

Analyse et Modélisation Multivariées*TP n°1: Régressions régularisées*

On pourra utiliser pour ce TP différents packages, parmi lesquels: ade4 , ClustOfVar , pls , glmnet ... et possiblement, par surcroît: FactoMineR (ou FactoShiny) et factoextra (pour de plus jolis graphiques).

Commandes utiles du package R: ade4

Pour recoder un tableau de variables qualitatives en indicatrices:

`acm.disjonctif(dataframe)`

Commandes utiles du package R: pls

Régression pls: `resultat = plsr(arguments)`

Exploration de l'objet résultat: `names(resultat)`

Contenus intéressants du résultat:

- coefficients = coefficients reconstitués des régresseurs dans le modèle de la variable dépendante, pour chacun des modèles, de 1 composante au nombre de composantes spécifié.

- loadings = coefficients des régresseurs dans les formules des composantes.

- scores = valeurs des composantes pour les individus.

- validation = paramètres et résultats de la validation (croisée ici). on peut les lister via la commande `names(resultat$validation)`: PRESS (predicted sum of squares) contient l'erreur quadratique totale de prédiction.

On peut en faire un graphique avec `barplot(resultat$validation$PRESS)`

Tracé des cercles de corrélations: `corrplot(resultat, identify=TRUE)`

Liens potentiellement utiles pour la validation croisée en R en général:

<https://www.geeksforgeeks.org/cross-validation-in-r-programming/>

<https://www.statology.org/how-to-perform-cross-validation-for-model-performance-in-r/>

<https://www.r-bloggers.com/2021/10/cross-validation-in-r-with-example/>

Données à utiliser:

Ce sont les données d'abondance, sur 1000 parcelles du bassin du fleuve Congo, de 27 espèces d'arbres (projet CoForTips), figurant dans le fichier *genus.csv* disponible sur le site des TP. Les abondances sont notées : *gen1* à *gen27*. Par ailleurs, on dispose de variables géographiques quantitatives (*latitude*, *longitude*, *altitude*, pluviométries annuelle et mensuelles), qualitatives (*geology* = type de sol), et de variables de photosynthèse (tous les indices EVI). Toutes ces variables sont mesurées sur des parcelles forestières dont la *surface* figure dans la dernière colonne.

Attention: la variable qualitative *forest* (type forestier) ne doit pas être utilisée ici.

L'objectif de votre travail sera de modéliser au mieux la densité *globale* de peuplement arboré par ces 27 espèces à l'aide des autres variables. Pour ce faire, nous considérerons successivement plusieurs techniques, dont nous comparerons les résultats.

Questions

1. Calculs de variables:

a) Calcul de la variable dépendante Y :

Sommer les abondances des 27 espèces et diviser cette somme par la surface de chaque parcelle. On obtient ainsi la densité de peuplement arboré sur la parcelle. On appellera cette variable Y : *treedensity*.

b) Calcul du tableau des variables explicatives:

Calculer les carrés des variables géographiques quantitatives. Formez les indicatrices de la variable *geology*. L'ensemble des variables explicatives est constitué des variables géographiques quantitatives, de leurs carrés, des indicatrices de la variable *geology* et des EVI.

Faites l'inventaire théorique de toutes les multi-colinéarités (combinaisons linéaires de variables qui s'annulent) présentées par l'ensemble des variables explicatives.

2. Première régression sur composantes principales:

a) Réaliser une ACP réduite globale des variables explicatives. Retenir les composantes qui ne sont pas du bruit.

b) Modéliser la densité à partir des composantes retenues. Vous pourrez éliminer les composantes qui n'ont pas de rôle statistiquement significatif (justifiez la validité des tests de Student ici). Donner le R^2 du modèle obtenu. Construire le graphe d'abscisse Y et d'ordonnée \hat{Y} , et commenter.

c) Retrouver les coefficients des variables originelles dans le prédicteur linéaire \hat{Y} .

d) Si besoin est, corriger la linéarité de la liaison en utilisant une transformation de type Log sur Y qui laisse les zéros invariants.

3. Seconde régression sur composantes principales:

Ici, on partitionne les variables explicatives en deux thèmes: *Photosynthèse* (les EVI) et *Géographie* (tout le reste).

Reprendre les questions 2-a à 2-d, mais en utilisant la *réunion* des composantes principales issues d'ACP *séparées* des thèmes *Géographie* et *Photosynthèse*. N.B. Si vous avez dû transformer Y (question 2-d), il est fort probable qu'il faille ici user de la même transformation.

4. Régression PLS:

a) Utiliser la régression PLS pour modéliser au mieux la densité (ou sa transformation). Vous utiliserez la validation croisée (de type Leave K out ou K-fold) pour déterminer le meilleur nombre de composantes.

b) Tenter au mieux d'interpréter les composantes retenues, et les plans qu'elles engendrent, exactement comme pour une ACP réduite.

c) Retrouver les coefficients des variables (et interactions) originelles dans le prédicteur linéaire \hat{Y} .

d) Comparer l'ajustement et les coefficients du modèle obtenu avec ceux des régressions sur composantes principales (on prêterait notamment attention aux signes de ces coefficients).

5. Régressions pénalisées:

a) Utiliser la régression ridge (package *glmnet*) pour modéliser au mieux la densité (ou sa transformation). Vous déterminerez le poids optimal de la pénalité par validation croisée.

b) Utiliser la régression LASSO (package *glmnet*) pour modéliser au mieux la densité (ou sa transformation). Vous déterminerez le poids optimal de la pénalité par validation croisée.

6. Synthèse & conclusions

a) Comparer la qualité de prédiction et les coefficients des modèles obtenus entre les différentes méthodes.

Vous pourrez produire un tableau synoptique de ces coefficients selon les méthodes, afin d'en faciliter

la comparaison:

<i>Coefficients</i> Variables	Méthode:	RCP globale	RCP thèmes	PLS	Ridge	LASSO
altitude						
longitude						
latitude						
...						
...						
...						
...						
	R²					
	PRESS					

b) Quels sont les avantages et inconvénients que vous avez pu constater des différentes méthodes?
Quel modèle retenez-vous en définitive?