Math 189: Homework 7

Math 189 Homework 7

Introduction

In this assignment you will study the USDA Women's Health data set. In 1985, the USDA commissioned a study of women's nutrition. Nutrient intake was measured for a random sample of 737 women aged 25-50 years. The variables may together represent facets of *health*. In this study, we are trying to determine the more underlying factors of the data. We must first use a scree plot to determine the least number of factors that would account for the majority of the variability. In other words, we are trying to group variables that have high correlation, as they likely contain the same information. Remember, these are not known factors like Iron or Vitamin A. We then must use both PCA (Principal Component Analysis) and the Maximum Likelihood Estimation Method to analyze these latent factors. PCA sets up these factors as a combination of the variables, while MLE models the variables as functions of the factors; both quite similar, but reversed. Then, after graphing the factors, we must try to gain knowledge about the correlation between similar variables and figure out what the factors themselves are.

Data

The USDA Women's Health Survey, commissioned by the USDA in 1985, is the dataset (nutrient.txt) we are using that contains five types of women's nutrient intakes which were measured from a random sample of 737 women aged 25-50 years in the United States. The five nutrients that we will be looking at are Calcium, Iron, Protein, Vitamin A, Vitamin C. These results were compiled into a dataset (nutrient.txt) which we will use for this study. https://rdrr.io/cran/profileR/man/nutrient.html

The variables measured in the dataset are as shown below:

1. Calcium (mg)
2. Iron (mg)
3. Protein (g)
4. Vitamin A (μg)
5. Vitamin C (mg)

Tasks

Analyze the dataset according to the following steps:

1. Explore the data graphically in order to investigate the correlations between variables. Make the case for correlation to a non-technical audience by using a level plot.

```r
fpath = paste0(getwd(), "/nutrient.txt")
nutrient = read.table(fpath)
nutrient$V1 = NULL
colnames(nutrient) = c("Calcium", "Iron", "Protein", "Vitamin A", "Vitamin C")
head(nutrient)
```

```
##   Calcium   Iron Protein Vitamin A Vitamin C
## 1  522.29 10.188  42.561    349.13    54.141
## 2  343.32  4.113  67.793    266.99    24.839
## 3  858.26 13.741  59.933    667.90   155.455
## 4  575.98 13.245  42.215    792.23   224.688
## 5 1927.50 18.919 111.316    740.27    80.961
## 6  607.58  6.800  45.785    165.68    13.050
```

```
library(lattice)
library(ellipse)
```

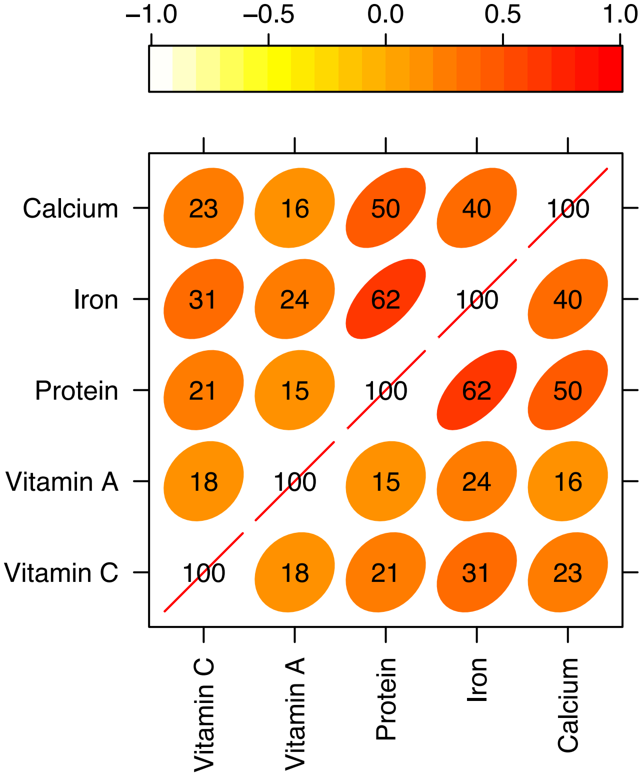
```
##
## Attaching package: 'ellipse'

## The following object is masked from 'package:graphics':
##
##     pairs
```

```
cor_nutrient <- cor(nutrient)


# Function to generate correlation plot
panel.corrgram <- function(x, y, z, subscripts, at, level = 0.9, label = FALSE, ...) {
     require("ellipse", quietly = TRUE)
     x <- as.numeric(x)[subscripts]
     y <- as.numeric(y)[subscripts]
     z <- as.numeric(z)[subscripts]
     zcol <- level.colors(z, at = at,  ...)
     for (i in seq(along = z)) {
         ell=ellipse(z[i], level = level, npoints = 50,
                     scale = c(.2, .2), centre = c(x[i], y[i]))
         panel.polygon(ell, col = zcol[i], border = zcol[i], ...)
     }
     if (label)
         panel.text(x = x, y = y, lab = 100 * round(z, 2), cex = 0.8,
                    col = ifelse(z < 0, "white", "black"))
 }

# generate correlation plot
print(levelplot(cor_nutrient[seq(5,1), seq(5,1)], at = do.breaks(c(-1.01, 1.01), 20),
           xlab = NULL, ylab = NULL, colorkey = list(space = "top"), col.regions=rev(heat.colors(100)),
           scales = list(x = list(rot = 90)),
           panel = panel.corrgram, label = TRUE))
```



