



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. 2023. hal-04263074v1

HAL Id: hal-04263074

<https://hal.science/hal-04263074v1>

Preprint submitted on 27 Oct 2023 (v1), last revised 8 Apr 2024 (v2)

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 (0000-0002-6667-5057)^{1*†}, Xavier Bry^{1†} and
Catherine Trottier^{1,2†}

¹IMAG, CNRS, Université de Montpellier, Montpellier, France.

²AMIS, Université Paul-Valéry Montpellier 3, F34000, Montpellier,
France.

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

†These authors contributed equally to this work.

Abstract

In a context of component-based multivariate model we propose to model the residual dependence of the responses. Each response of a response matrix is assumed to depend, through a Generalized Linear Model, on a set of explanatory variables, as well as on a set of additional covariates. Explanatory variables are partitioned into conceptually homogeneous variable groups, viewed as explanatory themes. Variables in themes are supposed many and redundant. Thus, generalized linear regression demands dimension reduction and regularization with respect to each theme. By contrast, additional covariates contain few variables, selected so as not to be too redundant, thus demanding no regularization. Regularization is performed searching each theme for an appropriate number of orthogonal components that both contribute to predict the responses and capture relevant structural information in themes. A small set of latent factors completes the model so as to model the covariance matrix of the linear predictors of the responses conditional on the components. To estimate the multiple-theme model, we present an algorithm combining thematic component-based model estimation and factor model estimation. This methodology is tested on simulated data and then applied to an agricultural ecology dataset.


Keywords: EM algorithm, Factor model, Latent variables, 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 (Meyer, 2009). In an ecological context for instance, the species co-occurrences that are not explained by the environmental variables demand to be modeled. With this aim in mind, we model the residual dependence of the responses assumed drawn from Generalized Linear Models (GLM, McCullagh and Nelder, 1989), by means of random numeric latent variables called factors, which we introduce in the linear predictors. Estimating the residual covariance matrix will allow us to identify groups of species. In a Bayesian context, Bystrova et al (2021) profit from the clustering properties of the Dirichlet process to group the species. Indeed, the species in the same cluster have the same rows in the residual correlation matrix. In the same vein, we propose to group the species sharing strong (positive or negative) residual correlations. To do this, we estimate the residual variance-covariance matrix of the linear predictors, and then perform clustering on a dissimilarity matrix built from the estimated correlation matrix. Identifying a group of correlated species is equivalent to identifying a square block of high absolute values in the correlation matrix, once the rows and columns are suitably reordered.

An approach, named Generalized Linear Latent Variable Model (GLLVM), has been proposed by Skrondal and Rabe-Hesketh (2004) to combine GLM with random latent variables. Unfortunately, when factors are involved, the log-likelihood cannot be derived from the GLLVM analytically. Several works propose to maximize this log-likelihood but, some of them suffer from a high computation time. We may cite for instance the works 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 the works using Bayesian Markov Chain Monte Carlo (MCMC) (Hui, 2016; Tikhonov et al, 2020). Few methods reduce the computation time by using a closed form approximation of the log-likelihood. The works employing a variational approximation (Hui et al, 2017), a Laplace approximation (Niku et al, 2017, 2019a) or an extended variational approximation (Korhonen et al, 2023) deserve mentioning. To better reduce the computation time, we propose to use a modeling approach based on Saidane et al (2013), which assumes that maximization should be performed through the EM algorithm after linearizing the and assuming the linearized model is approximately Gaussian. As demonstrated by Wolfinger and O’connell (1993), the linearization method proves to be an interesting alternative to estimate parameters in a context of intractable likelihood. However, modeling the responses also requires taking into account a large set of possibly highly correlated explanatory covariates, so that the GLLVM demands regularization. This can be carried out by means of component-based dimension reduction. This consists in assuming that there exists a small number of common latent explanatory dimensions, which we aim to capture through as many linear combinations of the explanatory variables, named components.

In the wake of Marx (1996), the Supervised Component-based Generalized Linear Regression (SCGLR) was proposed by Bry et al (2013) in order to bridge the multivariate GLM estimation with the component-based dimension reduction of the explanatory space. Moreover, SCGLR allows both to find strong interpretable

explanatory directions, and to produce regularized predictors in the high-dimensional framework (when the covariates outnumber the statistical units). Indeed, 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 relevant directions of the space spanned by the explanatory variables. To extend this method, [Bry et al \(2020b\)](#) developed THEME-SCGLR with the aim to search for the components in a thematic partitioning of the explanatory variables, hence referred to as “themes”. 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  package **SCGLR** is freely available at <https://github.com/SCnext/SCGLR>.

All the aforementioned extensions assume 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 this dependence, we may also need to take into account a thematic partitioning of the explanatory variables. We thus have to extend THEME-SCGLR to the GLLVM.

The paper is organized as follows. In Section 2, we recall the principle of the original SCGLR and its thematic extension. Section 3 presents our extension of SCGLR to the factor models. Section 4 details several simulation studies that illustrate the interest and the good 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 SCGLR with a partitioning of explanatory variables

2.1 Preliminary notations

The sequel contains mathematical developments which use notations listed hereafter:

- Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^N$ be vectors and $\mathbf{W} \in \mathbb{R}^{N \times N}$ be a symmetric positive definite matrix. 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 defines the linear correlation coefficient, denoted ρ .
- $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_P] \in \mathbb{R}^{N \times P}$ and $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_Q] \in \mathbb{R}^{N \times Q}$ being matrices. The space spanned by their column-vectors is denoted $\text{span}[\mathbf{A}, \mathbf{B}]$.
- Let w_n be the weight of unit n , and $\mathbf{W} = \text{diag}(w_n)_{n=1, \dots, N}$. Let \mathbb{R}^N be endowed with metric \mathbf{W} , and let $\mathbf{A} \in \mathbb{R}^{N \times P}$ be a matrix. The \mathbf{W} -orthogonal projector onto $\text{span}[\mathbf{A}]$ is given by $\mathbf{\Pi}_{\text{span}[\mathbf{A}]}^{\mathbf{W}} = \mathbf{A} (\mathbf{A}^T \mathbf{W} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{W}$. The cosine of the angle between a vector $\mathbf{b} \in \mathbb{R}^N$ and $\text{span}[\mathbf{A}]$ with respect to metric \mathbf{W} is given by $\cos_{\mathbf{W}}(\mathbf{b}, \text{span}[\mathbf{A}]) = \cos_{\mathbf{W}}(\mathbf{b}, \mathbf{\Pi}_{\text{span}[\mathbf{A}]}^{\mathbf{W}} \mathbf{b})$.

2.2 The original SCGLR context

In the framework of a multivariate GLM we 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 weakly or not 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 and possibly highly redundant covariates, considered as proxies to latent dimensions, which must be found and interpreted. Thus, the matrix \mathbf{X} demands dimension reduction and regularization, contrary to \mathbf{A} . To achieve this, SCGLR searches for explanatory components in \mathbf{X} jointly supervised by the response set. A component $\mathbf{f} \in \mathbb{R}^N$ writes $\mathbf{f} = \mathbf{X}\mathbf{u}$, where $\mathbf{u} \in \mathbb{R}^P$ is a loading vector. 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 γ_k and $\boldsymbol{\delta}_k$ are regression parameters. Component \mathbf{f} is common to all the responses and, for identification, we impose $\mathbf{u}^T\mathbf{u} = 1$. It was assumed originally that the responses are independent conditional on the explanatory variables, and consequently on \mathbf{f} .

2.3 The original SCGLR specific criterion

For 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 μ_{nk} the mean parameter for statistical unit n . The working variable w_{nk} associated with y_{nk} is then calculated as the first order development of h_k at point μ_{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 $\mathbb{V}[\boldsymbol{\zeta}_k] =: \mathbf{W}_k^{-1}$.

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 loading vector \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 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 produced by a common latent dimension. Assuming that \mathbf{X} consists of P standardized numeric variables, the associated SR measure ϕ 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} is the weight matrix (in practice, $\mathbf{W} = \frac{1}{N} \mathbf{I}_N$). Components will line up with a more or less thin bundle depending on whether $l \geq 1$ is greater or smaller respectively. The main objective is to focus on the most interpretable directions.

The SCGLR specific criterion, proposed by Bry et al (2020b), introduces 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 ϕ and $\psi_{\mathbf{A}}$ by solving

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

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.

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 $\mathbf{X}_1, \dots, \mathbf{X}_R$ plus a set of covariates \mathbf{A} , and that structurally relevant dimensions should be explicitly identified in the \mathbf{X}_r ’s. 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 quite 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 Θ . The maximization is carried out alternating two steps: (i) Given Θ , 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, 2022). (ii) Given all \mathbf{u}_r , maximize the criterion with respect to Θ . 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 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 new component \mathbf{f}_r^{h+1} must best complement both the existing ones 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, we impose that \mathbf{f}_r^{h+1} be orthogonal to \mathbf{F}_r^h , *i.e.* $\mathbf{F}_r^{hT} \mathbf{W} \mathbf{f}_r^{h+1} = \mathbf{0}$. We calculate every new component as the solution of program (2), with the additional constraint: $\Delta_r^h T \mathbf{u}_r^{h+1} = \mathbf{0}$, where $\Delta_r^h = \mathbf{X}_r^T \mathbf{W} \mathbf{F}_r^h$, and loop on r until overall convergence of the component system. 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 $\Delta_r^0 = \mathbf{0}$.

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 $\boldsymbol{\gamma}_k = (\gamma_{k1}, \dots, \gamma_{kR})^T$ 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 random latent factors $\mathbf{g}_n = (g_{n1}, \dots, g_{nJ})^T$

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

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 regression parameters 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, hence 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 rotation. 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 GLM with factors

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 latent variables \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 is not allowed. 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 working variable \mathbf{w}_k could 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 assume that the distribution of the working variables given \mathbf{F} , \mathbf{A} and \mathbf{G} is Gaussian, and view the factors as latent variables. The model pseudo-log-likelihood $l(\boldsymbol{\Theta}; \mathbf{W})$, where \mathbf{W} denotes the matrix of working 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 working variables have a Gaussian distribution, we calculate and then maximize the expectation of their complete log-likelihood $l(\boldsymbol{\Theta}; \mathbf{W}, \mathbf{G})$. Further details of the EM algorithm are given in [Appendix C](#).

3.3 The overall algorithm

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

3.4 The clustering steps

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

1. Estimate the residual covariance matrix $\boldsymbol{\Sigma} := \mathbf{B}^T \mathbf{B}$.
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 responses highly correlated, be it positively or negatively.
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, 2021](#)).

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** $\text{\textcircled{R}}$ 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 implemented 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). As shown by Chauvet et al (2019), the hyper-parameters must be chosen to avoid the components to be too close to the principal components ($s > 0.5$) or to be drawn 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 a situation of a more or less clearly separated 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 ξ 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 also calculated the RI and ARI of the partitions output by a competing $\text{\textcircled{R}}$ package in a context of binary data. For each simulation, one hundred samples have been generated. The $\text{\textcircled{R}}$ 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 ξ_5 are simulated independently. 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 ξ_1 and ξ_2 respectively, and \mathcal{M}_1 contains fifty unstructured noise variables drawn from a 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 ξ_3, ξ_4 and ξ_5 respectively, and \mathcal{M}_2 contains sixty unstructured noise variables drawn from a multivariate normal distribution. More formally, for all $i = 1, \dots, 5$, a variable x_p within a bundle is simulated as $x_p = \xi_i + \varepsilon_p$, where $\varepsilon_p \sim \mathcal{N}_N(0, 0.1\mathbf{I}_N)$. This generation leads to $P = 500$ explanatory variables.

