Extension to mixed models of the Supervised Component-based Generalised Linear Regression

Jocelyn Chauvet, Catherine Trottier, Xavier Bry, Frederic Mortier

Extension to mixed models of the Supervised Component-based Generalised Linear Regression

Jocelyn Chauvet[1] **Catherine Trottier**[1,2] **Xavier Bry**[1]
Frédéric Mortier[3]

[1] Institut Montpelliérain Alexander Grothendieck,
CNRS, Univ. Montpellier, France.
jocelyn.chauvet@umontpellier.fr, xavier.bry@umontpellier.fr
[2] Univ. Paul-Valéry Montpellier 3,
F34000, Montpellier, France.
catherine.trottier@univ-montp3.fr
[3] Cirad, UR BSEF, Campus International de Baillarguet
TA C-105/D - 34398 Montpellier,
frederic.mortier@cirad.fr

Abstract

We address the component-based regularisation of a multivariate Generalized Linear Mixed Model (GLMM). A set of random responses Y is modelled by a GLMM, using a set X of explanatory variables, a set T of additional covariates, and random effects used to introduce the dependence between statistical units. Variables in X are assumed many and redundant, so that regression demands regularisation. By contrast, variables in T are assumed few and selected so as to require 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. To estimate the model, we propose to maximise a criterion specific to the Supervised Component-based Generalised Linear Regression (SCGLR) within an adaptation of Schall's algorithm. This extension of SCGLR is tested on both simulated and real data, and compared to Ridge- and Lasso-based regularisations.

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

1 Data, Model and Problem

A set of q random responses $Y = \{y^1, \ldots, y^q\}$ is assumed explained by two different sets of covariates $X = \{x^1, \ldots, x^p\}$ and $T = \{t^1, \ldots, t^r\}$, and a random effect $\xi = \{\xi^1, \ldots, \xi^q\}$. Explanatory variables in X are assumed many and redundant while additional covariates in T are assumed selected so as to preclude redundancy. Explanatory variables in T are thus kept as such in the model. By contrast, X may contain several unknow structurally relevant dimensions $K < p$ important to model and predict Y, how many we do not know. X is thus to be searched for an appropriate number of orthogonal components that both capture relevant structural information in X and contribute to model Y.

Each y^k is modelled through a Generalised Linear Mixed Model (GLMM) [7] assuming conditional distributions from the exponential family. More specifically in this work, the n statistical units are not considered independent but partitioned into N groups. The random effects included in the GLMM aim at modelling the dependence of units within each group.

Over the last decades, component-based regularisation methods for Generalised Linear Models (GLM) have been developped. In the univariate framework, i.e. when $Y = \{y\}$, Bastien et al. [1] proposed an extension to GLM of the classical Partial Least Square (PLS) regression, combining generalised linear regressions of the dependent variable on each of the regressors considered separately. However, doing so, this method does not take into account the variance structure of the overall model when building a component. Still in the univariate framework, Marx [6] proposed a more appropriate Iteratively Reweighted Partial Least Squares (IRPLS) estimation that builds PLS components using the weighting matrix derived from the GLM. More recently, Bry et al. [2] extended the work by Marx [6] to the multivariate framework with a technique named Supervised Component-based Generalised Linear Regression (SCGLR). The basic principle of SCGLR is to build optimal components common to all the dependent variables. To achieve it, SCGLR introduces a new criterion which is maximised at each step of the Fisher Scoring Algorithm (FSA).

Besides, regularisation methods have already been developped for GLMM, in which the random effects allow to model complex dependence structure. Eliot et al. [3] proposed to extend the classical ridge regression to Linear Mixed Models (LMM). The Expectation-Maximisation algorithm they suggest includes a new step to find the best shrinkage parameter - in the Generalised Cross-Validation (GCV) sense - at each iteration. More recently, Groll and Tutz [4] proposed an L_1-penalised algorithm for fitting a high-dimensional GLMM, using Laplace approximation and efficient coordinate gradient descent.

Instead of using a penalty on the norm of the coefficient vector, we propose to base the regularisation of the GLMM estimation on SCGLR-type components.

2 Reminder on SCGLR with additional covariates

In this section, we consider the simplified situation where each y^k is modelled through a GLM (without random effect) and only one component is calculated ($K = 1$). Moreover, 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.

M' : transpose of any matrix (or vector) M.

The first conceptual basis of SCGLR consists in searching for an optimal component $f = Xu$ common to all the y's. Therefore, SCGLR adapts the classical FSA to predictors

having colinear X-parts. To be precise, for each $k \in \{1, \ldots, q\}$, the linear predictor writes:

$$\eta^k = (Xu)\gamma_k + T\delta_k$$

where γ_k and δ_k are the parameters associated with component $f = Xu$ and covariates T respectively. For identification, we impose $u'Au = 1$, where A may be any symmetric definite positive matrix. Assuming that both the y's and the n statistical units are independent, the likelihood function L can be written:

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

where L_k is the likelihood function relative to y^k. Owing to the product $\gamma_k u$, the "linearised model" (LM) on each step of the associated FSA for the GLM estimation is not indeed linear: an alternated least squares step was designed. Denoting z^k the classical working variables on each FSA's step and W_k^{-1} their variance-covariance matrix, the least squares on the LM consists in the following optimisation (see [2]):

$$\min_{u'Au=1} \sum_{k=1}^{q} \left\| z^k - \Pi^{W_k}_{\langle Xu,T\rangle} z^k \right\|^2_{W_k} \quad \Longleftrightarrow \quad \max_{u'Au=1} \sum_{k=1}^{q} \left\| \Pi^{W_k}_{\langle Xu,T\rangle} z^k \right\|^2_{W_k}$$

which is also equivalent to :

$$\max_{u'Au=1} \psi_T(u), \quad \text{with} \quad \psi_T(u) = \sum_{k=1}^{q} \left\| z^k \right\|^2_{W_k} \cos^2_{W_k}\left(z^k, \langle Xu, T\rangle\right) \tag{1}$$

The second conceptual basis of SCGLR consists in introducing a closeness measure of the component $f = Xu$ to the strongest structures in X. Indeed, ψ_T is a mere goodness-of-fit measure, and must be combined with a structural relevance measure to get regularisation. Consider a given weight-matrix W - e.g. $W = \frac{1}{n} I_n$ - reflecting the a priori relative importance of units, the most structurally relevant component would be the solution of:

$$\max_{u'Au=1} \phi(u), \quad \text{with} \quad \phi(u) = \left(\sum_{j=1}^{p} \langle Xu|x^j\rangle^{2l}_W \right)^{\frac{1}{l}} = \left(\sum_{j=1}^{p} (u'X'Wx^j x^{j'}WXu)^l \right)^{\frac{1}{l}} \tag{2}$$

Tuning parameter l allows to draw components towards more (greater l) or less (smaller l) local variable bundles as depicted on Figure 1.

To sum things up, s being a parameter that tunes the importance of the structural relevance relative to the goodness-of-fit, SCGLR attempts a trade-off between (1) and (2), solving:

$$\max_{u'Au=1} \left[\phi(u)\right]^s \left[\psi_T(u)\right]^{1-s} \tag{3}$$

3 Adapting SCGLR to grouped data

We propose to adapt SCGLR to grouped data, for which the independence assumption of the statistical units is no longer valid (e.g. longitudinal or spatial correlated data). The whithin-group dependence is modelled by a random effect. Consequently, each of the y^k is ultimately modelled through a GLMM. We call this adaptation "Mixed-SCGLR".

3.1 Single component model estimation

We first present the method's principle. Then we give a Bayesian justification of the involved Henderson systems. Finally, we present our algorithm's steps.

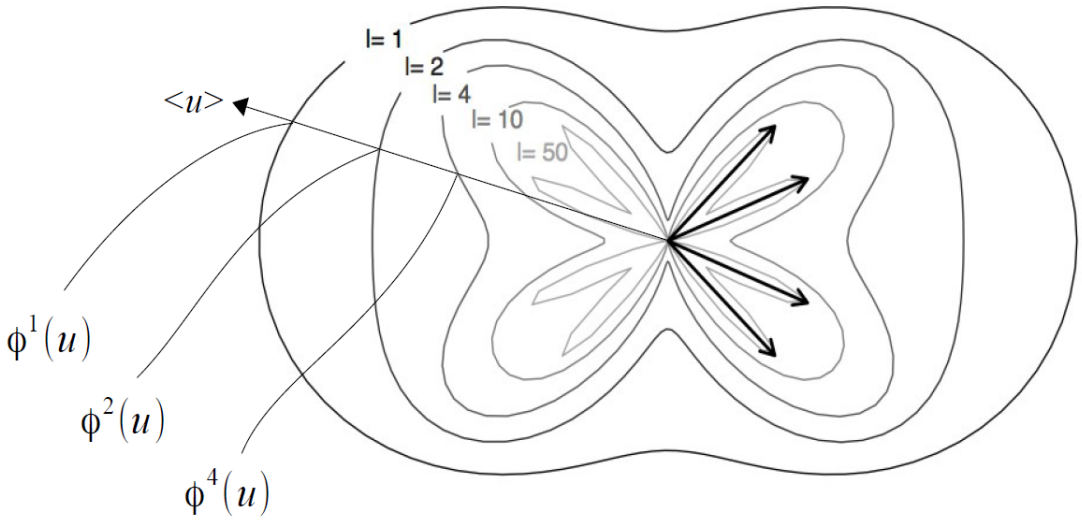

Figure 1: Polar representation of $\phi^l(u)$ according to the value of l, in the elementary case of four coplanar variables.

Principle We maintain predictors colinear in their X-parts. Introducing a random effect in each predictor and still imposing $u'Au = 1$, the linear predictors now write:

$$\forall k \in \{1, \ldots, q\}, \quad \eta_\xi^k = (Xu)\gamma_k + T\delta_k + U\xi^k \tag{4}$$

The random group effect is assumed different accross responses, yielding q random effects $\xi^1, \ldots, \xi^q$. These are assumed independent and normally distributed: $\mathcal{N}_N(0, D_k = \sigma_k^2 A_k)$, where N is the number of groups and A_k a known matrix ($A_k = I_N$ in general).

