- Analyse de données -
TP 3 corrigé

---

1 Introduction

1) Calculer la matrice de corrélation des 8 variables :

```
> round(cor(portable2),digits=2)
```

```
Commodite Acoustiq TpsCharge AutoVeille Autoconv PuisGSM SensiGSM SensiDCS
Commodite 1.00 -0.12 0.54 -0.03 -0.01 0.17 0.21 0.15
Acoustiq -0.12 1.00 -0.12 0.06 0.11 -0.43 0.23 0.09
TpsCharge 0.54 -0.12 1.00 0.19 -0.30 0.25 -0.18 -0.36
AutoVeille -0.03 0.06 0.19 1.00 0.18 -0.03 0.16 0.05
Autoconv -0.01 0.11 -0.30 0.18 1.00 -0.27 -0.04 -0.11
PuisGSM 0.17 -0.43 0.25 -0.03 -0.27 1.00 0.34 0.29
SensiGSM 0.21 0.23 -0.18 0.16 -0.04 0.34 1.00 0.93
SensiDCS 0.15 0.09 -0.36 0.05 -0.11 0.29 0.93 1.00
```

Quelles sont les variables les plus corrélées entre elles ? Les moins corrélées entre elles ? Quelles sont les variables les plus opposées ?

Le couple le plus corrélé ($r \simeq 1$) est `(SensiDCS, SensiGSM)` avec $0,93$
Le couple le moins corrélé ($r \simeq 0$) est `(AutoConv, Commodite)` avec $-0,01$.
Le couple le plus opposé ($r \simeq -1$) est `(PuisGSM, Acoustiq)` avec $-0,43$.

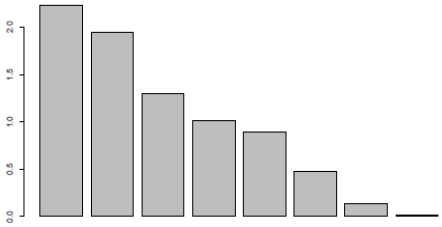
2 Première analyse en composantes principales

2) On fait une analyse en composantes principales des données centrées-réduites. On obtient les valeurs propres suivantes :

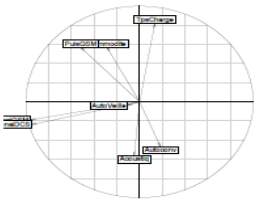
```
> round(acp$eig, digits=2)
[1] 2.23 1.94 1.30 1.01 0.89 0.47 0.13 0.02
```

3) *Faire une représentation en histogramme des valeurs propres. Combien de composantes principales faut-il retenir ? Quel est le pourcentage d'inertie totale expliquée par le sous-espace principal correspondant ? Si on ne retient que le premier plan principal, quel est le ratio expliqué ? La qualité de l'ACP est-elle bonne si on ne conserve que le premier plan principal ?*

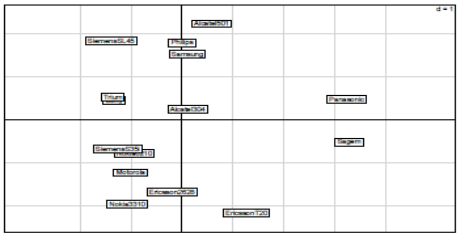
L'inertie totale est 8 et l'inertie portée par les 3 premiers axes est $2,23+1,94+1,30 = 5,47$. Cela représente 68% de l'inertie totale. Si on ne considère que les deux premières valeurs propres, on n'a que 52% de l'inertie totale. On voit que les valeurs propres ne décroissent pas très vite : la qualité de l'ACP ne sera pas très bonne.



4) *Représenter le cercle des corrélations du premier plan principal (axe 1 en abscisse, axe 2 en ordonnée), accompagné du tableau des corrélations entre les variables et les 2 (ou 3) premiers facteurs, la projection des individus sur le premier plan principal et les contributions (en 10000ièmes) de ces individus aux 2 (ou 3) premiers facteurs :*



```
> round(pca2$co,digits=2)
             Comp1 Comp2 Comp3 Comp4 Comp5
Commodite    -0.29  0.57 -0.46  0.23  0.52
Acoustiq     -0.05 -0.56 -0.44  0.58 -0.19
TpsCharge     0.13  0.82 -0.45  0.17 -0.14
AutoVeille   -0.11 -0.04 -0.66 -0.53 -0.49
Autoconv      0.18 -0.47 -0.38 -0.47  0.56
PuisGSM      -0.52  0.57  0.30 -0.31 -0.07
SensiGSM     -0.96 -0.19 -0.14  0.04 -0.01
SensiDCS     -0.95 -0.24  0.07  0.02  0.04
```



```
> inertia.dudi(pca2,ro=T)$row.abs
             Axis1 Axis2 Axis3 Axis4 Axis5
Philips          0  1096  1115   905  1342
Nokia3310      345  1340   389   178   519
Nokia6210      263   208   188  1118   793
Sony           536    65   161  2330   346
SiemensSL45    583  1168   381   327    65
EricssonT20    479  1637   549    66     0
SiemensS35i    478   163   118   302   311
Alcatel304       3    20     4     1  2090
Samsung          3   800  1687   513   543
Trium          549    96   243   461  1234
Motorola       303   526     2    90  1429
Panasonic     3127    80   356  1753    49
Alcatel501     102  1720   436  1634    49
Ericsson2628    12   987  4341    29   965
Sagem         3215    95    29   295   266
```

