



ANAD

Rapport du TP

Application de l'ACP et interprétation des résultats

Réalisé par :

- BOUAZIZ Sofiane
- OUSSAIDENE Smail

Groupe :

SID 01

Table des matières

1	Introduction.....	3
2	Lecture des données.....	3
3	Prétraitement des données	3
4	Visualisation et statistiques concernant le tableau de données	4
5	Application de l'ACP	5
6	Choix des axes	5
7	Signification des axes	7
8	Qualité de représentation globale	7
9	Qualité de représentation des variables	8
10	Contribution relative des individus	9
11	Qualité de représentation des individus.....	10
12	Interprétation des individus sur le plan engendré par les axes 1 et 2	11
13	Illustration de la qualité de représentation des individus	12
14	Détails supplémentaire sur l'interprétation des individus sur le premier plan.....	13
15	Conclusion	13

1 Introduction

L'analyse en composantes principales (ACP) est un outil extrêmement puissant de synthèse de l'information, très utile lorsque l'on est en présence d'une somme importante de données quantitatives à traiter et interpréter. C'est une méthode descriptive permettant de traiter des tableaux de données quantitatives de grandes dimensions. Son but est de résumer la grande quantité d'information contenue dans le tableau de données de départ, et cela dans un tableau de plus petite dimension.

Dans le cadre de ce TP, nous allons utiliser cette méthode pour interpréter les données issus des notes du S1 ICS.

2 Lecture des données

Nous avons utilisé la library « *readxl* » pour récupérer les données représentant les notes du S1 du fichier Excel *NotesPVS1-S2.xlsx*.

```
library("readxl")
data = read_excel("NotesPVS1-S2.xlsx", sheet=2) # 2 Represente le numéro
de la feuille du fichier excel
```

Voici ci-dessous, une prise d'écran représentant une partie des données récupérées :

Matricule	...2	...3	...4	...5	...6	...7	...8	Nom & Prénom	Groupe	'SYS1 * x 5'	'RES1 * x 4'	'ANUM * x 4'
<chr>	<lg1>	<lg1>	<lg1>	<lg1>	<lg1>	<lg1>	<lg1>	<chr>	<chr>	<chr>	<chr>	<chr>
1 18/0194	NA	NA	NA	NA	NA	NA	NA	SARAOUI ILHAM	G01	SYS1: 6.82 <~	9.75	ANUM: 3.48 <~
2 17/0015	NA	NA	NA	NA	NA	NA	NA	SI AHMED MASSI~	G01	SYS1: 5.86 <~	9.10	9.33
3 18/0317	NA	NA	NA	NA	NA	NA	NA	LARDJANI ABDEL~	G08	SYS1: 5.00 <~	7.90	9.16
4 18/0308	NA	NA	NA	NA	NA	NA	NA	HAMITOUCHE MON~	G03	SYS1: 5.71 <~	RES1: 4.60 <~	11.16
5 18/0255	NA	NA	NA	NA	NA	NA	NA	ZERROUK AHMED	G08	SYS1: 6.66 <~	RES1: 5.25 <~	11.15
6 18/0296	NA	NA	NA	NA	NA	NA	NA	BENDAHHANE HAM~	G03	SYS1: 3.10 <~	RES1: 6.10 <~	ANUM: 4.31 <~
7 17/0283	NA	NA	NA	NA	NA	NA	NA	LAIREDJ SIHAM	G09	SYS1: 4.58 <~	RES1: 4.20 <~	ANUM: 3.63 <~
8 18/0321	NA	NA	NA	NA	NA	NA	NA	MERBOUHI WAIL	G08	SYS1: 0.00 <~	RES1: 0.00 <~	ANUM: 0.00 <~
9 Fait à oued Sm~	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
10 Signatures	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Nous avons tiré quelques remarques :

1. Des colonnes non nécessaires pour notre étude ont été récupérées
2. Le nom des colonnes représentant les modules contient des caractères en plus
3. Les données récupérées contiennent des incohérences, en effet les moyennes en dessous de la note minimale sont représentées de cette façon : *nom du module : note obtenue < note minimale*
4. Quelques lignes dans notre ensemble de données contiennent des données manquantes
5. Le type des données n'est pas numérique

3 Prétraitement des données

3.1 Suppression des colonnes non nécessaires

Nous avons supprimé deux types de colonnes :

- Des colonnes rajoutées lors de la lecture des données et sont nommées par défaut sous cette forme « ... x »
- Des colonnes sortant du domaine de notre étude

Pour cela nous avons utilisé la library « *dplyr* »

```
data = select(data, -starts_with("...")) # supprimer les colonnes qui
commencent par ...
data = select(data, -c('Matricule', `Nom & Prénom`, 'Groupe', `Ne S1`,
`Rang S1`))
```

3.2 Renommage des colonnes

Nous avons utilisé la library « *plyr* » pour renommer les colonnes souhaitées de cette façon :

```
data = plyr::rename(data, c(`SYS1?x 5` = 'SYS1', `RES1?x 4` = 'RES1', `ANUM?x 4` = 'ANUM', `RO?x 3` = 'RO', `ORG?x 3` = 'ORG', `IGL?x 5` = 'IGL', `THP?x 4` = 'THP', `LANG1?x 2` = 'LANG', `MoyS1` = 'MoyS1'))
```

