Response mixture models based on supervised components: clustering floristic taxa

Julien Gibaud, Xavier Bry, Catherine Trottier, Frédéric Mortier, Maxime Réjou-Méchain

Response mixture models based on supervised components: clustering floristic taxa

Julien Gibaud [1], Xavier Bry [1], Catherine Trottier [1, 2], Frédéric Mortier [3, 4], and Maxime Réjou-Méchain [5]

[1] Institut Montpelliérain Alexander Grothendieck, CNRS, Université de Montpellier, France

[2] AMIS, Université Paul-Valéry Montpellier 3, F34000, Montpellier, France

[3] CIRAD, Forêts et Sociétés, Montpellier, France

[4] Forêts et Sociétés, Université de Montpellier, CIRAD, Montpellier, France

[5] AMAP, Université de Montpellier, IRD, CNRS, CIRAD, INRAE, Montpellier, France

---

Address for correspondence: Julien Gibaud, Institut Montpelliérain Alexander Grothendieck, CNRS, Université de Montpellier, Place Eugène Bataillon, 34095, Montpellier, France.

E-mail: `julien.gibaud@umontpellier.fr`.

Phone: (+33) 467 143 828.

Fax: Who still uses fax ?

Abstract: In this paper, we propose to cluster outcomes in order to identify groups

predicted by specific explanatory components. A response matrix $\boldsymbol{Y}$ is assumed to depend on a set $\boldsymbol{X}$ of explanatory variables, and a set $\boldsymbol{A}$ of additional covariates. Variables in $\boldsymbol{X}$ are supposed many and redundant, which implies some dimension reduction and regularization. By contrast, $\boldsymbol{A}$ contains few selected variables which are forced into the regression model, as they demand no regularization. The matrix $\boldsymbol{Y}$ is assumed partitioned into G unknown groups of responses. We suppose that the outcomes in each group are predictable from an appropriate number of specific orthogonal supervised components of $\boldsymbol{X}$. The classification is based on a mixture model of the responses. To estimate the model, we propose a criterion extending that of Supervised Component-based Generalized Linear Regression (SCGLR), a PLS-type method, and develop an algorithm combining those of SCGLR and EM estimation: response mixture SCGLR (rmSCGLR). This new methodology is tested on simulated data and then applied to a floristic ecology dataset.

---

1   Introduction

The climate change produces many ecosystem imbalances which might involve large extinctions of animal or plant taxa. In this context, the development of models which allow to predict the future of the biodiversity, has become a crucial issue. A number of advances have been made, in particular by extending Species Distribution Models (SDM, Guisan and Thuiller, 2005), which treat the taxa separately, to Joint

Species Distribution Models (JSDM, Pollock et al., 2014). JSDM allow to formalize the interdependence between taxa, and to understand its impact on the composition of communities. Besides, modeling responses (here, the abundances of taxa) requires taking into account a large set of possibly highly correlated explanatory covariates, which is the case of climatic variables, so SDM as JSDM demand regularization. This can be carried out by means of component-based dimension reduction. This consists in assuming that there is a small number of common latent explanatory dimensions, which we aim to capture through as many linear combinations of the explanatory variables, named components. Moreover, the case where the explanatory variables outnumber the observations (referred to as "high dimensional") is likely to become a new standard (Warton et al., 2015). In this paper, we aim to build components which can be interpreted as new and relevant synthetic climatic variables.

Elaborating on the Iteratively Reweighted Partial Least Squares (IRPLS) developed by Marx (1996), Bry et al. (2013) proposed a methodology called Supervised Component-based Generalized Linear Regression (SCGLR) which bridges the multivariate Generalized Linear Model (GLM) estimation, with the component-based dimension reduction of the explanatory space. Unlike methods as Partial Least Squares (PLS, Wold et al., 1984) regression or Reduced Rank Vector Generalized Linear Model (RRVGLM, Yee and Hastie, 2003), SCGLR optimizes a general and flexible trade-off criterion between the Goodness-of-Fit (GoF) of the model and the Structural Relevance (SR, Bry and Verron, 2015) of directions with respect to the explanatory variables. This methodology allows both to find strong interpretable explanatory directions, and to produce regularized predictors in the high-dimensional framework. Different extensions have recently been proposed to deal with data with a more complex structure (Chauvet et al., 2019; Bry et al., 2020a,b). An ® package **SCGLR** is

available at https://github.com/SCnext/SCGLR.

All the aforementioned extensions assume that all the responses are explained by the same latent dimensions. In our context, this might well not be the case: the responses are very different, and are thus likely to be modeled from explanatory dimensions which are, to some extent, specific. To overcome the limitation of former versions of SCGLR, we propose to extend it so as to find groups of response variables being modeled by the same specific explanatory dimensions. The clustering models or techniques classically used in statistical literature to identify groups do not consider the presence or abundance data as responses to explanatory variables (Dufrêne and Legendre, 1997; De Cáceres et al., 2010). In order to take the modeling of outcomes into account in the clustering, we propose to combine the SCGLR model with a Finite Mixture Model (FMM) of responses (see (McLachlan and Peel, 2004) for a reference book).

In a context of multiple and numerous response variables, we have to cluster the outcomes, and not the statistical units as in the original and classical FMM approach. The interest of response clustering has already been shown in several works, e.g. those of Monni and Tadesse (2009); Ovaskainen and Soininen (2011); Pledger and Arnold (2014); Mortier et al. (2015) and Hill et al. (2020). In our work, we use a modeling approach based on Dunstan et al. (2011, 2013), which assumes that all outcomes can be clustered into a small number of groups with respect to their responses to environmental gradients. In their model, the outcomes within a group share the same regression parameters with an intercept specific to each outcome. By contrast, we propose to entitle responses to their own regression parameters, and to define a group as a set of responses depending on the same common explanatory dimensions. To

help find these, the trade-off criterion of SCGLR had to be extended so as to preclude response groups from depending on too close explanatory sub-spaces.

The paper is organized as follows. In Section 2, we recall the principle of the original SCGLR. Section 3 presents the extension of SCGLR to a response FMM. Section 4 details two simulation studies that illustrate the interest of our work and the good performances of the proposed algorithm, highlighting the importance of a relevant selection of the hyper-parameters. Section 5 presents the results obtained on a floristic ecology dataset.

2   The original SCGLR

In this part, we consider the situation where there exists only one group of responses. For the sake of simplicity, we restrict ourselves to calculating a single component.

2.1   The SCGLR context

In the framework of a multivariate Generalized Linear Model (GLM) we consider K response-vectors, encoded in a response matrix $\boldsymbol{Y} = [\boldsymbol{y_1}, \ldots, \boldsymbol{y_K}] \in \mathbb{R}^{N \times K}$, to be predicted through explanatory variables partitioned in two sets. The first one $\boldsymbol{A} = [\boldsymbol{a_1}, \ldots, \boldsymbol{a_Q}] \in \mathbb{R}^{N \times Q}$ is a set of covariates that are only few and weakly or not redundant. These variables are *a priori* assumed to be interesting per se, and their marginal effects have to be taken into account explicitly in the model. The second group $\boldsymbol{X} = [\boldsymbol{x_1}, \ldots, \boldsymbol{x_P}] \in \mathbb{R}^{N \times P}$ is one of numerous and possibly highly redundant covariates, considered as proxies to latent dimensions, which must be found and interpreted. Thus, the matrix $\boldsymbol{X}$ demands dimension reduction and regularization.

To achieve this, SCGLR searches for explanatory components in $\boldsymbol{X}$ jointly supervised by the response set. A component $\boldsymbol{f} \in \mathbb{R}^N$ writes $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$, where $\boldsymbol{u} \in \mathbb{R}^P$ is a loading vector. For a single component model, the linear predictor associated with response $\boldsymbol{y_k}$ is then given by:

$$\boldsymbol{\eta_k} = (\boldsymbol{X}\boldsymbol{u})\,\boldsymbol{\gamma_k} + \boldsymbol{A}\boldsymbol{\delta_k},$$

where $\boldsymbol{\gamma_k}$ and $\boldsymbol{\delta_k}$ are regression parameters. Component $\boldsymbol{f}$ is common to all the responses, and for an identification purpose, we impose $\boldsymbol{u}^T\boldsymbol{M}^{-1}\boldsymbol{u} = 1$, where $\boldsymbol{M} \in \mathbb{R}^{P \times P}$ is a symmetric positive definite matrix. We assume that the responses are independent conditional on the explanatory variables, and consequently on $\boldsymbol{f}$.

2.2   Preliminary notations

The sequel contains mathematical developments which use notations listed hereafter:

- Let $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^p$ be vectors and $\boldsymbol{M} \in \mathbb{R}^{p \times p}$ be a symmetric positive definite matrix. The Euclidean scalar product between $\boldsymbol{u}$ and $\boldsymbol{v}$ with respect to metric $\boldsymbol{M}$ is given by $\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\boldsymbol{M}} = \boldsymbol{u}^T\boldsymbol{M}\boldsymbol{v}$. Likewise, $\cos_{\boldsymbol{M}}(\boldsymbol{u}, \boldsymbol{v}) = \dfrac{\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\boldsymbol{M}}}{\|\boldsymbol{u}\|_{\boldsymbol{M}} \|\boldsymbol{v}\|_{\boldsymbol{M}}}$ denotes the cosine of the angle between $\boldsymbol{u}$ and $\boldsymbol{v}$ with respect to metric $\boldsymbol{M}$.

- If $\boldsymbol{u}$ and $\boldsymbol{v}$ are centred and $\boldsymbol{M} = \boldsymbol{I_p}$, the cosine defines Pearson's correlation, denoted ρ. In this paper, unless otherwise stated, the correlation refers to Pearson's correlation.

- $\boldsymbol{U} = [\boldsymbol{u_1}, \ldots, \boldsymbol{u_p}]$ and $\boldsymbol{V} = [\boldsymbol{v_1}, \ldots, \boldsymbol{v_q}]$ being matrices. The space spanned by their column-vectors is denoted $\text{span}[\boldsymbol{U}, \boldsymbol{V}]$.

- Let w_i be the weight of unit i, and $\boldsymbol{W} = diag(w_i)_{i=1,\ldots,n}$. Let $\mathbb{R}^n$ be endowed with metric $\boldsymbol{W}$, and let $\boldsymbol{A} \in \mathbb{R}^{n \times p}$ be a matrix. The $\boldsymbol{W}$-orthogonal projector

onto span$[\boldsymbol{A}]$ is given by $\boldsymbol{\Pi}^{\boldsymbol{W}}_{\mathrm{span}[\boldsymbol{A}]} = \boldsymbol{A}\left(\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{W}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{W}$. Thus, the cosine of the angle between a vector $\boldsymbol{b} \in \mathbb{R}^n$ and span$[\boldsymbol{A}]$ with respect to metric $\boldsymbol{W}$ is given by $\cos_{\boldsymbol{W}}\left(\boldsymbol{b}, \mathrm{span}[\boldsymbol{A}]\right) = \cos_{\boldsymbol{W}}\left(\boldsymbol{b}, \boldsymbol{\Pi}^{\boldsymbol{W}}_{\mathrm{span}[\boldsymbol{A}]}\boldsymbol{b}\right)$.

- Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times p}$ be two real matrices. The Frobenius product is computed as $\langle \boldsymbol{A}, \boldsymbol{B}\rangle_{\mathrm{Frob}} = \mathrm{Tr}\left(\boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{B}\right)$, where Tr denotes the trace of a matrix. The unit orthogonal projector with respect to the Frobenius norm is given by $\boldsymbol{\varpi}^{\boldsymbol{W}}_{\mathrm{span}[\boldsymbol{A}]} = \boldsymbol{\Pi}^{\boldsymbol{W}}_{\mathrm{span}[\boldsymbol{A}]}/\sqrt{\mathrm{Tr}\left(\boldsymbol{\Pi}^{\boldsymbol{W}}_{\mathrm{span}[\boldsymbol{A}]}{}^{\boldsymbol{T}}\boldsymbol{\Pi}^{\boldsymbol{W}}_{\mathrm{span}[\boldsymbol{A}]}\right)}$.

2.3   Measuring the Goodness-of-Fit

For parameter estimation, we make use of the Fisher Scoring Algorithm (FSA) (refer to (McCullagh and Nelder, 1989) for a complete overview of GLM methodologies). Let $\boldsymbol{w}_{\boldsymbol{k}}$ be the working variable associated with the response $\boldsymbol{y}_{\boldsymbol{k}}$, and $\boldsymbol{W}_{\boldsymbol{k}}^{-1}$ its variance matrix. Contrary to Bastien et al. (2005), we weight the model based on components by $\boldsymbol{W}_{\boldsymbol{k}}^{-1}$. Indeed, in the spirit of Nelder and Wedderburn (1972), at iteration t, $\boldsymbol{w}_{\boldsymbol{k}}$ can be viewed as the response in the linearized model:

$$\boldsymbol{w}_{\boldsymbol{k}}^{(t)} = \left(\boldsymbol{X}\boldsymbol{u}\right)\boldsymbol{\gamma}_{\boldsymbol{k}} + \boldsymbol{A}\boldsymbol{\delta}_{\boldsymbol{k}} + \boldsymbol{\zeta}_{\boldsymbol{k}}^{(t)},$$

with $\mathbb{E}(\boldsymbol{\zeta}_{\boldsymbol{k}}^{(t)}) = 0$ and $\mathbb{V}(\boldsymbol{\zeta}_{\boldsymbol{k}}^{(t)}) = \boldsymbol{W}_{\boldsymbol{k}}^{-1(t)}$. Due to the product $\boldsymbol{u}\boldsymbol{\gamma}_{\boldsymbol{k}}$, this linearized model must be estimated through an alternated weighted least squares process, estimating in turn $\{\boldsymbol{\gamma}_{\boldsymbol{k}}, \boldsymbol{\delta}_{\boldsymbol{k}}\}$ and $\boldsymbol{u}$.

Let $\boldsymbol{\Pi}^{\boldsymbol{W}_{\boldsymbol{k}}}_{\mathrm{span}[\boldsymbol{f}, \boldsymbol{A}]}$ be the projection on span$[\boldsymbol{f}, \boldsymbol{A}]$ with respect to $\boldsymbol{W}_{\boldsymbol{k}}$. The loading vector $\boldsymbol{u}$ solution of the least squares minimization may alternatively be viewed as

the solution of the following optimization program:

$$\max_{\boldsymbol{u},\boldsymbol{u}^T\boldsymbol{M}^{-1}\boldsymbol{u}=1} \psi_{\boldsymbol{A}}(\boldsymbol{u}) := \sum_{k=1}^{K} \alpha_k \|\boldsymbol{w_k}\|_{\boldsymbol{W_k}}^2 \cos^2_{\boldsymbol{W_k}}\left(\boldsymbol{w_k}, \Pi^{\boldsymbol{W_k}}_{\mathrm{span}[\boldsymbol{f},\boldsymbol{A}]}\boldsymbol{w_k}\right),$$

where $\{\alpha_1, \ldots, \alpha_K\}$ is a weighting system reflecting the *a priori* relative importance of working variables. $\psi_{\boldsymbol{A}}$ is merely a Goodness-of-Fit (GoF) measure, and maximizing it does not lead to strong and interpretable components. The GoF measure must therefore be aptly combined with a measure of Structural Relevance (SR) to extract dimensions that are both meaningful and predictive, and achieve satisfactory regularization.

