Modélisation linéaire généralisée sparse sur composantes supervisées avec interactions.

X. Bry, C. Trottier

Depuis 2013, nous avons développé, pour les données de grande dimension, une méthodologie de modélisation linéaire généralisée multivariée (i.e. à réponses multiples) fondée sur des composantes supervisées : SCGLR (Supervised Component-based Generalised Linear Regression). SCGLR s'inscrit dans la filiation de la régression des moindres carrés partiels (PLS regression). Comme la régression PLS, la méthode est fondée sur la maximisation d'un critère combinant la qualité d'ajustement du modèle et la force des composantes explicatives qui permettent de réduire la dimension du sous-espace explicatif en en excluant le bruit.

SCGLR étend la régression PLS de multiples manières. Tout d'abord, en utilisant la vraisemblance du modèle pour mesurer sa qualité d'ajustement, elle permet l'extension aux modèles linéaires généralisés avec réponses multiples de types différents (binaires, poissonniennes, gaussiennes etc.). Ensuite, la mesure de force des composantes est étendue au-delà de leur variance, incluant un sous-critère directionnel supplémentaire. Enfin, elle prend en compte un partitionnement thématique des variables explicatives, permettant d'exploiter la complémentarité des groupes de variables explicatives en fournissant des composantes plus faciles à interpréter sur le plan conceptuel. La méthode s'est avérée un auxiliaire puissant de modélisation dans le domaine de l'écologie forestière, permettant de modéliser le peuplement arboré de la forêt équatoriale du bassin du Congo.

Des extensions de SCGLR ont déjà fait l'objet de deux thèses de doctorat. La première a consisté à introduire des effets aléatoires dans les modèles afin de prendre en compte la non-indépendance des observations. La seconde a consisté à introduire dans le modèle, outre les composantes, des variables aléatoires latentes, afin de modéliser la dépendance entre réponses.

Toutefois, un aspect essentiel des modèles écologiques n'est pas encore pris en compte : les effets d'interaction entre dimensions explicatives. Par exemple, le peuplement d'un site par une espèce résultant de l'adaptation de l'espèce au site, la modélisation de ce peuplement devrait faire intervenir les interactions entre les variables biologiques décrivant l'espèce et les variables environnementales caractérisant le site. Mais ces variables étant très nombreuses, le nombre de leurs interactions est *a priori* rédhibitoire, ce qui rend indispensable une sélection des variables et des dimensions dont les interactions sont utiles au cours du processus d'estimation du modèle. L'inclusion des interactions d'ordre quelconque (i.e. entre un nombre quelconque de groupes de variables) permet la modélisation de données tensorielles (Multiway Modeling). Par exemple, l'abondance de p espèces étant mesurée sur q sites à r dates, on dispose d'un "parallélépipède" de données à modéliser.

Nous proposons donc, dans le cadre de cette thèse, d'étendre SCGLR à la recherche et modélisation des interactions entre groupes de variables explicatives grâce à plusieurs mécanismes d'identification et quantification de ces interactions. En particulier, il s'agira de chercher dans les différents groupes de variables explicatives les dimensions dont les interactions jouent un rôle important. Cette recherche pose un problème dimensionnel aigu. Contenir l'explosion de la dimension est doublement nécessaire, car outre la gestion numérique de l'estimation, il est indispensable d'aboutir à une modélisation interprétable. Aussi cherchera-t-on à pénaliser le critère à optimiser de sorte à obtenir une sparsité adéquate du modèle à interactions. Des outils computationnels devront être développés pour atteindre ces objectifs, ainsi que des outils d'inférence statistique permettant la validation des composantes et effets estimés.

De solides compétences en programmation informatique (si possible sur un noeud de calcul) sont impératives de la part du / de la candidat/e.

La thèse sera encadrée par Xavier Bry (directeur) et Catherine Trottier (co-directrice).

Mots-clés: Modèles Linéaires Généralisés, SCGLR, Composantes, Interactions, Sparsité, Multiway.

Localisation: IMAG - Université de Montpellier, Campus Triolet

Contact: xavier.bry@umontpellier.fr, *cc* catherine-trottier@umontpellier.fr

Références bibliographiques

1. Bry X., Trottier C., Mortier F. and Cornu G. (2018): *Component-based regularisation of a multivariate GLM with a thematic partioning of the explanatory variables*. Statistical Modelling.
2. Bry X., Trottier C., Verron T. and Mortier F. (2013): *Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm*. Journal of Multivariate Analysis.
3. Cornu G., Mortier F., Trottier C. and Bry X. (2018): *SCGLR: Supervised Component Generalized Linear Regression. R package version 3.0*. https://CRAN.R-project.org/package=SCGLR
4. Bry X. (2004) : *Estimation empirique d'un modèle à variables latentes comportant des interactions* – RSA vol. 52 n°3, 2004
5. Bry X., Verron T. (2008) : *Modélisation factorielle des interactions entre deux ensembles d'observations: la méthode FILM (Factor Interaction Linear Modelling)*, Journal de la SFdS/RSA (Paris) 149 - N°2.
6. Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations* (1st ed.). Chapman and Hall, https://doi.org/10.1201/b18401
7. Smilde A. K., Westerhuis, J. A., Boqué R. (2000) : *Multiway multiblock component and covariates regression models*; Journal of Chemometrics, 14: 301–331.
8. Azcarate S., Gomes A., Muñoz A., Goicoechea H. (2024) : *Recent advances in multiway data modeling for classification issues* , in *Data Handling in Science and Technology*, Volume 33, Chapter 9, Pages 193-218. DOI https://doi.org/10.1016/B978-0-443-13261-2.00024-2.