Exploratory Factor Analysis

Yuchen Hu

11/26/2018

It could be noticed that we have a large set of covariates comparing to the number of observations we have. It's possible that a lot of covariates in our datasets are highly correlated, and they might be jointly characterized by some latent factors.

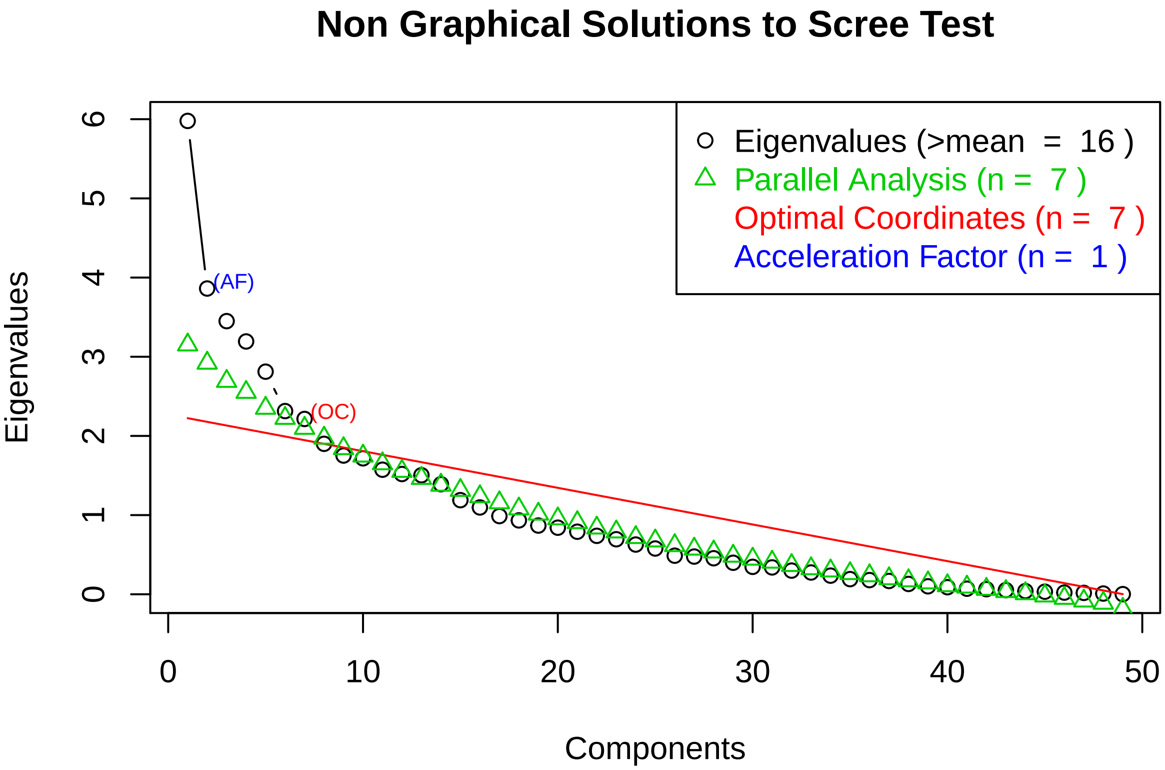
```
library(dplyr)
library(ggplot2)
library(data.table)
library(psych)
library(nFactors)
library(reshape2)
library(tidyr)

food <- read.csv("food_coded.csv",stringsAsFactors = FALSE)
#summary(food)

food <- read.csv("./data/food_coded_clean.csv", na = "nan")
food <- food %>% dplyr::select( - ends_with("_1"))
food <- food %>% dplyr::mutate_if(is.integer, as.factor) %>%
  dplyr::select_if(function(x) !is.character(x))
```

We start by filtering out all the numerical values in our dataframe. To decide the optimal number of factors to include in our model, we start by calculating the eigenvalues and ploting the scree plot.

```
food_numeric <- food[,-c(8,9,14,17,25,26,29,35,36,43,45,57)]
food_numeric <- food_numeric[complete.cases(food_numeric), ] %>%
  lapply(function(x) as.numeric(as.character(x)))
food_numeric <- as.data.frame(do.call(cbind, food_numeric))
ev <- eigen(cor(food_numeric))
ap <- parallel(subject=nrow(food_numeric),var=ncol(food_numeric),
               rep=100,cent=.05)
nS <- nScree(x=ev$values, aparallel=ap$eigen$qevpea)
plotnScree(nS)
```

