Une sur-pénalisation théoriquement fondée du critère AIC

Adrien Saumard [1] & Fabien Navarro [2]

[1] *CREST, ENSAI, Campus de Ker-Lann, Rue Blaise Pascal, BP 37203, 35172 Bruz, adrien.saumard@ensai.fr*

[2] *CREST, ENSAI, Campus de Ker-Lann, Rue Blaise Pascal, BP 37203, 35172 Bruz, fabien.navarro@ensai.fr*

Résumé. Le fait qu'une légère sur-pénalisation engendre une stabilisation des procédures de sélection de modèles est un phénomène bien connu des spécialistes. En effet, il a été remarqué depuis la fin des années 70 que l'ajout d'une petite quantité positive à des critères pénalisés classiques tels que AIC améliore dans les bons cas les résultats en prédiction, particulièrement pour les échantillons de taille petite ou modérée. La raison principale est que la sur-pénalisation tend à se prémunir contre le sur-apprentissage. Nous proposons la première stratégie générale et théoriquement fondée de sur-pénalisation et nous l'appliquons au critère AIC. De très bons résultats sont observés par simulation.

Mots-clés. Sélection de modèles, sur-pénalisation, critère d'information d'Akaike, estimation de densité, histogramme.

Abstract. Stabilization by over-penalization is a well-known phenomenon for specialists of model selection procedures. Indeed, it has been remarked for a long time that adding a small amount to classical penalized criteria such as AIC lead in good cases to an improvement of prediction performances, especially for moderate and small sample sizes. In particular, overfitting tends to be avoided. We propose here the first principled and general over-penalization strategy and apply it to AIC. Very good results are observed in simulations.

Keywords. Model selection, over-penalisation, Akaike's information criterion, density estimation, histogram.

1 Un phénomène connu mais mal compris

Un effet non-asymptotique subtil mais général en sélection de modèles est qu'une légère sur-pénalisation de l'espérance de la pénalité idéale entraîne un gain en performance sur un jeu fini de données.

En fait, l'idée de base qui correspond à corriger une pénalité dans le but d'améliorer ses performances sur des échantillons de petite taille ou de taille modérée est une thématique bien identifiée et déjà relativement classique dans le cas de l'estimation par maximum de vraisemblance. En effet, plusieurs corrections ont été proposées pour le critère AIC, dans des contextes statistiques divers, avec données indépendantes ou dans l'étude de séries temporelles (6, 8, 9, 10). Les quelques tentatives de validation théorique de ces modifications non-asymptotiques d'AIC restent néanmoins peu satisfaisantes, car trop restrictives ou trop peu informatives.

Birgé et Rozenholc in (6) ont proposé une approche théoriquement fondée de la sur-pénalisation d'AIC pour la sélection d'histogrammes en estimation de la densité, en se basant sur des résultats mathématiques précedemment obtenus par Castellan (7). Du point de vue de Birgé et Rozenholc (6), la sur-pénalisation d'AIC permet de prendre en compte la taille de la collection de modèles considérés, même lorsque celle-ci n'est que polynômiale. Le succès de cette théorie de la complexité de la collection de modèles, issue des travaux de Barron, Birgé et Massart (5), vient en grande partie du fait qu'elle permet d'expliquer dans un cadre général l'ajout de termes logarithmiques dans les pénalités classiques, lorsque la collection de modèles est exponentielle en la taille de l'échantillon. Cependant, cette théorie et en particulier les résultats de Castellan (7), ne permettent pas de prédire la forme de la correction souhaitée pour AIC. Ainsi, la forme précise de la modification d'AIC proposée par Birgé et Rozenholc (6) repose en fait sur une exploration massive par simulations, ce qui peut être vu comme une insuffisance de la théorie.

