Fiche TD avec le logiciel  : ter3

Le couplage de tableaux écologiques

S. Dray, D. Chessel

Cette fiche illustre l'utilisation des principales méthodes de couplage en écologie. Elle s'appuie essentiellement sur l'utilisation du paquet `ade4`.

Table des matières

1 Introduction

On considère un tableau $\mathbf{Y}$ (n x s) avec l'abondance de s espèces dans n sites. Un deuxième tableau $\mathbf{X}$ contient les mesures de p variables environnementales à n sites. Chaque tableau peut être traité par une analyse induisant deux triplets d'analyse simple $(\mathbf{X}, \mathbf{Q}_1, \mathbf{D})$ et $(\mathbf{Y}, \mathbf{Q}_2, \mathbf{D})$.

Il y a trois stratégies principales d'association de deux tableaux. Elles recouvrent, à cause des nombreux jeux de paramètres, un vaste ensemble de pratiques. Il suffit de saisir ces trois principes pour faire un choix, voire pour construire des associations originales dont on peut avoir besoin.

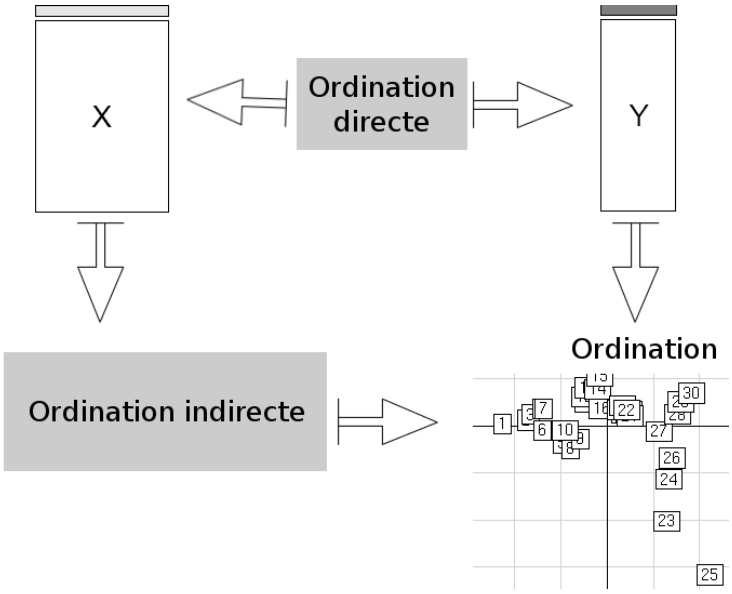
Le couplage de deux tableaux de données est une opération fondamentale en écologie statistique. On dispose d'une énorme littérature sur le sujet. Parmi les bases historiques on doit rappeler quelques opérations fondamentales.

1.1 Ordination directe et indirecte

En face du tableau $\mathbf{Y}$, qui peut être soumis à une méthode d'ordination, on retrouve de l'information mésologique contenue dans le tableau $\mathbf{X}$. Deux stratégies sont envisageables [Jongman et al., 1987] :

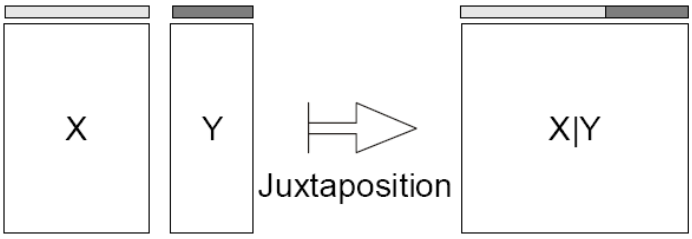
1. une ordination directe des relevés et des espèces le long du (ou des) gradient(s) de variables environnementales.
2. une ordination des relevés selon leur composition spécifique, suivie d'une interprétation du résultat obtenu à la lumière de connaissances extérieures sur les espèces ou les sites.

Whittaker [1967] qualifie ces deux stratégies, respectivement, *direct gradient analysis* et *indirect gradient analysis*.



1.2 Juxtaposition

Deux tableaux ayant les mêmes lignes sont simplement accolés pour former un nouveau tableau qui appelle une analyse simple.

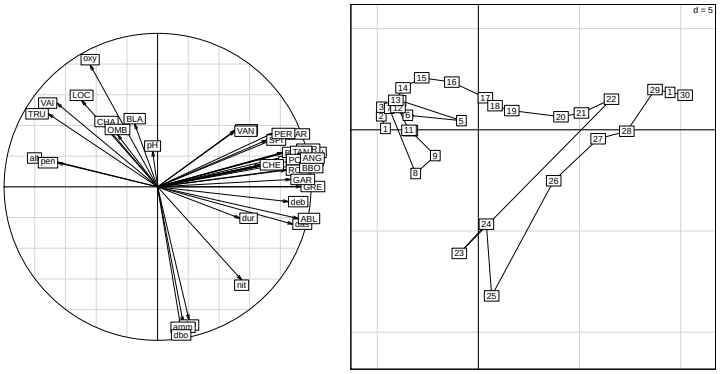


On associe un tableau 30 lignes-stations et 11 colonnes-variables (tableau mésologique) et un tableau 30 lignes-stations et 27 colonnes-espèces (tableau faunistique) :

```
library(ade4)
data(doubs)
tabdoubs <- cbind(doubs$mil, doubs$poi)
dim(tabdoubs)
[1] 30 38
```

On accole les deux tableaux et on soumet le résultat à une ACP normée :

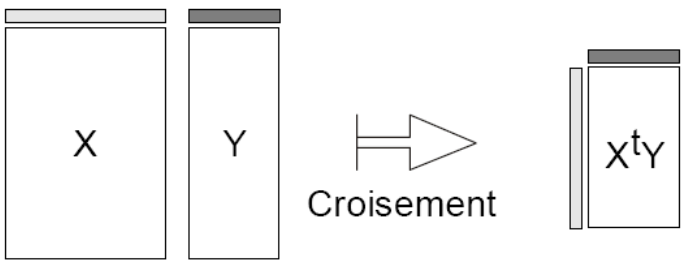
```
pcadoubs <- dudi.pca(tabdoubs, scannf = F, nf = 2)
s.corcircle(pcadoubs$co)

s.traject(pcadoubs$li)
s.label(pcadoubs$li, add.plot = T)
```



Cette approche ne fonctionne que si les inerties des deux tableaux sont comparables. Si l'un des deux l'emporte largement, il imposera son point de vue.

1.3 Croisement

Le tableau croisé utilise simplement un produit de matrices.



C'est la base des analyses de co-inertie que nous verrons plus loin. Il suffit que le produit de matrice ainsi défini ait un sens expérimental. Passer le tableau faunistique en pourcentage par espèce et transposer :

```
pcamil <- dudi.pca(doubs$mil, scannf = F, nf = 2)
profcol <- t(apply(doubs$poi, 2, function(x) x/sum(x)))
tabcroi <- as.data.frame(profcol %*% as.matrix(pcamil$tab))
```

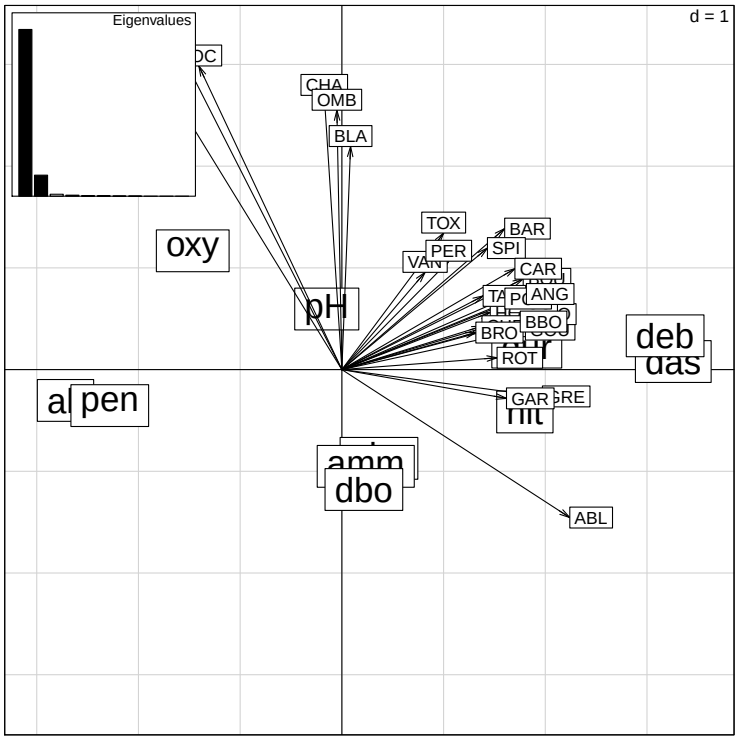
On obtient la position moyenne de chaque espèce sur chaque variable de milieu normalisée.

```
head(round(tabcroi, dig = 1))
```

