Régression linéaire généralisée sur composantes supervisées pour la modélisation jointe des réponses

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

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

Séminaire des doctorant.es

Sommaire

Motivations

Motivations écologiques

Dans un contexte de réchauffement climatique on cherche à

- Trouver les déterminants principaux des abondances d'espèces, parmi un grand nombre de variables
- Identifier des communautés d'espèces influencées par des déterminants communs

Motivations

Motivations écologiques

Dans un contexte de réchauffement climatique on cherche à

- Trouver les déterminants principaux des abondances d'espèces, parmi un grand nombre de variables
- Identifier des communautés d'espèces influencées par des déterminants communs

Traduction statistique

- Trouver des dimensions fortes permettant d'expliquer au mieux les réponses
  - ↬ Recherche de composantes supervisées
- Identifier des groupes de réponses partageant des dimensions explicatives propres à chaque groupe
  - ↬ Classification

Sommaire

Qu'est-ce qu'une composante ?

Notations

- $Y = [y_1, \ldots, y_q] \in \mathbb{R}^{n \times q}$ la matrice donnée des variables **réponses** (abondances d'espèces)
- $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n \times p}$ la matrice donnée des variables **explicatives**

Qu'est-ce qu'une composante ?

Notations

- $Y = [y_1, \ldots, y_q] \in \mathbb{R}^{n \times q}$ la matrice donnée des variables **réponses** (abondances d'espèces)
- $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n \times p}$ la matrice donnée des variables **explicatives**

Définition

Une composante est un vecteur f permettant de synthétiser l'information contenue dans les données et respectant

- $f_h = Xu_h$, pour tout $h = 1, \ldots, H$, et $F = [f_1, \ldots, f_H]$
- $f_h \perp f_g$, pour tout $h \neq g$

Qu'est-ce qu'une composante ?

Notations

- $Y = [y_1, \ldots, y_q] \in \mathbb{R}^{n \times q}$ la matrice donnée des variables **réponses** (abondances d'espèces)
- $X = [x_1, \ldots, x_p] \in \mathbb{R}^{n \times p}$ la matrice donnée des variables **explicatives**

Définition

Une composante est un vecteur f permettant de synthétiser l'information contenue dans les données et respectant

- $f_h = Xu_h$, pour tout $h = 1, \ldots, H$, et $F = [f_1, \ldots, f_H]$
- $f_h \perp f_g$, pour tout $h \neq g$

Exigences

- Les composantes doivent être proches des variables pour être interprétables
- Elles doivent bien prédire les réponses Y

Qu'est-ce qu'une composante supervisée ?

Les composantes supervisées sont des composantes permettant d'expliquer une variable réponse.

$$\forall k = 1, \ldots, q, \quad y_k \sim \mathbb{P}_{y_k}(\eta_k)$$

avec $\eta_k = F\gamma_k$, où $F = XU$ et où γ_k est un vecteur de coefficients (effets des composantes sur la réponse y_k). $\mathbb{P}_{y_k}$ est une loi de probabilité adaptée à la réponse.

Rappel

Nous voulons :

Rappel

Nous voulons :

- des composantes ...

Rappel

Nous voulons :

- des composantes ...
- ... qui puissent expliquer les réponses ...

Rappel

Nous voulons :

- des composantes ...
- ... qui puissent expliquer les réponses ...
- ... et qui soient interprétables.

SCGLR (Bry et al., 2018)

Pertinence structurelle (SR)

Le critère $\phi(u)$ mesure la "force" de la composante Xu (proximité aux variables x_j)

SCGLR (Bry et al., 2018)

Pertinence structurelle (SR)

Le critère $\phi(u)$ mesure la "force" de la composante Xu (proximité aux variables x_j)

Qualité d'ajustement (GoF)

Le critère $\psi(u, \theta)$ est la vraisemblance du modèle à composantes

$\rightarrow$ vraisemblance du modèle = probabilité que le modèle donne aux observations réalisées

SCGLR (Bry et al., 2018)

Pertinence structurelle (SR)

Le critère $\phi(u)$ mesure la "force" de la composante Xu (proximité aux variables x_j)

Qualité d'ajustement (GoF)

Le critère $\psi(u, \theta)$ est la vraisemblance du modèle à composantes

$\rightarrow$ vraisemblance du modèle = probabilité que le modèle donne aux observations réalisées

$$\Downarrow$$

Critère SCGLR

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

Le réel $s \in [0, 1]$ permet de régler un compromis entre SR et GoF

Méthodes d'estimation

Estimation de u

L'algorithme PING permet de résoudre les programmes de la forme

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

avec D la matrice de contrainte d'orthogonalité des composantes

Méthodes d'estimation

Estimation de u

L'algorithme PING permet de résoudre les programmes de la forme

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

avec D la matrice de contrainte d'orthogonalité des composantes

Estimation de θ

On trouve θ par maximum de vraisemblance, *i.e.* on résout

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

Sommaire

Qu'est-ce que la classification ?

