Extension de la régression linéaire généralisée sur composantes supervisées à la modélisation jointe des réponses

Julien GIBAUD[1], Xavier BRY[1] et Catherine TROTTIER[1,2]

[1] Institut Montpelliérain Alexander Grothendieck, CNRS, Univ. Montpellier, France.
[2] Univ. Paul-Valéry Montpellier 3, F34000, Montpellier, France.

Contact : julien.gibaud@umontpellier.fr, xavier.bry@umontpellier.fr et catherine.trottier@univ-montp3.fr.

Résumé

Dans ce travail, nous proposons d'étendre la méthode SCGLR, pour la rendre capable d'identifier des groupes de réponses expliquées par des composantes communes. À l'origine, SCGLR vise la construction de composantes explicatives dans un grand nombre de covariables, éventuellement fortement redondantes. Ces composantes sont supervisées conjointement par l'ensemble des réponses. Désormais, nous cherchons à identifier des groupes de réponses partageant les mêmes dimensions explicatives. Dans un cadre écologique par exemple, des communautés d'espèces devraient pouvoir être modélisées par des composantes propres à chaque communauté. Un algorithme est proposé afin d'estimer le modèle.

Mots clefs : SCGLR, mélange de réponses, algorithme EM, classification

Abstract

In this work, we propose to extend the SCGLR methodology, enabling it to identify clusters of responses sharing explanatory components. Originally, SCGLR was designed to find explanatory components in a large set of possibly highly redundant covariates, something much needed in a high-dimensional framework. These components are jointly supervised by all the responses. Henceforth, we aim at identifying clusters of responses sharing the same explanatory dimensions. In an ecological framework for instance, communities of species should be modeled by components which are characteristic of each community. An algorithm is proposed in order to estimate the model.

Keywords : SCGLR, response mixture, EM algorithm, clustering

1 Contexte

Les changements climatiques entraînent certains dérèglements des écosystèmes pouvant causer des extinctions d'espèces animales ou végétales. Dans ce contexte, le développement de modèles permettant de prédire le futur de la biodiversité est devenu un enjeu crucial. Récemment, de nombreuses avancées ont été faites dans ce domaine, en particulier par l'extension des modèles de distribution des espèces (Species Distribution Models, SDM), qui traitent les espèces séparément, à des modèles de distribution jointe (Joint Species Distribution Models, JSDM). Les JSDM permettent de formaliser l'interdépendance des espèces et de mieux comprendre son impact sur la composition des communautés. Par ailleurs, modéliser l'abondance des espèces requiert de prendre en compte un grand nombre de covariables explicatives souvent corrélées, ce qui impose une réduction de

dimension et la régularisation des modèles.

Dans leur article, Bry et *al.* [**?**] proposent une méthode - la régression linéaire généralisée sur composantes supervisées (Supervised Component-based Generalized Linear Regression, SCGLR) - combinant le modèle linéaire généralisé multivarié avec les méthodes à composantes permettant la réduction de dimension. SCGLR optimise un critère compromis entre la qualité d'ajustement (Goodness-of-Fit, GoF) et la proximité à des dimensions d'intérêt (Structural Relevance, SR) [**?**]. Cette technique ne trouve pas seulement des directions fortes et interprétables, elle produit aussi des prédicteurs régularisés, ce qui permet le traitement de données de grande dimension. Cependant, SCGLR suppose que l'ensemble des réponses dépend des mêmes dimensions explicatives. Pour nous affranchir de cette hypothèse, nous proposons d'étendre cette méthode aux mélanges sur les réponses. L'objectif est de trouver des classes de réponses (espèces) telles que toutes les réponses d'une classe soient modélisables par les mêmes dimensions explicatives.

2 Modélisation

Dans cette section, nous présentons la méthode SCGLR, puis son extension aux mélanges sur les réponses.

