Math 189 Homework 7: Factor Analysis of USDA Women's Health Survey

Yunchun Pan, Siqing Lyu, Nathan Ng, Pudan Xu

3/2/2021

Introduction

In this homework, we examine the USDA Women's Health Survey dataset **nutrient.txt**[1]. This dataset 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. There are 5 nutrients in the dataset: **Calcium (mg)**, **Iron (mg)**, **Protein (g)**, **Vitamin A (ug)**, and **Vitamin C (mg)**. These variables may represent facets of *health*.

In order to analyze the dataset, we first explore the data using a level plot to investigate the correlations between 5 nutrient variables. Secondly, we fit the factor model using both PCA and MLE, discuss the underlying assumptions for each method, and compare the proportion and the cumulative proportion of variance explained by PCs and common factors found by MLE. In addition, we use scree plots to visually show these reported proportions. Based on the results of comparison, we decide on which method to use and how many factors to include. Next, we observe the factor loading of each variable and interpret what it means for a variable to have a high or low loading. In the end, we use a scatter plot to show how the data is distributed in terms of the factors and examine the factor scores.

Our Work

We first load the USDA Women's Health Survey dataset into R to get the data we use throughout this assignment and save it as **nutrient**.

```
nutrient <- read.table("../Data/nutrient.txt", quote="\"", comment.char="")
nutrient$V1=NULL
colnames(nutrient)=c("Calcium (mg)", "Iron (mg)", "Protein (g)",
                     "Vitamin A (ug)", "Vitamin C (mg)")
head(nutrient)
```

```
##   Calcium (mg) Iron (mg) Protein (g) Vitamin A (ug) Vitamin C (mg)
## 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
```

Then, we calculate the correlation matrix **cor_nutrient** of **nutrient** and use a level plot to visualize **cor_nutrient** and investigate the correlations between 5 nutrient variables. If the color of an entry is close

[1]Source: The USDA Women's Health Survey dataset contains data extracted from the 1985 study commissioned by the USDA on women's nutrition.

to red and the entry has a oval shape, then correlation level between the corresponding two variables is close to 1, which means that two variables are strongly positively correlated. If the color of an entry is close to light orange and it has a round shape, then correlation level between the corresponding two variables is close to 0, which means that two nutrient variables are not really correlated.

```
# Load the library
library("lattice")
library("ellipse")
```

```
## Warning: package 'ellipse' was built under R version 3.6.3
```

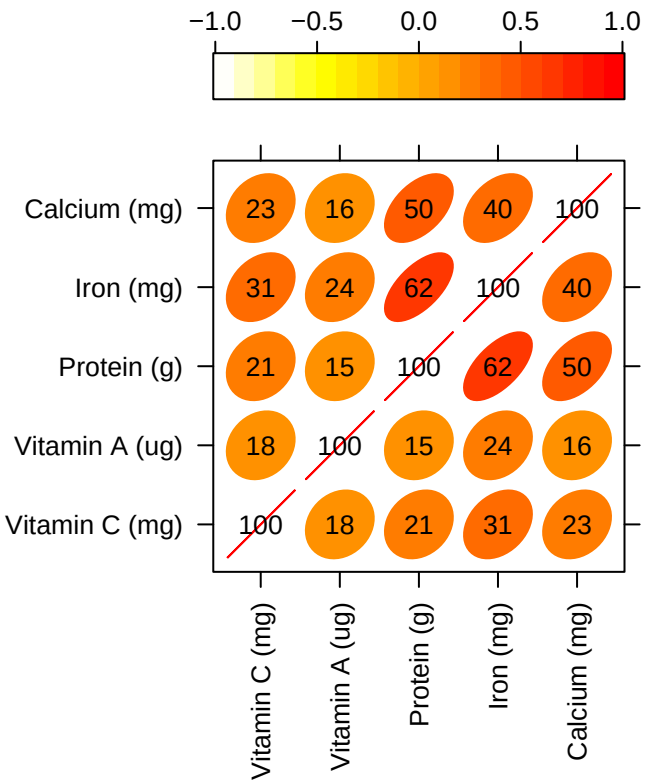
```
##
## Attaching package: 'ellipse'
```

```
## 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))
```



In the level plot above, the entries along the diagonal line have red colors and their shapes are straight lines, which suggests the expected result that the correlation of each nutrient variable with itself is always 1. Moreover, the correlation entry of Iron and Protein has a color that is closer to red and a more oval shape than any other entry. Hence, among all pairs of nutrient variables, Iron and Protein have the strongest positive correlation. What's more, positive correlations between the pair of Protein and Calcium and another pair of Iron and Calcium are fairly strong since corresponding entries also have colors close to red and oval shapes. Hence, Calcium, Iron, and Protein are strongly positively correlated, while all other pairs of nutrients are not very strongly correlated.