2.4   Measuring the Structural Relevance of components

Bry and Verron (2015) proposed the SR measure as a possible extension of the component's variance to measure the ability of a component to capture information in a set of variables containing latent structures such as variable-bundles. Informally, a bundle is a set of variables correlated "enough" to be viewed as produced by a common latent dimension. We call $\boldsymbol{W}$ the weight matrix reflecting the *a priori* relative importance of statistical units (typically, $\boldsymbol{W} = \frac{1}{N}\boldsymbol{I_N}$). Finally, consider component $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{u}$, where $\boldsymbol{u}$ is constrained by $\boldsymbol{u}^T\boldsymbol{M}^{-1}\boldsymbol{u} = 1$. Most often, we can take $\boldsymbol{M}^{-1} = \boldsymbol{I_P}$ although $\boldsymbol{M}^{-1}$ may take various forms according to the type of variables and structure of data (Bry et al., 2020b). Assuming that $\boldsymbol{X}$ consists of P standardized numeric variables, the associated SR measure ϕ is defined as the following generalized average of quadratic forms

$$\phi(\boldsymbol{u}) := \left(\frac{1}{P}\sum_{p=1}^{P} \langle \boldsymbol{X}\boldsymbol{u}, \boldsymbol{x_p}\rangle_{\boldsymbol{W}}^{2l}\right)^{1/l}.$$

The locality of a bundle of correlated variables is defined by the within-bundle correlation: the higher the correlation, the more local the bundle. The locality of the bundles to be tracked by components is tuned through the hyper-parameter $l \geq 1$. Components will line up with a more or less local bundle depending on whether l is greater or smaller, respectively. The main objective is to focus on the most interpretable directions.

2.5 The original SCGLR criterion

The SCGLR specific criterion, proposed by Bry et al. (2020b), introduced a hyper-parameter $s \in [0, 1]$ to tune the importance of the SR relative to the GoF. SCGLR attempts a trade-off between ϕ and $\psi_{\boldsymbol{A}}$ by solving:

$$\max_{\boldsymbol{u}, \boldsymbol{u}^T \boldsymbol{M}^{-1} \boldsymbol{u} = 1} s \ln(\phi(\boldsymbol{u})) + (1 - s) \ln(\psi_{\boldsymbol{A}}(\boldsymbol{u})).$$

When $s = 0$, the criterion identifies with the GoF, while at the other end, taking $s = 1$ makes it identify with the SR. Thus, increasing s intensifies both the focus of components on "strong" dimensions, and the regularization.

This compound criterion is quite general. Indeed, the GoF measure adapts any situation where a likelihood function is available for the model taking the components and $\boldsymbol{A}$ as covariates. Generally, this likelihood involves a vector of parameters. The maximization is carried out alternating two steps: (i) Given $\boldsymbol{u}$, maximize the criterion with respect to the parameter vector. This step is performed using a classical likelihood maximization algorithm relevant to the situation. (ii) Given the parameter vector, maximize the criterion with respect to $\boldsymbol{u}$ using a dedicated algorithm: PING

(for Projected Iterated Normed Gradient) recalled in Supplementary Material. PING is designed to maximize, at least locally, any criterion on the unit sphere (Chauvet et al., 2019; Bry et al., 2020a,b). We shall adapt the criterion and its maximization to the response-mixture model.

3   Response Mixture SCGLR

In this part, we formally express the combination of SCGLR with a FMM we use to build and estimate our response mixture component-based model.

3.1   The response mixture model

Let $\boldsymbol{Y} = [\boldsymbol{y_1}, \dots, \boldsymbol{y_K}] \in \mathbb{R}^{N \times K}$ be the response matrix. The responses are assumed to be modeled through a finite mixture of regression models, comprising G groups. The probability distribution function (pdf) of response $\boldsymbol{y_k}$ is thus:

$$L(\boldsymbol{y_k}; \boldsymbol{\theta_k}) = \sum_{g=1}^{G} p_g \prod_{n=1}^{N} d_k(y_{nk}; \mu_{nkg}),$$

where the nth individual of the kth response of the gth group has a pdf d_k belonging to the exponential family, with expectation μ_{nkg}. $\boldsymbol{\theta_k}$ denotes the vector of parameters, including the regression parameters $\boldsymbol{\gamma_{kg}}$ and $\boldsymbol{\delta_{kg}}$, as defined in Section 2.1, and p_g is the gth mixing probability with $p_g \in [0, 1]$ and $\sum_{g=1}^{G} p_g = 1$. Denoting h_k the kth canonical link function, we assume:

$$h_k(\mu_{nkg}) = (\boldsymbol{x_n^T} \boldsymbol{u_g}) \boldsymbol{\gamma_{kg}} + \boldsymbol{a_n^T} \boldsymbol{\delta_{kg}},$$

where $\boldsymbol{u_g}$ is the loading vector of the (first) component of group g, and $\boldsymbol{x_n}$ and $\boldsymbol{a_n}$ are the vectors composed by the nth rows of matrices $\boldsymbol{X}$ and $\boldsymbol{A}$ respectively. Thus, the responses in group g are predicted by component $\boldsymbol{f_g} = \boldsymbol{X}\boldsymbol{u_g}$, together with $\boldsymbol{A}$. For each $k = 1, \ldots, K$, d_k and h_k are chosen so as to suit the type of response $\boldsymbol{y_k}$ (e.g. binary, count, categorical, continuous etc.).

Conditional on the explanatory variables, the response variables are assumed independent. The group memberships of the responses being unknown, the model's log-likelihood

$$l(\boldsymbol{\Theta}; \boldsymbol{Y}) = \sum_{k=1}^{K} \ln\left(L(\boldsymbol{y_k}; \boldsymbol{\theta_k})\right),$$

where the set of parameters is $\boldsymbol{\Theta} = \{\boldsymbol{\theta_1}, \ldots, \boldsymbol{\theta_K}\}$, being difficult to maximize directly, we shall adopt the Expectation Maximization (EM, Dempster et al., 1977) algorithm to estimate the model's parameters.

Let z_{kg} be the latent indicator variable equal to 1 if the response $\boldsymbol{y_k}$ belongs to the gth group, and 0 otherwise. Let $\boldsymbol{z_k} = (z_{kg}; g = 1, \ldots, G)$ be the vector of group membership indicators of response $\boldsymbol{y_k}$, and let $\boldsymbol{Z} = [\boldsymbol{z_k}; k = 1, \ldots, K]$ be a $G \times K$ matrix. Conditional on $z_{kg} = 1$, the pdf of response $\boldsymbol{y_k}$ for unit n is $d_k(y_{nk}; \mu_{nkg})$. The model's complete log-likelihood writes:

$$l(\boldsymbol{\Theta}; \boldsymbol{Y}, \boldsymbol{Z}) = \sum_{k=1}^{K} \sum_{g=1}^{G} z_{kg} \ln\left(p_g \prod_{n=1}^{N} d_k(y_{nk}; \mu_{nkg})\right).$$

As mentioned above, step (i) of the combined criterion maximization boils down to maximizing the likelihood of the component-based model. Owing to the latent variable $\boldsymbol{Z}$, this step will be performed using the EM algorithm. The expectation of

the complete log-likelihood writes

$$\mathbb{E}\left[l(\boldsymbol{\Theta}; \boldsymbol{Y}, \boldsymbol{Z})|\boldsymbol{Y}; \boldsymbol{\Theta}'\right] = \sum_{k=1}^{K}\sum_{g=1}^{G}\alpha_{kg}\ln\left(p_g\prod_{n=1}^{N}d_k(y_{nk}; \mu_{nkg})\right).$$

The posterior probability is computed as

$$\alpha_{kg} := \frac{p_g\prod_{n=1}^{N}d_k(y_{nk}; \mu_{nkg})}{\sum_{r=1}^{G}p_r\prod_{n=1}^{N}d_k(y_{nk}; \mu_{nkr})}.$$

As a result, we use the algorithm presented in Supplementary Material to estimate the parameters of the response mixture model.

3.2   Calculating the components of the response groups

When clustering the responses according to their "common" SCGLR components, we must ensure that the explanatory subspaces spanned by the components associated to response clusters be reasonably separated (else, the algorithm may fail to converge). To achieve that, when calculating a component explanatory of a response cluster, we must prevent that it be too close to the explanatory subspaces of other clusters.

3.2.1   An additional sub-criterion to better separate explanatory sub-spaces

Let $\boldsymbol{F_{-g}} = \{\boldsymbol{f_1}, \ldots, \boldsymbol{f_{g-1}}, \boldsymbol{f_{g+1}}, \ldots, \boldsymbol{f_G}\}$ be the set of components from which the component of group g was removed. The space spanned by the component $\boldsymbol{f_g}$ may be uniquely represented by the orthogonal projector on it: $\boldsymbol{\varpi}_{\mathrm{span}[\boldsymbol{f_g}]}$. With this in mind, we propose to measure the separation of $\mathrm{span}[\boldsymbol{f_g}]$ from $\mathrm{span}[\boldsymbol{f_r}]$'s, for all $r \neq g$,

through the function of $\boldsymbol{u_g}$:

$$\varphi_{\boldsymbol{F_{-g}}}(\boldsymbol{u_g}) = 1 - \frac{1}{G-1}\sum_{r \neq g}\left\langle \boldsymbol{\varpi}_{\text{span}[\boldsymbol{f_g}]}, \boldsymbol{\varpi}_{\text{span}[\boldsymbol{f_r}]}\right\rangle_{\text{Frob}}. \tag{3.1}$$

Indeed, if for all $r \neq g$, $\text{span}[\boldsymbol{f_g}]$ is orthogonal to $\text{span}[\boldsymbol{f_r}]$ then the Frobenius product will be 0, so that the criterion will be equal to 1. At the other end, if for all $r \neq g$, $\text{span}[\boldsymbol{f_g}] = \text{span}[\boldsymbol{f_r}]$, the Frobenius product will be equal to 1, and the criterion to 0.

The new combined criterion we propose for group g is thus:

$$\max_{\boldsymbol{u_g}, \boldsymbol{u_g}^T \boldsymbol{M}^{-1} \boldsymbol{u_g}=1} s\ln(\phi(\boldsymbol{u_g})) + t\ln(\varphi_{\boldsymbol{F_{-g}}}(\boldsymbol{u_g})) + (1 - s - t)\ln(\psi_{\boldsymbol{A}}(\boldsymbol{u_g})), \tag{3.2}$$

where $s, t, (s+t) \in [0, 1]$.

3.2.2 Rank-1 component

Let us address step (ii) of the combined criterion maximization. The GoF measure applied to group g is given by

$$\psi_{\boldsymbol{A}}(\boldsymbol{u_g}) = \sum_{k=1}^{K}\alpha_{kg}\|\boldsymbol{w_{kg}}\|_{\boldsymbol{W_{kg}}}^2 \cos^2_{\boldsymbol{W_{kg}}}\left(\boldsymbol{w_{kg}}, \boldsymbol{\Pi}_{\text{span}[\boldsymbol{f_g}, \boldsymbol{A}]}^{\boldsymbol{W_{kg}}}\boldsymbol{w_{kg}}\right),$$

where the weights reflecting the degrees of membership to group g of responses are the posterior probabilities $\{\alpha_{1g}, \ldots, \alpha_{Kg}\}$. The functions ϕ and $\varphi_{\boldsymbol{F_{-g}}}$ are respectively given in section 2.4 and equation (3.1). The explicit expression of the criterion is given in Supplementary Material.

3.2.3 Higher rank components

We shall henceforth calculate the higher rank components. Let $\boldsymbol{f}_g^h = \boldsymbol{X}\boldsymbol{u}_g^h$ be the hth component of group g, and let $\boldsymbol{F}_g^h = [\boldsymbol{f}_g^1, \ldots, \boldsymbol{f}_g^h]$, where $h \leq H_g$, be the matrix of the first h components of group g. According to the local nesting principle (Bry et al., 2012), the new component $\boldsymbol{f}_g^{h+1}$ must best complement both the existing ones and $\boldsymbol{A}$, that is $\boldsymbol{A}_g^h := [\boldsymbol{F}_g^h, \boldsymbol{A}]$. So $\boldsymbol{f}_g^{h+1}$ has to be calculated using $\boldsymbol{A}_g^h$ as the new set of additional covariates. Moreover, to avoid linear redundancy of components, we impose that $\boldsymbol{f}_g^{h+1}$ be orthogonal to $\boldsymbol{F}_g^h$, *i.e.* $\boldsymbol{F}_g^{h\,T}\boldsymbol{W}\boldsymbol{f}_g^{h+1} = 0$.

We calculate every new component as the solution of program (3.2), with the additional constraint: $\boldsymbol{\Delta}_g^{h\,T}\boldsymbol{u}_g^{h+1} = 0$, where $\boldsymbol{\Delta}_g^h = \boldsymbol{X}^T\boldsymbol{W}\boldsymbol{F}_g^h$, and loop on g until overall convergence of the component system. Taking $\boldsymbol{A}_g^h = [\boldsymbol{F}_g^h, \boldsymbol{A}]$ and $\boldsymbol{F}_{-g} = \{\boldsymbol{F}_1^{H_1}, \ldots, \boldsymbol{F}_{g-1}^{H_{g-1}}, \boldsymbol{F}_{g+1}^{H_{g+1}}, \ldots, \boldsymbol{F}_G^{H_G}\}$, the sub-criteria become

$$\psi_{\boldsymbol{A}_g^h}\left(\boldsymbol{u}_g^{h+1}\right) = \sum_{k=1}^{K} \alpha_{kg}\|\boldsymbol{w}_{kg}\|_{\boldsymbol{W}_{kg}}^2 \cos_{\boldsymbol{W}_{kg}}^2\left(\boldsymbol{w}_{kg}, \Pi_{\mathrm{span}[\boldsymbol{f}_g^{h+1}, \boldsymbol{A}_g^h]}^{\boldsymbol{W}_{kg}} \boldsymbol{w}_{kg}\right)$$

and

$$\varphi_{\boldsymbol{F}_{-g}}\left(\boldsymbol{u}_g^{h+1}\right) = 1 - \frac{1}{G-1}\sum_{r \neq g}\left\langle \boldsymbol{\varpi}_{\mathrm{span}[\boldsymbol{F}_g^{h+1}]}, \boldsymbol{\varpi}_{\mathrm{span}[\boldsymbol{F}_r^{H_r}]} \right\rangle_{\mathrm{Frob}}.$$

For all $g = 1, \ldots, G$, the rank-1 component of group g is calculated using the same program with $\boldsymbol{A}_g^0 = \boldsymbol{A}$ and $\boldsymbol{\Delta}_g^0 = 0$.

