Régression linéaire généralisée sur composantes supervisées pour les modèles à facteurs latents

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

À l'origine, la Régression Linéaire Généralisée sur Composantes Supervisées (SCGLR) a été conçue pour trouver des composantes explicatives conjointement supervisées par un ensemble de réponses au sein de très nombreuses covariables redondantes, ce qui est nécessaire dans un contexte de grande dimension. Dans ce travail, nous proposons d'étendre la méthode SCGLR dans l'objectif de modéliser la matrice de variance-covariance des réponses de telle sorte que la corrélation entre ces réponses soit principalement expliquée par quelques facteurs. Nous cherchons à identifier des blocs dans la matrice de variance-covariance pour les réponses partageant des dépendances mutuelles. Dans un cadre écologique par exemple, nous nous intéressons aux relations statistiques entre les présences de différentes espèces. Un algorithme basé sur EM est proposé afin d'estimer le modèle.

Mots clefs: SCGLR, modèle à facteurs, algorithme EM, variables latentes.

Abstract

Originally, the Supervised Component-based Generalized Linear Regression (SCGLR) was designed to find explanatory components jointly supervised by a set of responses in many redundant covariates, something much needed in a high-dimensional framework. In this work, we propose to extend the SCGLR methodology with the objective of modeling the responses variance-covariance matrix in such a way that the correlation between these responses is mainly explained by few factors. We aim at identifying blocks in the variance-covariance matrix depicting the outcomes sharing mutual dependencies. In an ecological framework for instance, we study the statistical relations between the presences of different species. An algorithm based on EM is proposed in order to estimate the model.

Keywords: SCGLR, factor model, EM algorithm, latent variables.

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.* [1] 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 pertinence structurelle (Structural Relevance, SR) [2] mesurant la proximité des composantes supervisées à des dimensions d'intérêt. 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 les réponses sont indépendantes les unes des autres. Pour nous affranchir de cette hypothèse, nous proposons d'étendre cette méthode aux modèles à facteurs. L'objectif est de modéliser la matrice de variance covariance des réponses (espèces) afin d'identifier celles qui auraient des dépendances mutuelles.

2 Modélisation

Dans cette section, nous présentons la méthode SCGLR, puis son extension aux modèles à facteurs latents.

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 $X \in \mathbb{R}^{N\times P}$ et $A \in \mathbb{R}^{N\times R}$ les matrices correspondantes. Chaque réponse y^k fait l'objet d'un modèle linéaire généralisé (Generalized Linear Model, GLM) [8]. 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 son 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 $\psi_A(u) = \sum_{k=1}^{K} \left\|w^k\right\|_{W_k}^2 \cos^2_{W_k}\left(w^k, \Pi^{W_k}_{\text{vect}(f,A)} w^k\right)$. 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 (SR).

Dans ce travail nous utilisons une mesure particulière de SR : l'inertie duale généralisée (Variable Powered Inertia, VPI). On appelle W la matrice des poids *a priori* des observations (typiquement, $W = \frac{1}{N} I_N$) et on suppose les colonnes de X centrées et réduites. Nous voulons trouver une direction vect(u) proche d'un faisceau. En un mot, un faisceau est un ensemble de variables explicatives suffisamment corrélées pour être vues comme alignées autour de la même dimension latente. Pour cela, on pose $l \geq 1$ et la SR s'écrit :

$$\phi(u) = \left(\frac{1}{P}\sum_{p=1}^{P}\left(u^T X^T W x^p x^{pT} W X u\right)^l\right)^{1/l} = \left(\frac{1}{P}\sum_{p=1}^{P}\langle Xu, x^p\rangle_W^{2l}\right)^{1/l}.$$

Le paramètre l permet de trouver une composante proche d'un faisceau plus (l fort) ou moins (l faible) étroit de variables corrélées. De façon générale, quelle que soit la SR choisie, la métrique M de la contrainte $u^T M^{-1} u = 1$ est écrite 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 [6].