3.3 Correction des incohérences

Vous pouvez constater ci-dessous les incohérences citées plus haut :

11.76	13.00	10.16	7.40	10.90	11.00	THP: 4.90 < 5.00	13.33	10.25
7.94	14.05	11.68	8.00	12.80	10.30	5.50	14.20	10.23
SYS1: 6.54 < 6.92	8.95	15.82	9.68	10.15	13.80	THP: 4.50 < 5.00	13.50	10.18
12.88	12.63	ANUM: 6.82 < 8.31	9.05	10.60	8.15	8.80	13.72	10.15

Nous avons remédié à ce problème en utilisant la fonction *gsub* :

```
data$SYS1 = gsub("[A-Z0-9]+: ([0-9\\.]+) < [0-9\\.]+", "\\1", data$SYS1)
data$RES1 = gsub("[A-Z0-9]+: ([0-9\\.]+) < [0-9\\.]+", "\\1", data$RES1)
data$ANUM = gsub("[A-Z0-9]+: ([0-9\\.]+) < [0-9\\.]+", "\\1", data$ANUM)
data$RO = gsub("[A-Z0-9]+: ([0-9\\.]+) < [0-9\\.]+", "\\1", data$RO)
data$ORG = gsub("[A-Z0-9]+: ([0-9\\.]+) < [0-9\\.]+", "\\1", data$ORG)
data$IGL = gsub("[A-Z0-9]+: ([0-9\\.]+) < [0-9\\.]+", "\\1", data$IGL)
data$THP = gsub("[A-Z0-9]+: ([0-9\\.]+) < [0-9\\.]+", "\\1", data$THP)
data$LANG = gsub("[A-Z0-9]+: ([0-9\\.]+) < [0-9\\.]+", "\\1", data$LANG)
```

3.4 Suppression des données manquantes

Les dernières lignes de notre tableau de données contiennent des données manquantes

```
data = na.omit(data) #supprimer les données manquantes
```

3.5 Conversion des types de colonnes

Les moyennes des modules sont de type char

```
data = as.data.frame(apply(data, 2, as.numeric)) #conversion des type des colonnes en numeric
```

4 Visualisation et statistiques concernant le tableau de données :

Voici ci-dessous, une prise d'écran représentant une partie des données évoquée précédemment dans la section « 2 Lecture des données » après le prétraitement :

	SYS1	RES1	ANUM	RO	ORG	IGL	THP	LANG	MoyS1
235	6.82	9.75	3.48	8.58	10.65	8.85	7.35	9.98	7.94
236	5.86	9.10	9.33	8.25	11.15	8.35	2.35	10.48	7.78
237	5.00	7.90	9.16	10.75	8.30	7.80	2.80	12.16	7.50
238	5.71	4.60	11.16	8.00	10.00	7.60	4.10	11.67	7.44
239	6.66	5.25	11.15	9.50	7.50	9.20	3.70	6.03	7.43
240	3.10	6.10	4.31	6.18	5.90	6.80	2.30	13.50	5.45
241	4.58	4.20	3.63	4.62	4.20	8.30	3.00	4.29	4.76
242	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Avant de commencer notre étude, c'est préférable d'avoir des statistiques concernant notre tableau de données. Ces statistiques sont récupérées à l'aide de la commande suivante :

```
summary(data) #Avoir des statistique concernant le tableau de données
```

Voici ci-dessous, une prise d'écran représentant les statistiques de nos données :

SYS1		RES1		ANUM		RO	
Min.	: 0.00	Min.	: 0.00	Min.	: 0.00	Min.	: 0.000
1st Qu.:	9.51	1st Qu.:	10.71	1st Qu.:	11.84	1st Qu.:	9.262
Median	:11.38	Median	:13.10	Median	:14.16	Median	:11.120
Mean	:11.46	Mean	:12.98	Mean	:13.75	Mean	:11.034
3rd Qu.:	13.80	3rd Qu.:	15.19	3rd Qu.:	15.82	3rd Qu.:	12.665
Max.	:19.30	Max.	:18.85	Max.	:19.34	Max.	:17.900
ORG		IGL		THP		LANG	
Min.	: 0.00	Min.	: 0.00	Min.	: 0.000	Min.	: 0.00
1st Qu.:	10.30	1st Qu.:	10.00	1st Qu.:	5.825	1st Qu.:	11.49
Median	:11.74	Median	:11.20	Median	: 7.675	Median	:12.83
Mean	:11.48	Mean	:10.98	Mean	: 7.799	Mean	:12.68
3rd Qu.:	12.69	3rd Qu.:	12.20	3rd Qu.:	9.588	3rd Qu.:	14.16
Max.	:15.20	Max.	:14.65	Max.	:15.900	Max.	:17.84
MoyS1							
Min.	: 0.00						
1st Qu.:	10.27						
Median	:11.56						
Mean	:11.44						
3rd Qu.:	12.72						
Max.	:16.23						

5 Application de l'ACP

Nous avons utilisé la library « *FactoMineR* » pour appliquer l'ACP sur notre tableau de données :

```
library(FactoMineR)
acp = PCA(data[,-9]) #Toutes les colonnes à part la dernière qui represente MoyS1
```

6 Choix des axes

L'ACP nous permet de réduire la dimensionnalité de notre tableau de données, dans cette partie nous allons déterminer le nombre d'axes à retenir.

