M2 probabilités et statistiques des nouvelles donnés

Régression en grande dimension et sparsité

Djamila AZZOUZ

Professeur référent
Mohamed Hebiri

Année académique 2022 - 2023

I Partie 1 : Régularisation :

I.1 Question 1 :

Dans cette partie,on s'intéresse à la construction d'un modèle de prédiction qui est sous la forme :

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

Comme nous le savons, il ne s'agit pas juste de construire un modèle de prédiction mais plutôt avoir le meilleur modèle de prédiction, autrement dit, choisir les meilleures variables explicatives qui expriment bien la variables quantitatives qu'on souhaite prédire.

Pour cela, on introduit la sparsité qui fait référence à la présence de nombreuses variables explicatives qui ont une valeur nulle ou très faible pour la plupart des observations.

Il existe plusieurs techniques de régression en grande dimension et sparsité qui ont été développées, telles que la régression Lasso, la régression Ridge et la régression Elastic Net. Ces techniques utilisent des contraintes sur les coefficients de régression afin de sélectionner un sous-ensemble de variables explicatives qui sont les plus pertinents pour la prédiction de la variable cible, tout en réduisant les problèmes de surajustement.

- **Question a) : Estimation du vecteur de régression avec la méthode Elastic-Net**
  Avec la fonction Elastic-Net qu'on trouve dans la librairie **"sklearn"**, nous avons estimé le vecteur β, qui sera nul à partir de la 16 ème coordonnées.
- **Question b) : Traçage du chemin de régularisation Lasso**

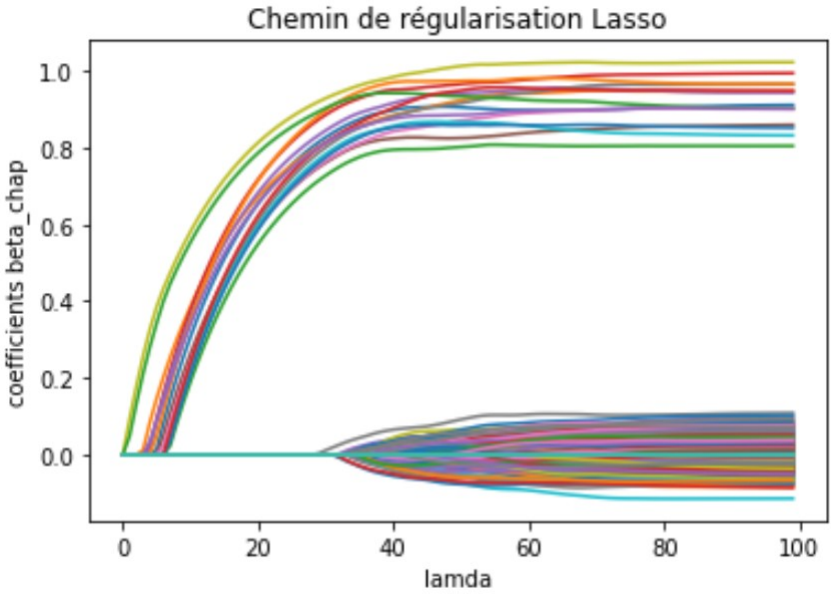


FIGURE 1 – Chemin de régularisation Lasso $(\hat{\beta}_j)$ pour chaque $j \in \{1, ..., p\}$

Observation :

- On remarque que, quand la valeur de lambda est **"faible"**, les coefficients de régression $\hat{\beta}_j$ pour tous $j \in \{1, ..., p\}$ sont élevés et donc le modèle est plus complexe.
- Quand la valeur de lambda est **"plus élevé"**, les coefficients de régression sont plus faibles et le modèle de régression est donc plus simple ce qui peut réduire le risque de surajustement.

- **Question c) : Détermination du paramètre de régularisation optimale pour les trois méthodes**
  A l'aide des fonctions **("Lasso","Ridge","Elastic")**, on construit notre modèle de régression avec chaque estimateur. Et après entraînement des ces modèles sur nos données d'apprentissages, nous obtenons :
  - La valeur optimale du paramètre de régularisation **Lasso** est $\hat{\lambda} = 0.08188060803523127$.
  - La valeur optimale du paramètre de régularisation **Ridge** est $\hat{\mu} = 10.0$

— La valeur optimale du paramètre de régularisation **Elastic-Net** est $\hat{\alpha} = 0.13283177124236364$

— **Question d) : Détermination de l'estimateur qui nous fournit une meilleure prédiction**

Pour savoir quel estimateur fournit une meilleure prédiction sur l'échantillon de test, nous devons d'abord entraîner les modèles en utilisant les données d'entraînement, puis évaluer leur performance sur les données de test en utilisant une mesure de performance appropriée, comme l'erreur quadratique moyenne (**MSE**) et/ ou le coefficient de détermination **(R^2)**.
En utilisant, les données test pour effectuer la prédiction sur chaque modèle construit à l'aide des estimateurs déjà mentionnés, et en choisissant comme critère de performance l'erreur quadratique moyenne le coefficient de détermination, nous obtenons :

— L'erreur quadratique avec le Lasso est : **MSE-Lasso= 1.1639885074539513**.