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

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

Afin de trouver les composantes d'ordre $h > 1$, nous notons $f^h = Xu^h$ la h-ième composante et $F^h = [f^1, \ldots, f^h]$ la matrice des h premières composantes, avec $h < H$. Par simplification, nous notons F la matrice des H composantes. Nous adoptons alors le principe d'emboîtement local (Local Nesting, LocNes) présenté par [3]. Suivant ce principe, la composante supplémentaire f^{h+1} doit venir compléter au mieux les composantes précédentes en plus de la matrice A, c'est à dire $A^h := [F^h, A]$. Ainsi, f^{h+1} est calculée en utilisant A^h comme matrice de covariables additionnelles. De plus, nous imposons que f^{h+1} soit orthogonale à F^h par la contrainte $F^{h\,T} W f^{h+1} = 0$. Cette maximisation sous contrainte est permise par l'algorithme du gradient normé projeté itéré (Projected Iterated Normed Gradient, PING) [7].

2.2 SCGLR pour modèle à facteurs

À l'origine, la méthode SCGLR fut développée dans un contexte de modèle linéaire généralisé, cependant, dans ce travail, nous nous limiterons à une matrice $Y \in \mathbb{R}^{N\times K}$ de réponses gaussiennes. Nous appelons y_n, f_n et a_n les vecteurs composés de la n-ième ligne des matrices Y, F et A respectivement. Chaque y_n est expliquée par f_n, a_n et par L variables latentes $g_n = (g_n^1, \ldots, g_n^L)^T$ appelées facteurs. Ainsi, le modèle s'écrit :

$$\underbrace{y_n}_{K\times 1} = \underbrace{\Gamma^T}_{K\times H}\underbrace{f_n}_{H\times 1} + \underbrace{\Delta^T}_{K\times R}\underbrace{a_n}_{R\times 1} + \underbrace{B^T}_{K\times L}\underbrace{g_n}_{L\times 1} + \underbrace{\varepsilon_n}_{K\times 1}, \tag{2}$$

où $\Gamma = [\gamma^1, \ldots, \gamma^K]$, $\Delta = [\delta^1, \ldots, \delta^K]$ et $B = [b^1, \ldots, b^K]$ sont des paramètres de régression et où $\varepsilon_n \sim \mathcal{N}(0, \Psi)$ avec $\Psi = \text{diag}(\sigma_k^2)_{k=1,\ldots,K}$, représente les erreurs indépendantes. De plus g_n est supposé suivre une loi $\mathcal{N}(0, I_L)$ et être indépendant de ε_m pour toutes valeurs de n et m. Ces hypothèses impliquent que le modèle est construit de telle manière que toute la corrélation entre

les réponses soit expliquée par les L facteurs.

Afin d'estimer les paramètres, nous devons maximiser la log-vraisemblance du modèle $l(\Theta; Y)$, où $\Theta = \{\Gamma, \Delta, B, \Psi\}$. Cependant, à cause des facteurs non observés, cette log-vraisemblance possède une expression complexe qui la rend difficile à maximiser. Ainsi, nous utilisons l'algorithme EM [4] pour estimer les paramètres. L'étape M de l'algorithme consiste à maximiser l'espérance conditionnelle de la log-vraisemblance complétée $\mathbb{E}[l(\Theta; Y, G)|Y; \Theta']$, cette espérance étant mise à jour dans l'étape E.

2.2.1 Étape E (Espérance conditionnelle)

Pour réaliser l'étape E de l'algorithme, nous devons calculer explicitement l'espérance conditionnelle de la log-vraisemblance complétée. Cette dernière s'écrit :

$$\mathbb{E}[l(\Theta; Y, G)|Y, \Theta'] = \sum_{n=1}^{N} \int \ln\left(f(y_n|g_n; \Theta) f(g_n; \Theta)\right) f(g_n|y_n; \Theta') dg_n.$$

