Compromis précision - temps de calcul appliqué au problème de régression linéaire

Maxime Brunin [1] & Christophe Biernacki [2] & Alain Celisse [3]

[1] *Université de Lille & Inria, maxime.brunin@inria.fr*
[2] *Université de Lille & Inria & CNRS, christophe.biernacki@math.univ-lille1.fr*
[3] *Université de Lille & Inria & CNRS, alain.celisse@math.univ-lille1.fr*

Résumé. Dans le cadre de la régression linéaire, notre objectif est de trouver un estimateur qui soit meilleur en terme de "précision" que l'estimateur des moindres carrés (EMC). Cet estimateur alternatif est construit à l'aide d'un algorithme de descente de gradient à pas fixe α et d'un temps d'arrêt. Ce temps d'arrêt assure la "précision" de cet estimateur alternatif. La perspective de ce travail sera d'étendre au cas d'estimateurs non explicites afin d'avoir en plus un gain en temps de calcul.

Mots-clés. régression linéaire, descente de gradient, temps d'arrêt.

Abstract. In linear regression, our goal is to find an estimator which performs better in terms of "accuracy" than the Ordinary Least Squares (OLS). This alternative estimator is built thanks to a gradient descent algorithm with fixed step α and a stopping time. This stopping time ensures the "accuracy" of this alternative estimator. The perspective of this work is to extend to the case of estimators which have no closed formula in order to have a gain in computation time.

Keywords. linear regression, gradient descent, stopping time.

1 Introduction

Le modèle linéaire que nous étudions est classique : $Y = X\theta^* + \epsilon$, avec $X \in \mathcal{M}_{n,d}(\mathbb{R})$ et $\text{rg}(X) = d$; $\theta^* \in \mathbb{R}^d$ est inconnu ; $\epsilon \in \mathbb{R}^n$ dont les coordonnées sont des variables indépendantes et sous-gaussiennes de paramètre et de variance σ^2. Nous nous plaçons donc dans le cas $n > d$.

Nous abordons ce problème de régression linéaire dans le cadre du compromis précision - temps de calcul. Nous proposons un estimateur qui soit meilleur que l'EMC $\hat{\theta} = \left(X^T X\right)^{-1} X^T Y$ en terme de "précision" sur les prévisions. Il existe des estimateurs alternatifs à l'EMC qui sont par exemple : l'estimateur ridge proposé par Hoerl (1970) dont l'erreur moyenne quadratique (EQM) est inférieure à celle de l'EMC pour un paramètre ridge λ suffisamment petit ; l'estimateur de James Stein (1961) dont l'erreur moyenne quadratique est inférieure à celle de l'EMC dès que le nombre de variables est supérieur ou égal à 3. Cependant, ces estimateurs ont les défauts suivants : l'estimateur ridge nécessite une

calibration de λ ; certaines coordonnées de l'estimateur de James Stein sont de mauvais estimateurs des coordonnées correspondantes de θ^*.

Nous utilisons une approche due à Wainwright (2014) qui utilise un temps d'arrêt pour améliorer la "précision" et le temps de calcul de l'estimateur construit grâce à un algorithme de descente de gradient dans le cadre des méthodes à noyaux. Cet algorithme de descente de gradient à pas fixe α est appliqué à la fonction $g(\theta) = \|Y - X\theta\|_{2,n}^2$ (pour $\theta \in \mathbb{R}^d$; $\forall x \in \mathbb{R}^n,\ \|x\|_{2,n}^2 = \sum_{i=1}^n x_i^2$) pour obtenir un estimateur à l'itération k de θ^* noté $\hat{\theta}^{(k)}$. La relation entre $\hat{\theta}^{(k+1)}$ et $\hat{\theta}^{(k)}$ est : $\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} - \alpha \nabla g(\hat{\theta}^{(k)})$ où $\nabla g(\hat{\theta}^{(k)})$ est le gradient de g au point $\hat{\theta}^{(k)}$; $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_n$ sont les valeurs propres de $\frac{1}{n} XX^T$; α appartient à l'intervalle $\left]0, \min\left(1, \frac{1}{\hat{\lambda}_1}\right)\right[$ pour assurer la convergence de la suite $\{\hat{\theta}^{(k)}\}_{k \in \mathbb{N}}$ vers $\hat{\theta}$.

