Ejercicio práctico Análisis Factorial

Modelos Lineales Generalizados

Leandro Gutierrez

1. Planteamiento del problema

Para este ejercicio nos enfocaremos en un set de datos que representa la calidad de distintos tipos de vino tinto portugués. Dicha calidad se determina en función de distintos atributos que caracterizan cada tipo de vino. Mediante el Análisis Factorial, exploraremos la posibilidad de clasificarlos en base a distintas caracteristicas del propio vino, tales como el porcentaje de alcohol o su densidad.

El subconjunto de variables del dataset original que utilizaremos son las siguientes:

- **residual.sugar**: la cantidad de azúcar que queda después de la fermentación, es raro encontrar vinos con menos de 1 gramo/litro y los vinos con más de 45 gramos/litro se consideran dulces.
- **density**: la densidad del vino se aproxima a la del agua en función del porcentaje de alcohol y del contenido de azúcar.
- **pH**: describe el grado de acidez o base de un vino en una escala de 0 (muy ácido) a 14 (muy básico); la mayoria de los vinos se sitúan entre 3 y 4 en la escala de pH.
- **alcohol**: el porcentaje de alcohol que contiene el vino.
- **citric.acid**: encontrado en pequeñas cantidades, el ácido citrico puede añadir frescura y sabor a los vinos.
- **volatile.acidity**: la cantidad de ácido acético en el vino, que en niveles demasiado altos puede producir un sabor desagradable a vinagre.

Podras encontrar el dataset en el apartado de 'Material Complementario', carpeta Data con el nombre: 4.2_PCA_AF_ejercicio.csv. Así pues, lo primero que haremos será cargar el dataset en R.

Asi pues, lo primero que haremos es cargar el dataset en R:

```
# leemos los datos
data <- read.csv("/Users/lgutierrez/Proyectos/master/M4/data/4.2_PCA_AF_ejercicio.csv", sep = ";")

# creamos un dataframe tibble
df <- as_tibble(data)

# creamos un auxiliar para no trabajar sobre nuestro dataframe original
df_aux <- df

# previsualizamos datos
head(df_aux)
```

```
## # A tibble: 6 x 12
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
##           <dbl>            <dbl>       <dbl>          <dbl>     <dbl>
## 1           7.4             0.7         0              1.9     0.076
## 2           7.8             0.88        0              2.6     0.098
## 3           7.8             0.76        0.04           2.3     0.092
## 4          11.2             0.28        0.56           1.9     0.075
```

```
## 5           7.4             0.7         0               1.9     0.076
## 6           7.4             0.66        0               1.8     0.075
## # i 7 more variables: free.sulfur.dioxide <dbl>, total.sulfur.dioxide <dbl>,
## #   density <dbl>, pH <dbl>, sulphates <dbl>, alcohol <dbl>, quality <int>
```

1.1 Preparación del dataset.

Tal y como podrás comprobar, el dataset contiene variables que no necesitamos para el ejercicio, por lo que hay que seleccionar únicamente las definidas en el apartado anterior.

