Analyse en Composantes Principales
Exploration multidimensionnelle des données

Dhafer Malouche

http://essai.academia.edu/DhaferMalouche
dhafer.malouche@me.com

Janvier 2014

Plan

Exemple de données

Exemple 1 : `decathlon`

- **Source :** package `FactoMineR`
- **Cadre des données :**
  - JO athènes 23/24 Août 2004
  - Decastar 25/26 Septembre 2004
- **Variables :**
  - 10 scores des athlètes (épreuve du décathlon)
  - Total des points, classement des athlètes
  - Variable compétition.

  12 Variables quantitatives, 1 Variable qualitative.
- **Individus :** 41 athlètes.

Exemple 1 : `decathlon`

Objectif

- Obtenir une typologie des athlètes par rapport à leurs performances.
- Définir des dimensions et des facteurs en commun entre les variables : un facteur de puissance, un facteur de rapidité...
- Interpréter la typologie par au score final et/ou rang de l'athlète.
- Y-a-t-il une différence entre les scores des JO d'Athènes et Decastar ?

Exemple 2 : Climat en Europe

- **Source :**
  `http://www.factominer.free.fr/book/temperature.csv`
- **Description :** Une investigation du climat dans différentes villes en Europe.
- **Variables :**
  - Température mensuelle (de Janvier à Décembre)
  - Les coordonnées GPS de chaque ville.
  - Amplitude thermale : Différence entre les températures maximales et minimales
  - Moyenne Annuelle.
  - Une variable qualitative : la direction (S, N, O, E).
- **Objectif :** Comprendre la variabilité de la température dans différentes villes européennes.

Exemple 3 : Données génomiques

- **Source :**
  `http://factominer.free.fr/book/chicken.csv`
- **Description :** 43 poulets ayant subis 6 régimes : régime normal (N), Jeûne pendant 16h (F16), Jeûne pendant 16h et puis se réalimenter pendant 5h (F16R5), (F16R16), (F48), (F48R24)
- **Variables :** Après le régime, on a effectué une analyse des gènes utilisant une puce ADN : 7407 expressions de gènes.
- **Objectif :**
  - Voir si les gènes s'expriment différemment selon le niveau de stress.
  - Combien de temps lui faut-il le poulet pour revenir à la situation normale ?

- Commande `princomp`
- Package `FactoMineR` : commande `PCA`, un Plugin `Rcmdr`
- Package `ade4` : commande `dudi.pca` et une interface `Tcl/Tk`
- Package pcaMethods : différents types de méthodes d'ACP.
- Package `rattle`
- Package `JGR`, `Deducer`..

Les données : Individus et variables

- $E = \{w_1, \ldots, w_n\}$ un échantillon de taille n.
- $x^1, \ldots, x^d$, d variables quantitatives.
- y une variable quantitative supplémentaire, $y = f(x^1, \ldots, x^d)$, f inconnue.
- A une variable qualitative supplémentaire à r modalités : $a_1, \ldots, a_r$.
- les observations :
  - quantitatives : $x_i^j = x^j(w_i)$, $\forall\, i = 1 \ldots n$ et $j = 1 \ldots d$.
  - supplémentaire quantitative : $y_i = y(w_i)$, $\forall\, i = 1 \ldots n$.
  - supplémentaire qualitative : $a_l = A(w_i)$, $\forall\, i = 1 \ldots n$ et $l = 1 \ldots r$.

$x^1, \ldots, x^d$ variables actives, y et A variables supplémentaires et $X = (x_i^j)$ est la matrice $n \times d$ des variables actives.

Les éléments de l'analyse

- p_i est le poids de chaque individu w_i dans l'échantillon E :

$$p_i \in [0,1],\ \sum_{i=1}^{n} p_i = 1$$

. Matrice des poids

$$D = \begin{pmatrix} p_1 & & 0 \\ & \ddots & \\ 0 & & p_n \end{pmatrix}$$

- Nuage des individus : chaque individu w_i est représenté par $x_i = {}^t(x_i^1, \ldots, x_i^d) \in \mathbb{R}^d$. $\mathcal{N}_I = \{ x_i \in \mathbb{R}^d,\ i = 1 \ldots n\}$.
- Nuage des variables : chaque variable x^j est représentée par $x^j = {}^t(x_1^j, \ldots, x_n^j) \in \mathbb{R}^n$. $\mathcal{N}_V = \{ x^j \in \mathbb{R}^n,\ j = 1 \ldots d\}$.

Géométries dans $\mathbb{R}^n$ et $\mathbb{R}^d$

- Centre de gravité de $\mathcal{N}_I$, $g = {}^t(\overline{x}^1, \ldots, \overline{x}^d) \in \mathbb{R}^d$ où

$$\overline{x}^j = \sum_{i=1}^{n} p_i x_i^j \text{ est la moyenne de } x^j$$

  Pour la suite X est une matrice centrée $\Rightarrow g = 0$.

- $\mathbb{R}^d$ est un espace euclidien muni du produit scalaire usuel :

$$< x_i, x_{i'} >= {}^t(x_i) x_{i'} = \sum_{j=1}^{d} x_i^j x_{i'}^j.$$

- $\mathbb{R}^n$ est un espace euclidien muni du produit scalaire :

$$< x^j, x^{j'} >= {}^t(x^j)\, D\, x^{j'} = \sum_{i=1}^{n} p_i\, x_i^j x_i^{j'}.$$

  C'est un produit scalaire de métrique D.

Variances, Covariances et Corrélations

- Variance de la variable x^j :

$$\text{Var}(x^j) = \sum_{i=1}^{n} p_i (x_i^j)^2$$

$\Rightarrow \text{Var}(x^j) = \| x^j \|^2$ dans $\mathbb{R}^n$.

- Covariance entre les variables x^j et $x^{j'}$

$$\text{Cov}(x^j, x^{j'}) = \sum_{i=1}^{n} p_i x_i^j x_i^{j'} = < x^j, x^{j'} > .$$

- Corrélation entre les variables x^j et $x^{j'}$

$$\text{Cor}(x^j, x^{j'}) = \frac{\text{Cov}(x^j, x^{j'})}{\sqrt{\text{Var}(x^j)\,\text{Var}(x^{j'})}} = \frac{< x^j, x^{j'} >}{\| x^j \| \| x^{j'} \|} = \cos\alpha$$

où α est l'angle entre les vecteurs x^j et $x^{j'}$ dans $\mathbb{R}^n$.

