

EXTENSION DE LA RÉGRESSION LINÉAIRE GÉNÉRALISÉE SUR COMPOSANTES SUPERVISÉES (SCGLR) AUX DONNÉES GROUPEES.

Jocelyn Chauvet ¹ & Catherine Trottier ¹ & Xavier Bry ¹ & Frédéric Mortier ²

¹ IMAG, Université de Montpellier - Place Eugène Bataillon - 34095 Montpellier cedex 5, jocelyn.chauvet@umontpellier.fr, catherine.trottier@univ-montp3.fr,

xavier.bry@univ-montp2.fr

² Cirad, UR BSEF, Campus International de Baillarguet - TA C-105/D - 34398 Montpellier, frederic.mortier@cirad.fr

Résumé. Nous proposons de construire des composantes permettant de régulariser un Modèle Linéaire Généralisé Mixte (GL2M) multivarié. Un ensemble de réponses aléatoires Y est modélisé par un GL2M, au moyen d'un ensemble X de variables explicatives, et d'un ensemble T de variables additionnelles. Les variables explicatives dans X sont supposées nombreuses et redondantes : il est donc nécessaire de régulariser la régression linéaire généralisée mixte. À l'inverse, les variables de T sont supposées peu nombreuses et sélectionnées de sorte à n'exiger aucune régularisation. La régularisation consiste ici à construire un nombre approprié de composantes orthogonales permettant tout à la fois une bonne modélisation de Y et l'extraction d'informations structurelles dans X . Pour cela, nous proposons d'insérer à chaque étape de l'algorithme de Schall permettant l'estimation d'un GL2M, l'optimisation d'un critère propre à SCGLR. Cette extension de la méthode SCGLR est testée et comparée à d'autres méthodes de régularisation de type Ridge et Lasso, sur données simulées et réelles.

Mots-clés. Modèles à composantes, GL2M multivarié, Effet aléatoire, Pertinence structurelle, Régularisation, SCGLR.

Abstract. We address component-based regularisation of a multivariate Generalized Linear Mixed Model. A set of random responses Y is modelled by a GLMM, using a set X of explanatory variables and a set T of additional covariates. Variables in X are assumed many and redundant : generalized linear mixed regression demands regularisation with respect to X . By contrast, variables in T are assumed few and selected so as to demand no regularisation. Regularisation is performed building an appropriate number of orthogonal components that both contribute to model Y and capture relevant structural information in X . We propose to optimize a SCGLR-specific criterion within a Schall's algorithm in order to estimate the model. This extension of SCGLR is tested on simulated and real data, and compared to Ridge- and Lasso-based regularisations.

Keywords. Component-models, Multivariate GLMM, Random effect, Structural Relevance, Regularisation, SCGLR.

1 Données, modélisation et problème.

Nous considérons un ensemble de q réponses aléatoires $Y_{n \times q} = \{y^1, \dots, y^q\}$, expliquées par deux ensembles de covariables $X_{n \times p} = \{x^1, \dots, x^p\}$ et $T_{n \times r} = \{t^1, \dots, t^r\}$ aux statuts bien différents. Les p variables explicatives contenues dans X sont supposées nombreuses et redondantes tandis que les r variables additionnelles contenues dans T sont supposées peu nombreuses et sélectionnées de sorte à éviter les redondances. Les variables dans T peuvent alors être conservées dans le modèle sans traitement particulier. Par contre, on suppose que seulement quelques dimensions $H < p$ dans X suffisent à capter la majorité de l'information utile pour modéliser Y . En se basant sur les structures fortes de X , il s'agira de construire un nombre approprié de composantes orthogonales, dans le but de prédire au mieux Y .

Considérant des distributions de réponses aléatoires appartenant à la famille exponentielle, chacune des y^k est modélisée selon un GL2M (McCulloch, Searle (2001)). Plus précisément, les situations considérées sont celles où les n unités statistiques ne sont pas indépendantes mais structurées en N groupes. L'effet aléatoire introduit aura précisément pour objectif de modéliser la dépendance à l'intérieur de chacun des groupes. Précisons de plus que les réponses Y sont supposées indépendantes conditionnellement à $X \cup T$.

