ProcMod

Christelle Melodelima & Eric Coissac

2018-07-06

Aims of the module

Expliquer un tableau multavariées decrivant des individus ou des sites (ci dessous dénommés individus) par un ensemble de tableaux eux-même multivariés décrivant les mêmes individus. Par exemple, expliquer les changement de communauté d'espèces entre différents sites géographiques à partir de tableaux de données climatiques, chimique, d'espèces d'autres groupes taxinomiques... Chaque tableau est considéré comme une variable explicative sans chercher à donner un rôle à chacune des variables qui le composent.

Model principes

L'idée est de s'appuyer sur les analyses procustéennes en les généralisant à plusieurs (k) tableaux. Pour mémoire, l'analyse procustéenne consiste à superposer deux nuages de points dans un espace de dimensions quelconque en réalisant trois operations:

1. une translation (centrage des données)
2. une rotation
3. une mise à l'échelle.

Dans notre cas, nous considérerons que les deux premières opérations ont pour seul but de projeter l'ensemble des tableaux réponse et explicatifs, dans un espace commun. La troisième operation d'homothétie servira de base à l'analyse de partition de la variance du tableau réponse. Cette approche a de forts liens avec les analyses de co-inerties développées par Chessel et al

Données en entrée

Les tableaux utilisés dans cette analyse doivent se projeter dans un espace orthogonal, ce qui implique aucune corrélation entre les colonnes des tableaux. Ils doivent tous décrire les mêmes individus et doivent donc tous avoir le même nombre de lignes n.

Tableaux de variables quelconques

Comme posé en préambule, l'analyse vise à mesurer l'effet global de k tableaux de variables sur un tableau réponse sans s'interresser à l'effet individuel de chacune des variables des différents tableaux explicatifs. Les tableaux utilisés peuvent donc sans perte d'information être projetés dans un espace orthogonal par une simple PCA.

Tableaux de distances

Il est possible de caractériser la dissimilarité entre les individus par une autre mesure que la corrélation (*cf* PCA ci-dessus) en estimant un tableau de distances par une méthode appropriée au type de données étudiées. Par exemple une distance de Bray-Curtis ou de Jaccard, lorsqu'il s'agit de comparer les communautés d'espèces présentes dans plusieurs sites.

Si la distance utilisée est une métrique, le tableau de distance pourra être projeter dans un espace orthogonal à $n - 1$ dimensions par une PCoA.

Si la distance utilisée n'est pas une métrique, il faudra recourrir à une méthode non paramétrique telque la NMDS ou alterer le tableau de distance pour le rendre diagonalisable.

Méthode de calculs du modèle procustéen

L'analyse procustéenne peut être assimilée à un modèle linéaire entre deux tableaux:

- Le tableau réponse : Y
- Le tableau explicatif : X

Si l'ensemble des variables de chaque tableau appartiennent à $\mathbb{R}$:

- L'opération de translation est réalisée via le centrage des variables des tableaux X et Y. Dorénavant, la notation X et Y représentera les tableau après centrage.
- La rotation pour projeter le tableau X sur Y est calculer à partir de la décomposition en valeurs singulières de la matrice $Y'X$ soit $Y'X = U\Lambda V'$. On définit la rotation de X sur Y de la manière suivante:

$$Rot(X|Y) = XVU'$$

- Le facteur d'homothétie a se calcule selon :

$$a = \frac{\sum diag(\Lambda)}{\sum diag(X'X)}$$

$\sum diag(\Lambda)$ est la co-inertie entre les matrices X et Y et $\sum diag(X'X)$ est l'inertie du tableau X. Co-inertie et inertie pouvant être assimilées à la covariance et à la variance de vecteurs. Le facteur a est donc l'équivalent en dimention 1 de la pente de la droite de régression linéaire de Y par X:

$$a = \frac{Cov_{XY}}{Var_X}$$

Nous proposons ici de généraliser l'analyse procustéen à k tableaux en résolvant la régression multiple du tableau de réponse Y par k matrices explicatives X_1, X_2, ... X_k.