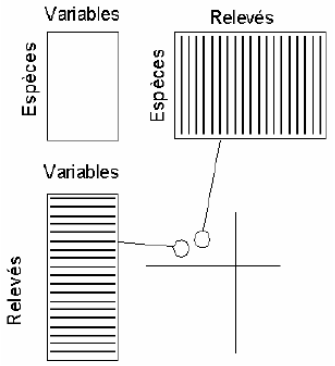
	das	alt	pen	deb	pH	dur	pho	nit	amm	oxy	dbo
CHA	-0.2	-0.2	-0.2	-0.1	0.8	0.3	-0.4	-0.3	-0.5	1.0	-0.6
TRU	-0.7	0.6	0.5	-0.5	0.3	-0.5	-0.5	-0.7	-0.5	0.8	-0.6
VAI	-0.5	0.4	0.3	-0.4	0.1	-0.3	-0.4	-0.5	-0.4	0.6	-0.5
LOC	-0.4	0.3	0.1	-0.4	0.1	-0.3	-0.4	-0.4	-0.4	0.5	-0.4
OMB	-0.1	-0.3	-0.2	0.1	0.6	0.5	-0.4	-0.4	-0.5	1.1	-0.5
BLA	0.0	-0.4	-0.3	0.1	0.8	0.3	-0.3	-0.1	-0.4	0.8	-0.5

On peut faire l'ACP non centrée de ce tableau. On peut également travailler sur les corrélations entre les deux tableaux. On obtient un biplot qui donne deux typologies de variables et leur lien :

```
scatter(dudi.pca(cor(doubs$mil, doubs$poi), scale = F, center = F,
    scan = F), clab.r = 2)
```



On a alors trois tableaux et on comprend que les relevés sont supplémentaires de deux manières dans l'analyse du tableau croisé :



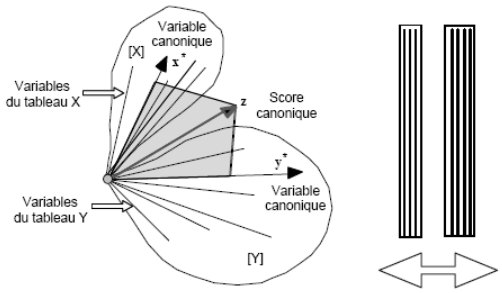
Le cas le plus célèbre d'un croisement de tableaux signifiant est celui des profils écologiques [Godron et al., 1968, Gounot, 1969]. Les variables de milieu sont toutes qualitatives et les variables floristiques sont toutes binaires (0 absence, 1 présence). Le tableau $\mathbf{X}$ est formé des indicatrices des classes de chaque variable de milieu. Le tableau $\mathbf{Y}$ est formé des indicatrices de présence de chaque espèce. Le tableau croisé $\mathbf{Y}^T\mathbf{X}$ a pour lignes les espèces et pour colonnes les modalités de milieu. Les cases contiennent le nombre de stations de chaque modalité de milieu contenant l'espèce. Sur une ligne, on trouve une juxtaposition de profils écologiques bruts. Romane [1972] a eu l'idée d'envoyer un tel tableau dans l'analyse des correspondances, idée qui sera retrouvée plus tard ...[Montaña and Greig-Smith, 1990]. On obtient un cas particulier de l'analyse de co-inertie [Mercier et al., 1992].

2 Analyse canonique des corrélations

L'analyse canonique des corrélations [Hotelling, 1936] est la plus ancienne et la plus connue des méthodes de couplage introduite en détail pour l'écologie par Gittins [1985]. Le fondement considère deux ACP normées. On appellera simplement $\mathbf{X}$ et $\mathbf{Y}$ les deux tableaux normalisés (moyennes nulles et variances unitaires par colonne).

On suppose que les deux paquets de variables (colonnes de $\mathbf{X}$ et $\mathbf{Y}$) sont sans redondances (la régression d'une variables de $\mathbf{Y}$ sur $\mathbf{X}$ ou d'une variable de $\mathbf{X}$ sur $\mathbf{Y}$ est définie sans problème) donc que les matrices de corrélations des deux paquets sont inversibles. L'analyse canonique est celle de ($\mathbf{Y}^T\mathbf{DX}$, $(\mathbf{X}^T\mathbf{DX})^{-1}$, $(\mathbf{Y}^T\mathbf{DY})^{-1}$). L'analyse recherche une combinaison de variables du tableau $\mathbf{X}$ de variance 1 et une combinaison de variables du tableau $\mathbf{Y}$ de variance 1 de corrélation (au carré) maximale. Elle est basée sur la diagonalisation de $(\mathbf{Y}^T\mathbf{DY})^{-1}\mathbf{Y}^T\mathbf{DX}(\mathbf{X}^T\mathbf{DX})^{-1}\mathbf{X}^T\mathbf{DY}$ et est disponible par la fonction `cancor` du package `stats`.

L'analyse canonique prend sa signification dans l'espace des variables. On peut la résumer par la figure :



La variable canonique du tableau **X** est la combinaison des variables de **X** la mieux prédite par une régression multiple sur les variables de **Y** tout comme la variable canonique du tableau **Y** est la combinaison des variables de **Y** la mieux prédite par une régression multiple sur les variables de **X**. On appelle score canonique la bissectrice des deux variables canoniques (somme normalisée). Il est capital de voir dans l'analyse canonique toutes les contraintes associées à deux régressions multiples. **Le nombre de variables (de X et de Y) doit être forcément limité par rapport au nombre d'individus.**

C'est pourquoi, sur les couplages de tableaux écologiques, elle a mauvaise presse. On ne peut l'utiliser que si le nombre de variables est faible par rapport au nombre d'individus.

```
pcafau <- dudi.pca(doubs$poi, scale = F, scannf = F, nf = 2)
cancor(pcamil$tab, pcafau$tab)$cor
```

```
 [1] 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
 [9] 1.0000000 0.9052173 0.7643280
```

Ce n'est pas la peine de continuer. Il existe une multitude de combinaisons linéaires des variables faunistiques parfaitement corrélées à des combinaisons linéaires de variables de milieu, mais cela ne peut rien nous apprendre. Par contre, on l'utilise avec un tableau et les coordonnées de l'analyse de l'autre ou avec deux ensembles de coordonnées [Barkham and Norris, 1970] :

```
can1 <- cancor(pcamil$l1, pcafau$l1)
can1$cor
```

```
[1] 0.7927896 0.6577477
```

```
cor(as.matrix(pcamil$l1) %*% can1$xcoef, as.matrix(pcafau$l1) %*%
    can1$ycoef)
```

```
             [,1]         [,2]
[1,] 7.927896e-01 8.832182e-17
[2,] 1.182255e-16 6.577477e-01
```

Une des premières applications de l'analyse canonique des corrélations en écologie est attribuable à Austin [1968] qui juge la procédure inadaptée :

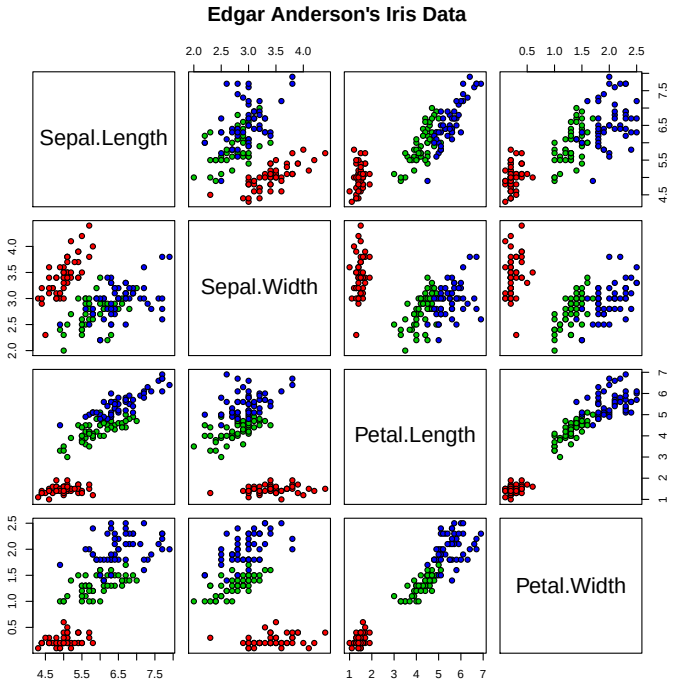
> *the mathematical model on which it is based, with its requirements for orthogonal correlation between vegetation and environment and complete linearity, appears to be too stringent.*

D'un point de vue pratique, elle est réhabilitée par l'analyse de données d'occurrences [Gimaret-Carpentier et al., 2003, Pélissier et al., 2002]. D'un point de vue théorique, l'analyse canonique reste une méthode fondamentale.