Pour cela, nous avons utilisé la commande suivante pour avoir plus d'informations concernant notre ACP :

```
summary(acp)
```

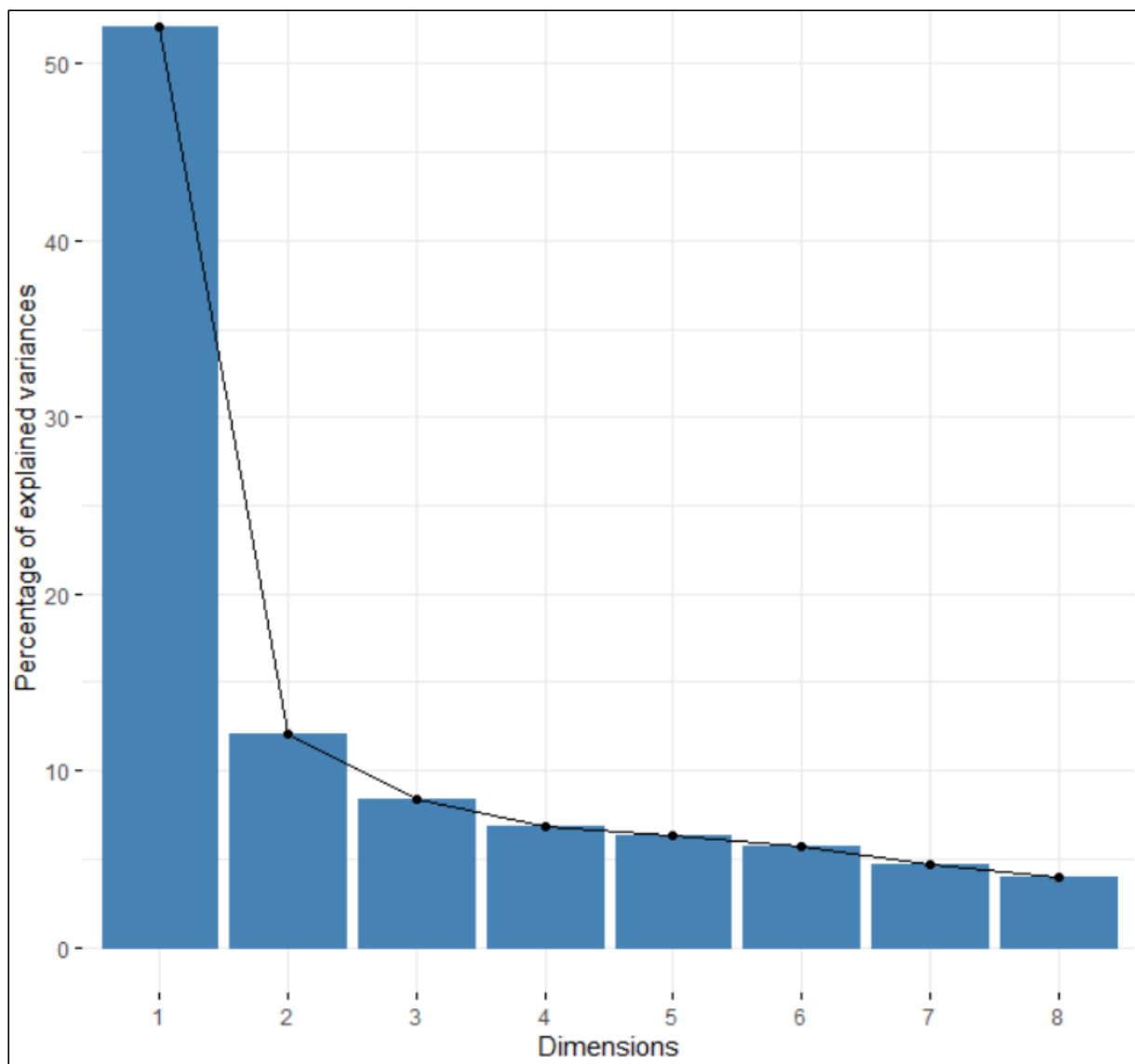
Les résultats obtenus concernant les valeurs propres sont donnés ci- dessous :

Eigenvalues								
	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8
Variance	4.164	0.962	0.671	0.548	0.503	0.457	0.373	0.320
% of var.	52.053	12.024	8.391	6.854	6.293	5.717	4.668	4.000
Cumulative % of var.	52.053	64.078	72.468	79.322	85.615	91.332	96.000	100.000

Nous avons utilisé la library « *factoextra* » pour représenter l'éboulis des valeurs propres :

```
library("factoextra")  
fviz_eig(acp) #Représentation de l'eboulis des valeurs propres
```

Nous avons obtenu le résultat suivant :



D'après l'éboulis des valeurs propres, le coude se trouve entre la 3ème et la 4ème valeur. Donc on retient les 3 premiers axes.

