

Fiche de Biostatistique – Stage 3

Schémas de dualité

D. Chessel, A.B. Dufour & J. Thioulouse

Résumé

La fiche décrit le principe général d'un schéma de dualité et passe en revue quelques usages de l'analyse en composantes principales.

Plan

1	SCHEMA DE DUALITE.....	2
1.1	Rang d'un schéma.....	3
1.2	Valeurs propres d'un schéma.....	3
1.3	Vecteurs principaux d'un schéma.....	4
1.4	Propriétés d'optimalité.....	6
1.5	Approximations de tableau.....	6
2	GRAPHES CANONIQUES.....	8
2.1	Procédure centrale.....	8
2.2	Décentrage.....	12
2.3	Redondance.....	18
2.4	Valeurs propres.....	22
2.5	Profils et biplot.....	26
2.6	Tableaux homogènes.....	32
3	MODELES PROBABILISTES.....	35
3.1	Loi normale bivariée.....	35
3.2	Erreur sphérique.....	37
4	LA CLASSE D'OBJETS <small>DUDI</small>	40
5	ANALYSES DES CORRESPONDANCES MULTIPLES.....	43
5.1	Définition.....	43
5.2	Ours.....	46
5.3	Compléments.....	49
5.4	ACM pondérée.....	50
5.5	Extension.....	51

1 Schéma de dualité

On connaît la pratique de l'ACP centrée ou normée décrite dans le chapitre "Géométrie de l'espace des variables". Il s'agit d'un cas particulier d'une classe d'objets qu'on appelle schémas de dualité. On en donne ici une version simplifiée sans insister sur les démonstrations pour en faciliter l'usage. L'idée d'origine est de P. Cazes, l'essentiel théorique est dans une langue vernaculaire (Cailliez and Pages 1976, Pagès et al. 1979, Cailliez 1984), il est appelé analyse générale par Greenacre (1984, Annexe A2 p. 346). Il est introduit en écologie par Escoufier (1987). Un schéma est constitué de trois éléments donnant un triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$.

\mathbf{X} est une matrice de données en général issue d'une matrice de données brutes \mathbf{Y} à l'aide d'une transformation préalable. \mathbf{X} a n lignes et p colonnes. Les n lignes de \mathbf{X} sont des vecteurs de \mathbb{R}^p tandis que les p colonnes de \mathbf{X} sont des vecteurs de \mathbb{R}^n . \mathbf{X}' est la transposée de la matrice \mathbf{X} , elle a p lignes et n colonnes.

\mathbf{Q} est la matrice d'un produit scalaire de \mathbb{R}^p , soit une matrice carrée symétrique qui définit la fonction :

$$(\mathbf{x}, \mathbf{y}) = \left(\begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}, \begin{bmatrix} y_1 \\ \vdots \\ y_p \end{bmatrix} \right) \in \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbf{x}'\mathbf{Q}\mathbf{y} = \langle \mathbf{x} | \mathbf{y} \rangle_{\mathbf{Q}} \in \mathbb{R}$$

\mathbf{D} est la matrice d'un produit scalaire de \mathbb{R}^n , soit une matrice carrée symétrique qui définit la fonction :

$$(\mathbf{x}, \mathbf{y}) = \left(\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \right) \in \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbf{x}'\mathbf{D}\mathbf{y} = \langle \mathbf{x} | \mathbf{y} \rangle_{\mathbf{D}} \in \mathbb{R}$$

L'essentiel est que ces quatre matrices s'assemblent car les produits :

$$\mathbf{X}\mathbf{Q}, \mathbf{Q}\mathbf{X}', \mathbf{X}'\mathbf{D} \text{ et } \mathbf{D}\mathbf{X}$$

ont un sens. On les dispose dans un schéma dit schéma de dualité :

$$\begin{array}{ccc} & \mathbf{Q} & \\ \mathbb{R}^p & \rightarrow & \mathbb{R}^{p*} \\ \mathbf{X}' \uparrow & & \downarrow \mathbf{X} \\ \mathbb{R}^{n*} & \leftarrow & \mathbb{R}^n \\ & \mathbf{D} & \end{array}$$

Rigoureusement, \mathbb{R}^{p*} est le dual de \mathbb{R}^p (ensemble des applications linéaires de \mathbb{R}^p dans \mathbb{R}), \mathbb{R}^{n*} est le dual de \mathbb{R}^n (ensemble des applications linéaires de \mathbb{R}^n dans \mathbb{R}), \mathbf{Q} est vue comme la matrice d'une application linéaire définie par $(\mathbf{Q}(\mathbf{x}))(\mathbf{y}) = \langle \mathbf{x} | \mathbf{y} \rangle_{\mathbf{Q}}$,

\mathbf{D} est vue comme la matrice d'une application linéaire définie par $(\mathbf{D}(\mathbf{x}))(\mathbf{y}) = \langle \mathbf{x} | \mathbf{y} \rangle_{\mathbf{D}}$ et \mathbf{X}^t est vue comme la matrice d'une application linéaire définie à l'aide du bidual. Pour l'utilisateur, la simplification :

$$\begin{array}{ccc}
 & \mathbf{Q} & \\
 & \boxed{p} \rightarrow \boxed{p} & \\
 (\mathbf{X}, \mathbf{Q}, \mathbf{D}) \Leftrightarrow & \mathbf{X}^t \uparrow & \downarrow \mathbf{X} \\
 & \boxed{n} \leftarrow \boxed{n} & \\
 & \mathbf{D} &
 \end{array}$$

suffit pour se souvenir que les produits de matrices :

$$\mathbf{X}'\mathbf{D}\mathbf{X}\mathbf{Q}, \mathbf{D}\mathbf{X}\mathbf{Q}\mathbf{X}', \mathbf{X}\mathbf{Q}\mathbf{X}'\mathbf{D} \text{ et } \mathbf{Q}\mathbf{X}'\mathbf{D}\mathbf{X}$$

ont un sens. On définit deux matrices $\mathbf{V} = \mathbf{X}'\mathbf{D}\mathbf{X}$ et $\mathbf{W} = \mathbf{X}\mathbf{Q}\mathbf{X}'$ qui s'insèrent par :

$$\begin{array}{ccc}
 & \mathbf{Q} & \mathbf{W} \\
 \boxed{p} \rightarrow \boxed{p} & & \boxed{n} \rightarrow \boxed{n} \\
 \leftarrow & & \leftarrow \\
 \mathbf{V} & & \mathbf{D}
 \end{array}$$

$\mathbf{V}\mathbf{Q}$ et $\mathbf{W}\mathbf{D}$ sont les opérateurs d'Escoufier. Le schéma a des propriétés théoriques très générales qui prennent des significations propres lorsqu'on utilise un ensemble de paramètres originaux. En particulier les ACP classiques, l'analyse des correspondances (AFC), les analyses de correspondances non symétriques (ANSC), les analyses discriminantes (AFD), l'analyse canonique (AC), les analyses de co-inertie (ACO), les analyses sur variables instrumentales (ACPVI), l'analyse des correspondances multiples (ACM), l'analyse factorielle multiple (AFM) et diverses extensions sont des conséquences directes des propriétés générales.

1.1 Rang d'un schéma

Le rang d'une matrice est la dimension du sous-espace engendré par ses lignes ou ses colonnes. Les matrices \mathbf{X} , \mathbf{X}' , $\mathbf{V}\mathbf{Q}$ et $\mathbf{W}\mathbf{D}$ ont même rang, lequel est inférieur ou égal au nombre de lignes de \mathbf{X} (n) et au nombre de colonnes de \mathbf{X} (p). Appelons r le rang commun ou rang du schéma.

1.2 Valeurs propres d'un schéma

Les matrices $\mathbf{V}\mathbf{Q}$ et $\mathbf{W}\mathbf{D}$ sont toujours diagonalisables. La première ($p \times p$) a p valeurs propres distinctes ou confondues dont les r premières sont non nulles. Ces valeurs propres sont positives et rangées par ordre décroissant :

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_r > \lambda_{r+1} = \dots = \lambda_p = 0$$

La seconde ($n \times n$) a n valeurs propres distinctes ou confondues dont les r premières sont non nulles. Ces valeurs propres sont positives et rangées par ordre décroissant :

$$\mu_1 \geq \mu_2 \geq \dots \geq \mu_r > \mu_{r+1} = \dots = \mu_n = 0$$

Les deux ensembles de valeurs propres non nulles sont identiques et on a toujours :

$$\mu_1 = \lambda_1, \mu_2 = \lambda_2, \dots, \mu_r = \lambda_r$$

En toute généralité, il est possible d'avoir des valeurs propres non nulles multiples, mais dans la pratique, ne serait-ce que pour des raisons numériques, ce cas est exclu. On a alors une seule série de valeurs propres non nulles qu'on appelle les valeurs propres du schéma :

$$\lambda_1 > \lambda_2 > \dots > \lambda_r > 0$$

1.3 Vecteurs principaux d'un schéma

Les matrices \mathbf{VQ} et \mathbf{WD} sont toujours diagonalisables et ont même valeurs propres non nulles. La première est \mathbf{Q} -symétrique (la matrice \mathbf{QVQ} est égale à sa transposée) et à ce titre définit une base de vecteurs propres \mathbf{Q} -orthonormée. Donc, il existe une matrice \mathbf{A} telle que :

$$\mathbf{VQA} = \mathbf{A}\Lambda \text{ et } \mathbf{A}'\mathbf{QA} = \mathbf{I}_p$$

$$\Lambda = \text{Diag}(\lambda_1, \lambda_2, \dots, \lambda_p)$$

On ne s'intéresse qu'aux r premiers (valeurs propres non nulles) qui sont définis de façon unique (valeurs propres distinctes) :

$$\mathbf{VQA}_r = \mathbf{A}_r\Lambda_r \text{ et } \mathbf{A}_r'\mathbf{QA}_r = \mathbf{I}_r$$

\mathbf{A}_r est une matrice à p lignes et r colonnes, chacune de ses colonnes étant un vecteur propre \mathbf{Q} -normé de \mathbb{R}^p . Λ_r est la matrice diagonale qui contient les valeurs propres.

La seconde est \mathbf{D} -symétrique (la matrice \mathbf{DWD} est égale à sa transposée) et à ce titre définit une base de vecteurs propres \mathbf{D} -orthonormée. Donc, il existe une matrice \mathbf{B} telle que :

$$\mathbf{WDB} = \mathbf{B}\mathbf{M} \text{ et } \mathbf{B}'\mathbf{DB} = \mathbf{I}_n$$

$$\mathbf{M} = \text{Diag}(\mu_1, \mu_2, \dots, \mu_n)$$

On ne s'intéresse qu'aux r premiers (valeurs propres non nulles) qui sont définis de façon unique (valeurs propres distinctes) :

$$\mathbf{WDB}_r = \mathbf{B}_r\Lambda_r \text{ et } \mathbf{B}_r'\mathbf{DB}_r = \mathbf{I}_r$$

\mathbf{B}_r est une matrice à n lignes et r colonnes, chacune de ses colonnes étant un vecteur propre \mathbf{D} -normé de \mathbb{R}^n . Λ_r est la matrice diagonale qui contient les mêmes valeurs propres.

Les vecteurs de \mathbf{A}_r sont les r axes principaux du schéma, les vecteurs de \mathbf{B}_r sont les r composantes principales du schéma.

Les matrices \mathbf{QV} et \mathbf{DW} sont toujours diagonalisables et ont même valeurs propres non nulles que les précédentes. La première est \mathbf{Q}^{-1} -symétrique (la matrice $\mathbf{Q}^{-1}\mathbf{QV}$ est égale à sa transposée) et à ce titre définit une base de vecteurs propres \mathbf{Q}^{-1} -orthonormée. Donc, il existe une matrice \mathbf{A}^* telle que :

$$\mathbf{QVA}^* = \mathbf{A}^* \Lambda \text{ et } \mathbf{A}^{*t} \mathbf{Q}^{-1} \mathbf{A} = \mathbf{I}_p$$

On ne s'intéresse qu'aux r premiers (valeurs propres non nulles) qui sont définis de façon unique (valeurs propres distinctes) :

$$\mathbf{QVA}_r^* = \mathbf{A}_r^* \Lambda_r \text{ et } \mathbf{A}_r^{*t} \mathbf{Q}^{-1} \mathbf{A}_r^* = \mathbf{I}_r$$

La seconde est \mathbf{D}^{-1} -symétrique (la matrice $\mathbf{D}^{-1}\mathbf{DW}$ est égale à sa transposée) et à ce titre définit une base de vecteurs propres \mathbf{D}^{-1} -orthonormée. Donc, il existe une matrice \mathbf{B}^* telle que :

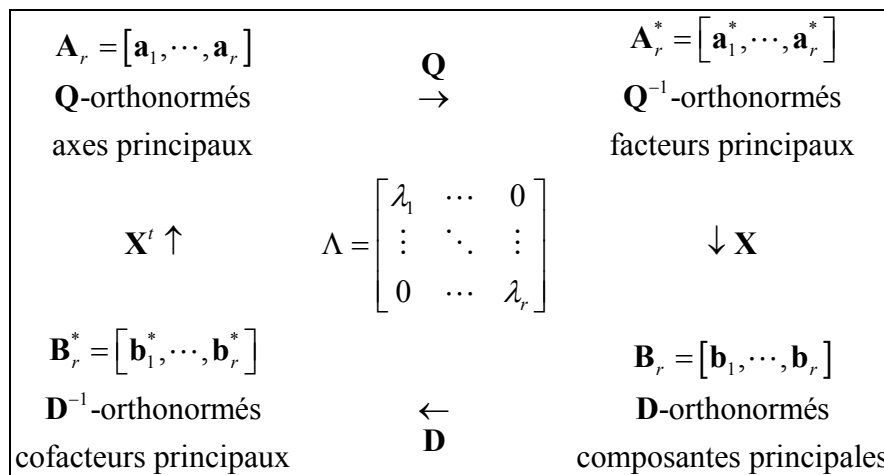
$$\mathbf{DWB}^* = \mathbf{B}^* \mathbf{M} \text{ et } \mathbf{B}^{*t} \mathbf{D}^{-1} \mathbf{B}^* = \mathbf{I}_n$$

On ne s'intéresse qu'aux r premiers (valeurs propres non nulles) qui sont définis de façon unique (valeurs propres distinctes) :

$$\mathbf{DWB}_r^* = \mathbf{B}_r^* \Lambda_r \text{ et } \mathbf{B}_r^{*t} \mathbf{D}^{-1} \mathbf{B}_r^* = \mathbf{I}_r$$

Les vecteurs de \mathbf{A}_r^* sont les r **facteurs principaux** du schéma, les vecteurs de \mathbf{B}_r^* sont les r **cofacteurs principaux** du schéma (Tenenhaus and Young 1985).

Il suffit pour retenir l'essentiel de mémoriser le schéma :



Enfin, il suffit de calculer un seul des quatre systèmes d'axes pour obtenir les trois autres. En effet, on a :

$$\boxed{1) \mathbf{A}_r^* = \mathbf{QA}_r \quad 2) \mathbf{B}_r = \mathbf{XA}_r^* \Lambda_r^{-\frac{1}{2}} \quad 3) \mathbf{B}_r^* = \mathbf{DB}_r \quad 4) \mathbf{A}_r = \mathbf{X}' \mathbf{B}_r^* \Lambda_r^{-\frac{1}{2}}}$$

1.4 Propriétés d'optimalité

Ces vecteurs ont des propriétés fondamentales qui s'écrivent sous forme de théorèmes généraux et qui se retrouvent sous des formes pratiques variées dans la suite.

1) Si on cherche un vecteur \mathbf{a} de \mathbb{R}^p \mathbf{Q} -normé qui maximise $\|\mathbf{XQa}\|_{\mathbf{D}}^2$, la solution est unique. Elle est obtenue pour $\mathbf{a} = \mathbf{a}_1$ et le maximum atteint est λ_1 . Si on cherche un nouveau vecteur \mathbf{a} de \mathbb{R}^p \mathbf{Q} -normé qui maximise à nouveau $\|\mathbf{XQa}\|_{\mathbf{D}}^2$ sous la contrainte $\langle \mathbf{a} | \mathbf{a}_1 \rangle_{\mathbf{Q}} = \mathbf{a}_1' \mathbf{Q} \mathbf{a} = 0$, la solution est unique. Elle est obtenue pour $\mathbf{a} = \mathbf{a}_2$ et le maximum atteint est λ_2 . Et ainsi de suite de proche en proche. Si on cherche un nouveau vecteur \mathbf{a} de \mathbb{R}^p \mathbf{Q} -normé qui maximise $\|\mathbf{XQa}\|_{\mathbf{D}}^2$ sous la contrainte $\langle \mathbf{a} | \mathbf{a}_1 \rangle_{\mathbf{Q}} = \dots = \langle \mathbf{a} | \mathbf{a}_{j-1} \rangle_{\mathbf{Q}} = 0$, la solution est unique. Elle est obtenue pour $\mathbf{a} = \mathbf{a}_j$ et le maximum atteint est λ_j .

On résume en disant que les vecteurs $\mathbf{a}_1, \dots, \mathbf{a}_j, \dots, \mathbf{a}_r$ maximisent successivement sous contrainte d'orthonormalité au sens de \mathbf{Q} la forme quadratique $\|\mathbf{XQa}\|_{\mathbf{D}}^2$.

2) Les vecteurs $\mathbf{a}_1^*, \dots, \mathbf{a}_j^*, \dots, \mathbf{a}_r^*$ maximisent successivement sous contrainte d'orthonormalité au sens de \mathbf{Q}^{-1} la forme quadratique $\|\mathbf{Xa}^*\|_{\mathbf{D}}^2$.

3) Les vecteurs $\mathbf{b}_1, \dots, \mathbf{b}_j, \dots, \mathbf{b}_r$ maximisent successivement sous contrainte d'orthonormalité au sens de \mathbf{D} la forme quadratique $\|\mathbf{X}'\mathbf{Db}\|_{\mathbf{Q}}^2$.

4) Les vecteurs $\mathbf{b}_1^*, \dots, \mathbf{b}_j^*, \dots, \mathbf{b}_r^*$ maximisent successivement sous contrainte d'orthonormalité au sens de \mathbf{D}^{-1} la forme quadratique $\|\mathbf{X}'\mathbf{b}^*\|_{\mathbf{D}}^2$.

5) Si on cherche un couple formé d'un vecteur \mathbf{a} de \mathbb{R}^p \mathbf{Q} -normé et d'un vecteur \mathbf{b} de \mathbb{R}^n \mathbf{D} -normé qui maximise $\langle \mathbf{XQa} | \mathbf{b} \rangle_{\mathbf{D}} = \langle \mathbf{X}'\mathbf{Db} | \mathbf{a} \rangle_{\mathbf{Q}}$, la solution est unique. Elle est obtenue pour $\mathbf{a} = \mathbf{a}_1$ et $\mathbf{b} = \mathbf{b}_1$ et le maximum atteint est $\sqrt{\lambda_1}$. Sous contrainte d'orthonormalité, le résultat s'étend aux couples suivants.

