Exercise 2

First of all, we import the data and specify what the columns represent. We further take a look at the dimension of the data.

```
frenchfood <- read.csv("~/StatLearn WS22/Übungszettel_WS2122/data/french-food.txt", sep="",
                       row.names = 1)
head(frenchfood)
```

```
##      X1  X2  X3   X4   X5  X6  X7
## CM2 332 428 354 1437  526 247 427
## OW2 293 559 388 1527  567 239 258
## MA2 372 767 562 1948  927 235 433
## CM3 406 563 341 1507  544 324 407
## OW3 386 608 396 1501  558 319 363
## MA3 438 843 689 2345 1148 243 341
```

```
colnames(frenchfood) <- c("bread", "vegetables", "fruit", "meat", "poultry", "milk", "wine")
dim(frenchfood)
```

```
## [1] 12  7
```

a) [1 point]

From the formula in the lecture, we know that for $d = 7$ we have $\Delta \geq 0$ when $p \leq 3$, so an orthogonal factor model with 3 or less factors makes sense.

Eigenvalue method:

```
cor_food <- cor(frenchfood)
eig_food <- eigen(cor_food)

eig_food$values
```

```
## [1] 4.3332373164 1.8302901700 0.6308364243 0.1283275007 0.0575561897
## [6] 0.0188486021 0.0009037968
```

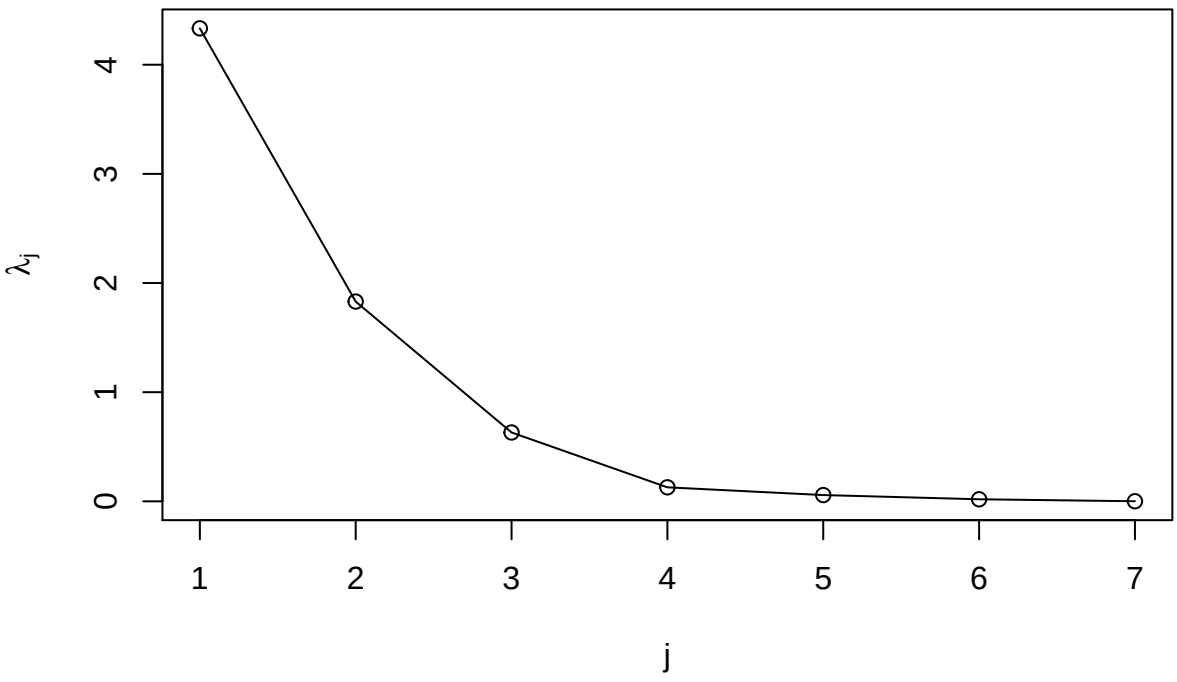
```
sum(eig_food$values > 1)
```

```
## [1] 2
```

The first two eigenvalues of the correlation matrix are greater than 1. The criterion thus suggests to use 2 factors.

Screeplot method: Here, the screeplot just plots the index of the eigenvalue vs the eigenvalue (of the correlation matrix).

```
plot(1:7, eig_food$values, type = "o", xlab = "j", ylab = expression(lambda[j]))
```



This does not give a clear answer, either 2 or 3 factors might be considered.