7 Signification des axes :

Les résultats obtenus concernant les variables de départ sont donnés ci- dessous :

Variables									
	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
SYS1	0.787	14.891	0.620	-0.207	4.461	0.043	0.057	0.480	0.003
RES1	0.770	14.220	0.592	-0.306	9.746	0.094	0.043	0.272	0.002
ANUM	0.773	14.355	0.598	-0.129	1.725	0.017	-0.084	1.054	0.007
RO	0.729	12.754	0.531	-0.100	1.041	0.010	-0.279	11.596	0.078
ORG	0.700	11.753	0.489	0.458	21.791	0.210	-0.178	4.698	0.032
IGL	0.636	9.706	0.404	0.114	1.362	0.013	0.726	78.427	0.526
THP	0.778	14.521	0.605	-0.286	8.505	0.082	-0.132	2.610	0.018
LANG	0.570	7.799	0.325	0.703	51.370	0.494	-0.076	0.863	0.006

Une variable X^j contribue à la construction d'un axe α , si sa contribution absolue par rapport à cet axe est supérieure à $\frac{1}{p}$ avec $p = 8$, ie : $\text{cab}^\alpha(j) > \frac{1}{p}$, $p = 8$. Mais comme les contributions absolues sous R sont données en pourcentage alors $\text{cab}^\alpha(j) > (\frac{100}{p}) = 12,5$, $p = 8$.

Le signe des variables sur l'axe α est déterminé par la projection de celle-ci sur l'axe α

7.1 1^{er} axe :

+	-
SYS1, RES1, ANUM, RO , THP	

L'axe 1 est un axe à effet taille, il mesure les notes obtenues en SYS1, RES1, ANUM, RO , THP.

7.2 2^{ème} axe :

+	-
ORG, LANG	

L'axe 2 est un axe à effet taille, il mesure les notes obtenues en ORG, LANG.

7.3 3^{ème} axe :

+	-
IGL	

L'axe 3 est un axe à effet taille, il mesure les notes obtenues en IGL.

8 Qualité de représentation globale :

8.1 Du plan engendré par les axes 1 et 2 :

$$I_{12} = \frac{\lambda_1 + \lambda_2}{\sum \lambda_i}$$

$I_{12} = 64,078$ %, donc 64,078 % de l'information est contenue dans le plan engendré par les axes 1 et 2.

8.2 Du plan engendré par les axes 1,2 et 3 :

$$I_{123} = \frac{\lambda_1 + \lambda_2 + \lambda_3}{\sum \lambda_i}$$

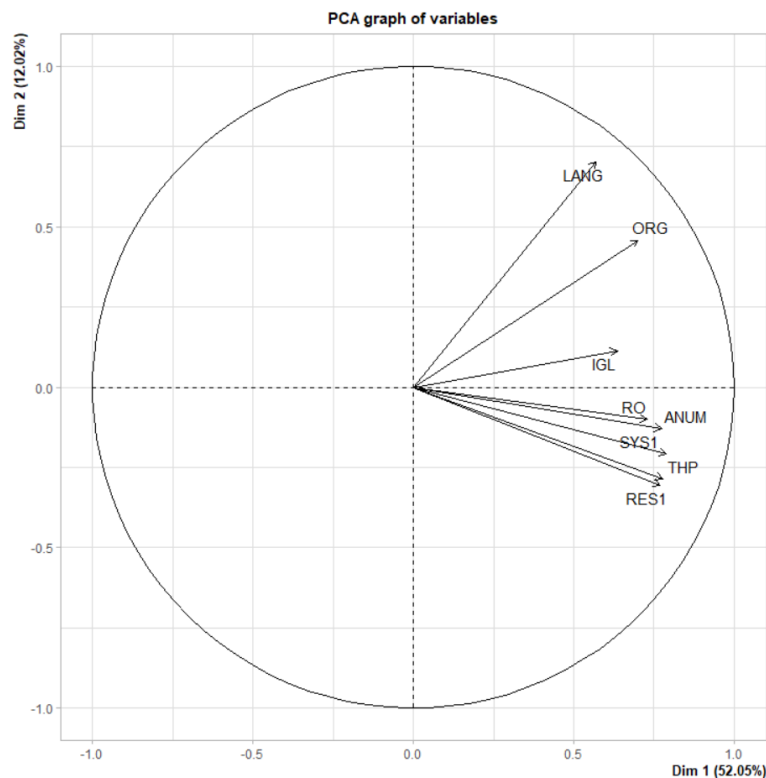
$I_{123} = 72,468$ %, donc 72,468 % de l'information est contenue dans le plan engendré par les axes 1, 2 et 3

9 Qualité de représentation des variables :

9.1 Sur le plan engendré par les axes 1 et 2 :

Variable	cos12	Qualité de représentation
SYS1	0,663	Moyennement bien représentée (0,663 proche de 0,67)
RES1	0,686	Bien représentée (0,686 > 0,67)
ANUM	0,615	Moyennement bien représentée
RO	0,541	Mal représentée (0,541 < 0,67)
ORG	0,699	Bien représentée
IGL	0,417	Mal représentée
THP	0,687	Bien représentée
LANG	0,819	Très bien représentée (0,819 proche de 1)

Comme on a effectué une ACP normée, car les modules sont de coefficient différent, alors on peut déterminer la qualité de représentation des variables sur le plan engendré par les axes 1 et 2 en utilisant le cercle de corrélation :



On remarque que :

- La projection des variables SYS1, RES1, ANUM, ORG, THP et LANG se retrouvent proches du cercle, alors ces variables sont bien représentées
- La projection des variables RO, IGL se retrouvent à l'intérieur du cercle, alors ces variables sont mal représentées

9.2 Sur le plan engendré par les axes 1,2 et 3 :

Variable	cos123	Qualité de représentation
SYS1	0,666	Moyennement bien représentée
RES1	0,688	Bien représentée
ANUM	0,622	Moyennement bien représentée
RO	0,619	Moyennement bien représentée
ORG	0,731	Bien représentée
IGL	0,943	Très bien représentée
THP	0,705	Bien représentée
LANG	0,825	Très bien représentée

On remarque que toutes les variables sont bien représentées sur le plan engendré par les axes 1,2 et 3.

