

L'ordination des tableaux écologiques

S. Dray & D. Chessel

Cette fiche illustre l'utilisation des principales méthodes d'ordination en écologie. Elle s'appuie essentiellement sur l'utilisation du paquet **ade4**. D'autres paquets offrent également des fonctions équivalentes ou complémentaires (*e.g.* **vegan**, **labdsv**...). On pourra trouver dans le *Task view Environmetrics* (<http://cran.r-project.org/src/contrib/Views/Environmetrics.html>) de l'information sur les paquets qui peuvent être utiles à un écologue.

Table des matières

1	Introduction	2
2	L'analyse d'un tableau floro-faunistique	3
2.1	Représenter des données	3
2.2	L'Analyse en composantes principales	4
2.3	L'Analyse factorielle des correspondances	7
2.3.1	Une procédure	8
2.3.2	Modèle Gaussien	8
2.3.3	Un artefact mathématique?	11
2.3.4	Moyennes réciproques	13
2.3.5	Variances réciproques	15
2.3.6	Retour dans le Doubs	17
3	L'analyse des variables environnementales	18
3.1	Variables quantitatives	18
3.2	Variables qualitatives	20
3.3	Le mélange	21
4	Information supplémentaire	21
	Références	23

1 Introduction

Depuis de nombreuses années, l'écologie est un champ favorable au développement et à l'application des méthodes d'analyses multivariées. Gauch [1982, p. 1] en explicite les raisons :

Community ecology concerns assemblages of plants and animals living together and the environmental and historical factors with which they interact. [...] Community data are multivariate because each sample site is described by the abundances of a number of species, because numerous environmental factors affect communities, and so on. [...] The application of multivariate analysis to community ecology is natural, routine and fruitful.

Malgré le développement de méthodes d'ordination et de classification par les phytocéologues, la famille des "analyses à vecteurs propres" (*eigenvectors analysis*), basées sur l'optimisation d'un critère à l'aide de diagonalisation de matrices, séduit rapidement les écologues [Noy-Meir and Whittaker, 1977] :

These methods were characterized by a higher level of formality and mathematical sophistication than the ad hoc ordination techniques developed directly by ecologists for ecological purposes; and their use was facilitated by the development of computers.

Qu'est ce que l'ordination ?

C'est le fait de *mettre les choses en ordre* (*putting things in order*) ce qui explique les titres des articles de Jackson and Somers [1991], Peet et al. [1988], Wartenberg et al. [1987], Palmer [1993]. Le terme est introduit en écologie par Goodall [1954].

Ordination is the collective term for multivariate techniques that arrange sites along axes on the basis of data on species composition. [Jongman et al., 1987]

The term 'ordination' derives from early attempts to order a group of objects, for example in time or along an environmental gradient. Nowadays the term is used more generally and refers to an 'ordering' in any number of dimensions (preferably few) that approximates some pattern of response of the set of objects. The usual objective of ordination is to help generate hypotheses about the relationship between the species composition at a site and the underlying environmental gradients. [Digby and Kempton, 1987]

La librairie **ade4** (**A**nalyse des **D**onnées **E**cologiques : méthodes **E**xploratoires et **E**uclidiennes en sciences de l'**E**nvironnement) regroupe un grand nombre de ces méthodes. Elle est basée sur la théorie du schéma de dualité [Escoufier, 1987] qui est implémentée par la classe **dudi**.

```
library(ade4)
apropos("dudi")
[1] "as.dudi"      "dist.dudi"    "dudi.acm"     "dudi.coa"
[5] "dudi.dec"     "dudi.fca"     "dudi.fpca"    "dudi.hillsmith"
[9] "dudi.mix"     "dudi.nsc"     "dudi.pca"     "dudi.pco"
[13] "inertia.dudi" "is.dudi"      "ktab.list.dudi" "print.dudi"
[17] "redo.dudi"    "scatter.dudi" "t.dudi"
```

2 L'analyse d'un tableau floro-faunistique

La situation : n sites sont décrits par l'abondance ou par un indice de présence-absence de s espèces, l'information est contenue dans le tableau \mathbf{Y} ($n \times s$). Comment obtenir un résumé de cette information afin de comprendre l'organisation des communautés spécifiques ?

2.1 Représenter des données

On utilise les données `doubs` (voir `?doubs`).

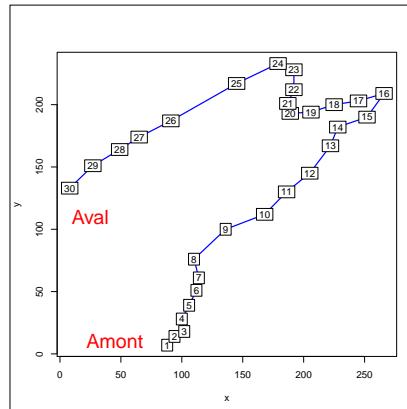
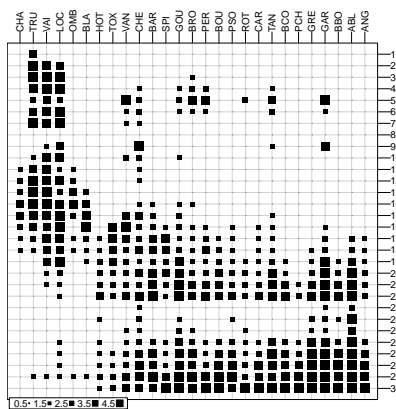
```
data(doubs)
names(doubs)
[1] "mil" "poi" "xy"
dim(doubs$poi)
[1] 30 27
```

Abondances de 27 espèces de poissons dans 30 sites :

```
table.value(doubs$poi, csize = 0.4)
```

30 sites dans l'espace géographique :

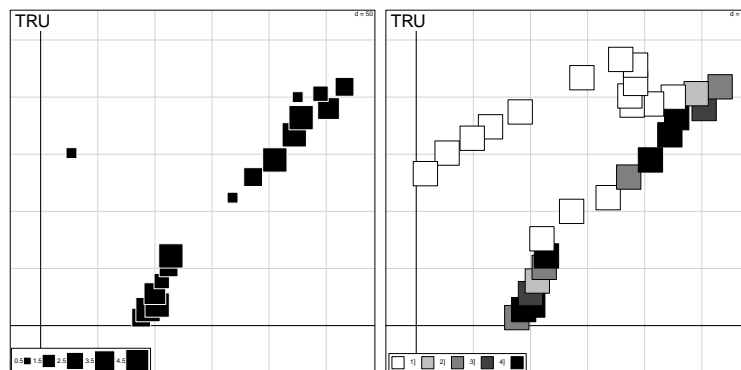
```
plot(doubs$xy, col = "blue", ty = "l", lwd = 2)
s.label(doubs$xy, add.plot = T)
text(45, 10, "Aval", cex = 2, col = "red")
text(25, 110, "Aval", cex = 2, col = "red")
```



Représenter la distribution de la truite dans l'espace géographique :

```
s.value(doubs$xy, doubs$poi[, names(doubs$poi) == "TRU"], sub = "TRU",
        csub = 2)

s.value(doubs$xy, doubs$poi[, names(doubs$poi) == "TRU"], method = "greylevel",
        sub = "TRU", csub = 2)
```



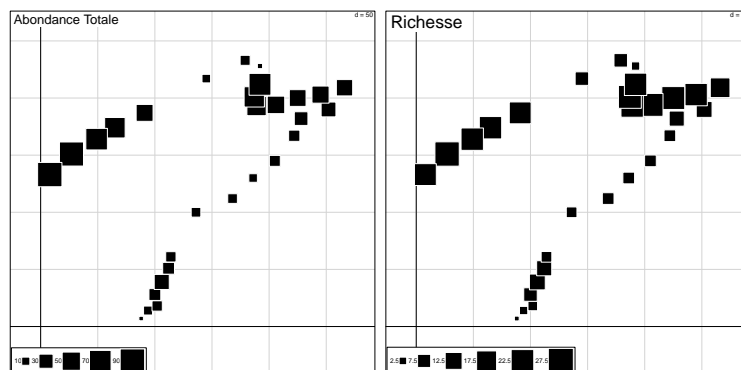
Représenter sur un graphique l'abondance de toutes les espèces :

```
par(mfrow = n2mfrow(ncol(doubs$poi)))
for (i in 1:ncol(doubs$poi)) s.value(doubs$xy, doubs$poi[, i], method = "greylevel",
  sub = names(doubs$poi)[i], csub = 2, cleg = 0)
```