De nombreuses méthodes de régularisation dans le cadre de la modélisation GLM (McCullagh, Nelder (1989)) ont été développées ces dernières années. Dans le cadre univarié, c'est à dire lorsque $Y = \{y\}$, Bastien et al. (2004) proposent de combiner, à la manière de PLS, les régressions linéaires généralisées de la variable dépendante sur chacun des régresseurs pris isolément. Cependant, la technique en question, appelée PLSGLR, semble insatisfaisante car elle ne tient pas compte de la variance induite par la modélisation GLM. Toujours dans le cadre univarié, Marx (1996) introduit la stratégie nommée IRPLS, qui tient compte de la matrice de poids des observations provenant de la modélisation GLM pour construire les composantes PLS. Dans la lignée de Marx (1996), Bry, Trottier et al. (2013) proposent SCGLR, pour Régression Linéaire Généralisée sur Composantes Supervisées. Cette extension multivariée de la méthode précédente recherche des composantes communes à tous les y^k . Pour les construire, un nouveau critère est introduit et maximisé dans chacune des étapes de l'algorithme des scores de Fisher (FSA).

Cependant, toutes ces méthodes se fondent sur une hypothèse importante : l'indépendance des n unités statistiques intervenant dans la modélisation. On se propose ici d'étendre la méthode SCGLR en y introduisant une structure de dépendance entre les unités statistiques via un effet aléatoire.

L'une des motivations de ce travail est la modélisation et la prédiction de l'abondance d'espèces d'arbres dans la forêt tropicale du bassin du Congo. En effet, les mesures d'abondance d'espèces étant naturellement organisées en groupes dans l'espace (les concessions forestières dans nos données), la structure de dépendance induite doit être prise en compte dans les modèles.

2 Retour sur SCGLR avec variables additionnelles.

Nous nous plaçons dans cette partie dans le cas où chacune des variables réponses y^k est modélisée selon un GLM. Pour simplifier, nous ne présenterons que la recherche de la première composante ($H = 1$).

Le premier fondement conceptuel de SCGLR consiste à chercher dans X une composante commune notée $f = Xu$, optimale pour l'ensemble des y^k . Les prédicteurs du FSA sont donc partiellement colinéaires :

$$\forall k \in \{1, \dots, q\}, \eta^k = (Xu)\gamma_k + T\delta_k,$$

avec, pour des questions d'identification, $u' Au = 1$ où A est une matrice symétrique définie positive. Si l'on suppose l'indépendance des réponses y^k et des n unités statistiques, la vraisemblance s'écrit simplement :

$$L(y|\eta) = \prod_{i=1}^n \prod_{k=1}^q L_k(y_i^k | \eta_i^k)$$

En raison du terme $\gamma_k u$, le modèle linéarisé construit à chaque étape du FSA n'est pas linéaire : une procédure de moindres carrés alternés est donc adoptée. En notant z^k les variables de travail classiques du FSA et W_k^{-1} leurs matrices de variance, le programme suivant doit être considéré :

$$\mathcal{Q}_1 : \max_{u' Au=1} \psi_T(u), \quad \text{avec} \quad \psi_T(u) = \sum_{k=1}^q \|z^k\|_{W_k}^2 \cos^2_{W_k}(z^k, \langle Xu, T \rangle)$$

Le deuxième fondement conceptuel de SCGLR consiste à introduire une mesure de la proximité de la composante $f = Xu$ aux structures fortes de X : la pertinence structurelle. Avec W la matrice des poids des n unités statistiques, la composante la plus pertinente serait solution du programme :

$$\mathcal{Q}_2 : \max_{u' Au=1} \phi(u), \quad \text{avec} \quad \phi(u) = \left(\sum_{j=1}^p \langle Xu | x^j \rangle_W^{2l} \right)^{\frac{1}{l}} = \left(\sum_{j=1}^p (u' X' W x^j x^{j'} W Xu)^l \right)^{\frac{1}{l}},$$

le paramètre $l \in [1, +\infty]$ permettant de régler la localité des faisceaux de variables visés.