It appears that there seems to be a high correlation between Iron and Protein and relatively strong correlation between calcium and protein in comparison to others.

2. Fit the factor model using both PCA and MLE, and compare the parameter estimates. Discuss the underlying assumptions for each method. Which results do you prefer, and why?

```
pca_result <- prcomp (nutrient, scale =TRUE)
eigen_val <- pca_result$sdev^2
pve <- eigen_val/sum(eigen_val)
```

```
n.factors <- 2

fa_fit <- factanal(nutrient, n.factors, rotation="varimax")
fa_fit
```

```
##
## Call:
## factanal(x = nutrient, factors = n.factors, rotation = "varimax")
##
## Uniquenesses:
##   Calcium      Iron   Protein Vitamin A Vitamin C
##     0.694     0.453     0.005     0.848     0.748
##
## Loadings:
##           Factor1 Factor2
## Calcium   0.466   0.298
## Iron      0.568   0.474
## Protein   0.989   0.131
## Vitamin A         0.378
## Vitamin C 0.151   0.479
##
##                Factor1 Factor2
## SS loadings       1.55   0.703
## Proportion Var    0.31   0.141
## Cumulative Var    0.31   0.451
##
## Test of the hypothesis that 2 factors are sufficient.
## The chi square statistic is 0.94 on 1 degree of freedom.
## The p-value is 0.331
```

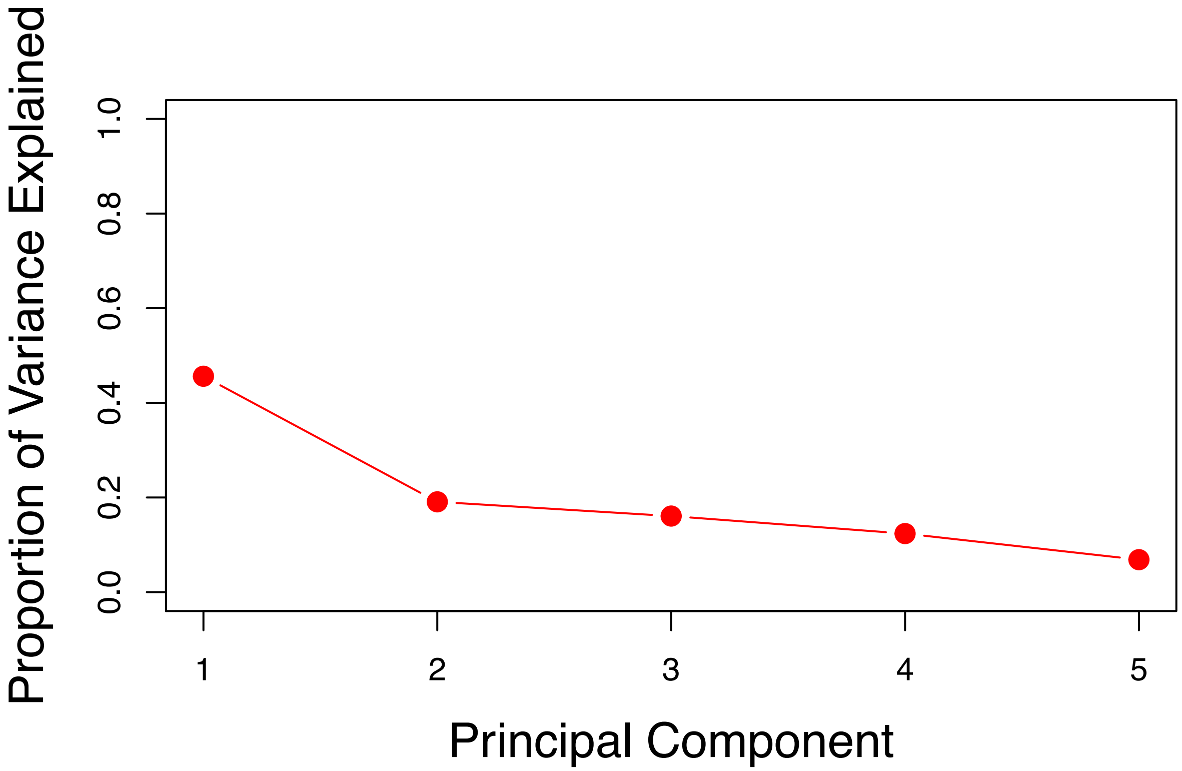
```
t(pca_result$rotation)[1:2,]
```

```
##        Calcium        Iron    Protein Vitamin A Vitamin C
## PC1  0.4725630  0.54314812  0.5370491 0.2724785 0.3449756
## PC2 -0.2658644 -0.09248598 -0.3476750 0.7825987 0.4329248
```

