M2 Écologie,Évolution,Biométrie 
UE Description Statistique des Structures Biologiques

L'ordination directe

D. Chessel

Notes de cours cssb6

Quelques principes de l'ordination écologique.

Table des matières

1 Introduction

L'ordination désigne l'art de mettre en ordre, la pratique et la méthodologie associée. Sous le titre *Ordination Methods for Ecologists*, un site fameux lui est entièrement consacré :

`http://ordination.okstate.edu/`

Le forum associé est ORDNEWS.

La question est posé par la notion de relevés écologiques (*phytosociological relevés* est une expression anglaise). Un segment tracé par une corde, une surface, un volume d'eau, d'air, de sol, de tout ce qu'on veut, un piège ou un intervalle de temps pour un observateur permet d'identifier des êtres vivants par catégories. On obtient ainsi des matrices relevés-espèces.

L'abondance des espèces (ou catégories au sens plus large) dans les relevés se fait par mesure de présence-absence, densité, biomasse, dénombrement, recouvrement, fréquence, ... En général, ces matrices sont de dimensions élevées (50, 500, 5000 espèces pour 10, 100,1000, 10000 relevés), contiennent beaucoup de zéros, ont des marges très hétérogènes. Chaque relevé a une composante aléatoire forte (historique, expérimentale) et n'a de sens que par rapport aux autres. L'ensemble des différences deux à deux entre espèces et entre relevés définit une structure qui se définit par des méthodes d'ordination ou de classification.

L'écologie des communautés a ainsi contribué grandement au développement de l'analyse multivariée. La littérature associée est énorme. Un pilier est la synthèse de R.H. Whittaker [34] basée sur la disposition des espèces sur les gradients environnementaux, modèle qui s'oppose à celui de l'assemblage des espèces de la phytosociologie [9]. On trouve ici quelques repères sur un débat complexe, renouvelé actuellement par l'usage des marqueurs moléculaires ou des traits biologiques.

L'ordination directe est celle qui part des mesures de milieu pour décrire l'organisation des taxons. H.G. Gauch [17] en a été un des défenseurs les plus connus.

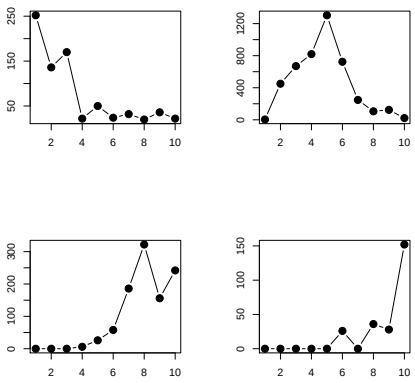
2 Courbes de réponse

Utiliser un jeu de données classiques [18] :

```
library(ade4)
data(santacatalina)
w <- t(santacatalina[c(6, 2, 3, 9), ])
```

Éditer ce petit tableau d'intérêt pédagogique. On y trouve la 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 de quelques espèces le long du gradient :

```
par(mfrow = c(2, 2))
invisible(apply(w, 2, plot, type = "b", ylab = "", xlab = "", pch = 19,
    cex = 1.5))
```

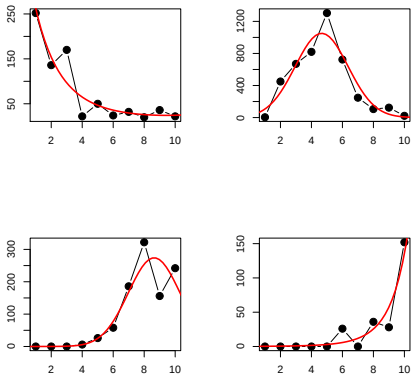


Les courbes de réponses sur un gradient sont en général **unimodales** et présentent un mode (*préférendum*) et un étalement (*amplitude*) caractéristique de l'espèce. On suppose qu'aux deux bouts du gradient des courbes monotones sont l'effet de l'amplitude d'échantillonnage. L'ordination *gaussienne* ajuste à ces observations des courbes du type :

$$y = A \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

On ferait aujourd'hui un glm :

```
fun1 <- function(k) {
    x <- 1:10
    y <- w[, k]
    glm0 <- glm(y ~ x + I(x^2), family = poisson)
    xnew <- seq(0, 11, le = 50)
    ynew <- predict(glm0, type = "response", newdata = list(x = xnew))
    plot(x, y, type = "b", ylab = "", xlab = "", pch = 19, cex = 1.5)
    lines(xnew, ynew, lwd = 2, col = "red")
}
par(mfrow = c(2, 2))
lapply(1:4, fun1)
```



Ce qui est important est l'absence de *linéarité* en terme de modèle sous forme de droite. Ceci a généré l'idée que les méthodes de l'algèbre *linéaire* étaient inadaptées (ce qui est absurde). Une question fort délicate est celle de la modélisation automatique des courbes de réponses dans les grands tableaux.

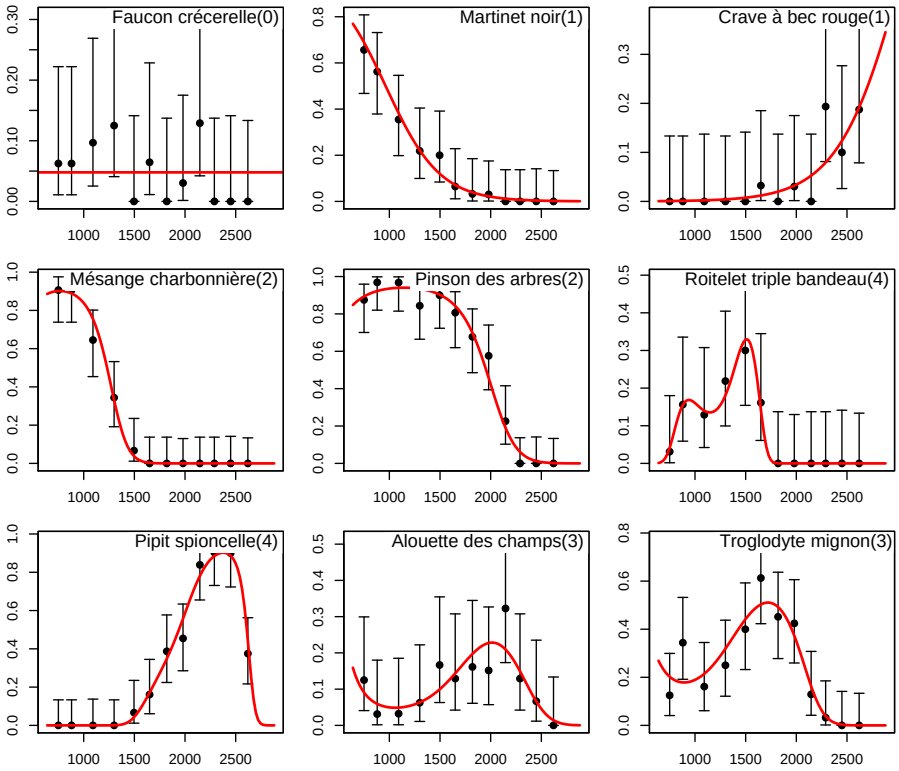
```
data(tarentaise)
source("http://pbil.univ-lyon1.fr/R/donnees/plot.freq.grad.R")
```

Le descriptif est dans :

`http://pbil.univ-lyon1.fr/R/pps/pps038.pdf`

La modélisation de la courbe de réponse en présence-absence des espèces d'oiseaux à l'altitude mette en évidence une réelle diversité de modèles vraisemblables, l'hésitation se concentrant chez les taxa les moins fréquents.

```
etud <- function(num, ncla = 12, possub = "topright") {
    x <- tarentaise$alti
    y <- tarentaise$ecol[, num]
    nomesp <- tarentaise$frnames[num]
    plot.freq.grad(x, y, ncla, show = F)
    if (sum(y) > 15) {
        x0 <- seq(min(x), max(x), le = 100)
        glm1 <- glm(y ~ x, family = binomial)
        glm2 <- glm(y ~ x + I(x^2), family = binomial)
        glm3 <- glm(y ~ x + I(x^2) + I(x^3), family = binomial)
        glm4 <- glm(y ~ x + I(x^2) + I(x^3) + I(x^4), family = binomial)
        ypred1 <- predict(glm1, newdata = list(x = x0), type = "response")
        ypred2 <- predict(glm2, newdata = list(x = x0), type = "response")
        ypred3 <- predict(glm3, newdata = list(x = x0), type = "response")
        ypred4 <- predict(glm4, newdata = list(x = x0), type = "response")
        ano0 <- anova(glm1, test = "Chisq")
        ano1 <- anova(glm1, glm2, test = "Chisq")
        ano2 <- anova(glm2, glm3, test = "Chisq")
        ano3 <- anova(glm3, glm4, test = "Chisq")
        if (ano3[2, 5] < 0.05)
            opt <- 4
        else if (ano2[2, 5] < 0.05)
            opt <- 3
        else if (ano1[2, 5] < 0.05)
            opt <- 2
        else if (ano0[2, 5] < 0.05)
            opt <- 1
        else opt <- 0
    }
    else opt <- 0
    if (opt == 0)
        abline(h = mean(y), lwd = 2, col = "red")
    else if (opt == 1)
        lines(x0, ypred1, lwd = 2, col = "red")
    else if (opt == 2)
        lines(x0, ypred2, lwd = 2, col = "red")
    else if (opt == 3)
        lines(x0, ypred3, lwd = 2, col = "red")
    else if (opt == 4)
        lines(x0, ypred4, lwd = 2, col = "red")
    nomesp = paste(nomesp, "(", opt, ")", sep = "")
    scatterutil.sub(nomesp, 2, possub)
    box()
}
par(mar = c(2.5, 2.5, 1, 1))
par(mfrow = c(3, 3))
etud(num = 39)
etud(6)
etud(88)
etud(8)
etud(11)
etud(54)
etud(84)
etud(36)
etud(26)
```



