n}$ la famille d'estimateurs linéaires associés :*

- *bruit Gaussien homoscédastique : on observe $Y = F + \varepsilon \in \mathbb{R}^n$ avec $\varepsilon \sim \mathcal{N}(0,\sigma^2)$,*
- *complexité polynomiale :* $\mathrm{Card}(\mathcal{M}_n) \leq C_{\mathcal{M}} n^\alpha$,
- *$\exists m_1, m_2 \in \mathcal{M}_n$ t.q.* $\mathrm{tr}(A_{m_1}) \geq n/2$, $\mathrm{tr}(A_{m_2}) \leq \sqrt{n}$ *et*

$$\forall i \in \{1,2\} \ , \quad \frac{1}{n}\left\|(I - A_{m_i})F\right\|^2 \leq \sigma^2 \sqrt{\frac{\ln(n)}{n}} \ ,$$

- *pour tout $m \in \mathcal{M}_n$,*

$$\mathrm{tr}(A_m^\top A_m) \leq \mathrm{tr}(A_m) \quad \textit{et} \quad |\!|\!| A_m |\!|\!| \leq 1 \ .$$

Alors, des constantes $n_0, L_\alpha > 0$ (seule L_α pouvant éventuellement dépendre de α) existent telles que si $n \geq n_0$, avec probabilité au moins $1 - 8C_{\mathcal{M}} n^{-2}$, si pour tout $C > 0$,

$$\widehat{m}_{\min}(C) \in \arg\min_{m\in\mathcal{M}_n} \left\{ \frac{1}{n}\left\|Y - \widehat{F}_m\right\|^2 + \frac{C\left(2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m)\right)}{n} \right\} \ ,$$

$$\forall C < \left(1 - L_\alpha \sqrt{\frac{\ln(n)}{n}}\right)\sigma^2 , \quad \mathrm{tr}\left(A_{\widehat{m}_{\min}(C)}\right) \geq \frac{n}{3}$$

$$\forall C > \left(1 + L_\alpha \frac{\sqrt{\ln(n)}}{n^{1/4}}\right)\sigma^2 , \quad \mathrm{tr}\left(A_{\widehat{m}_{\min}(C)}\right) \leq n^{3/4} \ .$$

Par ailleurs, dès lors que l'on a une estimation suffisamment précise de σ^2, on a une inégalité-oracle avec la pénalité C_L.

Théorème 3. *On fait les hypothèses du Théorème 2, et l'on suppose l'existence d'une constante $\kappa \geq 1$ telle que*

$$\forall m \in \mathcal{M}_n \, , \quad \frac{\mathrm{tr}(A_m)\sigma^2}{n} \leq \frac{\kappa}{n}\mathbb{E}\left[\left\|F - \widehat{F}_m\right\|^2\right] \ . \tag{22}$$

Alors, une constante n_1 dépendant uniquement de κ existe telle que si $n \geq n_1$, avec probabilité au moins $1 - 8C_{\mathcal{M}} n^{-2}$, pour tout

$$C \in \left[\sigma^2(1 - (\ln(n))^{-1}), \sigma^2(1 + (\ln(n))^{-1})\right]$$

et tout

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}_n} \left\{ \frac{1}{n}\left\|Y-\widehat{F}_m\right\|^2 + \frac{2C\,\mathrm{tr}(A_m)}{n} \right\} \ ,$$

on a l'inégalité-oracle

$$\frac{1}{n}\left\|F-\widehat{F}_{\widehat{m}}\right\|^2 \leq \left(1+\frac{40\kappa}{\ln(n)}\right) \inf_{m\in\mathcal{M}_n} \left\{ \frac{1}{n}\left\|F-\widehat{F}_m\right\|^2 \right\} + \frac{36(\kappa+\alpha+2)\ln(n)\sigma^2}{n} \ .$$

En particulier, la constante $\widehat{C}_{\min}$ fournie par l'algorithme de la Section 4.3 vérifie sur le même événement de grande probabilité l'hypothèse du Théorème 3, d'après le Théorème 2.

