Partiel M2 MIAGE AD - 29 mars 2021

À chaque inversion de matrice, donnez suffisamment d'étapes pour prouver que vous avez effectué le calcul sans assistance.

1 Matrices (8 points)

On considère la matrice $A_n = \begin{pmatrix} 1 & 1 & ... & 1 \\ 0 & 1 & ... & 1 \\ 0 & 0 & ... & 1 \end{pmatrix}$ dont tous les termes sous-diagonaux sont nuls, tous les autres valant 1.

Par exemple, $A_3 = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$ et $A_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

1. Justifiez l'inversibilité de A_n (pour tout n). (1 point)
2. Calculer A_2^{-1} puis A_3^{-1}. (2 points)
3. En déduire une hypothèse sur la forme générale de A_n^{-1}, et vérifiez la. (2 points)
4. Déterminez l'ensemble des vecteurs propres non nuls de A_n. (3 points)

2 Chaînes de Markov (8 points)

Un joueur d'échecs occasionnel installe sur son téléphone un programme pour jouer dans les transports. Ce programme dispose de 4 niveaux (numérotés de 1 à 4). On suppose que si le niveau réel du joueur est $n \in [1,4]$, la probabilité qu'il gagne contre le niveau m vaut $\frac{n}{n+m}$. On s'intéresse au nombre de parties jouées en moyenne avant d'atteindre le niveau 4, sachant que le joueur passe au niveau supérieur quand il gagne et revient au précédent quand il perd (pas de parties nulles).

Rappel : ce temps moyen s'écrit $t = Nc$ où $N = (I - Q)^{-1}$ avec Q la matrice des états transients, et c vecteur colonne ne contenant que des 1.

1. Modélisez la situation par une chaîne de Markov, en supposant $n = 2$. (1 point)
2. La chaîne est-elle irréductible (ergodique) ? Apériodique (régulière) ? Justifiez. (1.5 point)
3. Calculez la distribution stationnaire w telle que $wP = w$.
   Vérifiez la cohérence du résultat. (1.5 point)
4. Supposant qu'il commence au niveau 1, calculez le nombre moyen de parties jouées avant d'arriver au niveau 4. (3 points)
5. Proposez une amélioration permettant de modéliser aussi l'évolution du niveau du joueur. Est-ce toujours une chaîne de Markov homogène telle qu'étudiée en cours ? (1 point)

Rappel : une chaîne est homogène si les probabilités de transition ne dépendent pas du temps.

3 Arbres de décision (8 points)

Un joggeur décide chaque midi s'il va courir en fonction de quatre critères : la météo, la température, la vitesse du vent et son retard au boulot.

Construire un arbre de décision **binaire** selon la méthode vue en cours en utilisant l'indice de Gini : $I_G = 1 - \sum_{k=1}^{K} p_k^2$, avec p_k proportion de la classe k dans les données du noeud courant. Détaillez les étapes *[calculatrice]*, et dessinez l'arbre. (7 points)

Température	Vitesse (du vent)	Météo	Retard (au boulot)	Courir ?
15.2	10	soleil	non	oui
12.0	15	nuages	non	oui
10.0	10	pluie	oui	non
5.5	35	nuages	non	non
30.0	15	soleil	oui	non
7.0	25	pluie	non	non
18.0	12	pluie	non	oui
18.0	13	nuages	oui	oui
20.5	30	soleil	non	oui
20.0	20	pluie	oui	non

(#BonSensRequired) Vous paraît-il judicieux d'élaguer cet arbre ? Pourquoi ? (1 point).

4 Analyse en Composantes Principales (8 points)

Dans cet exercice on utilise un jeu de données indiquant la qualité de divers vins (blancs) en fonction de leurs caractéristiques physico-chimiques. La qualité (dernière colonne) est une note de 0 à 10. Autres variables :

- fixed acidity : acide tartrique
- volatile acidity : acide acétique ("vinaigre")
- citric acid : acide citrique
- residual sugar : sucres résiduels
- chlorides : chlorures (sel)
- free sulfur dioxide : dioxyde de soufre (non lié à d'autres molécules)
- total sulfur dioxide : dioxyde de soufre (toutes formes)
- density : densité
- pH : pH
- sulphates : sulfate de potassium
- alcohol : degré d'alcool

Les variables 6, 7 et 10 correspondent à des additifs (conservateurs), a priori sans bonnes propriétés gustatives.