2.1 Analyse discriminante

L'analyse discriminante est une méthode de couplage particulière faisant intervenir un tableau numérique et une partition des individus. L'analyse implémentée dans la fonction `discrimin` d'`ade4` recherche des coefficients des variables numériques pour obtenir une combinaison linéaire de variance unité qui maximise le rapport de la variance interclasse sur la variance totale (séparation des groupes). Un exemple célèbre (`?iris`) :

```
pairs(iris[1:4], main = "Edgar Anderson's Iris Data", pch = 21,
    bg = c("red", "green3", "blue")[unclass(iris$Species)])
```



L'analyse discriminante recherche une combinaison des variables permettant de séparer les espèces :

```
data(iris)
iris.pca <- dudi.pca(iris[, 1:4], scannf = F, nf = 2)
iris.dis <- discrimin(iris.pca, iris$Species, scannf = F, nf = 2)
iris.dis$eig
```

```
[1] 0.9698722 0.2220266
```

L'analyse discriminante est une analyse canonique entre le tableau de valeurs numériques et les indicatrices de classes :

```
iris.spe <- acm.disjonctif(data.frame(iris$Species))
iris.pca <- dudi.pca(iris[, 1:4], scannf = F, nf = 2)
iris.can <- cancor(iris.pca$tab, iris.spe, ycenter = F)
iris.can$cor^2
```

```
[1] 9.698722e-01 2.220266e-01 3.827166e-32
```

2.2 Analyse factorielle des correspondances

Le véritable individu statistique de l'AFC est la correspondance (case non-nulle de la table de contingence) qui donne son nom à la méthode. La valeur d'abondance correspond au poids d'une correspondance.

```
data(rpjdl)
fau <- rpjdl$fau
coarpjdl <- dudi.coa(fau, scannf = F, nf = 2)
fau[fau > 0][1:10]
[1] 1 1 1 1 1 1 1 1 1 1
```

Il y a 1639 correspondances. Ici, le poids associé à chaque correspondance vaut 1 car les données sont des présences-absences. Une espèce et un site sont associés à chaque correspondance :

```
espfac <- as.factor(names(fau)[col(as.matrix(fau))[fau > 0]])
sitfac <- as.factor(row.names(fau)[row(as.matrix(fau))[fau > 0]])
espdisj <- acm.disjonctif(data.frame(espfac))
sitdisj <- acm.disjonctif(data.frame(sitfac))
```

L'AFC est une analyse discriminante qui calcule un score pour les correspondances espèces pour maximiser la variance inter-sites. C'est également une analyse discriminante qui calcule un score pour les correspondances sites pour maximiser la variance inter-espèces :

```
acpesp <- dudi.pca(espdisj, scale = F, center = F, scannf = F, nf = 2)
acpsit <- dudi.pca(sitdisj, scale = F, center = F, scannf = F, nf = 2)
discsit <- discrimin(acpesp, sitfac, scannf = F, nf = 2)
discesp <- discrimin(acpsit, espfac, scannf = F, nf = 2)
coarpjdl$eig[1:5]
[1] 0.7532461 0.2929057 0.2293391 0.2046670 0.1572887
discsit$eig[1:6]
[1] 1.0000000 0.7532461 0.2929057 0.2293391 0.2046670 0.1572887
discesp$eig[1:6]
[1] 1.0000000 0.7532461 0.2929057 0.2293391 0.2046670 0.1572887
```

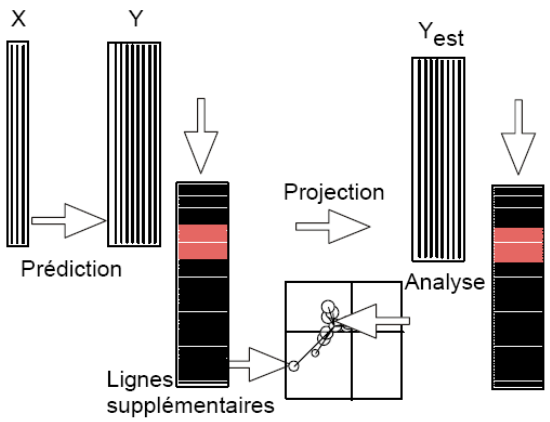
On a donc deux analyses discriminantes : la première calcule un score des sites de variance unité qui maximise la séparation des niches (variance des positions des espèces) ; la seconde calcule un score des espèces de variance unité qui maximise la diversité β (variance des positions des sites).

Une analyse discriminante est une analyse canonique. L'AFC est une double analyse discriminante et une analyse canonique :

```
canrpjdl <- cancor(espdisj, sitdisj, xcenter = F, ycenter = F)
canrpjdl$cor[1:6]^2
[1] 1.0000000 0.7532461 0.2929057 0.2293391 0.2046670 0.1572887
```

3 Analyses sur variables instrumentales

Ce sont des méthodes dissymétriques. Un tableau est formé de variables explicatives (dites instrumentales) et un tableau est formé de variables à étudier. Dans ces méthodes, on étudie le second en utilisant le premier. On parle en général d'analyses sur variables instrumentales. Fondations dans Rao [1964], voir la synthèse dans Lebreton et al. [1991]. Dans `ade4`, la fonction `pcaiv` permet de mettre en oeuvre cette stratégie.



X est le tableau des variables instrumentales. **Y** est le tableau à analyser. Chacune des variables de **Y** est prédite par une régression multiple sur les variables de **X**. Les modèles sont rangés dans le tableau $\mathbf{Y}_{est}$. Les variables de $\mathbf{Y}_{est}$ sont obtenues par projection des variables de **Y** sur le sous-espace engendré par les variables de **X**.

Cette opération est linéaire.

On considère alors que $\mathbf{Y}_{est}$ est un ensemble de lignes. Quand on passe de la ligne i du tableau **Y** à la ligne i du tableau $\mathbf{Y}_{est}$ l'opération *n'est plus linéaire*. On analyse $\mathbf{Y}_{est}$ et on projette en lignes supplémentaires celle de **Y**.

On fait une analyse de **Y** sous contrainte de **X**. Suivant **X** et **Y**, on a un ensemble de méthodes dites d'ordination sous contraintes ou d'analyses sur variables instrumentales.

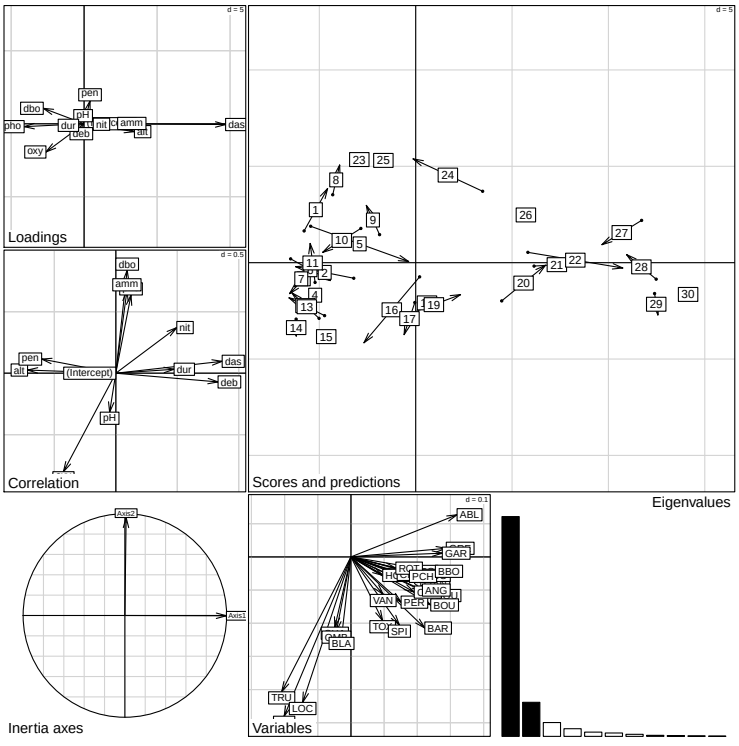
Si **Y** est analysé par une méthode induisant un triplet $(\mathbf{Y}, \mathbf{Q}_2, \mathbf{D})$, l'analyse de **Y** sous contrainte de **X** correspond à $(\mathbf{P_X Y}, \mathbf{Q}_2, \mathbf{D})$. $\mathbf{P_X}$ est le projecteur sur **X** et est égal à $\mathbf{P_X} = \mathbf{X}(\mathbf{X}^T\mathbf{D}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}$. Ces méthodes contiennent une étape de projection (régression) et requiert donc un nombre d'individus important par rapport au nombre de variables de **X**.