2.1 SCGLR

n individus sont décrits par K réponses y_k, $k = 1, \ldots, K$, ainsi que des covariables explicatives séparées en deux groupes : un groupe X de covariables *a priori* nombreuses et possiblement redondantes, et un autre A de covariables additionnelles peu nombreuses et faiblement, voire non-redondantes. On notera X et A les matrices correspondantes. Chaque réponse y_k fait l'objet d'un modèle linéaire généralisé (Generalized Linear Model, GLM) [**?**]. Pour la partie explicative du modèle, seule la matrice X requiert réduction de dimension et régularisation. À cette fin, SCGLR cherche dans X des composantes communes à l'ensemble des réponses. Une composante $f \in \mathbb{R}^n$ est donnée par $f = Xu$ où $u \in \mathbb{R}^p$ est un vecteur de coefficients. Le prédicteur linéaire associé à la réponse y_k est donné par

$$\eta_k = (Xu)\,\gamma_k + A\delta_k,$$

où γ_k et δ_k sont les paramètres de régression. La composante f est commune à l'ensemble des réponses y_k et pour assurer l'identifiabilité, nous imposons $u^T M^{-1} u = 1$, où $M \in \mathbb{R}^{p\times p}$ est une matrice symétrique définie positive. Nous supposons que les réponses sont indépendantes conditionnellement aux variables explicatives.

À cause du produit $u\gamma_k$, le modèle "linéarisé" à chaque étape de l'algorithme des scores de Fisher (Fisher Scoring Algorithm, FSA) pour l'estimation du GLM, n'est pas linéaire et doit être estimé de façon alternée sur u et sur $\{\gamma_k, \delta_k\}$. Soient w_k, la pseudo-réponse (ou variable de travail) associée à chaque étape du FSA, et W_k^{-1} sa matrice de variance-covariance. L'estimateur des moindres carrés de u est solution des programmes équivalents suivants :

$$\min_{u,u^T M^{-1}u=1} \sum_{k=1}^{K} \left\| w_k - \Pi^{W_k}_{\mathrm{vect}(f,A)} w_k \right\|^2_{W_k} \Leftrightarrow \max_{u,u^T M^{-1}u=1} \sum_{k=1}^{K} \left\| \Pi^{W_k}_{\mathrm{vect}(f,A)} w_k \right\|^2_{W_k} \Leftrightarrow \max_{u,u^T M^{-1}u=1} \psi_A(u),$$

avec $\psi_A(u) = \sum_{k=1}^{K} \|w_k\|^2_{W_k} \cos^2_{W_k}\left(w_k, \Pi^{W_k}_{\mathrm{vect}(f,A)} w_k\right)$. La quantité ψ_A est une mesure de GoF. Pour trouver des composantes fortes et interprétables, le GoF ne suffit pas. Il faut le combiner avec une

mesure de pertinence structurelle.

Soient un ensemble $N = \{N_1, \ldots, N_J\}$ de matrices symétriques semi-définies positives, un ensemble $\omega = \{\omega_1, \ldots, \omega_J\}$ de poids et un scalaire $l \geq 1$. La mesure de pertinence structurelle (SR) ϕ associée est donnée par

$$\phi(u) = \left(\sum_{j=1}^{J} \omega_j \left(u^T N_j u \right)^l \right)^{1/l}.$$

Les matrices N_j sont telles que les formes quadratiques $u^T N_j u$ mesurent la proximité du vecteur u à des structures de référence.

Dans ce travail nous utilisons une mesure particulière de SR : la variance de la composante. On appelle W la matrice des poids *a priori* des observations (typiquement, $W = \frac{1}{n} I_n$). On prend X centrée en colonne. Nous voulons trouver une direction vect(u) captant une inertie suffisante des observations. Pour cela, on pose : $N = \{X^T W X\}$, $\omega = \{1\}$ et $l = 1$. Ainsi la SR devient

$$\phi(u) = u^T X^T W X u = \|Xu\|_W^2 = \mathbb{V}(Xu).$$

Nous reconnaissons le critère maximisé, sous la contrainte $u^T M^{-1} u = 1$, par l'ACP de X avec la métrique M et la matrice des poids W. De façon générale, quelle que soit la SR choisie, la métrique M de la contrainte $u^T M^{-1} u = 1$ est choisie de la forme $M^{-1} = \tau I_n + (1 - \tau) X^T W X$, où $\tau \in [0, 1]$ est un paramètre de régularisation de type ridge [**?**].

