Chapitre 3: Analyse exploratoire multidimensionnelle (SVD & ACP)

Module : Mathématiques pour l'ingénieur

Filière : $GDNC_1$

Pr. A. Aberqi

Plan

I- De la SVD vers L'ACP

De la SVD vers L'ACP

Les données sont synthétisées dans la matrice de design

$$X = \begin{array}{c} \\ \mathbf{x}_1 \\ \vdots \\ \vdots \\ \mathbf{x}_n \\ \\ \end{array} \left(\begin{array}{ccccc} \mathbf{v}_1 & \cdots & \mathbf{v}_k & \cdots & \mathbf{v}_p \\ x_{11} & \cdots & x_{1k} & \cdots & x_{1p} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} & \cdots & x_{np} \end{array} \right).$$

où $\mathbf{x}_i$ représente l'individu i avec les valeurs $\mathbf{x}_{ik}$ prises par les différents descripteurs. $\mathbf{v}_k$

Métrique utilisée

La métrique qu'on va utiliser pour les individus est la distance Euclidienne. $x_i = (x_{i1}, ..., x_{ip})$ pour i=1,...,n.

$$d(x_i, x_j) = \sum_{k=1}^{k=p} (x_{ik} - x_{jk})^2 .$$

Pourtant, pour les variables, pour étudier la liaison entre eux, on étidierons le coefficient de correlation $r(v_k, v_{k'})$

- **La moyenne d'une variable** v_k est

$$m_k = \frac{1}{n} \sum_{i=1}^{i=n} (v_k)_i$$

- **La variance d'une variable** v_k est

$$\sigma_k^2 = \frac{1}{n} \sum_{i=1}^{i=n} ((v_k)_i - m_k)^2$$

Principe

Il s'agit d'une technique de réduction de dimension qui va nous permettre de projeter les données sur des sous-espaces afin de synthétiser l'information.
Pour faire cela, l'ACP va transformer des variables **corrélées** en un nouvel ensemble de variables **décorrélées** qui vont se présenter comme une combinaison linéaire des anciennes variables. Ces nouvelles variables seront appelées **axes principaux**. Ces axes correspondent a des directions de l'espace selon lesquelles **la variance est maximale**.

Centrer la matrice du design X

Les moyennes de chaque variables v_k soit égale à 0.

$$X_{centrée} = \begin{array}{c} \\ \mathbf{x}_1 \\ \vdots \\ \vdots \\ \mathbf{x}_n \\ \\ \end{array} \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_k & \cdots & \mathbf{v}_p \\ x_{11} - m_1 & \cdots & x_{1k} - m_k & \cdots & x_{1p} - m_p \\ \vdots & & \vdots & & \vdots \\ x_{i1} - m_1 & \cdots & x_{ik} - m_k & \cdots & x_{ip} - m_p \\ \vdots & & \vdots & & \vdots \\ x_{n1} - m_1 & \cdots & x_{nk} - m_k & \cdots & x_{np} - m_p \end{pmatrix}.$$

Réduire la matrice du design X

On réduit notre jeu de données, en dévisant par l'écart-type chaque variable. v_k a pur variance égale à 1.

$$X_{centrée-reduite} = \begin{array}{c} \\ \mathbf{x}_1 \\ \vdots \\ \vdots \\ \mathbf{x}_n \\ \\ \end{array} \begin{pmatrix} \mathbf{v}_1 & \cdots & \mathbf{v}_k & \cdots & \mathbf{v}_p \\ \frac{x_{11}-m_1}{s_1} & \dots & \frac{x_{1k}-m_k}{s_k} & \dots & \frac{x_{1p}-m_p}{s_p} \\ \vdots & & \vdots & & \vdots \\ \frac{x_{i1}-m_1}{s_1} & \dots & \frac{x_{ik}-m_k}{s_k} & \dots & \frac{x_{ip}-m_p}{s_p} \\ \vdots & & \vdots & & \vdots \\ \frac{x_{n1}-m_1}{s_1} & \dots & \frac{x_{nk}-m_k}{s_k} & \dots & \frac{x_{np}-m_p}{s_p} \end{pmatrix}.$$

Remarques

• Toutes les variables vont avoir la même variance (égale à 1) ce qui va éviter de tirer l'ACP vers les variables dont la variance est élevée simplement parce que les valeurs prises par cette dernière sont plus grandes.
• En revanche, si les données associées à une variable présentent un bruit important (mauvaise collecte des données, problème avec l'outil de mesure, ...) alors cette dernière aura une variance semblable à une variable qui serait elle plus informative. Il est donc important de s'assurer que toutes les variables sont informatives avant de faire cela.
• La distance entre deux points x_i et x_j devient

$$d(x_i, x_j) = \sum_{k=1}^{k=p} \frac{1}{s_k^2} \, (x_i)_k - (x_j)_k)^2 \, .$$

Comment ?