La différence entre les preuves de ces deux résultats et le cas des estimateurs des moindres carrés (Théorème 1 et Théorème 1 du premier cours) est principalement d'ordre technique : on a non pas deux mais quatre quantités aléatoires à concentrer (car $A_m^\top A_m \neq A_m$), et il faut faire attention dans la comparaison des termes de reste et espérances, où apparaissent aussi bien $\mathrm{tr}(A_m)$ et $\mathrm{tr}(A_m^\top A_m)$.

Notons que l'hypothèse (22) peut être supprimée. Ces deux théorèmes sont également valides pour un ensemble $\mathcal{M}_n$ infini dans le cas du choix du paramètre de régularisation $\lambda \in \,]0,+\infty[$ en régression ridge (à noyau fixe). Ces deux extensions du résultat de [5] seront présentes dans la prochaine version de [6].

5. Pénalités minimales et calibration en général

Les résultats obtenus dans le cas des estimateurs linéaires nous conduisent à reformuler l'heuristique de pente comme un cas particulier d'une heuristique plus large, visant à calibrer automatiquement des pénalités pour le problème de sélection d'estimateurs.

Tout d'abord, récapitulons les différents types de résultats qui peuvent être prouvés en lien avec ce problème. On note pen_0 une forme de pénalité minimale (telle que $\mathrm{pen}_{\min} = C^\star_{\min}\,\mathrm{pen}_0$), pen_1 une forme de pénalité optimale (telle que $\mathrm{pen}_{\mathrm{opt}} = C^\star\,\mathrm{pen}_1$), et pour tout $C>0$,

$$\widehat{m}_{\min}(C) \in \arg\min_{m\in\mathcal{M}_n} \left\{ P_n\gamma\left(\widehat{s}_m\right) + C\,\mathrm{pen}_0(m) \right\} \ .$$

(1) L'excès de risque de $\widehat{m}_{\min}(C)$ explose (en comparaison de celui de l'oracle) lorsque $C < C^\star_{\min}$.

(2) Inégalité oracle pour $\widehat{m}_{\min}(C)$ lorsque $C > C^\star_{\min}$.

(3) La complexité $\mathcal{C}_{\widehat{m}_{\min}(C)}$ «saute» brusquement au voisinage de $C = C^\star_{\min}$.

(4) La forme pen_0 de la pénalité minimale est connue ou estimable à l'aide des données uniquement.

(5) Inégalité oracle avec constante $1+o(1)$ lorsqu'on utilise la pénalité $\mathrm{pen}_{\mathrm{opt}}$.

(6) La forme pen_1 de la pénalité optimale est connue ou estimable à l'aide des données uniquement.

(7) Il y a un lien connu entre $C^\star_{\min}$ et $C^\star$, en particulier indépendant de $s^\star$.

Les deux premiers résultats correspondent à la définition théorique initiale d'une pénalité minimale. Ils ne sont pas nécessaires pour valider un algorithme de calibration de pénalité. Il est en revanche crucial d'être capable de montrer les points (3) et (4), qui seuls permettent d'estimer $C^\star_{\min}$ à l'aide des données.

Les points (5), (6) et surtout (7) sont quant à eux déterminants pour que l'estimation de $C^\star_{\min}$ ait un intérêt du point de vue de la sélection d'estimateurs finale. En particulier, il n'est pas évident de connaître une fonction f telle que $f(C^\star_{\min}) = C^\star$ hors du cas de l'heuristique de pente

$$\mathrm{pen}_0 = \mathrm{pen}_1 \quad \text{et} \quad f(t) = 2t$$

et du cas des estimateurs linéaires

$$\mathrm{pen}_0(m) = 2\,\mathrm{tr}(A_m) - \mathrm{tr}(A_m^\top A_m) \quad \mathrm{pen}_1(m) = 2\,\mathrm{tr}(A_m) \quad \text{et} \quad f(t) = t \ .$$

Au final, si l'on est capable de prouver (ou si l'on conjecture) les points (3) à (7), on obtient l'algorithme suivant.