ACP : Problématique

- Trouver un sous-espace vectoriel de dimension faible (2 ou 3) pour faire une représentation des graphiques des variables et des individus.
- Cette représentation a pour objectif de bien montrer
  - les relations existantes entre les variables : les *corrélations*.
  - les similarités et les éventuels sous-groupes dans l'échantillon.
- Chercher une suite de variables artificielles $C^1, \ldots, C^k$, $k << d$, qui
  - sont combinaisons linéaires des variables $x^1, \ldots, x^d$,
  - non-corrélées entre elles
  - et captent le maximum de variance du nuage des individus $\mathcal{N}_I$.

ACP : Deux approches, 1ère approche

- La variance totale du nuage des points :

$$I(\mathcal{N}_I) = \sum_{i=1}^{n} p_i d(0, x_i)^2 = \sum_{i=1}^{n} p_i \parallel x_i \parallel^2$$

- On va chercher $u \in \mathbb{R}^d$, $\parallel u \parallel = {}^t uu = 1$, tel que la variance totale de la projection de $\mathcal{N}_I$ sur $< u >$ soit maximale.

$$I(\mathcal{N}_I^{\perp}) = \sum_{i=1}^{n} p_i d\left(0, x_i^{\perp}\right)^2 = \sum_{i=1}^{n} p_i \parallel x_i^{\perp} \parallel^2$$

ACP : Deux approches, 1ère approche

- Or $x_i^{\perp} =< u, x_i > u = ({}^t x_i u)u$, $\| x_i^{\perp} \|^2 = ({}^t x_i u)^2$ et donc

$$I(\mathcal{N}_I^{\perp}) = \sum_{i=1}^{n} p_i ({}^t x_i u)^2 = {}^t u V u$$

où $V = {}^t XDX$ est la matrice variance.

- Notre problème, dans une ACP, est

$$\max_{u \in \mathbb{R}^d, \|u\|=1} {}^t u \, V \, u$$

- Solution u^1 sera appelée 1er facteur principale (*Loadings*).

ACP : Deux approches, 2ième approche

- Chercher une variable artificielle C, combinaison linéaire des variables $x^1, \ldots, x^d$, et qui soit de variance maximale.
- D'une façon équivalente : chercher les réels $u_1, \ldots, u_d$ tels que $C = \sum_{j=1}^{d} u_j x^j$ soit de variance maximale.
  Les u_j doivent vérifier $\sum_{j=1}^{d} (u_j)^2 = 1$.
- Notre problème, dans une ACP, est

$$\max_{u \in \mathbb{R}^d, \|u\|=1} {}^t u \, V \, u$$

- Solution C^1 sera appelée 1ère composante principale.

Solution

- Posons
$$\mathcal{L}(\lambda) = {}^t u\, V\, u = {}^t u\, V\, u - \lambda({}^t u u - 1)$$
- u est un maximum si
$$\frac{\partial}{\partial u}\, {}^t u\, V\, u = 0, \iff \frac{\partial}{\partial u}\mathcal{L}(\lambda) = 0$$
- Mais
$$\frac{\partial}{\partial u}\mathcal{L}(\lambda) = 2Vu - 2\lambda u$$
- Si u est un maximum de ${}^t u\, V\, u$ alors u est un vecteur propre de V de norme 1.
- V étant une matrice définie positive : V admet d valeurs propres positives
$$\lambda_1 \geq \lambda_2 \geq \ldots\ldots \geq \lambda_d.$$
- u est le vecteur propre associé à la plus grande valeur de V.

Solution...*suite*

- $u^1 = {}^t(u_1^1, \ldots, u_d^j) \in \mathbb{R}^d$ un vecteur propre, $\| u^1 \| = 1$, associé à λ_1.
- On a ${}^t(u^1)Vu^1 = \lambda_1 {}^t(u^1)u^1 = \lambda_1$.
- La coordonnée de chaque individu w_i sur $< u^1 >$ est ${}^t x_i u^1 = \sum_{j=1}^d x_i^j u_j^1$.
- Les coordonnées de tous les individus s'écrivent dans la matrice colonne suivante

$$C^1 = X\, u^1 = {}^t\left(\sum_{j=1}^d x_1^j u_j^1, \ldots, \sum_{j=1}^d x_n^j u_j^1\right) = \sum_{j=1}^d u_j^1 x^j$$

- $\mathrm{Var}(C^1) = {}^t(C^1)DC^1 = {}^t(u^1)\,{}^t XDXu^1 = {}^t(u^1)Vu^1 = \lambda_1$.

Précautions

- Comme $C^1 = u_1^1 x^1 + \ldots + u_j^1 x^j$, donc $x^1, \ldots, x^d$ doivent être mesurées avec la même unité.
- Comme $\mathrm{Var}(C^1)$ est maximale, donc il faut que les valeurs des variances de $x^1, \ldots, x^d$ ne soient pas très différentes les unes des autres.
- Solution : réduire toutes les variables $x^1, \ldots, x^d \longrightarrow z^1, \ldots, z^d$

$$z_i^j = \frac{x_i^j - \overline{x}^j}{\sigma^j}.$$

C'est l'Analyse en Composantes Principales normée.

- Par contre, si toutes les variables x^j sont des variables ordinales (enquêtes de satisfaction) il est souhaitable d'utiliser une ACP non-normée.

Les autres composantes principales

- Chercher un vecteur unitaire $u \in \mathbb{R}^d$ orthogonal à u^1 et tel que

$$\sum_{i=1}^{n} p_i \, d(x_i, x_i^{\perp})^2$$

soit minimale où $x_i^{\perp}$ pour tout $i = 1, \ldots, n$, est la projection orthogonale de x_i sur $< u >$.