3.3 The overall algorithm

The overall algorithm, presented in Supplementary Material, consists in alternating the following steps: (i) Given the current set of components, estimate the mixture

response parameters through the EM algorithm; (ii) Given the current group memberships of responses, calculate all the components of all the groups. The method thus implemented is named response mixture SCGLR (rmSCGLR).

To give our algorithm a good starting point, namely well separated response clusters and strong initial components, we use the **ClustOfVar** ® package (Chavent et al., 2012) to determine the G initial response groups, and then, the **pls** ® package (Mevik and Wehrens, 2007) in each group, to find the initial supervised components.

3.4   A hyper-parameter calibration heuristic

The hyper-parameters are calibrated using the Bayesian Information Criterion (BIC), defined by

$$\text{BIC} = -2l\left(\boldsymbol{\Theta}; \boldsymbol{Y}\right) + \ln\left(N\right) \times \left(\text{number of parameters}\right).$$

Keribin (2000) shows the reliability of BIC in a context of mixture model. The hyper-parameters are many $(s, l, t, G, H_1, \ldots, H_G)$, so that using BIC to compare all their combinations on a cross-product grid is out of the question in practice. We choose instead to study the effects of varying the hyper-parameters following a heuristic. Even if these parameters have different purposes, which can to some extent be dealt with sequentially, they are not completely independent. For instance: the higher s is, the higher H_g is likely to be be. In practice, we propose the following heuristic: first, we perform an optimization on the hyper-parameters s and l with standard SCGLR (e.g. without mixture) on a grid $(s, l) \in \{0.1, 0.2, \ldots, 0.9\} \times \{1, 2, 3, 4, 5\}$, calculating only one component. In a second step, we chose the number of groups by varying G

from 1 to 5, keeping s and l fixed to their previously optimized values, still calculating a single component. The decision of distinguishing the groups only through their first component is justified by the preliminary simulation study presented in Supplementary Material. Next, we implement forward selection to determine a suitable number of components in each group. We add one component in each group alternatively, and then choose the combination minimizing the BIC. We repeat that until the BIC rises. Finally, we vary the hyper-parameter t in $\{0.1, 0.2 \ldots, 0.9\}$, subject to the constraint $s + t \leq 1$, in order to better separate the components which might cause confusion between groups.

4   Simulation study

Two simulation studies have been implemented to assess the performance of rm-SCGLR. The first one, presented in Section 4.1, focuses on the identification of groups in a case of high correlations between latent variables spanning the explanatory spaces. In this simulation, we first present the components combination found by the previous heuristic for $s \in \{0.1, 0.3, 0.5\}$. Then we study the determination of the best value of hyper-parameter t. In the Supplementary Material, we present a preliminary simulation, in which we study the recovering of the true numbers of components, in a context of low correlation between the latent variables. In both simulations, we set $l = 4$ in order to facilitate the interpretation of components. For more information on the effects of hyper-parameters s and l, we refer the reader to (Chauvet et al., 2019; Bry et al., 2020a,b). The ® package **rmSCGLR** and the simulation codes are available at https://github.com/julien-gibaud/rmSCGLR.

In the simulation study, we use the Rand Index (RI, Rand, 1971) and the Adjusted Rand Index (ARI, Hubert and Arabie, 1985) to assess the correctness of the classification decisions. In addition, to measure the quality of the latent variables recovery, we calculate the square correlation between the latent variable $\boldsymbol{\xi}$ and the components:

$$\rho^{(}\boldsymbol{\xi}, .) = \max_{g,h} \rho\left(\boldsymbol{\xi}, \boldsymbol{f_g^h}\right)^2,$$

where $\boldsymbol{f_g^h}$ denotes the hth component of group g. The RI, ARI, square correlation and BIC are all given through mean values over a hundred samples.

4.1 Generation of the simulated data

The variables are simulated on $N = 100$ observations. Two latent variables $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$ are simulated with a correlation $\rho = 0.9$, while two others, $\boldsymbol{\xi_3}$ and $\boldsymbol{\xi_4}$, are simulated independent of any other. The $\boldsymbol{X}$ matrix consists in five blocks: $\boldsymbol{X} = [\boldsymbol{X_1}, \boldsymbol{X_2}, \boldsymbol{X_3}, \boldsymbol{X_4}, \boldsymbol{X_5}]$, where $\boldsymbol{X_1} \in \mathbb{R}^{N \times 20}$, $\boldsymbol{X_2} \in \mathbb{R}^{N \times 20}$, $\boldsymbol{X_3} \in \mathbb{R}^{N \times 10}$ and $\boldsymbol{X_4} \in \mathbb{R}^{N \times 10}$ are bundles of variables distributed about $\boldsymbol{\xi_1}, \boldsymbol{\xi_2}, \boldsymbol{\xi_3}$ and $\boldsymbol{\xi_4}$ respectively. More formally, for all $i = 1, \ldots, 4$, a variable $\boldsymbol{x_p}$ is simulated as $\boldsymbol{x_p} = \boldsymbol{\xi_i} + \boldsymbol{\varepsilon_p}$, where $\boldsymbol{\varepsilon_p} \sim \mathcal{N}_N(0, 0.1\boldsymbol{I_N})$. The $\boldsymbol{X_5}$ block contains only 40 unstructured noise variables constructed as $\boldsymbol{x_p} \sim \mathcal{N}_N(0, \boldsymbol{I_N})$. The response matrix $\boldsymbol{Y}$ is partitioned into two groups of responses only distinguished by their explanatory latent variables. The first group consists of Poisson and Gaussian responses whose linear predictors are combinations of $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_3}$, while the second group gathers Gaussian and binary responses with

linear predictors combining $\boldsymbol{\xi_2}$ and $\boldsymbol{\xi_4}$. The matrix $\boldsymbol{Y}$ is generated as:

$$\forall k = 1, \ldots, 20, \quad \boldsymbol{y_k} \sim \mathcal{N}_N \left(\boldsymbol{\mu} = \boldsymbol{\gamma_1^k} \boldsymbol{\xi_1} + \boldsymbol{\gamma_2^k} \boldsymbol{\xi_3}, \boldsymbol{\Sigma} = \boldsymbol{I_N} \right),$$

$$\forall k = 21, \ldots, 70, \quad \boldsymbol{y_k} \sim \mathcal{P} \left(\boldsymbol{\lambda} = \exp \left[0.25 \boldsymbol{\gamma_1^k} \boldsymbol{\xi_1} + 0.25 \boldsymbol{\gamma_2^k} \boldsymbol{\xi_3} \right] \right),$$

$$\forall k = 71, \ldots, 80, \quad \boldsymbol{y_k} \sim \mathcal{N}_N \left(\boldsymbol{\mu} = \boldsymbol{\gamma_1^k} \boldsymbol{\xi_2} + \boldsymbol{\gamma_2^k} \boldsymbol{\xi_4}, \boldsymbol{\Sigma} = \boldsymbol{I_N} \right),$$

$$\forall k = 81, \ldots, 100, \quad \boldsymbol{y_k} \sim \mathcal{B} \left(\boldsymbol{p} = \mathrm{logit}^{-1} \left[\boldsymbol{\gamma_1^k} \boldsymbol{\xi_2} + \boldsymbol{\gamma_2^k} \boldsymbol{\xi_4} \right] \right),$$

where for all k, $\boldsymbol{\gamma_1^k}$ and $\boldsymbol{\gamma_2^k}$ are uniformly generated, with $\boldsymbol{\gamma_1^k} \in [-4, 4]$ and $\boldsymbol{\gamma_2^k} \in [-2, 2]$.

4.2   Results and interpretation

Table 1 sums up the heuristic performed to find the best component combination for $s \in \{0.1, 0.3, 0.5\}$. We observe that the three values of s lead to detect two groups of responses. In this simulation, taking a higher value of s is not recommended. Indeed, as s increases, the components get closer to the principal components (Bry et al., 2020a). Thus, for $s > 0.5$, the first component of each group being drawn towards the same first principal component, they tend to be similar. This similarity hinders the distinction between groups. Performing a forward selection step and opting for the minimal value of BIC, we see that only $s = 0.1$ and $s = 0.3$ lead to the right combination of components. However, $s = 0.5$ leads to identify the true overall number of directions central to the explanatory bundles. Thus, in the sequel, the analysis is done with combinations $(H_1, H_2) = (2, 2)$ for $s = 0.1$ and $s = 0.3$, while we set $(H_1, H_2) = (3, 1)$ for $s = 0.5$.

On the last step of the heuristic, summed up in Table 2, we can see the impact of the hyper-parameter t. For $s = 0.1$, the RI and the ARI increase as t goes from 0

Table 1: Mean values of BIC over a hundred samples, for a high correlation value ($\rho = 0.9$) between the latent variables $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$, for $s \in \{0.1, 0.3, 0.5\}$ and different combinations of H_1 and H_2.

$s = 0.1$			$s = 0.3$			$s = 0.5$		
H_1	H_2	BIC	H_1	H_2	BIC	H_1	H_2	BIC
1	1	30802.37	1	1	31281.99	1	1	31862.52
2	1	29577.02	2	1	29896.88	2	1	30416.06
1	2	29538.69	1	2	29821.90	1	2	30431.90
2	**2**	**29091.21**	**2**	**2**	**29549.27**	**3**	**1**	**29593.27**
1	3	29513.46	1	3	29561.49	2	2	29811.69
3	2	30030.12	3	2	30296.23	4	1	30054.50
2	3	30108.98	2	3	30292.89	3	2	30450.21

to 0.4, and then decrease. Our criterion allows to distinguish two sub-spaces close to one another: for $t = 0.4$, the RI and the ARI values are respectively equal to 0.883 and 0.764 despite the high correlation between the first latent variables of the two groups. These observations are consistent with the BIC which decreases from $t = 0$ to $t = 0.4$. When t is too high, the RI and the ARI decrease, while the BIC increases, as observed for $t \geq 0.5$. In such cases, the weight of the separation sub-criterion φ is too heavy, and prevents the first components of the two groups to be close enough, which precludes the correct identification of the latent variables, hence of the groups. As a result, the square correlations between the rank-1 components and the corresponding latent variables are lower than 0.9 for $t \geq 0.4$. Moreover, when $t \geq 0.5$, the correlations $\rho^2(\boldsymbol{\xi_3}, .)$ and $\rho^2(\boldsymbol{\xi_4}, .)$ are higher than $\rho^2(\boldsymbol{\xi_1}, .)$ and $\rho^2(\boldsymbol{\xi_2}, .)$. The reason for this is that the latent variables $\boldsymbol{\xi_3}$ and $\boldsymbol{\xi_4}$ are found first.

For $s = 0.3$, we observe, likewise, that the best values of RI and ARI, corresponding to the minimal value of BIC, are reached for $t = 0.2$ but they are lower than that in the $s = 0.1$ case. As noticed by Chauvet et al. (2019), the thinner the bundles, the greater the value of s has to be, to recover the latent variables correctly. Here, indeed, the error variance being low ($\sigma^2 = 0.1$), the square correlations are, on the whole,

greater for $s = 0.3$ than for $s = 0.1$. As in the case $s = 0.1$, $\rho^2(\boldsymbol{\xi_1}, .)$ and $\rho^2(\boldsymbol{\xi_2}, .)$ decrease with t. However, contrary to the case $s = 0.1$, the increase of the square correlations $\rho^2(\boldsymbol{\xi_3}, .)$ and $\rho^2(\boldsymbol{\xi_4}, .)$ could not be observed, since t could not exceed 0.6.

Finally, for $s = 0.5$, we observe the dramatic effect of taking too many components in a group. For all values of t, the RI and the ARI are respectively close to 0.5 and 0. This indicates that for $s = 0.5$ and $(H_1, H_2) = (3, 1)$, the obtained classification is not better than a random one.

Table 2: Mean values of RI, ARI, square correlation and BIC over a hundred samples, for a high correlation value ($\rho = 0.9$) between the latent variables $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$, for $s \in \{0.1, 0.3, 0.5\}$, the optimized combination of components and t ranging from 0 to 0.8.

| | | | | group 1 | | group 2 | | |
s	t	RI	ARI	$\rho^2(\boldsymbol{\xi_1}, .)$	$\rho^2(\boldsymbol{\xi_3}, .)$	$\rho^2(\boldsymbol{\xi_2}, .)$	$\rho^2(\boldsymbol{\xi_4}, .)$	BIC
	0	0.860	0.718	0.971	0.950	0.963	0.927	29095.04
	0.1	0.861	0.721	0.970	0.951	0.955	0.907	29085.84
	0.2	0.865	0.729	0.966	0.939	0.938	0.888	28963.32
	0.3	0.870	0.738	0.931	0.889	0.913	0.878	28955.93
0.1	**0.4**	**0.883**	**0.764**	**0.899**	**0.889**	**0.893**	**0.874**	**28950.69**
	0.5	0.873	0.745	0.857	0.878	0.858	0.847	29531.91
	0.6	0.853	0.705	0.835	0.859	0.856	0.861	29705.92
	0.7	0.844	0.684	0.841	0.907	0.853	0.881	30302.17
	0.8	0.693	0.378	0.788	0.934	0.865	0.900	31805.47
	0	0.799	0.595	0.958	0.967	0.957	0.927	29497.32
	0.1	0.814	0.626	0.956	0.967	0.956	0.939	29493.86
	0.2	**0.815**	**0.629**	**0.957**	**0.956**	**0.965**	**0.970**	**29489.26**
0.3	0.3	0.812	0.623	0.957	0.958	0.955	0.959	29518.38
	0.4	0.796	0.591	0.951	0.958	0.950	0.951	29519.18
	0.5	0.794	0.589	0.919	0.915	0.917	0.911	29528.79
	0.6	0.792	0.582	0.813	0.795	0.814	0.815	29532.73
	0	0.560	0.039	0.948	0.918	0.948	0.911	29581.08
	0.1	0.572	0.054	0.945	0.896	0.951	0.897	29518.82
0.5	0.2	0.562	0.044	0.943	0.872	0.949	0.883	29541.87
	0.3	0.556	0.033	0.945	0.902	0.947	0.902	29531.97
	0.4	0.551	0.025	0.945	0.938	0.945	0.910	29606.57

For the sake of visualization, Figure 1 shows the correlation scatterplots of plane

