

PROJECTE ANÀLISI DE DADES: Wine Quality

David Anglada Rotger i Andreu Huguet Segarra

17/5/2019

Contents

1	Introducció	1
2	Descripció de la Base de Dades	1
3	Anàlisi de Components Principals (PCA)	3
3.1	Breu síntesi	3
3.2	PCA de les nostres dades.	3
4	Inferència Multivariant	7
5	Anàlisi Discriminant (LDA/QDA)	7
6	Anàlisi clúster	7
7	Conclusions	7
8	Referències	7

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 possitivament 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 (excellent).

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.*

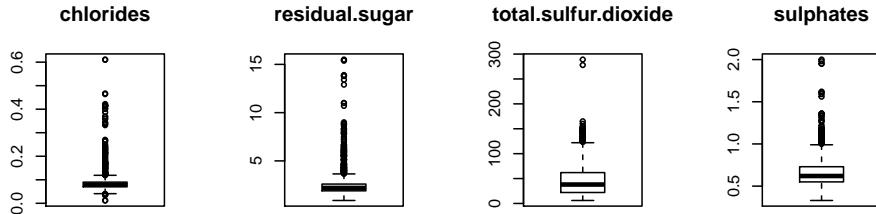
Pel que fa als *missing values*, no n'hi ha cap en aquesta base de dades, tal i com s'indica a la seva descripció oficial.

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

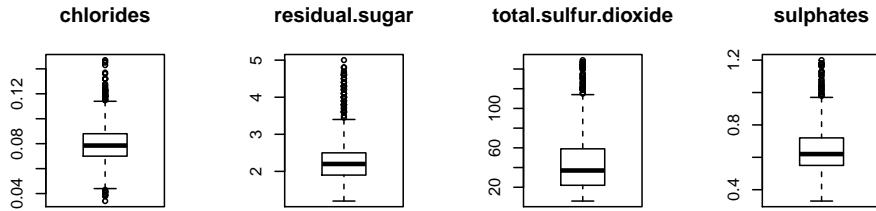
```
## fixed.acidity  volatile.acidity  citric.acid  residual.sugar
## Min. : 4.60  Min. :0.1200  Min. :0.000  Min. : 0.900
## 1st Qu.: 7.10 1st Qu.:0.3900  1st Qu.:0.090  1st Qu.: 1.900
## Median : 7.90 Median :0.5200  Median :0.260  Median : 2.200
## Mean   : 8.32 Mean  :0.5278  Mean  :0.271  Mean  : 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

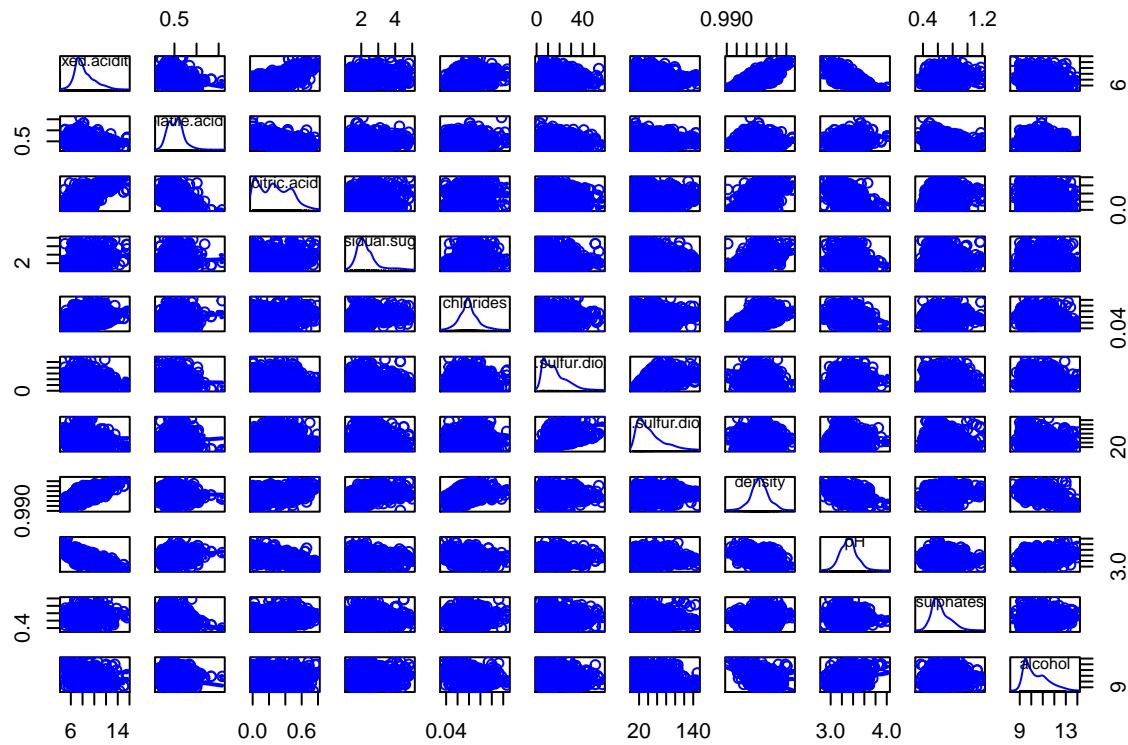
correlaciones.

