

# Analyse en composantes principales

Christophe Ambroise



## Section 1

### **Introduction**

# Introduction I

## Objectif

Les *méthodes factorielles* ont pour objectif de

- visualiser, et plus généralement,
- traiter des données multidimensionnelles,

## Redondance

La prise en compte simultanée de nombreuses variables est un problème difficile ; Heureusement, l'information apportée par ces variables est souvent redondante

## Une solution

Remplacer les variables initiales par un nombre réduit de nouvelles variables sans perdre trop d'information.

## Introduction II

### Principes

Par exemple, lorsque les variables sont toutes quantitatives, l'analyse en composantes principales (ACP) va chercher à résoudre ce problème en

- considérant que les nouvelles variables sont des combinaisons linéaires des variables initiales
- non corrélées

### Tableau original vers tableau synthétique

On passe d'un tableau original  $X$  à un tableau synthétique avec le même nombre de lignes mais un nombre de colonnes réduit  $C$ .

### Historique

Cette méthode a d'abord été développée par K. Pearson (1900) pour deux variables, puis par H. Hotelling (1933) qui l'a étendue à un nombre quelconque de variables.

## Maximisation de la variance des données projetées I

### Le nuage des individus

Le tableau de données  $X$  est une  $n \times p$  matrice réelle:

- chaque lignes  $X_i$  décrit un individu par  $p$  variables
- chaque colonne  $X^j$  décrit une variable par  $n$  individus

### Centrage de la matrice $X$

Le nuage des individus est centré autour du centre de gravité du nuage (ou vecteur des moyennes empiriques):

$$\bar{X} = \frac{1}{n} X_i = \begin{pmatrix} X_{i1} \\ X_{i2} \\ \dots \\ X_{ip} \end{pmatrix}$$

Sans perte de généralité nous supposerons que ce vecteur moyenne est le vecteur nul (il suffit de centrer la matrice  $X$  originale ).

## Maximisation de la variance des données projetées II

### Variance empirique

La variance empirique du nuage est la somme des variance de chaque variable:

$$\hat{\sigma}^2 = \sum_{j=1}^p \hat{\sigma}_j^2$$

avec  $\hat{\sigma}_j^2 = \frac{1}{n} \sum_i X_{ij}^2$

### Lien entre variance empirique et inertie

$$n\hat{\sigma}^2 = \sum_i \sum_{j=1}^p X_{ij}^2$$

peut s'interpréter aussi comme la somme des distances des individus au centre du nuage.

## Projection des individus I

### Projection sur l'axe $u_1$

La projection vectorielle du vecteur  $X_i$  sur la droite vectorielle de vecteur directeur  $u_1$  est défini par

$$c_{i1} u_1$$

où  $c_{i1} = \langle X_i, u_1 \rangle$  est la coordonnée de  $X_i$  dans la base  $\{u_1\}$ .

### Projection sur le sous espace vectoriel de base $u_1, \dots, u_d$

La projection vectorielle du vecteur  $X_i$  sur le s.e.v. de base  $u_1, \dots, u_d$  est défini par le vecteur

$$c_{i1} u_1 + \dots + c_{id} u_d$$

où  $c_{ik} = \langle X_i, u_k \rangle$  est la  $k$ ième coordonnées de  $X_i$  dans base.

## Matrice de variance covariance empirique et diagonalisation I

La matrice  $S = \frac{1}{n}X^tX$  est une estimation de la matrice de variance. C'est une matrice symétrique définie positive. En effet,  $S = S^t$  et

$$\mathbf{y}^t S \mathbf{y} = \mathbf{y}^t \frac{1}{n} \sum_i X_i X_i^t \mathbf{y} \quad (1)$$

$$= \frac{1}{n} \sum_i (\mathbf{y}^t X_i)(X_i^t \mathbf{y}) = \sum_i \|\mathbf{y}^t X_i\|^2 \geq 0 \quad (2)$$

### Interprétation

- les termes de la diagonale sont les variances empiriques:

$$\hat{\sigma}_j^2 = \frac{1}{n} \sum_i X_{ij}^2$$

- les termes hors diagonales sont les covariances empiriques

$$\hat{\rho}_{jk} = \frac{1}{n} \sum_i X_{ij} X_{ik}$$



## Variance projetée sur un axe I

On cherche à trouver  $\mathbf{v}$  tel que la projection des individus de  $X$  sur le vecteur  $\mathbf{v}$  (projection vectorielle) soit maximum:

$$\begin{cases} \max_{\mathbf{v}} \mathbf{v}^t S \mathbf{v}, \\ \mathbf{v}^t \mathbf{v} = 1. \end{cases}$$

avec  $S = \frac{1}{n} X^t X$

Si l'on exprime  $\mathbf{v}$  dans la base (orthonormée) des vecteurs propres de  $S$ ,

$$\mathbf{v} = \sum_{j=1}^p \alpha_j \mathbf{u}_j$$

alors le problème précédent devient

$$\begin{cases} \max_{\alpha_1, \dots, \alpha_d} (\sum_{j=1}^p \alpha_j \mathbf{u}_j)^t U D U^t (\sum_{j=1}^p \alpha_j \mathbf{u}_j), \\ \sum_j \alpha_j^2 = 1. \end{cases}$$

