SÉLECTION D'ESTIMATEURS AVEC DES PÉNALITÉS ALÉATOIRES

SYLVAIN ARLOT
CNRS

Les algorithmes (ou estimateurs) utilisés en apprentissage statistique dépendent généralement d'un ou plusieurs paramètres dont le choix est particulièrement important. Plus généralement, on a souvent à disposition toute une famille d'estimateurs, et se pose alors le problème d'en choisir un dont le risque est aussi petit que possible.

Une approche classique est la pénalisation : on choisit l'estimateur (le paramètre) qui minimise la somme du risque empirique et d'une «pénalité», l'objet de ce dernier terme étant d'éviter le sur-apprentissage.

Cet exposé aborde la question de construction de pénalités optimales, à l'aide des données uniquement, pour différents problèmes de sélection d'estimateurs en régression non-paramétrique.

Nous considérons dans une première partie (Section 1) le cas où la pénalité optimale est connue à constante multiplicative près. C'est notamment le cas pour la sélection d'estimateurs linéaires, où la pénalité optimale (C_L de Mallows) dépend de la variance du bruit, inconnue en général. En utilisant le concept de pénalité minimale, nous verrons comment obtenir une calibration optimale de la pénalité, à l'aide des données uniquement [AB09a].

Nous aborderons dans un deuxième temps (Section 2) le cas (courant) où la forme de la pénalité optimale est inconnue, comme en régression avec un bruit hétéroscédastique. Il est alors naturel d'utiliser une procédure de rééchantillonnage pour estimer cette pénalité, que l'on peut rendre algorithmiquement efficace en considérant un sous-échantillonnage de type «V-fold». Dans le cadre où les estimateurs candidats sont des régressogrammes, on dispose d'inégalités-oracle non-asymptotiques montrant que les pénalités aléatoires ainsi construites sont optimales et s'adaptent à l'hétéroscédasticité des données [Arl09, Arl08, AL11].

1. Pénalités minimales

On considère le cadre de la régression sur un plan d'expérience fixe : on observe

$$Y = (Y_1, \dots, Y_n) = F + \varepsilon \in \mathbb{R}^n \ ,$$

avec $\varepsilon_1, \dots, \varepsilon_n$ indépendantes et telles que pour tout $i \in \{1, \dots, n\}$, $\mathbb{E}[\varepsilon_i] = 0$ et $\mathbb{E}[\varepsilon_i^2] = \sigma^2$. L'objectif est de trouver, à l'aide des données, un $t \in \mathbb{R}^n$ dont la perte quadratique

$$\frac{1}{n} \|t - F\|_2^2 = \frac{1}{n} \sum_{i=1}^{n} (t_i - F_i)^2$$

Date: 23 mars 2011.

est minimale.

Bon nombre d'estimateurs classiques sont des *estimateurs linéaires*, c'est-à-dire de la forme

$$\widehat{F} = AY$$

avec $A \in \mathcal{M}_n(\mathbb{R})$ déterministe, par exemple :

- les moindres carrés (A est la matrice de projection orthogonale sur le modèle correspondant),
- la régression ridge (à noyau), en particulier les splines de lissage,
- les k-plus proches voisins,
- les estimateurs de Nadaraya-Watson.

On suppose donnée une famille de matrice $(A_m)_{m\in\mathcal{M}}$ et l'on considère le problème de sélection d'un estimateur parmi $(\widehat{F}_m)_{m\in\mathcal{M}}$ avec $\widehat{F}_m = A_m Y$. Ceci recouvre en particulier le problème de sélection de modèles ou du choix de paramètres dans les exemples ci-dessus (paramètre de régularisation, noyau, nombre k de voisins, largeur de bande, etc.).

