Analyse en composantes principales du comportement bancaire des clients

Quelques sorties R

Makhatch ABDULVAGABOV Floran (son nom) Samuel (son nom)

Janvier 2020

Table des matières

1 Présentation du document

Le jeu de données étudié permet d'analyser le comportement bancaire des 500 clients. Les données sont issues d'une enquête réalisée régulièrement par une banque pour créer de nouveaux produits afin de fidèliser les clients.

1.1 Présentation des données

Le jeu de données `cbg` contient 500 individus (clients) sur lesquels on mesure 10 variables :

- solde : solde moyen du compte courant sur les 12 derniers mois (en euros) ;
- mdecouv : montant cumulé des découverts sur le compte courant durant les 12 derniers mois (en euros) ;
- ncompte : nombre de comptes utilisés en plus du compte courant (par exemple les livrets . . .) ;
- memprunt : Montant total des retraits effectuées sur le livret d'épargne sur les 12 derniers mois (en euros) ;
- mdepot : montant total des versements effectuées sur le livret d'épargne lors des 5 dernières années (en euro) ;
- mretrait : montant total des retraits effectuées sur le livret d'épargne sur les 12 derniers mois (en euros) ;
- nbenf : nombre d'enfants de moins de 18 ans ;
- age : age du client enquêté ;
- csp : catégorie socio-professionnelle du client ;
- code : codification de la catégorie socio-professionnelle du client ;

