EXTENSION DE LA RÉGRESSION LINÉAIRE GÉNÉRALISÉE SUR COMPOSANTES SUPERVISÉES À UNE PARTITION THÉMATIQUE DES RÉGRESSEURS.

Catherine Trottier[1] & Xavier Bry[2] & Frédéric Mortier[3] & Guillaume Cornu[3] & Thomas Verron[4]

[1] *Université Paul Valéry Montpellier, Route de Mende - 34199 Montpellier Cedex 5*
catherine.trottier@univ-montp3.fr
[2] *I3M, Université de Montpellier, Place Eugène Bataillon CC 051 - 34095 Montpellier*
[3] *Cirad, UR Biens et Services des Ecosystèmes Forestiers tropicaux, Campus International de Baillarguet -TA C-105 / D - 34398 Montpellier*
[4] *SEITA, Imperial Tobacco Group, Centre de recherche SCR, 48 rue Danton - 45404 Fleury-les-Aubrais*

Résumé. Nous proposons de construire des composantes permettant de régulariser un Modèle Linéaire Généralisé (GLM) multivarié. Un ensemble de réponses aléatoires Y est supposé dépendre, via un GLM, d'un ensemble X de variables explicatives, ainsi que d'un ensemble T de covariables additionnelles. X est partitionné en R blocs $X_1, ..., X_R$, conceptuellement homogènes, considérés comme autant de thèmes explicatifs. Les variables dans chaque X_r sont supposées nombreuses et redondantes. Il est donc nécessaire de régulariser la régression linéaire généralisée dans chaque thème. À l'inverse, les variables de T sont supposées peu nombreuses et sélectionnées de sorte à n'exiger aucune régularisation. On procède à la régularisation en cherchant dans chaque thème un nombre approprié de composantes orthogonales permettant de modéliser Y tout en extrayant une information structurelle pertinente dans chaque thème. Nous proposons un critère très général mesurant la pertinence structurelle d'une composante dans un thème, que nous introduisons dans l'algorithme des scores de Fisher d'estimation du modèle. La méthode, nommée THEME-SCGLR, est testée sur simulations et appliquée à la modélisation de l'abondance des espèces d'arbres dans la forêt tropicale du bassin du Congo.

Mots-clés. Modèles à composantes, GLM multivarié, Pertinence structurelle, Régularisation, SCGLR.

Abstract. We address component-based regularisation of a Multivariate Generalized Linear Model. A set of random responses Y is assumed to depend, through a GLM, on a set X of explanatory variables, as well as on a set T of additional covariates. X is partitioned into R conceptually homogenous blocks $X_1, \dots, X_R$, viewed as explanatory *themes*. Variables in each X_r are assumed many and redundant. Thus, generalised linear regression demands regularisation with respect to each X_r. By contrast, variables in T are assumed selected so as to demand no regularisation. Regularisation is performed searching each X_r for an appropriate number of orthogonal components that both contribute to model Y and capture relevant structural information in X_r. We propose a very general criterion to measure structural relevance of a component in a block, and show how to take structural relevance into account within a Fisher's scoring-type algorithm in order to estimate the model. The method, named THEME-SCGLR, is tested on simulated data and applied to model the abundancies of tree species in the Congo bassin rain forest.

Keywords. Component-Models, Multivariate GLM, Structural Relevance, Regularisation, SCGLR.

1 Data, Model and Problem

A set of q random responses $Y = \{y^1, \ldots, y^q\}$ is assumed to depend on p numeric regressors, partitioned into R blocks $X_1, \ldots, X_R$, with : $\forall r$, $X_r = \{x_r^1, \ldots, x_r^{p_r}\}$, plus one block T of additional covariates. Let $X := [X_1, \ldots, X_R]$. X and T may include the indicator variables of nominal explanatory variables. Every X_r may contain several unknown structurally relevant dimensions important to predict Y, how many we do not know. Variables in T are assumed to have been selected so as to preclude redundancy, while variables in the X_r's have not : T gathers all explanatory variables to be kept as such in the model, whereas dimension reduction and regularisation are needed in the X_r's. Each X_r is thus to be searched for an appropriate number of orthogonal components that both capture relevant structural information in X_r and contribute to model Y.

Each y^k is modelled through a GLM (as defined in [1]) taking $X \cup T$ as regressor set. Moreover, the y's are assumed independent conditional on $X \cup T$. All variables are measured on the same n units.