## Variance projetée sur un axe II

$$\begin{cases} \max_{\alpha_1, \dots, \alpha_d} (\sum_{j=1}^p \alpha_j^2 \lambda_j), \\ \sum_j \alpha_j^2 = 1. \end{cases}$$

où  $\lambda_j$  est la  $j$ ème valeur propre.

### Solution

L'équation donne donc un barycentre sur la demi droite des réels positifs entre  $\lambda_1$  et  $\lambda_p$ . La valeur max du barycentre est  $\lambda_1$ , et elle est obtenue pour  $\alpha_1 = 1$  et  $\alpha_j = 0, \forall j \neq 1$  (car tous les  $\lambda_j$  sont positifs). Le vecteur solution est donc le vecteur propre de  $S$  associé à la plus grand valeur propre  $\lambda_1$ . La projection des  $X_i$  sur  $\mathbf{u}_1$  est la première composante principale

$$C^1 = (c_{11}, \dots, c_{n1})^t$$

## Variance projetée sur un sous espace vectorielle

On admettra que le sous espace vectoriel de dimension  $k$  qui maximise la variance projetée est le sous espace vectoriel engendré par les  $k$  premiers vecteur propres de  $S$ .

### Composantes principales

Les projections des  $X_i$  sur les vecteurs propres  $u_j$  sont les composantes principales:

$$C = (C^1 \dots C^p) = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1p} \\ \vdots & \vdots & \vdots & \vdots \\ c_{n1} & c_{n2} & \dots & c_{np} \end{pmatrix}$$

## Relation entre données originales et composantes principales

$$C = XU$$

## Formule de reconstitution

$$X = \sum_j C^j \mathbf{u}_j^t$$

La dernière relation montre que l'on peut reconstituer le tableau initial avec les composantes principales et les axes principaux. Cette relation est appelée formule de reconstitution. Si on se limite aux  $k$  ( $k < p$ ) premiers termes, on obtient une approximation du tableau initial :

## Qualité de la représentation

### Qualité globale

La qualité globale de représentation de l'ensemble initial  $X$  sur les  $k$  premières composantes principales est mesuré comme le pourcentage de variance expliquée :

$$\frac{\lambda_1 + \dots + \lambda_k}{\text{trace}(S)} 100.$$

## Contribution relative d'un axe à un individu

Sachant que l'inertie totale du nuage est  $\frac{1}{n} \sum_{i=1}^p \|X_i\|^2$ , la quantité  $\frac{1}{n} \|X_i\|^2$  représente la part d'inertie apportée par chaque  $X_i$ .

Après projection sur l'axe  $u_\alpha$ , l'inertie restante est donc  $\frac{1}{n} (c_\alpha^i)^2$ . Chacun des termes  $\frac{1}{n} (c_\alpha^i)^2$  représente donc la part de l'inertie initial  $\frac{1}{n} \|X_i\|^2$  qu'apportait l'individu  $i$ , conservée par l'axe  $\alpha$ . Le rapport de ces deux quantités est appelée *contribution relative* du  $\alpha$  axe factoriel à l'individu  $i$  et elle est notée  $COR(i, \alpha)$  :

$$COR(i, \alpha) = \frac{(c_\alpha^i)^2}{\|X_i\|^2}.$$

Cette quantité représente aussi le carré du cosinus de l'angle formé par l'individu  $X_i$  et par le vecteur  $u_\alpha$ . Si  $COR(i, \alpha)$  est proche de 1, l'individu est bien représenté par cet axe, si  $COR(i, \alpha)$  est au contraire proche de 0, l'individu est très mal représenté par cet axe.

$$QLT(i, k) = \frac{\sum_{\alpha=1}^k (c_\alpha^i)^2}{\|X_i\|^2} = \sum_{\alpha=1}^k COR(i, \alpha).$$

## Contribution relative d'un individu à un axe

En partant de la relation  $\lambda_\alpha = \frac{1}{n} \sum_{i=1}^n (c_\alpha^i)^2$ , on peut décomposer  $\lambda_\alpha$ , l'inertie conservée par l'axe  $\mathbf{u}_\alpha$ , selon les individus. On définit alors la contribution relative de l'individu  $i$  à l'axe  $\alpha$ , notée  $CTR(i, \alpha)$  : c'est la part d'inertie du  $\alpha$ axe pris en compte (ou expliquée) par l'individu  $i$ . Nous avons :

$$CTR(i, \alpha) = \frac{1}{n} \frac{(c_\alpha^i)^2}{\lambda_\alpha}.$$



## Section 2

### **Interprétation des nouvelles variables**

## Cercle des corrélations

- Chaque ancienne variable possède une corrélation avec les nouvelles variables.
- Ces corrélations sont utilisées pour interpréter les nouvelles variables en fonctions des anciennes.