Owing to the GLMM dependence structure, the FSA was adapted by Schall [9]. We, in turn, adapt Schall's algorithm to our component-based predictor (4), by introducing the following alternated procedure at each step:

- Given γ_k, δ_k, ξ^k and σ_k^2, we build the component $f = Xu$ by solving a (3)-type program, which attempts a compromise between goodness-of-fit and structural relevance criterions.
- Given u, z_ξ^k being the classical working variables of the Schall's algorithm and $W_{\xi,k}^{-1}$ their conditional variance-covariance matrix, parameters γ_k, δ_k and ξ^k are estimated by solving the following Henderson system, which, subsequently, allows us to estimate σ_k^2 :

$$\begin{pmatrix} (Xu)'W_{\xi,k}(Xu) & (Xu)'W_{\xi,k}T & (Xu)'W_{\xi,k}U \\ T'W_{\xi,k}(Xu) & T'W_{\xi,k}T & T'W_{\xi,k}U \\ U'W_{\xi,k}(Xu) & U'W_{\xi,k}T & U'W_{\xi,k}U + D_k^{-1} \end{pmatrix} \begin{pmatrix} \gamma_k \\ \delta_k \\ \xi^k \end{pmatrix} = \begin{pmatrix} (Xu)'W_{\xi,k}\, z_\xi^k \\ T'W_{\xi,k}\, z_\xi^k \\ U'W_{\xi,k}\, z_\xi^k \end{pmatrix} \tag{5}$$

  We chose Henderson's method [5] since it is quicker than EM, for instance.

A Bayesian justification of the Henderson systems The conditional distribution of the data, given the random effects, is supposed to belong to the exponential family, i.e. for each $k \in \{1, \ldots, q\}$, the conditional density of $Y_i^k \,|\, \xi^k$ may be written:

$$f_{Y_i^k|\xi^k}\left(y_i^k, \theta_i^k\right) = \exp\left\{\frac{y_i^k\theta_i^k - b_k\left(\theta_i^k\right)}{a_{k,i}(\phi_k)} + c_k\left(y_i^k, \phi_k\right)\right\}$$

Linearisation step : Denoting G_k the link function of variable y^k, g_k its first derivative and μ_k the conditional expectation (i.e. $\mu^k := \mathbb{E}(Y^k \,|\, \xi^k)$), the working variables are obtained as:

$$z_\xi^k = \eta_\xi^k + e^k, \quad \text{where} \quad e_i^k = (y_i^k - \mu_i^k) g_k(\mu_i^k)$$

Their conditional variance-covariance matrices are:

$$W_{\xi,k}^{-1} := \text{Var}\left(z_\xi^k \,|\, \xi^k\right) = \text{Diag}\left(\left[g_k\left(\mu_i^k\right)\right]^2 \text{Var}\left(Y_i^k \,|\, \xi^k\right)\right)_{i=1,\ldots,n}$$

Estimation step : Our estimation step is based on the following lemma about normal hierarchy:

Lemma 1 *Given*

$$y|\theta \sim \mathcal{N}(M\theta, R)$$
$$\theta \sim \mathcal{N}(\alpha, \Omega)$$

the posterior distribution is $\theta|y \sim \mathcal{N}(\hat{\theta}, C)$*, where* $C = \left(M'R^{-1}M + \Omega^{-1}\right)^{-1}$ *and* $\hat{\theta}$ *satisfies:*

$$C^{-1}\hat{\theta} = M'R^{-1}y + \Omega^{-1}\alpha. \tag{6}$$

Given u, we just apply Schall's method [9] with our regularised linear predictors (4), which is equivalent to consider the following modelling:

$$z_\xi^k|\gamma_k, \delta_k, \xi^k \sim \mathcal{N}\left((Xu)\gamma_k + T\delta_k + U\xi^k,\ W_{\xi,k}^{-1}\right) \tag{7}$$
$$\left(\gamma_k, \delta_k, \xi^k\right) \sim \mathcal{N}\left(\begin{pmatrix}\gamma_k^{(0)}\\ \delta_k^{(0)}\\ 0\end{pmatrix}, \begin{pmatrix}B_k^\gamma & 0 & 0\\ 0 & B_k^\delta & 0\\ 0 & 0 & D_k\end{pmatrix}\right)$$

We suggest to choose a noninformative prior distribution for parameters γ_k et δ_k (as inspired by Stiratelli et al. [10]) imposing $B_k^{\gamma\,-1} = B_k^{\delta\,-1} = 0$. Current estimates of parameters γ_k, δ_k and ξ^k are thus obtained solving (6), which is equivalent to the Henderson system (5).