3 Profils écologiques

Dans l'ordination proprement dite, l'objectif est moins de collectionner les modèles que de positionner les espèces les unes par rapport aux autres et de décrire le mode d'assemblage.

```
sco.distri(tarentaise$alti, tarentaise$ecol, inc = F, clab = 0.75)
```

La moyenne et la variance des positions occupées sont des outils simples et efficaces (figure 1). Mais la question du modèle de l'organisation du renouvellement des espèces dans le gradient est posée. On conserve plus d'information avec un profil écologique [19][20].

L'ensemble des profils écologiques d'une espèce définissent sa niche, l'ensemble des espèces ayant des profils écologiques voisins définissent des groupes écologiques. Ces concepts ont établi la notoriété de l'écologie sigmatiste dite *Zurich-Montpellier School*. S

```
profalti <- data.frame(t(apply(tarentaise$ecol, 2, function(x) tapply(x,
    tarentaise$envir$alti, sum))))
dim(profalti)
```

```
[1] 98 14
```

```
sum(profalti)
```

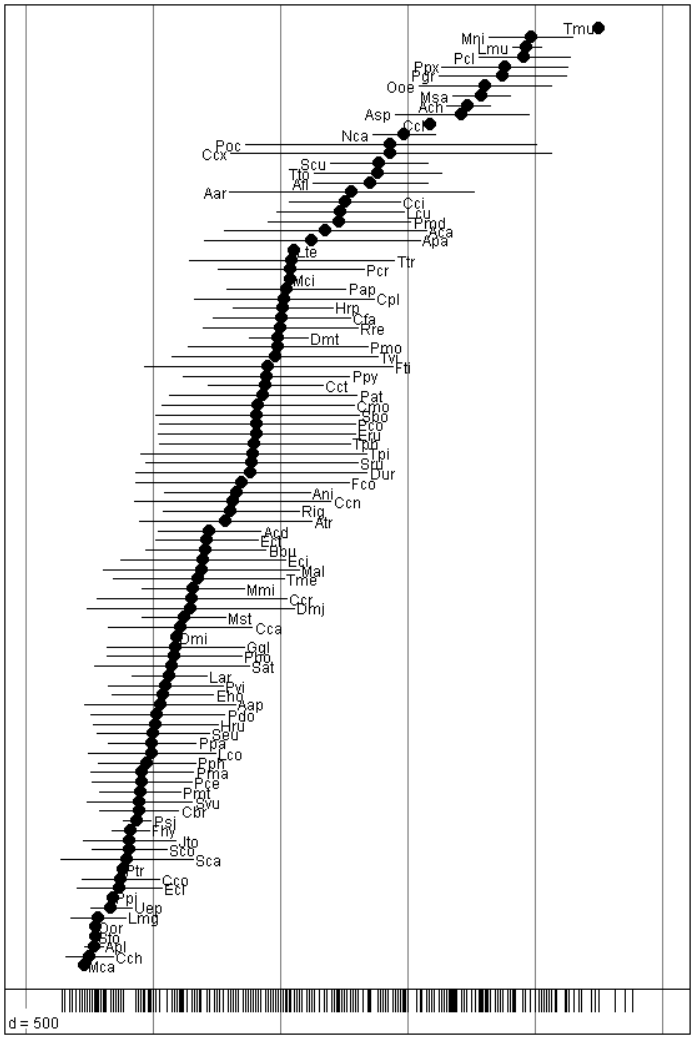


FIG. 1 – En bas : 376 relevés d'avifaune entre 500 et 3000 m d'altitude. Au dessus 98 espèces d'oiseaux représentées par la moyenne et l'écart-type des positions où elles sont présentes. C'est l'archétype de l'ordination des espèces par un gradient environnemental.

```
[1] 3444
```

Le tableau `profalti` contient simplement le nombre de présence (nombre de relevés où l'espèce est présente) dans chaque classe d'altitude. Ces distributions n'ont de sens que par rapport à la distribution des stations dans les mêmes classes :

```
ref <- unlist(as.list(table(tarentaise$envir$alti)))
ref
```

```
A0600 A0750 A0900 A1050 A1200 A1350 A1500 A1650 A1800 A1950 A2100 A2250 A2400 A2550
   17    31    24    21    24    25    35    26    28    31    29    31    30    24
```

Intervient ici un rupture très importante. Il y a deux manières de concevoir l'analyse d'un tel tableau et plus généralement les données floro-faunistiques. La première voit le relevé comme un individu et l'espèce comme une variable. La seconde voit l'occurence comme un individu, le relevé et l'espèce comme des variables. Ceci est très étonnant et très sous-estimé.

Considérons d'abord les relevés-individus. L'espèce est une variable. Le lien entre cette espèce et la variable de milieu est la table de contingence du profil brut.

```
profalti[1, ]
```

```
    A0600 A0750 A0900 A1050 A1200 A1350 A1500 A1650 A1800 A1950 A2100 A2250 A2400
Mal     6     6     1     7     3     4     7     2     0     1     0     0     0
    A2550
Mal     0
```

```
table(tarentaise$ecol[, 1], tarentaise$envir$alti)
```

```
    A0600 A0750 A0900 A1050 A1200 A1350 A1500 A1650 A1800 A1950 A2100 A2250 A2400
  0    11    25    23    14    21    21    28    24    28    30    29    31    30
  1     6     6     1     7     3     4     7     2     0     1     0     0     0
    A2550
  0    24
  1     0
```

```
chisq.test(tarentaise$ecol[, 1], tarentaise$envir$alti, sim = T)
```

```
        Pearson's Chi-squared test with simulated p-value (based on 2000
        replicates)
data:  tarentaise$ecol[, 1] and tarentaise$envir$alti
X-squared = 52.0068, df = NA, p-value = 0.0004998
```

Chaque ligne du tableau de profils se se compare à la même référence qui en donne la signification (sinon on ferait l'analyse des exigences écologiques des poteaux de téléphone). L'analyse d'un tel tableau est une analyse des correspondances décentrée [15][10] :

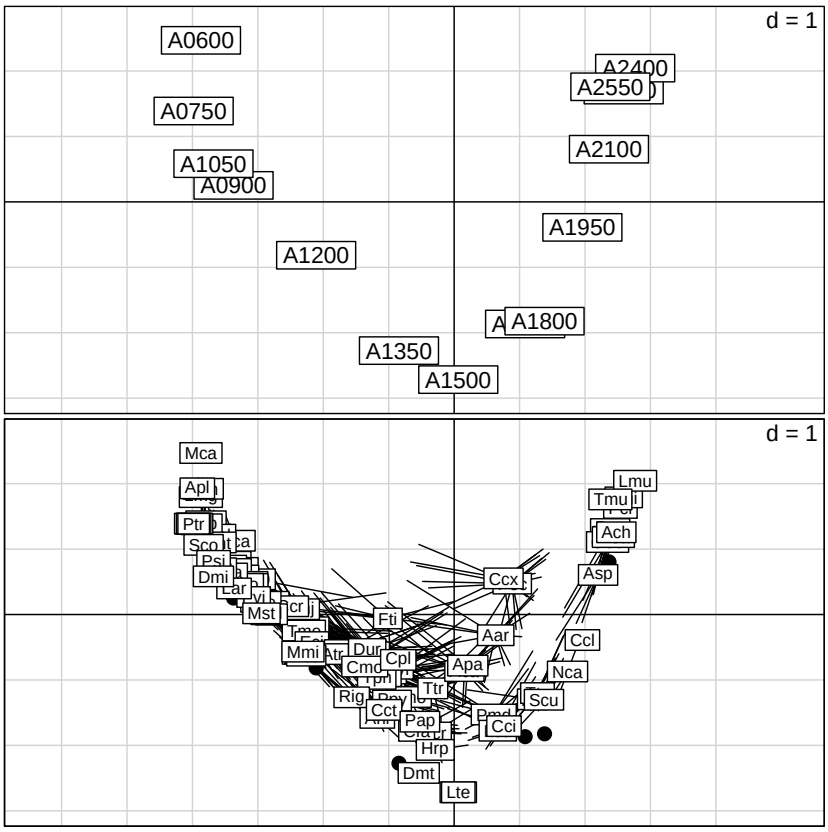
```
X <- as.data.frame(t(profalti))
w <- dudi.dec(X, ref, scan = F)
sum(apply(X, 2, function(x) chisq.test(x, p = ref/sum(ref), sim = T,
    B = 1)$statistic))
```

```
[1] 4710.321
```

```
sum(w$eig) * sum(ref)
```

```
[1] 4710.321
```

```
par(mfrow = c(2, 1))
s.label(w$li)
s.label(w$li, clab = 0, cpoi = 2)
s.distri(w$li, X, cell = 0, csta = 0.3, add.p = T, clab = 0.75,
    cpoi = 0)
```