Représenter l'abondance totale et la richesse spécifique :

```
s.value(doubs$xy, apply(doubs$poi, 1, sum), sub = "Abondance Totale",
  csub = 1.5)
```

```
rich = apply(ifelse(doubs$poi > 0, 1, 0), 1, sum)
s.value(doubs$xy, rich, sub = "Richesse", csub = 2)
```

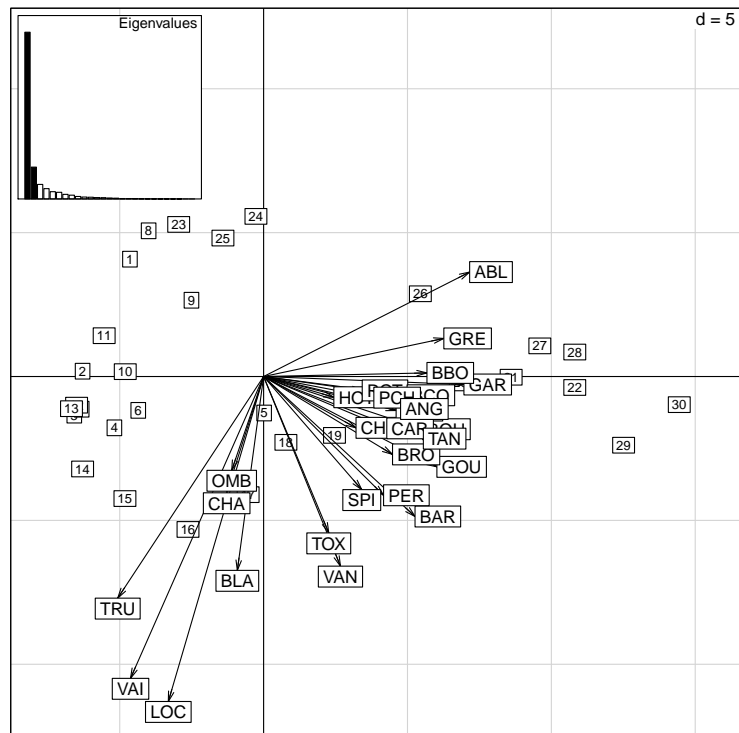


2.2 L'Analyse en composantes principales

L'ACP [Pearson, 1901, Hotelling, 1933] est utilisée dès 1954 par Goodall [1954] qui introduit le terme *ordination* en écologie. L'analyse est basée sur la diagonalisation de la matrice de covariances $(1/n)\mathbf{Y}^T\mathbf{Y}$ où le tableau $\mathbf{Y} = [y_{ij}]$ est préalablement centré par colonne (ACP centrée). Elle correspond au triplet $(\mathbf{Y}, \mathbf{I}_s, \frac{1}{n}\mathbf{I}_n)$ avec $y_{ij} = y_{ij} - \bar{y}_j$. Les facteurs synthétiques issus de cette procédure maximisent simultanément la variance des sites et la somme des covariances (au carré) des espèces.

Pour faire l'ACP, on utilise la fonction `dudi.pca` :

```
acp1 = dudi.pca(doubs$poi, scale = F, scannf = F, nf = 2)
scatter(acp1)
```



L'objet `acp1` est de la classe `dudi`. C'est une liste :

```
class(acp1)
[1] "pca" "dudi"
is.list(acp1)
[1] TRUE
acp1
Duality diagramm
class: pca dudi
$call: dudi.pca(df = doubles$poi, scale = F, scannf = F, nf = 2)
$nf: 2 axis-components saved
$rank: 27
eigen values: 42.75 8.158 3.723 2.641 1.861 ...
vector length mode content
1 $cw 27 numeric column weights
2 $lw 30 numeric row weights
3 $eig 27 numeric eigen values

data.frame nrow ncol content
1 $tab 30 27 modified array
2 $li 30 2 row coordinates
3 $li 30 2 row normed scores
4 $co 27 2 column coordinates
5 $ci 27 2 column normed scores
other elements: cent norm
```

L'analyse sépare deux types d'espèces. Peut-on en dire plus ?

```
s.label(acp1$co)
abline(0, 1/2, lwd = 3, lty = 2)
segments(1, 0, 0, 2)
abline(0, -2)
text(-1.4, 1.3, "Faune \n cyprinicole", cex = 2, col = "red")
text(0.8, -0.25, "Faune \n salmonicole", cex = 2, col = "blue")
```

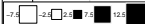


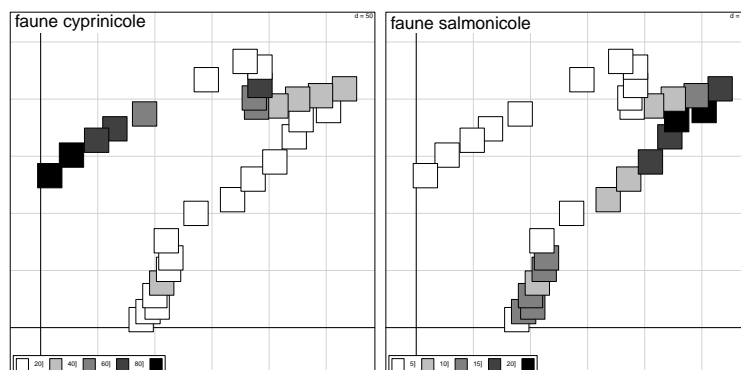
```
s.value(doubs$xy, acp1$li[, 1], sub = "Axe 1", csub = 2)

s.value(doubs$xy, acp1$li[, 2], sub = "Axe 2", csub = 2)

s.value(doubs$xy, apply(doubs$poi[, 7:27], 1, sum), sub = "faune cyprinicole",
        csub = 2, method = "greylevel")

s.value(doubs$xy, apply(doubs$poi[, 1:6], 1, sum), sub = "faune salmonicole",
        csub = 2, method = "greylevel")
```





Une modification préalable des données (profils ligne, profils colonne...) peut être envisagée en fonction de l'objectif de l'analyse et du type de données [Noy-Meir, 1973, Noy-Meir et al., 1975, Legendre and Gallagher, 2001].

2.3 L'Analyse factorielle des correspondances

Les premiers résultats obtenus en écologie à l'aide de l'ACP apparaissent très prometteurs [Noy-Meir and Whittaker, 1977] :

At some stage it may have seemed to be the definitive solution to the problem of community.

Cependant, le modèle inhérent à l'ACP est fortement critiqué. La possibilité fournie à l'utilisateur de choisir les pondérations ainsi que les transformations éventuelles du tableau n'apparaît pas comme un critère de souplesse de la méthode mais plutôt comme un désavantage :

The multiplicity of possibilities, among which the user had to choose subjectively, caused some ecologists to wonder how objective the objective methods really were. [Noy-Meir and Whittaker, 1977]

Une imposante littérature s'est attachée à critiquer le caractère linéaire de la méthode [Noy-Meir and Whittaker, 1977, Beals, 1973, Austin, 1976, Kessel and Whittaker, 1976]. En effet, les axes issus de l'ACP sont des combinaisons linéaires des variables initiales (les espèces) or, les courbes de réponse des espèces à une variable de milieu sont, la plupart du temps, curvilinéaires. Beals [1973] critique fermement l'ACP et incite au développement de méthodes d'ordination en adéquation avec les théories écologiques :

Principal components analysis, as a method of ordination to detect environmental influences, makes many unreal assumptions about ecological data. It does not take into account the normal-curve relationship between species success and environment, nor the ecological ambiguity of species absence in a stand.

L'apparition de l'analyse factorielle des correspondances (AFC) a permis de concilier la rigueur mathématique à la théorie écologique. Elle a été popularisée parallèlement par Benzécri [1969] comme une méthode d'analyse aux vecteurs propres (*eigenvectors analysis*), cas particulier de l'ACP, et par Hill [1973], sous le nom de *reciprocal averaging*, comme une extension des travaux de Whittaker [1967] sur l'analyse directe de gradients. L'ordination des tableaux floreo-faunistiques par l'AFC a permis d'annihiler certains problèmes inhérents à

l'ACP. Alors que l'ACP peut être vue comme une modélisation de l'abondance des espèces y_{ij} , l'AFC travaille sur les abondances relatives $\frac{1}{y_{++}} \frac{y_{ij}}{y_{i+} + y_{+j}} - 1$. Autrement dit, on travaille sur la composition relative plutôt que sur les abondances.