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

$$\begin{aligned} \forall k = 1, \dots, 20, \quad \mathbf{y}_k &\sim \mathcal{N}_N(\boldsymbol{\mu} = \gamma_{1k} \boldsymbol{\xi}_1 + \gamma_{2k} \boldsymbol{\xi}_2 + \mathbf{G} \mathbf{b}_k, \boldsymbol{\Sigma} = \sigma_k^2 \mathbf{I}_N) \\ \forall k = 21, \dots, 40, \quad \mathbf{y}_k &\sim \mathcal{P}(\boldsymbol{\lambda} = \exp[0.5\gamma_{1k} \boldsymbol{\xi}_4 + 0.5\gamma_{2k} \boldsymbol{\xi}_3 + \mathbf{G} \mathbf{b}_k]) \\ \forall k = 41, \dots, 50, \quad \mathbf{y}_k &\sim \mathcal{B}(\mathbf{p} = \text{logit}^{-1}[\gamma_{2k} \boldsymbol{\xi}_3 + \gamma_{3k} \boldsymbol{\xi}_2 + \mathbf{G} \mathbf{b}_k]), \end{aligned}$$

where for all k , σ_k^2 , γ_{1k} , γ_{2k} and γ_{3k} are uniformly generated, with $\sigma_k^2 \in [0.1, 0.2]$, $\gamma_{1k} \in [-4, 4]$, $\gamma_{2k} \in [-2, 2]$ and $\gamma_{3k} \in [-0.5, 0.5]$. In the linear predictors, we range the latent dimensions following their regression parameters.

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 E 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 dramatically when factors are involved in the model. This shows that, because 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 $J = 0$ or $J = 1$, the BIC leads to overestimate the number of components.

4.1.3 Varying the hyper-parameters and the variance within the 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 $\sigma_B^2 \in \{0.1, 0.2, 0.3\}$. Figure 1 shows the residual correlation matrices for the three values of σ_B^2 .

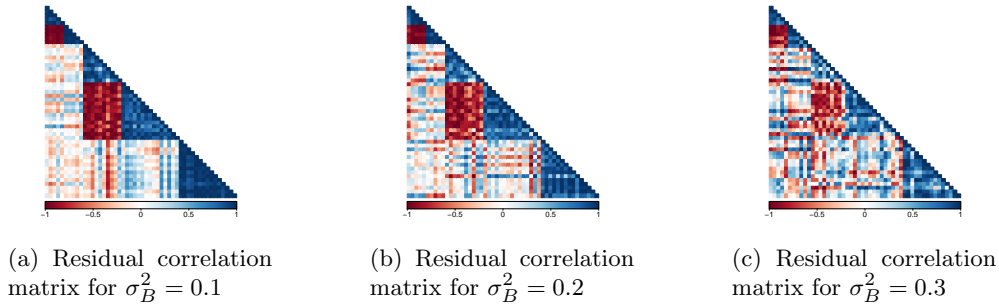


Fig. 1: Residual correlation matrices for different values of σ_B^2

Table 1 gives the results for $\sigma_B^2 = 0.1$. For the $s = 0.3$ and $s = 0.5$ cases, the values of RI and ARI are slightly better than the $s = 0.1$ case. Moreover, for $s = 0.3$ and $s = 0.5$, the maximum value for RI and ARI is reached for $l = 4$. This is in accordance with the hyper-parameters calibrated through the SCGLR [R](#) package. The Tables summing up the results for $\sigma_B^2 = 0.2$ and $\sigma_B^2 = 0.3$ are presented in Appendix F. 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 . We may also note that the difference between the values of RI and ARI across the hyper-parameters s and l tends to fade when σ_B^2 increases. 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 σ_B^2 is involved in the stochastic one. 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 close to the principal components, the two components of theme \mathbf{X}_2 are drawn between the three bundles and so, produce low square correlations with the latent dimensions. The interest of tuning the locality is shown by the gap between the results obtained for $l = 1$ and $l = 2$: in the latter case, the square correlations are dramatically better. Furthermore, ξ_3 being the less explanatory latent dimension, $\rho^2(\xi_3, \cdot)$ is always lower than the other square correlations.

Figure 2 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 [R](#) package `gllvm` ([Niku et al, 2019b](#)). This package offers three ways to perform GLLVM: using a variational approximation (VA, [Hui et al, 2017](#)), a Laplace approximation (LA, [Niku et al, 2017](#),

Table 1: Mean values of RI, ARI and square correlation over a 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	$\rho^2(\xi_{1,\cdot})$	$\rho^2(\xi_{2,\cdot})$	$\rho^2(\xi_{3,\cdot})$	$\rho^2(\xi_{4,\cdot})$
0.1	1	0.926	0.839	0.938	0.906	0.759	0.838
	2	0.927	0.839	0.980	0.959	0.810	0.935
	3	0.920	0.816	0.981	0.962	0.805	0.942
	4	0.928	0.837	0.979	0.966	0.816	0.945
	7	0.926	0.830	0.966	0.954	0.798	0.946
	10	0.927	0.835	0.967	0.955	0.792	0.945
0.3	1	0.944	0.876	0.973	0.936	0.735	0.753
	2	0.944	0.877	0.993	0.972	0.934	0.950
	3	0.946	0.881	0.987	0.974	0.938	0.965
	4	0.947	0.882	0.985	0.974	0.927	0.962
	7	0.943	0.875	0.984	0.974	0.911	0.964
	10	0.945	0.878	0.984	0.974	0.911	0.964
0.5	1	0.942	0.871	0.974	0.937	0.697	0.659
	2	0.944	0.875	0.994	0.972	0.943	0.946
	3	0.947	0.882	0.988	0.975	0.948	0.961
	4	0.948	0.884	0.986	0.975	0.946	0.967
	7	0.945	0.879	0.985	0.975	0.917	0.975
	10	0.944	0.877	0.985	0.975	0.911	0.969