5.1. **Algorithme de calibration de pénalités.** On suppose connues (estimables) les formes pen_0 et pen_1 des pénalités minimales et optimales, c'est-à-dire telles que

$$\mathrm{pen}_0 = \frac{1}{C^\star_{\min}}\,\mathrm{pen}_{\min} \quad \text{et} \quad \mathrm{pen}_1 = \frac{1}{C_\star}\,\mathrm{pen}_{\mathrm{opt}}$$

pour des constantes $C^\star_{\min}, C^\star > 0$. On suppose connue (ou estimable) une fonction f telle que

$$C^\star = f(C^\star_{\min}) \ ,$$

relation qui doit être vraie pour toute loi P (en particulier indépendamment de la valeur de $s^\star$. On suppose connues (ou estimables) les mesures de complexités $(\mathcal{C}_m)_{m\in\mathcal{M}_n}$ associées à chaque estimateur de la famille. L'algorithme s'écrit alors :

(1) pour tout $C > 0$, calculer

$$\widehat{m}_{\min}(C) \in \arg\min_{m\in\mathcal{M}_n} \left\{ P_n\gamma\left(\widehat{s}_m\right) + C\,\mathrm{pen}_0(m) \right\} \ ,$$

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

(3) choisir

$$\widehat{m} \in \arg\min_{m\in\mathcal{M}_n} \left\{ P_n\gamma\left(\widehat{s}_m\right) + f(\widehat{C}_{\min})\,\mathrm{pen}_1(m) \right\} \ ,$$

Notons que l'algorithme proposé par [8] pour le calcul exact de la trajectoire $(\widehat{m}(C))_{C>0}$ s'applique encore à ce cadre. Le premier point n'est donc pas trop coûteux dès lors que l'on a calculé $\left(P_n\gamma\left(\widehat{s}_m\right)\right)_{m\in\mathcal{M}_n}$ et $\left(\mathrm{pen}_0(m)\right)_{m\in\mathcal{M}_n}$.

5.2. **Résultats mathématiques existants.** En plus des résultats déjà mentionnés en Section 3.4, quelques résultats mathématiques partiels ont été prouvés autour du concept de pénalité minimale :

- Concentration de p_2 autour de son espérance pour les estimateurs par minimum de contraste borné [18].
- Explosion du risque si $C < C^\star_{\min}$:
  - pour les estimateurs des moindres carrés en régression, avec un bruit Gaussien, et les familles de modèles $\mathcal{M}_n$ de complexité exponentielle [16, Proposition 2], pour lesquelle $C^\star$ n'est pas connue précisément ;
  - résultat similaire pour les familles de complexité intermédiaire entre polynômiale et exponentielle lorsque $s^\star = 0$ [16, Proposition 3] ;
  - pour l'estimateur de Dantzig en estimation de densité, dans un cas particulier [13] ;
  - pour une méthode de seuillage pour l'estimation de l'intensité d'un processus de Poisson [39] ; notons que dans ce cas on dispose d'un intervalle pour $C^\star/C^\star_{\min} \in [1, 12]$.
- Explosion de la dimension si $C < C^\star_{\min}$:
  pour les estimateurs des moindres carrés en régression, avec un bruit Gaussien, et les familles de modèles $\mathcal{M}_n$ de complexité polynômiale ou exponentielle, en supposant $s^\star = 0$ et en prenant une pénalité multiplicative [9].

Le premier de ces résultats est à rapprocher des résultats obtenus pour l'heuristique de pente avec des contrastes généraux [40, 42]. Notons qu'il est suffisamment précis pour être utilisé comme élément de preuve dans le cas des régressogrammes en régression hétéroscédastique [8].

