Response clustering in component-based GLM

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

[1] IMAG, CNRS, Univ. Montpellier, France.
[2] AMIS, Univ. Paul-Valéry Montpellier 3, Montpellier, France.

ASMDA 2021

Outline

Outline

Motivations

Ecological motivations

In a context of global warming, we aim at

- Finding the main determinants of species observations, among a large number of explanatory variables
- Identifying species communities influenced by common determinants

Motivations

Ecological motivations

In a context of global warming, we aim at

- Finding the main determinants of species observations, among a large number of explanatory variables
- Identifying species communities influenced by common determinants

Statistical counterparts

- Finding strong dimensions allowing to explain the responses as best as possible
  - ↬ Searching for supervised components
- Identifying groups of responses with explanatory dimensions specific to each group
  - ↬ Clustering

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 (species)
- 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 (species)
- 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 $h = 1, \ldots, H$, and $F = [f_1, \ldots, f_H]$
- $f_h \perp f_g$, for all $h \neq g$

What is a supervised component ?

Notations

- Let $Y = [y_1, \dots, y_q] \in \mathbb{R}^{n \times q}$ be the matrix of responses (species)
- Let $X = [x_1, \dots, x_p] \in \mathbb{R}^{n \times p}$ 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 $h = 1, \dots, H$, and $F = [f_1, \dots, f_H]$
- $f_h \perp f_g$, 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., 2013)

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., 2013)

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., 2013)

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

Estimation steps

Iterate :

Estimation steps

Iterate :

Estimation of u given θ

The PING algorithm allows to solve a program of the form

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

where D is the constraint matrix of components' orthogonality

Estimation steps

Iterate :

Estimation of u given θ

The PING algorithm allows to solve a program of the form

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

where D is the constraint matrix of components' orthogonality

Estimation of θ given u

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

$$\nabla_\theta \psi(u, \theta) = 0$$

Outline

Mixture of responses (Dunstan et al., 2013)

More notations

- Let G be the number of groups
- Let z_{kg} be the latent dummy variable equal to 1 if the response y_k belongs to the group g
- Let p_g be the *a priori* probability of belonging to the group g

Mixture of responses (Dunstan et al., 2013)

More notations

- Let G be the number of groups
- Let z_{kg} be the latent dummy variable equal to 1 if the response y_k belongs to the group g
- Let p_g be the *a priori* probability of belonging to the group g

Model

Conditionally to $z_{kg} = 1$, $y_k \sim \mathbb{P}(\theta_g)$ with pdf $d(y_k; \theta_g)$.
The model likelihood writes : $\psi(u_1, \ldots, u_G; \Theta) = \prod_{k=1}^{q} \sum_{g=1}^{G} p_g d(y_k; \theta_g)$,
with θ_g including :

- the loading vectors u_g specific to each group
- the parameters of within cluster regression model of responses on components

Mixture of responses (Dunstan et al., 2013)

More notations

- Let G be the number of groups
- Let z_{kg} be the latent dummy variable equal to 1 if the response y_k belongs to the group g
- Let p_g be the *a priori* probability of belonging to the group g

Model

Conditionally to $z_{kg} = 1$, $y_k \sim \mathbb{P}(\theta_g)$ with pdf $d(y_k; \theta_g)$.
The model likelihood writes : $\psi(u_1, \ldots, u_G; \Theta) = \prod_{k=1}^{q} \sum_{g=1}^{G} p_g d(y_k; \theta_g)$,
with θ_g including :

- the loading vectors u_g specific to each group
- the parameters of within cluster regression model of responses on components

Problem : this likelihood is hard to maximize

Estimation (Dempster et al., 1977)

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

- Maximize a likelihood in the presence of latent variables
  $\rightarrow$ Here : the dummy variable z_{kg}
- Estimate the posterior distribution of each z_k conditional on the observations
  $\rightarrow$ Here : the posterior group membership probabilities of each response

Outline

A new enhanced criterion

The separation criterion

We introduce the criterion $\varphi(u_1, \dots, u_G)$ in order to separate the explanatory spaces