Dans le but d'évaluer la "précision" sur les prévisions de l'estimateur alternatif $\hat{\theta}^{(k)}$ ainsi obtenu, nous calculons l'estimateur $\hat{Y}^{(k)} = X\hat{\theta}^{(k)}$. Nous souhaitons stopper l'algorithme lorsque la "précision" de l'estimateur alternatif $\hat{\theta}^{(\hat{k}^*)}$ ($\hat{k}^*$ est le temps d'arrêt), évaluée par $\Delta\left(\hat{Y}^{(\hat{k}^*)}\right) = \frac{1}{n}\left\|\hat{Y}^{(\hat{k}^*)} - Y^*\right\|_{2,n}^2$ ($Y^* = X\theta^*$) est minimale.

Dans la suite de ce document, nous présentons l'estimateur alternatif $\hat{\theta}^{(\hat{k}^*)}$ puis un théorème garantissant la "précision" de cet estimateur. Des simulations permettent d'illustrer la bonne performance en matière de précision de cet estimateur.

2 Temps d'arrêt comme compromis biais - variance

2.1 Effet et interprétation du temps d'arrêt sur la prévision

Choisir un temps d'arrêt permet d'améliorer la "précision" sur les prévisions de $\hat{\theta}^{(k)}$ par rapport à l'EMC $\hat{\theta}$ car on a $\mathrm{E}\left[\Delta\left(\hat{Y}^{(\overline{k}^*)}\right)\right] < \mathrm{E}\left[\Delta\left(\hat{Y}\right)\right]$, où $\Delta(.) = \frac{1}{n}\|. - Y^*\|_{2,n}^2$; $\hat{Y} = X\hat{\theta}$; $\overline{k}^* = \underset{k \in \mathbb{N}}{\arg\min}\ \left\{\mathrm{E}\left[\Delta\left(\hat{Y}^{(k)}\right)\right]\right\}$. Nous travaillons dans cette partie sur $\mathrm{E}\left[\Delta\left(\hat{Y}^{(k)}\right)\right]$ mais notre but final est de travailler sur $\Delta\left(\hat{Y}^{(k)}\right)$. La propriété $\mathrm{E}\left[\Delta\left(\hat{Y}^{(\overline{k}^*)}\right)\right] < \mathrm{E}\left[\Delta\left(\hat{Y}\right)\right]$ peut être interprétée comme un compromis biais - variance. En effet, $\mathrm{E}\left[\Delta\left(\hat{Y}^{(k)}\right)\right]$ peut être décomposé classiquement en une somme du biais au carré et de la variance de $\hat{Y}^{(k)}$ notés $\mathrm{b}\left(\hat{Y}^{(k)}\right)^2$ et $\mathrm{var}\left(\hat{Y}^{(k)}\right)$ respectivement :

$$\mathrm{E}\left[\Delta\left(\hat{Y}^{(k)}\right)\right] = \underbrace{\frac{1}{n}\left\|S^k P^T\left(Y^{(0)} - Y^*\right)\right\|_{2,n}^2}_{\mathrm{b}\left(\hat{Y}^{(k)}\right)^2} + \underbrace{\frac{\sigma^2}{n}\mathrm{Tr}\left(\left(I_n - S^k\right)^2\right)}_{\mathrm{var}\left(\hat{Y}^{(k)}\right)},$$

où $\frac{1}{n} XX^T = P\Lambda P^T$ est obtenu par décomposition en valeurs singulières ; $S = I_n - \alpha\Lambda$.

Lorsque $k = 0$, la variance de $\hat{Y}^{(k)}$ est égale à 0 et le biais au carré de $\hat{Y}^{(k)}$ est potentiellement élevé $\left(\mathrm{b}\left(\hat{Y}^{(0)}\right)^2 = \frac{1}{n}\left\|Y^{(0)} - Y^*\right\|_{2,n}^2\right)$. A l'inverse, lorsque k tend vers $+\infty$, la variance de $\hat{Y}^{(k)}$ tend vers $\frac{\sigma^2 d}{n}$ et le biais au carré de $\hat{Y}^{(k)}$ tend vers 0.

La figure 1 montre que $\mathrm{E}\left[\Delta\left(\hat{Y}^{(\overline{k}^*)}\right)\right] < \mathrm{E}\left[\Delta\left(\hat{Y}\right)\right] = \lim_{k\to+\infty} \mathrm{E}\left[\Delta\left(\hat{Y}^{(k)}\right)\right]$ et illustre le compromis biais - variance expliqué ci-dessus que doit traduire le temps d'arrêt $\hat{k}^*$, approchant ici $\overline{k}^*$.

