Homework 2

Arturo Ortiz

2023-11-06

# Problem #1

The first and second component account for about 76% of the variance. Line 5 and 6 contribute the most for comp 1 and line 2 and 3 the most for comp 2

crim\_corr <- matrix(c(1.000, 0.402, 0.396, 0.301, 0.305, 0.339, 0.340, 0.402, 1.000, 0.618, 0.150, 0.135, 0.206, 0.183, 0.396, 0.618, 1.000, 0.321, 0.289, 0.363, 0.345, 0.301, 0.150, 0.321, 1.000, 0.846, 0.759, 0.661, 0.305, 0.135, 0.289, 0.846, 1.000, 0.797, 0.800, 0.339, 0.206, 0.363, 0.759, 0.797, 1.000, 0.736, 0.340, 0.183, 0.345, 0.661, 0.800, 0.736, 1.000), nrow = 7, byrow = TRUE)  
  
crim\_corr.pca <- princomp(covmat = crim\_corr)  
summary(crim\_corr.pca, loadings = TRUE)

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## Standard deviation 1.9492241 1.2256950 0.80610632 0.6000474 0.58237656  
## Proportion of Variance 0.5427821 0.2146183 0.09282963 0.0514367 0.04845178  
## Cumulative Proportion 0.5427821 0.7574004 0.85023003 0.9016667 0.95011851  
## Comp.6 Comp.7  
## Standard deviation 0.48502898 0.33751644  
## Proportion of Variance 0.03360759 0.01627391  
## Cumulative Proportion 0.98372609 1.00000000  
##   
## Loadings:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7  
## [1,] 0.276 0.365 0.882   
## [2,] 0.212 0.639 -0.258 -0.687   
## [3,] 0.295 0.512 -0.381 0.699 0.101   
## [4,] 0.438 -0.235 -0.102 0.619 0.318 -0.503  
## [5,] 0.456 -0.277 -0.113 0.290 0.785  
## [6,] 0.450 -0.178 -0.870   
## [7,] 0.436 -0.180 -0.770 0.233 -0.353

# Problem #2

urlremote <- urlRemote <- "https://raw.githubusercontent.com/"  
pathGithub <- pathGithub <- "EricBrownTTU/ISQS6350/main/"  
filename <- "protein.csv"  
protein <- read.csv(paste0(urlremote, pathGithub, filename))  
protein <- protein[,1:9]  
protein[,1:4]

## Country Red\_Meat White\_Meat Eggs  
## 1 Albania 10 1 1  
## 2 Austria 9 14 4  
## 3 Belgium 14 9 4  
## 4 Bulgaria 8 6 2  
## 5 Czechoslovakia 10 11 3  
## 6 Denmark 11 11 4  
## 7 East\_Germany 8 12 4  
## 8 Finland 10 5 3  
## 9 France 18 10 3  
## 10 Greece 10 3 3  
## 11 Hungary 5 12 3  
## 12 Ireland 14 10 5  
## 13 Italy 9 5 3  
## 14 The\_Netherlands 10 14 4  
## 15 Norway 9 5 3  
## 16 Poland 7 10 3  
## 17 Portugal 6 4 1  
## 18 Romania 6 6 2  
## 19 Spain 7 3 3  
## 20 Sweden 10 8 4  
## 21 Switzerland 13 10 3  
## 22 United\_Kingdom 17 6 5  
## 23 USSR 9 5 2  
## 24 West\_Germany 11 13 4  
## 25 Yugoslavia 4 5 1

## A

protein\_corr <- cor(protein[,-1])  
protein\_corr.pca <- princomp(protein[-1], cor = TRUE)  
summary(protein\_corr.pca, loadings = TRUE)

## Importance of components:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  
## Standard deviation 2.0123730 1.1931703 1.0200644 0.71777525 0.61533610  
## Proportion of Variance 0.5062057 0.1779569 0.1300664 0.06440016 0.04732981  
## Cumulative Proportion 0.5062057 0.6841626 0.8142290 0.87862917 0.92595899  
## Comp.6 Comp.7 Comp.8  
## Standard deviation 0.56927736 0.3676835 0.36477421  
## Proportion of Variance 0.04050959 0.0168989 0.01663253  
## Cumulative Proportion 0.96646858 0.9833675 1.00000000  
##   
## Loadings:  
## Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8  
## Red\_Meat 0.314 0.127 0.586 0.502 0.300 0.393 0.152 0.148  
## White\_Meat 0.318 0.395 -0.503 -0.290 0.304 0.326 0.455  
## Eggs 0.423 0.190 0.307 -0.411 -0.454 0.224 -0.512  
## Milk 0.372 0.394 -0.595 0.120 -0.429 0.123 0.363  
## Fish 0.148 -0.755 -0.162 -0.266 0.271 0.478   
## Cereals -0.438 0.241 0.369 -0.118 0.753 -0.169  
## Starchy\_Foods 0.306 -0.374 -0.445 0.339 0.526 -0.374 0.190  
## Pulses\_nuts\_oilseeds -0.418 -0.128 0.200 0.395 -0.400 -0.368 0.555

