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

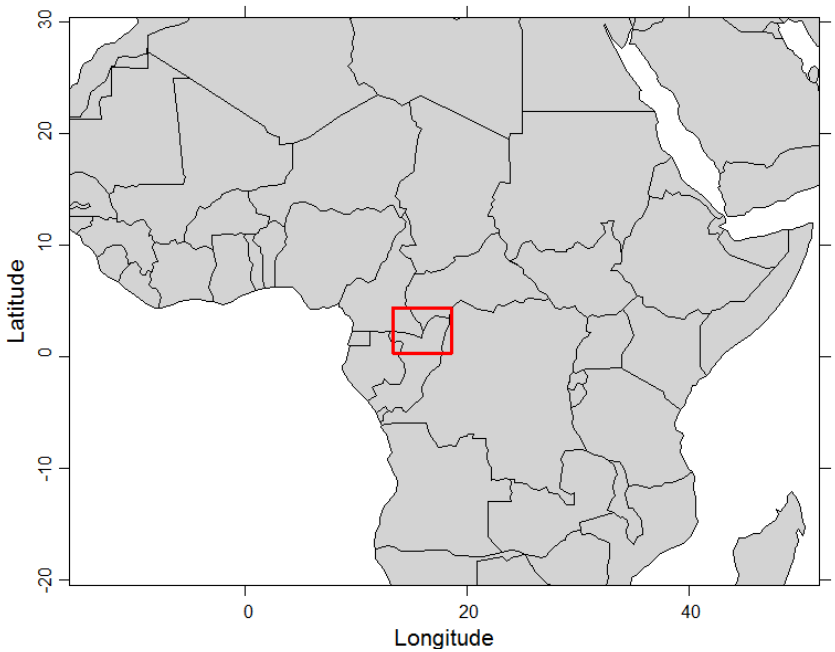
Problem

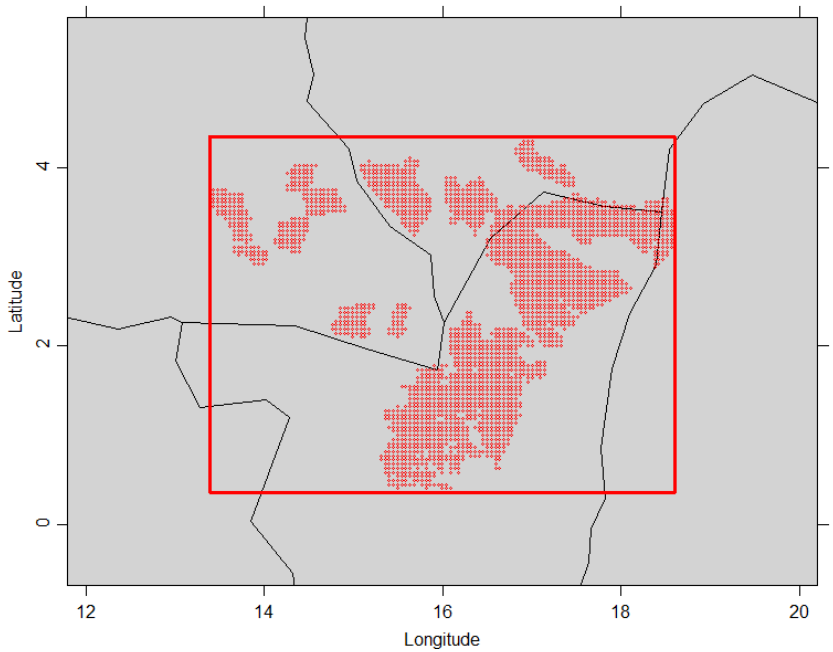
Model and predict the **abundance of tree species** in the tropical moist forest of the Congo-Basin

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 :

