Factor model in multiblock component-based GLM

Julien GIBAUD[1], Xavier BRY[1] and Catherine TROTTIER[1,2]

[1] IMAG, CNRS, Univ. Montpellier, France.
[2] AMIS, UPV Montpellier 3, Montpellier, France.

RJS 2022

Outline

Outline

Motivations

Ecological motivations

In a context of global warming, we aim at:

- Finding the main determinants of species observations, among a thematic partitioning of the explanatory variables
- Identifying groups of species sharing mutual dependencies

Motivations

Ecological motivations

In a context of global warming, we aim at:

- Finding the main determinants of species observations, among a thematic partitioning of the explanatory variables
- Identifying groups of species sharing mutual dependencies

Statistical counterparts

- Finding strong dimensions allowing to explain the responses as best as possible
  - ↬ Supervised components
- Identifying blocks in the responses' conditional variance-covariance matrix
  - ↬ Random latent variables: factors

Outline

What is a supervised component?

Notations

- Let $Y = [y_1, \ldots, y_q] \in \mathbb{R}^{n \times q}$ be the matrix of responses (GLM)
- Let $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n \times p}$ be the matrix of explanatory variables

What is a supervised component?

Notations

- Let $Y = [y_1, \ldots, y_q] \in \mathbb{R}^{n \times q}$ be the matrix of responses (GLM)
- Let $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n \times p}$ be the matrix of explanatory variables

Definition

A component is a vector $f \in \mathbb{R}^n$ linearly combining the explanatory variables, such that

- $f^h = Xu^h$, for $h = 1, \ldots, H$
- $f^h \perp f^g$, for all $h \neq g$

What is a supervised component?

Notations

- Let $Y = [y_1, \ldots, y_q] \in \mathbb{R}^{n \times q}$ be the matrix of responses (GLM)
- Let $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n \times p}$ be the matrix of explanatory variables

Definition

A component is a vector $f \in \mathbb{R}^n$ linearly combining the explanatory variables, such that

- $f^h = Xu^h$, for $h = 1, \ldots, H$
- $f^h \perp f^g$, for all $h \neq g$

Demands

- Components must be close to some explanatory variables to be interpreted
- Components must predict responses Y $\Rightarrow$ supervised components

SCGLR (Bry et al., 2020)

Structural Relevance (SR)

The criterion $\phi(u)$ measures the "strength" of the component $f = Xu$ (overall closeness to explanatory variables) under the constraint $\|u\|^2 = 1$

SCGLR (Bry et al., 2020)

Structural Relevance (SR)

The criterion $\phi(u)$ measures the "strength" of the component $f = Xu$ (overall closeness to explanatory variables) under the constraint $\|u\|^2 = 1$

Goodness-of-Fit (GoF)

The criterion $\psi(u, \theta)$ is the likelihood of the component model

SCGLR (Bry et al., 2020)

Structural Relevance (SR)

The criterion $\phi(u)$ measures the "strength" of the component $f = Xu$ (overall closeness to explanatory variables) under the constraint $\|u\|^2 = 1$

Goodness-of-Fit (GoF)

The criterion $\psi(u, \theta)$ is the likelihood of the component model

$$\Downarrow$$

The SCGLR combined criterion

$$\underset{u,\, \|u\|^2=1}{\operatorname{argmax}} \; s \ln\left(\phi(u)\right) + (1-s) \ln\left(\psi(u,\theta)\right)$$

The real $s \in [0, 1]$ allows to tune the trade-off between SR and GoF

THEME-SCGLR (Bry et al., 2020)

Notations

- Let $X = [X_1, \ldots, X_R] \in \mathbb{R}^{n \times p}$ be the matrix of R thematic subsets
- Let $f_r^h = X_r u_r^h$ be the hth component of theme X_r

THEME-SCGLR (Bry et al., 2020)

