



**HAL**  
open science

# Generalized linear model based on latent factors and supervised components

Julien Gibaud, Xavier Bry, Catherine Trottier

► **To cite this version:**

Julien Gibaud, Xavier Bry, Catherine Trottier. Generalized linear model based on latent factors and supervised components. *Computational Statistics*, 2024, 10.1007/s00180-024-01544-8. hal-04263074v3

**HAL Id: hal-04263074**

**<https://hal.science/hal-04263074v3>**

Submitted on 11 Jul 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Generalized linear model based on latent factors and supervised components

Julien Gibaud<sup>1\*†</sup>, Xavier Bry<sup>1†</sup> and Catherine Trottier<sup>1,2†</sup>

<sup>1</sup>IMAG, Université de Montpellier, CNRS, Montpellier, France.

<sup>2</sup>AMIS, Université Paul-Valéry Montpellier 3, F34000, Montpellier,  
France.

\*Corresponding author(s). E-mail(s): [julien.gibaud@umontpellier.fr](mailto:julien.gibaud@umontpellier.fr);

†These authors contributed equally to this work.

## Abstract

In a context of component-based multivariate modeling we propose to model the residual dependence of the responses. Each response of a response vector is assumed to depend, through a Generalized Linear Model, on a set of explanatory variables. The vast majority of explanatory variables are partitioned into conceptually homogeneous variable groups, viewed as explanatory themes. Variables in themes are supposed many and some of them are highly correlated or even collinear. Thus, generalized linear regression demands dimension reduction and regularization with respect to each theme. Besides them, we consider a small set of “additional” covariates not conceptually linked to the themes, and demanding no regularization. Supervised Component Generalized Linear Regression proposed to both regularize and reduce the dimension of the explanatory space by searching each theme for an appropriate number of orthogonal components, which both contribute to predict the responses and capture relevant structural information in themes. In this paper, we introduce random latent variables (a.k.a. factors) so as to model the covariance matrix of the linear predictors of the responses conditional on the components. To estimate the model, we present an algorithm combining supervised component-based model estimation with factor model estimation. This methodology is tested on simulated data and then applied to an agricultural ecology dataset.


**Keywords:** EM algorithm, factor model, generalized linear latent variable model, multivariate generalized linear model, supervised components

# 1 Introduction

In multivariate modeling, accounting for mutual dependencies between the responses is a rich ongoing research field in both theoretical and applied statistics. For instance, see [Bartholomew et al \(2011, Chapter 3\)](#) for the theoretical aspects; [Meyer \(2009\)](#), [Pollock et al \(2014\)](#), [Hui \(2017\)](#) for application to ecology and [Watkins \(2018\)](#) for a review in the quantitative psychology field. As recalled by [Ovaskainen et al \(2017\)](#) in an ecological context, the species co-occurrences are not only explained by the environmental variables but also partly by biological interactions between the species. As a result, species abundances can hardly be assumed independent conditional on the environmental variables. The distributions of the responses (here the abundances of species) will be assumed to belong in the exponential family. We shall model their residual dependence by introducing random latent variables in the linear predictors. These random latent variables are henceforth referred to as “factors”.

Generalized Linear Latent Variable Models (GLLVMs) have been proposed by [Skrovdal and Rabe-Hesketh \(2004\)](#) to combine Generalized Linear Models (GLM; [McCullagh and Nelder, 1989](#)) with factors. Unfortunately, when factors are involved, the log-likelihood includes an integral which has no closed form in general. In the particular case of Gaussian responses, the GLLVM is a classic factor model. Several methods for estimating GLLVMs suffer from a high computation time; see for instance those using the adaptive quadrature ([Rabe-Hesketh et al, 2002](#)), the Expectation Maximization algorithm (EM; [Dempster et al, 1977](#)) in conjunction with Monte Carlo integration ([Hui et al, 2015](#)) or those using Bayesian Markov Chain Monte Carlo (MCMC) ([Hui, 2016](#); [Tikhonov et al, 2020](#)). The methods using a variational approximation ([Hui et al, 2017](#)), a Laplace approximation ([Niku et al, 2017, 2019a](#)) or an extended variational approximation ([Korhonen et al, 2023](#)) reduce the computation time by using a closed form approximation of the log-likelihood. From an other perspective, a fitting approach proposed by [Saidane et al \(2013\)](#) assumes that maximization can be performed through the EM algorithm after linearizing the model and assuming the linearized model is approximately Gaussian. Originally introduced by [Schall \(1991\)](#), the linearization method gives, through empirical studies, an alternative to estimate parameters in a context of intractable likelihood whether in a Generalized Linear Mixed Model ([Chauvet et al, 2019](#)) or in a factor model context ([Saidane et al, 2013](#)).

Modeling the responses also requires taking into account a large set of possibly highly correlated explanatory covariates. The correlation between the covariates entails critical identification problems, so that the GLLVM demands regularization. This, together with an interpretable dimension reduction can be carried out by calculating a small number of explanatory deterministic latent variables. These deterministic latent variables, defined as linear combinations of the explanatory variables, are referred to as “components”. Different approaches propose to bridge the multivariate GLM estimation with dimension reduction of the explanatory space. Methods as Reduced Rank Vector Generalized Linear Model (RRVGLM; [Yee and Hastie, 2003](#)) or concurrent ordination ([van der Veen et al, 2023](#)) reduce the number of parameters to estimate by assuming that the set of explanatory variables can be replaced by a small number of their linear combinations. Alternatively, orthogonal components are constructed by

the Iteratively Reweighted Partial Least Squares (IRPLS; Marx, 1996) and the Supervised Component-based Generalized Linear Regression (SCGLR; Bry et al, 2013) in order to capture relevant information on the covariates for predicting the responses. SCGLR allows both to find interpretable explanatory components, and to produce regularized linear predictors in the high-dimensional framework, i.e. when the covariates outnumber the statistical units. To achieve that, 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 of the space spanned by the explanatory variables. Later, Bry et al (2020b) developed an extension of SCGLR, called THEME-SCGLR, with the aim to search for components in a thematic partitioning of the explanatory variables into groups, hence referred to as “themes”. Variables in a theme must have conceptual kinship, so that the components combining them linearly can be interpreted with respect to their common concept. For instance, precipitations and solar radiation measures can be gathered in a “climate” theme, soil measures in a “geology” theme, etc. Within each theme, the components are required to extract the information that is useful to predict the responses when associated with the components of the other themes. An **SCGLR**  package is freely available at <https://github.com/SCnext/SCGLR>.

All the extensions of SCGLR (Bry et al, 2013; Chauvet et al, 2019; Bry et al, 2020a,b; Gibaud et al, 2024) developed so far, have been assuming that the responses are independent conditional on the explanatory covariates. We now propose to overcome this limitation by allowing the responses to have some conditional dependence, which we model by introducing common factors into their linear predictors. Besides, we also take into account a thematic partitioning of the explanatory variables. To put it shortly, we refine THEME-SCGLR in order to model the conditional dependence between the responses drawn from a distribution belonging to the exponential family, as GLLVM does.

In a species-rich ecosystem context, to better understand the communities-specific characteristics of species, several works partition the responses into groups. Some methods cluster the responses with respect only to the values they take (Swaine and Whitmore, 1988). Others cluster responses on the basis of their regression coefficients on covariates (Dunstan et al, 2013; Mortier et al, 2015). Others still, base the clustering on the fact that the responses depend more or less on the same subset of covariates (Gibaud et al, 2024). In this paper, we propose to estimate the residual covariance matrix of the responses conditional on the explanatory components, and then, to use this matrix to partition the responses into groups of highly correlated responses (in square value). Indeed, high correlations, whatever their sign, could hint at interactions between species. More precisely, to group species, we perform clustering on a dissimilarity matrix built from the estimated residual correlation matrix of the linear predictors. Identifying a group of correlated species is equivalent to identifying a square block of high absolute values in this correlation matrix, once the rows and columns are suitably reordered.

The paper is organized as follows. In Section 2, we recall the principle of THEME-SCGLR. Section 3 presents our extension to the factor models. Section 4 details several simulation studies that illustrate the interest and the performances of the proposed

algorithm. Section 5 presents the results it yields on an agricultural ecology dataset. Finally, Section 6 provides a conclusion and a discussion.

## 2 A reminder of THEME-SCGLR

### 2.1 Preliminary notations

The sequel contains mathematical developments which use notations listed hereafter:

- Consider generally  $N$  statistical units, unit  $n$  being endowed with weight  $w_n$ , and let  $\mathbf{W} = \text{diag}(w_n)_{n=1,\dots,N}$ . Let  $\mathbf{a}$  and  $\mathbf{b}$  be vectors of  $\mathbb{R}^N$ , endowed with metric  $\mathbf{W}$ . The Euclidean scalar product between  $\mathbf{a}$  and  $\mathbf{b}$  with respect to metric  $\mathbf{W}$  is given by  $\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{W}} = \mathbf{a}^T \mathbf{W} \mathbf{b}$ .
- Likewise,  $\cos_{\mathbf{W}}(\mathbf{a}, \mathbf{b}) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle_{\mathbf{W}}}{\|\mathbf{a}\|_{\mathbf{W}} \|\mathbf{b}\|_{\mathbf{W}}}$  denotes the cosine of the angle between  $\mathbf{a}$  and  $\mathbf{b}$  with respect to metric  $\mathbf{W}$ . If  $\mathbf{a}$  and  $\mathbf{b}$  are centred and  $\mathbf{W} = \frac{1}{N} \mathbf{I}_N$ , the cosine is the linear correlation coefficient, denoted  $\rho$ .
- $\mathbf{Q} = [\mathbf{q}_1, \dots, \mathbf{q}_C] \in \mathbb{R}^{N \times C}$  and  $\mathbf{R} = [\mathbf{r}_1, \dots, \mathbf{r}_D] \in \mathbb{R}^{N \times D}$  being matrices, the concatenation of  $\mathbf{Q}$  and  $\mathbf{R}$  writes  $[\mathbf{Q}, \mathbf{R}]$ . The space spanned by the column-vectors of  $\mathbf{Q}$  is denoted  $\text{span}[\mathbf{Q}]$ .
- The  $\mathbf{W}$ -orthogonal projector onto  $\text{span}[\mathbf{Q}]$  is given by  $\mathbf{\Pi}_{\text{span}[\mathbf{Q}]}^{\mathbf{W}} = \mathbf{Q}(\mathbf{Q}^T \mathbf{W} \mathbf{Q})^{-1} \mathbf{Q}^T \mathbf{W}$ . The cosine of the angle between a vector  $\mathbf{b} \in \mathbb{R}^N$  and  $\text{span}[\mathbf{Q}]$  with respect to metric  $\mathbf{W}$  is given by  $\cos_{\mathbf{W}}(\mathbf{b}, \text{span}[\mathbf{Q}]) = \cos_{\mathbf{W}}(\mathbf{b}, \mathbf{\Pi}_{\text{span}[\mathbf{Q}]}^{\mathbf{W}} \mathbf{b})$ .

### 2.2 The original SCGLR method

In the framework of a multivariate GLM, consider  $K$  response vectors measured on  $N$  statistical units, encoded in a response matrix  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_K] \in \mathbb{R}^{N \times K}$ , to be predicted through explanatory variables partitioned in two groups. The first one  $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_Q] \in \mathbb{R}^{N \times Q}$  is a group of covariates that are few and not or at the most weakly redundant. These variables are assumed to be interesting per se, and their marginal effects have to be taken into account explicitly in the model. The second group  $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_P] \in \mathbb{R}^{N \times P}$  gathers numerous covariates, considered as proxies to latent dimensions, which must be found and interpreted. Some of them being highly correlated or even collinear, the model demands dimension reduction and regularization with respect to  $\mathbf{X}$ , and not to  $\mathbf{A}$ . To achieve this, SCGLR searches for explanatory components in  $\mathbf{X}$  jointly supervised by the responses. A component  $\mathbf{f} \in \mathbb{R}^N$  writes  $\mathbf{f} = \mathbf{X} \mathbf{u}$ , where  $\mathbf{u} \in \mathbb{R}^P$  is a vector of component coefficients. For a single component model, the linear predictor associated with response  $\mathbf{y}_k$  is given by

$$\boldsymbol{\eta}_k = (\mathbf{X} \mathbf{u}) \gamma_k + \mathbf{A} \boldsymbol{\delta}_k,$$

where  $\gamma_k$  and  $\boldsymbol{\delta}_k$  are regression parameters. Component  $\mathbf{f}$  is common to all the responses and, for identification, we impose  $\mathbf{u}^T \mathbf{M} \mathbf{u} = 1$ . As a general rule, we must have  $\mathbf{M} = \mathbf{M}_{\text{PCA}}^{-1}$ , where  $\mathbf{M}_{\text{PCA}}$  is the metric suitable for  $\mathbf{X}$ 's Principal Component Analysis (PCA). In this paper, assuming that  $\mathbf{X}$  consists of  $P$  standardized numeric variables, we have  $\mathbf{M} = \mathbf{I}_P$ . It is assumed in the original SCGLR method, that the

responses are independent conditional on the explanatory variables, and consequently on  $\mathbf{f}$ .

### 2.3 The original SCGLR specific criterion

For the parameter estimation, SCGLR takes advantage of the GLM background. Here, we make use of the Fisher Scoring Algorithm (FSA). Let  $h_k$  denote the canonical link function associated with the response  $\mathbf{y}_k$ ,  $h'_k$  its first derivative and  $\mu_{nk}$  the mean parameter for statistical unit  $n$ . The linear predictor is connected to the mean parameter through the link function,  $\eta_{nk} = h_k(\mu_{nk})$ . In the wake of [McCullagh and Nelder \(1989\)](#), the adjusted dependent variable  $w_{nk}$  associated with  $y_{nk}$  is then calculated as the first order development of  $h_k$  at point  $\mu_{nk}$

$$\begin{aligned} w_{nk} &= h_k(\mu_{nk}) + (y_{nk} - \mu_{nk}) h'_k(\mu_{nk}) \\ &= \eta_{nk} + \zeta_{nk}, \end{aligned}$$

where  $\zeta_{nk} = (y_{nk} - \mu_{nk}) h'_k(\mu_{nk})$ . In the spirit of [Nelder and Wedderburn \(1972\)](#), this development leads to the conditional linearized model expressed column-wise

$$\mathbf{w}_k = (\mathbf{X}\mathbf{u})\gamma_k + \mathbf{A}\boldsymbol{\delta}_k + \boldsymbol{\zeta}_k,$$

with  $\mathbb{E}[\boldsymbol{\zeta}_k] = 0$  and  $\mathbb{V}[\boldsymbol{\zeta}_k] =: \mathbf{W}_k^{-1}$ . An explicit expression of  $\mathbf{W}_k^{-1}$  is given in [Appendix D](#).

Due to the product  $\mathbf{u}\gamma_k$ , this linearized model derived from the FSA is not linear and must be estimated through an alternated weighted least squares process, estimating in turn  $\{\gamma_k, \boldsymbol{\delta}_k\}$  and  $\mathbf{u}$ .

Let  $\Pi_{\text{span}[\mathbf{X}\mathbf{u}, \mathbf{A}]}^{\mathbf{W}_k}$  be the projection on  $\text{span}[\mathbf{X}\mathbf{u}, \mathbf{A}]$  with respect to  $\mathbf{W}_k$ . As suggested by [Bry et al \(2013\)](#), the vector of component coefficients  $\mathbf{u}$  may be viewed as the solution of the optimization program  $\max_{\mathbf{u}^T \mathbf{u} = 1} \psi_{\mathbf{A}}(\mathbf{u})$ , where

$$\psi_{\mathbf{A}}(\mathbf{u}) := \sum_{k=1}^K \|\mathbf{w}_k\|_{\mathbf{W}_k}^2 \cos_{\mathbf{W}_k}^2 \left( \mathbf{w}_k, \Pi_{\text{span}[\mathbf{X}\mathbf{u}, \mathbf{A}]}^{\mathbf{W}_k} \mathbf{w}_k \right).$$

The sub-criterion  $\psi_{\mathbf{A}}$  is merely a Goodness-of-Fit (GoF) measure, and maximizing it does not lead to strong, regularizing and and interpretable components. The GoF measure must therefore be aptly combined with an other sub-criterion to achieve both meaningful and predictive dimension reduction, together with regularization.

[Bry and Verron \(2015\)](#) proposed such a sub-criterion, named Structural Relevance (SR) to measure the ability of a component to capture information in a set of variables containing structures such as variable-bundles. Informally, a bundle is a set of variables correlated “enough” to be viewed as proxies to a common latent dimension. The associated SR measure  $\phi$  is defined as the following generalized average of quadratic

forms of  $\mathbf{u}$

$$\phi(\mathbf{u}) := \left( \frac{1}{P} \sum_{p=1}^P \langle \mathbf{X}\mathbf{u}, \mathbf{x}_p \rangle_{\mathbf{W}}^{2l} \right)^{1/l}, \quad (1)$$

where  $\mathbf{W} = \frac{1}{N} \mathbf{I}_N$  is the weight matrix. Components will align with a more or less thin bundle depending on whether  $l \geq 1$  is greater or smaller respectively. The main purpose of introducing SR is to focus on more interpretable directions.

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

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

When  $s = 0$ , the criterion identifies with the GoF, while at the other end, taking  $s = 1$  makes it identify with the SR. Increasing  $s$  intensifies both the focus of components on “strong” dimensions, and the regularization, at the cost of some GoF. The explicit expression of the criterion is given in Appendix A. Moreover, Appendix C gives a proof of how SCGLR regularizes and shrinks the coefficients.