Finally, as mentioned in [9], given estimate $\widehat{\xi^k}$ for ξ^k, we have the following updates for the maximum likelihood estimation of the variance parameters σ_k^2, $k \in \{1, \ldots, q\}$:

$$\sigma_k^2 \longleftarrow \frac{\widehat{\xi^k}' A_k^{-1} \widehat{\xi^k}}{N - \frac{1}{\sigma_k^2} tr\left(A_k^{-1} C_k\right)} \quad \text{where} \quad C_k = \left(U'W_{\xi,k}U + D_k^{-1}\right)^{-1}$$

The algorithm Component $f = Xu$ is still found solving a (3)-type program, adapting the expression of ψ_T. Indeed, conditional on the random effects ξ^k, the working variables z_ξ^k are assumed normally distributed according to (7). We thus modify the previous goodness-of-fit measure, taking into account the variance of z_ξ^k conditional on ξ^k. In case of grouped data,

$$\psi_T(u) = \sum_{k=1}^{q} \left\|z_\xi^k\right\|_{W_{\xi,k}}^2 \cos_{W_{\xi,k}}^2\left(z_\xi^k, \langle Xu, T\rangle\right) \tag{8}$$

Step 1 Computation of the component

Set:

$$u^{[t]} = \underset{u'Au=1}{\arg\max}\ [\phi(u)]^s\ [\psi_T(u)]^{1-s} \quad \text{where } \psi_T \text{ is defined by (8)}$$
$$f^{[t]} = Xu^{[t]}$$

Step 2 Henderson systems

For each $k \in \{1, \ldots, q\}$, solve the following system:

$$\begin{pmatrix} f^{[t]\prime} W^{[t]}_{\xi,k} f^{[t]} & f^{[t]\prime} W^{[t]}_{\xi,k} T & f^{[t]\prime} W^{[t]}_{\xi,k} U \\ T' W^{[t]}_{\xi,k} f^{[t]} & T' W^{[t]}_{\xi,k} T & T' W^{[t]}_{\xi,k} U \\ U' W^{[t]}_{\xi,k} f^{[t]} & U' W^{[t]}_{\xi,k} T & U' W^{[t]}_{\xi,k} U + {D^{[t]}_k}^{-1} \end{pmatrix} \begin{pmatrix} \gamma_k \\ \delta_k \\ \xi^k \end{pmatrix} = \begin{pmatrix} f^{[t]\prime} W^{[t]}_{\xi,k}\, z^{k[t]}_\xi \\ T' W^{[t]}_{\xi,k}\, z^{k[t]}_\xi \\ U' W^{[t]}_{\xi,k}\, z^{k[t]}_\xi \end{pmatrix}$$

Call $\gamma^{[t]}_k$, $\delta^{[t]}_k$ and $\xi^{k[t]}$ the solutions.

Step 3 Updating variance parameters

For each $k \in \{1, \ldots, q\}$, compute:

$$\sigma^{2[t+1]}_k = \frac{{\xi^{k[t]}}' A_k^{-1} \xi^{k[t]}}{N - \frac{1}{\sigma^{2[t]}_k} tr\left(A_k^{-1} C^{[t]}_k\right)} \quad \text{and} \quad D^{[t+1]}_k = \sigma^{2[t+1]}_k A_k$$

Step 4 Updating working variables and weighting matrices

For each $k \in \{1, \ldots, q\}$, compute:

$$\begin{aligned} \eta^{k[t]} &= f^{[t]}\gamma^{[t]}_k + T\delta^{[t]}_k + U\xi^{k[t]} \\ \mu^{[t]}_{k,i} &= G_k^{-1}\left(\eta^{k[t]}_i\right),\ i = 1, \ldots, n \\ z^{k[t+1]}_i &= \eta^{k[t]} + \left(y^k_i - \mu^{[t]}_{k,i}\right) g_k\left(\mu^{[t]}_{k,i}\right),\ i = 1, \ldots, n \\ W^{[t+1]}_{\xi,k} &= Diag\left(\left(\left[g_k\left(\mu^{[t]}_{k,i}\right)\right]^2 \mathrm{Var}\left(Y^k_i \mid \xi^k\right)^{[t]}\right)^{-1}\right)_{i=1,\ldots,n} \end{aligned}$$

Step 1–4 are repeated until stability of u and parameters γ_k, δ_k and σ^2_k is reached.

Algorithm 1: Current iteration of the single component Mixed-SCGLR

3.2 Extracting higher rank components

Let $F^h = \left\{f^1, \ldots, f^h\right\}$ 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 additional covariates. Moreover, we must impose that f^{h+1} be orthogonal to F^h, i.e.:

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

Component $f^{h+1} := Xu^{h+1}$ is thus obtained solving:

$$\begin{cases} \max \quad [\phi(u)]^s\ [\psi_{T^h}(u)]^{1-s} \\ \text{subject to:} \quad u'Au = 1 \text{ and } D^{h\prime} u = 0 \end{cases} \tag{9}$$

where $\psi_{T^h}(u) = \sum_{k=1}^{q} \left\|z^k_\xi\right\|^2_{W_{\xi,k}} \cos^2_{W_{\xi,k}}\left(z^k_\xi, \langle Xu, T^h\rangle\right)$ and $D^h = X'WF^h$.

In Appendix, we give an algorithm to maximise, at least locally, any criterion on the unit sphere: the Projected Iterated Normed Gradient (PING) algorithm. Varying the initialisation allows us to increase confidence that the maximum reached is global. It allows us to build components of rank $h > 1$ by solving programs (9) and also component of rank $h = 1$ if we impose $T^h = T$ and $D^h = 0$ in the aforementioned program.