On va procéder de façon semblable à la **SVD**. Si on cherche à projeter les données sure un sous-espace de dimension s, on va commencer par chercher des directions $u_1, ..., u_s$ sur lesquelles on va maximiser la variance (ou l'inertie) du nuage de points. Ces directions seront appelées axes principaux avec cette même convention que pour la SVD

• l'axe défini par le vecteur u_1 est le sous-espace de dimension 1 qui maximise l'inertie du nuage du point après projection, • l'axe défini par le vecteur u_2, orthogonal à u_1 est le deuxième sous-espace de dimension 1 qui maximise l'inertie du nuage du point après projection.

• On continue avec u_3 qui est orthogonal à u_1 et u_2, et ainsi de suite.

Matrcie de correlation !! et ses valeurs propres

Mais comment obtenir ces vecteurs formellement ? Lorsque nous avons étudié la décomposition en valeurs singulières, nous nous sommes intéressés aux matricesX^TX et XX^T dont nous avons cherché les valeurs propres et les vecteurs propres. Et bien nous allons faire la même chose ici, mais nous ne travaillerons pas directement sur la matrice de design X mais plutôt sur sa version centrée-réduite normée X_{cenred}.
Soit $C \in \mathcal{M}_p(\mathbb{R})$

$$C = \frac{1}{n} X_{cenred}^T X_{cenred}$$

avec

$$c_{ij} = \frac{1}{n} \sum_{\ell=1}^{\ell=n} \left(\frac{x_{\ell i} - m_i}{s_i} \right) \left(\frac{x_{\ell j} - m_j}{s_j} \right) \in [-1, 1]$$

Matrice de corrélation

Dans la suite, nous noterons z_{ij} les éléments de la matrice $X_{cenred}/\sqrt{n}$ qui correspondent aux données centrées réduites normée de la matrice de design X afin d'éviter toute confusion. En définissant $C = Z^T Z$, on définit une matrice symétrique réelle, elle est donc orthogonalement semblable à une matrice diagonale, i.e. il existe donc une matrice orthogonale $U' \in \mathcal{M}_p(\mathbb{R})$ et une matrice diagonale $\Sigma_p = diag(\lambda_1, ..., \lambda_p)$ telle que $\lambda_1 \geq .... \geq \lambda_p$. On a

$$C = U'\Sigma_p U'^T$$

et pour tout $m \leq p$, $Cu'_m = \lambda_m u'_m$: où U' est formée des vecteurs propres de C. Donc le vecteur propre u'_m est associé à la valeur proper, λ_m qui est la même plus grande valeur propre.

I- Introduction et pratique de L'ACP

ACP est l'une des techniques les plus utilisées lorsqu'on part à la pêche à l'information dans des grand jeux de données. En d'autre terme lorsqu'on veut faire le data minining.
Voiçi un jeu de données avec des éléves disposant en lignes, est écrit par 4 variables en colonnes, leurs notes de Maths X_1, Phy X_2, Fran X_3 et Ang X_4.

Individus	Maths X_1	Phy X_2	Fran X_3	Ang X_4
Ahmed	10	8.5	8	10
Ali	15	4	4	14
Youssef	8	9	4	8
Sara	9	15	12	9
Yassmin	19	12	13	18
Fouad	2	8	15	3
Fati	14	10	14	14
Lili	5	8	6	6
Rim	4	3	7	5
Nour	8	12	3	8
Reda	2	4	2	3
Fares	1	3	4	4

Supposons que nous voulons explorer les données sans attente précise derrière la tête.

- Quel est le moyen le plus simple de faire ?
- Que direz-vous d'un nuage de points X_1 sur l'axe des x et X_2 sur l'axe des y ?.
- $> plot(X_1, X_2)$

Imaginons maintenant le même raisonnement sur le jeu de données plus de deux variables, pour trois variables, on peut utilisé le nuage de points 3D !
Mais que faire si nous avons plus de 3 variables ?
C'est là qu' apparaît l'intrêt d'analyse en composantes principales.
Conceptuellement, toutes les colonnes d'un jeu contiennent des informations potentiellement intéressantes. L' ACP crée un jeu artificiel avec un nombre de dimensions inférieure a celui du premier. La seule différence est que les premiers dimensions de l'ACP concentrent la majeure partie de l'information.