- **Ejercicio 1**: Selecciona las variables a utilizar definidas en el apartado anterior del dataset original.

```
# limpiamos el dataset
df_aux <- df_aux %>% select(residual.sugar, density, pH, alcohol, citric.acid, volatile.acidity)

# visualizamos summary
summary(df_aux)
```

```
##  residual.sugar      density             pH           alcohol
##  Min.   : 0.900   Min.   :0.9901   Min.   :2.740   Min.   : 8.40
##  1st Qu.: 1.900   1st Qu.:0.9956   1st Qu.:3.210   1st Qu.: 9.50
##  Median : 2.200   Median :0.9968   Median :3.310   Median :10.20
##  Mean   : 2.539   Mean   :0.9967   Mean   :3.311   Mean   :10.42
##  3rd Qu.: 2.600   3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:11.10
##  Max.   :15.500   Max.   :1.0037   Max.   :4.010   Max.   :14.90
##   citric.acid    volatile.acidity
##  Min.   :0.000   Min.   :0.1200
##  1st Qu.:0.090   1st Qu.:0.3900
##  Median :0.260   Median :0.5200
##  Mean   :0.271   Mean   :0.5278
##  3rd Qu.:0.420   3rd Qu.:0.6400
##  Max.   :1.000   Max.   :1.5800
```

```
# visualizamos glimpse del dataset
glimpse(df_aux)
```

```
## Rows: 1,599
## Columns: 6
## $ residual.sugar   <dbl> 1.9, 2.6, 2.3, 1.9, 1.9, 1.8, 1.6, 1.2, 2.0, 6.1, 1.8~
## $ density          <dbl> 0.9978, 0.9968, 0.9970, 0.9980, 0.9978, 0.9978, 0.996~
## $ pH               <dbl> 3.51, 3.20, 3.26, 3.16, 3.51, 3.51, 3.30, 3.39, 3.36,~
## $ alcohol          <dbl> 9.4, 9.8, 9.8, 9.8, 9.4, 9.4, 9.4, 10.0, 9.5, 10.5, 9~
## $ citric.acid      <dbl> 0.00, 0.00, 0.04, 0.56, 0.00, 0.00, 0.06, 0.00, 0.02,~
## $ volatile.acidity <dbl> 0.700, 0.880, 0.760, 0.280, 0.700, 0.660, 0.600, 0.65~
```

```
# summary(df_aux)
# df_aux[1, c('residual.sugar')] <- NA
# summary(df_aux)
```

Podemos observar que nuestro dataset sanitizado cuenta con 6 columnas y 1599 observaciones. Todas las variables son de tipo cuantitativas continuas y sus tipos de datos intrínsecos son `numeric`. No se observan valores nulos.

1.2 Análisis Factorial.

Una vez dispongas del dataset preparado, realiza el Análisis Factorial para 2 factores utilizando la función factanal.

```r
# realizamos el Análisis Factorial
facts <- factanal(df_aux, factors = 2)
facts
```

```
##
## Call:
## factanal(x = df_aux, factors = 2)
##
## Uniquenesses:
##   residual.sugar          density               pH          alcohol
##            0.874            0.005            0.681            0.654
##      citric.acid volatile.acidity
##            0.005            0.635
##
## Loadings:
##                   Factor1 Factor2
## residual.sugar              0.343
## density            0.225   0.972
## pH                -0.514  -0.234
## alcohol            0.194  -0.555
## citric.acid        0.987   0.147
## volatile.acidity -0.583   0.158
##
##                Factor1 Factor2
## SS loadings      1.675   1.471
## Proportion Var   0.279   0.245
## Cumulative Var   0.279   0.524
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 252.24 on 4 degrees of freedom.
## The p-value is 2.14e-53
```

- **Ejercicio 2**: Podrias indicar qué variables no están bien representadas por los factores? Justifica tu respuesta

```r
print('Como vimos en la teoria, la singularidad de cada variable corresponde a la proporción de la varia
factores, y por lo tanto una alta singularidad en una variable indica que los factores no representan bi
```

```
## [1] "Como vimos en la teoria, la singularidad de cada variable corresponde a la proporción de la vari
```

- **Ejercicio 3**: Imprime la singularidad de cada variable.

```r
print('Para justifiar la respuesta anterior veamos el output "uniqueness" de nuestra funcion "factanal"
```

```
## [1] "Para justifiar la respuesta anterior veamos el output \"uniqueness\" de nuestra funcion \"factan
```

```r
# vemos singularidad de cada variable
facts$uniqueness
```

```
##   residual.sugar          density               pH          alcohol
##        0.8736706        0.0050000        0.6814500        0.6542943
##      citric.acid volatile.acidity
##        0.0050000        0.6347011
```