L'idée de la pénalisation est de sélectionner

$$\widehat{m} \in \operatorname{argmin}_{m\in\mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + \operatorname{pen}(m) \right\} \quad , \tag{1}$$

où le terme $\operatorname{pen}(m)$ compense l'optimisme du risque empirique $n^{-1}\|\widehat{F}_m - Y\|^2$ comme estimateur du risque $n^{-1}\mathbb{E}[\|\widehat{F}_m - F\|^2]$. Une pénalité optimale (satisfaisant une inégalité-oracle si $\mathcal{M}$ n'est pas trop grand) est alors

$$\operatorname{pen}_{C_L}(m) := \frac{2\sigma^2 \operatorname{tr}(A_m)}{n}$$

[Mal73, C_L], qui dépend de la variance σ^2 du bruit, a priori inconnue.

Birgé et Massart [BM07] ont proposé un algorithme d'estimation de la constante σ^2 apparaissant devant la pénalité, qui ne suppose pas la connaissance préalable d'un modèle contenant la cible F. Cet algorithme repose sur l'«heuristique de pente», qui stipule que la pénalité optimale est égale au double de la pénalité minimale, et qui n'est justifiée que lorsque toutes les matrices A_m sont des projections orthogonales (c'est-à-dire, pour la sélection de modèles).

Nous en proposons une généralisation pour les estimateurs linéaires, en remplaçant l'heuristique de pente par l'utilisation du concept plus général de pénalité minimale :

Entrée : $(A_m)_{m\in\mathcal{M}}$ famille finie de matrices, $Y \in \mathbb{R}^n$

- $\forall C > 0$, calculer

$$\widehat{m}_0(C) \in \operatorname{argmin}_{m\in\mathcal{M}} \left\{ \frac{1}{n} \left\| \widehat{F}_m - Y \right\|^2 + C \frac{2\operatorname{tr}(A_m) - \operatorname{tr}(A_m^\top A_m)}{n} \right\} .$$

- Trouver $\widehat{C}$ autour duquel $C \mapsto \operatorname{tr}(A_{\widehat{m}_0(C)})$ «saute», par exemple tel que $\operatorname{tr}(A_{\widehat{m}_0(C)}) > 9n/10$ si $C < \widehat{C}(1-\delta_n)$ et $\operatorname{tr}(A_{\widehat{m}_0(C)}) < n/10$ si $C > \widehat{C}(1+\delta_n)$, pour un «petit» $\delta_n > 0$.

Sortie : $\widehat{m} \in \operatorname{argmin}_{m\in\mathcal{M}} \{ n^{-1}\|\widehat{F}_m - Y\|^2 + 2\widehat{C} n^{-1} \operatorname{tr}(A_m) \}$.

L'intuition justifiant cet algorithme est que

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

est une *pénalité minimale* : il y a sur-apprentissage si $\mathrm{pen} < \mathrm{pen}_{\min}$, tandis que le nombre de degrés de liberté $\mathrm{tr}(A_{\widehat{m}})$ de l'estimateur sélectionné est beaucoup plus faible si $\mathrm{pen} > \mathrm{pen}_{\min}$.

On a alors le résultat suivant (version simplifiée de l'énoncé prouvé dans [AB09a, AB09b]) :

Théorème 1. *Si* $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, $\mathrm{Card}(\mathcal{M}) \leq c_{\mathcal{M}} n^{\alpha_{\mathcal{M}}}$,

$$\forall m \in \mathcal{M} \;\; , \quad \left\|\!\left| A_m \right|\!\right| \leq 1 \quad et \quad \mathrm{tr}\left(A_m^\top A_m\right) \leq \mathrm{tr}(A_m)$$

$$\exists m_1, m_2 \in \mathcal{M} \;\; , \quad A_{m_1} = I_n \quad \mathrm{tr}(A_{m_2}) \leq \sqrt{n} \quad et \quad \|(I_n - A_{m_2})F\|^2 \leq \sigma^2 \sqrt{n \ln(n)}$$

alors, il existe des constantes absolues $L_1, L_2, L_3 > 0$ *telles qu'avec probabilité au moins* $1 - L_1 c_{\mathcal{M}} n^{-2}$, *on a*