- Solution : u^2 un vecteur propre associé à $\lambda_2 \leq \lambda_1$ et orthogonal à u^1.
- $C^2 = X\,u^2 = \sum_{j=1}^{d} u_j^2 x^j$ est la deuxième composante principale.
- C^1 et C^2 sont deux variables non corrélées :

$$\mathrm{Cov}(C^1, C^2) = {}^t(C^1)D(C^2) = {}^t(u^1)\,{}^tXDX(u^2) = \lambda_2\,{}^t(u^1)u^2 = 0$$

Propriétés

- On peut construire toutes les composantes principales $C^1, C^2, \ldots, C^d$ telles

$$C^k = X\,u^k = \sum_{j=1}^{d} u_j^k x^j, \quad \mathrm{Var}(C^k) = \lambda_k$$

où $u^1, \ldots, u^d$ est une base orthonormée de vecteurs propres de V :

$$< u^k, u^{k'} >= \delta_{kk'}\ ^1 \quad \text{et} \quad Vu^k = \lambda_k u^k.$$

- Posons $\mathbf{C} = (C^1 \mid \ldots \mid C^d)$ $(n \times d)$ et $\mathbf{u} = (u^1 \mid \ldots \mid u^d)$ $(d \times d)$:

$$\mathbf{C} = X\mathbf{u} = (C_i^k), \quad {}^t\mathbf{C}D\mathbf{C} = \Lambda \quad \text{et } \mathbf{u}\,{}^t\mathbf{u} = I_d \text{ où } \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_j \end{pmatrix}.$$

[1] Cette Conditionne entraîne que ${}^t\mathbf{u}\mathbf{u} = I_d \Rightarrow {}^t\mathbf{u} = \mathbf{u}^{-1}$ et donc $\mathbf{u}\,{}^t\mathbf{u} = I_d$

Reconstitution des données

- $\mathbf{C} = X\mathbf{u} \iff X = \mathbf{C}\,{}^t\mathbf{u}$

$\Rightarrow \forall,\ i = 1, \ldots, n,\ x_i = \sum_{k=1}^{d} C_i^k u^k.$

$\Rightarrow$ $(C_i^1, \ldots, C_i^d)$ sont les coordonnées de l'individu w_i dans la BON $\{u^1, \ldots, u^d\}$ et si $(x_i)_k^{\perp}$ est le projeté orthogonal de x_i sur $< u^k >$ (le $k-$ième axe principal) est

$$(x_i)_k^{\perp} = C_i^k u^k = \left(\sum_{j=1}^{d} x_i^j u_j^k \right) u^k.$$

- Pour représenter les individus sur le plan engendré par les deux premiers axes principaux, il suffit de considérer les deux premières colonnes de $\mathbf{C}$.

Qualité de la représentation

- $\mathcal{N}_I^k = \left\{(x_1)_k^\perp, \ldots, (x_n)_k^\perp\right\}$ est la projection du nuage $\mathcal{N}_I$ sur $< u^k >$.
- L'inertie totale du nuage $\mathcal{N}_I$ s'exprime par

$$I = \sum_{i=1}^{n} p_i d(0, x_i)^2 = \sum_{i=1}^{n} p_i \, {}^t(x_i) x_i = \mathrm{tr}({}^t XDX) = \sum_{j=1}^{d} (\sigma^j)^2 = \sum_{k=1}^{d} \lambda_k$$

- L'inertie du nuage $\mathcal{N}_I^k$ s'exprime par

$$I_k = \sum_{i=1}^{n} p_i d(0, (x_i)_k^\perp)^2 = \sum_{i=1}^{n} p_i \, {}^t\left((x_i)_k^\perp\right)(x_i)_k^\perp = {}^t(u^k) \, {}^t XDXu^k = \lambda_k.$$

- La qualité de la représentation sur $< u^k >$ se mesure par

$$\frac{\lambda_k}{\sum_{l=1}^{d} \lambda_l}.$$

Qualité de la représentation...

1. Contribution d'un individu w_i à l'inertie totale

$$\text{Ctr}(x_i) = \frac{p_i\, {}^t(x_i)x_i}{I} = \frac{\sum_{k=1}^{d}(C_i^k)^2}{\sum_{k=1}^{d}\lambda_k}$$

   $\text{Ctr}(x_i)$ permet d'indiquer la présence d'une observation aberrante.

2. Contribution d'un individu w_i à l'inertie du $k-$ième axe principal $< u^k >$

$$\text{Ctr}_k(x_i) = \frac{(C_i^k)^2}{\sum_{l=1}^{d}\lambda_l}$$

3. $\cos^2$ de l'angle entre x_i et $(x_i)_k^{\perp}$ mesure la qualité de la représentation de x_i sur $< u^k >$

$$\cos_k^2(x_i) = \frac{(C_i^k)^2}{{}^t(x_i)x_i}$$

Reconstitution des données

- $\mathbf{C} = X\mathbf{u} \iff X = \mathbf{C}\,{}^t\mathbf{u}$

$\Rightarrow \forall, j = 1, \ldots, d,\ x^j = \sum_{k=1}^{d} u_j^k C^k.$

- Comme $\{C^1, \ldots, C^d\}$ est une famille de vecteurs orthogonaux 2 à 2 dans $\mathbb{R}^n$ muni de la métrique D, $\forall\, k \neq k'$,

  $\mathrm{Cov}(C^k, C^{k'}) =< C^k, C^{k'} >= {}^t(C^k)D(C^{k'})$ et $\parallel C^k \parallel = \sqrt{\lambda_k}$.

  On considère la famille $\{Z^1, \ldots, Z^d\}$ où $\forall\, k = 1, \ldots, d$

$$Z^k = \frac{C^k}{\sqrt{\lambda_k}}$$

  $< Z^k >$ est appelé le $k-$ième axe unitaire.

- La projection orthogonale de x^j sur $< Z^k >$, s'exprime alors

$$(x^j)^{\perp} = u_j^k \sqrt{\lambda_k}\, Z^k.$$

  $\Rightarrow u_j^k \sqrt{\lambda_k}$ est la coordonnée de x^j sur $< Z^k >$.

