SCGLR : un package R pour la régression linéaire généralisée sur composantes supervisées.

C. Trottier[a,c], **G. Cornu**[b], **F. Mortier**[b] and **X. Bry**[a]

[a]Institut de Mathématiques et de Modélisation de Montpellier
Université Montpellier II
Place Eugène Bataillon CC 051 - 34095, Montpellier, France
xavier.bry@univ-montp2.fr
catherine.trottier@univ-montp2.fr

[b]UPR Biens et Services des Ecosystèmes Forestiers tropicaux
CIRAD
Campus International de Baillarguet, TA C-105/D 34398 Montpellier Cedex 5, France
frederic.mortier@cirad.fr
guillaume.cornu@cirad.fr

[c]Département MIAP
Université Montpellier III
Route de Mende, 34199 Montpellier Cedex 5, France
catherine.trottier@univ-montp3.fr

Mots clefs : Modèles linéaires généralisés multivariés, Modèles à composantes, Régression PLS, Algorithme des scores de Fisher.

Nous présentons un nouveau package *R*, nommé **SCGLR**, qui implémente une nouvelle méthode de régression sur composantes dans les modèles linéaires généralisés multivariés. L'enjeu statistique de cette méthode concerne la modélisation simultanée de plusieurs variables, dont les distributions appartiennent à la famille exponentielle et sont possiblement différentes, à l'aide de nombreuses variables explicatives le plus souvent redondantes, ce qui implique une régularisation de la régression.

L'approche classique du modèle linéaire généralisé se heurte à différentes limites : 1) elle ne permet pas la modélisation conjointe de plusieurs variables, 2) elle ne permet pas la prise en compte de nombreux régresseurs corrélés, 3) elle ne permet pas d'explorer l'espace des régresseurs en extrayant progressivement leur pouvoir prédictif.

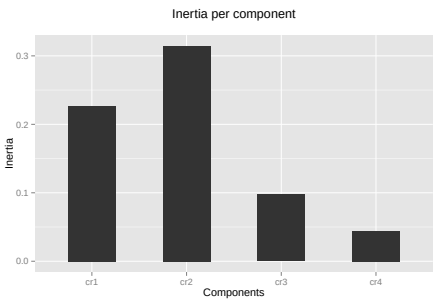
Notre méthode, SCGLR, procède d'une approche de type PLS pour laquelle l'information pertinente contenue dans les données est résumée par quelques composantes qui prédisent le mieux possible les variables à expliquer. Elle réalise ainsi un compromis entre la recherche des structures fortes dans l'espace des variables explicatives et la qualité d'ajustement du modèle. SCGLR est décrite dans [1] et résolue par un algorithme qui couple une technique de linéarisation, bien connue dans l'algorithme des scores de Fisher, avec un calcul des composantes de type PLS multivariée.

La méthode et le package sont illustrés par une application en écologie forestière (`data(genus)` associé au package *R*) où l'on s'intéresse à comprendre comment des communautés d'arbres se structurent en fonction de caractéristiques environnementales. Les variables à expliquer sont

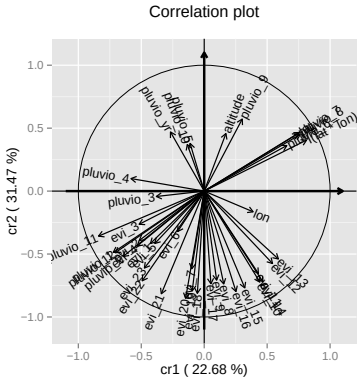
les abondances de 27 espèces observées en forêt tropicale humide du bassin du Congo. Ces abondances de chaque espèce se traduisent par des données de comptage. 40 variables environnementales géo-référencées constituent les variables explicatives. Elles mesurent autant des éléments physiques que des caractéristiques de la végétation. Ces informations ont été relevées sur un total de 1000 parcelles.

Le package **SCGLR** utilise une version ≥ 3.0 de *R*. Il contient un ensemble de fonctions dont les deux principales sont `scglr()` et `scglrCrossval()`. La fonction `scglr()` a pour objectif la construction des composantes d'une part, et l'estimation des paramètres de régression sur les composantes elles-mêmes, ou par leur intermédiaire sur les variables explicatives d'origine d'autre part. On obtient aussi les pourcentages et pourcentages cumulés de variance totale des régresseurs capturés par les composantes, ainsi que les déviances des modèles GLM de chaque variable à expliquer régressée sur les composantes.

Les méthodes `print`, `summary` et `plot` peuvent être appliquées à une sortie de la fonction `scglr()` (objet de classe *scglr*). Ainsi sur les données forestières pour 4 composantes :

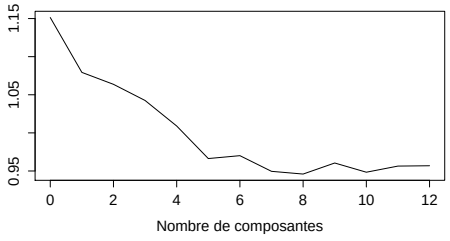


(a) Inertie par composante



(b) Correlation composantes 1,2

La fonction `scglrCrossval()` quant à elle, a pour but de sélectionner le nombre de composantes nécessaires par cross-validation.



Références

[1] Bry X, Trottier C, Verron T, Mortier F (2013). Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm. *Journal of Multivariate Analysis*, **119**(0), 47-60

[2] Marx, B. D. (1996). Iteratively Reweighted Partial Least Squares estimation for Generalized Linear Regression. *Technometrics*, **38**(4), 374-381