```
## 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.

```

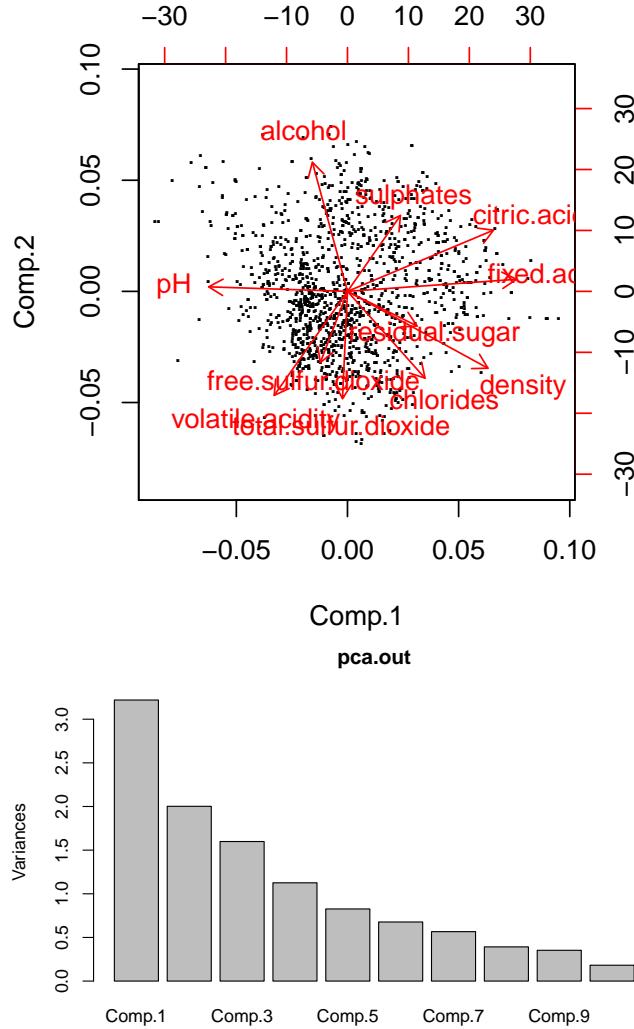
## Importance of components:
##                               Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
## Standard deviation      1.7943537 1.4149643 1.2644268 1.0610261 0.90896784
## Proportion of Variance 0.2927005 0.1820113 0.1453432 0.1023433 0.07511114
## Cumulative Proportion  0.2927005 0.4747118 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**.

4 Inferència Multivariant

5 Anàlisi Discriminant (LDA/QDA)

6 Anàlisi clúster

7 Conclusions

8 Referències