Modèle linéaire généralisé basé sur des facteurs latents et des composantes supervisées

Julien Gibaud, Xavier Bry, Catherine Trottier

▶ **To cite this version:**

Julien Gibaud, Xavier Bry, Catherine Trottier. Modèle linéaire généralisé basé sur des facteurs latents et des composantes supervisées. 54ème Journées de Statistiques de la Société Française de Statistique (SFDS(, Jul 2023, Bruxelles, Belgique. hal-04050330

HAL Id: hal-04050330
https://hal.science/hal-04050330

Submitted on 29 Mar 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Modèle linéaire généralisé basé sur des facteurs latents et des composantes supervisées

Julien Gibaud [1,2], Xavier Bry [1] et Catherine Trottier [1,2]

[1] *IMAG, CNRS, Université de Montpellier, France*
[2] *AMIS, Université Paul-Valéry Montpellier 3, France*
julien.gibaud@umontpellier.fr

Résumé. À l'origine, la Régression Linéaire Généralisée sur Composantes Supervisées (SCGLR) a été conçue pour trouver, au sein de très nombreuses covariables redondantes, des composantes explicatives conjointement supervisées par plusieurs réponses, ce qui est nécessaire dans un contexte de grande dimension. Plus tard, SCGLR fut améliorée pour chercher des composantes au sein de variables explicatives partitionnées en thèmes. Dans ce travail, nous proposons d'étendre cette méthode en modélisant la matrice de variance-covariance conditionnelle des réponses de telle sorte que la covariance conditionnelle des réponses soit principalement expliquée par quelques facteurs. Nous chercherons donc non seulement à extraire des thèmes des composantes explicatives, mais aussi à identifier des blocs dans la matrice de variance-covariance des réponses conditionnellement à ces composantes. Après la linéarisation du modèle, un algorithme combinant EM et celui de SCGLR thématique est proposé afin d'estimer les paramètres du modèle. Cette nouvelle méthodologie est testée sur des données simulées puis sur des données issues de l'écologie agricole.

Mots-clés. SCGLR, modèle à facteurs, algorithme EM, variables latentes

Abstract. Originally, the Supervised Component-based Generalized Linear Regression (SCGLR) was designed to extract, from a large set of redundant covariates, explanatory components jointly supervised by a set of responses, something much needed in a high-dimensional framework. Later on, SCGLR has been improved to search for components in explanatory variables divided up into thematic subsets. In this work, we propose to extend this methodology by modeling the responses' conditional variance-covariance matrix so that the conditional covariance of responses is mainly explained by few factors. Not only do we aim to extract explanatory components from the thematic subsets, but also to identify blocks in the conditional variance-covariance matrix of the responses. After linearising the model, we propose an algorithm combining EM with that of thematic SCGLR to estimate the model parameters. This new methodology is tested on simulated data and then applied to an agricultural ecology dataset.

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

1 Contexte

Dans leur article, Bry et al. (2013) 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, Bry and Verron, 2015) 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. Bry et al. (2020) proposent de chercher ces composantes dans un partitionnement thématique des variables explicatives. Dans un contexte écologique par exemple, les variables de température et de précipitation seraient considérées comme appartenant à deux thèmes distincts. Cependant, SCGLR suppose que les réponses sont indépendantes conditionnellement aux covariables. 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 conditionnelle des réponses afin d'identifier des groupes de réponses liées.

2 Modélisation

Dans cette section, nous présentons la méthode SCGLR, sa généralisation au partitionnement des variables explicatives, 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 $X \in \mathbb{R}^{N \times P}$ de covariables *a priori* nombreuses et possiblement redondantes, et un autre $A \in \mathbb{R}^{N \times Q}$ de covariables additionnelles peu nombreuses et faiblement, voire non-redondantes. Chaque réponse y_k fait l'objet d'un modèle linéaire généralisé (Generalized Linear Model, GLM, McCullagh and Nelder, 1989). 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. Dans un modèle à une composante, le prédicteur linéaire associé à la réponse y_k est ainsi 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 $u^T u = 1$.

