Wine

Anthony LEZIN

8/19/2020

I. Généralités et premières observations

A. Préparation des données

1. Chargement des données, renommage et observations

```
wine=read.table("https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data",sep=",")

wine.names=c("Class","Alcohol","Malic acid","Ash","Alcalinity of ash","Magnesium","Total phenols","Flava

wine.short=c("Class","Al","Mal.ac","Ash","Ash.Alc","Mg","Tot ph","Fl","Ph.Nonfl.","Proant","Col.int","Hu

names(wine)=wine.short
attach(wine)
```

Traduction française des données

```
wine.names.fr=c("Classe","Alcool","Acide malique","Cendres","Alcalinité des cendres","Magnesium","Phénol

#équivalences des noms
equivalences = cbind(wine.names,wine.short,wine.names.fr)
kable(data.frame(equivalences))
```

wine.names	wine.short	wine.names.fr
Class	Class	Classe
Alcohol	Al	Alcool
Malic acid	Mal.ac	Acide malique
Ash	Ash	Cendres
Alcalinity of ash	Ash.Alc	Alcalinité des cendres
Magnesium	Mg	Magnesium
Total phenols	Tot ph	Phénols totaux
Flavanoids	Fl	flavonoïdes
Nonflavanoid phenols	Ph.Nonfl.	phénols non flavonoïdes
Proanthocyanins	Proant	Proanthocyanidines
Color intensity	Col.int	Intensité de la couleur
Hue	Hue	Teinte
OD280/OD315 of diluted wines	OD	OD280 / OD315 des vins dilués
Proline	Prol.	Proline

```
rownames(equivalences)=c()
colnames(equivalences)=c()
```

Normalisation des données

```
# centrage et réduction
wine.sc_num=data.frame(scale(wine[,-1]))
wine.sc=cbind(as.factor(wine[,1]),wine.sc_num)
colnames(wine.sc)[1]="Class"
```

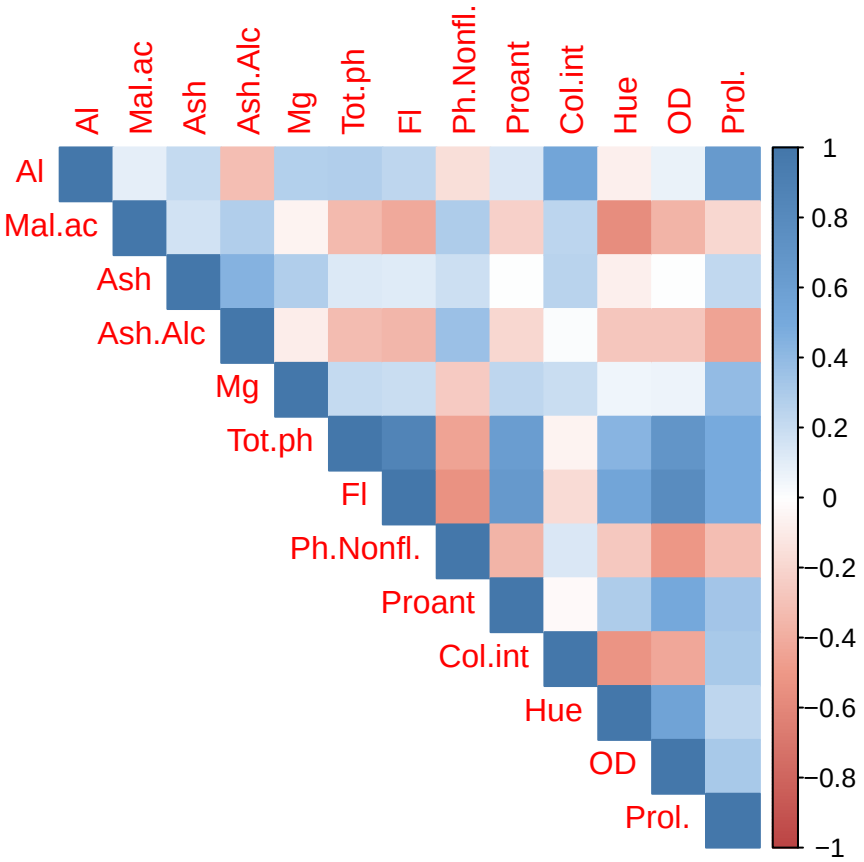
B. Premières observations

Voici le corrélogramme et la matrice de corrélation

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
C = cor(wine.sc_num)
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))
corrplot(C, method="color", col=col(200),type="upper")
```



