Ejercicio 5

Primero carguemos los datos

```
data <- read.delim("comida_francesa.txt")
```

Para elegir el número de factores, haremos un breve análisis de componentes principales y el número de factores que elegiremos serán el número de componentes principales que acumulen más del 85% de variabilidad (elección arbitraria).

Como nota, usaremos la función `prcomp()` en lugar de `princomp()` ya que la segunda utiliza el factor $1/n$ en lugar de $1/(n-1)$ para calcular el estimador de la matriz de correlaciones, lo que lo hace menos compatible con el resto de funciones de R[1].

```
pc <- prcomp(data[, 2:ncol(data)], center = T, scale = T)
summary(pc)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.082 1.3524 0.79560 0.35768 0.23903 0.13631 0.03151
## Proportion of Variance 0.619 0.2613 0.09043 0.01828 0.00816 0.00265 0.00014
## Cumulative Proportion  0.619 0.8803 0.97077 0.98904 0.99720 0.99986 1.00000
```

Como con tres componentes principales acumulamos el 88% de la varianza, usaremos dos factores para nuestro análisis.

```
fa_2 <- factanal(data[, 2:ncol(data)], factors = 2)
```

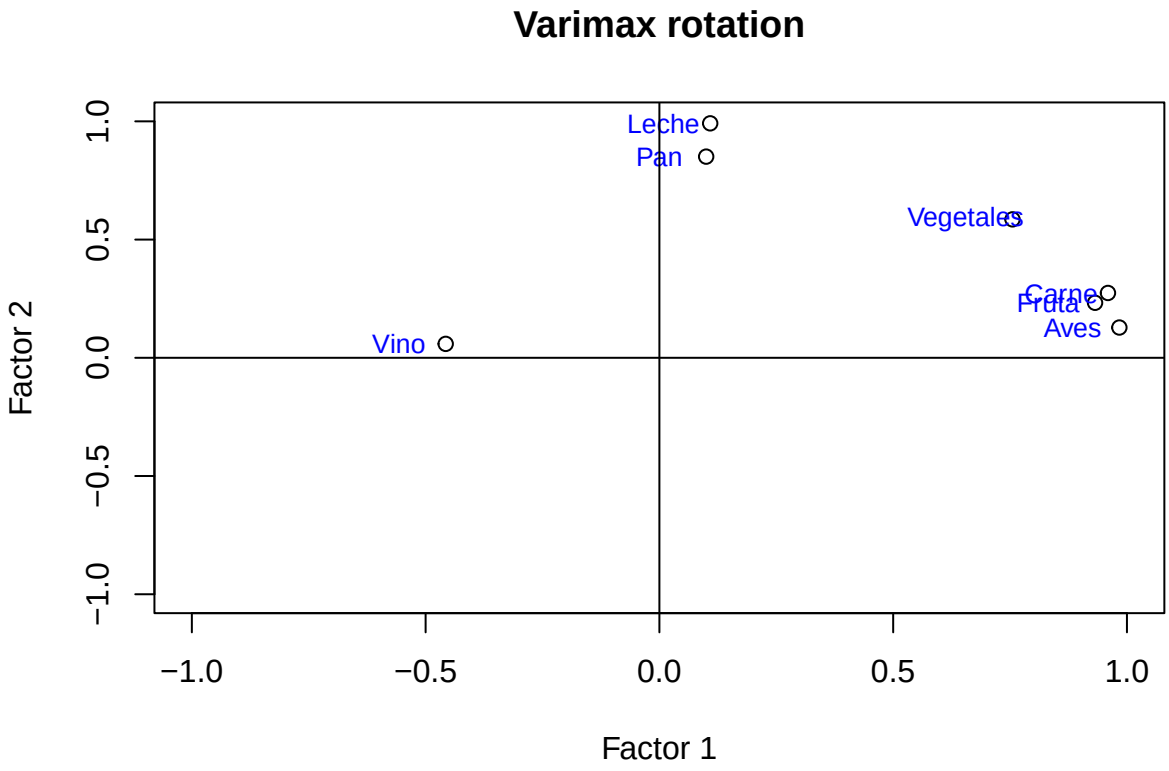
Interpretación

Grafiquemos los factores e interpretemos.

```
plot(fa_2$loadings[,1],
     fa_2$loadings[,2],
     xlab = "Factor 1",
     ylab = "Factor 2",
     ylim = c(-1,1),
     xlim = c(-1,1),
     main = "Varimax rotation")

text(fa_2$loadings[,1]-0.1,
     fa_2$loadings[,2],
      colnames(data[, 2:length(data)]),
      col="blue",
     cex = 0.8)
abline(h = 0, v = 0)
```