Ainsi, nous devons préalablement calculer les lois de $y_n|g_n$ et de $g_n|y_n$. La loi de $y_n|g_n$ est donnée par le modèle (2) tandis que la loi de $g_n|y_n$ est donnée par la règle de l'espérance conditionnelle des lois Gaussiennes multivariées :

$$\text{Si} \quad \begin{pmatrix} y_n \\ g_n \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \Gamma^T f_n + \Delta^T a_n \\ 0 \end{pmatrix}, \begin{pmatrix} B^T B + \Psi & B^T \\ B & I_L \end{pmatrix}\right),$$

alors $g_n|y_n \sim \mathcal{N}(\alpha(y_n - \Gamma^T f_n - \Delta^T a_n), I_L - \alpha B^T)$, où $\alpha = B(B^T B + \Psi)^{-1}$. Finalement, tout calcul fait, l'espérance de la log-vraisemblance complétée devient :

$$\begin{aligned}\mathbb{E}[l(\Theta; Y, G)|Y, \Theta'] = -\frac{1}{2}\Bigg\{ & N(K+L)\ln(2\pi) + N\sum_{k=1}^{K}\ln(\sigma_k^2) + \sum_{n=1}^{N}\mathbb{E}\left[g_n^T g_n|y_n; \Theta'\right] + \\ & \sum_{k=1}^{K}\frac{1}{\sigma_k^2}\left[\left\|y^k - F\gamma^k - A\delta^k\right\|^2 + b^{k\,T}\tilde{R}b^k - 2\left(\tilde{G}b^k\right)^T\left(y^k - F\gamma^k - A\delta^k\right)\right]\Bigg\},\end{aligned}$$

où les lignes de la matrice $\tilde{G}$ sont les moments d'ordre 1 :

$$\tilde{g}_n := \mathbb{E}(g_n|y_n; \Theta) = \alpha(y_n - \Gamma^T f_n - \Delta^T a_n) \tag{3}$$

et où $\tilde{R} = \sum_{n=1}^{N} \tilde{R}_n$ est la somme des moments d'ordre 2 :

$$\tilde{R}_n := \mathbb{E}(g_n g_n^T|y_n; \Theta) = \mathbb{V}(g_n|y_n; \Theta) + \mathbb{E}(g_n|y_n; \Theta)\mathbb{E}(g_n|y_n; \Theta)^T = I_L - \alpha B^T + \tilde{g}_n \tilde{g}_n^T. \tag{4}$$

2.2.2 Étape M (Maximisation)

Dans un objectif d'identification, nous avons besoin de contraindre la matrice B. Comme démontré par [5], si Ω est une matrice orthogonale, nous pouvons remplacer le produit $B^T g_n$ par $B_0^T g_{0n}$ dans lequel le nouveau facteur $g_{0n} = \Omega g_n$ est une rotation de l'ancien facteur g_n. Les conditions sur les moments respectées par les anciens facteurs sont aussi respectées par les nouveaux, autrement dit, $\mathbb{E}(g_{0n}) = \Omega\mathbb{E}(g_n) = 0$ et $\mathbb{V}(g_{0n}) = \Omega\mathbb{V}(g_n)\Omega^T = I_L$. De plus, les paramètres aussi subissent une rotation. Les nouveaux paramètres sont liés aux anciens par $B_0^T = B^T\Omega^T$. Étant donné que ces nouveaux paramètres et facteurs donnent lieu à la même distribution, ils ne peuvent être identifiés

à partir des observations que si des restrictions supplémentaires sont imposées. Nous choisissons d'imposer la matrice B contrainte :

$$B = \begin{pmatrix} b_1^1 & \dots & b_1^L & b_1^{L+1} & \dots & b_1^K \\ & \ddots & \vdots & \vdots & \ddots & \vdots \\ & & b_L^L & b_L^{L+1} & \dots & b_L^K \end{pmatrix},$$

où pour tout $k < l$, $k, l = 1, \dots, L$, $b_l^k = 0$ et où pour tout $l = 1, \dots, L$, $b_l^l > 0$.