## 2.4 THEME-SCGLR

Bry et al (2020b) refer to the “thematic model” as the conceptual model stating that variables in  $\mathbf{Y}$  are dependent on  $R$  themes  $\mathbf{X}_1, \dots, \mathbf{X}_R$ , and that structurally relevant dimensions should be explicitly identified in the  $\mathbf{X}_r$ ’s. Each theme is a set of covariates, related to the same concept, where some may be highly correlated or even collinear. Besides them, we consider a small set of “additional” covariates  $\mathbf{A}$  not conceptually linked to the themes, and demanding no dimension reduction. For a single component in each theme, the linear predictor associated with response  $\mathbf{y}_k$  is then given by

$$\boldsymbol{\eta}_k = (\mathbf{X}_1 \mathbf{u}_1) \gamma_{k1} + \dots + (\mathbf{X}_R \mathbf{u}_R) \gamma_{kR} + \mathbf{A} \boldsymbol{\delta}_k.$$

To achieve theme-specific regularization, the SCGLR criterion has to be adapted. Denoting  $\mathbf{f}_r = \mathbf{X}_r \mathbf{u}_r$  the (first) component of theme  $\mathbf{X}_r$ , we have  $\Pi_{\text{span}[\mathbf{f}_1, \dots, \mathbf{f}_R, \mathbf{A}]}^{\mathbf{W}_k} = \Pi_{\text{span}[\mathbf{f}_r, \mathbf{A}_r]}^{\mathbf{W}_k}$  where  $\mathbf{A}_r = [\mathbf{f}_1, \dots, \mathbf{f}_{r-1}, \mathbf{f}_{r+1}, \dots, \mathbf{f}_R, \mathbf{A}]$ . For each theme, the GoF measure thus becomes

$$\psi_{\mathbf{A}_r}(\mathbf{u}_r) := \sum_{k=1}^K \|\mathbf{w}_k\|_{\mathbf{W}_k}^2 \cos_{\mathbf{W}_k}^2 \left( \mathbf{w}_k, \Pi_{\text{span}[\mathbf{X}_r \mathbf{u}_r, \mathbf{A}_r]}^{\mathbf{W}_k} \mathbf{w}_k \right).$$

The SR measure remains the same  $\phi(\mathbf{u}_r)$  as given by Equation (1). Finally, the optimization program can be solved by iteratively maximizing in turn the trade-off criterion on every  $\mathbf{u}_r$

$$\forall r, \quad \max_{\mathbf{u}_r^T \mathbf{u}_r = 1} s \ln(\phi(\mathbf{u}_r)) + (1-s) \ln(\psi_{\mathbf{A}_r}(\mathbf{u}_r)). \quad (2)$$

This combined criterion is rather general. Indeed, the GoF measure adapts any situation where a likelihood function is available for the model taking the components and  $\mathbf{A}$  as covariates. Generally, this likelihood involves a vector of parameters  $\Theta$ . The maximization is carried out alternating two steps: (i) Given  $\Theta$ , maximize the criterion with respect to each  $\mathbf{u}_r$  using a dedicated algorithm: PING (for Projected Iterated Normed Gradient) recalled in Appendix B, designed to maximize, at least locally, any criterion on the unit sphere (Chauvet et al, 2019; Bry et al, 2020a,b; Gibaud et al, 2024). (ii) Given all  $\mathbf{u}_r$ , maximize the criterion with respect to  $\Theta$ . This step is performed using a classical likelihood maximization algorithm relevant to the situation.

## 2.5 Higher rank components

Consider step (i) of the combined criterion maximization, and suppose we want to extract a given number of components  $H_r \leq \text{rank}(\mathbf{X}_r)$  from each theme  $\mathbf{X}_r$ . Let  $\mathbf{f}_r^h = \mathbf{X}_r \mathbf{u}_r^h$  be the rank- $h$  component of theme  $\mathbf{X}_r$ , and let  $\mathbf{F}_r^h = [\mathbf{f}_r^1, \dots, \mathbf{f}_r^h]$ , where  $h \leq H_r$ , be the matrix of the first  $h$  components of theme  $\mathbf{X}_r$ . The model comprising all components writes

$$\begin{aligned} \boldsymbol{\eta}_k &= \sum_{h=1}^{H_1} (\mathbf{X}_1 \mathbf{u}_1^h) \gamma_{k1}^h + \dots + \sum_{h=1}^{H_R} (\mathbf{X}_R \mathbf{u}_R^h) \gamma_{kR}^h + \mathbf{A} \boldsymbol{\delta}_k \\ &= \mathbf{F}_1^{H_1} \boldsymbol{\gamma}_{k1} + \dots + \mathbf{F}_R^{H_R} \boldsymbol{\gamma}_{kR} + \mathbf{A} \boldsymbol{\delta}_k, \end{aligned}$$

where  $\boldsymbol{\gamma}_{k1}, \dots, \boldsymbol{\gamma}_{kR}$  are vectors of regression parameters.

The new component  $\mathbf{f}_r^{h+1}$  must best complement both all other components and  $\mathbf{A}$ , that is  $\mathbf{A}_r^h := [\mathbf{F}_1^{H_1}, \dots, \mathbf{F}_{r-1}^{H_{r-1}}, \mathbf{F}_r^h, \mathbf{F}_{r+1}^{H_{r+1}}, \dots, \mathbf{F}_R^{H_R}, \mathbf{A}]$ . So  $\mathbf{f}_r^{h+1}$  has to be calculated using  $\mathbf{A}_r^h$  as the new set of additional covariates. Moreover, to avoid linear redundancy of components within each theme, we impose that  $\mathbf{f}_r^{h+1}$  be orthogonal to  $\mathbf{F}_r^h$ , *i.e.*  $\mathbf{F}_r^h \mathbf{T} \mathbf{W} \mathbf{f}_r^{h+1} = \mathbf{0}$ . We calculate every new component as the solution of program (2), with the additional constraint:  $\boldsymbol{\Delta}_r^h \mathbf{T} \mathbf{u}_r^{h+1} = \mathbf{0}$ , where  $\boldsymbol{\Delta}_r^h = \mathbf{X}_r^T \mathbf{W} \mathbf{F}_r^h$ , and loop on  $r$  until convergence of the overall criterion. The first  $H_r$  components in the Partial Least Squares (PLS; Wold et al, 1984) regression of  $\mathbf{Y}$  on  $\mathbf{X}_r$  are taken as initial values of  $\mathbf{F}_r^h$ . For all  $r = 1, \dots, R$ , the rank-1 component of theme  $\mathbf{X}_r$  is calculated using the same program with  $\mathbf{F}_r^0 = \emptyset$  and  $\boldsymbol{\Delta}_r^0 = \mathbf{0}$ .

## 3 Extending SCGLR to a factor model

As mentioned above, step (ii) of the combined criterion maximization boils down to maximizing the likelihood of the component-based model. In this section, the components are thus taken as known. For the sake of simplicity, we shall consider the matrix  $\mathbf{F} = [\mathbf{F}_1^{H_1}, \dots, \mathbf{F}_R^{H_R}]$  as the new set of explanatory variables and  $\boldsymbol{\gamma}_k = (\boldsymbol{\gamma}_{k1}^T, \dots, \boldsymbol{\gamma}_{kR}^T)^T$  its vector of regression parameters associated with the response  $\mathbf{y}_k$ .



### 3.1 SCGLR in a factor model context

Let  $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_K] \in \mathbb{R}^{N \times K}$  be the response matrix. For unit  $n$ , each response is assumed to be linearly modeled using the components and additional covariates, plus  $J$  factors  $\mathbf{g}_n = (g_{n1}, \dots, g_{nJ})^T$ . So, the linear predictor for unit  $n$  and response  $y_k$  writes

$$\eta_{nk} = \mathbf{f}_n^T \boldsymbol{\gamma}_k + \mathbf{a}_n^T \boldsymbol{\delta}_k + \mathbf{g}_n^T \mathbf{b}_k,$$

where  $\mathbf{f}_n$  and  $\mathbf{a}_n$  are the vectors composed of the  $n$ th rows of matrices  $\mathbf{F}$  and  $\mathbf{A}$  respectively, and  $\mathbf{b}_k$  is the vector of loadings associated with  $\mathbf{g}_n$ . The factors are assumed drawn from a multivariate normal distribution  $\mathbf{g}_n \sim \mathcal{N}_J(0, \mathbf{I}_J)$  and independent across statistical units. This model is designed so that the  $J$  factors capture as much as possible of the covariance between the responses not accounted for by the components and additional covariates, i.e. their residual covariance. Denoting  $\mathbf{G} \in \mathbb{R}^{N \times J}$  the matrix containing all the realizations of factors, the linear predictor associated with the response  $\mathbf{y}_k$  expressed column-wise becomes

$$\boldsymbol{\eta}_k = \mathbf{F}\boldsymbol{\gamma}_k + \mathbf{A}\boldsymbol{\delta}_k + \mathbf{G}\mathbf{b}_k.$$

Let  $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_K] \in \mathbb{R}^{J \times K}$  be the loading matrix. Jöreskog (1969) notices that the loading matrix  $\mathbf{B}$  is defined up to an arbitrary orthogonal transformation. To guarantee the identification of the model, we choose to constrain the  $J \times J$  sub-matrix of  $\mathbf{B}$  to be an upper triangular matrix with positive diagonal elements (Geweke and Zhou, 1996). An advantage of the factor model is to yield the matrix  $\boldsymbol{\Sigma} = \mathbf{B}^T \mathbf{B} \in \mathbb{R}^{K \times K}$ , storing the residual covariances of the responses, in a parsimonious manner. Indeed, the number of factors retained may remain small with respect to the size of the covariance matrix.

### 3.2 Estimating the parameters of a GLLVM

Let  $\boldsymbol{\Theta} = \{\boldsymbol{\gamma}_k, \boldsymbol{\delta}_k, \mathbf{b}_k \mid k = 1, \dots, K\}$  be the set of parameters. The marginal log-likelihood of the model is obtained by integrating over factors  $\mathbf{g}_n$

$$\begin{aligned} l(\boldsymbol{\Theta}; \mathbf{Y}) &= \sum_{n=1}^N \ln(L(\mathbf{y}_n; \boldsymbol{\Theta})) \\ &= \sum_{n=1}^N \ln \left( \int \prod_{k=1}^K L(y_{nk} \mid \mathbf{g}_n; \boldsymbol{\Theta}) L(\mathbf{g}_n) d\mathbf{g}_n \right). \end{aligned}$$

In a context of non-Gaussian responses, the maximization of this log-likelihood cannot be obtained in closed form. In the spirit of Saidane et al (2013), the estimation of the parameters is performed in two steps: first, we linearize the model; then, we maximize the pseudo-likelihood of the linearized model under a Gaussian assumption.

#### 3.2.1 The linearization step

Temporarily considering the factors given, i.e. conditional on  $\mathbf{G}$ , the above log-likelihood is that of a classic multivariate GLM. The adjusted dependent variable  $\mathbf{w}_k$

can be viewed as the response in the linearized model

$$\mathbf{w}_k = \mathbf{F}\boldsymbol{\gamma}_k + \mathbf{A}\boldsymbol{\delta}_k + \mathbf{G}\mathbf{b}_k + \boldsymbol{\zeta}_k,$$

where  $\mathbb{E}[\mathbf{w}_k | \mathbf{G}] = \mathbf{F}\boldsymbol{\gamma}_k + \mathbf{A}\boldsymbol{\delta}_k + \mathbf{G}\mathbf{b}_k$  and  $\mathbb{V}[\mathbf{w}_k | \mathbf{G}] = \mathbb{V}[\boldsymbol{\zeta}_k] = \mathbf{W}_k^{-1}$ .

### 3.2.2 The estimation step


In this step, we take the distribution of the adjusted dependent variables given  $\mathbf{F}$ ,  $\mathbf{A}$  and  $\mathbf{G}$  to be Gaussian, and view the factors as latent variables. The model pseudo-log-likelihood  $l(\boldsymbol{\Theta}; \mathcal{W})$ , where  $\mathcal{W}$  denotes the matrix of adjusted dependent variables, being difficult to maximize directly, we use the EM algorithm to estimate the model parameters. Assuming, in the wake of [Wolfinger and O'connell \(1993\)](#), that the adjusted dependent variables have a Gaussian distribution, we calculate and then maximize the expectation of their complete log-likelihood  $l(\boldsymbol{\Theta}; \mathcal{W}, \mathbf{G})$ . Further details of the derived EM algorithm are given in [Appendix D](#).

## 3.3 The overall algorithm

The overall algorithm, presented in [Appendix E](#), consists in alternating the following steps: (i) On the current linearized model, given the current set of parameters and corresponding expected factor values, calculate all the components of all the themes iteratively through the PING algorithm. (ii) Given the current components, calculate the adjusted dependent variables of the linearized model and their variance matrix. (iii) Given the adjusted dependent variables, re-estimate the factor model parameters and expected factor values through the EM algorithm. The method thus implemented is named F-SCGLR (for Factor-SCGLR).

## 3.4 Posterior clustering of responses

Recall that one of the aims of this work is to group the responses according to their mutual residual dependencies. In other words, two responses having a high residual correlation (positive or negative) should be cast to the same cluster. To achieve this, we propose the following strategy

1. Estimate the residual covariance matrix  $\boldsymbol{\Sigma} := \mathbf{B}^T \mathbf{B}$  of the linear predictors.
2. Calculate the corresponding residual correlation matrix  $\mathbf{C}$  where  $C_{ij} := \Sigma_{ij} / \sqrt{\Sigma_{ii}\Sigma_{jj}}$ .
3. Calculate the associated dissimilarity matrix  $\mathbf{D}$  where  $D_{ij}^2 := 2(1 - C_{ij}^2)$ . The squared residual correlation is used in order to consider as close two highly correlated responses whether this correlation be positive or negative.
4. Perform Multidimensional Scaling (MDS; [Cox and Cox, 2008](#)) on the matrix  $\mathbf{D}$  to obtain a Euclidean representation of the responses (i.e. a set of coordinates in a Euclidean space) with respect to this distance structure. We use the function `cmdscale` of the `stats`  package ([R Core Team, 2023](#)).
5. Perform a K-means algorithm (taking as a starting point the output of a hierarchical clustering procedure) on the coordinates obtained on the previous step. We use the

**factoextra**  $\mathbb{R}$  package (Kassambara, 2017) where the function *hkmeans* runs the K-means and the function *fviz-nbclust* optimizes the number of clusters using the silhouette criterion.

## 4 Simulation study

Several simulation studies have been conducted to assess the performance of F-SCGLR. The first one focuses on the identification of the right combination of components and factors. The combination was calibrated across the cross-product grid  $(H_1, \dots, H_R, J) \in \{1, \dots, 4\}^R \times \{0, \dots, 5\}$  by minimizing the Bayesian Information Criterion (BIC; Schwarz, 1978) to get more parsimonious models. As shown by Chauvet et al (2019), the hyper-parameters must be chosen so as to prevent the components from being drawn too strongly towards the principal components ( $s > 0.5$ ) or towards too local bundle ( $l > 10$ ). Thus, the second simulation aims at studying the influence of the hyper-parameters  $s \in \{0.1, 0.3, 0.5\}$  and  $l \in \{1, 2, 3, 4, 7, 10\}$  in the situation of a more or less clearly separated response cluster pattern. In this simulation, 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 steps detailed in Section 3.4. In addition, to measure the quality of the latent dimensions recovery, we calculate the maximum square correlation between each true dimension, represented by their direction vector  $\xi$  used for simulation, and the components

$$\rho^2(\xi, \cdot) = \max_{r,h} \rho(\xi, f_r^h)^2,$$

where  $f_r^h$  denotes the  $h$ th component of theme  $\mathbf{X}_r$ . Finally, as reference values for comparison, we calculated among other the RI and ARI of the partitions output by a competing  $\mathbb{R}$  package in a context of Gaussian, Poisson or Bernoulli responses. For each simulation, five hundred samples have been generated. The developed  $\mathbb{R}$  package **FactorSCGLR** and the simulation codes are available at <https://github.com/julien-gibaud/FactorSCGLR>.

### 4.1 Simulation in a context of mixed distributions

#### 4.1.1 Generation of the simulated data

The variables are simulated on  $N = 100$  statistical units. Five latent dimensions  $\xi_1, \xi_2, \xi_3, \xi_4$  and  $\xi_5$  are independently simulated from a standard multivariate normal distribution. The  $\mathbf{X}$  matrix consists in two themes:  $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$ . The first theme  $\mathbf{X}_1 = [\mathcal{X}_1, \mathcal{X}_2, \mathcal{M}_1]$  is made of three blocks:  $\mathcal{X}_1 \in \mathbb{R}^{N \times 90}$  and  $\mathcal{X}_2 \in \mathbb{R}^{N \times 60}$  are bundles of variables distributed about  $\xi_1$  and  $\xi_2$  respectively, and  $\mathcal{M}_1$  contains fifty unstructured noise variables drawn from a standard multivariate normal distribution. Likewise, the second theme  $\mathbf{X}_2 = [\mathcal{X}_3, \mathcal{X}_4, \mathcal{X}_5, \mathcal{M}_2]$  is made of four blocks:  $\mathcal{X}_3 \in \mathbb{R}^{N \times 100}$ ,  $\mathcal{X}_4 \in \mathbb{R}^{N \times 80}$  and  $\mathcal{X}_5 \in \mathbb{R}^{N \times 60}$  are bundles of variables distributed about  $\xi_3, \xi_4$  and  $\xi_5$  respectively, and  $\mathcal{M}_2$  contains sixty unstructured noise variables drawn from a standard multivariate normal distribution. This generation leads to  $P = 500$  explanatory variables. More formally, for all  $i = 1, \dots, 5$ , a variable  $x_{ip}$  within a

bundle is simulated as  $\mathbf{x}_{ip} = \boldsymbol{\xi}_i + \boldsymbol{\varepsilon}_p$ , where  $\boldsymbol{\varepsilon}_p \sim \mathcal{N}_N(0, 0.1\mathbf{I}_N)$ . The bundles are thus generated rather thin. To study the influence of wider bundles on the components, we refer the reader to [Chauvet et al \(2019\)](#). The  $N$  realizations of the  $J = 3$  factors, simulated through  $\mathbf{g}_n \sim \mathcal{N}_J(0, \mathbf{I}_J)$ , are stored in matrix  $\mathbf{G} \in \mathbb{R}^{N \times J}$ . The matrix  $\mathbf{B} \in \mathbb{R}^{J \times K}$  of factor loadings is generated so as to exhibit a three-cluster pattern


$$\begin{aligned} \forall k = 1, \dots, 5, \quad \mathbf{b}_k &\sim \mathcal{N}_J(\boldsymbol{\mu}_1, \sigma_B^2 \mathbf{I}_J) \\ \forall k = 6, \dots, 10, \quad \mathbf{b}_k &\sim \mathcal{N}_J(-\boldsymbol{\mu}_1, \sigma_B^2 \mathbf{I}_J) \\ \forall k = 11, \dots, 20, \quad \mathbf{b}_k &\sim \mathcal{N}_J(\boldsymbol{\mu}_2, \sigma_B^2 \mathbf{I}_J) \\ \forall k = 21, \dots, 35, \quad \mathbf{b}_k &\sim \mathcal{N}_J(-\boldsymbol{\mu}_2, \sigma_B^2 \mathbf{I}_J) \\ \forall k = 36, \dots, 50, \quad \mathbf{b}_k &\sim \mathcal{N}_J(\boldsymbol{\mu}_3, \sigma_B^2 \mathbf{I}_J), \end{aligned}$$

where  $\sigma_B^2 = 0.1$ ,  $\boldsymbol{\mu}_1 = (2, 0, 0)^T$ ,  $\boldsymbol{\mu}_2 = (0, -1, 0)^T$  and  $\boldsymbol{\mu}_3 = (0, 0, 1.5)^T$ . Finally, the response matrix  $\mathbf{Y}$  is simulated as a mix of Gaussian, Poisson and Bernoulli distributions, with

$$\begin{aligned} \forall k = 1, \dots, 20, \quad \mathbf{y}_k &\sim \mathcal{N}_N(\boldsymbol{\mu} = \gamma_{k1}\boldsymbol{\xi}_1 + \gamma_{k2}\boldsymbol{\xi}_2 + \mathbf{G}\mathbf{b}_k, \boldsymbol{\Sigma} = \mathbf{I}_N) \\ &\text{where } \gamma_{k1} \sim \mathcal{U}[-4, 4] \text{ and } \gamma_{k2} \sim \mathcal{U}[-2, 2], \\ \forall k = 21, \dots, 40, \quad \mathbf{y}_k &\sim \mathcal{P}(\boldsymbol{\lambda} = \exp[\gamma_{k4}\boldsymbol{\xi}_4 + \gamma_{k3}\boldsymbol{\xi}_3 + \mathbf{G}\mathbf{b}_k]) \\ &\text{where } \gamma_{k4} \sim \mathcal{U}[-2, 2] \text{ and } \gamma_{k3} \sim \mathcal{U}[-1, 1], \\ \forall k = 41, \dots, 50, \quad \mathbf{y}_k &\sim \mathcal{B}(\mathbf{p} = \text{logit}^{-1}[\gamma_{k3}\boldsymbol{\xi}_3 + \gamma_{k2}\boldsymbol{\xi}_2 + \mathbf{G}\mathbf{b}_k]) \\ &\text{where } \gamma_{k3} \sim \mathcal{U}[-2, 2] \text{ and } \gamma_{k2} \sim \mathcal{U}[-0.5, 0.5]. \end{aligned}$$

In the linear predictors, we order the latent variables by decreasing magnitude of their regression parameter.