## B

The first and second components explains 50.62% and 68.42% of the variance, respectively.

pro\_eigen <- eigen(protein\_corr)  
lambda <- pro\_eigen$values  
100 \* round(cumsum(lambda)/sum(lambda), 4)

## [1] 50.62 68.42 81.42 87.86 92.60 96.65 98.34 100.00

## C

Variables with greater positive loading coefficients such as Eggs and Milk have a positive relationship with Component 1, while those with higher negative loading coefficients such as Cereals and Pulses\_nuts\_oilseeds have a negative relationship with Component 1. This data assists in determining which variables contribute more or less to Component 1.

Variables with greater positive loading coefficients such as White\_Meat and Cereals have a positive relationship with Component 2, while variables with higher negative loading coefficients such as Fish and Starchy\_Foods have a negative relationship with Component 2.

load\_co <- protein\_corr.pca$loadings  
round(load\_co[,1:2], 3)

## Comp.1 Comp.2  
## Red\_Meat 0.314 0.127  
## White\_Meat 0.318 0.395  
## Eggs 0.423 0.190  
## Milk 0.372 0.087  
## Fish 0.148 -0.755  
## Cereals -0.438 0.241  
## Starchy\_Foods 0.306 -0.374  
## Pulses\_nuts\_oilseeds -0.418 -0.128

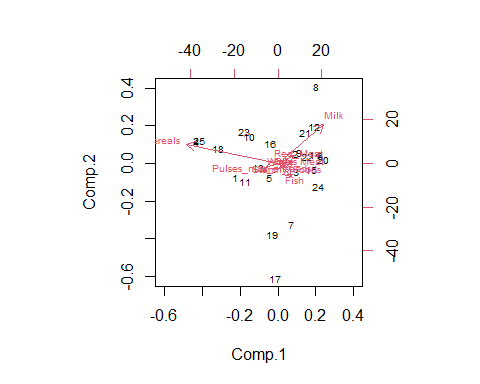
## D

Component 1 and Component 2 in the data show two separate eating habits. Component 1 is linked to a diet high in animal protein and low in starchy foods, where Component 2 is linked to a diet high in white meat, cereals, eggs, and fish but low in starchy foods and legumes/nuts/oilseeds. These interpretations can give insights into the foundation of the dietary data and can be utilized for more study or understanding of the examined population’s food preferences and habits.

## E

Czechoslovakia and East Germany are 5 and 7 respectively. Czechoslovakia appears to be in the middle of component 1 and component 2. East Germany appears to have a negative correlation for component 2, while having a positive correlation for component 1.

pro.pca <- princomp(protein[,2:9])  
biplot(pro.pca, cex = 0.6)



# Problem #3

data("heptathlon", package = "HSAUR2")  
hep = heptathlon  
hep$hurdles <- max(hep$hurdles) - hep$hurdles  
hep$run200m <- max(hep$run200m) - hep$run200m  
hep$run800m <- max(hep$run800m) - hep$run800m  
head(hep)

## hurdles highjump shot run200m longjump javelin run800m  
## Joyner-Kersee (USA) 3.73 1.86 15.80 4.05 7.27 45.66 34.92  
## John (GDR) 3.57 1.80 16.23 2.96 6.71 42.56 37.31  
## Behmer (GDR) 3.22 1.83 14.20 3.51 6.68 44.54 39.23  
## Sablovskaite (URS) 2.81 1.80 15.23 2.69 6.25 42.78 31.19  
## Choubenkova (URS) 2.91 1.74 14.76 2.68 6.32 47.46 35.53  
## Schulz (GDR) 2.67 1.83 13.50 1.96 6.33 42.82 37.64  
## score  
## Joyner-Kersee (USA) 7291  
## John (GDR) 6897  
## Behmer (GDR) 6858  
## Sablovskaite (URS) 6540  
## Choubenkova (URS) 6540  
## Schulz (GDR) 6411