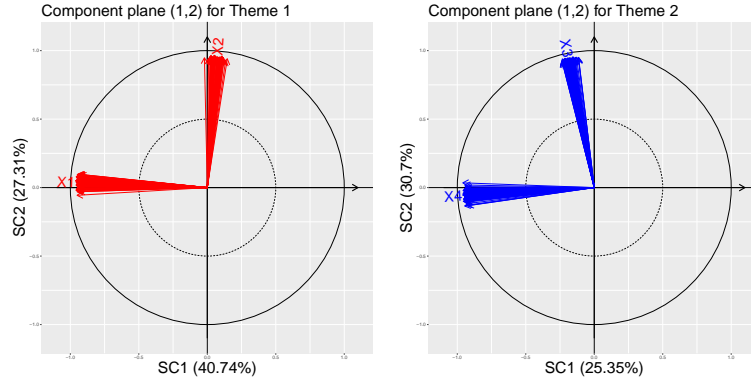
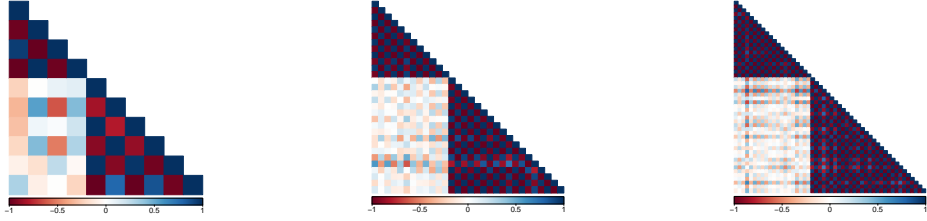


Fig. 2: Correlation scatterplot of plane (1,2) for the two themes with $s = 0.3$ and $l = 4$ obtained by the F-SCGLR algorithm. The red arrows represent the bundles \mathcal{X}_1 and \mathcal{X}_2 which explain the first theme. The blue ones represent the bundles \mathcal{X}_3 and \mathcal{X}_4 which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

2019a) or an extended variational approximation (EVA, Korhonen et al, 2023). To deal with the possibly high correlations among the predictors, the \mathbb{R} package `glvm` also proposes to shrink the regression parameters by performing a reduced rank regression with random slopes (van der Veen et al, 2023). The performances of all the methods are compared through the RI and ARI of the partitions output by the estimated residual correlation matrix, the Procrustes error between the true and estimated loading



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

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

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

Fig. 3: Residual correlation matrices for different values of K

matrix \mathbf{B} and between the true and estimated matrix of latent variables \mathbf{G} , and the computation time in seconds. 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. For the sake of simplicity, a bundle \mathbf{X} of ten variables distributed about the latent dimension $\boldsymbol{\xi}$ is generated. One categorical variable with three 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 regression coefficients 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 $\boldsymbol{\mu}_2 = (1.5, 0)^T$. The comparable \mathbb{R} package not allowing to consider different distribution families for the responses, we restricted the comparison to binary outcomes

$$\forall k = 1, \dots, K, \quad \mathbf{y}_k \sim \mathcal{B} \left(\mathbf{p} = \text{logit}^{-1} [\gamma_k \boldsymbol{\xi} + \mathbf{A} \boldsymbol{\delta}_k + \mathbf{G} \mathbf{b}_k] \right),$$

where for all k , γ_k and $\boldsymbol{\delta}_k$ are uniformly generated, with $\gamma_k \in [-4, 4]$ and $\boldsymbol{\delta}_k \in [-1, 1]$. Figure 3 shows the residual correlation matrices obtained for the three values of K .

4.2.2 Results of the comparison

In this simulation the \mathbb{R} package **SCGLR** calibrates the hyper-parameters to $s = 0.5$ and $l = 1$, while the BIC selects $H_1 = 1$ and $J = 2$. Table 2 sums up the RI, ARI, Procrustes error and computation time output by, on the one hand, our

Table 2: Mean values of RI, ARI, Procrustes error of the latent variables \mathbf{G} and their loadings \mathbf{B} and computation time over a hundred samples with $N \in \{100, 200, 300\}$ and $K \in \{10, 30, 50\}$.