Interprétation et qualités

- Rappelons que

$$\forall, j = 1, \ldots, d, \quad x^j = \sum_{k=1}^{d} u_j^k \sqrt{\lambda_k}\, Z^k$$

$$\forall, k = 1, \ldots, d, \quad C^k = \sum_{j=1}^{d} u_j^k x^j$$

$\Rightarrow$ $\mathrm{Var}(x^j) = \| x^j \|^2 = \sum_{k=1}^{d} (u_j^k)^2 \lambda_k$

$\Rightarrow$ $\mathrm{Ctr}_k(x^j \sim Z^k) = \dfrac{(u_j^k)^2 \lambda_k}{\mathrm{Var}(x^j)}$ est la contribution de Z^k dans l'explication de x^j.

Interprétation et qualités

- Comme $\sum_{j=1}^{d}(u_j^k)^2 = 1 \Rightarrow \lambda_k = \sum_{j=1}^{d}(u_j^k)^2\lambda_k$, et comme $\mathrm{Var}(C^k) = \lambda_k$ et $\mathrm{Var}((x^j)_k^{\perp}) = (u_j^k)^2\lambda_k$.

$\Rightarrow$ $\mathrm{Ctr}_k(C^k \sim x^j) = (u_j^k)^2$

- Calculons la corrélation entre x^j et C^k ?

$$\mathrm{Cor}(x^j, C^k) = \mathrm{Cor}(x^j, Z^k) = \frac{< x^j, Z^k >}{\sqrt{\mathrm{Var}(x^j)\,\mathrm{Var}(Z^k)}} = \frac{u_j^k\sqrt{\lambda_k}}{\sqrt{\mathrm{Var}(x^j)}}.$$

Si X est une matrice centrée-réduite (i.e., $\mathrm{Var}(x^j) = 1$, $\forall\, j$.

$$\mathrm{Cor}(x^j, C^k) = u_j^k\,\sqrt{\lambda_k}.$$

Qualité de représentation des variables

- On calcule de $\cos^2$ de l'angle entre x^j et $(x^j)_k^{\perp}$:

$$\cos^2\left(x^j,(x^j)_k^{\perp}\right)=\frac{\parallel (x^j)_k^{\perp}\parallel^2}{\parallel x^j\parallel^2}=\frac{(u_j^k)^2\lambda_k}{\mathrm{Var}(x^j)}=\left(\mathrm{Cor}(x^j,C^k)\right)^2.$$

- Dans le cas d'une ACP normée, $\mathrm{Var}(x^j)=\parallel x^j\parallel^2=1,\ \forall j$.
  - $\parallel (x^j)_k^{\perp}\parallel^2\leq\parallel x^j\parallel^2=1\Rightarrow$ la projection de x^j se trouve à l'intérieur d'un cercle de rayon : cercle de corrélation.
  - $\parallel (x^j)_k^{\perp}\parallel^2=\cos^2\left(x^j,(x^j)_k^{\perp}\right)=\left(\mathrm{Cor}(x^j,C^k)\right)^2=(u_j^k)^2\lambda_k$

$\Rightarrow$ Si $(x^j)^{\perp}$ est la projection de x^j sur le plan $< Z^1,Z^2 >$, alors $(x^j)^{\perp}$ se trouve à l'intérieure du cercle de corrélation.

Récapitulation

1. Centrer ou Centrer et réduire X, Calculer $V = {}^tXDX$
2. Diagonaliser V : Λ matrice diagonale des valeurs propres, u : matrices des vecteurs propres, ${}^t\mathbf{uu} = I_d$
3. Matrice coordonnées des individus dans $\{u^1, \ldots, u^d\}$:

$$\mathbf{C} = X\,\mathbf{u}$$

4. Matrice coordonnées des variables dans $\{Z^1, \ldots, Z^d\}$:

$$\mathbf{u}\sqrt{\Lambda}$$

Représentation d'une variable quantitatives dans $\mathbb{R}^n$.

1. $\mathbb{R}^n$ est l'espace des variables

2. $y = {}^t(y_1, \ldots, y_n)$ les observations d'une variable quantitative supplémentaire.
   La variable y a subi les mêmes transformations que $x^1, \ldots, x^d$: y est centrée (réduite) si $x^1, \ldots, x^d$ sont centrées (réduites).

3. On représente $(y^k)^\perp$ la projection orthogonale de y sur un $k-$ième axe unitaire $< Z^k >$

$$(y^k)^\perp = {}^t(y)DZ^k \cdot Z^k = \frac{1}{\sqrt{\lambda_k}} {}^t(y)DZ^k \cdot C^k$$

4. Comme $C^k = \sum_{j=1}^d u_j^k x^j$, $\Rightarrow (y^k)^\perp = \sum_{j=1}^d \frac{{}^t(y)DZ^k}{\sqrt{\lambda_k}} u_j^k \cdot x^j$.

Cœfficient de corrélation partiel

- Le cœfficient de corrélation partiel fourni par $(y^k)^\perp$ entre y et $x^1, \ldots, x^d$ est le cœfficient de corrélation entre y et $(y^k)^\perp$

$$R^k_{y|x^1,\ldots,x^d} = \text{Cor}(y, (y^k)^\perp) = \frac{< y, (y^k)^\perp >}{\| y \| \| (y^k)^\perp \|} = \frac{\| (y^k)^\perp \|}{\| y \|}$$

Comme $\| y \|^2 = \| (y^k)^\perp \|^2 + \| y - (y^k)^\perp \|^2$ (Pythagore).

Alors

$$\left(R^k_{y|x^1,\ldots,x^d}\right)^2 = 1 - \frac{\| y - (y^k)^\perp \|^2}{\| y \|^2}$$

mesure la *part* de $(y^k)^\perp$ dans l'explication de y.

- Si ACP-normée alors $(R^k_{y|x^1,\ldots,x^d})^2 = \| (y^k)^\perp \|^2$ et
- $\sum_{k=1}^d (R^k_{y|x^1,\ldots,x^d})^2$ mesure le cœfficient de corrélation partiel fourni par les composantes principales entre y et $x^1, \ldots, x^d$

Représentation d'une variable qualitative dans $\mathbb{R}^d$.