[10]:
```
library(FactoMineR); library(factoextra); library(gridExtra); library(corrplot)
options(repr.plot.width=15, repr.plot.height=9)
```

[2]:
```
data <- read.csv("~/winequality-white.csv", sep=";"); head(data); dim(data)
```

A data.frame: 6 × 12

	fixed.acidity <dbl>	volatile.acidity <dbl>	citric.acid <dbl>	residual.sugar <dbl>	chlorides <dbl>	free.sulfur.dio <dbl>
1	7.0	0.27	0.36	20.7	0.045	45
2	6.3	0.30	0.34	1.6	0.049	14
3	8.1	0.28	0.40	6.9	0.050	30
4	7.2	0.23	0.32	8.5	0.058	47
5	7.2	0.23	0.32	8.5	0.058	47
6	8.1	0.28	0.40	6.9	0.050	30

1. 4898 2. 12

1] Décrivez ce que fait la ligne de code ci-dessous. (1 points)

[3]:
```
res.pca <- PCA(data, quanti.sup=12, ncp=6, graph=FALSE)
```

2] Qu'est-ce qui est affiché par le code ci-dessous ?
Combien d'axes doit-on garder (au minimum) pour conserver 80% d'inertie ? (1.5 points)

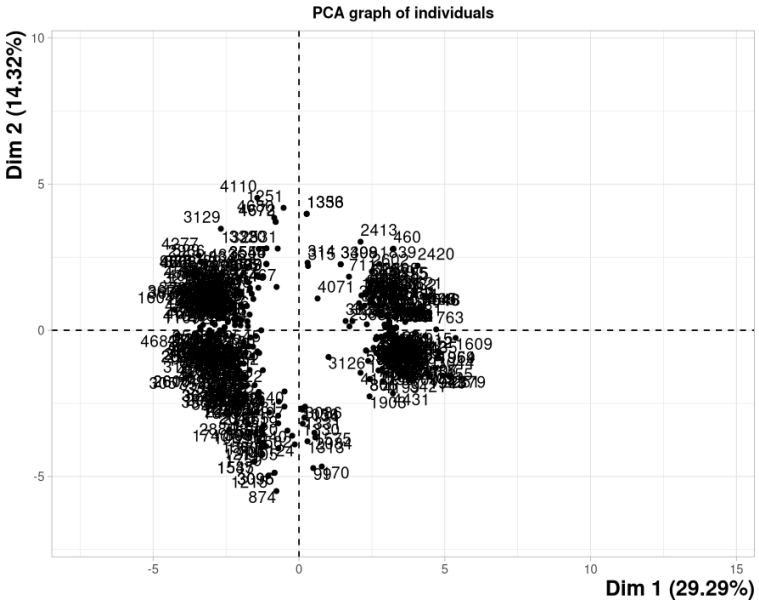
[4]:
```
res.pca$eig
```

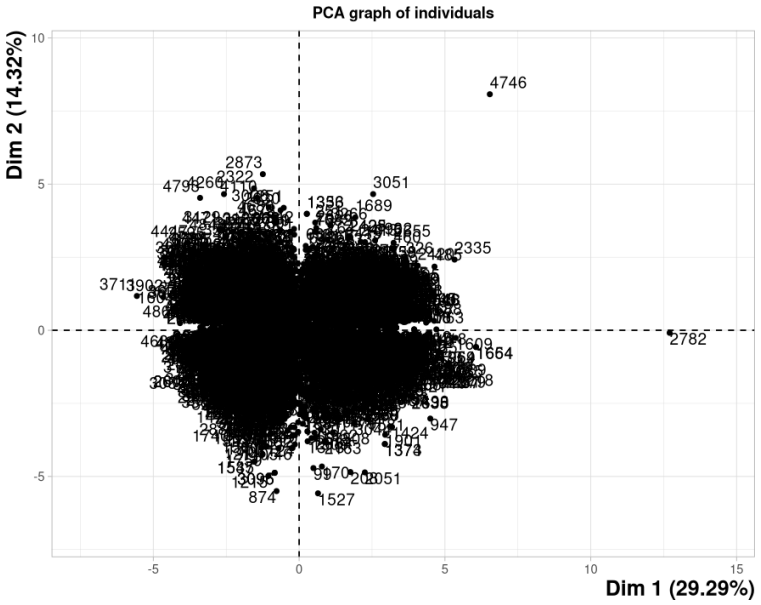