#### 4.1.2 Identification of the true model

In this simulation, the hyper-parameters are first calibrated through the **SCGLR**  package (e.g. without factors) and set to  $s = 0.3$  and  $l = 4$ . Appendix F sums up the results on a cross-product grid. As expected, the combination which minimizes the BIC is given by the true combination  $(H_1, H_2, J) = (2, 2, 3)$ . However, several points deserve mentioning. We observe, for all component combinations, that the values of BIC decrease drastically when enough factors are involved in the model. This shows that, as mutual dependencies may generally exist, the residual covariance should be modeled. When the model involves too many factors (when  $J = 4$  and  $J = 5$ ), the number of useful components is underestimated. Indeed, the variability of the model captured by the factors then contains a part of the variability otherwise captured by the components. In the opposite situation, when  $J = 0$ , the BIC leads to overestimate the number of components.

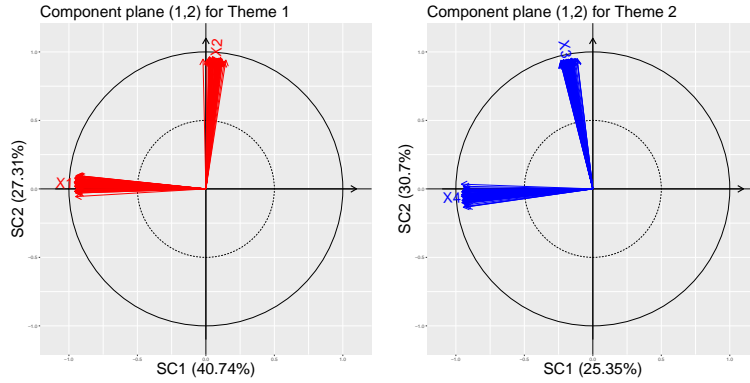
**Table 1:** Mean values of RI, ARI and square correlation over five hundred samples with  $\sigma_B^2 = 0.1$ ,  $s \in \{0.1, 0.3, 0.5\}$  and  $l \in \{1, 2, 3, 4, 7, 10\}$ .

$s$	$l$	RI	ARI	$\rho^2(\boldsymbol{\xi}_1, \cdot)$	$\rho^2(\boldsymbol{\xi}_2, \cdot)$	$\rho^2(\boldsymbol{\xi}_3, \cdot)$	$\rho^2(\boldsymbol{\xi}_4, \cdot)$
0.1	1	0.933	0.847	0.967	0.937	0.801	0.864
	2	0.920	0.818	0.988	0.970	0.887	0.931
	3	0.915	0.802	0.988	0.970	0.888	0.939
	4	0.931	0.842	0.983	0.970	0.888	0.956
	7	0.924	0.825	0.981	0.970	0.869	0.958
	10	0.918	0.812	0.980	0.970	0.869	0.958
0.3	1	0.920	0.820	0.975	0.940	0.720	0.739
	2	0.926	0.831	0.993	0.972	0.931	0.946
	3	0.926	0.836	0.985	0.972	0.927	0.967
	4	0.924	0.830	0.983	0.972	0.921	0.971
	7	0.925	0.834	0.978	0.972	0.904	0.968
	10	0.921	0.824	0.976	0.972	0.895	0.966
0.5	1	0.919	0.817	0.975	0.938	0.713	0.661
	2	0.921	0.823	0.994	0.972	0.928	0.950
	3	0.919	0.818	0.986	0.974	0.922	0.973
	4	0.920	0.820	0.984	0.974	0.920	0.975
	7	0.919	0.818	0.982	0.974	0.903	0.972
	10	0.920	0.821	0.982	0.974	0.903	0.972

### 4.1.3 Varying the hyper-parameters and the variance within the response clusters

Henceforth, keeping the true combination found by the BIC, we focus on the influence of the hyper-parameters  $s$  and  $l$  on the classification decision and latent dimension recovery. In order to compare the results in a context of more or less distinct cluster pattern, we vary the variance within the cluster by taking  $\sigma_B^2 \in \{0.1, 0.2, 0.3\}$ . As an illustration, Appendix G shows the residual correlation matrices.

Table 1 gives the results for  $\sigma_B^2 = 0.1$ . The tables summing up the results for  $\sigma_B^2 = 0.2$  and  $\sigma_B^2 = 0.3$  are presented in Appendix H. As expected, the higher the variance within the cluster, the weaker the values of RI and ARI for all the combinations of  $s$  and  $l$ . The main result about the square correlations is that the variance within the cluster does not have a relevant influence on the quality of the latent dimensions recovery. Indeed, the search for components is related to the deterministic part of the model, while  $\sigma_B^2$  is involved in the random part of the model. The square correlations, for  $s = 0.3$  and  $s = 0.5$  with  $l \geq 2$ , are greater than for  $s = 0.1$ . This observation is consistent with Chauvet et al (2019) who notice that the thinner the bundles, the greater the value of  $s$  has to be to recover the latent dimensions correctly. Here, indeed, the variance within the bundles is equal to 0.1 (thin bundles). However, the particular case of  $l = 1$  deserves mentioning. The components calculated with  $l = 1$  being closer to the principal components, the two components of theme  $\mathbf{X}_2$  are drawn in between the three bundles and so, produce low square correlations with the latent dimensions. The interest of tuning the locality parameter  $l$  is shown by the gap between the results obtained for  $l = 1$  and  $l = 2$ : in the latter case, the square correlations are higher. Furthermore,  $\boldsymbol{\xi}_3$  being the less explanatory latent dimension,  $\rho^2(\boldsymbol{\xi}_3, \cdot)$  is always lower than the other square correlations.



**Fig. 1:** Correlation scatterplot of plane (1,2) for the two themes with  $s = 0.3$  and  $l = 4$  obtained by the F-SCGLR algorithm. The red arrows represent the bundles  $\mathcal{X}_1$  and  $\mathcal{X}_2$  which explain the first theme. The blue arrows represent the bundles  $\mathcal{X}_3$  and  $\mathcal{X}_4$  which explain the second theme. The percentage of inertia captured by each component is given in parentheses.

Figure 1 shows the correlation scatterplots in the component planes (1, 2) for the first two themes. The components are almost perfectly aligned with the explanatory bundles. However, as observed in Table 1, the bundle  $\mathcal{X}_3$  seems slightly less correlated with the component  $f_2^2$  than the other bundles with their corresponding components.

## 4.2 Comparative study

To compare the different GLLVM implementations, we use the  $\mathbb{R}$  package `gllvm` (Niku et al, 2019b, 2023). This package offers three ways to fit GLLVM: using a variational approximation (VA; Hui et al, 2017), a Laplace approximation (LA; Niku et al, 2017, 2019a) or an extended variational approximation (EVA; Korhonen et al, 2023). The  $\mathbb{R}$  package `gllvm` also proposes the concurrent ordination (van der Veen et al, 2023) which performs a reduced rank regression in order to reduce the number of parameters to estimate.

SCGLR and concurrent ordination both address dimension reduction using a small number of latent variables. Both use linear combinations of the covariates (components) with random latent variables (factors). But in the two methods, the components do not pursue exactly the same goal, and so, are not calculated in the same way. Concurrent ordination calculates them maximizing the model likelihood, whereas SCGLR calculates them maximizing a criterion that combines this likelihood with the structural relevance of the components, which measures the information they capture in each theme. With a fixed coefficient model, the concurrent ordination does not perform regularization. Now, `gllvm` includes a random slope model option, which does regularize the coefficient estimation, much the same way as ridge regression does (Hoerl and Kennard, 1970). So, in case of collinearities in explanatory variables, this option should be used.

The performances of the methods are compared through the RI and ARI of the partitions output by the estimated residual correlation matrix, the Procrustes errors between the true and estimated loading matrix  $\mathbf{B}$  and between the true and estimated matrix of factors  $\mathbf{G}$ , latent dimension recovery, root mean square errors of the regression parameters of matrix  $\mathbf{A}$  and computation time in seconds. The results are output by our package **FactorSCGLR** performing the F-SCGLR method, on the one hand, and the package **gllvm** implementing the LA, VA or EVA approaches, on the other hand. We have compared the performances of **gllvm** with random slopes, without random slopes, and of **FactorSCGLR**, on simulated data with Gaussian, Poisson, and binary responses. Since the performances of **gllvm** with and without random slopes did not prove much different except for the computation time, the results for random slopes are presented in Appendix H.

Due to the excessive computation time of the Bayesian MCMC methods, the  $\mathbb{R}$  packages **boral** (Hui, 2016) and **Hmsc** (Tikhonov et al, 2020) are not tested in this article. Their performances are respectively discussed by Niku et al (2019b) and Pichler and Hartig (2021).

#### 4.2.1 Generation of the simulated data

The variables are simulated on  $N \in \{100, 200, 300\}$  statistical units. Two bundles of five variables distributed about the latent dimensions  $\xi_1$  and  $\xi_2$  respectively are generated. Ten unstructured noise variables complete the matrix  $\mathbf{X}$ . One categorical variable with eight levels is taken as only additional covariate  $\mathbf{A}$ . In this simulation,  $J = 2$  factors are simulated to model the residual covariance of the  $K \in \{10, 30, 50\}$  responses. The loadings of the factors are generated in order to get a two-cluster design

$$\begin{aligned} \forall k = 1, \dots, 0.4K, \quad \mathbf{b}_k &\sim \mathcal{N}_J \left( (-1)^k \boldsymbol{\mu}_1, 0.1\mathbf{I}_J \right) \\ \forall k = 0.4K + 1, \dots, K, \quad \mathbf{b}_k &\sim \mathcal{N}_J \left( (-1)^k \boldsymbol{\mu}_2, 0.1\mathbf{I}_J \right), \end{aligned}$$

where  $\boldsymbol{\mu}_1 = (0, 2)^T$  and  $\boldsymbol{\mu}_2 = (1.5, 0)^T$ . As an illustration, Appendix G shows the residual correlation matrices. For all  $k = 1, \dots, K$ , the simulated linear predictor for response  $y_k$  is thus given by

$$\boldsymbol{\eta}_k = \gamma_{k1}\boldsymbol{\xi}_1 + \gamma_{k2}\boldsymbol{\xi}_2 + \mathbf{A}\boldsymbol{\delta}_k + \mathbf{G}\mathbf{b}_k,$$

where  $\gamma_{k1}$ ,  $\gamma_{k2}$  and  $\boldsymbol{\delta}_k$  are uniformly generated, with  $\gamma_{k1} \in [-4, 4]$ ,  $\gamma_{k2} \in [-2, 2]$  and  $\boldsymbol{\delta}_k \in [-0.5, 0.5]$ . The package we want to compare F-SCGLR to, not allowing to consider different distribution families for the responses, we restrict the comparison to responses having the same distribution.

This comparative study is divided into three parts. First, the response variables are drawn from a normal distribution. In this context, the estimation of the parameters being the same for VA and EVA, our method is compared with LA and VA. In the second part, the response variables are Poisson distributed for which LA and VA are available in the **gllvm** package. The third part is dedicated to Bernoulli distributed responses. As detailed by Korhonen et al (2023), the VA method fails to give

a closed-form approximation of the log-likelihood with the logit link. However, the **FactorSCGLR** package is only implemented for the logit link. So, this part compares F-SCGLR with LA and EVA only. In all cases, the **gllvm** package runs with a reduced rank regression in order to compare the estimates of the components and factors.

In this simulation the **R** package **SCGLR** calibrates the hyper-parameters to  $s = 0.5$  and  $l = 4$ , while the BIC selects  $H_1 = 2$  and  $J = 2$ .

#### 4.2.2 Comparison results for the Gaussian distribution

Table 2 sums up the results for the Gaussian distribution. As expected, the three methods perform better when the number of either statistical units or responses increase. When  $K \neq 10$ , the three methods have classification indices close to 1. Similarly, the Procrustes errors of the loadings  $\mathbf{B}$  are small for the best combination of each method. F-SCGLR, LA and VA respectively reach 0.006, 0.029 and 0.041. The **gllvm** package has close Procrustes errors of the loadings  $\mathbf{G}$  for VA (0.122) and LA (0.124), while 0.047 is obtained by F-SCGLR. Indeed, since the concurrent ordination method implemented in the **gllvm** package does not deal with structural relevance, the components found by F-SCGLR better align with the latent dimensions than VA and LA. This observation remains valid across the simulations when using other distributions for the responses. The lowest value of RMSE is obtained by F-SCGLR, albeit a same order of magnitude is achieved by the three methods. For all cases, F-SCGLR is the fastest method. Moreover, we observe that LA runs faster than VA when  $K$  or  $N$  increase.

#### 4.2.3 Comparison results for the Poisson distribution

Table 3 sums up the results for the Poisson distribution. Unlike with the simulations involving Gaussian and Bernoulli distributions, F-SCGLR has the least good rate of classification for all  $N$  and  $K$ . Moreover, for  $N = 300$ , we observe that a higher number of responses may cause a deterioration of the classification of the compared methods. The Procrustes errors of  $\mathbf{B}$  and  $\mathbf{G}$  computed by F-SCGLR are greater than those of methods implemented in the **gllvm** package. The linearization might be to blame in the Poisson case, because the log link can cause instability in the estimation results. We observe that these errors are around twice as high for F-SCGLR. This causes the RI and ARI to be lower. The latent dimension recovery, although better than with VA and LA, appears to be greater when an other distribution is used for the response variables. In all cases, F-SCGLR obtains the highest RMSE. The estimation of the regression parameters of  $\mathbf{A}$  given by VA and LA are of the same order. In the Poisson distribution case, the computation time is significantly longer for the three methods.

#### 4.2.4 Comparison results for the Bernoulli distribution

Table 4 sums up the results for the Bernoulli distribution. We observe that, for  $N \in \{100, 200\}$ , F-SCGLR gives the best values of RI and ARI, followed by LA and EVA. When  $N = 300$  and  $K = 50$ , the three methods have a classification rate close to 1. The highest value obtained of the ARI is given by LA, followed by F-SCGLR and for LA. For all cases, the lowest Procrustes errors on  $\mathbf{B}$  and  $\mathbf{G}$  are reached by F-SCGLR. The minimal errors for  $\mathbf{B}$  are 0.055, 0.129 and 0.078 for F-SCGLR, EVA



**Table 2:** Mean values of RI, ARI, Procrustes error of the factors  $\mathbf{G}$  and their loadings  $\mathbf{B}$ , latent dimension recovery, root mean square error and computation time over five hundred samples with  $N \in \{100, 200, 300\}$  and  $K \in \{10, 30, 50\}$  for the Gaussian distribution.

$N$	100			200			300		
$K$	10	30	50	10	30	50	10	30	50
F-SCGLR									
RI	0.966	0.994	0.997	0.970	0.993	0.999	0.981	0.993	0.998
ARI	0.930	0.988	0.994	0.938	0.986	0.998	0.962	0.987	0.997
Proc $B$	0.020	0.022	0.022	0.010	0.010	0.010	0.007	0.006	0.007
Proc $G$	0.162	0.121	0.112	0.117	0.071	0.064	0.102	0.057	0.047
$\rho^2(\xi_1, \cdot)$	0.970	0.968	0.968	0.974	0.975	0.974	0.976	0.976	0.975
$\rho^2(\xi_2, \cdot)$	0.945	0.943	0.943	0.963	0.963	0.962	0.966	0.965	0.966
$RMSE_A$	0.772	0.862	0.771	0.563	0.554	0.571	0.441	0.467	0.448
Time	0.520	2.552	7.117	0.938	4.396	12.60	1.352	6.115	18.26
gllvm-VA									
RI	0.913	0.981	0.980	0.931	0.977	0.986	0.928	0.974	0.989
ARI	0.823	0.963	0.960	0.859	0.953	0.972	0.857	0.949	0.978
Proc $B$	0.164	0.067	0.056	0.122	0.058	0.048	0.174	0.061	0.041
Proc $G$	0.219	0.172	0.170	0.163	0.136	0.126	0.169	0.136	0.122
$\rho^2(\xi_1, \cdot)$	0.852	0.828	0.829	0.834	0.840	0.822	0.844	0.846	0.848
$\rho^2(\xi_2, \cdot)$	0.718	0.797	0.812	0.777	0.811	0.804	0.786	0.832	0.846
$RMSE_A$	0.793	0.902	0.777	0.590	0.562	0.606	0.464	0.497	0.490
Time	4.486	13.15	29.89	12.83	35.30	65.44	26.64	63.07	112.8
gllvm-LA									
RI	0.951	0.970	0.963	0.966	0.961	0.997	0.948	0.970	0.992
ARI	0.899	0.940	0.926	0.930	0.923	0.994	0.892	0.940	0.983
Proc $B$	0.060	0.047	0.050	0.029	0.040	0.033	0.051	0.035	0.034
Proc $G$	0.235	0.161	0.195	0.203	0.175	0.124	0.207	0.166	0.126
$\rho^2(\xi_1, \cdot)$	0.859	0.811	0.851	0.850	0.848	0.835	0.808	0.841	0.874
$\rho^2(\xi_2, \cdot)$	0.706	0.783	0.824	0.813	0.836	0.817	0.779	0.837	0.870
$RMSE_A$	0.818	0.879	0.798	0.614	0.582	0.600	0.479	0.513	0.500
Time	6.060	13.79	28.22	13.62	26.92	59.54	17.70	42.71	79.34

and LA respectively. However, we may note that all methods predict the factors with the same order of magnitude. In their best combinations, the values of the Procrustes error on  $\mathbf{G}$  are equal to 0.153 (F-SCGLR), 0.179 (EVA) and 0.163 (LA). For  $K = 50$  and  $N \in \{200, 300\}$ , the lowest value of RMSE is reached by LA while for the other cases F-SCGLR gives the lowest value. Across this simulations, EVA is the fastest for  $N = 100$ . For the other cases, F-SCGLR ran faster. As noted by [Korhonen et al \(2023\)](#), LA is relatively slow.

