Bornes de risque améliorées pour le Lasso

Arnak Dalalyan [1] & Mohamed Hebiri [2] & Johannes Lederer [3]

[1] *ENSAE-CREST, 3, Avenue Pierre Larousse, 92240 Malakoff, France,*
arnak.dalalyan@ensae.fr
[2] *Université Paris-Est, 5, boulevard Descartes, 77454 Marne-la-Vallée, France.*
Mohamed.Hebiri@univ-mlv.fr
[3] *Cornell University, 1188 Comstock Hall, Ithaca, NY 14853-2601, Etats-Unis*
johannesslederer@cornell.edu

Résumé. Malgré le nombre important de travaux concernant l'étude théorique de l'estimateur Lasso, la relation entre sa performance statistique et les corrélations entre les variables explicatives n'est pas très bien comprise. Le but de ce travail est de fournir de nouveaux résultats sur cette relation dans le cadre de la régression linéaire multiple. D'une part, nous montrons que l'incorporation dans le paramètre d'ajustement d'un indicateur simple de corrélation entre les variables conduit à des inégalités oracles exactes avec un terme résiduel optimal même dans des cas où les variables explicatives sont fortement corrélées. D'autre part, nous exhibons un exemple révélant que pour certaines matrices de design comportant des variables modérément corrélées, la qualité de la prédiction Lasso est médiocre quelle que soit la valeur du paramètre d'ajustement.

Mots-clés. Pénalité ℓ_1, prédiction, parcimonie, inégalité oracle

Abstract. Although the Lasso has been extensively studied, the relationship between its prediction performance and the correlations of the covariates is not fully understood. In this document, we give new insights into this relationship in the context of multiple linear regression. We show, in particular, that the incorporation of a simple correlation measure into the tuning parameter leads to a nearly optimal prediction performance of the Lasso even for highly correlated covariates. However, we also reveal that for moderately correlated covariates, the prediction performance of the Lasso can be mediocre irrespective of the choice of the tuning parameter.

Keywords. ℓ_1-penalty, prediction, sparsity, oracle inequality

1 Introduction

La présente étude porte sur la performance en terme de risque de prédiction de l'estimateur Lasso. Nous nous intéressons au modèle de la régression linéaire multiple avec un design déterministe.Plus précisément, nous supposons que les observations $y_1, \ldots, y_n \in \mathbb{R}^n$ et $\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p \in \mathbb{R}^n$ sont disponibles. De plus, nous supposons que pour un vecteur de

régression $\boldsymbol{\beta}^* \in \mathbb{R}^p$ et un niveau de bruit $\sigma^* > 0$ les variables aléatoires $y_i - \beta_1^*(\boldsymbol{x}^1)_i - \ldots - \beta_p^*(\boldsymbol{x}^p)_i$ sont i.i.d. gaussiennes de moyenne 0 et de variance σ^{*2}. En d'autres termes,

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \sigma^* \mathcal{N}_n(0, \mathbf{I}_n), \tag{1}$$

où $\boldsymbol{y} := (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ est le vecteur réponse, $\mathbf{X} := (\boldsymbol{x}^1, \ldots, \boldsymbol{x}^p) \in \mathbb{R}^{n \times p}$ est la matrice de design déterministe (sans perte de généralité, on suppose que $\|\boldsymbol{x}^j\|_2^2 \leq n$ pour tout $j \in \{1, \ldots, p\}$), $\xi \in \mathbb{R}^n$ est le vecteur de bruit, et $\mathbf{I}_n$ est la matrice identité de taille $n \times n$.

Rappelons que le Lasso (cf. Tibshirani (1996)) est défini comme solution du problème d'optimisation convexe

$$\hat{\boldsymbol{\beta}}_\lambda^{\text{Lasso}} \in \arg\min_{\boldsymbol{\beta}} \Big\{ \frac{1}{2n} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \Big\}. \tag{2}$$

Dans le document présent, nous allons répondre aux questions suivantes.