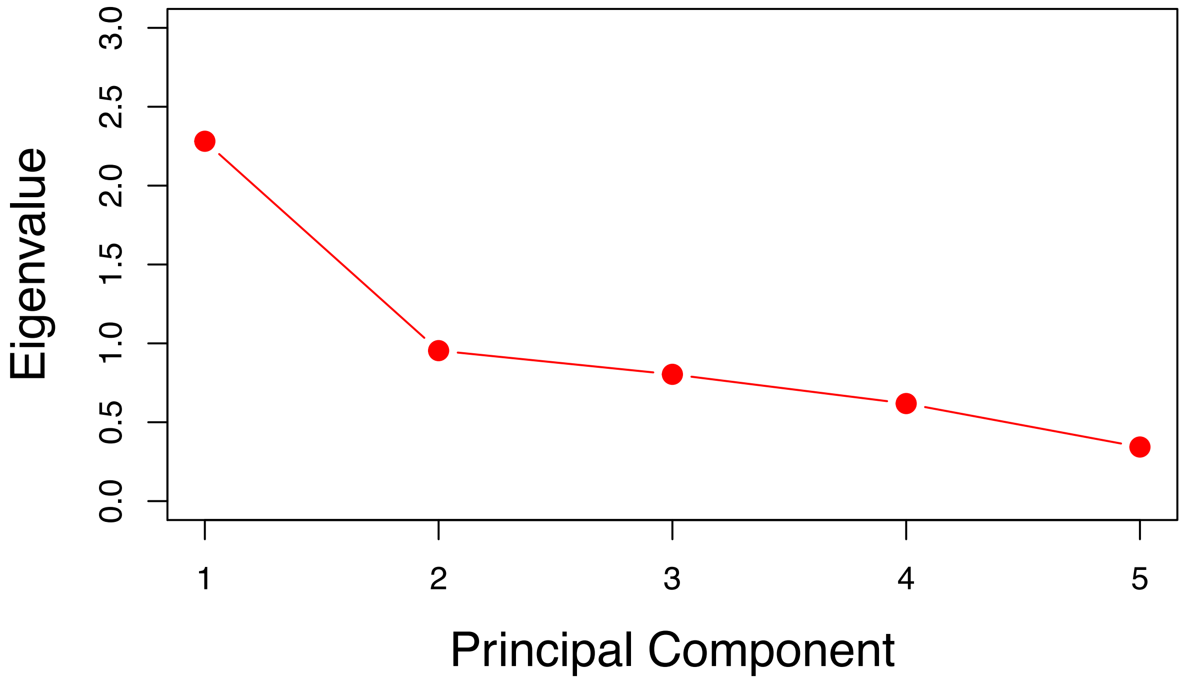
The assumptions for PCA is relatively weak, since the variables are needed to be continuous and ordinal variables might be accepted. It is also assumed that there are no significant outliers as outliers affect the covariance matrix. For the MLE method it is assumed that the data is independently sampled from a multivariate normal distribution. The method also has assumptions for mean and covariance on factors, mean and covariance on random errors, and covariance between factors and errors. We prefer using PCA over MLE because we found that PCA explains the variance better than MLE since MLE can have at most 2 factors for this dataset.