Pour construire un compromis entre le GoF et la SR, SCGLR introduit un réel $s \in [0, 1]$ et considère le programme de maximisation suivant :

$$\max_{u, u^T M^{-1} u = 1} \phi(u)^s \psi_A(u)^{1-s} \Leftrightarrow \max_{u, u^T M^{-1} u = 1} s \ln(\phi(u)) + (1 - s) \ln(\psi_A(u)).$$

2.2 MixRep-SCGLR

Dans cette section, nous proposons d'étendre SCGLR aux modèles de mélange sur les réponses. Nous considérons désormais que l'ensemble des réponses n'est pas modélisé par les mêmes dimensions explicatives. Nous cherchons à classifier les réponses dans des groupes tels que toutes les réponses d'un même groupe soient expliquées par les mêmes dimensions explicatives. Nous étudions le cas où toutes les réponses sont gaussiennes et nous supposons ici qu'il n'existe qu'une direction explicative par groupe.

Soit $\mathbf{Y} = [y_1, \ldots, y_K]$, l'ensemble des réponses. On note G le nombre de classes *a priori* du modèle et p_g la probabilité d'appartenance de chaque réponse à la classe g. Si y_k appartient au groupe g, alors y_k suit une loi normale multivariée $N_n(\mu_{kg}, \Sigma_{kg})$ avec $\mu_{kg} = (Xu_g)\gamma_{kg} + A\delta_{kg}$ et $\Sigma_{kg} = \sigma_{kg}^2 I_n$. La densité de y_k est donc :

$$f_Y(y_k; \Theta_k) = \sum_{g=1}^{G} \frac{p_g}{(2\pi\sigma_{kg}^2)^{n/2}} \exp\left(-\frac{1}{2\sigma_{kg}^2} \left\| y_k - (Xu_g)\gamma_{kg} - A\delta_{kg} \right\|^2 \right).$$

Conditionnellement aux variables explicatives, les réponses sont indépendantes. Ainsi,

$$f_{\mathbf{Y}}(\mathbf{y}; \Theta) = \prod_{k=1}^{K} f_Y(y_k; \Theta_k),$$

où l'ensemble des paramètres à estimer est $\Theta = \{p_1, \ldots, p_G, \gamma_{11}, \ldots, \gamma_{KG}, \delta_{11}, \ldots, \delta_{KG}, \sigma^2_{11}, \ldots, \sigma^2_{KG}\}$. Cette densité sert de mesure de GoF : $\psi_A(u, \Theta)$. Cependant, concernant Θ, la log-vraisemblance correspondante étant difficile à optimiser, nous utilisons l'algorithme EM [**?**] pour estimer les paramètres du modèle.

Soient z_{kg} la variable indicatrice latente valant 1 si la réponse y_k est dans le groupe g, le vecteur $Z_k = (z_{kg}\,;\, g = 1, \ldots, G)$ et la matrice $\mathbf{Z} = [Z_k\,;\, k = 1, \ldots, K]$. Conditionnellement à $z_{kg} = 1$, la réponse y_k suit une loi normale multivariée $N_n(\mu_{kg}, \Sigma_{kg})$. La log-vraisemblance complétée est donnée par

$$l(\Theta; \mathbf{Y}, \mathbf{Z}) = \ln\left(f_{\mathbf{YZ}}(\mathbf{y}, \mathbf{z}; \Theta)\right) = \ln\left(\prod_{k=1}^{K} f_{YZ}(y_k, z_k; \Theta_k)\right).$$

L'**étape M** de EM consiste à maximiser l'espérance conditionnelle de la log-vraisemblance complétée $\mathbb{E}\left[l(\Theta; \mathbf{Y}, \mathbf{Z})|\mathbf{Y}; \Theta'\right]$. Cette espérance conditionnelle est mise à jour dans l'**étape E**. Par ailleurs, nous prenons ici la variance de la composante comme SR. Ainsi, le critère à maximiser à l'**étape M** courante devient :

$$c(U, \gamma, \delta, \sigma^2) = s\sum_{g=1}^{G} \ln\left(\|Xu_g\|^2_W\right) + (1-s)\mathbb{E}\left[l(\Theta; \mathbf{Y}, \mathbf{Z})|\mathbf{Y}; \Theta'\right],$$

où $U = \{u_1, \ldots, u_G\}$, Xu_g étant la composante du groupe g, et $\gamma = \{\gamma_{11}, \ldots, \gamma_{KG}\}$, $\delta = \{\delta_{11}, \ldots, \delta_{KG}\}$ et $\sigma^2 = \{\sigma^2_{11}, \ldots, \sigma^2_{KG}\}$ sont les ensembles des paramètres à estimer.