```
cor_mat=matrix(round(C,digits=2),nrow=13)
row.names(cor_mat) =c("Al","Mal.ac","Ash","Ash.Alc","Mg","Tot ph","Fl","Ph.Nonfl.","Proant","Col.int","
colnames(cor_mat) =row.names(cor_mat)
kable(cor_mat)
```

	Al	Mal.ac	Ash	Ash.Alc	Mg	Tot ph	Fl	Ph.Nonfl.	Proant	Col.int	Hue	OD	Prol.
Al	1.00	0.09	0.21	-0.31	0.27	0.29	0.24	-0.16	0.14	0.55	-0.07	0.07	0.64
Mal.ac	0.09	1.00	0.16	0.29	-0.05	-0.34	-0.41	0.29	-0.22	0.25	-0.56	-0.37	-0.19
Ash	0.21	0.16	1.00	0.44	0.29	0.13	0.12	0.19	0.01	0.26	-0.07	0.00	0.22
Ash.Alc	-0.31	0.29	0.44	1.00	-0.08	-0.32	-0.35	0.36	-0.20	0.02	-0.27	-0.28	-0.44
Mg	0.27	-0.05	0.29	-0.08	1.00	0.21	0.20	-0.26	0.24	0.20	0.06	0.07	0.39
Tot ph	0.29	-0.34	0.13	-0.32	0.21	1.00	0.86	-0.45	0.61	-0.06	0.43	0.70	0.50
Fl	0.24	-0.41	0.12	-0.35	0.20	0.86	1.00	-0.54	0.65	-0.17	0.54	0.79	0.49
Ph.Nonfl.	-0.16	0.29	0.19	0.36	-0.26	-0.45	-0.54	1.00	-0.37	0.14	-0.26	-0.50	-0.31
Proant	0.14	-0.22	0.01	-0.20	0.24	0.61	0.65	-0.37	1.00	-0.03	0.30	0.52	0.33
Col.int	0.55	0.25	0.26	0.02	0.20	-0.06	-0.17	0.14	-0.03	1.00	-0.52	-0.43	0.32
Hue	-0.07	-0.56	-0.07	-0.27	0.06	0.43	0.54	-0.26	0.30	-0.52	1.00	0.57	0.24
OD	0.07	-0.37	0.00	-0.28	0.07	0.70	0.79	-0.50	0.52	-0.43	0.57	1.00	0.31
Prol.	0.64	-0.19	0.22	-0.44	0.39	0.50	0.49	-0.31	0.33	0.32	0.24	0.31	1.00

J'observe que :

- les **"Totals Phenols"** et les **"Flavanoids"** sont **extrêmement correlés**
- les **"OD"** sont **très corrélés¨** avec **"Totals Phenols"** ainsi que les **"Flavanoids"**
- Il y a une **bonne corrélation** entre les variables **"Alcohol"** et **"Proline"** d'une part, les **"Proanthocyanins"** avec les **"Totals Phenols"** ainsi que les **"Flavanoids"** d'autre part.

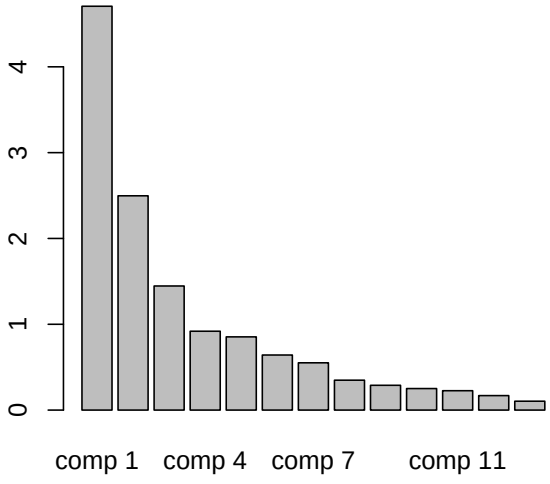
II. Sélection du nombre de composantes

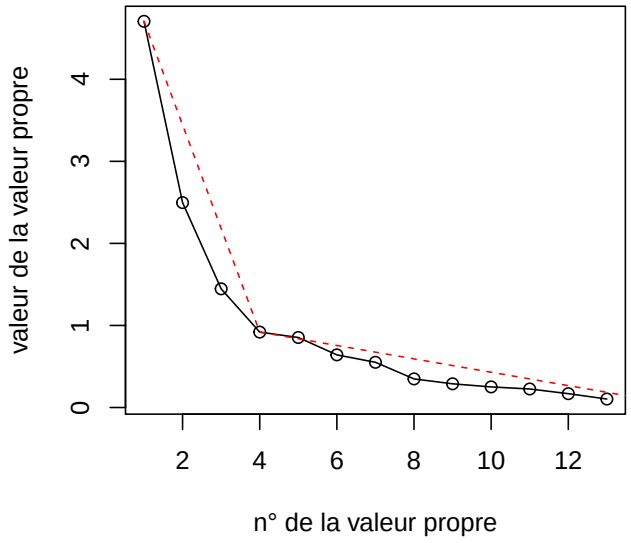
J'utilise le **critère de Kaiser** :
je ne retiens que les axes dont l'inertie est supérieure à l'inertie moyenne I/p.