1.5 Approximations de tableau

On considère deux schémas de dualité de mêmes paramètres à l'exclusion du tableau :

$$\begin{array}{ccccc}
 \boxed{p} & \xrightarrow{\mathbf{Q}} & \boxed{p} & \boxed{p} & \xrightarrow{\mathbf{Q}} & \boxed{p} \\
 \mathbf{X}' \uparrow & & \downarrow \mathbf{X} & \text{et } \mathbf{Z}' \uparrow & & \downarrow \mathbf{Z} \\
 \boxed{n} & \xleftarrow{\mathbf{D}} & \boxed{n} & \boxed{n} & \xleftarrow{\mathbf{D}} & \boxed{n}
 \end{array}$$

On peut alors utiliser la configuration définie par :

$$\begin{array}{ccc}
 \boxed{p} & \xrightarrow{\mathbf{Q}} & \boxed{p} \\
 \mathbf{Z}' \uparrow & & \downarrow \mathbf{X} \\
 \boxed{n} & \xleftarrow{\mathbf{D}} & \boxed{n}
 \end{array}$$

L'opérateur qui "fait le tour" $\mathbf{Z}'\mathbf{D}\mathbf{X}\mathbf{Q}$ n'est pas nécessairement diagonalisable mais cette matrice carrée a une trace (somme des éléments diagonaux) qui définit une application qui associe à un couple de tableaux un nombre :

$$(\mathbf{X}, \mathbf{Z}) \mapsto \text{Trace}(\mathbf{Z}'\mathbf{D}\mathbf{X}\mathbf{Q}) = \text{Trace}(\mathbf{X}\mathbf{Q}\mathbf{Z}'\mathbf{D})$$

C'est un produit scalaire dans l'ensemble des tableaux :

$$\langle \mathbf{X} | \mathbf{Z} \rangle_{\mathbf{D}, \mathbf{Q}} = \text{Trace}(\mathbf{Z}'\mathbf{D}\mathbf{X}\mathbf{Q}) = \text{Trace}(\mathbf{X}\mathbf{Q}\mathbf{Z}'\mathbf{D})$$

On a donc une distance entre deux tableaux :

$$d_{\mathbf{D}, \mathbf{Q}}^2(\mathbf{X}, \mathbf{Z}) = \|\mathbf{X} - \mathbf{Z}\|_{\mathbf{D}, \mathbf{Q}}^2 = \langle \mathbf{X} - \mathbf{Z} | \mathbf{X} - \mathbf{Z} \rangle_{\mathbf{D}, \mathbf{Q}} = \text{Trace}\left((\mathbf{X} - \mathbf{Z})' \mathbf{D} (\mathbf{X} - \mathbf{Z}) \mathbf{Q}\right)$$

Par exemple, si

$$\mathbf{D} = \begin{pmatrix} 1 \\ n \end{pmatrix} \mathbf{I}_n \quad \mathbf{Q} = \begin{pmatrix} 1 \\ p \end{pmatrix} \mathbf{I}_p$$

$$d_{\mathbf{D}, \mathbf{Q}}^2(\mathbf{X}, \mathbf{Z}) = \frac{1}{np} \sum_{i=1}^n \sum_{j=1}^p (x_{ij} - z_{ij})^2$$

Le schéma $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ permet d'écrire la décomposition exacte :

$$\boxed{\mathbf{X} = \sum_{k=1}^r \sqrt{\lambda_k} \mathbf{b}_k \mathbf{a}_k'}$$

Cela dérive simplement du fait que la ligne i du tableau, vecteur de \mathbb{R}^p , s'exprime par ses coordonnées dans la base des axes principaux, ou la colonne j du tableau, vecteur de \mathbb{R}^n , s'exprime par ses coordonnées dans la base des composantes principales. Les formules de transition, déjà rencontrées, donnent :

$$\boxed{\mathbf{X}\mathbf{Q}\mathbf{A}_r = \mathbf{B}_r \Lambda_r^{\frac{1}{2}}} \quad \text{et} \quad \boxed{\mathbf{X}'\mathbf{D}\mathbf{B}_r = \mathbf{A}_r \Lambda_r^{\frac{1}{2}}}$$

On peut dire que les coordonnées des projections des lignes sur les axes principaux sont, à une constante près, les composantes des composantes principales et dualement que les coordonnées des projections des colonnes sur les composantes principales sont, à une constante près, les composantes des axes principaux.

Il s'en suit que la meilleure reconstitution de \mathbf{X} par un tableau de rang 1 est :

$$\hat{\mathbf{X}}_1 = \sqrt{\lambda_1} \mathbf{b}_1 \mathbf{a}_1^t$$

que la meilleure reconstitution de \mathbf{X} par un tableau de rang 2 est :

$$\hat{\mathbf{X}}_2 = \sqrt{\lambda_1} \mathbf{b}_1 \mathbf{a}_1^t + \sqrt{\lambda_2} \mathbf{b}_2 \mathbf{a}_2^t$$

que la meilleure reconstitution de \mathbf{X} par un tableau de rang m est :

$$\mathbf{X} = \sum_{k=1}^m \sqrt{\lambda_k} \mathbf{b}_k \mathbf{a}_k^t$$

L'erreur de reconstitution est alors :

$$d_{\mathbf{D}, \mathbf{Q}}^2 \left(\mathbf{X}, \sum_{k=1}^m \sqrt{\lambda_k} \mathbf{b}_k \mathbf{a}_k^t \right) = \left\| \mathbf{X} - \sum_{k=1}^m \sqrt{\lambda_k} \mathbf{b}_k \mathbf{a}_k^t \right\|_{\mathbf{D}, \mathbf{Q}}^2 = \left\| \sum_{k=m+1}^r \sqrt{\lambda_k} \mathbf{b}_k \mathbf{a}_k^t \right\|_{\mathbf{D}, \mathbf{Q}}^2 = \sum_{k=m+1}^r \lambda_k$$

On dit qu'on a opéré la décomposition en valeurs singulières du tableau. Cette idée a été découverte par Eckart et Young (1936).

2 Graphes canoniques

2.1 Procédure centrale

On peut maintenant concevoir une routine unique de "diagonalisation d'un schéma de dualité" qui pour un triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ et le choix d'un entier f dit "nombre d'axes conservés" renvoie les f premières valeurs propres $\lambda_1 > \lambda_2 > \dots > \lambda_f$, les f premiers axes principaux $\mathbf{a}_1, \dots, \mathbf{a}_f$ et les f premières composantes principales $\mathbf{b}_1, \dots, \mathbf{b}_f$. L'entier f peut être choisi à la lecture du graphe des valeurs propres.

Pour celui qui veut écrire le programme, deux remarques sont utiles. La première porte sur la procédure de calcul sous-jacente. On peut utiliser soit la diagonalisation d'une matrice soit la décomposition en valeurs singulières.

La documentation de R est un résumé parfait du problème :

```
eigen                                package:base                        R Documentation
Spectral Decomposition of a Matrix
Description:
  Function `eigen` computes eigenvalues and eigenvectors by
  providing an interface to the EISPACK routines `RS`, `RG`, `CH`
  and `CG`.
  Function `La.eigen` uses the LAPACK routines DSYEV/DSYEV, DGEEV,
```


ZHEEV and ZGEEV.

Usage:

```
eigen(x, symmetric, only.values = FALSE)  
La.eigen(x, symmetric, only.values = FALSE,  
         method = c("dsyevr", "dsyev"))
```

Arguments:

x: a matrix whose spectral decomposition is to be computed.

symmetric: if 'TRUE', the matrix is assumed to be symmetric (or Hermitian if complex) and only its lower triangle is used. If 'symmetric' is not specified, the matrix is inspected for symmetry.

only.values: if 'TRUE', only the eigenvalues are computed and returned, otherwise both eigenvalues and eigenvectors are returned.

method: The LAPACK routine to use in the real symmetric case.

Details:

If 'symmetric' is unspecified, the code attempts to determine if the matrix is symmetric up to plausible numerical inaccuracies. It is faster and surer to set the value yourself.

'La.eigen' is preferred to 'eigen' for new projects, but its eigenvectors may differ in sign and (in the asymmetric case) in normalization. (They may also differ between methods and between platforms.)

The LAPACK routine DSYEVR is usually substantially faster than DSYEV: see <URL:
<http://www.cs.berkeley.edu/~demmel/DOE2000/Report0100.html>>. Most benefits are seen with an optimized BLAS system.

Computing the eigenvectors is the slow part for large matrices.

Using 'method="dsyevr"' requires IEEE 754 arithmetic. Should this not be supported on your platform, 'method="dsyev"' is used, with a warning.

Value:

The spectral decomposition of 'x' is returned as components of a list.

values: a vector containing the p eigenvalues of 'x', sorted in decreasing order, according to 'Mod(values)' if they are complex.

vectors: a p * p matrix whose columns contain the eigenvectors of 'x', or 'NULL' if 'only.values' is 'TRUE'.

For 'eigen(, symmetric = FALSE)' the choice of length of the eigenvectors is not defined by LINPACK. In all other cases the vectors are normalized to unit length.

Recall that the eigenvectors are only defined up to a constant: even when the length is specified they are still only defined up to a scalar of modulus one (the sign for real matrices).

References:

Smith, B. T., Boyle, J. M., Dongarra, J. J., Garbow, B. S., Ikebe, Y., Klema, V., and Moler, C. B. (1976). Matrix Eigensystems Routines - EISPACK Guide. Springer-Verlag Lecture Notes in

Computer Science.

Anderson. E. and ten others (1999) LAPACK Users' Guide. Third Edition. SIAM.

Available on-line at <URL:

http://www.netlib.org/lapack/lug/lapack_lug.html>.

See Also:

`svd', a generalization of `eigen'; `qr', and `chol' for related decompositions.

To compute the determinant of a matrix, the `qr' decomposition is much more efficient: `det'.

`capabilities' to test for IEEE 754 arithmetic.

Examples:

```
> a <- cbind(c(1,-1),c(-1,1))
> a
      [,1] [,2]
[1,]    1   -1
[2,]   -1    1
> eigen(a)
$values
[1] 2 0

$vectors
      [,1] [,2]
[1,] -0.7071 -0.7071
[2,]  0.7071 -0.7071
```

On aura soin de diagonaliser une matrice **symétrique** pour obtenir une base de vecteurs propres orthonormés. La procédure de base utilise une matrice symétrique et renvoie des vecteurs orthonormés dans la métrique canonique. On se contentera de "casser le schéma de dualité" du côté de la plus petite dimension en remplaçant **Q** ou **D** par une décomposition du type :

$$\begin{array}{ccc} \boxed{p} & \xrightarrow{\mathbf{I}_p} & \boxed{p} \\ \mathbf{H} \uparrow & & \downarrow \mathbf{H}^t \\ \boxed{p} & \xrightarrow{\mathbf{Q}} & \boxed{p} \\ \mathbf{X}^t \uparrow & & \downarrow \mathbf{X} \\ \boxed{n} & \xleftarrow{\mathbf{D}} & \boxed{n} \end{array}$$

Si **Q** est une matrice diagonale (ses termes diagonaux sont strictement positifs et les autres sont nuls) on se contente de noter :

$$\mathbf{H} = \mathbf{H}^t = \mathbf{Q}^{\frac{1}{2}} = \text{Diag}(\sqrt{q_{11}}, \dots, \sqrt{q_{pp}}) \Rightarrow \mathbf{H}^{-1} = \mathbf{Q}^{-\frac{1}{2}} = \text{Diag}\left(\frac{1}{\sqrt{q_{11}}}, \dots, \frac{1}{\sqrt{q_{pp}}}\right)$$

Si **Q** n'est pas diagonale, elle est symétrique (comme matrice d'un produit scalaire) et sa décomposition spectrale propose :

$$\mathbf{Q} = \mathbf{W}\Theta\mathbf{W}^t \Rightarrow \mathbf{H} = \Theta^{\frac{1}{2}}\mathbf{W}^t \Rightarrow \mathbf{H}^{-1} = \mathbf{W}\Theta^{-\frac{1}{2}}$$

On peut aussi utiliser la procédure de Choleski. Dans tous les cas, on diagonalisera alors, avec la procédure de base une matrice symétrique et on retrouvera les axes par :

$$\mathbf{HX'DXH'} = \mathbf{Z\Lambda Z'} \Rightarrow \mathbf{A} = \mathbf{H^{-1}Z}$$

Les composantes principales sont enfin obtenues par les formules de transition. Si la petite dimension est de l'autre côté, on applique la même procédure dans l'autre sens.

On obtient le même résultat avec la décomposition en valeurs singulières :

```
svd                                package:base                                R Documentation
Singular Value Decomposition of a Matrix
Description:
  Compute the singular-value decomposition of a rectangular matrix.
Usage:
  svd(x, nu = min(n, p), nv = min(n, p))
  La.svd(x, nu = min(n, p), nv = min(n, p), method = c("dgesdd", "dgesvd"))
Arguments:
  x: a matrix whose SVD decomposition is to be computed.
  nu: the number of left singular vectors to be computed. This
      must be one of `0`, `nrow(x)` and `ncol(x)`, except for
      `method = "dgesdd"`.
  nv: the number of right singular vectors to be computed. This
      must be one of `0` and `ncol(x)`.
  method: The LAPACK routine to use in the real case.
Details:
  The singular value decomposition plays an important role in many
  statistical techniques.
Value:
  The SVD decomposition of the matrix as computed by LINPACK,
      X = U D V',
  where U and V are orthogonal, V' means V transposed, and D is a
  diagonal matrix with the singular values D[i,i]. Equivalently, D
  = U' X V, which is verified in the examples, below.
  The components in the returned value correspond directly to the
  values returned by DSVDC.
  d: a vector containing the singular values of `x`.
  u: a matrix whose columns contain the left singular vectors of
      `x`.
  v: a matrix whose columns contain the right singular vectors of
      `x`.
  For `La.svd` the return value replaces `v` by `vt`, the
  (conjugated if complex) transpose of `v`.
```

References:

Dongarra, J. J., Bunch, J. R., Moler, C. B. and Stewart, G. W. (1978) LINPACK Users Guide. Philadelphia: SIAM Publications.

Anderson. E. and ten others (1999) LAPACK Users' Guide. Third Edition. SIAM.
 Available on-line at <URL:
http://www.netlib.org/lapack/lug/lapack_lug.html>.

Il conviendra alors de casser deux fois le schéma de départ, dans la même logique :

$$\begin{array}{ccc}
 \boxed{p} & \xrightarrow{\mathbf{I}_p} & \boxed{p} \\
 \mathbf{H} \uparrow & & \downarrow \mathbf{H}' \\
 \boxed{p} & \xrightarrow{\mathbf{Q}} & \boxed{p} \\
 \mathbf{X}' \uparrow & & \downarrow \mathbf{X} \\
 \boxed{n} & \xleftarrow{\mathbf{D}} & \boxed{n} \\
 \mathbf{K}' \uparrow & & \downarrow \mathbf{K} \\
 \boxed{n} & \xleftarrow{\mathbf{I}_n} & \boxed{n}
 \end{array}$$

On prendra la décomposition en valeurs singulières de :

$$\mathbf{KXH}' = \mathbf{UDV}'$$

On retrouvera les valeurs propres comme carrés des valeurs singulières et les axes et composantes par $\mathbf{B} = \mathbf{K}^{-1}\mathbf{U}$ et $\mathbf{A} = \mathbf{H}^{-1}\mathbf{V}$. On vérifie qu'on obtient ainsi des bases orthonormales pour les métriques choisies :

$$\mathbf{A}'\mathbf{Q}\mathbf{A} = \mathbf{A}'\mathbf{H}'\mathbf{H}\mathbf{A} = \mathbf{V}'\mathbf{V} = \mathbf{I}_p$$

2.2 Décentrage

A partir du schéma général, on peut définir des pratiques dont les ACP centrées et normées sont des cas particuliers.

Prenons l'exemple d'un tableau d'examen. On trouvera les données sur la carte Deug de la pile Data du logiciel ADE-4. 104 étudiants (Deug Mass 2° année) ont leur notes pour 9 matières : 1- Algèbre et Analyse des données (sur 100), 2- Analyse (sur 60), 3- Probabilités (sur 80), 4-Informatique (sur 60), 5-Dominante (Sociologie ou Economie sur 120), 6-Options (sur 40), 7-Ouvertures (sur 40), 8-Anglais (sur 40) et 9-Education physique et sportive (bonification <=15).

```

> data(deug)
> deug$tab[1:3,]
  Algebra Analysis Proba Informatic Economy Option1 Option2 English Sport
1      40      26.0   26      26.0   51.9      17      24      19.0  11.5
2      37      34.5   37      32.0   72.0      24      22      26.0  11.5
3      37      41.0   29      34.5   72.0      24      27      19.6  11.5
etc.
    
```

prcomp

package:mva

R Documentation

Principal Components Analysis

Description:

Performs a principal components analysis on the given data matrix and returns the results as an object of class `prcomp`.

Usage:

```
prcomp(x, retx = TRUE, center = TRUE, scale. = FALSE, tol = NULL)
```

Arguments:

- x: a matrix (or data frame) which provides the data for the principal components analysis.
- retx: a logical value indicating whether the rotated variables should be returned.
- center: a logical value indicating whether the variables should be shifted to be zero centered. Alternately, a vector of length equal the number of columns of `x` can be supplied. The value is passed to `scale`.
- scale: a logical value indicating whether the variables should be scaled to have unit variance before the analysis takes place. The default is `FALSE` for consistency with S, but in general scaling is advisable. Alternately, a vector of length equal the number of columns of `x` can be supplied. The value is passed to `scale`.
- tol: a value indicating the magnitude below which components should be omitted. (Components are omitted if their standard deviations are less than or equal to `tol` times the standard deviation of the first component.) With the default null setting, no components are omitted. Other settings for tol could be `tol = 0` or `tol = sqrt(.Machine\$double.eps)`, which would omit essentially constant components.

L'ACP normée est obtenue avec :

```
> pr0 <- prcomp(deug$tab, scale=T)
> pr0
Standard deviations:
[1] 1.7611 1.1675 1.0160 0.9665 0.8601 0.7581 0.7298 0.6615 0.5336
```

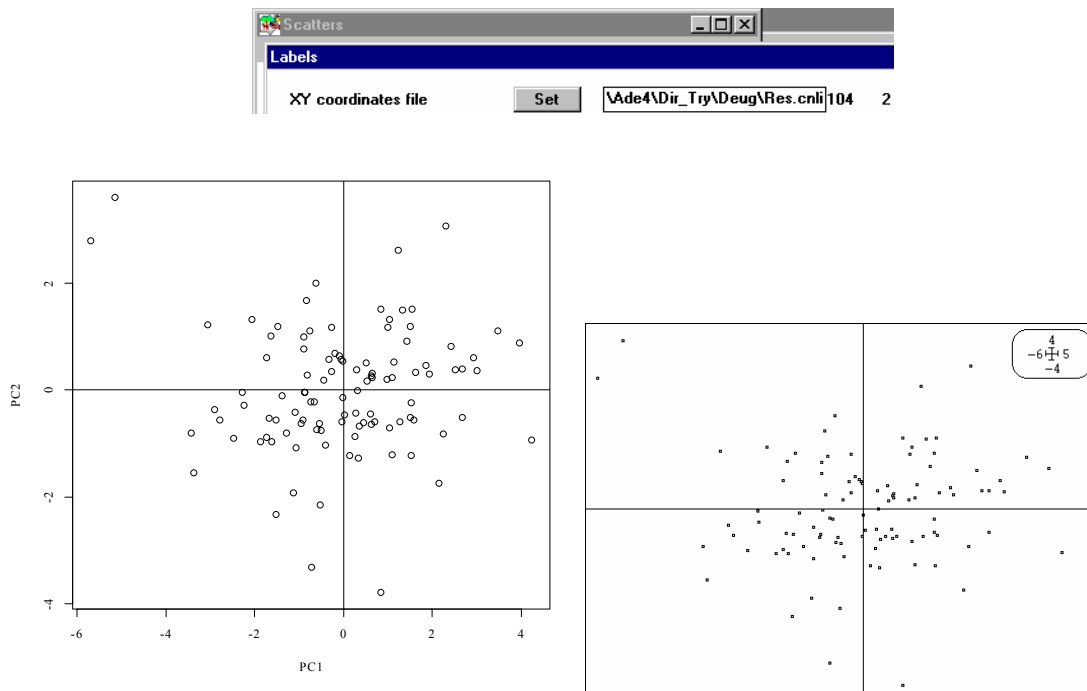
Dans le logiciel ADE4 2001 (<http://pbil.univ-lyon1.fr/ADE-4/ADE-4F.html>) :



Num.	Eigenval.	R.Iner.	R.Sum		Num.	Eigenval.	R.Iner.	R.Sum	
01	+3.1014E+00	+0.3446	+0.3446		02	+1.3630E+00	+0.1514	+0.4960	
03	+1.0323E+00	+0.1147	+0.6107		04	+9.3405E-01	+0.1038	+0.7145	
05	+7.3975E-01	+0.0822	+0.7967		06	+5.7467E-01	+0.0639	+0.8606	
07	+5.3254E-01	+0.0592	+0.9197		08	+4.3754E-01	+0.0486	+0.9684	
09	+2.8478E-01	+0.0316	+1.0000						

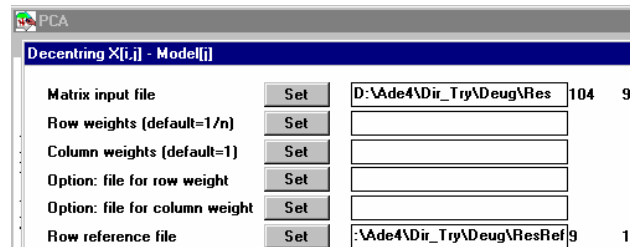
Le lien se fait par :

```
> pr0$sdev^2
[1] 3.1014 1.3630 1.0323 0.9341 0.7398 0.5747 0.5325 0.4375 0.2848
Référence: 50,30,40,30,60,20,20,20,0.
> names(pr0)
[1] "sdev"      "rotation"  "x"
> plot(pr0$x[,1:2])
> abline(h=0)
> abline(v=0)
```