— Coefficient de détermination avec Lasso est : **R^2-lasso = 0.9320499474531798**.

— L'erreur quadratique avec le Ridge est : **MSE-Ridge= 14.678063585155257**.

— Coefficient de détermination avec Ridge est : **R^2-Ridge = 0.1431400005155763**.

— L'erreur quadratique avec le Ridge est : **MSE-Elastic= 1.5640644722481518**.

— Coefficient de détermination avec Ridge est : **R^2-Elastic = 0.9086947487923707**.

Observation :
D'après les résultats obtenus, on constate que Le meilleur estimateur est le **Lasso**, car le risque quadratique associé est d'une valeur de 1.16 avec un coefficient de détermination très proche de 1, ce qui montre que notre modèle de prédiction est le meilleur avec cet estimateur.

I.2 Question 2 :

Dans cette partie, nous allons refaire les mêmes procédures que nous avons fait dans la première question, mais cette fois ci, on augmente le nombre de coefficients non nuls à estimer dans le modèle, ce qui revient à inclure plus de variables explicatives dans le modèle de prédiction.
Le chemin de régularisation Lasso obtenu est le suivant :

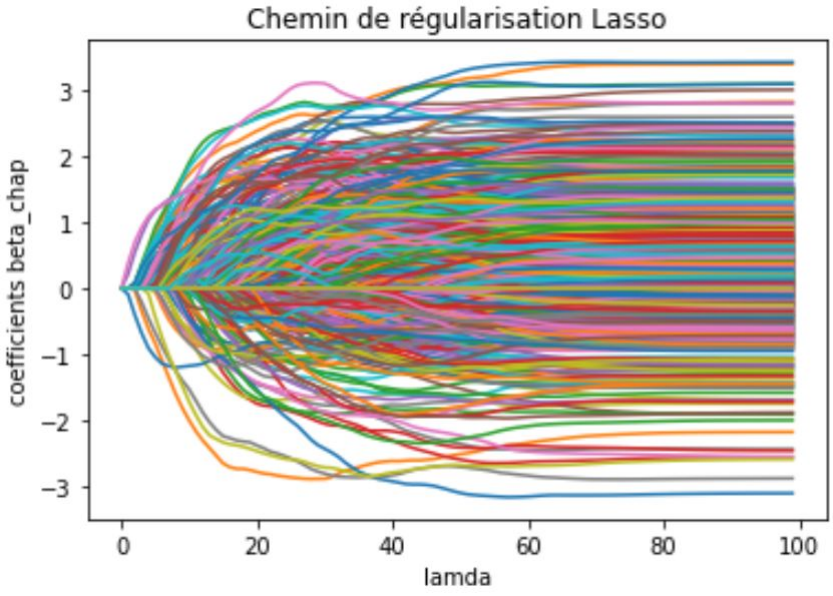


FIGURE 2 – Chemin de régularisation Lasso $(\hat{\beta}_j)$ pour chaque $j \in \{1, ..., p\}$

Observation :
D'après la figure au dessus, on remarque que plus qu'un coefficient $\hat{\beta}_j$ atteint une valeur proche de 0, plus qu'on peut considérer que la variable explicative correspondante n'a pas d'importance pour le modèle et peut être exclue.
Les valeurs optimales des paramètres dans ce cas sont :

— La valeur optimale du paramètre de régularisation **Lasso** est $\hat{\lambda} = 5.801466051534192$.

— La valeur optimale du paramètre de régularisation **Ridge** est $\hat{\mu} = 0.1$

— La valeur optimale du paramètre de régularisation **Elastic-Net** est $\hat{\alpha} = 1.1602932103068384$

Les résultats de performances obtenues sont les suivant :

— L'erreur quadratique avec le Lasso est : **MSE-Lasso= 1511.6851894593613**.

— Coefficient de détermination avec Lasso est : **R^2-lasso = -0.0141674602701376**.

— L'erreur quadratique avec le Ridge est : **MSE-Ridge= 1277.964261304868**.

— Coefficient de détermination avec Ridge est : **R^2-Ridge = 0.1431400005155763**.

— L'erreur quadratique avec le Ridge est : **MSE-Elastic= 1359.270985054973**.

— Coefficient de détermination avec Ridge est : **R^2-Elastic = 0.08808499789224855**.