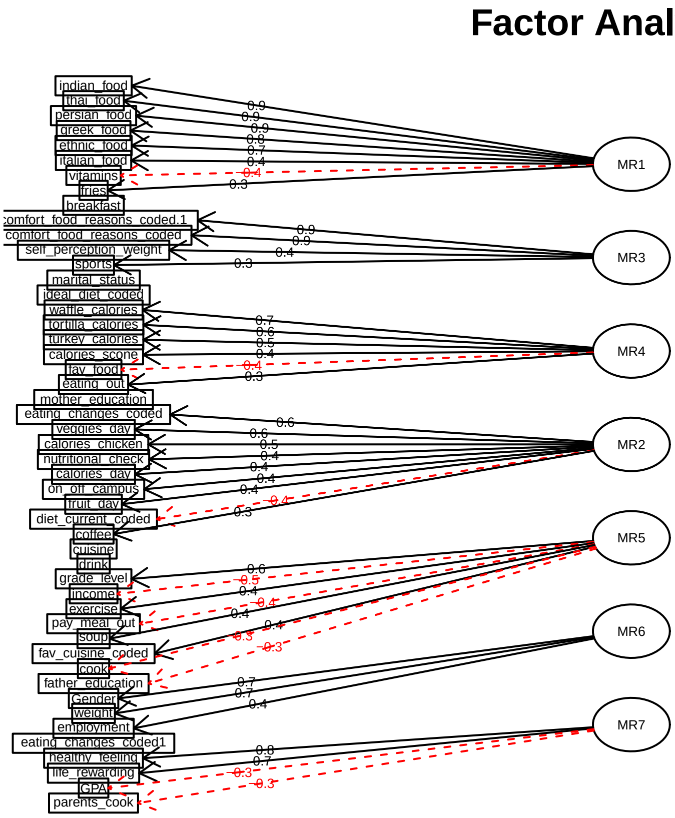


It seems that we may choose to include 7 factors in our model.

```
fa_numeric <- fa(food_numeric, 7)
summary(fa_numeric)
```

```
##
## Factor analysis with Call: fa(r = food_numeric, nfactors = 7)
##
## Test of the hypothesis that 7 factors are sufficient.
## The degrees of freedom for the model is 854  and the objective function was  28.53
## The number of observations was  58  with Chi Square =  1003.38  with prob <  0.00029
##
## The root mean square of the residuals (RMSA) is  0.07
## The df corrected root mean square of the residuals is  0.08
##
## Tucker Lewis Index of factoring reliability =  0.513
## RMSEA index =  0.126  and the 10 % confidence intervals are  0.039 NA
## BIC =  -2464.24
##  With factor correlations of
##       MR1   MR3   MR4   MR2   MR5   MR6   MR7
## MR1  1.00 -0.07  0.02  0.11 -0.10 -0.09  0.00
## MR3 -0.07  1.00 -0.01 -0.08 -0.02  0.04  0.07
## MR4  0.02 -0.01  1.00  0.16 -0.06  0.01 -0.07
## MR2  0.11 -0.08  0.16  1.00 -0.11 -0.12  0.07
## MR5 -0.10 -0.02 -0.06 -0.11  1.00  0.03 -0.10
## MR6 -0.09  0.04  0.01 -0.12  0.03  1.00 -0.07
## MR7  0.00  0.07 -0.07  0.07 -0.10 -0.07  1.00
```