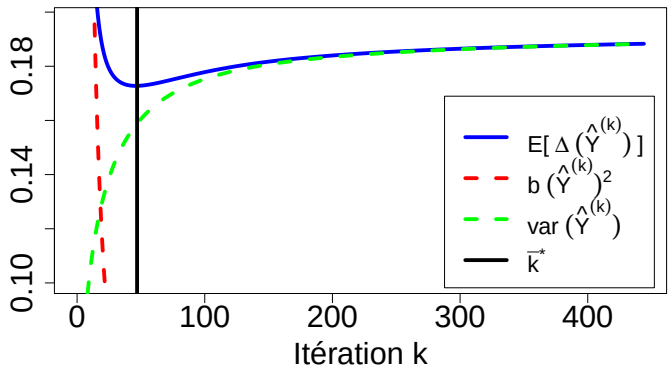

FIGURE 1 – Graphique de $\mathrm{E}\left[\Delta\left(\hat{Y}^{(k)}\right)\right]$, du biais au carré et de la variance de $\hat{Y}^{(k)}$ en fonction de k pour $n = 30$, $d = 20$ et $\alpha = \frac{9}{10}\min\left(1, \frac{1}{\hat{\lambda}_1}\right)$.

2.2 Contrôle du biais et de la variance

Nous travaillons sur $\Delta\left(\hat{Y}^{(k)}\right) = \frac{1}{n}\left\|\hat{Y}^{(k)} - Y^*\right\|_{2,n}^2$, qui mesure la qualité prédictive, plutôt que sur $\mathrm{E}\left[\Delta\left(\hat{Y}^{(k)}\right)\right]$. Nous cherchons à estimer $k^* = \arg\min_{k\in\mathbb{N}} \left\{\Delta\left(\hat{Y}^{(k)}\right)\right\}$ par le temps d'arrêt $\hat{k}^*$. Nous souhaitons que $\hat{k}^*$ minimise $\Delta\left(\hat{Y}^{(k)}\right)$. Comme il est difficile de minimiser $\Delta\left(\hat{Y}^{(k)}\right)$, nous minimisons un de ses majorants. Ce majorant se décompose en un terme de biais au carré B_k^2 et un terme de variance V_k grâce à l'inégalité (1). Nous retrouvons l'idée de la section 2.1 où $\hat{k}^*$ traduit un compromis biais - variance.

$$\Delta\left(\hat{Y}^{(k)}\right) \leq \underbrace{\frac{2}{n}\left\|E\left[\hat{Y}^{(k)}\right] - Y^*\right\|_{2,n}^2}_{B_k^2} + \underbrace{\frac{2}{n}\left\|\hat{Y}^{(k)} - E\left[\hat{Y}^{(k)}\right]\right\|_{2,n}^2}_{V_k}. \qquad (1)$$

On majore B_k^2 et V_k par $B_k^{2,\sup}$ et $V_k^{\sup}$ respectivement par les deux lemmes suivants. Le lemme 2.1 est un résultat de Wainwright (2014) tandis que le lemme 2.2 est une adaptation d'un lemme de Wainwright (2014) lorsque σ^2 n'est pas connu.

Lemme 2.1 *Si* $\|\theta^*\|_{2,d} \leq 1$ *et* $\theta_0 = 0$, $\forall k \geq 1$,

$$B_k^2 \leq \frac{1}{\mathrm{e}k\alpha} =: B_k^{2,sup}.$$

Lemme 2.2 *D'après l'inégalité de concentration de Wright (1973), sur un évènement* $\mathcal{A}_q$ *de grande probabilité,* $\forall k \in [\![1, \hat{k}^*]\!]$,

$$V_k \leq 5\sigma^2 k\alpha \left[R_K\left(\frac{1}{\sqrt{k\alpha}}\right)\right]^2 =: V_k^{sup},$$

où $\forall \epsilon > 0,\ R_K(\epsilon) = \sqrt{\frac{1}{n}\sum_{i=1}^d \min\left(\hat{\lambda}_i, \epsilon^2\right)}$.

2.3 Temps d'arrêt proposé et propriétés associées

Le temps d'arrêt $\hat{k}^*$ que nous proposons est défini par

$$\hat{k}^* = \max\left\{k \in \mathbb{N} : V_k^{\sup} \leq c\left(\frac{\sigma}{\hat{\sigma}}\right)^2 B_k^{2,\sup}\right\}, \tag{2}$$

où $\hat{\sigma}^2$ est un estimateur de σ^2.