3.1 Analyse des redondances

L'analyse des redondances [Wollenberg, 1977] est également appelée Analyse en Composantes Principales sur Variables Instrumentales (ACPVI). Cette méthode est basée sur une ACP du tableau **Y**. D'un point de vue pratique, l'ACPVI peut être vue comme une ACP des prédictions du tableau **Y** obtenues par régressions multiples sur les variables de **X**.

```
pcafau <- dudi.pca(doubs$poi, scale = F, scannf = F, nf = 2)
pcaivdoubs <- pcaiv(pcafau, doubs$mil, scannf = F, nf = 2)
pcaivdoubs
Principal Component Analysis with Instrumental Variables
call: pcaiv(dudi = pcafau, df = doubs$mil, scannf = F, nf = 2)
class: pcaiv dudi
$rank (rank)     : 11
$nf (axis saved) : 2

eigen values: 38.42 5.954 2.416 1.339 0.7431 ...

 vector length mode    content
 $eig   11     numeric eigen values
 $lw    30     numeric row weigths (from dudi)
 $cw    27     numeric col weigths (from dudi)
```

```
data.frame nrow ncol content
$Y          30   27   Dependant variables
$X          30   11   Explanatory variables
$tab        30   27   modified array (projected variables)

data.frame nrow ncol content
$c1         27   2    PPA Pseudo Principal Axes
$as         2    2    Principal axis of dudi$tab on PAP
$ls         30   2    projection of lines of dudi$tab on PPA
$li         30   2    $ls predicted by X

data.frame nrow ncol content
$fa         12   2    Loadings (CPC as linear combinations of X
$l1         30   2    CPC Constraint Principal Components
$co         27   2    inner product CPC - Y
$cor        12   2    correlation CPC - X

iner inercum inerC inercumC ratio R2    lambda
42.7 42.7    42.6  42.6     0.996 0.902 38.4
8.16 50.9    7.76  50.4     0.989 0.767 5.95
plot(pcaivdoubs)
```



Il existe deux possibilités pour interpréter une ACPVI. L'analyse recherche des coefficients (`fa`) des variables de $\mathbf{X}$. La combinaison linéaire obtenue est une composante principale ou composante explicative (`l1`) [Obadia, 1978]. La composante explicative maximise la somme des carrés de corrélations (si $\mathbf{Y}$ est analysé par une ACP normée) ou covariances (ACP centrée) avec les variables de $\mathbf{Y}$. Les colonnes de $\mathbf{Y}$ sont alors représentées par leurs corrélations ou covariances (`co`) avec la composante explicative. Les corrélations entre $\mathbf{X}$ et la composante explicative sont dans `cor`.

```
var(pcaivdoubs$l1)/30 * 29
```

```
            RS1         RS2
RS1 1.00000e+00 1.37436e-15
RS2 1.37436e-15 1.00000e+00
```

```
head(cov(doubs$poi, pcaivdoubs$l1)/30 * 29)
```

```
           RS1        RS2
CHA -0.3010175 -0.5167752
TRU -1.2944349 -0.9924101
VAI -1.2484550 -1.1721456
LOC -0.9025339 -1.0704726
OMB -0.2711770 -0.5454216
BLA -0.1758249 -0.5916634
```

```
head(pcaivdoubs$co)
```

```
         Comp1      Comp2
CHA -0.3010175 -0.5167752
TRU -1.2944349 -0.9924101
VAI -1.2484550 -1.1721456
LOC -0.9025339 -1.0704726
OMB -0.2711770 -0.5454216
BLA -0.1758249 -0.5916634
```

```
sum(pcaivdoubs$co[, 1]^2)
```

```
[1] 38.41774
```

```
pcaivdoubs$eig[1]
```

```
[1] 38.41774
```

La deuxième interprétation de l'ACPVI consiste à calculer un pseudo axe principal (`c1`). Les lignes de $\mathbf{Y}$ sont projetées sur les pseudo axes principaux. Ces projections `ls` sont des combinaisons des variables de $\mathbf{Y}$ maximisant la variance expliquée par $\mathbf{X}$. Les prédictions de ces projections par $\mathbf{X}$ sont contenues dans `li`.

```
t(as.matrix(pcaivdoubs$c1)) %*% as.matrix(pcaivdoubs$c1)
```

```
              CS1           CS2
CS1  1.000000e+00 -3.295975e-16
CS2 -3.295975e-16  1.000000e+00
```

```
head(as.matrix(pcafau$tab) %*% as.matrix(pcaivdoubs$c1))
```

```
         CS1         CS2
1 -4.5710259  3.84649959
2 -6.2312310 -0.20450758
3 -6.5447150 -1.60614068
4 -5.2474177 -1.76428284
5 -0.3650540  0.06064447
6 -4.4324506 -0.93028165
```

```
head(pcaivdoubs$ls)
```

```
       Axis1       Axis2
1 -4.5710259  3.84649959
2 -6.2312310 -0.20450758
3 -6.5447150 -1.60614068
4 -5.2474177 -1.76428284
5 -0.3650540  0.06064447
6 -4.4324506 -0.93028165
```

```
lmprovi <- lm(pcaivdoubs$ls[, 1] ~ as.matrix(doubs$mil))
predict(lmprovi)[1:5]
```

```
        1         2         3         4         5
-5.789858 -3.213591 -5.071321 -5.186859 -5.450346
```

```
pcaivdoubs$li[1:5, 1]
```

```
[1] -5.789858 -3.213591 -5.071321 -5.186859 -5.450346
```

```
sum(predict(lmprovi)^2)/30
```

```
[1] 38.41774
```

```
summary(lmprovi)
```

```
Call:
lm(formula = pcaivdoubs$ls[, 1] ~ as.matrix(doubs$mil))
Residuals:
    Min      1Q  Median      3Q     Max
-3.6070 -1.4567 -0.1575  0.6265  5.0853

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        0.6364407 28.5981622   0.022    0.982
```

```
as.matrix(doubs$mil)das  0.0070597  0.0041025   1.721    0.102
as.matrix(doubs$mil)alt  0.0129179  0.0132863   0.972    0.344
as.matrix(doubs$mil)pen  0.3811727  1.0449578   0.365    0.720
as.matrix(doubs$mil)deb -0.0001077  0.0018553  -0.058    0.954
as.matrix(doubs$mil)pH  -0.0407674  0.3727161  -0.109    0.914
as.matrix(doubs$mil)dur -0.0228790  0.0602193  -0.380    0.708
as.matrix(doubs$mil)pho -0.0476984  0.0283163  -1.684    0.109
as.matrix(doubs$mil)nit  0.0044965  0.0142265   0.316    0.756
as.matrix(doubs$mil)amm  0.0616064  0.0716665   0.860    0.401
as.matrix(doubs$mil)oxy -0.1201557  0.0927444  -1.296    0.211
as.matrix(doubs$mil)dbo -0.0737394  0.0535007  -1.378    0.185

Residual standard error: 2.638 on 18 degrees of freedom
Multiple R-squared: 0.9019,        Adjusted R-squared: 0.842
F-statistic: 15.05 on 11 and 18 DF,  p-value: 6.73e-07
```

L'ACPVI fournit donc un compromis entre l'analyse canonique (maximisation du carré de la corrélation multiple) et l'analyse en composantes principales (maximisation de la variance) en maximisant la variance expliquée (maximisation du produit). On retrouve cette information dans l'objet `pcaivdoubs` :

```
pcaivdoubs$param
iner inercum inerC inercumC ratio R2    lambda
42.7 42.7    42.6  42.6     0.996 0.902 38.4
8.16 50.9    7.76  50.4     0.989 0.767 5.95
```

L'analyse simple trouve des combinaisons des variables de $\mathbf{Y}$ de variance maximale (`iner` et `inercum` pour le cumul). Les valeurs propres de l'ACPVI (`lambda`) sont des variances expliquées (`lambda`) et correspondent au produit de la variance (`inerC`) par le carré de la corrélation multiple (`R2`). En maximisant un compromis (la variance expliquée), on rajoute une contrainte (prédiction par les variables de $\mathbf{X}$) et la maximisation de la variance n'est donc plus optimale (elle l'est pour l'analyse simple). On mesure l'importance de cette contrainte par le ratio des variances des combinaisons des variables de $\mathbf{Y}$ des deux analyses :

```
pcafau$eig[1]
```

```
[1] 42.74627
```

```
pcaivdoubs$eig[1]
```

```
[1] 38.41774
```

```
sum(pcaivdoubs$lw * pcaivdoubs$ls[, 1]^2)
```

```
[1] 42.59456
```

```
sum(pcaivdoubs$lw * pcaivdoubs$ls[, 1]^2)/pcafau$eig[1]
```