1. $\mathbb{R}^d$ l'espace des individus.
2. A est la variables qualitative à r modalités $\{A_1, \ldots, A_r\}$. On peut définir r sous-groupes d'individus $E_1, \ldots, E_r$ où
$$E_l = \{w_i \in E,\ A(w_i) = a_l\}$$
3. On définit $g_1, \ldots, g_r$ les centres de gravité resp. de $E_1, \ldots, E_r$: $\forall\, l = 1, \ldots, r$, g_l a pour coordonnées dans $\mathbb{R}^d$, $x_l = {}^t(x_l^1, \ldots, x_l^d)$ où
$$x_l^j = \sum_{w_i \in E_l} p_i x_i^j$$
4. On représente $(g_l)_k^{\perp}$ la projection orthogonale de g_l sur le $k-$ième axe principal $< u^k >$,
$$(g_l)_k^{\perp} = ({}^t g_l u^k) \cdot u^k$$

Représentation d'une variable qualitative dans $\mathbb{R}^d$.

- Décomposition de la variance :

  Variance Totale = Variance intra-classes + Variance inter-classes

- Appliquée à notre échantillon E

$$\sum_{i=1} p_i d(0, w_i)^2 = \sum_{l=1}^{r} q_l \sum_{w_i \in E_l} \frac{p_i}{q_l} d(w_i, g_l)^2 + \sum_{l=1}^{r} q_l d(0, g_l)^2$$

  où $q_l = \sum_{w_i \in E_l} p_i$.

- Pouvoir de séparation de la $k-$ième composante est

$$\Lambda_k = \frac{1}{\lambda_k} \sum_{l=1}^{r} q_l d(0, (g_l)_k^{\perp})^2$$

  Λ_k est le rapport entre la variance totale de E^k la projection de E sur $< u^k >$ et la variance inter-classe de E^k, $\Lambda_k \in [0, 1]$.

Individus supplémentaires

- w_s un individu supplémentaire qu'on a envie de le représenter sur un $k-$ième axe principal $< u^k >$.
- $\forall\, j = 1 \ldots d$, $x_s^j = x^j(w_s)$. Donc w_s a pour coordonnées dans $\mathbb{R}^d$: $x_s = {}^t(x_s^1, \ldots, x_s^d)$.
- D'abord, il faut faire subir à x_s les mêmes transformations que les x_i :

$$\text{centrer } x_s^j - \overline{x}^j \quad \text{ou centre et réduire } \frac{x_s^j - \overline{x}^j}{\sigma^j}$$

où $\overline{x}^j$ et σ^j sont resp. la moyenne et l'écart-type de la variable x^j (sans l'observation x_s).

- La représentation de w_s sur $< u^k >$ se fait par projection orthogonale :

$$(x_s)^{\perp} = ({}^t x_s u^k) \cdot u^k.$$

Les données `decathlon`

- Importation des données à partir du package

```
> library(FactoMineR)
> data(decathlon)
```

- Utilisation du package `FactoMineR`

```
> res.pca <- PCA(decathlon,
+ quanti.sup=11:12,quali.sup=13)
```

- 11 et 12 étant les numéros des colonnes des variables "rang" et "score obtenues" : variables traitées comme des variables quantitatives supplémentaires.
- 13 est le numéro de la variable "compétition" : variable qualitative supplémentaire.

Sortie PCA de FacoMineR

```
>  res.pca
**Results  for  the  Principal  Component  Analysis  (PCA))**
name description
1    "$eig" "eigenvalues"
2    "$var" "results  for  the  variables"
3    "$var$coord" "coordinates  of  the  variables"
4    "$var$cor" "correlations  variables  -  dimensions"
5    "$var$cos2" "cos2  for  the  variables"
6    "$var$contrib" "contributions  of  the  variables"
7    "$ind" "results  for  the  individuals"
8    "$ind$coord" "coord.  for  the  individuals"
9    "$ind$cos2" "cos2  for  the  individuals"
10  "$ind$contrib" "contributions  of  the  individuals"
11  "$quanti.sup" "results  for  the  supplementary  quantitative  variables"
12  "$quanti.sup$coord"  "coord.  of  the  supplementary  quantitative  variables"
13  "$quanti.sup$cor"     "correlations  suppl.  quanti.  variables  -  dimensions"
14  "$quali.sup" "results  for  the  supplementary  categorical  variables"
15  "$quali.sup$coord"    "coord.  of  the  supplementary  categories"
16  "$quali.sup$vtest"    "v-test  of  the  supplementary  categories"
17  "$call" "summary  statistics"
18  "$call$centre" "mean  for  the  variables"
19  "$call$ecart.type"    "standard  error  for  the  variables"
20  "$call$row.w" "weights  for  the  individuals"
21  "$call$col.w" "weights  for  the  variables"
```