3 Algorithme

Étape E (Espérance conditionnelle)

Pour réaliser l'**étape E** de l'algorithme, nous devons calculer explicitement l'espérance conditionnelle de la log-vraisemblance complétée. Tout calcul fait :

$$\begin{aligned}\mathbb{E}\left[l(\Theta; \mathbf{Y}, \mathbf{Z})|\mathbf{Y}; \Theta'\right] &= -\frac{nK}{2}\ln(2\pi) + \sum_{k=1}^{K}\sum_{g=1}^{G}\alpha_{kg}\ln(p_g) \\ &\quad -\frac{1}{2}\sum_{k=1}^{K}\sum_{g=1}^{G}\alpha_{kg}\left(n\ln(\sigma^2_{kg}) + \frac{1}{\sigma^2_{kg}}\|y_k - (Xu_g)\,\gamma_{kg} - A\delta_{kg}\|^2\right),\end{aligned}$$

avec $\alpha_{kg} = \mathbb{P}(Z_k = g|y_k; \Theta'_k) = \dfrac{p_g N_n(\mu_{kg}, \Sigma_{kg})}{\sum_{g'=1}^{G} p_{g'} N_n(\mu_{kg'}, \Sigma_{kg'})}$, où $Z_k = g$ est un raccourci pour signifier que 1 est à la g-ième place sur l'indicatrice.

Étape M (Maximisation)

L'**étape M** maximise sur Θ : $\mathbb{E}\left[l(\Theta; \mathbf{Y}, \mathbf{Z})|\mathbf{Y}; \Theta'\right]$, sous la contrainte $\sum_{g=1}^{G} p_g = 1$. Ainsi nous devons annuler le gradient en Θ du Lagrangien suivant :

$$L(\Theta, \lambda) = \mathbb{E}\left[l(\Theta; \mathbf{Y}, \mathbf{Z})|\mathbf{Y}; \Theta'\right] - \lambda\left(\sum_{g=1}^{G} p_g - 1\right).$$

On obtient,

$$\nabla_{p_g} L(\Theta, \lambda) = 0 \Leftrightarrow \hat{p}_g = \frac{1}{K}\sum_{k=1}^{K}\alpha_{kg}.$$

Pour estimer γ_{kg} et δ_{kg}, nous posons $T_g = [Xu_g, A]$ et $\theta_{kg} = (\gamma_{kg}, \delta_{kg})^T$. La contrainte ne dépendant pas de θ_{kg}, il suffit d'annuler le gradient de $\mathbb{E}\left[l(\Theta; \mathbf{Y}, \mathbf{Z})|\mathbf{Y}; \Theta'\right]$. On obtient,

$$\nabla_{\theta_{kg}} \mathbb{E}\left[l(\Theta; \mathbf{Y}, \mathbf{Z})|\mathbf{Y}; \Theta'\right] = 0 \Leftrightarrow \hat{\theta}_{kg} = \left(T_g^T W T_g\right)^{-1} \left(T_g^T W y_k\right).$$

On estime de même σ_{kg}^2 :

$$\nabla_{\sigma_{kg}^2} \mathbb{E}\left[l(\Theta; \mathbf{Y}, \mathbf{Z})|\mathbf{Y}; \Theta'\right] = 0 \Leftrightarrow \hat{\sigma}_{kg}^2 = \left\|y_k - T_g \hat{\theta}_{kg}\right\|_W^2.$$

