

# PROJECTE ANÀLISI DE DADES: Wine Quality

David Anglada Rotger i Andreu Huguet Segarra

17/5/2019

## Contents

<b>1</b>	<b>Introducció</b>	<b>1</b>
<b>2</b>	<b>Descripció de la Base de Dades</b>	<b>1</b>
<b>3</b>	<b>Anàlisi de Components Principals (PCA)</b>	<b>3</b>
3.1	Breu síntesi . . . . .	3
3.2	PCA de les nostres dades. . . . .	3
<b>4</b>	<b>Inferència Multivariant</b>	<b>7</b>
4.1	Breu Síntesi . . . . .	7
4.2	Inferència multivariant en les nostres dades. . . . .	8
<b>5</b>	<b>Anàlisi Discriminant (LDA/QDA)</b>	<b>12</b>
5.1	Breu síntesi . . . . .	12
<b>6</b>	<b>Anàlisi clúster</b>	<b>15</b>
6.1	Breu síntesi . . . . .	15
6.2	Anàlisi Clúster de les nostres dades . . . . .	16
<b>7</b>	<b>Conclusions</b>	<b>22</b>
<b>8</b>	<b>Referències</b>	<b>22</b>

## 1 Introducció

Avui en dia, el rang de consumidors del vi s'ha ampliat moltíssims, convertint-se amb una beguda que es pot trobar a totes les cases i, fins i tot, comparable amb la cervesa. Això ha provocat que l'interés per aquest producte hagi augmentat els darrers anys, provocant també un creixement de la indústria de la vineria. Com a conseqüència, el nombre d'investigacions i estudis que tenen com a finalitat la millora de la qualitat del vi o la pujada de les seves vendes s'ha disparat.

Un dels temes que preocupa més a aquest sector és la **certificació de qualitat**, un aspecte que depèn profundament de la catació i valoració d'enòlegs experts. Així doncs, la finalitat d'aquest estudi és estudiar quines variables afecten més a la qualitat final del vi i de quina manera influeixen.

Donat que algunes d'aquestes variables es poden controlar durant el procés de producció, serà interessant veure les que influeixen positivament amb l'objectiu de potenciar-les, o detectar les que influeixen negativament per poder intentar neutralitzar el seu efecte.

## 2 Descripció de la Base de Dades

Per la realització d'aquest estudi es disposa d'una base de dades amb 1599 entrades diferents. Cada una es correspon amb la certificació de qualitat d'un vi en particular (en particular, la base de dades la formen mostres de *vinho verde*, un dels vins més importants de tot Portugal), acompanyat de 11 variables fisicoquímiques

més. Aquestes variables són el resultat de tests objectius, menys la variable resposta, la qualitat, que és la mitjana de la valoració de 3 enòlegs experts. Cada expert va otorgar una puntuació entre el 0 (dolent) al 10 (excel.lent).

Les 12 variables del *dataset* són les següents:

- **fixed.acidity**: Acidesa fixada.
- **volatile.acidity**: Acidesa volàtil.
- **citric.acid**: Àcid cítric.
- **residual.sugar**: Sucre residual.
- **chlorides**: Clorurs.
- **free.sulfur.dioxide**: Diòxid de sofre lliure.
- **total.sulfur.dioxide**: Totat de diòxid de sofre.
- **density**: Densitat.
- **pH**: pH.
- **sulphates**: Sulfats.
- **alcohol**: Alcohol.
- **quality**: Qualitat. *Variable resposta*.

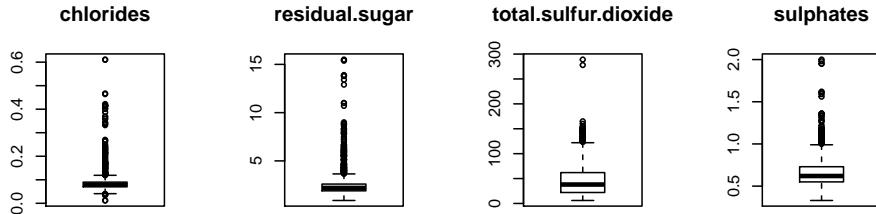
Pel que fa als *missing values*, no n'hi ha cap en aquesta base de dades, tal i com s'indica a la seva descripció oficial.

Abans de res, s'observen les característiques de cada una de les variables i els resultats són els següents.

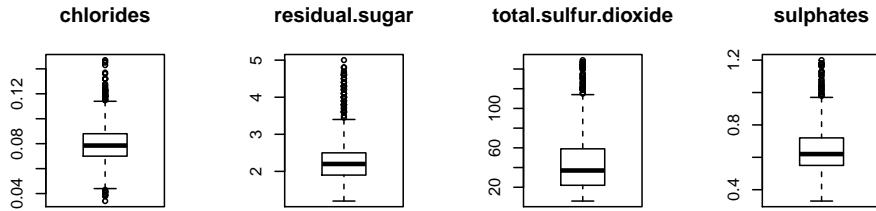
```
##   fixed.acidity  volatile.acidity  citric.acid  residual.sugar
##   Min. : 4.60    Min. :0.1200    Min. :0.000    Min. : 0.900
##   1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900
##   Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200
##   Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539
##   3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600
##   Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500
##   chlorides      free.sulfur.dioxide total.sulfur.dioxide
##   Min. :0.01200   Min. : 1.00     Min. : 6.00
##   1st Qu.:0.07000  1st Qu.: 7.00     1st Qu.:22.00
##   Median :0.07900  Median :14.00     Median :38.00
##   Mean   :0.08747  Mean   :15.87     Mean   :46.47
##   3rd Qu.:0.09000  3rd Qu.:21.00     3rd Qu.:62.00
##   Max.   :0.61100  Max.   :72.00     Max.   :289.00
##   density         pH            sulphates        alcohol
##   Min. :0.9901    Min. :2.740    Min. :0.3300    Min. : 8.40
##   1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500   1st Qu.: 9.50
##   Median :0.9968   Median :3.310   Median :0.6200   Median :10.20
##   Mean   :0.9967   Mean   :3.311   Mean   :0.6581   Mean   :10.42
##   3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300   3rd Qu.:11.10
##   Max.   :1.0037   Max.   :4.010   Max.   :2.0000   Max.   :14.90
```

En general, tots els valors semblen normals. L'únic que crida l'atenció és que les variables **residual.sugar**, **chlorides**, **sulphates** i **total.sulfur.dioxide** prenen valors molt concentrats a l'extrem esquerre, ja que el tercer quartil i la mediana estan molt allunyats del màxim. Això pot voler dir que hi ha algun outlier que convé eliminar perquè no afecti al nostre anàlisi. Per detectar-ho, es realitza un *boxplot* d'aquestes variables.

```
## Total outliers =
## [1] 163
## % Outliers =
## [1] 10.19387
```



S'observa que, com era d'esperar, hi ha nombrosos *outliers* en aquestes variables. Donat que el total d'outliers suposa el 10% de les dades, es procedeix a eliminar-los per evitar possibles influències negatives en l'anàlisi. S'observa ara que els valors de les variables estan més repartits.



### 3 Anàlisi de Components Principals (PCA)

#### 3.1 Breu síntesi

El primer anàlisi que de les dades que es farà és l'**Anàlisi de Components (PCA)**. Els objectius fonamentals d'aquest estudi són reduir el nombre de variables i realitzar una representació en dues dimensions de les dades (**biplot**). El que es busca són combinacions lineals de les variables de la forma  $F_i = a_{i1}X_1 + \dots + a_{iN}X_p$  tal que  $F_1, \dots, F_N$  (components principals) no estiguin correlats. Aquestes combinacions lineals ens dona la descomposició espectral de la matriu de les dades  $F = XA$ , on A és la matriu de vectors principals. Una vegada calculats aquests components, ens quedarem amb els suficients per explicar un 80% de la variància de les dades o, en el nostre cas, amb els que tinguin un **valor propi major que 1**.

Una vegada feta aquesta descomposició, es pot fer el **biplot** de les dades a partir de  $X = FA$ . En aquesta representació, val a dir que les distàncies euclidianes entre els punts aproximen les distàncies Mahalanobis entre les observacions reals. A més, la llargada de les fletxes que representen cada variable és una estimació de la seva desviació estàndard. Ara bé, un dels aspectes més interessants és que l'**angle** entre les diferents fletxes representa una estimació de la correlació entre les dues variables, és a dir:

- Si dues fletxes són gairebé paral·leles i amb el mateix sentit, les variables estaran correlades positivament.
- Si dues fletxes són gairebé paral·leles i amb sentits opositius, les variables estaran correlades negativament.
- Si dues fletxes són gairebé ortogonals, les variables no estaran correlades.

Un altre aspecte a tenir en compte, és si s'utilitza la matriu de covariàncies o la de correlacions pels càlculs. Donat que aquestes dades tenen variables amb unitats de mesura diferents, s'utilitzarà la **matriu de correlacions** (motiu pel qual ens quedarem amb els components principals amb valors propis majors que 1), que és invariant respecte les unitats de mesura.

En resum, els objectius d'aquest anàlisi seràn la representació **biplot** de les dades, la **correlació** entre les variables de la base de dades i els diferents **coeficients** dels components principals més representatius, ja que serà un indicador de les variables més importants.