3. Use a scree plot to decide on a dimension reduction, and justify your choice.

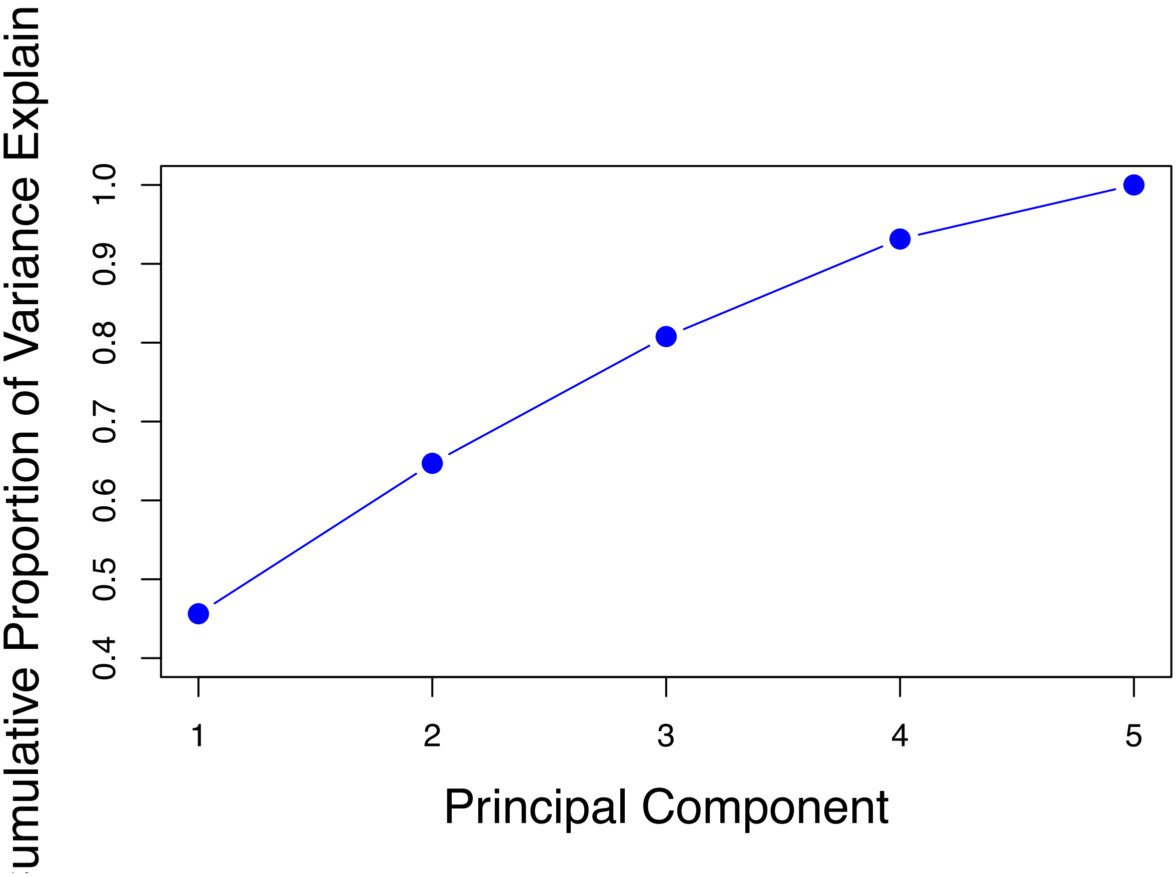
```
plot(pve, xlab=" Principal Component ",
  ylab=" Proportion of Variance Explained ",
  ylim=c(0,1), xaxt="n" ,type='b', col="red",
  cex=2,pch=20, cex.lab=1.5)
axis(1, at=c(1,2,3,4,5),labels=c(1,2,3,4,5))
```



```
#Scree plot for individual eigenvalues
plot(eigen_val  , xlab=" Principal Component ", ylab=" Eigenvalue ",
     ylim=c(0,3), xaxt="n" ,type=’b’, col="red", cex=2,
pch=20, cex.lab=1.5)
axis(1, at=c(1,2,3,4,5),labels=c(1,2,3,4,5))
```



```
#Scree plot for accumalated eigenvalues
plot(cumsum(pve), xlab=" Principal Component ",
  ylab ="Cumulative Proportion of Variance Explained ",
  ylim=c(0.4,1) , xaxt="n",type=’b’, col="blue", cex=2, pch=20, cex.lab=1.5)
axis(1, at=c(1,2,3,4,5),labels=c(1,2,3,4,5))
```