$$\begin{aligned}\text{cor}(X^j, C^k) &= \text{cor}(X^j, X\mathbf{u}_k) = \frac{\text{cov}(X^j, X\mathbf{u}_k)}{\sqrt{\mathbb{V}(X^j)\mathbb{V}(X\mathbf{u}_k)}} \\ &= \frac{\lambda_k u_k^j}{\sqrt{\frac{1}{n} \|X^j\|^2 \lambda_k}}.\end{aligned}$$

car

- $\mathbb{V}(X\mathbf{u}_k) = \mathbf{u}_k^\top \mathbf{S} \mathbf{u}_k = \lambda_k$  (voir calcul précédent)
- $\text{cov}(X^j, X\mathbf{u}_k) = \frac{1}{n} (X^j)^\top X\mathbf{u}_k$  est la  $j$ -ième coordonnée de  $\frac{1}{n} X^\top X\mathbf{u}_k = \lambda_k \mathbf{u}_k$ .

Si les variables ont été préalablement normalisées, on obtient

$$\text{cor}(X^j, C^k) = \sqrt{\lambda_k} u_k^j.$$

## Section 3

### Un exemple d'ACP

## Les données

Il s'agit du tableau de notes décrits. Rappelons que ces données regroupent les notes obtenues par neuf élèves dans les matières mathématiques, sciences, français, latin et dessin :

**Table 1:** Notes de 9 élèves

|          | math | scie | fran | lati | d.m |
|----------|------|------|------|------|-----|
| jean     | 6.0  | 6.0  | 5.0  | 5.5  | 8   |
| aline    | 8.0  | 8.0  | 8.0  | 8.0  | 9   |
| annie    | 6.0  | 7.0  | 11.0 | 9.5  | 11  |
| monique  | 14.5 | 14.5 | 15.5 | 15.0 | 8   |
| didier   | 14.0 | 14.0 | 12.0 | 12.5 | 10  |
| andré    | 11.0 | 10.0 | 5.5  | 7.0  | 13  |
| pierre   | 5.5  | 7.0  | 14.0 | 11.5 | 10  |
| brigitte | 13.0 | 12.5 | 8.5  | 9.5  | 12  |
| evelyne  | 9.0  | 9.5  | 12.5 | 12.0 | 18  |

## Centrage du tableau de données

Les moyennes des cinq variables sont respectivement 9.67, 9.83, 10.22, 10.05 et 11. Le tableau centré en colonne  $X$  est obtenu en soustrayant à chaque colonne la moyenne correspondante :

```
X<- scale(X,center=TRUE,scale = FALSE)
knitr::kable(X,format="latex", caption = "Tableau centré",digits = 2)
```

Table 2: Tableau centré

|          | math  | scie  | fran  | lati  | d.m |
|----------|-------|-------|-------|-------|-----|
| jean     | -3.67 | -3.83 | -5.22 | -4.56 | -3  |
| aline    | -1.67 | -1.83 | -2.22 | -2.06 | -2  |
| annie    | -3.67 | -2.83 | 0.78  | -0.56 | 0   |
| monique  | 4.83  | 4.67  | 5.28  | 4.94  | -3  |
| didier   | 4.33  | 4.17  | 1.78  | 2.44  | -1  |
| andré    | 1.33  | 0.17  | -4.72 | -3.06 | 2   |
| pierre   | -4.17 | -2.83 | 3.78  | 1.44  | -1  |
| brigitte | 3.33  | 2.67  | -1.72 | -0.56 | 1   |
| evelyne  | -0.67 | -0.33 | 2.28  | 1.94  | 7   |

## Matrice de variance

$$S = \frac{1}{9} X' X$$

```
n<-nrow(X)
p<-ncol(X)
S<-var(X)*(n-1)/n
knitr::kable(S,format="latex", caption = "Matrice de variance",digits = 2)
```

**Table 3:** Matrice de variance

|      | math  | scie | fran  | lati | d.m  |
|------|-------|------|-------|------|------|
| math | 11.39 | 9.92 | 2.66  | 4.82 | 0.11 |
| scie | 9.92  | 8.94 | 4.12  | 5.48 | 0.06 |
| fran | 2.66  | 4.12 | 12.06 | 9.29 | 0.39 |
| lati | 4.82  | 5.48 | 9.29  | 7.91 | 0.67 |
| d.m  | 0.11  | 0.06 | 0.39  | 0.67 | 8.67 |