#### 3.2 PCA de les nostres dades.

Abans de res, es construeix un nou *dataset* eliminant la variable resposta, és a dir, la variable **quality**. Tot seguit, es calcula la matriu de correlacions de les dades, així com el seu *scatter plot* per observar aquestes

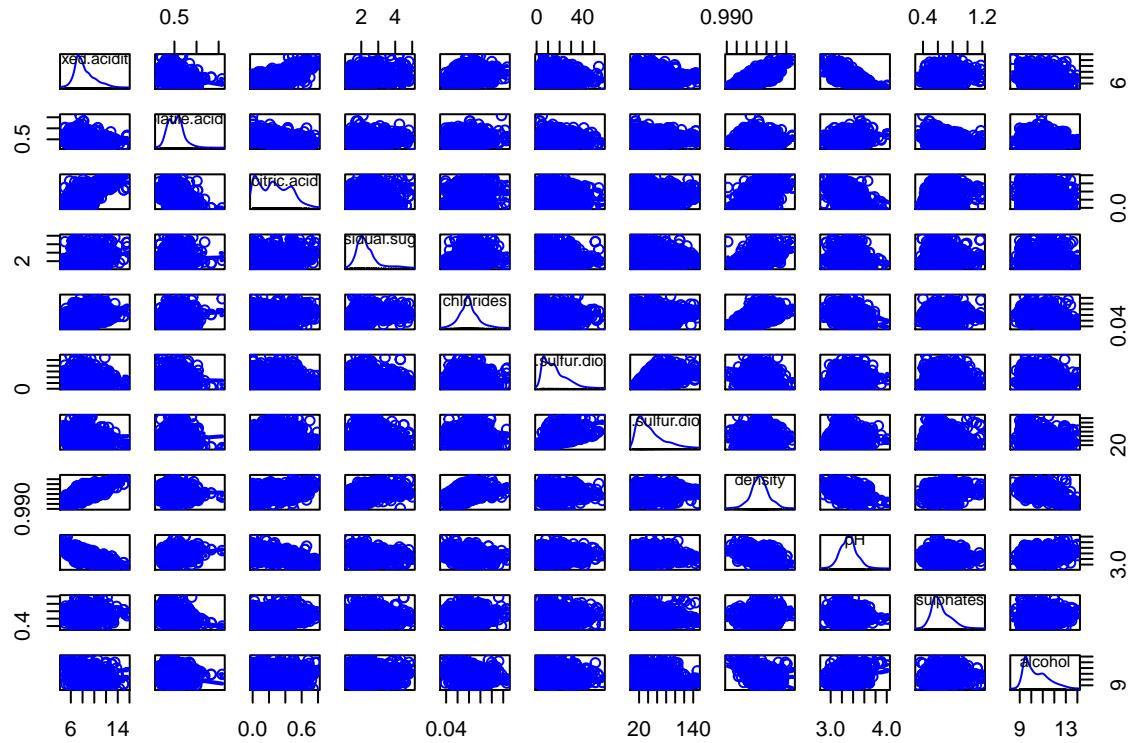
correlaciones.

```
## fixed.acidity volatile.acidity citric.acid
## fixed.acidity 1.0000000 -0.266994660 0.6970263272
## volatile.acidity -0.2669947 1.000000000 -0.5743546402
## citric.acid 0.6970263 -0.574354640 1.000000000
## residual.sugar 0.2562919 0.043882659 0.1830280846
## chlorides 0.2599116 0.103766699 0.1316719088
## free.sulfur.dioxide -0.1629395 -0.021098530 -0.0825786899
## total.sulfur.dioxide -0.1155196 0.084253502 0.0001528991
## density 0.6973400 -0.004704452 0.3891438503
## pH -0.7144390 0.236975992 -0.5267910623
## sulphates 0.2087144 -0.336055195 0.2913133654
## alcohol -0.0766056 -0.183823778 0.1181286119
## residual.sugar chlorides free.sulfur.dioxide
## fixed.acidity 0.25629194 0.25991163 -0.16293946
## volatile.acidity 0.04388266 0.10376670 -0.02109853
## citric.acid 0.18302808 0.13167191 -0.08257869
## residual.sugar 1.00000000 0.25629623 0.02164089
## chlorides 0.25629623 1.00000000 -0.00275134
## free.sulfur.dioxide 0.02164089 -0.00275134 1.00000000
## total.sulfur.dioxide 0.09024186 0.10708125 0.64615172
## density 0.37713107 0.41602098 -0.06504795
## pH -0.12650274 -0.24239591 0.11766371
## sulphates 0.04443907 -0.03004194 0.06341016
## alcohol 0.11790589 -0.24689180 -0.05470825
## total.sulfur.dioxide density pH
## fixed.acidity -0.1155195673 0.697340022 -0.714439012
## volatile.acidity 0.0842535020 -0.004704452 0.236975992
## citric.acid 0.0001528991 0.389143850 -0.526791062
## residual.sugar 0.0902418632 0.377131068 -0.126502743
## chlorides 0.1070812526 0.416020977 -0.242395906
## free.sulfur.dioxide 0.6461517202 -0.065047945 0.117663710
## total.sulfur.dioxide 1.0000000000 0.076698681 -0.004352554
## density 0.0766986812 1.000000000 -0.376163096
## pH -0.0043525544 -0.376163096 1.000000000
## sulphates -0.0379376697 0.131001431 -0.041190430
## alcohol -0.2332413483 -0.505537180 0.207558302
## sulphates alcohol
## fixed.acidity 0.20871441 -0.07660560
## volatile.acidity -0.33605519 -0.18382378
## citric.acid 0.29131337 0.11812861
## residual.sugar 0.04443907 0.11790589
## chlorides -0.03004194 -0.24689180
## free.sulfur.dioxide 0.06341016 -0.05470825
## total.sulfur.dioxide -0.03793767 -0.23324135
## density 0.13100143 -0.50553718
## pH -0.04119043 0.20755830
## sulphates 1.00000000 0.23540400
## alcohol 0.23540400 1.00000000
## [1] 24
## Loading required package: carData
##
```

```

## Attaching package: 'ade4'
## The following object is masked from 'package:FactoMineR':
##   reconst

```



S'observa, tant en la matriu de correlacions com en el plot de les variables que, en general, **no hi ha correlació entre elles**. De fet, la majoria de valors de la matriu de correlació són inferiors a 0.3. Només hi ha 4 casos que cal destacar:

- `fixed.acidity` i `density` estan positivament correlades: 0.697.
- `fixed.acidity` i `pH` estan negativament correlades: 0.714.
- `fixed.acidity` i `citric.acid` estan positivament correlades: 0.697.
- `free.sulfur.dioxide` i `free.sulfur.dioxide` estan positivament correlades: 0.646.

Ara bé, en cap d'aquests casos la correlació entre les variables en qüestió és major que 0.75. És a dir, que no són correlacions molt clares. Calculem ara els components principals que s'han explicat a l'apartat anterior.

```

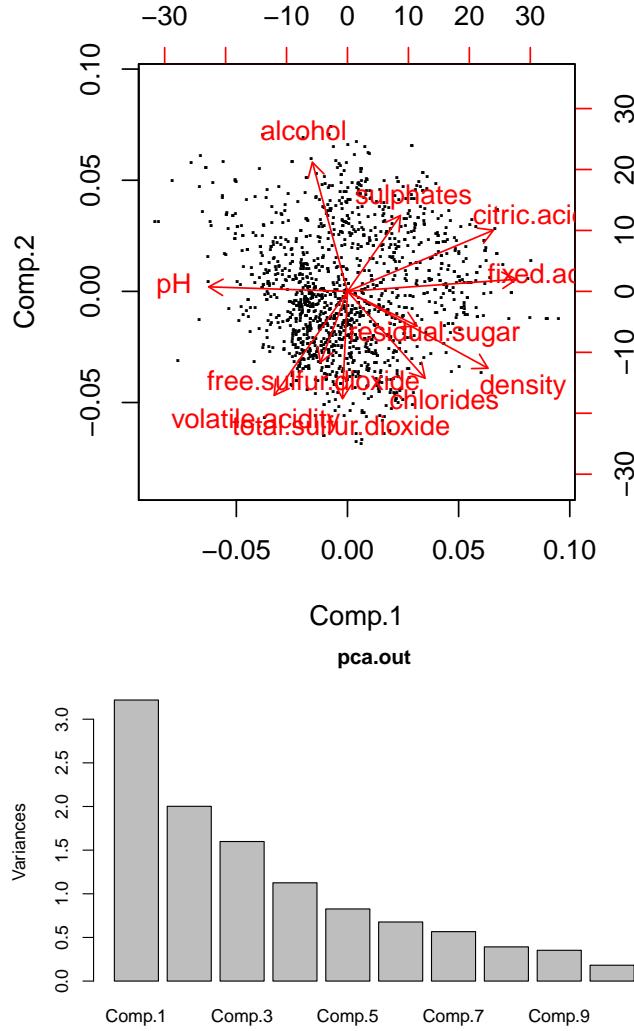
## Importance of components:
##                               Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## Standard deviation      1.7943537 1.4149643 1.2644268 1.0610261 0.90896784
## Proportion of Variance 0.2927005 0.1820113 0.1453432 0.1023433 0.07511114
## Cumulative Proportion  0.2927005 0.4747118 0.6200550 0.7223983 0.79750940
##                               Comp.6    Comp.7    Comp.8    Comp.9
## Standard deviation      0.82277278 0.75246762 0.62604171 0.59388657
## Proportion of Variance 0.06154137 0.05147341 0.03562984 0.03206375
## Cumulative Proportion  0.85905076 0.91052417 0.94615401 0.97821776
##                               Comp.10   Comp.11
## Standard deviation      0.4263341 0.240507554
## Proportion of Variance 0.0165237 0.005258535
## Cumulative Proportion  0.9947415 1.000000000

```

S'observa que, per a explicar almenys un 80% de la variabilitat de les dades, es necessiten almenys **5**

**components principals**, tot i que el cinquè ja té un valor propi menor que 1. Això és degut a què, com s'ha vist a la matriu de correlacions, les variables són força independents, amb pocs casos de correlacions importants, cosa que implica que no podem reduir molt el nombre de variables. Una altra aspecte a destacar és que **no hi ha cap valor propi nul**, fet que ens indica que no hi ha cap variable que sigui combinació lineal directe de les altres.