De plus, l'AFC réalise une ordination des espèces et des relevés de façon simultanée permettant de placer une espèce à la moyenne des relevés qu'elle occupe et un relevé à la moyenne des espèces qu'il contient. Selon Hill [1973], c'est le fait que l'AFC utilise une bonne ordination des espèces pour ordonner les relevés (et inversement) qui lui confère un avantage décisif par rapport à l'ACP. Le succès de l'utilisation de l'AFC en écologie a été reconnu par Digby and Kempton [1987]

Correspondence analysis (CA) is now probably the most popular ordination method among ecologists.

et confirmé par l'interprétation de Ter Braak [1985] qui présente l'AFC comme une approximation de modèle de réponse unimodale des espèces à une variable latente.

L'AFC a été et reste la méthode d'ordination la plus utilisée en écologie. Le modèle théorique de l'AFC est très riche au point d'être découvert et redécouvert par de nombreux auteurs. La méthode compte un grand nombre d'appellations différentes (*correspondence analysis, reciprocal averaging, dual scaling, canonical scoring...*), et est le sujet principal de plusieurs ouvrages Greenacre [1984], Nishisato [1980]. Elle n'en reste pas moins mystérieuse pour la plupart de ces utilisateurs.

2.3.1 Une procédure

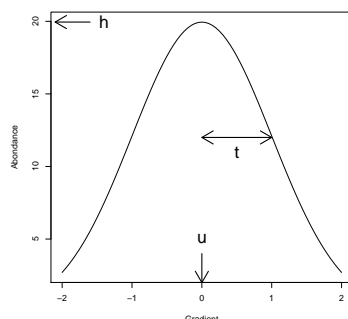
Soit le tableau \mathbf{Y} ($n \times s$). On calcule les fréquences conjointes $p_{ij} = y_{ij}/y_{++}$ et marginales $p_{i+} = y_{i+}/y_{++}$, $p_{+j} = y_{+j}/y_{++}$.

On note \mathbf{P} le tableau des p_{ij} et les matrices diagonales $\mathbf{D}_n = \text{diag}(p_{1+}, \dots, p_{n+})$ et $\mathbf{D}_s = \text{diag}(p_{+1}, \dots, p_{+s})$. Soit le tableau $\mathbf{Z} = \mathbf{D}_n^{-1} \mathbf{P} \mathbf{D}_s^{-1} - \mathbf{1}_{ns}$.

L'AFC est l'analyse de $(\mathbf{Z}, \mathbf{D}_s, \mathbf{D}_n)$ et diagonalise $\mathbf{D}_s^{1/2} \mathbf{Z}^T \mathbf{D}_n \mathbf{Z} \mathbf{D}_s^{1/2}$.

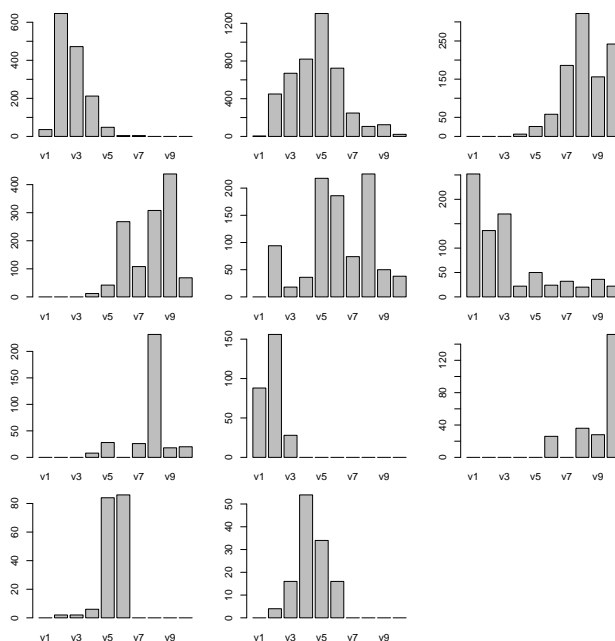
2.3.2 Modèle Gaussien

Whittaker [1967] formalise le concept de modèle unimodal : la courbe de réponse des espèces (abondance le long d'un gradient) est unimodale (un pic). On peut définir trois valeurs pour décrire ce type de courbe : l'optimum u , l'abondance à cet optimum h et la tolérance t . L'utilisation d'une fonction Gaussienne $y = h \exp\left(-\frac{(x-u)^2}{2t^2}\right)$ s'impose naturellement pour représenter ce type de réponse :



Utilisons les données **santacatalina** : Densité à l'hectare de 11 espèces d'arbres par classe de valeurs de l'humidité du sol (moyenne sur plusieurs stations par classe). Tracer les courbes de réponse des espèces le long du gradient :

```
data(santacatalina)
sant = santacatalina
par(mfrow = n2mfrow(nrow(sant)))
par(mar = c(2, 3, 1, 1))
for (i in 1:nrow(sant)) barplot(unlist(sant[i, ]))
```



On modélise les courbes de réponse à l'aide d'un modèle poissonnien : $\log(y) = b_0 + b_1x + b_2x^2$ avec $u = -b_1/2b_2$, $t = \sqrt{-1/2b_2}$ et $h = \exp(b_0 - b_1^2/4b_2)$:

```
x = 1:10
par(mfrow = n2mfrow(nrow(sant)))
par(mar = c(2, 3, 1, 1))
parmod = matrix(0, nrow(sant), 3)
colnames(parmod) = c("u", "t", "h")
for (i in 1:nrow(sant)) {
```

```

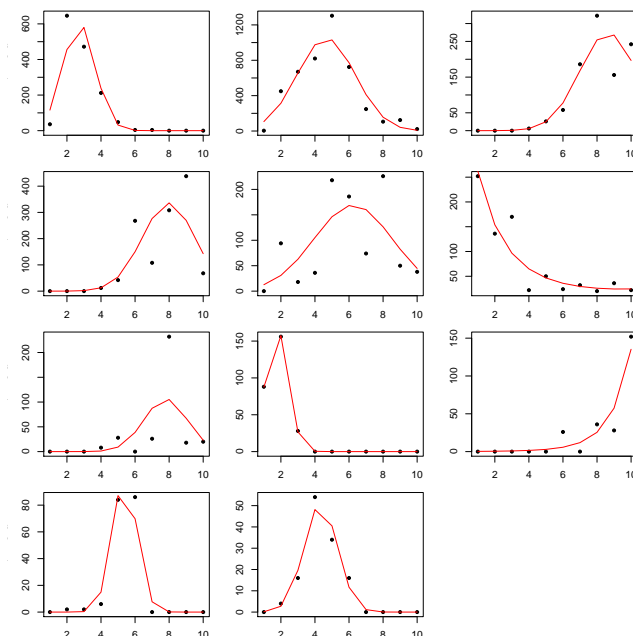
mod = glm(unlist(sant[i, ]) ~ x + I(x^2), family = "poisson")
b = coef(mod)
parmod[i, 1] = -b[2]/2/b[3]
parmod[i, 2] = sqrt(-1/2/b[3])
parmod[i, 3] = exp(b[1] - b[2]^2/4/b[3])
plot(x, unlist(sant[i, ]), pch = 20)
lines(x, predict(mod, type = "response"), col = "red")
}
parmod

```

```

      u      t      h
[1,] 2.712702 0.9399125 6.076750e+02
[2,] 4.659518 1.7099978 1.050744e+03
[3,] 8.644621 1.6642481 2.740513e+02
[4,] 7.970270 1.5501211 3.364258e+02
[5,] 6.243101 2.3006235 1.692681e+02
[6,] 9.419728      NaN 2.418191e+01
[7,] 7.792686 1.2595875 1.066068e+02
[8,] 1.748298 0.6511177 1.694943e+02
[9,] -9.401426      NaN 2.663113e-02
[10,] 5.389085 0.7096344 1.011999e+02
[11,] 4.338634 0.9664241 5.118840e+01