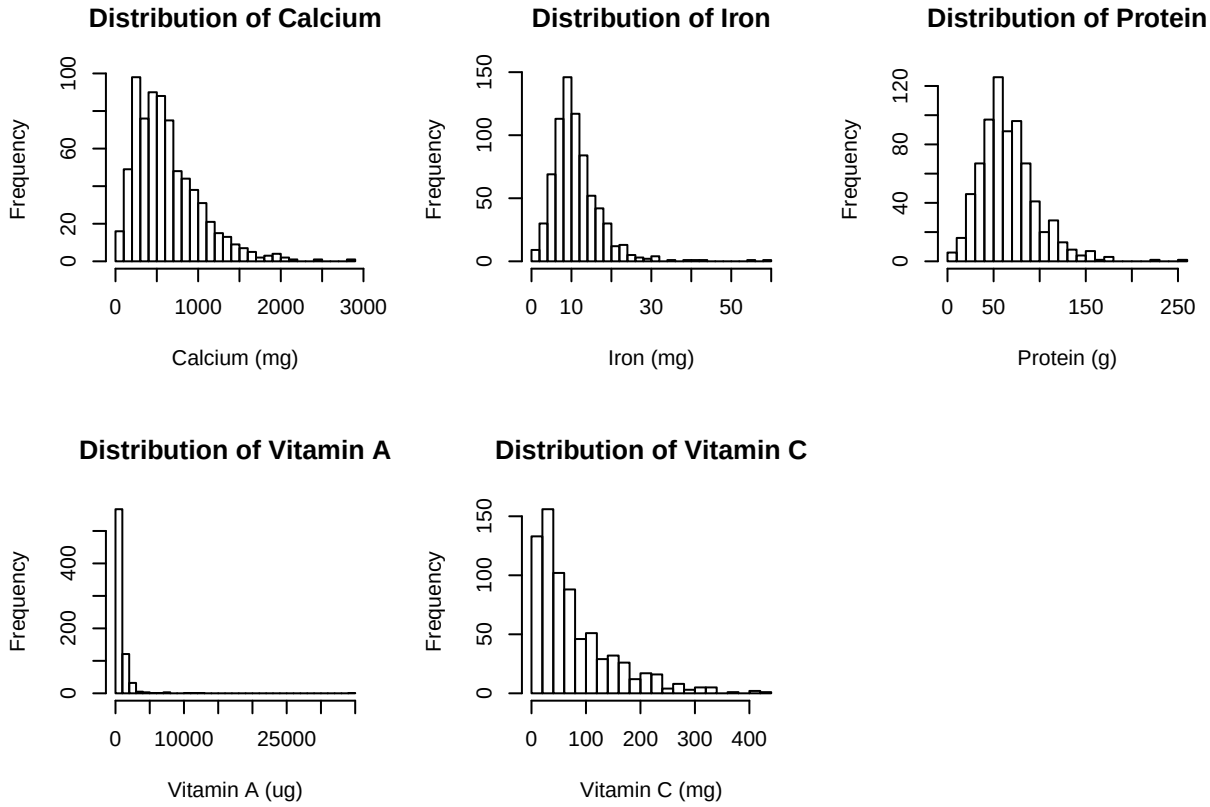
After investigating the correlations between 5 nutrient variables, we proceed to fit the factor model using both Principal Component Analysis (PCA) and Maximum Likelihood Estimation (MLE).

In order to perform PCA, we need to meet the following assumptions. The first assumption is that all variables are continuous. The second assumption is that there are no significant outliers in the data, since they would greatly affect the covariance matrix used in the PCA. If any of these two assumptions does not hold, then the results of PCA is not reliable. In this homework, we assume that these two assumptions are met.