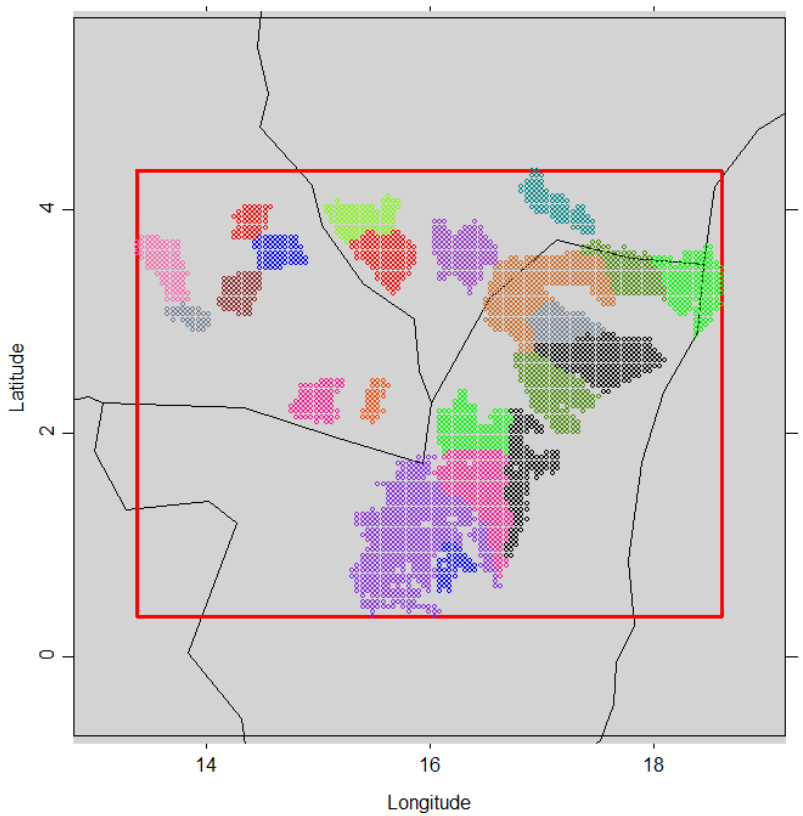
- $p = 56$ **explanatory variables**
- $r = 2$ **additional covariates**

Difficulties

- High level of correlation among the explanatory variables
  $\hookrightarrow$ **Regularisation** is needeed
- Spatial observations
  $\hookrightarrow$ 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 stucture
  $\hookrightarrow$ Within-group dependence modelled by a **random effect**
  $\hookrightarrow$ **Multivariate GLMM**
- High correlations among the explanatory variables
  $\hookrightarrow$ **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

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

Design matrices

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

Responses

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

Design matrices

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

Our linear predictors in the GLMM framework

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

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

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

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

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

with :

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

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

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

with :