Trouver la composante de chaque classe

Pour trouver les vecteurs de U, nous allons utiliser l'algorithme du gradient normé projeté itéré (Projected Iterated Normed Gradient, PING) [**?**] dans chaque classe. En effet, le critère global étant une somme de sous-critères relatifs aux classes, il suffit de maximiser isolément chaque sous-critère de classe. Ainsi, pour la classe g, le sous-critère est :

$$\begin{aligned} c(u_g, \gamma, \delta, \sigma^2) &= s \ln\left(\|Xu_g\|_W^2\right) + (1-s)\left[-\frac{nK}{2G}\ln(2\pi) + \ln(p_g)\sum_{k=1}^{K}\alpha_{kg}\right. \\ &\quad \left. -\frac{1}{2}\sum_{k=1}^{K}\alpha_{kg}\left(n\ln(\sigma_{kg}^2) + \frac{1}{\sigma_{kg}^2}\|y_k - (Xu_g)\gamma_{kg} - A\delta_{kg}\|^2\right)\right]. \end{aligned}$$

Algorithme

Algorithme MixRep-SCGLR

On initialise l'algorithme avec les valeurs initiales $u^{(0)}$, $\gamma^{(0)}$, $\delta^{(0)}$, $p_g^{(0)}$ et $t = 0$.
À l'itération $t+1$:

1. On estime les paramètres (hors U) par l'algorithme EM.
   À l'itération $m+1$:

   (a) **Étape E** :
   Pour $k = 1, \ldots, K$:
   Pour $g = 1, \ldots, G$:
   $$\alpha_{kg}^{(m+1)} = \frac{p_g^{(m)} N_n(\mu_{kg}, \Sigma_{kg})}{\sum_{g'=1}^{G} p_{g'}^{(m)} N_n(\mu_{kg'}, \Sigma_{kg'})}$$

   (b) **Étape M** :

   i. Pour $g = 1, \ldots, G$:
   $$p_g^{(m+1)} = \frac{1}{K}\sum_{k=1}^{K}\alpha_{kg}^{(m+1)}$$
   ii. Pour $k = 1, \ldots, K$:
   Pour $g = 1, \ldots, G$:
   $$\theta_{kg}^{(m+1)} = \left(T_g^{(t)T} W T_g^{(t)}\right)^{-1} T_g^{(t)T} W y_k$$
   $$\sigma_{kg}^{2(m+1)} = \left\|y_k - T_g^{(t)} \theta_{kg}^{(m+1)}\right\|_W^2$$

   Les paramètres $\gamma^{(t+1)}$, $\delta^{(t+1)}$ et $\sigma^{2(t+1)}$ sont alors respectivement égaux à $\gamma^{(m_{\max})}$, $\delta^{(m_{\max})}$ et $\sigma^{2(m_{\max})}$ obtenus à la convergence de l'algorithme EM.

2. On calcule $U = (u_g)_g$ à l'aide de l'algorithme PING.
   Pour $g = 1, \ldots, G$:
   $u_g^{(t+1)} = \underset{u, u^T M^{-1} u = 1}{\operatorname{argmax}} \; c(u_g^{(t)}, \gamma^{(t+1)}, \delta^{(t+1)}, \sigma^{2(t+1)})$
   $T_g^{(t+1)} = [X u_g^{(t+1)}, A]$

Lorsque l'algorithme a convergé, on peut classer les réponses parmi les groupes à l'aide des probabilités *a posteriori* d'inclusion. Une réponse y_k est dans le groupe g si

$$\alpha_{kg}^{(t_{\max})} > \alpha_{kg'}^{(t_{\max})}$$

pour tout $g' \neq g$.

L'algorithme fut testé avec succès sur différents jeux de données simulées.

Remerciements

Cette recherche a été soutenue par le projet GAMBAS financé par l'Agence Nationale de la Recherche (ANR-18-CE02-0025).

Références

[1] Xavier Bry, Catherine Trottier, Thomas Verron, and Frédéric Mortier. Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm. *Journal of Multivariate Analysis*, 119 : 47–60, 2013.

[2] Xavier Bry and Thomas Verron. THEME : THEmatic Model Exploration through multiple co-structure maximization. *Journal of Chemometrics*, 29(12) : 637–647, 2015.

[3] Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society : Series B (Methodological)*, 39(1) : 1–22, 1977.

[4] Arthur E. Hoerl and Robert W. Kennard. Ridge regression : applications to nonorthogonal problems. *Technometrics*, 12(1) : 69–82, 1970.

[5] Alston S. Householder. *The theory of matrices in numerical analysis*. Courier Corporation, 2013.

[6] P. McCullagh and J.A. Nelder. 1989, Generalized Linear Models, Chapman and Hall, New York, NY.