## Axes principaux d'inertie

La diagonalisation de la matrice de variance fournit les valeurs propres suivantes (rangées par ordre décroissant)

$$\lambda_1 = 28.2533, \lambda_2 = 12.0747, \lambda_3 = 8.6157, \lambda_4 = 0.0217, \lambda_5 = 0.0099.$$

et les vecteurs propres normés ou axes principaux d'inertie suivants

$$\mathbf{u}_1 = \begin{pmatrix} 0.51 \\ 0.51 \\ 0.49 \\ 0.48 \\ 0.03 \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} -0.57 \\ -0.37 \\ 0.65 \\ 0.32 \\ 0.11 \end{pmatrix}, \mathbf{u}_3 = \begin{pmatrix} -0.05 \\ -0.01 \\ 0.11 \\ 0.02 \\ -0.99 \end{pmatrix}, \mathbf{u}_4 = \begin{pmatrix} 0.29 \\ -0.55 \\ -0.39 \\ 0.67 \\ -0.03 \end{pmatrix}, \mathbf{u}_5 = \begin{pmatrix} -0. \\ 0. \\ -0. \\ 0. \\ -0. \end{pmatrix}$$

## Qualité de la représentation

- les inerties du nuage projeté sur les 5 axes sont égales aux valeurs propres.
- l'inertie du nuage est égale à  $\text{trace}(S)$ , c'est-à-dire aussi à la somme des valeurs propres, ici 48.975.
- les pourcentages d'inertie expliquée par chaque axe sont donc de 57.69, 24.65, 17.59, 0.04 et 0.02.
- Les pourcentages d'inertie expliquée par les sous-espaces principaux sont 57.69, 82.34, 99.94, 99.98 et 100.00.
- le nuage initial est pratiquement dans un espace de dimension 3.



## Composantes principales $C = XU$

```
U<-eigen(S)$vectors ; Lambda<-eigen(S)$values ; C = X%*%U
knitr::kable(C,format="latex",
             caption = "Composantes principales",digits = 2)
```

**Table 4:** Composantes principales

|          |       |       |       |       |       |
|----------|-------|-------|-------|-------|-------|
| jean     | -8.70 | 1.70  | 2.55  | -0.15 | -0.12 |
| aline    | -3.94 | 0.71  | 1.81  | -0.09 | 0.04  |
| annie    | -3.21 | -3.46 | 0.30  | 0.17  | 0.02  |
| monique  | 9.76  | -0.22 | 3.34  | -0.17 | 0.10  |
| didier   | 6.37  | 2.17  | 0.96  | 0.07  | -0.19 |
| andré    | -2.97 | 4.65  | -2.63 | -0.02 | 0.15  |
| pierre   | -1.05 | -6.23 | 1.69  | 0.12  | 0.04  |
| brigitte | 1.98  | 4.07  | -1.40 | 0.24  | 0.01  |
| evelyne  | 1.77  | -3.40 | -6.62 | -0.16 | -0.06 |

Ces composantes principales permettent d'obtenir, par exemple, les plans de représentation 1,2 et 1,3 suivants

## Contributions relatives des axes aux individus

```
COR<- C^2 / rowSums(X^2)
knitr::kable(COR,format="latex",
             caption = "Contribution relative des axes aux individus",
             digits = 2)
```

**Table 5:** Contribution relative des axes aux individus

|          |      |      |      |   |   |
|----------|------|------|------|---|---|
| jean     | 0.89 | 0.03 | 0.08 | 0 | 0 |
| aline    | 0.80 | 0.03 | 0.17 | 0 | 0 |
| annie    | 0.46 | 0.53 | 0.00 | 0 | 0 |
| monique  | 0.89 | 0.00 | 0.11 | 0 | 0 |
| didier   | 0.88 | 0.10 | 0.02 | 0 | 0 |
| andré    | 0.24 | 0.58 | 0.19 | 0 | 0 |
| pierre   | 0.03 | 0.91 | 0.07 | 0 | 0 |
| brigitte | 0.17 | 0.74 | 0.09 | 0 | 0 |
| evelyne  | 0.05 | 0.20 | 0.75 | 0 | 0 |

## Contributions relatives des individus aux axes

```
CTR<- 1/n* C^2 / matrix(eigen(S)$values,n,p,byrow = TRUE)
knitr::kable(CTR,format="latex",
             caption = "Contributions relatives des individus aux axes",
             digits = 2)
```

