Math 189: hw 7 solution

TA

March 2021

Factor Analysis of USDA Women's Health Survey

We 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. We seek a factor model, where the latent factors will explain the major nutritional features.

The task is to develop a factor model to better explain the main drivers of nutrition, and to see whether dimension reduction is possible.

Tasks

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.

Let us load necessary packages and read the data, then calculate the pairwise correlations.

```
rm(list = ls())
library(matrixStats)
library(lattice)
library(ellipse)
```

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

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

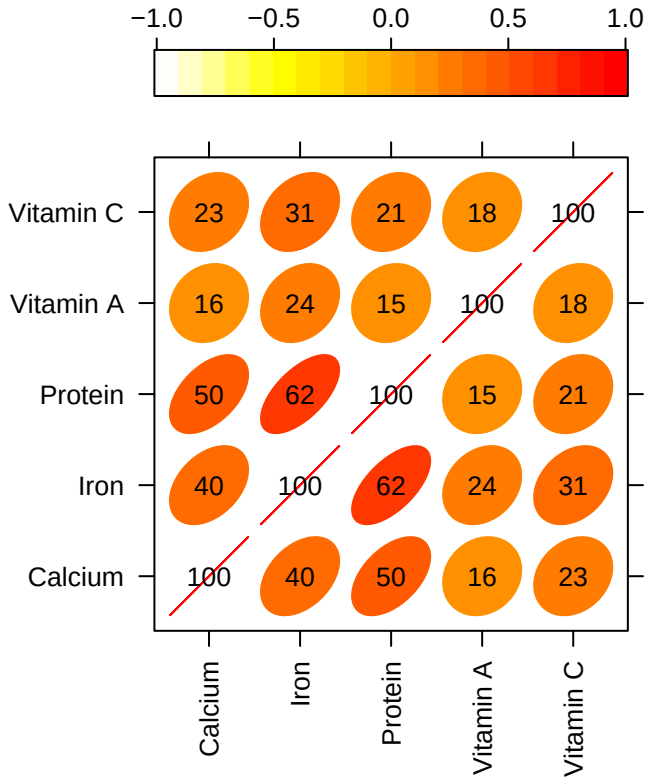
```
setwd("/Users/xopan/Documents/Courses/UCSD/Math 189/Homework/HW7")
nutrient = read.table("nutrient.txt")[, 2:6]
n = nrow(nutrient)
p = ncol(nutrient)
nutrient = scale(nutrient)
colnames(nutrient) = c("Calcium", "Iron", "Protein", "Vitamin A", "Vitamin C")
cor_df = cor(nutrient)
```

To visualize correlation using a level plot, we adopt the function **panel.corrgram** on lecture note 22.

```
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_df, 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 can be found from the level plot that Protein and Iron are highly correlated, and other variables are just weakly related.

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?

We first fit PCA, then estimate the factor model with 2 factors via MLE.

```
# PCA
pca_result <- prcomp(nutrient, scale =TRUE)
pca_result$rotation
```

```
##                 PC1         PC2         PC3         PC4         PC5
## Calcium   0.4725630 -0.26586441  0.07315734 -0.80510731 -0.22902090
## Iron      0.5431481 -0.09248598  0.04929366  0.53538298 -0.63825674
## Protein   0.5370491 -0.34767498  0.13665314  0.23826213  0.71781439
## Vitamin A 0.2724785  0.78259867  0.54601653 -0.07636022  0.09659118
```

```
## Vitamin C 0.3449756  0.43292481 -0.82183332 -0.05067110  0.12486157
```

```
eigen_val <- pca_result$sdev^2
pve <- eigen_val/sum(eigen_val)
pve
```

```
## [1] 0.45625099 0.19078083 0.16073078 0.12368272 0.06855467
```

```
# MLE for factor model
n.factors <- 2
fa_fit <- factanal(nutrient, n.factors, scores = "regression", rotation="varimax")
fa_fit$loadings
```