- **Ejercicio 4**: ¿Qué variables están contribuyendo más a cada uno de los factores? Justifica tu respuesta.

```r
print('Para poder determinar cuanto contribuye cada variable a cada factor encontrado utilizamos el outp
```

```
## [1] "Para poder determinar cuanto contribuye cada variable a cada factor encontrado utilizamos el ou
```

```
# vemos las cargas de cada variable
facts$loadings
```

```
##
## Loadings:
##                   Factor1 Factor2
## residual.sugar             0.343
## density            0.225   0.972
## pH                -0.514  -0.234
## alcohol            0.194  -0.555
## citric.acid        0.987   0.147
## volatile.acidity -0.583   0.158
##
##                 Factor1 Factor2
## SS loadings       1.675   1.471
## Proportion Var    0.279   0.245
## Cumulative Var    0.279   0.524
```

```
print('Este output corresponde a una tabla de cargas, donde podemos observar la contribución de cada va
```

```
## [1] "Este output corresponde a una tabla de cargas, donde podemos observar la contribución de cada v
```

- **Ejercicio 5**: ¿Qué proporción de la varianza está explicada por cada factor? Siguiendo la regla de Kaiser, mantendrias los dos factores?

```
print('Para determinar que proporción de la varianza es explicada por cada factor debemos analizar la s
```

```
## [1] "Para determinar que proporción de la varianza es explicada por cada factor debemos analizar la
```

1.3 Matriz de Residuos.

- **Ejercicio 6**: Imprime la matriz de residuos e interpreta los resultados. ¿Qué variables están mejor representadas en los factores según los valores de la matriz?

```
# obtenemos matriz de cargas
lambda <- facts$loadings

# obtenemos matriz de singularidades
psi <- diag(facts$uniquenesses)

# obtenemos matriz de correlaciones observada
s <- facts$correlation

# obtenemos matriz de correlaciones ajustada
sigma <- lambda %*% t(lambda) + psi

# obtenemos la matriz de residuos
round(s - sigma, 6)
```

```
##                  residual.sugar   density        pH   alcohol citric.acid
## residual.sugar        -0.000001  0.001064  0.043057  0.213892   -0.000384
## density                0.001064 -0.000002  0.000806 -0.000210    0.000003
## pH                     0.043057  0.000806  0.000001  0.175480   -0.000601
## alcohol                0.213892 -0.000210  0.175480 -0.000001    0.000593
## citric.acid           -0.000384  0.000003 -0.000601  0.000593   -0.000004
## volatile.acidity       0.003198 -0.000049 -0.028017 -0.001737   -0.000095
```

```
##                  volatile.acidity
## residual.sugar           0.003198
## density                 -0.000049
## pH                      -0.028017
## alcohol                 -0.001737
## citric.acid             -0.000095
## volatile.acidity        -0.000002
```

```
print('Para determinar que tan bien representan nuestros factores a las variables originales debemos obs
```

```
## [1] "Para determinar que tan bien representan nuestros factores a las variables originales debemos ob
```