Les strates d'altitude sont positionnée de telle manière qu'une espèce également présente dans chaque strate soit à l'origine (vérifier que la moyenne est nulle pour la distribution des points d'échantillonnage). Les espèces sont à la moyenne de leur propre distribution et forme un gradient continu. On peut voir dans le deuxième axe une réplique numérique du premier (*arch effect*). Voir aussi la focnction `eisera` dans la librairie `adehabitat` qui utilise le même concept en biologie des populations.

Considérons maintenant les occurrences-individus.

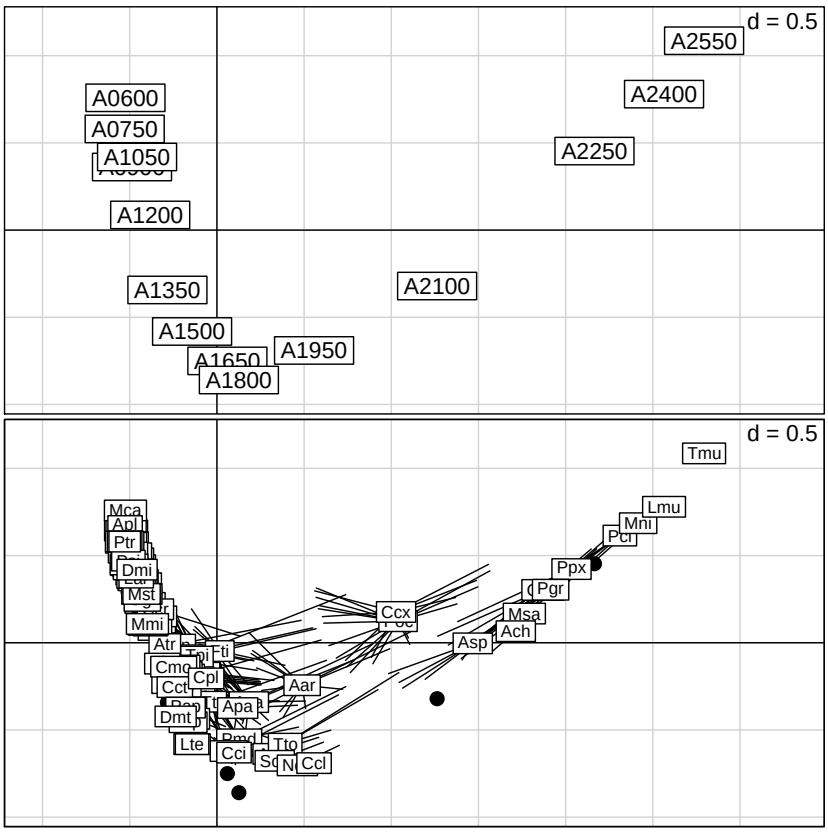
```
sum(profalti)
```

```
[1] 3444
```

On peut au contraire penser l'occurrence d'une espèce à un endroit donné comme un fait élémentaire. Il y a dans ce tableau 3444 occurrences, contact d'une espèce donnée à une altitude donnée. Ce que l'avifaune dit de l'altitude tient au lien

qui existe entre le nom de l'espèce et la classe d'altitude qui ici seule nous occupe. L'ornithologue dit pour exprimer ceci "donnez moi une liste d'espèce et je vous dirai où vous étiez et quand vous y étiez". L'analyse de ce lien se fait par l'analyse des correspondances du tableau qui est ici réellement une analyse canonique. Cette approche est très précise :

```
ww <- dudi.coa(X, scan = F)
par(mfrow = c(2, 1))
s.label(ww$li)
s.label(ww$li, clab = 0, cpoi = 2)
s.distri(ww$li, X, cell = 0, csta = 0.3, add.p = T, clab = 0.75,
    cpoi = 0)
```



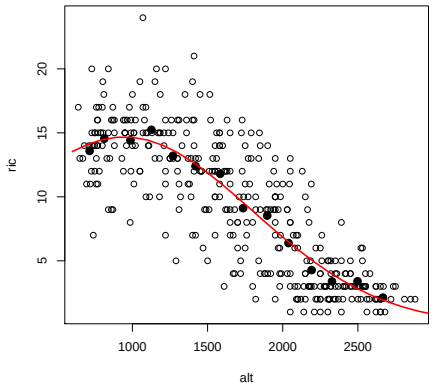
Les strates d'altitude sont positionnée de telle manière qu'en moyenne l'ensemble des occurrences (toute espèces confondues) soit à l'origine. Vérifier que la moyenne est nulle pour la distribution de la richesse des points d'échantillonnage :

```
ric <- apply(tarentaise$ecol, 1, sum)
sum(ww$li[, 1] * tapply(ric, tarentaise$envir$alti, sum))
```

La distortion par l'appauvrissement de la richesse en altitude est très sensible. L'analyse elle-même de la richesse est un problème classique de l'écologie statistique

```
alt <- tarentaise$alti
plot(alt, ric)
```

```
altmoy <- tapply(alt, tarentaise$envir$alti, mean)
ricmoy <- tapply(ric, tarentaise$envir$alti, mean)
points(altmoy, ricmoy, pch = 20, cex = 2)
glm1 <- glm(ric ~ alt + I(alt^2), family = poisson)
altut <- seq(600, 3000, by = 100)
lines(altut, predict(glm1, new = list(alt = altut), type = "response"),
    lwd = 2, col = "red")
```



Un modèle vrai pour un vrai gradient ou une vue de l'esprit de l'écologue ?

4 Ordination sur composantes

Le problème devient sérieux avec p variables et q espèces. C'est un des premiers problèmes qui a justifié le terme d'écologie statistique. Dans la rotonde du CEPE (Centre d'Etudes PhytoSociologiques et Ecologiques) devenu CEFE [1] (Centre d'Ecologie Fonctionelle et Evolutive) :

on a longtemps conservé des meubles remplis des cartes perforées de milliers de relevés de la végétation autour du bassin méditerranéen. C'était le monde de l'écothèque méditerranéenne. L'analyse des correspondances des tableaux de profils écologiques a été introduite [28] comme pratique *ad hoc* puis redécouverte [24] puis identifiée comme canonique [23]. Il s'agit cependant de stratégie de coinertie [12][14][16].

L'ordination sous contrainte, où le tableau de milieu prédit le tableau florofaunistique, est la méthode la plus populaire sous le nom de *Canonical Correspondance Analysis* ou simplement *CCA* [30][31] diffusé par le biais du logiciel

[1] Voir Du CEPE au CEFE... ou une brève histoire de la phytosociologie, entretien avec J. Blondel à `http://www.educ-envir.org/~euziere/science/article.php3?id_article=115`

commercial CANOCO [32] et disponible dans Rdans la librairie `vegan` de J. Oksanen, auteur d'une célèbre critique [25] contre le précédent. `ade4` en contient une version du type *AFCVI*, pour Analyse Factorielle des Correspondances sur Variables Instrumentales, autre point de vue sur une méthode de théorie complexe [11][22] (ce qui n'intéresse évidemment pas l'utilisateur).

Une idée plus simple et souvent efficace est de remplacer les variables de milieu par leurs composantes principales. Ceci est souvent efficace parce que les variables d'environnement contiennent une bonne part de redondance due à un enregistrement aussi exhaustif qu'il est possible.

