# S&DS363 Factor Analysis

Daniel Kim

4/17/2020

```
library(psych)
library(rela)
library(factoextra)
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
data <- read.csv("~/Downloads/food-texture.csv", row.names = "X")
```

2)

```
cor(data)
```

```
##                  Oil    Density     Crispy   Fracture    Hardness
## Oil       1.00000000 -0.7500240  0.5930863 -0.5337392 -0.09604521
## Density  -0.75002399  1.0000000 -0.6709460  0.5721324  0.10793720
## Crispy    0.59308631 -0.6709460  1.0000000 -0.8439650  0.41109340
## Fracture -0.53373917  0.5721324 -0.8439650  1.0000000 -0.37335844
## Hardness -0.09604521  0.1079372  0.4110934 -0.3733584  1.00000000
```

There seems to be a positive correlation between oil and crispiness of food which makes sense. There is also a positive correlation of crispiness and hardness which also makes sense. There is a negative correlation between hardness and fracture which makes sense for food.

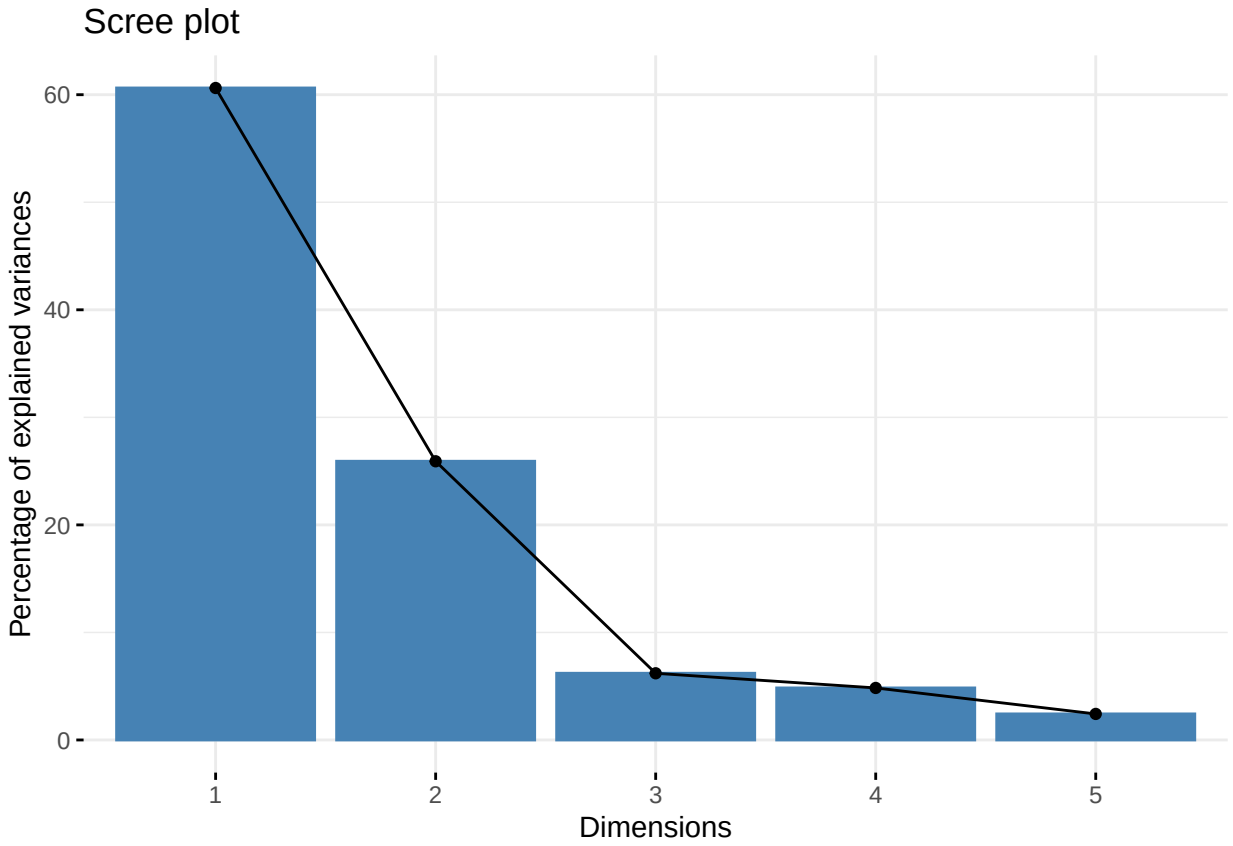
3)

```
KMO(data)
```

```
## Kaiser-Meyer-Olkin factor adequacy
## Call: KMO(r = data)
## Overall MSA =  0.71
## MSA for each item =
##      Oil  Density   Crispy Fracture Hardness
##     0.82     0.71     0.67     0.79     0.43
```

4)

```
food.pca <- prcomp(data, scale = TRUE)
fviz_eig(food.pca)
```




The elbow appears to form at around 3 dimensions meaning it might be ideal to use the first 2 principal components.