In order to perform MLE, we also need to meet certain assumptions so that the results of MLE is reliable. The first assumption is that observations in the data are independently and identically distributed. The second assumption is that the data are sampled from a multivariate normal distribution, which means that all variables are normally distributed. For simplicity, we assume that first assumption is met. To determine if assumption of normality holds for our data, we use the histograms below to visualize the distribution of each nutrient.

```
par(mfrow=c(2, 3))
hist(nutrient$`Calcium (mg)`, main="Distribution of Calcium",
     xlab="Calcium (mg)", breaks=25)
```

```
hist(nutrient$`Iron (mg)`, main="Distribution of Iron",
     xlab="Iron (mg)", breaks=25)
hist(nutrient$`Protein (g)`, main="Distribution of Protein",
     xlab="Protein (g)", breaks=25)
hist(nutrient$`Vitamin A (ug)`, main="Distribution of Vitamin A",
     xlab="Vitamin A (ug)", breaks=25)
hist(nutrient$`Vitamin C (mg)`, main="Distribution of Vitamin C",
     xlab="Vitamin C (mg)", breaks=25)
```



In the histograms shown above, Calcium, Iron, and Protein are roughly normally distributed, but each of their distributions is slightly skewed to the right. However, Vitamin A and Vitamin C have very right skewed distributions. Since distributions of Vitamin A and Vitamin C are clearly not normally distributed, the results of MLE might not be reliable and accurate.

We start with PCA method. Here, we calculate sample means and sample variances of 5 nutrient variables and save the results to **out1**. As the table below shows, sample mean and sample standard deviation differ largely from one nutrient consumption to another. But it is also important to note that intakes of these nutrients are recorded using different measurements, which are grams, milligrams, and micrograms.

```
out1 <- rbind(apply(nutrient, 2, mean),
              apply(nutrient, 2, var))
rownames(out1) <- c("Sample Mean","Sample Variance")
out1
```

```
##                Calcium (mg) Iron (mg) Protein (g) Vitamin A (ug)
## Sample Mean         624.0493  11.12990    65.80344       839.6353
```

```
## Sample Variance  157829.4439  35.81054   934.87688   2668452.3706
##                 Vitamin C (mg)
## Sample Mean           78.92845
## Sample Variance     5416.26408
```

As we can see, the women observed consume protein the most out of all 5 nutrients, with an average intake of 65.803 grams. Calcium is the second most consumed nutrient with an average intake of 624.049 milligrams. Vitamin C is the third most consumed nutrient with an average intake of 78.928 milligrams, while Iron is the fourth most consumed nutrient with an average intake of 11.130 milligrams. Vitamin A has the least average consumption of 839.635 micrograms by the women, but Vitamin A consumption has a significantly large variance of 2668452, or a standard deviation of 1633.5398 micrograms, which is much larger than the sample mean. Vitamin C consumption also has a fairly large variance of 5416.264, or a standard deviation of 73.595 milligrams. Calcium also has a fairly large variance of 157829.4439 and standard deviation of 397.277 milligrams, as well as Protein with a variance of 934.876 and standard deviation of 30.577 grams. Iron has the least variance of 35.810 and standard deviation of 5.984 milligrams. Vitamins seem to have very large spreads compared to their sample means, which can be shown from heavy right skews we see in histograms above.

Next, we standardize each nutrient variable to have mean 0 and standard deviation 1 and save the standardized data as **zscale_nutrient**. Then, we calculate the sample covariance matrix of **zscale_nutrient** and find the eigenvalues **pca_var** of the covariance matrix. We also report the proportion of total variance explained **pve** by each Principal Component (PC) and report the cumulative proportion explained.

```
ncols = dim(nutrient)[2]
zscale_nutrient <- nutrient
for (i in 1:ncols) {
  zscale_nutrient[,i] <- (nutrient[,i] - mean(nutrient[,i]))/sd(nutrient[,i])
}

pca_result <- eigen(cov(zscale_nutrient))
pca_var <- pca_result$values
pve <- pca_var/sum(pca_var)
out2 <- cbind(pca_var,pve,cumsum(pve))
colnames(out2) <- c("Eigenvalue","Proportion","Cumulative")
rownames(out2) <- c("PC1","PC2","PC3","PC4","PC5")
```

Then, we fit the factor model using MLE method. We first chose $m = 3$ as the number of common factors, but **factanal** function gives an error that 3 factors are too many for 5 variables. Therefore, we set the number of factors as 2 and enter the raw data of **nutrient** to estimate the factor model with MLE method. The estimated factor loadings are saved into variable **loading**.

```
# The n.factors denotes how many factors used in fitting
# Set the number of factors as 2
n.factors <- 2

# Fit factor model with MLE methods
# The rotation option denotes what rotation matrix is used
fa_fit <- factanal(nutrient, n.factors, rotation="varimax",scores = "regression")

# Factor loadings
loading <- fa_fit$loadings
```