A matrix: 11 × 3 of type dbl

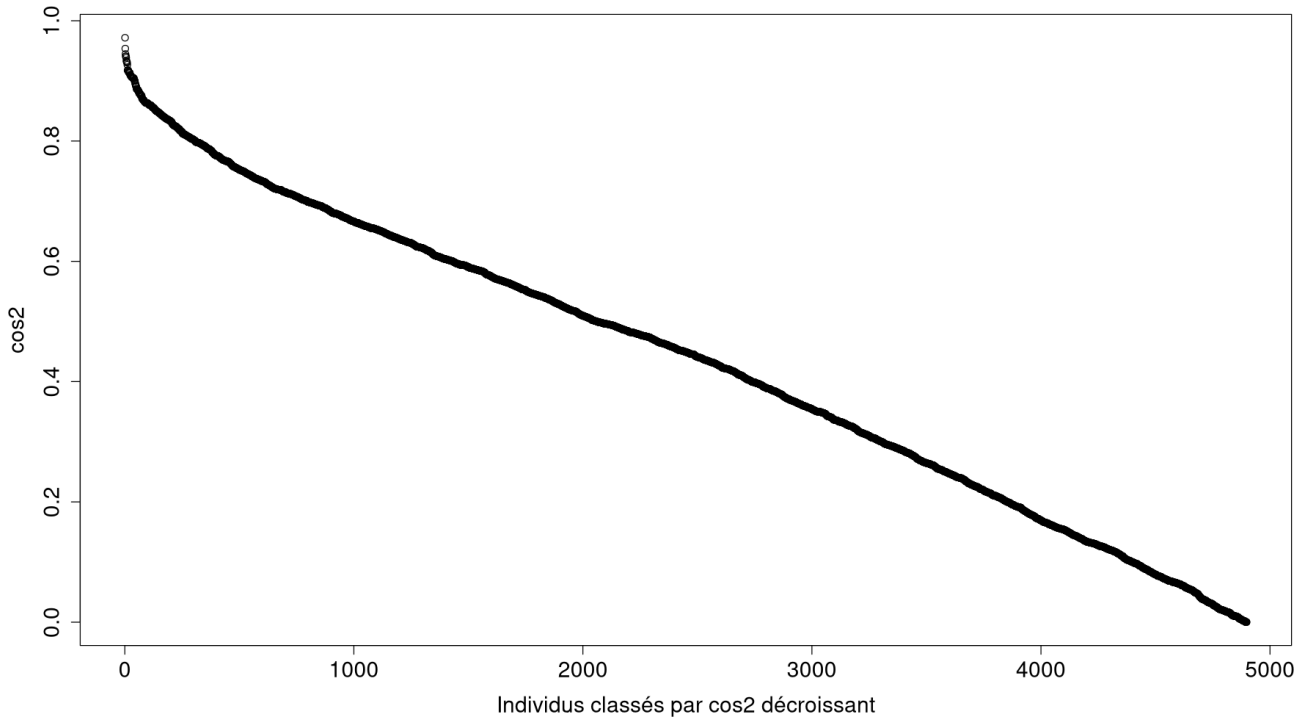
	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	3.22225389	29.293217	29.29322
comp 2	1.57523993	14.320363	43.61358
comp 3	1.22167134	11.106103	54.71968
comp 4	1.01852235	9.259294	63.97898
comp 5	0.97333458	8.848496	72.82747
comp 6	0.93874151	8.534014	81.36149
comp 7	0.72659802	6.605437	87.96692
comp 8	0.59935848	5.448713	93.41564
comp 9	0.41414367	3.764942	97.18058
comp 10	0.28948714	2.631701	99.81228
comp 11	0.02064909	0.187719	100.00000

3] Commentez le résultat du code ci-dessous. (2 points)

[5]:
```
grid.arrange(
  plot(res.pca, choix="ind", invisible="quali", select="cos2 500", unselect=1,
       cex.axis=1.5),
  plot(res.pca, choix="ind", invisible="quali", cex.axis=1.5), ncol=2)
s <- sort(rowSums(res.pca$ind$cos2[,1:2]), decreasing=TRUE)
plot(s, cex.axis=1.5, cex.lab=1.5, xlab="Individus classés par cos2␣
 ↪décroissant", ylab="cos2")
```