```
par(mfrow=c(1,2))
# réalisation de l'ACP
library('FactoMineR')
res.pca = PCA(wine.sc_num,graph=FALSE)

#valeurs propres et composantes
kable(res.pca$eig)
```

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	4.7058503	36.1988481	36.19885
comp 2	2.4969737	19.2074903	55.40634
comp 3	1.4460720	11.1236305	66.52997
comp 4	0.9189739	7.0690302	73.59900
comp 5	0.8532282	6.5632937	80.16229
comp 6	0.6416570	4.9358233	85.09812
comp 7	0.5510283	4.2386793	89.33680
comp 8	0.3484974	2.6807489	92.01754
comp 9	0.2888799	2.2221534	94.23970
comp 10	0.2509025	1.9300191	96.16972
comp 11	0.2257886	1.7368357	97.90655
comp 12	0.1687702	1.2982326	99.20479
comp 13	0.1033779	0.7952149	100.00000

```
barplot(res.pca$eig[,1])

#critère du coude
plot(1:13,res.pca$eig[1:13],pch=1, type="o", xlab="n° de la valeur propre", ylab="valeur de la valeur pr
x.coude=c(1,4,14)
y.coude=c(res.pca$eig[1],res.pca$eig[4],res.pca$eig[13])
points(x.coude,y.coude, type="l",col="red",lty=2)
```





En ACP normée, $I/p = 1$, je ne retiens donc que les axes associés à des valeurs propre supérieures à 1, c-à-d, les 3 premiers ici.

On peut d'ailleurs vérifier ce résultat en utilisant le **critère du coude** :
sur l'eboulis des valeurs propres, j'observe un décrochement (coude) suivi d'une décroissance régulière à partir de la 4ème valeur propre. Je sélectionne les axes avant le décrochement, donc les 3 premiers.

La proportion d'inertie expliquée par les 3 premiers axes est de de 66.5 %. Cela reste acceptable pour 14 variables.

–> Je me restreins désormais à une ACP à 3 composantes.

```
res.pca = PCA(wine.sc_num,ncp=3,graph=FALSE)
```

III. Le plan factoriel principal

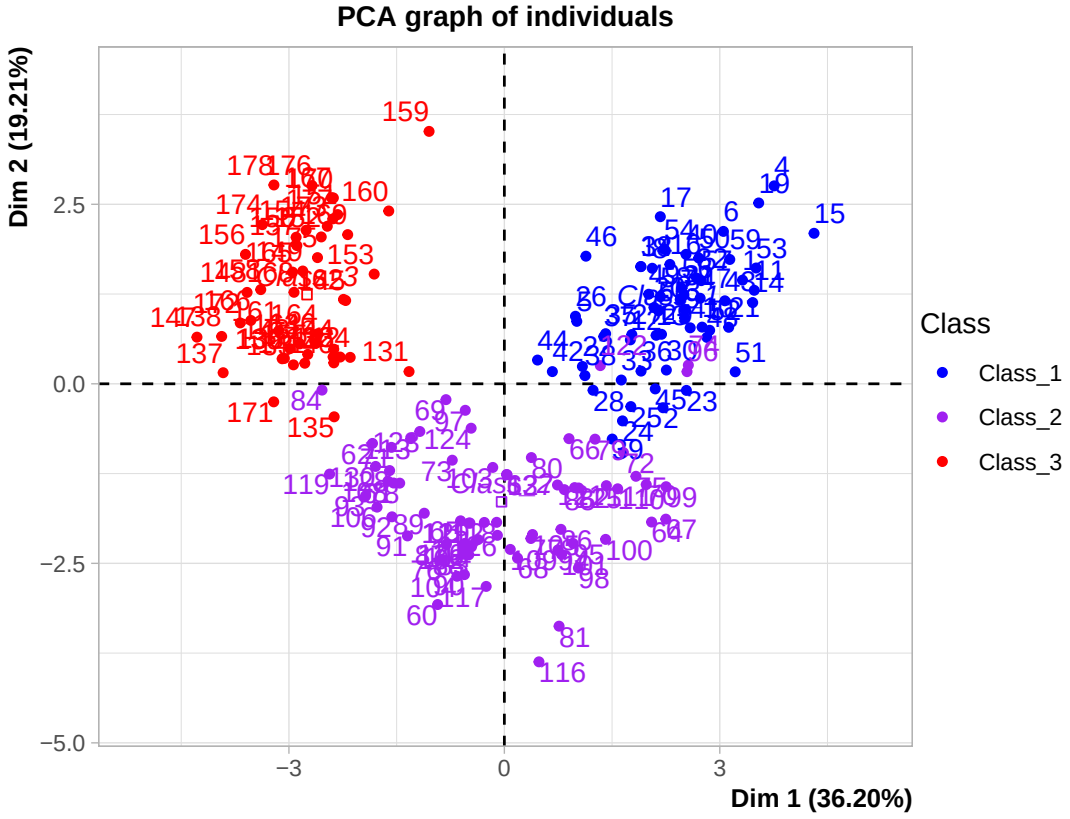
A. Description des axes selon les individus

Ajoutons les moyennes (par classe) à la liste des individus.