A continuació es presenta la representació *biplot* corresponent a aquests components principals.



El primer que cal dir d'aquesta representació és que només explica el 47.47% de la variabilitat de les dades, ja que s'hi veuen representats només els dos primers components principals. Així doncs, aquesta representació no és del tot fiable. Alguns fets que ho demostren són que, per exemple, les fletxes de les variables **residual.sugar** i **density** són pràcticament paral·leles i, en canvi, la correlació entre aquestes dues variables és de 0.377. Això podria ser degut a què els 2 primers components principals no representen gaire la variabilitat d'aquestes variables i es centren més en altres.

Tot i així, es veur representada molt representada la correlació negativa de la variable **fixed.acidity** i **pH**, fet que pot voler dir que tenen gran part de la seva variància explicada pels dos primers components. El mateix passa amb la correlació positiva de **fixed.acidity** i **citric.acid**.

A continuació, s'analitzen els coeficients dels components principals, per veure quines variables expliquen més.

```
##  
## Loadings:  
##
```

```
Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
```

```

## fixed.acidity      0.510           0.161  0.261
## volatile.acidity -0.221 -0.398  0.311 -0.240       0.198  0.633
## citric.acid       0.440  0.234 -0.183        0.177      -0.124
## residual.sugar    0.210 -0.132          -0.744  0.234  0.271 -0.272
## chlorides         0.234 -0.332  0.112 -0.242 -0.122 -0.863
## free.sulfur.dioxide -0.272 -0.640           0.105
## total.sulfur.dioxide -0.410 -0.549       0.146
## density           0.424 -0.294  0.110      -0.314  0.295 -0.177
## pH                -0.419           -0.272 -0.351      -0.451
## sulphates         0.160  0.290 -0.316 -0.168 -0.751       0.310
## alcohol            -0.106 0.494 -0.158 -0.461  0.285 -0.155  0.308
##                               Comp.8 Comp.9 Comp.10 Comp.11
## fixed.acidity       0.281  0.210  0.328   0.634
## volatile.acidity    0.244 -0.134 -0.345
## citric.acid         0.397 -0.307 -0.639
## residual.sugar     -0.379           -0.133  0.141
## chlorides          0.658 -0.226
## free.sulfur.dioxide -0.616  0.339
## density             0.286  0.104  0.255  -0.587
## pH                 0.556           0.327
## sulphates          -0.268 -0.132
## alcohol             0.294           0.328  -0.337
##
##                               Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8
## SS loadings          1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var       0.091  0.091  0.091  0.091  0.091  0.091  0.091  0.091
## Cumulative Var       0.091  0.182  0.273  0.364  0.455  0.545  0.636  0.727
##                               Comp.9 Comp.10 Comp.11
## SS loadings          1.000  1.000  1.000
## Proportion Var       0.091  0.091  0.091
## Cumulative Var       0.818  0.909  1.000

```

S'observa, com havíem intuit amb el *biplot*, que el primer component representa sobretot la variable **fixed.acidity**. També les variables **citric.acid**, **density** i **pH**. Això explica que les correlacions que havíem destacat abans es vegin prou representades en el biplot. Una fet que crida l'atenció és que la component que més representa la variable **residual.sugar** és la tercera. Això explica que la seva representació en el biplot no fos fiable. Un altre fet destacable és que les variables **free.sulfur.dioxide** i **total.sulfur.dioxide** no tenen representació en el primer component.

## 4 Inferència Multivariant

### 4.1 Breu Síntesi

En segon lloc, s'aplicaran alguns mètodes d'**Inferència Multivariant** en les dades. L'objectiu d'aquest estudi serà descobrir si hi ha diferències estadísticament significatives entre diversos grups de dades, agrupats per la qualitat dels vins. Per detectar aquestes possibles diferències es realitzarà el test de  $T^2$  de **Hotelling** (on assumirem igualtat de variàncies entre els grups) i un test de **t-student** (on no assumirem igualtat de variàncies entre els grups).

En ambdós casos, les hipòtesis dels test són les següents: siguin  $\mu_1$  i  $\mu_2$  els vectors de mitjanes de les variables de dos grups del conjunt de dades,

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Ara bé, l'estadístic de prova dels tests varia en funció de si s'assumeixen igualtat de variàncies o no en els grups. En cas de que sí, es té que

$$T^2 \text{ ha de seguir una } \frac{(n_1 + n_2 - 2)p}{n_1 + n_2 - p - 1} F_{p, n_1 + n_2 - p - 1}$$

on  $T^2 = [\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)]'((1/n_1 + 1/n_2)S_p)^{-1}[\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)]$ . En canvi, si no s'assumeix igualtat de variàncies, es té que

$$T^2 \text{ ha de seguir una } \chi^2$$

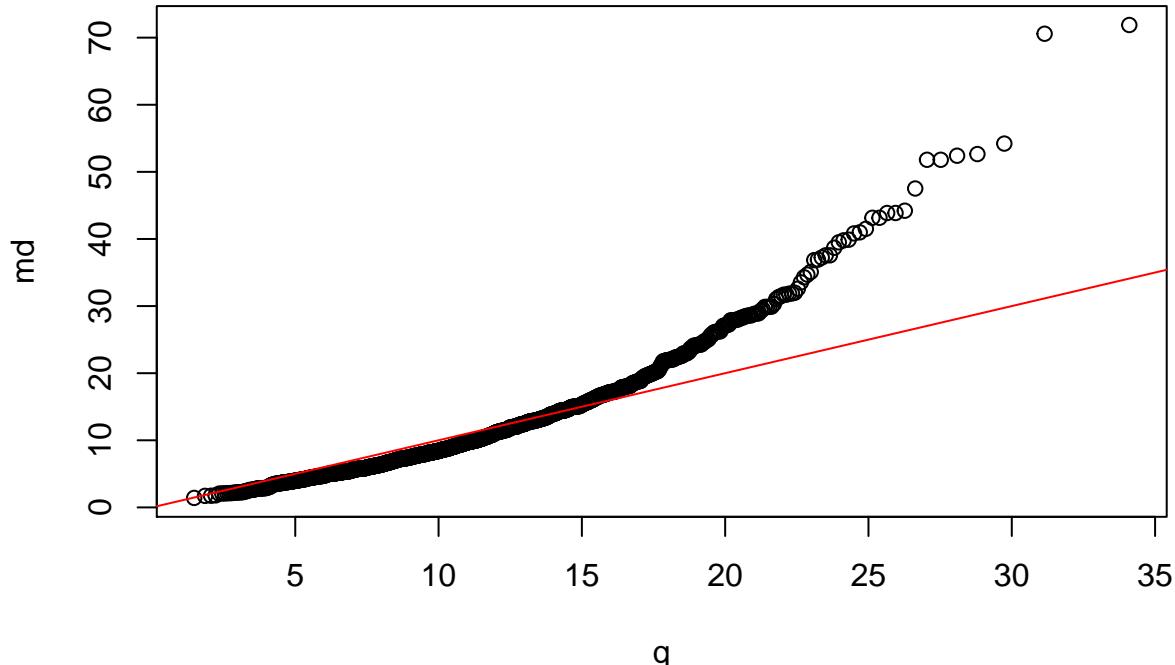
$$\text{on } T^2 = [\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)]' \left( \frac{1}{n_1} S_1 + \frac{1}{n_2} S_2 \right)^{-1} [\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)].$$

En ambdós casos, s'assumeix que les variables dels grups segueixen una distribució normal multivariant. Llavors, el primer que s'haurà de fer abans d'aplicar aquests tests a les nostres dades, serà mirar si podem assumir normalitat multivariant. Pel que fa a la divisió dels grups, es realitzaran dues divisions diferents: una de dos grups (qualitat inferior o igual a 5 i qualitat superior a 5) i una de tres grups (qualitat 3-4, qualitat 5-6, qualitat 7-8).