Since we set the number of factors as 2 for the factor model with MLE method, we focus on the proportion of total variance explained by each of first 2 PCs and the cumulative proportion explained by first 2 PCs. Here, we display the results from factor models with PCA and MLE and compare them.

```r
# Result from PCA
out2
```

```
##     Eigenvalue Proportion Cumulative
## PC1  2.2812550 0.45625099  0.4562510
## PC2  0.9539042 0.19078083  0.6470318
## PC3  0.8036539 0.16073078  0.8077626
## PC4  0.6184136 0.12368272  0.9314453
## PC5  0.3427734 0.06855467  1.0000000
```

```r
# Result from MLE
loading
```

```
##
## Loadings:
##                 Factor1 Factor2
## Calcium (mg)    0.466   0.298
## Iron (mg)       0.568   0.474
## Protein (g)     0.989   0.131
## Vitamin A (ug)          0.378
## Vitamin C (mg)  0.151   0.479
##
##                 Factor1 Factor2
## SS loadings       1.55   0.703
## Proportion Var    0.31   0.141
## Cumulative Var    0.31   0.451
```

By observing two tables above, we find that the proportion of variance explained by PC1 and PC2 are 45.625099% and 19.078083% respectively, which are higher than 31% and 14.1% variance explained by factor 1 and 2 from the factor model with MLE. Since each of first 2 PCs explains more proportion of variance than each factor in the model with MLE, the cumulative proportion of variance explained by PC1 and PC2 is 64.70318% and higher than 45.1% variance explained by both two factors in MLE method. Based on the finding that first 2 PCs can explain more proportion of variance than 2 factors of MLE method, we prefer to use PCA for making the factor model. But, it is important to note that this result of MLE might not be reliable since the assumption of normality does not hold for our data.

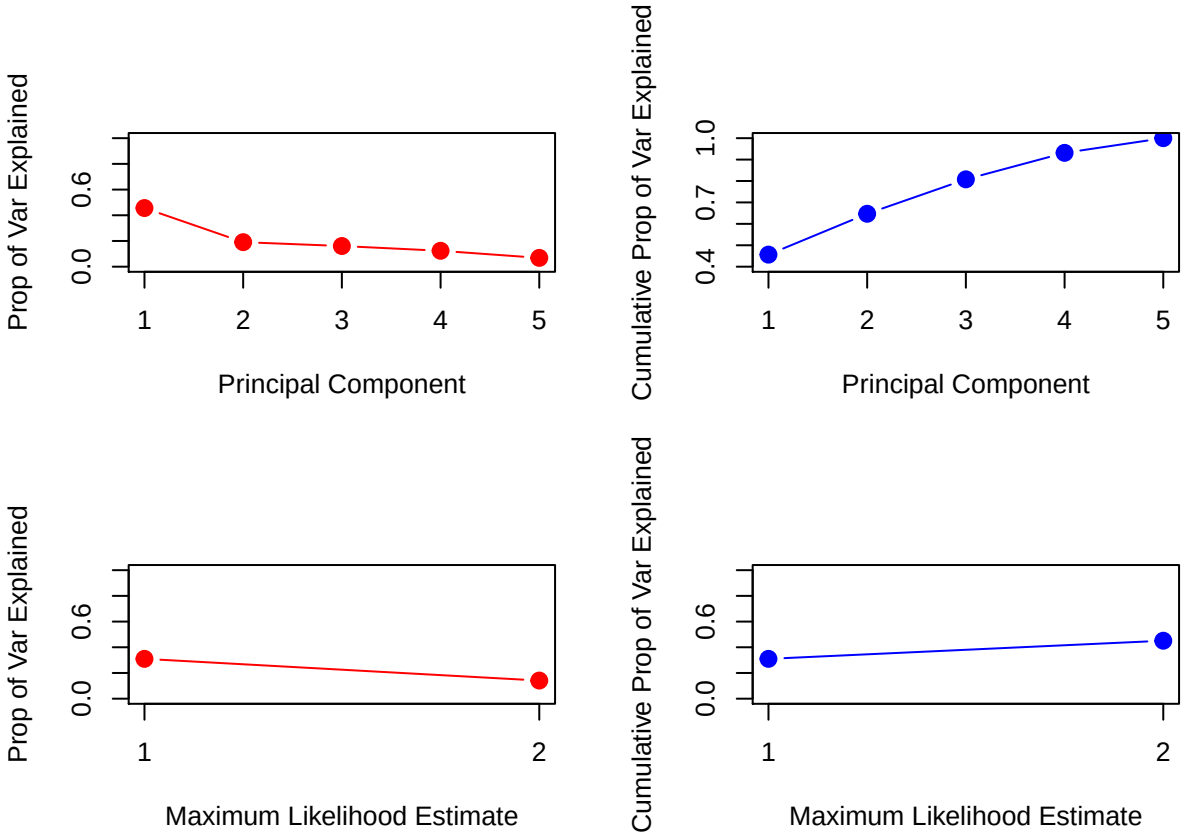
Here, we create scree plots to visualize the proportion of variance explained by each of 5 PCs and each common factor found by MLE method, as well as the cumulative proportions of variance explained by them.

```r
#Scree plot
par(mfrow=c(2,2))
plot(pve, xlab="Principal Component",
  ylab="Prop of Var Explained",
  ylim=c(0,1), xaxt="n" ,type='b', col="red",
  cex=2,pch=20, cex.lab=1)
axis(1, at=c(1,2,3,4,5),labels=c(1,2,3,4,5))

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

