Agrégation d'hold-out en régression linéaire parcimonieuse et robuste

Guillaume Maillard [1]

[1] *6 Avenue de la Fonte, 4364 Esch-sur-Alzette, Luxembourg, guillaume.maillard@uni.lu*

Résumé. Les méthodes de régression linéaire parcimonieuse ont en généréral un paramètre libre qui contrôle le degré de parcimonie. Le choix de ce paramètre est sujet à un compromis biais-variance. Ici, nous considérons l'utilisation de l'agrégation d'hold-out pour agréger les estimateurs appartenant à ces familles paramétrées, dans le cadre de la régression linéaire avec perte de Huber. L'agrégation d'hold-out (Agghoo) est une méthode consistant à moyenner des estimateurs sélectionnés par validation simple. Nous démontrons que l'agrégation d'hold-out vérifie une inégalité d'oracle non-asymptotique quand elle est appliquée aux estimateurs parcimonieux paramétrés par le nombre de coefficients non-nuls. En particulier, le résultat s'applique à une variante du Lasso proposée par Zou,Hastié et Tibshirani (2007). Des simulations permettent de comparer Agghoo et la validation croisée. Elles montrent la supériorité d'Agghoo par rapport à la validation croisée quand la dimension intrinsèque est élevée et que des variables non-pertinentes sont corrélées avec les variables prédictives.

Mots-clés. Sélection d'hyperparamètre, Régression parcimonieuse, Régression robuste, Agrégation, Validation Croisée, Lasso

Abstract. Sparse linear regression methods generally have a free hyperparameter which controls the amount of sparsity, and is subject to a bias-variance tradeoff. This article considers the use of Aggregated hold-out to aggregate over values of this hyperparameter, in the context of linear regression with the Huber loss function. Aggregated hold-out (Agghoo) is a procedure which averages estimators selected by hold-out (cross-validation with a single split). In the theoretical part of the article, it is proved that Agghoo satisfies a non-asymptotic oracle inequality when it is applied to sparse estimators which are parametrized by their zero-norm. In particular, this includes a variant of the Lasso introduced by Zou, Hastié and Tibshirani (2007). Simulations are used to compare Agghoo with cross-validation. They show that Agghoo performs better than CV when the intrinsic dimension is high and when there are confounders correlated with the predictive covariates.

Keywords. Hyperparameter selection, Sparse regression, Robust regression, Aggregation, Cross-validation, Lasso

1 Texte long

La régression linéaire consiste à prédire une variable cible Y à l'aide d'un vecteur X de variables auxiliaires et d'un *prédicteur linéaire* t, de la forme

$$t(x) = \langle \theta, x \rangle + q, \text{ où } q \in \mathbb{R} \text{ et } \theta \in \mathbb{R}^d.$$

L'erreur de prédiction est mesurée à l'aide d'une fonction ϕ du résidu $Y - t(X)$, ce qui permet de définir un *risque*

$$\mathcal{L}(t) = \mathbb{E}\left[\phi\left(Y - t(X)\right)\right],$$

et un *excès de risque*,

$$\ell(s,t) = \mathcal{L}(t) - \inf_f \mathcal{L}(f),$$

correspondant à l'écart entre le risque de t et le risque minimal (pris sur toutes les fonctions mesurables de X), supposé atteint en s.

L'objectif de la régression linéaire, envisagée sous l'angle de la *prédiction*, est alors de minimiser le risque ou, de façon équivalente, l'excès de risque. Pour ce faire, on suppose disposer d'un jeu de données constituée de n couples i.i.d $(X_i, Y_i)_{1 \leq i \leq n}$ suivant la même loi que (X, Y).

Le cas le plus classique est celui de la régression des moindres carrés, correspondant à la fonction $\phi(x) = x^2$. En *régression robuste*, on considère plutôt des fonctions de perte Lipschitz, qui résistent mieux à la contamination des données Y_i. La fonction de perte de Huber,

$$\phi_c(u) = \frac{u^2}{2}\mathbb{I}_{|u| \leq c} + c\left(|u| - \frac{c}{2}\right)\mathbb{I}_{|u| > c},$$

fait partie des plus populaires.

La possibilité de résoudre le problème de régression linéaire dépend de d et n. Pour la perte des moindre carrés, il est bien connu que le *risque minimax* de la régression linéaire est au moins de l'ordre de $\frac{d}{n}$. Dans le cas de la perte de Huber, nous avons démontré qu'il en allait de même tant que $d < n$ (Maillard, 2021).