Observation :
On remarque qu' en augmentant le nombre de variables explicatives inclus dans les modèles de prédictions obtenus pour chaque estimateur sont moins performant. Quoi qu'on remarque que le meilleur dans ce cas est celui avec l'estimateur Ridge car il a une valeur plus petite que Lasso et Elastic.

I.3 Question 3 :

Dans cette partie, nous allons diminuer la taille de notre échantillon à $n = 100$, et nous nous supposerons plus que nos variables explicatives sont indépendantes mais plutôt fortement corrélées. On refait les mêmes procédures que dans les deux précédentes parties et observons ce qui ce passe :
Le chemin de régularisation Lasso obtenu est le suivant :

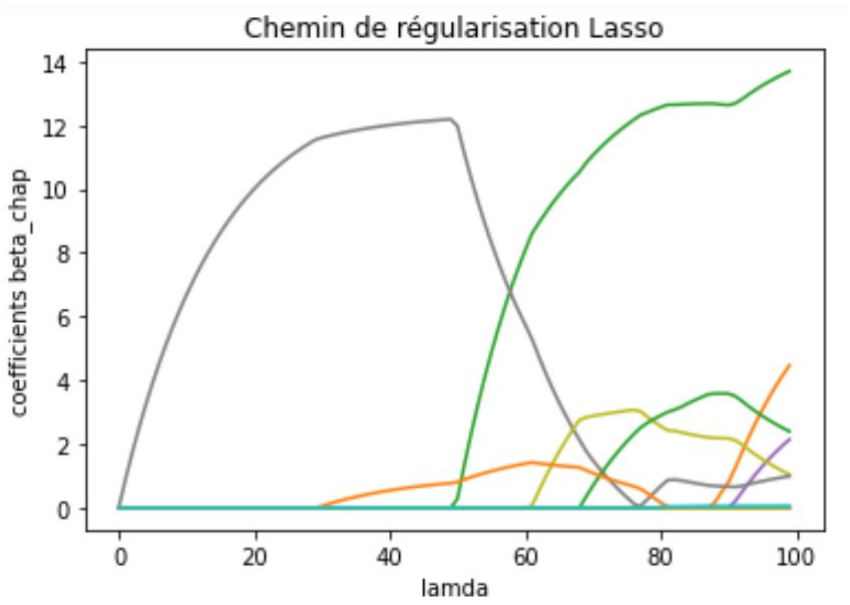


FIGURE 3 – Chemin de régularisation Lasso $(\hat{\beta}_j)$ pour chaque $j \in \{1, ..., p\}$

Les valeurs optimales des paramètres dans ce cas sont :

— La valeur optimale du paramètre de régularisation **Lasso** est $\hat{\lambda} = 1.8971431918052801$.

— La valeur optimale du paramètre de régularisation **Ridge** est $\hat{\mu} = 10.0$

— La valeur optimale du paramètre de régularisation **Elastic-Net** est $\hat{\alpha} = 0.08174542208766214$

Les résultats de performances obtenues sont les suivant :

— L'erreur quadratique avec le Lasso est : **MSE-Lasso= 23.28172514255651**.

— Coefficient de détermination avec Lasso est : **R^2-lasso = 0.8549151980970264**.

— L'erreur quadratique avec le Ridge est : **MSE-Ridge= 6.916175747391917**.

— Coefficient de détermination avec Ridge est : **R^2-Ridge = 0.9569004452164784**.

— L'erreur quadratique avec le Ridge est : **MSE-Elastic= 9.944877976965126**.

— Coefficient de détermination avec Ridge est : **R^2-Elastic = 0.9380264717325499**.

Observation :
On remarque qu'en diminuant l'échantillon avec les variables explicatives qui sont fortement corrélées, le meilleur estimateur est l'estimateur Ridge. En effet, son erreur quadratique est la plus petite, et son coefficient de détermination est plus proche de 1.

Les valeurs optimales des paramètres dans ce cas sont :

II Partie 2 : Application sur des données réelles

Nous allons appliquer ces méthodes d'estimations sur les données ozone que nous avons récupérés sur le site de **"data.gouv"**. Nous souhaitons donc prédire la concentration d'ozone maximale du lendemain (maxo3) en fonction des facteurs météorologiques, tel que la température (T), la nébulosité (Ne), la vitesse du vent(Vx) et la concentration d'ozone de la veille (maxO 3v) sur trois créneaux horaires différents (9h-12h et 15h).
Après la séparation de données en données d'entraînement et test , nous allons utiliser les différents estimateurs déjà mentionnés et entraîner nos données dessus pour pouvoir faire la prédiction et décider lesquels des estimateurs est meilleur dans notre cas.