Maths X_1	Phy X_2	Fran X_3	Ang X_4

$\Rightarrow$ Majeure partie des informations

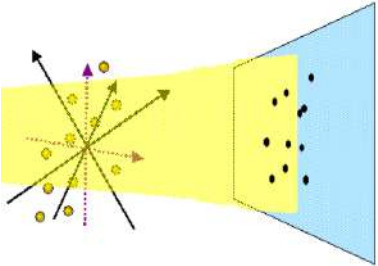
F_1	F_2	F_3	F_4
//////////	/////////		
/////////	//////		
////////	//////		
////////	////////		

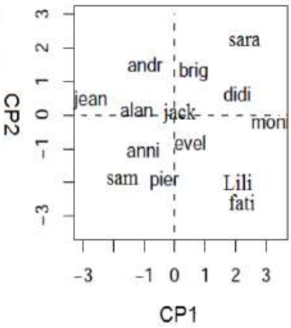
Dans le monde de l'ACP, l'information est appelée **inertie** et les dimensions sont appelées **facteurs** ou axes.

Analyse et Théorie de ACP

L'analyse en composantes principales, permet de projeter le nuage de points des individus sur des sous espaces de dimension petit (2 ou3), tel qu'on retient le plus d'informations possibles(à la droite de la matrice de corrélation et ses valeurs et vecteurs prores), en respectant au mieux :

- Les distances entre les individus (on regroupe ce qui sont proches les uns des autres)
- La structyre de corrélation entre variables.
- Les distances dans l'espace projeté entre les points doivent être les plus proches des distances dans l'espace d'origine.
- Les nouveaux axes sont appellés facteurs ou CP, et doivent être orthogonales et non corrélés.




sara
andr
brig
didi
jean
alan
jack
moni
anni
evel
sam
pier
Lili
fati
CP2
CP1
-3
-1
0
1
2
3

Inertie total du nuage de points N

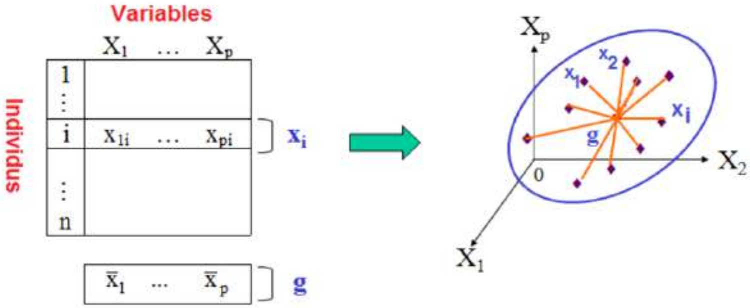
On va voir l'enertie total d'un nuage de points, qui sera un indicateur fort pour mesurer la dispersion des points du nuage autour de son centre de gravité.

Définition

$$I(N,g) = \frac{1}{n}\sum_{i=1}^{i=n} d^2(x_i,g)$$

$$= \frac{1}{n}\sum_{i=1}^{i=n}[\sum_{j=1}^{j=p}]|x_{ij} - \overline{x_j}|^2$$

$$= \sum_{j=1}^{j=p}(\frac{1}{n}\sum_{i=1}^{i=n}(x_{ij} - \overline{x_j})^2 = \sum_{j=1}^{j=p}\sigma_i^2$$

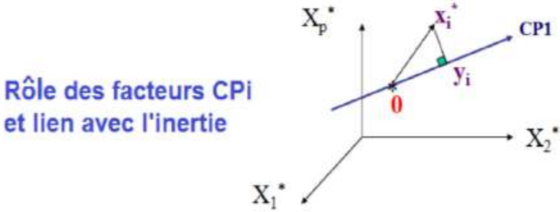

Variables
Individus
X_1 ... X_p
1
i
n
x_{1i} ... x_{pi}
x_i
$\bar{x}_1$ ... $\bar{x}_p$
g
X_p
X_2
X_1
0
x_1
x_2
x_i
g

1. Noté $I(N, g)$ et dépend des variances des variables.
2. Il mesure les dispersion du nuage N par rapport à son centre de gravité $g(\overline{x_1}, \overline{x_2}, ..., \overline{x_p})$
3. C'est la moyenne des distances entre les points et le centre de gravité g.
4. Lorsque cette inertie est faible, le points sont proches du centre de gravité.

Remarque

Pour neutraliser le problème des unités on remplace les données d'origine du tableau par les valeurs centrées-réduite (moyenne =0 et écart-type =1).

Après la neutralisation, l'inertie totale du nouveau nuage de point N^* par rapport à l'origine 0 est $I(N^*, 0) = p$ égale au nombre de variables du tableau.



$$d^2(x_i^*, 0) = d^2(y_i, 0) + d^2(x_i^*, y_i)$$

$$\frac{1}{n}\sum_{i=1}^{i=n} d^2(x_i^*, 0) = \frac{1}{n}\sum_{i=1}^{i=n} d^2(y_i, 0) + \frac{1}{n}\sum_{i=1}^{i=n} d^2(x_i^*, y_i)$$

Inertie totlae $I = p$ = Inertie expliquer par CP_1 + Inertie résiduelle.
Inertie expliquer par CP_1 à maximiser.
Inertie résiduelle à minimiser.