## 4.2 Inferència multivariant en les nostres dades.

### 4.2.1 Distribució normal multivariant.

Com s'ha dit anteriorment, abans de poder estudiar si hi ha diferències estadísticament significatives entre grups de dades, hem d'estar segurs que podem assumir normalitat en les variables. Per això es comprovarà que la relació entre les distàncies de les variables i els quartils corresponents a una  $\chi^2_p$  amb  $p$  graus de llibertat (número de variables, 11 en aquest cas) sigui lineal.



S'observa que la relació no és perfectament lineal però ho és suficient per assumir normalitat multivariant en les dades i, per tant, poder aplicar els tests de Hotelling i de *t*-student.

Abans de res, s'han dividit les dades en dos grups (qualitat major o menor que 5) i en tres grups (rangs de qualitat de 3-4, 5-6 i 7-8).

```
## [1] 55
## [1] 1189
## [1] 192
```

#### 4.2.2 Test Hotellin i *t*-student

En primer lloc, es realitza en test de Hotelling assumint igualtat de variàncies entre els dos grups de dades: un en què els vins tenen qualitat inferior o igual a 5 i un altre on tenen qualitat major que 5.

```
## Loading required package: mvtnorm
## Loading required package: ICS
##
## Hotelling's two sample T2-test
##
## data: db.menor5 and db.major5
## T.2 = 59.976, df1 = 11, df2 = 1424, p-value < 2.2e-16
## alternative hypothesis: true location difference is not equal to c(0,0,0,0,0,0,0,0,0,0,0,0)
```

Donat que s'obté un *p-value* < 0.05, s'ha de rebutjar la hipòtesi nula i, per tant, hi ha diferències estadísticament significatives entre les mitjanes dels vins de qualitat 5 o inferior amb els dels vins de qualitat 6 o major. Aquest resultat era d'esperar, ja que en cas contrari, voldria dir que la valoració dels jutges és 100% subjectiva, cosa que sabem que no és del tot certa. Observant les mitjanes, es comprova el resultat. S'observa que sobretot es difereix en les variables volatile.acidity, alcohol i total.sulfur.dioxide.

```
##      fixed.acidity    volatile.acidity    citric.acid
##          8.1339394        0.5871288       0.2242576
##      residual.sugar    chlorides    free.sulfur.dioxide
##          2.2890152        0.0821500      15.9196970
##      total.sulfur.dioxide density           pH
##          51.9863636        0.9969627      3.3202424
##      sulphates          alcohol
##          0.5923182        9.9371212
##
##      fixed.acidity    volatile.acidity    citric.acid
##          8.46597938       0.47485180     0.29257732
##      residual.sugar    chlorides    free.sulfur.dioxide
##          2.27957474       0.07713144     15.14432990
##      total.sulfur.dioxide density           pH
##          37.71134021      0.99638138      3.31976804
##      sulphates          alcohol
##          0.68436856       10.87420533
```

Es calcula a continuació la  $T^2$  sense assumir igualtat de variàncies. S'observa que el valor obtingut és força diferent a l'anterior i, per tant, es realitza un test *t*-student sense considerar les variàncies iguals.

```
##      [,1]
## [1,] 687.6322
## [1] "fixed.acidity"
## [1] 0.0002786998
```

```

## [1] "volatile.acidity"
## [1] 2.538612e-34
## [1] "citric.acid"
## [1] 6.329806e-12
## [1] "residual.sugar"
## [1] 0.7727862
## [1] "chlorides"
## [1] 4.646153e-09
## [1] "free.sulfur.dioxide"
## [1] 0.1322358
## [1] "total.sulfur.dioxide"
## [1] 4.870603e-19
## [1] "density"
## [1] 4.597627e-10
## [1] "pH"
## [1] 0.9527719
## [1] "sulphates"
## [1] 2.560078e-42
## [1] "alcohol"
## [1] 1.586652e-73

```

Analitzant els *p-values*, s'observa que només hi ha diferències significatives entre les variables **fixed.acidity**, **volatile.acidity**, **citric.acid**, **chlorides**, **total.sulfur.dioxide**, **density**, **sulphates** i **alcohol**. En la resta de variables, no hi ha diferències significatives. Així doncs, aquestes són les variables que principalment permetran dir si un vi és bo o és dolent.

A continuació es realitzarà el mateix anàlisi però comparant els 3 grups especificats anteriorment 2 a 2. Donat que s'ha vist que no es pot assumir igualtat de variables, es realitzarà directament el test *t-student*. En primer lloc, comparem els vins dolents (qualitat 3-4) amb els vins bons (qualitat 7-8).

```

## [1] "fixed.acidity"
## [1] 0.001429792
## [1] "volatile.acidity"
## [1] 1.339201e-13
## [1] "citric.acid"
## [1] 1.896413e-11
## [1] "residual.sugar"
## [1] 0.7525926
## [1] "chlorides"
## [1] 0.01613708
## [1] "free.sulfur.dioxide"
## [1] 0.2377745
## [1] "total.sulfur.dioxide"
## [1] 0.6667922
## [1] "density"
## [1] 0.002130184
## [1] "pH"
## [1] 5.571237e-05
## [1] "sulphates"
## [1] 3.283174e-12
## [1] "alcohol"
## [1] 1.372715e-15

```

Així doncs, entre aquests dos grups, s'observen diferències estadísticament significatives en les variables **fixed.acidity**, **volatile.acidity**, **citric.acid**, **chlorides**, **density**, **sulphates**, **alcohol** i **pH**. Es veu que ara hi ha diferències en el pH i no n'hi ha en els nivells totals de Diòxid de Sofre, a diferència d'abans.

Tot i així, sembla bastant objectiu que un vi sigui de baixa qualitat a que sigui d'alta qualitat. Es comparen ara els vins dolents (3-4) amb els vins mitjans (5-6).

```
## [1] "fixed.acidity"
## [1] 0.1515529
## [1] "volatile.acidity"
## [1] 4.451386e-07
## [1] "citric.acid"
## [1] 0.0003648193
## [1] "residual.sugar"
## [1] 0.3948942
## [1] "chlorides"
## [1] 0.5007721
## [1] "free.sulfur.dioxide"
## [1] 0.002802706
## [1] "total.sulfur.dioxide"
## [1] 0.0004207809
## [1] "density"
## [1] 0.7522975
## [1] "pH"
## [1] 0.0006541628
## [1] "sulphates"
## [1] 0.006650157
## [1] "alcohol"
## [1] 0.5517309
```

En aquest cas, s'observen diferències en només 6 variables (abans n'hi havia en 8): **volatile.acidity**, **citric.acid**, **free.sulfur.dioxide**, **total.sulfur.dioxide**, **pH** i **sulphates**. Així doncs, sembla que les diferències entre vins de mitjana qualitat i de baixa qualitat no estan tant clares (tot i que hi segueixen estant).

Per últim, es comparen els vins mitjans (5-6) amb els vins bons (7-8).

```
## [1] "fixed.acidity"
## [1] 0.0004874425
## [1] "volatile.acidity"
## [1] 6.868122e-25
## [1] "citric.acid"
## [1] 3.047428e-14
## [1] "residual.sugar"
## [1] 0.2428893
## [1] "chlorides"
## [1] 3.56122e-05
## [1] "free.sulfur.dioxide"
## [1] 0.004392452
## [1] "total.sulfur.dioxide"
## [1] 3.892771e-15
## [1] "density"
## [1] 1.97578e-08
## [1] "pH"
## [1] 0.04874981
## [1] "sulphates"
## [1] 3.701585e-28
## [1] "alcohol"
## [1] 4.796577e-41
```

En aquest cas, veiem diferències estadísticament significatives en totes les variables tret del `residual.sugar`. És a dir, són els grups més diferents i, per tant, quan un vi es bo, es nota.

## 5 Anàlisi Discriminant (LDA/QDA)

### 5.1 Breu síntesi

En tercer lloc, es farà un **Anàlisi Discriminant (LDA/QDA)**. Els objectius principals d'aquest estudi són L'anàlisi de discriminant lineal (LDA en anglès) consisteix en trobar una funció discriminant tal que decideixi a quina classe  $C_k$  pertany una observació  $x$ .

En LDA, assumim que les dades de les diferents  $k$  classes provenen d'una distribució gaussiana i la seva funció de densitat es pot escriure com:

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)$$

on, per cada classe  $k$ ,  $\mu_k$  és la mitjana de la classe  $k$  i  $\sigma$  és la variància comuna de totes les dades (un altre requisit que assumim quan fem LDA).

