Régression linéaire sur composantes supervisées pour les modèles à facteurs latents

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

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

JDS 2021

Plan

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 possédant des dépendances mutuelles

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 possédant des dépendances mutuelles

Traduction statistique

- Trouver des dimensions fortes permettant d'expliquer au mieux les réponses
  - ↪ Recherche de composantes supervisées
- Identifier des blocs de réponses liées en modélisant la matrice de variance covariance
  - ↪ Modèles à facteurs latents

Sommaire

Qu'est-ce qu'une composante ?

Notations

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

Qu'est-ce qu'une composante ?

Notations

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

Définition

Une composante est un vecteur f permettant de synthétiser l'information contenue dans les données X 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 supervisée ?

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

$$\forall k = 1, \dots, q, \quad y_k = F\gamma_k + \varepsilon_k, \quad \varepsilon_k \sim \mathcal{N}(0, \sigma_k^2 I_n)$$

où $F = XU$ et γ_k est un vecteur de coefficients (effets des composantes sur la réponse y_k).

Souhaits

Nous voulons :

Souhaits

Nous voulons :

- des composantes ...

Souhaits

Nous voulons :

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

Souhaits

Nous voulons :

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

SCGLR (Bry et al., 2020)

Pertinence structurelle (SR)

Le critère $\phi(u)$ mesure la "force" de la composante Xu (proximité aux variables x_j) sous la contrainte $\|u\|^2 = 1$

SCGLR (Bry et al., 2020)

Pertinence structurelle (SR)

Le critère $\phi(u)$ mesure la "force" de la composante Xu (proximité aux variables x_j) sous la contrainte $\|u\|^2 = 1$

Qualité d'ajustement (GoF)

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

SCGLR (Bry et al., 2020)

Pertinence structurelle (SR)

Le critère $\phi(u)$ mesure la "force" de la composante Xu (proximité aux variables x_j) sous la contrainte $\|u\|^2 = 1$

Qualité d'ajustement (GoF)

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

$$\Downarrow$$

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

On alterne :

Méthodes d'estimation

On alterne :

Estimation de u avec θ fixé

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

On alterne :

Estimation de u avec θ fixé

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

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

avec D la matrice de contrainte d'orthogonalité des composantes

Estimation de θ avec u fixé

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

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

Sommaire

Modèles à facteurs latents

Encore plus de notations

- L le nombre de facteurs latents
- $g_i \sim \mathcal{N}(0, I_L)$, $i = 1, \ldots n$, le vecteur aléatoire des facteurs latents
- $B \in \mathbb{R}^{L \times q}$ la matrice déterministe des pondérations des facteurs
- $\varepsilon_i \sim \mathcal{N}(0, \Psi)$ le $i$ème vecteur des erreurs, indépendant des facteurs avec $\Psi = \text{diag}\left(\sigma_k^2\right)_k$

Modèles à facteurs latents

Encore plus de notations

- L le nombre de facteurs latents
- $g_i \sim \mathcal{N}(0, I_L)$, $i = 1, \ldots n$, le vecteur aléatoire des facteurs latents
- $B \in \mathbb{R}^{L \times q}$ la matrice déterministe des pondérations des facteurs
- $\varepsilon_i \sim \mathcal{N}(0, \Psi)$ le $i$ème vecteur des erreurs, indépendant des facteurs avec $\Psi = \operatorname{diag}\left(\sigma_k^2\right)_k$

Modèle

Le modèle pour chaque ligne est donné par $y_i = B^T g_i + \varepsilon_i$
de vraisemblance $f(Y; B, \Psi) = \prod_{i=1}^n f(y_i; B, \Psi)$

Modèles à facteurs latents

Encore plus de notations

- L le nombre de facteurs latents
- $g_i \sim \mathcal{N}(0, I_L)$, $i = 1, \dots n$, le vecteur aléatoire des facteurs latents
- $B \in \mathbb{R}^{L \times q}$ la matrice déterministe des pondérations des facteurs
- $\varepsilon_i \sim \mathcal{N}(0, \Psi)$ le $i$ème vecteur des erreurs, indépendant des facteurs avec $\Psi = \operatorname{diag}\left(\sigma_k^2\right)_k$