Notations

- Let $X = [X_1, \ldots, X_R] \in \mathbb{R}^{n \times p}$ be the matrix of R thematic subsets
- Let $f_r^h = X_r u_r^h$ be the hth component of theme X_r

New demands

- Components must be explicitly identified in the themes
- Components must be orthogonal to the other components within the theme
- Components must again predict responses Y

THEME-SCGLR (Bry et al., 2020)

Notations

- Let $X = [X_1, \ldots, X_R] \in \mathbb{R}^{n \times p}$ be the matrix of R thematic subsets
- Let $f_r^h = X_r u_r^h$ be the hth component of theme X_r

New demands

- Components must be explicitly identified in the themes
- Components must be orthogonal to the other components within the theme
- Components must again predict responses Y

The THEME-SCGLR combined criterion

$$\underset{\forall r, \|u_r\|^2=1}{\text{argmax}} \; s \sum_{r=1}^{R} \ln\left(\phi(u_r)\right) + (1-s) \ln\left(\psi(u_1, \ldots, u_R, \theta)\right)$$

Estimation steps

Iterate:

Estimation steps

Iterate:

Estimation of u_r given θ

The PING algorithm allows to solve a program of the form

$$\begin{cases} \max\limits_{u_r} \quad c(u_r), \\ \text{s.t.} \quad \|u_r\|^2 = 1 \quad \text{and} \quad D^T u_r = 0, \end{cases}$$

where D is the constraint matrix of components' orthogonality

Estimation steps

Iterate:

Estimation of u_r given θ

The PING algorithm allows to solve a program of the form

$$\begin{cases} \max\limits_{u_r} \quad c(u_r), \\ \text{s.t.} \quad \|u_r\|^2 = 1 \quad \text{and} \quad D^T u_r = 0, \end{cases}$$

where D is the constraint matrix of components' orthogonality

Estimation of θ given all u_r

Maximize the likelihood on θ, *e.g.* solve

$$\nabla_\theta \, \psi(u_1, \ldots, u_R, \theta) = 0$$

Outline

Factor models

More notations

- Let J be the number of factors
- Let $g_i \sim \mathcal{N}(0, I_J)$, $i = 1, \dots n$, be the random vector of factors
- Let $B \in \mathbb{R}^{L \times q}$ be the loading matrix of factors
- Let ε_i be the error vector

Factor models

More notations

- Let J be the number of factors
- Let $g_i \sim \mathcal{N}(0, I_J)$, $i = 1, \ldots n$, be the random vector of factors
- Let $B \in \mathbb{R}^{L \times q}$ be the loading matrix of factors
- Let ε_i be the error vector

Classic linear factor model

The model expressed row-wise is given by $y_i = B^T g_i + \varepsilon_i$
of likelihood $L(Y; B) = \prod_{i=1}^{n} L(y_i; B)$

Factor models

More notations

- Let J be the number of factors
- Let $g_i \sim \mathcal{N}(0, I_J)$, $i = 1, \dots n$, be the random vector of factors
- Let $B \in \mathbb{R}^{L \times q}$ be the loading matrix of factors
- Let ε_i be the error vector

Classic linear factor model

The model expressed row-wise is given by $y_i = B^T g_i + \varepsilon_i$
of likelihood $L(Y; B) = \prod_{i=1}^{n} L(y_i; B)$

Problem 1: The model is not unique

Factor models

More notations

- Let J be the number of factors
- Let $g_i \sim \mathcal{N}(0, I_J)$, $i = 1, \dots n$, be the random vector of factors
- Let $B \in \mathbb{R}^{L \times q}$ be the loading matrix of factors
- Let ε_i be the error vector

Classic linear factor model

The model expressed row-wise is given by $y_i = B^T g_i + \varepsilon_i$
of likelihood $L(Y; B) = \prod_{i=1}^{n} L(y_i; B)$

