Analyse en Composantes Principales (ACP)

Résumé

Méthode factorielle de réduction de dimension pour l'exploration statistique de données quantitatives complexes. Construction du modèle statistique associé, estimation. Représentations graphiques des individus, des variables et simultanée ; qualité de représentation.

Travaux pratiques *de complexité croissante par l'études de données de températures puis de données socio-économiques cubiques.*

Retour au plan du cours.

1 introduction

Lorsqu'on étudie simultanément un nombre important de variables quantitatives (ne serait-ce que 4 !), comment en faire un graphique global ? La difficulté vient de ce que les individus étudiés ne sont plus représentés dans un plan, espace de dimension 2, mais dans un espace de dimension plus importante (par exemple 4). L'objectif de l'Analyse en Composantes Principales (ACP) est de revenir à un espace de dimension réduite (par exemple 2) en déformant le moins possible la réalité (cf. l'introduction élémentaire à l'ACP). Il s'agit donc d'obtenir le résumé le plus pertinent possible des données initiales.

C'est la matrice des variances-covariances (ou celle des corrélations) qui va permettre de réaliser ce résumé pertinent, parce qu'on analyse essentiellement la dispersion des données considérées. De cette matrice, on va extraire, par un procédé mathématique adéquat, les facteurs que l'on recherche, en petit nombre. Ils vont permettre de réaliser les graphiques désirés dans cet espace de petite dimension (le nombre de facteurs retenus), en déformant le moins possible la configuration globale des individus selon l'ensemble des variables initiales (ainsi remplacées par les facteurs).

C'est l'interprétation de ces graphiques qui permettra de comprendre la structure des données analysées. Cette interprétation sera guidée par un certain nombre d'indicateurs numériques et graphiques, appelés aides à l'interprétation, qui sont là pour aider l'utilisateur à faire l'interprétation la plus juste et la plus objective possible.

L'analyse en Composantes Principales (ACP) est un grand classique de l"analyse des données" en France pour l'étude exploratoire ou la compression d'un grand tableau $n \times p$ de données quantitatives. Le livre de Jolliffe (2002)[2] en détaille tous les aspects et utilisations de façon exhaustive. Elle est introduite ici comme l'estimation des paramètres d'un modèle, afin de préciser la signification statistique des résultats obtenus. L'ACP est illustrée dans ce chapitre à travers l'étude de données élémentaires. Elles sont constituées des moyennes sur dix ans des températures moyennes mensuelles de 32 villes françaises. La matrice initiale $\mathbf{X}$ est donc (32×12). Les colonnes sont l'observation à différents instants d'une même variable ; elles sont homogènes et il est inutile de les réduire.

L'ACP joue dans ce cours un rôle central ; cette méthode sert de fondement théorique aux autres méthodes de statistique multidimensionnelle dites *factorielles* qui en apparaissent comme des cas particuliers. Cette méthode est donc étudiée en détail et abordée avec différents niveaux de lecture. La première section présente les grands principes de façon très élémentaire, voire intuitive, tandis que les suivantes explicitent les expressions matricielles des résultats.

D'un point de vue plus "mathématique", l'ACP correspond à l'approximation d'une matrice (n, p) par une matrice de même dimensions mais de rang $q < p$ (cf. rappels d'algèbre linéaire) ; q étant souvent de petite valeur 2, 3 pour la construction de graphiques facilement compréhensibles.

2 Espaces vectoriels

2.1 Notations

Soit p *variables statistiques* réelles X^j $(j = 1, \ldots, p)$ observées sur n individus i $(i = 1, \ldots, n)$ affectés des poids w_i :

$$\forall i = 1, \ldots, n \ : w_i \ > \ 0 \text{ et } \sum_{i=1}^{n} w_i = 1 \ ;$$

$$\forall i = 1, \ldots, n \ : x_i^j \ = \ X^j(i), \text{ mesure de } X^j \text{ sur le } i^{ème} \text{ individu.}$$

Ces mesures sont regroupées dans une matrice $\mathbf{X}$ d'ordre $(n \times p)$.

	X^1	$\cdots$	X^j	$\cdots$	X^p
1	x_1^1	$\cdots$	x_1^j	$\cdots$	x_1^p
$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	x_i^1	$\cdots$	x_i^j	$\cdots$	x_i^p
$\vdots$	$\vdots$		$\vdots$		$\vdots$
n	x_n^1	$\cdots$	x_n^j	$\cdots$	x_n^p

- À chaque individu i est associé le vecteur $\mathbf{x}_i$ contenant la i-ème ligne de $\mathbf{X}$ mise en colonne. C'est un élément d'un espace vectoriel noté E de dimension p ; nous choisissons $\mathbb{R}^p$ muni de la base canonique $\mathcal{E}$ et d'une métrique de matrice $\mathbf{M}$ lui conférant une structure d'espace euclidien : E est isomorphe à $(\mathbb{R}^p, \mathcal{E}, \mathbf{M})$; E est alors appelé *espace des individus*.
- À chaque variable X^j est associé le vecteur $\mathbf{x}^j$ contenant la j-ème colonne *centrée* (la moyenne de la colonne est retranchée à toute la colonne) de $\mathbf{X}$. C'est un élément d'un espace vectoriel noté F de dimension n ; nous choisissons $\mathbb{R}^n$ muni de la base canonique $\mathcal{F}$ et d'une métrique de matrice $\mathbf{D}$ diagonale des *poids* lui conférant une structure d'espace euclidien : F est isomorphe à $(\mathbb{R}^n, \mathcal{F}, \mathbf{D})$ avec $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$; F est alors appelé *espace des variables*.

2.2 Métrique des poids

L'utilisation de la métrique des poids dans l'espace des variables F donne un sens très particulier aux notions usuelles définies sur les espaces euclidiens. Ce paragraphe est la clé permettant de fournir les interprétations en termes statistiques des propriétés et résultats mathématiques.

Moyenne empirique de X^j :	$\overline{x^j}$	$=$	$\left\langle \mathbf{X}e^j, \mathbf{1}_n \right\rangle_{\mathbf{D}} = e^{j'}\mathbf{X}'\mathbf{D}\mathbf{1}_n$
Barycentre des individus :	$\overline{\mathbf{x}}$	$=$	$\mathbf{X}'\mathbf{D}\mathbf{1}_n$.
Matrice des données centrées :	$\overline{\mathbf{X}}$	$=$	$\mathbf{X} - \mathbf{1}_n\overline{\mathbf{x}}'$.
Écart-type de X^j :	σ_j	$=$	$(\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^j)^{1/2} = \left\|\mathbf{x}^j\right\|_{\mathbf{D}}$.
Covariance de X^j et X^k :	$\mathbf{x}^{j'}\mathbf{D}\mathbf{x}^k$	$=$	$\left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}$.
Matrice des covariances :	$\mathbf{S}$	$=$	$\sum_{i=1}^n w_i(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})'$
		$=$	$\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}$.
Corrélation de X^j et X^k :	$\frac{\left\langle \mathbf{x}^j, \mathbf{x}^k \right\rangle_{\mathbf{D}}}{\left\|\mathbf{x}^j\right\|_{\mathbf{D}}\left\|\mathbf{x}^k\right\|_{\mathbf{D}}}$	$=$	$\cos\theta_{\mathbf{D}}(\mathbf{x}^j, \mathbf{x}^k)$.