```
##
## 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
```

The assumptions for PCA: (1) variables are continuous; (2) there's a linear relationship between all variables; (3) sampling adequacy; (4) no significant outliers.

The assumptions for factor model: (1) factors and random errors have zero means; (2) factors have identity covariance matrix; (3) random errors have diagonal covariance matrix; (4) factors and random errors are uncorrelated.

Grading: The assumptions for PCA are not included in lecture, so don't deduct points for this, but the assumptions for factor model should be mentioned.

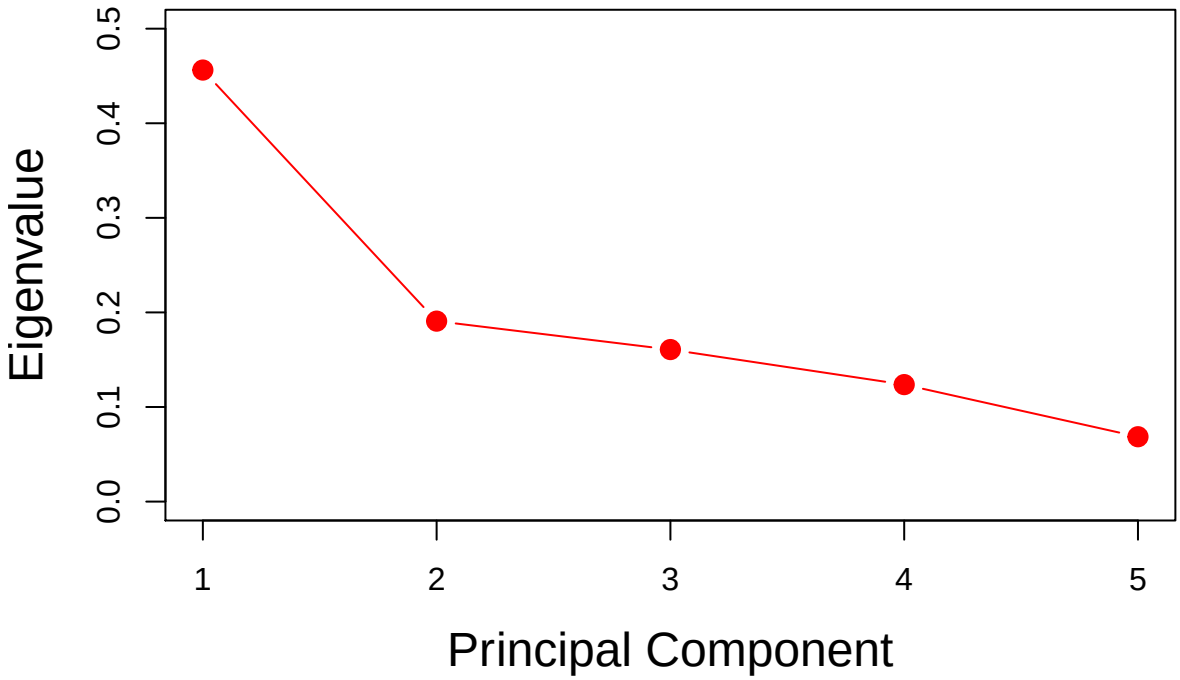
We prefer PCA. With 2 factors, the factor model can only explain 45% of the total variance, but with 2 principal components, we are able to cover 65% variance, and 3 principal components can explain 81% variance. PCA gives us more information about the original data, and is more suitable for further analysis.

Students can also choose a factor model, if their reasons make sense.

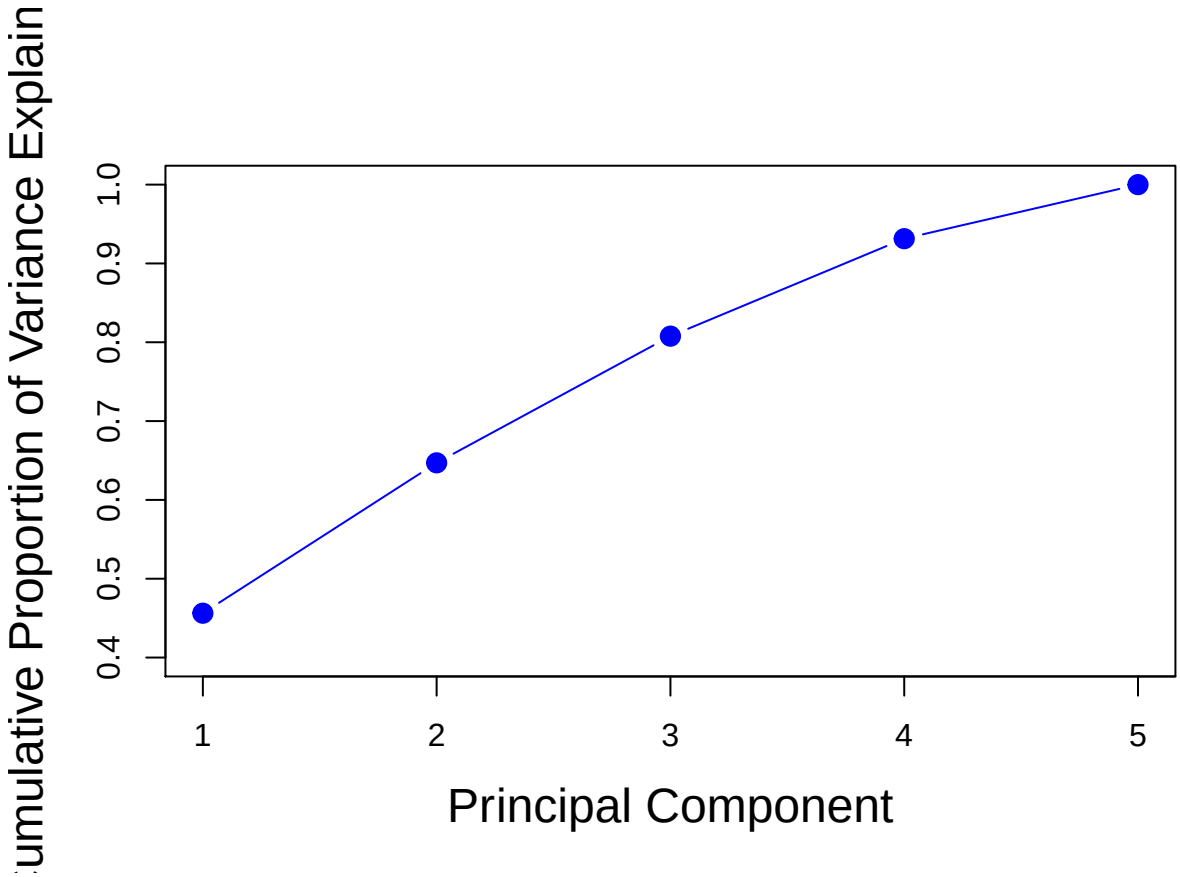
3. Use a scree plot to decide on a dimension reduction, and justify your choice.

We draw scree plots for individual eigenvalues and accumulative eigenvalues as below.