El mètode assigna cada observació  $x$  a la classe la qual  $P_k(x)$  sigui major. Per aquesta tasca de maximització podem considerar la tasca equivalent de trobar el màxim prenent el logaritme i aleshores ens queda la funció discriminant que volíem:

$$\delta_k(x) = x \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

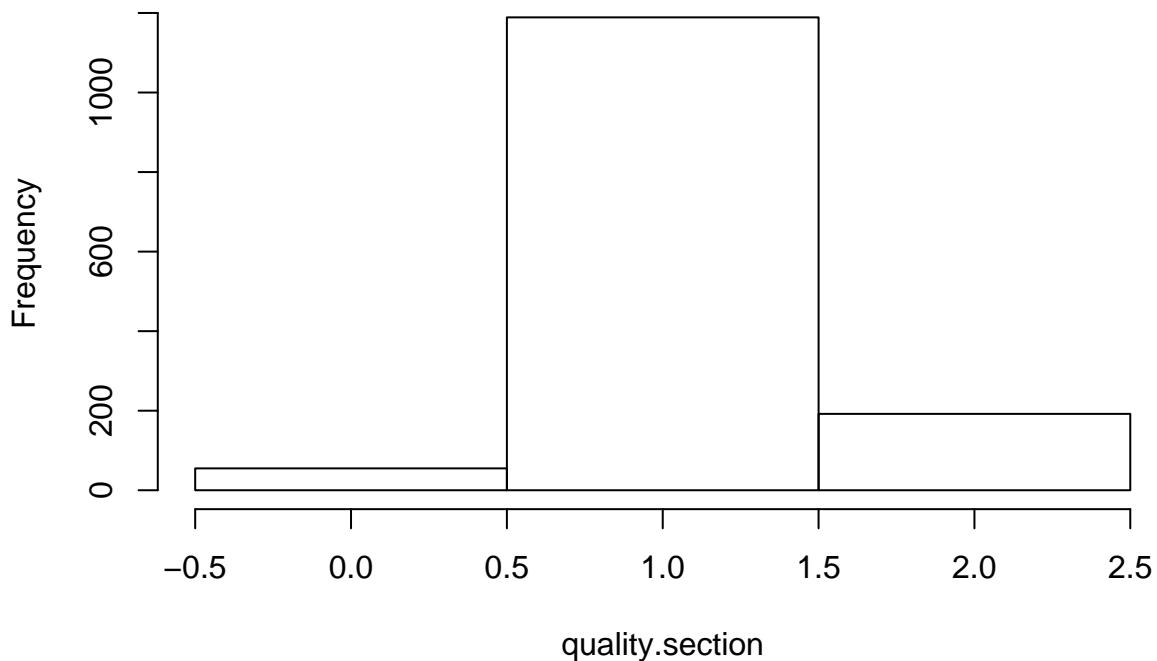
Com veiem, és una funció lineal en  $x$ .

En el nostre cas, però, ens interessaran les dades de distribucions Guassianes multivariants. L'equació del discriminant canvia introduint-hi notació vectorial:

$$\delta_k(\mathbf{x}) = \mathbf{x}^\top \Sigma^{-1} \vec{\mu}_k - \frac{1}{2} \vec{\mu}_k^\top \Sigma^{-1} \vec{\mu}_k + \log(\pi_k)$$

Veiem això en R. Farem el LDA amb la llibreria `MASS`. Introduirem una nova variable `quality.section` que agruparà els vins en dolents (qualificació entre 0 i 3), mediocres (entre 4 i 6) i bons (entre 7 i 10).

## Histogram of quality.section



```
library(MASS)
lda.out <- lda(quality.section ~ data.matrix(db))
lda.out$means

##   data.matrix(db)fixed.acidity data.matrix(db)volatile.acidity
## 0                 7.921818          0.7284545
## 1                 8.254079          0.5369092
## 2                 8.792708          0.4038542
##   data.matrix(db)citric.acid data.matrix(db)residual.sugar
## 0                 0.1556364         2.371818
## 1                 0.2488730         2.272035
## 2                 0.3676042         2.332292
##   data.matrix(db)chlorides data.matrix(db)free.sulfur.dioxide
## 0                 0.08212727        12.09091
## 1                 0.08014130        15.94113
## 2                 0.07431250        13.75000
##   data.matrix(db)total.sulfur.dioxide data.matrix(db)density
## 0                  33.69091        0.9967000
## 1                  46.73003        0.9967708
## 2                  32.08333        0.9958769
##   data.matrix(db)pH data.matrix(db)sulphates data.matrix(db)alcohol
## 0                 3.398545        0.5736364        10.20545
## 1                 3.320202        0.6283431        10.27869
## 2                 3.296146        0.7466146        11.53238
```

Com veiem, les mitjanes de totes les variables excepte `density` són prou diferenciades. Podem veure com de bé ha discriminat el LDA amb la matriu de confusió:

```
lda.pred <- predict(lda.out)
(lda.confusiontable <- table(quality.section, lda.pred$class))
```

```

## quality.section    0     1     2
##                 0     7    48     0
##                 1    10   1119    60
##                 2     0    105    87
sum(diag(lda.confusiontable))/sum(lda.confusiontable)

## [1] 0.8447075

```

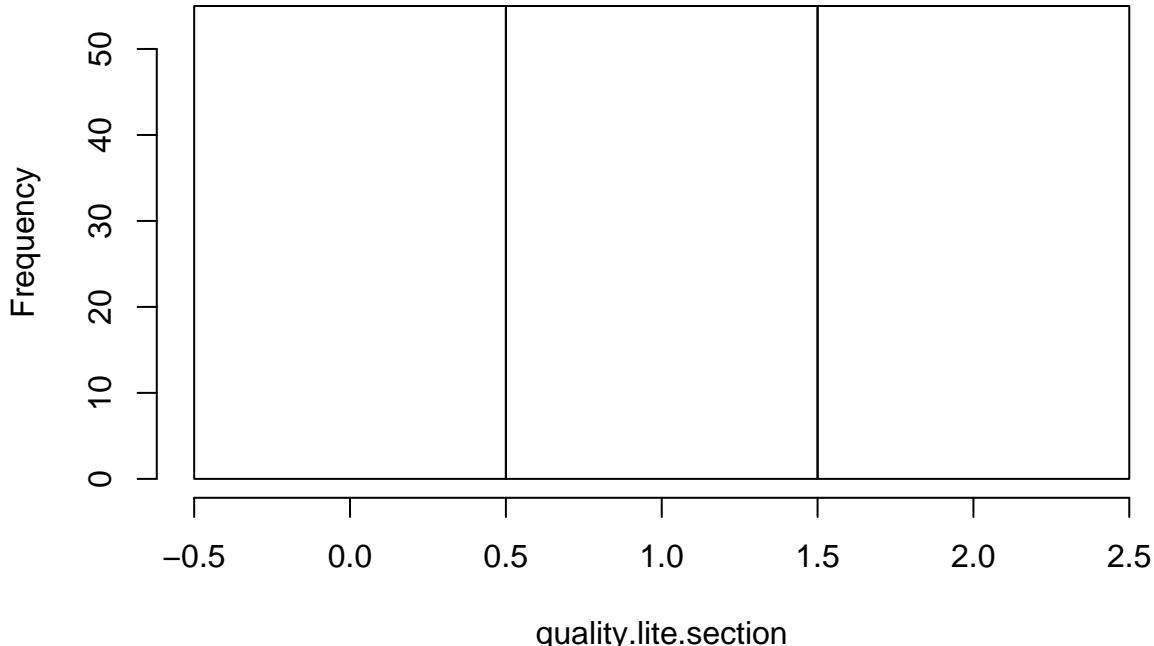
Veiem que distingeix perfectament els vins bons dels dolents però entre mediocres i els extrems hi ha més discrepància. Això es deu a la gran predominància de vins mediocres dins les dades, que els ha donat un paper principal i l'anàlisi s'ha centrat a obtenir millors resultats agrupant vins mediocres. Podríem baixar aquest biaix igualant les dades dels tres conjunts mitjançant la reducció de la classe mediocre i fer múltiples models amb les diferents dades.

```

##    0     1     2
##   55 1189 192

```

### Histogram of quality.lite.section



Ara, amb aquesta nova distribució més equitativa de les dades, tornem a fer un LDA.

```

df.lite <- data.frame(db.lite)
lda.lite.out <- lda(quality.lite.section ~ ., df.lite)
lda.lite.out$means

##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 0      7.921818        0.7284545   0.1556364      2.371818 0.08212727
## 1      8.476364        0.5484545   0.2525455      2.475455 0.08101818
## 2      9.538182        0.3724545   0.4343636      2.420000 0.07805455
##   free.sulfur.dioxide total.sulfur.dioxide density      pH sulphates
## 0          12.09091       33.69091 0.9967000 3.398545 0.5736364
## 1          14.98182       47.05455 0.9973067 3.294364 0.6545455
## 2          12.25455       30.05455 0.9963829 3.242000 0.7405455

```

```

##      alcohol
## 0 10.20545
## 1 10.10909
## 2 11.42364

Veiem com ha funcionat ara:

lda.lite.pred <- predict(lda.lite.out)
( lda.lite.confusiontable <- table(quality.lite.section, lda.lite.pred$class))

##
## quality.lite.section  0   1   2
##                      0 38 13  4
##                      1 10 34 11
##                      2  1  8 46
sum(diag(lda.lite.confusiontable))/sum(lda.lite.confusiontable)

## [1] 0.7151515

```

## 6 Anàlisi clúster

### 6.1 Breu síntesi

Un altre dels anàlisis que realitzarem sobre les dades és l'**Anàlisi Clúster**, l'anàlisi d'agrupaments. Tal i com indica el seu nom, amb aquest estudi es buscaran els grups naturals que hi ha dins de les dades. Aquests grups estan formats per observacions “semblants” entre elles, on aquest nivell de semblança s’ha de definir prèviament i s’ha d’entendre com la distància entre les observacions. A més, aquest estudi ens permet reduir el conjunt de dades de  $n$  observacions a  $k$  grups prou semblants.