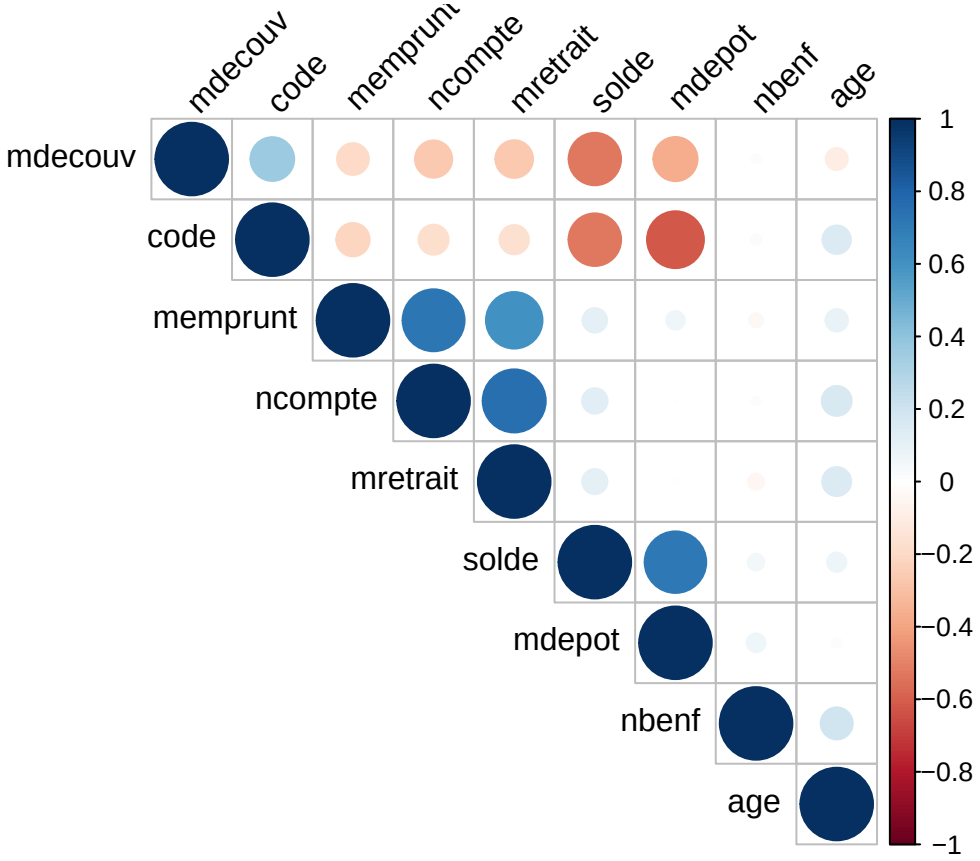
1. artisant - commerçant

2. cadre
3. employé
4. ouvrier
5. retraité
6. autre ;

```
##      solde            mdecouv          ncompte          memprunt
##  Min.   :   2.89   Min.   :   0.0   Min.   : 0.000   Min.   :    0.0
##  1st Qu.: 208.63   1st Qu.:   0.0   1st Qu.: 1.000   1st Qu.:    0.0
##  Median : 681.07   Median : 168.1   Median : 2.000   Median :  910.1
##  Mean   :1415.91   Mean   : 475.5   Mean   : 2.628   Mean   : 7685.7
##  3rd Qu.:2759.82   3rd Qu.: 951.1   3rd Qu.: 3.000   3rd Qu.: 9445.5
##  Max.   :7546.98   Max.   :1986.3   Max.   :10.000   Max.   :80735.9
##      mdepot            mretrait          nbenf            age
##  Min.   :     1.52   Min.   :    0.0   Min.   :0.000   Min.   :18.00
##  1st Qu.:   788.48   1st Qu.:  825.3   1st Qu.:0.000   1st Qu.:28.00
##  Median :  3389.36   Median : 1846.9   Median :0.000   Median :37.00
##  Mean   : 21416.31   Mean   : 3356.2   Mean   :0.532   Mean   :40.97
##  3rd Qu.: 31974.96   3rd Qu.: 4160.7   3rd Qu.:1.000   3rd Qu.:53.00
##  Max.   :145879.13   Max.   :20250.1   Max.   :7.000   Max.   :78.00
##                  csp           code
##  artisan-commercant: 38   Min.   :1.000
##  autre             :113   1st Qu.:3.000
##  cadre             : 80   Median :4.000
##  employe           : 77   Mean   :3.906
##  ouvrier           :114   3rd Qu.:5.000
##  retraite          : 78   Max.   :6.000
```

On voit que le troisème quartile pour la variable nbenf est de 1 et la médianne vaut 0. Donc 75% des clients de la banque ont au plus un enfant et la moitié n'en a aucun. En moyenne l'âge des clients est de 37 ans et la catégorie socio-professionnelle la plus représentée est ouvrier. Un quart des clients a plus de 3 comptes.

1.2 Analyse des corrélations



2 Réalisation de l'ACP

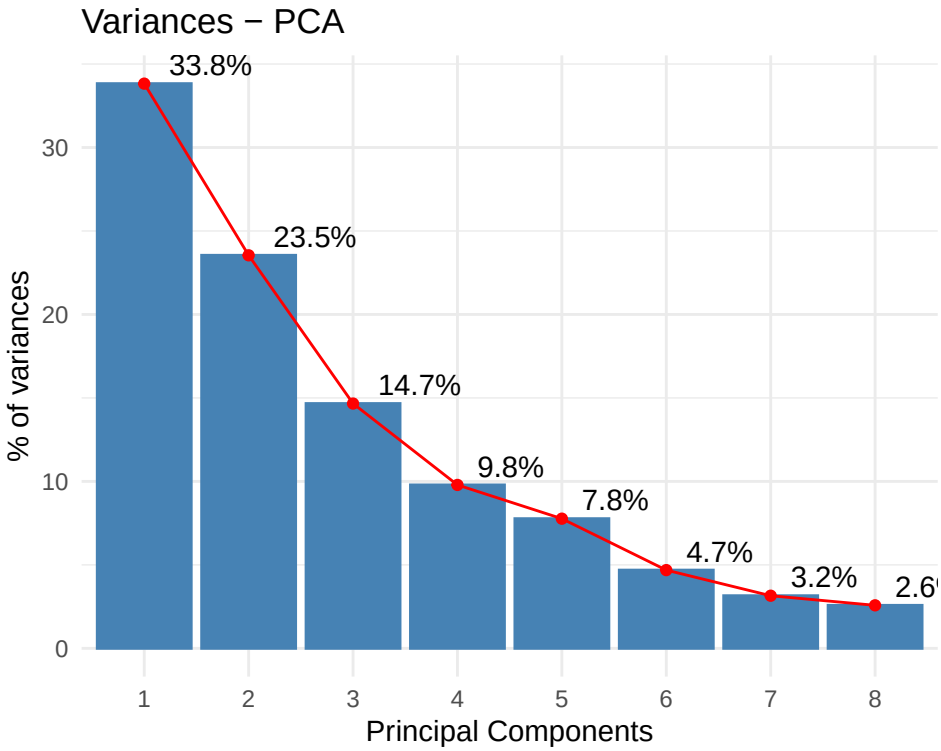
Pour réaliser l'analyse en composantes principales de ce jeux de données, les variables csp et code ne seront pas prises en compte (déclarées qualitative supplémentaire).