10 Contribution relative des individus :

L'individu **BOUAZIZ Sofiane** est représenté par le rang **4**, et l'individu **OUSSAIDENE Smail** est représenté par le rang **6**.

Nous allons se contenter pour l'étude des individus sur ces deux derniers.

Les résultats obtenus concernant les individus sont donnés ci- dessous :

Individuals (the 10 first)											
	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2	
1	5.288	4.967	2.449	0.883	-0.511	0.112	0.009	-1.154	0.820	0.048	
2	4.775	4.610	2.109	0.932	-0.337	0.049	0.005	0.640	0.252	0.018	
3	5.079	4.734	2.224	0.869	-0.141	0.009	0.001	-0.935	0.538	0.034	
4	4.846	4.369	1.894	0.813	-1.305	0.732	0.073	-0.586	0.211	0.015	
5	4.208	3.814	1.444	0.822	-1.029	0.455	0.060	-0.348	0.075	0.007	
6	4.500	3.927	1.530	0.762	-0.693	0.206	0.024	-1.171	0.844	0.068	
7	4.174	3.489	1.208	0.699	-1.667	1.193	0.159	0.082	0.004	0.000	
8	3.995	3.710	1.366	0.862	0.222	0.021	0.003	-0.644	0.256	0.026	
9	3.630	3.349	1.113	0.851	-0.323	0.045	0.008	0.882	0.478	0.059	
10	3.666	3.435	1.171	0.878	-0.183	0.014	0.002	-1.097	0.741	0.090	

Un individu X_i contribue à la construction d'un axe α , si sa contribution absolue par rapport à cet axe est supérieure à $\frac{1}{n}$ avec $n = 242$, ie : $\text{cab}^\alpha(i) > (\frac{1}{n})$, $n = 242$. Mais comme les contribution absolu sous R sont donné en pourcentage alors $\text{cab}^\alpha(i) > (\frac{100}{n}) = 0,413$, $n = 242$.

Le signe des individus sur l'axe α est déterminé par la projection de celui-ci sur l'axe α

10.1 1 er axe :

+	-
SYS1, RES1, ANUM, RO, THP I ₄ , I ₆	

Les individus I₄, I₆ ont obtenu de bonnes notes en SYS1, RES1, ANUM, RO et THP.

10.2 2 ème axe :

+	-
ORG, LANG	I ₄

L'individu I_4 n'a pas obtenu de bonnes notes en ORG et LANG comparable à la moyenne de la promotion.

Remarque :

On remarque que l'individu I_4 a de mauvaises notes en ORG et LANG, malgré que ses notes dans ces modules soient au-dessus de la moyenne de la promotion, et ceci revient à la mauvaise qualité de représentation de l'individu I_4 sur l'axe 2 ($\cos^2(4)$ est très loin de 1).

10.3 3^{ème} axe :

+	-
IGL	I_6

L'individu I_6 n'a pas obtenu de bonnes notes en IGL comparable à la moyenne de la promotion.

Remarque :

On remarque que l'individu I_6 a une mauvaise note en IGL, malgré que sa note dans ce module est au-dessus de la moyenne de la promotion, et ceci revient à la mauvaise qualité de représentation de l'individu I_6 sur l'axe 3 ($\cos^3(6)$ est très loin de 1).

11 Qualité de représentation des individus :

11.1 Sur le plan engendré par l'axe 1 :

Individus	\cos^2	Qualité de représentation
I_4	0,813	Bien représenté
I_6	0,762	Bien représenté