Quelles sont les variables qui déterminent la première composante principale ? La proximité des ces variables sur le cercle est-elle cohérente avec ce que l'on avait remarqué en 1) ?

La première composante principale est déterminée négativement par `SensiGSM` et `SensiDCS`. On le voit sur le cercle, mais aussi par les corrélations, respectivement $-0,96$ et $-0,95$. Il n'y a pas de variable fortement corrélée positivement avec cet axe. Comme ces variables sont très près du cercle, on peut déduire de leur proximité dans le cercle de corrélation qu'elles sont aussi très corrélées l'une à l'autre. En effet, ce sont les deux variables notées à la question 1 comme ayant une corrélation de $0,93$.

Quels sont les 2 individus qui déterminent le plus la première composante principale ? Quelle est leur contribution aux axes ? Comparer ces contributions au poids respectif de ces individus. Quel est le lien entre les variables de la question précédente et ces 2 individus ?

Les individus qui déterminent la première composante principale sont `Panasonic` et `+Sagem+`. On le voit sur la projection des individus mais surtout par leur contribution aux axes (qui sont données ici en 10000 ièmes) :

- `Panasonic : 0,3127`
- `Sagem : 0; 3215`

Ces valeurs sont à comparer avec le poids de chaque individu, qui est ici $1/15 = 0,07$. Elles sont très au dessus des valeurs normales, puisque les contributions des individus somment à 1 et que ces deux éléments à eux seuls en représentent presque les deux tiers.

D'une part le premier axe est déterminé par `SensiGSM` et `SensiDCS`. D'autre part, il est déterminé par les deux individus `Panasonic` et `Sagem`. Il est en fait facile de vérifer que ces deux portables ont une note hors norme pour ces deux variables, ce qui explique les caractéristiques de la première composante principale.

3 Nouvelle ACP

On décide de faire une nouvelle analyse en composantes principales sur les variables centrées-réduites en passant les 2 téléphones mobiles perturbateurs en éléments supplémentaires.

5) *Calculer les mêmes caractéristiques que dans la section précédente (valeurs propres, corrélations entre variables et facteurs, projection des individus sur le premier plan principal et contributions des individus aux axes, qualité de représentation des individus) :*

```
> round(acp2$co, digits=2)
          Comp1 Comp2 Comp3 Comp4 Comp5
Commodite -0.24 -0.68 0.12 -0.63 0.19
Acoustiq 0.71 -0.04 0.21 0.32 0.55
TpsCharge -0.63 -0.67 0.22 0.13 0.18
AutoVeille 0.10 -0.40 0.70 0.41 -0.36
Autoconv 0.24 0.38 0.70 -0.45 -0.16
PuisGSM -0.78 -0.14 -0.22 0.20 -0.19
SensiGSM 0.75 -0.59 -0.11 0.02 -0.06
SensiDCS 0.74 -0.38 -0.42 -0.12 -0.34
```

On trouve les variables qui déterminent les axes en regardant les coefficients de corrélation. On se fixe une limite par exemple à $0,60$. Le premier axe est caractérisé par `SensiGSM` (corrélation de $0,75$), `SensiDCS` $(0,74)$ et `Acoustiq` $(0,71)$ d'une part ;
`PuisGSM` $(-0,78)$ et `TpsCharge` $(-0,63)$ d'autre part.
Le deuxième axe est plus lié à `Commodite` $(-0,68)$, `TpsCharge` $(-0;67)$ et `SensiGSM` $(-0,59)$.
Le troisième axe correspond à `AutoVeille` $(0,70)$ et `AutoConv` $(0,70)$.

```
> abs(inertia.dudi(pca3,ro=T)$row.rel)
            Axis1 Axis2 Axis3 Axis4 Axis5 con.tra
Philips      4699 418 3162 1069 2 1004
Nokia3310    6658 97 1091 566 218 653
Nokia6210    2605 59 1224 2295 3342 438
Sony         1144 3855 891 1102 2893 594
SiemensSL45    54 7997 243 135 1 799
EricssonT20 1061 3267 3923 42 647 1106
SiemensS35i 6209 1901 144 630 153 334
Alcatel304  1850 395 252 6119 757 629
```

```
Samsung     4940 544 2143 282 1409 895
Trium         44 1058 1191 3056 3307 497
Motorola    5929 41 270 3709 11 447
Alcatel501  8060 289 7 4 697 1355
Ericsson2628 627 5147 3366 707 103 1247
```