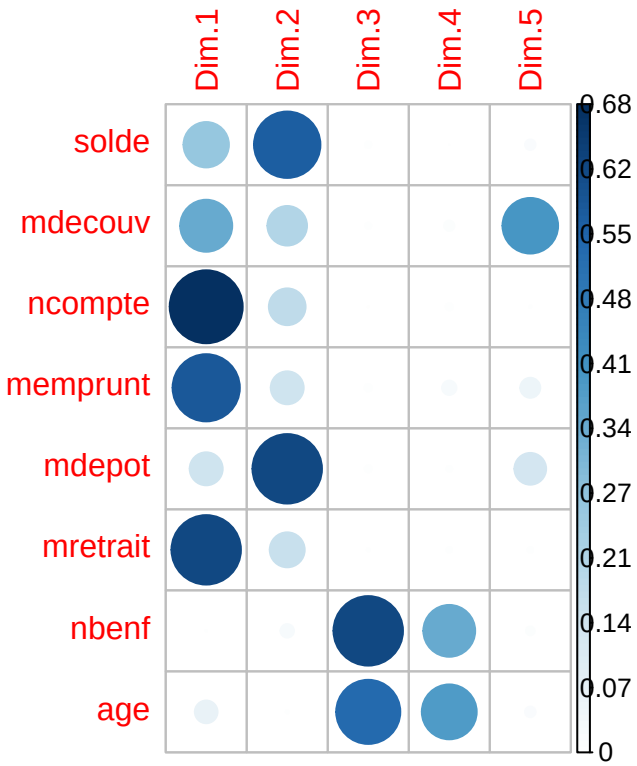
```
res.pca<-PCA(cbg, scale.unit = TRUE,  quali.sup=9:10, graph=F)
```

2.1 Choix de la dimension

Pour choisir le nombre de composantes principales, on peut s'aider des figures et du tableau suivants.

	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	2.7059279	33.824098	33.82410
Dim.2	1.8833722	23.542153	57.36625
Dim.3	1.1730582	14.663227	72.02948
Dim.4	0.7830071	9.787589	81.81707
Dim.5	0.6214060	7.767574	89.58464
Dim.6	0.3748707	4.685884	94.27053
Dim.7	0.2521917	3.152396	97.42292
Dim.8	0.2061662	2.577078	100.00000