Pel que fa als algoritmes que s'utilitzen per trobar aquests grups, bàsicament els dividirem en 3 tipus: **jeràrquics**, **no jeràrquics** i els mètodes **model-based**. En relació als primers, val a dir que, un cop una observació s’ha assignat a un grup, ja no canvia de grup. L’algoritme més utilitzat és l’**aglomeració jeràrquica**, que simplement va ajuntant a cada pas les observacions/grups que estiguin a menor distància. Pel càclul de la distància entre dos grups ja creats, es pot considerar:

- **Single linkage**: La distància entre les observacions més properes entre els dos grups. Sensible als outliers.
- **Complete linkage**: La distància entre les observacions més llunyanes entre els dos grups. Sensible als outliers.
- **Average linkage**: La mitjana de la distància entre les observacions dels dos grups. Menys sensible als outliers.
- **Centroid distance**: Es defineix la distància entre grups com  $d_{rs}^2 = \sum_{j=1}^p (\bar{x}_{rj} - \bar{x}_{sj})^2$ , on  $r$  i  $s$  són els dos grups. Menys sensible als outliers.
- **Ward’s criterion**: Es defineix  $\Delta = \frac{n_r n_s}{n_r + n_s} d_{rs}^2$  i s’agrupen els dos grups amb  $\Delta$  mínim. Menys sensible als outliers.

Per altra banda, en els algoritmes **no jeràrquics**, una observació pot anar canviant de grup a mesura que avança el procediment. L’algoritme més utilitzat en aquest cas és el  **$K$ -Means**. La idea és anar calculant pas a pas els centres de cada grup i anar ajustant-los a les dades. Es comença amb un conjunt de centres aleatòri i s’assignen les observacions més properes a cada centre. A cada pas es recalculen els centres i es reassiguen les dades. Quan en un pas ja no hi ha més reassignacions, s’acaba. En nombre de grups  $K$  és fixat.

Per últim, així com tant els algoritmes jeràrquics o no jeràrquics no feien cap tipus d’assumpció sobre la distribució de les dades, en els mètodes **model-based** s’utilitzen models probabilístics concrets per agrupar les dades. La idea és que s’entenen les dades com un *finite mixture model*

$$g(x|\pi, \theta) = \pi_1 f_1(x|\theta_1) + \cdots + \pi_k f_k(x|\theta_k)$$

on les  $f_i$  són les distribucions de probabilitat de cada un dels  $k$  grups, que normalment es té  $f_i \sim N(\mu, \Sigma)$ . A partir d'aquesta assumpció, es calculen les probabilitats de que cada observació  $x_i$  pertanyi al grup  $i$  com

$$\frac{\pi_i f_i(x_j|\theta_i)}{\sum_{i=1}^k \pi_i f_i(x_j|\theta_i)}$$

El procediment en aquests algoritmes probabilístics és estimar el *finite mixture model* amb màxima versemblança, calcular les probabilitats *a-posteriori* de cada observació a cada grup i assignar-la al grup on tingui una probabilitat més elevada.

## 6.2 Anàlisi Clúster de les nostres dades

Tal i com ja s'havia fet abans, s'utilitzarà un *dataset* on s'ha extret la variable resposta **quality**. El primer que es fa és l'estandardització de les dades per aconseguir que totes tinguin mitjana zero, anomenada  $X_s$ . A continuació, es calcula la matriu de distàncies  $D$  euclidianes entre les observacions de les dades i s'observa que la distància més gran és 12.98895.

A continuació, es procedeix a realitzar-se l'anàlisi clúster de les dades amb cada un dels mètodes explicats a l'apartat anterior.

### 6.2.1 Anàlisi Cluster Jeràrquic

En primer lloc, es realitza l'aglomeració jeràrquica de les dades amb 4 tipus diferents de distàncies entre grups (indicades com a títol de cada *plot*). Donat que la variable resposta d'aquest anàlisi és la qualitat del vi, s'haurien de veure agrupacions d'observacions amb qualitats similars, ja que s'entén que si dues observacions tenen la mateixa qualitat, són observacions properes. Així doncs, donat que la qualitat s'ha valorat de l'1 al 10, s'haurien de detectar un nombre menor que 10 de grups clars

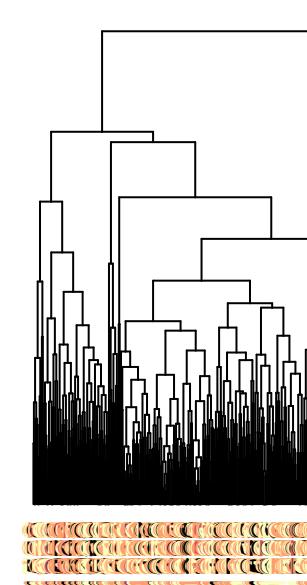
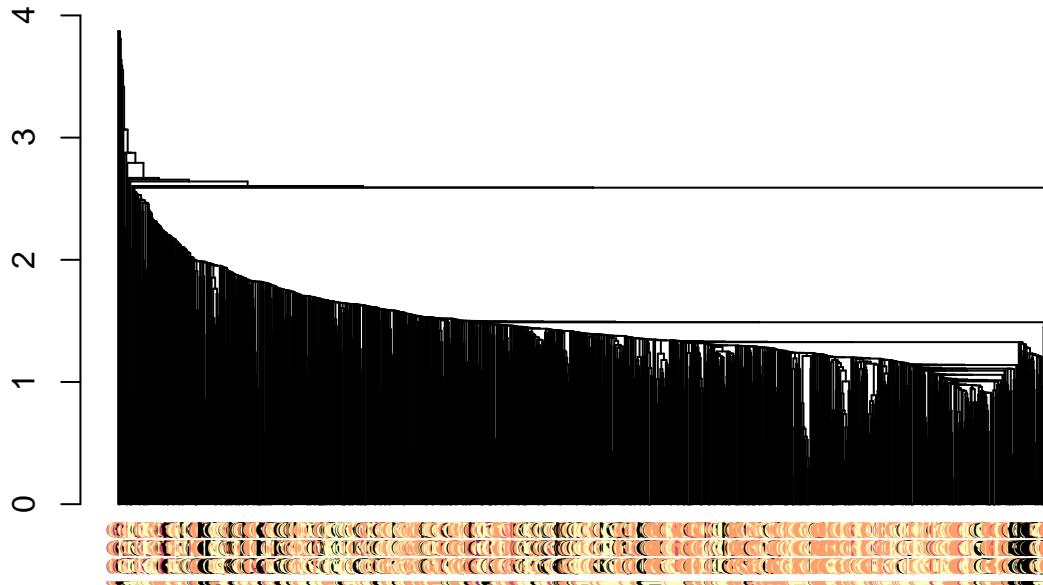
```
## Warning: package 'colorspace' was built under R version 3.5.2
## Loading required package: viridisLite
## Warning: package 'dendextend' was built under R version 3.5.2
##
## -----
## Welcome to dendextend version 1.12.0
## Type citation('dendextend') for how to cite the package.
##
## Type browseVignettes(package = 'dendextend') for the package vignette.
## The github page is: https://github.com/talgalili/dendextend/
##
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/issues
## Or contact: <tal.galili@gmail.com>
##
## To suppress this message use: suppressPackageStartupMessages(library(dendextend))
## -----
##
## Attaching package: 'dendextend'
```

```

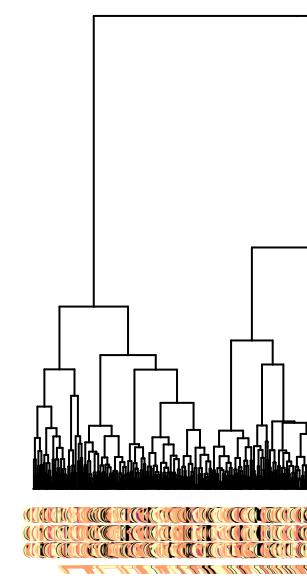
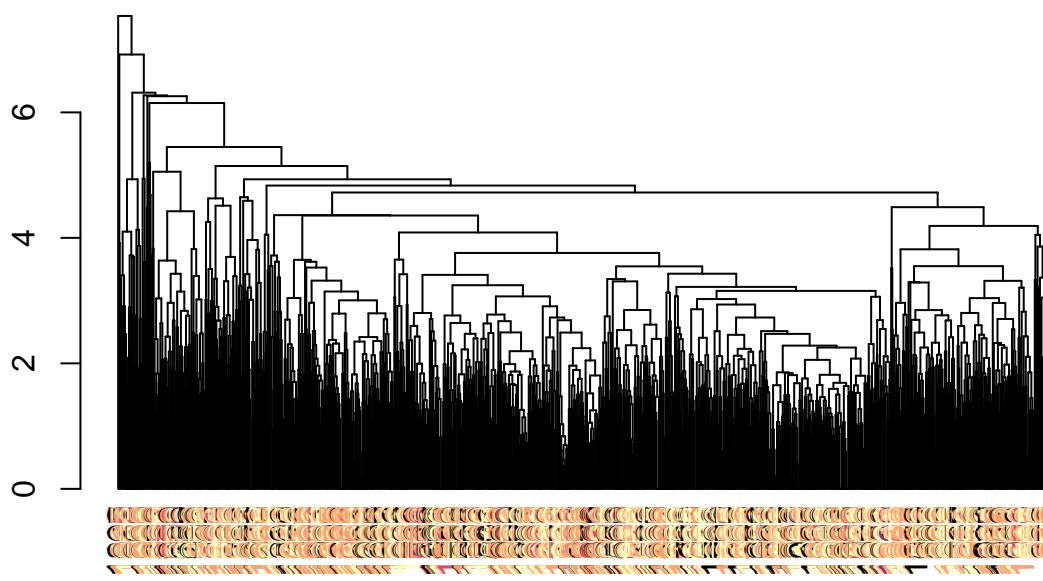
## The following object is masked from 'package:stats':
##
##     cutree