There has been attempts to deal with the particular case of $R = 1$ with T empty. In the univariate situation $Y = \{y\}$, Bastien et al. [2] combined generalised linear regression with univariate Partial Least Squares. In the multiple-y context, Bry [3] proposed an extension to GLM of Thematic Component Analysis. In our view, both methods lack consistency in estimation weightings. Still in the univariate situation, Marx [4] proposed a more consistent Iteratively Reweighted Partial Least Squares estimation. More recently, Bry et al. [5] extended the work by Marx [4] with a technique named Supervised Component-based Generalised Linear Regression (SCGLR). The basic principle of SCGLR is to replace the weighted least squares step of the Fisher's Scoring Algorithm (FSA) with an extended Partial Least Squares step. That way, component-based regularisation was introduced into GLM estimation. In this work, we propose to extend SCGLR by :

1. Introducing additional covariates.
2. Extending the notion of *structural relevance* of a component, so as to track various kinds of structures.
3. Extending SCGLR to the multiple-explanatory-block situation.

Let us use the following notations :

- Π_E^M := orthogonal projector on space *E*, with respect to some metric M.
- $\langle X \rangle$:= space spanned by the column-vectors of *X*.

2 Adapting the FSA to Estimate a Multivariate GLM with Partially Common Predictor

Let us assume only one block X. Searching for common components in X to explain several y's, we first have to adapt the classical FSA to predictors colinear in their *X*-parts :

$$\forall k = 1, \ldots, q : \ \eta_k \ = \ X\gamma_k u + T\delta_k$$

For identification, we impose $u'Au = 1$, where A may be any symetric definite positive matrix. In view of the conditional independence assumption, and independence of units :

$$l(y|\eta) = \prod_{i=1}^{n} \prod_{k=1}^{q} l_k(y_{ki}|\eta_{ki})$$

Due to the product $\gamma_k u$, the linearized model on each step of the deduced FSA is not linear and estimation has thus to be done through an alternated least squares step. Denoting z_k the classical working variables on each FSA's step, the solution of the following program is sought :

$$Q : \min_{f \in \langle X \rangle} \sum_k \|z_k - \Pi_{\langle f,T \rangle}^{W_k} z_k\|_{W_k}^2 \ ,$$

which is equivalent to program Q' :

$$Q' : \max_{u'Au=1} \psi(u) \ , \qquad \text{where} \qquad \psi(u) = \sum_k \|z_k\|_{W_k}^2 \cos_{W_k}^2(z_k \, , \, \langle Xu, T \rangle) \tag{1}$$

In order to later deal with multiple X_r's, we have yet to replace Q' by another equivalent program :

$$Q'' : \max_{\forall r, \, u_r' A_r u_r = 1} \psi(u_1, \ldots, u_R)$$

where $A_1, \ldots, A_R$ are any given symetric definite positive matrices, and $\psi(u_1, \ldots, u_R)$ is equal to :

$$\sum_k \|z_k\|_{W_k}^2 \cos_{W_k}^2(z_k \, , \, \langle X_1 u_1, \ldots, X_R u_R, T \rangle) \tag{2}$$

$\psi(u_1, \ldots, u_R)$ is a goodness-of-fit measure, now to be combined with some structural relevance measure to get regularisation.

3 Structural Relevance

Consider a given weight matrix W, e.g. $W = n^{-1}I_n$, reflecting the a priori importance of units. Let X be an $n \times p$ variable-block endowed with a $p \times p$ metric matrix M. Component $f = Xu$ is constrained by : $\|u\|^2_{M^{-1}} = 1$ (M^{-1} will thus be our choice of the aforementioned matrix A). We may consider various measures of structural relevance, according to the type of structure we want f to align with. Among them is the variable powered inertia, defined as follows.

We impose $\|f\|^2_W = 1$ through $M = (X'WX)^{-1}$. Let $l \geq 1$.
For a block X consisting of p standardised numeric variables x^j :

$$\phi(u) = \left(\sum_{j=1}^{p} \cos^{2l}(Xu, x^j)\right)^{\frac{1}{l}} = \left(\sum_{j=1}^{p} (u'X'Wx^j x^{j'} WXu)^l\right)^{\frac{1}{l}}$$

For $l = 1$, we get the part of X's variance captured by component f.
More generally, tuning parameter l allows to draw components towards more (greater l) or less (smaller l) local variable bundles. Fig. 1 graphs $\phi^l(v)$ in polar coordinates ($z(\theta) = \phi^l(e^{i\theta})e^{i\theta}$; $\theta \in [0, 2\pi]$) for various values of l in the elementary case of 4 coplanar variables x. One can see how the value of l tunes the locality of bundles considered.

For a block X consisting of p categorical variables X^j, each of which is coded through the set of its centred indicator variables (less one to avoid singularity of $X^{j'}WX^j$), we take :

$$\phi(u) = \left(\sum_{j=1}^{p} \cos^{2l}(Xu\,,\, \langle X^j \rangle)\right)^{\frac{1}{l}} = \left(\sum_{j=1}^{p} \langle Xu | \Pi^W_{X^j} Xu \rangle^l_W\right)^{\frac{1}{l}} .$$

4 THEME-SCGLR

We shall first consider the simpler case of a single explanatory block ($R = 1$), and then turn to the general case.