## 5 Analysis of an agricultural ecology dataset

### 5.1 Data description


We apply F-SCGLR to the dataset available following the link <https://doi.org/10.15454/AJZUQN>. The sample we consider gives the observation of  $K = 12$  agrobiodiversity variables over  $N = 54$  winter cereal fields in the French Vallées et Coteaux de Gascogne. The agrobiodiversity is reported through three carabid beetle variables (two abundances and a Shannon index), three vascular plant

**Table 3:** Mean values of RI, ARI, Procrustes error of the factors  $\mathbf{G}$  and their loadings  $\mathbf{B}$ , latent dimension recovery, root mean square error and computation time over five hundred samples with  $N \in \{100, 200, 300\}$  and  $K \in \{10, 30, 50\}$  for the Poisson distribution.

$N$	100			200			300		
$K$	10	30	50	10	30	50	10	30	50
F-SCGLR									
RI	0.826	0.881	0.881	0.872	0.824	0.855	0.865	0.866	0.855
ARI	0.651	0.763	0.765	0.743	0.650	0.713	0.727	0.735	0.714
Proc $B$	0.307	0.286	0.341	0.257	0.388	0.366	0.260	0.349	0.423
Proc $G$	0.512	0.550	0.579	0.456	0.517	0.548	0.429	0.507	0.526
$\rho^2(\xi_1, \cdot)$	0.966	0.958	0.966	0.974	0.974	0.963	0.976	0.966	0.976
$\rho^2(\xi_2, \cdot)$	0.816	0.909	0.876	0.825	0.892	0.849	0.808	0.900	0.889
$RMSE_{\mathbf{A}}$	0.914	1.141	1.528	0.886	1.015	1.387	0.959	0.943	1.233
Time	4.373	13.51	29.22	8.053	24.44	63.03	10.66	32.76	102.7
gllvm-VA									
RI	0.939	0.942	0.951	0.940	0.886	0.904	0.900	0.900	0.886
ARI	0.874	0.884	0.903	0.877	0.773	0.811	0.797	0.800	0.774
Proc $B$	0.157	0.152	0.121	0.170	0.233	0.215	0.244	0.205	0.245
Proc $G$	0.273	0.249	0.236	0.253	0.287	0.281	0.289	0.270	0.289
$\rho^2(\xi_1, \cdot)$	0.817	0.815	0.817	0.852	0.818	0.827	0.816	0.808	0.823
$\rho^2(\xi_2, \cdot)$	0.579	0.730	0.759	0.708	0.789	0.782	0.692	0.766	0.782
$RMSE_{\mathbf{A}}$	0.708	0.789	0.715	0.476	0.619	0.511	0.363	0.412	0.501
Time	8.893	35.98	78.44	33.92	101.8	206.4	73.93	223.2	438.9
gllvm-LA									
RI	0.915	0.920	0.917	0.919	0.858	0.918	0.941	0.918	0.896
ARI	0.825	0.840	0.838	0.832	0.716	0.839	0.878	0.836	0.794
Proc $B$	0.104	0.164	0.176	0.141	0.266	0.205	0.148	0.211	0.198
Proc $G$	0.278	0.280	0.271	0.247	0.353	0.272	0.245	0.280	0.263
$\rho^2(\xi_1, \cdot)$	0.832	0.755	0.818	0.850	0.852	0.787	0.825	0.850	0.801
$\rho^2(\xi_2, \cdot)$	0.704	0.721	0.767	0.828	0.782	0.778	0.803	0.857	0.799
$RMSE_{\mathbf{A}}$	0.718	0.753	0.756	0.455	0.522	0.486	0.363	0.430	0.403
Time	100.7	228.2	456.4	160.4	480.0	933.4	280.3	671.6	1557

variables (richness, relative cover and a Shannon index) and six axes from correspondence analyses (CA) performed on presence-absence data of carabid species and plant species respectively. The three abundance and richness responses are assumed to be Poisson random variables while the other responses are considered normally distributed. To model the agrobiodiversity, we have  $P = 21$  variables partitioned into  $R = 4$  themes and  $Q = 1$  additional covariate. The first theme  $\mathbf{X}_1$  characterizes the pest control through four variables. Six farming intensity variables make up the second theme  $\mathbf{X}_2$ . The third and fourth themes  $\mathbf{X}_3$  and  $\mathbf{X}_4$  gather six and five variables representing the landscape heterogeneity related to semi-natural covers and to the crop mosaic, respectively. The binary categorical variable coding the observation year (2016 or 2017) is taken as the additional covariate put into matrix  $\mathbf{A}$ . For more information about this dataset, we refer the reader to [Dufлот et al \(2022\)](#).


## 5.2 Results and interpretation

As in Section 4, we need to calibrate the hyper-parameters. We first tune  $s$  and  $l$  through the SCGLR  package, then find the best combination of component and

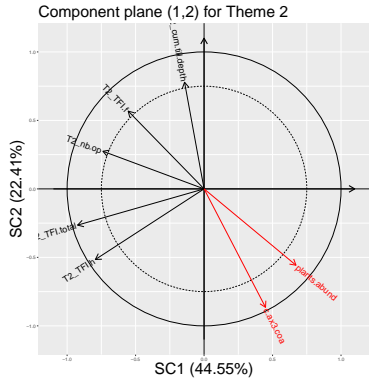
**Table 4:** Mean values of RI, ARI, Procrustes error of the factors  $\mathbf{G}$  and their loadings  $\mathbf{B}$ , latent dimension recovery, root mean square error and computation time over five hundred samples with  $N \in \{100, 200, 300\}$  and  $K \in \{10, 30, 50\}$  for the Bernoulli distribution.

$N$	100			200			300		
$K$	10	30	50	10	30	50	10	30	50
F-SCGLR									
RI	0.852	0.824	0.851	0.920	0.950	0.971	0.957	0.975	0.991
ARI	0.693	0.649	0.706	0.834	0.900	0.943	0.910	0.950	0.982
Proc $B$	0.171	0.194	0.225	0.096	0.087	0.090	0.068	0.055	0.056
Proc $G$	0.478	0.283	0.224	0.440	0.234	0.176	0.424	0.217	0.153
$\rho^2(\xi_1, \cdot)$	0.966	0.970	0.969	0.973	0.975	0.974	0.974	0.976	0.976
$\rho^2(\xi_2, \cdot)$	0.946	0.945	0.943	0.963	0.963	0.963	0.965	0.966	0.966
$RMSE_A$	1.423	1.763	1.820	0.807	0.967	1.009	0.657	0.765	0.788
Time	3.954	5.828	7.126	1.072	3.520	9.712	1.384	4.807	12.69
gllvm-EVA									
RI	0.655	0.719	0.724	0.691	0.890	0.945	0.721	0.940	0.969
ARI	0.325	0.439	0.449	0.394	0.781	0.890	0.451	0.880	0.939
Proc $B$	1.051	0.670	0.579	1.120	0.551	0.238	1.132	0.358	0.129
Proc $G$	0.664	0.393	0.337	0.644	0.291	0.214	0.637	0.265	0.179
$\rho^2(\xi_1, \cdot)$	0.798	0.839	0.870	0.815	0.849	0.847	0.829	0.812	0.842
$\rho^2(\xi_2, \cdot)$	0.483	0.620	0.738	0.537	0.719	0.750	0.616	0.710	0.780
$RMSE_A$	1.971	1.914	1.952	1.291	1.195	1.110	0.980	0.885	0.857
Time	1.454	2.806	4.601	2.623	12.56	43.61	3.966	38.36	107.9
gllvm-LA									
RI	0.691	0.759	0.780	0.878	0.938	0.962	0.859	0.997	0.990
ARI	0.389	0.520	0.567	0.749	0.876	0.925	0.719	0.995	0.981
Proc $B$	1.072	1.321	1.439	1.137	0.655	0.398	1.132	0.078	0.094
Proc $G$	0.707	0.349	0.261	0.519	0.250	0.192	0.507	0.230	0.163
$\rho^2(\xi_1, \cdot)$	0.812	0.824	0.834	0.851	0.835	0.792	0.810	0.844	0.851
$\rho^2(\xi_2, \cdot)$	0.432	0.620	0.664	0.620	0.742	0.749	0.637	0.791	0.820
$RMSE_A$	2.533	2.196	1.911	1.422	0.947	0.706	1.634	0.516	0.527
Time	174.3	479.9	825.5	180.7	227.5	591.1	146.9	333.3	623.3

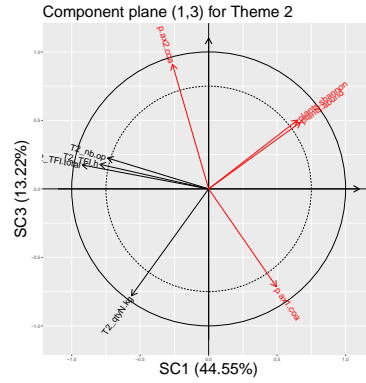
factor numbers using the BIC. In view of the small number of explanatory variables in each theme, we only allow the number of components to reach  $H_r = 3$ . We thus minimize the BIC on the cross-product grid  $(H_1, H_2, H_3, H_4, J) \in \{0, \dots, 3\}^4 \times \{0, \dots, 5\}$  with the  $s$  and  $l$  values previously found.

The cross-validation performed by the **SCGLR**  package recommends tuning hyper-parameters to  $s = 0.5$  and  $l = 1$  for this agricultural ecology dataset. Then, the component combination minimizing the BIC is  $(H_1, H_2, H_3, H_4) = (0, 3, 0, 0)$ , meaning that only the farming intensity theme was found relevant for the prediction of the agrobiodiversity. [Duflot et al \(2022\)](#) make the assumption that agrobiodiversity is predictable from the farming intensity (theme  $\mathbf{X}_2$ ) and the landscape heterogeneity (themes  $\mathbf{X}_3$  and  $\mathbf{X}_4$ ). The combination found by the BIC validates this hypothesis as to the effect of the farming intensity and the non-effect of the pest control in the prediction of the agrobiodiversity. However, the landscape heterogeneity themes proved useless for this prediction here.

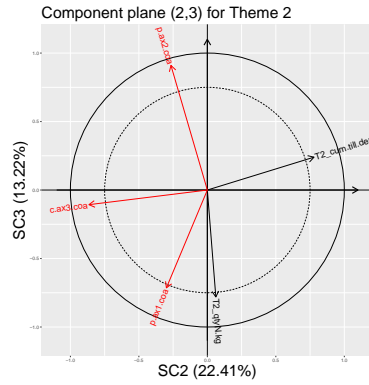
Let us now try to interpret the components of the second theme. The first component  $\mathbf{f}_2^1$  is correlated ( $\rho = -0.924$ ,  $\rho = -0.794$  and  $\rho = -0.738$ ) with a bundle of three



(a) Component plane (1,2) of the farming intensity theme



(b) Component plane (1,3) of the farming intensity theme

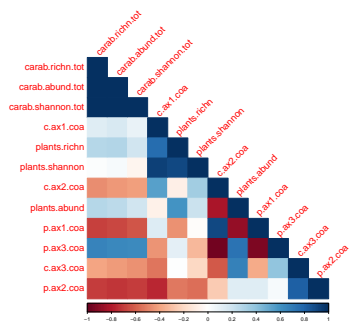


(c) Component plane (2,3) of the farming intensity theme

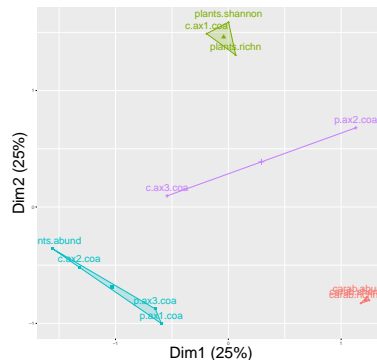
**Fig. 2:** Correlation plots of F-SCGLR plane (1,2), (2,3) and (1,3) of the second theme (farming intensity). The black arrows represent the theme’s covariates while the red arrows are the linear predictors of the responses. The plot displays only variables having a cosine greater than 0.75 with the plane. The percentage of inertia captured by each component is given in parentheses.

variables, of which “TFI.total” and “TFI.h” represent a treatment frequency index of herbicides, and “nb.op” is the total number of operations conducted by the farmers. The second component  $f_2^2$  is correlated ( $\rho = 0.779$ ) with the variable “cum.till.depth” measuring the cumulative tillage depth. The quantity of nitrogen denoted “qtyN.kg” is the most correlated explanatory variable ( $\rho = -0.781$ ) with the last component  $f_3^3$ . Figure 2 represents the correlation plots of the second theme (farming intensity).

In this agricultural ecology dataset, three factors are recommended by the BIC, to model the residual variance-covariance matrix. By applying the clustering steps given in Section 3.4, four groups of responses are identified. The first group is composed



(a) Residual correlation matrix for the agricultural ecology dataset. The response variables are ordered by group.




(b) Plane (1,2) derived from the residual correlation matrix by MDS. The clusters are identified through their colors.

**Fig. 3:** The residual correlation matrix alongside the Euclidean representation output by the MDS.

by the three measures of the carabids. The second group gathers the first CA axis of the carabids, the plant richness and the plant Shannon diversity index. The third group gathers the carabids' second CA axis, the plant cover and the first and third plants' CA axis. Finally, the fourth group contains the carabids' third CA axis and the plants' second CA axis. Figure 3 shows the residual correlation values alongside their Euclidean representation output by the MDS. The carabids' measure group having very high residual correlations, the distances between the responses composing it are close to 0 leading to a very compact group (in red) on the first principal plane of the MDS. On the contrary, the weaker the residual correlations, the wider the groups are scattered on the graph.


## 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. THEME-SCGLR extended SCGLR to a thematic partition of the explanatory variables, allowing to make better use of the complementary between the explanatory themes, both statistically when fitting the model, and conceptually when interpreting the components. F-SCGLR refines THEME-SCGLR in a major way: using factors besides components, it models the residual variance-covariance matrix of the responses with parsimony. This matrix can then be used for clustering, enabling to identify groups of linked responses.

In our simulation study, F-SCGLR proved to behave as expected regarding response clusters. Whenever the clusters were reasonably distinct, the original partition was recovered. We compared F-SCGLR with other methods in situations allowing the comparison. With Normal and Bernoulli-distributed responses, F-SCGLR performed better than the competing methods, which was not the case with Poisson-distributed responses. This might be due to some instability introduced in the linearized model by the log link function. Whatever the dispersion of the regression coefficients within the clusters, F-SCGLR provided components aligned with the simulated latent dimensions underlying the explanatory variables. Our  package refines the package **gllvm** in two ways: (i) Components having enough SR allow both an interpretable dimension reduction and regularization, which is mandatory whenever the explanatory variables are not linearly independent, e.g. in a high dimensional situation. Besides, the coefficient shrinkage implied in this regularization improves prediction. (ii) Responses with different distribution families are allowed in the response set. Applying the method to the agricultural ecology dataset, we found four groups of responses. The first group gathers the measures of the carabids. The other groups are composed by a mix between the plant variables and the axes output by correspondence analyses. However, even though a strong residual covariance between responses may hint at a biological interaction between species (Pollock et al, 2014), Poggiato et al (2021) recall that the residual correlations cannot distinguish the biotic from the abiotic effects. Besides, performing F-SCGLR revealed that the treatment by herbicides, the operations conducted by the farmers, the tillage depth and the quantity of nitrogen are the variables most involved in the prediction of the agrobiodiversity.

In this research, some limitations have been reached. Using the EM algorithm on each step of the overall algorithm involves a high number of iterations. Due to the absence of consensus about the maximization of the log-likelihood, we think that more research on this topic is necessary. SCGLR and its extensions have too many hyperparameters, which make it necessary so far to use heuristics for their calibration. Moreover, only Bernoulli, Binomial, Gaussian and Poisson distributions are currently handled in the **FactorSCGLR** package. The package should be improved by adding different distributions as Negative Binomial, Zero Inflated Poisson, Tweedie, Gamma, Beta or Exponential, which are allowed in the **gllvm** package, among others. Finally, for distributions allowing it, an (extended) variational approximation approach to criterion maximization should be implemented.

## Declarations

- There is no potential conflict of interest.
- This research was supported by the GAMBAS project funded by the French Agence Nationale de la Recherche (ANR-18-CE02-0025).
- The agricultural ecology dataset is available following the link <https://doi.org/10.15454/AJZUQN>.
- The  package **FactorSCGLR** and the simulation codes are available at <https://github.com/julien-gibaud/FactorSCGLR>.
- All the authors are contributed equally to this work.

## Appendix A Analytical expression of the SCGLR-specific criterion and its derivative

The specific criteria, which SCGLR maximizes to compute the  $(h + 1)$ -th vector of component coefficients, writes

$$\phi(\mathbf{u}) = \left( \sum_{j=1}^J \omega_j (\mathbf{u}^T \mathbf{X}^T \mathbf{N}_j \mathbf{X} \mathbf{u})^l \right)^{1/l} \quad (\text{A1})$$

and

$$\psi_{A_h}(\mathbf{u}) = \sum_{k=1}^K \|\mathbf{w}_k\|_{W_k}^2 \cos_{W_k}^2(\mathbf{w}_k, \text{span}[\mathbf{X}\mathbf{u}, \mathbf{A}^h]), \quad (\text{A2})$$

where for all  $j = 1, \dots, J$ ,  $\mathbf{N}_j$  is a symmetric semi-definite positive matrix. These matrices are chosen such that the quadratic forms  $\mathbf{u}^T \mathbf{X}^T \mathbf{N}_j \mathbf{X} \mathbf{u}$  measure the closeness of the vector of component coefficients, or equivalently the corresponding component, to some reference structures in the data.