```

### Single linkage



### Average linkage



En el cas del *Single linkage*, sembla ser que hi ha un grup molt gran i petits grupets. Això podria significar que la variabilitat de la qualitat de les observacions és petita, és a dir, que tenim moltes puntuacions iguals. Si es volguéssin agrupar les dades en un nombre més reduït de grups seguint aquesta agrclomeració, quedaría un grup amb un porcentatge molt elevat de les observacions i grups amb un percentatge molt petit. Així doncs, en aquest cas, no podem dir que resultin detectables les diferents qualitats de vi en aquesta agrupació. Per poder diferenciar un nombre raonable de grups, s'hauria de tallar a una distància de més de 3.

En el cas del *Complete linkage*, sembla ser que hi ha unes agrupacions de tamanys més iguals. De fet, es

poden veure clarament tres grups, que es podria suposar que són agrupacions de vi de alta qualitat, vi de baixa qualitat i vi de qualitat mitjana. En aquest cas doncs, a diferència de l'altre, sembla que sí que són detectables les diferents qualitats del vi. Per poder diferenciar aquests 3 grups clars, hauríem de tallar a una distància de més de 10.

En el cas del *Average linkage*, es té un comportament semblant al de *Single linkage*: s'acaba formant un grup molt gran i grups més reduïts. Així que tampoc serveix per detectar les diferents qualitats del vi. Per poder diferenciar un nombre raonable de grups, s'hauria de tallar a una distància de més de 6.

Per últim, en el cas del *Ward's criterion* és el cas on més clar es veu que es formen grups mes o menys del mateix tamany i iguals. En concret, es veu que es formen 4 grups clars i, per tant, són detectables les diferents qualitats del vi. Per poder diferenciar un nombre raonable de grups, s'hauria de tallar a una distància de més de 40.

Ara bé, donat que s'han marcat de colors diferents les observacions, tenint en compte la seva puntuació, s'observa que a cada grup hi ha barreja de puntuacions. Això no és preocupant, ja que sabem que la puntuació ha estat posada de manera relativament subjectiva i pot no estar perfectament correlacionada amb el valor de les dades, que és en el que es basa aquest algoritme per fer els grups. Per poder diferenciar aquests 4 grups clars, s'hauria de tallar a una distància de més de 3. En definitiva, s'observa que dues observacions de qualitats prou diferents poden caure dins el mateix clúster. Un altre detall a remarcar, és que hi ha moltes observacions de la mateixa qualitat.

A continuació, s'analitza en més profunditat com han agrupat aquests algoritmes, tallant l'arbre en cada cas per formar un nombre en concret de clústers. En el primer cas, el *Single Linkage*, s'ha tallat l'arbre per obtenir 3 clústers i el resultat és bastant dolent, ja que com s'havia comentat abans, es crea un clúster molt grans i dos de molt petits (de 2 i 1 observació). A més, el clústers petits contenen observacions amb qualitat 5, que és la més abundant al clúster gran.

```
## clusters.closest
##   1   2   3
## 1433   2   1

##
## clusters.closest   3   4   5   6   7   8
##                 1   6   48  603  584  176   16
##                 2   0   0    2   0    0    0
##                 3   1   0    0   0    0    0
```

En el cas del *Complete Linkage*, s'observa que si es creen tres grups, les mides dels grups són bastant similars. Tot i que tots els grups contenen moltes observacions de qualitat 5 i 6, s'observa que les observacions de qualitat més baixa s'han concentrat al clúster 1 i les de qualitat més alta al clúster 2. Així doncs, s'ha realitzat una separació més o menys dependent de la qualitat. A més, en el clúster “d'alta qualitat” és on menys observacions de qualitat 5 hi ha i en el clúster de “baixa qualitat” és on més n'hi ha.

```
## clusters.farthest
##   1   2   3
## 473  432  531

##
## clusters.farthest   3   4   5   6   7   8
##                 1   5   29  260  161  18    0
##                 2   2    7  103  199  111  10
##                 3   0   12  242  224   47    6
```

En el cas de l'*Average Linkage*, es té una situació semblant a la del *Single Linkage*: un grup molt gran i dos de petits amb 3 i 1 observacions. Per tant, no es detecten bé les diferents qualitats del vi.

```
## clusters.average
##   1   2   3
```

```

## 1432    3    1
##
## clusters.average   3   4   5   6   7   8
##           1   6   48  603  583  176   16
##           2   1   0   2   0   0   0
##           3   0   0   0   1   0   0

```

Per últim, en el cas del *Ward Criterion*, s'observa que es poden crear 4 grups de tamanys similars. A més, al primer grup es concentren gairebé totes les observacions de pitjor qualitat (3 i 4), en el segon les de qualitat una mica millor (4 i 5), en el tercer és on abunden més les de qualitat mitjana-alta (6 i 7) i en l'últim és on hi ha la gran majoria de qualitat alta (7 i 8). Per tant, aquest és el cas on més clarament s'agrupa segons la qualitat del vi.

```

## clusters.ward
##   1   2   3   4
## 464 348 275 349
##
## clusters.ward   3   4   5   6   7   8
##           1   5   29  247  164   18   1
##           2   1   11  224   99   12   1
##           3   1   4   68  139   59   4
##           4   0   4   66  182   87   10

```

Una aproximació de l'error en el cas del *Ward Criterion* és 44,84%.

```
## [1] 0.448468
```

En conclusió, si s'estableixen com a criteris de distància entre grups el *Ward Criterion*, un algoritme d'agrupació jeràrquic és capaç d'agrupar les observacions més o menys per qualitat. També s'observa que les qualitats altes o les qualitats baixes són les que millor es detecten, cosa que fa pensar que ha d'haver-hi detalls prou significants en les dades fisicoquímiques per diferenciar un vi dolent d'un vi molt bo.

### 6.2.2 Anàlisi Clúster No Jeràrquic

En segon lloc, s'aplicarà un algoritme no jeràrquic per agrupar les observacions. En concret, s'aplicarà l'algoritme **K-Means** amb  $K = 4$ , per poder-ho comparar amb l'aglomeració jeràrquica amb el criteri de Ward utilitzada anteriorment.

```

## [1] 484 345 313 294
##
##   fixed.acidity volatile.acidity citric.acid residual.sugar   chlorides
## 1      -0.4242339          0.65166691 -0.76034028     -0.2364532  0.1137701
## 2     -0.1109607          0.07591068  0.02159113      0.2272295  0.3277178
## 3      1.3943290          -0.71384660  1.24462415      0.3969197  0.3447877
## 4     -0.6558310          -0.40191152 -0.09867893     -0.2999548 -0.9389316
##
##   free.sulfur.dioxide total.sulfur.dioxide   density       pH
## 1        -0.44482721         -0.4515851 -0.02771249  0.29565966
## 2         1.02822937          1.3133876  0.28608984 -0.07093096
## 3        -0.51134331         -0.5030402  0.87737138 -0.92961052
## 4        0.07009418          -0.2622446 -1.22416800  0.58619046
##
##   sulphates   alcohol
## 1 -0.4264400 -0.4497398
## 2 -0.1825842 -0.5678806
## 3  0.5090933  0.1585502
## 4  0.3742936  1.2379818

```

```

## [1] 0.3687725
##
##      1   2   3   4
##  3   5   1   1   0
##  4  33   8   4   3
##  5 264 214  82  45
##  6 168 109 157 150
##  7  14   12  64  86
##  8   0    1   5  10

```

S'observa que les dades es divideixen en 4 grups de tamanys similars. Pel que fa a la distribució d'observacions dintre de cada un dels grups tenim:

- Grup 1: Qualitat **baixa** (3-4): La majoria d'observacions de baixa qualitat es concentren al primer grup. També veiem que s'hi concentren moltes observacions de qualitat 5, cosa que fa pensar que el pas de qualitat 4 a qualitat 5 és més subjectiu que objectiu. Pel que fa a les variables que més es tenen en compte per calcular el centre d'aquest grup són la **volatile.acidity**, que es té en compte positivament (és a dir, com més, pitjor sembla ser el vi) i el **citric.acid**, que es té en compte positivament (és a dir, com menys, pitjor és el vi).
- Grup 2: Qualitat **baixa-mitja** (4-5): La majoria d'observacions de qualitat 4 ja hem dit que estan al primer grup, però en el segon grup es concentren la majoria de les observacions restants de qualitats 4 i 5. S'observa que el que més es té en compte en aquest segon grup és les dues variables que es refereixen al Diòxid de Sofre. Les variables que abans marcaven una baixa qualitat del vi, no es tenen tant en compte en aquest grup.
- Grup 3: Qualitat **mitja-alta** (6-7): Una gran part de les observacions de qualitat 6-7 es concentra en aquest grup. Pel que fa a les variables que es destaquen, s'observa que es valora negativament la **volatile.acidity**, fet lògic ja que en els vins de baixa qualitat es valorava positivament, i es valora molt positivament el **citric.acid**, contrari també als vins de baixa qualitat. Un altre variable que sembla ser important en aquest grup és el **fixed.acidity**, que es valora positivament.
- Grup 4: Qualitat **alta** (7-8): Per últim, la majoria d'observacions de qualitat alta es concentren en aquest grup, on es valoren sobretot l'alcohol (variable **alcohol**), positivament, i la densitat (variable **density**) negativament. Ara bé, crida l'atenció que en aquest grup es valori negativament la densitat i en el grup anterior, de qualitat mitjana-alta, aquesta mateixa variable es valori positivament. El mateix passa amb la variable **fixed.acidity**. Ara bé, la variable **volatile.acidity**, igual que abans, es valora negativament.

Totes aquestes inconsistències en la valoració de cada una de les variables a cada un dels grups es pot assignar a la subjectivitat de la valoració.

S'observa que, en aquest cas, la variabilitat entre els clusters comparada amb la variabilitat total de les dades és molt més reduïda. De fet, una aproximació de l'error és  $\frac{\text{BetweenSS}}{\text{TSS}}$ , és a dir, la suma de quadrats entre els clústers entre la suma de quadrats total. En aquest cas, és 36,9%, inferior a l'aconseguida en el *Ward's Criterion*. Per tant, amb aquestes dades, *K-Means* funciona millor.

```

kmeans.maxs <- apply(compar.kmeans, 1, which.max)
kmeans.misclass <- 0
for (k in 1:4) kmeans.misclass <- kmeans.misclass + sum(compar.kmeans[, -kmeans.maxs[k]])
(kmeans.err <- kmeans.misclass/sum(compar.kmeans))

## [1] 0.5389972

```

### 6.2.3 Anàlisi Clúster *Model-Based*

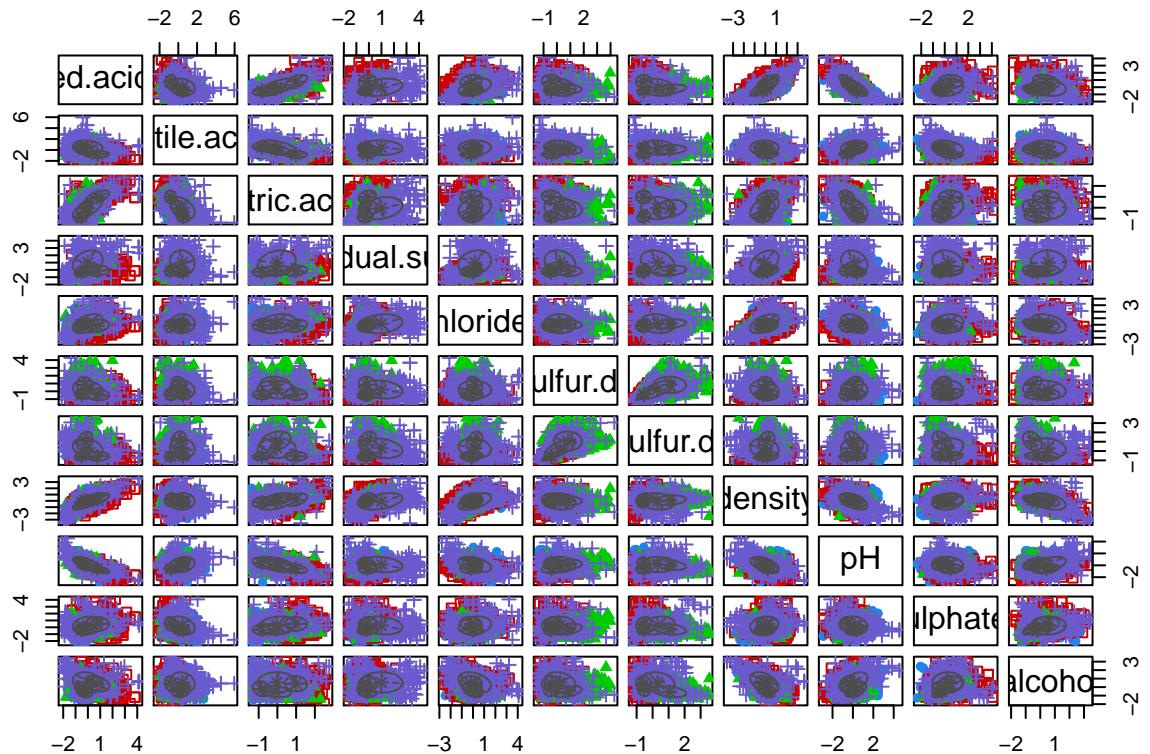
Per últim, anem a agrupar les observacions seguint un model amb 4 mixtures, per poder-ho comparar amb la resta. El primer que s'observa és que els tamanys dels grups són bastant similars. A continuació s'analitzen cada un dels grups:

- Grup 1: Qualitat **baixa-mitja** (4-5): La majoria d'observacions de qualitat 4 o 5 es concentren en aquest primer grup. Una de les variables que es té més en compte és **citric.acid**, negativament. Això ja d'havia observat en l'anàlisi anterior, en el grup dels vins de qualitat baixa. Igual que abans, també es valora positivament el **volatile.acidity**. El **fixed acidity** es valora negativament (abans s'havia vist que en els vins bons es valorava positivament).
- Grup 2: Qualitat **alta** (7-8): La majoria de vins d'aquest grup són d'alta qualitat (7-8). Tal i com havíem vist abans, es valoren positivament el **fixed.acidity** i el **citric.acid**, cosa que encaixa amb el que s'ha vist al grup anterior, de vins de qualitat més baixa. Observem que també, a diferència del grup anterior, el **volatile.acidity** es valora negativament. En l'anàlisi del *K-Means* s'havia vist que la variable **density** es valorava positivament en els bons vins i, en aquest cas, veiem que es compleix.
- Grup 3: Qualitat **mitja-alta** (5-6-7): Els vins de qualitat mitja es concentren en aquest grup, on hi ha poc vins de qualitat extremes (baixes o altes). Igual que abans, observem que tenen un pes important positivament les variables relacionades amb el Diòxid de Sofre. Donat que la qualitat és mitjana, les variables que estem detectant que marquen la qualitat del vi, no es tenen tant en compte en aquest grup.
- Grup 4: Qualitat **baixa** (3-4): Els pitjors vins es concentren en aquest grup, tot i que és el grup de tamany menor. Tornem a observar que es valora positivament la variable **volatile.acidity**, igual que en l'altre grup de vins de baixa qualitat. Un detall que cria l'atenció és la valoració positiva de l'alcohol en aquest grup. Aquest fet s'explica perquè s'observa que també s'han inclòt vins de qualitat molt bona en aquest grup, cosa que ha fet que pugui la mitjana.

Tal i com s'ha comentat abans, el fet que apareguin vins de qualitat màxima i mínima en un mateix grup és degut a la subjectivitat de la valoració.

```
## Warning: package 'mclust' was built under R version 3.5.2
## Package 'mclust' version 5.4.3
## Type 'citation("mclust")' for citing this R package in publications.

## [,1]      [,2]      [,3]      [,4]
## fixed.acidity -0.50377858  0.7749196 -0.25584831 -0.01093973
## volatile.acidity  0.48863434 -0.8148384  0.06790509  0.42530349
## citric.acid    -0.83129135  0.9890591 -0.09214020 -0.03653834
## residual.sugar -0.30473586 -0.1227827 -0.11380340  1.00636542
## chlorides       -0.03418033 -0.1702727  0.13151016  0.15940590
## free.sulfur.dioxide -0.44501120 -0.3934486  0.81647402  0.22995934
## total.sulfur.dioxide -0.54597632 -0.5583936  1.07017890  0.30679826
## density         -0.22415412  0.1989290  0.09904686 -0.09164716
## pH              0.38546216 -0.5765211  0.10620521  0.12703662
## sulphates      -0.37199750  0.4913392 -0.26133821  0.25990220
## alcohol        -0.19986770  0.4046527 -0.54372108  0.55198607
```



```
## 
##      3   4   5   6   7   8
##  1   1 22 215 185  26  0
##  2   1   6  83 203 110 11
##  3   0   6 233 121   8  0
##  4   5 14  74  75  32  5
## [1] 0.494429
```

## 7 Conclusions

## 8 Referències