5)

**Factor Analysis using Maximum Likelihood**

```
fact1 <- factanal(data, factors = 2)
fact1
```

```
##
## Call:
## factanal(x = data, factors = 2)
##
## Uniquenesses:
##      Oil  Density   Crispy Fracture Hardness
##    0.334    0.156    0.042    0.256    0.407
##
## Loadings:
##          Factor1 Factor2
## Oil      -0.816
## Density   0.919
## Crispy   -0.745   0.635
## Fracture  0.645  -0.573
## Hardness          0.764
##
##                Factor1 Factor2
## SS loadings      2.490   1.316
```

```
## Proportion Var   0.498   0.263
## Cumulative Var   0.498   0.761
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 0.27 on 1 degree of freedom.
## The p-value is 0.603
```

```
repro1 <- fact1$loadings%*%t(fact1$loadings)
repro1
```

```
##                  Oil    Density     Crispy   Fracture    Hardness
## Oil       0.66613977 -0.7500246  0.5956994 -0.5155194 -0.09526886
## Density  -0.75002460  0.8444745 -0.6698646  0.5796718  0.10825742
## Crispy    0.59569942 -0.6698646  0.9577762 -0.8439652  0.41108842
## Fracture -0.51551938  0.5796718 -0.8439652  0.7439766 -0.37339100
## Hardness -0.09526886  0.1082574  0.4110884 -0.3733910  0.59305393
```

```
resid1 <- fact1$cor-repro1
round(resid1,2)
```

```
##            Oil Density Crispy Fracture Hardness
## Oil       0.33    0.00   0.00    -0.02     0.00
## Density   0.00    0.16   0.00    -0.01     0.00
## Crispy    0.00    0.00   0.04     0.00     0.00
## Fracture -0.02   -0.01   0.00     0.26     0.00
## Hardness  0.00    0.00   0.00     0.00     0.41
```

```
#get root-mean squared residuals
len <- length(resid1[upper.tri(resid1)])
RMSR1 <- sqrt(sum(resid1[upper.tri(resid1)]^2)/len)
RMSR1
```

```
## [1] 0.006304819
```

```
sum(rep(1,len)[abs(resid1[upper.tri(resid1)])>0.05])/len
```

```
## [1] 0
```

### Perform Factor Analysis using iterative PCA with Varimax Rotation

```
#this uses the fa() function in the psych package.  Note that this fails with only 2 factors
fact2 <- fa(data, nfactors=3, rotate="varimax", SMC=FALSE, fm="pa")
fact2
```

```
## Factor Analysis using method =  pa
## Call: fa(r = data, nfactors = 3, rotate = "varimax", SMC = FALSE, fm = "pa")
## Standardized loadings (pattern matrix) based upon correlation matrix
##            PA1   PA3   PA2   h2    u2 com
## Oil       0.32  0.88 -0.07 0.88 0.118 1.3
## Density  -0.53 -0.65  0.16 0.73 0.273 2.1
## Crispy    0.83  0.40  0.28 0.93 0.066 1.7
## Fracture -0.76 -0.34 -0.26 0.76 0.238 1.7
## Hardness  0.22 -0.12  0.96 0.98 0.020 1.1
##
##                        PA1  PA3  PA2
## SS loadings           1.71 1.48 1.10
## Proportion Var        0.34 0.30 0.22
```

```
## Cumulative Var        0.34 0.64 0.86
## Proportion Explained  0.40 0.35 0.26
## Cumulative Proportion 0.40 0.74 1.00
##
## Mean item complexity =  1.6
## Test of the hypothesis that 3 factors are sufficient.
##
## The degrees of freedom for the null model are  10  and the objective function was  3.33 with Chi Squa
## The degrees of freedom for the model are -2  and the objective function was  0.01
##
## The root mean square of the residuals (RMSR) is  0.01
## The df corrected root mean square of the residuals is  NA
##
## The harmonic number of observations is  50 with the empirical chi square  0.04  with prob <  NA
## The total number of observations was  50  with Likelihood Chi Square =  0.61  with prob <  NA
##
## Tucker Lewis Index of factoring reliability =  1.094
## Fit based upon off diagonal values = 1
## Measures of factor score adequacy
##                                                  PA1  PA3  PA2
## Correlation of (regression) scores with factors   0.91 0.90 0.98
## Multiple R square of scores with factors          0.84 0.81 0.96
## Minimum correlation of possible factor scores     0.67 0.63 0.92
```

```
#get reproduced correlation matrix
repro2 <- fact2$loadings%*%t(fact2$loadings)
#residual correlation matrix
resid2 <- cor(data)-repro2
round(resid2,2)
```

```
##            Oil Density Crispy Fracture Hardness
## Oil       0.12    0.00  -0.01    -0.01     0.00
## Density   0.00    0.27  -0.01    -0.01     0.00
## Crispy   -0.01   -0.01   0.07     0.00     0.00
## Fracture -0.01   -0.01   0.00     0.24     0.00
## Hardness  0.00    0.00   0.00     0.00     0.02
```

```
#get root-mean squared residuals - again, in output above
len <- length(resid2[upper.tri(resid2)])
RMSR3 <- sqrt(sum(resid2[upper.tri(resid2)]^2)/len)
RMSR3
```

```
## [1] 0.006706366
```