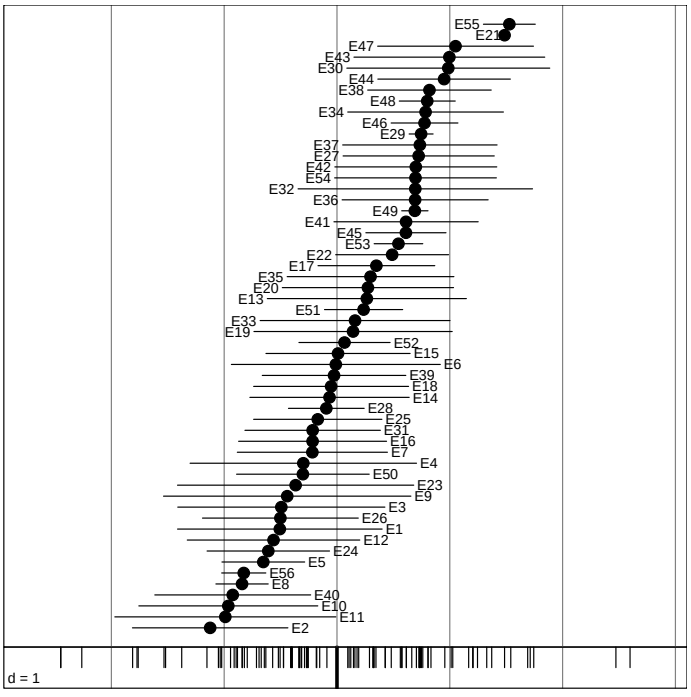
Donnons quelques exemples. On trouve une enquête phytososiologique à vocation d'aménagement [13] dans :

`http://pbil.univ-lyon1.fr/R/pps/pps053.pdf`

```
data(mafragh)
X <- mafragh$mil
names(X) <- c("Arg", "Lim", "Sab", "K2O", "Mg", "Na1", "K", "Cond",
    "Capa", "Na2", "Alt")
round(100 * cor(X), 0)
```

```
     Arg Lim Sab K2O  Mg Na1   K Cond Capa Na2 Alt
Arg  100 -63 -72  44  19  29  54   33   44  26 -15
Lim  -63 100  -8 -24  -2   3 -32    0    1   2   8
Sab  -72  -8 100 -34 -22 -38 -41  -40  -56 -34   9
K2O   44 -24 -34 100  34  25  57   26   23  24 -18
Mg    19  -2 -22  34 100  41  41   34   13  36 -22
Na1   29   3 -38  25  41 100  57   74   31  71 -33
K     54 -32 -41  57  41  57 100   42   28  37 -33
Cond  33   0 -40  26  34  74  42  100   41  96 -41
Capa  44   1 -56  23  13  31  28   41  100  29 -25
Na2   26   2 -34  24  36  71  37   96   29 100 -40
Alt  -15   8   9 -18 -22 -33 -33  -41  -25 -40 100
```

```
Xpca <- dudi.pca(X, scal = T, scan = F)
sco.distri(Xpca$l1[, 1], mafragh$flo, clab = 0.75)
```



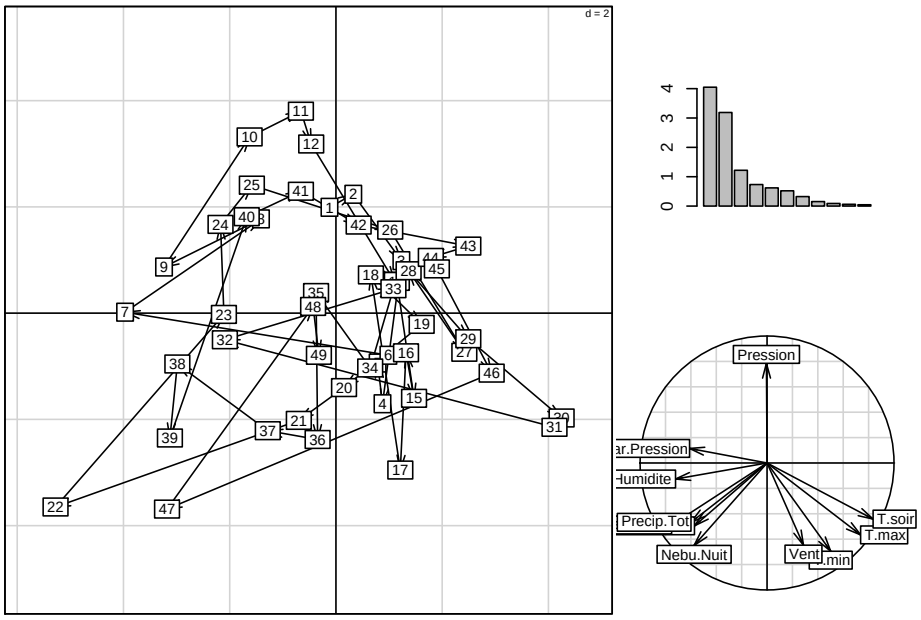
Vérifier que 41% de la variabilité de toutes les variables est prise en compte dans la première variable de synthèse. On obtient bien une ordination sur cette variable sans savoir cependant si elle est optimale, voire même valide. Deux conceptions sont en jeu. Soit un relevé est un échantillon d'une association (comme le *Scirpetum maritimi*) soit un relevé est une collection d'espèces indépendantes qui sont là pour des raisons écologiques et historiques. La figure conforte le deuxième point de vue mais ne le justifie pas.

L'expérience [33] rapportée dans :

`http://pbil.univ-lyon1.fr/R/pps/pps034.pdf`

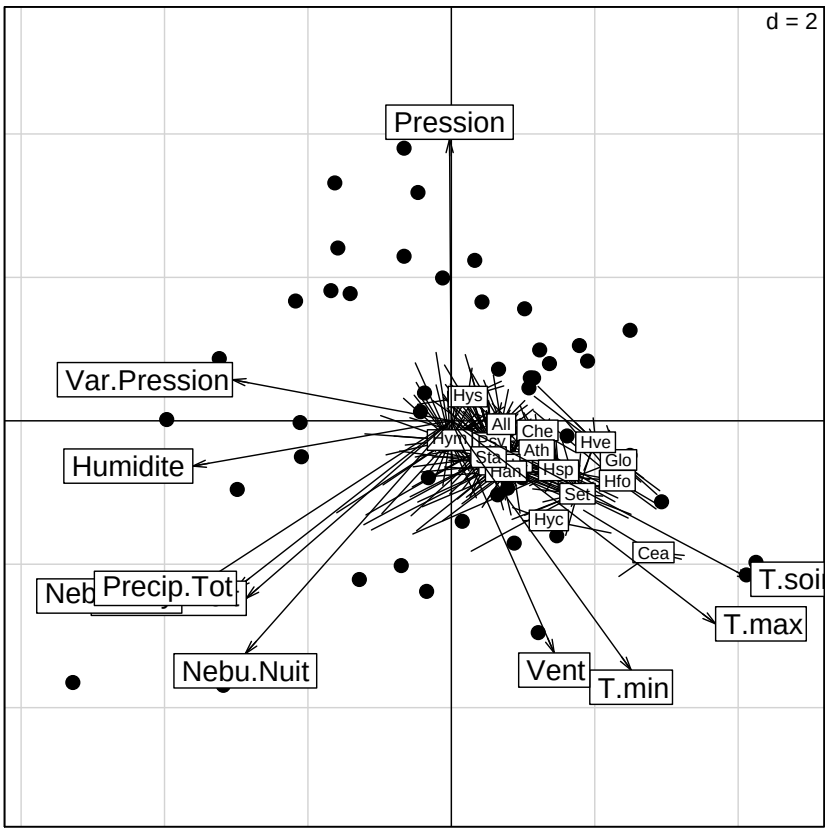
est plus claire.

```
data(trichometeo)
X <- trichometeo$meteo
Xpca <- dudi.pca(X, scal = T, scan = F)
layout(matrix(c(1, 1, 1, 1, 2, 3), nrow = 2))
s.traject(Xpca$li, clab = 0)
s.label(Xpca$li, add.p = T, clab = 1.25)
barplot(Xpca$eig)
s.corcircle(Xpca$co, clab = 1.25)
```



Sur le plan, observer le mouvement de rotations dans les suites de nuits successives. Beau, chaud, orageux et on recommence. Le cercle de corrélation est explicite. Le plan contient toute l'information des variables de milieu. Le but de l'expérience s'exprime d'un coup dans :