Attention : Par souci de simplicité des notations, on désigne toujours par $\mathbf{x}^j$ les colonnes de la matrice **centrée** $\overline{\mathbf{X}}$. On considère donc que des vecteurs "variables" sont toujours centrés.

Ainsi, lorsque les variables sont centrées et représentées par des vecteurs de F :

- la *longueur* d'un vecteur représente un *écart-type*,
- le *cosinus* d'un angle entre deux vecteurs représente une *corrélation*.

2.3 Objectifs

Les objectifs poursuivis par une ACP sont :

- la représentation graphique "optimale" des individus (lignes), minimisant les déformations du nuage des points, dans un sous-espace E_q de dimension q $(q < p)$,
- la représentation graphique des variables dans un sous-espace F_q en explicitant au "mieux" les liaisons initiales entre ces variables,
- la réduction de la dimension (compression), ou approximation de X par un tableau de rang q $(q < p)$.

Les derniers objectifs permettent d'utiliser l'ACP comme préalable à une autre technique préférant des variables orthogonales (régression linéaire) ou un nombre réduit d'entrées (réseaux neuronaux).

Des arguments de type géométrique dans la littérature francophone, ou bien de type statistique avec hypothèses de normalité dans la littérature anglo-saxonne, justifient la définition de l'ACP. Nous adoptons ici une optique intermédiaire en se référant à un modèle "allégé" car ne nécessitant pas d'hypothèse "forte" sur la distribution des observations (normalité). Plus précisément, l'ACP admet des définitions équivalentes selon que l'on s'attache à la représentation des individus, à celle des variables ou encore à leur représentation simultanée.

3 Modèle

Les notations sont celles du paragraphe précédent :

- $\mathbf{X}$ désigne le tableau des données issues de l'observation de p variables *quantitatives* X^j sur n individus i de *poids* w_i,
- E est l'espace des individus muni de la base canonique et de la métrique

de matrice $\mathbf{M}$,
- F est l'espace des variables muni de la base canonique et de la métrique des poids $\mathbf{D} = \text{diag}(w_1, \ldots, w_n)$.

De façon générale, un modèle s'écrit :

$$\textbf{Observation} = \textbf{Modèle} + \textbf{Bruit}$$

assorti de différents types d'hypothèses et de contraintes sur le modèle et sur le bruit.

En ACP, la matrice des données est supposée être issue de l'observation de n vecteurs aléatoires indépendants $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, de même matrice de covariance $\sigma^2\mathbf{\Gamma}$, mais d'espérances différentes $\mathbf{z}_i$, toutes contenues dans un sous-espace affine de dimension q ($q < p$) de E. Dans ce modèle, $E(\mathbf{x}_i) = \mathbf{z}_i$ est un paramètre spécifique attaché à chaque individu i et appelé *effet fixe*, le modèle étant dit *fonctionnel*. Ceci s'écrit en résumé :

$$\begin{array}{l} \{\mathbf{x}_i\,; i = 1, \ldots, n\},\ n \text{ vecteurs aléatoires indépendants de } E, \\ \mathbf{x}_i = \mathbf{z}_i + \boldsymbol{\varepsilon}_i\,, i = 1, \ldots, n \text{ avec } \left\{ \begin{array}{l} E(\boldsymbol{\varepsilon}_i) = 0,\ \text{var}(\boldsymbol{\varepsilon}_i) = \sigma^2\mathbf{\Gamma}, \\ \sigma > 0 \text{ inc. } \mathbf{\Gamma} \text{ rég. et connue,} \end{array} \right. \\ \exists A_q,\ \text{sous-espace affine de dim. } q \text{ de } E \text{ tel que } \forall i, \mathbf{z}_i \in A_q\ (q < p). \end{array} \tag{1}$$

Soit $\overline{\mathbf{z}} = \sum_{i=1}^n w_i \mathbf{z}_i$. Les hypothèses du modèle entraînent que $\overline{\mathbf{z}}$ appartient à A_q. Soit donc E_q le sous-espace vectoriel de E de dimension q tel que :

$$A_q = \overline{\mathbf{z}} + E_q.$$

Les paramètres à estimer sont alors E_q et $\mathbf{z}_i, i = 1, \ldots, n$, éventuellement σ ; $\mathbf{z}_i$ est la part systématique, ou *effet*, supposée de rang q ; éliminer le bruit revient donc à réduire la dimension.

Si les $\mathbf{z}_i$ sont considérés comme *aléatoires*, le modèle est alors dit *structurel* ; on suppose que $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ est un échantillon statistique i.i.d. Les unités statistiques jouent des rôles symétriques, elles ne nous intéressent que pour l'étude des relations entre les variables. On retrouve alors le principe de l'analyse en facteurs (ou en facteurs communs et spécifiques, ou *factor analysis*).

3.1 Estimation

PROPOSITION 1. — *L'estimation des paramètres de (1) est fournie par l'ACP de* $(\mathbf{X}, \mathbf{M}, \mathbf{D})$ *c'est-à-dire par la décomposition en valeurs singulières de* $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$ *:*

$$\widehat{\mathbf{Z}_q} = \sum_{k=1}^{q} \lambda_k^{1/2} \mathbf{u}^k \mathbf{v}^{k\prime} = \mathbf{U}_q \mathbf{\Lambda}^{1/2} \mathbf{V}_q'.$$

Preuve

Sans hypothèse sur la distribution de l'erreur, une estimation par les moindres carrés conduit à résoudre le problème :

$$\min_{E_q, \mathbf{z}_i} \left\{ \sum_{i=1}^n w_i \left\| \mathbf{x}_i - \mathbf{z}_i \right\|_{\mathbf{M}}^2 \ ;\ \dim(E_q) = q, \mathbf{z}_i - \overline{\mathbf{z}} \in E_q \right\}. \tag{2}$$