N	100			200			300		
K	10	30	50	10	30	50	10	30	50
F-SCGLR									
RI	0.845	0.869	0.895	0.922	0.951	0.980	0.950	0.976	0.986
ARI	0.679	0.740	0.792	0.839	0.903	0.960	0.898	0.952	0.973
Proc B	0.156	0.165	0.173	0.098	0.082	0.084	0.067	0.055	0.054
Proc G	0.429	0.216	0.156	0.429	0.202	0.138	0.422	0.196	0.133
Time	0.532	1.721	4.812	0.834	3.187	7.982	1.221	3.960	11.11
gllvm-EVA									
RI	0.687	0.775	0.785	0.721	0.792	0.762	0.726	0.809	0.763
ARI	0.386	0.550	0.571	0.456	0.583	0.518	0.466	0.617	0.521
Proc B	0.982	0.670	0.615	1.021	0.713	0.664	0.948	0.635	0.621
Proc G	0.578	0.300	0.223	0.544	0.256	0.182	0.518	0.231	0.166
Time	1.436	5.182	11.02	4.167	18.20	44.76	8.995	36.86	88.99
gllvm-VA									
RI	0.657	0.702	0.704	0.727	0.739	0.716	0.725	0.779	0.770
ARI	0.337	0.402	0.403	0.468	0.476	0.426	0.455	0.556	0.534
Proc B	0.737	0.633	0.635	0.677	0.618	0.668	0.624	0.614	0.618
Proc G	0.709	0.296	0.244	0.697	0.245	0.180	0.702	0.224	0.165
Time	1.384	8.80	20.18	3.645	22.78	47.58	8.665	38.34	79.65
gllvm-shrinkage									
RI	0.606	0.700	0.765	0.648	0.735	0.698	0.65	0.697	0.685
ARI	0.238	0.405	0.532	0.331	0.475	0.393	0.327	0.398	0.369
Proc B	1.028	0.826	0.682	1.055	0.724	0.694	1.051	0.710	0.676
Proc G	0.622	0.311	0.209	0.638	0.249	0.170	0.615	0.241	0.145
Time	1.237	3.584	7.493	3.640	21.97	49.61	8.464	52.66	111.9
gllvm-LA									
RI	0.598	0.699	0.689	0.652	0.720	0.734	0.710	0.749	0.761
ARI	0.231	0.398	0.375	0.323	0.440	0.461	0.433	0.496	0.516
Proc B	1.068	1.291	1.326	1.224	0.851	0.755	1.165	0.688	0.632
Proc G	0.650	0.352	0.292	0.557	0.257	0.186	0.501	0.230	0.168
Time	32.88	69.09	127.0	46.70	124.8	262.8	53.95	182.5	386.2

package **FactorSCGLR** performing the F-SCGLR method, and, on the other hand, the package **gllvm** implementing the VA, LA and EVA approaches.

We observe that, for all combinations of N and K , F-SCGLR gives the best values of RI and ARI, followed by EVA, VA, the shrinkage approach and then LA. Indeed, the highest values obtained of the ARI are respectively: 0.973, 0.617, 0.556, 0.532 and 0.516. Unlike with our package, which performs better when the number of either statistical units or responses increase, a higher number of responses may cause a deterioration of the classification of the compared methods. Likewise for all cases, the lowest Procrustes errors on \mathbf{B} and \mathbf{G} are reached by F-SCGLR. The ordination performances are not ranked in the same order as those obtained for the classification. The F-SCGLR, VA, EVA, LA and shrinkage methods respectively have an error on \mathbf{B} equal to 0.054, 0.614, 0.615, 0.632 and 0.676. However, we may note that all methods could well predict the latent variables. In their best combinations, the values of the Procrustes error on \mathbf{G} are equal to 0.133 (F-SCGLR) and 0.168 (LA). When the number of responses increases, the Procrustes error on \mathbf{B} for the F-SCGLR, shrinkage


and LA methods may get bigger while the Procrustes error on \mathbf{G} decreases for all methods. Across the simulations, F-SCGLR is the fastest of the compared methods. The longest computation time (less than 12 seconds) occurred for $N = 300$ and $K = 50$, while EVA, VA, shrinkage and LA ran for 89, 80, 112 and 386 seconds respectively in that case. However, contrary to [Korhonen et al \(2023\)](#), we did not observe that EVA ran faster than VA: in the $K = 10$ and $(N, K) = (300, 50)$ cases, the computation time of VA was lower. The conclusions of our respective works agree nevertheless that 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 one Shannon index), three vascular plant variables (one richness, one relative cover and one Shannon index) and six axes of 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 samples of Poisson random variables while the other responses are considered normally distributed. To model the agrobiodiversity, we have $P = 21$ variables partitioned into four themes and $Q = 1$ additional covariable. 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 considered 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 we find the best combination of number of components and factors according to the BIC. However, due to 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 **SCGLR**  package recommends tuning hyper-parameters to $s = 0.5$ and $l = 1$ for this agricultural ecology dataset. Henceforth, 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. [Dufлот 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 here useless to this prediction.

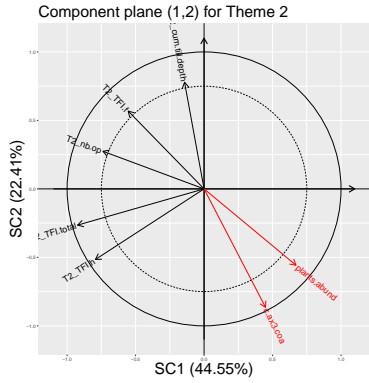
We henceforth try to interpret the components of the second theme. The first component f_2^1 is correlated ($\rho = -0.924$, $\rho = -0.794$ and $\rho = -0.738$) with a bundle of three variables, of which “TFI.total” and “TFI.h” represent a treatment frequency index of herbicides, and “nb.op” is the 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_2^3 . Figure 4 represents the correlation plots of the second theme.

In this agricultural ecology dataset, three factors are recommended, according to 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 by the three measures of the carabids. The second group gathers the first axis of the CA of the carabids, the plant richness and the plant Shannon diversity index. The carabids’ second CA axis, the plant cover and the first and third plants’ CA axes make up the third group. Finally, the fourth group contains the carabids’ third CA axis and the plants’ second CA axis. Figure 5 shows the residual correlation values alongside the Euclidean representation of the responses 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 on the graph, the weaker the residual correlations, the wider the groups are scattered.

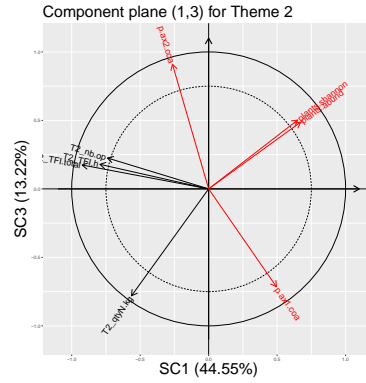
6 Conclusion and discussion

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 extends 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: through a factor model, it models the residual variance-covariance matrix of the responses using a small number of latent factors. This matrix can be used as a basis for clustering, enabling to identify groups of linked responses.