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

### A.1 The structural relevance measure

In practice, we take either the variance component or the variable powered inertia (VPI). In the first case, the SR and its gradient are easily given by

$$\phi(\mathbf{u}) = \|\mathbf{X}\mathbf{u}\|_{\mathbf{W}}^2 \quad \text{and} \quad \nabla_{\mathbf{u}}\phi(\mathbf{u}) = 2\mathbf{X}^T \mathbf{W} \mathbf{X} \mathbf{u}.$$

The explicit expression of the VPI is

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

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

$$\begin{aligned} \nabla_{\mathbf{u}}\phi(\mathbf{u}) &= \frac{1}{l} \left[ \nabla_{\mathbf{u}} \left( \frac{1}{P} \sum_{p=1}^P \langle \mathbf{X}\mathbf{u}, \mathbf{x}_p \rangle_{\mathbf{W}}^{2l} \right) \right] \left[ \frac{1}{P} \sum_{p=1}^P \langle \mathbf{X}\mathbf{u}, \mathbf{x}_p \rangle_{\mathbf{W}}^{2l} \right]^{1/l-1} \\ &= \frac{1}{l} \left[ \frac{1}{P} \sum_{p=1}^P 2l \mathbf{X}^T \mathbf{W} \mathbf{x}_p \langle \mathbf{X}\mathbf{u}, \mathbf{x}_p \rangle_{\mathbf{W}}^{2l-1} \right] \phi(\mathbf{u})^{1-l} \\ &= \frac{2}{P} \phi(\mathbf{u})^{1-l} \mathbf{X}^T \mathbf{W} \sum_{p=1}^P \langle \mathbf{X}\mathbf{u}, \mathbf{x}_p \rangle_{\mathbf{W}}^{2l-1} \mathbf{x}_p. \end{aligned}$$

## A.2 The goodness of fit measure

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

$$\text{span}[\mathbf{X}\mathbf{u}, \mathbf{A}_h] = \text{span}[\boldsymbol{\chi}_k^h \mathbf{u}, \mathbf{A}_h] \quad \text{with} \quad \boldsymbol{\chi}_k^h = \Pi_{\text{span}[\mathbf{A}_h]^\perp}^{\mathbf{W}_k} \mathbf{X}.$$

Since  $\text{span}[\boldsymbol{\chi}_k^h]$  is orthogonal to  $\text{span}[\mathbf{A}_h]$ ,

$$\Pi_{\text{span}[\mathbf{X}\mathbf{u}, \mathbf{A}_h]}^{\mathbf{W}_k} = \Pi_{\text{span}[\boldsymbol{\chi}_k^h \mathbf{u}, \mathbf{A}_h]}^{\mathbf{W}_k} = \Pi_{\text{span}[\boldsymbol{\chi}_k^h \mathbf{u}]}^{\mathbf{W}_k} + \Pi_{\text{span}[\mathbf{A}_h]}^{\mathbf{W}_k}.$$

Consequently, by classical Euclidean statistical concepts, we get

$$\begin{aligned} & \cos_{\mathbf{W}_k}^2(\mathbf{w}_k, \text{span}[\mathbf{X}\mathbf{u}, \mathbf{A}_h]) \\ &= \cos_{\mathbf{W}_k}(\mathbf{w}_k, \text{span}[\mathbf{X}\mathbf{u}, \mathbf{A}_h]) \cos_{\mathbf{W}_k}(\mathbf{w}_k, \text{span}[\mathbf{X}\mathbf{u}, \mathbf{A}_h]) \\ &= \left[ \frac{\left\| \Pi_{\text{span}[\mathbf{X}\mathbf{u}, \mathbf{A}_h]}^{\mathbf{W}_k} \mathbf{w}_k \right\|_{\mathbf{W}_k}}{\|\mathbf{w}_k\|_{\mathbf{W}_k}} \right] \left[ \frac{\left\langle \mathbf{w}_k, \Pi_{\text{span}[\mathbf{X}\mathbf{u}, \mathbf{A}_h]}^{\mathbf{W}_k} \mathbf{w}_k \right\rangle_{\mathbf{W}_k}}{\left\| \Pi_{\text{span}[\mathbf{X}\mathbf{u}, \mathbf{A}_h]}^{\mathbf{W}_k} \mathbf{w}_k \right\|_{\mathbf{W}_k}} \right] \\ &= \frac{\left\langle \mathbf{w}_k, \left( \Pi_{\text{span}[\boldsymbol{\chi}_k^h \mathbf{u}]}^{\mathbf{W}_k} + \Pi_{\text{span}[\mathbf{A}_h]}^{\mathbf{W}_k} \right) \mathbf{w}_k \right\rangle_{\mathbf{W}_k}}{\|\mathbf{w}_k\|_{\mathbf{W}_k}^2} \\ &= \frac{\left\langle \mathbf{w}_k, \Pi_{\text{span}[\boldsymbol{\chi}_k^h \mathbf{u}]}^{\mathbf{W}_k} \mathbf{w}_k \right\rangle_{\mathbf{W}_k}}{\|\mathbf{w}_k\|_{\mathbf{W}_k}^2} + \frac{\left\langle \mathbf{w}_k, \Pi_{\text{span}[\mathbf{A}_h]}^{\mathbf{W}_k} \mathbf{w}_k \right\rangle_{\mathbf{W}_k}}{\|\mathbf{w}_k\|_{\mathbf{W}_k}^2}. \end{aligned}$$

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

$$\begin{aligned} \psi_{\mathbf{A}_h}(\mathbf{u}) &= \sum_{k=1}^K \|\mathbf{w}_k\|_{\mathbf{W}_k}^2 \cos_{\mathbf{W}_k}^2(\mathbf{w}_k, \text{span}[\mathbf{X}\mathbf{u}, \mathbf{A}_h]) \\ &= \sum_{k=1}^K \left( \left\langle \mathbf{w}_k, \Pi_{\text{span}[\boldsymbol{\chi}_k^h \mathbf{u}]}^{\mathbf{W}_k} \mathbf{w}_k \right\rangle_{\mathbf{W}_k} + \left\langle \mathbf{w}_k, \Pi_{\text{span}[\mathbf{A}_h]}^{\mathbf{W}_k} \mathbf{w}_k \right\rangle_{\mathbf{W}_k} \right). \end{aligned}$$

Now,

$$\begin{aligned} & \left\langle \mathbf{w}_k, \Pi_{\text{span}[\boldsymbol{\chi}_k^h \mathbf{u}]}^{\mathbf{W}_k} \mathbf{w}_k \right\rangle_{\mathbf{W}_k} \\ &= \mathbf{w}_k^T \mathbf{W}_k \Pi_{\text{span}[\boldsymbol{\chi}_k^h \mathbf{u}]}^{\mathbf{W}_k} \mathbf{w}_k \\ &= \mathbf{w}_k^T \mathbf{W}_k \boldsymbol{\chi}_k^h \mathbf{u} \left( \mathbf{u}^T \boldsymbol{\chi}_k^h{}^T \mathbf{W}_k \boldsymbol{\chi}_k^h \mathbf{u} \right)^{-1} \mathbf{u}^T \boldsymbol{\chi}_k^h{}^T \mathbf{W}_k \mathbf{w}_k \\ &= \frac{\mathbf{u}^T \boldsymbol{\chi}_k^h{}^T \mathbf{W}_k \mathbf{w}_k \mathbf{w}_k^T \mathbf{W}_k \boldsymbol{\chi}_k^h \mathbf{u}}{\mathbf{u}^T \boldsymbol{\chi}_k^h{}^T \mathbf{W}_k \boldsymbol{\chi}_k^h \mathbf{u}}. \end{aligned}$$



Let

$$\mathbf{a}_k := \mathbf{X}_k^h T \mathbf{W}_k \mathbf{w}_k \mathbf{w}_k^T \mathbf{W}_k \mathbf{X}_k^h, \quad \mathbf{b}_k := \mathbf{X}_k^h T \mathbf{W}_k \mathbf{X}_k^h$$

and

$$c_k := \left\langle \mathbf{w}_k, \Pi_{\text{span}[\mathbf{A}_h]}^{\mathbf{W}_k} \mathbf{w}_k \right\rangle_{\mathbf{W}_k}.$$

We finally have

$$\psi_{\mathbf{A}_h}(\mathbf{u}) = \sum_{k=1}^K \left( \frac{\mathbf{u}^T \mathbf{a}_k \mathbf{u}}{\mathbf{u}^T \mathbf{b}_k \mathbf{u}} + c_k \right)$$

and

$$\nabla_{\mathbf{u}} \psi_{\mathbf{A}_h}(\mathbf{u}) = 2 \sum_{k=1}^K \frac{(\mathbf{u}^T \mathbf{b}_k \mathbf{u}) \mathbf{a}_k \mathbf{u} - (\mathbf{u}^T \mathbf{a}_k \mathbf{u}) \mathbf{b}_k \mathbf{u}}{(\mathbf{u}^T \mathbf{b}_k \mathbf{u})^2}.$$

## Appendix B The PING algorithm

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

$$\begin{cases} \max_{\mathbf{u}} C_h(\mathbf{u}), \\ \text{s.t. } \mathbf{u}^T \mathbf{M} \mathbf{u} = 1 \quad \text{and} \quad \mathbf{\Delta}_h^T \mathbf{u} = 0, \end{cases} \quad (\text{B3})$$

where  $C_h$  is a function of  $\mathbf{u}$  to maximize and  $\mathbf{\Delta}_h$  an additional constraint matrix. In the SCGLR context,  $C_h(\mathbf{u})$  is the specific criterion and  $\mathbf{\Delta}_h$  the orthogonal constraint matrix. We rewrite this optimization program by setting  $\mathbf{v} = \mathbf{M}^{1/2} \mathbf{u}$ ,  $G_h(\mathbf{v}) = C_h(\mathbf{M}^{-1/2} \mathbf{v})$  and  $\mathbf{E}_h = \mathbf{M}^{-1/2} \mathbf{\Delta}_h$ .

$$\begin{cases} \max_{\mathbf{v}} G_h(\mathbf{v}), \\ \text{s.t. } \mathbf{v}^T \mathbf{v} = 1 \quad \text{and} \quad \mathbf{E}_h^T \mathbf{v} = 0. \end{cases} \quad (\text{B4})$$

To solve (B4), we must equate to zero the gradient of the following Lagrangian

$$\mathcal{L}(\mathbf{v}, \lambda, \boldsymbol{\eta}) = G_h(\mathbf{v}) - \lambda(\mathbf{v}^T \mathbf{v} - 1) - \boldsymbol{\eta}^T \mathbf{E}_h^T \mathbf{v}.$$

Setting  $\Gamma_h(\mathbf{v}) = \nabla_{\mathbf{v}} G_h(\mathbf{v})$ , we have

$$\nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}, \lambda, \boldsymbol{\eta}) = 0 \Leftrightarrow \Gamma_h(\mathbf{v}) - 2\lambda \mathbf{v} - \mathbf{E}_h \boldsymbol{\eta} = 0 \quad (\text{B5})$$

$$\Leftrightarrow \mathbf{v} = \frac{1}{2\lambda} (\Gamma_h(\mathbf{v}) - \mathbf{E}_h \boldsymbol{\eta}). \quad (\text{B6})$$

Multiplying (B5) by  $\mathbf{E}_h^T$

$$2\lambda \underbrace{\mathbf{E}_h^T \mathbf{v}}_{=0} = \mathbf{E}_h^T \Gamma_h(\mathbf{v}) - \mathbf{E}_h^T \mathbf{E}_h \boldsymbol{\eta} \Leftrightarrow \mathbf{E}_h^T \Gamma_h(\mathbf{v}) = \mathbf{E}_h^T \mathbf{E}_h \boldsymbol{\eta}$$

$$\Leftrightarrow \boldsymbol{\eta} = (\mathbf{E}_h^T \mathbf{E}_h)^{-1} \mathbf{E}_h^T \Gamma_h(\mathbf{v}). \quad (\text{B7})$$

Substituting (B7) in (B6), we get

$$\begin{aligned} \mathbf{v} &= \frac{1}{2\lambda} \left( \Gamma_h(\mathbf{v}) - \mathbf{E}_h (\mathbf{E}_h^T \mathbf{E}_h)^{-1} \mathbf{E}_h^T \Gamma_h(\mathbf{v}) \right) \\ &= \frac{1}{2\lambda} \left( \mathbf{I} - \mathbf{E}_h (\mathbf{E}_h^T \mathbf{E}_h)^{-1} \mathbf{E}_h^T \right) \Gamma_h(\mathbf{v}) \\ &= \frac{1}{2\lambda} \mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}), \end{aligned}$$

where  $\mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} = \mathbf{I} - \mathbf{E}_h (\mathbf{E}_h^T \mathbf{E}_h)^{-1} \mathbf{E}_h^T$ . Finally, the constraint  $\|\mathbf{v}\|^2 = 1$  gives

$$\mathbf{v} = \frac{\frac{1}{2\lambda} \mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v})}{\left\| \frac{1}{2\lambda} \mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}) \right\|} = \frac{\mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v})}{\left\| \mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}) \right\|},$$

which suggests the basic iteration of the PING algorithm

$$\mathbf{v}^{(t+1)} = \frac{\mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}^{(t)})}{\left\| \mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}^{(t)}) \right\|}. \quad (\text{B8})$$

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  $(\mathbf{v}^{(t)}, \mathbf{v}^{(t+1)})$  is a direction of ascent. In other words, that

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

By construction, we know that on every iteration  $t$  of the algorithm,  $\mathbf{v}^{(t)}$  is orthogonal to  $\text{span}[\mathbf{E}_h]$ . Thus, since for all  $t$ ,  $\mathbf{v}^{(t)} = \mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} \mathbf{v}^{(t)}$ , we have

$$\begin{aligned} \left\langle \mathbf{v}^{(t+1)} - \mathbf{v}^{(t)}, \Gamma_h(\mathbf{v}^{(t)}) \right\rangle &= \left\langle \mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} (\mathbf{v}^{(t+1)} - \mathbf{v}^{(t)}), \Gamma_h(\mathbf{v}^{(t)}) \right\rangle \\ &= \left\langle \mathbf{v}^{(t+1)} - \mathbf{v}^{(t)}, \mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}^{(t)}) \right\rangle. \end{aligned}$$

Now, Equation (B8) implies that

$$\mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}^{(t)}) = \mathbf{v}^{(t+1)} \left\| \mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}^{(t)}) \right\|.$$

So,

$$\text{sgn} \left( \left\langle \mathbf{v}^{(t+1)} - \mathbf{v}^{(t)}, \Gamma_h(\mathbf{v}^{(t)}) \right\rangle \right) = \text{sgn} \left( \left\langle \mathbf{v}^{(t+1)} - \mathbf{v}^{(t)}, \mathbf{v}^{(t+1)} \right\rangle \right)$$

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

Finally,

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

Although iteration (B8) follows a direction of ascent, it does not guarantee that function  $G_h$  actually increases on every step. Indeed, we may go too far in such a direction, and overshoot the maximum. However, let us consider

$$\boldsymbol{\kappa}^{(t)} = \frac{\mathbf{\Pi}_{\operatorname{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}^{(t)})}{\left\| \mathbf{\Pi}_{\operatorname{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}^{(t)}) \right\|}.$$

Staying close enough to the current starting point on the arc  $(\mathbf{v}^{(t)}, \boldsymbol{\kappa}^{(t)})$  ensures that function  $G_h$  increases on every iteration. Indeed, let  $\boldsymbol{\varpi}$  be the plane tangent to the unit sphere on  $\mathbf{v}^{(t)}$  and let  $\boldsymbol{w}$  denote the unit-vector tangent to arc  $(\mathbf{v}^{(t)}, \boldsymbol{\kappa}^{(t)})$  on  $\mathbf{v}^{(t)}$ . Then, there exists  $\tau > 0$  such that,  $\boldsymbol{w} = \tau \mathbf{\Pi}_{\boldsymbol{\varpi}} \boldsymbol{\kappa}^{(t)}$ , and

$$\left\langle \boldsymbol{w}, \boldsymbol{\kappa}^{(t)} \right\rangle = \tau \left\langle \mathbf{\Pi}_{\boldsymbol{\varpi}} \boldsymbol{\kappa}^{(t)}, \boldsymbol{\kappa}^{(t)} \right\rangle = \tau \cos^2(\boldsymbol{\kappa}^{(t)}, \boldsymbol{\varpi}) > 0.$$

However, staying too close to the current starting point can impact the convergence speed of the algorithm to reach the maximum. We avoid that by using a one dimensional maximization function (e.g. Gauss-Newton) to find the maximum of  $G_h$  on the arc  $(\mathbf{v}^{(t)}, \boldsymbol{\kappa}^{(t)})$ , and take it as  $\mathbf{v}^{(t+1)}$ . We consider two possible generic iterations for the PING algorithm to 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

---

- 1: **while** not convergence **do**
  - 2:    $\boldsymbol{\kappa}^{(t)} \leftarrow \frac{\mathbf{\Pi}_{\operatorname{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}^{(t)})}{\left\| \mathbf{\Pi}_{\operatorname{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}^{(t)}) \right\|}$
  - 3:   Use a Newton-Raphson unidimensional maximization procedure to find the maximum of  $G_h(\mathbf{v})$  on the arc  $(\mathbf{v}^{(t)}, \boldsymbol{\kappa}^{(t)})$  and take it as  $\mathbf{v}^{(t+1)}$
  - 4:    $t \leftarrow t + 1$
  - 5: **end while**
- 

## Appendix C How SCGLR regularizes and shrinks

We first recall the general context given in Appendix A and Appendix B. We aim at calculating  $\mathbf{f} = \mathbf{X}\mathbf{u}$  the  $(h+1)$ -th component, subject to the constraint  $\mathbf{u}^T \mathbf{M}\mathbf{u} = 1$ ,

---

**Algorithm 2** Alternative PING algorithm
 

---

```

1:  $\mathbf{m} \leftarrow \frac{\Pi_{\text{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}^{(t)})}{\|\Pi_{\text{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}^{(t)})\|}$ 
2: while  $G_h(\mathbf{m}) < G_h(\mathbf{v}^{(t)})$  do
3:    $\mathbf{m} \leftarrow \frac{\mathbf{v}^{(t)} + \mathbf{m}}{\|\mathbf{v}^{(t)} + \mathbf{m}\|}$ 
4: end while
5:  $\mathbf{v}^{(t+1)} \leftarrow \mathbf{m}$ 
6:  $t \leftarrow t + 1$ 

```

---

where  $M$  is a Euclidean metric suitable for the variables in  $\mathbf{X}$ . Let  $\mathbf{F}^h = [\mathbf{f}^1, \dots, \mathbf{f}^h]$  be the matrix concatenating the previous components calculated in  $\mathbf{X}$  and  $\mathbf{A}_h = [\mathbf{F}^h, \mathbf{A}]$  the additional covariates.