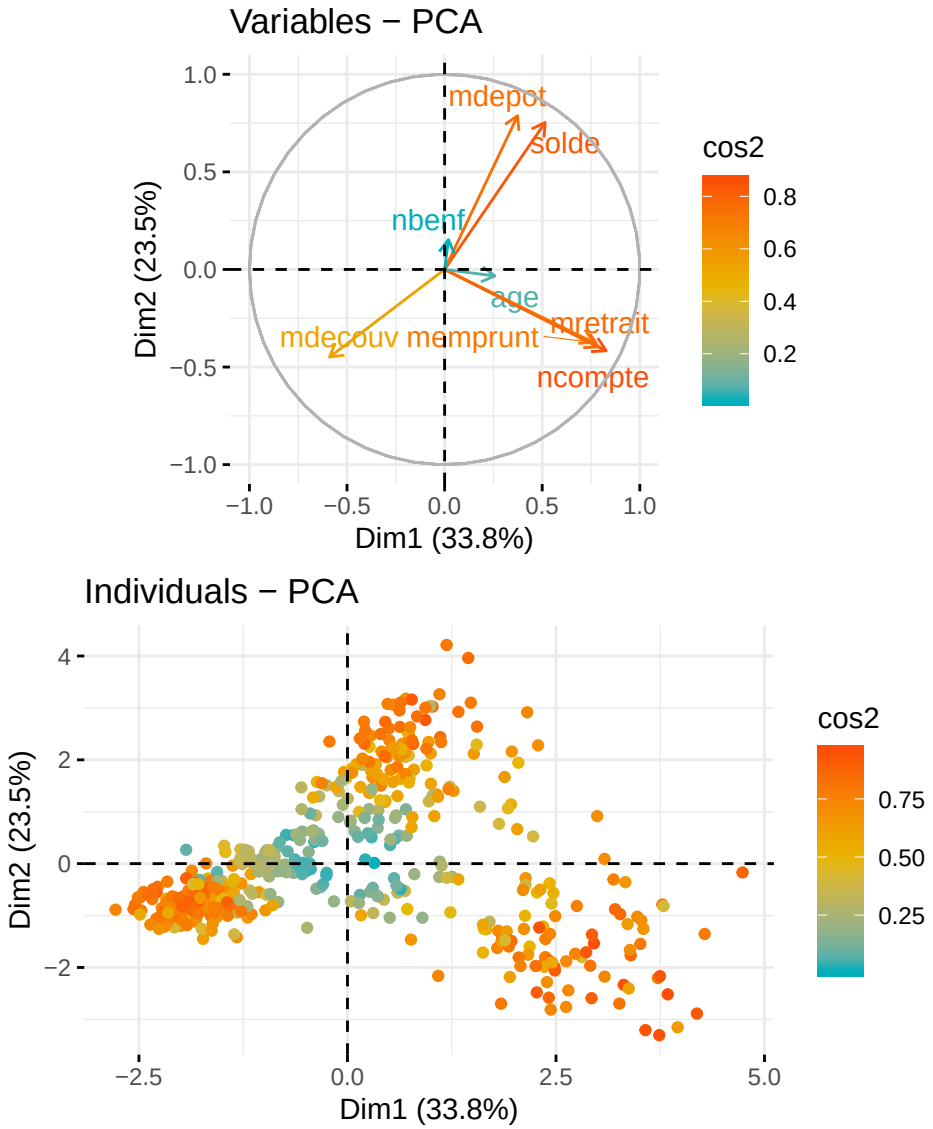




Grâce au tableau on voit qu'avec 4 dimensions on conserve presque 82% de la variance et sur le graph on voit que la dimension 5 n'apporte de que 7.8% de variance. De plus, sur la figure des contribution des variables aux dimensions, on voit que toutes les variables contribuent principalement aux 4 premières dimensions (sauf mdecouv qui contribue un peu à la dimension 5). Donc on peut s'arrêter aux 4 dimensions.