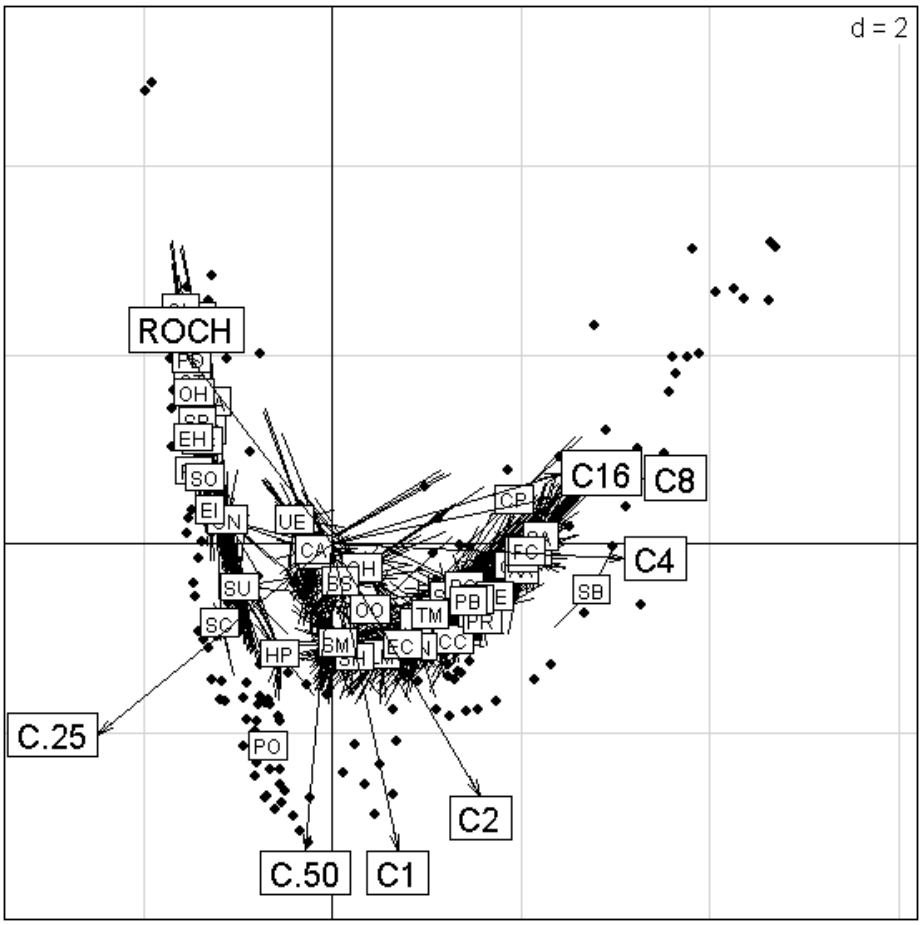
```
Y <- as.data.frame(log(trichometeo$fau + 1))
s.label(Xpca$li, clab = 0, cpoi = 2)
s.distri(Xpca$li, Y, clab = 0.75, cell = 0, cstar = 0.3, add.p = T)
s.arrow(5 * Xpca$co, clab = 1.25, add.p = T)
```



L'ornithologie a été une des premières disciplines à suivre la phyto-écologie dans la voie de l'ordination. Suivront tous les autres domaines, en particulier ceux de l'hydrobiologie. Les données [27] sont décrites dans :

`http://pbil.univ-lyon1.fr/R/pps/pps048.pdf`

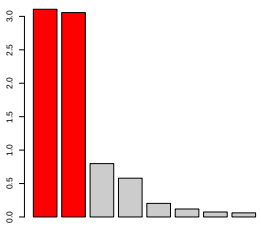
```
data(rpjdl)
X <- rpjdl$mil
Y <- rpjdl$fau
Xpca <- dudi.pca(X, scal = T, scan = F)
s.label(Xpca$li, clab = 0, cpoi = 1.5)
s.distri(Xpca$li, Y, clab = 0.75, cell = 0, cstar = 0.3, add.p = T)
s.arrow(3.5 * Xpca$co, clab = 1.25, add.p = T)
```



Il y a dans cette figure très peu d'intention *a priori* et beaucoup d'informations *a posteriori*. L'art de traiter les données consiste souvent à se *laisser faire* par les résultats. On voit ici que le jeu de données mélange deux mondes, celui des milieux ouverts, à gauche et celui des milieux forestiers à droite. Il y a articulation entre les deux par la strate buissonnante et les espèces ubiquistes. Mais l'essentiel est une partition.

En fait, il y a ici, un fait redoutable. Les deux premières valeurs propres sont très voisines :

```
barplot(Xpca$eig, col = c(rep("red", 2), rep(grey(0.8), 6)))
```



Le plan existe, il définit la structure. Les axes sont sans importance, ils portent le plan et tout couple de vecteurs orthogonaux dans ce plan ferait l'affaire. Les gradients n'existent pas ! Mieux vaut ne pas les interpréter.

5 Ordination sous contrainte

Quelques principes sont indiqués sur un exemple. Ordonner les espèces sur un gradient, soit par modélisation de courbes de réponse, soit par *averaging* reste aisé. Quand la variable est qualitative, le passage aux profils écologiques s'impose. Quand la mesure du gradient est multivariée la réduction de dimensions ramène aux cas précédents. Mais la question est toujours de savoir si les composantes principales sont les meilleures variables d'ordination.

On peut concevoir, en effet, qu'une combinaison de variables peut beaucoup varier sans beaucoup ordonner alors qu'une autre, moins variable, pourrait mieux discriminer les niches. On peut même demander à une variable de ne s'occuper que de séparer les moyennes par espèces sous la seule contrainte d'être de variance unité. On atteint là des difficultés statistiques sans commune mesure avec les compétences de l'utilisateur de base qui va se soumettre aux aléas des comités de lecture, de la mode ou des logiciels disponibles.

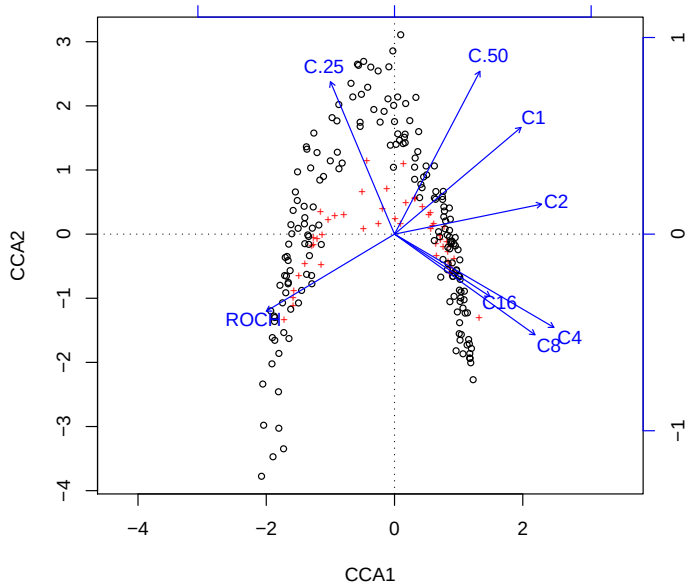
La CCA est disponible par les fonctions `cca` d'`ade4` et de `vegan`. Il faut pour comparer les deux éviter les conflits :

```
library(vegan)
```

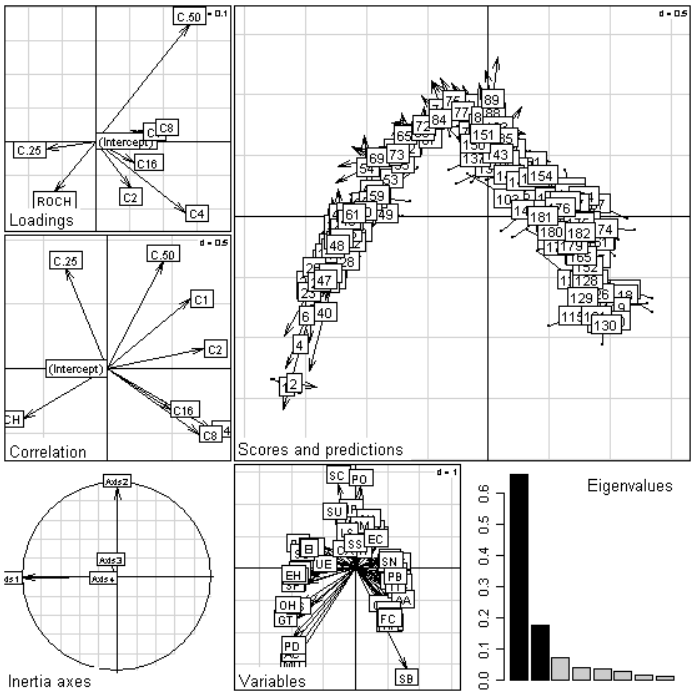
```
Attachement du package : 'vegan'
        The following object(s) are masked from package:ade4 :

         cca
```

```
ccavegan <- cca(Y, X)
plot(ccavegan)
ccavegan
```

```
Call:
cca(X = Y, Y = X)
              Inertia Rank
Total           4.468
Constrained     1.045    8
Unconstrained   3.423   50
Inertia is mean squared contingency coefficient

Eigenvalues for constrained axes:
   CCA1    CCA2    CCA3    CCA4    CCA5    CCA6    CCA7    CCA8
0.66170 0.17729 0.07054 0.03891 0.03616 0.02865 0.01793 0.01395

Eigenvalues for unconstrained axes:
    CA1    CA2    CA3    CA4    CA5    CA6    CA7    CA8
0.2237 0.1779 0.1641 0.1539 0.1489 0.1416 0.1323 0.1222
(Showed only 8 of all 50 unconstrained eigenvalues)
```