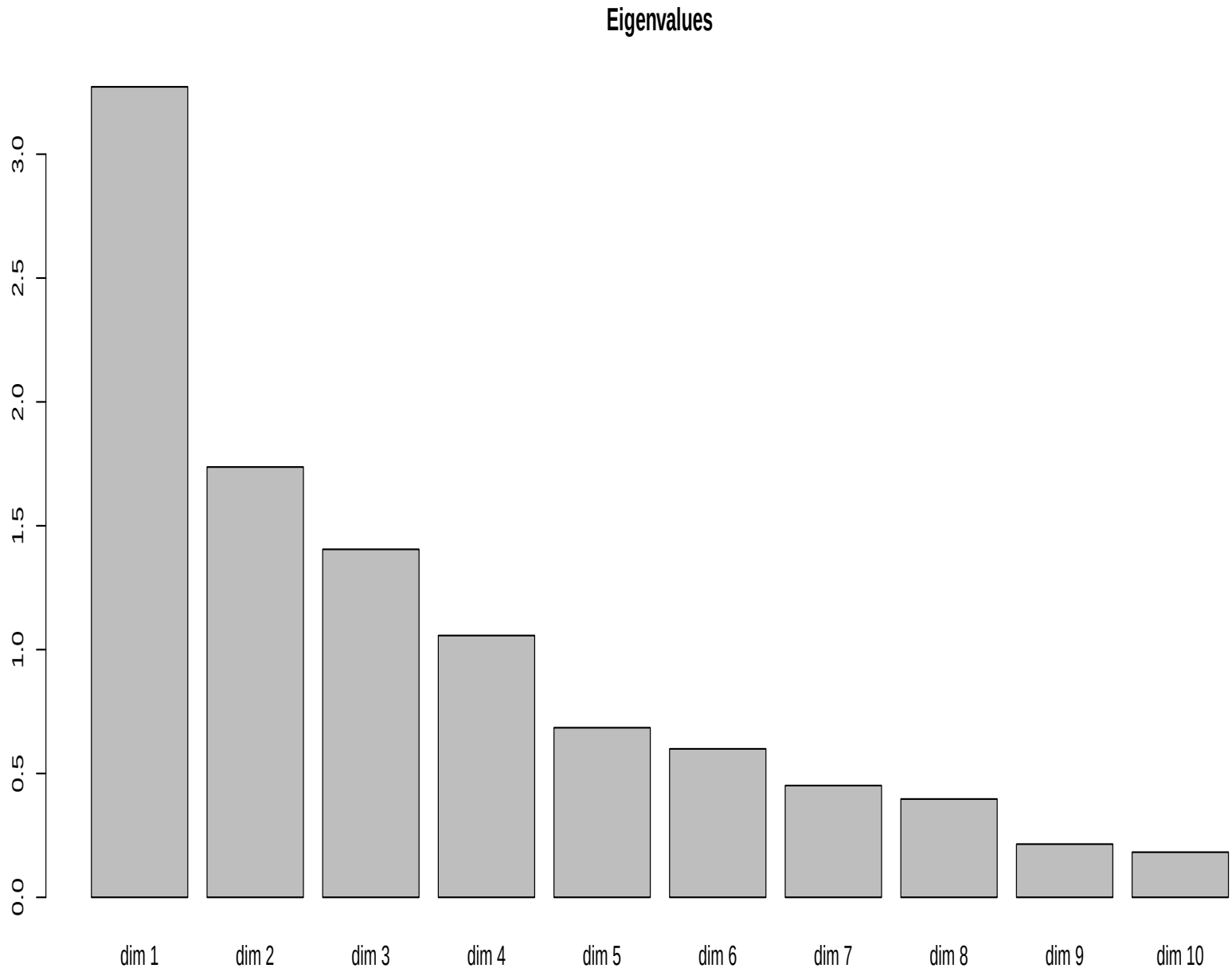
Choix du nombre des axes

- On examine la part de variance de chaque axe :

```
> round(res.pca$eig,3)
        eigenvalue percentage of variance cumulative percentage of variance
comp 1       3.272                 32.719                            32.719
comp 2       1.737                 17.371                            50.090
comp 3       1.405                 14.049                            64.140
comp 4       1.057                 10.569                            74.708
comp 5       0.685                  6.848                            81.556
comp 6       0.599                  5.993                            87.548
comp 7       0.451                  4.512                            92.061
comp 8       0.397                  3.969                            96.030
comp 9       0.215                  2.148                            98.178
comp 10      0.182                  1.822                           100.000
```

- On trace un graphe appelé ébouli des valeurs propres (scree plot) :

```
>  barplot(res.pca$eig[,1],main="Eigenvalues",
+     names.arg=paste("dim",1:nrow(res.pca$eig)))
```


Eigenvalues
0.0
0.5
1.0
1.5
2.0
2.5
3.0
dim 1
dim 2
dim 3
dim 4
dim 5
dim 6
dim 7
dim 8
dim 9
dim 10

Etude du nuage des points

- Les coordonnées des individus sur les quatre premiers axes principaux :

  ```
  > res.pca$ind$coord[,1:4]
  ```

- La qualité de la représentation de chaque individu sur les quatre premiers axes principaux.