The structural relevance  $\phi$  of the component, defined by Equation (A1), could be expressed as  $\phi(\mathbf{u}) = \|\mathbf{X}\mathbf{u}\|_{\mathbf{W}}^2 \tilde{\phi}(\mathbf{u})$ , where

$$\tilde{\phi}(\mathbf{u}) = \left( \sum_{j=1}^J \omega_j \left[ \left( \frac{\mathbf{X}\mathbf{u}}{\|\mathbf{X}\mathbf{u}\|_{\mathbf{W}}} \right)^T N_j \left( \frac{\mathbf{X}\mathbf{u}}{\|\mathbf{X}\mathbf{u}\|_{\mathbf{W}}} \right) \right]^l \right)^{1/l}.$$

The likelihood  $\psi_{\mathbf{A}_h}$  of the model fitted on  $\mathbf{f}$ , all the previous component and additional covariates is defined by Equation (A2). It is important to note that  $\tilde{\phi}$  and  $\psi_{\mathbf{A}_h}$  are zero-degree homogeneous functions of  $\mathbf{u}$ , that is, for all  $t \in \mathbb{R}$ ,  $\tilde{\phi}(t\mathbf{u}) = \tilde{\phi}(\mathbf{u})$  and  $\psi_{\mathbf{A}_h}(t\mathbf{u}) = \psi_{\mathbf{A}_h}(\mathbf{u})$ .

The current vector of component coefficients  $\mathbf{u}^\star$  is solution of the following optimization program

$$\begin{cases} \max_{\mathbf{u} \in \mathbb{R}^P} C_h(\mathbf{u}), \\ \text{s.t. } \mathbf{u}^T M \mathbf{u} = 1 \quad \text{and} \quad (\mathbf{F}^h)^T \mathbf{W} \mathbf{X} \mathbf{u} = \mathbf{0}, \end{cases} \quad (\text{C9})$$

where

$$C_h(\mathbf{u}) = \phi(\mathbf{u})^s \psi_{\mathbf{A}_h}(\mathbf{u})^{1-s} = \|\mathbf{X}\mathbf{u}\|_{\mathbf{W}}^{2s} \tau_h(\mathbf{u}),$$

where

$$\tau_h(\mathbf{u}) = \tilde{\phi}(\mathbf{u})^s \psi_{\mathbf{A}_h}(\mathbf{u})^{1-s}$$

and  $s \in ]0, 1]$ . Clearly,  $\tau_h$  is a zero-degree homogeneous function of  $\mathbf{u}$ .

Let  $\mathcal{E} = \{\mathbf{v} \in \mathbb{R}^P \mid \mathbf{f} = \mathbf{X}\mathbf{u}^\star = \mathbf{X}\mathbf{v}\}$  be the affine space of vectors giving the component  $\mathbf{f}$ . If we suppose that  $\mathbf{X}$  is full-rank, the vector of component coefficients  $\mathbf{u}^\star$  is the only vector giving the component  $\mathbf{f}$  (i.e.  $\mathcal{E} = \{\mathbf{u}^\star\}$ ). Thus, the equation  $\mathbf{X}\mathbf{v} = \mathbf{X}\mathbf{u}^\star$  implies that  $\mathbf{v} = \mathbf{u}^\star$ , and then  $\|\mathbf{v}\|_M = \|\mathbf{u}^\star\|_M$ . Now, assuming that  $\mathbf{X}$  is not full-rank, we have

**Lemma 1.** *For all  $\mathbf{v} \in \mathcal{E} \setminus \{\mathbf{u}^\star\}$ , we have  $\|\mathbf{v}\|_M > \|\mathbf{u}^\star\|_M$ .*

*Proof.* Since  $\mathbf{X}$  is not full-rank, there is an infinity of vectors giving the component  $\mathbf{f}$ . Let  $\mathbf{v} \neq \mathbf{u}^*$  be any one of them,  $\mathbf{f} = \mathbf{X}\mathbf{u}^* = \mathbf{X}\mathbf{v}$ . By setting  $\tilde{\mathbf{v}} = \mathbf{v}/\|\mathbf{v}\|_M$ , we have  $\|\tilde{\mathbf{v}}\|_M = 1$  and  $(\mathbf{F}^h)^T \mathbf{W} \mathbf{X} \tilde{\mathbf{v}} = \mathbf{0}$ . Since  $\mathbf{u}^*$  is the solution of the optimization program (C9), we obtain

$$C_h(\tilde{\mathbf{v}}) \leq C_h(\mathbf{u}^*) \Leftrightarrow \|\mathbf{X}\tilde{\mathbf{v}}\|_{\mathbf{W}}^{2s} \tau_h(\tilde{\mathbf{v}}) \leq \|\mathbf{X}\mathbf{u}^*\|_{\mathbf{W}}^{2s} \tau_h(\mathbf{u}^*).$$

Since  $\tau_h$  is zero-degree homogeneous, we have  $\tau_h(\tilde{\mathbf{v}}) = \tau_h(\mathbf{v}) = \tau_h(\mathbf{u}^*)$ . As  $C_h$  is positive and its maximum most generally non-zero, so is  $\tau_h$ , and we get

$$\begin{aligned} \|\mathbf{X}\tilde{\mathbf{v}}\|_{\mathbf{W}}^{2s} \leq \|\mathbf{X}\mathbf{u}^*\|_{\mathbf{W}}^{2s} &\Leftrightarrow \frac{\|\mathbf{X}\mathbf{v}\|_{\mathbf{W}}^{2s}}{\|\mathbf{v}\|_M^{2s}} \leq \frac{\|\mathbf{X}\mathbf{u}^*\|_{\mathbf{W}}^{2s}}{\|\mathbf{u}^*\|_M^{2s}} \\ &\Leftrightarrow \|\mathbf{v}\|_M \geq \|\mathbf{u}^*\|_M. \\ &\Leftrightarrow \|\mathbf{v}\|_M \geq 1. \end{aligned} \tag{C10}$$

Now, the affine space  $\mathcal{E}$  is tangent to the unit sphere at point  $\mathbf{u}^*$ . Indeed, if we suppose that  $\mathbf{v} \neq \mathbf{u}^*$  belongs to the intersection of  $\mathcal{E}$  with the unit sphere, the vector  $\mathbf{w} = (\mathbf{v} + \mathbf{u}^*)/2$  still belongs to  $\mathcal{E}$ . Moreover, the unit sphere being a strictly convex set, we have  $\|\mathbf{w}\|_M < 1$ . But this contradicts Equation (C10).

Since the affine space  $\mathcal{E}$  is tangent to the unit sphere,  $\mathbf{v}$  could be expressed as  $\mathbf{v} = \mathbf{u}^* + \Pi_{\text{span}[\mathbf{u}^*]^\perp}^M \mathbf{v}$ . We thus have

$$\|\mathbf{v}\|_M^2 = \left\| \mathbf{u}^* + \Pi_{\text{span}[\mathbf{u}^*]^\perp}^M \mathbf{v} \right\|_M^2 = \|\mathbf{u}^*\|_M^2 + \left\| \Pi_{\text{span}[\mathbf{u}^*]^\perp}^M \mathbf{v} \right\|_M^2.$$

Finally, since  $\mathbf{v} \neq \mathbf{u}^*$ , we get

$$\|\mathbf{v}\|_M^2 > \|\mathbf{u}^*\|_M^2 \Leftrightarrow \|\mathbf{v}\|_M > \|\mathbf{u}^*\|_M.$$

This concludes the proof. □

For the sake of visualization, Lemma 1 is represented by Figure C1.

Now, the linear predictor associated with response  $\mathbf{y}_k$  writes

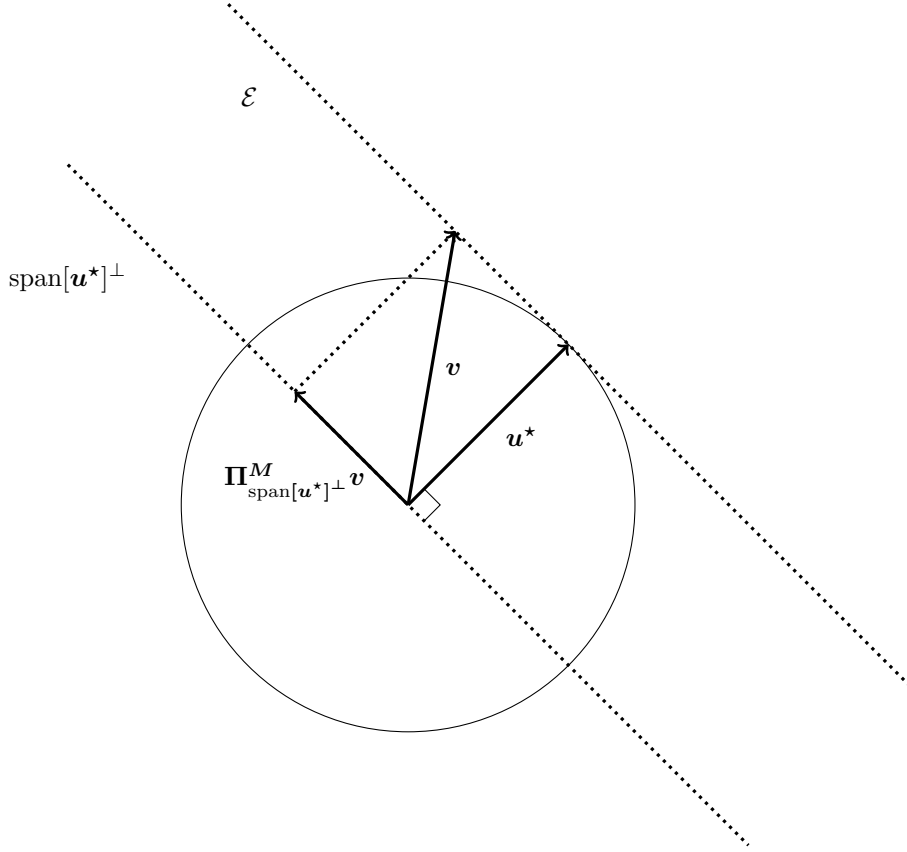
$$\boldsymbol{\eta}_k = (\mathbf{X}\mathbf{u}^*)\gamma_k + \mathbf{A}_h \boldsymbol{\delta}_k.$$

If  $\mathbf{X}$  is not full rank, the linear predictor could be expressed as

$$\boldsymbol{\eta}_k = (\mathbf{X}\mathbf{v})\gamma_k + \mathbf{A}_h \boldsymbol{\delta}_k,$$

where  $\mathbf{v} \neq \mathbf{u}^*$ . Denoting  $\boldsymbol{\beta}_k^{\mathbf{u}^*} = \mathbf{u}^* \gamma_k$  and  $\boldsymbol{\beta}_k^{\mathbf{v}} = \mathbf{v} \gamma_k$ , we have

**Theorem 2.** *For all  $\mathbf{v} \in \mathcal{E} \setminus \{\mathbf{u}^*\}$ , if  $\gamma_k \neq 0$  then  $\|\boldsymbol{\beta}_k^{\mathbf{v}}\|_M > \|\boldsymbol{\beta}_k^{\mathbf{u}^*}\|_M$ .*



**Fig. C1:** Expression of the vector  $v$  with respect to  $u^*$  and its orthogonal projection.

*Proof.* From Lemma 1, we have

$$\|v\|_M > \|u^*\|_M \Rightarrow |\gamma_k| \|v\|_M > |\gamma_k| \|u^*\|_M \Rightarrow \|\beta_k^v\|_M > \|\beta_k^{u^*}\|_M.$$

□

Informally, when  $X$  is not full rank, SCGLR gives the smallest vector of coefficients with respect to its component.

## Appendix D The EM algorithm

Consider the linearized model where the factors are unknown. We shall use the EM algorithm to estimate the parameters. The previous developments lead to the conditional linearized model

$$w_k = F\gamma_k + A\delta_k + Gb_k + \zeta_k,$$

where  $\mathbb{E}[\mathbf{w}_k | \mathbf{G}] = \mathbf{F}\boldsymbol{\gamma}_k + \mathbf{A}\boldsymbol{\delta}_k + \mathbf{G}\mathbf{b}_k$  and

$$\mathbb{V}[\mathbf{w}_k | \mathbf{G}] = \mathbb{V}[\boldsymbol{\zeta}_k] = \mathbf{W}_k^{-1} = \text{diag}(v_{nk}^{-1})_{n=1, \dots, N},$$

with  $v_{nk}^{-1} := a_{nk}(\phi_k)v_k(\mu_{nk})h'_k(\mu_{nk})^2$ ,  $a_{nk}$  and  $v_k$  being known functions and  $\phi_k$  being the dispersion parameter related to  $\mathbf{y}_k$ . The linearized model expressed row-wise writes

$$\mathbf{w}_n = \boldsymbol{\Gamma}^T \mathbf{f}_n + \boldsymbol{\Delta}^T \mathbf{a}_n + \mathbf{B}^T \mathbf{g}_n + \boldsymbol{\zeta}_n,$$

where  $\boldsymbol{\Gamma} = [\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_K]$ ,  $\boldsymbol{\Delta} = [\boldsymbol{\delta}_1, \dots, \boldsymbol{\delta}_K]$ ,  $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_K]$ , and where  $\mathbf{w}_n$ ,  $\mathbf{f}_n$ ,  $\mathbf{a}_n$  and  $\mathbf{g}_n$  are the vectors composed of the  $n$ th rows of matrices  $\boldsymbol{\mathcal{W}}$ ,  $\mathbf{F}$ ,  $\mathbf{A}$  and  $\mathbf{G}$  respectively. The expectation and the variance of  $\mathbf{w}_n$  are given by

$$\mathbb{E}[\mathbf{w}_n] = \boldsymbol{\Gamma}^T \mathbf{f}_n + \boldsymbol{\Delta}^T \mathbf{a}_n \quad \text{and} \quad \mathbb{V}[\mathbf{w}_n] = \mathbf{B}^T \mathbf{B} + \boldsymbol{\Upsilon}_n^{-1},$$

where

$$\boldsymbol{\Upsilon}_n^{-1} = \text{diag}(v_{nk}^{-1})_{k=1, \dots, K}.$$

Denoting  $\boldsymbol{\Theta} = \{\boldsymbol{\Gamma}, \boldsymbol{\Delta}, \mathbf{B}\}$  the set of parameters, the complete log-likelihood writes

$$\begin{aligned} l(\boldsymbol{\Theta}; \boldsymbol{\mathcal{W}}, \mathbf{G}) &= \ln(L(\boldsymbol{\mathcal{W}}, \mathbf{G}; \boldsymbol{\Theta})) \\ &= \sum_{n=1}^N \ln(L(\mathbf{w}_n | \mathbf{g}_n; \boldsymbol{\Theta})) + \ln(L(\mathbf{g}_n; \boldsymbol{\Theta})) \\ &= \sum_{n=1}^N \left[ -\ln\left((2\pi)^{K/2} \det(\boldsymbol{\Upsilon}_n^{-1})^{1/2}\right) \right. \\ &\quad \left. - \frac{1}{2} (\mathbf{w}_n - \boldsymbol{\Gamma}^T \mathbf{f}_n - \boldsymbol{\Delta}^T \mathbf{a}_n - \mathbf{B}^T \mathbf{g}_n)^T \boldsymbol{\Upsilon}_n (\mathbf{w}_n - \boldsymbol{\Gamma}^T \mathbf{f}_n - \boldsymbol{\Delta}^T \mathbf{a}_n - \mathbf{B}^T \mathbf{g}_n) \right. \\ &\quad \left. - \ln\left((2\pi)^{J/2}\right) - \frac{1}{2} \mathbf{g}_n^T \mathbf{g}_n \right] \\ &= -\frac{1}{2} \sum_{n=1}^N \left[ \sum_{k=1}^K \ln(v_{nk}^{-1}) + \mathbf{g}_n^T \mathbf{g}_n + (K+J) \ln(2\pi) \right. \\ &\quad \left. + \sum_{k=1}^K v_{nk} (w_{nk} - \mathbf{f}_n^T \boldsymbol{\gamma}_k - \mathbf{a}_n^T \boldsymbol{\delta}_k - \mathbf{g}_n^T \mathbf{b}_k)^2 \right]. \end{aligned}$$

## D.1 The expectation (E) step

We first calculate the expectation of the complete log-likelihood conditional on the data  $\boldsymbol{\mathcal{W}}$

$$\mathbb{E}[l(\boldsymbol{\Theta}; \boldsymbol{\mathcal{W}}, \mathbf{G}) | \boldsymbol{\mathcal{W}}; \boldsymbol{\Theta}'] =$$

$$\sum_{n=1}^N \int \ln(L(\mathbf{w}_n | \mathbf{g}_n; \Theta) L(\mathbf{g}_n; \Theta) L(\mathbf{g}_n | \mathbf{w}_n; \Theta')) d\mathbf{g}_n.$$

We first need to find the law of  $\mathbf{g}_n | \mathbf{w}_n$ . Assuming  $\mathbf{w}_n$  is approximately Gaussian, the random vector  $(\mathbf{w}_n^T, \mathbf{g}_n^T)^T$  is Gaussian, and

$$\begin{pmatrix} \mathbf{w}_n \\ \mathbf{g}_n \end{pmatrix} \sim \mathcal{N} \left( \begin{pmatrix} \Gamma^T \mathbf{f}_n + \Delta^T \mathbf{a}_n \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{B}^T \mathbf{B} + \Upsilon_n^{-1} & \mathbf{B}^T \\ \mathbf{B} & \mathbf{I}_J \end{pmatrix} \right).$$