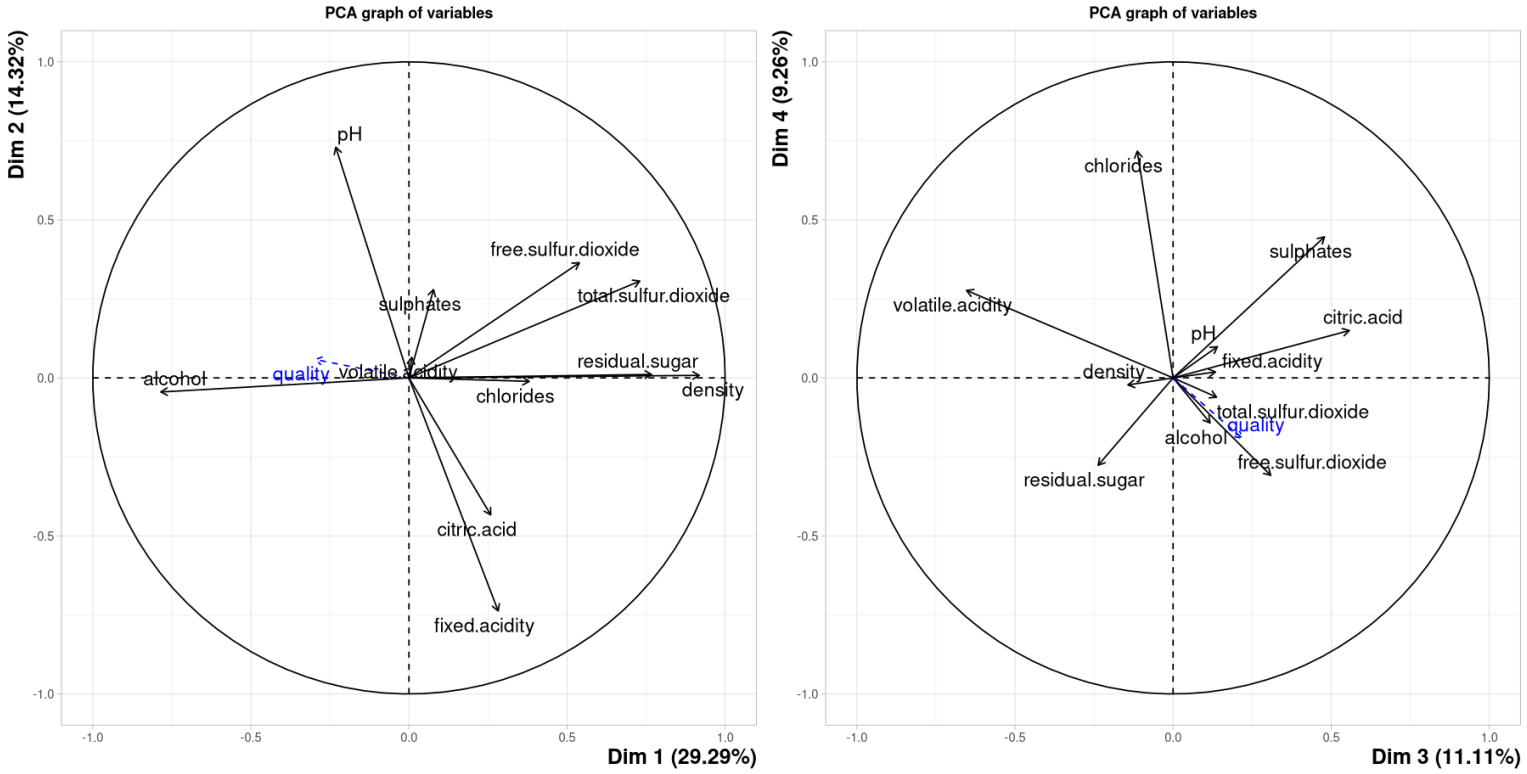

PCA graph of individuals
Dim 2 (14.32%)
Dim 1 (29.29%)