Modèle

Le modèle pour chaque ligne est donné par $y_i = B^T g_i + \varepsilon_i$
de vraisemblance $f(Y; B, \Psi) = \prod_{i=1}^n f(y_i; B, \Psi)$

Problème 1 : Modèle non identifiable

Modèles à facteurs latents

Encore plus de notations

- L le nombre de facteurs latents
- $g_i \sim \mathcal{N}(0, I_L)$, $i = 1, \ldots n$, le vecteur aléatoire des facteurs latents
- $B \in \mathbb{R}^{L \times q}$ la matrice déterministe des pondérations des facteurs
- $\varepsilon_i \sim \mathcal{N}(0, \Psi)$ le $i$ème vecteur des erreurs, indépendant des facteurs avec $\Psi = \text{diag}\left(\sigma_k^2\right)_k$

Modèle

Le modèle pour chaque ligne est donné par $y_i = B^T g_i + \varepsilon_i$
de vraisemblance $f(Y; B, \Psi) = \prod_{i=1}^n f(y_i; B, \Psi)$

Problème 1 : Modèle non identifiable

Problème 2 : Vraisemblance compliquée à maximiser

Identification (Saidane, 2006)

Problème d'identification

Soit Ω une matrice orthogonale ($\Omega^T\Omega = I$). Le modèle peut se réécrire

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

avec $\mathbb{E}\left[\Omega g_i\right] = \Omega \mathbb{E}\left[g_i\right] = 0$
et $\mathbb{V}\left[\Omega g_i\right] = \Omega \mathbb{V}\left[g_i\right] \Omega^T = I_L$

$\rightarrow$ On obtient la même distribution

Identification (Saidane, 2006)

Problème d'identification

Soit Ω une matrice orthogonale ($\Omega^T\Omega = I$). Le modèle peut se réécrire

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

avec $\mathbb{E}\left[\Omega g_i\right] = \Omega \mathbb{E}\left[g_i\right] = 0$
et $\mathbb{V}\left[\Omega g_i\right] = \Omega \mathbb{V}\left[g_i\right] \Omega^T = I_L$

$\rightarrow$ On obtient la même distribution

On impose une forme particulière à la matrice B :

$$B = \begin{pmatrix} b_1^1 & \dots & b_1^L & b_1^{L+1} & \dots & b_1^q \\ & \ddots & \vdots & \vdots & \ddots & \vdots \\ & & b_L^L & b_L^{L+1} & \dots & b_L^q \end{pmatrix}.$$

Méthode d'estimation

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

- Maximiser une vraisemblance en présence de variables latentes
  $\rightarrow$ Ici : les facteurs latents
- Estimer les paramètres du modèle
  $\rightarrow$ Ici : la matrice B ainsi que la matrice Ψ

SCGLR avec modèles à facteurs latents

Modèle à composantes supervisées et facteurs latents

$$Y = \underbrace{F\Gamma}_{\text{déterministe}} + \underbrace{GB + \varepsilon}_{\text{stochastique}}, \quad \varepsilon \sim \mathcal{N}(0, \otimes_n \Psi)$$

avec $F = XU$ la matrice des composantes, $\Gamma = [\gamma_1, \ldots, \gamma_q]$ et G la matrice des facteurs latents.

La vraisemblance est $f(Y; F, \Gamma, B, \Psi) = \prod_{i=1}^{n} f(y_i; F, \Gamma, B, \Psi)$

SCGLR avec modèles à facteurs latents

Modèle à composantes supervisées et facteurs latents

$$Y = \underbrace{F\Gamma}_{\text{déterministe}} + \underbrace{GB + \varepsilon}_{\text{stochastique}}, \quad \varepsilon \sim \mathcal{N}(0, \otimes_n \Psi)$$

avec $F = XU$ la matrice des composantes, $\Gamma = [\gamma_1, \ldots, \gamma_q]$ et G la matrice des facteurs latents.
La vraisemblance est $f(Y; F, \Gamma, B, \Psi) = \prod_{i=1}^{n} f(y_i; F, \Gamma, B, \Psi)$