Formal test based on normality assumption: we can compute p-values for models with 2 or 3 factors.

```
sapply(2:3,
       function(nf){
         paste( "pvalue for ", nf , "factors:" , factanal(frenchfood, factors = nf )$PVAL)})
```

```
## [1] "pvalue for  2 factors: 0.00303255302486333"
## [2] "pvalue for  3 factors: 0.0199073009373262"
```

The formal test still rejects the null hypothesis ($p < 0.02$). But 3 is the maximal amount of factors that we may use for $d = 7$ variables.

b) [0.5 points]

```
fa_ml <- factanal(frenchfood, factors = 3, rotation = "varimax", scores = "regression")
fa_ml
```

```
##
## Call:
## factanal(x = frenchfood, factors = 3, scores = "regression",     rotation = "varimax")
##
## Uniquenesses:
##      bread vegetables      fruit       meat    poultry       milk       wine
##      0.005      0.082      0.049      0.005      0.005      0.005      0.437
##
## Loadings:
```

```
##            Factor1 Factor2 Factor3
## bread       0.230   0.869   0.433
## vegetables  0.772   0.548  -0.144
## fruit       0.882   0.179  -0.376
## meat        0.950   0.223  -0.211
## poultry     0.988          -0.116
## milk        0.146   0.984
## wine       -0.337   0.107   0.662
##
##                Factor1 Factor2 Factor3
## SS loadings      3.439   2.123   0.851
## Proportion Var   0.491   0.303   0.122
## Cumulative Var   0.491   0.795   0.916
##
## Test of the hypothesis that 3 factors are sufficient.
## The chi square statistic is 9.85 on 3 degrees of freedom.
## The p-value is 0.0199
```

c) [1.5 points]

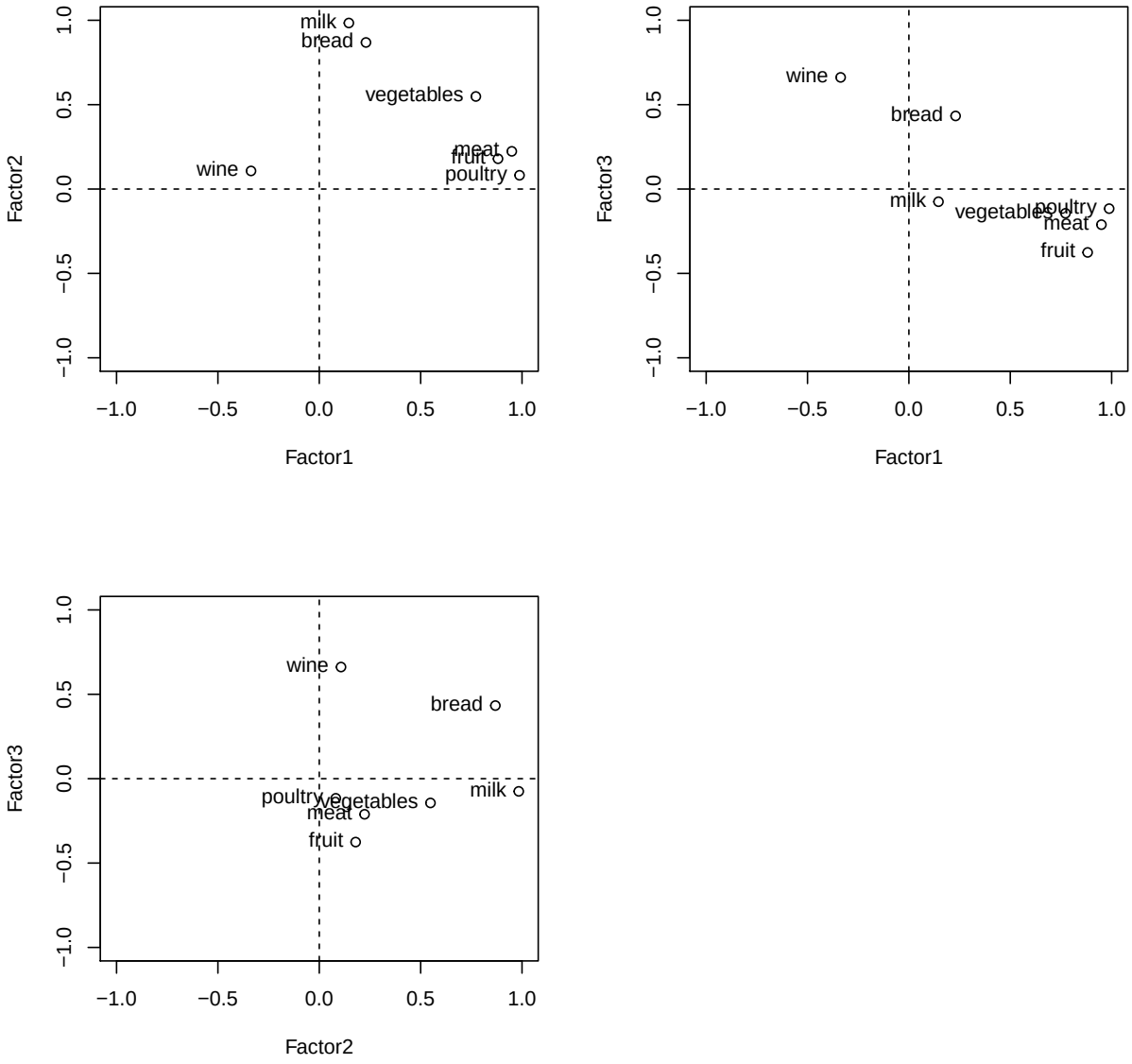
```
par(mfrow=c(2,2))
plot(fa_ml$loadings[,1:2], xlim=c(-1,1),ylim=c(-1,1))
text(fa_ml$loadings[,1:2], label=colnames(frenchfood), pos=2)
abline(v=0, lty=2)
abline(h=0, lty=2)

plot(fa_ml$loadings[,c(1,3)], xlim=c(-1,1),ylim=c(-1,1))
text(fa_ml$loadings[,c(1,3)], label=colnames(frenchfood), pos=2)
abline(v=0, lty=2)
abline(h=0, lty=2)

plot(fa_ml$loadings[,2:3], xlim=c(-1,1),ylim=c(-1,1))
text(fa_ml$loadings[,2:3], label=colnames(frenchfood), pos=2)
abline(v=0, lty=2)
abline(h=0, lty=2)

fa_ml$loadings
```

```
##
## Loadings:
##            Factor1 Factor2 Factor3
## bread       0.230   0.869   0.433
## vegetables  0.772   0.548  -0.144
## fruit       0.882   0.179  -0.376
## meat        0.950   0.223  -0.211
## poultry     0.988          -0.116
## milk        0.146   0.984
## wine       -0.337   0.107   0.662
##
##                Factor1 Factor2 Factor3
## SS loadings      3.439   2.123   0.851
## Proportion Var   0.491   0.303   0.122
## Cumulative Var   0.491   0.795   0.916
```