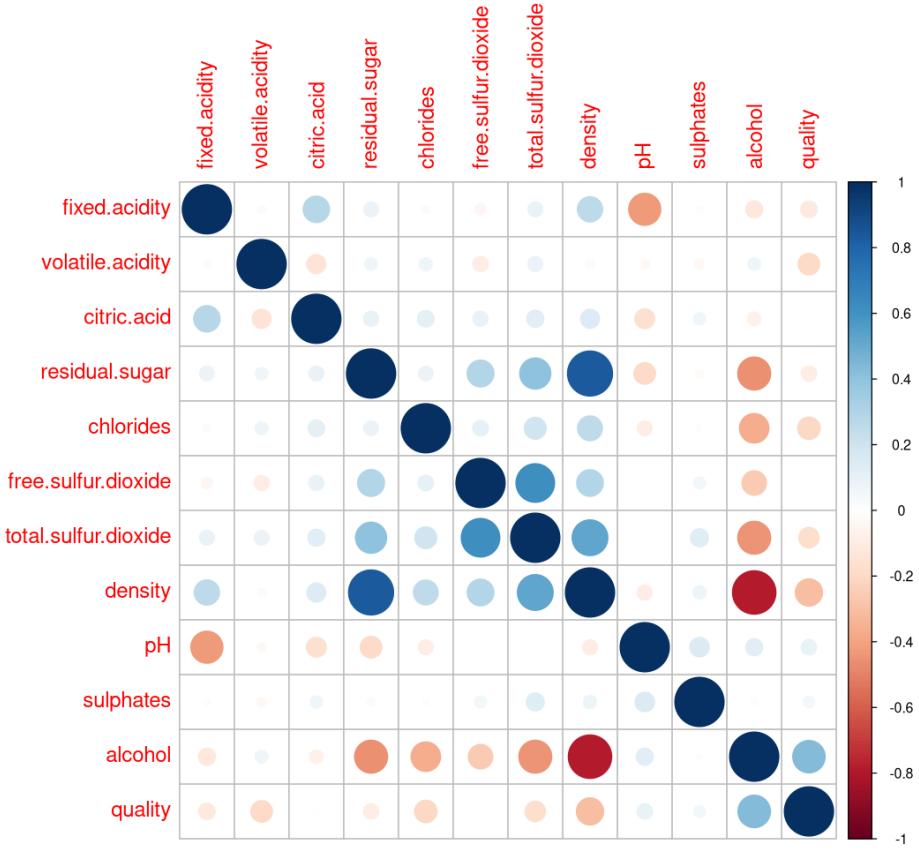

PCA graph of individuals
Dim 2 (14.32%)
Dim 1 (29.29%)


cos2
Individus classés par cos2 décroissant

4] Commentez les graphes ci-dessous. (3.5 points)

[6]:
```
grid.arrange(
  plot(res.pca, choix="var", axes=c(1,2), cex.axis=1.5, cex=1.2),
  plot(res.pca, choix="var", axes=c(3,4), cex.axis=1.5, cex=1.2), ncol=2)
corrplot(cor(data), tl.cex=1.2)
```





5 Analyse Factorielle des Correspondances (8 points)

On reprend le jeu de données de l'exercice précédent, en se focalisant sur les variables "quality" et "free.sulfur.dioxide".

- Rappel : la distance du χ^2 entre une répartition d'entiers $n_1, \ldots, n_k$ et l'effectif théorique $t_1, \ldots, t_k$ s'écrit $d_{\chi^2} = \sum_{i=1}^{k} \frac{(n_i - t_i)^2}{t_i}$.
- Indication : à forte dose, le dioxyde de soufre libre masque le goût du vin.

1. On donne le tableau des effectifs théoriques en cas d'indépendance.
   Expliquez comment celui-ci a été obtenu. (2 points)