1.4 Interpretación de los factores.

- **Ejercicio 7**: Ajusta tres modelos factoriales, uno sin rotación, uno con rotación varimax y uno con rotación promax, y haz una gráfica de dispersión del factor 1 y el 2 para cada uno de ellos. Representa el valor de cada punto con el nombre de la variable.

```
# creamos 3 modelos distintos modificando la rotación
modelo.none <- factanal(df_aux, factors = 2, rotation="none")
modelo.varimax <- factanal(df_aux, factors = 2, rotation="varimax")
modelo.promax <- factanal(df_aux, factors = 2, rotation="promax")

modelo.none
```

```
##
## Call:
## factanal(x = df_aux, factors = 2, rotation = "none")
##
## Uniquenesses:
##   residual.sugar          density               pH          alcohol
##            0.874            0.005            0.681            0.654
##      citric.acid volatile.acidity
##            0.005            0.635
##
## Loadings:
##                  Factor1 Factor2
## residual.sugar    0.302   0.188
## density           0.824   0.562
## pH               -0.536   0.177
## alcohol          -0.234  -0.539
## citric.acid       0.825  -0.560
## volatile.acidity -0.322   0.511
##
##                Factor1 Factor2
## SS loadings      1.897   1.249
## Proportion Var   0.316   0.208
## Cumulative Var   0.316   0.524
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 252.24 on 4 degrees of freedom.
## The p-value is 2.14e-53
```

```
modelo.varimax
```

```
##
## Call:
```

```
## factanal(x = df_aux, factors = 2, rotation = "varimax")
##
## Uniquenesses:
##   residual.sugar          density               pH          alcohol
##            0.874            0.005            0.681            0.654
##      citric.acid volatile.acidity
##            0.005            0.635
##
## Loadings:
##                   Factor1 Factor2
## residual.sugar              0.343
## density            0.225   0.972
## pH                -0.514  -0.234
## alcohol            0.194  -0.555
## citric.acid        0.987   0.147
## volatile.acidity -0.583   0.158
##
##                 Factor1 Factor2
## SS loadings       1.675   1.471
## Proportion Var    0.279   0.245
## Cumulative Var    0.279   0.524
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 252.24 on 4 degrees of freedom.
## The p-value is 2.14e-53
```