Soit $\overline{\mathbf{X}} = \mathbf{X} - \mathbf{1}_n \overline{\mathbf{x}}'$ la matrice centrée et $\mathbf{Z}$ la matrice $(n \times p)$ dont les lignes sont les vecteurs $(\mathbf{z}_i - \overline{\mathbf{z}})'$.

$$\sum_{i=1}^n w_i \left\| \mathbf{x}_i - \mathbf{z}_i \right\|_{\mathbf{M}}^2 = \sum_{i=1}^n w_i \left\| \mathbf{x}_i - \overline{\mathbf{x}} + \overline{\mathbf{z}} - \mathbf{z}_i \right\|_{\mathbf{M}}^2 + \left\| \overline{\mathbf{x}} - \overline{\mathbf{z}} \right\|_{\mathbf{M}}^2 \ ;$$

le problème (2) conduit alors à prendre $\widehat{\overline{\mathbf{z}}} = \overline{\mathbf{x}}$ et devient équivalent à résoudre :

$$\min_{\mathbf{Z}} \left\{ \left\| \overline{\mathbf{X}} - \mathbf{Z} \right\|_{\mathbf{M},\mathbf{D}} \ ; \mathbf{Z} \in \mathcal{M}_{n,p},\ \text{rang}(\mathbf{Z}) = q \right\}. \tag{3}$$

La fin de la preuve est une conséquence immédiate du théorème d'approximation matricielles (cf. rappels d'algèbre linéaire). □

- Les $\mathbf{u}^k$ sont les vecteurs propres $\mathbf{D}$-orthonormés de la matrice $\overline{\mathbf{X}}\mathbf{M}\overline{\mathbf{X}}'\mathbf{D}$ associés aux valeurs propres λ_k rangées par ordre décroissant.
- Les $\mathbf{v}_k$, appelés *vecteurs principaux*, sont les vecteurs propres $\mathbf{M}$-orthonormés de la matrice $\overline{\mathbf{X}}'\mathbf{D}\overline{\mathbf{X}}\mathbf{M} = \mathbf{S}\mathbf{M}$ associés aux mêmes valeurs propres ; ils engendrent des s.e.v. de dimension 1 appelés axes principaux.

Les estimations sont donc données par :

$$\begin{aligned}
\widehat{\overline{\mathbf{z}}} &= \overline{\mathbf{x}}, \\
\widehat{\mathbf{Z}_q} &= \sum_{k=1}^{q} \lambda^{1/2} \mathbf{u}^k \mathbf{v}^{k\prime} = \mathbf{U}_q \mathbf{\Lambda}^{1/2} \mathbf{V}_q' = \overline{\mathbf{X}} \widehat{\mathbf{P}_q}', \\
\text{où } \widehat{\mathbf{P}_q} &= \mathbf{V}_q \mathbf{V}_q' \mathbf{M} \text{ est la matrice de projection} \\
&\quad \mathbf{M}\text{-orthogonale sur } \widehat{E_q}, \\
\widehat{E_q} &= \text{vect}\{\mathbf{v}^1, \ldots, \mathbf{v}^q\}, \\
\widehat{E_2} &\text{ est appelé plan principal}, \\
\widehat{\mathbf{z}_i} &= \widehat{\mathbf{P}_q} \mathbf{x}_i + \overline{\overline{x}}.
\end{aligned}$$

Remarques

1. Les solutions sont emboîtées pour $q = 1, \ldots, p$:

$$E_1 = \text{ vect}\{\mathbf{v}^1\} \subset E_2 = \text{ vect}\{\mathbf{v}^1, \mathbf{v}^2\} \subset E_3 = \text{ vect}\{\mathbf{v}^1, \mathbf{v}^2, \mathbf{v}^3\} \subset \ldots$$

2. Les espaces principaux sont uniques sauf, éventuellement, dans le cas de valeurs propres multiples.
3. Si les variables ne sont pas homogènes (unités de mesure différentes, variances disparates), elles sont préalablement réduites :

$$\widetilde{\overline{\mathbf{X}}} = \overline{\mathbf{X}} \mathbf{\Sigma}^{-1/2} \text{ où } \mathbf{\Sigma} = \text{ diag } (\sigma_1^2, \ldots, \sigma_p^2), \text{ avec } \sigma_j^2 = \text{Var } (X^j) \text{ ;}$$

$\widetilde{\mathbf{S}}$ est alors la matrice $\mathbf{R} = \mathbf{\Sigma}^{-1/2} \mathbf{S} \mathbf{\Sigma}^{-1/2}$ des *corrélations*.

Sous l'hypothèse que la distribution de l'erreur est gaussienne, une estimation par maximum de vraisemblance conduit à la même solution.

3.2 Autre définition

On considère p variable statistiques *centrées* $X^1, \ldots, X^p$. Une *combinaison linéaire* de coefficients $\mathbf{f}_j$ de ces variables,

$$\mathbf{c} = \sum_{j=1}^{p} f_j \mathbf{x}^j = \overline{\mathbf{X}} \mathbf{f},$$

définit une nouvelle variable centrée C qui, à tout individu i, associe la "mesure"

$$C(i) = (\mathbf{x}_i - \overline{\mathbf{x}})' \mathbf{f}.$$

PROPOSITION 2. — *Soient p variables quantitatives centrées $X^1, \ldots, X^p$ observées sur n individus de poids w_i ; l'ACP de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$ est aussi la recherche des q combinaisons linéaires normées des X^j, non corrélées et dont la somme des variances soit maximale.*

- Les vecteurs $\mathbf{f}^k = \mathbf{M}\mathbf{v}^k$ sont les *facteurs principaux*. Ils permettent de définir les combinaisons linéaires des X^j optimales au sens ci-dessus.
- Les vecteurs $\mathbf{c}^k = \overline{\mathbf{X}}\mathbf{f}^k$ sont les *composantes principales*.
- Les variables C^k associées sont centrées, non corrélées et de variance λ_k ; ce sont les *variables principales* ;

$$\begin{aligned}
\text{cov}(C^k, C^\ell) &= (\overline{\mathbf{X}}\mathbf{f}^k)' \mathbf{D} \overline{\mathbf{X}} \mathbf{f}^\ell = \mathbf{f}^{k\prime} \mathbf{S} \mathbf{f}^\ell \\
&= \mathbf{v}^{k\prime} \mathbf{M} \mathbf{S} \mathbf{M} \mathbf{v}^\ell = \lambda_\ell \mathbf{v}^{k\prime} \mathbf{M} \mathbf{v}^\ell = \lambda_\ell \delta_k^\ell.
\end{aligned}$$

- Les $\mathbf{f}^k$ sont les vecteurs propres $\mathbf{M}^{-1}$-orthonormés de la matrice $\mathbf{MS}$.
- La matrice

$$\mathbf{C} = \overline{\mathbf{X}}\mathbf{F} = \overline{\mathbf{X}}\mathbf{M}\mathbf{V} = \mathbf{U}\mathbf{\Lambda}^{1/2}$$

  est la matrice des composantes principales.
- Les axes définis par les vecteurs $\mathbf{D}$-orthonormés u^k sont appelés *axes factoriels*.