4 Simulation study in the canonical Gaussian case

A simple simulation study is conducted to characterise the relative performances of Mixed-SCGLR and Ridge- and Lasso-based regularisations in the LMM framework [3, 4]. We focus on the multivariate case, i.e. several $y's$, with many redundant explanatory variables. To do so, two random responses $Y = \left\{y^1, y^2\right\}$ are generated, and explanatory variables X are simulated so as to contain three independent bundles of variables: X_1, X_2 and X_3. Each explanatory variable is assumed normally distributed with mean 0 and variance 1. The level of redundancy within each bundle is tuned with parameter $\tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. To be precise, correlations among explanatory variables whithin bundle X_j are:

$$corr(X_j) = \tau \mathbf{1}\mathbf{1}' + (1 - \tau)I$$

Besides, bundle X_1 (15 variables) models and predicts only y^1, bundle X_2 (10 variables) only y^2, while bundle X_3 (5 variables) is designed to be a bundle of noise. Considering no additional covariates ($T = 0$), we thus simulate Y as :

$$\begin{cases} y^1 = X\beta_1 + U\xi^1 + \varepsilon^1 \\ y^2 = X\beta_2 + U\xi^2 + \varepsilon^2 \end{cases} \tag{10}$$

We consider the case of $N = 10$ groups and $R = 10$ units per group. Consequently, the random-effect design matrix can be written: $U = I_N \otimes \mathbf{1}_R$. All variables whithin each bundle are assumed to contribute homogeneously to predict Y. Then our choice of the fixed parameters are:

$$\beta_1 = (\underbrace{0.3, \ldots, 0.3}_{5 \text{ times}}, \underbrace{0.4, \ldots, 0.4}_{5 \text{ times}}, \underbrace{0.5, \ldots, 0.5}_{5 \text{ times}}, \underbrace{0, \ldots, 0}_{15 \text{ times}})'$$

$$\beta_2 = (\underbrace{0, \ldots, 0}_{15 \text{ times}}, \underbrace{0.3, .., 0.3}_{3 \text{ times}}, \underbrace{0.4, .., 0.4}_{4 \text{ times}}, \underbrace{0.5, .., 0.5}_{3 \text{ times}}, \underbrace{0, \ldots, 0}_{5 \text{ times}})'$$

Finally, residual variability and within groups variability are fixed to $\sigma_k^2 = 1$. We thus simulate random effects and noise respectively as: $\xi^k \sim \mathcal{N}_N(0, \sigma_k^2 I_N)$ and $\varepsilon^k \sim \mathcal{N}_n(0, \sigma_k^2 I_n)$, where $k \in \{1, 2\}$ and $n = NR$.

On the whole, $M = 500$ simulations are conducted for each value of τ, according to model (10). Simulation m provides two fixed-effects estimations: $\hat{\beta}_1^{(m)}$ and $\hat{\beta}_2^{(m)}$. Unlike Mixed-SCGLR, both LMM-Ridge and (G)LMM-Lasso are not designed for multivariate responses: estimations are computed separately in these cases. Consequently, for each method, we decide to retain only the one which provides the lower relative error. Their mean over M simulations (MLRE) is defined as:

$$\text{MLRE} = \frac{1}{M} \sum_{m=1}^{M} \min \left(\frac{\left\| \hat{\beta}_1^{(m)} - \beta_1 \right\|^2}{\|\beta_1\|^2}, \frac{\left\| \hat{\beta}_2^{(m)} - \beta_2 \right\|^2}{\|\beta_2\|^2} \right)$$

In Table 1, we summarise the optimal regularisation parameters selected via cross-validation. Corresponding MLRE's are presented in Table 2 to which we added the results provided without regularisation. As expected, in both Ridge and Lasso regularisations,

the shrinkage parameter value increases with τ. On the other hand, the greater τ, the more Mixed-SCGLR (with $l = 4$ as recommended in [8]) focuses on the main structures in X that contribute to model Y. The average value of s is approximatively 0.5, which means that there is no significant preference between goodness-of-fit and structural relevance. Except for $\tau = 0.1$, Mixed-SCGLR provides the most precise fixed effect estimates despite the sophistication of the dependence structure and the high level of correlation among explanatory variables. Indeed, if there are no actual bundles in X ($\tau \simeq 0$), looking for structures in X may lead Mixed-SCGLR to be slightly less accurate. Conversely, the stronger the structures (high τ), the more efficient our method.