Idée

Classifier des objets statistiques dans des groupes

$\rightarrow$ Objets : individus, variables, réponses...

Qu'est-ce que la classification ?

Exemples

Il existe plusieurs méthodes de classification :

- Méthodes de partitionnement
  - → Ex : K-means, CAH,...
- Méthodes par probabilité d'appartenance
  - → Ex : modèle de mélange,...

Qu'est-ce que la classification ?

Exemples

Il existe plusieurs méthodes de classification :

- Méthodes de partitionnement
  - → Ex : K-means, CAH,...
- Méthodes par probabilité d'appartenance
  - → Ex : modèle de mélange,...

Objectifs

- Objets à grouper : réponses
- Méthode : modèle de mélange

Sommaire

Modèle de mélange de réponses

Encore plus de notations

- G le nombre de groupes
- z_{kg} la variable indicatrice latente qui vaut 1 si la réponse y_k est dans le groupe g
- p_g la probabilité *a priori* qu'une réponse soit dans le groupe g

Modèle de mélange de réponses

Encore plus de notations

- G le nombre de groupes
- z_{kg} la variable indicatrice latente qui vaut 1 si la réponse y_k est dans le groupe g
- p_g la probabilité *a priori* qu'une réponse soit dans le groupe g

Modèle

Conditionnellement à $z_{kg} = 1$, $y_k \sim \mathbb{P}(\theta_g)$ de densité $d(y_k; \theta_g)$.
La vraisemblance du modèle s'écrit : $\prod_{k=1}^{q} \sum_{g=1}^{G} p_g d(y_k; \theta_g)$,
avec θ_g comprenant :

- les vecteurs u_g propres à chaque groupe
- les γ_{kg} (effets de la composante $f_g = Xu_g$ sur y_k)

Modèle de mélange de réponses

Encore plus de notations

- G le nombre de groupes
- z_{kg} la variable indicatrice latente qui vaut 1 si la réponse y_k est dans le groupe g
- p_g la probabilité *a priori* qu'une réponse soit dans le groupe g

Modèle

Conditionnellement à $z_{kg} = 1$, $y_k \sim \mathbb{P}(\theta_g)$ de densité $d(y_k; \theta_g)$.
La vraisemblance du modèle s'écrit : $\prod_{k=1}^{q} \sum_{g=1}^{G} p_g d(y_k; \theta_g)$,
avec θ_g comprenant :

- les vecteurs u_g propres à chaque groupe
- les γ_{kg} (effets de la composante $f_g = Xu_g$ sur y_k)

Problème : Vraisemblance compliquée à maximiser

Méthode d'estimation (Dempster et al., 1977)

On utilise l'algorithme Expectation-Maximization (EM).
Il permet de :

- Maximiser une vraisemblance en présence de variables latentes
  - → Ici : la variable nominale de groupe z_{kg}
- Estimer la loi conditionnelle de chaque z_k sachant les observations
  - → Ici : les probabilités *a posteriori* d'appartenance aux groupes de chaque réponse

Response-Mixture SCGLR

Vraisemblance

$$\psi(u_1, \ldots, u_G, Y; \Xi) = \prod_{k=1}^{q} \sum_{g=1}^{G} p_g \mathcal{N}_n(\mu_{kg}, \Sigma_g), \quad \mu_{kg} = X u_g \gamma_{kg}, \quad \Sigma_g = \sigma_g^2 I$$

Response-Mixture SCGLR

Vraisemblance

$$\psi(u_1, \ldots, u_G, Y; \Xi) = \prod_{k=1}^{q} \sum_{g=1}^{G} p_g \mathcal{N}_n(\mu_{kg}, \Sigma_g), \quad \mu_{kg} = X u_g \gamma_{kg}, \quad \Sigma_g = \sigma_g^2 I$$

Critère à maximiser

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

Response-Mixture SCGLR

Vraisemblance

$$\psi(u_1, \ldots, u_G, Y; \Xi) = \prod_{k=1}^{q} \sum_{g=1}^{G} p_g \mathcal{N}_n(\mu_{kg}, \Sigma_g), \quad \mu_{kg} = X u_g \gamma_{kg}, \quad \Sigma_g = \sigma_g^2 I$$