Ainsi, si d est d'ordre n, l'excès de risque est minoré par une constante, donc il n'existe pas d'estimateur consistant de θ. Cependant, il arrive en pratique que d soit d'ordre n, voire (beaucoup) plus grand que n. Il faut alors trouver moyen de réduire la dimension effective du problème afin de se ramener à un modèle pour lequel l'estimation est possible. En régression parcimonieuse, l'on suppose que seul un petit nombre de variables sont utiles pour prédire Y. On cherche donc des prédicteurs linéaires $t : x \mapsto \langle \theta, x \rangle + q$, tels que le nombre de coefficients non nuls de θ, $\|\theta\|_0 = |\{i : \theta_i \neq 0\}|$, soit petit. La quantité $\|\theta\|_0$ permet alors de mesurer la "complexité" du paramètre θ, et ainsi d'envisager un compromis biais-variance, entre le nombre de coefficients non nuls autorisés et le risque empirique atteint.

1.1 Agrégation d'hold-out

Ici, nous considérerons une suite d'estimateurs de régression linéaires

$$\hat{t}_k : x \mapsto \langle \hat{\theta}_k, x \rangle + \hat{q}_k$$

indexée par le paramètre $\left\|\hat{\theta}_k\right\|$, i.e tels que $\left\|\hat{\theta}_k\right\| = k$. Nous supposons aussi que le paramètre $\hat{q}_k$ est obtenu par minimisation du risque empirique à $\hat{\theta}_k$ fixé: presque sûrement, pour tout jeu de données D_n,

$$\hat{q}_k(D_n) \in \operatorname*{argmin}_{q \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \phi_c \left(Y_i - \langle \theta, X_i \rangle - q \right).$$

Ces deux conditions sont vérifiées par les algorithmes gloutons qui ajoutent (ou enlèvent) une variable à chaque étape, ainsi que par une variante du Lasso proposée par Zou, Hastié et Tibshirani (2007) . Le paramètre k est alors laissé au choix de l'utilisateur.

Ici, nous envisageons d'agréger les estimateurs $(\hat{t}_k)_{1 \leq k \leq K}$ en utilisant l'*agrégation d'hold-out* définie par Maillard, Arlot et Lerasle (2021). Il s'agit de moyenner plusieurs estimateurs $(\hat{t}_{\hat{k}_i})_{1 \leq i \leq V}$ obtenus par validation simple, pour divers découpages des données en échantillon d'entraînement $(X_j)_{j \in T_i}$ et échantillon de validation $(X_j)_{j \notin T_i}$. Nous définissons ainsi l'estimateur

$$\hat{t}^{ag} = \frac{1}{V} \sum_{i=1}^{V} \hat{t}_{\hat{k}_i}.$$

Nous supposerons par la suite que les ensembles finis T_i sont de même cardinal n_t.

L'idée d'agréger des estimateurs en régression linéaire parcimonieuse a déjà connu un certain succès. Ici, il s'agit de combiner agrégation et calibration du paramètre k.

1.2 Théorie

Afin d'analyser théoriquement les performances de l'agrégation d'hold-out, nous faisons trois hypothèses.

- La loi conditionnelle de Y sachant X doit suffisamment "charger" l'intervalle $\left[s(X) - \frac{c}{2}, s(X) + \frac{c}{2}\right]$.
- La partie linéaire des prédicteurs $\hat{t}_k$ ne doit pas être excessivement grande, en ce sens que $\langle \hat{\theta}_k, X - EX \rangle$ et $\langle \hat{\theta}_k, X_i - EX \rangle$ sont supposés majorés en norme L^1 par un polynôme Ln_t^{α}.
- L'hypothèse cruciale concerne le rapport entre la "norme d'Orlicz" $\|\langle \theta, X \rangle\|_{\psi_1}$ et la norme L^2, $\|\langle \theta, X \rangle\|_{L^2}$, pour des vecteurs θ tels que $\|\theta\|_0 \leq K$. Ce rapport doit être au plus d'ordre $\frac{\sqrt{n_t}}{\log^{\beta} n_t}$, pour un certain $\beta > 0$.

Sous ces trois hypothèses, nous avons le résultat suivant.

Theorem 1.1 *Il existe des constantes r_0, r_1, r_2 telles que*

$$\mathbb{E}\left[\ell\big(s,\hat{t}^{ag}\big)\right] \leq r_0 \mathbb{E}\left[\min_{1\leq k\leq K} \ell(s,\hat{t}_k)\right] + r_1 \frac{c^2 \log n_t}{\eta(n-n_t)} + r_2 \frac{cLK\log K}{n_t^2\sqrt{n-n_t}}.$$

En choisissant $n_t = \lfloor \frac{n}{2} \rfloor$ et $K \leq n$ (par exemple), on constate que le terme de reste est d'ordre $\frac{\log n}{n}$, ce qui est optimal dans le cas $d >> \log n$.