```



On estime les trois paramètres pour chaque espèce. La procédure ne fonctionne pas pour deux espèces dont les courbes de réponse sont fortement tronquées. Dans la plupart des cas, la situation est plus complexe : on ne connaît pas le gradient *a priori*. On permute les sites :

```

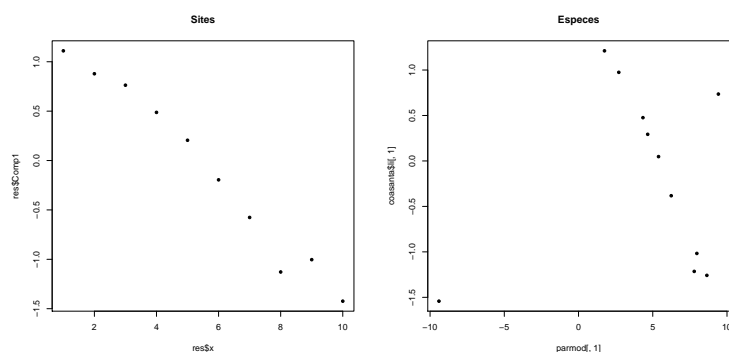
santper = santacatalina[, sample(1:ncol(santacatalina))]
coasanta = dudi.coa(santper, scannf = F, nf = 1)

res = merge(data.frame(x = 1:10, row.names = paste("v", 1:10, sep = "")),
            coasanta$co, by = "row.names")
plot(res$x, res$Comp1, main = "Sites", pch = 20)

plot(parmod[, 1], coasanta$li[, 1], main = "Especes", pch = 20)

```

L'AFC a retrouvé le gradient dans les données. Les résultats sont bons sauf pour les deux espèces dont la distribution est fortement tronquée. L'AFC fournit *une solution approchée de la solution donnée par la technique du maximum de vraisemblance dans le cas de modèles de réponse unimodale à une variable sous-jacente* [Ter Braak, 1985].

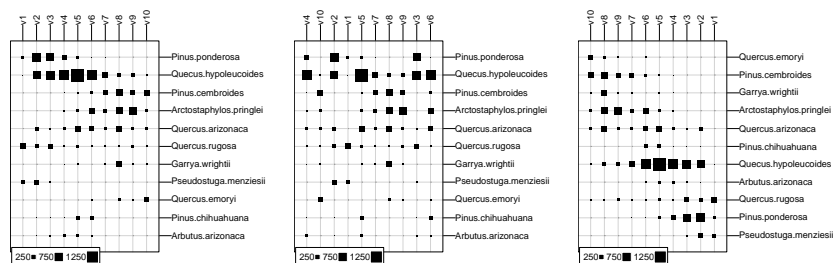


Le tableau est réorganisé :

```
table.value(sant, csize = 1.5)
```

```
table.value(santper, csize = 1.5)
```

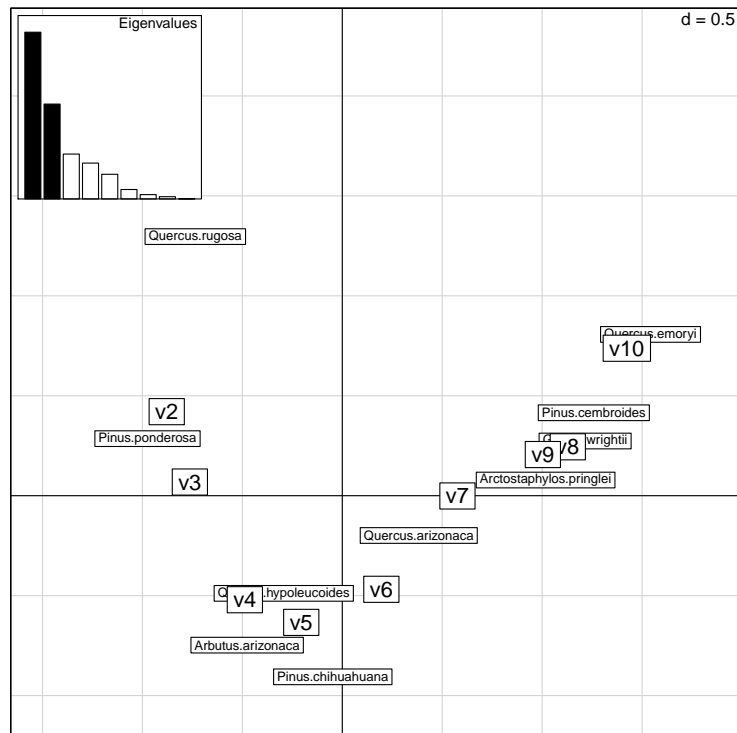
```
table.value(santper[order(coasanta$li[, 1]), order(coasanta$co[, 1])], csize = 1.5)
```



2.3.3 Un artefact mathématique ?

On reprend l'analyse précédente en conservant deux axes cette fois-ci :

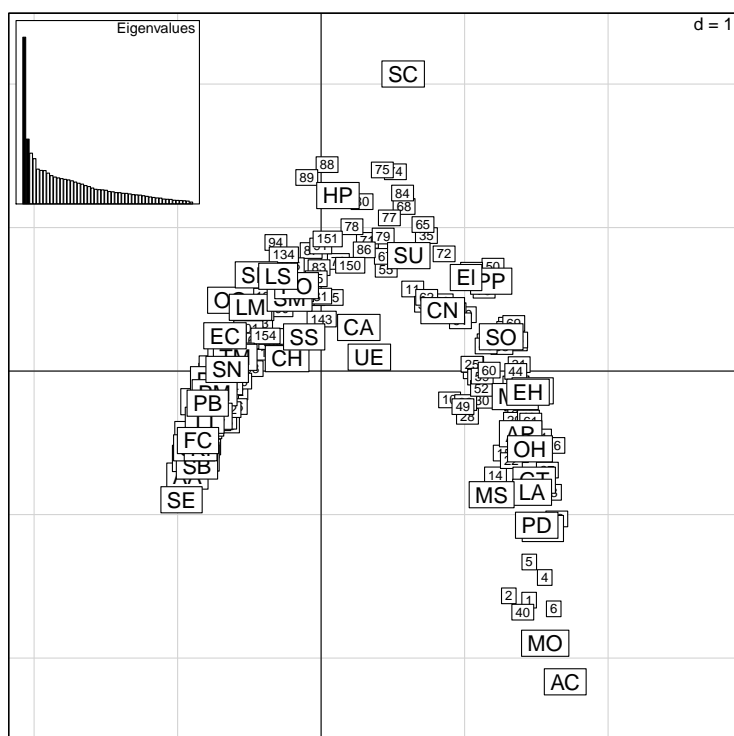
```
coasanta2 = dudi.coa(sant, scannf = F, nf = 2)
scatter(coasanta2)
```



L'AFC a la fâcheuse tendance de replier les gradients sur deux dimensions. Hill and Gauch [1980] proposent une méthode *ad hoc* pour éliminer ce problème : le deuxième axe est divisé en segments, puis les scores sont recentrés par segment. La méthode n'est pas acceptée par tous et a entraînée une vive polémique [Jackson and Somers, 1991, Peet et al., 1988, Wartenberg et al., 1987]. On y perd le caractère euclidien des méthodes d'analyse de données. Une autre approche est proposée par De'ath [1999].

Dans certains cas, une structure en fer à cheval n'est pas un artefact mais de l'information [Lebreton et al., 1988]. Ici, deux gradients (milieu ouvert, milieu fermé) :

```
data(rpjdl)
coarpjdl = dudi.coa(rpjdl$fau, scannf = F, nf = 2)
scatter(coarpjdl)
```



2.3.4 Moyennes réciproques

L'AFC correspond également à une procédure itérative dite des moyennes réciproques (*reciprocal averaging*) [Hill, 1973] :

1. Affecter un score aléatoire aux espèces.
2. Calculer la moyenne conditionnelle par site : un site est placé à la moyenne des espèces qu'il contient.
3. Calculer la moyenne conditionnelle par espèce de ce code site : une espèce est placée à la moyenne des sites qu'elle occupe.
4. Centrer et réduire les deux scores en utilisant les pondérations marginales.
5. Répéter les étapes 2-4 jusqu'à la convergence.