Appliqué au cas de la validation croisée ou des méthodes classiques de rééchantillonnage—qui peuvent se voir artificiellement comme des méthodes de pénalisation, en ajoutant et retranchant à ces critères le risque empirique des estimateurs—cet effet de sur-pénalisation est à l'origine du fait que le biais de ces méthodes classiques d'estimation du risque tend à accroître leur performance pour des échantillons petits à modérés, même si cela entraîne leur sous-optimalité asymptotique (1, 2, 3). Concernant la pénalisation par rééchantillonnage, une stratégie de sur-pénalisation adéquate en pratique correspond à choisir une constante de calibration légèrement supérieure à la valeur asymptotique optimale (1, 2, 4). Cependant, aucune heuristique ne permet actuellement de choisir de manière automatique et générale le bon niveau de sur-pénalisation.

2 Une stratégie générale de sur-pénalisation

Nous proposons une correction naturelle du critère AIC pour la sélection d'histogrammes en estimation de la densité par maximum de vraisemblance. Dans cette approche, la forme de la sur-pénalisation se déduit directement de l'ordre de grandeur des déviations de l'excès de risque, c'est-à-dire la divergence de Kullback-Leibler, autour de sa moyenne. De telles estimées sont déduites de l'établissement d'inégalités de concentration pour l'excès de risque.

Plus précisement, supposons que l'on dispose d'un échantillon $(Z_1, ..., Z_n)$ i.i.d. de loi $P^{\otimes n}$ sur $\mathcal{Z}^n$, où $\mathcal{Z}$ est un sous-ensemble de $\mathbb{R}^d$, $d \geq 1$. On suppose aussi que P admet une densité f_* par rapport à la mesure de Lebesgue restreinte à $\mathcal{Z}$.

On cherche donc à estimer f_* par un histogramme (sous-)régulier, en sélectionnant parmi plusieurs candidats.

Lorsqu'une partition m de $\mathcal{Z}$ est fixée, le projeté (au sens des moindres carrés mais aussi au sens de la divergence de Kullback) f_m de f_* sur m vaut

$$f_m = \sum_{I \in m} \frac{P(I)}{\mathrm{Leb}(I)} \mathbb{1}_I \ .$$

L'estimateur $\hat{f}_m$ du maximum de vraisemblance sur l'ensemble des histogrammes portés

par la partition m vaut quant à lui,

$$\hat{f}_m = \sum_{I \in m} \frac{P_n(I)}{\mathrm{Leb}(I)} \mathbb{1}_I \ .$$

On s'intéresse à la concentration de l'excès de risque de cet estimateur sur le modèle m, donnée par $\mathcal{K}\left(f_m, \hat{f}_m\right)$ où $\mathcal{K}(f,g) := \int_{\mathcal{Z}} f(x) \ln\left(\frac{f(x)}{g(x)}\right) dx$ est la divergence de Kullback de g par rapport à f, et aussi à la concentration du risque empirique, donnée ici par $\mathcal{K}\left(\hat{f}_m, f_m\right)$. On démontre le théorème suivant.

Théorème 1. *Soient α, A_+, A_- et A_Λ des constantes positives et soit m une partition finie de $\mathcal{Z}$. Le cardinal de m est noté D_m. Supposons que*

$$0 < A_\Lambda \leq D_m \inf_{I \in m} \{P(I)\} \quad \text{et} \quad 1 < D_m \leq A_+ \frac{n}{\ln n} \leq n \ .$$

Si

$$\frac{(\alpha+1) A_+}{A_\Lambda} < \tau = 12 - 3\sqrt{15} < 0.39 \ ,$$

ou si $n \geq n_0(\alpha, A_+, A_\Lambda)$ pour $n_0(\alpha, A_+, A_\Lambda)$ assez grand, alors il existe $A_0 > 0$, dépendant seulement de α, A_+ et A_Λ, telle que en posant

$$\varepsilon_n^+(m) = \max\left\{ \sqrt{\frac{D_m \ln n}{n}}; \sqrt{\frac{\ln n}{D_m}}; \frac{\ln n}{D_m} \right\}$$