Critère à maximiser

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

Méthode d'estimation

On alterne itérativement ces deux maximisations :

- On trouve $\Xi = \{(p_g)_g, (\gamma_{kg})_{kg}, (\sigma_g^2)_g\}$ à l'aide de l'algorithme EM
- Pour tout g, on trouve u_g à l'aide de l'algorithme PING

Simulations

Simulation des réponses

On simule quatre variables latentes φ_i, pour $i = 1, \ldots, 4$.
Les variables latentes φ_1 et φ_2 sont corrélées à 0.5.
On simule deux groupes de réponses de la manière suivante :

$$\forall i = 1, \ldots, 40, \quad y_i = \gamma_{i,1}\varphi_1 + \gamma_{i,2}\varphi_3 + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, I_n)$$

$$\forall i = 41, \ldots, 100, \quad y_i = \gamma_{i,1}\varphi_2 + \gamma_{i,2}\varphi_4 + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, I_n)$$

où, pour tout i, $-4 < \gamma_{i,1} < 4$
et, pour tout i, $-2 < \gamma_{i,2} < 2$.

Simulations

Simulation des réponses

On simule quatre variables latentes φ_i, pour $i = 1, \ldots, 4$.
Les variables latentes φ_1 et φ_2 sont corrélées à 0.5.
On simule deux groupes de réponses de la manière suivante :

$$\forall i = 1, \ldots, 40, \quad y_i = \gamma_{i,1}\varphi_1 + \gamma_{i,2}\varphi_3 + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, I_n)$$

$$\forall i = 41, \ldots, 100, \quad y_i = \gamma_{i,1}\varphi_2 + \gamma_{i,2}\varphi_4 + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, I_n)$$

où, pour tout i, $-4 < \gamma_{i,1} < 4$
et, pour tout i, $-2 < \gamma_{i,2} < 2$.

Simulation des variables explicatives

La matrice X contient 30 variables divisées en cinq blocs :
$X = [B_1, B_2, B_3, B_4, B_5]$
Pour $i = 1, \ldots, 4$, B_i est un faisceau de cinq variables centrées en φ_i.
B_5 est un bloc de dix variables sans rôle prédictif.

Résultats

Probabilité *a posteriori* d'inclusion

- $\forall i = 1, \ldots, 40$, $\mathbb{P}(y_i \text{ est dans le groupe } 1) > 0.999$
- $\forall i = 41, \ldots, 100$, $\mathbb{P}(y_i \text{ est dans le groupe } 2) > 0.999$