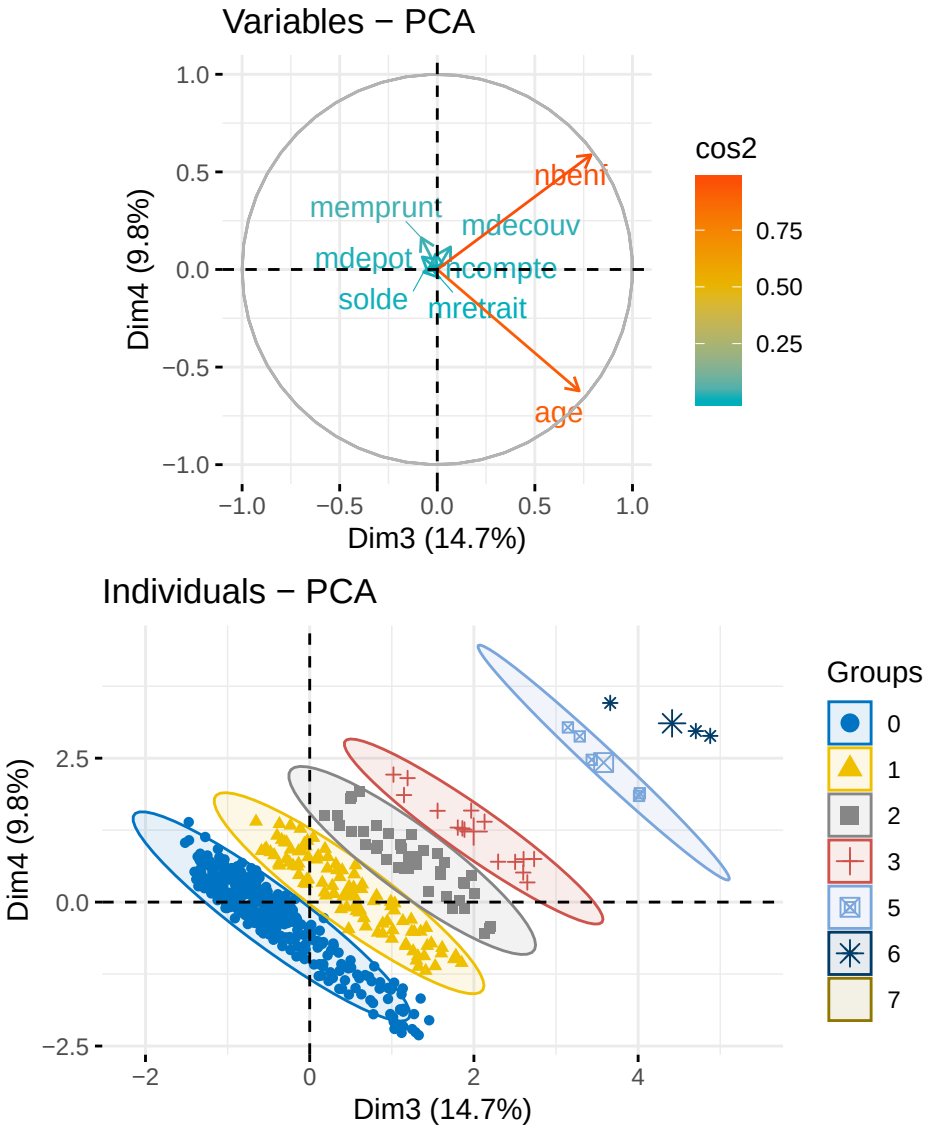
2.2 Cercles de correlation et nuages de points

2.2.1 Premier plan



Dans le premier plan (dimension 1-2), on voit sur le cercle de corrélations que les variables mdepot et solde sont corrélées entre elles. Ce résultat est assez intuitif, car plus on dépose d'argent sur le compte, plus le solde sera important. Il est aussi natuel de constater que la variable mdecouv est corrélée négativement à ces deux variables, car le fait d'être à découvert est opposé au fait d'avoir un solde important. On remarque aussi que les variables mretrait, memprunt et ncompte sont très corrélées entre elles. On pourrait supposer qu'une personne qui retire beaucoup d'argent est susceptible d'avoir plusieur comptes et faire des emprunts.
En outre, presque tous les vecteurs sont bient représentés sur ce plan (sauf le nombre d'enfants et l'âge), on le voit car ils sont assez proches de la circonférence du cercle et leur gradient de couleurs est dans les couleurs chaudes (ce qui correspont au cos2 proche de 1)
Sur le nuage de points, on voit qu'on pourrait séparer les individus en trois groupes dans ce plan. Le premier groupe des individus sont souvent à découvert et n'ont pas un solde élevé, le deuxième qui a tout à l'opposé et le troisième constitué des individus qui font souvent des emprunts et prélévements et qui ont plusieurs comptes

2.2.2 Deuxième plan



Dans le second plan (dimension 3-4), seuls les vecteurs nbenf, age sont bien représentés sur le cercle de corrélation.
Sur le nuage de points on distingue clairement plusieurs groupes d'individus (entourés par les ellipses). En superposant ce résultat avec le cercle de corrélations, on voit que la diretion dans laquelle les groupes sont divisés est la même que la direction du vecteur nbenf (nombre d'enfants). Ce qui explique aussi le caractere dicontinu.
À l'intérieur de chaque groupe (ellipse), on voit que les points ont une distrubition qui à l'aire d'être centrée sur le centre de l'ellipse. De même que précédemment en regardant le cercle de corrélation on constate que la direction dans laquelle les points sont distribuée est la même que la direction du vecteur age (on pourrait par exemple supposer que cette distribution suit une loi gaussienne).