```
wine.sc_moy=aggregate(wine.sc[, 2:14], list(wine.sc$Class), mean)
names(wine.sc_moy)=names(wine.sc)
wine.sc_moy2=rbind(wine.sc_moy,wine.sc)
```

J'observe les individus obtenus par l'ACP en "séparant" les éléments selon leurs classes.

```
res.pca = PCA(wine.sc,ncp=3,quali.sup=1,graph=FALSE)
plot(res.pca,habillage=1,col.hab=c("blue","purple","red"),choix="ind")
```



C'est quasi parfait !
Le 1er plan de l'ACP semble parfaitement **départager les individus** selon leur **classes**.

Axe 1

Le signe des éléments sur l'axe 1 **oppose** parfaitement les individus des **classes 1** et **3**. PLus précisémment :

- les individus de la classe 1 ont des coordonnées positives sur cet axe et sont donc positivement corrélés avec cet axe.
- les individus de la classe 3 ont des coordonnées négatives sur axe et sont donc négativement corrélés avec cet axe.
- les individus de la classe 2, eux, semblent équitablement répartis autour de 0 sur cet axe.

Par ailleurs, les valeurs des individus sur cet axe semblent assez fidèlement départager les 3 classes, étant donné qu'à part quelques éléments atypiques par classe, on observe que :

- les individus de la classe 3 ont des valeurs appartenant à [-4.5;-2.5]
- les individus de la classe 2 ont des valeurs appartenant à [-2.5;2]
- les individus de la classe 1 ont des valeurs appartenant à [2;4.5]

Axe 2

Le signe des éléments sur l'axe 2 **oppose** quasi parfaitement le groupe des individus des **classes 1 et 3** à celui des indovidus de la **classe 2**, plus précisémment :

les individus de la classe 1 et 3 (respectivement de la classe 2) ont des coordonnées positives (respectivement négatives) sur cet axe et sont donc positivement (respectivement négativement) corrélés avec cet axe.

Individus bien représentés sur les axes par classe

Le graphique précédent semble montrer que certains individus suivants sont assez bien représentés par les axes.
Il s'agirait des individus 51, 23 pour la **classe 1**, 116, 17, 81 pour la **classe 2** et 137, 171 pour la **classe 3**.

```
li_ind=c(51,23,116,17,81,137,171)
A=NULL
for (i in 1:length(li_ind)){
  A=rbind(A,round(res.pca$ind$cos2[i,],digits=2))
}
rownames(A)=paste("individu", li_ind[1:7],sep=" ")
kable(A)
```

	Dim.1	Dim.2	Dim.3
individu 51	0.69	0.13	0.00
individu 23	0.43	0.01	0.36
individu 116	0.57	0.10	0.09
individu 17	0.60	0.32	0.00
individu 81	0.14	0.11	0.58
individu 137	0.60	0.29	0.03
individu 171	0.52	0.12	0.08

L'ensemble de ces points semblent convenablement représenter leurs axes, mais j'estime que la contribution sur les autres axes est encore trop importante.

Je décide procéder autrement en créant une fonction qui détermine les points possédant une contribution significative sur les axes sélectionnés et non significative sur les autres.
Ainsi, dans l'exemple ci-dessous : au moins 70 % pour l'axe représenté par le point et moins de 7% pour les autres axes.