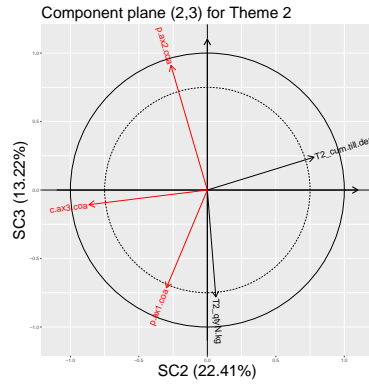
In our simulation study, F-SCGLR proved to behave as expected regarding response clusters. Whenever the clusters were reasonably distinct, the original partitions were recovered. Whatever the dispersion of the regression coefficients within the clusters, it provided components aligned with the simulated latent dimensions. Our  package outperforms the package `gllvm` in three ways: (i) The thematic model allows to find supervised components, thus reducing the dimension in a context of possibly numerous explanatory variables. (ii) Responses with different distribution families are



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



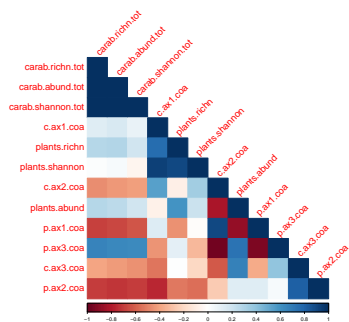
(b) Component plane (1,3) for the farming intensity theme



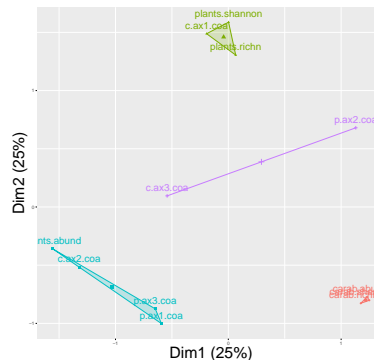
(c) Component plane (2,3) for the farming intensity theme

Fig. 4: Correlation plots of F-SCGLR plane (1,2), (2,3) and (1,3) of the second theme (farming intensity). The black arrows represent the covariates while the red ones 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.

allowed. (iii) The performances of our package are better in terms of computation time and of cluster detection. On the agricultural ecology dataset, we found four groups of responses. Due to very high residual correlations, the first one gathers the measures of the carabids. The other groups are composed by a mix between the plant variables and the axes of the 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. Moreover, performing F-SCGLR, we revealed that the treatment by herbicides, the operations conducted by the farmers,



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




(b) Plane (1,2) of the MDS derived from the residual correlation matrix where the groups are colored.

Fig. 5: The residual correlation matrix alongside the Euclidean representation of the responses output by the MDS.

the tillage depth and the quantity of nitrogen are the most involved variables in the prediction of the agrobiodiversity.

In this research, some limitations have been reached. The use of 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 researches in this topics need to be effected. SCGLR and its extensions suffer of a high number of hyper-parameters involving the use of heuristics to well calibrate the latter. Only Bernoulli, Binomial, Gaussian and Poisson distributions can currently be 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.

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 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 needs to maximize for computing the $(h + 1)$ -th loading-vector write

$$\phi(\mathbf{u}) = \left(\sum_{j=1}^p \omega_j (\mathbf{u}^T \mathbf{N}_j \mathbf{u})^l \right)^{1/l}$$

and

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

To facilitate the computation of the loading-vector, 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 power 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 VPI is

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

To calculate the gradient we use the classical rules of derivation

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

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

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

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

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

Consequently, by classical Euclidean statistical concepts, we have

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

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

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

Now,

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

Let,

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

and

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

Finally, we 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

$$\begin{cases} \max_{\mathbf{u}} J_h(\mathbf{u}), \\ \text{s.t. } \mathbf{u}^T \mathbf{M}^{-1} \mathbf{u} = 1 \quad \text{and} \quad \mathbf{\Delta}_h^T \mathbf{u} = 0, \end{cases} \quad (\text{B1})$$

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

$$\begin{cases} \max_{\mathbf{v}} G_h(\mathbf{v}), \\ \text{s.t. } \mathbf{v}^T \mathbf{v} = 1 \quad \text{and} \quad \mathbf{E}_h^T \mathbf{v} = 0. \end{cases} \quad (\text{B2})$$

To solve (B2), 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

$$\nabla_{\mathbf{v}} \mathcal{L}(\mathbf{v}, \lambda, \boldsymbol{\eta}) = 0 \Leftrightarrow \Gamma_h(\mathbf{v}) - 2\lambda \mathbf{v} - \mathbf{E}_h \boldsymbol{\eta} = 0 \quad (\text{B3})$$

$$\Leftrightarrow \mathbf{v} = \frac{1}{2\lambda} (\Gamma_h(\mathbf{v}) - \mathbf{E}_h \boldsymbol{\eta}). \quad (\text{B4})$$

Multiplying (B3) by \mathbf{E}_h^T

$$\begin{aligned} 2\lambda \underbrace{\mathbf{E}_h^T \mathbf{v}}_{=0} &= \mathbf{E}_h^T \Gamma_h(\mathbf{v}) - \mathbf{E}_h^T \mathbf{E}_h \boldsymbol{\eta} \Leftrightarrow \mathbf{E}_h^T \Gamma_h(\mathbf{v}) = \mathbf{E}_h^T \mathbf{E}_h \boldsymbol{\eta} \\ &\Leftrightarrow \boldsymbol{\eta} = (\mathbf{E}_h^T \mathbf{E}_h)^{-1} \mathbf{E}_h^T \Gamma_h(\mathbf{v}). \end{aligned} \quad (\text{B5})$$