La typologie des étudiants se fait autour de l'étudiant moyen. Si on raisonne en termes ordinaires, on dit "j'ai tant de points d'avance ici, tant de points de retard là, ...".

```
> ref <- c(50,30,40,30,60,20,20,20,0)
> ref
[1] 50 30 40 30 60 20 20 20 0
> pr1 <- prcomp(deug$tab,center=ref)
> pr1$sdev
[1] 20.271 16.320 8.428 7.716 6.877 5.873 4.563 3.931 3.219
> pr1$sdev^2
[1] 410.91 266.33 71.03 59.53 47.29 34.49 20.82 15.45 10.36
```



Total inertia: 927.216

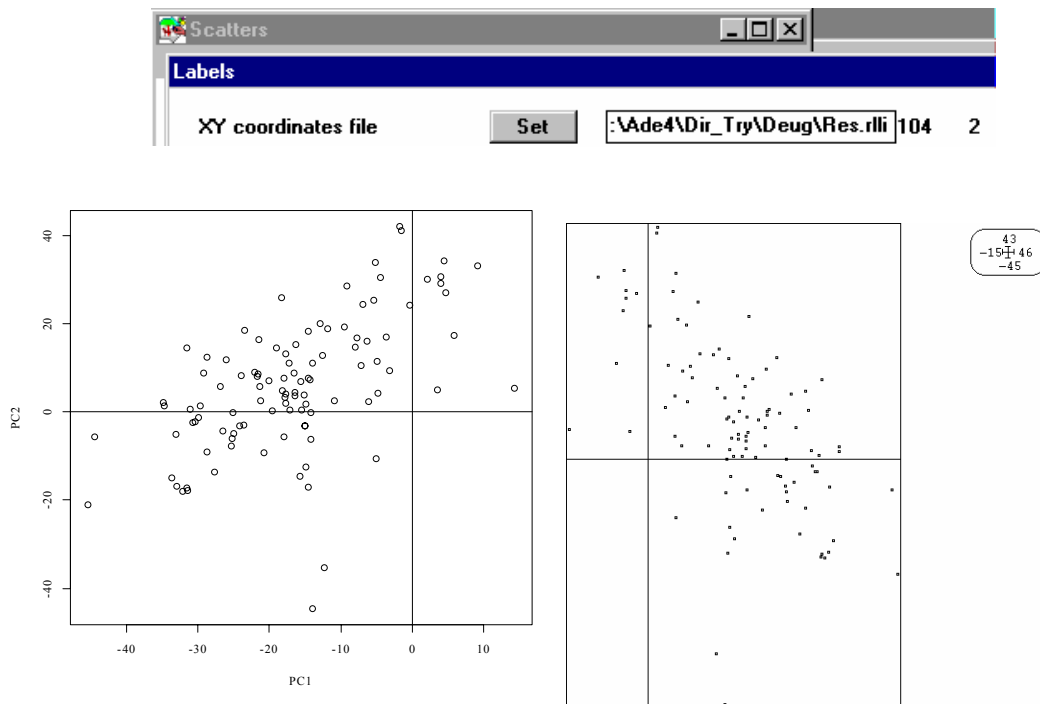
Num.	Eigenval.	R.Iner.	R.Sum	Num.	Eigenval.	R.Iner.	R.Sum
01	+4.0696E+02	+0.4389	+0.4389	02	+2.6377E+02	+0.2845	+0.7234
03	+7.0346E+01	+0.0759	+0.7992	04	+5.8958E+01	+0.0636	+0.8628
05	+4.6838E+01	+0.0505	+0.9133	06	+3.4162E+01	+0.0368	+0.9502
07	+2.0617E+01	+0.0222	+0.9724	08	+1.5302E+01	+0.0165	+0.9889
09	+1.0263E+01	+0.0111	+1.0000				

Une toute petite difficulté :

The root-mean-square for a column is obtained by computing the square-root of the sum-of-squares of the non-missing values in the column divided by the number of non-missing values **minus one**.

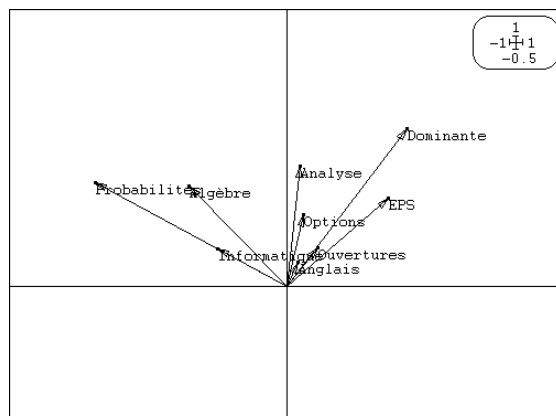
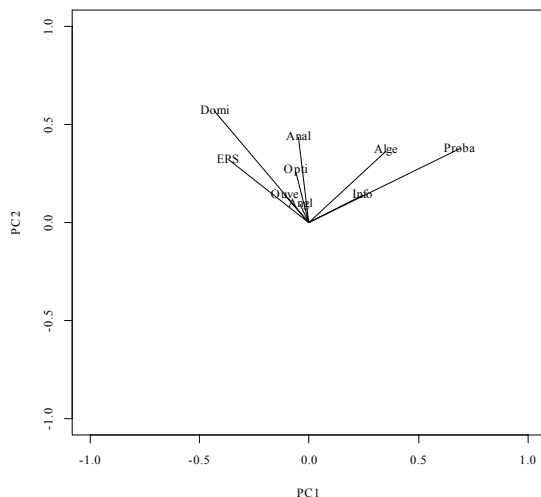
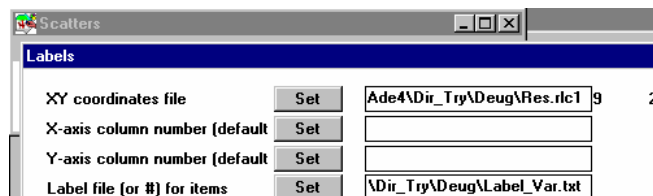
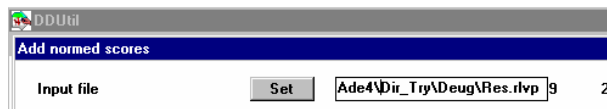
```
> 406.96*104/103
[1] 410.9 (n-1) est une obligation, dommage !

> plot(pr1$x[,1:2])
> abline(h=0)
> abline(v=0)
```



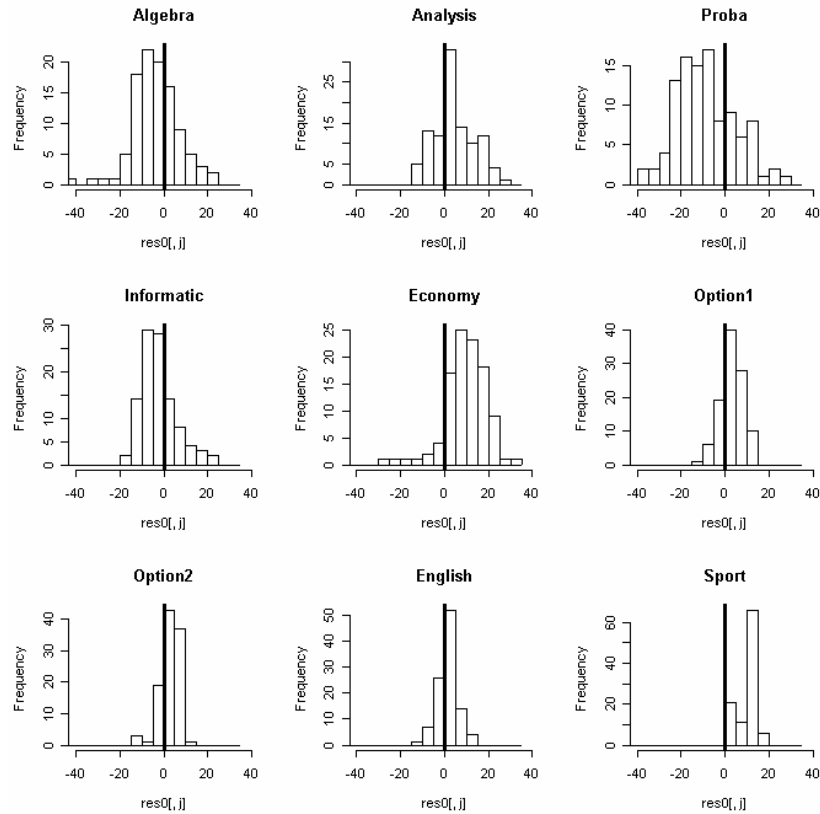
Bien noter que si u est propre $-u$ aussi, donc que l'orientation d'un axe est aléatoire. Qu'est-ce qu'on apprend ?

```
> labelvar
c("Alge", "Anal", "Proba", "Info", "Domi", "Opti", "Ouve", "Angl", "EPS")
> plot(pr1$rotation[,1:2], type="n", xlim=c(-1,1), ylim=c(-1,1))
> segments(0,0,pr1$rotation[,1],pr1$rotation[,2])
> text(pr1$rotation[,1],pr1$rotation[,2],labelvar)
```



L'explication :

```
res0 <- t(t(deug$tab)-ref)
par(mfrow=c(3,3))
for(j in 1:9){
  hist(res0[,j],xlim=c(-40,40),breaks=seq(-45,35,by=5),
       main=names(deug$tab)[j],right=F)
  abline(v=0,lwd=3)
}
```



Il est bien connu que certains professeurs sont plus sympathiques que d'autres.

La procédure prcomp de R utilise le schéma de gauche, celle de PCA utilise celui de droite :

$\boxed{p=9}$ $\mathbf{X}' \uparrow$	$\mathbf{Q} = \mathbf{I}_9$ \rightarrow	$\boxed{p=9}$ $\downarrow \mathbf{X} = [y_{ij} - \alpha_j]$	$\boxed{p=9}$ $\mathbf{X}' \uparrow$	$\mathbf{Q} = \mathbf{I}_9$ \rightarrow	$\boxed{p=9}$ $\downarrow \mathbf{X} = [y_{ij} - \alpha_j]$
$\boxed{n=104}$	\leftarrow	$\mathbf{D} = \frac{1}{(n-1)} \mathbf{I}_n$	$\boxed{n=104}$	\leftarrow	$\mathbf{D} = \frac{1}{n} \mathbf{I}_n$

Les deux sont très voisins. Ce qui est plus variable est l'interprétation donnée aux calculs :

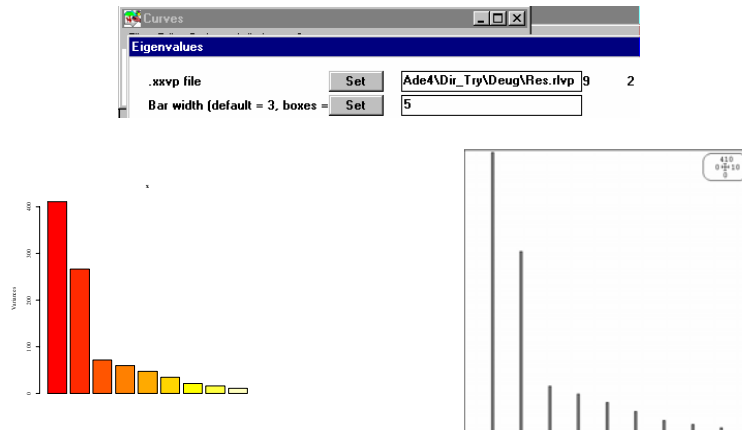
scale: a logical value indicating whether the variables should be scaled to have unit variance before the analysis takes place. The default is `FALSE' for consistency with S, **but in general scaling is advisable**. Alternately, a vector of length equal the number of columns of `x' can be supplied. The value is passed to `scale'.

Ceci indique que l'interprétation en terme de modélisation des données est privilégiée. L'analyse fabrique un modèle des données par les données (Eckart and Young 1936) :

$$\hat{\mathbf{X}}_2 = \sqrt{\lambda_1} \mathbf{b}_1 \mathbf{a}_1' + \sqrt{\lambda_2} \mathbf{b}_2 \mathbf{a}_2'$$

La note d'un étudiant dans une matière est le nombre de points nécessaires pour la moyenne corrigé par deux produits du type : score du professeur \times score de l'étudiant. En utilisant une représentation géométrique conservant les échelles, donc sans changement d'échelle ni explicite ni implicite, l'interprétation en termes de géométrie est privilégiée. Un nuage de 104 points est projeté sur un plan pour le voir.

> plot(pr1)

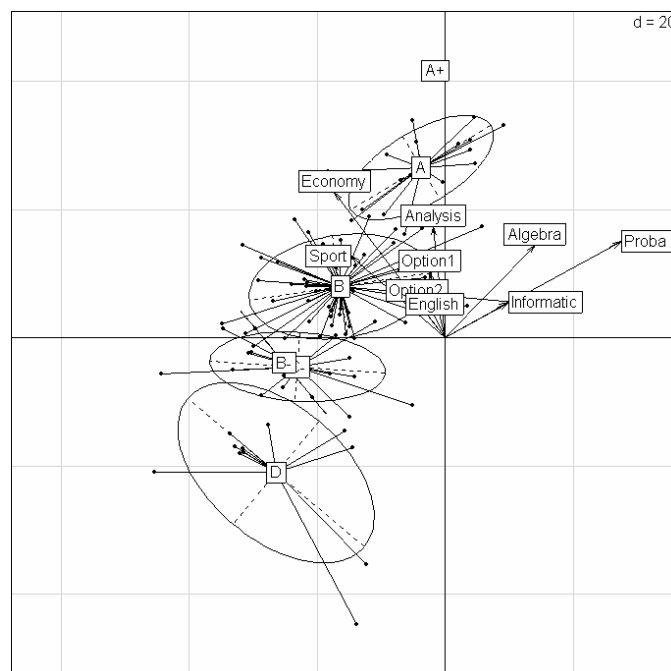


La dimension 2 s'impose par le graphe des valeurs propres.

Pour refaire les calculs avec le package ade4 pour R :

<http://cran.r-project.org/src/contrib/PACKAGES.html#ade4>)

```
data(deug)
# decentred PCA
pcal <- dudi.pca(deug$stab, scal = FALSE, center = deug$cent,
  scan = FALSE)
s.class(pcal$li, deug$result)
s.arrow(40 * pcal$c1, add.plot = TRUE)
```



2.3 Redondance

Si on veut faire disparaître les différences (sensibles !) de système de notation (en moyenne et en amplitude), l'ACP normée s'impose.

$$\begin{array}{ccc}
 \boxed{p} & \xrightarrow{\mathbf{Q} = \mathbf{I}_p} & \boxed{p} \\
 \mathbf{X}^t \uparrow & & \downarrow \mathbf{X} \\
 \boxed{n} & \xleftarrow{\mathbf{D} = \frac{1}{(n-1)} \mathbf{I}_n} & \boxed{n}
 \end{array}
 \qquad
 \begin{array}{ccc}
 \boxed{p} & \xrightarrow{\mathbf{Q} = \mathbf{I}_p} & \boxed{p} \\
 \mathbf{X}^t \uparrow & & \downarrow \mathbf{X} \\
 \boxed{n} & \xleftarrow{\mathbf{D} = \frac{1}{n} \mathbf{I}_n} & \boxed{n}
 \end{array}$$

$$x_{ij} = \frac{y_{ij} - m_j}{\sqrt{\frac{1}{(n-1)} \sum_{i=1}^n (y_{ij} - m_j)^2}}
 \qquad
 x_{ij} = \frac{y_{ij} - m_j}{\sqrt{\frac{1}{n} \sum_{i=1}^n (y_{ij} - m_j)^2}}$$

Les deux schémas donnent des résultats identiques. Une composante principale est alors un score des lignes de norme 1, centré, qui maximise :

$$\|\mathbf{XDa}\|_{\mathbf{Q}}^2 = \sum_{j=1}^p \text{corr}^2(\mathbf{a}, \mathbf{X}^j)$$

princomp package:mva R Documentation

Principal Components Analysis

Description:

``princomp'` performs a principal components analysis on the given data matrix and returns the results as an object of class ``princomp'`.

``loadings'` extracts the loadings.

``screeplot'` plots the variances against the number of the principal component. This is also the ``plot'` method.

Usage:

```

princomp(x, cor = FALSE, scores = TRUE, covmat = NULL,
         subset = rep(TRUE, nrow(as.matrix(x))))
loadings(x)
plot(x, npcs = min(10, length(x$sdev)),
     type = c("barplot", "lines"), ...)
screeplot(x, npcs = min(10, length(x$sdev)),
          type = c("barplot", "lines"), ...)

print(x, ...) summary(object) predict(object, ...)

```

Arguments:

`x`: a matrix (or data frame) which provides the data for the principal components analysis.

`cor`: a logical value indicating whether the calculation should use the correlation matrix or the covariance matrix.

`scores`: a logical value indicating whether the score on each principal component should be calculated.

`covmat`: a covariance matrix, or a covariance list as returned by ``cov.wt'`, ``cov.mve'` or ``cov.mcd'`. If supplied, this is used

rather than the covariance matrix of `x'.

subset: a vector used to select rows (observations) of the data matrix `x'.

x, object: an object of class `"princomp"`, as from `princomp()`'.

npcs: the number of principal components to be plotted.

type: the type of plot.

...: graphics parameters.

Details:

The calculation is done using `eigen` on the correlation or covariance matrix, as determined by `cor`. This is done for compatibility with the S-PLUS result. A preferred method of calculation is to use `svd` on `x`, as is done in `prcomp`.

Note that the default calculation uses divisor `N` for the covariance matrix.

The `print` method for these objects prints the results in a nice format and the `plot` method produces a scree plot.

Value:

`princomp` returns a list with class `"princomp"` containing the following components:

sdev: the standard deviations of the principal components.

loadings: the matrix of variable loadings (i.e., a matrix whose columns contain the eigenvectors).

center: the means that were subtracted.

scale: the scalings applied to each variable.

n.obs: the number of observations.

scores: if `scores = TRUE`, the scores of the supplied data on the principal components.

call: the matched call.

References:

Mardia, K. V., J. T. Kent and J. M. Bibby (1979). *Multivariate Analysis*, London: Academic Press.

Venables, W. N. and B. D. Ripley (1997, 9). *Modern Applied Statistics with S-PLUS*, Springer-Verlag.

```
> pr2 <- princomp(res,cor=T)
```

```
> pr2
```

```
Call:
```

```
princomp(x = res, cor = T)
```

Standard deviations:

```
Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9  
1.7611 1.1675 1.0160 0.9665 0.8601 0.7581 0.7298 0.6615 0.5336
```

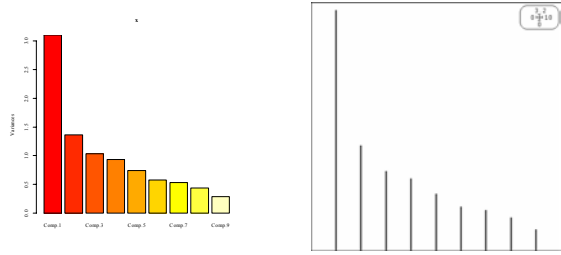
9 variables and 104 observations.