	(G)LMM-Lasso	LMM-Ridge	Mixed-SC(G)LR	
	Optimal shrinkage parameter	Optimal shrinkage parameter	Optimal number of components K	Optimal tuning parameter s
$\tau = 0.1$	63.4	24.1	25	0.53
$\tau = 0.3$	111.2	53.7	5	0.53
$\tau = 0.5$	171.3	73.2	3	0.51
$\tau = 0.7$	220.6	78.2	2	0.51
$\tau = 0.9$	254.9	84.9	2	0.52

Table 1: Optimal regularisation parameter values obtained by cross-validation over 500 simulations

	LMM	(G)LMM-Lasso	LMM-Ridge	Mixed-SC(G)LR
$\tau = 0.1$	0.141	0.083	0.090	0.094
$\tau = 0.3$	0.340	0.180	0.124	0.105
$\tau = 0.5$	0.686	0.413	0.150	0.059
$\tau = 0.7$	1.571	0.913	0.189	0.061
$\tau = 0.9$	5.022	2.431	0.261	0.050

Table 2: Mean Lower Relative Errors (MLRE's) associated with the optimal parameter values

In order to highlight the power of Mixed-SCGLR for model interpretation, we represent on Figure 2 the correlation scatterplots obtained for $\tau = 0.5$, $l = 4$, $s = 0.51$, and $K = 3$. It clearly appears that y^1 is explained by the first bundle and y^2 by the second. The third component calculated catches the third bundle, which appears to play no explanatory role.

5 Real data example in the canonical Poisson case

Genus is a dataset built from the "CoForChange" study, which was conducted on $n = 2600$ developped plots divided into $N = 22$ forest concessions. It gives the abundance of $q = 94$ common tree genera in the tropical moist forest of the Congo-Basin, $p = 56$ geo-referenced environmental variables, and $r = 2$ other covariates which describe geology and anthropogenic interference. Geo-referenced environmental variables are used to represent:

- 29 physical factors linked to topography, rainfall, or soil moisture,
- 25 photosynthesis activity indicators obtained by remote sensing: EVI (Enhanced Vegetation Index), NIR (Near InfraRed channel index) and MIR (Mid-InfraRed channel index)),

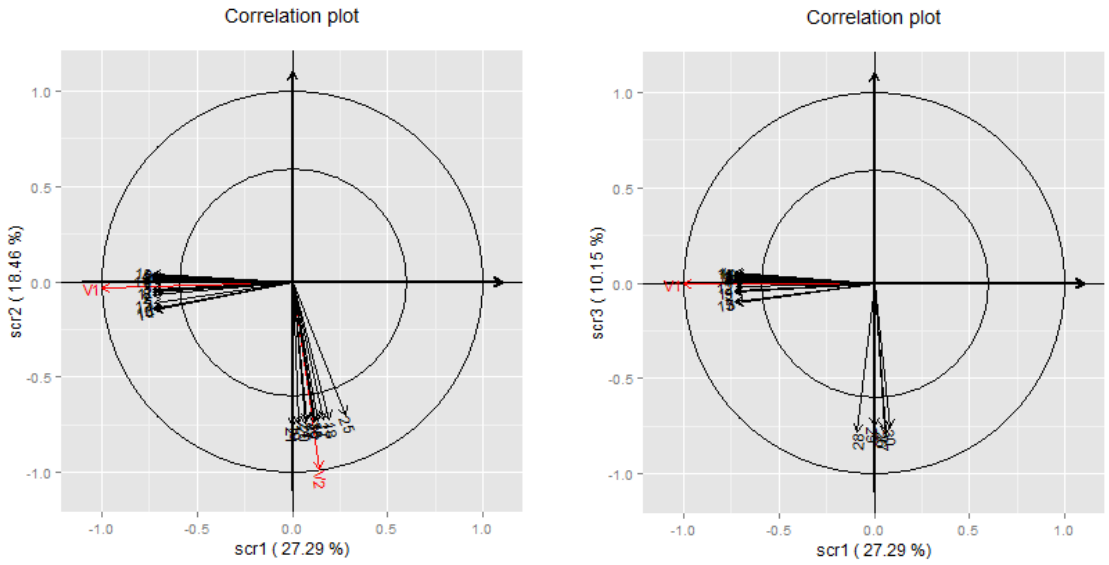

Figure 2: Correlation scatterplots given by Mixed-SCGLR method on the simulated data. The left hand side and the right hand side plots respectively show component planes $(1, 2)$ and $(1, 3)$. Both X-part linear predictors related to y^1 and y^2 are considered supplementary variables.

- 2 indicators which describe stand of trees height.

Physical factors are many and redundant (monthly rainfalls are highly correlated, for instance, and related to the geographic location). So are all photosynthesis activity indicators. Therefore, all these variables are put in X. By contrast, as geology and anthropogenic interference are weakly correlated and interesting per se, we put them in the set T of additional covariates.

It must be noted that the abundances of species given in *Genus* are count data. For each random response $y^1, \ldots, y^q$ we thus choose a Poisson regression with log link:

$$\forall k \in \{1, \ldots, q\}, \quad y^k \sim \mathcal{P}\left(\exp\left[\sum_{j=1}^{K}\left(Xu^j\right)\gamma_j + T\delta_k + U\xi^k\right]\right)$$

Among a series of parameter choices, values $l = 4$ and $s = 0.5$ prove to yield components very close to interpretable variable bundles. We therefore keep these parameter values in order to find - through a cross-validation procedure - the number of components K which minimises the Average Normalised Root Mean Square Error (AveNRMSE) defined as:

$$\text{AveNRMSE} = \frac{1}{q}\sum_{k=1}^{q}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i^k - \hat{y}_i^k}{\bar{y}^k}\right)^2}$$

On Figure 3, we plot the AveNRMSE's for $K \in \{0, 1, \ldots, 25\}$. As one can see, the best models are the ones with 12, 14 and 16 components. We retain the most parcimonious of them, i.e the one with 12 components. Two examples of correlation scatterplots we obtain are given on Figure 4, in which the X-parts of linear predictors are considered supplementary variables.