et

$$\varepsilon_n^-(m) = \max\left\{ \sqrt{\frac{D_m \ln n}{n}}; \sqrt{\frac{\ln n}{D_m}} \right\} \ ,$$

on obtient, sur un évènement de probabilité au moins égale à $1 - 4n^{-\alpha}$,

$$\left(1 - A_0 \varepsilon_n^-(M)\right) \frac{D_m - 1}{2n} \leq \mathcal{K}\left(f_m, \hat{f}_m\right) \leq \left(1 + A_0 \varepsilon_n^+(M)\right) \frac{D_m - 1}{2n} \ ,$$

$$\left(1 - A_0 \varepsilon_n^-(M)\right) \frac{D_m - 1}{2n} \leq \mathcal{K}\left(\hat{f}_m, f_m\right) \leq \left(1 + A_0 \varepsilon_n^+(M)\right) \frac{D_m - 1}{2n} \ .$$

On considère maintenant une collection de modèles $\mathcal{M}_n$ formée de partitions finie de $\mathcal{Z}$, et de cardinal $\#(\mathcal{M}_n) \leq n^a$ for a positive constant α. Sur la base du précédent théorème précédent, on propose la stratégie de sélection de modèles suivante, qui est en fait une surpénalisation du critère AIC,

$$\hat{m} \in \arg\min_{m \in \mathcal{M}_n} \left\{ \frac{-1}{n} \sum_{i=1}^n \ln\left(\hat{f}_m(Z_i)\right) + \left(1 + \hat{C} \max\left\{ \sqrt{\frac{D_m \ln n}{n}}; \sqrt{\frac{\ln n}{D_m}}; \frac{\ln n}{D_m} \right\}\right) \frac{D_m - 1}{n} \right\} \ ,$$

où $\hat{C}$ est une constante positive dont la valeur est apprise de part les données par une procédure auxiliaire qu'il serait trop fastidieux de détailler ici. Il est à noter que dans la plupart des exemples une constante $\hat{C}$ préfixée égale à 1 ou 2 donne déjà des très bons

résultats, souvent meilleurs que les corrections classiques d'AIC pour des échantillons de taille petite ou modérée.

Des résultats précis de simulations viendront de plus confirmer lors de l'exposé le très bon comportement de la méthode en pratique.

Cette stratégie de sur-pénalisation a enfin un fort potentiel de généralisation, car les quantités permettant de définir le niveau optimal de sur-pénalisation sont définies dans le cadre général de la M-estimation.

Bibliographie

[1] S. Arlot. V-fold cross-validation improved: V-fold penalization, Feb. 2008. URL `http://hal.archives-ouvertes.fr/hal-00239182/en/`. arXiv:0802.0566v2.

[2] S. Arlot. Model selection by resampling penalization. *Electron. J. Stat.*, 3:557–624, 2009.

[3] S. Arlot and A. Célisse. A survey of cross-validation procedures for model selection. *Stat. Surv.*, 4:40–79, 2010.

[4] S. Arlot and M. Lerasle. V-fold cross-validation and V-fold penalization in least-squares density estimation, Oct. 2012. arXiv:1210.5830.

[5] A. Barron, L. Birgé, and P. Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999. ISSN 0178-8051.

[6] L. Birgé and Y. Rozenholc. How many bins should be put in a regular histogram. *ESAIM Probab. Stat.*, 10:24–45 (electronic), 2006.

[7] G. Castellan. Modified Akaike's criterion for histogram density estimation. *Technical report ♯99.61, Université Paris-Sud*, 1999.

[8] G. Claeskens and N. L. Hjort. *Model selection and model averaging*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2008.

[9] C. M. Hurvich and C.-L. Tsai. Model selection for least absolute deviations regression in small samples. *Statist. Probab. Lett.*, 9(3):259–265, 1990.

[10] N. Sugiura. Further analysts of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, 7(1):13–26, 1978.