Problem 1: The model is not unique

Problem 2: Likelihood difficult to maximize

Problem 1: The model is not unique

Identification problem

Let Ω be an orthogonal matrix ($\Omega^T\Omega = I$). The model also writes

$$y_i = B^T g_i + \varepsilon_i = B^T \Omega^T \Omega g_i + \varepsilon_i$$

with $\mathbb{E}\left[\Omega g_i\right] = \Omega \mathbb{E}\left[g_i\right] = 0$
and $\mathbb{V}\left[\Omega g_i\right] = \Omega \mathbb{V}\left[g_i\right] \Omega^T = I_J$

Problem 1: The model is not unique

Identification problem

Let Ω be an orthogonal matrix ($\Omega^T \Omega = I$). The model also writes

$$y_i = B^T g_i + \varepsilon_i = B^T \Omega^T \Omega g_i + \varepsilon_i$$

with $\mathbb{E}[\Omega g_i] = \Omega \mathbb{E}[g_i] = 0$
and $\mathbb{V}[\Omega g_i] = \Omega \mathbb{V}[g_i] \Omega^T = I_J$
$\rightarrow$ We get the same distribution !!!

Problem 1: The model is not unique

Identification problem

Let Ω be an orthogonal matrix ($\Omega^T\Omega = I$). The model also writes

$$y_i = B^T g_i + \varepsilon_i = B^T \Omega^T \Omega g_i + \varepsilon_i$$

with $\mathbb{E}\left[\Omega g_i\right] = \Omega\mathbb{E}\left[g_i\right] = 0$
and $\mathbb{V}\left[\Omega g_i\right] = \Omega\mathbb{V}\left[g_i\right]\Omega^T = I_J$
$\rightarrow$ We get the same distribution !!!

Geweke and Zhou (1996) assure the uniqueness of the solution by imposing a upper triangle constraint on the matrix B:

$$B = \begin{pmatrix} b_{11} & \dots & b_{1J} & b_{1,J+1} & \dots & b_{1q} \\ & \ddots & \vdots & \vdots & \ddots & \vdots \\ & & b_{JJ} & b_{J,J+1} & \dots & b_{Jq} \end{pmatrix}.$$

Problem 2: Likelihood difficult to maximize

We perform the Expectation-Maximization (EM) algorithm.

It allows to:

Problem 2: Likelihood difficult to maximize

We perform the Expectation-Maximization (EM) algorithm.
It allows to:

- Maximize a likelihood in presence of random latent variables
  → Here: the factors

Problem 2: Likelihood difficult to maximize

We perform the Expectation-Maximization (EM) algorithm.
It allows to:

- Maximize a likelihood in presence of random latent variables
  $\rightarrow$ Here: the factors
- Estimate the model's parameters
  $\rightarrow$ Here: the matrix B

Outline

Factors SCGLR

Component-based model with factors

$$\eta = \underbrace{(X_1 u_1)\,\gamma_1 + \cdots + (X_R u_R)\,\gamma_R}_{\text{deterministic}} + \underbrace{G B}_{\text{stochastic}}$$

where the $X_r u_r$'s are the components, $\Gamma = [\gamma_1, \ldots, \gamma_R]$ the regression parameters and G the realizations of the factors.

The likelihood writes $L(Y; u_1, \ldots, u_R, \Gamma, B)$

Factors SCGLR

Component-based model with factors

$$\eta = \underbrace{(X_1 u_1)\,\gamma_1 + \cdots + (X_R u_R)\,\gamma_R}_{\text{deterministic}} + \underbrace{G B}_{\text{stochastic}}$$

where the $X_r u_r$'s are the components, $\Gamma = [\gamma_1, \ldots, \gamma_R]$ the regression parameters and G the realizations of the factors.
The likelihood writes $L(Y; u_1, \ldots, u_R, \Gamma, B)$