Vegetables, fruit, meat and poultry are loaded highest on factor 1.

Bread and milk are loaded highest on factor 2, vegetables are also loaded quite high on factor 2.

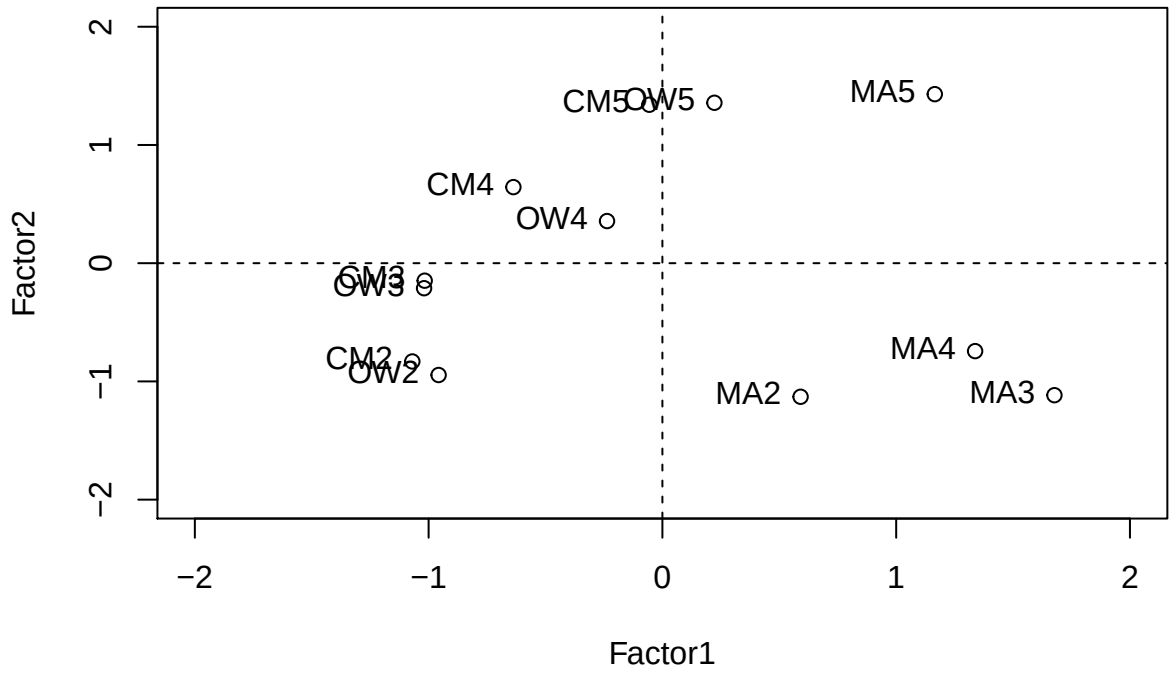
Wine is loaded highest on factor 3.

Interpretation: Factor 1 corresponds to food that is relatively pricey compared to the amount of energy it provides. Factor 2 corresponds to cheap foods that are filling. Factor 3 only corresponds to wine, a luxury food, not providing any energy.

d) [1 point]

```
par(mfrow=c(1,1))
plot(fa_ml$scores[,1:2], xlim=c(-2,2),ylim=c(-2,2))
text(fa_ml$scores[,1:2], label=rownames(frenchfood), pos=2)
```

```
abline(v=0, lty=2)
abline(h=0, lty=2)
```



The families with 5 kids have high factor scores for factor 2, i.e. they buy more bread and milk than families with less children do. Probably because it is an easy and affordable way to feed a lot of children. Further: all manager families have positive scores on factor 1. (more money spend on the pricier foods)