

# Supervised Component-based Generalised Linear Regression (SCGLR) for **grouped data**

**Jocelyn CHAUVET**

Joint work with Xavier Bry, Catherine Trottier & Frédéric Mortier

JDS 2016

June 02



- 1 Motivation
- 2 The Mixed-SCGLR method
- 3 Simulation study
- 4 Application to real data

# 1. Motivation

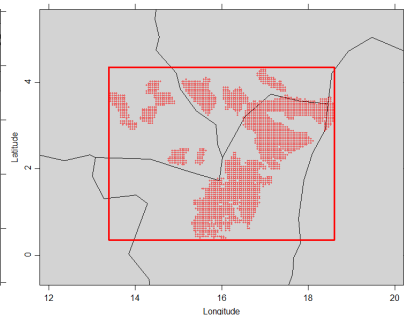
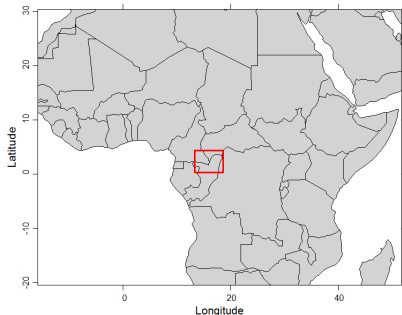
## Problem

Model and predict the **abundance of tree species** in the tropical moist forest of the Congo-Basin

# 1. Motivation

## Problem

Model and predict the **abundance of tree species** in the tropical moist forest of the Congo-Basin



On each plot :

- $q = 94$  common tree species (abundance : count data)

In order to model and explain it :

- $p = 56$  explanatory variables
- $r = 2$  additional covariates

On each plot :

- $q = 94$  common tree species (abundance : count data)

In order to model and explain it :

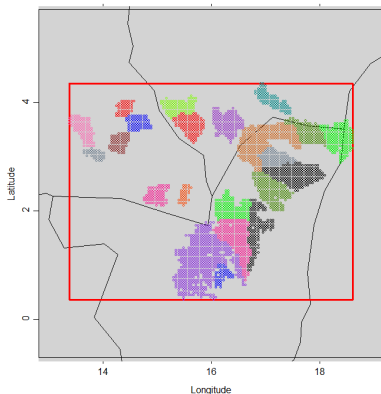
- $p = 56$  explanatory variables
- $r = 2$  additional covariates

## Difficulties

- High level of correlation among the explanatory variables  
↳ **Regularisation** is needed
- Spatial observations  
↳ Necessity to take account of the **dependence structure**

## Previous version of SCGLR

The 2615 plots **WERE ASSUMED INDEPENDENT**, although they are partitioned in 22 natural groups (forest concessions)



## Our new version of SCGLR

- Takes into account the dependence structure
  - ↳ Within-group dependence modelled by a **random effect**
  - ↳ **Multivariate GLMM**
- High correlations among the explanatory variables
  - ↳ **Supervised component-based regularisation**



## 2. The Mixed-SCGLR method

- General principle
- Construction of the component
- Algorithm

We focus on the single component model estimation

## Responses

- $\mathbf{Y}_{n \times q}$  : matrix of  $q$  responses  $y^1, \dots, y^q$

## Design matrices

- $\mathbf{X}_{n \times p}$  : explanatory variables (many and redundant)
- $\mathbf{T}_{n \times r}$  : additional covariates (few, no redundancy)
- $\mathbf{U}_{n \times N}$  : design matrix of the random effects

## Responses

- $Y_{n \times q}$  : matrix of  $q$  responses  $y^1, \dots, y^q$

## Design matrices

- $X_{n \times p}$  : explanatory variables (many and redundant)
- $T_{n \times r}$  : additional covariates (few, no redundancy)
- $U_{n \times N}$  : design matrix of the random effects

## Our linear predictors in the GLMM framework

For each  $k \in \{1, \dots, q\}$ ,

$$\eta_{\xi}^k = (X\mathbf{u})\gamma_k + T\delta_k + U\xi_k$$

$$\xi_k \stackrel{\text{ind.}}{\sim} \mathcal{N}_N(0, D_k = \sigma_k^2 Id_N), \text{ with } N \text{ the number of groups}$$

## Link function $g$

$$\eta_{\xi}^k = g\left(\mu_{\xi}^k\right) \text{ with } \mu_{\xi}^k = \mathbb{E}\left(Y^k \mid \xi_k\right)$$