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

"Linearised" model

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

"Linearised" model

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

Alternated procedure

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

Henderson systems

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

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

and

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

Goodness-Of-Fit

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

Goodness-Of-Fit

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

Structural Relevance

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

Goodness-Of-Fit

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

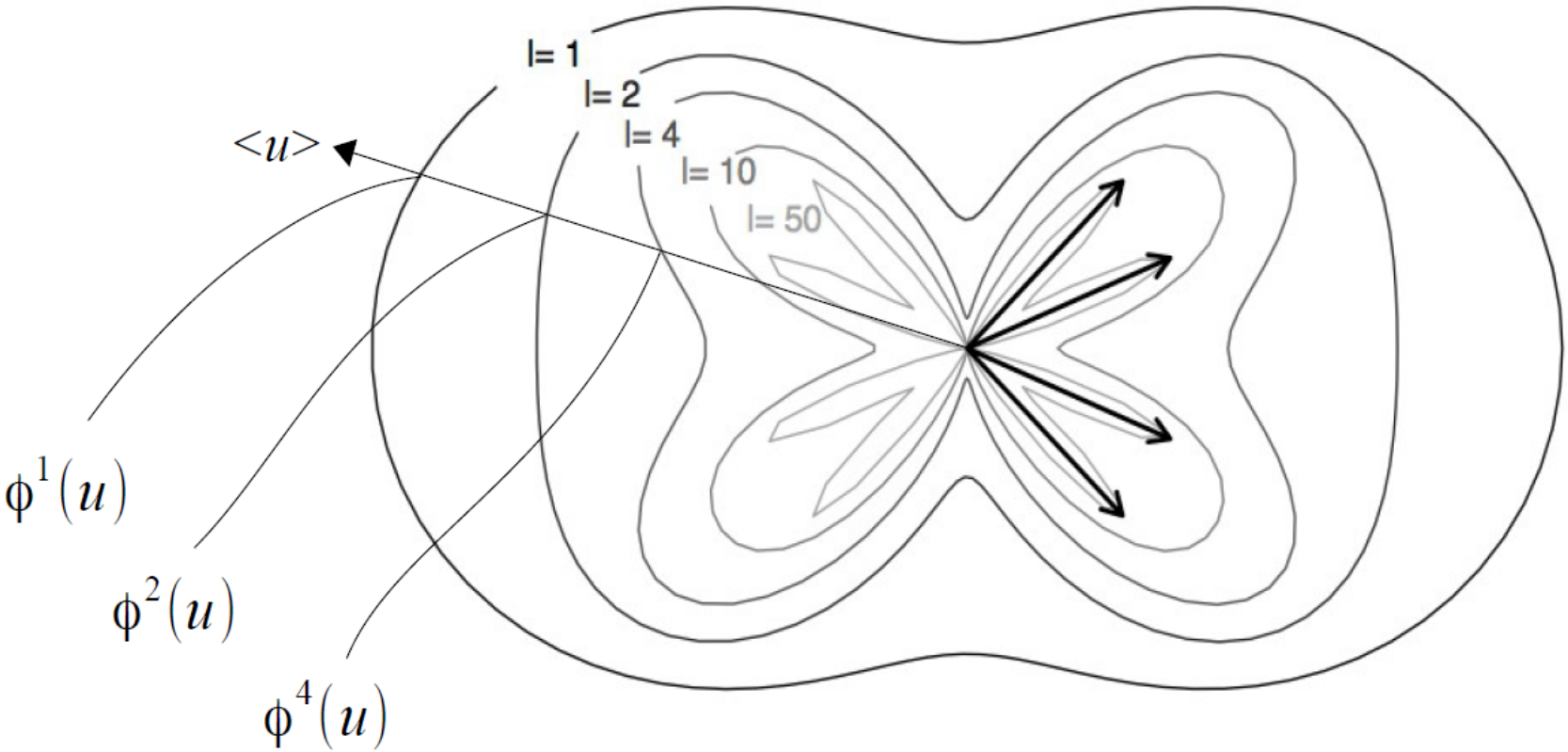
Structural Relevance

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

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_ξ^k, the weighting matrices W_k, and the variances σ_k^2

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

$u \leftarrow \underset{u'u=1}{\arg\max} \left[\psi(u)\right]^{1-s} \left[\phi(u)\right]^s$

$f \leftarrow Xu$

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

Estimate parameters γ_k, δ_k, ξ_k (Henderson system) and σ_k^2

Update z_ξ^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 : $\boldsymbol{Y} = \left[\boldsymbol{y^1}|\boldsymbol{y^2}\right]$

Random responses

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

Fixed effects

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

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

- 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

$\boldsymbol{N = 10}$ groups, $\boldsymbol{R = 10}$ units per group ($n = 100$ individuals)

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

Random effects

$\boldsymbol{N = 10}$ groups, $\boldsymbol{R = 10}$ units per group ($n = 100$ individuals)

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

Model

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

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

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

Number of simulations

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

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 λ

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 λ

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	LMM-Ridge	Mixed-SC(G)LR	
	shrinkage parameter $\lambda^{\star}_{\text{lasso}}$	shrinkage parameter $\lambda^{\star}_{\text{ridge}}$	numpber of component $K^{\star}$	tuning parameter $s^{\star}$
$\tau = 0.1$	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)