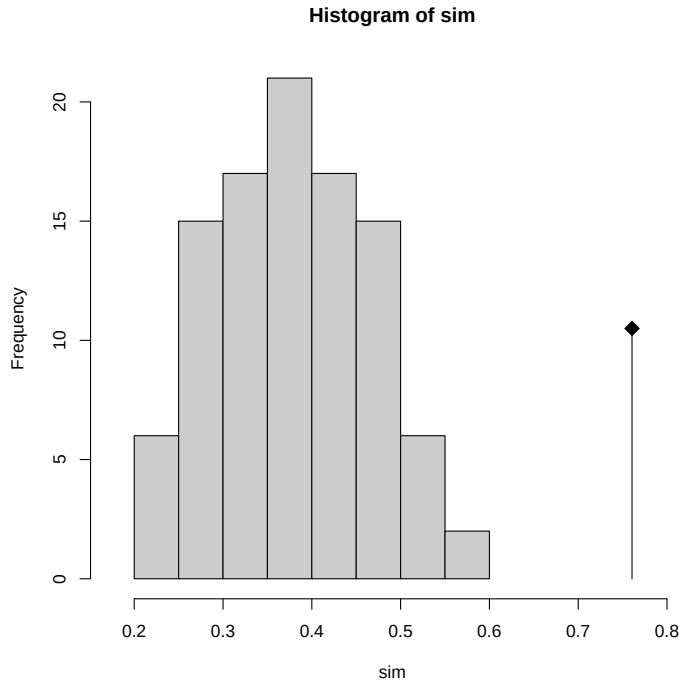
```
[1] 0.9964509
```

Finalement, un test par permutation est disponible pour tester ce lien :

```
testpcaiv <- randtest(pcaivdoubs)
testpcaiv
```

```
Monte-Carlo test
Call: randtest.pcaiv(xtest = pcaivdoubs)
Observation: 0.7605909

Based on 99 replicates
Simulated p-value: 0.01
Alternative hypothesis: greater

    Std.Obs Expectation    Variance
4.538202774 0.379338736 0.007057595
```

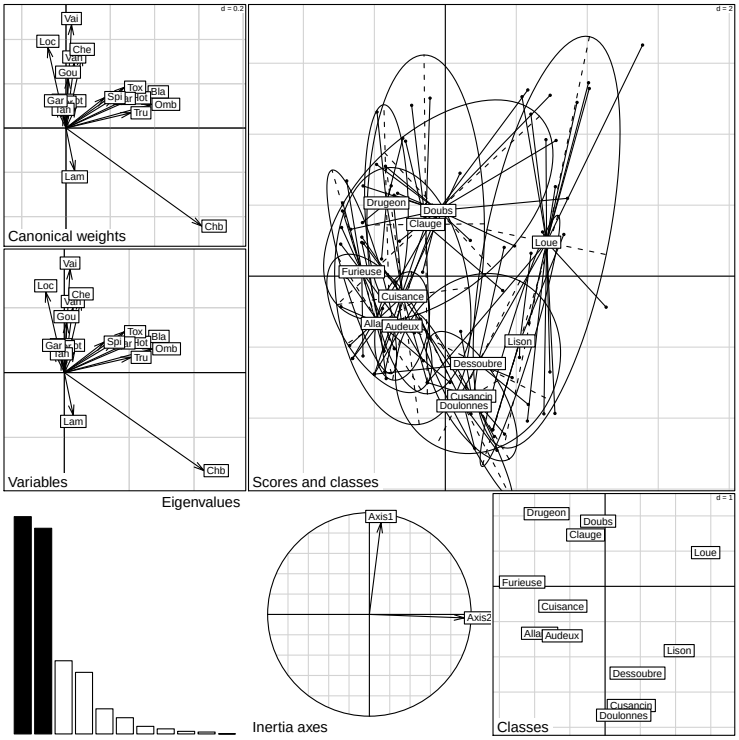
```
plot(testpcaiv)
```



3.2 Analyses inter et intra-classes

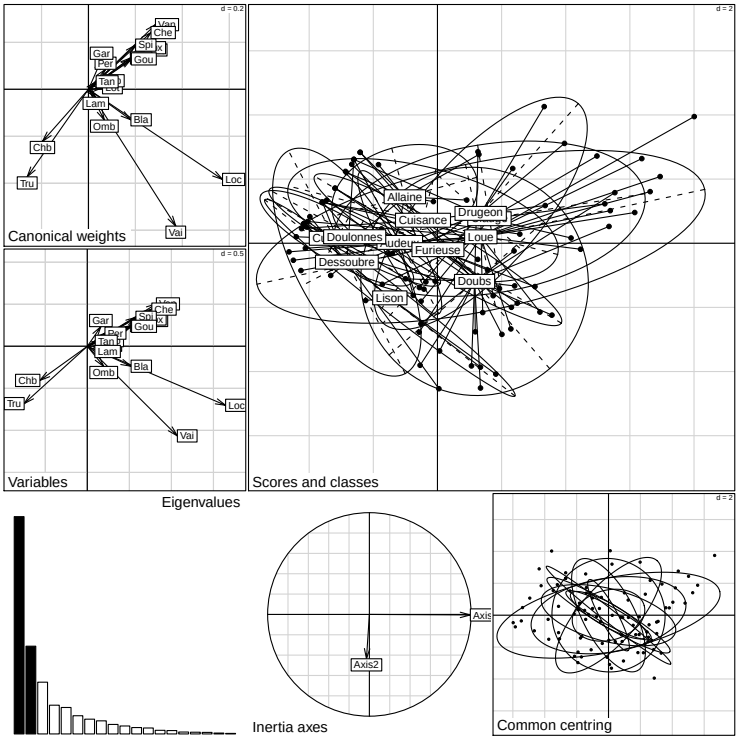
Les analyses inter et intra-classes sont des cas particuliers des analyses sur variables instrumentales où le tableau $\mathbf{X}$ contient une variable qualitative. Dans ce cas, prédire $\mathbf{Y}$ par $\mathbf{X}$ revient à remplacer la valeur d'une variable pour individu par la moyenne des individus de la même classe pour la même variable. L'analyse inter-classe est l'analyse de ce tableau de moyennes. Elle recherche des combinaisons des variables de $\mathbf{Y}$ maximisant la variance inter-classes.

```
data(jv73)
jv.pca.poi <- dudi.pca(jv73$poi, scal = F, scannf = F, nf = 2)
jv.bet.poi <- between(jv.pca.poi, jv73$fac.riv, scannf = F, nf = 2)
plot(jv.bet.poi)
```



Dans l'article original, Rao [1964] propose des analyses orthogonales afin d'éliminer un effet. L'analyse correspond au triplet $(\mathbf{P}_{\perp\mathbf{X}}\mathbf{Y}, \mathbf{Q}_2, \mathbf{D})$. $\mathbf{P}_{\perp\mathbf{X}}$ est le projecteur orthogonal à $\mathbf{X}$ et est égal à $\mathbf{P}_{\perp\mathbf{X}} = \mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{D}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{D}$. Au lieu d'analyser les prédictions de $\mathbf{Y}$ sur $\mathbf{X}$, on analyse les résidus. Dans le cas où $\mathbf{X}$ contient une variable qualitative, l'analyse orthogonale est une analyse intra-classes qui recherche des combinaisons des variables de $\mathbf{Y}$ maximisant la variance intra-classes.

```
jv.wit.poi <- within(jv.pca.poi, jv73$fac.riv, scannf = F, nf = 2)
plot(jv.wit.poi)
```

