TP2 – Régression en grande dimension et sparsité

BARTHE Alexandre – ZULFICAR Eric

22 Décembre2021

Introduction

On suppose le modèle linéaire suivant :

$$Y = X\beta + \eta$$

avec η un vecteur gaussien standard de dimension $n = 1000$, β un vecteur de dimension $p = 5000$ et X une matrice de features de taille $n \times p$.

L'objectif de ce TP est de comparer les méthodes de pénalisation suivantes : *Lasso, Ridge, ElasticNet*. Afin de procéder à cette comparaison nous allons estimer notre modèle à l'aide d'échantillons d'apprentissage contenant deux tiers des données.

Table des matières

1 — Régularisation

Question 1.

On suppose dans cette question que :

— $\beta_1 = ... = \beta_{15} = 1$ et $\beta_i = 0$ pour $i > 15$

— $\forall i,$ les X_i sont i.i.d suivant des gausiennes standards

a. Tracer le chemin de régularisation du Lasso

À l'aide de la méthode *lasso_path* de la classe *linear_model* de *scikit-learn* nous obtenons la figure suivante (Figure 1) :

On constate rapidement que plus le coefficient de régularisation augmente, le poids des coefficients tend vers zéro. Ce qui signifie que l'on est incapable de sélectionner des features avec un α trop grand ($\alpha > 1$).

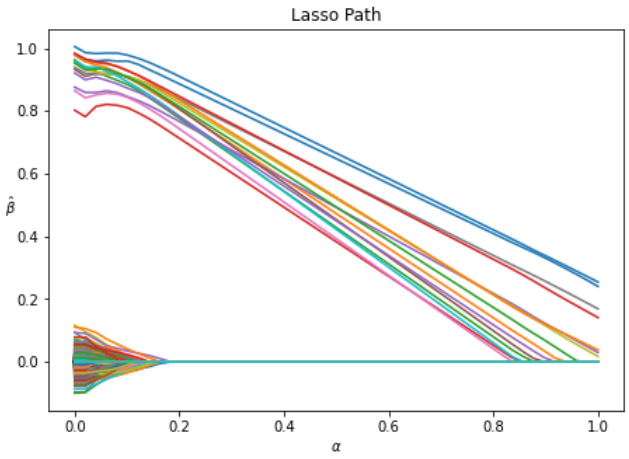

FIGURE 1 – Chemin de régularisation du LASSO

Nous pouvons remarquer que le Lasso se rapproche de notre modèle lorsque α est proche de 0.15. Plus précisément, il alloue des poids très faibles à certains coefficients lorsque α est faible, et si α est supérieur à 0.2 les poids qui étaient originellement proche de 1 tendent vers 0.

b. Déterminer la valeur optimale du paramètre de régularisation pour les trois méthodes sur le jeu d'apprentissage. Quel estimateur fournit la meilleure prédiction sur l'échantillon test ?

Afin d'évaluer le meilleur paramètres, nous avons employé la méthode de validation croisée sur l'échantillon d'apprentissage. Plus précisemment, nous avons utilisé la méthode *RandomSearchCV* de la classe *model_selection* de *scikit-learn*.

Celle-ci permet, sur l'espace des paramètres, de tirer aléatoirement des paramètres et de chercher lequel est le plus performant selon un score. Cela permet, contrairement à une validation croisée classique, de ne pas tester tous les paramètres.

C'est donc un procédé plus rapide. En revanche, nous perdons en précision.

Nous avons la grille de paramètres [1] suivant :

Modèle	l1_ratio	α
Lasso		$(\frac{i}{100})_{i=0,..,50}$
Ridge		$(\frac{i}{100})_{i=0,..,50}$
ElasticNet	$(\frac{i}{100})_{i=0,..,50}$	$(\frac{i}{50})_{i=0,..,50}$

TABLE 1 – grille de paramètres

Nous effectuons notre validation croisée sur cinq échantillons en testant au total 75 paramètres. Les paramètres que nous retenons, selon les modèles sont :

Modèle	l1_ratio	α	score = MSE
Lasso		0.10204	1.31
Ridge		0.5	15.41
ElasticNet	0.48979	0.12244	1.99

TABLE 2 – Meilleurs paramètres pour le cas 1

Sur l'échantillon d'entraînement, le Lasso est le plus performant, à tâtons avec l'ElasticNet, lorsque l'indice de sparsité est très élevé.