Question 1. Dans le cadre de l'estimation parcimonieuse, il existe des méthodes (Dalalyan et Tsybakov (2007), Rigollet et Tsybakov (2011), *etc.*) dont le risque de prédiction décroît vers zéro à la vitesse s^*/n ($s^* = \|\boldsymbol{\beta}^*\|_0$), dite vitesse rapide, sans aucune condition sur les corrélations entre les variables explicatives. En revanche les vitesses rapides pour le Lasso sont obtenues sous des contraintes relativement fortes sur les corrélations. Ces contraintes sont exprimées en terme de la condition d'isométrie restreinte, les valeurs propres restreintes, la cohérence mutuelle ou encore la condition d'incompatibilité (cf. van de Geer and Buhlmann (2009)). Par conséquent, nous ne savons pas si les vitesses rapides sont valables pour le Lasso quelle que soit les corrélations entre les variables. Cette question reste ouverte même si l'on autorise λ dépendre du bruit $\boldsymbol{\xi}$ et du vrai vecteur de régression $\boldsymbol{\beta}^*$.

Question 2. Pour des vecteurs parcimonieux $\boldsymbol{\beta}^*$ ayant pour support $J^* = \{j \in \{1, \ldots, p\} : \beta_j^* \neq 0\}$ et pour des variables fortement corrélées dans le sense où toutes les variables non-pertinentes $\{\boldsymbol{x}^j : j \notin J^*\}$ sont proches du sous-espace linéaire engendré par les variables pertinentes $\{\boldsymbol{x}^j : j \in J^*\}$, les résultats empiriques suggèrent que la perte de prédiction minimale est obtenue pour les valeurs de λ considérablement plus petites que la valeur universelle. Est-il possible de d'incorporer dans λ la géométrie des variables explicatives par le biais d'un indicateur simple, qui permettrait d'obtenir les vitesses rapides pour les variables très corrélées?

Dans toute la suite, on notera $\ell_n(\boldsymbol{\beta}, \boldsymbol{\beta}') := \frac{1}{n}\|\mathbf{X}(\boldsymbol{\beta} - \boldsymbol{\beta}')\|_2^2$ la perte de prédiction. Pour tout $T \subset [p]$, on note T^c et $|T|$ l'ensemble complémentaire $[p] \setminus T$ et le cardinal de T, respectivement. Pour toute matrice $\mathbf{A} \in \mathbb{R}^{p \times q}$ et pour tout $T \subset [q]$, $\mathbf{A}_T$ désigne la matrice obtenue de $\mathbf{A}$ en supprimant les colonnes appartenant à T^c. Pour un vecteur $\boldsymbol{u} \in \mathbb{R}^p$ et un ensemble $T \subset [p]$, $\boldsymbol{u}_T$ est le vecteur obtenu de $\boldsymbol{u}$ en supprimant les coordonnées appartenant à T^c. Pour deux vecteurs $\boldsymbol{u}$ et $\boldsymbol{u}'$, l'opération binaire $\odot$ désigne le produit élément-par-élément, $\boldsymbol{u} \odot \boldsymbol{u}' = (u_1 u_1', \ldots, u_p u_p')^\top$. On écrit $\mathbf{1}_p$ (resp. $\mathbf{0}_p$) pour le vecteur de $\mathbb{R}^p$ dont

toutes les coordonnées sont égales à un (resp. zéro). Pour tout sous-ensemble T de $[p]$, on note V_T le sous-espace linéaire de $\mathbb{R}^n$ engendré par les colonnes de $\mathbf{X}_T$. On note Π_T le projecteur orthogonal sur V_T et par ρ_T la distance Euclidienne (normalisée) maximale entre les colonnes de $\mathbf{X}$ et le sous-espace V_T, c'est-à-dire $\rho_T = n^{-1/2} \max_{j\in[p]} \|(\mathbf{I}_n - \Pi_T)\boldsymbol{x}^j\|_2$.