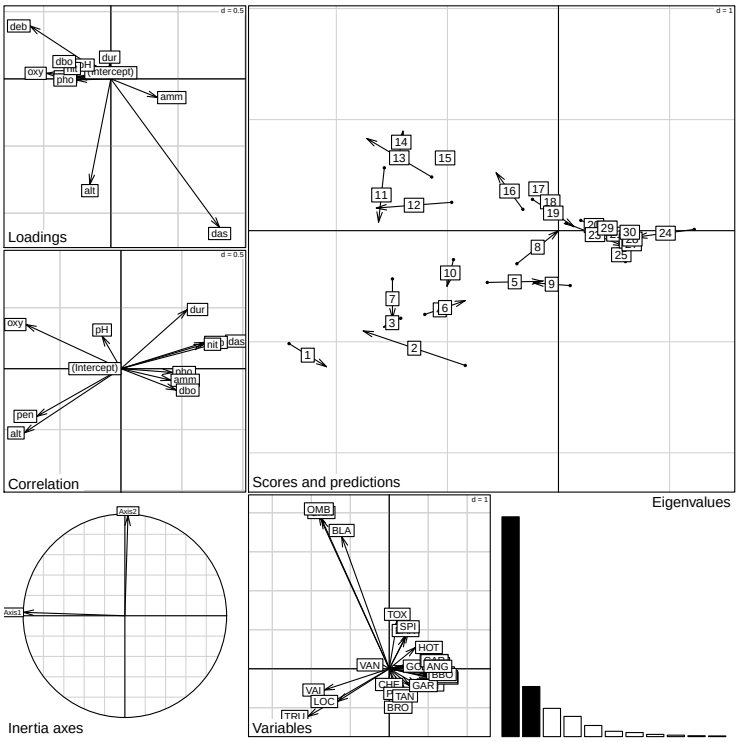


3.3 Analyse canonique des correspondances

L'analyse canonique des correspondances [Ter Braak, 1987, ter Braak, 1986] est la méthode la plus utilisée en écologie. Elle est également connue sous le nom d'Analyse Factorielle des Correspondances sur Variables Instrumentales (AFCVI,[Chessel et al., 1987, Lebreton et al., 1988a,b]). C'est une AFC de $\mathbf{Y}$ sous contrainte de $\mathbf{X}$. On peut utiliser la fonction `cca` ou `pcaiv`.

```
coafau <- dudi.coa(doubs$poi, scannf = F, nf = 2)
ccadoubs <- pcaiv(coafau, doubs$mil, scannf = F, nf = 2)
ccadoubs
Principal Component Analysis with Instrumental Variables
call: pcaiv(dudi = coafau, df = doubs$mil, scannf = F, nf = 2)
class: pcaiv dudi
$rank (rank)      : 11
$nf (axis saved) : 2

eigen values: 0.5345 0.1218 0.0687 0.04917 0.02709 ...

 vector length mode    content
 $eig   11     numeric eigen values
 $lw    30     numeric row weigths (from dudi)
 $cw    27     numeric col weigths (from dudi)

 data.frame nrow ncol content
 $Y         30   27   Dependant variables
 $X         30   11   Explanatory variables
 $tab       30   27   modified array (projected variables)

 data.frame nrow ncol content
 $c1        27   2    PPA Pseudo Principal Axes
 $as        2    2    Principal axis of dudi$tab on PAP
 $ls        30   2    projection of lines of dudi$tab on PPA
 $li        30   2    $ls predicted by X

 data.frame nrow ncol content
```

```
$fa        12   2    Loadings (CPC as linear combinations of X
$l1        30   2    CPC Constraint Principal Components
$co        27   2    inner product CPC - Y
$cor       12   2    correlation CPC - X

iner  inercum inerC inercumC ratio R2    lambda
0.601 0.601   0.597 0.597    0.994 0.895 0.535
0.144 0.745   0.142 0.739    0.992 0.857 0.122
plot(ccadoubs)
```



L'analyse recherche des coefficients (`fa`) des variables de **X**. La combinaison linéaire obtenue est une composante principale sous contrainte (`l1`). C'est un score des relevés de variance unité, combinaison linéaire des variables de milieu. Les espèces (`co`) sont positionnées à la moyenne des relevés. L'analyse maximise la variance des moyennes conditionnelles. Cette vision est parfaitement adaptée à la vision de la niche écologique et des gradients environnementaux sur lesquels se séparent les niches des espèces. Il existe un deuxième point de vue.

La deuxième interprétation de l'AFCVI consiste à calculer un pseudo axe principal (`c1`). Les lignes de **Y** (sites) sont projetées sur les pseudo axes principaux et positionnés à la moyenne des espèces qu'ils contiennent (`ls`). Les prédictions de ces projections par **X** sont contenues dans `li`.

3.3.1 Analyse discriminante

Green [1974, 1971] présente une utilisation de l'analyse discriminante multiple dont l'objectif est d'identifier les facteurs environnementaux séparant au mieux les niches des espèces. L'analyse discriminante constitue, selon Green, le modèle statistique approprié au concept de niche écologique multidimensionnelle développé par Hutchinson. Il présente ainsi les avantages de la méthode :

- L'analyse discriminante est un modèle d'analyse multivariée et le seul adapté à ce type de problème.
- Les fonctions discriminantes sont des combinaisons linéaires des variables de milieu et aucune supposition n'est faite sur le type de relation espèces-milieu.
- L'analyse repose seulement sur les présences et ne fait aucune hypothèse concernant l'absence d'une espèce dans un relevé dont il précise l'ambiguïté :

> *If the species is absent, there are three possible interpretations :(i) The species cannot live there ; that is, its niche does not include that point. (ii) The species can live there, but never had the opportunity for zoogeographic reasons. (iii) The species can and does live there, but the sample failed, by chance, to include a representative of that species.*

Le lien entre l'analyse canonique des correspondances (ACC) et l'analyse discriminante (AD) est clairement explicité dans Lebreton et al. [1988a]. L'ACC correspond à l'AD par la variable 'nom d'espèce' du tableau de variables environnementales dupliquées pour l'ensemble des correspondances (cases non nulles) du tableau $\mathbf{Y}$. Ter Braak and Verdonschot [1995] admettent le lien entre l'AD et l'ACC et précisent la différence fondamentale entre les deux méthodes :

> *In summary, the main difference between CCA and discriminant analysis is that the unit of the statistical analysis in discriminant analysis is the individual, whereas it is the site in CCA.*

```
espfac <- as.factor(names(fau)[col(as.matrix(fau))[fau > 0]])
mil <- rpjdl$mil
mildup <- mil[row(as.matrix(fau))[fau > 0], ]
dim(mil)
```

```
[1] 182   8
```

```
dim(mildup)
```

```
[1] 1639    8
```

```
pcamil <- dudi.pca(mildup, scannf = F, nf = 2)
dis.cca <- discrimin(pcamil, espfac, scannf = F, nf = 2)
cca1 <- cca(fau, mil, scannf = F, nf = 2)
cca1$eig[1:5]
```

```
[1] 0.66169822 0.17729322 0.07053920 0.03890881 0.03616246
```

```
dis.cca$eig[1:5]
```

```
[1] 0.66169822 0.17729322 0.07053920 0.03890881 0.03616246
```

3.3.2 Moyennes réciproques et régression

Ter Braak [1986] donne un algorithme itératif pour obtenir les résultats d'une ACC. C'est une extension de l'algorithme du *reciprocal averaging* de Hill [1973] dans lequel une étape de régression est ajoutée :

1. Affecter un score aléatoire aux sites.
2. Calculer la moyenne conditionnelle par espèce de ce code site : une espèce est placée à la moyenne des sites qu'elle occupe.
3. Calculer la moyenne conditionnelle par site : un site est placé à la moyenne des espèces qu'il contient.

4. Faire la régression pondérée (par la richesse des sites) du score des sites sur les variables de milieu.
5. Centrer et réduire les deux scores en utilisant les pondérations marginales.
6. Répéter les étapes 2-5 jusqu'à la convergence.

On obtient le premier axe de l'ACC. Les axes suivants sont obtenus par la même procédure, en y ajoutant une contrainte d'orthogonalité.

```
recavgcca = function(x, x2, eps = 1e-11) {
    x <- as.matrix(x)
    x2 <- as.matrix(x2)
    x <- x/sum(x)
    poili <- apply(x, 1, sum)
    poicol <- apply(x, 2, sum)
    profcol <- t(t(x)/poicol)
    proflig <- x/poili
    c1 <- matrix(rnorm(ncol(x)), ncol(x), 1)
    c1 <- scalewt(c1, poicol)
    l1 <- matrix(rnorm(nrow(x)), nrow(x), 1)
    l1 <- scalewt(l1, poili)
    x2 <- scalewt(x2, poili)
    epsi <- 1
    while (epsi > eps) {
        l1old <- l1
        c1 <- t(profcol) %*% l1
        l1 <- proflig %*% c1
        l1 <- as.matrix(predict(lm(l1 ~ x2, weights = poili)))
        l1 <- scalewt(l1, poili)
        epsi <- sum((l1 - l1old)^2)
        print(epsi)
    }
    c1 <- scalewt(c1, poicol)
    return(list(l1 = as.vector(l1), c1 = as.vector(c1)))
}
resdoubs <- recavgcca(doubs$poi[-8, ], doubs$mil[-8, ])
```

```
[1] 52.51998
[1] 0.5972074
[1] 0.005800127
[1] 7.484128e-05
[1] 1.083614e-06
[1] 1.684352e-08
[1] 2.735242e-10
[1] 4.603921e-12
```

```
ccadoubs <- cca(doubs$poi, doubs$mil, scannf = F, nf = 2)
resdoubs$l1[1:5]
```

```
[1] -3.3181915 -1.1474946 -1.9417172 -1.5101749 -0.8770918
```

```
ccadoubs$l1[1:5, 1]
```

```
[1] -3.3181916 -1.1474945 -1.9417172 -1.5101749 -0.8770917
```

```
resdoubs$c1[1:5]
```

```
[1] -1.742064 -2.088414 -1.670491 -1.324354 -1.806333
```

```
ccadoubs$c1[1:5, 1]
```

```
[1] -1.742064 -2.088414 -1.670491 -1.324354 -1.806333
```

```
poili <- apply(doubs$poi[-8, ], 1, sum)/sum(doubs$poi[-8, ])
proflig <- as.matrix(doubs$poi[-8, ]/sum(doubs$poi[-8, ]))/poili
as.vector(proflig %*% resdoubs$c1)[1:5]
```

```
[1] -2.0884143 -1.7580917 -1.5741687 -1.0337281 -0.1464066
```

```
ccadoubs$ls[1:5, 1]
```

```
[1] -2.0884144 -1.7580917 -1.5741687 -1.0337281 -0.1464065
```