A la différence de Wainwright (2014), le temps d'arrêt $\hat{k}^*$ ne dépend que des données. On devrait choisir le paramètre c tel que $\hat{k}^*$ soit proche de k^*. Ce choix de c fait cependant l'objet d'un travail en cours. On peut malgré tout interpréter c utilement comme un compromis biais - variance car il est approximativement égal au rapport entre $V_k^{\sup}$ et $B_k^{2,\sup}$ en k^*. Nos travaux en cours indiquent que c devrait dépendre de n tandis que Wainwright (2014) le fixe arbitrairement à $\frac{5}{4\mathrm{e}}$ jusqu'à présent. Nous nous concentrons ici d'étudier empiriquement l'influence de c sur $\mathrm{E}\left[\Delta\left(\hat{Y}^{(\hat{k}^*)}\right)\right]$ sur des simulations en section 2.4.

D'autres temps d'arrêt existent pour stopper un algorithme de descente de gradient : Reiß (2016) définit un temps d'arrêt τ qui garantit que la distance euclidienne de Y à la valeur prédite à l'itération $k = \tau$ ne soit pas inférieure à un seuil et qui imite le comportement d'un temps d'arrêt analogue à $\hat{k}^*$.

Nous présentons maintenant notre résultat principal concernant les propriétés de $\hat{k}^*$.

Théorème 2.1 *On suppose que* $\|\theta^*\|_{2,d} \leq 1$ *et* $\theta_0 = 0$. *Etant donné le temps d'arrêt* $\hat{k}^*$ *défini par* (2), *il existe des constantes* $c_1, c_2 \in \mathbb{R}_+^*$ *telles que sur l'évènement* $\mathcal{A}_q \subset \Omega_q$ *de probabilité au moins* $P(\Omega_q) - c_1 \exp\left(-c_2 n\varepsilon_{\sigma_q^1}^4\right)$ $\left(\Omega_q = \{|\,\hat{\sigma} - \sigma\,| \leq q\sigma\}\,;\ \sigma_q^1 = (1-q)\sigma\right)$:

(a) $\forall k \in [\![1, \hat{k}^*]\!]$

$$\Delta\left(\hat{Y}^{(k)}\right) \leq \frac{1}{\mathrm{e}k\alpha}\left[1 + c\left(\frac{\sigma}{\hat{\sigma}}\right)^2\right].$$

(b) $\forall k \geq 0$,

$$\mathrm{E}\left[\Delta\left(\hat{Y}^{(k)}\right)\right] \geq \frac{\sigma^2}{4}\underbrace{(k\alpha)^2\left(R_K\left(\frac{1}{\sqrt{k\alpha}}\right)\right)^4}_{f(k)},$$

où $\varepsilon_\sigma = \inf\left\{\varepsilon > 0 : R_K(\varepsilon) \leq \sqrt{\frac{c}{5\mathrm{e}}}\frac{\varepsilon^2}{\sigma}\right\}$ *;* f *est croissante sur* $\mathbb{N}$ *vérifiant* $f(k) \underset{k\to+\infty}{\longrightarrow} 1$.

Nous interprétons ce théorème de la façon suivante : l'assertion (a) signifie que la distance de $\hat{Y}^{(k)}$ à Y^* tend à décroître au moins jusqu'au temps d'arrêt $\hat{k}^*$ car elle est majorée par une fonction décroissante de k. L'assertion (b) signifie que $\mathrm{E}\left[\Delta\left(\hat{Y}^{(k)}\right)\right]$ se dégrade si l'on choisit une itération k trop grande (en particulier $k > \hat{k}^*$) car elle est minorée par une fonction croissante de k qui tend vers $\frac{\sigma^2}{4}$ lorsque k tend vers $+\infty$.

2.4 Simulations

Nous générons, $\forall i \in [\![1, n]\!]$, la i$^{\mathrm{e}}$ ligne de X notée x_i par $\mathcal{N}(0, \Sigma)$ avec Σ une matrice diagonale ayant sa plus petite valeur propre égale à 1, sa plus grande valeur propre égale à 10 et les autres valeurs propres choisies uniformément entre 1.1 et 9.9 ; $\forall j \in [\![1, d]\!]$, $(\theta^*)_j = \frac{1}{\sqrt{d}} \Rightarrow \|\theta^*\|_{2,d} = 1$; $\epsilon \sim \mathcal{N}(0, \sigma^2)$ où σ^2 est choisi tel que le rapport signal à bruit "SNR" est égal 4 : $\mathrm{SNR}^2 = \frac{\|X\theta^*\|_{2,n}^2}{\|\epsilon\|_{2,n}^2} \approx \frac{(\theta^*)^T\Sigma\theta^*}{\sigma^2}$; $\alpha = \frac{9}{10}\min\left(1, \frac{1}{\hat{\lambda}_1}\right)$.