$(1, 2)$ for each group. We can see that the components $\boldsymbol{f}$ are well aligned with the corresponding simulated latent variables $\boldsymbol{\xi}$, except for $\boldsymbol{f_2^2}$, which slightly deviates from $\boldsymbol{\xi_4}$. Due to the high correlation between $\boldsymbol{\xi_1}$ and $\boldsymbol{\xi_2}$, the bundles $\boldsymbol{X_1}$ (in red) and $\boldsymbol{X_2}$ (in blue) are both well aligned with the first component of each group.
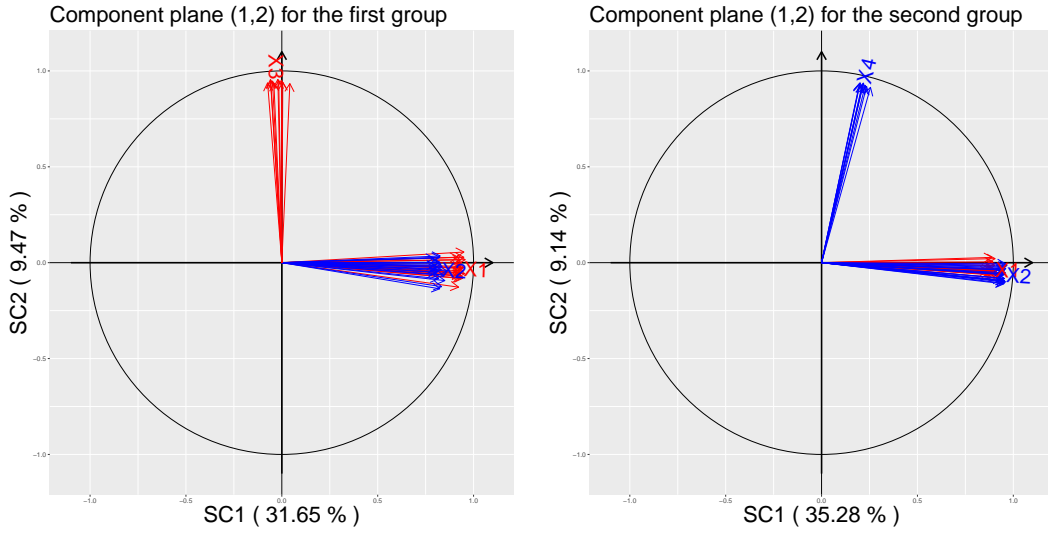

Figure 1: Correlation scatterplot of plane $(1,2)$ for the two groups with $s = 0.1$ and $t = 0.4$ obtained by the rmSCGLR algorithm. The red arrows represent the bundles $\boldsymbol{X_1}$ and $\boldsymbol{X_3}$ which explain the first group. The blue ones represent the bundles $\boldsymbol{X_2}$ and $\boldsymbol{X_4}$ which explain the second group. The percentage of inertia captured by each component is given in parentheses.

As reference values for comparison, we also calculated the RI and ARI of the partitions output by, on the one hand, the ® package **ClustOfVar**, employed to initialize our algorithm, and, on the other hand, the ® package **ecomix** implementing the approach proposed by Dunstan et al. (2011, 2013). However, **ecomix** not allowing to consider different distribution families for the responses, we restricted the comparison to the case of Gaussian responses. Thus, with the previous generated data, we have twenty responses in the first group and ten in the second one. Table 3 presents the results. As expected, in a context of component-based model, the **ecomix** classification does not outperform the random classification. The classification output by **ClustOfVar**

is slightly better, but only provides a good starting point for **rmSCGLR**, which leads to high values of RI and ARI. We may note that **rmSCGLR** offers a greater classification performance than in the case of mixed distribution families.

Table 3: Mean values of RI and ARI over a hundred samples for the ℝ packages **rmSCGLR**, **ClustOfVar** and **ecomix**.

rmSCGLR		ClustOfVar		ecomix	
RI	ARI	RI	ARI	RI	ARI
0.931	0.864	0.551	0.130	0.496	0.031

5   Analysis of a floristic ecology dataset

5.1   Data description

We apply rmSCGLR to the *CoForTaxa* dataset available on demand at http://dx.doi.org/10.18167/DVN1/UCNCA7. The sample we consider gives the abundances of $K = 193$ floristic taxa in the Congo basin rainforest over a $N = 1571$ $10{\times}10$-km^2 grid cells across central Africa. To predict abundances, we have $P = 24$ climatic variables and $Q = 3$ non-climatic additional variables. $\boldsymbol{X}$ consists of all the climatic variables, i.e.: eleven temperature variables coded "C1",...,"C11", eight precipitation variables coded "C12",...,"C19", three climatic water deficit variables coded "sumCWD", "maxCWD" and "MCWD" respectively, one climatic water balance coded "meanCWB" and one evapotranspiration variable coded "meanET0". Figure 2 shows the correlation plot given by the Principal Component Analysis (PCA) of the climatic variables. Since it appears that the explanatory variables exhibit a clear bundle structure, a methodology such as SCGLR is necessary to regularize the model estimation and reduce the dimension of the explanatory space. Besides, the

non-climatic variables, i.e. the soil type (Harmonized World Soil Database, "HWSD") and the human-induced forest-disturbance intensity index ("Anthr2"), as well as its logarithm ("logAnthr2") to account for nonlinear effects, are few and weakly correlated with the variables in $\boldsymbol{X}$ as well as between themselves, and interesting per se. We shall then consider them as additional explanatory variables, and gather them in matrix $\boldsymbol{A}$. The response variables are assumed to be Poisson random variables, independent conditional on $\boldsymbol{X}$ and $\boldsymbol{A}$. Moreover, the variable corresponding to the number of plots within each grid cell is taken as the offset of the Poisson regression. For more information about the *CoForTaxa* dataset, we refer the reader to Réjou-Méchain et al. (2021).
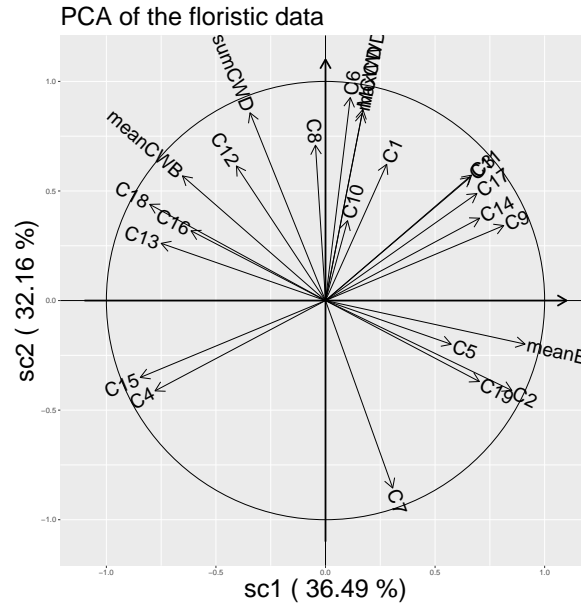

Figure 2: Component plane (1,2) of the explanatory climatic variables obtained through PCA. The percentage of inertia captured by each component is given in parentheses.

5.2   Hyper-parameter calibration

We present the results obtained when following the parameter-varying scheme presented in Section 3.4. As noticed by Réjou-Méchain et al. (2021), the tuning parameters $s = 0.1$ and $l = 1$ allow to optimize SCGLR on *CoForTaxa* dataset. Here, thanks to the heuristic, $G = 3$ groups are retained to carry on with the analysis, using the previously found values of the tuning parameters. Starting with one component per group, we increment the number of components by one in each group alternately. Only adding one in the third group improves the criterion. When Réjou-Méchain et al. (2021) applied the basic SCGLR (without response mixture) to these data, three relevant components were found. The combination $H = (1, 1, 2)$ thus does not seem irrelevant. To get a refined model with this combination of components, the tuning parameter t needs to be raised to 0.5 to allow to better distinguish the groups, and minimize the BIC.

5.3   Results and interpretation

The rmSCGLR methodology aims at identifying response groups having specific explanatory spaces. However, one may have to put a heavy weight on the separation criterion, e.g. $(t = 0.5)$, in order to identify such groups. As a result, the associated explanatory spaces may first be too artificially separated, which would degrade the quality of prediction. So, in an exploratory first step, we may use rmSCGLR merely to identify groups and their most specific explanatory dimensions, which we analyze so as to identify group-specific explanatory variables. Then, in a predictive second step, we perform classical SCGLR on each of the identified groups to fine-tune the prediction, and validate the findings provided by rmSCGLR.

Applying rmSCGLR led here to three groups of taxa. Two of them were associated with a single explanatory component, and the last one with two components. The groups respectively comprise 44, 67 and 82 taxa. The contents of the groups are given in Supplementary Material. We sum up the first two groups in Table 4, stating the explanatory variables most correlated with the components. Table 4 does not deal with the third group, as this one appears in the sequel to be something of a "junk" group with no homogeneous interpretation.

Table 4: Lists of explanatory variables most correlated with the component in each of the first two groups. Only correlations over 0.8 are given.

Groups	Explanatory variables	Correlation
1	C7, sumCWD, MCWD, maxCWD,	0.956, -0.955, -0.885, -0.880
2	C2, meanET0, C18, C19	-0.930, -0.929, 0.925, -0.862

The component of the first group is highly correlated with the variable "C7" (difference between the maximum of temperature of the warmest month and the minimum of temperature of the coldest month), and with the three climatic water deficit variables: "sumCWD", "MCWD" and "maxCWD". Thus, the abundances of taxa composing the first group would be linked to a gradient of temperature, and sensitive to a water deficit. The component of the second group is highly correlated with "C2" (the mean diurnal range), "meanET0" (the mean monthly evapotranspiration) and with "C18" and "C19" (the precipitations of the warmest quarter and the coldest quarter, respectively). This component is very similar to the first component found if we apply SCGLR on all the responses ($\rho = -0.965$). According to Réjou-Méchain et al. (2021), this component is highly related to a regional floristic gradient contrasting areas with a cool and light-deficient dry season (coastal Gabon) and areas with high evapotranspiration rates (northern limit of the central African forests). The components of the third group fail to be aligned with any bundle of variables. The correspond-

ing scatterplot is given in Supplementary Material. As mentioned by Réjou-Méchain et al. (2021), a majority of taxon abundances may relate with climate only by chance. Thus, by contrast to the first and second group, where the abundances are linked to water deficit or precipitation, the taxa composing the third group are not connected with any specific gradient but with various combinations of climatic variables.

These results are validated by separately performing SCGLR on each group. In the first group, SCGLR finds a single component, highly correlated ($\rho = 0.960$) with $\boldsymbol{f_1^1}$ of rmSCGLR. Three components are calculated by SCGLR to best predict the second group. However, on Figure 3a, we can see that all the linear predictors of the taxa's abundances composing the second group are highly correlated with the bundle found by $\boldsymbol{f_2^1}$ of rmSCGLR. The second and third components only provide a secondary improvement in predicting the abundances. The correlation between the first SCGLR-component of the second group and rmSCGLR's $\boldsymbol{f_2^1}$ is equal to -0.991. As expected for the third group of taxa, Figure 3b shows no particular correlation pattern between the linear predictors and any bundle, which highlights the absence of specific climatic gradient in this group's explanatory space. The planes spanned by the higher rank components are given in Supplementary Material.

Let us evaluate the benefits obtained by taking into account the clustering found by rmSCGLR. In (Réjou-Méchain et al., 2021), the quality of prediction was given by the mean of cross-validation Mean Squared Prediction Errors (MSPE), and we shall use the same index for comparison. We shall thus compare: (i) the prediction error we get with SCGLR on all taxa, named MSPE_{all}, (ii) the prediction error obtained with SCGLR on the three groups separately, named MSPE_1, MSPE_2 and MSPE_3 respectively, with their weighted mean named $\text{MSPE}_{\text{mean}}$, and (iii) the mean of the pre-

(a) Component plane $(1, 2)$ for the group 2

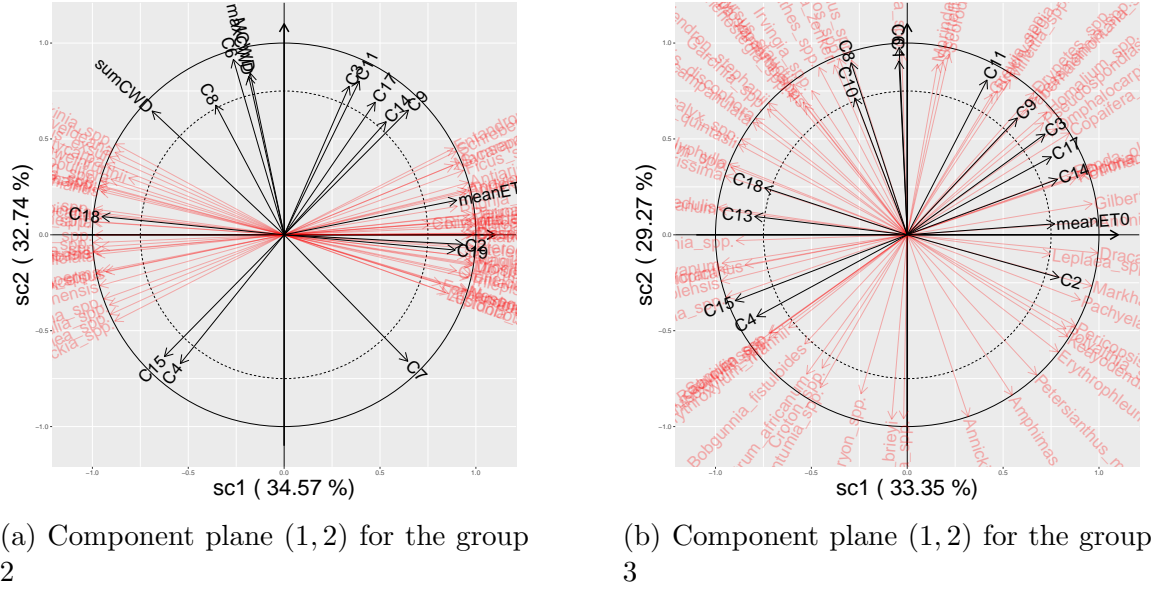(b) Component plane $(1, 2)$ for the group 3

Figure 3: Correlation scatterplots of plane (1,2) with linear predictors for the second and third separated groups obtained by the SCGLR algorithm. The black arrows represent the covariates. The red ones are the linear predictors of the responses. The plot displays only variables having a cosine over 0.75. The percentage of inertia captured by each component is given in parentheses.

diction error on random partitions into three groups of taxa, obtained over a hundred samples, named $\text{MSPE}_{\text{random}}$. The prediction error of SCGLR on all taxa was calculated by Réjou-Méchain et al. (2021), and found to be $\text{MSPE}_{\text{all}} = 3.23$. SCGLR, performed separately on the first and second groups, gave the following prediction errors: $\text{MSPE}_1 = 3.07$ and $\text{MSPE}_2 = 2.94$ respectively, which indicates an improved quality of prediction. However, the prediction error of the third group rises to $\text{MSPE}_3 = 3.41$, which indicates that group 3 is composed by taxa the abundances of which are poorly predictable from the sheer observed climatic variables. Finally, the mean prediction error of SCGLR accounting for the partition is: $\text{MSPE}_{\text{mean}} = 3.17$. The mean prediction error accounting for a random 3-group partition is: $\text{MSPE}_{\text{random}} = 3.20$. This shows that rmSCGLR was able to, if only slightly, better capture the explanatory structure of the floristic data. It should be noted that prediction of taxa abundances

from merely such climatic variables is usually poor (Beale et al., 2008).