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

```
recavg = function(x, eps = 1e-11) {
  x = as.matrix(x)
  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)
  epsi = 1
  while (epsi > eps) {
    ciold = c1
    l1 = proflig %*% ciold
    c1 = t(profcol) %*% l1
  }
}
```

```

      c1 = scalewt(c1, poicol)
      epsi = sum((c1 - c1old)^2)
      print(epsi)
    }
    l1 = scalewt(l1, poili)
    return(list(l1 = as.vector(l1), c1 = as.vector(c1)))
  }
  rasant = recavg(sant)

[1] 9.261274
[1] 0.9863195
[1] 0.5787305
[1] 0.2614993
[1] 0.09529892
[1] 0.03202909
[1] 0.01047896
[1] 0.003400200
[1] 0.001100830
[1] 0.0003562514
[1] 0.0001152976
[1] 3.732037e-05
[1] 1.208157e-05
[1] 3.911434e-06
[1] 1.266401e-06
[1] 4.10034e-07
[1] 1.327629e-07
[1] 4.298711e-08
[1] 1.391882e-08
[1] 4.506801e-09
[1] 1.459268e-09
[1] 4.725007e-10
[1] 1.529925e-10
[1] 4.953793e-11
[1] 1.604005e-11
[1] 5.193662e-12

rasant$l1
[1] -1.25349053 -0.37678917  1.61848961  1.30827398  0.49198040 -0.94584109
[7]  1.56244114 -1.55759924  1.98341843 -0.06117552 -0.61227354

coasanta2$l1[, 1]
[1] -1.25349098 -0.37678839  1.61848895  1.30827385  0.49198071 -0.94584316
[7]  1.56244071 -1.55760179  1.98341714 -0.06117407 -0.61227235

rasant$c1
[1] -1.4284060 -1.1298615 -0.9809768 -0.6269903 -0.2641725  0.2513029  0.7409905
[8]  1.4510294  1.2906108  1.8309823

coasanta2$c1[, 1]
[1] -1.4284086 -1.1298620 -0.9809768 -0.6269897 -0.2641717  0.2513035  0.7409905
[8]  1.4510291  1.2906105  1.8309815

poili = apply(sant, 1, sum)/sum(sant)
proflig = as.matrix(sant/sum(sant))/poili
as.vector(proflig %*% rasant$c1)
[1] -0.97466049 -0.29297479  1.25846771  1.01725755  0.38254291 -0.73544606
[7]  1.21488691 -1.21112305  1.54222046 -0.04756697 -0.47607718

coasanta2$li[, 1]
[1] -0.97466069 -0.29297445  1.25846742  1.01725749  0.38254305 -0.73544697
[7]  1.21488671 -1.21112418  1.54221989 -0.04756633 -0.47607665

```

L'AFC place donc les espèces à la moyenne des sites qu'elles occupent, et les sites à la moyenne des espèces qu'ils contiennent. On retrouve cette propriété dans la procédure graphique `scatter` associée à la méthode `dudi.coa` :

```

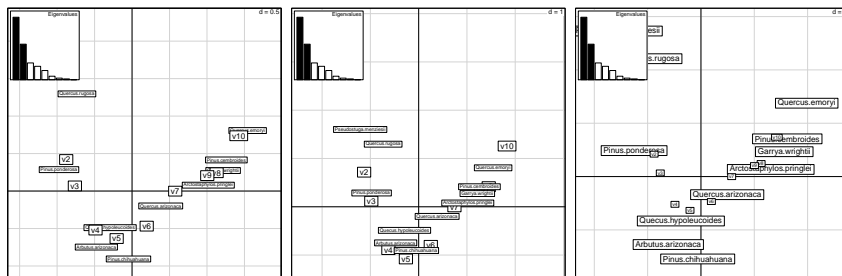
args(scatter.coa)
function (x, xax = 1, yax = 2, method = 1:3, clab.row = 0.75,
         clab.col = 1.25, posieig = "top", sub = NULL, csub = 2, ...)
NULL

scatter(coasanta2)

scatter(coasanta2, method = 2)

```

```
scatter(coasanta2, method = 3)
```



On a trois représentations fondamentales : la carte classique, les deux cartes doubles. Dans la première, on a la moitié de la seconde et la moitié de la troisième. Dans la seconde, les lignes sont positionnées avec des scores de variance unité (11) et les colonnes à la moyenne de leur distribution dans les lignes ont des coordonnées de variance λ (co). Dans la troisième, les colonnes sont positionnées avec des scores de variance unité (c1) et les lignes à la moyenne de leur distribution dans les colonnes ont des coordonnées de variance λ (1i). Les trois ont un sens [Oksanen, 1987].

2.3.5 Variances réciproques

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.

```
fau = rpjdl$fau
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))
```

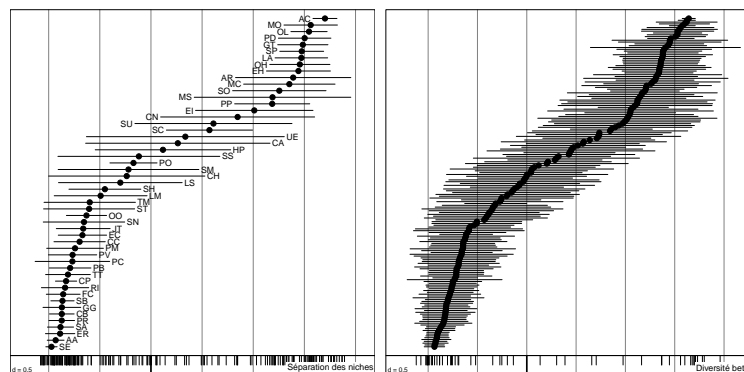
L'AFC est une analyse discriminante qui calcule un score pour les correspondances espèces pour maximiser la variance inter-sites. C'est également une analyse discriminante qui calcule un score pour les correspondances sites pour maximiser la variance inter-espèces :

```
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). Deux analyses en une, deux représentations graphiques :

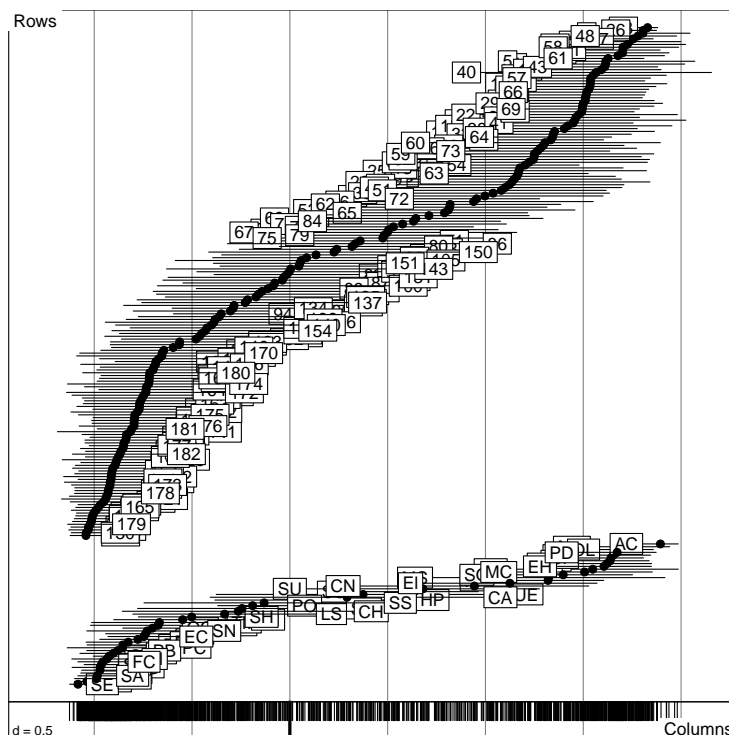
```
sco.distri(coarpjdl$l1[, 1], fau, sub = "Séparation des niches")
```

```
sco.distri(coarpjdl$c1[, 1], data.frame(t(fau)), clab = 0, sub = "Diversité beta")
```



Ces deux points de vue sont réconciliés par Thioulouse and Chessel [1992]. Au *reciprocal averaging*, on ajoute le *reciprocal scaling* : les correspondances sont positionnées par un score de variance unité qui maximise en même temps la séparation des sites et des espèces :