```
plot(pve, xlab=" Principal Component ", ylab=" Eigenvalue ", ylim=c(0,0.5), 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))
```




```
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(1,2,3,4,5,6),labels=c(1,2,3,4,5,6))
```



We believe a good model should be able to explain at least 50% variance, and the desired dimension cannot be too close to the original dimension. Based on these criteria, the appropriate number of components should be either 2 or 3.

Grading: If the student chooses the factor model, then the number of factor should be 2. A model with only 1 factor is not appropriate.

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?

```
loading <- fa_fit$loadings[,1:2]
loading
```

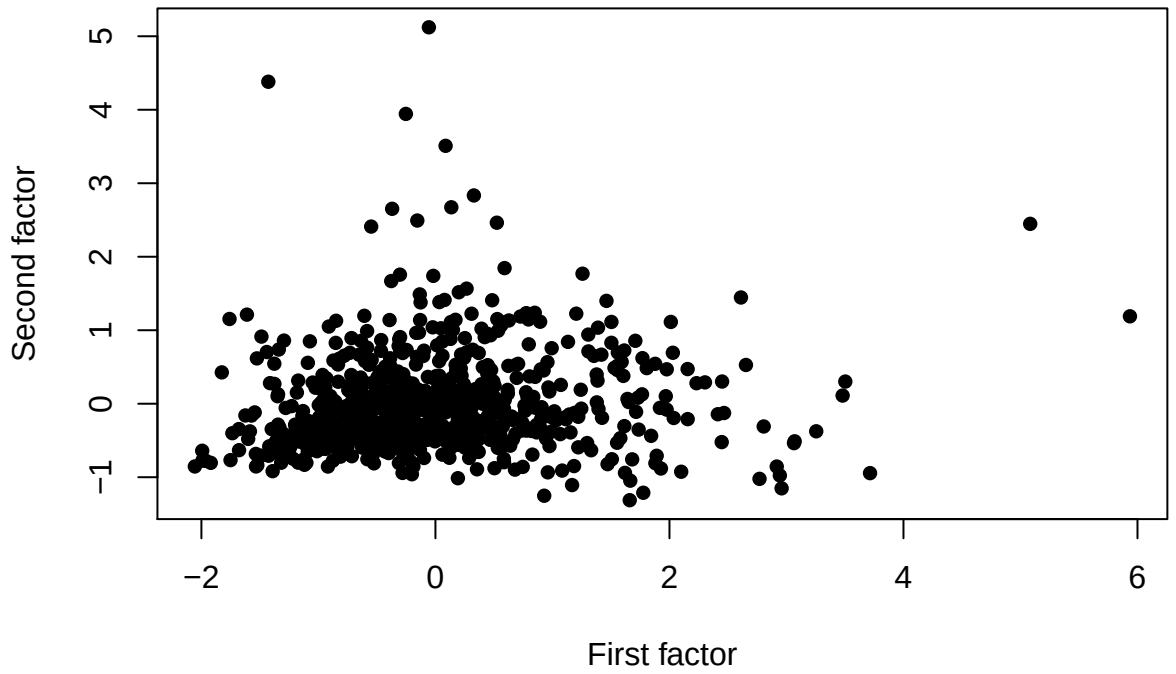
```
##              Factor1   Factor2
## Calcium   0.46622983 0.2984038
## Iron      0.56750463 0.4743206
## Protein   0.98885175 0.1310440
## Vitamin A 0.09839555 0.3777310
## Vitamin C 0.15105719 0.4787664
```

The first factor assigns high loading to Calcium, Iron and Protein, and the second factor puts more weights on Iron, Vitamin A and Vitamin C. Maybe the first factor is associated with meat, and the second factor is about vegetable or fruit.

Grading: They should get credits as long as their stories make sense.

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

```
score = fa_fit$scores
plot(score[, 1], score[, 2], pch = 16, xlab="First factor", ylab="Second factor")
```



Based on the scatter plot of factor scores, most points are grouped in the left-bottom region, and we assume this is a normal region for these two factors. We detect extreme values with either high intakes for one of the factors, or unbalanced intakes, and we suggest a closer look at these individuals for health-related reasons.

However, we should also emphasize that the two-factor model can only explain about 45% variance of the

original data. This dimension reduction approach via factor model is not very reliable due to the miss of information.

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

We apply two dimension reduction techniques to the data: PCA and factor analysis, and we are able to reduce the dimensionality from 5 to 2 (or 3). Through a rough analysis based on factor loadings, we detect two factors related to (1) Calcium, Iron and Protein, and (2) Iron, Vitamin A and Vitamin C accordingly. The interpretation of these factors may be an interesting topic in nutritional science.

The scale of the dataset is fairly small, and it's desirable to include more nutritional features for a comprehensive study.

Grading: The interpratation of factor model is a hard question, so be lenient on this part.