La contrainte n'imposant rien sur les paramètres Γ, Δ et Ψ, nous maximisons normalement sur ces derniers. Ainsi, pour tout $k = 1, \dots, K$, la maximisation de $\mathbb{E}[l(\Theta; Y, G)|Y, \Theta']$ sur $\{\gamma^k, \delta^k, \sigma^k\}$ donne :

$$\begin{pmatrix} \hat{\gamma}^k \\ \hat{\delta}^k \end{pmatrix} = \left([F, A]^T [F, A]\right)^{-1} [F, A]^T \left(y^k - \tilde{G} b^k\right), \tag{5}$$

$$\hat{\sigma}_k^2 = \frac{1}{N} \left\{ \left\| y^k - F\gamma^k - A\delta^k \right\|^2 + b^{k\,T} \tilde{R} b^k - 2 \left(\tilde{G} b^k\right)^T \left(y^k - F\gamma^k - A\delta^k\right) \right\}. \tag{6}$$

Désormais, nous devons maximiser sur le vecteur b^k sous la contrainte. Pour tout $k = 1, \dots, L$, posons $b^k = (b_{1:k}^{k\,T}, \mathbf{0}^T)^T$, où $b_{1:k}^k = (b_1^k, \dots, b_k^k)^T$ est un vecteur de longueur k et $\mathbf{0}$ le vecteur nul de longueur $L - k$. Ainsi, après maximisation, on a pour tout $k = 1, \dots, L$,

$$\hat{b}_{1:k}^k = \left(\tilde{R}_{1:k}^{1:k}\right)^{-1} \left(\tilde{G}^{1:k}\right)^T \left(y^k - F\gamma^k - A\delta^k\right), \tag{7}$$

où $\tilde{R}_{1:k}^{1:k}$ est la sous matrice de taille $k \times k$ de $\tilde{R}$ et où $\tilde{G}^{1:k}$ est la matrice composée des k premières colonnes de $\tilde{G}$. De la même manière, pour $k = L + 1, \dots, K$, on a :

$$\hat{b}^k = \tilde{R}^{-1} \tilde{G}^T \left(y^k - F\gamma^k - A\delta^k\right). \tag{8}$$

2.2.3 Contrainte sur le nombre maximal de facteurs

Nous appelons $\Sigma = B^T B + \Psi$ la matrice de variance-covariance de y_n. Il existe seulement $K(K+1)/2$ éléments distincts dans Σ, cependant il y a LK éléments dans B plus K éléments dans Ψ. La contrainte impose *a priori* $L(L-1)/2$ éléments nuls sur la matrice B. Finalement, $B^T B + \Psi$ possède $LK + K - L(L-1)/2$ éléments distincts. Pour déterminer ces paramètres, nous devons avoir $K(K+1)/2 \geq LK + K - L(L-1)/2$ ou, autrement dit, $L \leq (2K + 1 - \sqrt{8K+1})/2$. Nous donnons quelques exemples de nombres maximaux de facteurs en fonction du nombre de réponses :

K	1	2	3	4	5	6	7	8	9	10
L max	0	0	1	1	2	3	3	4	5	6

Table 1: Nombre maximal de facteurs

En pratique, pour simplifier la structure de variance-covariance résiduelle des réponses, le nombre de facteurs restera faible.

3 Algorithme

Des essais numériques, sur données simulées et réelles, impliquant l'Algorithme 1 seront exposés lors de la présentation orale de ces travaux de recherche.

Algorithm 1: SCGLR pour les modèles à facteurs

while *not convergence* **do**

Répéter les étapes (3) et (4) puis les étapes (5), (6), (7) et (8) jusqu'à la convergence de l'algorithme EM

$\Theta^{(t+1)} = \underset{\Theta}{\operatorname{argmax}}\ l(\Theta^{(t)}; Y)$

Répéter sur le nombre de composantes la maximisation du critère (1) à l'aide de l'algorithme PING

$\forall h = 1, \ldots, H, \quad f^{h\,(t+1)} = X u^{h\,(t+1)}$

$t \leftarrow t + 1$

end

Remerciements

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

References

[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] Xavier Bry, Thomas Verron, and Pierre Cazes. Exploring a physico-chemical multi-array explanatory model with a new multiple covariance-based technique: Structural equation exploratory regression. *Analytica chimica acta*, 642(1-2): 45–58, 2009.

[4] 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.

[5] John Geweke and Guofu Zhou. Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies*, 9(2): 557–587, 1996.

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

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

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