Total inertia: 9

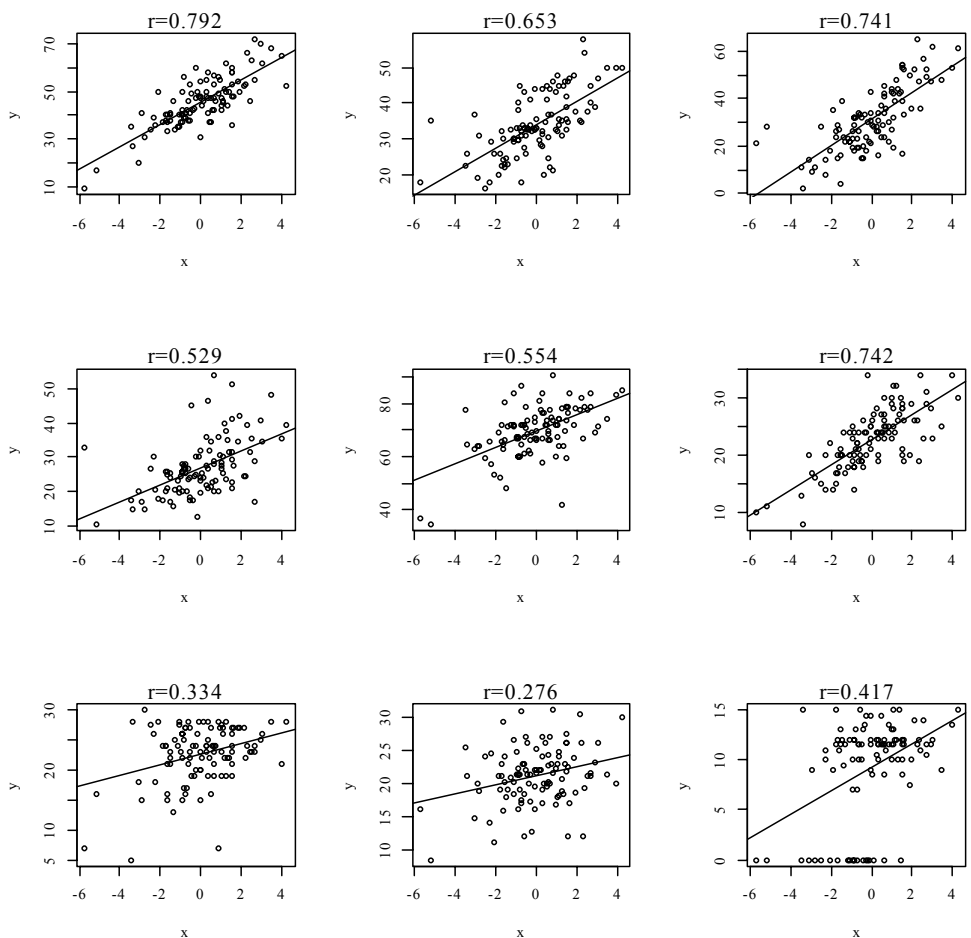
Num.	Eigenval.	R.Iner.	R.Sum	Num.	Eigenval.	R.Iner.	R.Sum
01	+3.1014E+00	+0.3446	+0.3446	02	+1.3630E+00	+0.1514	+0.4960
03	+1.0323E+00	+0.1147	+0.6107	04	+9.3405E-01	+0.1038	+0.7145
05	+7.3975E-01	+0.0822	+0.7967	06	+5.7467E-01	+0.0639	+0.8606
07	+5.3254E-01	+0.0592	+0.9197	08	+4.3754E-01	+0.0486	+0.9684
09	+2.8478E-01	+0.0316	+1.0000				

```
> names(pr2)
[1] "sdev" "loadings" "center" "scale" "n.obs" "scores" "call"
> pr2$sdev^2
Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
3.1014 1.3630 1.0323 0.9341 0.7398 0.5747 0.5325 0.4375 0.2848
```



Il reste un seul axe. Le graphe canonique, celui qui exprime le théorème cité, est alors :

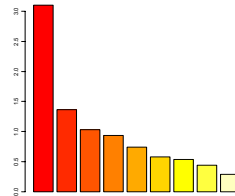
```
> par(mfrow=c(3,3))
> plotreg
function(y,x){
  plot(x,y)
  abline(lm(y~x))
  mtext(paste("r=",round(cor(x,y),digits=3),sep=""))
}
> apply(res, 2, plotreg, x=pr2$scores[,1])
```



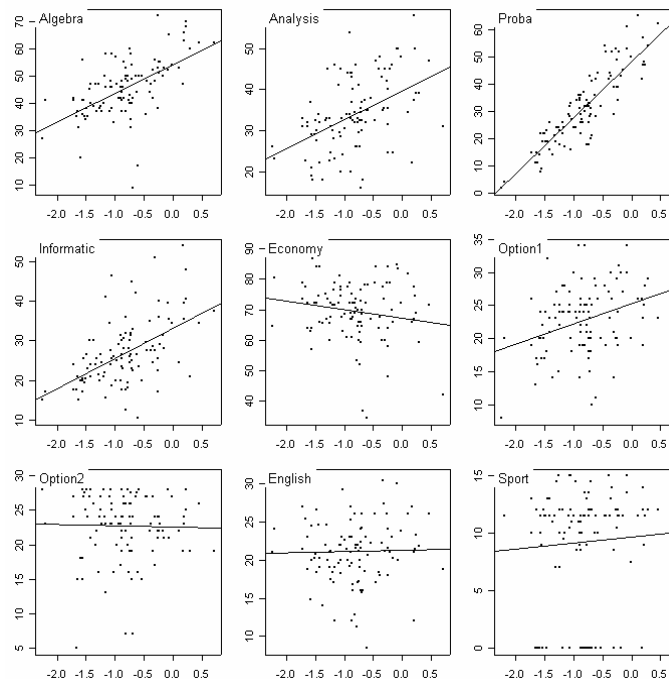
```
> pr2$loadings[,1]*pr2$sdev[1]
      V1      V2      V3      V4      V5      V6      V7      V8      V9
0.7925 0.6532 0.7410 0.5287 0.5539 0.7416 0.3336 0.2755 0.4172
```

Réduire ou éliminer la redondance est un des objectifs de l'analyse des données. L'abondance de variables souvent redondantes est un obstacle majeur à la modélisation statistique. Les coordonnées factorielles lorsqu'elles sont interprétées sont de bons substituts aux mesures. Les mêmes calculs dans ade4 :

```
acp1 <- dudi.pca(deug$stab)
```

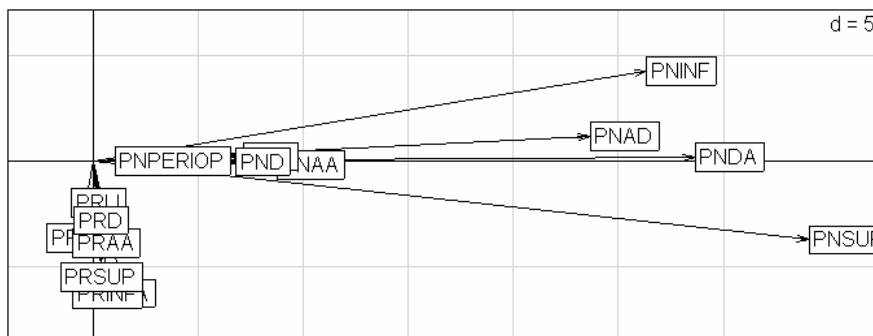


```
Select the number of axes: 1
score(pca1)
```

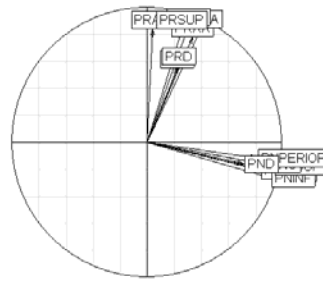


Les données de J.M. Lascaux (1996) donnent des exemples remarquables : 306 truites - 15 variables de coloration de la robe. ACP centrée (en haut), normée (en bas) :

```
data(lascaux)
s.arrow(dudi.pca(lascaux$colo, scal=F, scan=F) $co)
```



```
s.arrow(dudi.pca(lascaux$colo, scal=T, scan=F)$co)
```



Il y a simplement deux variables : le plus simple est de sommer par paquets. Dans d'autre cas, la redondance n'est pas un parasite mais l'essentiel de la recherche. La même figure est utilisée pour caractériser la cohérence des jurys de classement : les juges sont alors les variables.

2.4 Valeurs propres

Quelle que soit la transformation préalable, on se retrouve avec un tableau \mathbf{X} comportant n lignes et p colonnes. La géométrie du nuage des lignes est basée sur le produit scalaire :

$$\langle \mathbf{a} | \mathbf{b} \rangle = a_1 b_1 + \dots + a_p b_p = 1 a_1 b_1 + \dots + 1 a_p b_p$$

La géométrie du nuage des colonnes est basée sur le produit scalaire :

$$\langle \mathbf{a} | \mathbf{b} \rangle = \frac{1}{n} (a_1 b_1 + \dots + a_n b_n) = \frac{1}{n} a_1 b_1 + \dots + \frac{1}{n} a_n b_n$$

C'est l'option habituelle dans les cas simples. Mais en toute généralité, on peut choisir d'utiliser dans l'espace des lignes :

$$\langle \mathbf{a} | \mathbf{b} \rangle_{\omega} = \omega_1 a_1 b_1 + \dots + \omega_p a_p b_p$$

et dans l'espace des colonnes :

$$\langle \mathbf{a} | \mathbf{b} \rangle_{\pi} = \pi_1 a_1 b_1 + \dots + \pi_n a_n b_n$$

Ceci a un sens si et seulement si les coefficients ω_j et π_i sont positifs. Il y a autant de coefficients, qui interviennent dans la géométrie de l'espace des lignes que de colonnes et réciproquement. Les ω_j sont dits poids des colonnes et les π_i sont dits poids des lignes. L'analyse est caractérisée alors par un triplet formé du tableau transformé \mathbf{Y} et de deux matrices diagonales (cette présentation étant faite pour la cohérence mathématique du total) qui s'écrivent :

$$\mathbf{Y} = [y_{ij}] \quad 1 \leq i \leq n, 1 \leq j \leq p$$

L'inertie projetée est maximum pour le premier axe principal. Ce maximum atteint est la première valeur propre (λ_1). On cherche un second axe, orthogonal au précédent qui maximise sous cette contrainte la même quantité. On trouve le deuxième axe principal. Le maximum atteint est la seconde valeur propre (λ_2). Et ainsi de suite. Si on raisonne sur le nuage des vecteurs colonnes, on a le même objectif et les mêmes résultats :

$$\sum_{j=1}^p \omega_j d^2(P_j, p_j) + \sum_{j=1}^p \omega_j d^2(p_j, O) = \sum_{j=1}^p \omega_j d^2(P_j, O)$$

Minimum Maximum Constante

La première composante principale donne le maximum λ_1 , la seconde le nouveau maximum λ_2 , etc. L'inertie totale du nuage des lignes comme l'inertie totale du nuage des colonnes se décompose de manière unique :

$$\sum_{i=1}^n \pi_i d^2(M_i, O) = \sum_{j=1}^p \omega_j d^2(P_j, O) = \lambda_1 + \lambda_2 + \dots + \lambda_r$$

Le graphe des valeurs propres représente donc la manière dont se prend en compte la variabilité (progressivement jusqu'à 100%) ou comment on s'approche des données par automodélisation (jusqu'à 0%).

Les cas typiques sont :

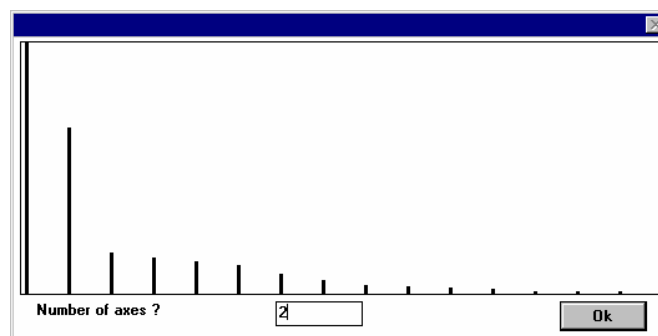
Carte Voiture (PCA: Correlation matrix PCA) :

Num.	Eigenval.	R.Iner.	R.Sum	Num.	Eigenval.	R.Iner.	R.Sum
01	+4.6560E+00	+0.7760	+0.7760	02	+9.1522E-01	+0.1525	+0.9285
03	+2.4043E-01	+0.0401	+0.9686	04	+1.0271E-01	+0.0171	+0.9857
05	+6.4656E-02	+0.0108	+0.9965	06	+2.0961E-02	+0.0035	+1.0000

93% de la variabilité totale s'exprime sur un plan (dimension 2).

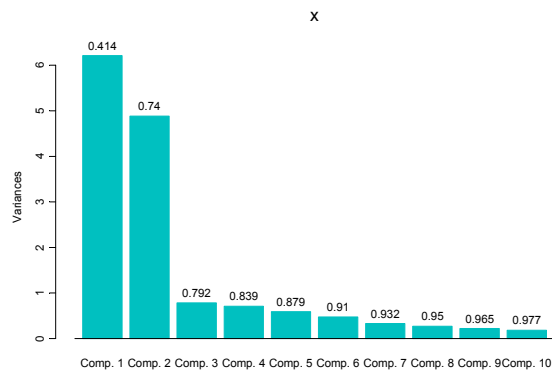
Carte Rhone (PCA: Correlation matrix PCA) :

Num.	Eigenval.	R.Iner.	R.Sum	Num.	Eigenval.	R.Iner.	R.Sum
01	+6.2743E+00	+0.4183	+0.4183	02	+4.1409E+00	+0.2761	+0.6943
03	+1.0082E+00	+0.0672	+0.7616	04	+8.5920E-01	+0.0573	+0.8188
05	+7.6219E-01	+0.0508	+0.8697	06	+6.6565E-01	+0.0444	+0.9140
07	+4.4427E-01	+0.0296	+0.9436	08	+3.1199E-01	+0.0208	+0.9644
09	+1.7067E-01	+0.0114	+0.9758	10	+1.4584E-01	+0.0097	+0.9855
11	+9.7799E-02	+0.0065	+0.9921	12	+6.5818E-02	+0.0044	+0.9965
13	+2.9567E-02	+0.0020	+0.9984	14	+2.3569E-02	+0.0016	+1.0000
15	+0.0000E+00	+0.0000	+1.0000				



On garde 69% d'inertie sur un plan sans chercher à augmenter ce taux: le nuage est en dimension 2.

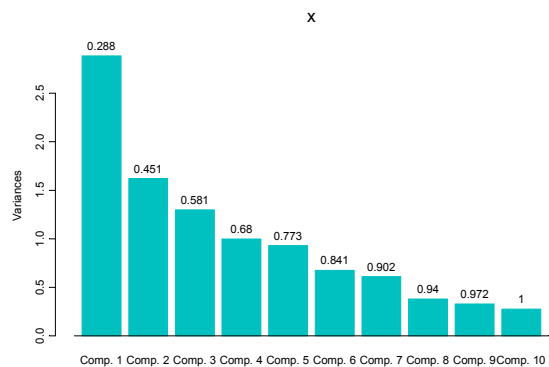
306 truites - 15 variables de coloration de la robe (Lascaux 1996) :



```
plot(princomp(log(color+1), cor=T))
```

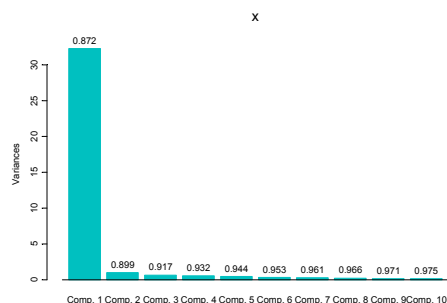
On garde 74% d'inertie sur un plan sans chercher à augmenter ce taux: le nuage est en dimension 2.

306 truites - 10 variables méristiques (ibidem):



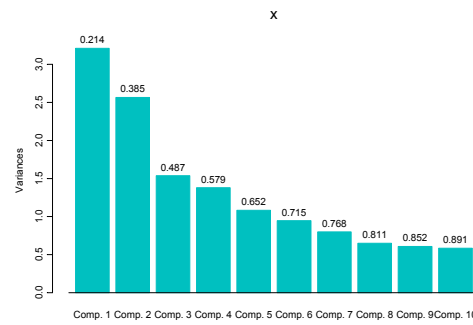
On ne garde qu'un axe, l'analyse souligne la faible cohérence de ce type de variable. Ce n'est pas une mauvaise analyse, au contraire. Ces variables sont très peu corrélées et c'est un fait essentiel.

306 truites - 37 variables morphométriques (ibidem) :



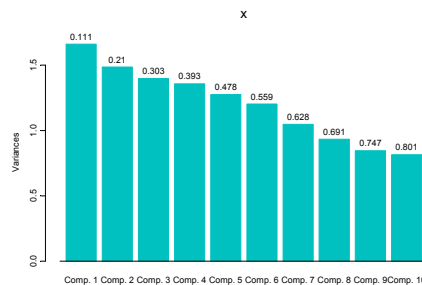
On ne garde qu'un axe, expression de l'effet taille. L'analyse illustre une trivialité et ce n'est pas la concentration de l'inertie dans peu de dimensions qui fait son intérêt.

306 truites - 15 variables ornementales binaires (ibidem) :



On garde deux facteurs.

```
plot(princomp(rmvnorm(n=100, d=15), cor=T))
```



L'ACP d'un jeu de données aléatoires ne donne aucune structure. Dès que les valeurs propres sont régulièrement décroissantes, on est dans la partie non organisée des données (pas forcément sans signification expérimentale par ailleurs).

2.5 Profils et biplot

Les données de profils sont une structure commune de données. Dans le tableau de départ \mathbf{Y} , chaque ligne est une distribution de fréquences notée $y_{ij} = f_{j/i}$ (la fréquence de la colonne j sachant qu'on est sur la ligne i). Le terme de fréquence est pris au sens large, fréquence des électeurs du canton i ayant voté pour le candidat j , fréquence de l'allèle j dans la population i , fréquence des proies de la catégorie j dans l'estomac i , etc.

Dans tous les cas $\sum_{j=1}^p f_{j/i} = 1$. Les moyennes par colonnes pour la pondération uniforme $q_j = \frac{1}{n} \sum_{i=1}^n f_{j/i}$ vérifient également $\sum_{j=1}^p q_j = 1$. Le tableau \mathbf{X} centré qui en découle $x_{ij} = y_{ij} - q_j$ a donc la propriété particulière :

$$\mathbf{X}\mathbf{1}_p = \mathbf{0}$$

Le schéma suivant sera appelé ACP centrée sur profils :

$$\begin{array}{ccc}
 \boxed{p} & \mathbf{Q} = \mathbf{I}_p & \boxed{p} \\
 \mathbf{X}^t \uparrow & \rightarrow & \downarrow \mathbf{X} = [f_{j/i} - q_j] \\
 \boxed{n} & \mathbf{D} = \frac{1}{n} \mathbf{I}_n & \boxed{n}
 \end{array}$$

Les vecteurs $\mathbf{1}_p$ et $\mathbf{1}_n$ sont respectivement axe principal et composante principale. La somme des composantes d'un autre axe principal est donc nulle par orthogonalité (propriété particulière des données en profils) comme la somme des composantes d'une autre composante principale donc nulle par orthogonalité (propriété commune à toutes les ACP centrées). Les deux systèmes de coordonnées factorielles sont centrés.

La relation de transition prend alors une signification particulière.

$$\mathbf{L}_r = \mathbf{XQA}_r = \mathbf{B}_r \Lambda_r^{\frac{1}{2}} \Rightarrow L_k(i) = \sum_{j=1}^p (f_{j/i} - q_j) a_{jk} = \sum_{j=1}^p f_{j/i} a_{jk} + \sum_{j=1}^p q_j a_{jk}$$

\mathbf{L}_r est le tableau des coordonnées factorielles des lignes sur les axes principaux. $L_k(i)$ est la coordonnée de l'individu i sur l'axe k . a_{jk} est la composante de rang j de l'axe k . A une constante près ($\sum_{j=1}^p q_j a_{jk}$) le point i est positionné sur l'axe k à la moyenne des composantes ($\sum_{j=1}^p f_{j/i} a_{jk}$) du vecteur. En dimension 3, l'opération donne *exactement* la représentation triangulaire.