C'est également le cas sur l'échantillon test :

	Lasso	Ridge	ElasticNet
MSE	1.3	14.35	1.67

TABLE 3 – Erreur quadratique moyenne sur l'échantillon de test pour le cas 1

Graphiquement (voir Figure 2), ce qui marque le plus est la différence entre le Ridge et les deux autres modèles. On voit que le Ridge est beaucoup plus linéaire mais loin de l'estimation attendue.

1. Nous garderons les notations de scikit-learn pour avoir un tableau plus lisible

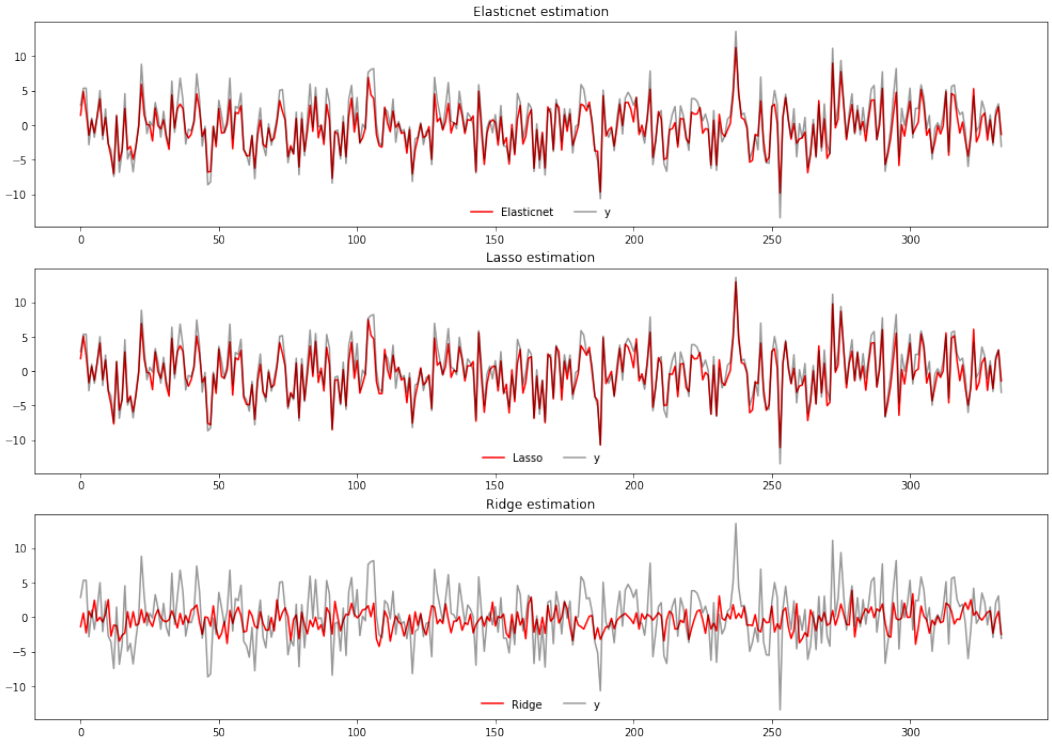

FIGURE 2 – Estimation sur l'échantillon de test pour le cas 1

Question 2.

— $\beta_1 = ... = \beta_{1500} = 1$ $\beta_i = 0$ pour $i > 1500$

— $\forall i,$ les X_i sont i.i.d suivant des gausiennes standards

a. Déterminer la valeur optimale du paramètre de régularisation pour les trois méthodes sur le jeu d'apprentissage. Quel estimateur fournit la meilleure prédiction sur l'échantillon test ?

De la même façon qu'à la question précédente, nous utiliseront la grille de paramètre (Table 1) pour notre *RandomSearchCV*.