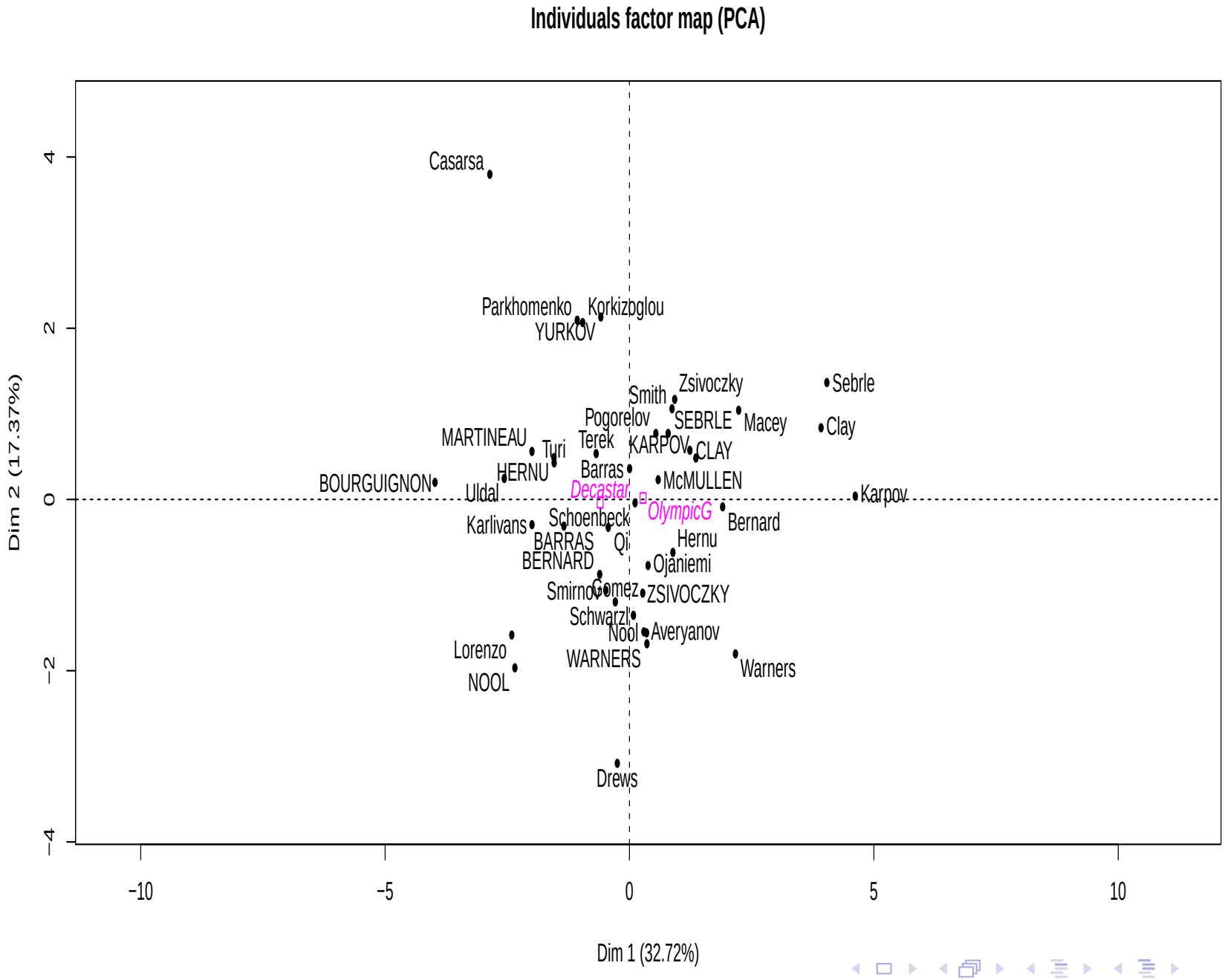
  ```
  > res.pca$ind$cos2[,1:4]
  ```

- La contribution de chaque individu dans la variance des quatre premiers axes principaux.

  ```
  > res.pca$ind$contrib[,1:4]
  ```

- Pour représenter les individus sur les deux premiers axes principaux.

  ```
  >  plot(res.pca,choix="ind",axes=1:2)
  ```


Individuals factor map (PCA)
Dim 2 (17.37%)
Dim 1 (32.72%)
Casarsa
Parkhomenko
Korkizoglou
YURKOV
Zsivoczky
Smith
Sebrle
Pogorelov
SEBRLE
Macey
Clay
MARTINEAU
Turi
Terek
KARPOV
CLAY
HERNU
Barras
BOURGUIGNON
McMULLEN
Uldal
Decastar
OlympicG
Karpov
Karlivans
Schoenbeck
Bernard
BARRAS
Qi
Hernu
BERNARD
Ojaniemi
Smirnov
Gomez
ZSIVOCZKY
Schwarzl
Nool
Averyanov
Lorenzo
WARNERS
Warners
NOOL
Drews

Etude de la représentation des variables

- Les coordonnées des variables sur les quatre premiers axes principaux :

  ```
  > res.pca$var$coord[,1:4]
  ```

- La qualité de la représentation de chaque variable sur les quatre premières axes principales.

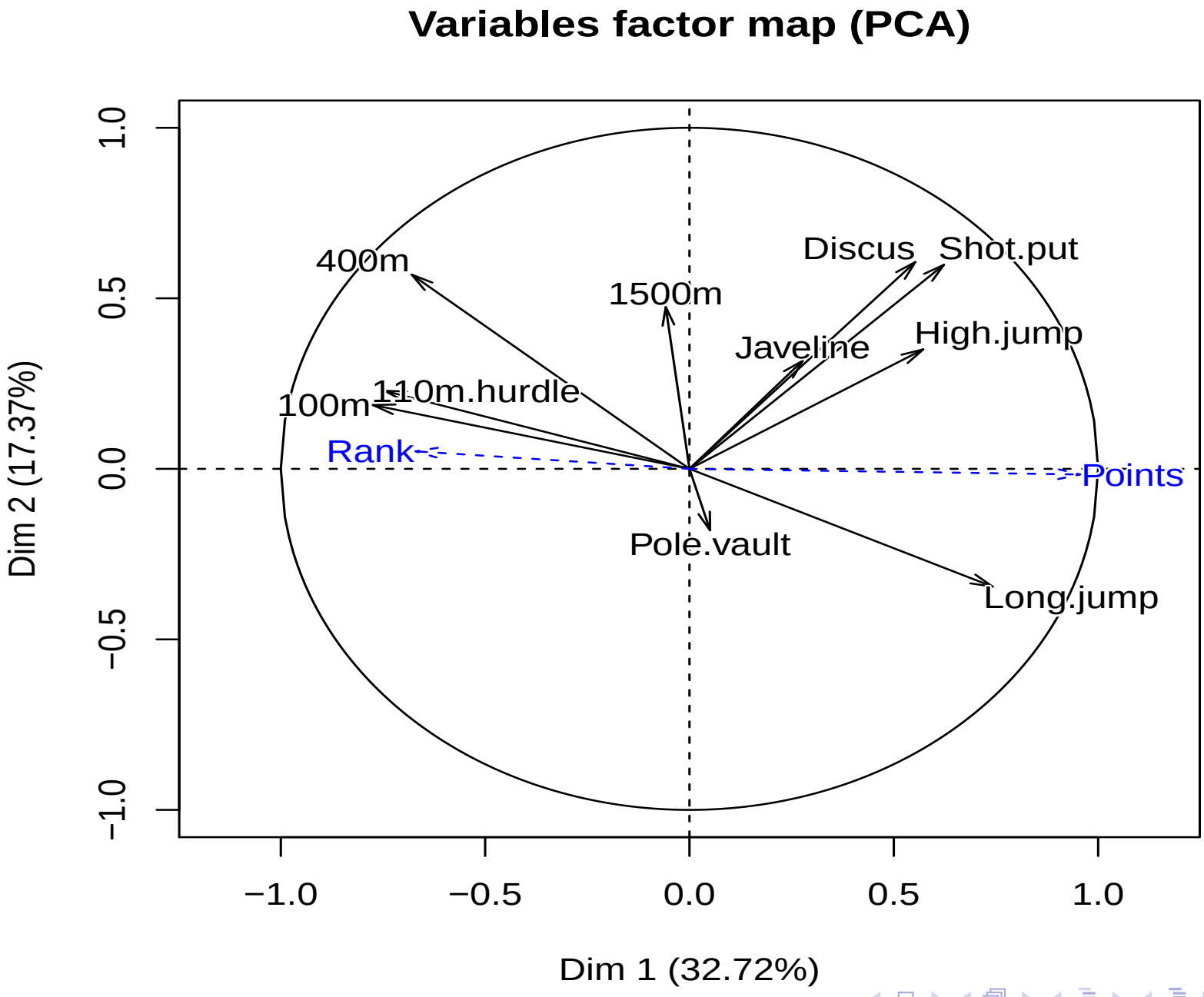
  ```
  > res.pca$ind$cos2[,1:4]
  ```

- La contribution de chaque variable dans la variance des quatre premières composantes principales.

  ```
  > res.pca$var$contrib[,1:4]
  ```