Exemple dans la carte Europe de la pile Data de Ade-4 (Collectif 1989) :

	1978			1986		
	Primaire	Secondaire	Tertiaire	Primaire	Secondaire	Tertiaire
Belgique	32	359	609	28	291	681
Danemark	79	319	602	59	282	659
Espagne	206	372	422	161	320	519
France	92	368	540	73	313	614
Grèce	320	297	383	285	281	434
Irlande	206	320	474	157	287	556
Italie	155	381	464	109	331	560
Luxembourg	62	392	546	40	330	630
Pays-Bas	54	330	616	49	255	696
Portugal	313	348	339	217	348	435
Allemagne	61	444	495	53	409	538
Royaume-Uni	28	390	582	25	309	666

Le premier tableau passé en pourcentage donne :

```

> euro78
      pri  sec  ter
Belgique 0.032 0.359 0.609
Danemark 0.079 0.319 0.602
Espagne  0.206 0.372 0.422
France   0.092 0.368 0.540
...
Pays-Bas 0.054 0.330 0.616
Portugal 0.313 0.348 0.339
Allemagne 0.061 0.444 0.495
Royaume-Uni 0.028 0.390 0.582

> pr0 <- princomp(euro78)
    
```

```

> pr0
Call:
princomp(x = euro78)

Standard deviations:
  Comp.1   Comp.2   Comp.3
1.318e-01 4.453e-02 4.715e-10

 3 variables and 12 observations.
> names(pr0)
[1] "sdev"      "loadings"  "center"    "scale"     "n.obs"     "scores"    "call"
> pr0$loadings
  Comp.1   Comp.2   Comp.3
pri -0.7540 -0.3133 0.5774
sec 0.1057 0.8096 0.5774
ter 0.6483 -0.4963 0.5774 Donner la valeur exacte pour 0.5774

```

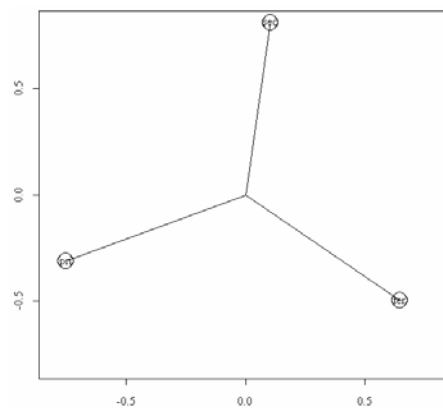
Les "loadings" sont les axes principaux.

```

> t(pr0$loadings)%*%pr0$loadings
      Comp.1   Comp.2   Comp.3
Comp.1 1.000e+00 -5.573e-17 1.975e-16
Comp.2 -5.573e-17 1.000e+00 -7.890e-17
Comp.3 1.975e-16 -7.890e-17 1.000e+00

> plot(a1,a2,cex=3,xlim=c(-0.8,0.8),ylim=c(-0.8,0.8))
> segments(0,0,a1,a2)
> text(a1,a2,names(euro78))

```

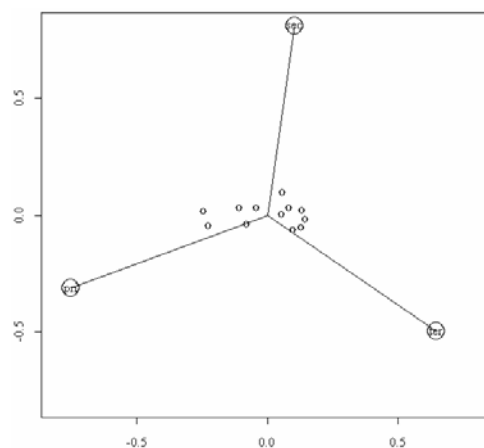


Le triangle est équilatéral. Les axes principaux sont orthogonaux à $\mathbf{1}_3$. Ils définissent le plan parallèle au plan $x + y + z = 1$ et passant par l'origine. Les composantes des axes sont les produits scalaires avec les vecteurs de la base canonique et on voit la projection des vecteurs de la base canonique sur le plan des axes principaux.

```

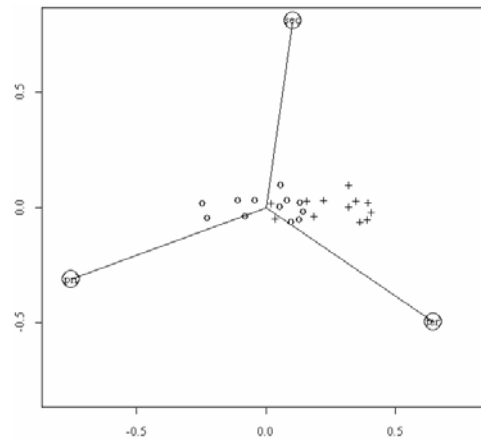
> points(pr0$scores[,1:2])

```



Cette figure est un biplot (graphe double) superposant la projection des points du nuage des lignes du tableau centré et de la base canonique de l'espace. Il est ici logique de représenter les points avant le centrage.

```
> points(as.matrix(euro78) %*% pr0$loadings[, 1:2], pch="+")
```

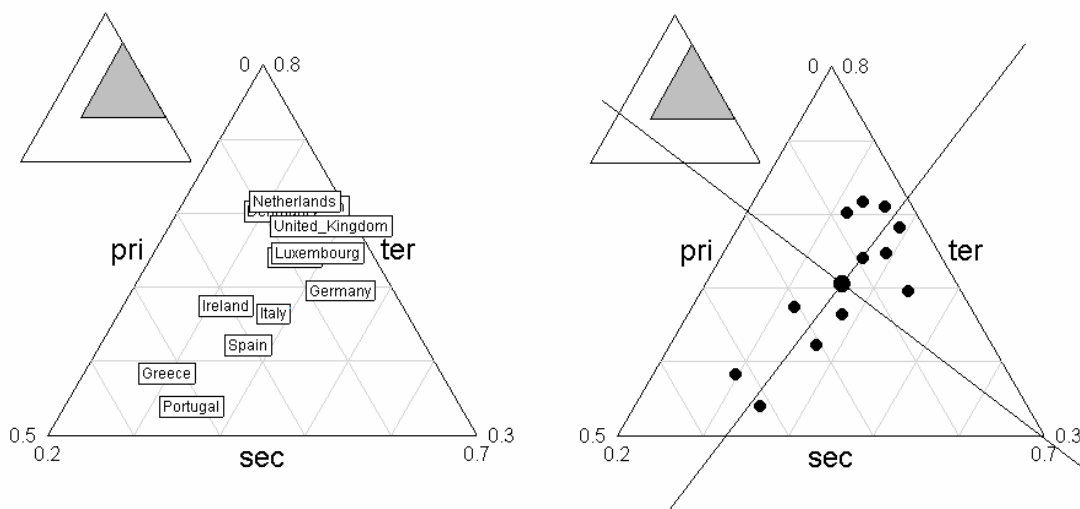


Les deux représentations sont décalées de :

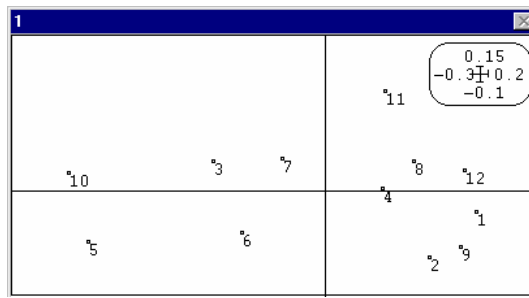
```
> pr0$center
  pri  sec  ter
0.134 0.360 0.506
> pr0$center %*% pr0$loadings[, 1:2] ( \sum_{j=1}^p q_j a_{jk} )
      Comp.1   Comp.2
[1,] 0.2651 -0.001656
```

En dimension 3 (3 catégories, les données sont dans un plan), la représentation triangulaire s'impose. Un exercice tout à fait excentrique : placer sur ce dessin les axes principaux.

```
> data(euro123)
> triangle.plot (euro123$in78, row.names(euro123$in78), clab=0.75)
> triangle.plot (euro123$in78, addaxes=T, cpo=2)
```

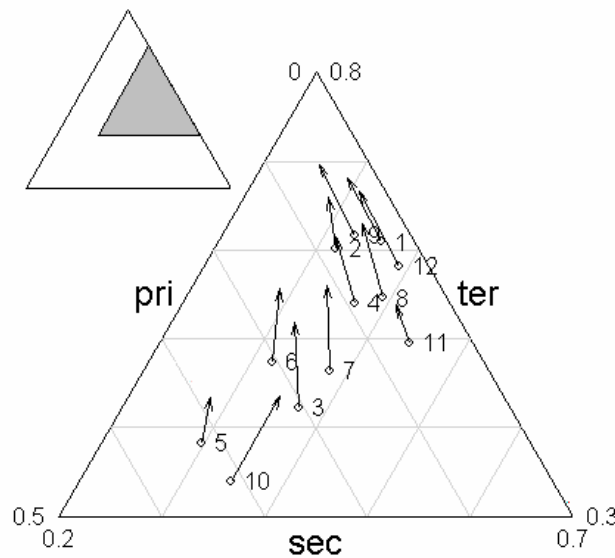


On fait ici le contraire du standard : projeter sur le plan des données les axes principaux au lieu de projeter les données sur les axes principaux. L'avantage est de "voir" ce que l'on fait dans une carte factorielle :



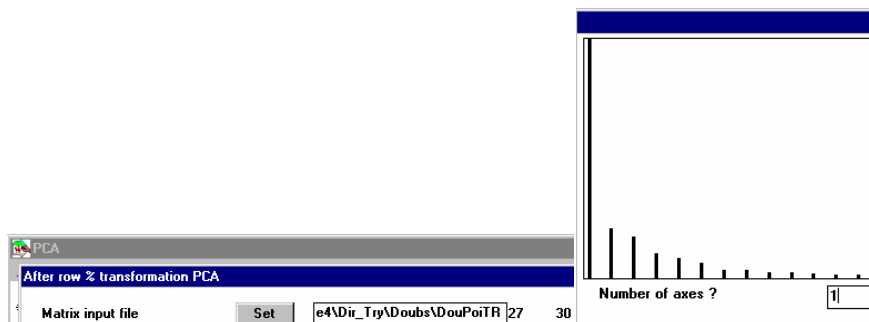
On retrouvera l'exemple dans la problématique *K*-tableaux. Question de fond : typologie d'évolutions ou évolution d'une typologie ?

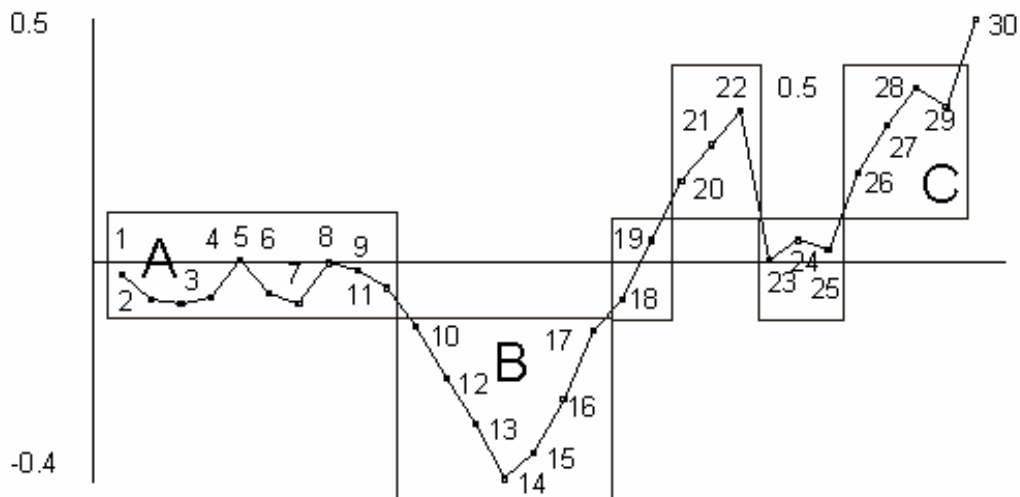
```
> triangle.biplot(euro123$in78, euro123$in86)
```



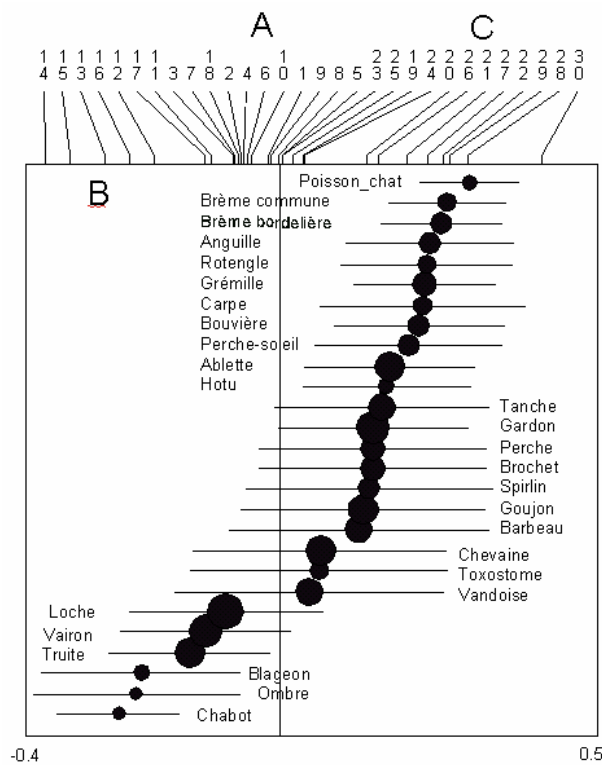
Les méthodes développées autour de ces questions sont nombreuses. On retiendra pour l'instant que la représentation triangulaire s'étend en dimension quelconque par le biais des relations d'averaging (lignes à la moyenne des catégories colonnes).

Exemple : carte Doubs de la pile Data du logiciel ADE-4 (Verneaux 1973). On considère les profils de distribution de 27 espèces dans 30 stations, on garde un axe.



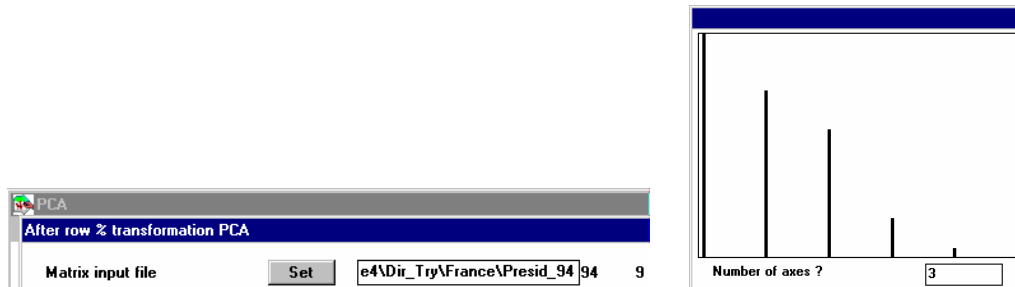


Scores des catégories obtenus par les composantes du premier axe principal représentées dans l'espace (ordre d'apparition sur le gradient amont-aval).

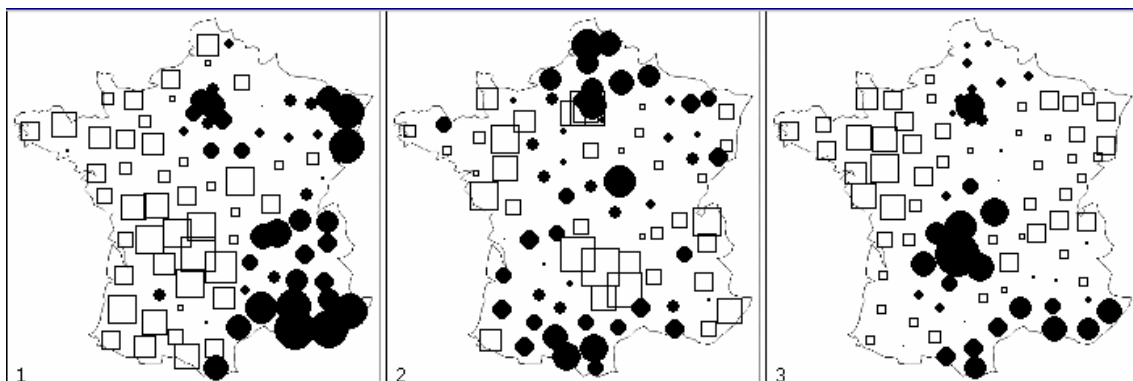
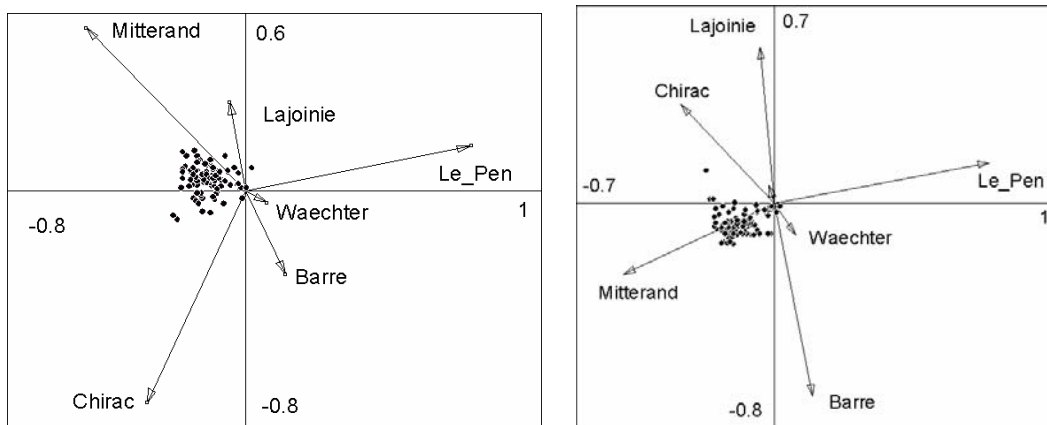


Scores des catégories obtenus par les composantes du premier axe principal. Scores des profils (espèces) représentés par leur moyenne (cercle) et leur variance (un écart-type) rangés par moyenne croissante. Le 0 donne la position de l'espèce indifférente (profil uniforme). Le score des catégorie maximise la variance des moyennes par espèce.

Exemple : carte France de la pile Data du logiciel ADE-4. On considère les profils de distribution de 94 départements pour l'élection présidentielle (Premier tour, avril 1988), on garde trois axes.



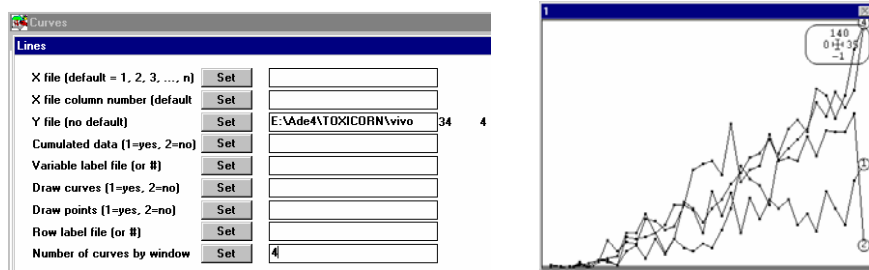
Les axes principaux sont des vecteurs à 9 composantes (candidats). Ils permettent de représenter les départements à la moyenne des distributions de fréquences (averaging). A gauche plan des axes 1 et 2, à droite plan des axes 1 et 3. L'expression d'un maximum de la variabilité ne cache pas la présence d'un compromis.



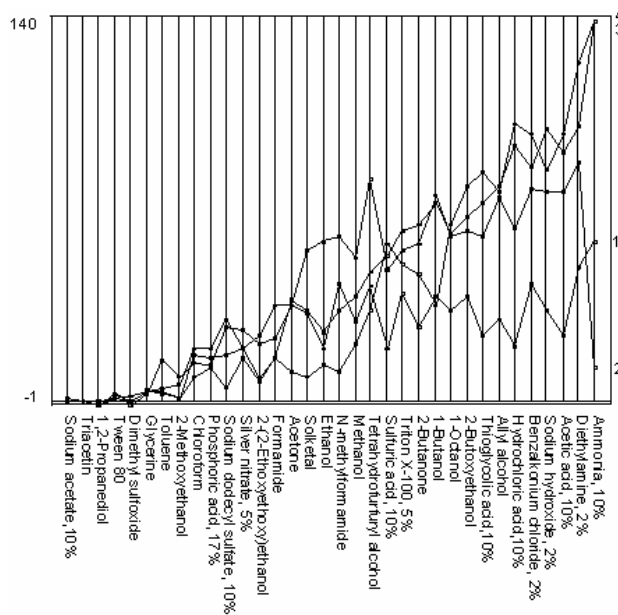
Cartographie des coordonnées centrées (expression des composantes régionales de l'opinion).