4.1 Dealing with a Single Explanatory Block

In order to regularise the regression corresponding to program Q' at each step of the FSA, we consider program :

$$R : \max_{u'M^{-1}u=1} \psi(u)^{1-s}\phi^s(u) \tag{3}$$

where $\psi(u)$ is given by (1) and s is a parameter tuning the relative importance of the structural relevance with respect to the goodness of fit. The product-form of the criterion is a straightforward way to make the solution insensitive to “size effects” of $\phi(u)$ and $\psi(u)$.

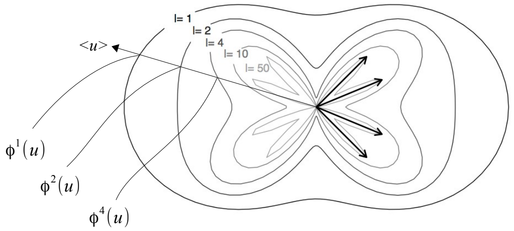

FIGURE 1 – Polar representation of the variable powered inertia according to the value of l.

Rank 1 component : is obtained by solving program (3) instead of performing the current step of the modified FSA given by (1). We developped an algorithm to maximise, at least locally, any criterion on the unit-sphere : the iterated normed gradient algorithm.

Rank $h > 1$ **component** : Let $F^h := \{f^1, \ldots, f^h\}$ be the set of the first h components. An extra component f^{h+1} must best complement the existing ones plus T, i.e. $T^h := F^h \cup T$. So f^{h+1} must be calculated using T^h as a block of extra-covariates. Moreover, we must impose that f^{h+1} be orthogonal to F^h, i.e. :

$$F^{h\prime} W f^{h+1} = 0 \tag{4}$$

4.2 Dealing with $R > 1$ Explanatory Blocks

Rank 1 component : Estimating the model in section 2 led to currently solving program Q''. Introducing structural relevance in it, we will now solve :

$$R'' : \max_{\forall r,\, u_r' M_r^{-1} u_r = 1} \psi(u_1, \ldots, u_R)^{1-s} \prod_{r=1}^{R} \phi^s(u_r) \tag{5}$$

where $\psi(u_1, \ldots, u_R)$ is given by (2). (5) can be solved by iteratively solving :

$$R_r : \max_{u_r' M_r^{-1} u_r = 1} \psi(u_r)^{(1-s)} \phi^s(u_r)$$

where $\psi(u_r)$ is calculated with $\tilde{T}_r = T \cup \{f_s; s \neq r\}$.

Rank $h > 1$ **component** : Suppose we want H_r components in X_r. $\forall r \in \{1, \ldots, R\}, \forall l < H_r$, let $F_r^l := \{f_r^h \,;\, h = 1, \ldots, l\}$. Component f_r^{h+1} must best complement the existing components (i.e. rank $< h+1$ ones in X_r and *all* components of all other blocks) plus T, i.e. : $T_r^h := F_r^h \cup_{s \neq r} F_s^{h_s} \cup T$. Informally, the algorithm consists in currently calculating all H_r components in X_r, as done in section 4.1 taking $T \cup_{s \neq r} F_s^{H_s}$ as extra-covariates, and then loop on r until overall convergence of the component-system is reached.

4.3 More material to present

Previous sections show the main ideas of THEME-SCGLR, but we shall also discuss :

- how to deal with mixed-type covariates, by adjusting metric M,
- how to get regression coefficients on the original variables from regression coefficients on the components,
- the principles we used for model assessment and model selection.

Finally, we will apply THEME-SCGLR on both simulated data and a real data set of abundancies of tree species in the Congo Bassin rain forest.

5 Conclusion

THEME-SCGLR is a powerful tradeoff between multivariate GLM estimation (which cannot afford many and redundant explanatory variables) and PCA-like methods (which take no explanatory model into account). Given a thematic model of the phenomenon under attention, it provides robust predictive models based on interpretable components. It also allows, through the exploration facilities it offers, to gradually refine the design of the model.

Bibliographie

[1] McCullagh, P., et Nelder, J.A. (1989). *Generalized linear models*, Chapman and Hall, New York, New York, USA.
[2] P. Bastien, V. Esposito Vinzi, M. Tenenhaus, “PLS generalized linear regression," *Computational Statistics & Data Analysis*, vol. 48(1), pp. 17-46, 2004.
[3] X. Bry, “Extension de l’Analyse en Composantes Thématiques univariée au modèle linéaire généralisé", *RSA* vol. 54(3), 2006
[4] B.D. Marx. “Iteratively Reweighted Partial Least Squares estimation for Generalized Linear Regression," *Technometrics*, vol. 38(4), pp. 374-381, 1996.
[5] X. Bry and C. Trottier and T. Verron and F. Mortier. “Supervized Component-based Generalized Linear Regression using a PLS-extended variant of the Fisher scoring algorithm," *Journal of Multivariate Analysis*, vol. 119, pp. 47-60, 2013.