```r
# Recherche de points représentatifs des axes
pts_car=function(j,a,b,c){
li=c(1,2,3,1,2)
A=NULL
  for (i in 1:178){
    if (round(res.pca$ind$cos2[i,],digits=2)[li[j]]>= 0.7 && round(res.pca$ind$cos2[i,],digits=2)[li[j+
      {A=cbind(A,c(i,round(res.pca$ind$cos2[i,],digits=2)))
      }
  }
  #t(A)
return(t(A))
}

# Points à fore contribution sur l'axe 1
kable(pts_car(1,0.7,0.07,0.07))
```

	Dim.1	Dim.2	Dim.3
21	0.70	0.04	0.01
23	0.74	0.00	0.01
36	0.75	0.01	0.04
137	0.78	0.00	0.00
141	0.71	0.01	0.03
161	0.72	0.04	0.00
163	0.73	0.03	0.06
166	0.77	0.05	0.01
171	0.82	0.01	0.06

```r
# Points à fore contribution sur l'axe 2
kable(pts_car(2,0.07,0.7,0.7))
```

	Dim.1	Dim.2	Dim.3
81	0.04	0.83	0.01
104	0.04	0.74	0.06
107	0.02	0.70	0.03
117	0.01	0.80	0.01

```r
# Points à fore contribution sur l'axe 3
kable(pts_car(3,0.07,0.07,0.7))
```

	Dim.1	Dim.2	Dim.3
26	0.05	0.05	0.77
122	0.05	0.00	0.75

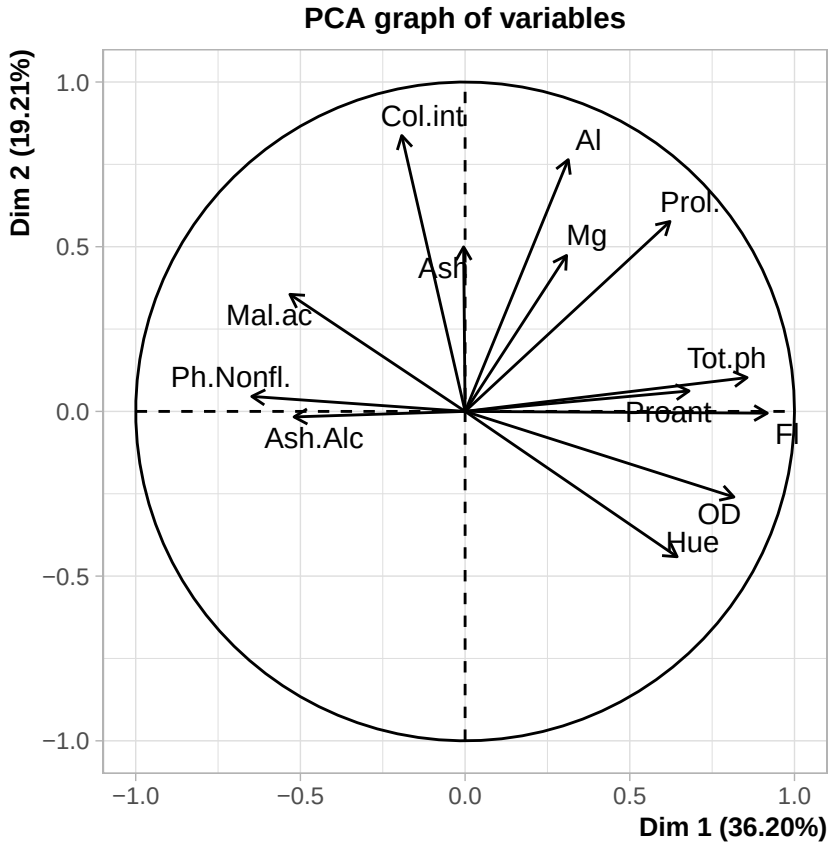
J'étudierai ces points ultérieurement après avoir émis une hypothèse sur les variables caractérisant les axes principaux sélectionnés par l'ACP.

Pour achever l'étude de ce graphique, je relève le fait que certains individus semblent atypiques. Il s'agit des indidividus "159" (pour la classe 3), "81, 116, 122, 74" (pour la classe 2) et "4, 19 et 15" à la rigueur (pour la classe 1).

B. Description des axes selon les variables

1. Le plan principal

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



Je vais désormais analyser le nuage des variables.
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.

Axe 1

L'axe 1 semble opposer 2 groupes de variables :

- **"Flavanoids", "Total.phenols", "OD 280.OD 315.of.diluted.wines"** et dans une moindre mesure **"Photocyanins"** et **"Hue"**
- **"Nonflavanoid.phenols"** et peut-être **"Malic Acid"** ainsi que **"Ash Alcanity"**

Pour toutes ces variables, la contribution sur l'axe 1 est assez important (tous les "|cos|" sont supérieurs à 0.5) contrairement à celle sur l'axe 2 qui semble peu élevée.

Le 1er groupe est positivement corrélé avec l'axe 1 contrairement au 2nd groupe.
Cela indique que les individus faisant partie du 1er groupe possèdent des valeurs des variables au-dessus de la

moyenne.

C'est le cas de **tous les individus** de la **classe 1**.

C'est le **phénomène contraire** pour ceux de la **classe 3** puisque l'ensemble des individus est négativement corrélé à cet axe. Ainsi, tous les individus faisant partie du 2nd groupe possèdent des valeurs de variables de ce groupe en-dessous de la moyenne.

C'est le cas de tous les individus de la **classe 3**.

Axe 2

l'axe 2 est positivement avec le groupe des variables **"Color.intensity"**, **"Alcohol"**,**"Ash"** et dans une moindre mesure **"Magnesium"**. Néanmoins, les 2 dernières variables ont une contribution peu élevée sur cet axe.

Il semblerait que la quasi totatlité des individus de le **classe 2** possèdent des valeurs de variables de ce groupe en-dessous de la moyenne.

Rq !

' Les axes 1 et 2 étant décorrélés par construction, il en de même pour les variables bien représentées par ces axes (par ex. **"Flavanoids"** pour l'axe 1 et **"Color Intensity"** pour l'axe 2).
Mêmes causes et mêmes effets avec les **"Flavanoids"** qui est totalement décorrélé du 2nd axe et donc de **"Ash"**.

On peut vérifier ces faits ici :