Nous effectuons notre validation croisée sur cinq échantillons en testant au total 75 paramètres. Les paramètres que nous retenons, selon les modèles sont :

Modèle	l1_ratio	α	score = MSE
Lasso		0.01020	1556.14
Ridge		0	1264.98
ElasticNet	0	0.061224	1265.34

TABLE 4 – Meilleurs paramètres pour le cas 2

Sur l'échantillon d'entraînement, de nouveau, l'ElasticNet offre de bonne performance, En revanche vu que le l1_ratio est nul, c'est simplement un Ridge avec un coefficient de pénalisation à

0.06. De façon équivalente, notre estimateur Ridge est en réalité une régression linéaire classique.

Un paramètrage plus fin serait donc plus intéressant.

Dans le cas de grande dimension avec un indice de sparsité plus faible, le Lasso performe beaucoup moins bien. Ce qui fut le cas en réintérant l'expérience.
Même conclusion sur l'échantillon test :

	Lasso	Ridge	ElasticNet
MSE	1718.33	1230.51	1230.58

TABLE 5 – Erreur quadratique moyenne sur l'échantillon de test pour le cas 2

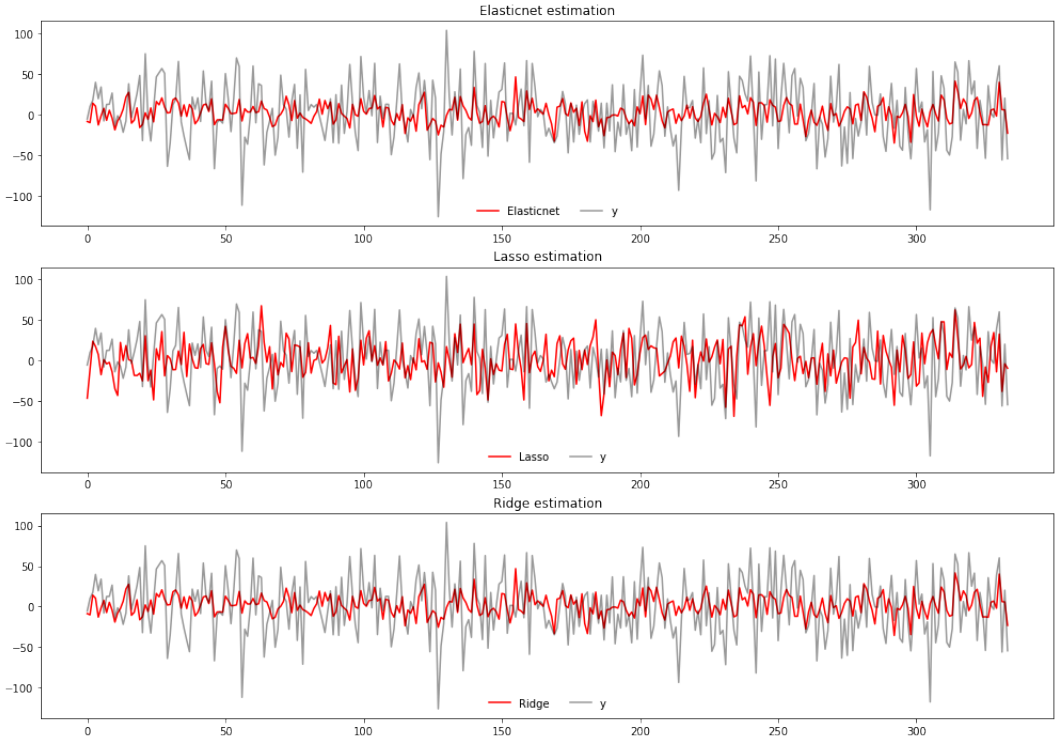

FIGURE 3 – Estimation sur l'échantillon de test pour le cas 2

Visuellement, on peut voir l'effet du Ridge sur notre régression. Il aplati notre prédiction afin de garder une certaine stabilité.

Question 3.