Les autres résultats concernent majoritairement l'explosion du risque lorsqu'on sous-pénalise (et sont donc d'un intérêt surtout théorique, montrant la nécessité d'une condition sur la pénalité pour obtenir une inégalité-oracle). De plus, la constante optimale $C^\star$ n'étant pas connue précisément dans les cadres où se situent ces travaux, plusieurs étapes restent à franchir avant de valider complètement un algorithme de calibration comme celui proposé plus haut.

6. Aspects pratiques

Pour tous les problèmes pratiques reliés aux algorithmes de calibration de pénalités par pénalités minimales, nous renvoyons à l'article [12] qui présente également un package matlab[3] pour la mise en œuvre de ces algorithmes.

[3] Capushe, téléchargeable librement à l'adresse `http://www.math.univ-toulouse.fr/~maugis/CAPUSHE.html`

7. Conclusion

Pour une discussion du problème de la surpénalisation, on pourra consulter [3, Chapitre 11] et [4, Section 6.3.2].

Références

[1] Hirotugu Akaike. Statistical predictor identification. *Ann. Inst. Statist. Math.*, 22 :203–217, 1970.

[2] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

[3] Sylvain Arlot. *Resampling and Model Selection*. PhD thesis, University Paris-Sud 11, December 2007. http ://tel.archives-ouvertes.fr/tel-00198803/.

[4] Sylvain Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3 :557–624 (electronic), 2009.

[5] Sylvain Arlot and Francis Bach. Data-driven calibration of linear estimators with minimal penalties. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 46–54, 2009.

[6] Sylvain Arlot and Francis Bach. Data-driven calibration of linear estimators with minimal penalties, September 2009. arXiv :0909.1884v1.

[7] Sylvain Arlot and Alain Celisse. Segmentation of the mean of heteroscedastic data via cross-validation. *Statistics and Computing*, pages 1–20, 2010. 10.1007/s11222-010-9196-x.

[8] Sylvain Arlot and Pascal Massart. Data-driven calibration of penalties for least-squares regression. *J. Mach. Learn. Res.*, 10 :245–279 (electronic), 2009.

[9] Yannick Baraud, Christophe Giraud, and Sylvie Huet. Gaussian model selection with an unknown variance. *Ann. Statist.*, 37(2) :630–672, 2009.

[10] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local Rademacher complexities. *Ann. Statist.*, 33(4) :1497–1537, 2005.

[11] Jean-Patrick Baudry. *Model selection for clustering. Choosing the number of classes.* PhD thesis, University Paris XI, December 2009. http ://tel.archives-ouvertes.fr/tel-00461550/.

[12] Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope Heuristics : Overview and Implementation. Technical Report 7223, INRIA, 2010. hal-00461639.

[13] Karine Bertin, Erwann Le Pennec, and Vincent Rivoirard. Adaptive density estimation. Technical report, arXiv, 2009. arXiv :0905.0884.

[14] Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3) :203–268, 2001.

[15] Lucien Birgé and Pascal Massart. A generalized cp criterion for gaussian model selection. Technical report, Universités de Paris 6 et Paris 7, 2001. Prépublication 647, 39 pages.

[16] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2) :33–73, 2007.

[17] Gilles Blanchard and Pascal Massart. Discussion : "Local Rademacher complexities and oracle inequalities in risk minimization" [Ann. Statist. **34** (2006), no. 6, 2593–2656] by V. Koltchinskii. *Ann. Statist.*, 34(6) :2664–2671, 2006.

[18] Stéphane Boucheron and Pascal Massart. A high dimensional Wilks phenomenon. *Probab. Theory Related Fields*, March 2010.

[19] Claire Caillerie and Bertrand Michel. Model selection for simplicial approximation. Technical Report 6981, INRIA, 2009. http ://hal.inria.fr/inria-00402091/.

[20] Y. Cao and Y. Golubev. On oracle inequalities related to smoothing splines. *Math. Methods Statist.*, 15(4) :398–414 (2007), 2006.