4 Analyse de co-inertie

L'analyse de co-inertie [Dolédec and Chessel, 1994] est la seule à tolérer des dimensions quelconques des deux côtés et est pratiquement la seule utilisable si les variables de milieu sont qualitatives (c'est alors l'effectif des modalités qui définit la dimension du tableau $\mathbf{X}$). C'est la plus simple puisqu'en gros elle fait une double analyse d'inertie des tableaux et garantit que les deux systèmes de coordonnées sont les plus cohérents possibles. Si on considère les deux triplets d'analyse simple $(\mathbf{X}, \mathbf{Q}_1, \mathbf{D})$ et $(\mathbf{Y}, \mathbf{Q}_2, \mathbf{D})$, l'analyse de co-inertie correspond à $(\mathbf{X}^T\mathbf{D}\mathbf{Y}, \mathbf{Q}_2, \mathbf{Q}_1)$. La figure 1 de Dray et al. [2003] donne le schéma général dans le cas de deux ACP normées :

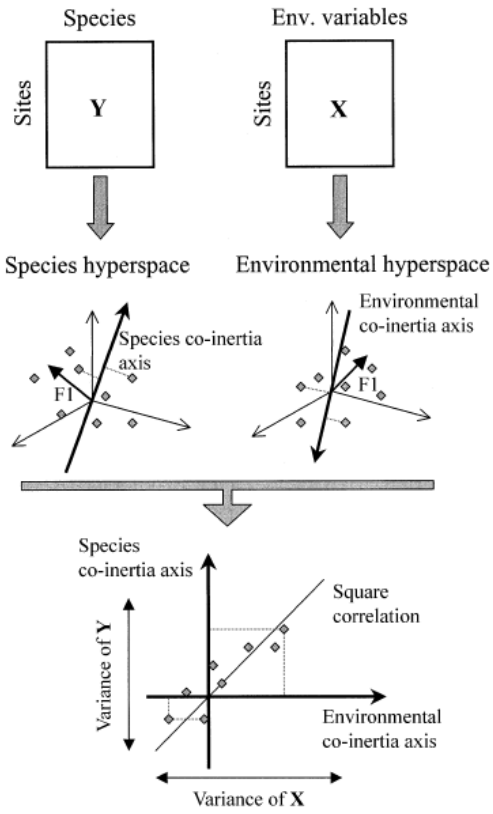
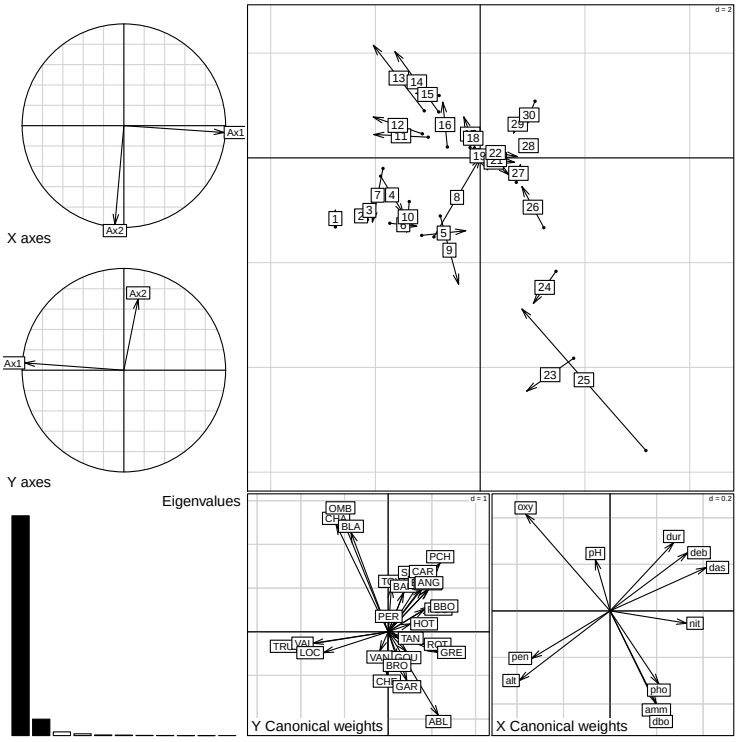


FIG. 1. Principles of co-inertia analysis (COIA). The two ecological data tables $\mathbf{X}$ and $\mathbf{Y}$ produce two representations of the sites in two hyperspaces. Separate analyses find axes maximizing inertia in each hyperspace (F1 [first factorial axis]). COIA aims to find a couple of co-inertia axes on which the sites are projected. COIA maximizes the square covariance between the projections of the sites on the co-inertia axes.

L'analyse recherche un axe de co-inertie espèces (`l1`) et un axe de co-inertie milieu (`c1`). Le tableau $\mathbf{X}$ est projeté sur l'axe de co-inertie milieu, les coordonnées des sites sont dans `lX`. Le tableau $\mathbf{Y}$ est projeté sur l'axe de co-inertie espèces, les coordonnées des sites sont dans `lY`. L'analyse maximise la co-inertie entre `lX` et `lY`.

```
coafau <- dudi.coa(doubs$poi, scannf = F, nf = 2)
pcamil <- dudi.pca(doubs$mil, row.w = coafau$lw, scannf = F, nf = 2)
```

```
coidoubs <- coinertia(pcamil, coafau, scannf = F, nf = 2)
coidoubs
Coinertia analysis
call: coinertia(dudiX = pcamil, dudiY = coafau, scannf = F, nf = 2)
class: coinertia dudi
$rank (rank)     : 11
$nf (axis saved) : 2
$RV (RV coeff)   : 0.636319

eigen values: 2.342 0.175 0.03947 0.01908 0.00658 ...

  vector length mode    content
1 $eig   11     numeric eigen values
2 $lw    27     numeric row weigths (crossed array)
3 $cw    11     numeric col weigths (crossed array)

   data.frame nrow ncol content
1  $tab       27   11   crossed array (CA)
2  $li        27   2    Y col = CA row: coordinates
3  $l1        27   2    Y col = CA row: normed scores
4  $co        11   2    X col = CA column: coordinates
5  $c1        11   2    X col = CA column: normed scores
6  $lX        30   2    row coordinates (X)
7  $mX        30   2    normed row scores (X)
8  $lY        30   2    row coordinates (Y)
9  $mY        30   2    normed row scores (Y)
10 $aX        2    2    axis onto co-inertia axis (X)
11 $aY        2    2    axis onto co-inertia axis (Y)
plot(coidoubs)
```



Dans le cas de deux ACP normées, l'analyse de co-inertie est équivalente à l'analyse inter-batteries de Tucker [1958]. Elle recherche deux combinaisons de variables de covariance maximale. On optimise ainsi un produit car $cov^2(x,y) = cor^2(x,y)var(x)var(y)$. $cor^2(x,y)$ est maximisée par l'analyse canonique des corrélations, $var(x)$ est maximisée par l'analyse simple du premier tableau et

$var(y)$ par l'analyse du second tableau. La co-inertie est un compromis entre ces trois analyses. On retrouve cette information dans le tableau suivant :

```
summary(coidoubs)
```

```
Eigenvalues decomposition:
        eig     covar      sdX       sdY      corr
1 2.3416297 1.5302384 2.366115 0.7591393 0.8519259
2 0.1749618 0.4182844 1.533416 0.3336151 0.8176473
Inertia & coinertia X:
    inertia      max     ratio
1  5.598498 5.727595 0.9774606
12 7.949863 8.153563 0.9750170

Inertia & coinertia Y:
     inertia       max     ratio
1  0.5762925 0.6009926 0.9589011
12 0.6875916 0.7453635 0.9224915

RV:
 0.636319
```

Finalement, un test par permutation est disponible pour tester ce lien :

```
testcoi <- randtest(coidoubs)
```

```
Warning: non uniform weight. The results from simulations
are not valid if weights are computed from analysed data.
```

```
testcoi
```

```
Monte-Carlo test
Call: randtest.coinertia(xtest = coidoubs)
Observation: 0.636319

Based on 999 replicates
Simulated p-value: 0.004
Alternative hypothesis: greater

    Std.Obs Expectation    Variance
4.433784322 0.239246936 0.008020281
```