## Working variables - classic local order 1 linearisation

$$g\left(y^k\right) \simeq z_{\xi}^k = \eta_{\xi}^k + \varepsilon_k$$

with :

$$\begin{cases} \mathbb{E}\left(\varepsilon_k \mid \xi_k\right) = 0 \\ \mathbb{V}\left(\varepsilon_k \mid \xi_k\right) \stackrel{\text{not.}}{=} W_k^{-1} \end{cases}$$

## Link function $g$

$$\eta_{\xi}^k = g\left(\mu_{\xi}^k\right) \text{ with } \mu_{\xi}^k = \mathbb{E}\left(Y^k \mid \xi_k\right)$$

## Working variables - classic local order 1 linearisation

$$g\left(y^k\right) \simeq z_{\xi}^k = \eta_{\xi}^k + \varepsilon_k$$

with :

$$\begin{cases} \mathbb{E}\left(\varepsilon_k \mid \xi_k\right) = 0 \\ \mathbb{V}\left(\varepsilon_k \mid \xi_k\right) \stackrel{\text{not.}}{=} W_k^{-1} \end{cases}$$

## "Linearised" model

$$z_{\xi}^k = (X u) \gamma_k + T \delta_k + U \xi_k + \varepsilon_k, \quad \text{with } \varepsilon_k \sim \mathcal{N}\left(0, W_k^{-1}\right)$$

## "Linearised" model

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

## Alternated procedure

- Given  $\gamma_k$ ,  $\delta_k$ ,  $\xi_k$  and  $\sigma_k^2$ , we calculate the component  $f = Xu$
- Given  $u$ , we estimate  $\gamma_k$ ,  $\delta_k$ ,  $\xi_k$  and  $\sigma_k^2$   
Several possibilities :
  - i Direct maximum likelihood estimation
  - ii Expectation-Maximisation algorithm
  - iii ...
  - iv **Henderson system** (more efficient)

## Henderson systems

Given  $\mathbf{f} = \mathbf{X}\mathbf{u}$ , for each  $k \in \{1, \dots, q\}$  :

$$\begin{pmatrix} f'W_k f & f'W_k T & f'W_k U \\ T'W_k f & T'W_k T & T'W_k U \\ U'W_k f & U'W_k T & U'W_k U + D_k^{-1} \end{pmatrix} \begin{pmatrix} \gamma_k \\ \delta_k \\ \xi_k \end{pmatrix} = \begin{pmatrix} f'W_k z_{\xi}^k \\ T'W_k z_{\xi}^k \\ U'W_k z_{\xi}^k \end{pmatrix}$$

and

$$\sigma_k^2 \leftarrow \frac{\xi_k' \xi_k}{N - \frac{1}{\sigma_k^2} \text{tr}(C_k)}$$

## Goodness-Of-Fit

$$\min RSS \iff \max_{u'u=1} \left\{ \psi(u) = \sum_{k=1}^q \left\| \Pi_{\langle Xu, T \rangle} z_{\xi}^k \right\|_{W_k}^2 \right\}$$



## Goodness-Of-Fit

$$\min RSS \iff \max_{u'u=1} \left\{ \psi(u) = \sum_{k=1}^q \left\| \Pi_{\langle Xu, T \rangle} z_{\xi}^k \right\|_{W_k}^2 \right\}$$

## Structural Relevance

- Classic dual PCA :  $\max_{u'u=1} \sum_{j=1}^p \rho^2(Xu, x^j)$

- Generalisation :  $\max_{u'u=1} \left\{ \phi(u) = \left( \sum_{j=1}^p [\rho^2(Xu, x^j)]^l \right)^{\frac{1}{l}} \right\}$

## Goodness-Of-Fit

$$\min RSS \iff \max_{u'u=1} \left\{ \psi(u) = \sum_{k=1}^q \left\| \Pi_{\langle Xu, T \rangle} z_{\xi}^k \right\|_{W_k}^2 \right\}$$