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 du programme d'optimisation suivant :

$$\max_{u,u^Tu=1} \psi_A(u) := \sum_{k=1}^{K} \|w_k\|^2_{W_k} \cos^2_{W_k}\left(w_k, \Pi^{W_k}_{\text{span}[f,A]} w_k\right).$$

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 span$[Xu]$ proche d'un faisceau de covariables (i.e. un ensemble de variables explicatives suffisamment corrélées pour être vues comme alignées autour de la même dimension latente). Pour cela, on considère un paramètre $l \geq 1$ et on formule la SR comme :

$$\phi(u) = \left(\frac{1}{P}\sum_{p=1}^{P} \langle Xu, x_p\rangle^{2l}_W\right)^{1/l}. \tag{1}$$

Le paramètre l permet de trouver une composante proche d'un faisceau plus (l fort) ou moins (l faible) étroit de variables corrélées.

Pour construire un compromis entre le GoF et la SR, SCGLR les combine en utilisant un paramètre $s \in [0, 1]$ permettant de régler leurs poids relatifs et considère la maximisation suivante :

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

2.2 Généralisation à un partitionnement thématique des variables explicatives

Bry et al. (2020) définissent le "modèle thématique" comme étant le modèle conceptuel indiquant que la matrice réponse Y dépend de R thèmes $X_1, \ldots, X_R$ plus un ensemble de covariables A, et que les dimensions structurellement pertinentes doivent être explicitement identifiées dans chaque X_r. Le prédicteur linéaire associé à la réponse y_k devient :

$$\eta_k = (X_1u_1)\gamma_{k1} + \cdots + (X_Ru_R)\gamma_{kR} + A\delta_k.$$

Pour ouvrir la voie à la régularisation des thèmes, nous devons adapter le critère compromis de SCGLR. En appelant $f_r = X_r u_r$ la composante du thème X_r, nous avons $\Pi^{W_k}_{\text{span}[f_1,\ldots,f_R,A]} = \Pi^{W_k}_{\text{span}[f_r,A_r]}$ où $A_r = [f_1, \ldots, f_{r-1}, f_{r+1}, \ldots, f_R, A]$. La mesure de GoF devient donc :

$$\psi_{A_r}(u_r) := \sum_{k=1}^{K} \|w_k\|^2_{W_k} \cos^2_{W_k}\left(w_k, \Pi^{W_k}_{\text{span}[f_r,A_r]} w_k\right),$$

où les covariables supplémentaires devront être spécifiques à chaque occasion. La mesure de SR reste inchangée en prenant $\phi(u_r)$ comme en (1). Ainsi on obtient un compromis entre le GoF et la SR spécifique à chaque thème. Enfin, le programme d'optimisation peut être résolu en maximisant itérativement le critère compromis sur chaque u_r :

$$\forall r, \quad \max_{u_r, u_r^T u_r = 1} s \ln(\phi(u_r)) + (1-s)\ln(\psi_{A_r}(u_r)). \tag{2}$$

Afin de trouver les composantes d'ordre $h > 1$, nous notons $f_r^h = X_r u_r^h$ la h-ième composante du thème X_r et $F_r^h = [f_r^1, \ldots, f_r^h]$, où $h \leq H_r$, la matrice des h premières composantes du thème X_r. La nouvelle composante f_r^{h+1} doit venir compléter au mieux les composantes précédentes en plus de la matrice A, c'est à dire $A_r^h := [F_1^{H_1}, \ldots, F_{r-1}^{H_{r-1}}, F_r^h, F_{r+1}^{H_{r+1}}, \ldots, F_R^{H_R}, A]$. Ainsi, f_r^{h+1} est calculée en utilisant A_r^h comme nouvelle matrice de covariables additionnelles. De plus, nous imposons que f_r^{h+1} soit orthogonale à F_r^h par la contrainte ${F_r^h}^T W f_r^{h+1} = 0$. Cette maximisation sous contrainte est permise par l'algorithme du gradient normé projeté itéré (Projected Iterated Normed Gradient, PING).