La stratégie SCGLR se propose alors de faire une synthèse entre les programmes $\mathcal{Q}_1$ et $\mathcal{Q}_2$, et de choisir la composante $f = Xu$ solution du programme :

$$\mathcal{Q} : \max_{u' Au=1} [\phi(u)]^s [\psi_T(u)]^{1-s}$$

Le paramètre s permet de régler l'importance relative accordée à la pertinence structurelle (programme $\mathcal{Q}_2$) par rapport à la qualité d'ajustement (programme $\mathcal{Q}_1$).

3 Adaptation de SCGLR aux données groupées.

Nous proposons une adaptation de la méthode SCGLR aux données groupées, pour lesquelles l'hypothèse d'indépendance des unités statistiques n'est plus valable. On peut penser par exemple aux données structurées en groupes dans l'espace ou aux données répétées dans le temps (données longitudinales). Nous modélisons la dépendance induite par un effet aléatoire, d'où une modélisation via un GL2M de chacune des réponses y^k .

3.1 Construction de la première composante.

Nous proposons de conserver la propriété de colinéarité partielle (sur X) des prédicteurs. L'injection d'un effet aléatoire dans chaque prédicteur permettra de traduire la structure de dépendance des unités statistiques. La matrice U contenant les données d'appartenance aux groupes, les prédicteurs s'écrivent alors :

$$\forall k \in \{1, \dots, q\}, \quad \eta_{\xi}^k = (Xu)\gamma_k + T\delta_k + U\xi_k, \text{ avec } u' Au = 1.$$

Les effets aléatoires $\xi_1, \dots, \xi_q$ sont supposés indépendants, et pour tout $k \in \{1, \dots, q\}$, ξ_k est supposé distribué selon une loi $\mathcal{N}_N(0, D_k = \sigma_k^2 Id)$, avec N le nombre de groupes.

La complexification de la structure de variance ne permet plus d'appliquer l'algorithme des scores de Fisher, qui suppose l'indépendance des unités statistiques. Nous avons alors envisagé d'adapter l'algorithme de Schall (1991) en y introduisant, à chaque étape, la procédure alternée suivante :

- À u fixé, on estime les paramètres γ_k , δ_k et σ_k^2 par un système de Henderson (Henderson (1975)).
- À γ_k , δ_k et σ_k^2 fixés, on construit la composante $f = Xu$.

En effet, les variables de travail qui interviennent dans notre algorithme peuvent s'écrire sous la forme :

$$z_{\xi}^k = (Xu)\gamma_k + T\delta_k + U\xi_k + e^k.$$

À u fixé, les variables de travail z_{ξ}^k sont considérées, à la manière de Schall, comme modélisées par un L2M. Une estimation courante des paramètres γ_k , δ_k et σ_k^2 est alors accessible par la résolution d'un système de Henderson, dont une justification dans ce cadre se trouve notamment dans Stiratelli et al. (1984), au moyen d'un raisonnement bayésien.

Quant à la construction de la composante $f = Xu$, elle est toujours solution d'un programme pouvant se mettre sous la forme $\mathcal{Q}$, mais il est nécessaire de modifier l'expression $\psi_T(u)$. Conditionnellement à l'effet aléatoire ξ_k , la variable de travail z_{ξ}^k possède une

structure de GLM. Nous proposons alors de modifier le critère de qualité d'ajustement en tenant compte de la variance de z_ξ^k conditionnellement à ξ_k . Pour cela, nous posons :

$$\psi_T(u) = \sum_{k=1}^q \|z_\xi^k\|_{W_{\xi,k}}^2 \cos^2_{W_{\xi,k}}(z_\xi^k, \langle Xu, T \rangle), \quad \text{avec} \quad W_{\xi,k}^{-1} := \mathbb{V}(z_\xi^k | \xi_k)$$