```
score(coarpjdl)
```



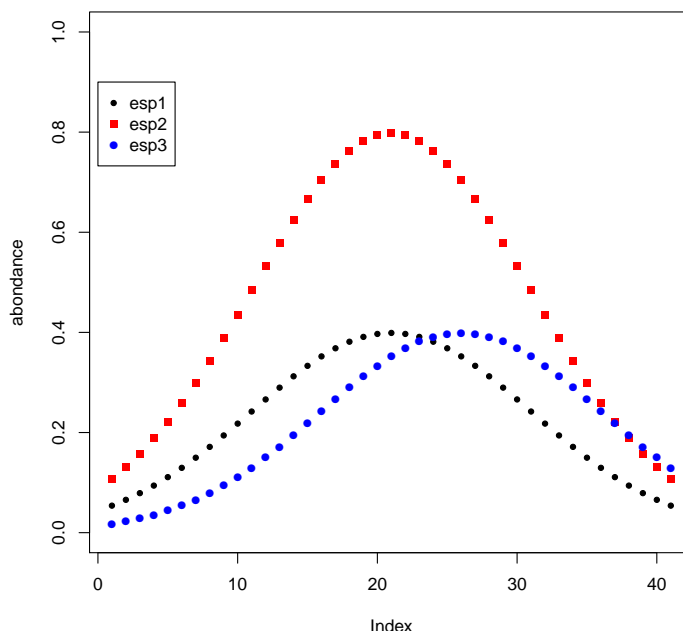
La caractéristique de l'AFC est d'être une double analyse discriminante. L'avantage est de ne pas préciser dans quel sens doit se faire la discrimination. L'inconvénient est que si on doit privilégier un point de vue, l'autre sera en oeuvre automatiquement. L'analyse non-symétrique des correspondances `dudi.nsc` peut-être une alternative [Lauro and D'Ambra, 1984, Gimaret-Carpentier et al., 1998, Kroonenberg and Lombardo, 1999].

2.3.6 Retour dans le Doubs

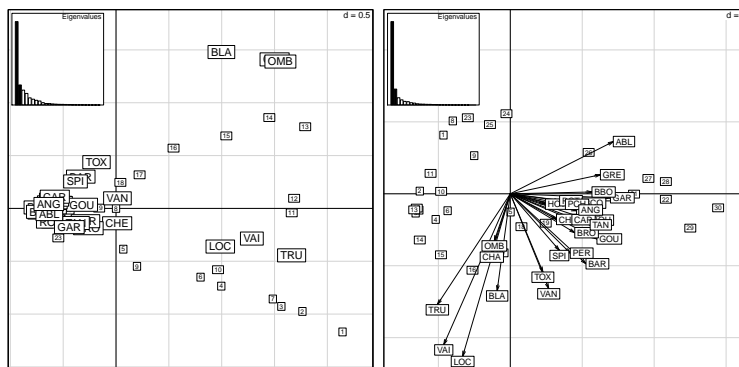
L'AFC travaille sur les compositions relatives, l'ACP sur les données brutes :

```
esp1 = dnorm(seq(-2, 2, by = 0.1))
esp2 = 2 * esp1
esp3 = dnorm(seq(-2, 2, by = 0.1), mean = 0.5)
sim = data.frame(esp1, esp2, esp3)
plot(esp1, ylab = "abundance", pch = 20, ylim = c(0, 1))
points(esp2, col = "red", pch = 15)
points(esp3, col = "blue", pch = 16)
legend(0, 0.9, c("esp1", "esp2", "esp3"), pch = c(20, 15, 16), col = c("black",
"red", "blue"))
dist.dudi(dudi.pca(sim, scale = F, scannf = F, nf = 2), amongrow = F)
      esp1      esp2
esp2 0.11728479
esp3 0.08851314 0.16513491

dist.dudi(dudi.coa(sim, scannf = F, nf = 2), amongrow = F)
      esp1      esp2
esp2 0.0000000
esp3 0.4353979 0.4353979
```



```
coadoubs = dudi.coa(doubs$poi, scannf = F, nf = 2)
scatter(coadoubs)
```



Commentaires ?

3 L'analyse des variables environnementales

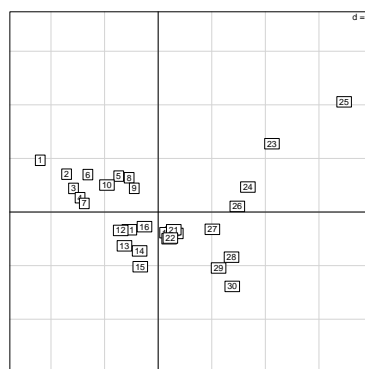
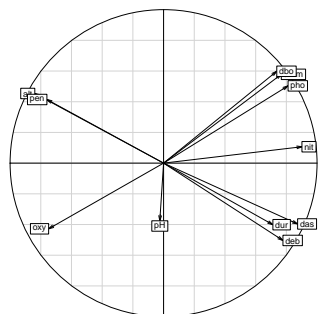
La description de la flore ou de la faune d'un certain nombre de sites s'accompagne la plupart du temps d'une description du milieu : les valeurs de p variables environnementales sont mesurées pour les n sites. L'information est contenue dans le tableau \mathbf{X} ($n \times p$).

3.1 Variables quantitatives

L'ACP sur matrice de corrélation (ACP normée) est utilisée lorsque les variables sont quantitatives. L'analyse est basée sur la diagonalisation de la matrice de corrélations $(1/n)\mathbf{X}^T\mathbf{X}$ où le tableau \mathbf{X} est préalablement normalisé par colonne. Elle correspond au triplet $(\mathbf{X}, \mathbf{I}_s, \frac{1}{n}\mathbf{I}_n)$ avec $x_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$. Les facteurs synthétiques issus de cette procédure maximisent simultanément la variance des sites et la somme des corrélations (au carré) des variables.

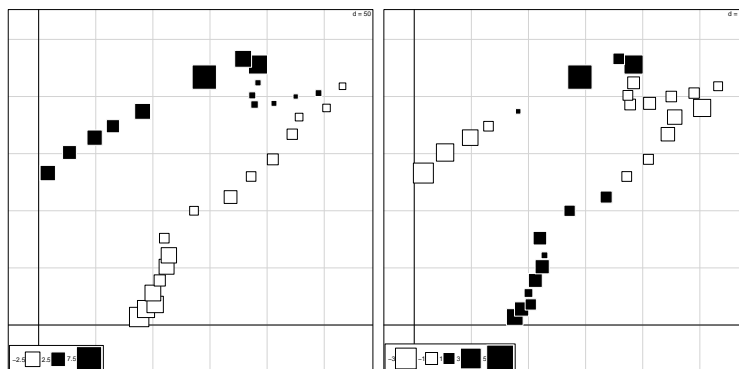
```
pcamil = dudi.pca(doubs$mil, scannf = F, nf = 2)
s.corcircle(pcamil$co)
```

```
s.label(pcamil$li)
```



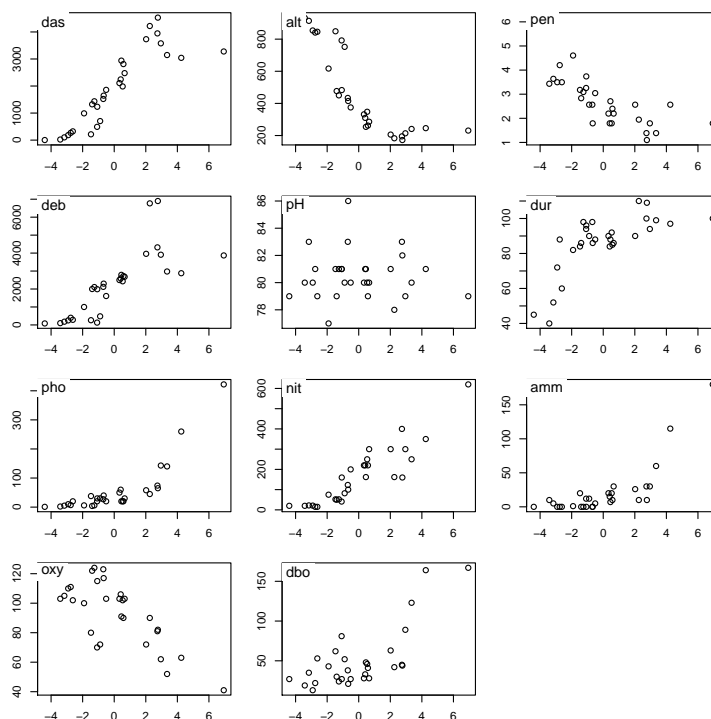
```
s.value(doubs$xy, pcamil$li[, 1])
```

```
s.value(doubs$xy, pcamil$li[, 2])
```



On a une structure amont-aval classique : (altitude et pente décroissante, débit augmentant avec la distance à la source, charge progressive en éléments minéraux et organiques). La structure est légèrement perturbée par la présence de pollution.