Les valeurs optimales des paramètres dans ce cas sont :

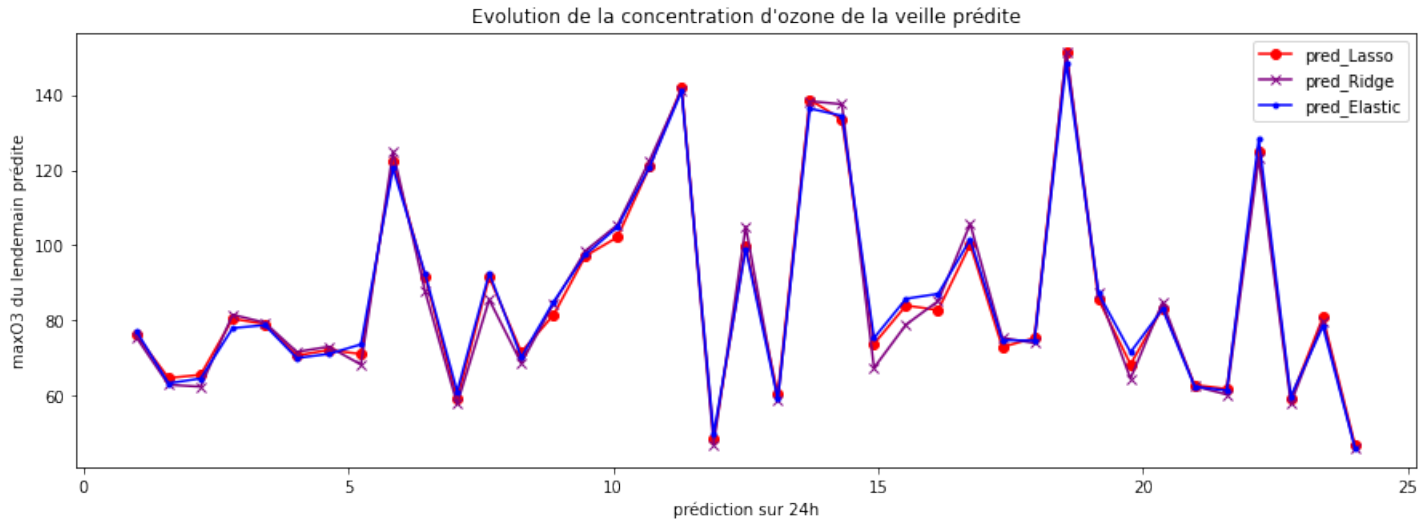
- La valeur optimale du paramètre de régularisation **Lasso** est $\hat{\lambda} = 0.15016870502424345$.
- La valeur optimale du paramètre de régularisation **Ridge** est $\hat{\mu} = 0.1$
- La valeur optimale du paramètre de régularisation **Elastic-Net** est $\hat{\alpha} = 0.032204196033846645$

Les résultats de performances obtenues sont les suivant :

- L'erreur quadratique avec le Lasso est : **MSE-Lasso= 1.018224205482406**.
- Coefficient de détermination avec Lasso est : **R^2-lasso = 0.9964177999059141**.
- L'erreur quadratique avec le Ridge est : **MSE-Ridge= 2.36575659599712**.
- Coefficient de détermination avec Ridge est : **R^2-Ridge = 0.9916770653701458**.
- L'erreur quadratique avec le Ridge est : **MSE-Elastic= 1.869813916574837**.
- Coefficient de détermination avec Ridge est : **R^2-Elastic = 0.9934218342563323**.

Observation :
On remarque que le meilleur estimateur pour avoir une meilleure prédiction de la concentration d'ozone du lendemain est l'estimateur**"LAsso"**. En effet, son erreur quadratique moyenne est la plus petite que les deux autres et même son coefficient de détermination est très proche de 1.

La figure ci- dessus montre les prédictions obtenues pour la concentration d'ozone du lendemain avec les trois estimateurs utilisés :



On remarque que les valeurs prédites avec les trois estimateurs sont très proches, mais en choisissant comme critère de performance l'erreur quadratique moyenne, on en déduit que l'estimateur lasso est celui qui nous fournit

une meilleure prédiction.
En observant le graphe au dessus, en constate que la concentration d'ozone atteint son pic entre 16h-20h ce qui signifie que le taux de pollution dans l'air est très haut, contrairement à sa concentration qui est très base entre 12-15h.

Conclusion : En conclusion, l'estimateur Ridge est préférable dans le cas où le nombre de variables explicatives est élevé ou la taille de l'échantillon est réduite et il y a une forte corrélation entre les variables. D'autre part, l'estimateur Lasso est plus adapté dans le cas où la dimension de l'échantillon est très grande. Il est important de prendre en compte ces considérations lors de la sélection de l'estimateur de régression pour obtenir les meilleures performances de prédiction possibles avec les meilleurs variables explicatives qui ont une grande influence sur notre variable qu'on souhaite prédire.