2.6 Tableaux homogènes

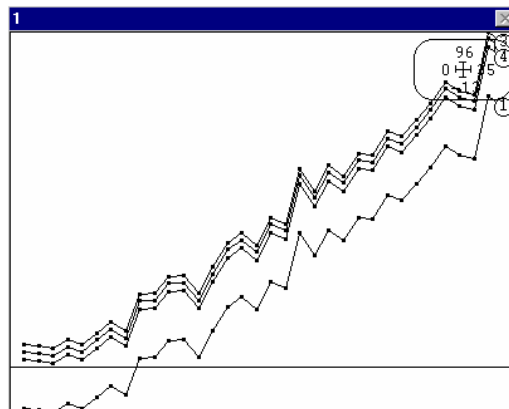
Ils comportent dans chaque cellule du tableau la mesure d'une même variable. L'ACP peut être vue comme un moyen de modéliser ces données. La question se comprend bien en toxicologie. Utiliser la carte Toxicorn (Jacobs and Martens 1990) :



Compléter avec Graph1D: Labels et assembler :

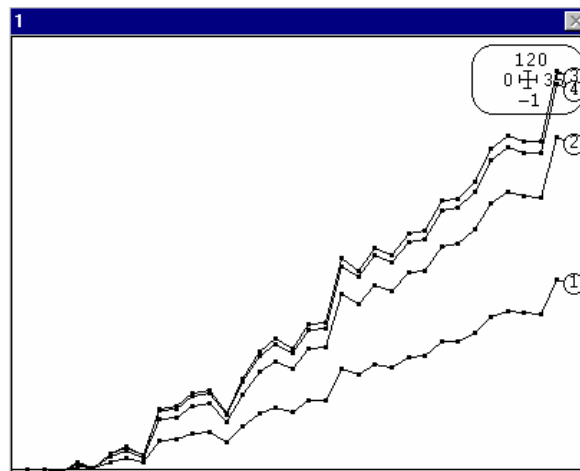


Chaque produit a un niveau de toxicité et chaque durée d'exposition (4, 24, 48 et 72 heures noté 1, 2, 3 et 4) a un effet, les deux manifestement se multiplient l'un l'autre. En effet un modèle additif donne (HTA: Double centring additive) :

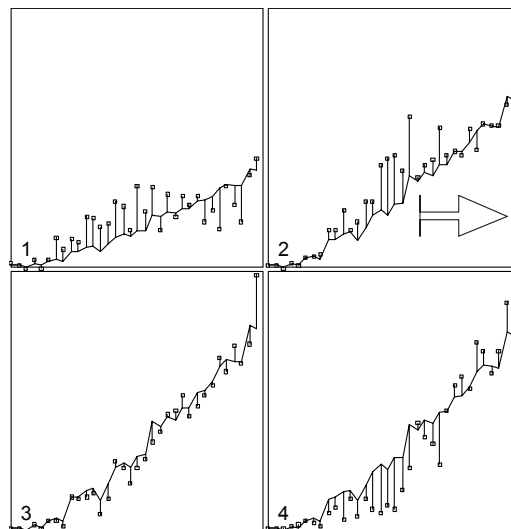


On veut trouver un score des lignes α_i et un score des colonnes β_j pour que l'écart au modèle $\sum_{i=1}^n \sum_{j=1}^p (x_{ij} - \alpha_i \beta_j)^2$ soit minimum. Comme le modèle $\frac{\alpha_i}{k} k \beta_j$ a les mêmes performances que le modèle $\alpha_i \beta_j$, on impose la recherche du modèle $k \alpha_i \beta_j$ avec les contraintes $\sum_{j=1}^p \beta_j^2 = 1$ et $\frac{1}{n} \sum_{i=1}^n \alpha_i^2 = 1$. k est alors un facteur d'échelle libre et la

solution est unique. La solution provient directement de l'ACP sans modification préalable du tableau brut (HTA: Double centring multiplicative) :



On peut alors superposer données et modèle :



Une des expériences donne un résultat qui nuit fortement à la modélisation de l'ensemble. A retenir : l'équation fondamentale est :

$$\text{Données} = \text{Evidences} + \text{Structures} + \text{Erreur}$$

Les évidences sont les propriétés des données immédiatement accessibles : certaines espèces sont rares et d'autres sont abondantes. Les individus les plus grands ont de plus grandes oreilles. Si on laisse les évidences dans les données, on dira que l'analyse enfonce des portes ouvertes. Les structures sont des propriétés associées aux comparaisons multiples entre éléments. Une technique comme l'utilisation du premier facteur de la décomposition en valeurs singulières participe selon les cas de l'élimination d'une évidence ou de la définition d'un fait peu apparent.

3 Modèles probabilistes

3.1 Loi normale bivariée

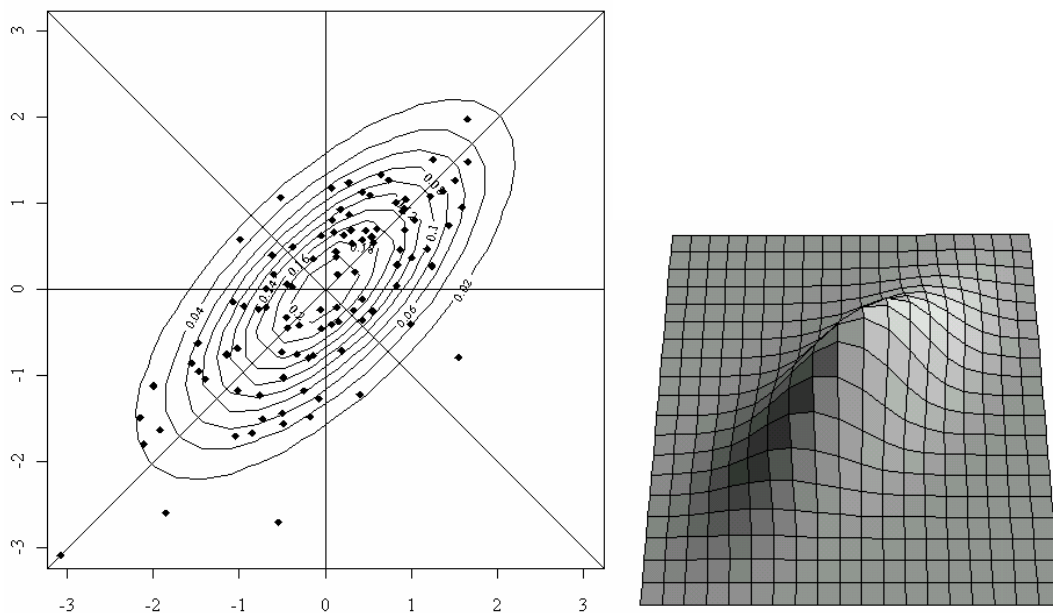
```

echa<-rmvnorm(100,mu=c(0,0),sigma=matrix(c(1,0.7,0.7,1),2,2))
echa<-as.data.frame(echa)
names(echa)<-c("x","y")
attach(echa)
par(mai=c(1,1,1,1))
plot(x,y,xlim=c(-3,3),ylim=c(-3,3),type="n")

xg<-seq(-3,3,le=20)
yg<-seq(-3,3,le=20)
xyg<-expand.grid(x=xg,y=yg)
xyg[1:5,]
z<-dmvnorm(xyg,mu=c(0,0),sigma=matrix(c(1,0.7,0.7,1),2,2))
z<-matrix(z,nrow=20)

contour(xg,yg,z)
points(x,y,pch=18)
abline(v=0)
abline(h=0)
abline(0,1)
abline(0,-1)
> persp(xg,yg,z,box=F,theta=0,phi=70,expand = 0.5, ltheta = 120, shade = 0.75)

```



On a créé un échantillon aléatoire simple d'une loi normale multivariée de moyenne (0,0) et de matrice de variances-covariances :

$$\mathbf{C} = \begin{bmatrix} 1 & 0.7 \\ 0.7 & 1 \end{bmatrix}$$

La densité s'écrit : $g(x, y) = \frac{1}{\sqrt{2\pi^2 \det(\mathbf{C})}} \exp\left(-\frac{1}{2} \begin{bmatrix} x & y \end{bmatrix} \mathbf{C}^{-1} \begin{bmatrix} x \\ y \end{bmatrix}\right)$

```

> C <- matrix(c(1,0.7,0.7,1),2,2)
> C
      [,1] [,2]
[1,]  1.0  0.7
[2,]  0.7  1.0
> Cm1 <- solve(C,diag(1,2))%*%C

```

```
> Cm1%%C
      [,1] [,2]
[1,]  1.0  0.7
[2,]  0.7  1.0
> det(C)
[1] 0.51

> (1/sqrt(2*pi*pi*0.51)*exp(-1*c(1,2)%%Cm1%%c(1,2)/2))
      [,1]
[1,] 0.02587
> dmnorm(c(1,2),c(0,0),C)
[1] 0.02578
```

On estime le vecteur des moyennes (vraies valeurs 0 et 0) :

```
> apply(echa,2,mean)
      x      y
-0.02606 -0.07586
```

On estime la matrice des variances-covariances (vraie valeur C) :

```
> var(echa)
      x      y
x 0.9016 0.7003
y 0.7003 1.0303
```

On estime les vecteurs propres normés de C. Les vraies valeurs sont :

```
> eigen(C)
$values
[1] 1.7 0.3

$vectors
      [,1] [,2]
[1,] 0.7071 0.7071
[2,] 0.7071 -0.7071
```

Les valeurs estimées sont :

```
> eigen(var(echa))
$values
[1] 1.6692 0.2627

$vectors
      y      x
x 0.6740 0.7387
y 0.7387 -0.6740
```

L'ACP centrée est la méthode d'estimation de ces axes principaux :

```
> prcomp(echa)
Standard deviations:
[1] 1.2920 0.5126

Rotation:
      PC1      PC2
x 0.6740 -0.7387
y 0.7387 0.6740

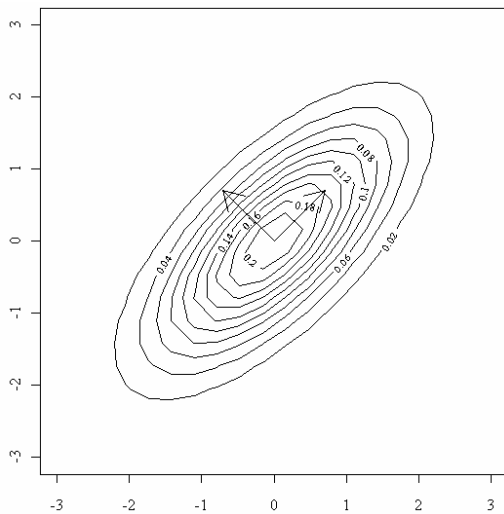
> pca0$sdev
Comp. 1 Comp. 2
 1.31 0.5514
> pca0$loadings
Comp. 1 Comp. 2
x -0.744 0.668
y -0.668 -0.744
```

La population :

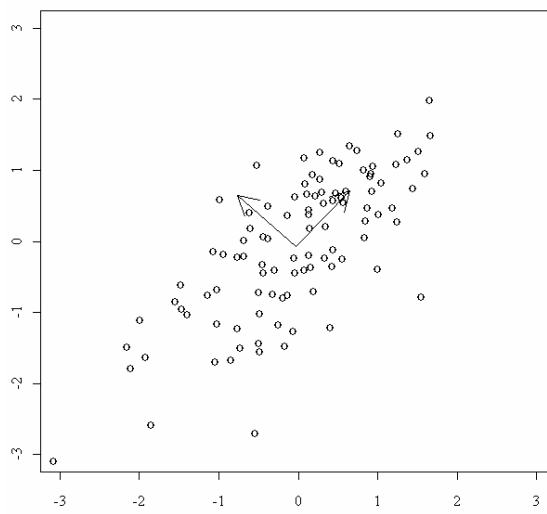
```
> par(mai=c(1,1,1,1))
> plot(x,y,xlim=c(-3,3),ylim=c(-3,3),type="n")
> contour(xg,yg,z)
> arrows(0,0,1/sqrt(2),1/sqrt(2))
> arrows(0,0,-1/sqrt(2),1/sqrt(2))
```

L'échantillon :

```
> pca0 <- prcomp(echa)
> plot(echa,xlim=c(-3,3),ylim=c(-3,3))
> moy <- apply(echa,2,mean)
> arrows(moy[1],moy[2],moy[1]+pca0$rotation[1,1],moy[1]+pca0$rotation[2,1])
> arrows(moy[1],moy[2],moy[1]+pca0$rotation[1,2],moy[1]+pca0$rotation[2,2])
```



Population



Echantillon

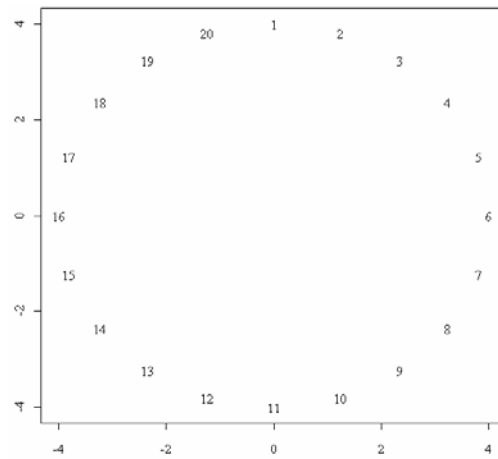
La même procédure (diagonalisation d'une matrice de covariances) prend une toute autre signification. C'est l'origine de conflits d'écoles.

3.2 Erreur sphérique

Une autre manière (Besse et al. 1986) de voir l'ACP peut se présenter par l'exemple suivant.

```
> x0 <- 4 * sin(seq(0, 1.9 * pi, le = 20))
> y0 <- 4 * cos(seq(0, 1.9 * pi, le = 20))
> plot(x0, y0, type = "n")
> text(x0, y0)
```

20 points sont sur un cercle.



```

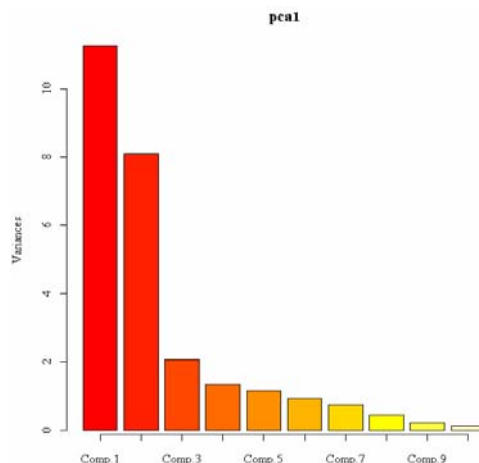
> library(MASS)
> error <- mvrnorm(20, mu=rep(0,10), Sigma = diag(10))

> error
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]
[1,] -1.05375 -0.52481 -0.28698  0.3521068  1.4085 -1.52848 -0.5539 -1.17653
[2,]  0.23963  0.74546  1.13050 -1.5022218 -0.4505  0.41332 -0.7328  0.61716
[3,]  1.13614 -2.15204  0.41437  0.0800738 -0.9318  0.05744 -0.7718 -0.06062
[4,] -0.05693 -0.06618 -2.30162 -2.2500917  0.0465 -0.04287 -0.6095 -0.35151
[5,] ...
> error[, 1] <- error[, 1] + x0
> error[, 2] <- error[, 2] + y0
> error[1:4, ]
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]
> error[, 1] <- error[, 1] + x0
> error[, 2] <- error[, 2] + y0
> error[1:4, ]
      [,1]      [,2]      [,3]      [,4]      [,5]      [,6]      [,7]      [,8]      [,9]
[1,] -1.054  3.475 -0.2870  0.35211  1.4085 -1.52848 -0.5539 -1.17653 -1.4154
[2,]  1.476  4.550  1.1305 -1.50222 -0.4505  0.41332 -0.7328  0.61716  1.0869
[3,]  3.487  1.084  0.4144  0.08007 -0.9318  0.05744 -0.7718 -0.06062  0.2953
[4,]  3.179  2.285 -2.3016 -2.25009  0.0465 -0.04287 -0.6095 -0.35151  0.7131
...
> pca1 <- princomp(error)
> pca1
Call:
princomp(x = error)

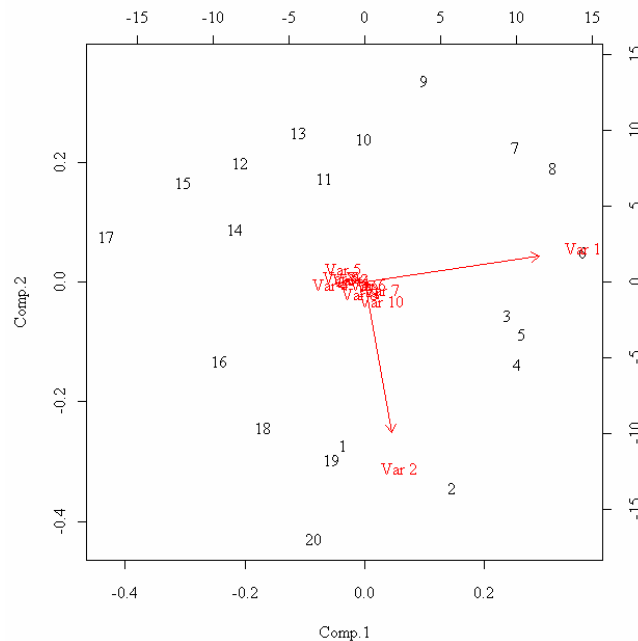
Standard deviations:
  Comp.1  Comp.2  Comp.3  Comp.4  Comp.5  Comp.6  Comp.7  Comp.8  Comp.9
Comp.10
  3.3523  2.8438  1.4376  1.1588  1.0798  0.9698  0.8639  0.6635  0.4648
  0.3705

10 variables and 20 observations.
>plot(pca1)

```



```
>biplot(pca1)
```



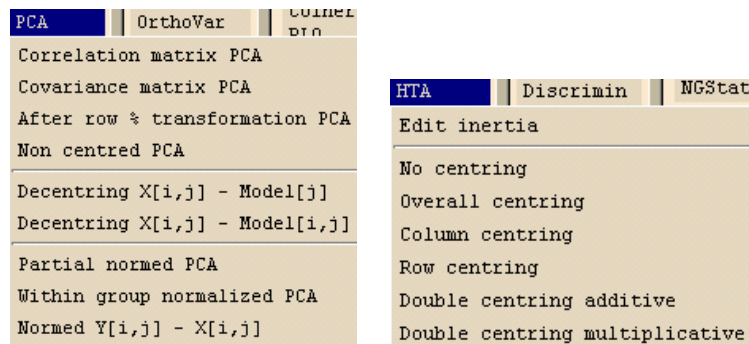
La ligne i du tableau est un échantillon aléatoire d'une loi normale de moyenne μ_i et de variance $\sigma^2 \mathbf{I}_p$. Le nuage des moyennes est un modèle du nuage observé, chaque point se trouvant autour de sa position avec une erreur gaussienne de même variance dans toutes les directions. Les moyennes se trouvent dans un sous-espace de dimension donnée (ici un plan de dimension 2). Il convient d'estimer la dimension du sous-espace, les positions des points modèles dans ce sous-espace et le paramètre σ^2 . On voit sur le graphe des valeurs propres l'estimation de la dimension, sur le biplot l'estimation du plan (ici le plan des deux premières variables) et sur ce plan l'estimation de la position des points (20 points sur un cercle). La métrique identité utilisée est alors un modèle de l'erreur (écart entre le point théorique et sa réalisation). Changer de métrique, c'est changer le modèle de l'erreur. De multiples développements sont associés à ce point de vue.

En conclusion, on peut noter ceci. L'utilisateur dit "le tableau est soumis à une analyse en composantes principales" veut dire que le tableau a été envoyé dans une procédure :

```
> prcomp
function (x, retx = TRUE, center = TRUE, scale. = FALSE, tol = NULL)
{
  x <- as.matrix(x)
  x <- scale(x, center = center, scale = scale.)

  s <- svd(x, nu = 0)
  if (!is.null(tol)) {
    rank <- sum(s$d > (s$d[1] * tol))
    if (rank < ncol(x))
      s$v <- s$v[, 1:rank, drop = FALSE]
  }
  s$d <- s$d/sqrt(max(1, nrow(x) - 1))
  dimnames(s$v) <- list(colnames(x), paste("PC", seq(len = ncol(s$v)),
    sep = ""))
  r <- list(sdev = s$d, rotation = s$v)
  if (retx)
    r$x <- x %*% s$v
  class(r) <- "prcomp"
  r
}
```

La procédure donne des valeurs propres (ou des valeurs singulières), des vecteurs et des scores. Le modèle conceptuel qui justifie cette procédure est multiple (algébrique : changement de base, numérique : modèle multiplicatif du tableau, probabiliste : estimation de modèles, géométrique : projections d'inertie maximale). Dans le résumé de l'opération par le triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$, on peut faire varier \mathbf{X} (centrage, non centrage, décentrage, normalisation), on peut faire varier \mathbf{Q} et \mathbf{D} (pondérations, normes non diagonales) :



On peut enfin exprimer les résultats de multiples façons. Dans la suite, nous passerons en revue des "réglages" de ce modèle général utiles dans des circonstances variées.