Nous étudions à présent l'influence de c sur la "précision" des prévisions de $\hat{\theta}^{(\hat{k}^*)}$ en traçant $\mathrm{E}\left[\Delta\left(\hat{Y}^{(\hat{k}^*)}\right)\right]$ et $\mathrm{E}\left[\Delta\left(\hat{Y}\right)\right]$ en fonction de n à d fixé. Nous effectuons ces simulations pour deux valeurs de c : $c = \frac{5}{4\mathrm{e}} \approx 0.46$ est la valeur arbitraire choisie par Wainwright (2014) et $c = 2.5$.

La figure 2 montre que, pour une valeur de c favorisant la variance ($c = 2.5$), on a $\mathrm{E}\left[\Delta\left(\hat{Y}^{(\hat{k}^*)}\right)\right] < \mathrm{E}\left[\Delta\left(\hat{Y}\right)\right]$, ce qui n'est pas le cas pour une valeur de c favorisant le biais ($c \approx 0.46$). De plus, la figure 2 montre que la différence entre $\mathrm{E}\left[\Delta\left(\hat{Y}\right)\right]$ et $\mathrm{E}\left[\Delta\left(\hat{Y}^{(\hat{k}^*)}\right)\right]$ varie avec n donc le fait que le paramètre c dépende de n semble expérimentalement confirmé.

3 Conclusion

Nous avons proposé un estimateur $\hat{\theta}^{(\hat{k}^*)}$, construit à partir d'un temps d'arrêt, qui soit plus "précis" que l'EMC pour une valeur de c bien réglée favorisant la variance de $\hat{Y}^{(\hat{k}^*)}$.

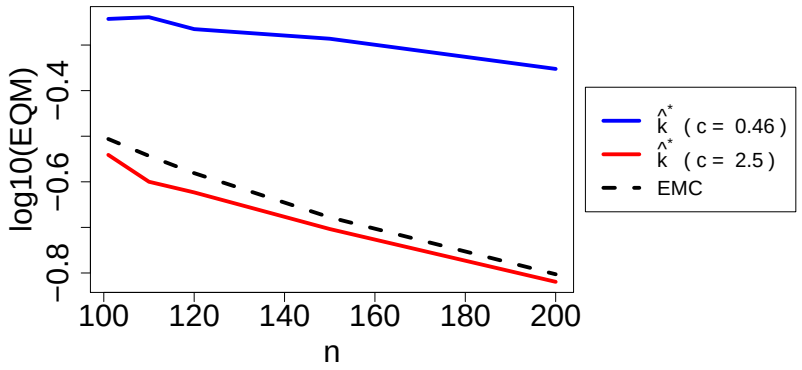

FIGURE 2 – Graphique de $\log 10$ de $\mathrm{E}\left[\Delta\left(\hat{Y}^{(\hat{k}^*)}\right)\right]$ (pour $c \approx 0.46$ et $c = 2.5$) et $\log 10$ de $\mathrm{E}\left[\Delta\left(\hat{Y}\right)\right]$ en fonction de n pour $d = 100$.

Cependant, le temps de calcul de $\hat{\theta}^{(\hat{k}^*)}$ est supérieur à l'EMC car l'EMC a une formule explicite. Nous avons prouvé que le temps d'arrêt améliore la précision de $\hat{\theta}^{(\hat{k}^*)}$ sur les prévisions. En perspective, nous devons trouver la dépendance en n de c afin de calibrer c. De plus, nous voulons étendre l'utilisation des temps d'arrêts à des cas où $\hat{\theta}$ n'a pas de formules explicites. Dans ce cadre, les temps d'arrêts peuvent améliorer simultanément la précision et le temps de calcul.

Bibliographie

[1] A. E. HOERL AND R. W. KENNARD (1970), Ridge Regression : Biased Estimation for Nonorthogonal Problems, *Technometrics*, vol. 12, n° 1, pp. 55 - 67.
[2] W. JAMES AND C. STEIN (1961), Estimation with Quadratic Loss, *Proc. Fourth Berkeley Symp. on Math. Statist. and Prob.*, vol. 1, pp. 361 - 379.
[3] M. REIß, M. HOFFMANN, G. BLANCHARD (2016), Optimal adaptation for early stopping in statistical inverse problems, *Preprint arXiv :1606.07702*.
[4] M. J. WAINWRIGHT AND G. RASKUTTI (2014), Early Stopping and Non-parametric Regression : An Optimal Data-dependent Stopping Rule, *JMLR*, vol. 15, n° 1, pp. 335 - 366.
[5] F. T. WRIGHT (1973), A Bound on Tail Probabilities for Quadratic Forms in Independent Random Variables Whose Distributions are not Necessarily Symmetric, *Ann. Prob.*, vol. 1, n° 6, pp. 1068 - 1070.