**Table 6:** Contributions relatives des individus aux axes

|          |      |      |      |      |      |
|----------|------|------|------|------|------|
| jean     | 0.30 | 0.03 | 0.08 | 0.11 | 0.15 |
| aline    | 0.06 | 0.00 | 0.04 | 0.04 | 0.02 |
| annie    | 0.04 | 0.11 | 0.00 | 0.15 | 0.00 |
| monique  | 0.37 | 0.00 | 0.14 | 0.15 | 0.11 |
| didier   | 0.16 | 0.04 | 0.01 | 0.03 | 0.40 |
| andré    | 0.03 | 0.20 | 0.09 | 0.00 | 0.25 |
| pierre   | 0.00 | 0.36 | 0.04 | 0.07 | 0.02 |
| brigitte | 0.02 | 0.15 | 0.03 | 0.30 | 0.00 |
| evelyne  | 0.01 | 0.11 | 0.56 | 0.14 | 0.04 |

## Analyse dans $R^n$

Les vecteurs  $\mathbf{d}^\alpha$ , composantes principales associées aux différentes variables, sont formés des coordonnées de toutes les variables pour un même axe  $\mathbf{v}_\alpha$  et vérifient la relation

$$\mathbf{d}^\alpha = \sqrt{\lambda_\alpha} \mathbf{u}_\alpha.$$

On obtient

```
D<- U * matrix(sqrt(Lambda),p,p,byrow=TRUE)
knitr::kable(D,format="latex",
              caption = "Variables",digits = 2)
```

**Table 7:** Variables

|      |       |       |       |       |
|------|-------|-------|-------|-------|
| 2.73 | 1.97  | -0.15 | -0.04 | 0.06  |
| 2.69 | 1.29  | -0.04 | 0.08  | -0.05 |
| 2.62 | -2.26 | 0.32  | 0.06  | 0.04  |
| 2.58 | -1.12 | 0.07  | -0.10 | -0.05 |
| 0.16 | -0.39 | -2.91 | 0.01  | 0.00  |

## Analyse dans $R^n$

Il est souvent préférable de représenter la projection des variables initiales normées. Il suffit de diviser chaque ligne du tableau précédent par la norme de la variables correspondante

$$\|\mathbf{x}^j\|^2 = \frac{1}{9} \sum_{i=1}^9 (x_i^j)^2.$$

Les  $\|\mathbf{x}^j\|$  correspondent en fait aux écarts-type des variables. On obtient respectivement 3.37, 2.99, 3.47, 2.81 et 2.94

```
F <- D / sqrt((1/n*colSums(X^2)))
knitr::kable(F,format="latex",
  caption = "Variables normées",digits = 2)
```

**Table 8:** Variables normées

|      |       |       |       |       |
|------|-------|-------|-------|-------|
| 0.81 | 0.58  | -0.04 | -0.01 | 0.02  |
| 0.90 | 0.43  | -0.01 | 0.03  | -0.02 |
| 0.75 | -0.65 | 0.09  | 0.02  | 0.01  |
| 0.92 | -0.40 | 0.02  | -0.04 | -0.02 |
| 0.06 | -0.13 | -0.99 | 0.00  | 0.00  |

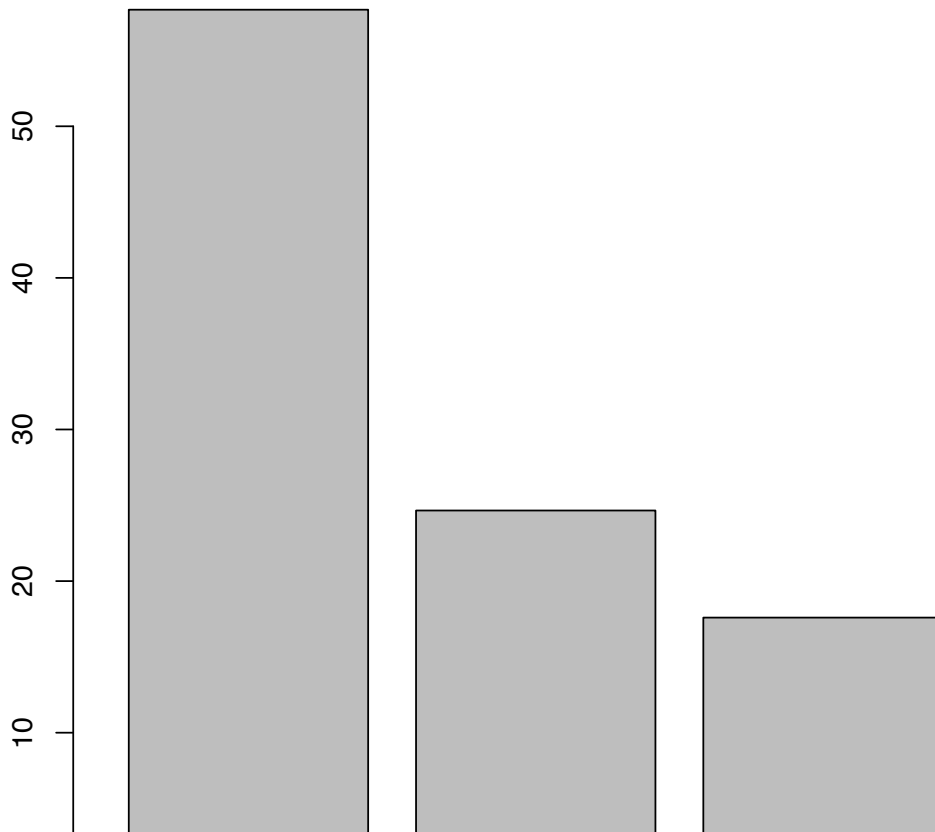
## L'ACP avec FactoMineR

```
library(FactoMineR)  
res.pca<-PCA(X, scale.unit=FALSE, ncp=5, graph=FALSE)
```