Using the conditioning rule of the multivariate Gaussian, we get

$$\mathbf{g}_n | \mathbf{w}_n \sim \mathcal{N}(\alpha_n (\mathbf{w}_n - \Gamma^T \mathbf{f}_n - \Delta^T \mathbf{a}_n), \mathbf{I}_J - \alpha_n \mathbf{B}^T),$$

where  $\alpha_n = \mathbf{B}(\mathbf{B}^T \mathbf{B} + \Upsilon_n^{-1})^{-1}$ . The moments of the random variable  $\mathbf{g}_n | \mathbf{w}_n$  are given by

$$\begin{aligned} \tilde{\mathbf{g}}_n &:= \mathbb{E}[\mathbf{g}_n | \mathbf{w}_n; \Theta] \\ &= \alpha_n (\mathbf{w}_n - \Gamma^T \mathbf{f}_n - \Delta^T \mathbf{a}_n) \end{aligned}$$

and

$$\begin{aligned} \tilde{\mathbf{R}}_n &:= \mathbb{E}[\mathbf{g}_n \mathbf{g}_n^T | \mathbf{w}_n; \Theta] \\ &= \mathbb{V}[\mathbf{g}_n | \mathbf{w}_n; \Theta] + \mathbb{E}[\mathbf{g}_n | \mathbf{w}_n; \Theta] \mathbb{E}[\mathbf{g}_n | \mathbf{w}_n; \Theta]^T \\ &= \mathbf{I}_J - \alpha_n \mathbf{B}^T + \tilde{\mathbf{g}}_n \tilde{\mathbf{g}}_n^T. \end{aligned}$$

Finally, we get the following explicit form of the expectation of the complete log-likelihood

$$\begin{aligned} &\mathbb{E}[l(\Theta; \mathcal{W}, \mathcal{G}) | \mathcal{W}, \Theta'] \\ &= -\frac{1}{2} \sum_{n=1}^N \left\{ (K+J) \ln(2\pi) + \sum_{k=1}^K \ln(v_{nk}^{-1}) + \right. \\ &\quad \left. \mathbb{E} \left[ \mathbf{g}_n^T \mathbf{g}_n + \sum_{k=1}^K v_{nk} (w_{nk} - \mathbf{f}_n^T \boldsymbol{\gamma}_k - \mathbf{a}_n^T \boldsymbol{\delta}_k - \mathbf{g}_n^T \mathbf{b}_k)^2 | \mathbf{w}_n; \Theta' \right] \right\} \\ &= -\frac{1}{2} \sum_{n=1}^N \left\{ (K+J) \ln(2\pi) + \sum_{k=1}^K \ln(v_{nk}^{-1}) + \mathbb{E}[\mathbf{g}_n^T \mathbf{g}_n | \mathbf{w}_n; \Theta'] + \right. \\ &\quad \left. \mathbb{E} \left[ \sum_{k=1}^K v_{nk} \left( (w_{nk} - \mathbf{f}_n^T \boldsymbol{\gamma}_k - \mathbf{a}_n^T \boldsymbol{\delta}_k)^2 + \mathbf{b}_k^T (\mathbf{g}_n \mathbf{g}_n^T) \mathbf{b}_k - \right. \right. \right. \\ &\quad \left. \left. \left. 2(w_{nk} - \mathbf{f}_n^T \boldsymbol{\gamma}_k - \mathbf{a}_n^T \boldsymbol{\delta}_k) \mathbf{g}_n^T \mathbf{b}_k \right) | \mathbf{w}_n; \Theta' \right] \right\} \end{aligned}$$



$$\begin{aligned}
&= -\frac{1}{2} \sum_{n=1}^N \left\{ (K+J) \ln(2\pi) + \sum_{k=1}^K \ln(v_{nk}^{-1}) + \mathbb{E}[\mathbf{g}_n^T \mathbf{g}_n \mid \mathbf{w}_n; \Theta'] + \right. \\
&\quad \left. \sum_{k=1}^K v_{nk} \left[ (w_{nk} - \mathbf{f}_n^T \boldsymbol{\gamma}_k - \mathbf{a}_n^T \boldsymbol{\delta}_k)^2 + \mathbf{b}_k^T \tilde{\mathbf{R}}_n \mathbf{b}_k - \right. \right. \\
&\quad \left. \left. 2(w_{nk} - \mathbf{f}_n^T \boldsymbol{\gamma}_k - \mathbf{a}_n^T \boldsymbol{\delta}_k) \tilde{\mathbf{g}}_n^T \mathbf{b}_k \right] \right\} \\
&= -\frac{1}{2} \left\{ N(K+J) \ln(2\pi) + \sum_{n=1}^N \sum_{k=1}^K \ln(v_{nk}^{-1}) + \sum_{n=1}^N \mathbb{E}[\mathbf{g}_n^T \mathbf{g}_n \mid \mathbf{w}_n; \Theta'] + \right. \\
&\quad \left. \sum_{k=1}^K \left[ \|\mathbf{w}_k - \mathbf{F} \boldsymbol{\gamma}_k - \mathbf{A} \boldsymbol{\delta}_k\|_{\mathbf{W}_k}^2 + \mathbf{b}_k^T \left( \sum_{n=1}^N v_{nk} \tilde{\mathbf{R}}_n \right) \mathbf{b}_k - \right. \right. \\
&\quad \left. \left. 2(\tilde{\mathbf{G}} \mathbf{b}_k)^T \mathbf{W}_k (\mathbf{w}_k - \mathbf{F} \boldsymbol{\gamma}_k - \mathbf{A} \boldsymbol{\delta}_k) \right] \right\},
\end{aligned}$$

where the rows of the matrix  $\tilde{\mathbf{G}}$  are composed of the  $\tilde{\mathbf{g}}_n^T$ 's.

## D.2 The maximization (M) step

This step maximizes the conditional expectation of the complete log-likelihood with respect to  $\Theta$ , subject to the upper triangular constraint on matrix  $\mathbf{B}$ . However, for all  $k$ , the parameters  $\boldsymbol{\gamma}_k$  and  $\boldsymbol{\delta}_k$  are not concerned by the constraint. Denoting  $\boldsymbol{\beta}_k^T = (\boldsymbol{\gamma}_k^T, \boldsymbol{\delta}_k^T)$  and  $\tilde{\mathbf{X}} = [\mathbf{F}, \mathbf{A}]$ , the first order conditions of the maximization yield

$$\begin{aligned}
&\nabla_{\boldsymbol{\beta}_k} \mathbb{E}[l(\Theta; \mathcal{W}, \mathbf{G}) \mid \mathcal{W}, \Theta'] = 0 \\
&\Leftrightarrow \nabla_{\boldsymbol{\beta}_k} \left\{ \|\mathbf{w}_k - \tilde{\mathbf{X}} \boldsymbol{\beta}_k\|_{\mathbf{W}_k}^2 - 2(\tilde{\mathbf{G}} \mathbf{b}_k)^T \mathbf{W}_k (\mathbf{w}_k - \tilde{\mathbf{X}} \boldsymbol{\beta}_k) \right\} = 0 \\
&\Leftrightarrow \tilde{\mathbf{X}}^T \mathbf{W}_k (\mathbf{w}_k - \tilde{\mathbf{X}} \boldsymbol{\beta}_k) - \tilde{\mathbf{X}}^T \mathbf{W}_k \tilde{\mathbf{G}} \mathbf{b}_k = 0 \\
&\Leftrightarrow \tilde{\mathbf{X}}^T \mathbf{W}_k \tilde{\mathbf{X}} \boldsymbol{\beta}_k = \tilde{\mathbf{X}}^T \mathbf{W}_k (\mathbf{w}_k - \tilde{\mathbf{G}} \mathbf{b}_k) \\
&\Rightarrow \hat{\boldsymbol{\beta}}_k = \left( \tilde{\mathbf{X}}^T \mathbf{W}_k \tilde{\mathbf{X}} \right)^{-1} \tilde{\mathbf{X}}^T \mathbf{W}_k (\mathbf{w}_k - \tilde{\mathbf{G}} \mathbf{b}_k).
\end{aligned}$$

If a response is drawn from a Gaussian law  $\mathbf{y}_k \sim \mathcal{N}(\tilde{\mathbf{X}} \boldsymbol{\beta}_k, \sigma_k^2 \mathbf{I}_N)$ , the residual variance  $\sigma_k^2$  must be estimated. In that case,

$$\begin{aligned}
&\nabla_{\sigma_k^2} \mathbb{E}[l(\Theta; \mathcal{W}, \mathbf{G}) \mid \mathcal{W}, \Theta'] = 0 \\
&\Leftrightarrow \nabla_{\sigma_k^2} \left\{ N \ln(\sigma_k^2) + \frac{1}{\sigma_k^2} \left[ \|\mathbf{w}_k - \tilde{\mathbf{X}} \boldsymbol{\beta}_k\|^2 + \mathbf{b}_k^T \left( \sum_{n=1}^N \tilde{\mathbf{R}}_n \right) \mathbf{b}_k \right. \right. \\
&\quad \left. \left. - 2(\tilde{\mathbf{G}} \mathbf{b}_k)^T (\mathbf{w}_k - \tilde{\mathbf{X}} \boldsymbol{\beta}_k) \right] \right\} = 0
\end{aligned}$$

$$\begin{aligned}
&\Leftrightarrow N - \frac{1}{\sigma_k^2} \left\{ \left\| \mathbf{w}_k - \tilde{\mathbf{X}} \beta_k \right\|^2 + \mathbf{b}_k^T \left( \sum_{n=1}^N \tilde{\mathbf{R}}_n \right) \mathbf{b}_k \right. \\
&\quad \left. - 2 \left( \tilde{\mathbf{G}} \mathbf{b}_k \right)^T \left( \mathbf{w}_k - \tilde{\mathbf{X}} \beta_k \right) \right\} = 0 \\
&\Rightarrow \hat{\sigma}_k^2 = \frac{1}{N} \left\{ \left\| \mathbf{w}_k - \tilde{\mathbf{X}} \beta_k \right\|^2 + \mathbf{b}_k^T \left( \sum_{n=1}^N \tilde{\mathbf{R}}_n \right) \mathbf{b}_k \right. \\
&\quad \left. - 2 \left( \tilde{\mathbf{G}} \mathbf{b}_k \right)^T \left( \mathbf{w}_k - \tilde{\mathbf{X}} \beta_k \right) \right\}.
\end{aligned}$$

Now, we need to estimate the vector  $\mathbf{b}_k$  under the upper triangular constraint. For each  $k = 1, \dots, J$ , let  $\mathbf{b}_k^T = (\mathbf{b}_{1:k,k}^T, \mathbf{0}^T)$  be the regression parameters, where  $\mathbf{b}_{1:k,k}^T = (b_{1k}, \dots, b_{kk})$  is a vector of length  $k$  to be estimated and  $\mathbf{0}$  is a null vector of length  $(J - k)$ . In this case, we define  $(\tilde{\mathbf{R}}_n)_{1:k,1:k}$  as the sub-matrix of size  $k \times k$  of  $\tilde{\mathbf{R}}_n$  and  $\tilde{\mathbf{G}}_{1:k}$  as the matrix composed by the first  $k$  columns of  $\tilde{\mathbf{G}}$ . The maximization yields

$$\begin{aligned}
&\nabla_{\mathbf{b}_{1:k,k}} \mathbb{E}[l(\Theta; \mathcal{W}, \mathbf{G}) \mid \mathcal{W}, \Theta'] = 0 \\
&\Leftrightarrow \nabla_{\mathbf{b}_{1:k,k}} \left\{ \mathbf{b}_{1:k,k}^T \left[ \sum_{n=1}^N v_{nk} \left( \tilde{\mathbf{R}}_n \right)_{1:k,1:k} \right] \mathbf{b}_{1:k,k} \right. \\
&\quad \left. - 2 \left( \tilde{\mathbf{G}}_{1:k} \mathbf{b}_{1:k,k} \right)^T \mathbf{W}_k \left( \mathbf{w}_k - \tilde{\mathbf{X}} \beta_k \right) \right\} = 0 \\
&\Leftrightarrow \left( \tilde{\mathbf{G}}_{1:k} \right)^T \mathbf{W}_k \left( \mathbf{w}_k - \tilde{\mathbf{X}} \beta_k \right) - \left[ \sum_{n=1}^N v_{nk} \left( \tilde{\mathbf{R}}_n \right)_{1:k,1:k} \right] \mathbf{b}_{1:k,k} = 0 \\
&\Rightarrow \hat{\mathbf{b}}_{1:k,k} = \left[ \sum_{n=1}^N v_{nk} \left( \tilde{\mathbf{R}}_n \right)_{1:k,1:k} \right]^{-1} \left( \tilde{\mathbf{G}}_{1:k} \right)^T \mathbf{W}_k \left( \mathbf{w}_k - \tilde{\mathbf{X}} \beta_k \right).
\end{aligned}$$

Likewise, for  $k = J + 1, \dots, K$ , the estimate  $\hat{\mathbf{b}}_k$  is given by

$$\hat{\mathbf{b}}_k = \left[ \sum_{n=1}^N v_{nk} \tilde{\mathbf{R}}_n \right]^{-1} \tilde{\mathbf{G}}^T \mathbf{W}_k \left( \mathbf{w}_k - \tilde{\mathbf{X}} \beta_k \right).$$

### D.3 The algorithm

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

## Appendix E The overall F-SCGLR algorithm

Algorithm 4 consists in alternating the following steps: (i) Given the current set of parameters, calculate all the components of all the themes iteratively through the

---

**Algorithm 3** The EM algorithm applied to factor models
 

---

```

1: while not convergence do
2:   Expectation step
3:   for  $n = 1, \dots, N$  do
4:      $\alpha_n^{(t+1)} = \mathbf{B}^{(t)} (\mathbf{B}^{(t)T} \mathbf{B}^{(t)} + \mathbf{\Upsilon}_n^{-1})^{-1}$ 
5:      $\tilde{\mathbf{g}}_n^{(t+1)} = \alpha_n^{(t+1)} (\mathbf{w}_n - \mathbf{\Gamma}^{(t)T} \mathbf{f}_n - \mathbf{\Delta}^{(t)T} \mathbf{a}_n)$ 
6:      $\tilde{\mathbf{R}}_n^{(t+1)} = \mathbf{I}_J - \alpha_n^{(t+1)} \mathbf{B}^{(t)T} + \tilde{\mathbf{g}}_n^{(t+1)} \tilde{\mathbf{g}}_n^{(t+1)T}$ 
7:   end for
8:   Maximization step
9:   for  $k = 1, \dots, K$  do
10:     $\beta_k^{(t+1)} = (\tilde{\mathbf{X}}^T \mathbf{W}_k \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \mathbf{W}_k (\mathbf{w}_k - \tilde{\mathbf{G}}^{(t+1)} \mathbf{b}_k^{(t)})$ 
11:   end for
12:   if Gaussian then
13:      $\sigma_k^{2(t+1)} = \frac{1}{N} \left\{ \left\| \mathbf{w}_k - \tilde{\mathbf{X}} \beta_k^{(t+1)} \right\|^2 + \mathbf{b}_k^{(t)T} \left( \sum_{n=1}^N \tilde{\mathbf{R}}_n^{(t+1)} \right) \mathbf{b}_k^{(t)} - \right.$ 
14:        $\left. 2 \left( \tilde{\mathbf{G}}^{(t+1)} \mathbf{b}_k^{(t)} \right)^T \left( \mathbf{w}_k - \tilde{\mathbf{X}} \beta_k^{(t+1)} \right) \right\}$ 
15:   end if
16:   if  $k \leq J$  then
17:      $\mathbf{b}_{1:k,k}^{(t+1)} =$ 
18:        $\left[ \sum_{n=1}^N v_{nk} \left( \tilde{\mathbf{R}}_n^{(t+1)} \right)_{1:k,1:k} \right]^{-1} \left( \tilde{\mathbf{G}}_{1:k}^{(t+1)} \right)^T \mathbf{W}_k \left( \mathbf{w}_k - \tilde{\mathbf{X}} \beta_k^{(t+1)} \right)$ 
19:   else
20:      $\mathbf{b}_k^{(t+1)} = \left[ \sum_{n=1}^N v_{nk} \tilde{\mathbf{R}}_n^{(t+1)} \right]^{-1} \tilde{\mathbf{G}}^{(t+1)T} \mathbf{W}_k \left( \mathbf{w}_k - \tilde{\mathbf{X}} \beta_k^{(t+1)} \right)$ 
21:   end if
22:    $t \leftarrow t + 1$ 
23: end while

```

---

PING algorithm. (ii) Given the current components, calculate the adjusted dependent variables of the linearized model and their variance matrix. (iii) Given the adjusted dependent variables, estimate the factor model parameters through the EM algorithm.

## Appendix F Identification of the true model

Table F1 sums up the identification diagnostics on a cross-product grid.

## Appendix G Figures of the residual correlation matrices

Figure G2 shows the residual correlation matrices for the three values of  $\sigma_B^2$  presented in Section 4.1.3. Figure G3 shows the residual correlation matrices obtained for the three values of  $K$  presented in Section 4.2.1.

---

**Algorithm 4** The F-SCGLR algorithm
 

---

```

1: while not convergence do
2:   Compute the components through the PING algorithm
3:    $\forall r = 1, \dots, R, \forall h = 1, \dots, H_r, \mathbf{f}_r^{h(t+1)} = \mathbf{X}_r \mathbf{u}_r^{h(t+1)}$ 
4:   Compute the adjusted dependent variables through the IRLS algorithm
5:    $\boldsymbol{\eta}_k^{(t+1)} = \mathbf{F}^{(t+1)} \boldsymbol{\gamma}_k^{(t)} + \mathbf{A} \boldsymbol{\delta}_k^{(t)} + \mathbf{G} \mathbf{b}_k^{(t)}$ 
6:    $\mu_{nk}^{(t+1)} = h_k^{-1} \left( \eta_{nk}^{(t+1)} \right), \forall n = 1, \dots, N$ 
7:    $w_{nk}^{(t+1)} = \eta_{nk}^{(t+1)} + h'_k \left( \mu_{nk}^{(t+1)} \right) \left( y_{nk} - \mu_{nk}^{(t+1)} \right), \forall n = 1, \dots, N$ 
8:    $\mathbf{W}_k^{(t+1)} = \text{diag} \left( \left[ a_{nk}(\phi_k) v_k \left( \mu_{nk}^{(t+1)} \right) h'_k \left( \mu_{nk}^{(t+1)} \right)^2 \right]^{-1} \right)_{n=1, \dots, N}$ 
9:   Compute the model parameter through the EM algorithm
10:   $\Theta^{(t+1)} = \underset{\Theta}{\text{argmax}} l(\Theta^{(t)}; \mathcal{W})$ 
11:  Increment
12:   $t \leftarrow t + 1$ 
13: end while

```

---

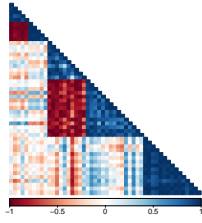
**Table F1:** Mean values of BIC over five hundred samples for  $(H_1, H_2) \in \{1, 2, 3, 4\}^2$  and  $J$  ranging from 0 to 5. The lowest values are in bold font.