## Structural Relevance

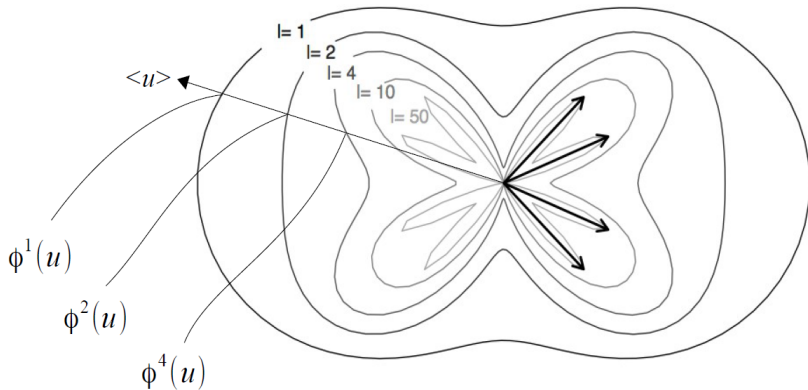
- Classic dual PCA :  $\max_{u'u=1} \sum_{j=1}^p \rho^2(Xu, x^j)$

- Generalisation :  $\max_{u'u=1} \left\{ \phi(u) = \left( \sum_{j=1}^p [\rho^2(Xu, x^j)]^l \right)^{\frac{1}{l}} \right\}$

## Compromise between both criterions

$$\max_{u'u=1} [\psi(u)]^{1-s} [\phi(u)]^s \quad \text{with } s \in [0, 1]$$

## Geometry of the Structural Relevance criterion



## Mixed-SCGLR algorithm (single component)

**Data :**  $X, T, Y$

**Result :** Single component model estimation

Initialise the working variables  $z_{\xi}^k$ , the weighting matrices  $W_k$ , and the variances  $\sigma_k^2$

**while**  $\langle$  convergence not achieved  $\rangle$  **do**

$$u \leftarrow \arg \max_{u'u=1} [\psi(u)]^{1-s} [\phi(u)]^s$$

$$f \leftarrow Xu$$

**for**  $k = 1$  to  $q$  **do**

Estimate parameters  $\gamma_k, \delta_k, \xi_k$  (Henderson system) and  $\sigma_k^2$

Update  $z_{\xi}^k$  and  $W_k$

**end**

**end**

## 3. Simulation study

- Data simulation
- Comparison with Ridge- and Lasso- based regularisations

The simulation and comparison are limited to the Gaussian case

## Random responses

Multivariate framework with only two responses :  $\mathbf{Y} = [y^1 | y^2]$

## Random responses

Multivariate framework with only two responses :  $Y = [y^1 | y^2]$

## Fixed effects

- No additional covariates :  $T = \emptyset$
- 30 explanatory variables  $\mathcal{N}(0, 1)$  :

$$X = \underbrace{x^1 \dots x^{15}}_{\substack{\text{bundle } X_1 \\ \hookrightarrow \text{predict } y^1}} \quad \underbrace{x^{16} \dots x^{25}}_{\substack{\text{bundle } X_2 \\ \hookrightarrow \text{predict } y^2}} \quad \underbrace{x^{26} \dots x^{30}}_{\substack{\text{bundle } X_3 \\ \hookrightarrow \text{noise}}}$$

- Within each bundle :

$$\text{cor}(x^j, x^k) = \begin{cases} 1 & \text{si } j = k \\ \tau & \text{si } j \neq k \end{cases} \quad \text{with } \tau \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$$

## Random effects

$N = 10$  groups,  $R = 10$  units per group ( $n = 100$  individuals)

$\hookrightarrow$  Design matrix  $U = Id_N \otimes \mathbf{1}_R$



## Random effects

$N = 10$  groups,  $R = 10$  units per group ( $n = 100$  individuals)  
 $\hookrightarrow$  Design matrix  $U = Id_N \otimes \mathbb{1}_R$

## Model

$$\begin{cases} y^1 = X\beta_1 + U\xi_1 + \varepsilon_1 \\ y^2 = X\beta_2 + U\xi_2 + \varepsilon_2 \end{cases}$$

with,  $\forall k \in \{1, 2\}, \forall i \in \{1, \dots, n\}$  :

- $\xi_{k,i} \sim \mathcal{N}(0, 1)$
- $\varepsilon_{k,i} \sim \mathcal{N}(0, 1)$

## Number of simulations

$M = 100$  samples for each value of  $\tau$ .

## LMM-Ridge (Univariate framework)



Eliot, M., Ferguson, J., Reilly, M.P. and Foulkes, A.S. (2011)  
*Ridge Regression for Longitudinal Biomarker Data.*


- ↪ Estimation : EM algorithm
- ↪ GCV at each step to find the best shrinkage parameter  $\lambda$

## LMM-Ridge (Univariate framework)

 Eliot, M., Ferguson, J., Reilly, M.P. and Foulkes, A.S. (2011) *Ridge Regression for Longitudinal Biomarker Data*.