2.3 SCGLR thématique pour modèles à facteurs

Dans cette section, les composantes sont considérées comme connues. Ainsi, pour des raisons de simplicité, nous prenons la matrice $F = [F_1^{H_1}, \ldots, F_R^{H_R}]$ comme la nouvelle matrice des variables explicatives et $\gamma_k = (\gamma_{k1}, \ldots, \gamma_{kR})^T$ le vecteur des paramètres de régression associé à la réponse y_k. Pour une unité statistique n, chaque réponse est modélisée linéairement à l'aide de J variables latentes aléatoires $g_n = (g_{n1}, \ldots, g_{nJ})^T$ appelées facteurs :

$$\eta_{nk} = f_n^T \gamma_k + a_n^T \delta_k + g_n^T b_k,$$

où f_n et a_n sont les vecteurs composés par les n-ième lignes des matrices F et A respectivement, et b_k le vecteur des paramètres de régression de g_n. Les facteurs suivent une loi normale multivariée $g_n \sim \mathcal{N}_J(0, I_J)$ et on suppose qu'ils sont indépendants entre les unités statistiques. Le modèle est construit de telle manière que les facteurs capturent la covariance entre les réponses conditionnellement aux composantes et covariables supplémentaires. En notant $G \in \mathbb{R}^{N \times J}$ la matrice contenant toutes les réalisations des facteurs, le prédicteur linéaire associé à la réponse y_k exprimé en colonne devient donc :

$$\eta_k = F\gamma_k + A\delta_k + Gb_k.$$

Soit $B = [b_1, \ldots, b_K] \in \mathbb{R}^{J \times K}$ la matrice des coefficients des facteurs. Comme montré par Geweke and Zhou (1996), cette matrice n'est pas unique. Pour garantir l'identification du modèle, nous devons imposer que la sous matrice de B de taille $J \times J$ soit un triangle supérieur avec des éléments diagonaux tous positifs. Un avantage principal des modèles à facteurs est d'exprimer la matrice $\Sigma = B^T B \in \mathbb{R}^{K \times K}$ des covariances conditionnelles des réponses de manière parcimonieuse. En effet, outre la contrainte triangulaire, le nombre de facteurs choisi est petit.

Dans un objectif de clarté, la figure 1 présente de manière graphique le modèle thématique avec facteurs.

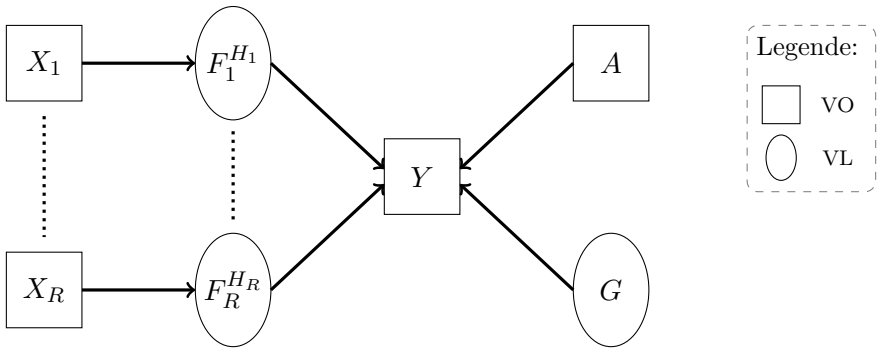

Figure 1: Représentation graphique de SCGLR thématique pour modèles à facteurs. L'intervention d'un groupe de variables sur un autre est représentée par une flèche. Les variables observées (VO) sont placées dans des carrés tandis que les ovales contiennent les variables latentes (VL).