Substituting (B5) in (B4), we get

$$\mathbf{v} = \frac{1}{2\lambda} \left(\Gamma_h(\mathbf{v}) - \mathbf{E}_h (\mathbf{E}_h^T \mathbf{E}_h)^{-1} \mathbf{E}_h^T \Gamma_h(\mathbf{v}) \right)$$

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

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

$$\mathbf{v}^{(t+1)} = \frac{\mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}^{(t)})}{\left\| \mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}^{(t)}) \right\|}. \quad (\text{B6})$$

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, show 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 (B6) 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,

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

Finally,

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

Although iteration (B6) 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

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

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 \mathbf{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, $\mathbf{w} = \tau \mathbf{\Pi}_{\boldsymbol{\varpi}} \boldsymbol{\kappa}^{(t)}$, and

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

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 type) to find the maximum of G_h on the arc $(\mathbf{v}^{(t)}, \boldsymbol{\kappa}^{(t)})$, and take it as $\mathbf{v}^{(t+1)}$. Therefore, we propose two possible generic iterations for the PING algorithm, which deal with this problem. Algorithm 1 and Algorithm 2 present these alternatives. The first one should be preferred, but is less easy to program.

Algorithm 1 PING algorithm

- 1: **while** not convergence **do**
 - 2: $\boldsymbol{\kappa}^{(t)} \leftarrow \frac{\mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}^{(t)})}{\left\| \mathbf{\Pi}_{\text{span}[\mathbf{E}_h]^\perp} \Gamma_h(\mathbf{v}^{(t)}) \right\|}$
 - 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 The EM algorithm

We are now dealing with the linearized model, where the factors are latent. So, we shall use the EM algorithm to estimate the parameters. The previous developments lead to the conditional 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

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

Algorithm 2 Alternative PING algorithm

```

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

```

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

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

where $\mathbf{\Gamma} = [\gamma_1, \dots, \gamma_K]$, $\mathbf{\Delta} = [\delta_1, \dots, \delta_K]$, $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_K]$, 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 \mathcal{W} , \mathbf{F} , \mathbf{A} and \mathbf{G} respectively. The expectation and the variance are given by $\mathbb{E}[\mathbf{w}_n] = \mathbf{\Gamma}^T \mathbf{f}_n + \mathbf{\Delta}^T \mathbf{a}_n$ and $\mathbb{V}[\mathbf{w}_n] = \mathbf{B}^T \mathbf{B} + \mathbf{\Upsilon}_n^{-1}$, where

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

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

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

C.1 The expectation (E) step

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

$$\begin{aligned} \mathbb{E}[l(\Theta; \mathcal{W}, \mathcal{G}) \mid \mathcal{W}; \Theta'] &= \\ & \sum_{n=1}^N \int \ln(L(\mathbf{w}_n \mid \mathbf{g}_n; \Theta) L(\mathbf{g}_n; \Theta)) L(\mathbf{g}_n \mid \mathbf{w}_n; \Theta') d\mathbf{g}_n. \end{aligned}$$

Thus, we need to first find the law of $\mathbf{g}_n \mid \mathbf{w}_n$. Since the random vector $(\mathbf{w}_n^T, \mathbf{g}_n^T)^T$ is Gaussian, we have

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

Thanks to the conditioning rule of the multivariate Gaussian, we get

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

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

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

and

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

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

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

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

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

C.2 The maximization (M) step

The maximization step maximizes the conditional expectation of the complete log-likelihood with respect to $\boldsymbol{\Theta}$, 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

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

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

$$\nabla_{\sigma_k^2} \mathbb{E}[l(\boldsymbol{\Theta}; \mathcal{W}, \mathbf{G}) \mid \mathcal{W}, \boldsymbol{\Theta}'] = 0$$

$$\begin{aligned}
&\Leftrightarrow \nabla_{\sigma_k^2} \left\{ N \ln(\sigma_k^2) + \frac{1}{\sigma_k^2} \left[\left\| \mathbf{w}_k - \tilde{\mathbf{X}} \boldsymbol{\beta}_k \right\|^2 + \mathbf{b}_k^T \left(\sum_{n=1}^N \tilde{\mathbf{R}}_n \right) \mathbf{b}_k \right. \right. \\
&\quad \left. \left. - 2 \left(\tilde{\mathbf{G}} \mathbf{b}_k \right)^T \left(\mathbf{w}_k - \tilde{\mathbf{X}} \boldsymbol{\beta}_k \right) \right] \right\} = 0 \\
&\Leftrightarrow N - \frac{1}{\sigma_k^2} \left\{ \left\| \mathbf{w}_k - \tilde{\mathbf{X}} \boldsymbol{\beta}_k \right\|^2 + \mathbf{b}_k^T \left(\sum_{n=1}^N \tilde{\mathbf{R}}_n \right) \mathbf{b}_k \right. \\
&\quad \left. - 2 \left(\tilde{\mathbf{G}} \mathbf{b}_k \right)^T \left(\mathbf{w}_k - \tilde{\mathbf{X}} \boldsymbol{\beta}_k \right) \right\} = 0 \\
&\Leftrightarrow \sigma_k^2 = \frac{1}{N} \left\{ \left\| \mathbf{w}_k - \tilde{\mathbf{X}} \boldsymbol{\beta}_k \right\|^2 + \mathbf{b}_k^T \left(\sum_{n=1}^N \tilde{\mathbf{R}}_n \right) \mathbf{b}_k \right. \\
&\quad \left. - 2 \left(\tilde{\mathbf{G}} \mathbf{b}_k \right)^T \left(\mathbf{w}_k - \tilde{\mathbf{X}} \boldsymbol{\beta}_k \right) \right\}.
\end{aligned}$$