## Variances expliquées

```
eigvalues<-data.frame(res.pca$eig)  
barplot(eigvalues$percentage.of.variance, names.arg=row.names(eigvalues),mai
```

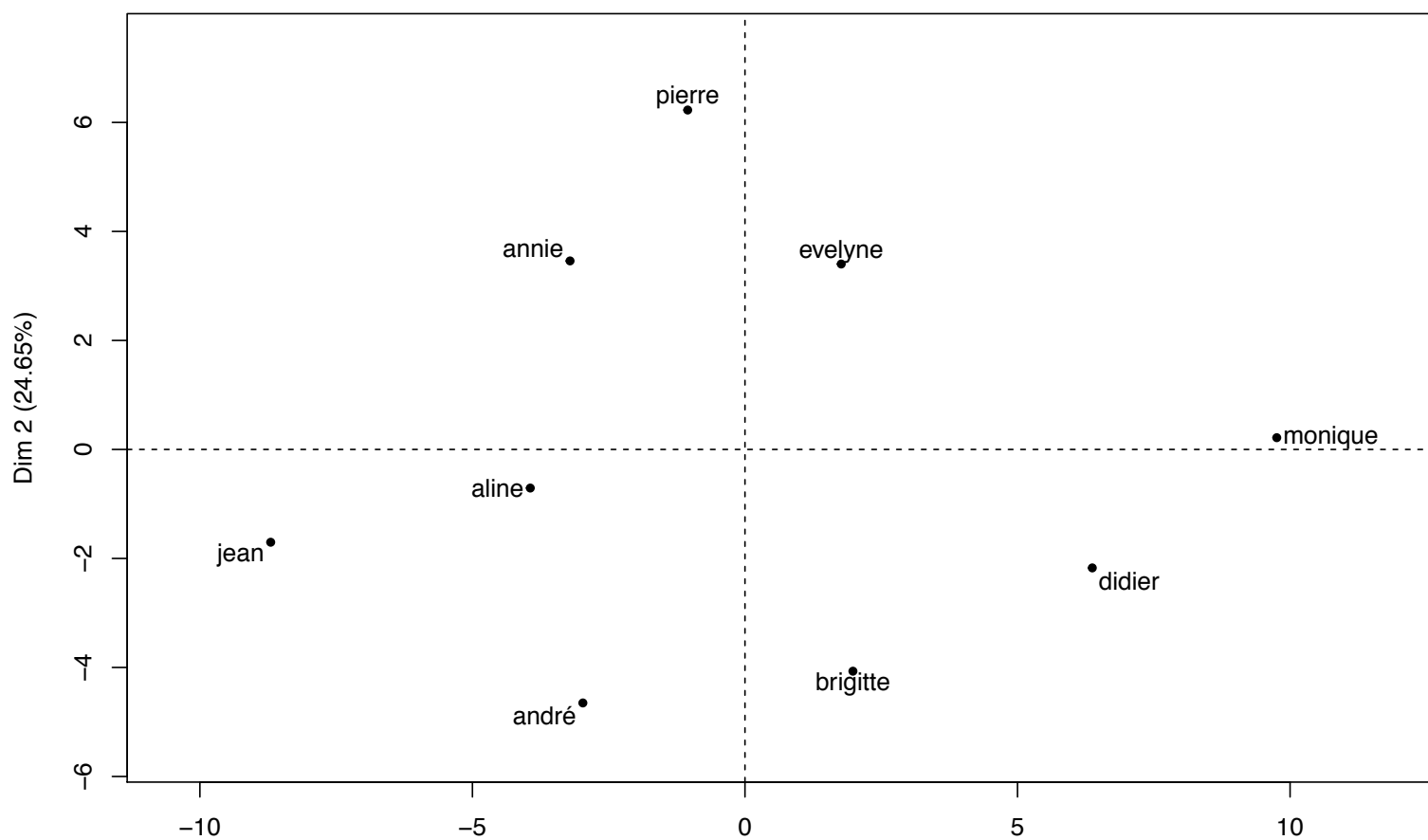
pourcentage de variance par axe



## Représentation des individus

```
plot(res.pca,choix="ind")
```

Individuals factor map (PCA)