3 Estimation du modèle

Soit $\Theta = \{\gamma_k, \delta_k, b_k; k = 1, \ldots, K\}$ l'ensemble des paramètres. La log-vraisemblance du modèle est obtenue en intégrant sur les facteurs g_n :

$$\begin{aligned} l(\Theta; Y) &= \sum_{n=1}^{N} \ln\left(L(y_n; \Theta)\right) \\ &= \sum_{n=1}^{N} \ln\left(\int \prod_{k=1}^{K} L(y_{nk}|g_n; \Theta) L(g_n)\, dg_n\right). \end{aligned}$$

Cependant, dans un contexte de réponses non Gaussiennes, la maximisation de cette log-vraisemblance n'est pas possible. Dans l'esprit de Saidane et al. (2013), nous avons choisi d'effectuer l'estimation en deux étapes.

3.1 Linéarisation du modèle

Nous considérons d'abord que les facteurs sont connus, c'est à dire que, conditionnellement à G, nous nous plaçons dans le cadre des GLM multivariés classiques. Soient h_k la fonction de lien canonique associée à la réponse y_k, h'_k sa première dérivée et μ_{nk} le paramètre d'espérance pour l'unité statistique n. La pseudo réponse w_{nk} est alors calculée comme le développement limité au premier ordre de h_k au point μ_{nk} :

$$\begin{aligned} w_{nk} &= h_k\left(\mu_{nk}\right) + \left(y_{nk} - \mu_{nk}\right) h'_k\left(\mu_{nk}\right) \\ &= \eta_{nk} + \zeta_{nk}, \end{aligned}$$

où $\zeta_{nk} = \left(y_{nk} - \mu_{nk}\right) h'_k\left(\mu_{nk}\right)$. Ce développement conduit alors au modèle linéarisé conditionnel suivant :

$$w_k = F\gamma_k + A\delta_k + Gb_k + \zeta_k,$$

où $\mathbb{E}[w_k|G] = F\gamma_k + A\delta_k + Gb_k$ et $\mathbb{V}[w_k|G] = \mathbb{V}[\zeta_k] = W_k^{-1}$.

3.2 Estimation à l'aide de l'algorithme EM

La deuxième étape consiste à supposer Gaussiennes les pseudo réponses désormais connues. Les facteurs aléatoires étant latents, la log-vraisemblance du modèle linéarisé $l(\Theta; \mathcal{W})$, où $\mathcal{W}$ désigne la matrice des pseudo réponses, possède une expression complexe qui la rend difficile à maximiser. Ainsi, nous utilisons l'algorithme EM (Dempster et al., 1977) pour estimer les paramètres. Nous calculons puis maximisons l'espérance de la log-vraisemblance complétée $l(\Theta; \mathcal{W}, G)$ conditionnellement aux pseudo réponses.

4 Algorithme

Algorithm 1: SCGLR thématique pour les modèles à facteurs

while *not convergence* **do**

Calculer les composantes grâce à la maximisation du critère (2) à l'aide de l'algorithme PING

$$\forall r = 1, \ldots, R, \forall h = 1, \ldots, H_r, \quad f_r^{h\,(t+1)} = X_r u_r^{h\,(t+1)}$$

Calculer les pseudo réponses à l'aide du FSA

$$\eta_k^{(t+1)} = F^{(t+1)} \gamma_k^{(t)} + A\delta_k^{(t)} + Gb_k^{(t)}$$

$$\mu_{nk}^{(t+1)} = h_k^{-1}\left(\eta_{nk}^{(t+1)}\right), \forall n = 1, \ldots, N$$

$$w_{nk}^{(t+1)} = \eta_{nk}^{(t+1)} + h_k'\left(\mu_{nk}^{(t+1)}\right)\left(y_{nk} - \mu_{nk}^{(t+1)}\right), \forall n = 1, \ldots, N$$

$$W_k^{(t+1)} = \operatorname{diag}\left(\left[a_{nk}(\phi_k) v_k\left(\mu_{nk}^{(t+1)}\right) h_k'\left(\mu_{nk}^{(t+1)}\right)^2\right]^{-1}\right)_{n=1,\ldots,N}$$