4 La classe d'objets dudi

Le schéma de dualité définit le contenu précis des objets dans la librairie ade4. Un schéma est constitué de trois éléments donnant un triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$

- \mathbf{X} un tableau à n lignes et p colonnes, considéré comme p points de \mathbb{R}^n (vecteurs colonnes) ou n points de \mathbb{R}^p (vecteurs lignes) ;
- \mathbf{Q} une matrice diagonale de dimension $p \times p$ contenant les poids des p colonnes de \mathbf{X} et utilisée comme produit scalaire dans \mathbb{R}^p ;
- \mathbf{D} une matrice diagonale de dimension $n \times n$ contenant les poids des n lignes considérés comme et utilisée comme produit scalaire dans \mathbb{R}^n .

Un objet de la classe dudi est une liste qui comprend un triplet de départ :

- tab un data frame avec n lignes et p colonnes
- cw les poids des colonnes, vecteur à p composantes
- lw les poids des lignes, vecteurs à n composantes

puis les dérivés de la diagonalisation (ou de la décomposition en valeurs singulières) du schéma :

- eig les valeurs propres dans u vecteur d'au plus min (n,p) composantes
- nf f le nombre d'axes conservés (imposé ou introduit après jugement sur le graphe des valeurs propres)
- c1 les axes principaux placés en colonnes dans un data frame avec p lignes et f colonnes
- l1 les composantes principales placées en colonnes dans un data frame avec n lignes et f colonnes

co les coordonnées des colonnes placées dans un data frame avec p lignes et f colonnes (chaque ligne de ce tableau désigne une colonne du tableau de départ, chaque colonne est une composante principale, chaque valeur est une coordonnée de la projection de la colonne du tableau de départ sur une composante principale)

li les coordonnées des lignes placées en colonnes dans un data frame avec n lignes et f colonnes (chaque ligne de ce tableau désigne une ligne du tableau de départ, chaque colonne est un axe principal, chaque valeur est une coordonnée de la projection de la ligne du tableau de départ sur un axe principal)

Par exemple, on reconnaîtra :

```
data(euro123)
unclass(dudi.pca(euro123$in78,scal=FALSE,scan=F))

$tab
      pri    sec    ter
Belgium -0.102 -0.001  0.103
Denmark -0.055 -0.041  0.096
Spain    0.072  0.012 -0.084
France  -0.042  0.008  0.034
Greece   0.186 -0.063 -0.123
Ireland  0.072 -0.040 -0.032
Italy    0.021  0.021 -0.042
Luxembourg -0.072 0.032  0.040
Netherlands -0.080 -0.030 0.110
Portugal  0.179 -0.012 -0.167
Germany  -0.073  0.084 -0.011
United_Kingdom -0.106 0.030  0.076 Tableau centré mais non normé par colonne

$scw
[1] 1 1 1 Poids des colonnes unitaires

$lw
[1] 0.08333 0.08333 0.08333 0.08333 0.08333 0.08333 0.08333 0.08333 0.08333
[10] 0.08333 0.08333 0.08333 Poids des lignes uniformes

$eig
[1] 0.017363 0.001983 Deux valeurs propres, la troisième est nulle

$rank
[1] 2 Le rang est 2

$nf
[1] 2 Le nombre d'axes conservés est 2

$c1
      CS1    CS2
pri  0.7540 -0.3133
sec -0.1057  0.8096
ter -0.6483 -0.4963 composantes des 2 axes principaux dans l'espace des 3 variables, CS veut dire Column Score, appelés aussi loadings ou poids des variables

$l1
      RS1    RS2
Belgium -1.0896 -0.4486
Denmark -0.7542 -1.4285
Spain    0.8157  0.6479
France  -0.4140  0.0620
Greece   1.7200 -1.0832
Ireland  0.6015 -0.8772
Italy    0.3100  0.7022
Luxembourg -0.6345  0.6426
Netherlands -0.9749 -1.2086
Portugal  1.8555  0.3838
Germany  -0.4309  2.1635
United_Kingdom -1.0045  0.4442 composantes des 2 composantes principales dans l'espace des individus, RS veut Row Score, appelé aussi coordonnée normalisée ou coordonnée de norme 1
```

```

$co
      Comp1   Comp2
pri  0.09935 -0.01395
sec -0.01392  0.03605
ter -0.08543 -0.02210 coordonnées des variables sur les deux composantes
principales

$li
      Axis1   Axis2
Belgium   -0.14358 -0.019974
Denmark   -0.09938 -0.063610
Spain     0.10748  0.028849
France    -0.05456  0.002761
Greece    0.22664 -0.048234
Ireland   0.07926 -0.039061
Italy     0.04084  0.031268
Luxembourg -0.08360  0.028613
Netherlands -0.12847 -0.053820
Portugal  0.24450  0.017089
Germany   -0.05679  0.096340
United_Kingdom -0.13237  0.019779 coordonnées des individus sur les deux axes
principaux, appelées aussi scores de variance lambda, ou score normé à la
valeur propre, ou coordonnée de variance lambda

$call      l'ordre de création de l'objet
dudi.pca(df = euro123$in78, scale = FALSE, scannf = F)

$cent      le vecteur de centrage, ici les moyennes (ACP centrée)
      pri   sec   ter
0.134 0.360 0.506

$norm      le vecteur de normalisation, ici unité (ACP non normée)
[1] 1 1 1

```

La fonction print.dudi résume avec :

```

dudi.pca(euro123$in78,scal=FALSE,scan=F)
Duality diagramm
class: pca dudi
$call: dudi.pca(df = euro123$in78, scale = FALSE, scannf = F)

$nf: 2 axis-components saved
$rank: 2
eigen values: 0.01736 0.001983
  vector length mode  content
1 $cw      3      numeric column weights
2 $lw     12      numeric row weights
3 $eig     2      numeric eigen values

  data.frame nrow ncol content
1 $tab     12    3    modified array
2 $li      12    2    row coordinates
3 $ll      12    2    row normed scores
4 $co      3     2    column coordinates
5 $cl      3     2    column normed scores
other elements: cent norm

```

Toutes les analyses d'un tableau donne un objet de ce type. La connaissance du modèle permet d'utiliser les variantes sans problèmes. Les plus utiles sont :

Fonction	Analyse	Référence
dudi.pca	Principal component Analysis	1
dudi.coa	Correspondence Analysis	2
dudi.acm	Multiple Correspondence Analysis	3
dudi.fca	Fuzzy Correspondence Analysis	4
dudi.mix	Mixture of numeric and factors	5
dudi.nsc	Non Symetric Correspondence Analysis	6
dudi.dec	Decentred Correspondence Analysis	7

Les fonctions dudi. 1—équivalent de prcomp/princomp, 2—Greenacre (1984), 3—Tenenhaus and Young (1985), 4—Chevenet et al.(1994), 5—Hill and Smith (1976), Kiers (1994) 6—Kroonenberg and Lombardo (1999), 7—Dolédec et al. (1995).

5 Analyses des correspondances multiples

5.1 Définition

L'analyse des correspondances multiples est destinée à la synthèse d'un tableau de variables qualitatives vu comme la juxtaposition de paquets d'indicatrices des classes (tableau disjonctif complet) :

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	0	1	0	1	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1
2	0	1	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1
3	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	1
4	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1
5	1	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0
6	1	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	1	0	1	0
7	0	0	1	0	1	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1
8	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	1	0
9	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
10	0	1	0	1	0	0	0	1	0	0	1	0	1	0	0	1	0	0	1	0	0
11	0	1	0	0	1	0	1	0	0	1	0	0	1	0	0	0	1	0	0	1	0
12	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	1	0	0	1	0
13	0	1	0	1	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	1	0
14	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	1	0
15	1	0	0	0	1	0	1	0	0	0	1	0	1	0	0	1	0	0	1	0	0
16	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0
17	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0
18	0	1	0	0	1	0	0	1	0	0	0	1	0	1	0	1	0	0	0	1	0
19	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	1
20	0	1	0	0	1	0	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0
21	1	0	0	0	1	0	1	0	0	0	1	0	1	0	0	1	0	0	1	0	0
22	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0
23	0	1	0	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0

Chaque tableau fait une typologie simple des individus (partition). L'objectif de l'ACM est de donner un point de vue coordonné sur toutes ces partitions. On note :

$$X = [X_1, X_2, \dots, X_v]$$

le tableau disjonctif complet. v est le nombre de variables qualitatives. n est le nombre d'individus (lignes) et m_1, m_2, \dots, m_v sont les nombres de modalités de chaque variable. Le nombre de modalités total est $m = m_1 + m_2 + \dots + m_v$. On peut considérer que l'ACM fait partie des analyses à un tableau (X).

On suppose que chaque ligne porte le poids $1/n$ (comme dans une ACP) ou le poids p_i ($\sum_{i=1}^n p_i = 1$). \mathbf{D} est comme d'habitude la diagonale des poids des lignes (matrice $n \times n$). On compte pour chaque modalité le nombre d'individus l'utilisant (le nombre de porteurs) en faisant :

$$\mathbf{X}^t \mathbf{1}_n = [n_{11}, n_{12}, \dots, n_{1m_1}, n_{21}, n_{22}, \dots, n_{2m_2}, \dots, n_{v1}, n_{v2}, \dots, n_{vm_v}]$$

Il vient simplement que :

$$n_{11} + n_{12} + \dots + n_{1m_1} = n_{21} + n_{22} + \dots + n_{2m_2} = \dots = n_{v1} + n_{v2} + \dots + n_{vm_v} = n$$

Au lieu de compter les porteurs d'une modalité, on peut aussi sommer les poids des porteurs de cette modalité. Pour une pondération uniforme, le poids d'une modalité est la fréquence ($f_{ij} = n_{ij}/n$) du nombre de porteurs et on a :

$$f_{11} + f_{12} + \dots + f_{1m_1} = f_{21} + f_{22} + \dots + f_{2m_2} = \dots = f_{v1} + f_{v2} + \dots + f_{vm_v} = 1$$

mais en général, les poids des modalités sont dans :

$$[f_{11}, f_{12}, \dots, f_{1m_1}, f_{21}, f_{22}, \dots, f_{2m_2}, \dots, f_{v1}, f_{v2}, \dots, f_{vm_v}] = \mathbf{X}^t \mathbf{D} \mathbf{1}_n$$

Les poids des modalités sont écrits sur la diagonale d'une matrice diagonale :

$$\mathbf{D}_m = \text{Diag}(f_{11}, f_{12}, \dots, f_{1m_1}, f_{21}, f_{22}, \dots, f_{2m_2}, \dots, f_{v1}, f_{v2}, \dots, f_{vm_v})$$

La somme de ces éléments vaut donc v . \mathbf{D}_m est une matrice à m lignes et m colonnes. Sa trace vaut v . On note $\mathbf{1}_{nm}$ la matrice à n lignes et m colonnes ne contenant que des 1. L'ACM (synthèse incontournable (Tenenhaus and Young 1985) est l'analyse du schéma :

$$\begin{array}{ccc} & \frac{1}{v} \mathbf{D}_m & \\ \boxed{m} & \rightarrow & \boxed{m} \\ \mathbf{Y}^t \uparrow & & \downarrow \mathbf{Y} = \mathbf{X} \mathbf{D}_m^{-1} - \mathbf{1}_{nm} \\ \boxed{n} & \leftarrow & \boxed{n} \\ & \mathbf{D} & \end{array}$$

Le paramétrage de ce schéma a été choisi pour ses bonnes propriétés. En effet, on considère les vecteurs indicateurs de variables dans l'ensemble des modalités :

$$[\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_v] = \begin{bmatrix} 1 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

On a toujours $\mathbf{X}\mathbf{w}_k = \mathbf{1}_n$ (il y a une modalité présente et une seule pour chaque individu et chaque variable) donc $(\mathbf{X}\mathbf{D}_m^{-1} - \mathbf{1}_{nm})\frac{1}{v}\mathbf{D}_m\mathbf{w}_k = \mathbf{0}_n$ et les vecteurs \mathbf{w}_k sont propres pour 0. Il y a donc dans cette analyse au moins v valeurs propres nulles. Les axes principaux sont des vecteurs à m composantes formés de $m = m_1 + m_2 + \dots + m_v$ "morceaux". Chaque segment est un codage numérique des modalités de la variable correspondante. Par orthogonalité aux vecteurs \mathbf{w}_k , ces scores des modalités sont centrés (pour les poids des modalités) et $\frac{1}{v}\mathbf{D}_m$ -normés, donc de variance moyenne égale à 1. Ils maximisent successivement :

$$\left\| (\mathbf{X}\mathbf{D}_m^{-1} - \mathbf{1}_{nm})\frac{1}{v}\mathbf{D}_m\mathbf{a} \right\|_{\mathbf{D}}^2 = \left\| (\mathbf{X} - \mathbf{1}_{nm}\mathbf{D}_m)\frac{1}{v}\mathbf{a} \right\|_{\mathbf{D}}^2 = \left\| \frac{1}{v}\mathbf{X}\mathbf{a} \right\|_{\mathbf{D}}^2$$

Or le vecteur $\mathbf{1}_n$ est composante principale pour la valeur propre 0 ($(\mathbf{D}_m^{-1}\mathbf{X}^t - \mathbf{1}_{mn})\mathbf{D}\mathbf{1}_n = 0$) donc les autres composantes principales sont centrées et $\left\| \frac{1}{v}\mathbf{X}\mathbf{a} \right\|_{\mathbf{D}}^2$ est la variance du score des individus obtenu en attribuant la moyenne des scores des modalités portées.

Inversement, une composante principale est un score des individus de moyenne 0 et variance 1 qui maximise :

$$\left\| (\mathbf{D}_m^{-1}\mathbf{X}^t - \mathbf{1}_{mn})\mathbf{D}\mathbf{b} \right\|_{\frac{1}{v}\mathbf{D}_m}^2 = \left\| \mathbf{D}_m^{-1}\mathbf{X}^t\mathbf{D}\mathbf{b} \right\|_{\frac{1}{v}\mathbf{D}_m}^2$$

c'est-à-dire la moyenne de la variance du score des modalités obtenu en attribuant à chaque modalité la moyenne des scores des individus porteurs.

L'objectif fondamental de l'ACM est d'obtenir des scores numériques des individus qui maximisent les pourcentages de variance expliquée, en moyenne, pour toutes les variables qualitatives (rapport de corrélation).

L'ACM fait la synthèse des liens entre les variables qualitatives par l'examen des liens entre les variables qualitatives \mathbf{q}^j et un code numérique de synthèse appelée variable synthétique.

En ACP normée, on a trouvé une variable synthétique (la première composante principale, concrètement la première coordonnée normée des lignes) qui maximise :

$$\sum_{j=1}^p \text{cor}^2(\mathbf{x}^j, \mathbf{y})$$

En ACM, on a trouvé une variable synthétique (la première composante principale, concrètement la première coordonnée normée des lignes) qui maximise :

$$\frac{1}{v} \sum_{j=1}^v \eta^2(\mathbf{q}^j, \mathbf{y})$$

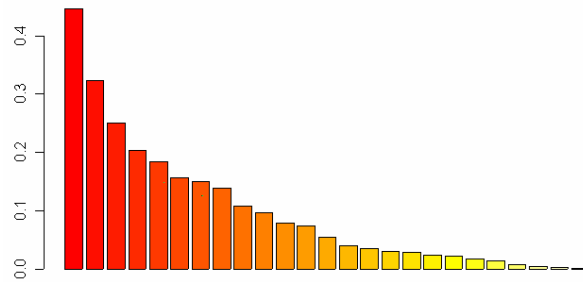
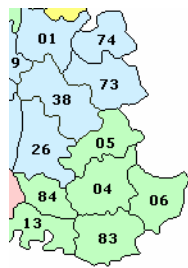
Cette manière de dépouiller l'ACM est due à Saporta (1975).

5.2 Ours

Les données ont un intérêt pédagogique. Elles ont été compilées par G. Erome(1989). Les Alpes sont découpées en 38 régions par l'Inventaire National Forestier. Pour chacune d'entre elle, on note :

- 1) **altit** l'importance de la tranche altitudinale habitée par l'ours (1/ moins de 50% de la surface entre 800 et 2000 m, 2/ entre 50 et 70%. 3/ plus de 70%) ;
- 2) **deniv** l'importance de la dénivelée moyenne par carrés de 50 km² (1/ moins de 700m, 2/ entre 700 et 900 m, 3/plus de 900 m) ;
- 3) **cloiso** le cloisonnement du massif (1/ une grande vallée ou une crête isole au moins un quart du massif, 2/ une grande vallée ou une crête isole moins d'un quart du massif est isolé, 3/ le massif ne présente pas de coupure) ;
- 4) **domain** l'importance du domaine forestier en contact avec le massif (1/ moins de 400 km², 2/ entre 400 et 1000 km², 3/ plus de 1000 km²) ;
- 5) **boise** le taux de boisement (1/ moins de 30%, 2/ entre 30 et 50%, 3/ plus de 50%) ;
- 6) **hetra** l'importance des hêtraies et forêts mixtes (1/ moins de 5%, 2/ entre 5 et 10%, 3/ plus de 10% du massif) ;
- 7) **favor** l'importance des forêts favorables, hêtraies, forêts mixtes, sapinières et pessières (1/ moins de 5%, 2/ entre 5 et 10%, 3/ plus de 10% du massif) ;
- 8) **inexp** l'importance des forêts inexploitées (1/ moins de 4%, 2/ entre 4 et 8%, 3/ plus de 8% de la surface totale) ;
- 9) **citat** la présence de l'ours avant sa disparition (1/ aucune citation depuis 1840, 2/ 1 à 3 citations avant 1900 et aucune après, 3/ 4 citations avant 1900 et aucune après, 4/ 4 citations ou plus avant 1900 et 1 citation ou plus entre 1900 et 1940).

10) **depart** Le département du district (**AHP**/Alpes-de-Haute-Provence
AM/Alpes-Maritimes **D**/Drôme **HP**/Hautes-Alpes **HS**/Haute-Savoie
I/Isère **S**/Savoie).