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

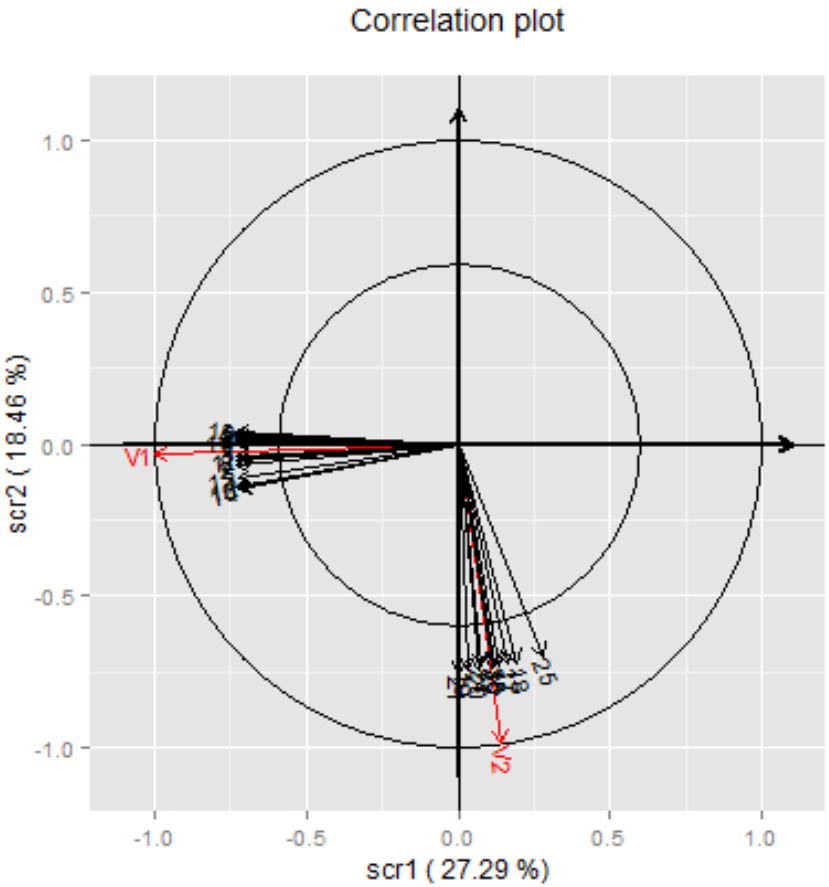
Robust comparison criterion : Mean Upper Relative Error (MURE)

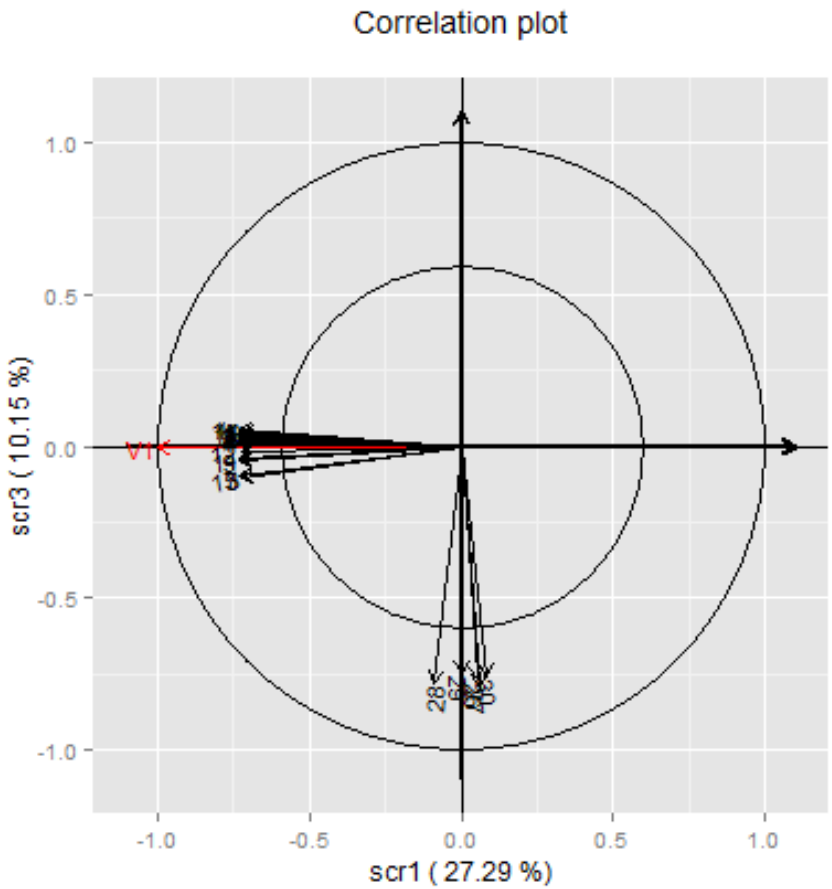
$$\text{MURE} = \frac{1}{M}\sum_{m=1}^{M} \max\left(\frac{\left\|\hat{\beta}_1^{(m)} - \beta_1\right\|^2}{\|\beta_1\|^2}, \frac{\left\|\hat{\beta}_2^{(m)} - \beta_2\right\|^2}{\|\beta_2\|^2}\right)$$

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).