4 Graphiques

4.1 Individus

Les graphiques obtenus permettent de représenter "au mieux" les distances euclidiennes inter-individus mesurées par la métrique $\mathbf{M}$.

4.1.1 Projection

Chaque individu i représenté par $\mathbf{x}_i$ est approché par sa projection $\mathbf{M}$-orthogonale $\widehat{\mathbf{z}_i}^q$ sur le sous-espace $\widehat{E_q}$ engendré par les q premiers vecteurs

principaux $\{\mathbf{v}^1, \ldots, \mathbf{v}^q\}$. En notant $\mathbf{e}_i$ un vecteur de la base canonique de E, la coordonnée de l'individu i sur $\mathbf{v}^k$ est donnée par :

$$\left\langle \mathbf{x}_i - \overline{\mathbf{x}}, \mathbf{v}^k \right\rangle_{\mathbf{M}} = (\mathbf{x}_i - \overline{\mathbf{x}})' \mathbf{M} \mathbf{v}^k = \mathbf{e}_i' \overline{\mathbf{X}} \mathbf{M} \mathbf{v}^k = c_i^k.$$

PROPOSITION 3. — *Les coordonnées de la projection* $\mathbf{M}$*-orthogonale de* $\mathbf{x}_i - \overline{\mathbf{x}}$ *sur* $\widehat{E_q}$ *sont les* q *premiers élément de la* i*-ème ligne de la matrice* $\mathbf{C}$ *des composantes principales.*

4.1.2 Qualités

La "qualité globale" des représentations est mesurée par la *part de dispersion expliquée* :

$$r_q = \frac{\text{tr}\mathbf{S}\mathbf{M}\widehat{\mathbf{P}_q}}{\text{tr}\mathbf{S}\mathbf{M}} = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}.$$

Remarque. — La dispersion d'un nuage de points unidimensionnel par rapport à sa moyenne se mesure par la variance. Dans le cas multidimensionnel, la dispersion du nuage $\mathcal{N}$ par rapport à son barycentre $\overline{\mathbf{x}}$ se mesure par l'*inertie*, généralisation de la variance :

$$I_g(\mathcal{N}) = \sum_{i=1}^{n} w_i \left\| \mathbf{x}_i - \overline{\mathbf{x}} \right\|_{\mathbf{M}}^2 = \left\| \overline{\mathbf{X}} \right\|_{\mathbf{M},\mathbf{D}}^2 = \text{ tr } (\overline{\mathbf{X}}' \mathbf{D} \overline{\mathbf{X}} \mathbf{M}) = \text{ tr } (\mathbf{S}\mathbf{M}).$$

La qualité de la représentation de chaque x_i est donnée par le cosinus carré de l'angle qu'il forme avec sa projection :

$$\left[\cos \theta(\mathbf{x}_i - \overline{\mathbf{x}}, \widehat{\mathbf{z}_i}^q)\right]^2 = \frac{\left\| \widehat{\mathbf{P}_q}(\mathbf{x}_i - \overline{\mathbf{x}}) \right\|_{\mathbf{M}}^2}{\left\| \mathbf{x}_i - \overline{\mathbf{x}} \right\|_{\mathbf{M}}^2} = \frac{\sum_{k=1}^{q} (c_i^k)^2}{\sum_{k=1}^{p} (c_i^k)^2}.$$

Pour éviter de consulter un tableau qui risque d'être volumineux (n lignes), les étiquettes de chaque individu sont affichées sur les graphiques avec des caractères dont la *taille est fonction de la qualité*. Un individu très mal représenté est à la limite de la lisibilité.

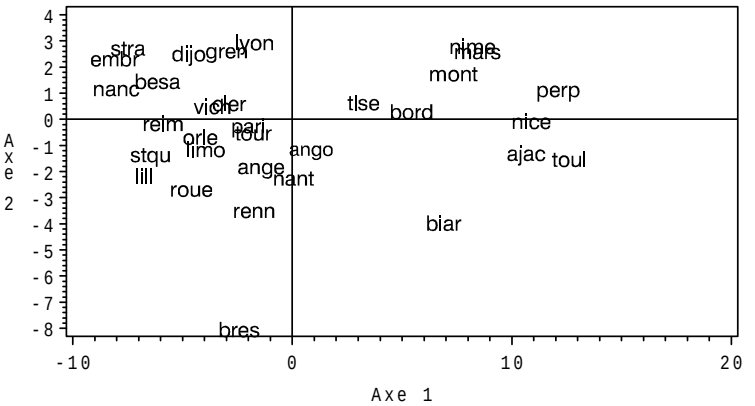

FIGURE 1 – *Températures : premier plan des individus.*

4.1.3 Contributions

Les contributions de chaque individu à l'inertie de leur nuage

$$\gamma_i = \frac{w_i \left\| \mathbf{x}_i - \overline{\mathbf{x}} \right\|_{\mathbf{M}}^2}{\text{tr}\mathbf{S}\mathbf{M}} = \frac{w_i \sum_{k=1}^{p} (c_i^k)^2}{\sum_{k=1}^{p} \lambda_k},$$

ainsi qu'à la variance d'une variable principale

$$\gamma_i^k = \frac{w_i (c_i^k)^2}{\lambda_k},$$

permettent de déceler les observations les plus *influentes* et, éventuellement, aberrantes. Ces points apparaissent visiblement lors du tracé des diagrammes-boîtes parallèles des composantes principales qui évitent ainsi une lecture fastidieuse de ce tableau des contributions. En effet, ils se singularisent aussi comme "outliers" ou atypiques hors de la boîte (au delà des moustaches) correspondant à une direction principale. Les individus correspondants, considérés comme *individus supplémentaires*, peuvent être éliminés lors d'une nouvelle analyse.