La matrice de corrélation va nous permettre de réaliser le résumé d'information.

1. Dans la matrice de corrélation, on va extraire à l'aide de ces vecteurs propres, les facteurs que l'on recherche en petit (2 ou3).
2. Les facteurs vont permettre de réaliser des projections désirées du nuage dans cet espace de petit dimension, en déformant le moins possible la configuration global des individus.
3. L'interpretation des graphiques dans le nouveau espace de petit dimension qui permettera de comprendre la structure des données analysées.

- La première composante principale CP_1 passe par le centre de gravité 0 du nuage de points N^*.
- Le facteur CP_1 est engendré par le vecteur propre de la matrice de corrélations associée à la plus grande valeur propre λ_1.
- Le CP_2 associée à $\lambda_2 < \lambda_1$ doit être non corrélée et perpendiculaire à CP_1.
- Le pourcentage d'information, (variable totale) expliqué en facteurs CP_i

$$\frac{\lambda_i}{\sum_{j=1}^{j=n} \lambda_j} \times 100\%$$

- Pour avoir une meilleure qualité de l'ACP, on détermine le nombre des facteurs qui conservent un pourcentage cumulé plus de 70% de la variance totale.
- La qualité de représentation sur le plan principal (CP_1, CP_2)

$$\frac{\lambda_1 + \lambda_2}{\sum_{j=1}^{j=n} \lambda_j} \times 100\%$$

Application à l'exemple introductif sous R

Charger le package FactoMineR et le Tableau des données
$>$ Charger les packages nécessaire
$>$ install.packages("FactoMineR") Pour effectuer l'ACP
$>$ install.packages("factoextra") Pour visualiser les résultats
$>$ install.packages("readxl") pour les fichiers excel
$>$ library(FactoMineR)
$>$ library(readxl)

Chargement de jeu de données Excel de mon bureau !

```
> jeu="C:/Users/ahmed/Desktop/baseacp.xlsx"
> donnees=read_excel(jeu)
> head(donnees)
# A tibble: 6 × 5
  Individus `X1:Math` `X2:Phy` `X3:Fran` `X4:Ang`
  <chr>         <dbl>    <dbl>     <dbl>    <dbl>
1 Ahmed            10        5         8       10
2 ali              15        4         4       14
3 youssef           8        9         4        8
4 sara              9       15        12        9
5 yassine          19       12        13       18
6 fouad             2        8        15        3
```

Chargement de jeu de données Excel de mon bureau !

> mat_cor=donnees[,2 :5]

>print(mat_cor)

A tibble: 12 × 4

	`X1:Math`	`X2:Phy`	`X3:Fran`	`X4:Ang`
	<dbl>	<dbl>	<dbl>	<dbl>
1	10	5	8	10
2	15	4	4	14
3	8	9	4	8
4	9	15	12	9
5	19	12	13	18
6	2	8	15	3
7	14	10	14	14
8	5	8	6	6
9	4	3	7	5
.0	8	12	3	8
.1	2	4	2	3
.2	1	3	4	4

Matrice de corrélations

> **matrice_correlation=cor(mat_cor)**
> **print(matrice_correlation)**

```
> matrice_correlation=cor(mat_cor)
> print(matrice_correlation)
          X1:Math    X2:Phy    X3:Fran    X4:Ang
X1:Math 1.0000000 0.4315985 0.3403421 0.9933463
X2:Phy  0.4315985 1.0000000 0.4875729 0.4048940
X3:Fran 0.3403421 0.4875729 1.0000000 0.3499793
X4:Ang  0.9933463 0.4048940 0.3499793 1.0000000
```

Analyse de la matrice de corrélations

Remarquons que toutes **les corrélations linéaires sont positives**, ce qui signifie que toutes les variables varient, en moyenne, dans le meme sens. La corrélation est forte entre $(X1 = MATH)$ et $(X4 = ANGL)$; c-à-d que les élèves qui ont obtenu de bonnes notes en MATH peuvent également avoir de bonnes notes en ANGL. La faible corrélation entre $(X1 = MATH)$ et $(X3 = FRAN)$ montre la grande rupture qui existe dans l'enseignement de ces deux matières.

Réaliser l'ACP

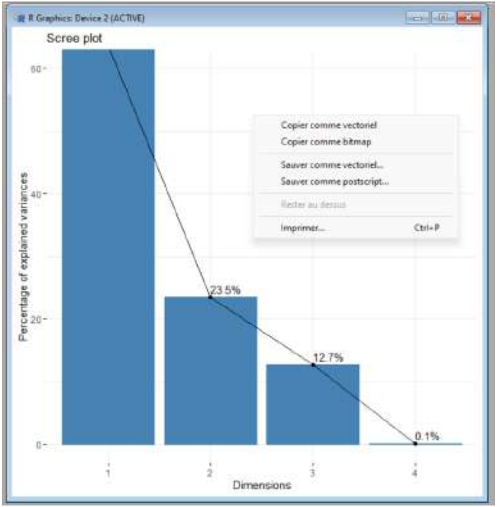
> res.acp <- PCA(donnees[,2 :5], scale.unit = TRUE, graph = FALSE)

- ncp= nombre des facteurs
- donnees $[, 2 : 5]$ pour éviter la première variable qualitative.