[1] 1




```
fa_2
```

```
##
## Call:
## factanal(x = data[, 2:ncol(data)], factors = 2)
##
## Uniquenesses:
##       Pan Vegetales     Fruta     Carne      Aves     Leche      Vino
##     0.266     0.087     0.078     0.005     0.016     0.005     0.788
##
## Loadings:
##           Factor1 Factor2
## Pan        0.100   0.851
## Vegetales  0.755   0.585
## Fruta      0.932   0.233
## Carne      0.959   0.274
## Aves       0.984   0.128
## Leche      0.109   0.992
## Vino      -0.457
##
##                Factor1 Factor2
## SS loadings      3.557   2.199
## Proportion Var   0.508   0.314
## Cumulative Var   0.508   0.822
##
## Test of the hypothesis that 2 factors are sufficient.
```

```
## The chi square statistic is 23.47 on 8 degrees of freedom.
## The p-value is 0.00282
```

Como podemos observar, el ruido (Uniqueness) de las variables es relativamente bajo para todas, salvo por el vino que es casi del 0.8. Lo que en general nos indica que no es un mal modelo de entrada ya que nuestros factores pueden representar las muestras sin la necesidad de corregir con un error grande (la matriz diagonal Ψ en la fórmula $\Sigma = \Lambda\Lambda^T + \Psi$).

Otra señal de que no tenemos un mal modelo es que podemos observar en la consola de R que la varianza acumulada (Cumulative Var) de nuestro modelo es del 82.2%, por lo que a ojo es una buena representación o resumen del total de los datos.

Finalmente, analicemos las cargas para nuestros factores habiendo usado la rotación varimax (que es la estándar en la función `factanal()`):

- Para el primer factor observamos que:
  1. Los gastos en vegetales, fruta, carne y aves componen principalmente este factor. Por lo que en este factor podríamos decir que se compacta la información al respecto de los gastos en alimentos ricos en proteína ó fibra.
  2. El pan y la leche son casi ignorados por el factor.
  3. El vino juega un papel casi importante, pero en comparación de los alimentos ricos en proteína o fibra es bastante menor, por lo que tampoco lo tomaríamos en cuenta.
- Para el segundo factor tenemos:
  1. El pan y la leche son las principales variables que componen a este factor. Entonces este factor nos compacta información al respecto de cuánto dinero gastan las familias en pan y en leche.
  2. Salvo los vegetales, todas las demás variables son bastante ignoradas en este factor.
  3. Los vegetales, al igual que el vino para el factor anterior, según criterios podrían jugar un papel relevante en la información dada por este factor. Nosotros la ignoraremos y concentraremos la interpretación en gastos en pan y leche.

Rotación

Para nuestro análisis anterior, usamos la rotación *varimax* que es la que viene por defecto en la función `factana()`, pero hagamos el contraste con otros dos análisis que usen otras rotaciones. Usemos en un análisis la rotación *promax* y en otro no hagamos ninguna rotación y veamos si hay alguna mejoría de la interpretación.

Esperaríamos encontrar un modelo con cargas o muy cercanas a 0, o muy cercanas a 1. Usando *varimax* podríamos decir que es un buen modelo ya que sólo teníamos una carga por factor que estaba cerca del 0.5 (el vino para el factor 1 y los vegetales para el factor 2), pero veamos si alguna de estas rotaciones mejora la interpretación de la información.

```
fa_2.none <- factanal(data[, 2:ncol(data)], factors = 2, rotation = "none")
fa_2.promax <- factanal(data[, 2:ncol(data)], factors = 2, rotation = "promax")
fa_2
```

```
##
## Call:
## factanal(x = data[, 2:ncol(data)], factors = 2)
##
## Uniquenesses:
##       Pan Vegetales     Fruta     Carne      Aves     Leche      Vino
##     0.266     0.087     0.078     0.005     0.016     0.005     0.788
##
## Loadings:
##           Factor1 Factor2
## Pan        0.100   0.851
## Vegetales  0.755   0.585
```

```
## Fruta      0.932   0.233
## Carne      0.959   0.274
## Aves       0.984   0.128
## Leche      0.109   0.992
## Vino      -0.457
##
##                Factor1 Factor2
## SS loadings      3.557   2.199
## Proportion Var   0.508   0.314
## Cumulative Var   0.508   0.822
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 23.47 on 8 degrees of freedom.
## The p-value is 0.00282
```

```
fa_2.promax
```

```
##
## Call:
## factanal(x = data[, 2:ncol(data)], factors = 2, rotation = "promax")
##
## Uniquenesses:
##       Pan Vegetales     Fruta     Carne      Aves     Leche      Vino
##     0.266     0.087     0.078     0.005     0.016     0.005     0.788
##
## Loadings:
##           Factor1 Factor2
## Pan                 0.866
## Vegetales  0.717   0.416
## Fruta      0.960
## Carne      0.983
## Aves       1.032  -0.123
## Leche               1.011
## Vino      -0.498   0.181
##
##                Factor1 Factor2
## SS loadings      3.718   1.994
## Proportion Var   0.531   0.285
## Cumulative Var   0.531   0.816
##
## Factor Correlations:
##         Factor1 Factor2
## Factor1   1.000  -0.378
## Factor2  -0.378   1.000
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 23.47 on 8 degrees of freedom.
## The p-value is 0.00282
```