$J = 0$					$J = 1$				
$H_2/H_1$	1	2	3	4	$H_2/H_1$	1	2	3	4
1	44808	42432	39535	37904	1	40616	36292	50764	28050
2	38240	36456	35118	33532	2	49517	<b>24509</b>	57579	38288
3	36405	34927	33718	32797	3	52801	26072	62440	37556
4	36028	34423	33182	<b>32202</b>	4	25759	25976	28371	29019
$J = 2$					$J = 3$				
$H_2/H_1$	1	2	3	4	$H_2/H_1$	1	2	3	4
1	23756	24077	22492	23051	1	20201	16324	16910	16733
2	21206	<b>20065</b>	22313	22777	2	18917	<b>15515</b>	15685	15881
3	20826	20899	21136	21718	3	19006	15533	15763	16042
4	21084	20487	20369	20405	4	19077	15959	16115	16305
$J = 4$					$J = 5$				
$H_2/H_1$	1	2	3	4	$H_2/H_1$	1	2	3	4
1	18661	<b>16837</b>	16941	17165	1	16436	<b>16034</b>	16217	16342
2	17014	16888	16979	17363	2	16516	16103	16181	16585
3	17284	16971	17149	17337	3	16487	16542	16390	16816
4	17435	17071	17261	17634	4	16690	16706	16892	18027

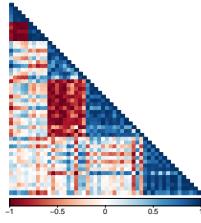
## Appendix H Additional simulation studies

### H.1 Additional simulation for Section 4.1.3

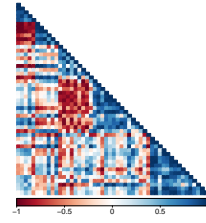
The Tables summing up the results for  $\sigma_B^2 = 0.2$  and  $\sigma_B^2 = 0.3$  are presented in Table H2 and Table H3.



(a) Residual correlation matrix for  $\sigma_B^2 = 0.1$

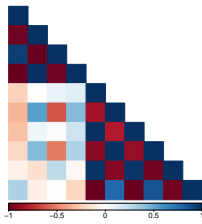


(b) Residual correlation matrix for  $\sigma_B^2 = 0.2$

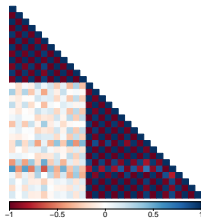


(c) Residual correlation matrix for  $\sigma_B^2 = 0.3$

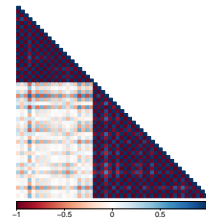
**Fig. G2:** Heatmap of the residual correlation matrices for various values of  $\sigma_B^2$ . The color intensity (irrespective of the color itself) reveals three response clusters, each gathering responses having a high residual correlation in absolute value.



(a) Residual correlation matrix for  $K = 10$



(b) Residual correlation matrix for  $K = 30$



(c) Residual correlation matrix for  $K = 50$

**Fig. G3:** Heatmap of the residual correlation matrices for various values of  $K$ . The color intensity (irrespective of the color itself) reveals three response clusters, each gathering responses having a high residual correlation in absolute value.

## H.2 Additional simulation for Section 4.2

The Tables summing up the results for Gaussian, Poisson and Bernoulli distribution with random slopes are presented in Table H4, Table H5 and Table H6. The performances of **gllvm** with and without random slopes did not prove much different except for the computation time, which was longer with random slopes. In Gaussian and Bernoulli cases, F-SCGLR exhibits better performances.

## References

- Bartholomew DJ, Knott M, Moustaki I (2011) Latent variable models and factor analysis: A unified approach, Third Edition. John Wiley & Sons
- Bry X, Verron T (2015) THEME: THEmatic Model Exploration through multiple co-structure maximization. J Chemom 29:637–647. <https://doi.org/10.1002/cem.2759>

**Table H2:** Mean values of RI, ARI and square correlation over five hundred samples with  $\sigma_B^2 = 0.2$ ,  $s \in \{0.1, 0.3, 0.5\}$  and  $l \in \{1, 2, 3, 4, 7, 10\}$ .

$s$	$l$	RI	ARI	$\rho^2(\xi_{1,\cdot})$	$\rho^2(\xi_{2,\cdot})$	$\rho^2(\xi_{3,\cdot})$	$\rho^2(\xi_{4,\cdot})$
0.1	1	0.749	0.369	0.965	0.934	0.836	0.869
	2	0.740	0.339	0.989	0.970	0.901	0.938
	3	0.740	0.339	0.983	0.969	0.909	0.943
	4	0.737	0.330	0.984	0.970	0.910	0.943
	7	0.738	0.334	0.978	0.964	0.899	0.950
	10	0.738	0.334	0.969	0.959	0.899	0.950
0.3	1	0.746	0.358	0.975	0.940	0.717	0.735
	2	0.746	0.359	0.993	0.972	0.920	0.932
	3	0.743	0.352	0.986	0.972	0.901	0.957
	4	0.739	0.339	0.979	0.971	0.884	0.962
	7	0.742	0.350	0.976	0.970	0.868	0.957
	10	0.742	0.349	0.974	0.970	0.864	0.957
0.5	1	0.743	0.359	0.975	0.939	0.715	0.659
	2	0.743	0.350	0.993	0.972	0.911	0.940
	3	0.741	0.345	0.986	0.973	0.881	0.969
	4	0.740	0.343	0.981	0.974	0.863	0.967
	7	0.738	0.338	0.978	0.973	0.859	0.962
	10	0.738	0.339	0.974	0.968	0.859	0.962

Bry X, Trottier C, Verron T, et al (2013) Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm. *J Multivar Anal* 119:47–60. <https://doi.org/10.1016/j.jmva.2013.03.013>

Bry X, Simac T, El Ghachi SE, et al (2020a) Bridging data exploration and modeling in event-history analysis: the supervised-component Cox regression. *Math Popul Stud* 27(3):139–174. <https://doi.org/10.1080/08898480.2018.1553413>

Bry X, Trottier C, Mortier F, et al (2020b) Component-based regularization of a multivariate GLM with a thematic partitioning of the explanatory variables. *Stat Modelling* 20(1):96–119. <https://doi.org/10.1177/1471082X18810114>

Chauvet J, Trottier C, Bry X (2019) Component-Based Regularization of Multivariate Generalized Linear Mixed Models. *J Comput Graph Stat* 28(4):909–920. <https://doi.org/10.1080/10618600.2019.1598870>

Cox MAA, Cox TF (2008) Multidimensional scaling. In: *Handbook of data visualization*. Springer, p 315–347

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* 39:1–22. <https://doi.org/10.1111/j.2517-6161.1977.tb01600.x>

Duflot R, San-Cristobal M, Andrieu E, et al (2022) Farming intensity indirectly reduces crop yield through negative effects on agrobiodiversity and key ecological functions. *Agric Ecosyst Environ* 326:107810. <https://doi.org/10.1016/j.agee.2021.107810>

**Table H3:** Mean values of RI, ARI and square correlation over five hundred samples with  $\sigma_B^2 = 0.3$ ,  $s \in \{0.1, 0.3, 0.5\}$  and  $l \in \{1, 2, 3, 4, 7, 10\}$ .

$s$	$l$	RI	ARI	$\rho^2(\xi_{1,\cdot})$	$\rho^2(\xi_{2,\cdot})$	$\rho^2(\xi_{3,\cdot})$	$\rho^2(\xi_{4,\cdot})$
0.1	1	0.674	0.154	0.963	0.931	0.817	0.867
	2	0.673	0.155	0.988	0.969	0.892	0.935
	3	0.675	0.162	0.985	0.969	0.907	0.942
	4	0.676	0.165	0.986	0.970	0.907	0.953
	7	0.674	0.153	0.981	0.968	0.888	0.955
	10	0.682	0.180	0.981	0.968	0.873	0.954
0.3	1	0.668	0.146	0.974	0.938	0.711	0.722
	2	0.668	0.143	0.993	0.971	0.923	0.935
	3	0.667	0.157	0.987	0.972	0.898	0.958
	4	0.671	0.156	0.981	0.970	0.875	0.957
	7	0.671	0.153	0.973	0.970	0.843	0.953
	10	0.670	0.157	0.970	0.970	0.842	0.956
0.5	1	0.672	0.154	0.975	0.938	0.719	0.662
	2	0.670	0.149	0.994	0.972	0.924	0.945
	3	0.670	0.158	0.984	0.973	0.875	0.961
	4	0.670	0.158	0.981	0.974	0.850	0.956
	7	0.670	0.165	0.974	0.973	0.837	0.954
	10	0.671	0.168	0.972	0.973	0.836	0.950

**Table H4:** Mean values of RI, ARI, Procrustes error of the factors  $\mathbf{G}$  and their loadings  $\mathbf{B}$ , latent dimension recovery, root mean square error and computation time over five hundred samples with  $N = 200$  and  $K = 30$  for the Gaussian distribution with random slopes.

	gllvm-VA		gllvm-LA	
	randomB = "LV"	randomB = "P"	randomB = "LV"	randomB = "P"
RI	0.966	0.872	0.953	0.962
ARI	0.931	0.747	0.906	0.924
Proc $B$	0.081	0.204	0.038	0.041
Proc $G$	0.103	0.191	0.160	0.191
$\rho^2(\xi_{1,\cdot})$	0.852	0.742	0.829	0.832
$\rho^2(\xi_{2,\cdot})$	0.707	0.419	0.796	0.717
$RMSEA$	0.577	0.650	0.598	0.609
Time	116.4	75.04	305.4	384.5

**Table H5:** Mean values of RI, ARI, Procrustes error of the factors  $\mathbf{G}$  and their loadings  $\mathbf{B}$ , latent dimension recovery, root mean square error and computation time over five hundred samples with  $N = 200$  and  $K = 30$  for the Poisson distribution with random slopes.

	gllvm-VA		gllvm-LA	
	randomB = "LV"	randomB = "P"	randomB = "LV"	randomB = "P"
RI	0.772	0.840	0.877	0.913
ARI	0.554	0.686	0.756	0.827
Proc $B$	0.571	0.299	0.139	0.363
Proc $G$	0.482	0.346	0.250	0.253
$\rho^2(\xi_{1,\cdot})$	0.670	0.634	0.775	0.811
$\rho^2(\xi_{2,\cdot})$	0.641	0.488	0.770	0.685
$RMSEA$	2.906	1.299	0.497	0.761
Time	176.5	119.3	643.7	518.8

**Table H6:** Mean values of RI, ARI, Procrustes error of the factors  $\mathbf{G}$  and their loadings  $\mathbf{B}$ , latent dimension recovery, root mean square error and computation time over five hundred samples with  $N = 200$  and  $K = 30$  for the Bernoulli distribution with random slopes.

	gllvm-EVA		gllvm-LA	
	randomB = "LV"	randomB = "P"	randomB = "LV"	randomB = "P"
RI	0.880	0.909	0.942	0.928
ARI	0.763	0.817	0.884	0.857
Proc $B$	0.646	0.578	0.494	0.541
Proc $G$	0.291	0.280	0.252	0.255
$\rho^2(\xi_{1,\cdot})$	0.864	0.821	0.808	0.791
$\rho^2(\xi_{2,\cdot})$	0.671	0.554	0.758	0.651
$RMSE_A$	1.162	1.145	0.783	0.788
Time	13.03	12.89	486.5	522.0

107810

Dunstan PK, Foster SD, Hui FK, et al (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. <https://doi.org/10.1007/s13253-013-0146-x>

Geweke J, Zhou G (1996) Measuring the pricing error of the arbitrage pricing theory. *Rev Financ Stud* 9(2):557–587. <https://doi.org/10.1093/rfs/9.2.557>

Gibaud J, Bry X, Trottier C, et al (2024) Response mixture models based on supervised components: Clustering floristic taxa. *Statistical Modelling* 24(3):226–244. <https://doi.org/10.1177/1471082X221115525>

Hoerl AE, Kennard RW (1970) Ridge regression: applications to nonorthogonal problems. *Technometrics* 12(1):69–82. <https://doi.org/10.1080/00401706.1970.10488635>

Hubert L, Arabie P (1985) Comparing partitions. *J Classif* 2(1):193–218. <https://doi.org/10.1007/BF01908075>

Hui FK (2016) boral—Bayesian ordination and regression analysis of multivariate abundance data in R. *Methods Ecol Evol* 7:744–750. <https://doi.org/10.1111/2041-210X.12514>

Hui FK (2017) Model-based simultaneous clustering and ordination of multivariate abundance data in ecology. *Comput Stat Data Anal* 105:1–10. <https://doi.org/10.1016/j.csda.2016.07.008>

Hui FK, Taskinen S, Pledger S, et al (2015) Model-based approaches to unconstrained ordination. *Methods Ecol Evol* 6:399–411. <https://doi.org/10.1111/2041-210X.12236>



- Hui FK, Warton DI, Ormerod JT, et al (2017) Variational approximations for generalized linear latent variable models. *J Comput Graph Stat* 26(1):35–43. <https://doi.org/10.1080/10618600.2016.1164708>
- Jöreskog KG (1969) A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika* 34(2):183–202
- Kassambara A (2017) Package ‘factoextra’. <http://www.sthdat.com/english/rpkgs/factoextra>
- Korhonen P, Hui FK, Niku J, et al (2023) Fast and universal estimation of latent variable models using extended variational approximations. *Stat Comput* 33(26). <https://doi.org/10.1007/s11222-022-10189-w>
- Marx BD (1996) Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics* 38(4):374–381. <https://doi.org/10.1080/00401706.1996.10484549>
- McCullagh P, Nelder JA (1989) *Generalized Linear Models*. Chapman and Hall
- Meyer K (2009) Factor-analytic models for genotype  $\times$  environment type problems and structured covariance matrices. *Genet Sel Evol* 41(21). <https://doi.org/10.1186/1297-9686-41-21>
- Mortier F, Ouédraogo DY, Claeys F, et al (2015) Mixture of inhomogeneous matrix models for species-rich ecosystems. *Environmetrics* 26(1):39–51. <https://doi.org/10.1002/env.2320>
- Nelder JA, Wedderburn RWM (1972) *Generalized Linear Models*. *J R Stat Soc A* 135(3):370–384. <https://doi.org/10.2307/2344614>
- Niku J, Warton DI, Hui FK, et al (2017) Generalized linear latent variable models for multivariate count and biomass data in ecology. *J Agric Biol Environ Stat* 22(4):498–522. <https://doi.org/10.1007/s13253-017-0304-7>
- Niku J, Brooks W, Herliansyah R, et al (2019a) Efficient estimation of generalized linear latent variable models. *PloS one* 14(5):e0216129. <https://doi.org/10.1371/journal.pone.0216129>
- Niku J, Hui FK, Taskinen S, et al (2019b) gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in R. *Methods Ecol Evol* 10(12):2173–2182. <https://doi.org/10.1111/2041-210X.13303>
- Niku J, Brooks W, Herliansyah R, et al (2023) gllvm: Generalized Linear Latent Variable Models. R package version 1.4.3
- Ovaskainen O, Tikhonov G, Norberg A, et al (2017) How to make more out of community data? A conceptual framework and its implementation as models and software.

- Ecol Lett 20(5):561–576. <https://doi.org/10.1111/ele.12757>
- Pichler M, Hartig F (2021) A new joint species distribution model for faster and more accurate inference of species associations from big community data. *Methods Ecol Evol* 12(11):2159–2173. <https://doi.org/10.1111/2041-210X.13687>
- Poggiato G, Münkemüller T, Bystrova D, et al (2021) On the interpretations of joint modeling in community ecology. *Trends Ecol Evol* 36(5):391–401. <https://doi.org/10.1016/j.tree.2021.01.002>
- Pollock LJ, Tingley R, Morris WK, et al (2014) Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods Ecol Evol* 5:397–406. <https://doi.org/10.1111/2041-210X.12180>
- R Core Team (2023) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL <https://www.R-project.org/>
- Rabe-Hesketh S, Skrondal A, Pickles A (2002) Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal* 2(1):1–21
- Rand WM (1971) Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc* 66(336):846–850. <https://doi.org/10.1080/01621459.1971.10482356>
- Saidane M, Bry X, Lavergne C (2013) Generalized linear factor models: A new local EM estimation algorithm. *Comm Stat Theory Meth* 42(16):2944–2958. <https://doi.org/10.1080/03610926.2013.790450>
- Schall R (1991) Estimation in generalized linear models with random effects. *Biometrika* 78(4):719–727. <https://doi.org/10.1093/biomet/78.4.719>
- Schwarz G (1978) Estimating the dimension of a model. *Ann Stat* 6(2):461–464
- Skrondal A, Rabe-Hesketh S (2004) Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models. Chapman and Hall/CRC
- Swaine MD, Whitmore TC (1988) On the definition of ecological species groups in tropical rain forests. *Vegetatio* 75:81–86. <https://doi.org/10.1007/BF00044629>
- Tikhonov G, Opedal ØH, Abrego N, et al (2020) Joint species distribution modelling with the R-package Hmsc. *Methods Ecol Evol* 11:442–447. <https://doi.org/10.1111/2041-210X.13345>
- van der Veen B, Hui FKC, Hovstad KA, et al (2023) Concurrent ordination: Simultaneous unconstrained and constrained latent variable modelling. *Methods Ecol Evol* 14(2):683–695. <https://doi.org/10.1111/2041-210X.14035>

- Watkins MW (2018) Exploratory Factor Analysis: A Guide to Best Practice. *J Black Psychol* 44(3):219–246. <https://doi.org/10.1177/0095798418771807>
- Wold S, Ruhe A, Wold H, et al (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. <https://doi.org/10.1137/0905052>
- Wolfinger R, O’connell M (1993) Generalized linear mixed models a pseudo-likelihood approach. *J Stat Comput Sim* 48(3-4):233–243. <https://doi.org/10.1080/00949659308811554>
- Yee TW, Hastie TJ (2003) Reduced-rank vector generalized linear models. *Stat modelling* 3(1):15–41