Combined criterion

$$\underset{\forall r,\, \|u_r\|^2=1}{\operatorname{argmax}} \; s \sum_{r=1}^{R} \ln\left(\phi(u_r)\right) + (1-s)\ln\left(L(Y; u_1, \ldots, u_R, \Gamma, B)\right)$$

Factors SCGLR

Component-based model with factors

$$\eta = \underbrace{(X_1 u_1)\,\gamma_1 + \cdots + (X_R u_R)\,\gamma_R}_{\text{deterministic}} + \underbrace{G\,B}_{\text{stochastic}}$$

where the $X_r u_r$'s are the components, $\Gamma = [\gamma_1, \ldots, \gamma_R]$ the regression parameters and G the realizations of the factors.
The likelihood writes $L(Y; u_1, \ldots, u_R, \Gamma, B)$

Combined criterion

$$\underset{\forall r,\, \|u_r\|^2=1}{\text{argmax}} \; s \sum_{r=1}^{R} \ln\left(\phi(u_r)\right) + (1-s)\ln\left(L(Y; u_1, \ldots, u_R, \Gamma, B)\right)$$

Estimation steps

The overall algorithm consists in alternating the following steps:

- We find $\{\Gamma, B\}$ through the EM algorithm
- We find all u_r through the PING algorithm

Outline

Deterministic simulation

Response variables

$Y = [y_1, \dots, y_{50}]$ is composed by 20 Gaussian responses, 20 Poisson responses and 10 Bernoulli responses

Deterministic simulation

Response variables

$Y = [y_1, \dots, y_{50}]$ is composed by 20 Gaussian responses, 20 Poisson responses and 10 Bernoulli responses

Explanatory variables

$$X_1 = [\underbrace{x_1, \dots, x_{60}}_{:=\mathcal{X}_1} \,|\, \underbrace{x_{61}, \dots, x_{100}}_{:=\mathcal{X}_2}] \text{ and } X_2 = [\underbrace{x_{101}, \dots, x_{160}}_{:=\mathcal{X}_3} \,|\, \underbrace{x_{161}, \dots, x_{200}}_{:=\mathcal{X}_4}]$$

- Theme X_1 is composed by two explanatory bundles
- Theme X_2 is composed by two explanatory bundles
- Bundles are sets of correlated variables

Stochastic simulation

Factors

We generate 3 factors to model the covariance between the responses

Stochastic simulation

Factors

We generate 3 factors to model the covariance between the responses

Loading matrix

We generate the matrix B of the form

$$B = [\underbrace{b_1, \ldots, b_5}_{\sim\mathcal{N}(\mu_1, 0.1 I_3)} \mid \underbrace{b_6, \ldots, b_{10}}_{\sim\mathcal{N}(-\mu_1, 0.1 I_3)} \mid \underbrace{b_{11}, \ldots, b_{20}}_{\sim\mathcal{N}(\mu_2, 0.1 I_3)} \mid \underbrace{b_{21}, \ldots, b_{35}}_{\sim\mathcal{N}(-\mu_2, 0.1 I_3)} \mid \underbrace{b_{36}, \ldots, b_{50}}_{\sim\mathcal{N}(\mu_3, 0.1 I_3)}],$$

where $\mu_1 = (2, 0, 0)^T$, $\mu_2 = (0, -1, 0)^T$ and $\mu_3 = (0, 0, 1.5)^T$.

Simulation: variance-covariance matrix

The variance-covariance matrix of the responses $B^T B$ becomes

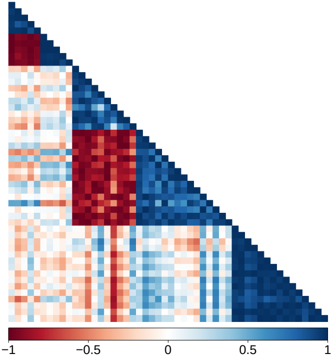


Figure 1: Conditional variance-covariance matrix

Simulation: variance-covariance matrix

The variance-covariance matrix of the responses $B^T B$ becomes

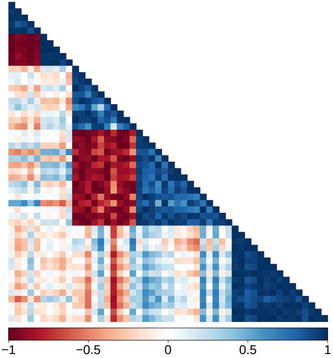


Figure 1: Conditional variance-covariance matrix

Question: How to identify the blocks ?