11.2 Sur le plan engendré par l'axe 2 :

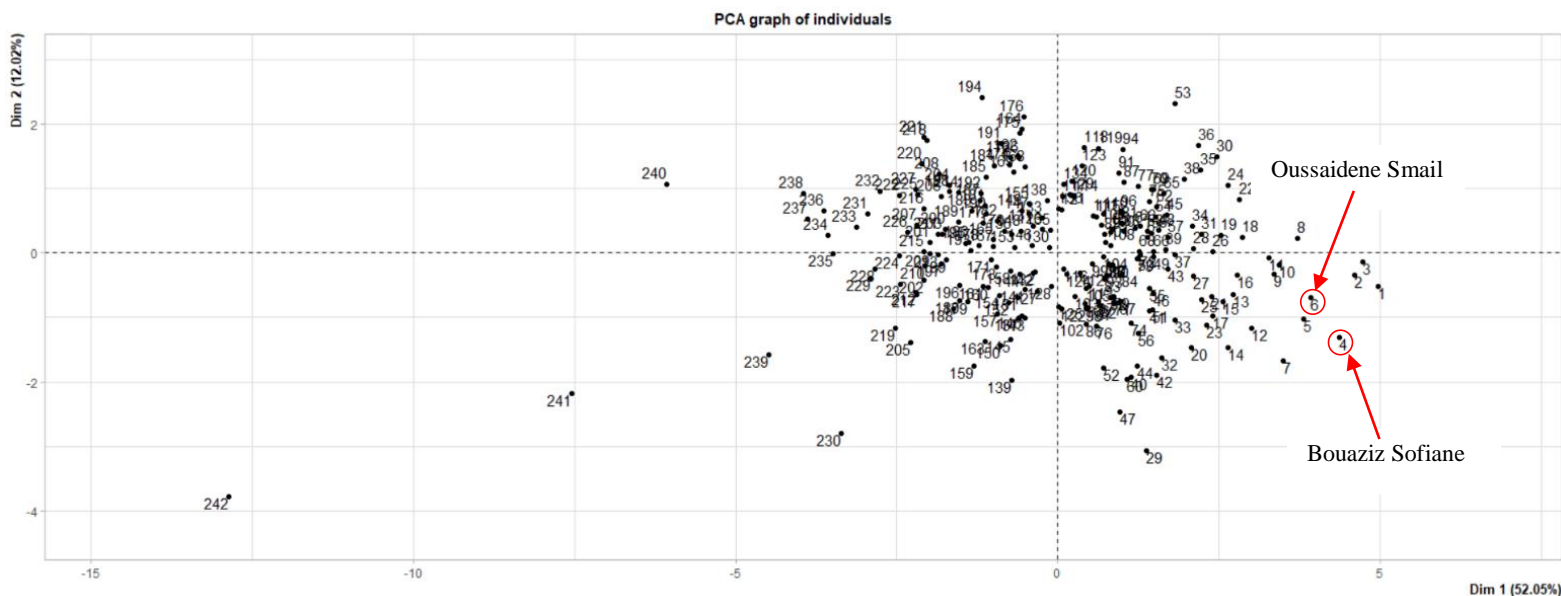
Individus	\cos^2	Qualité de représentation
I_4	0,073	Très mal représenté
I_6	0,024	Très mal représenté

11.3 Sur le plan engendré par les axes 1 et 2 :

Individus	\cos^2	Qualité de représentation
I_4	0,886	Bien représenté
I_6	0,789	Bien représenté

12 Interprétation des individus sur le plan engendré par les axes 1 et 2 :

La projection des individus sur le plan engendré par les axes 1 et 2 est donnée ci- dessous :



Pour l'individu I₄ :

- En projetant l'individu sur le 1^{er} axe, on remarque qu'il est loin de l'origine (la moyenne) et se situe dans le côté positif de l'axe. Et en se basant sur la signification des axes donnée précédemment, on peut conclure que l'individu I₄ a obtenu de bonnes notes en SYS1, RES1, ANUM, RO et THP.
- En projetant l'individu sur le 2^{ème} axe, on remarque qu'il est loin de l'origine (la moyenne) et se situe dans le côté négatif de l'axe. Et en se basant sur la signification des axes donnée précédemment, on peut conclure que l'individu I₄ n'a pas obtenu de bonnes notes en ORG et LANG comparable à la moyenne de la promotion.

Pour l'individu I₆ :

- En projetant l'individu sur le 1^{er} axe, on remarque qu'il est loin de l'origine (la moyenne) et se situe dans le côté positif de l'axe. Et en se basant sur la signification des axes donnée précédemment, on peut conclure que l'individu I₆ a obtenu de bonnes notes en SYS1, RES1, ANUM, RO et THP.
- En projetant l'individu sur le 2^{ème} axe, on remarque qu'il est proche de l'origine (la moyenne). Donc il est un point moyen qui est mal représenté par l'axe 2.

Remarque :