Étape 3 : Résultats de l'ACP

Visualiser l'inertie expliquée par les composantes principales :

```
> fviz_eig(res.acp, addlabels = TRUE, ylim = c(0, 60))
```



Valeurs propres et pourcentage des variances

> summary(res.acp)

ou bien > res.acp$eig

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4
Variance	2.546	0.941	0.507	0.006
% of var.	63.660	23.514	12.678	0.148
Cumulative % of var.	63.660	87.174	99.852	100.000

Facteurs à retenir

Les facteurs à retenir sont le premier et le deuxième puisque le pourcentage de la variance totale qui est conservée sur le plan principale engendrée par ces 2 facteurs est :

$$\frac{\lambda_1 + \lambda_2}{\sum_{i=1}^{4} \lambda_i} = \frac{2.546 + 0.941}{2.546 + 0.941 + 0.507 + 0.006} \times 100\% = 87.174\% > 70\%.$$

Remarque

- Le facteur CP_1 est engendré par le vecteur propre de la matrice de corrélations associé à la plus grande valeur propre λ_1.
- Le CP_2 associée à $\lambda_2 < \lambda_1$ doit être non corrélée et perpendiculaire à CP_1.
- Le pourcentage d'information, (variable totale) expliqué en facteurs CP_i
$$\frac{\lambda_i}{\sum_{j=1}^{j=n} \lambda_j} \times 100\%$$
- Pour avoir une meilleure qualité de l'ACP, on détermine le nombre des facteurs qui conservent un pourcentage cumulé plus de 70% de la variance totale.
- La qualité de représentation sur le plan principal (CP_1, CP_2)
$$\frac{\lambda_1 + \lambda_2}{\sum_{j=1}^{j=n} \lambda_j} \times 100\%$$

Coordonnées des variables (matrice des composantes)

Cette matrice donne les corrélations variables-facteurs.
> summary(res.acp) ou bien > res.acpvarcoord

Variables

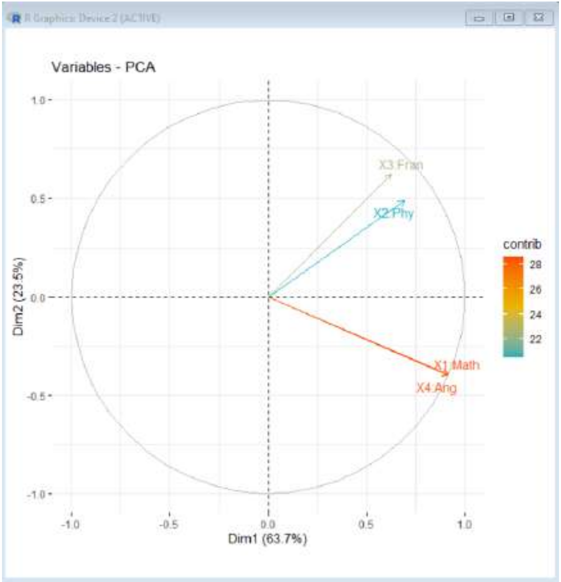
	Dim.1	ctr	cos2	Dim.2	ctr	cos2	D
X1:Math	0.917	33.046	0.841	-0.394	16.484	0.155	0
X2:Phy	0.693	18.834	0.480	0.487	25.214	0.237	-0
X3:Fran	0.627	15.426	0.393	0.622	41.184	0.387	0
X4:Ang	0.912	32.694	0.833	-0.401	17.118	0.161	0

Coordonnées des variables (matrice des composantes)

Remarque

On voit que le premier facteur CP1 est positivement corrélé et assez fortement avec chacune des 4 variables : plus un élève obtient de bonnes notes dans chacune des 4 disciplines, plus il a un score d'élève sur l'axe CP1, réciproquement : plus ces notes sont mauvaises, plus son score est négatif. En ce sens, CP1 représente le niveau général des étudiants. En ce qui concerne l'axe CP2, il oppose, d'une part, X2 et X3 (corrélations positives), d'autre part, X1 et X4 (corrélations positives).