Classification steps

We perform several *a posteriori* steps:

Classification steps

We perform several *a posteriori* steps:

1. We estimate the variance-covariance matrix $B^T B$

Classification steps

We perform several *a posteriori* steps:

1. We estimate the variance-covariance matrix $B^T B$
2. We estimate the correlation matrix

Classification steps

We perform several *a posteriori* steps:

1. We estimate the variance-covariance matrix $B^T B$
2. We estimate the correlation matrix
3. We calculate a dissimilarity matrix from the correlations: $d(x, y) = 2(1 - \text{cor}^2(x, y))$

Classification steps

We perform several *a posteriori* steps:

1. We estimate the variance-covariance matrix $B^T B$
2. We estimate the correlation matrix
3. We calculate a dissimilarity matrix from the correlations: $d(x, y) = 2(1 - \text{cor}^2(x, y))$
4. We perform the Multidimensional Scaling (MDS) on the dissimilarity matrix

Classification steps

We perform several *a posteriori* steps:

1. We estimate the variance-covariance matrix $B^T B$
2. We estimate the correlation matrix
3. We calculate a dissimilarity matrix from the correlations: $d(x, y) = 2(1 - \text{cor}^2(x, y))$
4. We perform the Multidimensional Scaling (MDS) on the dissimilarity matrix
5. We perform the K-means on the output of the MDS

Results

Bundles recovery

Components of theme 1	
$\mathrm{cor}^2(\mathcal{X}_1, f_1^1)$	0.984
$\mathrm{cor}^2(\mathcal{X}_2, f_1^2)$	0.979

Components of theme 2	
$\mathrm{cor}^2(\mathcal{X}_3, f_2^1)$	0.975
$\mathrm{cor}^2(\mathcal{X}_4, f_2^2)$	0.983

Results

Bundles recovery

Components of theme 1	
$\text{cor}^2(\mathcal{X}_1, f_1^1)$	0.984
$\text{cor}^2(\mathcal{X}_2, f_1^2)$	0.979

Components of theme 2	
$\text{cor}^2(\mathcal{X}_3, f_2^1)$	0.975
$\text{cor}^2(\mathcal{X}_4, f_2^2)$	0.983

Correctness of classification steps

- Rand Index: 0.948
- Adjusted Rand Index: 0.904

Results

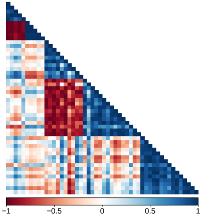
Bundles recovery

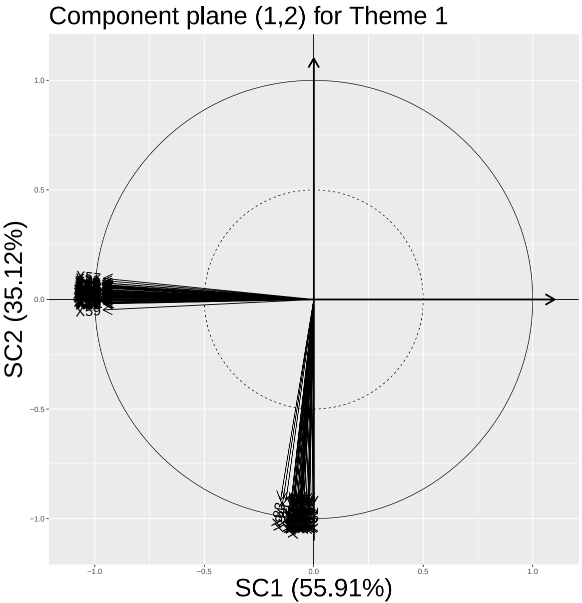
Components of theme 1	
$\mathrm{cor}^2(\mathcal{X}_1, f_1^1)$	0.984
$\mathrm{cor}^2(\mathcal{X}_2, f_1^2)$	0.979