4.1.4 Individus supplémentaires

Il s'agit de représenter, par rapport aux axes principaux d'une analyse, des individus qui n'ont pas participé aux calculs de ces axes. Soit $\mathbf{s}$ un tel vecteur, il doit être centré, éventuellement réduit, puis projeté sur le sous-espace de représentation. Les coordonnées sont fournies par :

$$\left\langle \mathbf{v}^k, \mathbf{V}_q \mathbf{V}_q' \mathbf{M}(\mathbf{s}-\overline{\mathbf{x}}) \right\rangle_{\mathbf{M}} = \mathbf{v}^{k\prime} \mathbf{M} \mathbf{V}_q \mathbf{V}_q' \mathbf{M}(\mathbf{s}-\overline{\mathbf{x}}) = \mathbf{e}^{k\prime} \mathbf{V}_q' \mathbf{M}(\mathbf{s}-\overline{\mathbf{x}}).$$

Les coordonnées d'un individu supplémentaire dans la base des vecteurs principaux sont donc :

$$\mathbf{V}_q' \mathbf{M}(\mathbf{s}-\overline{\mathbf{x}}).$$

4.2 Variables

Les graphiques obtenus permettent de représenter "au mieux" les corrélations entre les variables (cosinus des angles) et, si celles-ci ne sont pas réduites, leurs variances (longueurs).

4.2.1 Projection

Une variable X^j est représentée par la projection $\mathbf{D}$-orthogonale $\widehat{\mathbf{Q}_q \mathbf{x}^j}$ sur le sous-espace F_q engendré par les q premiers axes factoriels. La coordonnée de $\mathbf{x}^j$ sur $\mathbf{u}^k$ est :

$$\begin{aligned} \left\langle \mathbf{x}^j, \mathbf{u}^k \right\rangle_{\mathbf{D}} = \mathbf{x}^{j\prime} \mathbf{D} \mathbf{u}^k &= \frac{1}{\sqrt{\lambda_k}} \mathbf{x}^{j\prime} \mathbf{D} \overline{\mathbf{X}} \mathbf{M} \mathbf{v}^k \\ &= \frac{1}{\sqrt{\lambda_k}} \mathbf{e}^{j\prime} \overline{\mathbf{X}}' \mathbf{D} \overline{\mathbf{X}} \mathbf{M} \mathbf{v}^k = \sqrt{\lambda_k} v_j^k. \end{aligned}$$

PROPOSITION 4. — *Les coordonnées de la projection $\mathbf{D}$-orthogonale de $\mathbf{x}^j$ sur le sous-espace F_q sont les q premiers éléments de la j-ème ligne de la matrice $\mathbf{V}\mathbf{\Lambda}^{1/2}$.*

4.2.2 Qualité

La qualité de la représentation de chaque $\mathbf{x}^j$ est donnée par le cosinus carré de l'angle qu'il forme avec sa projection :

$$\left[\cos\theta(\mathbf{x}^j, \widehat{\mathbf{Q}_q \mathbf{x}^j})\right]^2 = \frac{\left\| \widehat{\mathbf{Q}_q \mathbf{x}^j} \right\|_{\mathbf{D}}^2}{\|\mathbf{x}^j\|_{\mathbf{D}}^2} = \frac{\sum_{k=1}^q \lambda_k (v_k^j)^2}{\sum_{k=1}^p \lambda_k (v_k^j)^2}.$$

4.2.3 Corrélations variables — facteurs

Ces indicateurs aident à l'interprétation des axes factoriels en exprimant les corrélations entre variables principales et initiales.

$$\mathrm{cor}(X^j, C^k) = \cos\theta(\mathbf{x}^j, \mathbf{c}^k) = \cos\theta(\mathbf{x}^j, \mathbf{u}^k) = \frac{\left\langle \mathbf{x}^j, \mathbf{u}^k \right\rangle_{\mathbf{D}}}{\|\mathbf{x}^j\|_{\mathbf{D}}} = \frac{\sqrt{\lambda_k}}{\sigma_j} v_j^k \ ;$$

ce sont les éléments de la matrice $\mathbf{\Sigma}^{-1/2}\mathbf{V}\mathbf{\Lambda}^{1/2}$.

4.2.4 Cercle des corrélations

Dans le cas de variables réduites $\widetilde{\mathbf{x}}^j = \sigma_j^{-1}\mathbf{x}^j$, $\left\|\widetilde{\mathbf{x}}^j\right\|_{\mathbf{D}} = 1$, les $\widetilde{\mathbf{x}}^j$ sont sur la sphère unité $\mathcal{S}_n$ de F. L'intersection $\mathcal{S}_n \cap F_2$ est un cercle centré sur l'origine et de rayon 1 appelé *cercle des corrélations*. Les projections de $\widetilde{\mathbf{x}}^j$ et $\mathbf{x}^j$ sont colinéaires, celle de $\widetilde{\mathbf{x}}^j$ étant à l'intérieur du cercle :

$$\left\| \widehat{\mathbf{Q}_2 \widetilde{\mathbf{x}}^j} \right\|_{\mathbf{D}} = \cos\theta(\mathbf{x}^j, \widehat{\mathbf{Q}_2 x_j}) \leq 1.$$

Ainsi, plus $\widehat{\mathbf{Q}_2 \widetilde{\mathbf{x}}^j}$ est proche de ce cercle, meilleure est la qualité de sa représentation. Ce graphique est commode à interpréter à condition de se méfier des échelles, le cercle devenant une ellipse si elles ne sont pas égales. Comme pour les individus, la taille des caractères est aussi fonction de la qualité des représentations.

4.3 Biplot

À partir de la décomposition en valeurs singulières de $(\overline{\mathbf{X}}, \mathbf{M}, \mathbf{D})$, on remarque que chaque valeur

$$x_i^j - \overline{\mathbf{x}}^j = \sum_{k=1}^p \sqrt{\lambda_k} \mathbf{u}_i^k \mathbf{v}_k^j = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{V}'\right]_i^j$$

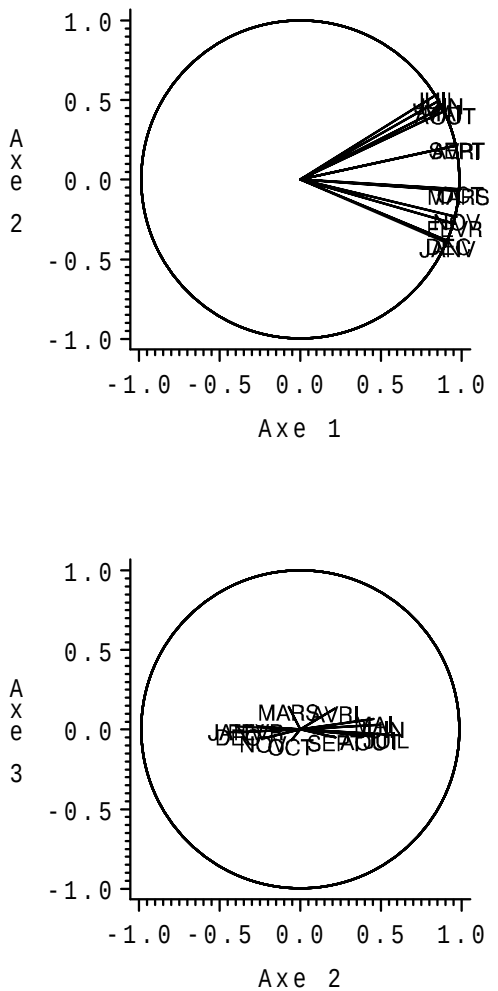

FIGURE 2 – *Températures : Premier et deuxième plan des variables.*

s'exprime comme produit scalaire usuel des vecteurs

$$\mathbf{c}_i = \left[\mathbf{U}\mathbf{\Lambda}^{1/2}\right]_i \text{ et } \mathbf{v}^j \text{ ou encore } \mathbf{u}_i \text{ et } \left[\mathbf{V}\mathbf{\Lambda}^{1/2}\right]_j .$$