```
sco.quant(pcamil$li[, 1], doubs$mil)
```



3.2 Variables qualitatives

Si l'information est collectée sous la forme de variables qualitatives, le tableau est analysé par une analyse des correspondances multiples [Tenenhaus and Young, 1985]. L'analyse est disponible par la fonction `dudi.acm`. Elle maximise la variance des sites et la somme des rapports de corrélation. On transforme le tableau de milieu `doubs$mil` en tableau de variables qualitatives :

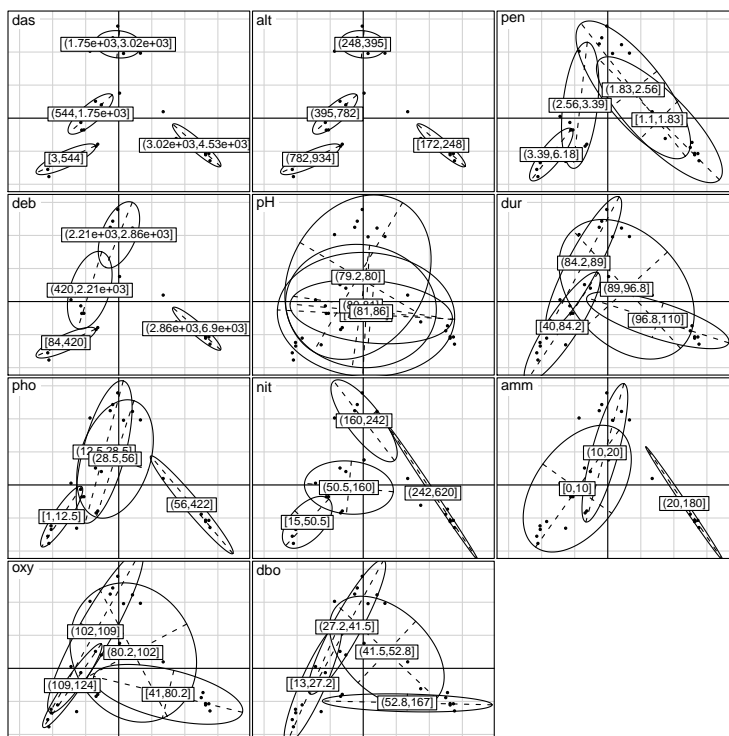
```
milqual = as.data.frame(lapply(doubs$mil, function(x) factor(cut(x,
  br = unique(quantile(x, seq(0, 1, le = 5))), inc = T))))
summary(milqual)
```

das		alt		pen		deb	
[3,544]	:8	[172,248]	:8	[1.1,1.83]	:8	[84,420]	:8
(544,1.75e+03]	:7	(248,395]	:7	(1.83,2.56]	:8	(420,2.21e+03]	:7
(1.75e+03,3.02e+03]	:7	(395,782]	:7	(2.56,3.39]	:6	(2.21e+03,2.86e+03]	:7
(3.02e+03,4.53e+03]	:8	(782,934]	:8	(3.39,6.18]	:8	(2.86e+03,6.9e+03]	:8

pH		dur		pho		nit		amm	
[77,79.2]	:8	[40,84.2]	:8	[1,12.5]	:8	[15,50.5]	:8	[0,10]	:17
(79.2,80]	:8	(84.2,89]	:7	(12.5,28.5]	:7	(50.5,160]	:8	(10,20]	:6
(80,81]	:9	(89,96.8]	:7	(28.5,56]	:7	(160,242]	:6	(20,180]	:7
(81,86]	:5	(96.8,110]	:8	(56,422]	:8	(242,620]	:8		

oxy		dbo	
[41,80.2]	:8	[13,27.2]	:8
(80.2,102]	:8	(27.2,41.5]	:7
(102,109]	:6	(41.5,52.8]	:7
(109,124]	:8	(52.8,167]	:8

```
acmdoubs = dudi.acm(milqual, scannf = F, nf = 2)
scatter(acmdoubs)
```



voir aussi `sco.boxplot` pour une représentation graphique.

3.3 Le mélange

Si le tableau contient des variables quantitatives et qualitatives, on peut utiliser les fonctions `dudi.mix` et `dudi.hillsmith` [Kiers, 1994, Hill and Smith, 1976] qui sont un compromis de l'ACP et l'ACM.

4 Information supplémentaire

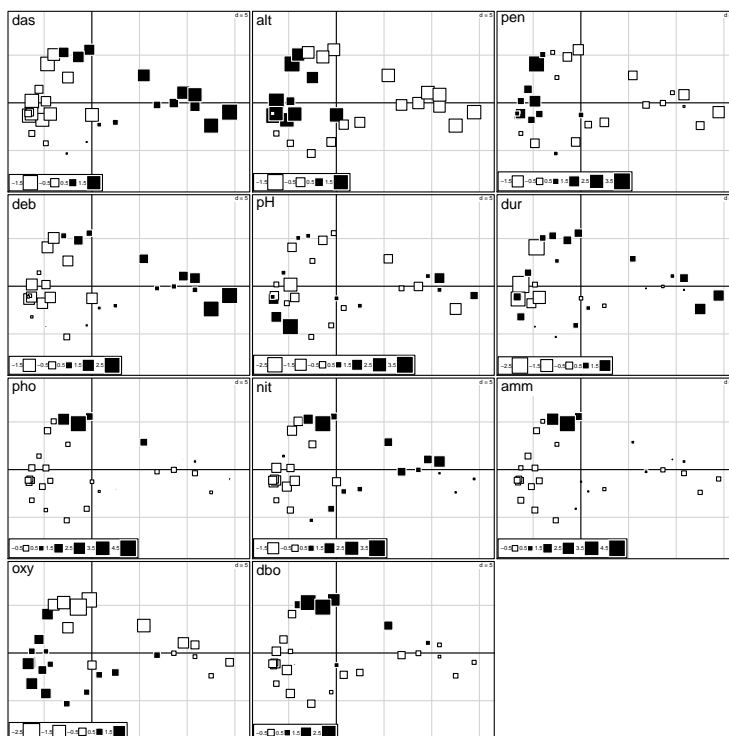
Gauch and Wentworth [1976, p. 17] évoque un point important :

a major purpose is interpretation of community relationships to environment, and not simply the representation of numerical relationships among samples or species in a hyperspace with a limited number of axes

Les éléments pour interpréter l'ordination d'un tableau floro-faunistique peuvent être calculatoire (corrélation, régression) ou graphique (représentation des variables de milieu sur le plan factoriel) :

```
cor(acp1$li, doubs$mil)
      das      alt      pen      deb      pH      dur
Axis1 0.8169072 -0.677420477 -0.57154952  0.78476404 -0.04933278 0.44764032
Axis2 0.1136704  0.001438816  0.09308912 -0.01373031 -0.25240170 0.03839049
      pho      nit      amm      oxy      dbo
Axis1 0.1148596 0.4686041 0.08400315 -0.4033553 0.08547836
Axis2 0.5371312 0.3092481 0.55786620 -0.6559081 0.70244937