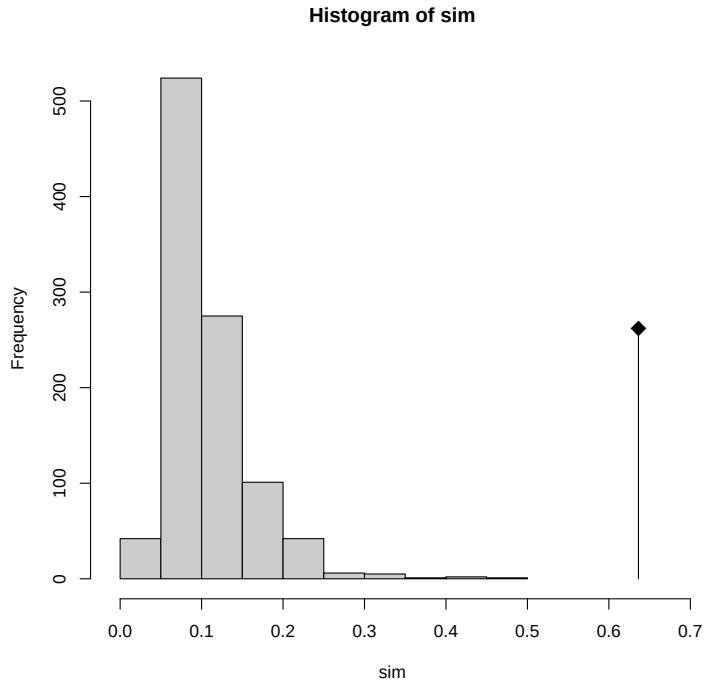
```
testcoi <- randtest(coidoubs, fixed = 2)
```

```
Warning: non uniform weight. The results from permutations
are valid only if the row weights come from the fixed table.
The fixed table is table Y : doubs$poi
```

```
testcoi
```

```
Monte-Carlo test
Call: randtest.coinertia(xtest = coidoubs, fixed = 2)
Observation: 0.636319

Based on 999 replicates
Simulated p-value: 0.001
Alternative hypothesis: greater

     Std.Obs  Expectation     Variance
10.206679828  0.106251700  0.002697075
```

```
plot(testcoi)
```



Dans le cas où les variables de milieu sont qualitatives (**X** analysé par une ACM) et **Y** (présence-absence) analysé par une AFC, la co-inertie entre **X** et **Y** est équivalente à l'analyse des profils écologique proposée par [Godron et al., 1968, Gounot, 1969].

```
poi01 <- ifelse(doubs$poi > 0, 1, 0)
milqual <- as.data.frame(lapply(doubs$mil, function(x) factor(cut(x,
    br = unique(quantile(x, seq(0, 1, le = 5))), inc = T))))
mildisj <- acm.disjonctif(milqual)
tabeco <- t(poi01) %*% as.matrix(mildisj)
coatabeco <- dudi.coa(data.frame(tabeco), scannf = F, nf = 2)
coatabeco$eig[1:5]
```

```
[1] 0.190390809 0.038979510 0.013573210 0.004476645 0.002325330
```

```
coa01 <- dudi.coa(data.frame(poi01), scannf = F, nf = 2)
acmdoubs <- dudi.acm(milqual, row.w = coa01$lw, scannf = F, nf = 2)
coi01 <- coinertia(coa01, acmdoubs, scannf = F, nf = 2)
coi01$eig[1:5]
```

```
[1] 0.190390809 0.038979510 0.013573210 0.004476645 0.002325330
```

Le couplage des tableaux est donc un univers assez complexe. Chacun des tableaux supporte sa propre analyse. Le couplage des deux peut se faire suivant trois principes généraux. Le schéma retenu pour le couple peut enfin être interprété de diverses façons. Il s'en suit une grande diversité d'expression. Fondamentalement, il y a plusieurs manières de voir et d'exprimer l'essentiel. Prévoir une réflexion préalable à toute analyse pour définir les objectifs et intégrer les propriétés connues des données, éviter les conseils impérieux de ceux qui savent ce qu'il faut faire et faire des essais préalables sans se soucier d'une expression définitive.

Références

M.P. Austin. An ordination study of a chalk grassland community. *Journal of Ecology*, 56 :739–757, 1968.

J.P. Barkham and J.M. Norris. Multivariate procedures in an investigation of vegetation and soil relations of two beach woodlands, costwold hills, england. *Ecology*, 51 :630–639, 1970.

D. Chessel, J.D. Lebreton, and N. G. Yoccoz. Propriétés de l'analyse canonique des correspondances. une utilisation en hydrobiologie. *Revue de Statistique Appliquée*, 35 :55–72, 1987.

S. Dolédec and D. Chessel. Co-inertia analysis : an alternative method for studying species-environment relationships. *Freshwater Biology*, 31 :277–294, 1994.

S. Dray, D. Chessel, and J. Thioulouse. Co-inertia analysis and the linking of ecological tables. *Ecology*, 84(11) :3078–3089, 2003.

C. Gimaret-Carpentier, S. Dray, and J.P. Pascal. Large-scale biodiversity pattern analyses of the endemic tree flora of the western ghats (india) using canonical correlation analysis of point data. *Ecography*, 26 :429–444, 2003.

R. Gittins. *Canonical analysis, a review with applications in ecology*. Springer-Verlag, Berlin, 1985.

M. Godron, P. Daget, L. Emberger, E. Le Floch, J. Poissonet, C. Sauvage, and J.P. Wacquant. *Relevé méthodique de la végétation et du milieu*. Editions du CNRS, Paris, 1968.

M. Gounot. *Méthodes d'étude quantitative de la végétation*. Masson, Paris, 1969.

R.H. . Green. Multivariate niche analysis with temporally varying environmental factors. *Ecology*, 55. 73-83 :55. 73–83, 1974.

R.H. Green. A multivariate statistical approach to the hutchinsonian niche : bivalve molluscs of central canada. *Ecology*, 52 :543–556, 1971.

M.O. Hill. Reciprocal averaging : an eigenvector method of ordination. *Journal of Ecology*, 61 :237–249, 1973.

H. Hotelling. Relations between two sets of variates. *Biometrika*, 28 :321–377, 1936.

R.H. Jongman, C.J.F. ter Braak, and O.F.R. van Tongeren. *Data analysis in community and landscape ecology*. Pudoc, Wageningen, 1987.

J.D. Lebreton, D. Chessel, R. Prodon, and N. G. Yoccoz. L'analyse des relations espèces-milieu par l'analyse canonique des correspondances. i. variables de milieu quantitatives. *Acta Œcologica, Œcologia Generalis*, 9 :53–67, 1988a.

J.D. Lebreton, M. Richardot-Coulet, D. Chessel, and N. G. Yoccoz. L'analyse des relations espèces-milieu par l'analyse canonique des correspondances . ii variables de milieu qualitatives. *Acta Œcologica, Œcologia Generalis*, 9 :137–151, 1988b.

J.D. Lebreton, R. Sabatier, G. Banco, and A.M. Bacou. Principal component and correspondence analyses with respect to instrumental variables : an overview of their role in studies of structure-activity and species- environment relationships. In J. Devillers and W. Karcher, editors, *Applied Multivariate Analysis in SAR and Environmental Studies*, pages 85–114. Kluwer Academic Publishers, 1991.

P Mercier, D. Chessel, and S. Dolédec. Complete correspondence analysis of an ecological profile data table : a central ordination method. *Acta Œcologica*, 13 :25–44, 1992.

C. Montaña and P. Greig-Smith. Correspondence analysis of species by environmental variable matrices. *Journal of Vegetation Science*, 1 :453–460, 1990.

J. Obadia. L'analyse en composantes explicatives. *Revue de Statistique Appliquée*, 24 :5–28, 1978.

R. Pélissier, S. Dray, and D. Sabatier. Within-plot relationships between tree species occurrences and hydrological soil constraints : an example in french guiana investigated through canonical correlation analysis. *Plant Ecology*, 162 :143–156, 2002.

C.R. Rao. The use and interpretation of principal component analysis in applied research. *Sankhya A*, 26 :329–359, 1964.

F. Romane. Utilisation de l'analyse multivariable en phytoécologie. *Investigaciõn pesquera*, 36 :131–139, 1972.

C.J.F. ter Braak. Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67 :1167–1179, 1986.

C.J.F. Ter Braak. Canonical correspondence analysis : a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67 :1167–1179, 1986.

C.J.F. Ter Braak. The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, 69 :69–77, 1987.

C.J.F. Ter Braak and P.F.M. Verdonschot. Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sciences*, 57 : 255–289, 1995.

L.R. . Tucker. An inter-battery method of factor analysis. *Psychometrika*, 23 : 111–136, 1958.

R.H. Whittaker. Gradient analysis of vegetation. *Biological Reviews*, 42 :207–264, 1967.

A.L. Wollenberg. Redundancy analysis, an alternative for canonical analysis. *Psychometrika*, 42 :207–219, 1977.