$$\forall C < \sigma^2 \left(1 - L_2(\alpha_{\mathcal{M}} + 2)\sqrt{\frac{\ln(n)}{n}}\right) \;\; , \quad \mathrm{tr}(A_{\widehat{m}_0(C)}) \geq \frac{9n}{10}$$

$$\forall C > \sigma^2 \left(1 + L_2(\alpha_{\mathcal{M}} + 2)\sqrt{\frac{\ln(n)}{n}}\right) \;\; , \quad \mathrm{tr}(A_{\widehat{m}_0(C)}) \leq \frac{n}{10}$$

$$\frac{1}{n}\left\|\widehat{F}_{\widehat{m}} - F\right\|^2 \leq \left(1 + \frac{1}{\ln(n)}\right) \inf_{m \in \mathcal{M}} \left\{ \frac{1}{n}\left\|\widehat{F}_m - F\right\|^2 \right\} + \frac{L_3(\alpha_{\mathcal{M}} + 2)^2 \sigma^2 \left(\ln(n)\right)^2}{n} \;\; .$$

Il est intéressant de noter que dans les cas des moindres carrés et des k-plus proches voisins, on retrouve l'heuristique de pente $\mathrm{pen}_{\mathrm{opt}}(m) \approx 2\,\mathrm{pen}_{\min}(m)$ car $\mathrm{tr}\left(A_m^\top A_m\right) = \mathrm{tr}(A_m)$; en revanche, dans le cas général, on a seulement

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

Des expériences numériques montrent que pour les différents exemples d'estimateurs linéaires mentionnés, l'algorithme proposé conduit à un risque final plus faible que la validation croisée «10-fold» et la validation croisée généralisée [CW79], la différence étant souvent importante [AB09a, AB09b].

2. Pénalisation par rééchantillonnage et pénalisation «V-fold»

Il est très courant en pratique d'avoir des données hétéroscédastiques, pour lesquelles la variance σ^2 du bruit varie avec la position d'observation. La forme de la pénalité idéale est alors inconnue, et l'on peut montrer théoriquement qu'il est nécessaire de l'estimer pour avoir une inégalité-oracle optimale [Arl10]. La même difficulté surgit pour le problème général de sélection d'estimateurs, dès qu'on considère un risque non-quadratique ou des estimateurs non-linéaires. L'objet de cette section est de montrer que l'on peut résoudre ce problème par rééchantillonnage, ou par sous-échantillonnage «V-fold».

On se place dans le cadre de la régression sur un plan d'expérience aléatoire[1]. On observe un échantillon $D_n = (X_i, Y_i)_{1 \leq i \leq n} \in (\mathcal{X} \times \mathbb{R})^n$ de variables aléatoires indépendantes et de même loi P, et l'on note $P_n := n^{-1} \sum_{i=1}^n \delta_{(X_i,Y_i)}$ sa mesure empirique. On note $\mathbb{S}$ l'ensemble des application mesurables $\mathcal{X} \mapsto \mathbb{R}$ et l'on cherche $t \in \mathbb{S}$ dont la perte

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

est minimale, où $\gamma : \mathbb{S} \times (\mathcal{X} \times \mathbb{R}) \mapsto [0, +\infty[$ est une fonction de contraste, par exemple le contraste des moindres carrés $\gamma(t;(x,y)) := (t(x)-y)^2$. Pour tout $t \in \mathbb{S}$, on définit également la perte relative

$$\ell \left(s^\star, t \right) = P\gamma(t) - \inf_{u \in \mathbb{S}} P\gamma(u) \geq 0 \ .$$