A new enhanced criterion

The separation criterion

We introduce the criterion $\varphi(u_1, \ldots, u_G)$ in order to separate the explanatory spaces

Maximized criterion

$$\underset{\forall g,\, \|u_g\|^2=1}{\operatorname{argmax}} \; s \sum_{g=1}^{G} \ln\left(\phi(u_g)\right) + t \ln\left(\varphi(u_1, \ldots, u_G)\right) + (1 - s - t) \ln\left(\psi(u_1, \ldots, u_G; \Theta)\right)$$

A new enhanced criterion

The separation criterion

We introduce the criterion $\varphi(u_1, \ldots, u_G)$ in order to separate the explanatory spaces

Maximized criterion

$$\underset{\forall g,\, \|u_g\|^2=1}{\operatorname{argmax}} \; s \sum_{g=1}^{G} \ln\left(\phi(u_g)\right) + t \ln\left(\varphi(u_1, \ldots, u_G)\right) + (1 - s - t) \ln\left(\psi(u_1, \ldots, u_G; \Theta)\right)$$

Estimation

We alternate on the two maximization steps :

- Find Θ through the EM algorithm
- Find u_g through the PING algorithm, for all g

Outline

Simulation study

Random responses

$$Y = [\underbrace{y_1, \ldots, y_{20}}_{\text{G1 : Gaussian}} \mid \underbrace{y_{21}, \ldots, y_{50}}_{\text{G2 : Poisson}} \mid \underbrace{y_{51}, \ldots, y_{100}}_{\text{G3 : Bernoulli}}]$$

Y is composed in 3 groups of responses

Simulation study

Random responses

$$Y = [\underbrace{y_1, \ldots, y_{20}}_{\text{G1 : Gaussian}} \mid \underbrace{y_{21}, \ldots, y_{50}}_{\text{G2 : Poisson}} \mid \underbrace{y_{51}, \ldots, y_{100}}_{\text{G3 : Bernoulli}}]$$

Y is composed in 3 groups of responses

Explanatory variables

$$X = [\underbrace{x_1, \ldots, x_{50}}_{\text{X1 : predict G1}} \mid \underbrace{x_{51}, \ldots, x_{90}}_{\text{X2 : predict G2}} \mid \underbrace{x_{91}, \ldots, x_{120}}_{\text{X3 : predict G3}} \mid \underbrace{x_{121}, \ldots, x_{140}}_{\text{X4 : predict G1}} \mid \underbrace{x_{141}, \ldots, x_{150}}_{\text{X5 : predict G2}} \mid \underbrace{x_{151}, \ldots, x_{200}}_{\text{X6 : noise}}]$$

- X is composed in 5 bundles plus 1 set of noise
- Bundles $X1$, $X3$ and $X5$ are weakly correlated (cor $= 0.5$)

Results obtained with the best classification index

Hyperparameters

The optimized parameters are $s = 0.1$ and $t = 0.6$

Results obtained with the best classification index

Hyperparameters

The optimized parameters are $s = 0.1$ and $t = 0.6$

Posterior group membership probabilities

- $\forall i = 1, \ldots, 20$, $\mathbb{P}(y_i$ belongs to the group 1$) > 0.9$
- $\forall i = 21, \ldots, 50$, $\mathbb{P}(y_i$ belongs to the group 2$) > 0.9$
- $\forall i = 51, \ldots, 100$, $\mathbb{P}(y_i$ belongs to the group 3$) > 0.9$

Results obtained with the best classification index

Hyperparameters

The optimized parameters are $s = 0.1$ and $t = 0.6$

Posterior group membership probabilities

- $\forall i = 1, \ldots, 20$, $\mathbb{P}(y_i$ belongs to the group 1$) > 0.9$
- $\forall i = 21, \ldots, 50$, $\mathbb{P}(y_i$ belongs to the group 2$) > 0.9$
- $\forall i = 51, \ldots, 100$, $\mathbb{P}(y_i$ belongs to the group 3$) > 0.9$