4. Application to real data

The *Genus* dataset

- $n = 2615$ developped 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

$\hookrightarrow$ Poisson regression with log link

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

Modelisation

Abundance of tree species : count data

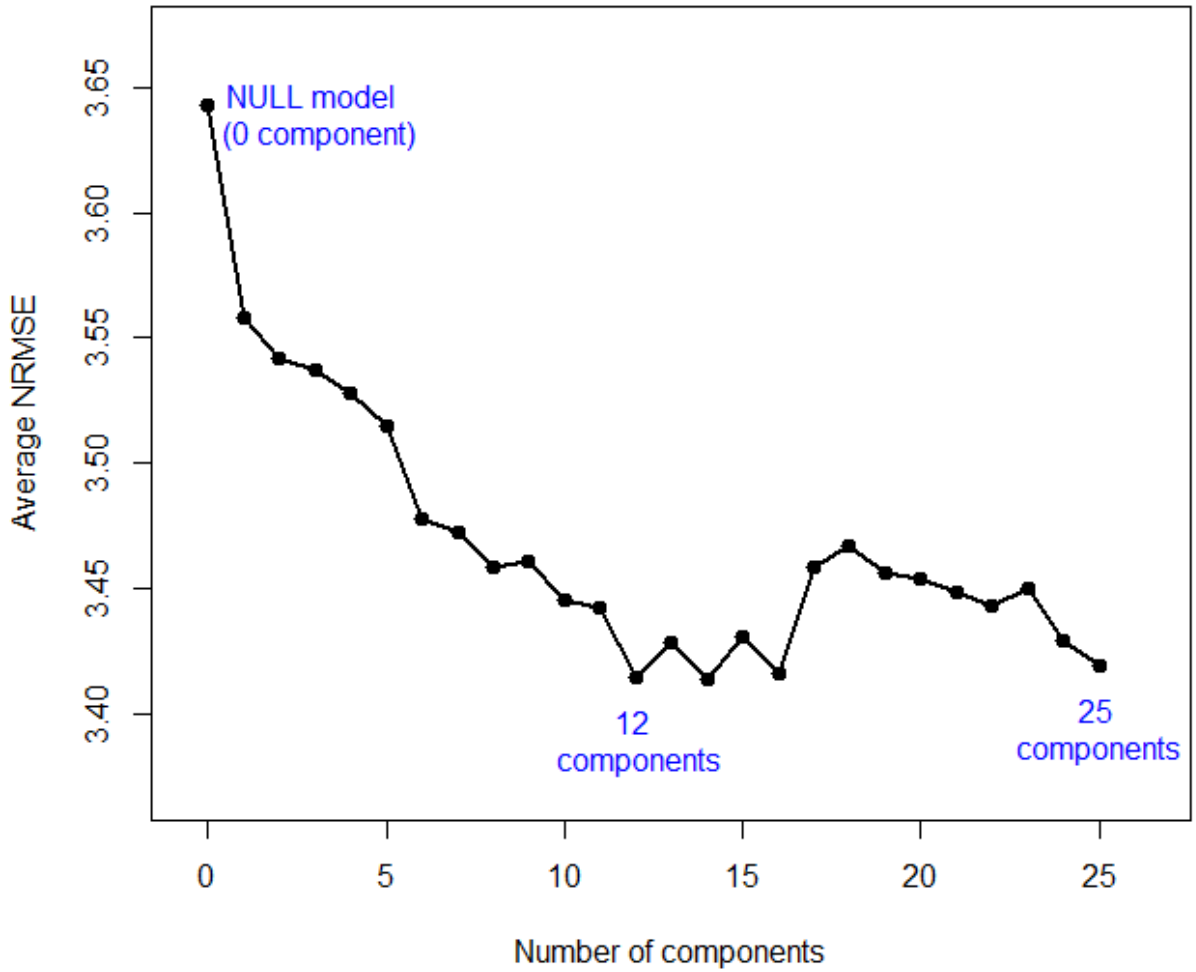
$\hookrightarrow$ Poisson regression with log link