On suppose donnée une famille d'estimateurs $\left(\widehat{s}_m \right)_{m \in \mathcal{M}}$, c'est-à-dire que pour tout $m \in \mathcal{M}$, $\widehat{s}_m = \widehat{s}_m(P_n) \in \mathbb{S}$, et l'on cherche à sélectionner $\widehat{m}(P_n)$ tel que $\ell \left(s^\star, \widehat{s}_{\widehat{m}(P_n)}(P_n) \right)$ est minimale. Une procédure de choix d'estimateur par pénalisation (1) s'écrit alors

$$\widehat{m} \in \operatorname{argmin}_{m \in \mathcal{M}} \left\{ P_n \gamma \left(\widehat{s}_m(P_n) \right) + \operatorname{pen}(m) \right\} \ , \tag{2}$$

et la pénalité idéale est celle qui conduit (2) à minimiser la perte, c'est-à-dire,

$$\operatorname{pen}_{\mathrm{id}}(m) := (P - P_n) \gamma \left(\widehat{s}_m(P_n) \right) \ . \tag{3}$$

2.1. **Pénalisation par rééchantillonnage.** Comme en général on ne connaît pas la forme de $\operatorname{pen}_{\mathrm{id}}(m)$ ou de son espérance (sauf cas particulier comme celui de la Section 1), nous proposons d'utiliser comme pénalité un estimateur par rééchantillonnage [Efr79] de l'espérance de (3) pour chaque $m \in \mathcal{M}$. En notant

$$P_n^W := \frac{1}{n} \sum_{i=1}^n W_i \delta_{(X_i,Y_i)}$$

la mesure empirique d'un rééchantillon[2], on obtient la pénalité

$$\operatorname{pen}_{\mathrm{reech}}(m) := C_W \mathbb{E}_W \left[(P_n - P_n^W) \gamma \left(\widehat{s}_m(P_n^W) \right) \right] \ , \tag{4}$$

où $\mathbb{E}_W [\cdot]$ désigne l'espérance relativement à l'alea de W (c'est-à-dire, conditionnellement à D_n), et $C_W > 0$ est à choisir convenablement en fonction de la loi de W. Dans le cas du bootstrap, $C_W = 1$ et la pénalité (4) a été proposée initialement par Efron [Efr83] ; voir [Arl09] pour le cas général.

Lorsque pour tout $m \in \mathcal{M}$, $\widehat{s}_m$ est un régressogramme[3], on peut prouver que sous des hypothèses «raisonnables» sur P autorisant l'hétéroscédasticité [Arl09], la procédure définie par (2) et (4) satisfait une inégalité-oracle de la forme

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

[1]Les mêmes procédures peuvent être définies dans un cadre général qui comprend par exemple la classification supervisée et l'estimation de densité [Arl09, Ler11, Arl08]

[2]Cette notation se généralise naturellement au cas d'un «rééchantillon à poids échangeables», où l'on suppose que $W \in [0, +\infty[^{\ n}$ est indépendant de D_n, et de loi invariante par permutation de ses coordonnées.

[3]c'est-à-dire, une fonction constante par morceaux obtenue par minimisation du risque empirique

avec probabilité au moins $1-K_1 n^{-2}$ où $K_1 > 0$ est une constante. Des simulations numériques permettent de confirmer les bonnes performances pratiques de cette procédure (en termes de risque) en régression hétéroscédastique.

2.2. **Pénalisation «V-fold».** Le défaut principal de la pénalité (4) est son coût de calcul très élevé. D'un point de vue pratique, on est donc conduit à approcher l'espérance (4) par une moyenne empirique sur un nombre $B \geq 1$ de réalisations indépendantes de W.

Alternativement, nous proposons de procéder à un sous-échantillonnage inspiré de la validation croisée «V-fold», qui a un coût algorithmique similaire, de l'ordre de V fois le coût de calcul de $\widehat{s}_m(P_n)$. Formellement, on considère une partition $\mathcal{B} = (B_j)_{1 \leq j \leq V}$ de $\{1, \dots, n\}$ (aussi régulière que possible), et pour tout $j \in \{1, \dots, V\}$ on note

$$P_n^{(j)} := \frac{1}{\mathrm{Card}(B_j)} \sum_{i \in B_j} \delta_{(X_i, Y_i)} \quad \text{et} \quad P_n^{(-j)} := \frac{1}{n - \mathrm{Card}(B_j)} \sum_{i \notin B_j} \delta_{(X_i, Y_i)}$$