6   Conclusion and discussion

The original SCGLR was designed to regularize GLM estimation and reduce the explanatory dimension through components, so as to decompose the linear predictor in an interpretable way. It allowed to find strong and interpretable supervised components common to response variables, by achieving a trade-off between Goodness-of-Fit and a Structural Relevance measure. The response mixture SCGLR extends SCGLR in two major ways: (i) Through a mixture model on the response variables, it identifies groups of responses that can be predicted from group-specific components. Doing so, this method improves both the prediction quality of the response groups, and the interpretation of what explains the responses. In our ecological framework, we detected communities of taxa sensitive to specific gradients of climate variables. (ii) It extends the criterion to be maximized by introducing a separation sub-criterion, which allows to specify sub-spaces which components had better keep away from. In the context of Response Mixture, this sub-criterion helped distinguish the groups by better separating their explanatory sub-spaces.

In our simulation study, rmSCGLR proved to behave as expected regarding groups. In a context of very close explanatory sub-spaces, it recovered the original groups, and provided components aligned with the latent variables. On the floristic ecology dataset, we found three communities of taxa. The first one is linked to a gradient of temperature, while the second one is connected to a regional floristic gradient contrasting two main areas. The last group gathers the taxa related to no specific gra-

dient, but to many combinations of the observed climatic variables. More predictive climatic components could likely be generated after removing these taxa.

In this research, some limitations have been reached. Just as the original SCGLR, our method does not allow to deal with a thematic partition of the explanatory variables. To overcome this limitation, we could extend THEME-SCGLR (Bry et al., 2020b) to a response mixture. For instance, the temperature and precipitation variables would be seen as pertaining to two distinct themes and each community of taxa would be predicted by common components with respect to each theme. In the spirit of Durif et al. (2018), another way of refining our model would be to create sparse components with intent to select relevant climatic variables. By definition, the heuristic presented in Section 3.4 does not guarantee to find the best values of the hyper-parameters. Several parameter-varying schemes could be implemented and the results compared. Hutter et al. (2015) propose a review of works allowing to best optimize the hyper-parameters.

Acknowledgements

References

Bastien, P., Vinzi, V. E., and Tenenhaus, M. (2005). PLS generalised linear regression. *Computational Statistics & data analysis*, **48**(1), 17–46.

Beale, C. M., Lennon, J. J., and Gimona, A. (2008). Opening the climate envelope reveals no macroscale associations with climate in European birds. *Proceedings of the National Academy of Sciences*, **105**(39), 14908–14912.

Bry, X. and Verron, T. (2015). THEME: THEmatic Model Exploration through multiple co-structure maximization. *Journal of Chemometrics*, **29**(12), 637–647.

Bry, X., Redont, P., Verron, T., and Cazes, P. (2012). THEME-SEER: a multidimensional exploratory technique to analyze a structural model using an extended covariance criterion. *Journal of Chemometrics*, **26**(5), 158–169.

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.

Bry, X., Simac, T., El Ghachi, S. E., and Antoine, P. (2020a). Bridging data exploration and modeling in event-history analysis: the supervised-component Cox regression. *Mathematical Population Studies*, **27**(3), 139–174.

Bry, X., Trottier, C., Mortier, F., and Cornu, G. (2020b). Component-based regularization of a multivariate GLM with a thematic partitioning of the explanatory variables. *Statistical Modelling*, **20**(1), 96–119.

Chauvet, J., Trottier, C., and Bry, X. (2019). Component-Based Regularization of Multivariate Generalized Linear Mixed Models. *Journal of Computational and Graphical Statistics*, **28**(4), 909–920.

Chavent, M., Simonet, V. K., Liquet, B., and Saracco, J. (2012). ClustOfVar: An R Package for the Clustering of Variables. *Journal of Statistical Software*, **50**(13), 1–16.

De Cáceres, M., Legendre, P., and Moretti, M. (2010). Improving indicator species analysis by combining groups of sites. *Oikos*, **119**(10), 1674–1684.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, **39**(1), 1–22.

Dufrêne, M. and Legendre, P. (1997). Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecological monographs*, **67**(3), 345–366.

Dunstan, P. K., Foster, S. D., and Darnell, R. (2011). Model based grouping of species across environmental gradients. *Ecological Modelling*, **222**(4), 955–963.

Dunstan, P. K., Foster, S. D., Hui, F. K., and Warton, D. I. (2013). Finite mixture of regression modeling for high-dimensional count and biomass data in ecology. *Journal of agricultural, biological, and environmental statistics*, **18**(3), 357–375.

Durif, G., Modolo, L., Michaelsson, J., Mold, J. E., Lambert-Lacroix, S., and Picard, F. (2018). High dimensional classification with combined adaptive sparse PLS and logistic regression. *Bioinformatics*, **34**(3), 485–493.

Guisan, A. and Thuiller, W. (2005). Predicting species distribution: offering more than simple habitat models. *Ecology letters*, **8**(9), 993–1009.

Hill, N., Woolley, S. N., Foster, S., Dunstan, P. K., McKinlay, J., Ovaskainen, O., and Johnson, C. (2020). Determining marine bioregions: A comparison of quantitative approaches. *Methods in Ecology and Evolution*, **11**(10), 1258–1272.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, **2**(1), 193–218.

Hutter, F., Lücke, J., and Schmidt-Thieme, L. (2015). Beyond manual tuning of hyperparameters. *KI-Künstliche Intelligenz*, **29**(4), 329–337.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 49–66.

Marx, B. D. (1996). Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*, **38**(4), 374–381.

McCullagh, P. and Nelder, J. (1989). Generalized Linear Models, Chapman and Hall, New York, NY.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Mevik, B.-H. and Wehrens, R. (2007). The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software*, **18**(i02).

Monni, S. and Tadesse, M. G. (2009). A stochastic partitioning method to associate high-dimensional responses and covariates. *Bayesian Analysis*, **4**(3), 413–436.

Mortier, F., Ouédraogo, D.-Y., Claeys, F., Tadesse, M. G., Cornu, G., Baya, F., Benedet, F., Freycon, V., Gourlet-Fleury, S., and Picard, N. (2015). Mixture of inhomogeneous matrix models for species-rich ecosystems. *Environmetrics*, **26**(1), 39–51.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, **135**(3), 370–384.

Ovaskainen, O. and Soininen, J. (2011). Making more out of sparse data: hierarchical modeling of species communities. *Ecology*, **92**(2), 289–295.

Pledger, S. and Arnold, R. (2014). Multivariate methods using mixtures: Correspondence analysis, scaling and pattern-detection. *Computational Statistics & Data Analysis*, **71**, 241–261.

Pollock, L. J., Tingley, R., Morris, W. K., Golding, N., O'Hara, R. B., Parris, K. M., Vesk, P. A., and McCarthy, M. A. (2014). Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, **5**(5), 397–406.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, **66**(336), 846–850.

Réjou-Méchain, M., Mortier, F., Bastin, J.-F., Cornu, G., Barbier, N., Bayol, N., Bénédet, F., Bry, X., Dauby, G., Deblauwe, V., et al. (2021). Unveiling African rainforest composition and vulnerability to global change. *Nature*, **593**, 90–94.

Warton, D. I., Foster, S. D., De'ath, G., Stoklosa, J., and Dunstan, P. K. (2015). Model-based thinking for community ecology. *Plant Ecology*, **216**(5), 669–682.

Wold, S., Ruhe, A., Wold, H., and Dunn, III, W. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific and Statistical Computing*, **5**(3), 735–743.

Yee, T. W. and Hastie, T. J. (2003). Reduced-rank vector generalized linear models. *Statistical modelling*, **3**(1), 15–41.

7   Supplementary Materials

7.1   The PING algorithm

The Projected Iterated Normed Gradient (PING) algorithm is an extension of the Power Iteration algorithm. To find the hth component, we use the PING algorithm which aims at solving any program of the form:

$$
\begin{cases}
\underset{u}{\operatorname{argmax}} \quad J_h(u), \\
\text{s.c.} \quad u^T M^{-1} u = 1 \quad \text{et} \quad \Delta_h^T u = 0.
\end{cases}
\tag{7.1}
$$

We rewrite this program by posing $v = M^{-1/2}u$, $G_h(v) = J_h(M^{1/2}v)$ and $E_h = M^{1/2}\Delta_h$.

$$
\begin{cases}
\underset{v}{\operatorname{argmax}} \quad G_h(v), \\
\text{s.c.} \quad v^T v = 1 \quad \text{et} \quad E_h^T v = 0.
\end{cases}
\tag{7.2}
$$

To solve (7.2), we must equate to zero the gradient of the following Lagrangian:

$$
L(v, \lambda, \eta) = G_h(v) - \lambda(v^T v - 1) - \eta^T E_h^T v.
$$

Setting $\Gamma_h(v) = \nabla_v G_h(v)$, we have

$$
\nabla_v L(v, \lambda, \eta) = 0 \Leftrightarrow \Gamma_h(v) - 2\lambda v - E_h \eta = 0
\tag{7.3}
$$

$$
\Leftrightarrow v = \frac{1}{2\lambda} \left(\Gamma_h(v) - E_h \eta \right).
\tag{7.4}
$$

Multiplying (7.3) by E_h^T:

$$2\lambda \underbrace{E_h^T v}_{=0} = E_h^T \Gamma_h(v) - E_h^T E_h \eta \Leftrightarrow E_h^T \Gamma_h(v) = E_h^T E_h \eta$$

$$\Leftrightarrow \eta = \left(E_h^T E_h\right)^{-1} E_h^T \Gamma_h(v). \tag{7.5}$$

Substituting (7.5) in (7.4), we get:

$$\begin{aligned}
v &= \frac{1}{2\lambda} \left(\Gamma_h(v) - E_h \left(E_h^T E_h\right)^{-1} E_h^T \Gamma_h(v)\right) \\
&= \frac{1}{2\lambda} \left(I - E_h \left(E_h^T E_h\right)^{-1} E_h^T\right) \Gamma_h(v) \\
&= \frac{1}{2\lambda} \Pi_{\mathrm{span}[E_h]^\perp} \Gamma_h(v),
\end{aligned}$$

where $\Pi_{\mathrm{span}[E_h]^\perp} = I - E_h \left(E_h^T E_h\right)^{-1} E_h^T$. Finally, the constraint $\|v\|^2 = 1$ gives

$$v = \frac{\frac{1}{2\lambda} \Pi_{\mathrm{span}[E_h]^\perp} \Gamma_h(v)}{\left\|\frac{1}{2\lambda} \Pi_{\mathrm{span}[E_h]^\perp} \Gamma_h(v)\right\|} = \frac{\Pi_{\mathrm{span}[E_h]^\perp} \Gamma_h(v)}{\left\|\Pi_{\mathrm{span}[E_h]^\perp} \Gamma_h(v)\right\|},$$

which suggests the basic iteration of the PING algorithm:

$$v^{(t+1)} = \frac{\Pi_{\mathrm{span}[E_h]^\perp} \Gamma_h(v^{(t)})}{\left\|\Pi_{\mathrm{span}[E_h]^\perp} \Gamma_h(v^{(t)})\right\|}. \tag{7.6}$$

Let us show that the basic iteration of the PING algorithm follows a direction of ascent. One way to do this is to show that the direction given by the arc $(v^{(t)}, v^{(t+1)})$ is a direction of ascent. In other words, show that:

$$\left\langle v^{(t+1)} - v^{(t)}, \Gamma_h\left(v^{(t)}\right)\right\rangle \geq 0.$$

By construction, we know that on every iteration t of the algorithm, $v^{(t)}$ is orthogonal to $\mathrm{span}[E_h]$. Thus, since for all t, $v^{(t)} = \Pi_{\mathrm{span}[E_h]^\perp} v^{(t)}$, we have

$$
\begin{aligned}
\left\langle v^{(t+1)} - v^{(t)}, \Gamma_h\left(v^{(t)}\right) \right\rangle &= \left\langle \Pi_{\mathrm{span}[E_h]^\perp}\left(v^{(t+1)} - v^{(t)}\right), \Gamma_h\left(v^{(t)}\right) \right\rangle \\
&= \left\langle v^{(t+1)} - v^{(t)}, \Pi_{\mathrm{span}[E_h]^\perp}\Gamma_h\left(v^{(t)}\right) \right\rangle.
\end{aligned}
$$

Now, the equation (7.6) implies that

$$
\Pi_{\mathrm{span}[E_h]^\perp}\Gamma_h(v^{(t)}) = v^{(t+1)}\left\|\Pi_{\mathrm{span}[E_h]^\perp}\Gamma_h(v^{(t)})\right\|.
$$

So,

$$
\begin{aligned}
\mathrm{sgn}\left(\left\langle v^{(t+1)} - v^{(t)}, \Gamma_h\left(v^{(t)}\right) \right\rangle\right) &= \mathrm{sgn}\left(\left\langle v^{(t+1)} - v^{(t)}, v^{(t+1)} \right\rangle\right) \\
&= \mathrm{sgn}\left(\left\|v^{(t+1)}\right\|^2 - \left\langle v^{(t)}, v^{(t+1)} \right\rangle\right) \\
&= \mathrm{sgn}\left(1 - \cos\left(v^{(t)}, v^{(t+1)}\right)\right).
\end{aligned}
$$

Finally,

$$
\left\langle v^{(t+1)} - v^{(t)}, \Gamma_h\left(v^{(t)}\right) \right\rangle \geq 0.
$$

Although iteration (7.6) follows a direction of ascent, it does not guarantee that function G actually increases on every step. Indeed, we may go too far in such a direction, and overshoot the maximum. However, let us consider

$$
\kappa^{(t)} = \frac{\Pi_{\mathrm{span}[E_h]^\perp}\Gamma_h(v^{(t)})}{\left\|\Pi_{\mathrm{span}[E_h]^\perp}\Gamma_h(v^{(t)})\right\|}.
$$

Staying close enough to the current starting point on the arc $(v^{(t)}, \kappa^{(t)})$ ensures that

function G increases on every iteration. With this aim in mind, let ϖ be the plane tangent to the unit sphere on $v^{(t)}$ and let w denote the unit–vector tangent to arc $(v^{(t)}, \kappa^{(t)})$ on $v^{(t)}$. Then, there exists $\tau > 0$ such that, $w = \tau \Pi_\varpi \kappa^{(t)}$, and

$$\left\langle \varpi, \kappa^{(t)} \right\rangle = \tau \left\langle \Pi_\varpi \kappa^{(t)}, \kappa^{(t)} \right\rangle = \tau \cos^2(\kappa^{(t)}, \varpi) > 0.$$

Although staying close enough to the current starting point on the arc $(v^{(t)}, \kappa^{(t)})$ ensures the increase of function G, staying too close can impact the convergence speed of the algorithm to reach the maximum. On the other hand, going too far from the starting point can cause the divergence of the algorithm. Therefore, we propose two possible generic iterations for the PING algorithm, which deal with this problem. Algorithm 1 and Algorithm 2 present these alternatives. The first one should be preferred, but is less easy to program.

