ProjetFleuranceMoran

Gabriel Moran et Paul Fleurance

9 décembre 2019

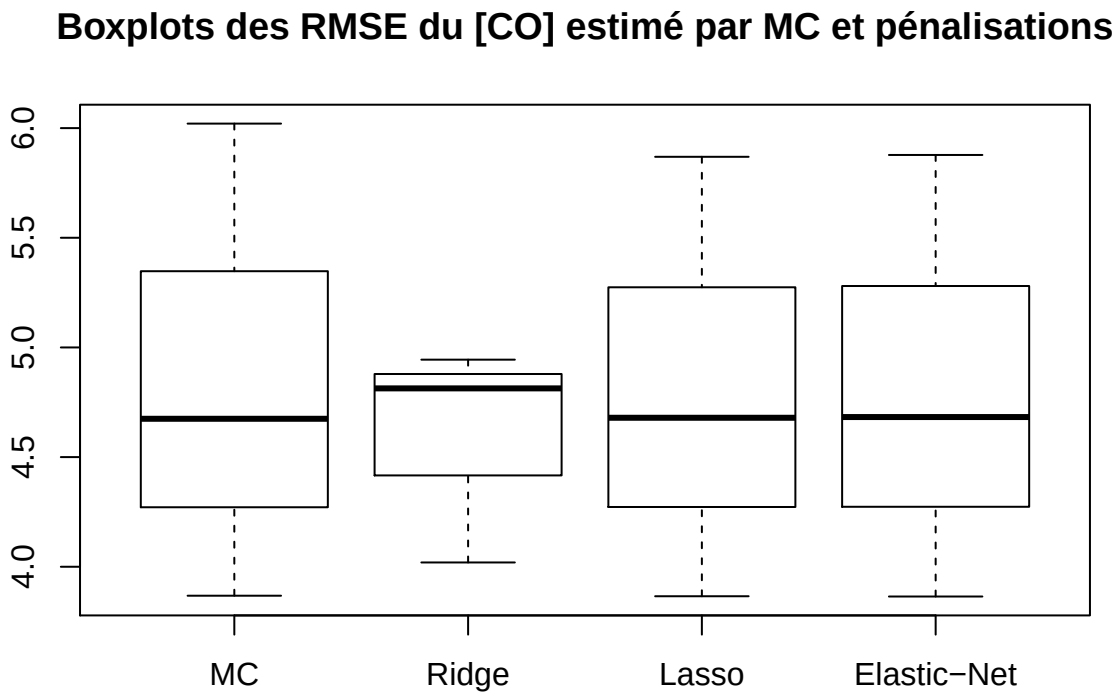
Préambule

Pour rappel, nous nous intéressons à la construction d'un modèle permettant de prédire la concentration de CO à partir des valeurs des capteurs et autres interférences potentielles. Pour une présentation plus ample des données et du problème envisagé, se référer au document précédent.

Premières approches de construction de modèle

Comme expliqué dans le document précédent, les variables explicatives sont fortement corrélées. Cela pose problème puisque, en notant X la concaténation des p variables explicatives, une des hypothèses fondamentales en régression linéaire est l'inversibilité de X^TX. Or, lorsque des variables sont fortement corrélées, X^TX se rapproche d'une matrice inversible, ce qui fait diminuer la précision des estimateurs des coefficients du modèle. Nous avons donc envisagé d'utiliser les méthode de pénalisations (Lasso, Ridge et Elastic-Net). Cependant, nous avons tout de même commencer par utiliser la méthode des moindres carrés (MC) pour ensuite tenter une sélection de variables par recherche pas à pas (méthodes "stepwise","backward" et "forward").

Le modèle construit produit des erreurs de l'ordre de grandeur de la variable cible (CO) et les trois méthodes de sélection de variables sélectionnent l'ensemble des variables, ce qui n'apporte aucune valeur ajoutée. Nous avons donc ensuite utiliser les méthodes de pénalisations (avec un alpha=0.5 pour elastic-net). Or, une validation croisée sur chacun des modèles construits produits des erreurs du même ordre de grandeur que le modèle des moindres carrés, malgrès la forte corrélation des variables explicatives (RMSE est la racine de l'erreur quadratique moyenne):