— n = 100, p = 50

— $\beta_1 = \beta_2 = 10$, $\beta_3 = \beta_4 = 5$ $\beta_5 ... = \beta_{14} = 1$ $\beta_i = 0$ pour $i > 14$

— $\forall i,$ les X_i suivent des gausiennes standards dont la matrice de covariance est :

$$\sum = \begin{pmatrix} 1 & 0.7 & \cdots & 0.7^{p-1} \\ 0.7 & 1 & \cdots & 0.7^{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ 0.7^{p-1} & 0.7^{p-2} & \cdots & 1 \end{pmatrix}$$

b. Déterminer la valeur optimale du paramètre de régularisation pour les trois méthodes sur le jeu d'apprentissage. Quel estimateur fournit la meilleure prédiction sur l'échantillon test ?

De nouveau, nous reprendrons la même grille de paramètres (Table 1). Nous effectuons notre validation croisée sur dix échantillons en testant au total 150 paramètres. Les paramètres retenus, selon les modèles, sont :

Modèle	l1_ratio	α	score = MSE
Lasso		0.11224	2.15
Ridge		0.5	4.85
ElasticNet	0.46938	0.02040	3.29

TABLE 6 – Meilleurs paramètres pour le cas 3

Sur l'échantillon d'entraînement, la Lasso est le plus performant. Passer vers une grille de validation classique semblerait plus optimal vue la dimension.

Néanmoins, ce résultat est en contradiction avec les résultats théorique du Lasso, qui est censé avoir plus de difficulté avec des variables corrélées. De plus, au cours des expériences, le Lasso revient à plusieurs reprise comme le plus précis des trois à la fois sur l'échantillon de test et d'entraînement.

Sur cet échantillon de test, les trois estimateurs semblent produire des résultats similaires :

	Lasso	Ridge	ElasticNet
MSE	11.6	13.87	12.39

TABLE 7 – Erreur quadratique moyenne sur l'échantillon de test pour le cas 3

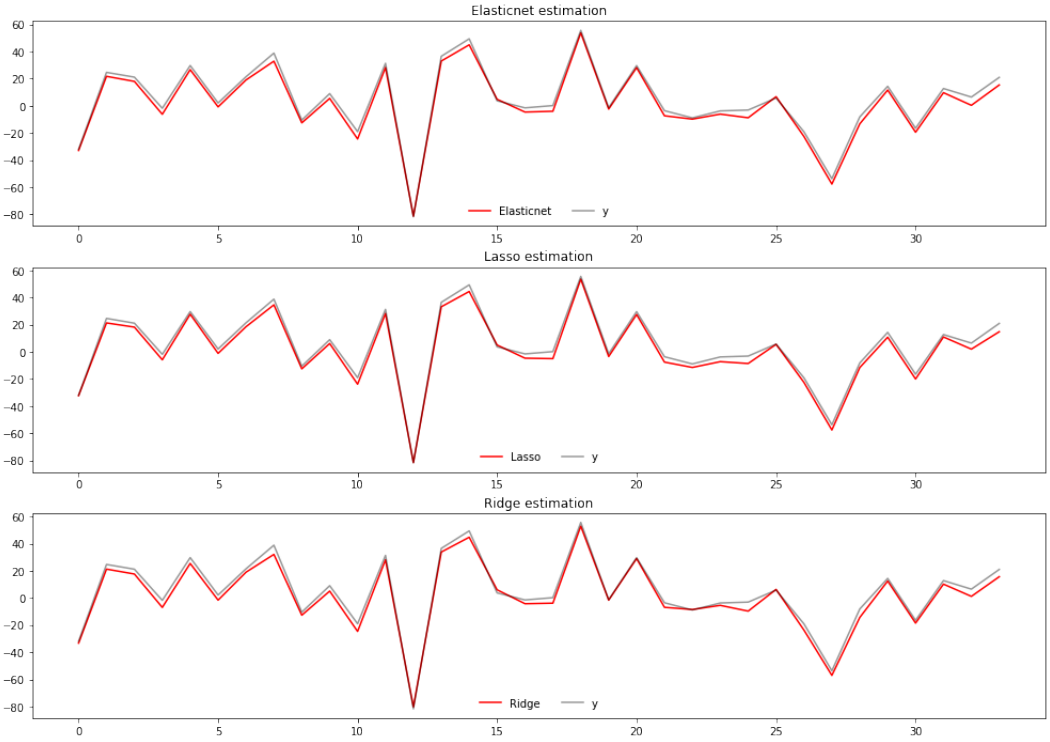

FIGURE 4 – Estimation sur l'échantillon de test pour le cas 3

Sur plusieurs simulation, on voit que le Lasso est l'ElasticNet sont les plus précis.