Résultats

Probabilité *a posteriori* d'inclusion

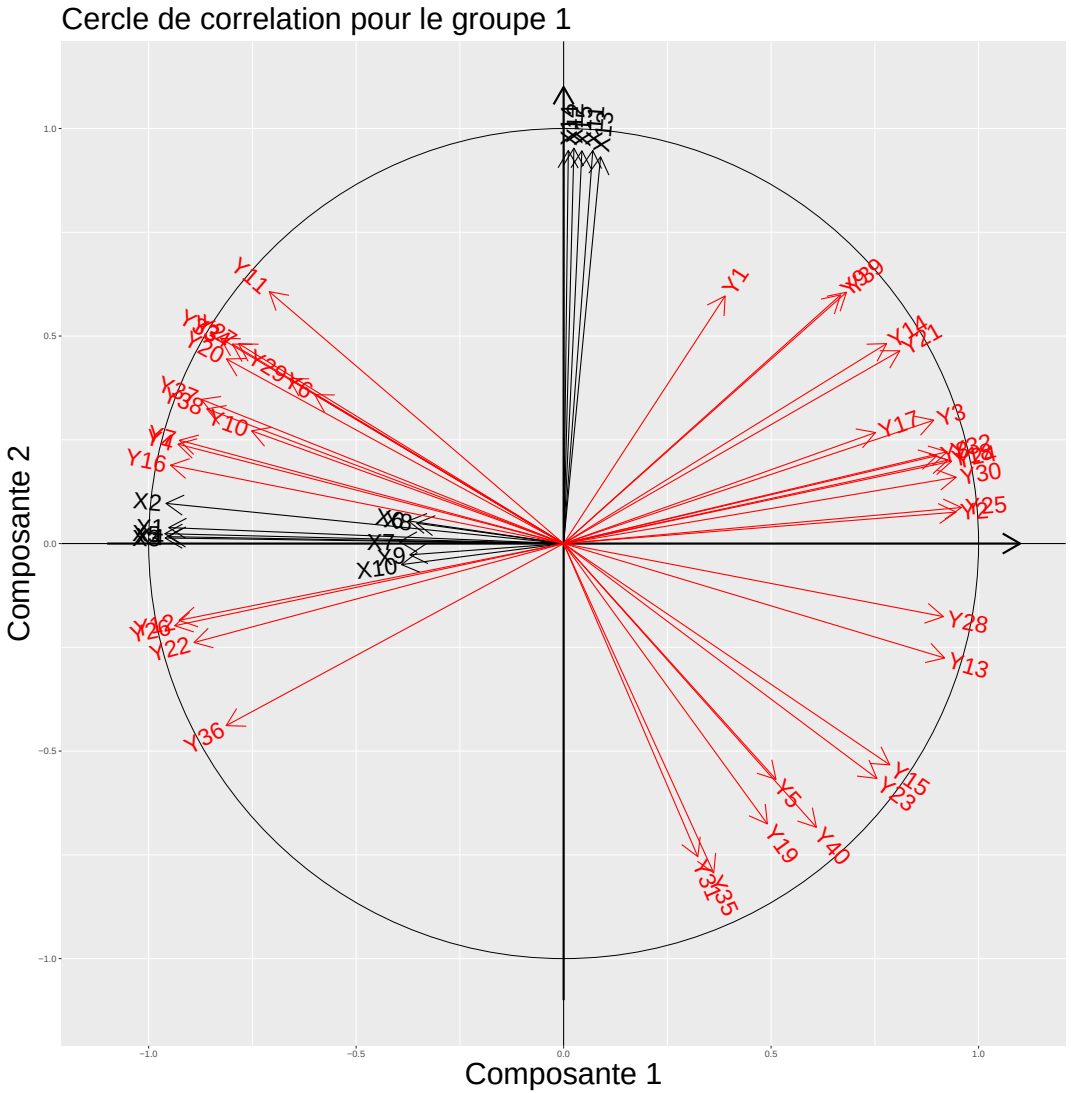
- $\forall i = 1, \ldots, 40$, $\mathbb{P}(y_i \text{ est dans le groupe 1}) > 0.999$
- $\forall i = 41, \ldots, 100$, $\mathbb{P}(y_i \text{ est dans le groupe 2}) > 0.999$