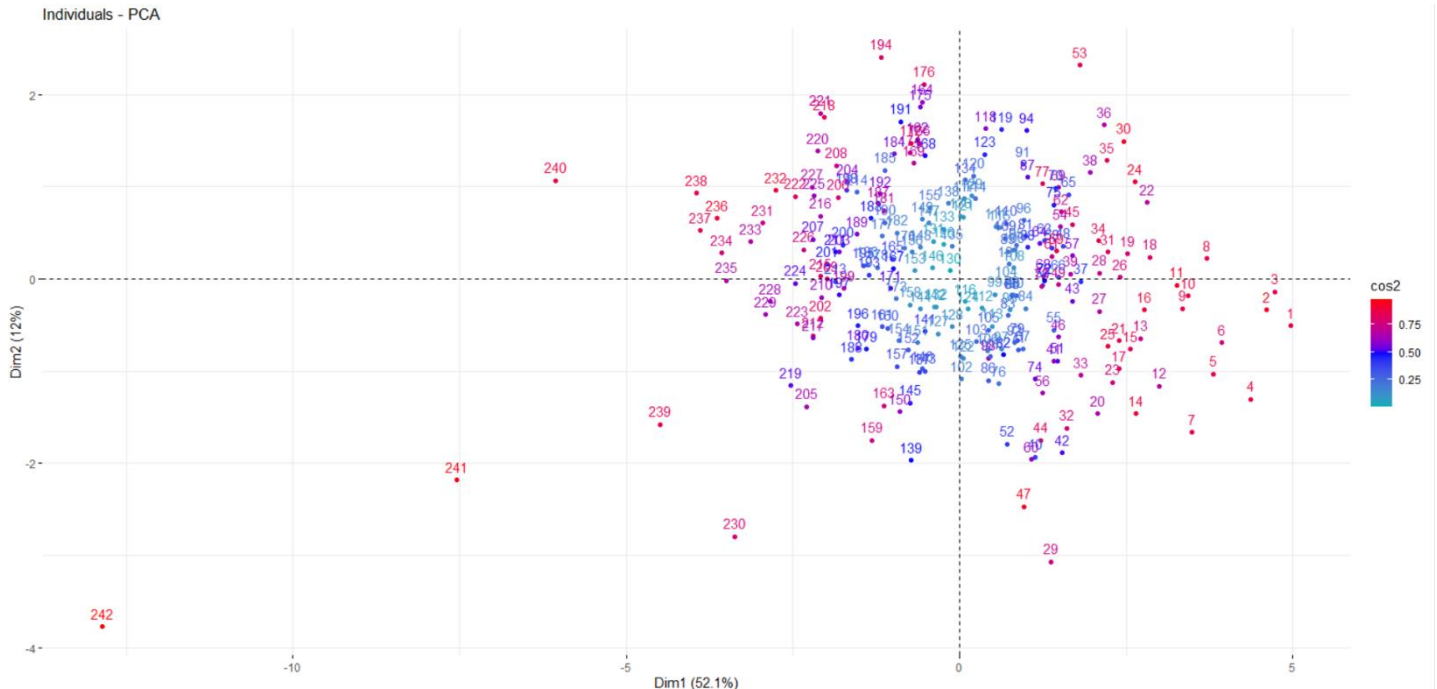
On remarque que les résultats obtenus dans l'interprétation graphique coïncident avec les interprétations obtenues dans les sections « 10 contribution relative des individus » et « 11 Qualité de représentation des individus »

13 Illustration de la qualité de représentation des individus :

Pour visualiser la qualité de représentation des individus sur le plan factoriel engendré par l'axe 1 et 2, nous avons utilisé la library « *factoextra* ».

```
#cos2 : represente la qualité de representation des individus sur le plan  
factoriel engendré par l'axe 1 et 2  
fviz_pca_ind(acp, col.ind = "cos2", gradient.cols = c("#00AFBB", "blue",  
"red"))
```

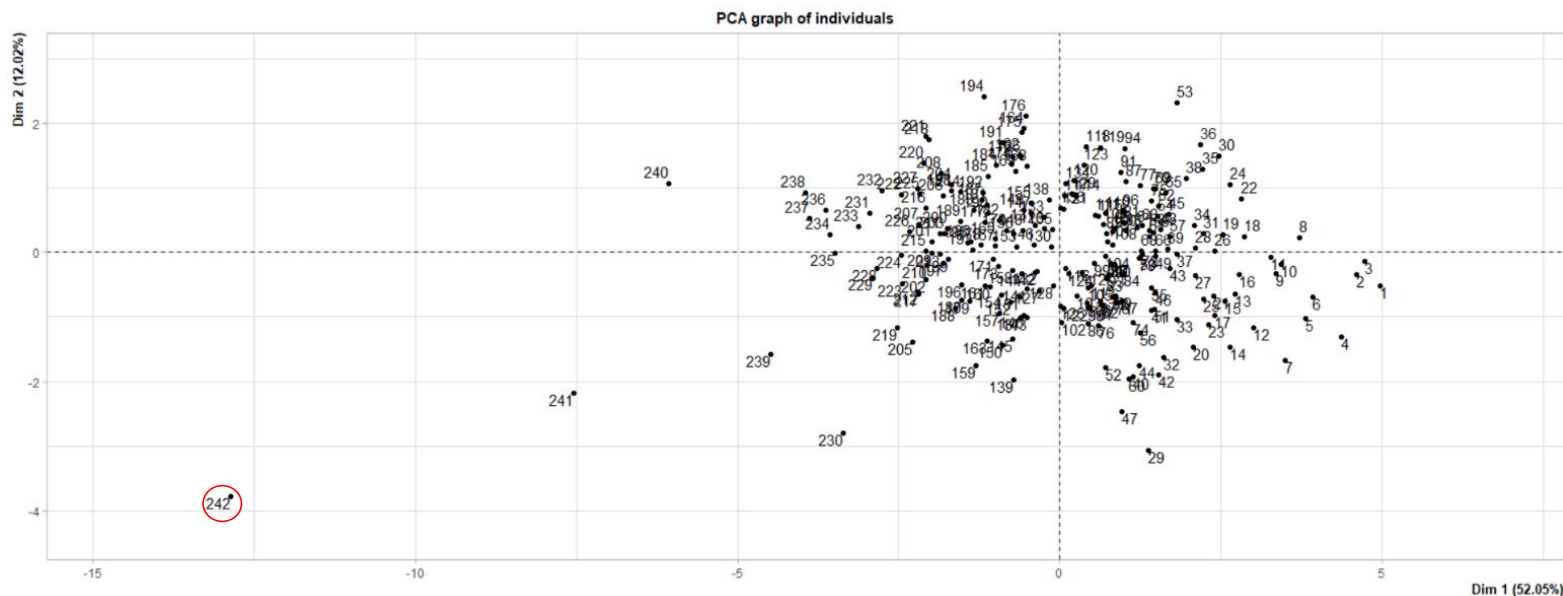
Le résultat obtenu est donné ci- dessous :