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 $(J - k)$ *a priori* fixed. 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

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

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

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

C.3 The algorithm

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

Algorithm 3 The EM algorithm applied to factor models with GLM

```

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:        $\left[ \sum_{n=1}^N v_{nk} \left( \tilde{\mathbf{R}}_n^{(t+1)} \right)_{1:k,1:k} \right]^{-1} \left( \tilde{\mathbf{G}}_{1:k}^{(t+1)} \right)^T \mathbf{W}_k (\mathbf{w}_k - \tilde{\mathbf{X}} \beta_k^{(t+1)})$ 
19:   else
20:      $\mathbf{b}_k^{(t+1)} = \left[ \sum_{n=1}^N v_{nk} \tilde{\mathbf{R}}_n^{(t+1)} \right]^{-1} \tilde{\mathbf{G}}^{(t+1)T} \mathbf{W}_k (\mathbf{w}_k - \tilde{\mathbf{X}} \beta_k^{(t+1)})$ 
21:   end if
22:    $t \leftarrow t + 1$ 
23: end while

```

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

Appendix E Identification of the true model

Table E1 sums up the results on a cross-product grid.

Appendix F Additional simulation studies

The Tables summing up the results for $\sigma_B^2 = 0.2$ and $\sigma_B^2 = 0.3$ are presented in Table F2 and Table F3.

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 working 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 E1: Mean values of BIC over a 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	79051	74909	63621	70120	1	34205	54295	28680	30613
2	57546	55794	46597	45896	2	29705	30369	26841	25463
3	54710	53330	43731	41406	3	31065	26012	24542	25592
4	44658	43938	42169	39733	4	34943	26930	24520	25369
$J = 2$					$J = 3$				
H_2/H_1	1	2	3	4	H_2/H_1	1	2	3	4
1	21265	19833	19907	21087	1	25655	17235	17474	17673
2	20303	18678	21227	20157	2	19150	15915	16197	16378
3	20601	20361	22026	20565	3	19050	16059	16341	16588
4	20774	19022	19309	19496	4	19329	16308	16640	16808
$J = 4$					$J = 5$				
H_2/H_1	1	2	3	4	H_2/H_1	1	2	3	4
1	18356	16025	16237	16556	1	16584	16212	16462	16706
2	16364	16058	16287	16651	2	16647	16387	16628	16911
3	16534	16242	16484	16919	3	16852	16661	17020	17167
4	16835	16424	16761	17103	4	17211	16849	17279	17412

References

- 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>
- 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

Table F2: Mean values of RI, ARI and square correlation over a 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	0.797	0.515	0.939	0.909	0.770	0.830
	2	0.788	0.494	0.986	0.967	0.837	0.934
	3	0.789	0.505	0.985	0.969	0.848	0.948
	4	0.790	0.501	0.984	0.970	0.842	0.950
	7	0.785	0.493	0.977	0.969	0.839	0.942
	10	0.787	0.494	0.973	0.970	0.825	0.941
0.3	1	0.795	0.519	0.973	0.936	0.720	0.738
	2	0.801	0.527	0.992	0.972	0.937	0.944
	3	0.801	0.528	0.988	0.974	0.920	0.964
	4	0.800	0.532	0.985	0.975	0.908	0.966
	7	0.803	0.538	0.977	0.975	0.904	0.950
	10	0.806	0.538	0.977	0.975	0.878	0.956
0.5	1	0.796	0.524	0.974	0.937	0.703	0.657
	2	0.804	0.535	0.993	0.973	0.946	0.943
	3	0.798	0.522	0.985	0.975	0.938	0.955
	4	0.792	0.514	0.981	0.975	0.919	0.954
	7	0.799	0.527	0.978	0.976	0.919	0.957
	10	0.796	0.520	0.978	0.976	0.902	0.952

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>

Bystrova D, Poggiato G, Bektaş B, et al (2021) Clustering species with residual covariance matrix in joint species distribution models. *Front Ecol Evol* 9. <https://doi.org/10.3389/fevo.2021.601384>

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>

Table F3: Mean values of RI, ARI and square correlation over a 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	0.701	0.264	0.949	0.921	0.760	0.827
	2	0.706	0.288	0.987	0.970	0.838	0.921
	3	0.704	0.278	0.986	0.972	0.824	0.944
	4	0.704	0.274	0.982	0.971	0.829	0.946
	7	0.705	0.292	0.975	0.964	0.825	0.945
	10	0.709	0.302	0.967	0.966	0.817	0.944
0.3	1	0.699	0.280	0.974	0.938	0.715	0.726
	2	0.701	0.281	0.991	0.972	0.917	0.945
	3	0.698	0.289	0.988	0.975	0.921	0.965
	4	0.701	0.290	0.987	0.975	0.904	0.950
	7	0.700	0.288	0.974	0.977	0.886	0.944
	10	0.697	0.291	0.970	0.977	0.871	0.944
0.5	1	0.699	0.277	0.974	0.938	0.706	0.654
	2	0.706	0.287	0.992	0.973	0.927	0.941
	3	0.705	0.288	0.987	0.975	0.920	0.949
	4	0.704	0.287	0.983	0.975	0.913	0.938
	7	0.703	0.294	0.973	0.977	0.904	0.956
	10	0.700	0.296	0.971	0.977	0.899	0.950

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>

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 (2022) Response mixture models based on supervised components: Clustering floristic taxa. *Stat Modelling* 0(0). <https://doi.org/10.1177/1471082X221115525>

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, 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.sthda.com/english/rpks/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>
- 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>
- 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 (2021) 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>
- 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
- 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>
- 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>