Algorithm 1: PING algorithm

---

while *not convergence* **do**

$$\kappa^{(t)} \leftarrow \frac{\Pi_{\text{span}[E_h]^\perp} \Gamma_h(v^{(t)})}{\left\| \Pi_{\text{span}[E_h]^\perp} \Gamma_h(v^{(t)}) \right\|}$$

Use a Newton-Raphson unidimensional maximization procedure to find the maximum of $G_h(v)$ on the arc $(v^{(t)}, \kappa^{(t)})$ and take it as $v^{(t+1)}$

$t \leftarrow t + 1$

end

Algorithm 2: Alternative PING algorithm

---

while *not convergence* **do**

$$m \leftarrow \frac{\Pi_{\text{span}[E_h]^\perp} \Gamma_h(v^{(t)})}{\left\| \Pi_{\text{span}[E_h]^\perp} \Gamma_h(v^{(t)}) \right\|}$$

while $G_h(m) < G_h(v^{(t)})$ **do**

$$m \leftarrow \frac{v^{(t)} + m}{\| v^{(t)} + m \|}$$

end

$v^{(t+1)} \leftarrow m$

$t \leftarrow t + 1$

end

7.2   The EM algorithm

Owing to the latent variable Z, this step will be performed using the EM algorithm. The M step of the EM algorithm consists in maximizing with respect to Θ the conditional expectation of the complete log-likelihood $\mathbb{E}\left[l(\Theta; Y, Z)|Y; \Theta'\right]$. The solution replaces then Θ', and the conditional expectation is updated in the E step.

7.2.1   The expectation (E) step

The expectation of the complete log-likelihood writes

$$\mathbb{E}\left[l(\Theta; Y, Z)|Y; \Theta'\right] = \sum_{k=1}^{K}\sum_{g=1}^{G} \alpha_{kg} \ln\left(p_g \prod_{n=1}^{N} d_k\left(y_{nk}; \mu_{nkg}\right)\right).$$

The posterior probability is computed as

$$\alpha_{kg} := \mathbb{P}(z_k = g|y_k; \theta_k) = \frac{p_g \prod_{n=1}^{N} d_k(y_{nk}; \mu_{nkg})}{\sum_{r=1}^{G} p_r \prod_{n=1}^{N} d_k(y_{nk}; \mu_{nkr})},$$

with $z_k = g$ meaning that the gth coordinate of the vector z_k equals 1. As noticed by Dunstan et al. (2013), the α_{kg}'s are likely to be very close to either 0 or 1, and this polarization grows with the number of observations. In this case, the EM algorithm is liable to get stuck and does not provide a satisfactory exploration of the parameter space. We thus shrank the α_{kg}'s, for five iterations of the EM algorithm only, using

$$\alpha_{kg}^* = \frac{2\tau\alpha_{kg} - \tau + 1}{2\tau - \tau G + G} \quad \text{where} \quad \tau = \frac{1 - 0.8G}{0.8(2 - G) - 1}.$$

The previous formula prevents any α_{kg}^* from being greater than 0.8 or lower than $(1 - 0.8)/(G - 1)$, while maintaining the sum-to-one constraint.

7.2.2   The maximization (M) step

The maximization step maximizes the conditional expectation of the complete log-likelihood with respect to Θ, subject to the constraint $\sum_{g=1}^{G} p_g = 1$. The maximization with respect to p_g yields:

$$\hat{p}_g = \frac{1}{K} \sum_{k=1}^{K} \alpha_{kg}.$$

The estimates of the regression parameters γ_{kg} and δ_{kg} are obtained as the solutions of:

$$\nabla_{(\gamma_{kg}, \delta_{kg})} \sum_{n=1}^{N} \ln(d_k(y_{nk}; \mu_{nkg})) = 0.$$

This equation characterizes the maximum likelihood estimate of the GLM of y_k in each group g. This estimate can be obtained as the fixed point of the FSA.

Assuming the response variable y_k belongs to the gth group, the working variable associated with y_{nk} is calculated as:

$$w_{nkg} = h_k\left(\mu_{nkg}\right) + \left(y_{nk} - \mu_{nkg}\right) h_k'\left(\mu_{nkg}\right) = \eta_{nkg} + \zeta_{nkg},$$

where

$$\zeta_{nkg} = \left(y_{nk} - \mu_{nkg}\right) h_k'\left(\mu_{nkg}\right).$$

In view of the conditional independence assumption, the variance matrix for w_{nkg} is

$$\mathbb{V}\left[w_{kg}\right] = W_{kg}^{-1} = \mathrm{diag}\left(a_{nk}(\phi_k) v_k\left(\mu_{nkg}\right) h_k'\left(\mu_{nkg}\right)^2\right)_{n=1,\dots,N},$$

where a_{nk} and v_k are known functions and ϕ_k is the dispersion parameter of y_k. Thus, to optimize the regression parameters, we perform a generalized least square step on

the linearized model defined by:

$$w_{kg} = (X u_g) \gamma_{kg} + A \delta_{kg} + \zeta_{kg},$$

with $\mathbb{E}(\zeta_{kg}) = 0$ and $\mathbb{V}(\zeta_{kg}) = W_{kg}^{-1}$.

As a result of the aforementioned developments, we shall use Algorithm 3 to estimate the parameters of the response mixture model.

Algorithm 3: The EM algorithm adapted to the response mixture

---

Input : $A_g := [f_g, A]$

while *not convergence* **do**

    Expectation step

    for $k = 1, \ldots, K$ **do**

        for $g = 1, \ldots, G$ **do**

$$\alpha_{kg}^{(t+1)} = \frac{p_g^{(t)} \prod_{n=1}^{N} d_k(y_{nk}; \mu_{nkg}^{(t)})}{\sum_{r=1}^{G} p_r^{(t)} \prod_{n=1}^{N} d_k(y_{nk}; \mu_{nkr}^{(t)})}$$

        end

    end

    Maximization step

    for $g = 1, \ldots, G$ **do**

$$p_g^{(t+1)} = \tfrac{1}{K} \sum_{k=1}^{K} \alpha_{kg}^{(t+1)}$$

        for $k = 1, \ldots, K$ **do**

$$\left(\gamma_{kg}^{(t+1)}, \delta_{kg}^{(t+1)\,T} \right)^T = \left(A_g^T W_{kg}^{(t)} A_g \right)^{-1} A_g^T W_{kg}^{(t)} w_{kg}^{(t)}$$

$$\eta_{kg}^{(t+1)} = f_g \gamma_{kg}^{(t+1)} + A\delta_{kg}^{(t+1)}$$

$$\mu_{nkg}^{(t+1)} = h_k^{-1}\left(\eta_{nkg}^{(t+1)} \right), \forall n = 1, \ldots, N$$

$$w_{nkg}^{(t+1)} = \eta_{nkg}^{(t+1)} + h_k'\left(\mu_{nkg}^{(t+1)} \right)\left(y_{nk} - \mu_{nkg}^{(t+1)} \right), \forall n = 1, \ldots, N$$

$$W_{kg}^{(t+1)} = \operatorname{diag}\left(\left[a_{nk}(\phi_k) v_k\left(\mu_{nkg}^{(t+1)} \right) h_k'\left(\mu_{nkg}^{(t+1)} \right)^2 \right]^{-1} \right)_{n=1,\ldots,N}$$

        end

    end

    $t \leftarrow t + 1$

end

7.3   Analytical expression of the specific criterion

The specific criterion which SCGLR maximizes to compute the $(h+1)$th loading-vector u^{h+1} writes

$$J(u) = \phi(u)^s \varphi(u)^t \psi_{A^h}(u)^{1-s-t},$$

with

$$
\begin{cases}
\phi(u) = \left(\sum_{j=1}^{p} \omega_j (u^T N_j u)^l \right)^{1/l} \\
\varphi(u) = 1 - \dfrac{1}{G} \sum_{g=1}^{G} \left\langle \varpi_{[F^h, Xu]}, \varpi_{E_g} \right\rangle \\
\psi_{A^h}(u) = \sum_{k=1}^{K} \|w_k\|_{W_k}^2 \cos_{W_k}^2 \left(w_k, \mathrm{span}[Xu, A^h] \right),
\end{cases}
\tag{7.7}
$$

where $F^h = [Xu^1, \ldots, Xu^h]$ and $A^h = [F^h, A]$.

To facilitate the computation of the loading-vector, we hereafter give an analytical expression of each sub-criterion and of its gradient.

7.3.1   The structural relevance measure

The general form of the structural relevance (SR) is $\phi(u)$ written in (7.7). However, in practice, we take either the variance component (VC) or the variable power inertia (VPI). In the first case, the SR and its gradient are easily given by

$$\phi(u) = \|Xu\|_W^2 \quad \text{and} \quad \nabla_u \phi(u) = 2X^T W X u.$$

The explicit expression of VPI is

$$\phi(u) = \left(\frac{1}{p} \sum_{j=1}^{p} \langle Xu, x_j \rangle_W^{2l} \right)^{1/l}.$$

To calculate the gradient we use the classical rules of derivation:

$$\nabla_u \phi(u) = \frac{1}{l} \left[\nabla_u \left(\frac{1}{p} \sum_{j=1}^{p} \langle Xu, x_j \rangle_W^{2l} \right) \right] \left[\frac{1}{p} \sum_{j=1}^{p} \langle Xu, x_j \rangle_W^{2l} \right]^{1/l-1}$$

$$= \frac{1}{l} \left[\frac{1}{p} \sum_{j=1}^{p} 2l X^T W x_j \langle Xu, x_j \rangle_W^{2l-1} \right] \phi(u)^{1-l}$$

$$= \frac{2}{p} \phi(u)^{1-l} X^T W \sum_{j=1}^{p} \langle Xu, x_j \rangle_W^{2l-1} x_j.$$

7.3.2   The goodness of fit measure

We aim at expressing $\psi_{A_h}(u)$ as a function of quadratic forms. To achieve that, we decompose the projection on the regression space as follows:

$$\text{span}[Xu, A_h] = \text{span}[\mathcal{X}_k^h u, A_h] \quad \text{with} \quad \mathcal{X}_k^h = \Pi_{\text{span}[A_h]^{\perp}}^{W_k} X.$$

Since $\text{span}[\mathcal{X}_k^h]$ is orthogonal to $\text{span}[A_h]$,

$$\Pi_{\text{span}[Xu, A_h]}^{W_k} = \Pi_{\text{span}[\mathcal{X}_k^h u, A_h]}^{W_k} = \Pi_{\text{span}[\mathcal{X}_k^h u]}^{W_k} + \Pi_{\text{span}[A_h]}^{W_k}.$$

Consequently, by classical Euclidean derivations, we have

$$\cos^2_{W_k}\left(w_k,\ \mathrm{span}[Xu, A_h]\right)$$

$$= \cos_{W_k}\left(w_k,\ \mathrm{span}[Xu, A_h]\right) \cos_{W_k}\left(w_k,\ \mathrm{span}[Xu, A_h]\right)$$

$$= \left[\frac{\left\|\Pi^{W_k}_{\mathrm{span}[Xu, A_h]} w_k\right\|_{W_k}}{\|w_k\|_{W_k}}\right] \left[\frac{\left\langle w_k, \Pi^{W_k}_{\mathrm{span}[Xu, A_h]} w_k\right\rangle_{W_k}}{\|w_k\|_{W_k}\left\|\Pi^{W_k}_{\mathrm{span}[Xu, A_h]} w_k\right\|_{W_k}}\right]$$

$$= \frac{\left\langle w_k, \left(\Pi^{W_k}_{\mathrm{span}[\mathcal{X}^h_k u]} + \Pi^{W_k}_{\mathrm{span}[A_h]}\right) w_k\right\rangle_{W_k}}{\|w_k\|^2_{W_k}}$$

$$= \frac{\left\langle w_k, \Pi^{W_k}_{\mathrm{span}[\mathcal{X}^h_k u]} w_k\right\rangle_{W_k}}{\|w_k\|^2_{W_k}} + \frac{\left\langle w_k, \Pi^{W_k}_{\mathrm{span}[A_h]} w_k\right\rangle_{W_k}}{\|w_k\|^2_{W_k}}.$$

The goodness of fit measure $\psi_{A_h}(u)$ then writes more explicitly

$$\psi_{A_h}(u) = \sum_{k=1}^{K} \|w_k\|^2_{W_k} \cos^2_{W_k}\left(w_k,\ \mathrm{span}[Xu, A_h]\right)$$

$$= \sum_{k=1}^{K} \left(\left\langle w_k, \Pi^{W_k}_{\mathrm{span}[\mathcal{X}^h_k u]} w_k\right\rangle_{W_k} + \left\langle w_k, \Pi^{W_k}_{\mathrm{span}[A_h]} w_k\right\rangle_{W_k}\right).$$

Now,

$$\left\langle w_k, \Pi^{W_k}_{\mathrm{span}[\mathcal{X}^h_k u]} w_k\right\rangle_{W_k} = w_k^T W_k \Pi^{W_k}_{\mathrm{span}[\mathcal{X}^h_k u]} w_k$$

$$= w_k^T W_k \mathcal{X}^h_k u \left(u^T \mathcal{X}^{h\,T}_k W_k \mathcal{X}^h_k u\right)^{-1} u^T \mathcal{X}^{h\,T}_k W_k w_k$$

$$= \frac{u^T \mathcal{X}^{h\,T}_k W_k w_k w_k^T W_k \mathcal{X}^h_k u}{u^T \mathcal{X}^{h\,T}_k W_k \mathcal{X}^h_k u}.$$

Let,

$$a_k := \mathcal{X}_k^{hT} W_k w_k w_k^T W_k \mathcal{X}_k^h, \quad b_k := \mathcal{X}_k^{hT} W_k \mathcal{X}_k^h \quad \text{and} \quad c_k := \left\langle w_k, \Pi_{\text{span}[A_h]}^{W_k} w_k \right\rangle_{W_k}.$$

Finally,

$$\psi_{A_h}(u) = \sum_{k=1}^{K} \left(\frac{u^T a_k u}{u^T b_k u} + c_k \right) \quad \text{and} \quad \nabla_u \psi_{A_h}(u) = 2 \sum_{k=1}^{K} \frac{u^T b_k u a_k u - u^T a_k u b_k u}{\left(u^T b_k u\right)^2}.$$