Critère à maximiser

$$\underset{u, \Gamma, B, \Psi, \text{ s.c. } \|u\|^2=1,\, D^T u=0}{\operatorname{argmax}} \quad s \ln(\phi(u)) + (1-s) \ln(f(Y; F, \Gamma, B, \Psi))$$

SCGLR avec modèles à facteurs latents

Modèle à composantes supervisées et facteurs latents

$$Y = \underbrace{F\Gamma}_{\text{déterministe}} + \underbrace{GB + \varepsilon}_{\text{stochastique}}, \quad \varepsilon \sim \mathcal{N}(0, \otimes_n \Psi)$$

avec $F = XU$ la matrice des composantes, $\Gamma = [\gamma_1, \ldots, \gamma_q]$ et G la matrice des facteurs latents.

La vraisemblance est $f(Y; F, \Gamma, B, \Psi) = \prod_{i=1}^{n} f(y_i; F, \Gamma, B, \Psi)$

Critère à maximiser

$$\underset{u,\Gamma,B,\Psi,\ \text{s.c.}\ \|u\|^2=1,\ D^T u=0}{\operatorname{argmax}} \quad s \ln(\phi(u)) + (1-s) \ln(f(Y; F, \Gamma, B, \Psi))$$

Méthode d'estimation

On alterne itérativement ces deux maximisations :

- On trouve $\{\Gamma, B, \Psi\}$ à l'aide de l'algorithme EM
- On trouve u à l'aide de l'algorithme PING

Sommaire

Simulation déterministe 1

Variables réponses

$Y = [y_1, \ldots, y_{50}]$ est composée de 50 réponses gaussiennes

Simulation déterministe 1

Variables réponses

$Y = [y_1, \ldots, y_{50}]$ est composée de 50 réponses gaussiennes

Variables explicatives

$$X = [\underbrace{x_1, \ldots, x_{10}}_{\text{faisceau X1}} \mid \underbrace{x_{11}, \ldots, x_{15}}_{\text{faisceau X2}} \mid \underbrace{x_{16}, \ldots, x_{30}}_{\text{bruit}}]$$

- X est composée de deux faisceaux et d'un ensemble de variables de bruit
- Les faisceaux prédisent les réponses

Simulation stochastique 1

Facteurs

On simule 2 facteurs latents expliquant les réponses

Simulation stochastique 1

Facteurs

On simule 2 facteurs latents expliquant les réponses

Matrice B

$$B = \begin{pmatrix} \mathcal{U}_{[1,1.5]}^{\otimes 20} & \mathbf{0} & \mathcal{U}_{[-1.5,-1]}^{\otimes 10} \\ \mathbf{0} & \mathcal{U}_{[1,1.5]}^{\otimes 20} & \mathcal{U}_{[-1.5,-1]}^{\otimes 10} \end{pmatrix}$$

Simulation stochastique 1

Facteurs

On simule 2 facteurs latents expliquant les réponses

Matrice B

$$B = \begin{pmatrix} \mathcal{U}_{[1,1.5]}^{\otimes 20} & \mathbf{0} & \mathcal{U}_{[-1.5,-1]}^{\otimes 10} \\ \mathbf{0} & \mathcal{U}_{[1,1.5]}^{\otimes 20} & \mathcal{U}_{[-1.5,-1]}^{\otimes 10} \end{pmatrix}$$

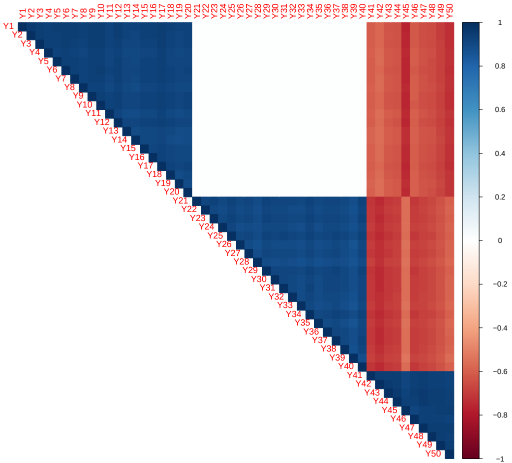
Variance résiduelle

$\Psi = \mathrm{diag}\left(\sigma_k^2\right)_{k=1,\ldots,50}$ où

$\forall k = 1, \ldots, 50, \quad \sigma_k^2 \sim \mathcal{U}_{[0.1,0.2]}$

Simulation 1 : matrice de variance covariance

La matrice de variance covariance $B^T B + \Psi$ devient



Résultats

Corrélations

Composantes	
$\mathrm{cor}^2(\varphi_1, f_1)$	0.984
$\mathrm{cor}^2(\varphi_2, f_2)$	0.979