[21] Pierre Connault. *Calibration d'algorithmes de type Lasso*. PhD thesis, Université Paris-Sud. En cours de rédaction.

[22] Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, 31(4) :377–403, 1978/79.

[23] Bradley Efron. How biased is the apparent error rate of a prediction rule ? *J. Amer. Statist. Assoc.*, 81(394) :461–470, 1986.

[24] Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory*, 47(5) :1902–1914, 2001.

[25] Vladimir Koltchinskii. Local Rademacher complexities and oracle inequalities in risk minimization. *Ann. Statist.*, 34(6) :2593–2656, 2006.

[26] Marc Lavielle. Using penalized contrasts for the change-point problem. *Signal Proces.*, 85(8) :1501–1510, 2005.

[27] Émilie Lebarbier. Detecting multiple change-points in the mean of a Gaussian process by model selection. *Signal Proces.*, 85 :717–736, 2005.

[28] Matthieu Lerasle. *Rééchantillonnage et sélection de modèles optimale pour l'estimation de la densité de variables indépendantes ou mélangeantes*. PhD thesis, INSA de Toulouse, June 2009.

[29] Matthieu Lerasle. Optimal model selection for stationary data under various mixing conditions. arXiv :0911.1497, October 2010.

[30] Ker-Chau Li. From Stein's unbiased risk estimates to the method of generalized cross validation. *Ann. Statist.*, 13(4) :1352–1377, 1985.

[31] Ker-Chau Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation : discrete index set. *Ann. Statist.*, 15(3) :958–975, 1987.

[32] Fernando Lozano. Model selection using Rademacher penalization. In *Proceedings of the 2nd ICSC Symp. on Neural Computation (NC2000). Berlin, Germany*. ICSC Academic Press, 2000.

[33] Colin L. Mallows. Some comments on C_p. *Technometrics*, 15 :661–675, 1973.

[34] Pascal Massart. A non-asymptotic theory for model selection. In *European Congress of Mathematics*, pages 309–323. Eur. Math. Soc., Zürich, 2005.

[35] Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[36] Matthieu Lerasle. Optimal model selection in density estimation. arXiv :0910.1654v2, 2009.

[37] Cathy Maugis and Bertrand Michel. A non asymptotic penalized criterion for Gaussian mixture model selection. Technical Report 6549, INRIA, 2008.

[38] Cathy Maugis and Bertrand Michel. Slope heuristics for variable selection and clustering via Gaussian mixtures. Technical Report 6550, INRIA, 2008.

[39] Patricia Reynaud-Bouret and Vincent Rivoirard. Calibration of thresholding rules for poisson intensity estimation. Technical report, arXiv, 2009. arXiv :0904.1148.

[40] Adrien Saumard. *Estimation par Minimum de Contraste Régulier et Heuristique de Pente en Sélection de Modèles*. PhD thesis, Université de Rennes 1, October 2010.

[41] Adrien Saumard. Nonasymptotic quasi-optimality of AIC and the slope heuristics in maximum likelihood estimation of density using histogram models. hal-00512310, September 2010.

[42] Adrien Saumard. Optimal upper and lower bounds for the true and empirical excess risks in heteroscedastic least-squares regression. hal-00512304, September 2010.

[43] Adrien Saumard. The slope heuristics in heteroscedastic regression. hal-00512306, September 2010.

[44] Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

[45] Gideon Schwarz. Estimating the dimension of a model. *Ann. Statist.*, 6(2) :461–464, 1978.

[46] Grace Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

Cnrs – Équipe Sierra, Laboratoire d'Informatique de l'École Normale Supérieure, (CNRS/ENS/INRIA UMR 8548), INRIA - 23 avenue d'Italie - CS 81321, 75214 PARIS Cedex 13 - France

E-mail address: sylvain.arlotRETIRERCECI@ens.fr

URL: http://www.di.ens.fr/~arlot/