Pour $q = 2$, la quantité $\widehat{z}_i^j$ en est une approximation limitée aux deux premiers termes.

Cette remarque permet d'interpréter deux autres représentations graphiques en ACP projetant *simultanément* individus et variables.

1. la représentation *isométrique ligne* utilise les matrices $\mathbf{C}$ et $\mathbf{V}$; elle permet d'interpréter les distances entre individus ainsi que les produits scalaires entre un individu et une variable qui sont, dans le premier plan principal, des approximations des valeurs observées $X^j(\omega_i)$;
2. la représentation *isométrique colonne* utilise les matrices $\mathbf{U}$ et $\mathbf{V}\mathbf{\Lambda}^{1/2}$; elle permet d'interpréter les angles entre vecteurs variables (corrélations) et les produits scalaires comme précédemment.

Remarques

1. Dans le cas fréquent où $\mathbf{M} = \mathbf{I}_p$ et où les variables sont réduites, le point représentant X^j, en superposition dans l'espace des individus se confond avec un pseudo individu supplémentaire qui prendrait la valeur 1 (écart-type) pour la variable j et 0 pour les autres.
2. En pratique, ces différents types de représentations (simultanées ou non) ne diffèrent que par un changement d'échelle sur les axes ; elles sont très voisines et suscitent souvent les mêmes interprétations. L'usage théoriquement abusif fait finalement superposer les deux représentations isométriques lignes et colonnes.

5 Choix de dimension

La qualité des estimations auxquelles conduit l'ACP dépend, de façon évidente, du choix de q, c'est-à-dire du nombre de composantes retenues pour reconstituer les données, ou encore de la dimension du sous-espace de représentation.

De nombreux critères de choix pour q ont été proposés dans la littérature. Nous présentons ici ceux, les plus courants, basés sur une heuristique et un reposant sur une quantification de la stabilité du sous-espace de représentation.

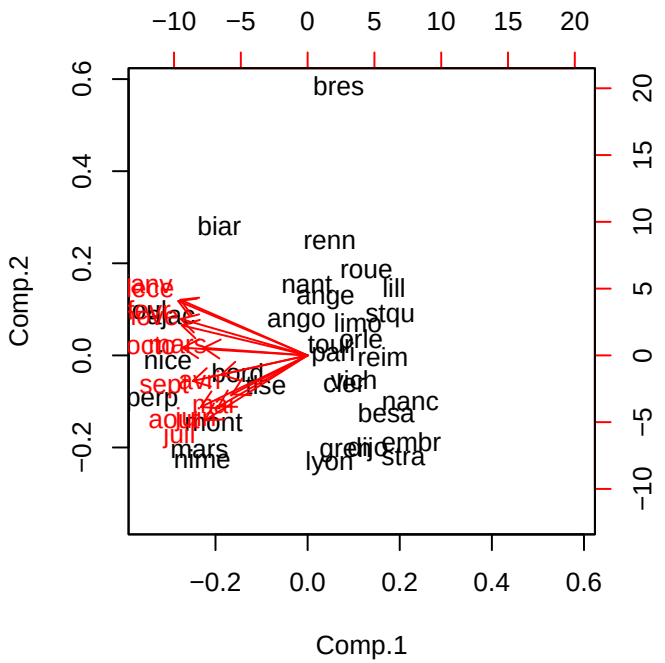

FIGURE 3 – Températures : Représentation simultanée ou biplot du premier plan.

D'autres critères, non explicités, s'inspirent des pratiques statistiques décisionnelles ; sous l'hypothèse que l'erreur admet une distribution *gaussienne*, on peut exhiber les lois asymptotiques des valeurs propres et donc construire des tests de nullité ou d'égalité de ces dernières. Malheureusement, outre la nécessaire hypothèse de normalité, ceci conduit à une procédure de tests emboîtés dont le niveau global est incontrôlable. Leur utilisation reste donc heuristique.

5.1 Part d'inertie

La "qualité globale" des représentations est mesurée par la *part d'inertie expliquée* :

$$r_q = \frac{\sum_{k=1}^{q} \lambda_k}{\sum_{k=1}^{p} \lambda_k}.$$

La valeur de q est choisie de sorte que cette part d'inertie expliquée r_q soit supérieure à une valeur seuil fixée a priori par l'utilisateur. C'est souvent le seul critère employé.

5.2 Règle de Kaiser

On considère que, si tous les éléments de Y sont indépendants, les composantes principales sont toutes de variances égales (égales à 1 dans le cas de l'ACP réduite). On ne conserve alors que les valeurs propres supérieures à leur moyenne car seules jugées plus "informatives" que les variables initiales ; dans le cas d'une ACP réduite, ne sont donc retenues que celles plus grandes que 1. Ce critère, utilisé implicitement par SAS/ASSIST, a tendance à surestimer le nombre de composantes pertinentes.

5.3 Éboulis

C'est le graphique (figures 4) présentant la décroissance des valeurs propres. Le principe consiste à rechercher, s'il existe, un "coude" (changement de signe dans la suite des différences d'ordre 2) dans le graphe et de ne conserver que les valeurs propres jusqu'à ce coude. Intuitivement, plus l'écart $(\lambda_q - \lambda_{q+1})$ est significativement grand, par exemple supérieur à $(\lambda_{q-1} - \lambda_q)$, et plus on peut être assuré de la stabilité de $\widehat{E_q}$.