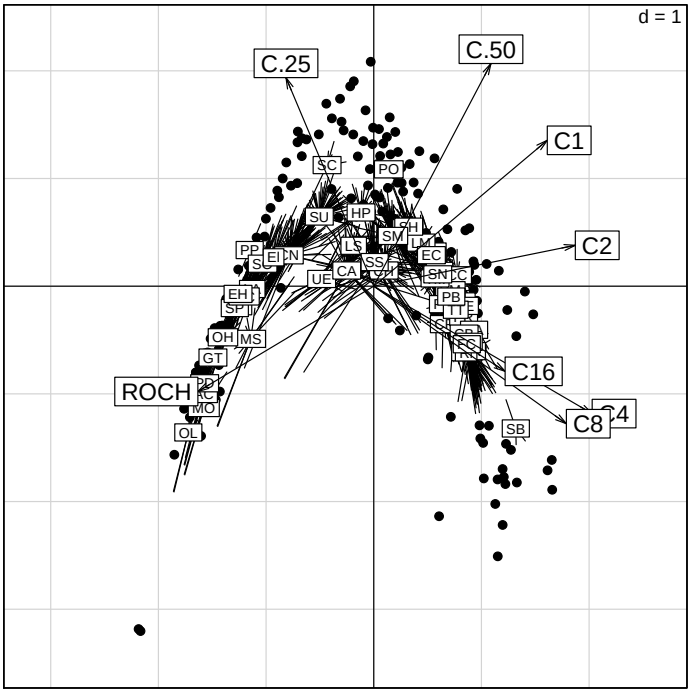

```
coarpjdl <- dudi.coa(Y, scan = F, nf = 4)
ccaade4 <- pcaiv(coarpjdl, X, scan = F)
plot(ccaade4)
```



On pourrait croire que ça n'a pas grand-chose à voir. Il faut quelques explications. Dans `vegan` la *CCA* est une méthode indépendante. Dans `ade4`, c'est un

cas particulier d'*ACPVI*. Mais les deux sont très voisines. Pour refaire le plot de `vegan` :

```
s.label(ccaade4$l1, clab = 0, cpoi = 1.5)
s.distri(ccaade4$l1, Y, clab = 0.75, cell = 0, cstar = 0.3, add.p = T)
s.arrow(2.5 * ccaade4$cor[-1, ], clab = 1.25, add.p = T)
```



Il n'y a pas identité pour des détails de conception qui importent peu mais l'intention est la même et l'interprétation aussi. Les deux figures expriment un même principe et s'appelle des triplots (représentation simultanée des relevés, des espèces et des variables). On pourrait penser que l'ACP du tableau de milieu fournissait directement ce résultat. Ce n'est pas vrai. Il y a deux modifications très importantes. La première touche les pondérations.

En ACP chaque relevé a le même poids et, quand on parle de moyenne, variance, covariance, corrélation il s'agit de la notion ordinaire (en $\frac{1}{n}$ ou $\frac{1}{n-1}$, suivant les logiciels).

En ACC (*CCA*), qui dérive de l'AFC, le tableau faunistique (`rpjdl$fau`) définit le poids des sites. Un site a une importance proportionnelle à l'abondance des espèces qui y sont. On est dans le point de vue : l'environnement est celui des occurrences. Un site avec 50 espèces compte 10 fois plus qu'un site à 5 espèces parce qu'il décrit l'environnement de 10 fois plus de taxa. Comparer `w$lw[1]` et `sum(rpjdl$fau[1,])/sum(rpjdl$fau)`. Ici un site contient entre 3 et 20 espèces.

La seconde touche à l'objectif. Bien distinguer :

ACP-PCA Trouver une composante principale (score des sites) qui maximise la somme de ses carrés de corrélation avec les variables. Peu importe sa variance (on ne change pas une corrélation en changeant les unités), on la prend unitaire par convention.

AFC-COA Trouver une composante (score des sites) qui maximise la variance des moyennes par espèce. Évidemment si on multiplie par 2 le score, les moyennes sont multipliées par 2 et la quantité recherchée par 4. On impose la variance unité, on impose la pondération marginale.

ACC-CCA Trouver une composante (score des sites) qui maximise la variance des moyennes par espèce, comme en AFC, avec les mêmes contraintes mais en imposant en outre d'être une combinaison linéaire des variables de milieu. C'est pourquoi on dit ordination sous contrainte ou AFC sur Variables Instrumentales.

On ne peut pas comparer les statistiques de l'ACP avec celle de l'ACC parce qu'on a changé de pondération. On peut comparer, par contre celle de l'ACP et de l'ACC. Mais in faut rajouter une aide à l'interprétation.

L'AFC est par essence symétrique. Quand on ordonne les espèces par les sites, on ordonne aussi les sites par les espèces. Quand on introduit un tableau, on casse relativement la symétrie mais pas tout à fait. Bien distinguer :

AFC-COA Trouver un axe (score des espèces) qui maximise la variance des moyennes par sites. Évidemment si on multiplie par 2 le score, les moyennes sont multipliées par 2 et la quantité recherchée par 4. On impose la variance unité, on impose la pondération marginale (sur la marge espèces).

AFC-COA Trouver un axe (score des espèces) de variance unité, pour la pondération marginale (comme ci-dessus) qui donne un score par moyenne pour les sites (comme-ci-dessus) qui a une variance (comme ci-dessus) **EXPLIQUÉE** maximale. Ceci veut dire que le critère optimisé est un compromis entre la variance (qui ne peut dépasser 1 par contrainte) et le pourcentage de variance expliquée (le R^2 de la régression multiple du score sur les variables de milieu). Cette régression est pondérée par les poids des sites. Le critère est

$$Variance_{explique} = Variance \times R^2$$

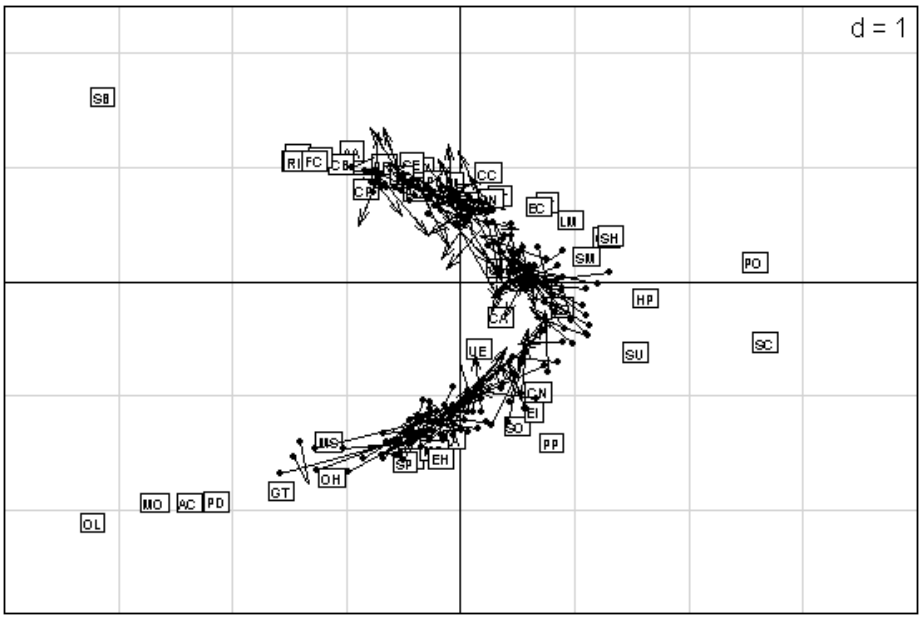
On trouve ça dans :

```
ccaade4$param
```

```
iner  inercum inerC inercumC ratio R2    lambda
0.753 0.753   0.747 0.747    0.992 0.885 0.662
0.293 1.05    0.275 1.02     0.977 0.645 0.177
```

Le score 1 de l'AFC fait une variance de 0.753 (variance des moyennes par sites). Le score 1 de la CCA ne peut pas faire mieux mais fait une variance de 0.747, soit 99.2 % de l'optimum. Le score de l'ACC est prédictible à 88.5% par les variables environnementales. La valeur propre est le produit de la variance (0.747) par le R^2 (0.885) : c'est la valeur propre de rang 1 de cette analyse.

```
s.label(ccaade4$c1, 2, 1, clab = 0.5, xlim = c(-4, 4))
s.label(ccaade4$ls, 2, 1, clab = 0, cpoi = 1, add.p = TRUE)
s.match(ccaade4$ls, ccaade4$li, 2, 1, clab = 0, add.p = T)
```



On renforce ainsi l'idée d'un gradient et de l'arch effect qui l'accompagne. On voit que l'AFC et sa dérivée, la CCA, renforce l'idée du continuum [21] au détriment de celle de classification pour laquelle elle peut servir de moyen d'expression [29]. Il ne saurait y avoir de méthodes statistiques totalement neutres.