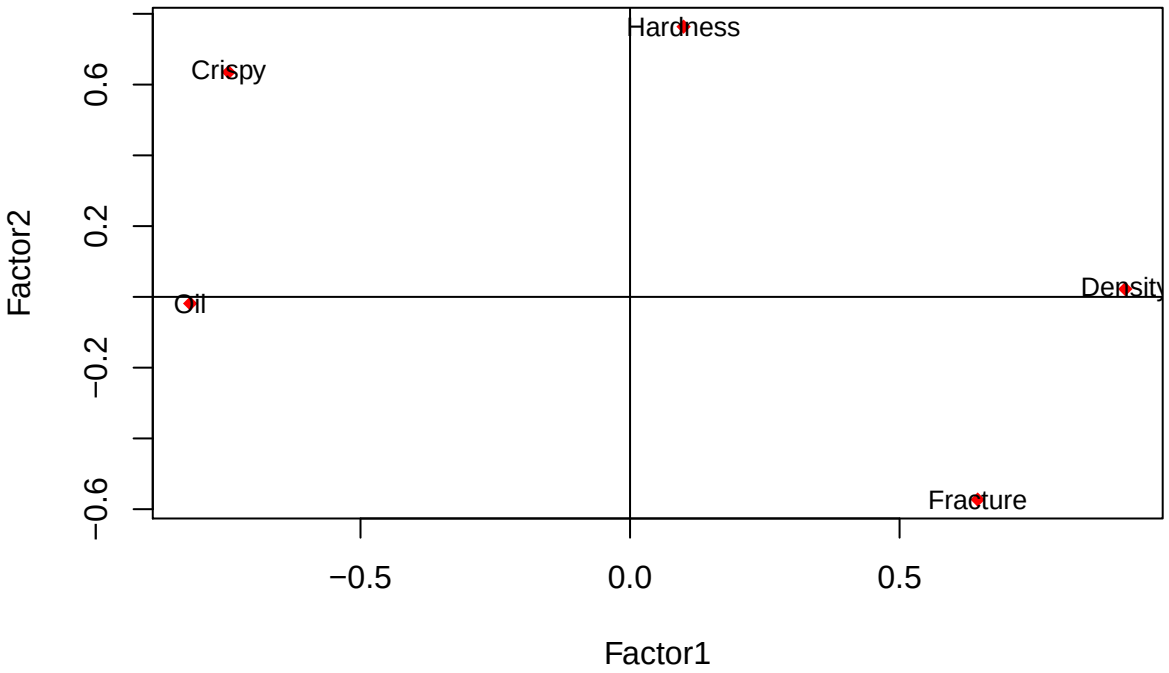
```
#get proportion of residuals greater than 0.05 in absolute value
sum(rep(1,len)[abs(resid2[upper.tri(resid2)])>0.05])/len
```

```
## [1] 0
```

The root square mean residual is slightly lower for the method using maximum likelihood estimation.

6)

```
plot(fact1$loadings, pch=18, col='red')
abline(h=0)
abline(v=0)
text(fact1$loadings, labels=names(data),cex=0.8)
```




Taking a look on the figures above is appears that factor 1 accounts for pastry, which is dense and can be bend a lot before it breaks. Whereas factor 2 accounts for pastry that crispy and hard to break apart. So if we need to names these factors we would probably call them soft pastry (factor 1) and hard pastry (factor 2).