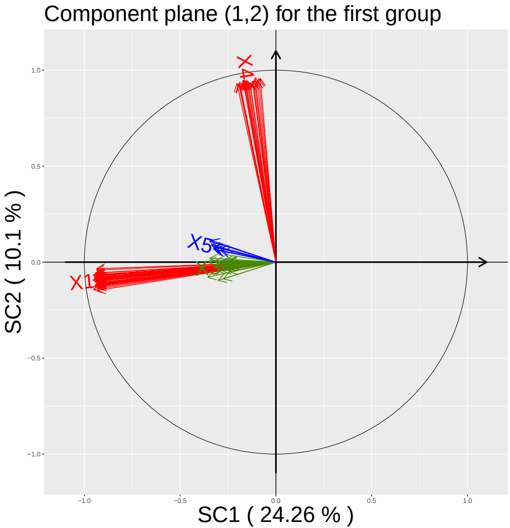
Correlations

group 1	
$\mathrm{cor}^2(X_1, f_1)$	0.960
$\mathrm{cor}^2(X_4, f_2)$	0.966

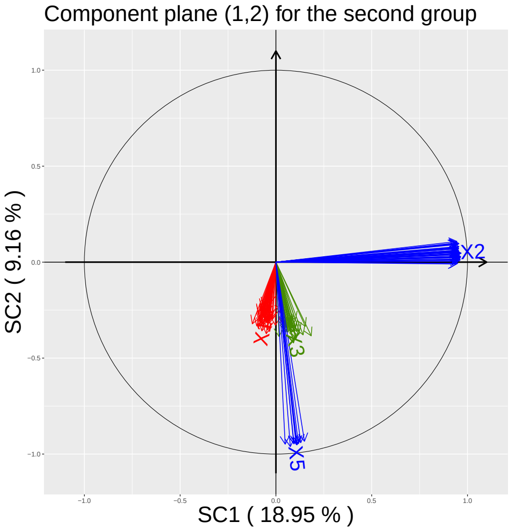
group 2	
$\mathrm{cor}^2(X_2, f_1)$	0.993
$\mathrm{cor}^2(X_5, f_2)$	0.911

group 3	
$\mathrm{cor}^2(X_3, f_1)$	0.987

Correlation scatterplot for the first group



Correlation scatterplot for the second group



Real data

The *Genus* dataset

- 27 species abundances (Y matrix)
- 39 explanatory variables (X matrix)
  - ↬ subset of 13 rainfall variables ("pluvio")
  - ↬ subset of 23 photosynthesis variables ("evi")
  - ↬ subset of 3 location variables

Real data

The *Genus* dataset

- 27 species abundances (Y matrix)
- 39 explanatory variables (X matrix)
  - subset of 13 rainfall variables ("pluvio")
  - subset of 23 photosynthesis variables ("evi")
  - subset of 3 location variables

Results

Groups	Responses	Explanatory variables
1	Y_4	"pluvio1", "pluvio12"
2	Y_8, Y_{19}	"pluvio7", "pluvio6"
3	Y_1, Y_3, Y_5, Y_7, Y_{11}, Y_{12}, Y_{13}, Y_{14}, Y_{16} Y_{21}, Y_{24}, Y_{25}, Y_{26}, Y_{27}	"pluvio8", "pluvio1"
4	Y_9	"pluvio10", "evi13"
5	Y_6, Y_{15}, Y_{18}, Y_{22}, Y_{23}	"pluvio7", "pluvio8"
6	Y_2, Y_{10}, Y_{17}, Y_{20}	"pluvio7", "pluvio11"

Conclusion

We have :

- Extended SCGLR to response mixture
- Enhanced the combined criterion
- Developed an algorithm able to find groups of responses predicted by specific explanatory spaces

Acknowledgments and bibliography

Thank you for your attention

Bry, X., Trottier, C., Verron, T., and Mortier, F. (2013). Supervised component generalized linear regression using a PLS-extension of the Fisher scoring algorithm. *Journal of Multivariate Analysis*, 119 : 47–60.

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

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