```
c(cor(wine.sc$Fl,wine.sc$Col.int),cor(wine.sc$Fl,wine.sc$Ash))
```

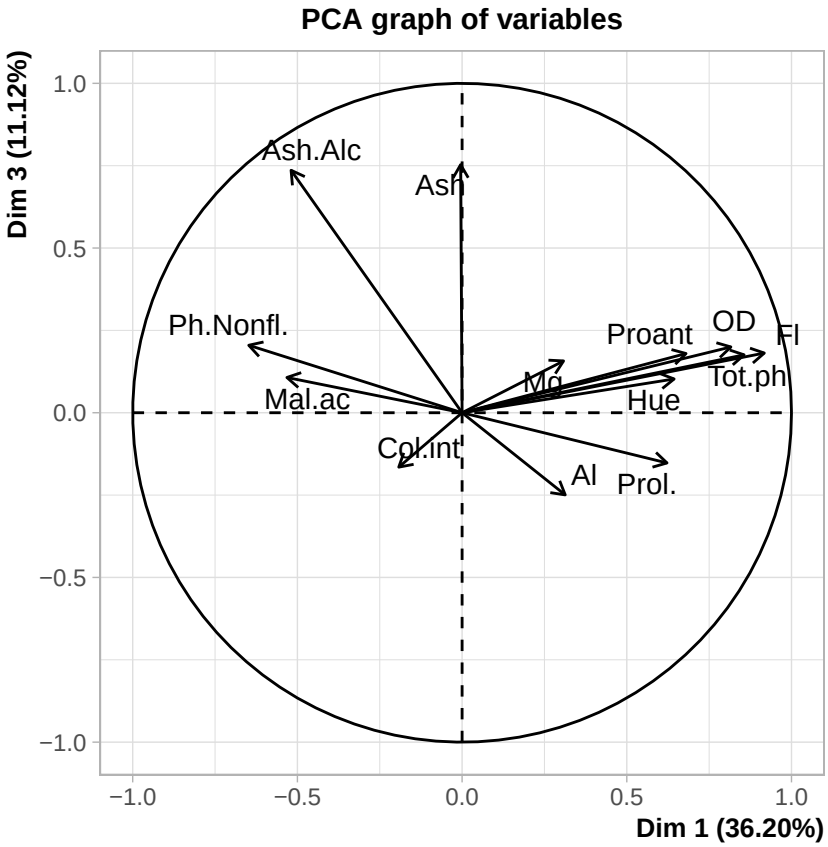
```
## [1] -0.1723794  0.1150773
```

Les coefficients sont relativement proches de 0.

2. Les plans secondaires

plan contenant les composantes 1 et 3