Le calcul des coefficients d'échelles a_i est réalisé par la même approche que pour ceux d'une régression linéaire multiple, mais ici à partir de la matrice d'inertie et de co-inertie des tableaux Y et X_i:

Calcul de l'inertie d'une matrice et de co-inertie entre deux matrices

Soit CoI_{XY} la cointertie entre les tableaux X et Y et I_X l'inertie du tableau X. CoI_{XY} se calcule à partir de la décomposition en valeur singulière de la matrice $Y'X = U\Lambda V'$.

$$CoI_{XY} = \frac{\sum diag(\Lambda)}{n-1}$$

Comme pour la Covariance, $CoI_{XY} = CoIYX$ et $CoI_{XX} = I_X$

Calcul du coefficient de corrélation entre deux matrices

Par analogie on définit R_{XY} comme le coefficient de corrélation entre deux matrices X et Y de la manière suivante :

$$R_{XY} = \frac{CoI_{XY}}{\sqrt{I_X I_Y}}$$

Calcul des facteurs d'échelle en procuste multiple à k tableaux

On note à partir de maintenant :

- $X = \{X_1, X_2, ... X_k\}$ l'ensemble des k tableaux explicatifs centrés
- Y le tableau réponse centré
- $M \in \{Y\} \cup X$ tel que M est la matrice de plus grande dimension.
- CoI_{YX} la matrice colonne des co-inerties entre Y et chacun des éléments de X
- CoI_{XX} la matrice d'inertie, co-inerties entre tous les éléments de X
- $a = \{a_1\, a_2, ... a_k\}$ l'ensemble des coefficients d'échelle associés à chacun des éléments de X dans le modèle procustéen multiple.

a se calcule par :

$$a = (CoI_{XX})^{-1} CoI_{YX}$$

Les prédictions du modèle peuvent donc s'écrirent:

$$\widehat{Rot(Y|M)} = a_1 Rot(X_1|M) + a_2 Rot(X_2|M) + \ ... \ + a_k Rot(X_k|M)$$

Interaction entre deux tableaux explicatifs

L'interaction entre deux tableaux explicatifs X_i et X_j est estimée de manière analogue à l'interaction dans un modèle linéaire multiple. Le principe est de poser que a_i le facteur d'échelle associé à X_i est une fonction affine de X_j et symétriquement a_j est une fonction affine de X_i. Pour réaliser ce calcul il est nécessaire de projeter les deux matrices explicatives sur M.

$$a_i = (b_i Rot(X_j|M) + c_i) a_j = (b_j Rot(X_i|M) + c_j)$$

Dans le cas de deux tableaux, le modèle procustéen expliquant Y par X_1, X_2 et l'interaction de X_1 et X_2 peut donc s'écrire :

$$\begin{aligned}
\widehat{Rot(Y|M)} =& a_1 \cdot Rot(X_1|M) + a_2 \cdot Rot(X_2|M) \\
=& (b_1 Rot(X_2|M) + c_1) \cdot Rot(X_1|M) + \\
& (b_2 Rot(X_1|M) + c_2) \cdot Rot(X_2|M) \\
=& c_1 Rot(X_1|M) + c_2 Rot(X_2|M) + (b_1 + b_2) Rot(X_1|M) \cdot Rot(X_2|M)
\end{aligned}$$

En renommant les facteurs d'échelle les prédictions s'écrivent:

$$\widehat{Rot(Y|M)} = a_1 Rot(X_1|M) + a_2 Rot(X_2|M) + a_{1,2} Rot(X_1|M) \cdot Rot(X_2|M)$$

Ce qui revient à construire un nouveau modèle sans interaction incluant un tableau explicatif supplémentaire

$$X_{1,2} = Rot(X_1|M) \cdot Rot(X_2|M)$$

Ce principe peut être généralisé à l'interaction entre plus de deux tableaux.

Partition de l'inertie de Y selon le modèle procustéen

La variation total SCT est définie comme suit:

$$SCT = I_Y * (n-1)$$

La variation résiduelle non expliquée par le modèle est :

$$SCR = \sum (Rot(Y|M) - \widehat{Rot(Y|M)})^2$$

Selon l'approche de Scherrer on propose de définir SCE_i la contribution de X_i à la variation de Y peut s'écrire :

$$SCE_i = a_i \; R_{X_iY} \; SCT$$

On propose de tester la signignificativité de l'effet de chacun des tableaux explicatifs par une méthode de permutation des lignes de chacune des matrices explicatives.