- ↪ Estimation : EM algorithm
- ↪ GCV at each step to find the best shrinkage parameter  $\lambda$

## GLMM-Lasso (Univariate framework)

 Groll, A. and Tutz, G. (2014) *Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation*.

- ↪ Laplace approximation of the likelihood
- ↪ Estimation : efficient coordinate gradient descent

## Optimal regularisation parameters (10-folds CV)

	(G)LMM-Lasso shrinkage parameter $\lambda_{\text{lasso}}^*$	LMM-Ridge shrinkage parameter $\lambda_{\text{ridge}}^*$	Mixed-SC(G)LR number of component $K^*$		tuning parameter $s^*$
$\tau = 0.1$	65	24	25		0.50
$\tau = 0.3$	92	54	5		0.58
$\tau = 0.5$	124	73	3		0.70
$\tau = 0.7$	163	78	3		0.73
$\tau = 0.9$	175	85	2		0.80

## Robust comparison criterion : Mean Upper Relative Error (MURE)

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

## Robust comparison criterion : Mean Upper Relative Error (MURE)

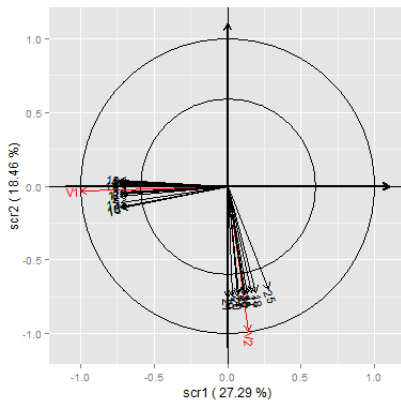
$$\text{MURE} = \frac{1}{M} \sum_{m=1}^M \max \left( \frac{\|\hat{\beta}_1^{(m)} - \beta_1\|^2}{\|\beta_1\|^2}, \frac{\|\hat{\beta}_2^{(m)} - \beta_2\|^2}{\|\beta_2\|^2} \right)$$

## MURE's associated with the optimal parameter values

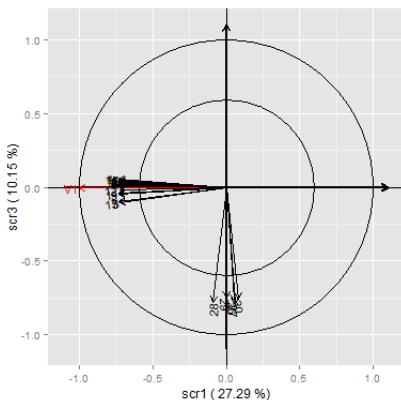
	LMM (no regularisation)	(G)LMM- Lasso	LMM- Ridge	Mixed- SC(G)LR
$\tau = 0.1$	0.12	0.05	0.08	0.12
$\tau = 0.3$	0.33	0.12	0.13	0.10
$\tau = 0.5$	0.61	0.20	0.16	0.07
$\tau = 0.7$	1.32	0.25	0.20	0.06
$\tau = 0.9$	4.62	0.26	0.31	0.05

## Example of scatterplots for $\tau = 0.5$ : Component planes (1,2) & (1,3).

Correlation plot



Correlation plot



## 4. Application to real data

### The *Genus* dataset

- $n = 2615$  developed plots, divided in  $N = 22$  forest concessions (considered as groups)
- $q = 94$  common tree genera (responses  $Y$ )
- $p = 56$  explanatory variables ( $X$ )
- $r = 2$  additional covariates ( $T$ )



## Modelisation

Abundance of tree species : count data

↪ Poisson regression with log link

$$\forall k \in \{1, \dots, q\}, \quad y^k \sim \mathcal{P} \left( \exp \left[ \sum_{j=1}^{K^*} (X u^j) \gamma_{k,j} + T \delta_k + U \xi_k \right] \right)$$