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

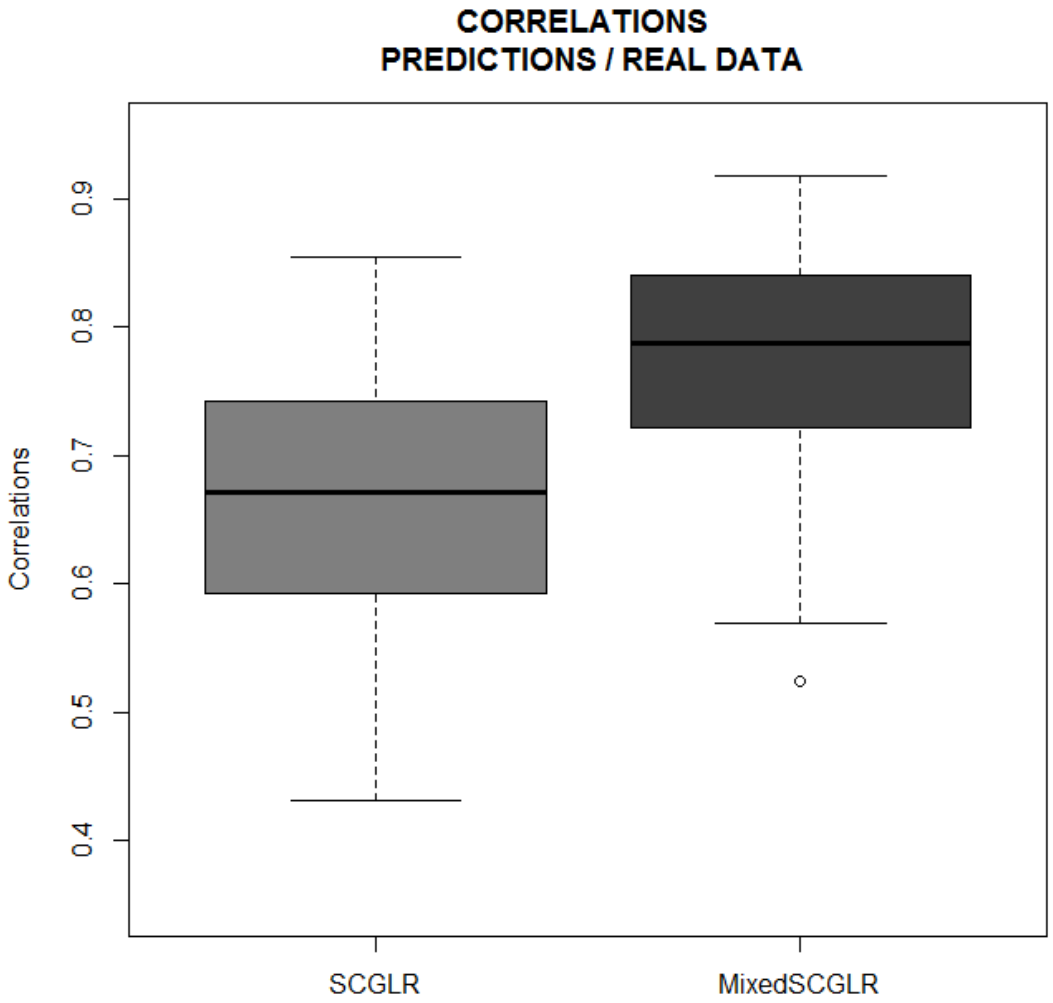
Optimal number of components $K^\star$

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

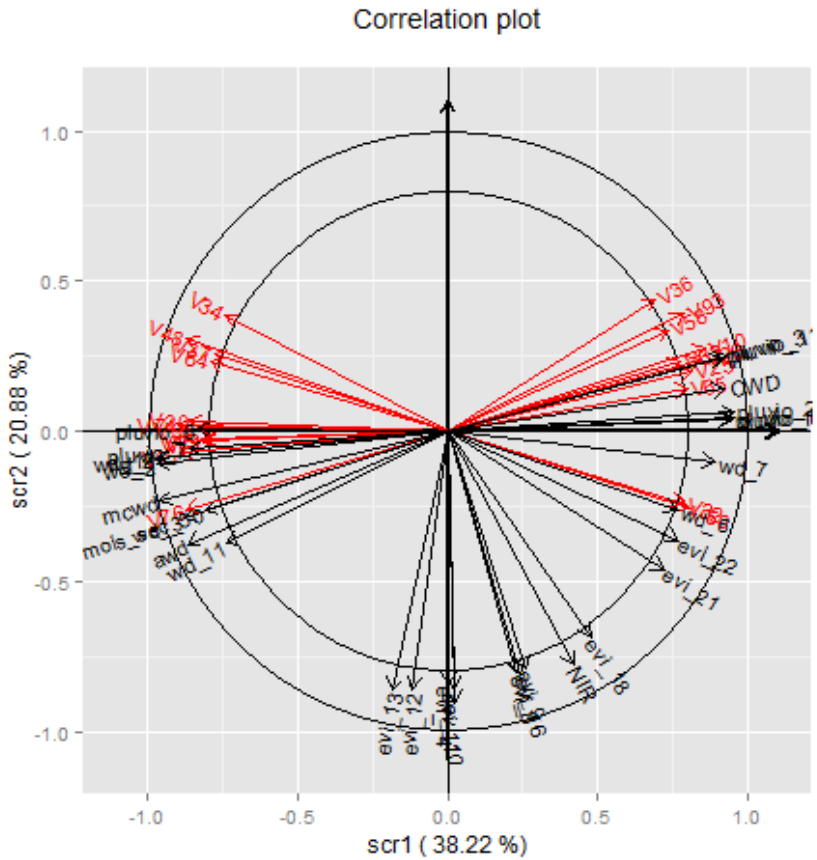