How to build a procrustean model

```
library(ProcMod)
```

Loading of the demo data

They consist in two MOTUs tables describing bacterial and eukaryiote communities at 21 sites accros the eastern coast of Australia.

```
data("eukaryotes")
data("bacteria")
```

At each site environmental data are also available to describe soil chemistry, climat and site location.

```
data("soil")
data("climat")
data("geography")
```

Processing of the MOTUs data

Using the *vegan* package MOTUs frequencies are transformed according to Hellinger by take the square root of the MOTUs relative frequencies at each site

```
library(vegan)
#> Le chargement a nécessité le package : permute
#> Le chargement a nécessité le package : lattice
#> This is vegan 2.5-2

bac.hellinger = decostand(bacteria,method ="hellinger")
euk.hellinger = decostand(eukaryotes,method ="hellinger")
```

Bray Curtis distances among sites are computed according to both the communities

```
bac.bray = vegdist(bac.hellinger,method = "bray")
euk.bray = vegdist(euk.hellinger,method = "bray")
```

Processing the environmental data

Soil and climatic data are centered and rescaled for having the same influence

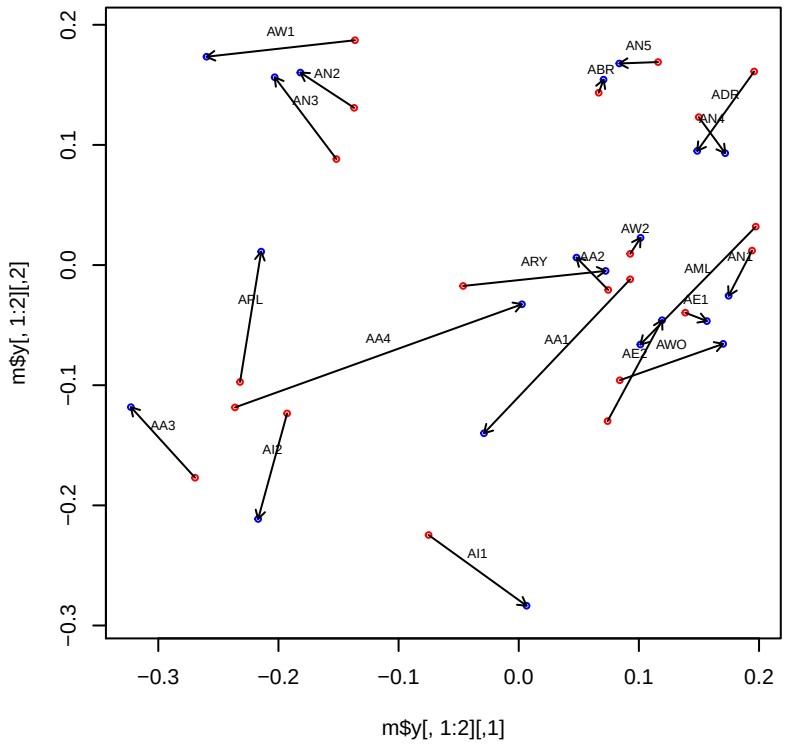
```
soil.rescaled   = scale(soil,   center = TRUE, scale = TRUE)
climat.rescaled = scale(climat, center = TRUE, scale = TRUE)
```

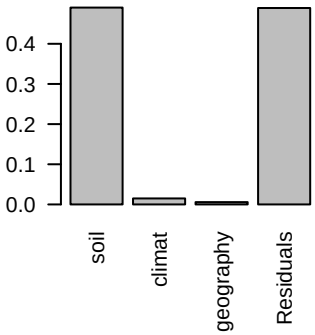
Assembling the data for the model

```
data = procmod.frame(euk       = euk.bray,
                     bac       = bac.bray,
                     climat    = climat,
                     soil      = soil,
                     geography = geography)
```

```
euk.pm = pm(formula = euk ~ soil + climat + geography, data = data)
euk.pm
#>
#> Call:
#> pm(formula = euk ~ soil + climat + geography, data = data)
#>
#> Coefficients:
#>         soil       climat    geography
#> 0.1411894993 0.0002849650 0.0006001613
```

```
plot(euk.pm)
```






```
anova(euk.pm)
#> Analysis of Variance Table
#>
#> Response: euk
#>            Df  Sum Sq Mean Sq F value    Pr(>F)
#> soil        1 1.09067 1.09067 18.0414 0.0004843 ***
#> climat      1 0.03318 0.03318  0.5489 0.4683427
#> geography   1 0.01281 0.01281  0.2118 0.6508585
#> Residuals 18 1.08817 0.06045
#> ---
#> Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```