Analyse du nuage de variables (cercle des corrélations)



Analyse du nuage de variables (cercle des corrélations)

Sur le cercle des corrélations, les principes de lecture sont les suivants :

- Plus une variable possède une qualité de représentation élevée dans l'ACP, plus sa flèche est longue,
- Plus deux variables sont corrélées, plus leurs flèches pointent dans la même direction (dans le cercle de corrélation, le coefficient de corrélation est symbolisé par les angles géométriques entre les flèches),
- Plus une variable est proche d'un axe principal de l'ACP, plus elle est liée à lui. Cette dernière règle permet généralement de donner un sens concret aux axes de l'ACP.

Analyse du nuage de variables (cercle des corrélations) pour notre exemple

Remarquons que les 4 variables sont bien représentées, car elles sont proches du cercle. Toutes les variables sont assez éloignées de O, les variables, et donc les angles qu'elles forment, n'ont pas été trop déformées dans la projection. Toutes les variables occupent une zone assez restreinte à l'intérieur du cercle des corrélations. L'angle maximum entre deux variables est inférieur à $90\circ$. Ceci suggère que toutes les variables sont corrélées positivement entre elles (cosinus positive). Les notes des 2 matières (MATH et ANGL) sont plus liées entre elles qu'avec les autres matières. Ceci suggère l'existence de qualités communes (ou de goûts communs) pour réussir dans ces matières. On peut faire des remarques identiques pour PHYS et FRAN. L'écart entre ces deux matières et les précédentes suggère l'existence de qualités différentes (ou de goûts différents) pour réussir ces deux groupes de matières.

Conclusion

En conclusion : Le cercle des corrélations permet de voir, parmi les anciennes variables, les groupes de variables très corrélées entre elles. Donc son étude est plus simple et plus informative que l'analyse directe de la matrice de corrélation.

Qualité de représentation d'une variable dans le plan principal (CP1,CP2) : (voir cos2

Par exemple, la qualité de représentation de X1 est :

$$QLT = \frac{(0.917)^2 + (-0.394)^2}{(0.917)^2 + (-0.394)^2 + (0.022)^2 + (0.055)^2} = 99.65\%$$

Plus la $QLT \geq 60\%$, plus la variable est bien représentée.
$cos2/dim1 + cos2/dim2 + cos2/dim3 + cos2/dim4 = 0.9965$

Contribution d'une variable à la formation d'un facteur :

Par exemple, la contribution de X1, à la formation de Dim1 est :

$$CTR = \frac{(0.917)^2}{(0.917)^2 + (0.693)^2 + (0.627)^2 + (0.912)^2} = 33.03\%$$

(voir la colonne 1, dans les coordonnées des variables)

Interprétation

Remarque

Remarquons la contribution de X1 et X4 à la construction de CP1, ce qui est clair d'après le tableau des corrélations variables-facteurs : (X1 et X4 sont positivement et fortement corrélés à CP1 par 0,917 et 0,912). Par contre, on voit la contribution de X2 et X3 à la construction de CP2.

Coordonnées des individus :

Remarque

Le but ici est de fournir des images planes de dimension 2 approchées du nuage d'individus situés dans l'espace de dimension 4.

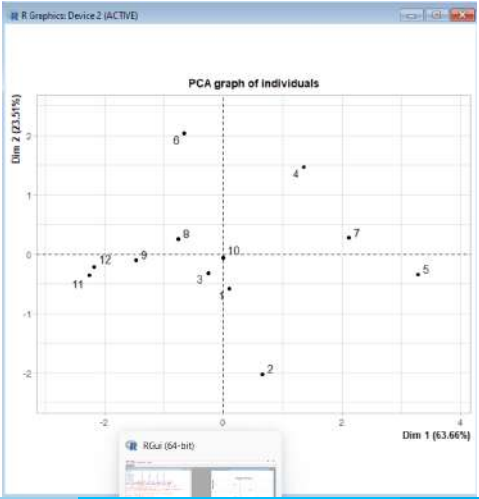
IMPORTANT

L'ensemble des projections de tous les points du nuage d'individus sur son premier axe factoriel U1 appelé premier facteur, sur les individus, constitue une nouvelle variable. On montre que cette variable se confond, 'a la norme près, avec la première composante principale CP1 obtenue dans la projection du nuage de variables. Donc, l'interprétation des axes du graphique ci-dessus est par définition celle des composantes principales.

Visualisation des individus dans le plan factoriel

Visualisation des individus dans le plan factoriel :

```
> fviz_pca_ind(res.acp, geom.ind = "point", addEllipses =
TRUE,legend.title = "Espèce")
```



représentation sur chaque axe (voir cos2), coordonnées des individus

Individuals (the 10 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	c
1	0.866	0.107	0.037	0.015	-0.590	3.086	0.464	0.624	6.3
2	2.163	0.666	1.453	0.095	-2.029	36.493	0.881	0.330	1.7
3	0.891	-0.252	0.209	0.080	-0.313	0.866	0.123	-0.795	10.3
4	2.136	1.362	6.066	0.406	1.460	18.891	0.467	-0.760	9.4
5	3.309	3.285	35.308	0.986	-0.344	1.051	0.011	0.194	0.6
6	2.321	-0.652	1.391	0.079	2.036	36.746	0.770	0.899	13.2
7	2.230	2.121	14.724	0.905	0.270	0.646	0.015	0.630	6.5
8	0.872	-0.755	1.864	0.750	0.248	0.544	0.081	-0.358	2.1
9	1.643	-1.463	7.003	0.793	-0.099	0.087	0.004	0.741	9.0
10	1.529	-0.001	0.000	0.000	-0.063	0.036	0.002	-1.527	38.3