La qualité de la représentation d'un individu sur un axe se mesure par le cosinus carré de l'angle entre l'individu et l'axe. En se reportant à la table ci-dessus, on voit que les deux individus vraiment mal représentés sur le premier axe sont `Trium` $(0,0044)$ et `SiemensSL45` $(0,0054)$.

Pour trouver les individus mal représentés sur le sous-espace formé par les trois premiers axes principaux, il faut calculer le cosinus carré des individus avec le sous espace formé par les trois premiers axes principaux. Ce cosinus carré est en fait la somme des valeurs données dans les trois premières colonnes, puisque les cosinus carrés sont additifs (sous-espaces orthogonaux). En additionant les trois colonnes, on trouve :

```
> inertia.dudi(pca3,ro=T)$row.cum[,3]
Philips Nokia3310 Nokia6210 Sony SiemensSL45 EricssonT20 SiemensS35i Alcatel304 Samsung
Trium Motorola Alcatel501 Ericsson2628
8279 7846 3888 5889 8295 8251 8255 2498 7627 2293 6240 8356 9140
```

Les deux téléphones les moins bien représentés sont donc `Trium` $(0,2293)$ et `Alcatel304` $(0,2498)$.
On remarque que la somme des cosinus carrés n'est pas exactement 1 (erreurs d'arrondis) ; par exemple pour `Motorola`, on a : $0,5929+0,0041+0,0270+0,3709+0,0011+0,0447 = 1,0407$.

6) *Expliquer en quoi la nouvelle analyse est différente de la précédente et pourquoi elle est utile. Combien de valeurs propres faut-il retenir, et quelle proportion d'inertie totale cela représente-t-il ?*

Cette nouvelle ACP est différente car les téléphones `Panasonic` et `Sagem` ont été retirés de l'analyse. On espère donc que les axes principaux ne seront plus "tirés" vers ces deux éléments, qui avaient une trop forte influence. Il est toujours possible d'interpréter ces deux éléments (et ils sont d'ailleurs représentés sur la projection des individus sur le premier axe principal). Pour cela, il suffira de calculer leurs coordonnées sur chaque axe principal à partir des coordonnées de la composante principale. Pour résumer, ces éléments n'ont pas servi à l'analyse, mais on peut voir comment ils se comportent vis-à-vis des résultats de cette analyse.
Si on retient 3 valeurs propres, cela correspond à une inertie de $2,73+1,72+1,32 = 5,77$, soit 72% de l'inertie totale. On a donc une légère augmentation de la qualité globale de la représentation.

7) On représente les individus supplémentaires. *Les 2 individus supplémentaires sont-ils bien représentés sur le plan principal ?*

```
> round(acp2.sup$lisup, digits=2)
Axis1 Axis2 Axis3 Axis4 Axis5 Axis6 Axis7 Axis8
Panasonic -2.43 2.34 1.99 -0.12 3.52 -2.43 0.03 0.92
Sagem -2.13 3.59 2.04 0.52 1.15 -3.08 -0.36 -0.20
```

Les coordonnées des deux éléments supplémentaires sur la projection des individus sur le premier plan principal sont comparables aux autres. Pour savoir si les éléments sont biens représentés par les 3 premiers axes principaux, il faut calculer leur cosinus carré par rapport au sous espace. Pour cela, on élève les données ci-dessus au carré et on divise chaque ligne par sa somme, ce qui donne les cosinus carrés des angles par rapport aux 3 premiers axes (en 10000ièmes) :

Pour cela, on calcule leur cosinus carré :

```
> acp2cos2sup=acp2.sup$lisup^2/apply(acp2.sup$lisup^2,1,sum)

> round(acp2cos2sup[,1 :3]*10000)
          Axis1 Axis2 Axis3
Panasonic 1708 1592 1151
Sagem     1379 3934 1265
```

Les données que nous recherchons sont obtenues en sommant les deux lignes ci-dessus, donnant respectivement 4451 et 6578. Les deux téléphones sont donc plutôt mal représentés (respectivement $0,4451$ et $0,6578$ sur les trois premiers axes), mais ils sont quand même dans la moyenne au regard des autres cosinus carrés.