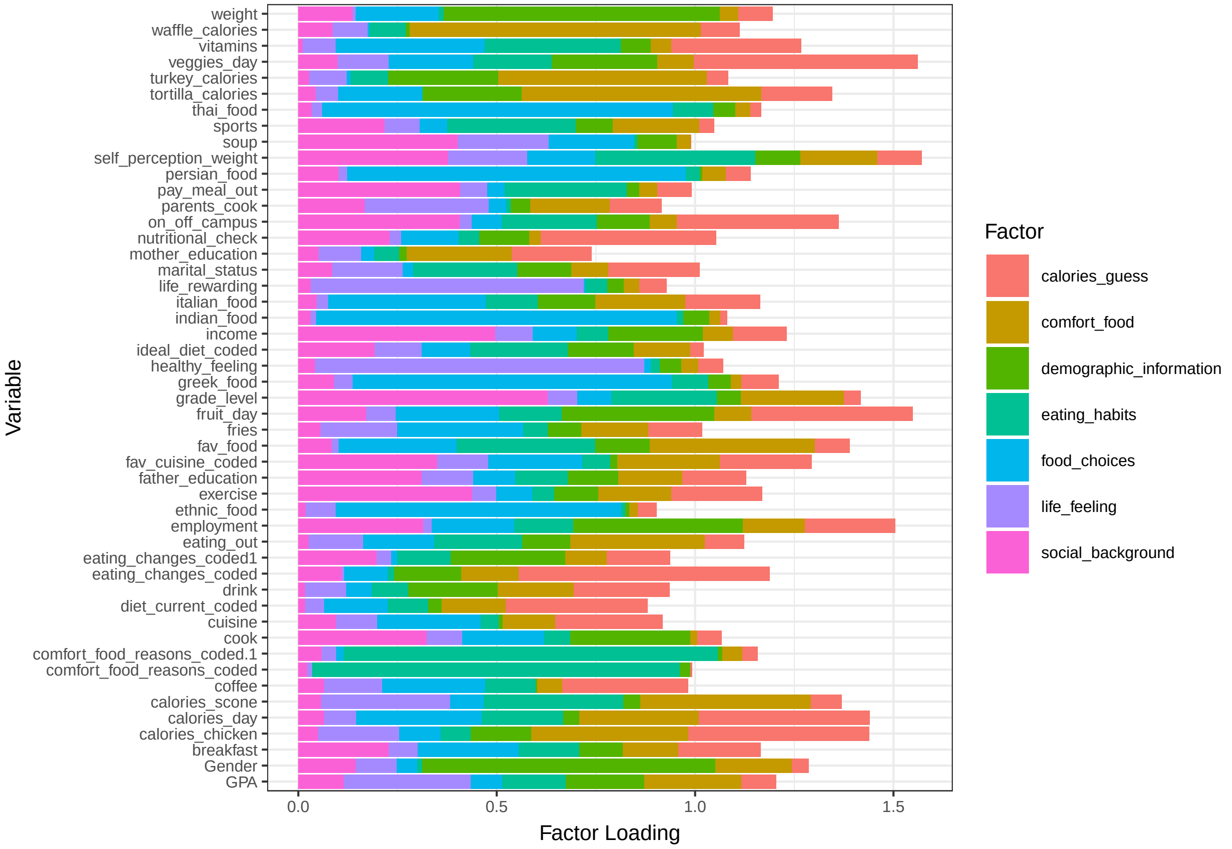
```
fa.diagram(fa_numeric,side=1)
```



From the factor diagram, we may figure out some interpretation for the latent factors by the main variables contributing to them. For factor 1, it's mainly composed of variables explaining the individual preference towards different types of cuisine. For factor 2, it's mainly composed of variables related to individual's daily eating habits. Factor 3 explains individual's choice in terms of comfort food, and factor 4 characterizes how well people did in guessing the calories of different food. Factor 5 focuses on some more general information on the individual's social background, and factor 6 explains individual's demographic information. Finally, factor 7 reflects how people feel about their life.

The dimension could thus be reduced by replacing the covariates by the factor loadings. According to the interpretation of the variables, we define the 7 factors as "food_choices", "eating_habits", "comfort_food", "calories_guess", "social_background", "demographic_information", and "life_feeling" respectively.

```
# plot the result
gathering <- as.data.frame(matrix(nrow = dim(food_numeric)[2], ncol =7))
gathering$Variable <- colnames(food_numeric)
for (i in 1:7) {
  for (j in 1:dim(food_numeric)[2]) {
    gathering[j, i] <- fa_numeric$loadings[j, i]
  }
}
colnames(gathering) <- c("food_choices", "eating_habits", "comfort_food",
                         "calories_guess", "social_background",
                         "demographic_information", "life_feeling", "Variable")
gathering <- gathering %>% gather("Factor", "Value", 1:7)
ggplot(gathering, aes(Variable, abs(Value), fill=Factor)) +
  geom_bar(stat="identity") + coord_flip() +
  ylab("Factor Loading") +
theme_bw(base_size=8)
```


weight
waffle_calories
vitamins
veggies_day
turkey_calories
tortilla_calories
thai_food
sports
soup
self_perception_weight
persian_food
pay_meal_out
parents_cook
on_off_campus
nutritional_check
mother_education
marital_status
life_rewarding
italian_food
indian_food
income
ideal_diet_coded
healthy_feeling
greek_food
grade_level
fruit_day
fries
fav_food
fav_cuisine_coded
father_education
exercise
ethnic_food
employment
eating_out
eating_changes_coded1
eating_changes_coded
drink
diet_current_coded
cuisine
cook
comfort_food_reasons_coded.1
comfort_food_reasons_coded
coffee
calories_scone
calories_day
calories_chicken
breakfast
Gender
GPA
Variable
Factor Loading
Factor
calories_guess
comfort_food
demographic_information
eating_habits
food_choices
life_feeling
social_background