3.2 Construction des composantes suivantes.

Supposons avoir construit les h premières composantes, et posons $F^h = \{f^1, \dots, f^h\}$. La $(h+1)^{\text{ème}}$ composante est conçue pour compléter au mieux les h composantes précédentes et T , autrement dit $T^h = F^h \cup T$. De plus, nous imposons l'orthogonalité entre f^{h+1} et F^h , via la contrainte :

$$F^{h'} W f^{h+1} = 0$$

La composante $f^{h+1} = Xu^{h+1}$ est donc obtenue en résolvant le programme suivant, dont u^{h+1} est solution :

$$\begin{cases} \text{Maximiser} & [\phi(u)]^s [\psi_{T^h}(u)]^{1-s} \\ \text{sous les contraintes :} & u' Au = 1 \text{ et } (F^{h'} W X)u = 0 \end{cases}$$

4 Tests numériques.

Pour illustrer le potentiel de la méthode, nous présenterons dans un premier temps des résultats sur données simulées et réelles impliquant une modélisation GL2M (notamment des données groupées poissonniennes, lien log).

Des comparaisons seront également effectuées avec d'autres méthodes de régularisation qui, elles, ne font pas appel à la construction de variables latentes (nos composantes), mais consistent à appliquer une pénalisation sur la norme des coefficients de régression. Deux de ces méthodes ont été retenues :

- Celle proposée par Eliot et al. (2011), qui étend la régression ridge au cas des modèles linéaires mixtes (L2M) univariés, non généralisés.
- Celle proposée par Groll et Tutz (2014), qui consiste à injecter une pénalisation en norme L^1 lors de l'ajustement de modèles GL2M, toujours dans le cadre univarié.

5 Conclusion.

L'extension de SCGLR proposée est un bon compromis entre une modélisation GL2M (possédant de bonnes qualités prédictives dans le cadre de données groupées mais très instable si les régresseurs sont fortement redondants) et des stratégies de type GLM sur composantes principales (qui ne tiennent pas compte des variables réponses pour la construction des composantes). Les qualités de SCGLR sont préservées, et l'extension présentée

permet de réagir face à des données groupées en produisant des modèles prédictifs robustes basés sur des composantes interprétables.

Bibliographie

- [1] C.E. McCulloch, S.R. Searle (2001). *Generalized, Linear, and Mixed Models*, John Wiley & Sons, New-York, USA.
- [2] P. McCullagh, J.A. Nelder (1989). *Generalized linear models*, Chapman and Hall, New York, USA.
- [3] P. Bastien, V. Esposito Vinzi, M. Tenenhaus (2004). *PLS generalized linear regression*, Computational Statistics & Data Analysis, 48(1) :17-46.
- [4] B. D. Marx (1996). *Iteratively reweighted partial least squares estimation for generalized linear regression*, Technometrics, 38(4) :374–381.
- [5] X. Bry, C. Trottier, T. Verron, F. Mortier (2013). *Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm*, Journal of Multivariate Analysis, 119(C) :47–60.
- [6] R. Schall (1991). *Estimation in generalized linear models with random effects*, Biometrika, 78(4) :719–727.
- [7] C.R. Henderson (1975). *Best linear unbiased estimators and prediction under a selection model*, Biometrics, 31 :423–447.
- [8] R. Stiratelli, N. Laird, J.H. Ware. *Random-Effects Models for Serial Observations with Binary Response*, Biometrics, 40(4) :961-971.
- [9] M. Eliot, J. Ferguson, M.P. Reilly and A.S. Foulkes (2011). *Ridge Regression for Longitudinal Biomarker Data*, The International Journal of Biostatistics, 7(1) :1-11.
- [10] A. Groll, G. Tutz (2014). *Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation*, Statistics and Computing, 24(2) :137-154.