7.3.3   The separation sub-criterion

The general form of the separation sub-criterion is $\varphi(u_g^{h+1})$ given in (7.7). We apply this formula to the G explanatory spaces $F_1^{H_1}, \ldots, F_G^{H_G}$ of sizes $H_1, \ldots, H_G$ respectively. We want to separate F_g of F_r for all $r \neq g$. The sub-criterion becomes:

$$
\begin{aligned}
\varphi_{F_{-g}}(u_g^{h+1}) &= 1 - \frac{1}{G-1} \sum_{r \neq g} \left\langle \varpi_{\text{span}[F_g^h, X u_g^{h+1}]}, \varpi_{\text{span}[F_r^{H_r}]} \right\rangle_{\text{Frob}} \\
&= 1 - \frac{1}{G-1} \sum_{r \neq g} \left\langle \frac{\Pi_{\text{span}[F_g^h, X u_g^{h+1}]}}{\sqrt{h+1}}, \frac{\Pi_{\text{span}[F_r^{H_r}]}}{\sqrt{H_r}} \right\rangle_{\text{Frob}} \\
&= 1 - \frac{1}{\sqrt{h+1}(G-1)} \sum_{r \neq g} \sum_{i=1}^{h+1} \sum_{j=1}^{H_r} \frac{\left\langle X u_g^i, X u_r^j \right\rangle_W^2}{\left\| X u_g^i \right\|_W^2 \left\| X u_r^j \right\|_W^2 \sqrt{H_r}}.
\end{aligned}
$$

Let,

$$
\begin{cases}
d_{rgj} = 2 X^T W X u_r^j \left\langle X u_g^{h+1}, X u_r^j \right\rangle_W \left\| X u_g^{h+1} \right\|_W^2 \\[2mm]
e_{rgj} = 2 \left\langle X u_g^{h+1}, X u_r^j \right\rangle_W^2 X^T W X u_g^{h+1} \\[2mm]
f_{rgj} = \left(\left\| X u_g^{h+1} \right\|_W^2 \right)^2 \left\| X u_r^j \right\|_W^2 \sqrt{H_r}
\end{cases}
$$

The gradient of the quotient becomes:

$$\nabla_{u_g^{h+1}} \left(\frac{\langle X u_g^{h+1}, X u_r^j \rangle_W^2}{\left\| X u_g^{h+1} \right\|_W^2 \left\| X u_r^j \right\|_W^2 \sqrt{H_r}} \right) = \frac{d_{rgj} - e_{rgj}}{f_{rgj}}$$

Then, we compute the gradient of φ:

$$\nabla_{u_g^{h+1}} \varphi_{F_{-g}}(u_g^{h+1}) = -\frac{1}{\sqrt{h+1}(G-1)} \sum_{r \neq g} \sum_{j=1}^{H_r} \frac{d_{rgj} - e_{rgj}}{f_{rgj}}.$$

7.4   The rmSCGLR algorithm

Algorithm 4: rmSCGLR algorithm

---

while *not convergence* **do**

    Update mixture's parameters with the EM algorithm

    $\Theta^{(n+1)} = \arg\max_{\Theta} l(\Theta^{(n)}; Y, Z)$

    Update loading vectors with the PING algorithm

    for $g = 1, \ldots, G$ **do**

        for $h = 1, \ldots, H_g$ **do**

$$u_g^{h\,(n+1)} = \max_{\substack{u_g^{h\,T} M^{-1} u_g^h = 1 \\ \Delta_g^{h-1\,T} u_g^h = 0}} s \ln\left(\phi\left(u_g^h\right)\right) + t \ln\left(\varphi_{F_{-g}}\left(u_g^h\right)\right) + (1-s-t) \ln\left(\psi_A\left(u_g^h\right)\right)$$

        end

    end

    $n \leftarrow n + 1$

end

At the end, we can classify the responses according to their posterior probabilities. A response y_k is assigned to cluster g if

$$\alpha_{kg}^{(n_{\max})} > \alpha_{kr}^{(n_{\max})}, \forall r \neq g.$$

7.5 Preliminary simulation study

This simulation is devoted to recovering the true numbers of components, in a context of low correlation between the latent variables spanning the explanatory spaces. We assume unrealistically that the number of groups is known. s is fixed to 0.1, in order to study the behavior of the results when we vary the number of components per group and the weight t of the separation sub-criterion φ.

7.5.1 *Varying the numbers of components*

Three latent variables ξ_1, ξ_3 and ξ_5 are simulated with a pairwise correlation $\rho = 0.5$. Two more latent variables ξ_2 and ξ_4 are independently simulated. The X matrix consists in six blocks $X = [X_1, X_2, X_3, X_4, X_5, X_6]$, where $X_1 \in \mathbb{R}^{N \times 50}$, $X_2 \in \mathbb{R}^{N \times 40}$, $X_3 \in \mathbb{R}^{N \times 30}$, $X_4 \in \mathbb{R}^{N \times 20}$ and $X_5 \in \mathbb{R}^{N \times 10}$ are bundles aligned with ξ_1, ξ_2, ξ_3, ξ_4 and ξ_5, respectively. The X_6 block contains a set of 50 unstructured noise variables. The response matrix Y is partitioned into three groups of responses. The first group is composed of Gaussian responses, the expectations of which are linear combinations of ξ_1 and ξ_4. The second group gathers Poisson responses whose linear predictors are combinations of ξ_2 and ξ_5. The third group is made of binary responses depending only on ξ_3. The matrix Y is generated as:

$$\forall k = 1, \ldots, 20, \quad y_k \sim \mathcal{N}_N \left(\mu = \gamma_1^k \xi_1 + \gamma_2^k \xi_4, \Sigma = I_N \right),$$

$$\forall k = 21, \ldots, 70, \quad y_k \sim \mathcal{P} \left(\lambda = \exp \left[0.25 \gamma_1^k \xi_2 + 0.25 \gamma_2^k \xi_5 \right] \right),$$

$$\forall k = 71, \ldots, 100, \quad y_k \sim \mathcal{B} \left(p = \mathrm{logit}^{-1} \left[\gamma_1^k \xi_3 \right] \right),$$

where for all k, γ_1^k and γ_2^k are uniformly simulated such that $\left|\gamma_1^k\right| \in [2,4]$ and $\left|\gamma_2^k\right| \in [1,2]$.

7.5.2   Results and interpretation

The results of rmSCGLR on this preliminary simulation are given in Table 5. H_g denotes the number of components calculated in group g, and several triplets $H = (H_1, H_2, H_3)$ are tried. For none of these do we observe a clear difference of the RI and ARI across values of t. This was expected, for in this simulation, the explanatory subspaces are only weakly redundant. So, the separation sub-criterion φ proves almost useless here, and has practically no impact on the results.

For $(H_1, H_2, H_3) = (1,1,1)$, the lowest values of RI and ARI are respectively 0.980 and 0.958, without the help of rank $h > 1$ components. The first component of each group perfectly recovers the latent explanatory variable which has the largest effect in the linear predictor of its responses. No component is aligned with the latent variable ξ_4. The latent variable ξ_5 having a correlation of 0.5 with ξ_1 and ξ_3, we find that $\sqrt{\rho^2(\xi_5, .)} \simeq 0.5$ for all values of t.

Taking $(H_1, H_2, H_3) = (2,2,1)$ does not improve the RI and ARI. We notice that the latent variable ξ_5 is not as well recovered as the other latent variables, owing to the small size of the X_5 bundle. However, the BIC is considerably reduced, which illustrates the importance of taking the right number of components to correctly predict the responses.

The last case, where $(H_1, H_2, H_3) = (1,3,1)$ highlights the importance of getting a truly explanatory and strong first component in each group, and of not calculating too

many components in a group. Like in the former cases, the third group is perfectly recovered using the true number of explanatory components $H_3 = 1$. But some confusion arises between the first two groups. Indeed, the extra component f_2^3 of the second group is drawn towards the heaviest bundle X_1. Then, the responses predictable from X_1 tend to be scattered between the first and the second groups instead of being assigned to the first one, which causes a decrease of RI and ARI. Furthermore, owing to the correlation between ξ_1 and ξ_5, the components of the second group cannot be properly aligned with these latent variables. When $t = 0.8$ the weight on the separation criterion φ is heavy enough to recover ξ_1, ξ_2 and ξ_5 in the second group, and ξ_4 in the first group.

To sum up this simulation, we observe that the part played by the first component in recovering the groups is crucial. Indeed, in the first case, the groups are determined by the first component only. In the second case, their prediction is completed by further rank components. However, in the third case, we see that calculating too many components may lead to impede group recovery.

Table 5: Mean values of RI and square correlations between latent variables and supervised components, over a hundred samples, for a weak pairwise correlation value ($\rho = 0.5$) between the latent variables ξ_1, ξ_3 and ξ_5, and for various numbers H_g of components per group.

				group 1		group 2		group 3	
H	t	RI	ARI	$\rho^2(\xi_1,.)$	$\rho^2(\xi_4,.)$	$\rho^2(\xi_2,.)$	$\rho^2(\xi_5,.)$	$\rho^2(\xi_3,.)$	BIC
	0	0.992	0.983	0.971	0.030	0.980	0.309	0.976	33525.56
	0.1	0.986	0.970	0.962	0.037	0.978	0.303	0.969	33580.84
	0.2	0.985	0.967	0.965	0.033	0.976	0.317	0.972	33435.54
1	0.3	0.987	0.972	0.968	0.037	0.978	0.314	0.973	33577.05
1	0.4	0.991	0.980	0.971	0.032	0.980	0.297	0.975	33435.80
1	0.5	0.980	0.958	0.960	0.036	0.974	0.298	0.961	33612.22
	0.6	0.992	0.983	0.960	0.043	0.979	0.295	0.974	33631.08
	0.7	0.994	0.987	0.954	0.046	0.983	0.295	0.975	33837.14
	0.8	0.992	0.983	0.944	0.044	0.979	0.298	0.964	34304.05
	0	0.984	0.966	0.968	0.921	0.975	0.816	0.966	29945.48
	0.1	0.983	0.964	0.971	0.938	0.977	0.809	0.971	29878,00
	0.2	0.989	0.977	0.974	0.951	0.979	0.835	0.975	29838.60
2	0.3	0.994	0.988	0.974	0.952	0.981	0.865	0.978	29783.77
2	0.4	0.993	0.984	0.968	0.946	0.981	0.876	0.975	29936.19
1	0.5	0.991	0.981	0.957	0.934	0.981	0.856	0.972	30150.95
	0.6	0.984	0.966	0.944	0.928	0.976	0.844	0.960	30348.66
	0.7	0.997	0.993	0.932	0.946	0.983	0.864	0.976	30733.32
	0.8	0.983	0.965	0.916	0.925	0.973	0.827	0.971	31131.42
	0	0.878	0.750	0.871	0.264	0.945	0.514	0.978	30483.70
	0.1	0.874	0.742	0.856	0.214	0.956	0.506	0.965	30245.37
	0.2	0.859	0.712	0.858	0.230	0.970	0.555	0.932	30020.63
1	0.3	0.871	0.776	0.853	0.242	0.969	0.545	0.946	31090.38
3	0.4	0.868	0.724	0.839	0.370	0.961	0.580	0.980	30052.19
1	0.5	0.876	0.748	0.804	0.308	0.977	0.585	0.977	30322.50
	0.6	0.891	0.774	0.806	0.320	0.976	0.656	0.977	30815.20
	0.7	0.882	0.759	0.732	0.353	0.975	0.657	0.974	30572.94
	0.8	0.790	0.592	0.877	0.772	0.956	0.614	0.963	33790.50

Figure 4 shows the correlation scatterplots in the component planes $(1, 2)$ for the first two groups. As for the first simulation, the components are almost perfectly aligned with the explanatory bundles. Because of the weak correlation between ξ_1, ξ_3 and ξ_5, the three bundles X_1, X_3 and X_5 are visible on the same component for each of the two groups.
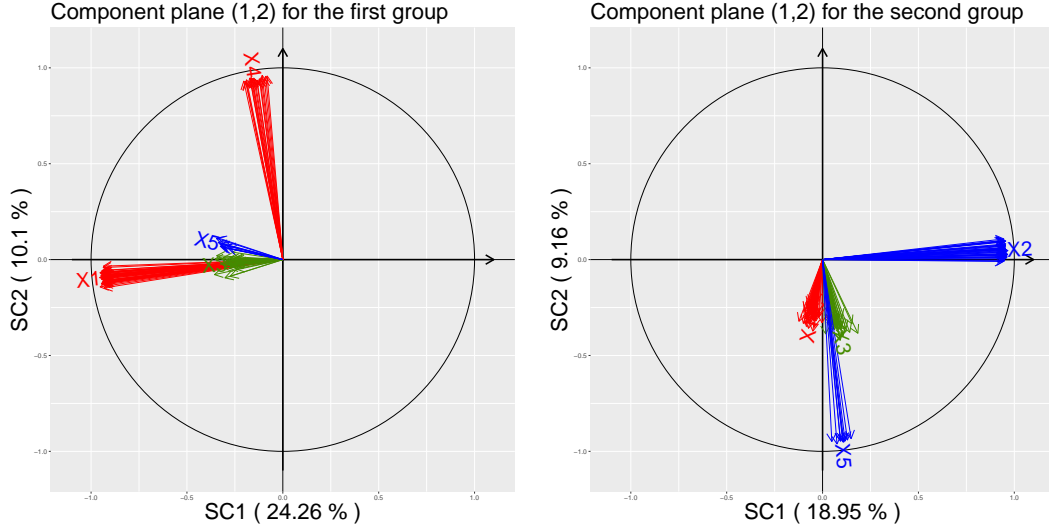
Figure 4: Correlation scatterplots of plane (1,2) for the first two groups of the second simulation, with $(H_1, H_2, H_3) = (2, 2, 1)$, obtained by rmSCGLR. The red arrows represent the bundles X_1 and X_4, explanatory of the first group. The blue ones represent the bundles X_2 and X_5, explanatory of the second group. The green bundle X_3 is explanatory of the third group. The percentage of inertia captured by each component is given in parentheses.

7.6   Groups of taxa

Table 6: Here is the list of the taxa used in this study (the family classification follows Angiosperm Phylogeny Group III).