```r
plot(c(.31,.141), xlab="Maximum Likelihood Estimate",
  ylab="Prop of Var Explained",
  ylim=c(0,1), xaxt="n" ,type='b', col="red",
  cex=2,pch=20, cex.lab=1)
axis(1, at=c(1,2),labels=c(1,2))

# Cumulative Proportion plot
plot(c(0.31, 0.451), xlab="Maximum Likelihood Estimate",
  ylab ="Cumulative Prop of Var Explained",
  ylim=c(0,1) , xaxt="n",type='b', col="blue", cex=2, pch=20, cex.lab=1)
axis(1, at=c(1,2),labels=c(1,2))
```



Those plots again show that the proportion of variance explained by each of first 2 PCs is higher than the proportion explained by each factor found by MLE method. Also, the cumulative proportion of variance explained by PC1 and PC2 appears visually to be higher than the proportion explained by both two factors found by MLE method. So, we prefer to fit the factor model using PCA. Since most variance, approximately 80% variance, can be explained by first 3 PCs and we want to include as few PCs as possible to explain as much variance as possible, we decide to fit the factor model using PCA and only include first 3 PCs.

After comparing the proportions of variance explained by first 2 PCs in PCA and two factors found by MLE, we take a closer look at the loading of every individual variable in each factor.