Calculer les paramètres à l'aide l'algorithme EM

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

Incrémenter

$$t \leftarrow t + 1$$

end

L'algorithme 1 alterne le calcul des composantes et celui des autres paramètres via EM. Cet algorithme est implémenté dans la librairie R **FactorSCGLR** disponible en suivant le lien https://github.com/julien-gibaud/FactorSCGLR. Des essais numériques, sur données simulées et réelles, que nous nous contenterons de résumer ici, seront exposés lors de la présentation orale de ce travail.

5 Expériences numériques

Différentes expériences numériques sont réalisées pour tester les performances de cette approche. Des réponses partageant de fortes corrélations conditionnelles positives ou négatives demanderaient à être groupées. Ainsi, grâce au positionnement multidimensionnel et à une distance fondée sur la matrice de corrélation conditionnelle issue de la matrice de variance-covariance conditionnelle modélisée par le modèle à facteurs, nous proposons d'identifier ces groupes au travers du RI et de l'ARI. Ainsi dans une première expérience, nous modélisons une matrice réponse incluant plusieurs distributions prédite par un partitionnement des variables explicatives en deux thèmes distincts. Dans une deuxième expérience où les variables réponses sont toutes binaires, nous comparons ensuite le temps de calcul, le RI et l'ARI de notre méthode avec la librairie R **gllvm** (Niku et al., 2019). Cependant, dans cette expérience, dans un objectif de comparaison, nous nous limitons à un petit nombre de variables explicatives encapsulées dans un seul thème.

Cette nouvelle approche a ensuite été testée sur un jeu de données constitué de mesures (abondance, richesse, diversité) réalisées sur des communautés de carabidés et de plantes vasculaires dans des champs de céréales des Vallées et Coteaux de Gascogne. Pour prédire cette biodiversité agricole, des variables explicatives réparties en quatre thèmes ont été récoltées. Le potentiel de prédation, l'intensité fermière et l'hétérogénéité paysagère liée aux couverts semi-naturels et à la mosaïque des cultures sont les thèmes incorporés dans la modélisation. Après utilisation de la méthode, il apparait que seul le thème d'intensité fermière est pertinent dans la prédiction de l'agrobiodiversité. Plus particulièrement, ce sont les variables explicatives représentant le traitement par herbicides, le nombre d'opérations effectuées par les fermiers, la profondeur du labourage et la quantité d'azote qui sont le plus impliquées dans cette prédiction. Les groupes de réponses identifiés grâce à la matrice de corrélation conditionnelle sont au nombre de quatre. Le premier regroupe des mesures faites sur les carabidés possédant de très fortes corrélations conditionnelles tandis que les autres groupes sont composés d'un mélange entre plantes et carabidés.

Remerciements

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

References

Bry, X., Trottier, C., Mortier, F., and Cornu, G. (2020). Component-based regularization of a multivariate GLM with a thematic partitioning of the explanatory variables. *Statistical Modelling*, 20(1):96–119.

Bry, X., Trottier, C., Verron, T., and Mortier, F. (2013). Supervised component general-

ized linear regression using a PLS-extension of the Fisher scoring algorithm. *Journal of Multivariate Analysis*, 119:47–60.

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

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Geweke, J. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies*, 9(2):557–587.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall.

Niku, J., Hui, F. K., Taskinen, S., and Warton, D. I. (2019). gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods in Ecology and Evolution*, 10(12):2173–2182.

Saidane, M., Bry, X., and Lavergne, C. (2013). Generalized linear factor models: A new local EM estimation algorithm. *Communications in Statistics-Theory and Methods*, 42(16):2944–2958.