2 Un contre-exemple

Afin de répondre à la première question de l'introduction, considérons l'exemple suivant. Soit $n \geq 2$ un entier et soit m la partie entière de $\sqrt{2n}$. On définit $\mathbf{X} \in \mathbb{R}^{n\times 2m}$ par

$$\mathbf{X} = \sqrt{\frac{n}{2}} \begin{pmatrix} \mathbf{1}_m^\top & \mathbf{1}_m^\top \\ \mathbf{I}_m & -\mathbf{I}_m \\ \mathbf{0}_{(n-m-1)\times m} & \mathbf{0}_{(n-m-1)\times m} \end{pmatrix}.$$

Pour éviter les complications techniques sans importance, on considère ici que le bruit suit la loi de Rademacher, c'est-à-dire $\mathbf{P}(\boldsymbol{\xi} = \boldsymbol{s}) = 2^{-n}$ pour tout $\boldsymbol{s} \in \{\pm 1\}^n$ (en conséquence, $\sigma^* = 1$). Soit le vrai vecteur de régression $\boldsymbol{\beta}^* \in \mathbb{R}^{2m}$ défini par $\beta_1^* = \beta_{m+1}^* = 1$ et $\beta_j^* = 0$ pour tout $j \in [2m] \setminus \{1, m+1\}$.

Proposition 1. *Quel que soit $\lambda > 0$, la perte de prédiction du Lasso $\hat{\boldsymbol{\beta}}_\lambda^{\text{Lasso}}$ vérifie l'inégalité $\mathbf{P}\Big(\ell_n(\hat{\boldsymbol{\beta}}_\lambda^{\text{Lasso}}, \boldsymbol{\beta}^*) \geq \frac{1}{2\sqrt{2n}}\Big) \geq \frac{1}{2}$.*

Cet exemple est particulièrement instructif pour au moins trois raisons. Premièrement, il montre que les vitesses rapides sont inatteignables pour le Lasso même si les corrélations entre les variables sont bien loin de ± 1. Deuxièmement, le résultat ci-dessus étant valable pour tout $\lambda > 0$, il montre que la situation ne peut être sauvée par un choix malin de λ même si l'on utilisait toute l'information contenue dans $\boldsymbol{\beta}^*$. Troisièmement, le résultat est valable pour un niveau de parcimonie très faible: la norme ℓ_0 de $\boldsymbol{\beta}^*$ est égale à 2.

3 Résultats principaux

Les contributions principales de ce travail peuvent être séparées en deux catégories: des résultats fournissant des vitesses dites lentes pour le Lasso et des résultats établissant des vitesses rapides. Dans les deux cas, les bornes de risques établies tirent profit des corrélations entre les variables explicatives.

3.1 Vitesses "lentes"

Théorème 1. *Soit $T \subset [p]$ et soit $\delta > 0$ une constante. Si $\lambda \geq \sigma^* \rho_T \sqrt{2\log(p/\delta)/n}$, alors le Lasso (2) vérifie*

$$\ell_n(\hat{\boldsymbol{\beta}}_\lambda^{\text{Lasso}}, \boldsymbol{\beta}^*) \leq \inf_{\bar{\boldsymbol{\beta}}\in\mathbb{R}^p} \Big\{\ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + 4\lambda\|\bar{\boldsymbol{\beta}}\|_1\Big\} + \frac{2\sigma^{*2}(|T| + 2\log(1/\delta))}{n}$$

avec une probabilité supérieure à $1 - 2\delta$.

Ce qui rend ce théorème particulièrement intéressant en le différenciant des résultats existants est la présence du facteur ρ_T dans la borne inférieure de λ. Il est clair que ce facteur est toujours entre 0 et 1. Il est proche de zéro lorsque les variables non pertinentes sont proches des variables pertinentes. Cette observation implique le résultat suivant.

Proposition 2. *Soit* $T_n \subset [p]$ *tel que les variables* $\{\boldsymbol{x}^j : j \in [p]\}$ *sont proche du sous-espace linéaire engendré par* $\{\boldsymbol{x}^j : j \in T_n\}$, *dans le sense où* $\rho_{T_n} \lesssim n^{-r}$ *pour une constante* $r > 0$. *Alors, en choisissant* $\lambda \geq c\sigma^* \sqrt{\log(p)/n^{2r+1}}$ *pour une constante* $c > 0$ *suffisamment grande, le Lasso* (2) *vérifie*

$$\ell_n(\hat{\boldsymbol{\beta}}^{\text{Lasso}}_\lambda, \boldsymbol{\beta}^*) \lesssim \left(\sqrt{\frac{\log(p)}{n^{2r+1}}} \, \|\boldsymbol{\beta}^*\|_1 \right) \bigvee \frac{|T_n|}{n} \tag{3}$$

avec une probabilité proche de 1. En particulier, si les variables $\{\boldsymbol{x}^j : j \notin J^*\}$ *sont à une distance euclidienne bornée par une constante de Vect*$\{\boldsymbol{x}^j : j \in J^*\}$, *alors* $r = 1/2$ *et, par conséquent, le Lasso atteint la vitesse rapide* s/n *à des facteurs logarithmiques près, à condition que* λ *soit de l'ordre de* $\sqrt{\log(p)}/n$.