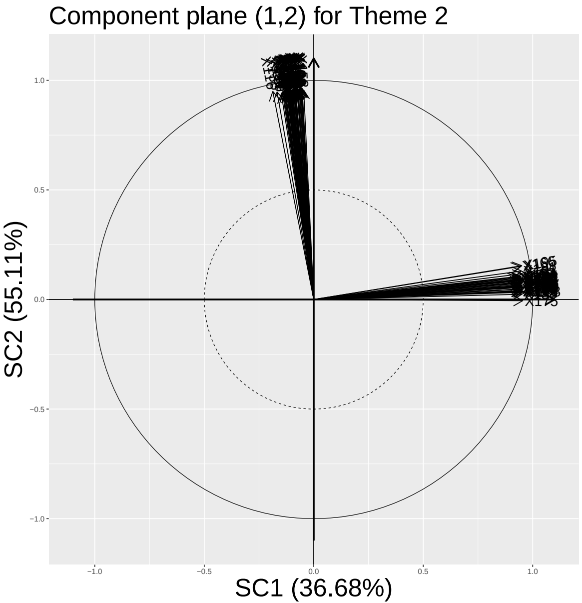
Components of theme 2	
$\mathrm{cor}^2(\mathcal{X}_3, f_2^1)$	0.975
$\mathrm{cor}^2(\mathcal{X}_4, f_2^2)$	0.983

Correctness of classification steps

- Rand Index: 0.948
- Adjusted Rand Index: 0.904




Component plane (1,2) for Theme 1
SC2 (35.12%)
SC1 (55.91%)
1.0
0.5
0.0
−0.5
−1.0
X57
X59


Component plane (1,2) for Theme 2
SC2 (55.11%)
SC1 (36.68%)
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

Real data

The *Genus* dataset

- 27 species abundances (Y matrix)
- 36 explanatory variables (X matrix)
  - ↬ subset of 23 photosynthesis variables “evi” (Theme 1)
  - ↬ subset of 13 rainfall variables “pluvio” (Theme 2)

Real data

The *Genus* dataset

- 27 species abundances (Y matrix)
- 36 explanatory variables (X matrix)
  - ↬ subset of 23 photosynthesis variables “evi” (Theme 1)
  - ↬ subset of 13 rainfall variables “pluvio” (Theme 2)

Results

Clusters	Responses
1	$Y_1,\ Y_5,\ Y_7,\ Y_9,\ Y_{12},\ Y_{15},\ Y_{26},\ Y_{27}$
2	$Y_2,\ Y_8,\ Y_{23},\ Y_{24}$
3	$Y_3,\ Y_{13}$
4	$Y_4,\ Y_{19}$
5	$Y_6,\ Y_{16},\ Y_{22},\ Y_{25}$
6	$Y_{10},\ Y_{18},\ Y_{20}$
7	$Y_{11},\ Y_{14}$
8	$Y_{17},\ Y_{21}$