les mesures empiriques des sous-échantillons respectifs $D_n^{(j)} = (X_i, Y_i)_{i \in B_j}$ et $D_n^{(-j)} = (X_i, Y_i)_{i \notin B_j}$. L'idée est d'appliquer l'heuristique de rééchantillonnage comme pour (4), avec $P_n^W = P_n^{(-J)}$ où J est une variable uniforme sur $\{1, \dots, V\}$ indépendante de D_n. On obtient alors la pénalisation «V-fold», définie par (2) avec la pénalité

$$\mathrm{pen}_{\mathrm{VF}}(m) := C_V \times \frac{1}{V} \sum_{j=1}^{V} \left(P_n - P_n^{(-j)} \right) \gamma \left(\widehat{s}_m(P_n^{(-j)}) \right) \quad . \tag{6}$$

Lorsque l'espérance de $\mathrm{pen}_{\mathrm{id}}(m)$ varie comme $1/n$ avec la taille de l'échantillon, on montre que $C_V = V - 1$ conduit à une estimation sans biais de $\mathbb{E}[\mathrm{pen}_{\mathrm{id}}(m)]$ [Arl08].

Comme pour les pénalités par rééchantillonnage, on peut montrer (sous des hypothèses similaires sur P) des inégalités-oracle de la forme (5) pour la pénalisation «V-fold» (6), pour les régressogrammes [Arl08] et en estimation de densité [AL11].

Notons que l'on peut également voir la pénalité «V-fold» comme une manière de corriger le biais de la validation croisée «V-fold» (voir [AC10] au sujet de la validation croisée). Cette correction du biais est fondamentale car elle permet d'avoir une inégalité-oracle optimale, comme (5), tandis que la validation croisée «V-fold» est sous-optimale d'un facteur multiplicatif $\kappa(V) > 1$ lorsque $V = \mathcal{O}(1)$ [Arl08].

Des expériences numériques mettent en évidence les bonnes performances des pénalités «V-fold» en termes de risque, même lorsque $V = 5$ ou 10. Une autre qualité importante est que le choix de V est beaucoup plus simple avec ces pénalités «V-fold» qu'avec la validation croisée «V-fold», pour laquelle le choix optimal de V semble très variable, et souvent contre-intuitif [Arl08].

Références

[AB09a] 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.

[AB09b] Sylvain Arlot and Francis Bach. Data-driven calibration of linear estimators with minimal penalties, September 2009. arXiv :0909.1884.

[AC10] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statist. Surv.*, 4 :40–79, 2010.

[AL11] Sylvain Arlot and Matthieu Lerasle. V-fold penalization for least-squares density estimation, 2011. Work in progress, to appear soon.

[Arl08] Sylvain Arlot. V-fold cross-validation improved : V-fold penalization, February 2008. arXiv :0802.0566.

[Arl09] Sylvain Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3 :557–624 (electronic), 2009.

[Arl10] Sylvain Arlot. Choosing a penalty for model selection in heteroscedastic regression, June 2010. arXiv :0812.3141.

[BM07] Lucien Birgé and Pascal Massart. Minimal penalties for Gaussian model selection. *Probab. Theory Related Fields*, 138(1-2) :33–73, 2007.

[CW79] 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.

[Efr79] Bradley Efron. Bootstrap methods : another look at the jackknife. *Ann. Statist.*, 7(1) :1–26, 1979.

[Efr83] Bradley Efron. Estimating the error rate of a prediction rule : improvement on cross-validation. *J. Amer. Statist. Assoc.*, 78(382) :316–331, 1983.

[Ler11] Matthieu Lerasle. Optimal model selection in density estimation. *Ann. Inst. H. Poincaré Probab. Statist.*, 2011. Accepted. arXiv :0910.1654.

[Mal73] Colin L. Mallows. Some comments on C_p. *Technometrics*, 15 :661–675, 1973.

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

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

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