La dépendance de λ en T suggérée par le Theorem 1 et la Proposition 2 peut engendrer des difficultés computationnelles supplémentaires. Dans certaines applications, l'ensemble T est prédéterminé. Dans d'autres applications, malheureusement, on doit parcourir l'ensemble des T possibles pour trouver celui qui minimise ρ_T. La Proposition 3 ci-après fournit une borne de risque différente permettant d'échapper à la minimisation par rapport à T dans les cas favorables.

Proposition 3. *Soit* $T \subset [p]$ *et soit* $\delta > 0$. *Soit*[1] $\nu_T = \inf_{\boldsymbol{u} \in \mathbb{R}^{|T|}} \frac{\sqrt{|T|} \cdot \|\mathbf{X}_T \boldsymbol{u}\|_2}{\sqrt{n}\|\boldsymbol{u}\|_1}$. *Si* $\lambda \geq \sigma^* \rho_T \sqrt{8\log(p/\delta)/n}$, *le Lasso* (2) *satisfait*

$$\ell_n(\hat{\boldsymbol{\beta}}^{\text{Lasso}}_\lambda, \boldsymbol{\beta}^*) \leq 16\rho_T^2 \|\boldsymbol{\beta}^*\|_1^2 + \frac{4\sigma^{*2}(|T| + 2\log(1/\delta))}{n} + \frac{2|T|\lambda^2}{\nu_T^2}$$

avec une probabilité supérieure à $1 - 2\delta$.

Comme les résultats précédents de cette section, la Proposition 3 indique que les corrélations peuvent être exploitées pour adapter le paramètre λ au design $\mathbf{X}$ par le biais de la quantité ρ_T. Mais, contrairement aux résultats précédent, la Proposition 3 montre que les vitesses rapides sont atteintes par le Lasso pour les variables fortement corrélées en utilisant la valeur universelle du paramètre d'ajustement.

[1] Il découle de l'inégalité de Cauchy-Schwarz que ν_T est supérieure ou égal à la plus petite valeur singulière de $\frac{1}{\sqrt{n}}\mathbf{X}_T$.

3.2 Vitesses "rapides"

Le but de cette section est de présenter une version améliorée des inégalités d'oracle parcimonieuses établies par Bickel, Ritov et Tsybakov (2009), voir également Koltchinskii, K. Lounici, and A. Tsybakov (2011) et Sun et Zhang (2012). A cette fin, pour tout $T \subset [p]$, on introduit les poids

$$\omega_j(T, \mathbf{X}) = \frac{1}{\sqrt{n}} \, \|(\mathbf{I}_n - \Pi_T)\boldsymbol{x}^j\|_2. \tag{4}$$

Comme les $\boldsymbol{x}^j$ ont une norme ℓ_2 inférieure ou égale à $\sqrt{n}$, les poids $\omega_j(T, \mathbf{X})$ sont tous entre zéro et un. De plus, ils s'annulent lorsque $\boldsymbol{x}^j$ appartient au sous-espace linéaire engendré par $\{\boldsymbol{x}^\ell, \ell \in T\}$. En particulier, $\omega_j(T, \mathbf{X}) = 0$ pour tout $j \in T$. En utilisant ces poids et tout $\gamma > 0$, on définit l'ensemble

$$\mathcal{C}_0(T, \gamma, \boldsymbol{\omega}) = \Big\{ \boldsymbol{\delta} \in \mathbb{R}^p : \|(\mathbf{1}_p - \gamma^{-1}\boldsymbol{\omega})_{T^c} \odot \boldsymbol{\delta}_{T^c}\|_1 < \|\boldsymbol{\delta}_T\|_1 \Big\}.$$