```r
fa_fit$loadings
```

```
##
## Loadings:
```

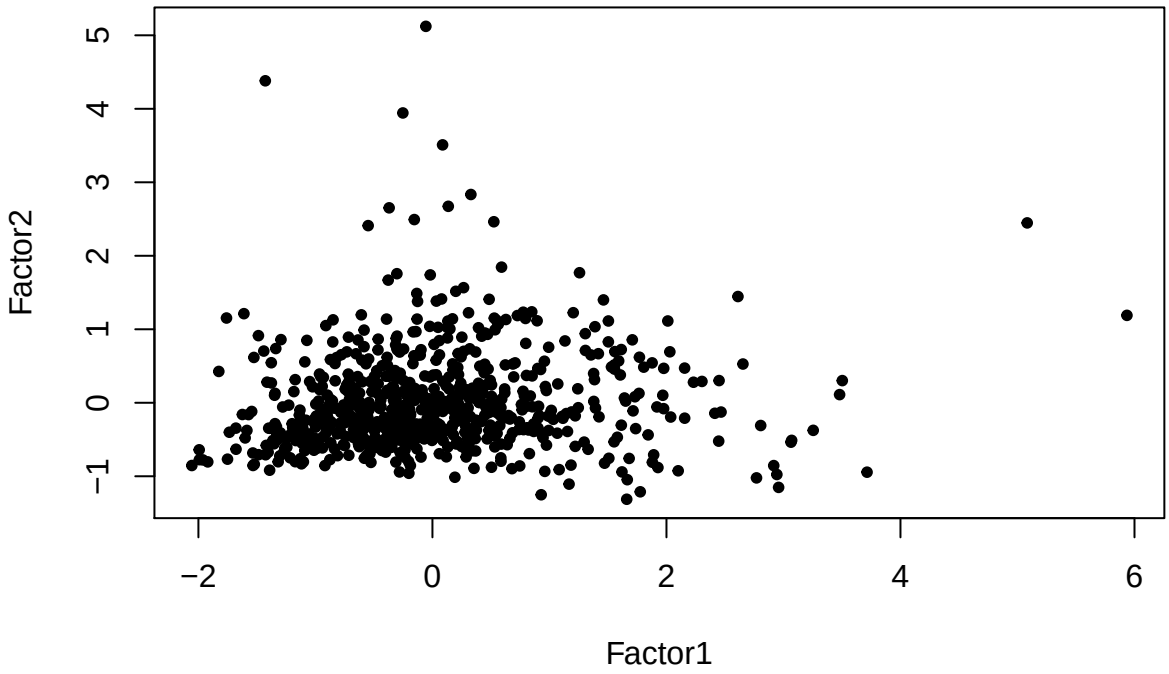
```
##                Factor1 Factor2
## Calcium (mg)   0.466   0.298
## Iron (mg)      0.568   0.474
## Protein (g)    0.989   0.131
## Vitamin A (ug)         0.378
## Vitamin C (mg) 0.151   0.479
##
##                Factor1 Factor2
## SS loadings       1.55   0.703
## Proportion Var    0.31   0.141
## Cumulative Var    0.31   0.451
```

In the first factor, Protein has the highest loading of 0.989, and Iron and Calcium have fairly high loadings of 0.568 and 0.466 respectively. However, Vitamin A and Vitamin C have very low loadings. Their loadings indicate that the first nutritional factor mainly determines Protein intake and partially determines the intake of Iron and Calcium, and it does not strongly determine the intake of Vitamins. In the second factor, both Iron and Vitamin C have the highest loadings of 0.474 and 0.479, and Vitamin A also has a fairly high loading of 0.378. In contrast, Calcium and Protein have low loadings of 0.298 and 0.131 respectively. Therefore, the second factor mainly determines the intake of Iron and Vitamin C and partially determines the intake of Vitamin A, and it does not significantly determine the intake of Calcium and Protein.

To sum up, there seems to be an underlying factor that mainly determines the intake of Protein and partially determines the intake of Calcium and Iron. There also seems to be an underlying factor that determines the intake of Vitamins and Iron. Analysis of factor loadings suggests that there are food groups or diets that are rich in Protein, Iron, and Calcium, while there are other food groups or diets that are rich in Vitamins and Iron. However, since the assumption of normality does not hold for our data, the analysis and the result above might not be reliable.

Now, we make a scatter plot to examine the factor scores. The scatterplot below shows how the data is distributed in terms of the factors.