## A

heptathlon\_scale <- scale(heptathlon[,1:8])  
hep.dist <- dist(heptathlon\_scale[1:8,])  
hep.dist

## Joyner-Kersee (USA) John (GDR) Behmer (GDR)  
## John (GDR) 2.170057   
## Behmer (GDR) 2.141493 1.709198   
## Sablovskaite (URS) 3.391144 1.864215 1.992928  
## Choubenkova (URS) 3.480483 2.340148 2.037346  
## Schulz (GDR) 4.039431 2.726377 2.183686  
## Fleming (AUS) 3.912060 2.830780 2.219066  
## Greiner (USA) 4.252664 2.732432 2.876330  
## Sablovskaite (URS) Choubenkova (URS) Schulz (GDR)  
## John (GDR)   
## Behmer (GDR)   
## Sablovskaite (URS)   
## Choubenkova (URS) 1.657821   
## Schulz (GDR) 1.667671 2.129201   
## Fleming (AUS) 1.835097 2.621110 1.721234  
## Greiner (USA) 1.768674 2.996317 1.817969  
## Fleming (AUS)  
## John (GDR)   
## Behmer (GDR)   
## Sablovskaite (URS)   
## Choubenkova (URS)   
## Schulz (GDR)   
## Fleming (AUS)   
## Greiner (USA) 1.443197

## B

mds\_result <- cmdscale(hep.dist, k = 2) eigenvalues <- eigen(crossprod(mds\_result))$values proportions <- eigenvalues / sum(eigenvalues) target\_proportion <- 0.80 cumulative\_proportion <- cumsum(proportions) min\_dimensions <- sum(cumulative\_proportion < target\_proportion) + 1

cov\_hep <- cov(heptathlon[, 1:8])  
hep\_ev <- eigen(cov\_hep)  
hep\_lambda <- hep\_ev$values  
round(hep\_lambda, 4)

## [1] 323202.4895 30.1257 9.6106 0.5920 0.2177 0.0419  
## [7] 0.0084 0.0006

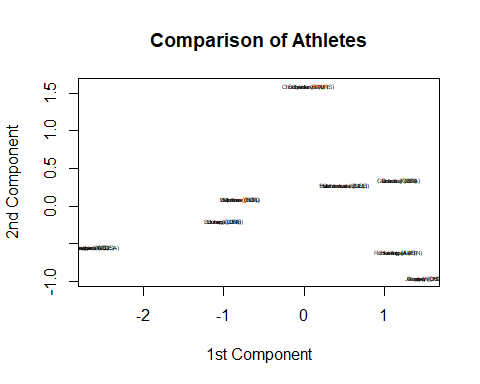
ev <- hep\_ev$vectors  
round(ev, 4)

## [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]  
## [1,] 0.0012 -0.0193 -0.0442 -0.1277 0.0299 0.9778 0.1563 -0.0058  
## [2,] -0.0001 0.0006 0.0052 0.0340 -0.0654 -0.0111 0.0745 -0.9944  
## [3,] -0.0021 -0.0827 0.0523 -0.9783 0.1110 -0.1199 -0.0680 -0.0442  
## [4,] 0.0015 0.0173 0.0252 -0.1165 -0.9906 0.0194 -0.0154 0.0600  
## [5,] -0.0008 -0.0011 0.0333 0.0542 0.0085 0.1648 -0.9815 -0.0739  
## [6,] -0.0016 -0.3275 -0.9424 -0.0174 -0.0283 -0.0442 -0.0398 -0.0064  
## [7,] 0.0113 -0.9408 0.3247 0.0925 -0.0188 0.0060 0.0175 0.0068  
## [8,] -0.9999 -0.0099 0.0050 0.0028 -0.0018 0.0015 0.0013 0.0004

cumulative\_proportion\_cov <- 100 \* round(cumsum(hep\_lambda) / sum(hep\_lambda), 4)

## C

newheptathlon.mds = cmdscale(hep.dist)  
plot(newheptathlon.mds, type = "n", main = "Comparison of Athletes", xlab = "1st Component", ylab = "2nd Component")  
text(newheptathlon.mds, labels = rownames(heptathlon), cex = 0.4)



## D

hep.cmd <- cmdscale(hep.dist, k = 2)  
plot(hep.cmd, type = "n", main = "Multidimensional Scaling of Heptathlon Data", xlab = "1st Component", ylab = "2nd Component")  
text(hep.cmd, labels = rownames(heptathlon), cex = 0.8)