Corrélation

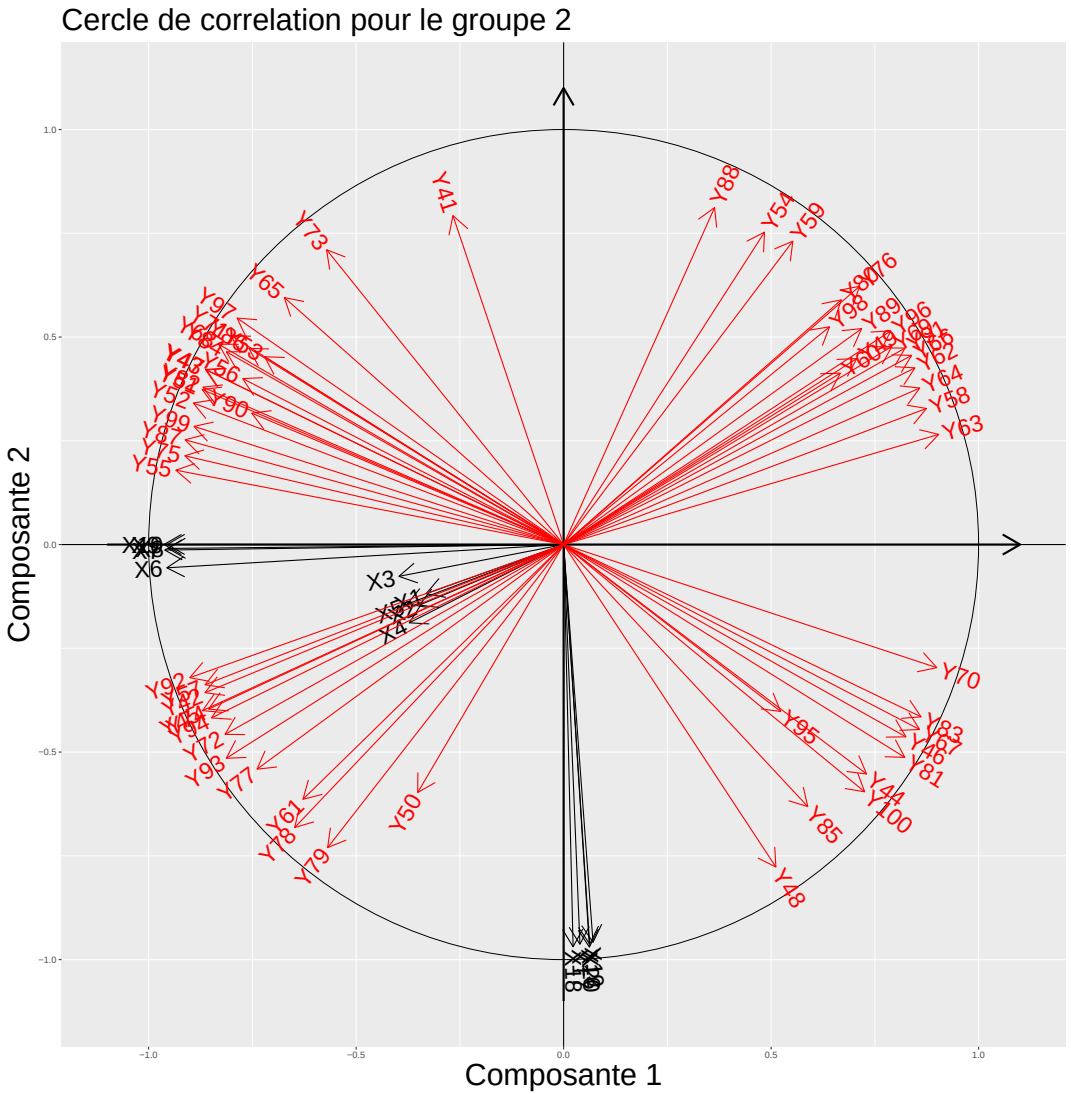
groupe 1	
$\rho^2(\varphi_1, f_1^1)$	0.982
$\rho^2(\varphi_3, f_1^2)$	0.966

groupe 2	
$\rho^2(\varphi_2, f_2^1)$	0.972
$\rho^2(\varphi_4, f_2^2)$	0.984

Résultats


Cercle de correlation pour le groupe 1
Composante 1
Composante 2
1.0
0.5
0.0
−0.5
−1.0
Y11
Y1
Y17
Y3
Y30
Y25
X2
X6
X7
X9
X10
Y28
Y13
Y36
Y15
Y23
X5
Y19
Y40
Y35

Résultats


Cercle de correlation pour le groupe 2
Composante 1
Composante 2
1.0
0.5
0.0
−0.5
−1.0
X6
X3
Y41
Y73
Y65
Y88
Y54
Y59
Y90
Y63
Y70
Y95
Y50
Y85
Y48
Y79
Y78
Y61
Y77
Y100
Y44
Y81

Autres simulations

Simulations

On reprend les simulations précédentes avec une corrélation de 0.9 entre φ_1 et φ_2.

Autres simulations

Simulations

On reprend les simulations précédentes avec une corrélation de 0.9 entre φ_1 et φ_2.

Probabilité *a posteriori* d'inclusion

- $\mathbb{P}(y_i \text{ est dans le groupe } 1) \approx 0.5$
- $\mathbb{P}(y_i \text{ est dans le groupe } 2) \approx 0.5$

Autres simulations

Simulations

On reprend les simulations précédentes avec une corrélation de 0.9 entre φ_1 et φ_2.

Probabilité *a posteriori* d'inclusion

- $\mathbb{P}(y_i \text{ est dans le groupe } 1) \approx 0.5$
- $\mathbb{P}(y_i \text{ est dans le groupe } 2) \approx 0.5$

Corrélation

groupe 1	
$\rho^2(\varphi_1, f_1^1)$	0.776
$\rho^2(\varphi_3, f_1^2)$	0.378

groupe 2	
$\rho^2(\varphi_2, f_2^1)$	0.818
$\rho^2(\varphi_4, f_2^2)$	0.348

Sommaire

Perspectives

- Étendre la méthode aux GLM
- Ajout d'un critère encourageant l'éloignement des sous-espaces explicatifs des groupes

Remerciements et Références

Merci de votre attention

Bry, X., Trottier, C., Mortier, F., and Cornu, G. (2018). Component-based regularization of a multivariate GLM with a thematic partitioning of the explanatory variables. *Statistical Modelling*, pages 1471082X18810114.

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.