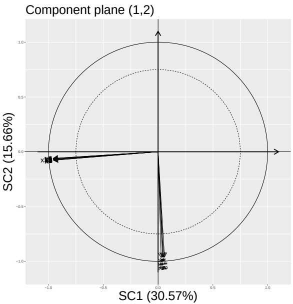
Facteurs	
$\mathrm{cor}^2(g_1, \tilde{g}_1)$	0.983
$\mathrm{cor}^2(g_2, \tilde{g}_2)$	0.969

Résultats

Corrélations

Composantes	
$\mathrm{cor}^2(\varphi_1, f_1)$	0.984
$\mathrm{cor}^2(\varphi_2, f_2)$	0.979

Facteurs	
$\mathrm{cor}^2(g_1, \tilde{g}_1)$	0.983
$\mathrm{cor}^2(g_2, \tilde{g}_2)$	0.969

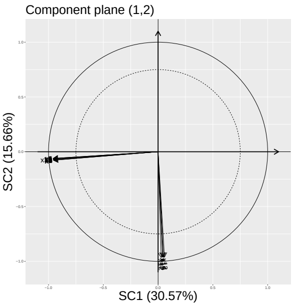


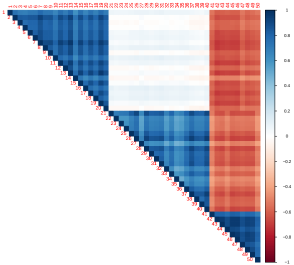
Résultats

Corrélations

Composantes	
$\mathrm{cor}^2(\varphi_1, f_1)$	0.984
$\mathrm{cor}^2(\varphi_2, f_2)$	0.979

Facteurs	
$\mathrm{cor}^2(g_1, \tilde{g}_1)$	0.983
$\mathrm{cor}^2(g_2, \tilde{g}_2)$	0.969





Simulation 2

Simulation

On refait la même simulation avec deux facteurs mais avec B de la forme :

$$B = [\underbrace{b_1, \dots, b_{15}}_{\sim\mathcal{N}(\mu_1, 0.01 I_2)} \mid \underbrace{b_{16}, \dots, b_{30}}_{\sim\mathcal{N}(\mu_2, 0.01 I_2)} \mid \underbrace{b_{31}, \dots, b_{40}}_{\sim\mathcal{N}(\mu_3, 0.01 I_2)} \mid \underbrace{b_{41}, \dots, b_{50}}_{\sim\mathcal{N}(\mu_4, 0.01 I_2)}],$$

où $\mu_1 = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} -0.5 \\ 0.5 \end{pmatrix}$, $\mu_3 = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix}$ et $\mu_4 = \begin{pmatrix} -0.5 \\ -0.5 \end{pmatrix}$.

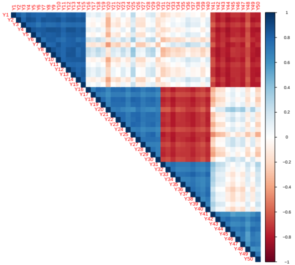
Simulation 2

Simulation

On refait la même simulation avec deux facteurs mais avec B de la forme :

$$B = [\underbrace{b_1, \ldots, b_{15}}_{\sim\mathcal{N}(\mu_1, 0.01 I_2)} \mid \underbrace{b_{16}, \ldots, b_{30}}_{\sim\mathcal{N}(\mu_2, 0.01 I_2)} \mid \underbrace{b_{31}, \ldots, b_{40}}_{\sim\mathcal{N}(\mu_3, 0.01 I_2)} \mid \underbrace{b_{41}, \ldots, b_{50}}_{\sim\mathcal{N}(\mu_4, 0.01 I_2)}],$$

où $\mu_1 = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} -0.5 \\ 0.5 \end{pmatrix}$, $\mu_3 = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix}$ et $\mu_4 = \begin{pmatrix} -0.5 \\ -0.5 \end{pmatrix}$.