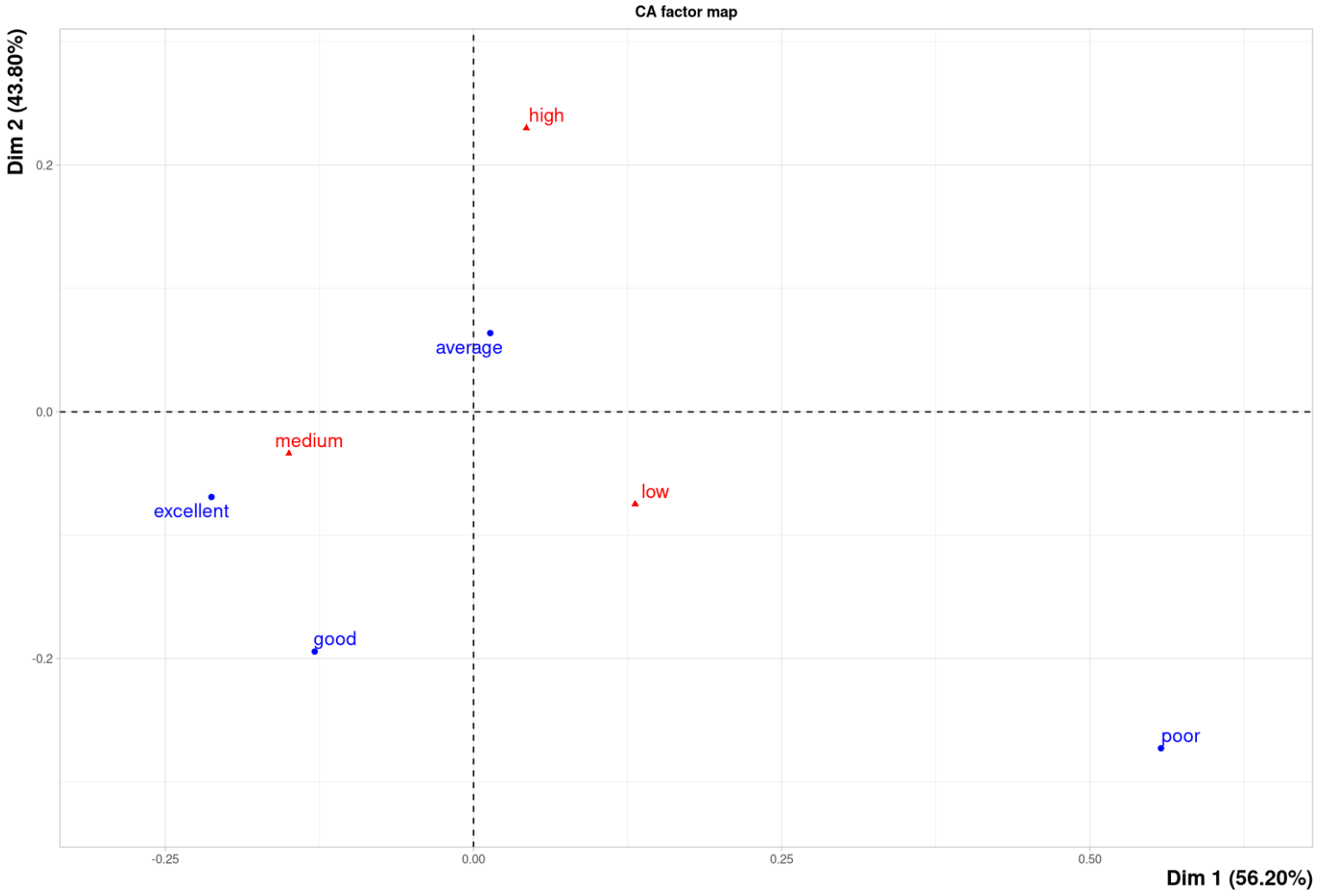
quality \ free.sulfur.dioxide	high	low	medium
average	695.5	1471.5	1488
excellent	34.2	72.5	73.3
good	167.5	354.3	358.2
poor	34.8	73.7	74.5

2. Calculer la distance du χ^2 avec le tableau réel ci-dessous. *[calculatrice]*.
   Combien y a-t-il de degrés de liberté, et pourquoi ? (2 points)
3. En déduire l'inertie totale du nuage des profils ligne ou colonne.
   Vérifiez à l'aide du dernier affichage page suivante. (1 point)
4. Commentez les résultats de l'AFC affichés ici. (3 points)

[7]:
```
q <- data[["quality"]]
q[q == 3 | q == 4] = "poor" ; q[q == 5 | q == 6 ] = "average"
q[q == 7] = "good" ; q[q == 8 | q == 9] = "excellent"
f <- data[["free.sulfur.dioxide"]]
ind1 <- f < 30 ; ind2 <- (f >= 30 & f < 50) ; ind3 <- f >= 50
f[ind1] = "low" ; f[ind2] = "medium" ; f[ind3] = "high"
df <- data.frame("quality"=q, "free.sulfur.dioxide"=f)
t <- table(df); t
```

```
           free.sulfur.dioxide
quality     high  low medium
  average    789 1430   1436
  excellent   27   60     93
  good        94  353    433
  poor        22  129     32
```

[8]:
```
res.CA <- CA(t, graph=FALSE); plot(res.CA, cex.axis=1.5, cex=1.2)
```



[9]:
```
res.CA$eig
```

A matrix: 2 × 3 of type dbl

	eigenvalue	percentage of variance	cumulative percentage of variance
dim 1	0.01639384	56.2013	56.2013
dim 2	0.01277602	43.7987	100.0000