Nous avons remarqué que :

- Les individus très proches de l'origine ont une très mauvaise qualité de représentation sur le 1 er plan factoriel et ils sont coloriés par un bleu clair car $\cos^2 < 0,25$
- Les individus proches de l'origine ont une mauvaise qualité de représentation sur le 1 er plan factoriel et ils sont coloriés par la couleur bleue car $0,25 < \cos^2 < 0,5$
- Les individus assez loin de l'origine ont une qualité de représentation moyenne sur le 1 er plan factoriel et ils sont coloriés par la couleur violet car $0,5 < \cos^2 < 0,75$
- Les individus loin de l'origine ont une bonne qualité de représentation sur le 1 er plan factoriel et ils sont coloriés par la couleur rouge car $\cos^2 > 0,75$

14 Détails supplémentaire sur l'interprétation des individus sur le premier plan



On remarque que l'individu 242 est très éloigné du nuage des individus et loin de l'origine (la moyenne) et se situe dans le côté négatif des deux axes. Et en se basant sur la signification des axes donnée précédemment, on peut conclure que l'individu I₄ a obtenu de très mauvaises notes en SYS1, RES1, ANUM, RO et THP (1^{er} axe) et en ORG et LANG (2^{eme} axe).

Après inspection des données :

SYS1: 0.00 < 6.92	RES1: 0.00 < 7.85	ANUM: 0.00 < 8.31	RO: 0.00 < 6.66	ORG: 0.00 < 6.94	IGL: 0.00 < 6.62	THP: 0.00 < 5.00	LANG1: 0.00 < 7.66	0.00
-------------------	-------------------	-------------------	-----------------	------------------	------------------	------------------	--------------------	------

Nous avons constaté que l'individu 242 a eu des zéro dans tous les modules, ce qui signifie qu'il était un abondant.

15 Conclusion :

Pour conclure, on a pu réduire la dimensionalité du problème de départ à partir de 8 variables en 3 dimensions avec une conservation de 72,468 % de l'information totale. A partir de l'ACP faite on a pu répondre à la problématique et on peut dégager les résultats suivants :

- Les variables SYS1, RES1, ANUM, RO et THP représentant les notes obtenues dans ces modules sont linéairement dépendantes
- Les variables ORG et LANG représentant les notes obtenues dans ces modules sont linéairement dépendantes

Annexe : Script complet

Ci-dessous vous trouverez le script complet implémenté :

```
library("readxl")
data = read_excel("NotesPVS1-S2.xlsx", sheet=2) # 2 Represente le numÃ©ro
de la feuille du fichier excel
library(dplyr)

data = select(data, -starts_with("...")) # supprimer les colonnes qui
commencent par ...
data = select(data, -c('Matricule', `Nom & Prénom`, 'Groupe', `Ne S1`,
`Rang S1`))

data = plyr::rename(data, c(`SYS1 * x 5` = 'SYS1', `RES1 * x 4` = 'RES1',
`ANUM * x 4` = 'ANUM', `RO * x 3` = 'RO', `ORG * x 3` = 'ORG', `IGL * x
5` = 'IGL', `THP * x 4` = 'THP', `LANG1 * x 2` = 'LANG', `Moy S1` =
'MoyS1'))

data$SYS1 = gsub("[A-Z0-9]+: ([0-9\\.]+) < [0-9\\.]+", "\\1", data$SYS1)
data$RES1 = gsub("[A-Z0-9]+: ([0-9\\.]+) < [0-9\\.]+", "\\1", data$RES1)
data$ANUM = gsub("[A-Z0-9]+: ([0-9\\.]+) < [0-9\\.]+", "\\1", data$ANUM)
data$RO = gsub("[A-Z0-9]+: ([0-9\\.]+) < [0-9\\.]+", "\\1", data$RO)
data$ORG = gsub("[A-Z0-9]+: ([0-9\\.]+) < [0-9\\.]+", "\\1", data$ORG)
data$IGL = gsub("[A-Z0-9]+: ([0-9\\.]+) < [0-9\\.]+", "\\1", data$IGL)
data$THP = gsub("[A-Z0-9]+: ([0-9\\.]+) < [0-9\\.]+", "\\1", data$THP)
data$LANG = gsub("[A-Z0-9]+: ([0-9\\.]+) < [0-9\\.]+", "\\1", data$LANG)

data = na.omit(data) #supprimer les donnÃ©es manquantes
data = as.data.frame(apply(data, 2, as.numeric)) #conversion des type des
colonnes en numeric

summary(data) #Avoir des statistique concernant le tableau de donnÃ©es

library(FactoMineR)
acp = PCA(data[,-9]) #Toutes les colonnes à part la dernière qui represente
MoyS1
summary(acp)

install.packages("factoextra")
library("factoextra")
fviz_eig(acp) #Représentation de l'eboulis des valeurs propres

#cos2 : represente la qualité de representation des individus sur le plan
factoriel engendré par l'axe 1 et 2
fviz_pca_ind(acp, col.ind = "cos2", gradient.cols = c("#00AFBB", "blue",
"red"))
```