AveNRMSE's as a function of the number of components

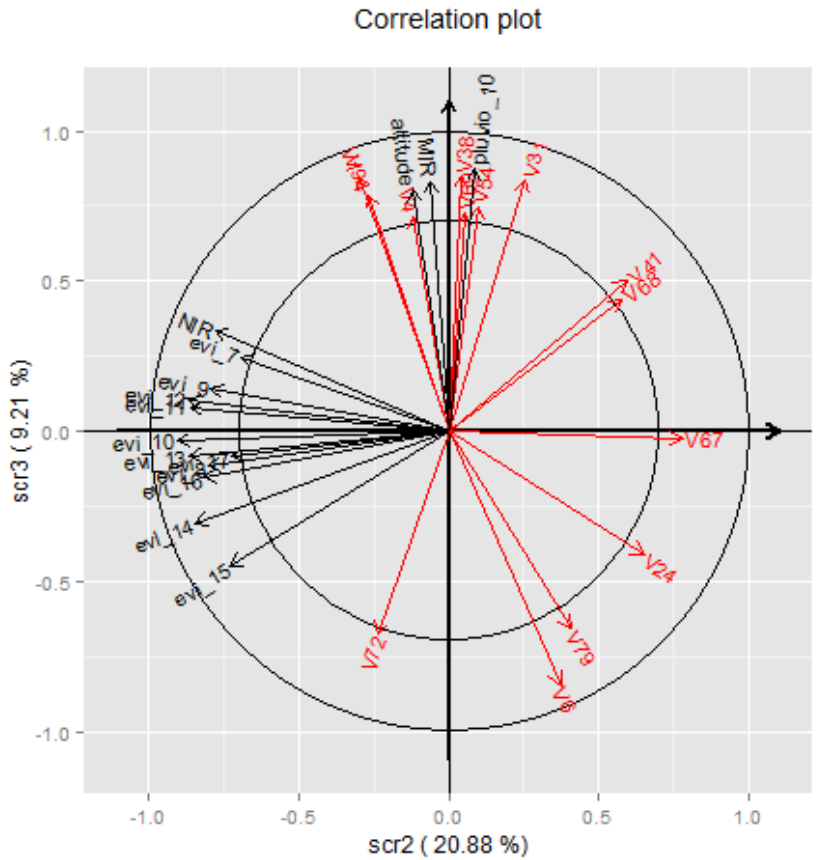


Improvement in correlations of predictions and responses



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 unbiaised 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.