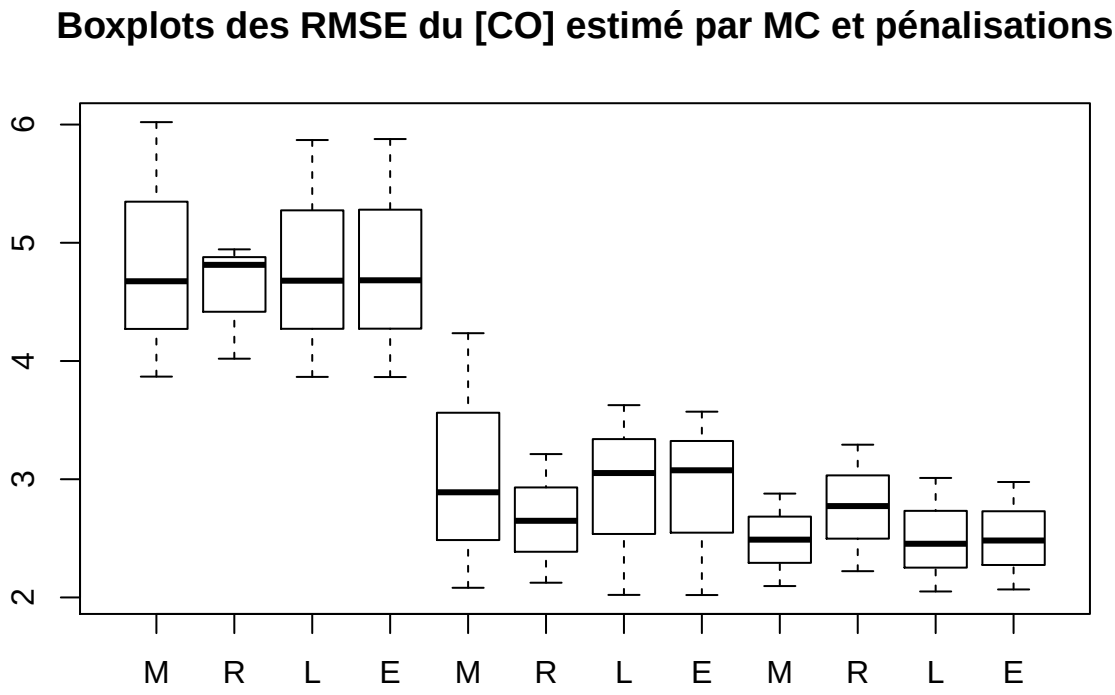
Par ailleurs, la méthode Lasso, qui favorise l'annulation de certains coefficients, ne fait aucune sélection de variables. Arrivés là, nous nous sommes rendu compte qu'il fallait sérieusement repenser notre modèle, notamment par des pré-traitements de nos données.

Autres approches par pré-traitements

Une des caractéristiques principales de nos données est la non-linéarité des signaux des capteurs. Une autre est la forte redondance des données de certaines variables : la température et le flux ne changent que tous les 5s, CO toute les 15 min, la tension du radiateur toute les 25s, et ce pour environ 3 instances par seconde. L'extraction de caractéristiques permettrait de répondre à ces problèmes. Tout d'abord, le radiateur ne sert qu'à déclencher les signaux des capteurs et sa tension n'apporte aucune information. Ensuite, en ce qui concerne les capteurs, l'amplitude des signaux produits résume entièrement l'information utile de ceux-ci. Si nous extrayons les amplitudes des signaux, nous aurons non seulement des données plus interprétables, mais aussi une forte réduction de redondance. Une de nos approches à donc été de fractionner les données par signal, en extraire l'amplitude et de moyenner les autres variables, en éliminant la tension du radiateur.

Une première visualisation de la matrice de corrélation, nous montre une forte réduction de corrélation entre capteurs de différents modèles, ce qui est rassurant, et accentue la corrélation entre les capteurs FIS et [CO] d'une part et Figaro et l'humidité d'autre part. La sélection de variables par recherche pas à pas nous donne pour les trois méthodes un nouveau modèle, constitué de la température, du flux et certains capteurs issus majoritairement du modèle FIS.

Par validation croisée, on observe clairement la supériorité du pouvoir prédictif des modèles construits (par MC(M), Ridge(R), Lasso(L) et Elastic-Net(E)) à partir des données transformées (4 du milieu), et encore plus pour les modèles construits uniquement à partir des variables sélectionnées précédemment (4 derniers) :



Il semblerait que le modèle construit par MC pour les données transformées sur variables sélectionnées ait le meilleur pouvoir prédictif. Nous nous pencherons dès lors sur des possibilités d'autres pré-traitements ainsi que d'autres méthodes de régression en la présence de variables non-linéaires, notamment le KNN ou le groupe-Lasso.