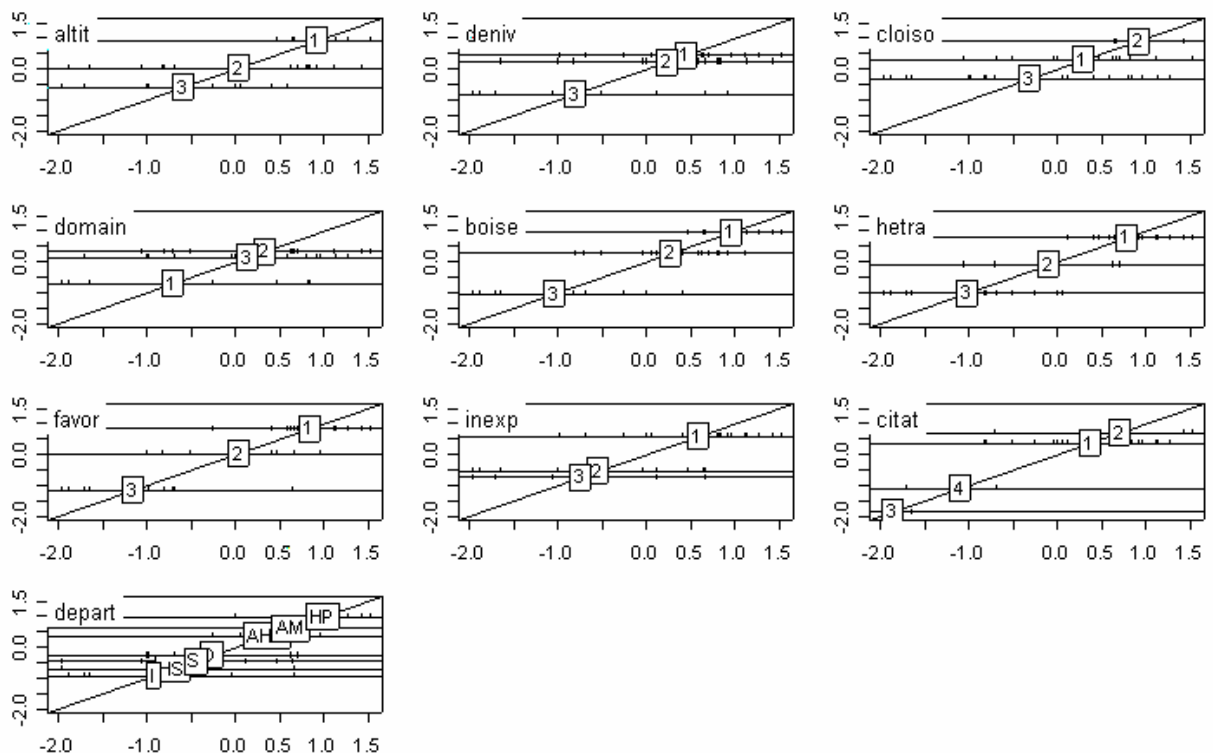


```
> acm1 <- dudi.acm(ours)
Select the number of axes: 2

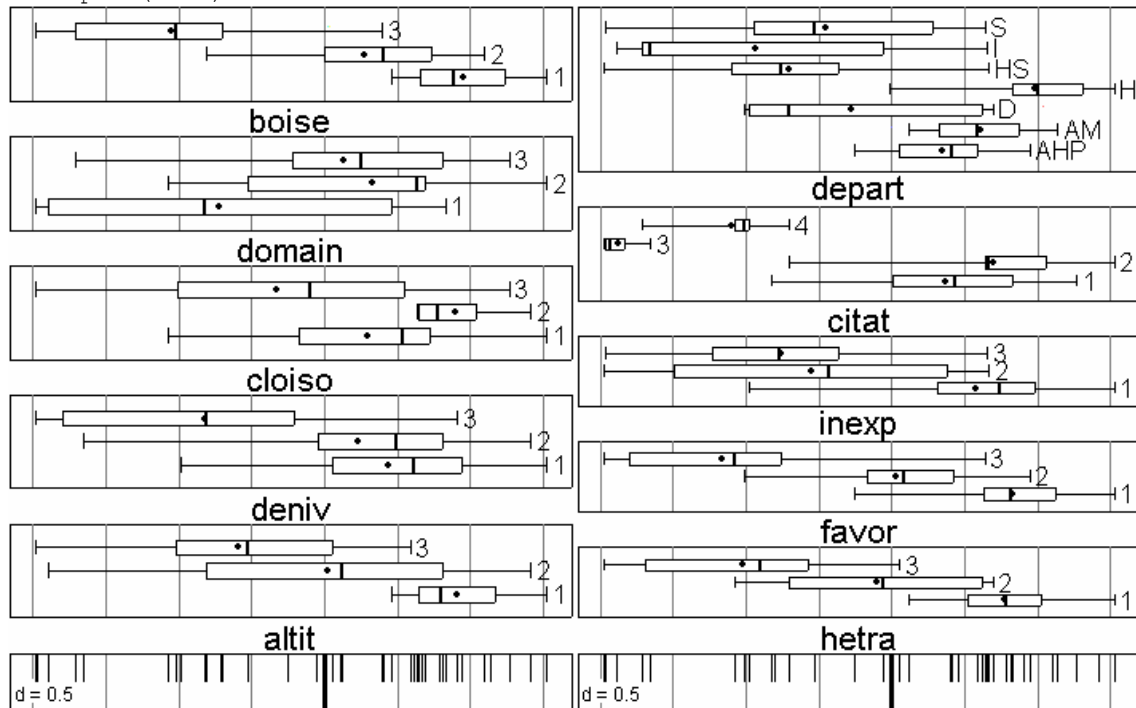
> for(i in c(1:8,10)) print(chisq.test(x=ours[,i],y=ours[,9]))

> acm1$scr
      RS1   RS2  p-value du khi2
altit 0.2951 0.2366      0.17
deniv 0.2793 0.1548      0.35
cloiso 0.1739 0.3509      0.03
domain 0.1670 0.6531 2e-5
boise 0.6468 0.1621 6e-05
hetra 0.6859 0.1095 0.006
favor 0.6674 0.1026 0.005
inexp 0.3755 0.2967      0.16
citat 0.6945 0.5662 ---
depart 0.4726 0.5956 0.05

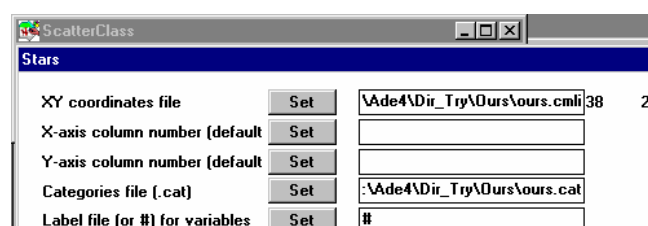
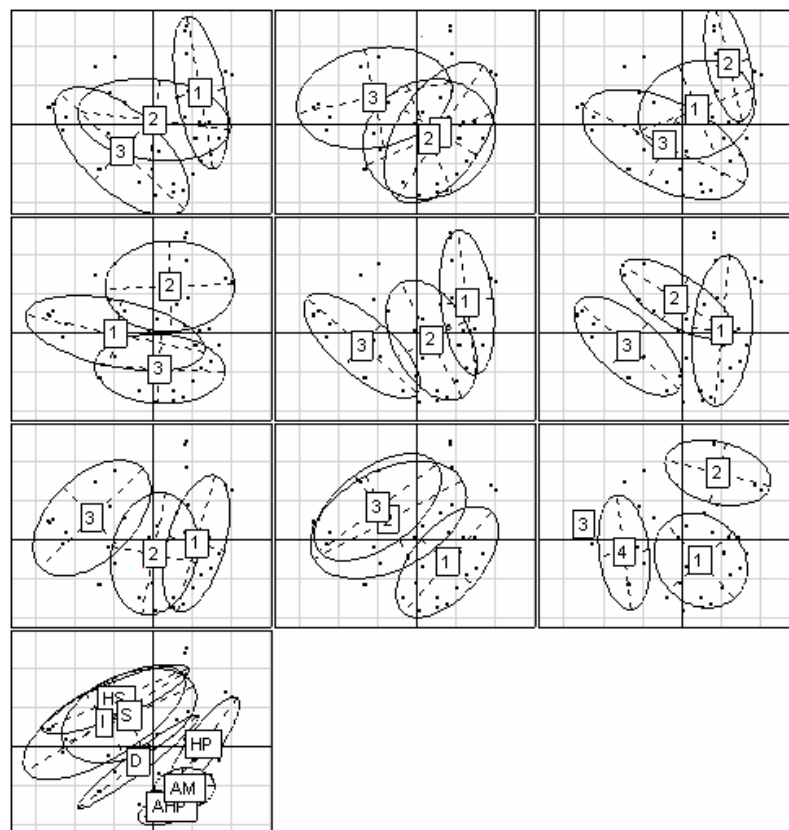
> score(acm1)
```

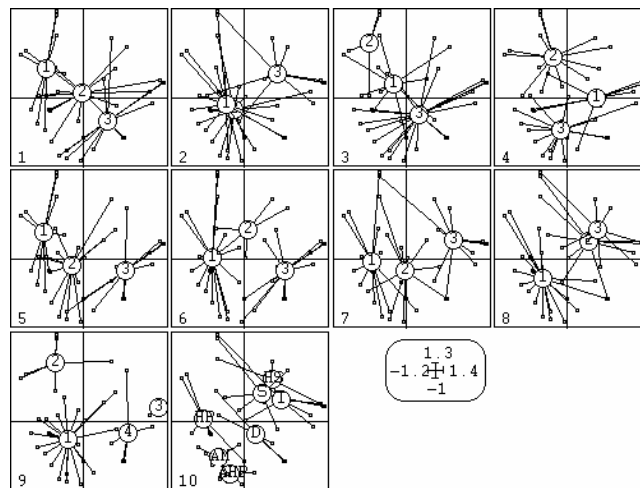


```
> boxplot(acm1)
```



```
> scatter(acm1)
```





Identifier le principe de la construction de ces figures.

Expliciter les relations entre les systèmes de coordonnées donnés par les deux logiciels (comparer la figure de la page 10 donnée dans R et la figure de la page 12 donnée par ADE-4).

5.3 Compléments

```
> ours01 <- acm.disjonctif(ours)
```

Identifier le contenu de ours01.

```
> oursburt <- acm.burt(ours,ours)
```

Identifier le contenu de oursburt (on l'appelle le tableau de BURT). Faire l'analyse des correspondances simples (AFC) du tableau de BURT.

Expliquer la nature et l'origine de la relation entre les scores des lignes et des colonnes de l'AFC du tableau de Burt.

```
> plot(cor$cs[,1],mcal$cs[,1])
> plot(cor$cs[,2],mcal$cs[,2])
```

Expliciter la nature de la relation entre les scores des lignes (ou des colonnes) de l'AFC du tableau de Burt et les scores des modalités de l'ACM du tableau.

Expliciter la nature de la relation entre les valeurs propres de l'AFC du tableau de Burt et celles de l'ACM du tableau : la démonstration est dans Saporta (1990, p. 224).

Faire l'analyse des correspondances simples (AFC) du tableau disjonctif complet.

Expliciter la nature de la relation entre les valeurs propres de l'AFC du tableau disjonctif et celles de l'ACM du tableau.

Expliciter la nature de la relation entre les scores des lignes (respectivement des colonnes) de l'AFC du tableau disjonctif et les scores des lignes (respectivement des modalités) de l'ACM du tableau. L'ACM peut être définie comme l'AFC du tableau

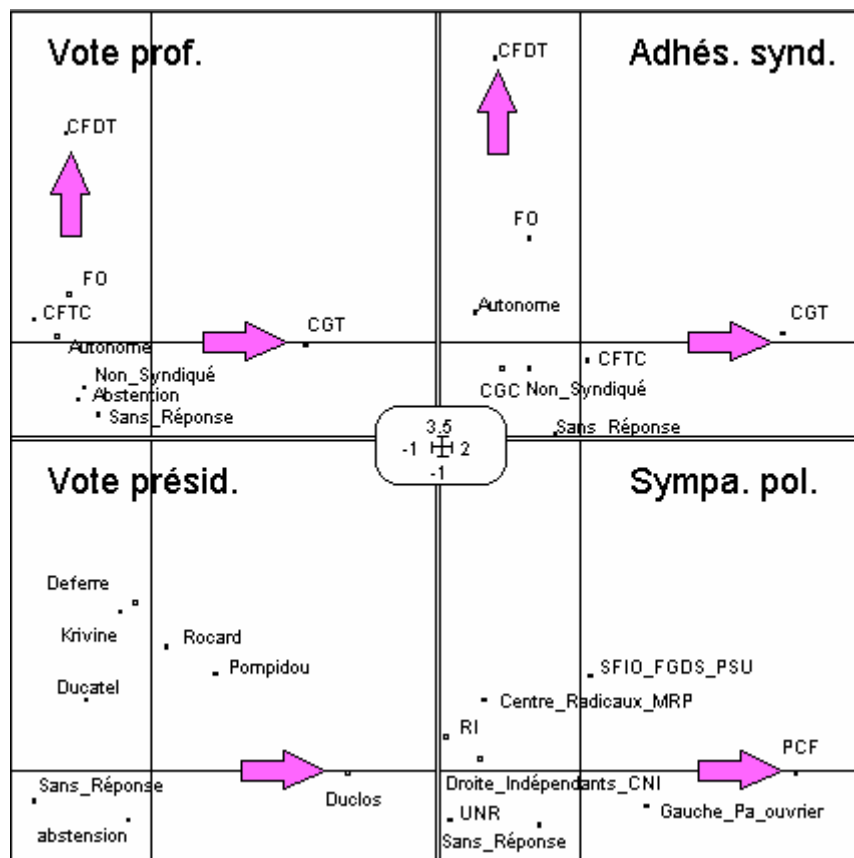
disjonctif complet. L'équivalence de l'AFC du tableau de Burt et celle de l'AFC du tableau disjonctif complet est complètement détaillée dans Lebart et al. (1995, p. 126).

5.4 ACM pondérée

Utiliser la carte de données Enquete du logiciel ADE-4. Les données sont proposées et étudiées dans Rouanet et Le Roux (1993, p. 283 et suivantes) p. 283 et suivantes. Lire le fichier ouv par CategVar: Read Categ File. Faire l'ACM (MCA: Multiple Correspondence Analysis) de ce fichier pondéré par les effectifs du fichier ouveff :

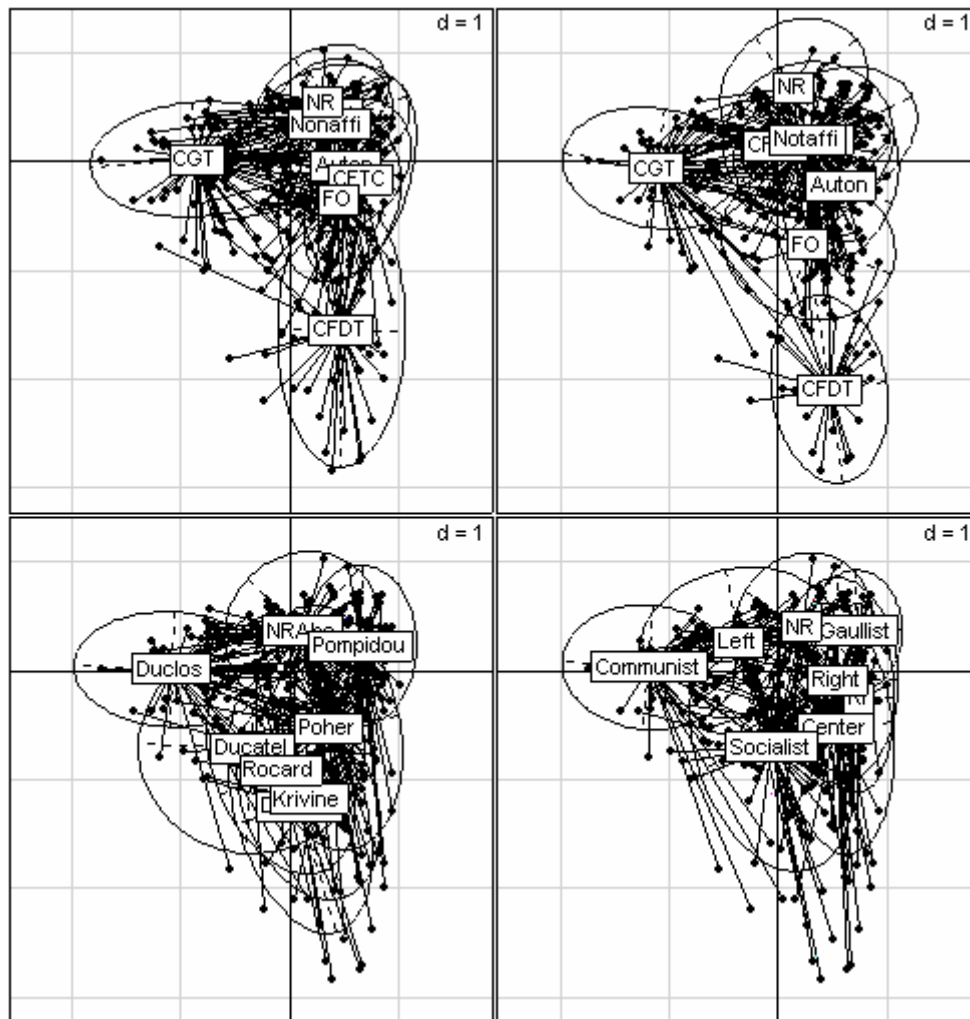
Num.	Eigenval.	R.Iner.	R.Sum	Num.	Eigenval.	R.Iner.	R.Sum
01	+6.1132E-01	+0.0873	+0.0873	02	+4.9123E-01	+0.0702	+0.1575
03	+4.1617E-01	+0.0595	+0.2170	04	+3.7331E-01	+0.0533	+0.2703
05	+3.4053E-01	+0.0486	+0.3189	06	+3.2813E-01	+0.0469	+0.3658
07	+3.0663E-01	+0.0438	+0.4096	08	+2.9978E-01	+0.0428	+0.4524

Représenter la carte des modalités multifenêtrée par variables (Scatters: Labels) :



Reprendre le même exercice dans R à partir du travail de Le Roux et Rouanet (1997)

```
data(worksurv)
acm1 <- dudi.acm(worksurv, row.w=attr(worksurv,"counts"), scan=F)
par(mfrow=c(2,2))
apply(worksurv,2, function(x) s.class(acm1$li,factor(x),
  attr(worksurv,'counts')))
```



5.5 Extension

Le concept de schéma de dualité permet de considérer l'ACP normée et l'ACM comme deux cas particuliers de l'analyse mixte qui mélange les deux types de variables. Voir, si la question est posée la fonction `dudi.mix`.

Références

- Besse, P., H. Caussinus, L. Ferre, and J. Fine. 1986. Some guidelines for principal component analysis. Pages 23-30 in I. A. f. S. Computing., editor. COMPSTAT 1986. Physica-Verlag, Heidelberg.
- Cailliez, F. 1984. Analyse des données. Les presses de l'Universités de Montréal.
- Cailliez, F., and J. P. Pages. 1976. Introduction à l'analyse des données. SMASH, 9 rue Duban, 75016 Paris.

- Chevenet, F., S. Dolédec, and D. Chessel. 1994. A fuzzy coding approach for the analysis of long-term ecological data. *Freshwater Biology* **31**:295-309.
- Collectif. 1989. *Encyclopaedia Universalis, Symposium, Les chiffres du Monde*. Encyclopaedia Universalis, Paris.
- Dolédec, S., D. Chessel, and J. M. Olivier. 1995. L'analyse des correspondances décentrée: application aux peuplements ichtyologiques du haut-Rhône. *Bulletin Français de la Pêche et de la Pisciculture* **336**:29-40.
- Eckart, C., and G. Young. 1936. The approximation of one matrix by another of lower rank. *Psychometrika* **1**:211-218.
- Erome, G. 1989. L'ours brun dans les Alpes françaises. Historique de sa disparition. Centre Ornithologique Rhône-Alpes, Villeurbanne.
- Escouffier, Y. 1987. The duality diagramm : a means of better practical applications. Pages 139-156 in P. Legendre and L. Legendre, editors. *Development in numerical ecology*. NATO advanced Institute , Serie G .Springer Verlag, Berlin.
- Greenacre, M. J. 1984. *Theory and applications of correspondence analysis*. Academic Press, London.
- Hill, M. O., and A. J. E. Smith. 1976. Principal component analysis of taxonomic data with multi-state discrete characters. *Taxon* **25**:249-255.
- Jacobs, G. A., and M. A. Martens. 1990. Quantification of eye irritation based upon in vitro changes of corneal thickness. *ATLA* **17**:255-262.
- Kiers, H. A. L. 1994. Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika* **56**:197-212.
- Kroonenberg, P. M., and R. Lombardo. 1999. Nonsymmetric correspondence analysis: a tool for analysing contingency tables with a dependence structure. *Multivariate Behavioral Research* **34**:367-396.
- Lascaux, J. M. 1996. Analyse de la variabilité morphologique de la truite commune (*Salmo trutta* L.) dans les cours d'eau du bassin pyrénéen méditerranéen. Thèse de doctorat en sciences agronomiques, INP Toulouse.
- Le Roux, B., and H. Rouanet. 1997. Interpreting axes in multiple correspondence analysis: method of the contributions of points and deviation. Pages 197-220 in B. J. and M. Greenacre, editors. *Visualization of categorical data*. Academic Press, London.
- Lebart, L., A. Morineau, and M. Piron. 1995. *Statistique exploratoire multidimensionnelle*. Dunod, Paris.
- Pagès, J. P., F. Cailliez, and Y. Escouffier. 1979. Analyse factorielle : un peu d'histoire et de géométrie. *Revue de Statistique Appliquée* **27**:5-28.

- Rouanet, H., and B. Le Roux. 1993. *Analyse des données multidimensionnelles*. Dunod, paris.
- Saporta, G. 1975. *Liaisons entre plusieurs ensembles de variables et codage de données qualitatives*. Thèse de 3^o cycle, Université Pierre et Marie Curie, Paris VI.
- Saporta, G. 1990. *Probabilités, analyse des données et statistique*. Technip, Paris.
- Tenenhaus, M., and F. W. Young. 1985. 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.
- Verneaux, J. 1973. *Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs. Essai de biotypologie*. Thèse d'état, Besançon.