1.3 Simulations

Dans cette section, nous comparons l'agrégation d'hold-out à la validation croisée par des simulations numériques. Soient r, s des entiers naturels tels que $rs \leq d$. Soient $(Z_i^0)_{1\leq i\leq r}, (Z_{i,j})_{1\leq i\leq r, 1\leq j\leq s}, (W_i)_{1\leq i\leq d-rs}$ des vecteurs gaussiens standards. Nous générons des variables $X \in \mathbb{R}_d$ dont les coordonnées sont définies par

$$X_{is+j} = \sqrt{0.8} Z_{i+1}^0 + \sqrt{0.2} Z_{i+1,j}$$

pour $i \in [|0; r-1|]$ et $j \in [|1; s|]$, et $X_i = W_{i-rs}$ pour $rs \leq i \leq d$. Les variables Y sont générées suivant la loi conditionnelle Cauchy$(\langle w_*, X\rangle, 0.3)$, où $w_* = \frac{3u}{\|Xu\|_{L^2}}$ et $u = \big(\mathbb{I}_{s|j-1}\mathbb{I}_{j\leq rs}\big)_{1\leq j\leq d}$. Ainsi, chaque "bloc" de s variables contient une variable prédictive et $s-1$ variables superflues, ayant chacune une corrélation de 0.8 avec la variable prédictive.

La figure 1 représente la performance d'Agghoo et de la validation croisée, où la valeur 1 correspond au choix optimal de k (*oracle de sélection*) et où le risque est estimé à l'aide de la perte Huber ϕ_2.

On constate que l'agrégation d'hold-out surpasse la validation croisée quand r et s sont suffisament grands, c'est à dire quand la dimension "intrinsèque" $\|w_*\|_0$ du problème et le niveau de corrélation sont suffisamment élevés.

Bibliographie

Maillard, G. (2021). Aggregated hold-out for sparse linear regression with a robust loss function, *arXiv e-prints*.

Maillard, G. and Arlot, S. and Lerasle, M. (2021). Aggregated Hold-out, *Journal of Machine Learning Research*, 22, pp. 1-55.

Zou, H. and Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the Lasso, *Annals of Statistics*, 35, pp. 2173-2192.

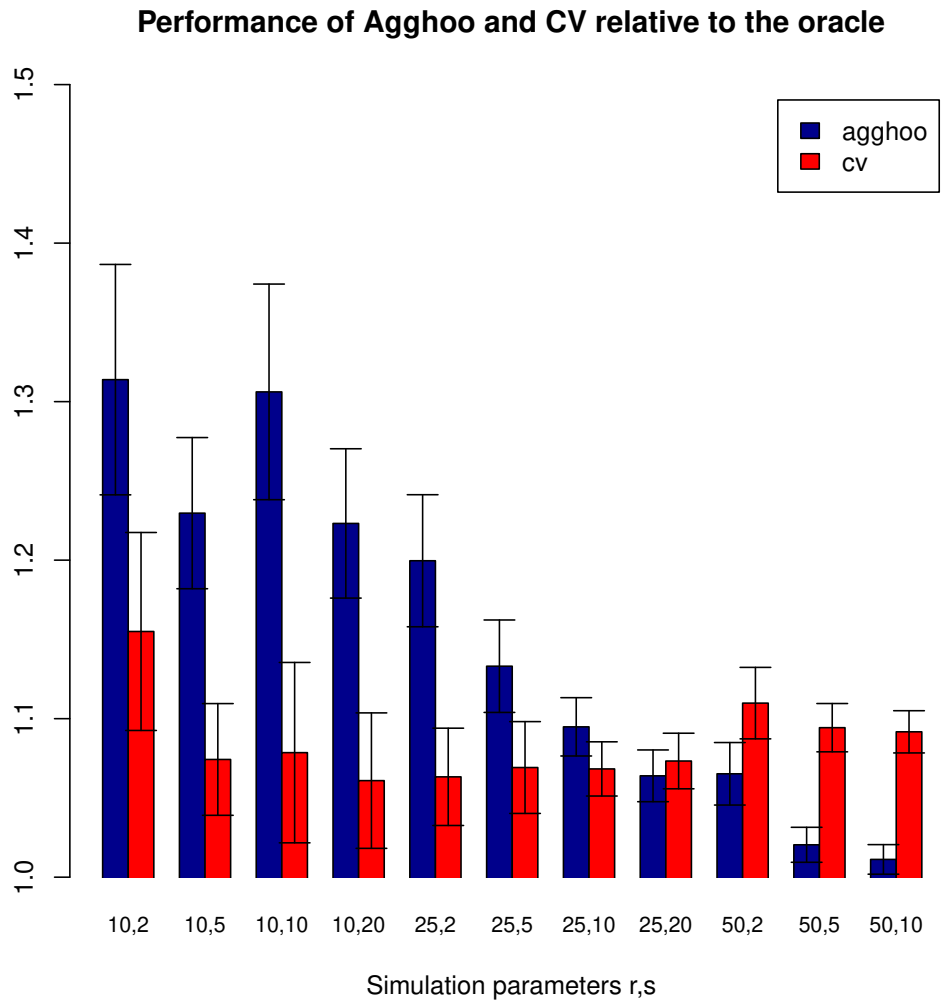

Figure 1: Risque estimé d'Agghoo et de la validation croisée, divisé par celui de l'oracle.