We are further going to reduce our dimension to 2 because we are able still able to explain more than 50% of the variance. Additionally, we are able to explain more proportion of variance than 1, but only slightly less than 3 dimensions or higher.

4. Examine the factor loadings, and discuss in your report which variables have high or low loadings. Can you associate an interpretation to your factors?

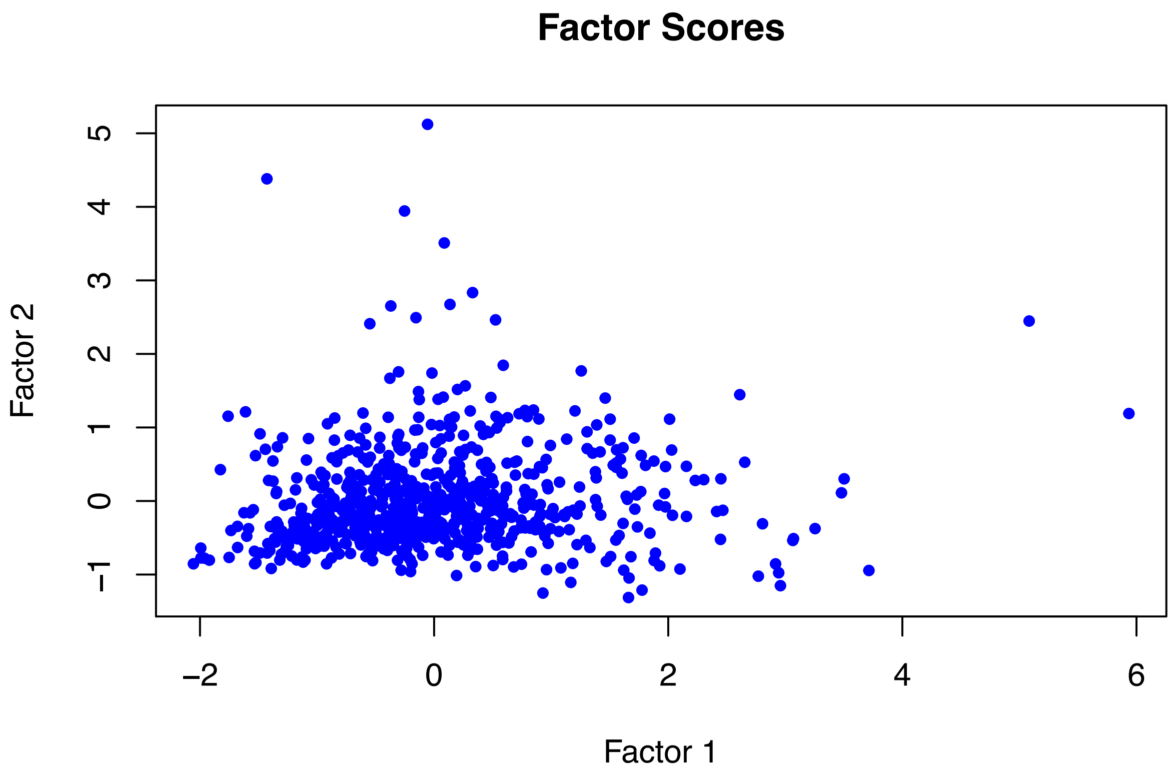
```
# Factor loadings
loading <- fa_fit$loadings[,1:2]
t(loading)
```

```
##           Calcium      Iron   Protein  Vitamin A Vitamin C
## Factor1 0.4662298 0.5675046 0.9888518 0.09839555 0.1510572
## Factor2 0.2984038 0.4743206 0.1310440 0.37773096 0.4787664
```

The first factor loads heavily on Calcium, Iron, and Protein. The second factor loads heavily on Vitamin A and Vitamin C. This gives us a good interpretation of common factors. The second factor indicates the common variation of Vitamins and the first factor indicates the common variation for Calcium, Iron and Protein. The first factor can be associated with individuals that eat meat and the second factor can be associated with individuals that are vegetarian or vegan.