Group	Family	Genus	Species
1	Huaceae	Afrostyrax	lepidophyllus
1	Fabaceae	Afzelia	spp.
1	Fabaceae	Albizia	ferruginea
1	Fabaceae	Albizia	spp.
1	Gentianaceae	Anthocleista	spp.
1	Fabaceae	Anthonotha	spp.
1	Phyllanthaceae	Antidesma	spp.
1	Fabaceae	Aphanocalyx	spp.
1	Fabaceae	Aubrevillea	kerstingii
1	Zygophyllaceae	Balanites	wilsoniana
1	Passifloraceae	Barteria	spp.
1	Lauraceae	Beilschmiedia	spp.
1	Malvaceae	Bombax	spp.
1	Malvaceae	Ceiba	pentandra
1	Cannabaceae	Celtis	spp.
1	Sapotaceae	Chrysophyllum	spp.
1	Annonaceae	Cleistopholis	spp.
1	Malvaceae	Cola	spp.
1	Boraginaceae	Cordia	spp.
1	Fabaceae	Detarium	macrocarpum

Continued on next page

Continued from previous page

Group	Family	Genus	Species
1	Fabaceae	Dialium	spp.
1	Euphorbiaceae	Discoglypremna	caloneura
1	Malvaceae	Duboscia	spp.
1	Arecaceae	Elaeis	guineensis
1	Malvaceae	Eribroma	oblongum
1	Hypericaceae	Harungana	madagascariensis
1	Annonaceae	Hexalobus	spp.
1	Ulmaceae	Holoptelea	grandis
1	Meliaceae	Khaya	spp.
1	Irvingiaceae	Klainedoxa	spp.
1	Meliaceae	Lovoa	trichilioides
1	Malvaceae	Mansonia	altissima
1	Urticaceae	Myrianthus	arboreus
1	Apocynaceae	Picralima	nitida
1	Sapotaceae	Pouteria	spp.
1	Malvaceae	Pterygota	spp.
1	Euphorbiaceae	Ricinodendron	heudelotii
1	Malvaceae	Sterculia	spp.
1	Olacaceae	Strombosiopsis	spp.
1	Myrtaceae	Syzygium	spp.
1	Combretaceae	Terminalia	superba
1	Fabaceae	Tetrapleura	tetraptera

Continued from previous page

Group	Family	Genus	Species
1	Malvaceae	Triplochiton	scleroxylon
1	Lamiaceae	Vitex	spp.
2	Clusiaceae	Allanblackia	spp.
2	Apocynaceae	Alstonia	spp.
2	Fabaceae	Angylocalyx	spp.
2	Anisophylleaceae	Anisophyllea	spp.
2	Moraceae	Antiaris	toxicaria
2	Fabaceae	Aubrevillea	platycarpa
2	Burseraceae	Aucoumea	klaineana
2	Sapotaceae	Autranella	congolensis
2	Sapotaceae	Baillonella	toxisperma
2	Fabaceae	Bikinia	spp.
2	Sapindaceae	Blighia	spp.
2	Sapotaceae	Breviea	sericea
2	Burseraceae	Canarium	schweinfurthii
2	Myristicaceae	Coelocaryon	spp.
2	Rubiaceae	Corynanthe	pachyceras
2	Fabaceae	Cylicodiscus	gabunensis
2	Burseraceae	Dacryodes	spp.
2	Fabaceae	Daniellia	spp.
2	Achariaceae	Dasylepis	seretii

Continued from previous page

Group	Family	Genus	Species
2	Malvaceae	Desplatsia	spp.
2	Ebenaceae	Diospyros	crassiflora
2	Fabaceae	Distemonanthus	benthamianus
2	Meliaceae	Entandrophragma	angolense
2	Meliaceae	Entandrophragma	candollei
2	Meliaceae	Entandrophragma	cylindricum
2	Meliaceae	Entandrophragma	utile
2	Vochysiaceae	Erismadelphus	exsul
2	Bignoniaceae	Fernandoa	adolfi
2	Moraceae	Ficus	spp.
2	Fabaceae	Gilbertiodendron	spp.
2	Euphorbiaceae	Gymnanthes	inopinata
2	Irvingiaceae	Irvingia	grandifolia
2	Lepidobotryaceae	Lepidobotrys	staudtii
2	Euphorbiaceae	Macaranga	spp.
2	Rhamnaceae	Maesopsis	eminii
2	Sapotaceae	Manilkara	spp.
2	Phyllanthaceae	Margaritaria	discoidea
2	Moraceae	Milicia	excelsa
2	Moraceae	Morus	mesozygia
2	Urticaceae	Musanga	cecropioides
2	Rubiaceae	Nauclea	spp.

Continued from previous page

Group	Family	Genus	Species
2	Malvaceae	Nesogordonia	spp.
2	Fabaceae	Newtonia	spp.
2	Picrodendraceae	Oldfieldia	africana
2	Salicaceae	Oncoba	spp.
2	Chrysobalanaceae	Parinari	spp.
2	Fabaceae	Pentaclethra	eetveldeana
2	Euphorbiaceae	Plagiostyles	africana
2	Combretaceae	Pteleopsis	hylodendron
2	Fabaceae	Pterocarpus	spp.
2	Violaceae	Rinorea	spp.
2	Burseraceae	Santiria	spp.
2	Oleaceae	Schrebera	arborea
2	Myristicaceae	Scyphocephalium	mannii
2	Anacardiaceae	Sorindeia	spp.
2	Clusiaceae	Symphonia	globulifera
2	Sapotaceae	Synsepalum	spp.
2	Ochnaceae	Testulea	gabonensis
2	Fabaceae	Tetraberlinia	bifoliolata
2	Euphorbiaceae	Tetrorchidium	didymostemon
2	Sapotaceae	Tieghemella	africana
2	Moraceae	Treculia	spp.
2	Meliaceae	Trichilia	spp.

Continued from previous page

Group	Family	Genus	Species
2	Anacardiaceae	Trichoscypha	spp.
2	Sapotaceae	Tridesmostemon	omphalocarpoides
2	Moraceae	Trilepisium	madagascariense
2	Dipterocarpaceae	Trillesanthus	excelsus
3	Fabaceae	Amphimas	spp.
3	Annonaceae	Annickia	spp.
3	Annonaceae	Anonidium	mannii
3	Rhizophoraceae	Anopyxis	klaineana
3	Euphorbiaceae	Anthostema	aubryanum
3	Anacardiaceae	Antrocaryon	spp.
3	Fabaceae	Berlinia	spp.
3	Fabaceae	Bobgunnia	fistuloides
3	Fabaceae	Brachystegia	spp.
3	Rubiaceae	Brenania	brieyi
3	Phyllanthaceae	Bridelia	spp.
3	Fabaceae	Calpocalyx	spp.
3	Meliaceae	Carapa	spp.
3	Sapotaceae	Chrysophyllum	lacourtianum
3	Fabaceae	Copaifera	spp.
3	Olacaceae	Coula	edulis
3	Euphorbiaceae	Croton	spp.

Continued from previous page

Group	Family	Genus	Species
3	Fabaceae	Cryptosepalum	spp.
3	Olacaceae	Diogoa	zenkeri
3	Ebenaceae	Diospyros	spp.
3	Asparagaceae	Dracaena	spp.
3	Putranjivaceae	Drypetes	spp.
3	Annonaceae	Duguetia	spp.
3	Fabaceae	Erythrophleum	spp.
3	Erythroxylaceae	Erythroxylum	mannii
3	Fabaceae	Eurypetalum	spp.
3	Fabaceae	Fillaeopsis	discophora
3	Apocynaceae	Funtumia	spp.
3	Clusiaceae	Garcinia	spp.
3	Fabaceae	Gilbertiodendron	dewevrei
3	Malvaceae	Grewia	spp.
3	Salicaceae	Homalium	spp.
3	Fabaceae	Hylodendron	gabunense
3	Fabaceae	Hymenostegia	spp.
3	Irvingiaceae	Irvingia	spp.
3	Fabaceae	Julbernardia	spp.
3	Phyllanthaceae	Keayodendron	bridelioides
3	Meliaceae	Leplaea	spp.
3	Sapotaceae	Letestua	durissima

Continued from previous page

Group	Family	Genus	Species
3	Ochnaceae	Lophira	alata
3	Calophyllaceae	Mammea	africana
3	Chrysobalanaceae	Maranthes	spp.
3	Bignoniaceae	Markhamia	spp.
3	Fabaceae	Millettia	spp.
3	Rubiaceae	Morinda	lucida
3	Fabaceae	Neochevalierodendron	stephanii
3	Ochnaceae	Ochna	spp.
3	Ixonanthaceae	Ochthocosmus	spp.
3	Sapotaceae	Omphalocarpum	spp.
3	Olacaceae	Ongokea	gore
3	Fabaceae	Pachyelasma	tessmannii
3	Pandaceae	Panda	oleosa
3	Rubiaceae	Pausinystalia	spp.
3	Fabaceae	Pentaclethra	macrophylla
3	Fabaceae	Pericopsis	elata
3	Lecythidaceae	Petersianthus	macrocarpus
3	Fabaceae	Piptadeniastrum	africanum
3	Annonaceae	Polyalthia	suaveolens
3	Fabaceae	Prioria	spp.
3	Anacardiaceae	Pseudospondias	spp.
3	Myristicaceae	Pycnanthus	angolensis

Continued from previous page

Group	Family	Genus	Species
3	Simaroubaceae	Quassia	spp.
3	Apocynaceae	Rauvolfia	spp.
3	Rubiaceae	Rothmannia	spp.
3	Euphorbiaceae	Sapium	spp.
3	Fabaceae	Scorodophloeus	zenkeri
3	Achariaceae	Scottellia	spp.
3	Lecythidaceae	Scytopetalum	klaineanum
3	Bignoniaceae	Spathodea	campanulata
3	Fabaceae	Stachyothyrsus	staudtii
3	Myristicaceae	Staudtia	kamerunensis
3	Fabaceae	Stemonocoleus	micranthus
3	Combretaceae	Strephonema	spp.
3	Olacaceae	Strombosia	spp.
3	Apocynaceae	Tabernaemontana	spp.
3	Fabaceae	Tessmannia	spp.
3	Fabaceae	Tetraberlinia	polyphylla
3	Phyllanthaceae	Uapaca	spp.
3	Annonaceae	Xylopia	aethiopica
3	Annonaceae	Xylopia	hypolampra
3	Annonaceae	Xylopia	quintasii
3	Rutaceae	Zanthoxylum	spp.

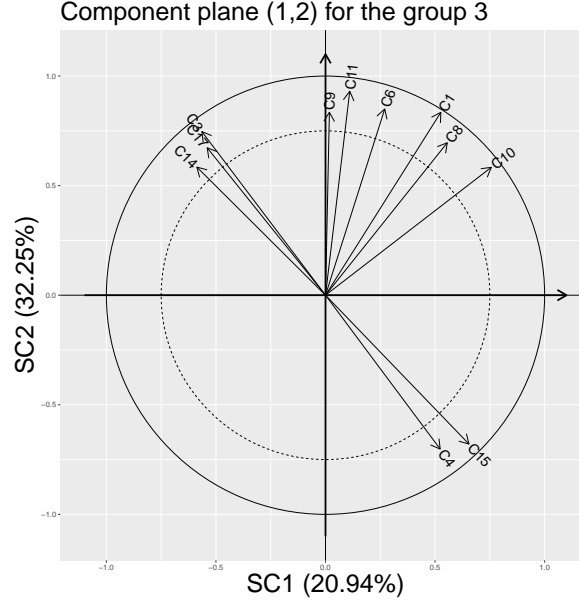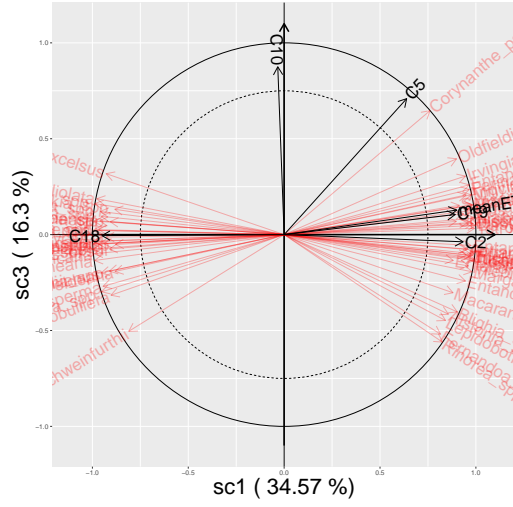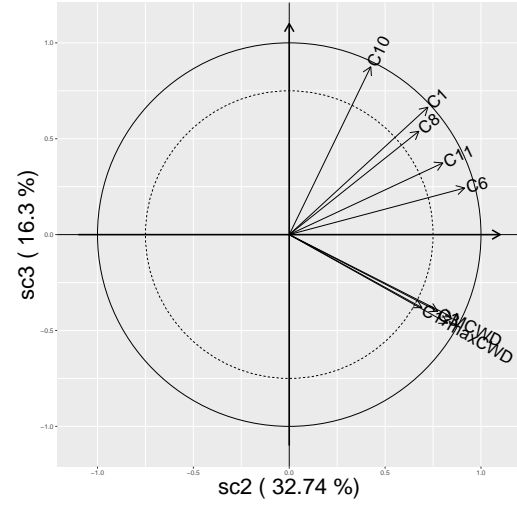## 7.7 Correlation plot of the third group

Figure 5: Component plane (1,2) of group 3 output by rmSCGLR on the *CoForTaxa* dataset, with optimal hyper-parameter $(s, l, t) = (0.1, 1, 0.5)$. The plot displays only variables having cosine greater than 0.75. The percentage of inertia captured by each component is given in parentheses.
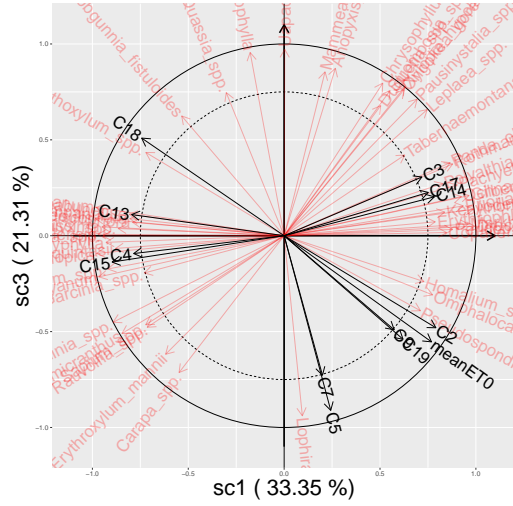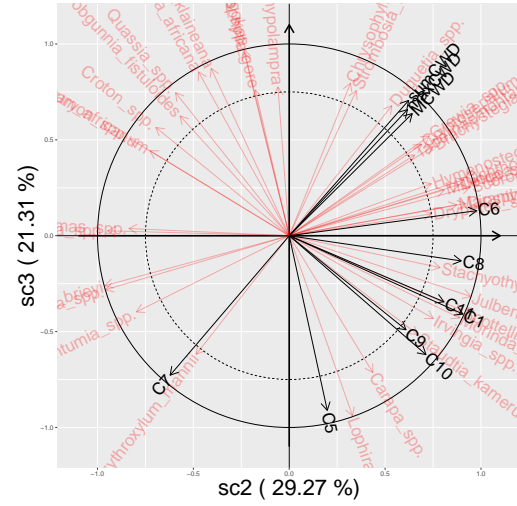
7.8   Higher rank correlation plot

(a) Component plane $(1, 3)$ of the group 2

(b) Component plane $(2, 3)$ of the group 2

(c) Component plane $(1, 3)$ of the group 3

(d) Component plane $(2, 3)$ of the group 3

Figure 6: Correlation scatterplots of planes (1,3) and (2,3) with linear predictors obtained by applying SCGLR to the second and third groups separately. The blacks arrows represent the covariates. The red ones represent the linear predictors. The plot displays only variables having a cosine greater than 0.75 with the plane. The percentage of inertia captured by each component is given in parentheses.