- Para el caso en el que no hacemos rotación tenemos:

```
fa_2.none$loadings[,1] # Cargas del primer factor
```

```
##        Pan  Vegetales      Fruta      Carne       Aves      Leche       Vino
##  0.6017038  0.9555287  0.8781203  0.9253912  0.8548299  0.6949917 -0.3244171
```

```
fa_2.none$loadings[,2] # Cargas del segundo factor
```

```
##          Pan    Vegetales        Fruta        Carne         Aves        Leche
##  0.609806864 -0.002118175 -0.388920828 -0.373125523 -0.503448531  0.715579085
##         Vino
##  0.327289904
```

1. En el primer factor tenemos bastantes atributos cercanos a 1, que son vegetales, fruta, carne y aves. Pero desafortunadamente tenemos muchas más variables con carga cercana a 0.5 (tres atributos, a saber pan, leche y vino).
2. Mientras que en el segundo factor ocurre algo peor, ya que ni siquiera tenemos cargas cernanas a 1, sólo la de vegetales que es cercana a 0. Todas las demás están cerca del 0.5 y no nos ayudan a una buena y condensada interpretación de los datos.
3. Podemos concluir que esta rotación no es mejor que la rotación *varimax*, por mucho.

- Para la rotación promax tenemos:

```
fa_2.promax$loadings[,1] # Cargas del primer factor
```

```
##         Pan   Vegetales       Fruta       Carne        Aves       Leche
## -0.02466404  0.71684006  0.96021163  0.98328858  1.03199363 -0.03723432
##        Vino
## -0.49776922
```

```
fa_2.promax$loadings[,2] # Cargas del segundo factor
```

```
##           Pan     Vegetales         Fruta         Carne          Aves
##  0.8657128472  0.4163619554  0.0004790853  0.0367788886 -0.1228212636
##         Leche          Vino
##  1.0110207218  0.1811405475
```

1. Curiosamente, para ambos factores sucedió lo mismo, y es que son bastante análogos a los factores de cuando usamos la rotación *varimax*.

Por buscar diferencias, para el primer factor la carga a los vegetales es menor que en la varimax, pero a cambio de eso, la carga a las variables que no aportaban al factor son aún menores (las de pan y leche). Mientras que para el segundo factor podríamos decir que este sí es un poco mejor (muy poco a decir verdad) que su análogo en la rotación varimax; comparemos las cargas

```
fa_2$loadings[,2]        # Cargas del segundo factor con rotación varimax
```

```
##        Pan  Vegetales      Fruta      Carne       Aves      Leche       Vino
## 0.85082428 0.58542890 0.23268939 0.27419588 0.12801962 0.99159433 0.05889403
```

```
fa_2.promax$loadings[,2] # Cargas del segundo factor con rotación promax
```

```
##           Pan     Vegetales         Fruta         Carne          Aves
##  0.8657128472  0.4163619554  0.0004790853  0.0367788886 -0.1228212636
##         Leche          Vino
##  1.0110207218  0.1811405475
```

Podemos ver que en general las cargas que aportaban interpretación se hicieron un poco más pequeñas o un poco más grandes (según si eran más cercanas a 0 o a 1 respectivamente), y la que no aportaba interpretación (vegetales), bajó su carga considerablemente haciendo que afecte un poco menos en la interpretación.

2. Podemos concluir que no hay grandes diferencias entre estas dos rotaciones, y es bastante subjetivo quedarse con una o con la otra. Si nos preguntan a nosotros, quizás optaríamos por esta, la rotada con *promax* porque en general hace un poco más extremos los valores a 0 y algunos cercanos a 1 los hace más grandes.

Podemos concluir que a grandes rasgos no, las rotaciones que usamos en este documento no encontraron una diferencia significativa para la interpretación de los datos. Sin embargo, por preferencias subjetivas podríamos elegir que la rotación promax mejoró un poco la interpretación de los datos, pero siguen teniendo exactamente la misma relevancia e interpretación todos los atributos en las cargas de ambos factores para las dos rotaciones.

[1] https://stats.stackexchange.com/questions/242864/princomp-outputs-seemingly-wrong-pca-scores-with-cor-true-input-argument

[2] Para la realización de este ejercicio fue de mucha ayuda el siguiente artículo

https://www.geo.fu-berlin.de/en/v/soga/Geodata-analysis/factor-analysis/A-simple-example-of-FA/index.html