6 Conclusion

Mixed-SCGLR is a powerful trade-off between multivariate GLMM estimation (which cannot afford many and redundant explanatory variables) and PCA-like methods (which

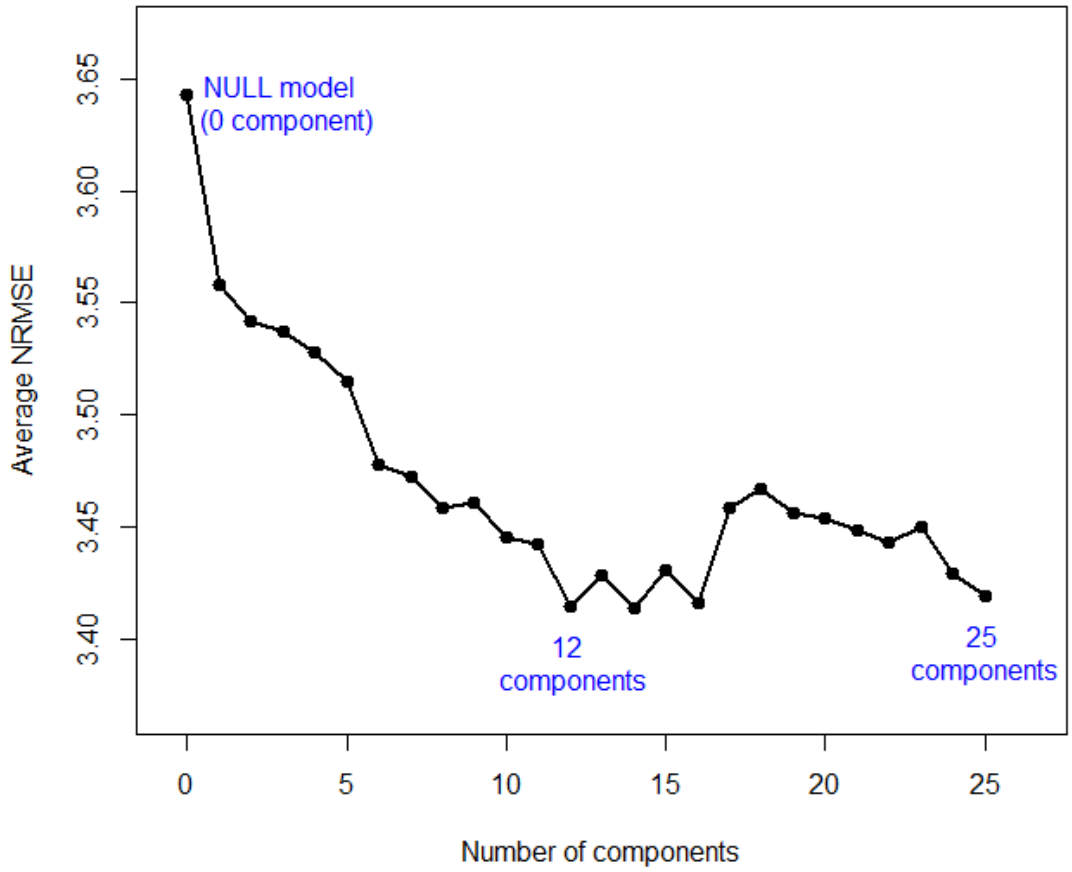

Figure 3: AveNRMSE's as a function of the number of components, obtained by a cross-validation procedure. The "null model" does not include any explanatory variables in X, but only additional covariates in T.

take no account of responses in building components). While preserving the qualities of the plain version of SCGLR, the mixed one performs well on grouped data, and provides robust predictive models based on interpretable components. Compared to penalty-based approaches as Ridge or Lasso, the orthogonal components built by Mixed-SCGLR reveal the multidimensional explanatory and predictive structures, and greatly facilitate the interpretation of the model.

Acknowledgement

The extended data *Genus* required the arrangement and the inventory of 140.000 developped plots across four countries : Central African Republic, Gabon, Cameroon and Democratic Republic of Congo. The authors thank the members of the CoForTips project for the use of this data.