Analyse des représentations des individus dans le plan factoriel

Ainsi, l'axe des abscisses représente le niveau général des étudiants, alors que celui des ordonnées représente leur profil. En effet, un étudiant appartenant au groupe 1 (4e quart du plan f) possède en général des notes meilleures dans les matières X1 et X4 avec des capacités déterminées en X2 et X3 ; c'est le cas par exemple d' 2 (Yassine) et 5(Fati).
Par opposition, un étudiant appartenant au groupe 4, c'est un étudiant qui a en général des notes faibles dans toutes les matières ; c'est le cas de 11, 9 et 12. Donc, le premier axe (axe horizontal) oppose les élèves qui ont globalement de bonnes notes à ceux qui ont généralement de mauvaises notes. Quant au deuxième, il oppose les élèves ayant globalement des très bonnes notes en X2 et X4 à ceux qui ont obtenu de faibles notes dans ces disciplines.

Qualité de représentation d'un individu dans le plan principal (CP1,CP2) :

Par exemple, la qualité de la représentation du premier individu est :

$$QLT = \frac{(0.107)^2 + (-0.590)^2}{(0.107)^2 + (-0.590)^2 + (-0.624)^2 + (0.044)^2} = 47.88\%$$

Plus $QLT \geq 60\%$ plus l'individu est bien repésenté. Chose qui n'est pas vrai pour ind 1, 3 et 10.

Contribution d'un individu à la formation d'un facteur :

Par exemple, la contribution du troisième individu à la formation de CP1 :

$$CTR = \frac{(-0.252)^2}{(0.107)^2 + ... + (-0.001)^2} = 0.21\%$$

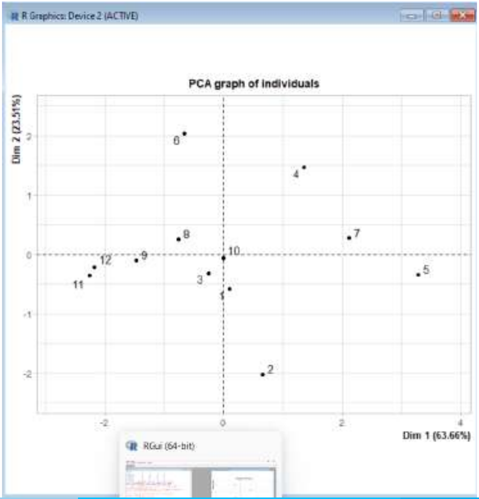
CONCLUSION

Conclusion : nous remarquons que l'ACP a l'avantage d'une part de résumer l'ensemble des variables initiales corrélées en un nombre réduit de facteurs non corrélés. D'autre part, elle nous a permis de mettre en évidence des similarités ou oppositions entre variables et individus.

Visualisation des individus dans le plan factoriel

Visualisation des individus dans le plan factoriel :

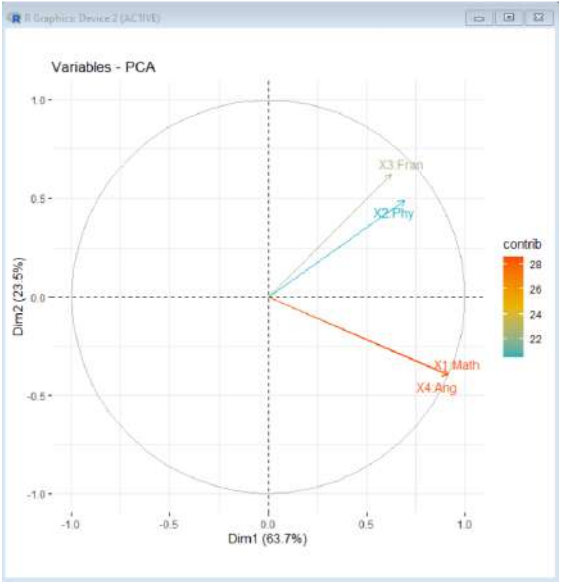
> fviz_pca_ind(res.acp, geom.ind = "point", addEllipses = TRUE,legend.title = "Espèce")