6 Ordination planifiée

Ces quelques indications sont extraites d'un débat qui dure depuis un demi-siècle. Ce débat a été dominé, entre autres, par I. Noy-Meir ou encore M.P. Austin (1966[6] 1971[5] 1980[3] 1986[4]) qui y reconnaît un problème d'interface [2]. Ce débat n'est pas clos.

On peut noter qu'en général on confond au sein des covariables (variables environnementales) celles qui dérivent des intentions (plan d'observations) et celles qui dérivent des mesures (variables observées). Où et quand, par exemple, sont définies par l'auteur, le temps qu'il fait, le débit de la rivière ou la profondeur du sol sont des observations.

L'ordination sous contrainte est bien adaptée aux covariables pilotées et permet de répondre aux questions

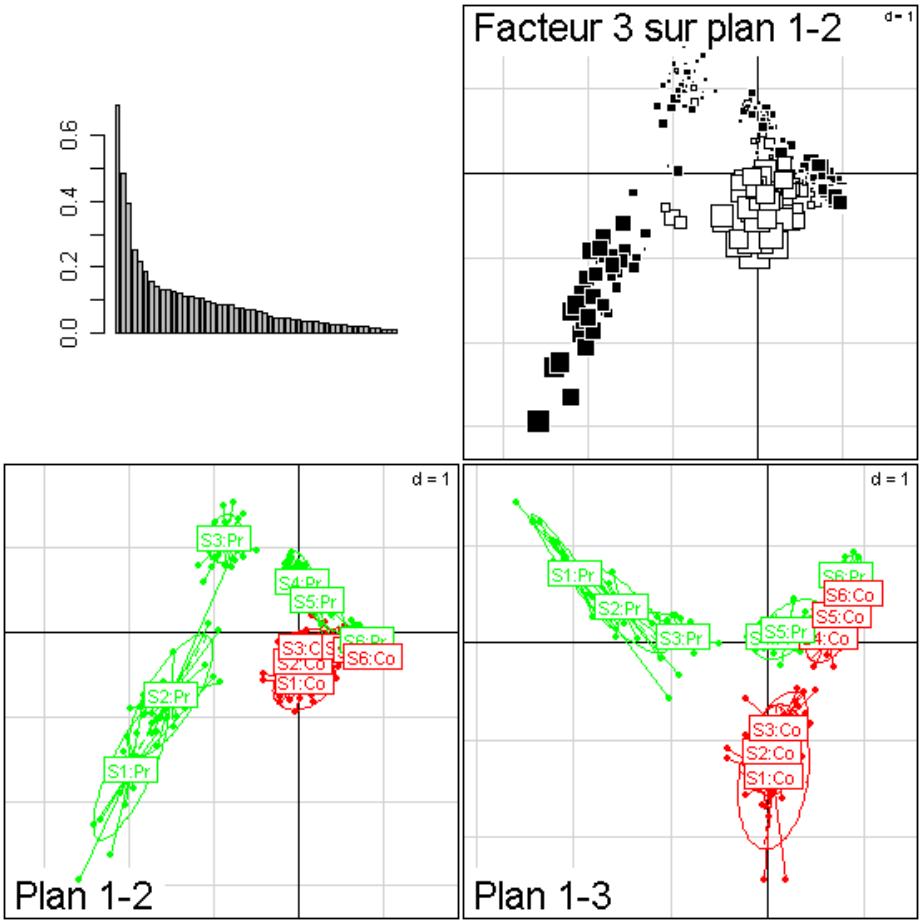
1. Quelle part de structure dépend du pilotage des observations ?
2. Quelle part de structure n'en dépend pas ?

On utilise les données décrites dans :

`http://pbil.univ-lyon1.fr/R/pps/pps052.pdf`

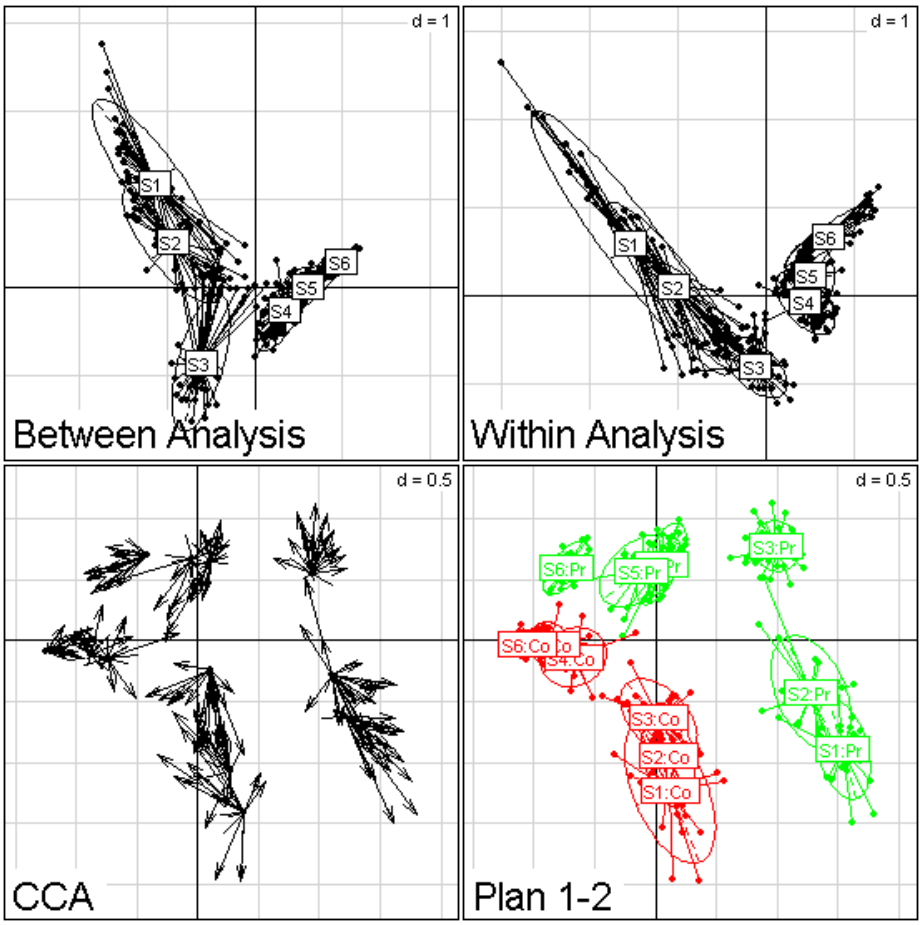
```
data(avimedi)
fac <- avimedi$plan$str:avimedi$plan$reg
colfac <- rep(c("green", "red"), 6)
par(mfrow = c(2, 2))
coa1 <- dudi.coa(avimedi$fau, scan = FALSE, nf = 3)
barplot(coa1$eig)
s.value(coa1$li, coa1$li[, 3], cleg = 0, sub = "Facteur 3 sur plan 1-2",
```

```
    csub = 2)
s.class(coa1$li, fac, col = colfac, sub = "Plan 1-2", csub = 2)
s.class(coa1$li, fac, xax = 1, yax = 3, col = colfac, sub = "Plan 1-3",
    csub = 2)
par(mfrow = c(1, 1))
```



La structure est une fleur étrange à deux branches (deux régions) et une ordination (6 types de végétation ou strates). On peut la voir de manière précise. Voir uniquement les différences entre strates, cacher les différences entre régions, mettre en évidence un modèle additif :

```
par(mfrow = c(2, 2))
bet1 <- between(coa1, avimedi$plan$str, scan = FALSE)
s.class(bet1$ls, avimedi$plan$str, sub = "Between Analysis", csub = 2)
wit1 <- within(coa1, avimedi$plan$reg, scan = FALSE)
s.class(wit1$li, avimedi$plan$str, sub = "Within Analysis", csub = 2)
pcaiv1 <- pcaiv(coa1, avimedi$plan, scan = FALSE)
s.match(pcaiv1$li, pcaiv1$ls, clab = 0, sub = "CCA", csub = 2)
s.class(pcaiv1$ls, fac, col = colfac, sub = "Plan 1-2", csub = 2)
par(mfrow = c(1, 1))
```