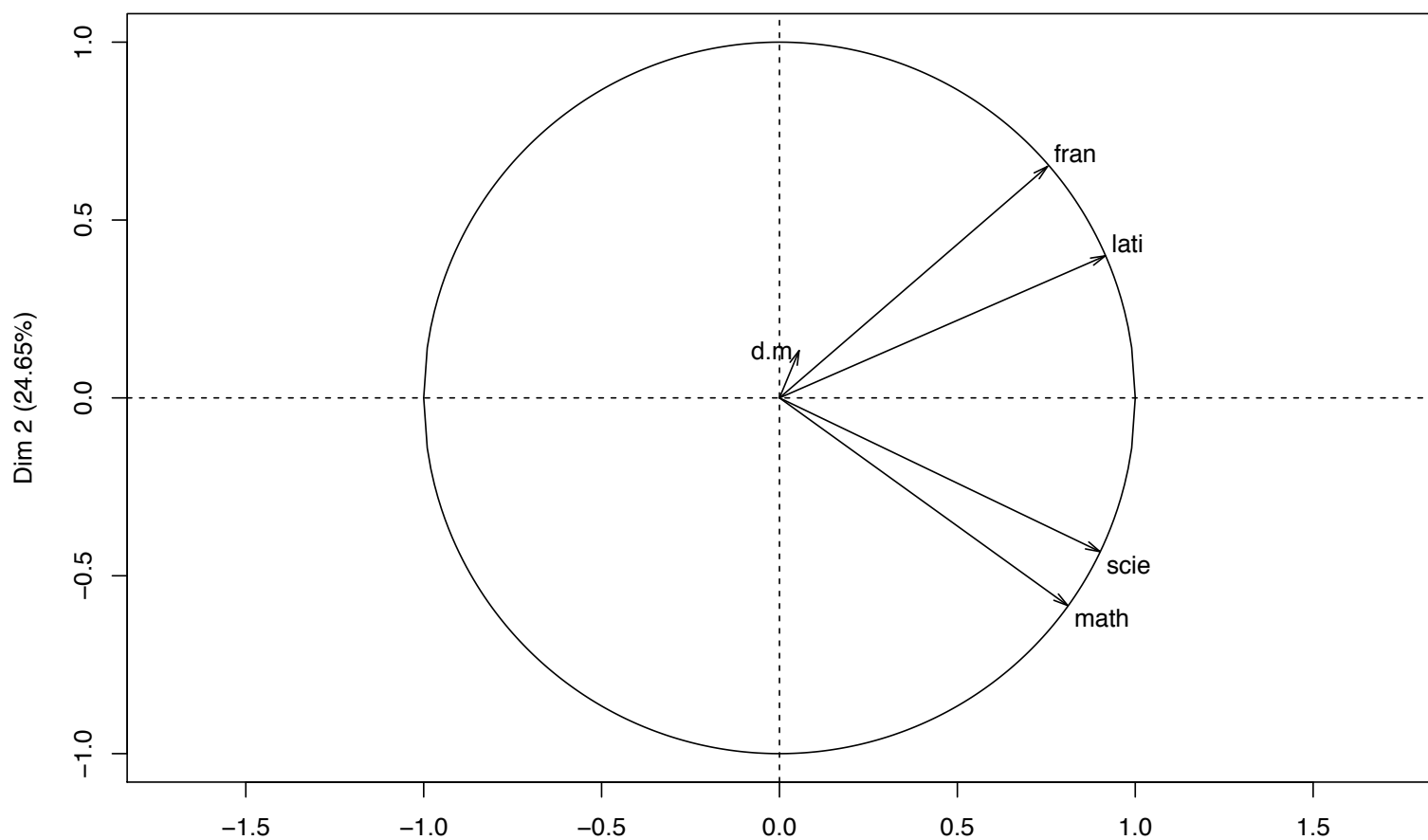




## Représentation des variables

```
plot(res.pca,choix="varcor")
```

Variables factor map (PCA)



## Contribution relative des axes aux individus

```
knitr::kable(res.pca$ind$cos2,format="latex",  
  caption = "Contribution relative des axes aux individus",  
  digits = 2)
```

**Table 9:** Contribution relative des axes aux individus

|          | Dim.1 | Dim.2 | Dim.3 | Dim.4 | Dim.5 |
|----------|-------|-------|-------|-------|-------|
| jean     | 0.89  | 0.03  | 0.08  | 0     | 0     |
| aline    | 0.80  | 0.03  | 0.17  | 0     | 0     |
| annie    | 0.46  | 0.53  | 0.00  | 0     | 0     |
| monique  | 0.89  | 0.00  | 0.11  | 0     | 0     |
| didier   | 0.88  | 0.10  | 0.02  | 0     | 0     |
| andré    | 0.24  | 0.58  | 0.19  | 0     | 0     |
| pierre   | 0.03  | 0.91  | 0.07  | 0     | 0     |
| brigitte | 0.17  | 0.74  | 0.09  | 0     | 0     |
| evelyne  | 0.05  | 0.20  | 0.75  | 0     | 0     |