```
score_mle <- fa_fit$scores
plot(score_mle, pch=20)
```



There does not seem to be any clear separate clusters that could indicate any pattern. Instead, we observe a large cluster mainly between -2 and 2 in the first factor and 1 to -1 in the second factor. There are also outliers scattered in the graph. Some outliers have a large factor 1 and a small factor 2, while other outliers have a large factor 2 and a small factor 1. These outliers might reflect significantly unbalanced nutrient intakes. For the outliers that have a large factor 1 and a small factor 2, the women observed might consume a significant amount of Protein, a high amount of Iron and Calcium, but very little Vitamins. For the outliers that have a small factor 1 and a large factor 2, the women observed might eat a large amount of Vitamins and Iron, but do not eat much Protein or Calcium. We may need to look closer into these outliers to better understand why they are outliers and determine if they are significant or meaningful.

However, this graph may not be a reliable representation of the actual observed data. First, the assumption of normality does not hold for our data, hence the analysis and the graph above might not be reliable. Moreover, since the cumulative proportion of the variance explained by factor model using MLE method is only 0.451, less than half of the data's variance is explained by MLE method, which could lead us to unreliable results. Instead, a model that explains a majority of the data's variability would be better for reflecting the actual data, such as PCA.

Conclusion

First, we explore the data graphically in order to investigate the correlations between 5 nutrient variables, Calcium, Iron, Protein, Vitamin A, and Vitamin C by using a level plot. According to the result, the correlation entry of Iron and Protein has a color that is closer to red and a more oval shape than any other entry. Hence, among all pairs of nutrient variables, Iron and Protein have the strongest positive correlation. What's more, positive correlations between the pair of Protein and Calcium and another pair of Iron and Calcium are fairly strong since corresponding entries also have colors close to red and oval shapes. Therefore, Calcium, Iron, and Protein are strongly positively correlated, while all other pairs of nutrients are not very strongly correlated.

Next, we proceed to fit the factor model using both Principal Component Analysis (PCA) and Maximum Likelihood Estimation (MLE). There are two assumptions that need to be met so that results of PCA are reliable. The first assumption is that all variables are continuous. The second assumption is that there are no significant outliers in the data, since they would greatly affect the covariance matrix used in the PCA. For simplicity, we just assume that they hold for our data. There are also two assumptions that need to be checked for MLE. The first assumption is that observations in the data are independently and identically distributed. The second assumption is that the data are sampled from a multivariate normal distribution, which means that all variables are normally distributed. For simplicity, we assume that the first assumption is met. To check the second assumption, we visualize the distribution of 5 nutrients. From the plots, we find that Calcium, Iron, and Protein are roughly normally distributed, but each of their distributions is slightly skewed to the right. However, since distributions of Vitamin A and Vitamin C are clearly not normally distributed, the second assumption does not hold for our data and the results of MLE might not be reliable and accurate.

To fit model using PCA, we calculate sample means and sample variances of 5 nutrient variable. The women observed consume protein the most out of all 5 nutrients. Calcium is the second most consumed, and Vitamin C is the third most consumed, while Iron is the fourth most consumed. Vitamin A has the least average consumption by the women. Also, we standardize each nutrient variable to have mean 0 and standard deviation 1, and then calculate the sample covariance matrix and find the eigenvalues of the covariance matrix. Next, we fit the factor model using MLE method, setting the number of factors as 2. By comparing the proportion of variance explained by each of first 2 PCs with proportion explained by each common factor found by MLE, we find that the proportion of variance explained by each of first 2 PCs is higher than that explained by factor 1 and 2 found by MLE. Therefore, the cumulative proportion of variance explained by PC1 and PC2 is also higher than that explained by both two factors in MLE method. Based on this finding, we prefer to use PCA for making the factor model. But, it is important to note that the result of MLE might not be reliable since the assumption of normality does not hold for our data.

Then, we create scree plots to visualize the proportion of variance explained by each of 5 PCs and each common factor found by MLE method, as well as the cumulative proportions of variance explained by them. The plots again show that the proportion of variance explained by each of first 2 PCs is higher than the proportion explained by each factor found by MLE method. Also, the cumulative proportion of variance explained by PC1 and PC2 appears visually to be higher than the proportion explained by both two factors found by MLE method. So, we prefer to fit the factor model using PCA. Since most variance, approximately 80% variance, can be explained by first 3 PCs and we want to include as few PCs as possible to explain as much variance as possible, we decide to fit the factor model using PCA and only include first 3 PCs.

Next, we take a closer look at the loading of every individual variable in each factor found by MLE. In the first factor, Protein has the highest loading, and Iron and Calcium have fairly high loadings, while Vitamin A and Vitamin C have very low loadings. Their loadings indicate that the first nutritional factor mainly determines Protein intake and partially determines the intake of Iron and Calcium, and it does not strongly determine the intake of Vitamins. In the second factor found by MLE, both Iron and Vitamin C have the highest loadings, and Vitamin A also has a fairly high loading, while Calcium and Protein have low loadings. So, the second factor mainly determines the intake of Iron and Vitamin C and partially determines the intake of Vitamin A, and it does not significantly determine the intake of Calcium and Protein.

In the end, we make a scatter plot to examine the factor scores. We observe a large cluster mainly between -2 and 2 in the first factor and 1 to -1 in the second factor. There are also outliers scattered in the graph. For the outliers that have a large factor 1 and a small factor 2, the women observed might consume a significant amount of Protein, a high amount of Iron and Calcium, but very little Vitamins. For the outliers that have a small factor 1 and a large factor 2, the women observed might eat a large amount of Vitamins and Iron, but do not eat much Protein or Calcium. However, this graph may not be a reliable representation of the actual observed data. First, the assumption of normality does not hold for our data. Moreover, since the cumulative proportion of the variance explained by two factors found by MLE method is less than 50%, this could lead us to unreliable results. Instead, a model that explains a majority of the data's variability would be better for reflecting the actual data, such as PCA.