```
plot(res.pca,choix="varcor",axes=c(1,3))
```



L'axe 1 oppose toujours le groupe contenant les variables "Flavanoids","Total.phenols","OD 280.OD 315.of.diluted.wines" et dans une moindre mesure "Photocyanins" et "Hue" avec le groupe contenant les variables "Nonflavanoid.phenols" et dans une moindre mesure "Malic.acid".

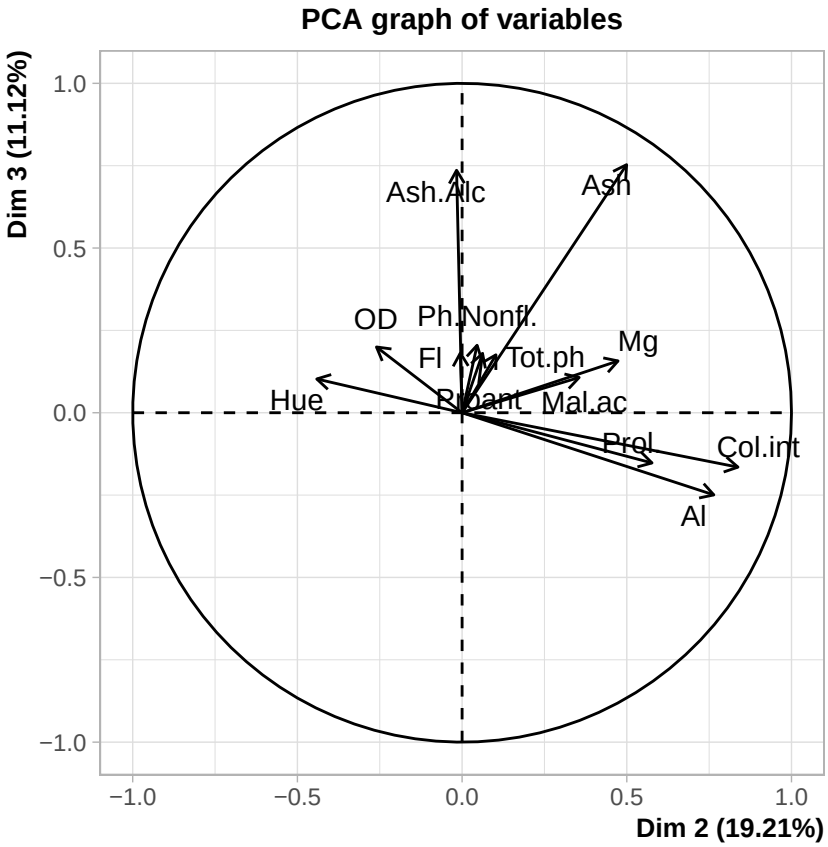
L'axe 3, lui, semble être caractérisé par 2 variables :

- "Ash" (dont la contribution semble bien plus importante que sur l'axe 2 dans l'étude du premier plan factoriel)
- "Alacanity.of.ash"

Ces variables sont positivement correllés avec l'axe. Il est par ailleurs intéressant de noter que cet axe contient peu d'informations sur les autres variables.

plan contenant les composantes 2 et 3

```
plot(res.pca,choix="varcor",axes=c(2,3))
```



Les principales variables influant dans l'axe 2 possèdent une corrélation positive avec cet axe.
Ce sont : **"Color.intensity", "Alcohol","Proline"** (dans une moindre mesure). **"Magnesium"** et **"Hue"** semblent désormais apporter une contribution plus faible.

L'axe 3, lui, semble toujours être caractérisé par 2 variables :

- **"Ash"** (dont la contribution semble plus importante que sur l'axe 2)
- **"Alacanity.of.ash"** (quasiment "colinéaire" à l'axe)

Ces variables sont positivement corrélés avec l'axe. Ce dernier contient toujours très peu d'informations sur les autres variables.

IV. Interprétation des résultats

Les variables étant centrées. Le signe et les valeurs des points typiques fournissent des éléments facilitant l'interprétation des axes.

```
# points typiques de la 1ère composante :
#pts_car(1,0.7,0.07,0.07)
wine_sc_dig2_1=round(wine.sc[c(pts_car(1,0.7,0.07,0.07)[,1]),-1],digits = 2)
tab1=cbind(wine.sc[c(pts_car(1,0.7,0.07,0.07)[,1]),1],wine_sc_dig2_1)
colnames(tab1)=wine.short
kable(tab1)
```

	Class	Al	Mal.ac	Ash	Ash.Alc	Mg	Tot ph	Fl	Ph.Nonfl.	Proant	Col.int	Hue	OD	Pro
21	1	1.30	-0.63	-0.32	-1.05	1.84	1.13	1.14	-0.98	0.89	0.26	0.58	1.55	0.1
23	1	0.87	-0.43	-0.02	-0.87	0.09	0.50	0.85	-0.74	0.17	-0.54	0.67	1.96	0.9

	Class	Al	Mal.ac	Ash	Ash.Alc	Mg	Tot ph	Fl	Ph.Nonfl.	Proant	Col.int	Hue	OD	Pro
36	1	0.59	-0.47	0.16	0.30	0.02	0.65	0.95	-0.82	0.47	0.02	0.36	1.21	0.5
137	3	-0.92	2.13	0.63	0.45	-0.75	-1.46	-1.56	1.35	-1.38	-0.52	-0.91	-1.89	-0.0
141	3	-0.09	0.42	1.22	0.45	-0.26	-1.21	-1.53	1.35	-1.47	-0.20	-0.82	-0.42	-0.4
161	3	-0.79	1.34	0.05	0.45	-0.82	0.01	-1.11	1.11	-0.96	1.12	-1.74	-1.45	-0.7
163	3	-0.19	0.84	0.78	0.75	0.44	-1.03	-1.43	1.91	-1.10	0.23	-0.38	-0.71	-0.5
166	3	0.90	1.81	-0.39	0.90	-0.82	-1.62	-1.56	1.27	-0.77	0.67	-0.78	-1.21	-0.7
171	3	-0.99	0.62	-0.17	-0.15	-0.26	-1.67	-1.54	0.31	-1.50	0.19	-1.30	-1.10	-0.7