1 — Conclusion

Ce que l'on peut observer avec nos trois modèles, c'est que selon les cas, le Lasso (cas 1) ou le Ridge (cas 2) sont des choix optimaux. L'elasticNet est quant à lui, toujours dans l'entre deux est un estimateur qui compensent les manques de ces deux estimateurs. En effet, on peut le voir dans chacun des trois cas présentés, la précision de l'ElasticNet est toujours proche de la meilleure obtenue.

2 — Données réelles

Nous avons décidé d'étudier la base de données suivantes : *Gas sensor array under flow modulation* disponible sur *UCI - Machine Learning*.

L'objectif de ces données est de pouvoir détecter le taux d'éthanol et d'acétone de la concoction à l'aide des données de capteurs de gaz dans la fréquence respiratoire. Les capteurs envoient 7500 signals.

On utilisera également l'id du capteur dans notre régression, il est sans doute possible de selon la situation un capteur est meilleur qu'un autre, donc les distinguer pourrait avoir un impact sur la précision.

Les variables cibles sont données par leurs noms respectives : *eth_conc* et *ace_conc*. Leurs valeurs sont comprise entre 0 et 1.

Enfin, on dénote 928 observations dans notre base de données. Nous sommes donc dans un cas où $p >> n$.

De la même façon, nous utiliserons la même grille de paramètre pour une validation croisée sur cinq échantillons en testant 30 paramètres.

Avant entraînement du modèle, nous normalisons les données (également les catégorielles transformé avec du OneHotEncoding). Nous obtenons les résultats suivants :

Modèle	l1_ratio	α	score = MSE
Lasso		0.01020	0.06
Ridge		0.43878	0.05
ElasticNet	0.04082	0.16327	0.06

Table 8 – Meilleurs paramètres pour nos modèles - données Gas sensor

	Lasso	Ridge	ElasticNet
MSE	0.07376	0.048046	0.07094

Table 9 – Erreur quadratique moyenne sur l'échantillon de test - données Gas sensor

Le premier constat que l'on peut faire est que le Ridge est beaucoup plus précis. La pénalisation Lasso reste bonne mais moins performantes. On peut donc supposer qu'il y a des variables corrélées où que le taux de pénalisation n'est pas assez bon pour la sélection de variables. Ou encore, un indice de sparsité faible.

L'ElasticNet rejoint les performance du Lasso à contrario des exemples précédents où il avait des performances similaires au meilleur.

Étant donnée que nous faisons une régression multivariée, il est difficile de représenter graphiquement notre estimation. Nous avons donc afficher les concoctions réels et estimés dans un nuage de points pour le meilleur modèle :

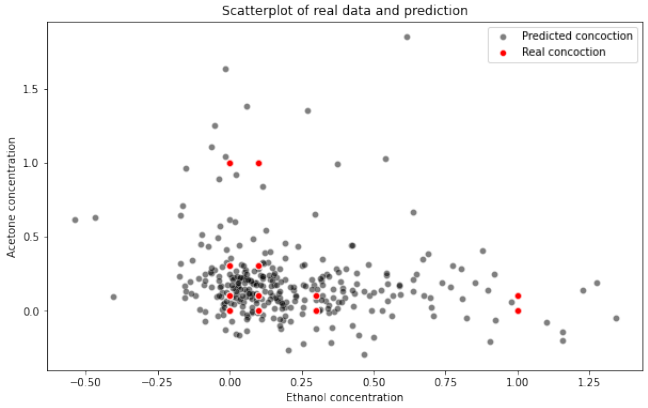

FIGURE 5 – Nuage de point entre concoctions prédites et réelles

Peu de chose en ressort, mais l'on voit que nos estimations, selon nos variables explicatives, sont assez représentatives de la réalité et dénote une fois de plus de la consistance de notre estimateur.