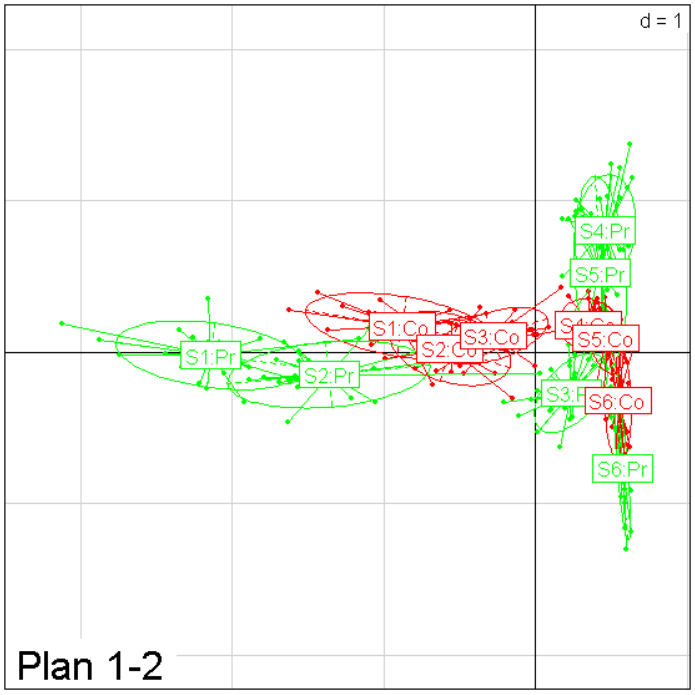
On trouve des contraintes positives dans `pcaiv` (`between` est la cas particulier d'un seul facteur), des contraintes négatives dans `pcaivortho` (`within` est la cas particulier d'un seul facteur). On peut insérer les deux dans la `cca` de `vegan`. La structure d'ade4 est plus ouverte pour le choix d'un sous-espace de projection. Par exemple, on a cherché des contraintes du type $A \cap B^{\perp}$ (A inter B orthogonal). Ce qui dépend deux facteurs A sans dépendre du facteur B, contrainte encore plus forte que le B sachant A du modèle linéaire ordinaire (la part de B dans $A+B$). Cet espace connu des mathématiciens [1] apparaît en statistique dans la méthode LONGI [26] et fournit la solution du problème posé par la séparation entre composante environnementale et composante spatiale [8]. On rajoute une petite fonction :

```
disj <- function(a) {
    if (!is.factor(a))
        stop("factor expected")
    n <- length(a)
    m <- nlevels(a)
    adi <- matrix(0, n, m)
    for (i in 1:n) adi[i, a[i]] <- 1
    return(adi)
}
ainterbortho <- function(a, b) {
    if (!is.factor(a))
        stop("a: factor expected")
    if (!is.factor(b))
```

```
        stop("b: factor expected")
    if (length(a) != length(b))
        stop("length non matched")
    A <- disj(a)
    B <- disj(b)
    qrA <- qr(A)
    qrB <- qr(B)
    if (qrA$rank == 0)
        return(list(dim = 0, base = NULL))
    if (qrB$rank == length(a))
        return(list(dim = 0, base = NULL))
    A0 <- qr.Q(qrA, complet = F)[, 1:qrA$rank]
    B0 <- qr.Q(qrB, complet = T)[, ((qrB$rank + 1):(length(a)))]
    C0 <- t(A0) %*% B0
    eig0 <- eigen(t(C0) %*% C0, sym = T)
    dim <- sum(((1 - eig0$values)^2) < 1e-07)
    base <- B0 %*% eig0$vectors[, 1:dim]
    return(as.data.frame(base))
}
```

On peut alors voir une illustration de la biogéographie insulaire [35] :

```
w <- ainterbortho(avimedi$plan$str, avimedi$plan$reg)
pcaiv3 <- pcaiv(coa1, w, scan = F)
s.class(pcaiv3$ls, fac, col = colfac, sub = "Plan 1-2", csub = 2)
```



La Corse se place à l'intérieur de la Provence, image de la différence entre les structures insulaire et continentale. La compétition limitée par un nombre d'espèces plus faible, le chevauchement des niches plus grand et d'autres hypothèses sont en jeu [7].

En guise de conclusion, on peut penser que le débat de l'ordination écologique a fourni des éléments théoriques intéressant d'autres domaines manipulant des données abondantes et fortement multidimensionnelles.

Références

[1] S.N. Afriat. Orthogonal and oblique projectors and the characteristics of pairs of vector spaces. *Proceedings of the Cambridge Philosophical Society, Mathematical and Physical Sciences*, 53 :800–816., 1957.

[2] M. P. Austin. Spatial prediction of species distribution : an interface between ecological theory and statistical modelling. *Ecological Modelling*, 157(2-3) :101–118, 2002.

[3] M.P. Austin. Searching for a model for use in vegetation analysis. *Vegetatio*, 42 :11–21, 1980.

[4] M.P. Austin. The theoretical basis of vegetation science. *Trends in Ecology and Evolution*, 1 :161–164, 1986.

[5] M.P. Austin and I. Noy-Meir. The problem of non-linearity in ordination : experiments with two-gradient models. *Journal of Ecology*, 59 :763–773, 1971.

[6] M.P. Austin and L. Orlóci. Geometric models in ecology ii an evaluation of some ordination techniques. *Journal of Ecology*, 54 :217–227, 1966.

[7] J. Blondel, D. Chessel, and B. Frochot. Bird species impoverishment, niche expansion, and density inflation in mediterranean island habitats. *Ecology*, 69 :1899–1917, 1988.

[8] D. Borcard, P. Legendre, and P. Drapeau. Partialling out the spatial component of ecological variation. *Ecology*, 73 :1045–1055, 1992.

[9] J. Braun-Blanquet. *Plant sociology, the study of plant communities*. Mc Graw Hill Book, New York, 1932.

[10] C. Calenge and A-B. Dufour. Eigenanalysis of selection ratios from animal radio-tracking data. *Ecology*, 87 :2349–2355, 2006.

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

[12] D. Chessel and P. Mercier. Couplage de triplets statistiques et liaisons espèces-environnement. In J.D. Lebreton and B. Asselain, editors, *Biométrie et Environnement*, pages 15–44. Masson, Paris, 1993.

[13] G. de Belair and M. Bencheikh-Lehocine. Composition et déterminisme de la végétation d'une plaine côtière marécageuse : La mafragh (annaba, algérie). *Bulletin d'Ecologie*, 18 :393–407, 1987.

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

[15] S. Dolédec, D. Chessel, and J.M. Olivier. L'analyse des correspondances décentrée : application aux peuplements ichtyologiques du haut-rhône. *Bulletin Français de la Pêche et de la Pisciculture*, 336 :29–40, 1995.

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

[17] H.G. Jr. Gauch. *Multivariate analysis in community ecology*. Cambridge University Press, Cambridge, 1982.

[18] H.G. Jr Gauch, G.B. Chase, and R.H. Whittaker. Ordination of vegetation samples by gaussian species distributions. *Ecology*, 55 :1382–1390, 1974.

[19] M. Gounot. Contribution à l'étude des groupements végétaux messicoles et rudéraux de la tunisie. *Annales du Service botanique et agronomique de la Direction générale de l'agriculture. Tunisie*, 31 :1–282, 1958.

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

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

[22] 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.

[23] 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.

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

[25] Jari Oksanen and Peter R. Minchin. Instability of ordination results under changes in input data order : explanations and remedies. *Journal of Vegetation Science*, 8 :447–454, 1997.

[26] J. Pontier, A.B. Dufour, and M. Normand. *Le modèle euclidien en analyse des données*. SMA, édition Ellipses, Bruxelles, 1990.

[27] R. Prodon and J.D. Lebreton. Breeding avifauna of a mediterranean succession : the holm oak and cork oak series in the eastern pyrénées. 1 : Analysis and modelling of the structure gradient. *Oïkos*, 37 :21–38, 1981.

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

[29] G. Roux and M. Roux. A propos de quelques méthodes de classification en phytosociologie. *Revue de Statistique Appliquée*, XV :59–72, 1967.

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

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

[32] C.J.F. Ter Braak. Canoco - a fortran program for canonical commnity ordination by [partial][detrended][canonical] correspondence analysis and redundancy analysis. Technical report, Version 2.1, TNO Institute of Applied Computer Science, Wageningen, 1987. Software documentation.

[33] P. Usseglio-Polatera and Y. Auda. Influence des facteurs météorologiques sur les résultats de piégeage lumineux. *Annales de Limnologie*, 23 :65–79, 1987.

[34] R.H. Whittaker. *Handbook of vegetation science. Part V. Ordination and classification of communities*. Dr. W. Junk b.v., The Hague, 1973.

[35] N. Yoccoz and D. Chessel. Ordination sous contraintes de relevés d'avifaune : élimination d'effets dans un plan d'observations à deux facteurs. *Compte rendu hebdomadaire des séances de l'Académie des sciences. Paris, D*, III :307 : 189–194, 1988.