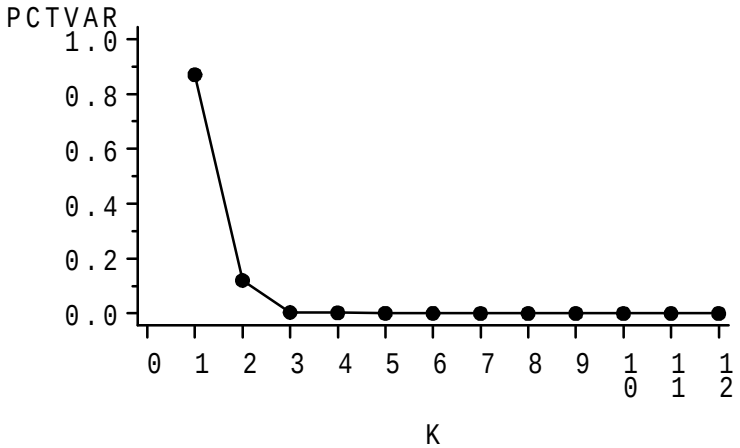

FIGURE 4 – *Températures : éboulis des valeurs propres.*

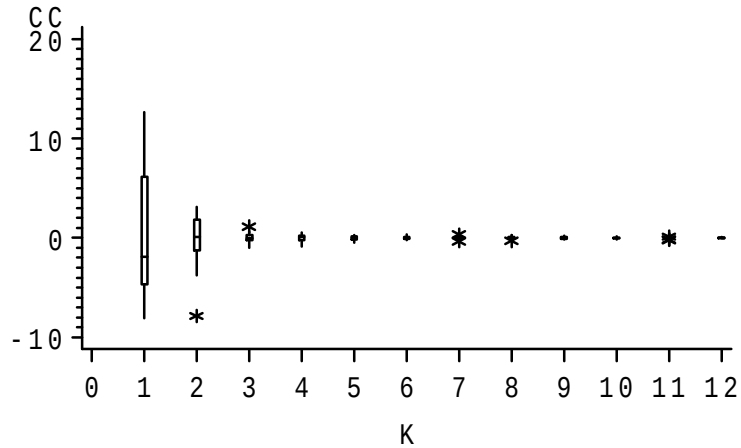

FIGURE 5 – *Températures : composantes en boîtes.*

5.4 Diagrammes boîtes

Un graphique (figure 5) présentant, en parallèle, les diagrammes boîtes des variables principales illustre bien leurs qualités : stabilité lorsqu'une grande boîte est associée à de petites moustaches, instabilité en présence d'une petite boîte, de grandes moustaches et de points isolés. Intuitivement, on conserve les premières "grandes boîtes". Les points isolés ou "outliers" désignent les points à forte contribution, ou potentiellement influents, dans une direction principale. Ils nécessitent une étude clinique : une autre analyse dans laquelle ils sont déclarés supplémentaires (poids nuls) afin d'évaluer leur impact sur l'orientation des axes.

5.5 Stabilité

La présentation de l'ACP, comme résultat de l'estimation d'un modèle, offre une autre approche au problème du choix de dimension. La qualité des estimations est évaluée de façon habituelle en statistique par un risque moyen quadratique définissant un critère de stabilité du sous-espace de représentation. Il est défini comme l'espérance d'une distance entre le modèle "vrai" et l'estimation qui en est faite. Besse (1992)[1] propose d'étudier la qualité de l'estimation du sous-espace de représentation $\widehat{E_q}$ en considérant la fonction perte :

$$L_q = Q(E_q, \widehat{E_q}) = \frac{1}{2}\left\|\mathbf{P}_q - \widehat{\mathbf{P}_q}\right\|^2_{\mathbf{M,D}} = q - \mathrm{tr}\mathbf{P}_q\widehat{\mathbf{P}_q},$$

où Q mesure la distance entre deux sous-espaces par la distance usuelle entre les matrices de projection qui leur sont associées. C'est aussi la somme des carrés des coefficients de corrélation canonique entre les ensembles de composantes ou de variables principales qui engendrent respectivement E_q et son estimation $\widehat{E_q}$.

Un risque moyen quadratique est alors défini en prenant l'espérance de la fonction perte :

$$R_q = EQ(E_q, \widehat{E_q}). \tag{4}$$

Sans hypothèse sur la distribution de l'erreur, seules des techniques de réechantillonnage (bootstrap, jackknife) permettent de fournir une estimation de ce risque moyen quadratique. Leur emploi est justifié, car le risque est invariant par permutation des observations, mais coûteux en temps de calcul.

On se pose donc la question de savoir pour quelles valeurs de q les représentations graphiques sont fiables, c'est-à-dire stables pour des fluctuations de l'échantillon. Besse (1992)[1] propose d'utiliser une approximation de l'estimateur par jackknife ; elle fournit, directement à partir des résultats de l'A.C.P. (valeurs propres et composantes principales), une estimation satisfaisante du risque :

$$\widehat{R_{JKq}} = \widehat{R_{\mathbf{P}q}} + O((n-1)^{-2}).$$

$\widehat{R_{\mathbf{P}q}}$ est une approximation analytique de l'estimateur jackknife qui a pour expression :

$$\widehat{R_{\mathbf{P}q}} = \frac{1}{n-1} \sum_{k=1}^{q} \sum_{j=q+1}^{p} \frac{\frac{1}{n} \sum_{i=1}^{n} (c_i^k)^2 (c_i^j)^2}{(\lambda_j - \lambda_k)^2} \tag{5}$$

où c_i^j désigne le terme général de la matrice des composantes principales $\mathbf{C}$.

Ce résultat souligne l'importance du rôle que joue l'écart $(\lambda_q - \lambda_{q+1})$ dans la stabilité du sous-espace de représentation. Le développement est inchangé dans le cas d'une ACP réduite ; de plus, il est valide tant que

$$n > \frac{\|\mathbf{S}\|_2^2}{\inf\{(\lambda_k - \lambda_{k+1}); k = 1, \ldots, q\}}.$$

La figure 6 montrent la stabilité du sous-espace de représentation en fonction de la dimension q pour l'A.C.P. des données de températures. Comme souvent, le premier axe est très stable tandis que le premier plan reste fiable. Au delà, les axes étant très sensibles à toute perturbation des données, ils peuvent être associés à du bruit. Ces résultats sont cohérents avec les deux critères graphiques précédents mais souvent, en pratique, le critère de stabilité conduit à un choix de dimension plus explicite.