## Modelisation

Abundance of tree species : count data

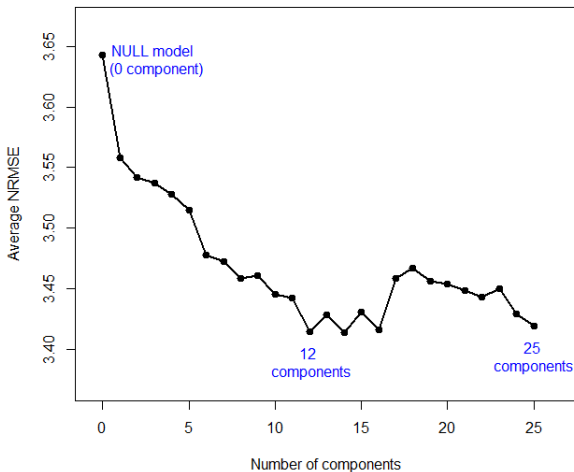
↪ Poisson regression with log link

$$\forall k \in \{1, \dots, q\}, \quad y^k \sim \mathcal{P} \left( \exp \left[ \sum_{j=1}^{K^*} (X u^j) \gamma_{k,j} + T \delta_k + U \xi_k \right] \right)$$

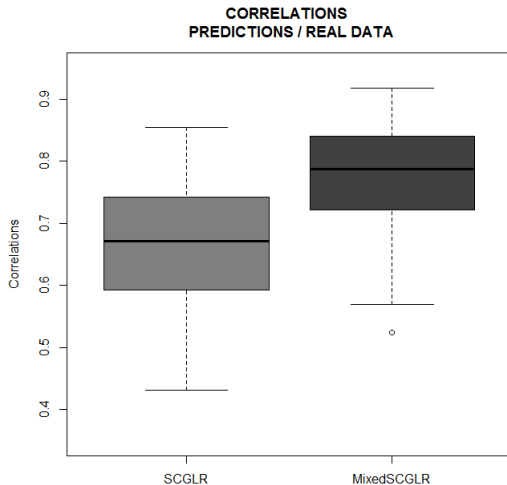
Optimal number of components  $K^*$

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

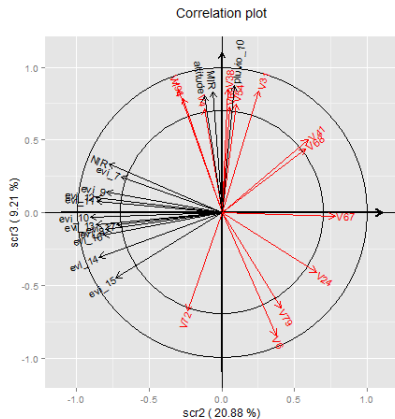
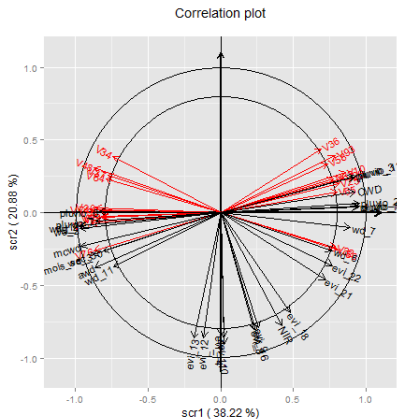
## AveNRMSE's as a function of the number of components



## Improvement in correlations of predictions and responses



## Component planes (1,2) & (2,3)



## Mixed-SCGLR





- is a powerful trade-off between
  - i **multivariate GLMM**
  - ii **component-based methods**
- reveals the **multidimensional explanatory and predictive structures**
- facilitates the **interpretation** of the model

## Mixed-SCGLR





- is a powerful trade-off between
  - i **multivariate GLMM**
  - ii **component-based methods**
- reveals the **multidimensional explanatory and predictive structures**
- facilitates the **interpretation** of the model

## Further development

- Mixed-SCGLR for **longitudinal data** :  $AR(p)$  random effect
- **Spatial autocorrelation** random effect

-  Bastien, P., Esposito Vinzi, V. and Tenenhaus, M. (2004) *PLS generalized linear regression*. Computational Statistics & Data Analysis, **48**, 17–46.
-  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.
-  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.
-  Groll, A. and Tutz, G. (2014) *Variable Selection for Generalized Linear Mixed Models by L1-Penalized Estimation*. Statistics and Computing, **24**, 137–154.



-  Henderson, C.R. (1975) *Best linear unbiased estimators and prediction under a selection model*. *Biometrics*, **31**, 423—447.
-  Marx, B.D. (1996) *Iteratively reweighted partial least squares estimation for generalized linear regression*. *Technometrics*, **38**, 374—381.
-  McCulloch, C.E. and Searle, S.R (2001) *Generalized, Linear, and Mixed Models*. John Wiley & Sons.
-  Schall, R. (1991) *Estimation in generalized linear models with random effects*. *Biometrika*, **78**, 719—727.