5. Examine the factor scores by scatter plots or pairwise scatter plots. Is there a story to tell from these results?

```r
fac_sc = factanal(nutrient,n.factors,rotation="varimax",scores="regression")
fac_1 = fac_sc$scores[,1]
fac_2 = fac_sc$scores[,2]
plot(fac_1,fac_2,main="Factor Scores",xlab="Factor 1", ylab="Factor 2",pch=20, col="blue")
```



The graph seems to be heavily focused below factor points 2 for both factor1 and factor2 which shows that this is where the normal range is. However, there seems to be some high and unbalanced nutrient intakes based on the scatterplot which will need further attention and can be seen as outliers. Since the proportion variance explained by factor1 and factor2 is less than 50%, the graph cannot be seen as reliable.

Methods & Analysis

To investigate the correlations between variables we used a level-plot. In level-plots, ellipses mean that there is a very strong correlation and circles mean that there is a weak correlation. Upon investigating the using a level-plot, we found that there seems to be a high correlation between Iron and Protein and relatively strong correlation between calcium and protein in comparison to others. We then proceeded to fit the factor model using both PCA and MLE and compared the parameter estimates. PCA is assumed to be relatively weak, since the variables are needed to be continuous and ordinal variables might be accepted. We also assumed that there are no significant outliers as outliers affect the covariance matrix. Using the MLE method we assume that the data is independently sampled from a multivariate normal distribution. We also have assumptions for mean and covariance on factors, mean and covariance on random errors, and covariance between factors and errors. We prefer using PCA over MLE because we found that PCA explains the variance better than MLE since MLE can only use 2 factors for this dataset. We then proceeded to use a scree plot to determine the level of reduction on the dimension. We are further going to reduce our dimension to 2 because we are able still able to explain more than 50% of the variance. Additionally, we are able to explain more proportion of variance than 1, but only slightly less than 3 dimensions or higher. Upon processing the scree plots we further examined the factor loadings. The first factor loads heavily on Calcium, Iron, and Protein. The second factor loads heavily on Vitamin A and Vitamin C. This gives us a good interpretation of common factors. The second factor indicates the common variation of Vitamins and the first factor indicates the common variation for Calcium, Iron and Protein. The first factor can be associated with individuals that eat meat and the second factor can be associated with individuals that are vegetarian or vegan. We then examined the factor scores using a scatter plot. We noticed that the graph seems to be heavily focused below factor points 2 for both factor1 and factor2 which shows that this is where the normal range is. However, there seems to be some high and unbalanced nutrient intakes based on the scatterplot which will need further attention and can be seen as outliers. Since the proportion variance explained by factor1 and factor2 is less than 50%, the graph cannot be seen as reliable.

Conclusion & Results

6. Summarize your findings and try to tell a nice story with this data analysis.

From the PCA, the first loading vector places more weight on Calcium, Iron, and Protein with much less weight on Vitamin A and Vitamin C. The second loading vector places most of its weight on Vitamin A and Vitamin C and much less weight on the other three features. The MLE method produces similar results for the first loading vector but slightly varies in the second. MLE's second loading factor indicates increased weight on Vitamin A, Vitamin C, and Iron. The Calcium, Iron and Protein are located close to each other and that the Vitamin A and Vitamin C variables are far from the other three which indicates their correlations. Calcium, Iron and Protein are high in meat and poultry foods while Vitamin A and Vitamin C are high in vegetables and fruits. Women who are vegetarians and vegans are likely to have less Calcium, Iron and Protein and high Vitamin A and Vitamin C while women who eat primarily meat and poultry have comparably high Calcium, Iron and Protein and low Vitamin A and C.