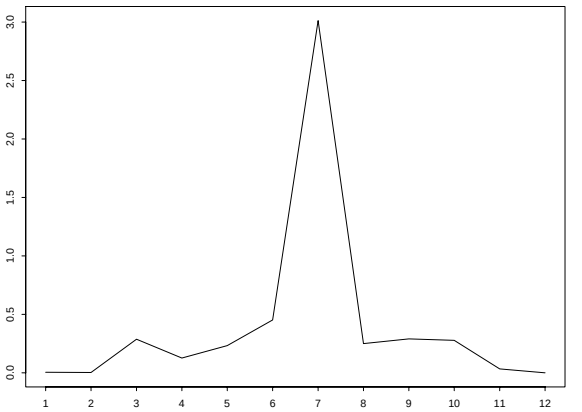

FIGURE 6 – *Températures : stabilité des sous-espaces.*

6 Interprétation

Les macros SAS utilisées, de même que la plupart des logiciels, proposent, ou autorisent, l'édition des différents indicateurs (contributions, qualités, corrélations) et graphiques définis dans les paragraphes précédents.

- Les *contributions* permettent d'identifier les individus très influents pouvant déterminer à eux seuls l'orientation de certains axes ; ces points sont vérifiés, caractérisés, puis éventuellement considérés comme *supplémentaires* dans une autre analyse.
- Il faut choisir le nombre de composantes à retenir, c'est-à-dire la dimension des espaces de représentation.
- Les axes factoriels sont interprétés par rapport aux variables initiales bien représentées.
- Les graphiques des individus sont interprétés, en tenant compte des qualités de représentation, en termes de regroupement ou dispersions par rapport aux axes factoriels et projections des variables initiales.

Les quelques graphiques présentés suffisent, dans la plupart des cas, à l'interprétation d'une ACP classique et évitent la sortie volumineuse, lorsque n est grand, des tableaux d'aide à l'interprétation (contributions, cosinus carrés). On échappe ainsi à une critique fréquente, et souvent justifiée, des anglo-saxons vis-à-vis de la pratique française de "l'analyse des données" qui, paradoxalement, cherche à "résumer au mieux l'information" mais produit plus de chiffres en sortie qu'il n'y en a en entrée !

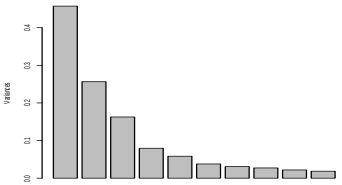

FIGURE 7 – *Souris : éboulis des dix premières valeurs propres de l'ACP.*

Remarque. — L'ACP est une technique *linéaire* optimisant un critère *quadratique* ; elle ne tient donc pas compte d'éventuelles liaisons non linéaires et présente une forte sensibilité aux valeurs extrêmes.

7 Exemple : Données génomiques

L'éboulis des premières valeurs propres (figure 7) conduit à considérer trois dimensions représentant environ les deux tiers de l'inertie globale mais nous limiterons l'interprétation un peu sommaire au premier plan.

La figure 8 représente conjointement souris et gènes (biplot). Dans le cadre de cette ACP, il est cohérent de rechercher quels sont les 25% des gènes contribuant le plus à la définition de l'espace propre à trois dimensions jugé pertinent. Avec cette sélection, la représentation des variables ainsi restreinte à 30 gènes est plus facilement lisible sur le plan factoriel. Pour des données plus volumineuses (puces pangénomiques) d'autres outils (version parcimonieuse ou creuse de l'ACP) sont à considérer.

Le premier plan (Fig. 8) doit être interprété globalement puisque sa deuxième bissectrice sépare exactement les souris `WT` des souris `PPAR`. Les gènes à coordonnées négatives sur l'axe 2 et positives sur l'axe1 sont sensiblement plus exprimés chez les souris `WT`, en particulier `CYP3A11`, `CYP4A10`, `CYP4A14`, `THIOL`, `PMDCI`, `GSTpi2`, `L.FABP` et `FAS` (négatif sur les deux axes). À l'inverse, les gènes à forte coordonnée négative sur l'axe 2 s'expriment davantage chez les souris `PPAR`, par exemple, `S14` et `CAR1`. Ceci est en partie connu des biologistes.

Sur cette représentation, seules les souris `WT` présentent des comportement

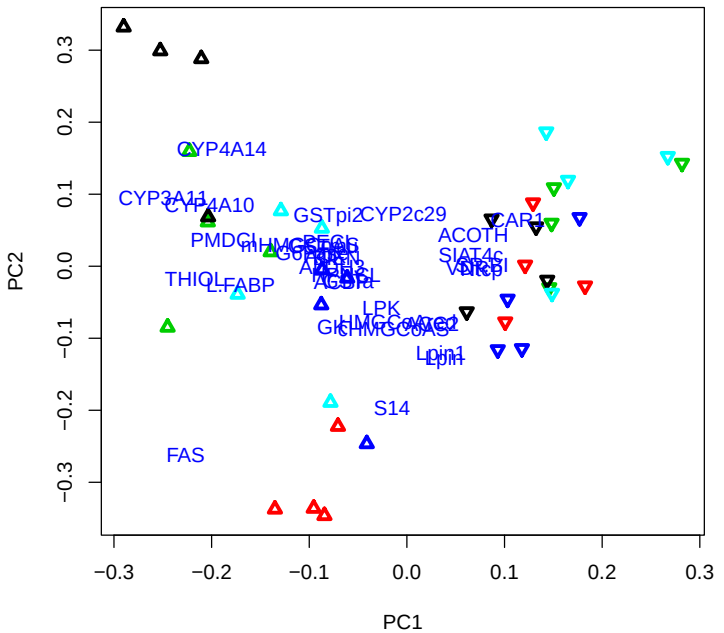

FIGURE 8 – *Représentations conjointe sur le premier plan principal. Les souris identifiés par leur génotype (*`WT` *triangles vers le haut,* `PPAR` *vers le bas) et leur régime (principalement noir-*`dha` *et rouge-*`efad`*).*

sensiblement différents au regard des régimes. Le phénomène le plus marquant est l'opposition, chez ces souris `WT`, entre les régimes `dha` (triangles noirs), dont les coordonnées sont toutes positives, et `efad` (triangles rouges), dont les coordonnées sont toutes négatives. Les gènes les plus exprimés dans le premier cas (régime `dha` chez les souris WT) sont `CYP3A11`, `CYP4A10`, `CYP4A14` ; dans le second cas (régime `efad` chez les mêmes souris), il s'agit des gènes `FAS` et `S14`. Parmi ces régulations, on note une opposition entre les `CYP4A`, connus pour être impliqués dans le catabolisme des acides gras, et les gènes `FAS` et `S14` impliqués eux dans la synthèse des lipides. Par ailleurs, la régulation de `CYP3A11` par le DHA a déjà été décrite dans la littérature.

Références

[1] P.C. Besse, *PCA stability and choice of dimensionality*, Statistics & Probability Letters **13** (1992), 405–410.

[2] I. Jolliffe, *Principal Component Analysis*, 2nd edition éd., Springer-Verlag, 2002.