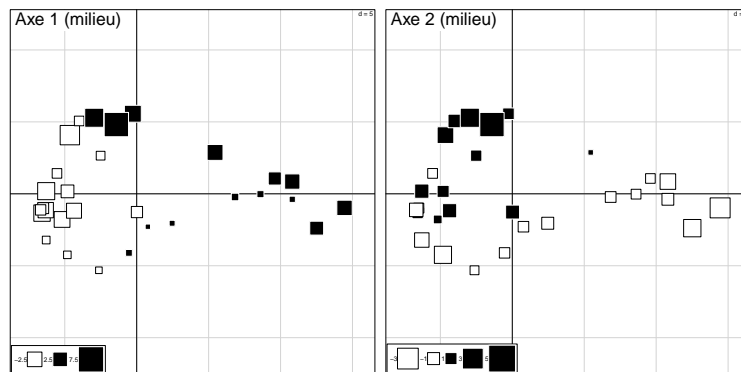
par(mfrow = n2mfrow(ncol(doubs$mil)))
for (i in 1:ncol(doubs$mil)) s.value(acp1$li, scale(doubs$mil[,
i]), sub = names(doubs$mil)[i], csub = 2)
```



Représenter les scores de l'ACP sur le milieu sur le plan factoriel de l'analyse du tableau faunistique :

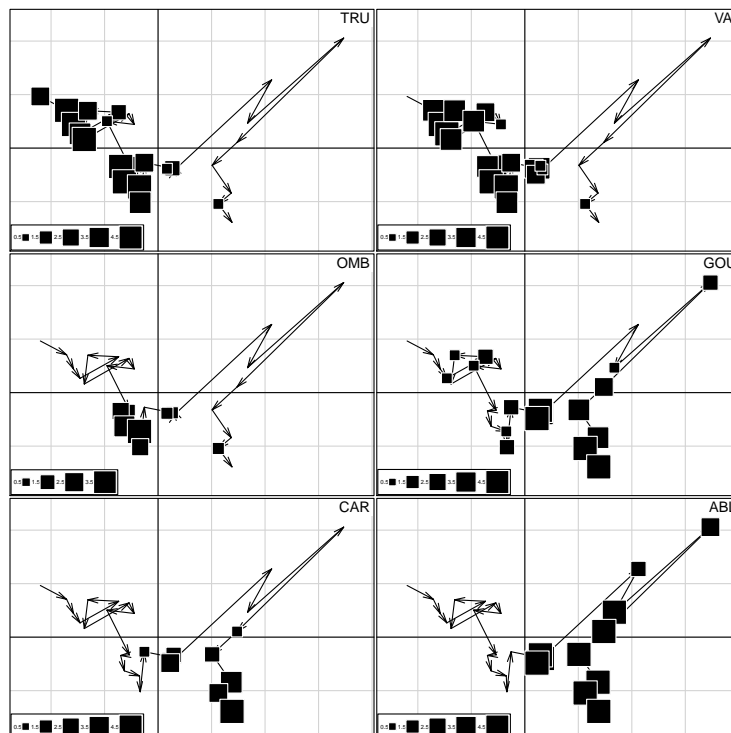
```
s.value(acp1$li, pcamil$li[, 1], sub = "Axe 1 (milieu)", csub = 2)
```

```
s.value(acp1$li, pcamil$li[, 2], sub = "Axe 2 (milieu)", csub = 2)
```



Représenter la distribution des espèces sur le plan factoriel de l'analyse du milieu :

```
par(mfrow = c(3, 2))
for (i in c(2, 3, 5, 13, 19, 26)) {
  s.traject(pcamil$li, clab = 0, cpoi = 0, sub = names(doubs$poi)[i],
    possub = "topright", csub = 2)
  s.value(pcamil$li, doubs$poi[, i], add.plot = T)
}
```



Références

- M.P. Austin. On non-linear species response models in ordination. *Vegetatio*, 33 :33–41, 1976.
- E.W. Beals. Ordination : mathematical elegance and ecological naïveté. *Journal of Ecology*, 61 :23–35, 1973.
- J.P. Benzécri. Statistical analysis as a tool to make patterns emerge from data. In S. Watanabe, editor, *Methodologies of pattern recognition*, pages 35–60. Academic Press, New York, 1969.
- G. De'ath. Extended dissimilarity : a method of robust estimation of ecological distances from high beta diversity data. *Plant Ecology*, 144 :191–199, 1999.
- P. G. N. Digby and R. A. . Kempton. *Multivariate Analysis of Ecological Communities*. Chapman and Hall, Population and Community Biology Series, London, 1987.
- Y. Escoufier. The duality diagramm : a means of better practical applications. In P. Legendre and L. Legendre, editors, *Development in numerical ecology*, pages 139–156. NATO advanced Institute , Serie G .Springer Verlag, Berlin, 1987.
- H.G. Jr. Gauch. *Multivariate analysis in community ecology*. Cambridge University Press, Cambridge, 1982.
- H.G. Jr. Gauch and T.R. Wentworth. Canonical correlation analysis as an ordination technique. *Vegetatio*, 33 :17–22, 1976.
- C. Gimaret-Carpentier, D. Chessel, and J.P. Pascal. Non-symmetric correspondence analysis : an alternative for community analysis with species occurrences data. *Plant Ecology*, 138 :97–112, 1998.
- D.W. Goodall. Objective methods for the classification of vegetation iii. an essay in the use of factor analysis. *Australian Journal of Botany*, 2 :304–324, 1954.
- M.J. Greenacre. *Theory and applications of correspondence analysis*. Academic Press, London, 1984.
- M.O. Hill. Reciprocal averaging : an eigenvector method of ordination. *Journal of Ecology*, 61 :237–249, 1973.
- M.O. Hill and H.G. Gauch. Detrended correspondence analysis : an improved ordination technique. *Vegetatio*, 42 :47–58, 1980.
- M.O. Hill and A.J.E. Smith. Principal component analysis of taxonomic data with multi-state discrete characters. *Taxon*, 25 :249–255, 1976.
- H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24 :498–520, 1933.
- D.A. Jackson and K.M. Somers. Putting things in order : the ups and downs of detrended correspondence analysis. *The American Naturalist*, 137 :707–712, 1991.

- R.H. Jongman, C.J.F. ter Braak, and O.F.R. van Tongeren. *Data analysis in community and landscape ecology*. Pudoc, Wageningen, 1987.
- S.R. Kessel and R.H. Whittaker. Comparison of three ordination techniques. *Vegetatio*, 32 :21–29, 1976.
- H.A.L. Kiers. Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika*, 56 :197–212, 1994.
- P.M. Kroonenberg and R. Lombardo. Nonsymmetric correspondence analysis : a tool for analysing contingency tables with a dependence structure. *Multivariate Behavioral Research*, 34 :367–396, 1999.
- N. Lauro and L. D’Ambra. L’analyse non symétrique des correspondances. In E. Diday, M. Jambu, L. Lebart, J. Pages, and R. Tomassone, editors, *Data Analysis and Informatics, III*, pages 433–446. Elsevier, North-Holland, 1984.
- 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 Oecologica, Ecologia Generalis*, 9 :53–67, 1988.
- P. Legendre and E.D. Gallagher. Ecological meaningful transformations for ordination of species data. *Oecologia*, 129 :271–280, 2001.
- S. Nishisato. *Analysis of categorical data : dual scaling and its applications*. University of Toronto Press, London, 1980.
- I. Noy-Meir. Data transformations in ecological ordination. i. some advantages of non-centering. *Journal of Ecology*, 61 :329–341, 1973.
- I. Noy-Meir and R.H. Whittaker. Continuous multivariate methods in community analysis : some problems and developments. *Vegetatio*, 33 :79–98, 1977.
- I. Noy-Meir, D. Walker, and W.T. Williams. Data transformation in ecological ordination. ii on the meaning of data standardization. *Journal of Ecology*, 63 :779–800, 1975.
- J. Oksanen. Problems of joint display of species and site scores in correspondence analysis. *Vegetatio*, 72 :51–57, 1987.
- M. Palmer. Putting things in even better order : the advantages of canonical correspondence analysis. *Ecology*, 74 :2215–2230, 1993.
- K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2 :559–572, 1901.
- R.K. Peet, R.G. Knox, J.S. Case, and R.B. Allen. Putting things in order : the advantages of detrended correspondence analysis. *The American Naturalist*, 131 :924–934, 1988.
- M. Tenenhaus and F.W. Young. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, 50 :91–119, 1985.

- C.J.F. Ter Braak. Correspondence analysis of incidence and abundance data : properties in terms of a unimodal response model. *Biometrics*, 41 :859–873, 1985.
- J. Thioulouse and D. Chessel. A method for reciprocal scaling of species tolerance and sample diversity. *Ecology*, 73 :670–680, 1992.
- D. Wartenberg, S. Ferson, and F.J. Ohlf. Putting things in order : a critique of detrended correspondence analysis. *The American Naturalist*, 129 :434–448, 1987.
- R.H. Whittaker. Gradient analysis of vegetation. *Biological Reviews*, 42 :207–264, 1967.