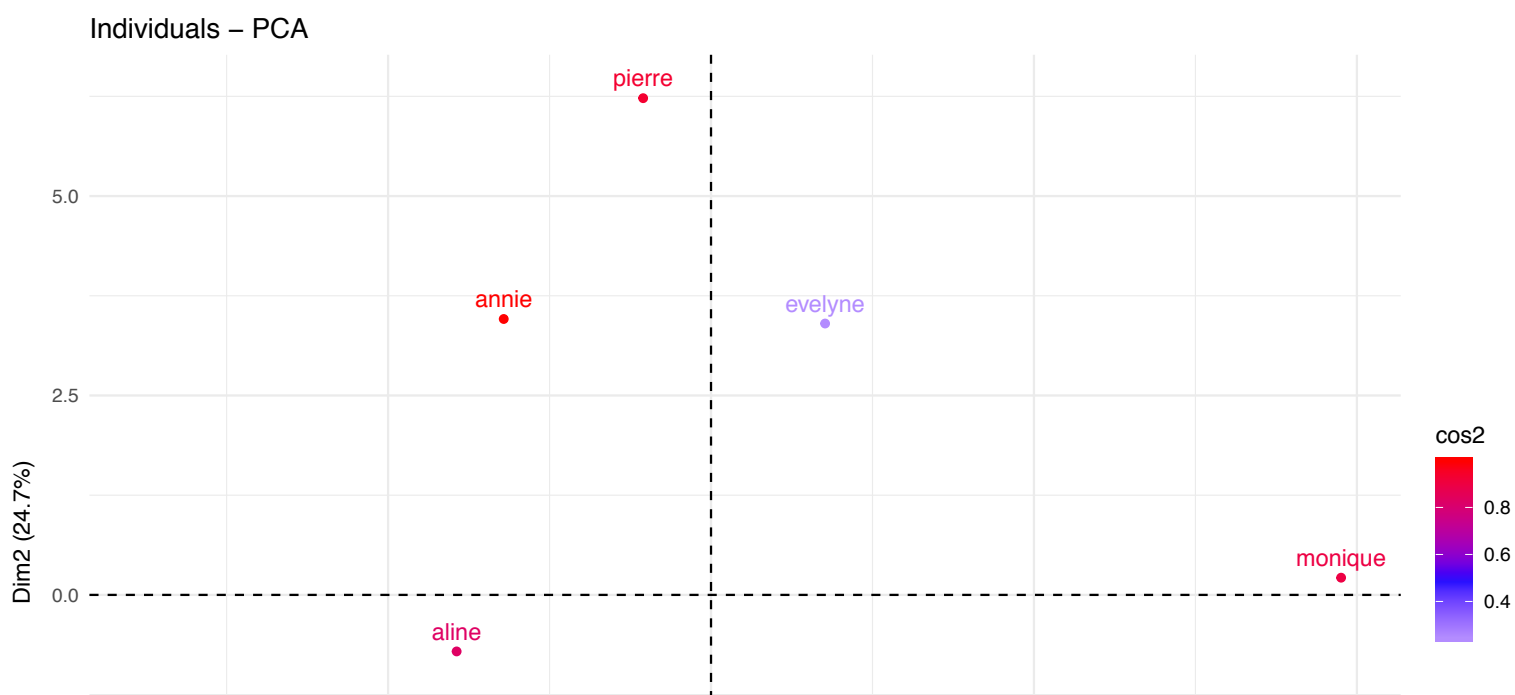
## Contribution relative des axes aux individus

```
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at ht
```

```
fviz_pca_ind(res.pca, col.ind="cos2") + scale_color_gradient2(low="white", mi  
high="red", midpoint=0.50) + theme_minimal()
```



## Contribution des individus aux axes

```
knitr::kable(res.pca$var$contrib,format="latex",  
             caption = "Contribution des individus aux axes",  
             digits = 2)
```

**Table 10:** Contribution des individus aux axes

|      | Dim.1 | Dim.2 | Dim.3 | Dim.4 | Dim.5 |
|------|-------|-------|-------|-------|-------|
| math | 26.47 | 32.14 | 0.26  | 8.34  | 32.78 |
| scie | 25.70 | 13.84 | 0.02  | 30.59 | 29.85 |
| fran | 24.24 | 42.30 | 1.17  | 15.50 | 16.79 |
| lati | 23.49 | 10.45 | 0.05  | 45.45 | 20.56 |
| d.m  | 0.09  | 1.27  | 98.50 | 0.12  | 0.02  |

## Contribution des individus aux axes

```
fviz_contrib(res.pca, choice = "ind", axes = 1)
```