```
# points typiques de la 2nde composante :
#pts_car(2,0.07,0.7,0.07)
wine_sc_dig2_2=round(wine.sc[c(pts_car(2,0.07,0.7,0.07)[,1]),-1],digits = 2)
tab2=cbind(wine.sc[c(pts_car(2,0.07,0.7,0.07)[,1]),1],wine_sc_dig2_2)
colnames(tab2)=wine.short
kable(tab2)
```

	Class	Al	Mal.ac	Ash	Ash.Alc	Mg	Tot ph	Fl	Ph.Nonfl.	Proant	Col.int	Hue	OD	Pro
81	2	-1.23	-1.27	-1.34	-0.15	-0.96	0.20	0.23	-0.50	-0.28	-1.10	1.85	0.72	-1.4
104	2	-1.45	-0.55	-1.77	0.00	-0.96	0.33	-0.39	0.07	-0.30	-1.29	-0.08	-0.24	-1.0
107	2	-0.92	-0.54	-0.90	-0.15	-1.38	-1.03	0.00	0.07	0.07	-0.72	0.19	0.79	-0.7
117	2	-1.45	-0.78	-1.37	0.39	-0.96	-0.50	-0.43	-0.50	-0.11	-1.34	-0.03	1.01	-0.8

```
# points typiques de la 3ème composante :
#pts_car(3,0.07,0.07,0.7)
wine_sc_dig2_3=round(wine.sc[c(pts_car(3,0.07,0.07,0.7)[,1]),-1],digits = 2)
tab3=cbind(wine.sc[c(pts_car(3,0.07,0.07,0.7)[,1]),1],wine_sc_dig2_3)
colnames(tab3)=wine.short
kable(tab3)
```

	Class	Al	Mal.ac	Ash	Ash.Alc	Mg	Tot ph	Fl	Ph.Nonfl.	Proant	Col.int	Hue	OD	Prol.
26	1	0.06	-0.26	3.11	1.65	1.70	0.54	0.65	0.87	0.57	-0.64	0.75	0.83	0.26
122	2	-1.77	-0.26	3.15	2.70	1.35	1.41	3.05	0.87	0.49	0.41	-0.12	1.52	-0.90

bien qu'identifié comme "bon représentant" de l'axe 3, l'élément 122 n'est pas représentatif de la classe 2, car au vu du graohiques des individus, il s'avère atypique pour cette classe.

Les observations confirment nos précédentes hypothèses : "la classe d'un vin dépend de la composition chimique qui le compose".

Ainsi :

- l'axe 1 oppose les vins des classes 1 et 3
- l'axe 2 précise la nature des vins des classe 2
- l'axe 3 caractérise les vins extrêmement riches en "Ash" et "Ash Alcanity"

De manière plus précise, on a les caractérisations suivantes :

- les vins de **classe 1** sont **très riches** en **"Flavanoids"** et en **"OD280/OD315 of diluted wines"**. Ils sont **assez riches** en **"Total phenols"** (l'individu 161 semble atypique), **"Proanthocyanins",**

“Hue” et **“Proline”**.
Par ailleurs ils s’avèrent **assez pauvres** en **“Malic acid”** et **“Nonflavanoid. phenols”**.

- c’est exactement **le contraire** pour ceux de la **classe 3**.
- les vins de classe 2 sont **pauvres** en **“Ash” et “Magnesium”**. Ils ont en général une **“Color Intensity”** ainsi qu’un degré **“d’Alcohol”** plus faible que la moyenne.

Récapitulatif

élements chimiques	*classe* 1	*classe* 2	*classe* 3
Flavanoids	très riche		très pauvre
OD280/OD315	très riche		très pauvre
Total phenols	riche		très pauvre
Proanthocyanins	riche		très pauvre
Hue	riche		très pauvre
Proline	riche		très pauvre
Ash		pauvre	
Alcalinity of ash		pauvre	
Magnesium		pauvre	
Color Intensity		pauvre	
Alcohol		pauvre	