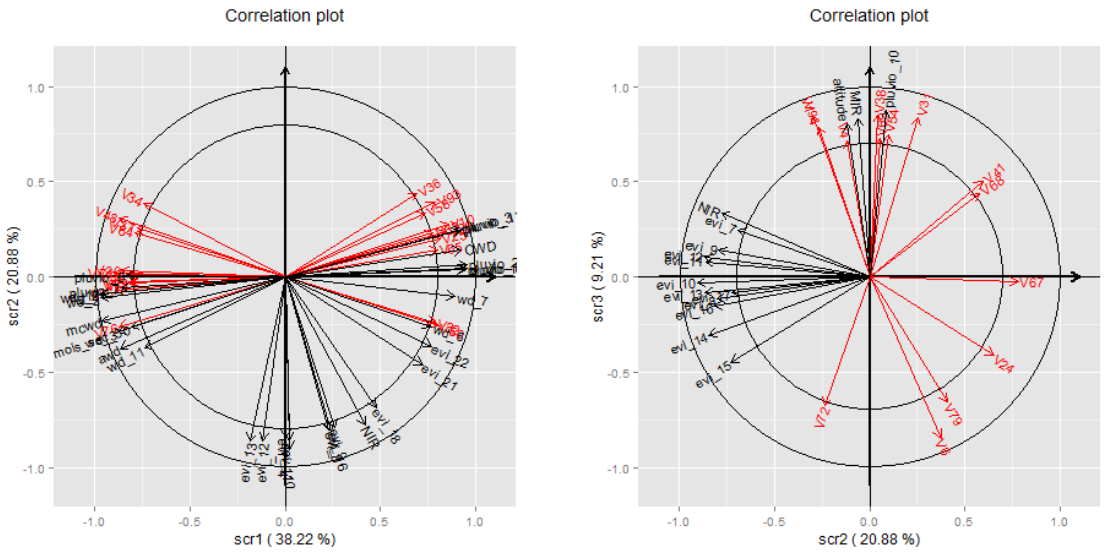

Figure 4: Two examples of correlations scatterplots on data *Genus*. The left hand side plot displays only variables having cosine greater than 0.8 with component plane $(1, 2)$. It reveal two patterns: a "rain-wind"-pattern driven by the Pluvio's and Wd's variables and a photosynthesis-pattern driven by the Evi's. On the right hand side, we plot variables having cosine greater than 0.7 with component plane $(2, 3)$. Component 3 reveals a bundle driven by variables Altitude, MIR and Pluvio10, which prove important to model and predict several y's.

Appendix

The Projected Iterated Normed Gradient (PING) is an extension of the iterated power algorithm, solving any program which has the form:

$$\max_{\substack{u'Au=1 \\ D'u=0}} h(u) \tag{11}$$

Note that putting $v := A^{1/2}u$, $g(v) := h(A^{-1/2}v)$ and $C := A^{-1/2}D$, program (11) is strictly equivalent to program (12):

$$\max_{\substack{v'v=1 \\ C'v=0}} g(v) \tag{12}$$

In our framework, the particular case $C = 0$ (no extra-orthogonality contrain) allows us to find the first rank component. Denoting $\Pi_{C^\perp} := I - (C'C)^{-1}C'$ and $\Gamma(v) := \nabla_v g(v)$, a Lagrange multiplier- based reasoning gives the basic iteration of the PING algorithm:

$$v^{[t+1]} = \frac{\Pi_{C^\perp}\Gamma\left(v^{[t]}\right)}{\left\|\Pi_{C^\perp}\Gamma\left(v^{[t]}\right)\right\|} \tag{13}$$

Despite the fact that iteration (13) follows a direction of ascent, it does not guarantee that g actually increases on every step. Algorithm PING therefore repeats the following steps until convergence of v is reached:

Step 1 Set: $\kappa^{[t]} = \dfrac{\Pi_{C^\perp}\Gamma\left(v^{[t]}\right)}{\left\|\Pi_{C^\perp}\Gamma\left(v^{[t]}\right)\right\|}$

Step 2 A Newton-Raphson unidimensional maximisation procedure is used to find the maximum of $g(v)$ on the arc $\left(v^{[t]}, \kappa^{[t]}\right)$ and take it as $v^{[t+1]}$.

References

[1] Bastien, P., Esposito Vinzi, V. and Tenenhaus, M. (2004) *PLS generalized linear regression.* Computational Statistics & Data Analysis, **48**, 17–46.

[2] Bry, X., Trottier, C., Verron, T. and Mortier, F. (2013) *Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm.* Journal of Multivariate Analysis, **119**, 47-60.

[3] Eliot, M., Ferguson, J., Reilly, M.P. and Foulkes, A.S. (2011) *Ridge Regression for Longitudinal Biomarker Data.* The International Journal of Biostatistics, **7**, 1–11.

[4] Groll, A. and Tutz, G. (2014) *Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation.* Statistics and Computing, **24**, 137–154.

[5] Henderson, C.R. (1975) *Best linear unbiaised estimators and prediction under a selection model.* Biometrics, **31**, 423-447.

[6] Marx, B.D. (1996) *Iteratively reweighted partial least squares estimation for generalized linear regression.* Technometrics, **38**, 374-381.

[7] McCulloch, C.E. and Searle, S.R (2001) *Generalized, Linear, and Mixed Models.* John Wiley & Sons.

[8] Mortier, F., Trottier, C., Cornu, G., Bry, X. (2015) *SCGLR - An R Package for Supervised Component Generalized Linear Regression.*

[9] Schall, R. (1991) *Estimation in generalized linear models with random effects.* Biometrika, **78**, 719-727.

[10] Stiratelli, R., Laird, N., and Ware, J.H. (1984) *Random-Effects Models for Serial Observations with Binary Response.* Biometrics, **40**, 961–971.