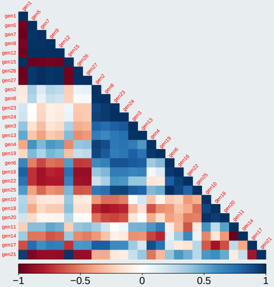
Real data

The *Genus* dataset

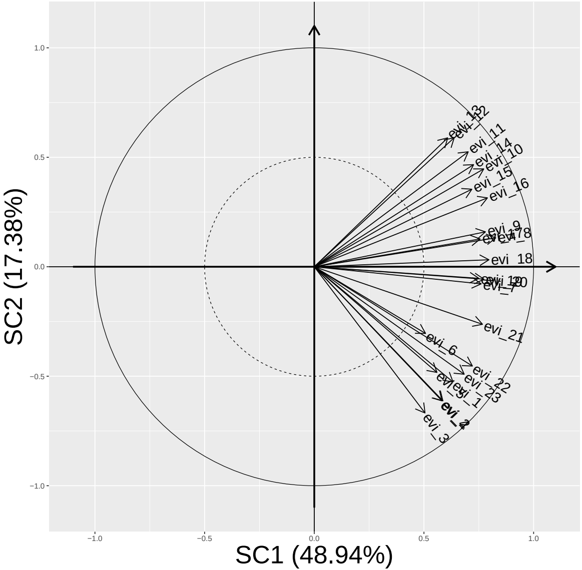
- 27 species abundances (Y matrix)
- 36 explanatory variables (X matrix)
  - ↬ subset of 23 photosynthesis variables "evi" (Theme 1)
  - ↬ subset of 13 rainfall variables "pluvio" (Theme 2)

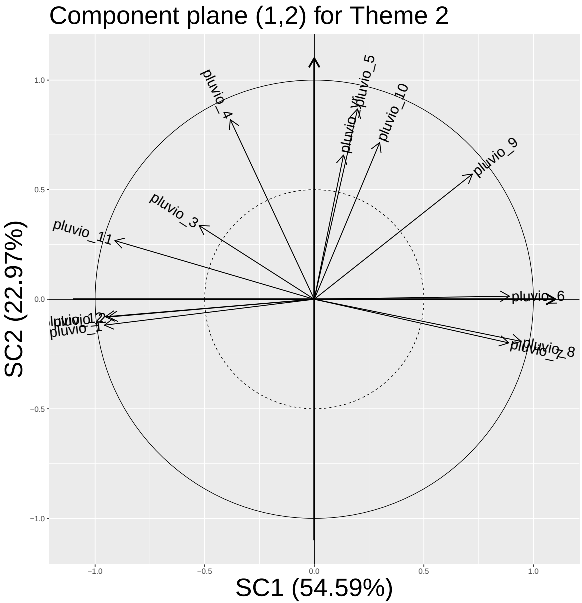
Results

Clusters	Responses
1	$Y_1,\ Y_5,\ Y_7,\ Y_9,\ Y_{12},\ Y_{15},\ Y_{26},\ Y_{27}$
2	$Y_2,\ Y_8,\ Y_{23},\ Y_{24}$
3	$Y_3,\ Y_{13}$
4	$Y_4,\ Y_{19}$
5	$Y_6,\ Y_{16},\ Y_{22},\ Y_{25}$
6	$Y_{10},\ Y_{18},\ Y_{20}$
7	$Y_{11},\ Y_{14}$
8	$Y_{17},\ Y_{21}$



Component plane (1,2) for Theme 1




Component plane (1,2) for Theme 2
SC2 (22.97%)
SC1 (54.59%)
pluvio_4
pluvio_5
pluvio_10
pluvio_9
pluvio_3
pluvio_11
pluvio_6
pluvio_8
pluvio_7
pluvio_12
pluvio_1
1.0
0.5
0.0
−0.5
−1.0

Conclusion

We have:

- Extended SCGLR to the factor model
- Developed an algorithm allowing to find relevant components and to model the variance-covariance matrix

Conclusion

We have:

- Extended SCGLR to the factor model
- Developed an algorithm allowing to find relevant components and to model the variance-covariance matrix

Perspectives

We want to:

- Add new distributions for the responses
- Better identify blocks in the variance-covariance matrix

Acknowledgments and references

Thank you for your attention

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

Geweke, J. and Zhou, G. (1996). Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies*, 9(2):557–587.