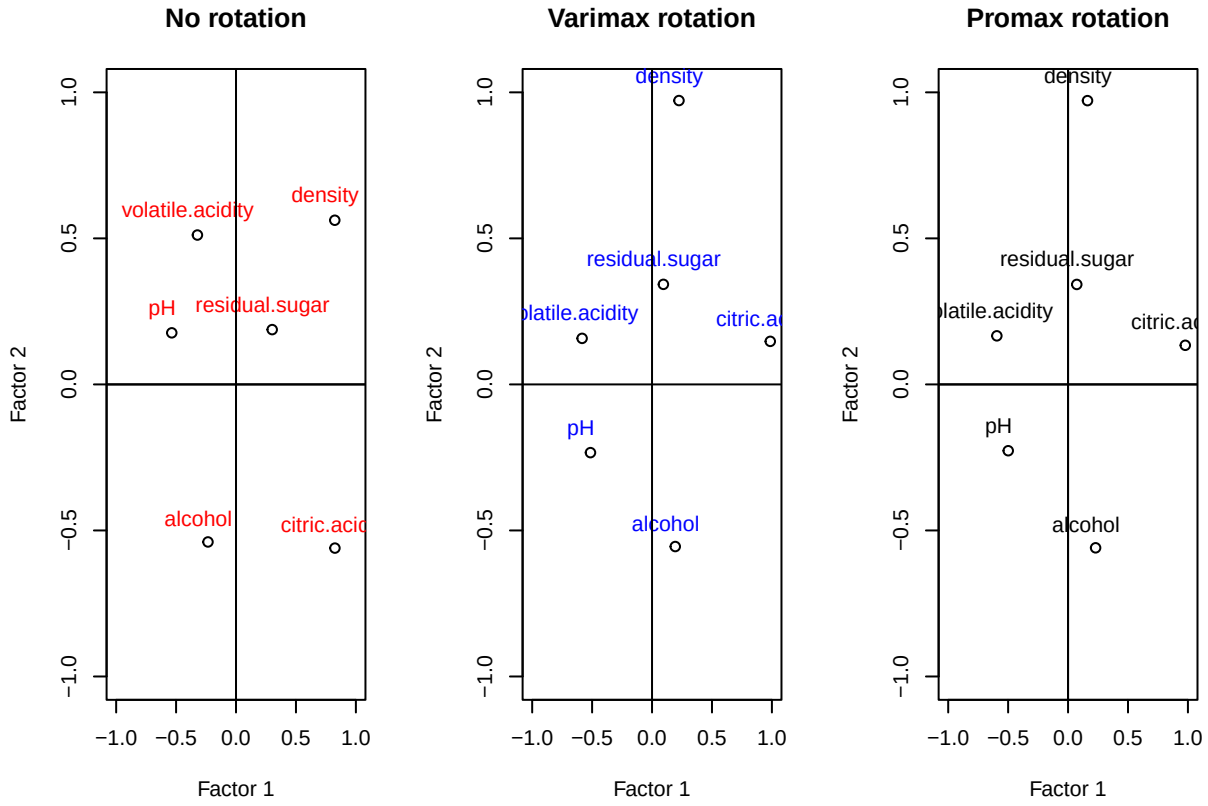
```
modelo.promax
```

```
##
## Call:
## factanal(x = df_aux, factors = 2, rotation = "promax")
##
## Uniquenesses:
##   residual.sugar          density               pH          alcohol
##            0.874            0.005            0.681            0.654
##      citric.acid volatile.acidity
##            0.005            0.635
##
## Loadings:
##                   Factor1 Factor2
## residual.sugar              0.342
## density            0.162   0.972
## pH                -0.499  -0.227
## alcohol            0.230  -0.560
## citric.acid        0.978   0.134
## volatile.acidity -0.594   0.166
##
##                 Factor1 Factor2
## SS loadings       1.643   1.471
## Proportion Var    0.274   0.245
## Cumulative Var    0.274   0.519
##
## Factor Correlations:
##         Factor1 Factor2
## Factor1  1.0000  0.0783
```

```
## Factor2  0.0783  1.0000
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 252.24 on 4 degrees of freedom.
## The p-value is 2.14e-53
```

```r
# definiemos output gráfico (3 gráficos en 1 fila)
par(mfrow = c(1,3))

# primer gráfico: sin rotación
plot(modelo.none$loadings[,1],
     modelo.none$loadings[,2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1,1),
     xlim = c(-1,1),
     main = "No rotation")
abline(h = 0, v = 0)

# texto de color rojo para el gráfico segundo
text(modelo.none$loadings[,1]-0.08,
     modelo.none$loadings[,2]+0.08,
      colnames(df_aux),
      col="red")
abline(h = 0, v = 0)

# segundo gráfico: rotacion = varimax
plot(modelo.varimax$loadings[,1],
     modelo.varimax$loadings[,2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1,1),
     xlim = c(-1,1),
     main = "Varimax rotation")

# texto de color azul para el gráfico segundo
text(modelo.varimax$loadings[,1]-0.08,
     modelo.varimax$loadings[,2]+0.08,
      colnames(df_aux),
      col="blue")
abline(h = 0, v = 0)

# tercer gráfico: rotacion = promax
plot(modelo.promax$loadings[,1],
     modelo.promax$loadings[,2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1,1),
     xlim = c(-1,1),
     main = "Promax rotation")
abline(h = 0, v = 0)

# texto de color rojo para el gráfico segundo
text(modelo.promax$loadings[,1]-0.08,
```

```
     modelo.promax$loadings[,2]+0.08,
      colnames(df_aux))
abline(h = 0, v = 0)
```



- **Ejercicio 8**: Interpreta los resultados. Podrias indicar qué caracteristicas representan mejor al factor 1 y al factor 2 y como se podría interpretar en función del significado que contienen? Si tuvieras que darle un nombre comercial a cada uno de los dos factores, que nombres les otorgarías?

```
print('Tomando la rotación Varimax, la default para el analisis factorial mediante la función "factanal
```

```
## [1] "Tomando la rotación Varimax, la default para el analisis factorial mediante la función \"factana
```

1.4 Puntuación del del ejercicio

Este ejercicio se puntuará con 10 puntos, siendo el mí?nimo necesario para superar la prueba de 5 puntos.

La puntuación es la siguiente:

- Ejercicio 1: 0.5 punto
- Ejercicio 2: 0.75 puntos
- Ejercicio 3: 0.75 puntos
- Ejercicio 4: 1.5 puntos
- Ejercicio 5: 1 puntos
- Ejercicio 6: 1 punto
- Ejercicio 7: 1.5 puntos
- Ejercicio 8: 3 puntos