Visualisation des variables code R

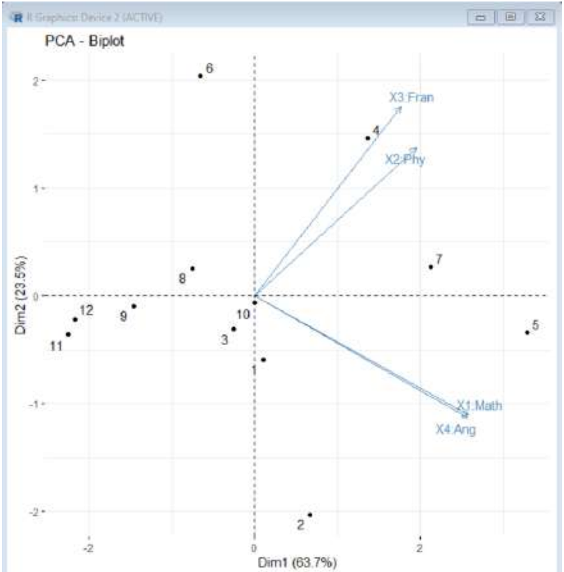
```
> plot.PCA(res.acp, axes=c(1,2), choix="var")
> # 3.3. Visualisation des variables
> fviz_pca_var(res.acp,
+              col.var = "contrib", # Coloration par contribution
+              gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
+              repel = TRUE)
>
```

Visualisation des variables, cercle des corrélations



Visualisation biplot (individus + variables)

> fviz_pca_biplot(res.acp,repel = TRUE,legend.title = "Espèce")



Coordonnées des varibales :

> **dimdesc(res.acp, axes=c(1,2))**

```
> #Coordonn'ees des varibales:
> dimdesc(res.acp, axes=c(1,2))
$Dim.1

Link between the variable and the continuous variables (R-square)
=================================================================
         correlation      p.value
X1:Math    0.9173248 2.644387e-05
X4:Ang     0.9124212 3.497636e-05
X2:Phy     0.6925316 1.255257e-02
X3:Fran    0.6267406 2.918993e-02

$Dim.2

Link between the variable and the continuous variables (R-square)
=================================================================
         correlation    p.value
X3:Fran     0.622375 0.03067679

> |
```

Vecteurs et valeurs propres, Percentages de variance

> **summary(res.acp)**

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4
Variance	2.546	0.941	0.507	0.006
% of var.	63.660	23.514	12.678	0.148
Cumulative % of var.	63.660	87.174	99.852	100.000

Qualité de représentation sur chaque axe (voir cos2), coordonnées et contributions des variables

Variables

	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
X1:Math	0.917	33.046	0.841	-0.394	16.484	0.155	0.022	0.097	0.000
X2:Phy	0.693	18.834	0.480	0.487	25.214	0.237	-0.532	55.853	0.283
X3:Fran	0.627	15.426	0.393	0.622	41.184	0.387	0.469	43.350	0.220
X4:Ang	0.912	32.694	0.833	-0.401	17.118	0.161	0.060	0.700	0.004

Qualité de représentation sur chaque axe (voir cos2), coordonnées et contributions des individus

Individuals (the 10 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	c
1	0.866	0.107	0.037	0.015	-0.590	3.086	0.464	0.624	6.3
2	2.163	0.666	1.453	0.095	-2.029	36.493	0.881	0.330	1.7
3	0.891	-0.252	0.209	0.080	-0.313	0.866	0.123	-0.795	10.3
4	2.136	1.362	6.066	0.406	1.460	18.891	0.467	-0.760	9.4
5	3.309	3.285	35.308	0.986	-0.344	1.051	0.011	0.194	0.6
6	2.321	-0.652	1.391	0.079	2.036	36.746	0.770	0.899	13.2
7	2.230	2.121	14.724	0.905	0.270	0.646	0.015	0.630	6.5
8	0.872	-0.755	1.864	0.750	0.248	0.544	0.081	-0.358	2.1
9	1.643	-1.463	7.003	0.793	-0.099	0.087	0.004	0.741	9.0
10	1.529	-0.001	0.000	0.000	-0.063	0.036	0.002	-1.527	38.3

Un peu plus sur R

Resulta= PCA(données[,2 :5],...)

name description

1. "$eig" "eigenvalues"
2 . "$var" "results for the variables"
3. "varcoord" "coord. for the variables"
4. "varcor" "correlations variables - dimensions"
5. "varcos2" "cos2 for the variables"
6. "varcontrib" "contributions of the variables"
7. "$ind" "results for the individuals"
8. "indcoord" "coord. for the individuals"
9. "indcos2" "cos2 for the individuals"
10. "indcontrib" "contributions of the individuals"
11. "$call" "summary statistics"
12. "$call$centre" "mean of the variables"
13. "$call$ecart.type" "standard error of the variables"
14. "$call$row.w" "weights for the individuals"
15. "$call$col.w" "weights for the variables"

Fin