- Pour représenter les individus sur les deux premières composantes principales.

  ```
  >  plot(res.pca,choix="var",axes=1:2)
  ```


Variables factor map (PCA)
Dim 2 (17.37%)
Dim 1 (32.72%)
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
400m
1500m
Discus
Shot.put
Javeline
High.jump
110m.hurdle
100m
Rank
Points
Pole.vault
Long.jump

Relation variables ACP

- Test de corrélation composantes principales et variables actives.

```
  >  dimdesc(res.pca)
$Dim.1
$Dim.1$quanti
           correlation      p.value
Points       0.9561543 0.000000e+00
Long.jump    0.7418997 2.849886e-08
...
$Dim.2
$Dim.2$quanti
           correlation      p.value
Discus       0.6063134 2.650745e-05
Shot.put     0.5983033 3.603567e-05
```

- Test des moyennes des coordonnées pour la variable quantitative supplémentaire (Test d'ANOVA).

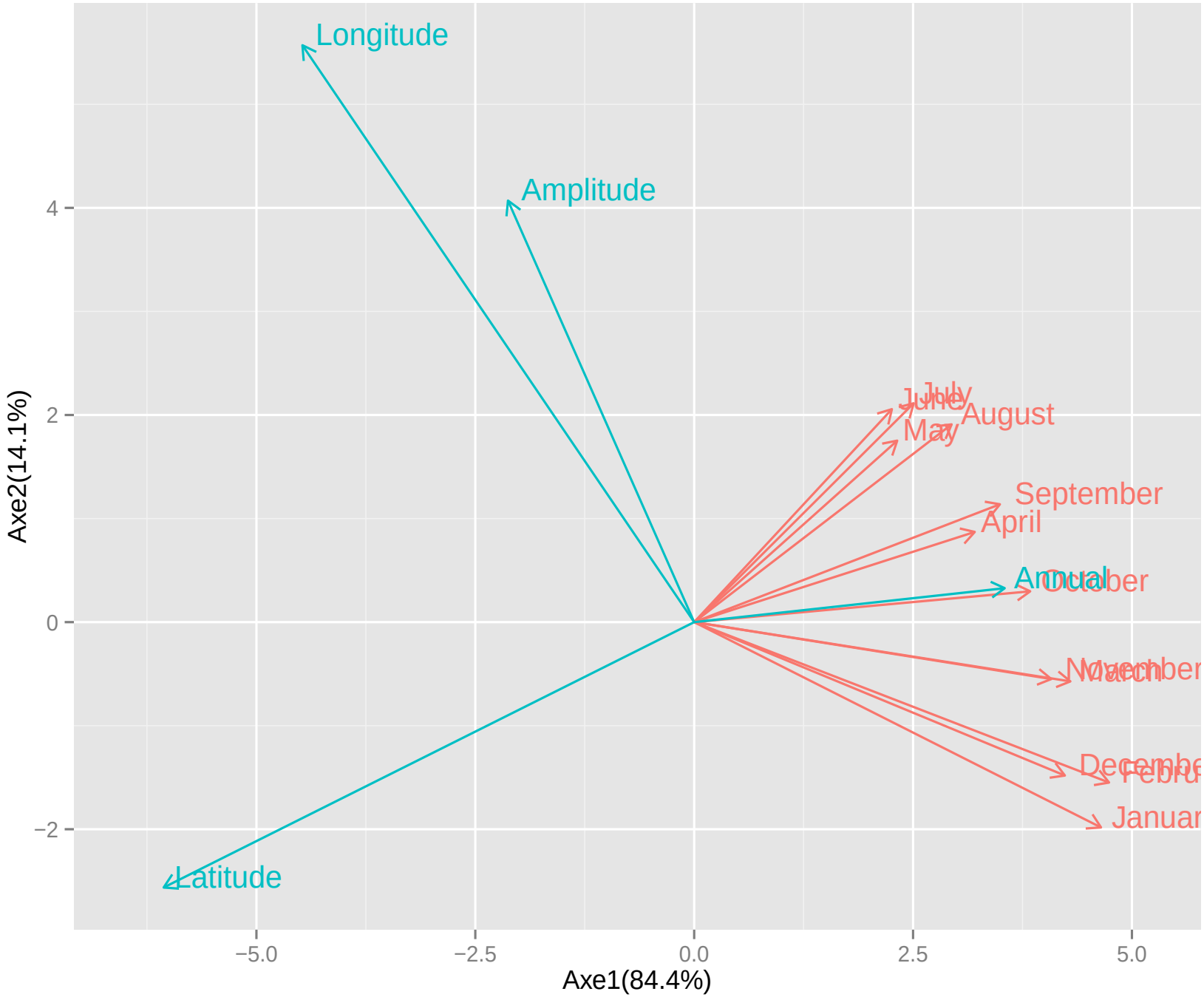
```
$Dim.1$quali
                    R2   p.value
Competition 0.05110487 0.1552515

$Dim.1$category
           Estimate   p.value
OlympicG  0.4393744 0.1552515
Decastar -0.4393744 0.1552515
```

Les données `temperature`

- Positionner les 23 capitales selon leurs températures moyennes mensuelles.
- Les variables sont de même unités : on ne réduit pas les données.
- Mise en œuvre sous `R`

```
>  library(FactoMineR)
>  temperature  <-  read.table("temperature.csv",
+  header=TRUE,sep=";",dec=".",row.names=1)
>  res<-PCA(temperature,scale.unit=F,
+ ind.sup=24:35,quanti.sup=13:16,quali.sup=17)
```


Longitude
Amplitude
Latitude
July
June
August
May
September
April
Annual
October
November
March
December
February
January
4
2
0
−2
−5.0
−2.5
0.0
2.5
5.0
Axe2(14.1%)
Axe1(84.4%)