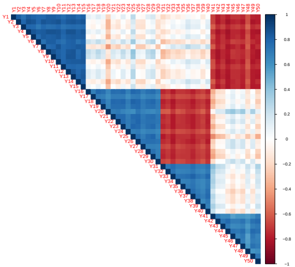
Simulation 2

Simulation

On refait la même simulation avec deux facteurs mais avec B de la forme :

$$B = [\underbrace{b_1, \dots, b_{15}}_{\sim \mathcal{N}(\mu_1, 0.01 I_2)} \mid \underbrace{b_{16}, \dots, b_{30}}_{\sim \mathcal{N}(\mu_2, 0.01 I_2)} \mid \underbrace{b_{31}, \dots, b_{40}}_{\sim \mathcal{N}(\mu_3, 0.01 I_2)} \mid \underbrace{b_{41}, \dots, b_{50}}_{\sim \mathcal{N}(\mu_4, 0.01 I_2)}],$$

où $\mu_1 = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} -0.5 \\ 0.5 \end{pmatrix}$, $\mu_3 = \begin{pmatrix} 0.5 \\ -0.5 \end{pmatrix}$ et $\mu_4 = \begin{pmatrix} -0.5 \\ -0.5 \end{pmatrix}$.



Question : Comment identifier les blocs ?

K-means

On procède à plusieurs étapes *a posteriori* :

K-means

On procède à plusieurs étapes *a posteriori* :

1. On estime la matrice de variance covariance

K-means

On procède à plusieurs étapes *a posteriori* :

1. On estime la matrice de variance covariance
2. On calcule une matrice de distance à partir de la matrice de variance covariance

K-means

On procède à plusieurs étapes *a posteriori* :

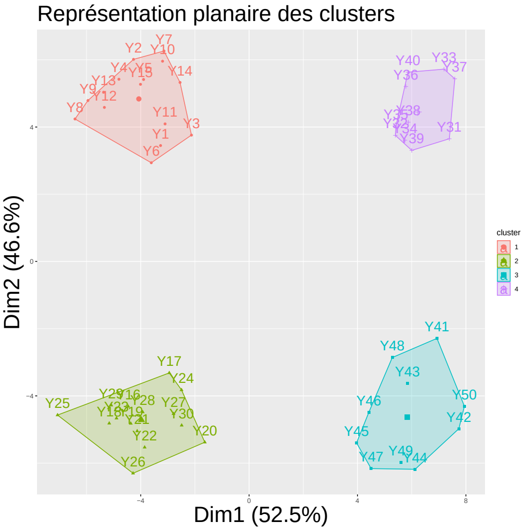
1. On estime la matrice de variance covariance
2. On calcule une matrice de distance à partir de la matrice de variance covariance
3. On effectue un K-means sur la matrice de distance

K-means

On procède à plusieurs étapes *a posteriori* :

1. On estime la matrice de variance covariance
2. On calcule une matrice de distance à partir de la matrice de variance covariance
3. On effectue un K-means sur la matrice de distance
4. On choisit le nombre de clusters maximisant le critère de silhouette

Résultats avec quatre clusters



Conclusion et perspectives

Conclusion

Nous avons :

- Étendu SCGLR aux modèles à facteurs latents
- Développé un algorithme capable de trouver des composantes supervisées et de modéliser la matrice de variance covariance

Conclusion et perspectives

Conclusion

Nous avons :

- Étendu SCGLR aux modèles à facteurs latents
- Développé un algorithme capable de trouver des composantes supervisées et de modéliser la matrice de variance covariance

Perspectives

Nous voulons :

- Étendre cette modélisation à un partitionnement thématique des variables explicatives
- Affiner la détection de blocs dans la matrice de variance covariance

Remerciements et Références

Merci de votre 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.

Saidane, M. (2006). *Modèles à facteurs conditionnellement hétéroscédastiques et à structure markovienne cachée pour les séries financières*. PhD thesis, Université Montpellier II-Sciences et Techniques du Languedoc.