Définition 1 (Facteur de compatibilité pondéré). Pour tout vecteur $\boldsymbol{\omega} \in \mathbb{R}^p$ dont les éléments sont non-négatifs, on appelle facteur de compatibilité pondéré la quantité

$$\bar{\kappa}_{T,\gamma,\boldsymbol{\omega}} = \inf_{\boldsymbol{\delta} \in \mathcal{C}_0(T,\gamma,\boldsymbol{\omega})} \frac{|T| \cdot \|\mathbf{X}\boldsymbol{\delta}\|_2^2}{n\big\{\|\boldsymbol{\delta}_T\|_1 - \|(\mathbf{1}_p - \gamma^{-1}\boldsymbol{\omega})_{T^c} \odot \boldsymbol{\delta}_{T^c}\|_1\big\}^2}.$$

Le facteur de compatibilité pondéré avec les poids $\boldsymbol{\omega}$ définis dans (**??**) sont particulièrement utiles pour décrire le comportement la performance du Lasso mesurée par la perte de prédiction. Ce facteur permet d'obtenir des vitesses rapides pour le Lasso sous des conditions bien plus faibles que celles utilisées dans les travaux antérieurs.

Théorème 2. *Soit $\delta \in (0, 1)$ un niveau de tolérance donné. Si pour un $\gamma > 1$, le paramètre d'ajustement du Lasso vérifie $\lambda = \gamma\sigma^*\sqrt{2\log(p/\delta)/n}$, alors sur un événement de probabilité au moins $1 - 2\delta$, l'inégalité suivante est vérifiée:*

$$\ell_n(\hat{\boldsymbol{\beta}}^{\text{Lasso}}_\lambda; \boldsymbol{\beta}^*) \leq \inf_{\bar{\boldsymbol{\beta}} \in \mathbb{R}^p, T \subset [p]} \Big\{ \ell_n(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + 4\lambda\|\bar{\boldsymbol{\beta}}_{T^c}\|_1 + \frac{4\sigma^{*2}|T|\log(p/\delta)}{n} \cdot r_{n,p,T} \Big\}, \tag{5}$$

où le terme résiduel est donné par $r_{n,p,T} = \log^{-1}(p/\delta) + 2|T|^{-1} + \gamma^2\bar{\kappa}^{-1}_{T,\gamma,\boldsymbol{\omega}}$.

La différence principale entre l'inégalité (5) et les bornes de risques connues dans la littérature réside dans la présence de ρ_T^2 dans le numérateur du dernier terme. Ce facteur est toujours plus petit que 1. Cependant, afin d'arriver à ce résultat amélioré, nous avons remplacé le facteur de compatibilité usuel par le facteur de compatibilité pondéré $\bar{\kappa}_{T,\gamma,\boldsymbol{\omega}}$ by $\bar{\kappa}_{T,\gamma,\bar{\boldsymbol{\omega}}}$ et avons diminué λ par le facteur ρ_T. Comme prouvé dans Dalalyan, Hebiri et Lederer (2014), cette amélioration permet, entre autres, de démontrer que l'estimateur des moindres carrés pénalisé par la variation totale atteint la vitesse optimale $1/n$ sur la classe des signaux constantes par morceaux.

Bibliographie

[1] P. Bickel, Y. Ritov, and A. Tsybakov. (2009), Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, 37(4):1705–1732.
[2] A. S. Dalalyan and A. B. Tsybakov. (2007), Aggregation by exponential weighting and sharp oracle inequalities. In Learning theory (COLT2007), Lecture Notes in Comput. Sci., Vol. 4539, 97–111.
[3] V. Koltchinskii, K. Lounici, and A. Tsybakov. (2011), Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. Ann. Statist., 39(5):2302–2329.
[4] P. Rigollet and A. Tsybakov. (2011), Exponential Screening and optimal rates of sparse estimation. *Ann. Statist.*, 39(2), 731–771.
[5] T. Sun and C.-H. Zhang. Scaled sparse linear regression. (2012), Biometrika, 99(4):879–898.
[6] R. Tibshirani. (1996), Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1), 267–288.
[7] S. van de Geer and P. Buhlmann. (2009), On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.*, 3, 1360–1392.
[8] A. S. Dalalyan, M. Hebiri, and J. Lederer (2014). On the Prediction Performance of the Lasso. *arXiv* 1402.1700.