ST557: HOMEWORK 4

Brian Cervantes Alvarez
May 21, 2024

Question 1

Part A

Models 1 and 2 both highlight important factors: Weight, Height, SBP, AAI, FEV, DSST, Physact, and Atrophy.

PC1 in both models combines anthropometric and physiological factors, while PC2 focuses on cognitive and respiratory aspects. In Model 2, PC2 includes physical activity and atrophy, and PC3 introduces more variation related to cognitive and respiratory factors, as well as physical activity and atrophy.

Consistent variables that I noticed between both models: Weight, Height, SBP, AAI, FEV, and DSST consistently impact both models.

Hence, the principal components represent a mix of physical health, cognitive function, and physiological factors. PC1 generally captures overall health and physiological status, while subsequent PCs reveal details such as cognitive performance and specific physical attributes.

```
# Load the data
physio <- read.csv('PhysioData.csv')

# Extract the correlation matrix
correlationMatrix <- as.matrix(physio)

# Principal Component Factor Analysis

# Factors = 2 & 3
pcaResult2 <- prcomp(correlationMatrix, rank = 2)
pcaResult3 <- prcomp(correlationMatrix, rank = 3)

# Variance explained
print(summary(pcaResult2)$importance)
```

```
                            PC1       PC2
Standard deviation     0.6530516 0.4725061 0.3382827 0.3050721 0.2968584
Proportion of Variance 0.3552700 0.1859800 0.0953300 0.0775300 0.0734100
Cumulative Proportion  0.3552700 0.5412500 0.6365800 0.7141100 0.7875200

Standard deviation     0.2743533 0.2351406 0.2144054 0.1907133 0.1822645
Proportion of Variance 0.0627000 0.0460600 0.0382900 0.0303000 0.0276700
Cumulative Proportion  0.8502200 0.8962800 0.9345700 0.9648700 0.9925500

Standard deviation     0.09459717 3.361295e-17
Proportion of Variance 0.00745000 0.000000e+00
Cumulative Proportion  1.00000000 1.000000e+00
```

```
print(summary(pcaResult3)$importance)
```

```
                            PC1       PC2       PC3
Standard deviation     0.6530516 0.4725061 0.3382827 0.3050721 0.2968584
Proportion of Variance 0.3552700 0.1859800 0.0953300 0.0775300 0.0734100
Cumulative Proportion  0.3552700 0.5412500 0.6365800 0.7141100 0.7875200

Standard deviation     0.2743533 0.2351406 0.2144054 0.1907133 0.1822645
Proportion of Variance 0.0627000 0.0460600 0.0382900 0.0303000 0.0276700
Cumulative Proportion  0.8502200 0.8962800 0.9345700 0.9648700 0.9925500

Standard deviation     0.09459717 3.361295e-17
Proportion of Variance 0.00745000 0.000000e+00
Cumulative Proportion  1.00000000 1.000000e+00
```

```
# Loadings for factors  = 2 & 3
loadings2 <- pcaResult2$rotation
loadings3 <- pcaResult3$rotation
print(loadings2)
```

```
                PC1         PC2
weight  -0.381143834  0.08752747
height  -0.530688629  0.08658642
physact  0.009944638 -0.10659706
ldl      0.273985249 -0.06137024
alb      0.003366355 -0.03627835
crt     -0.324579480  0.31786338
plt      0.402387871 -0.04795723
sbp      0.161546655  0.51069170
aai     -0.144510926 -0.55164730
fev     -0.426639096 -0.15666626
dsst    -0.005095845 -0.44008213
atrophy -0.040315746  0.28587062
```

```
print(loadings3)
```

```
                PC1         PC2         PC3
weight  -0.381143834  0.08752747 -0.11087595
height  -0.530688629  0.08658642  0.00798927
physact  0.009944638 -0.10659706  0.72758107
ldl      0.273985249 -0.06137024 -0.27710141
alb      0.003366355 -0.03627835 -0.29683779
crt     -0.324579480  0.31786338 -0.03686026
plt      0.402387871 -0.04795723  0.03251985
sbp      0.161546655  0.51069170  0.21349093
aai     -0.144510926 -0.55164730  0.04151764
fev     -0.426639096 -0.15666626  0.09280287
dsst    -0.005095845 -0.44008213 -0.26487373
atrophy -0.040315746  0.28587062 -0.40605645
```

Part B

```
# Residual Matrix for factors 2 & 3
residualMatrix2 <- correlationMatrix - (pcaResult2$rotation %*% t(pcaResult2$rotation))
residualMatrix3 <- correlationMatrix - (pcaResult3$rotation %*% t(pcaResult3$rotation))
print(residualMatrix2)
```

```
               weight       height     physact           ldl           alb
weight    0.847068320  0.337911425 -0.01491694  0.113370143  0.051196513
height    0.337911425  0.710872372  0.07944245 -0.005835126  0.093126964
physact  -0.014916945  0.079442450  0.98853817 -0.041023262  0.010863053
ldl       0.113370143 -0.005835126 -0.04102326  0.921165776  0.121384656
alb       0.051196513  0.093126964  0.01086305  0.121384656  0.998672549
crt       0.102458395  0.165083863  0.01078488 -0.022917696  0.056908832
plt       0.008118457 -0.077065852 -0.01812162  0.083593055 -0.067389756
sbp       0.027333323 -0.033429683  0.05501439 -0.043139441  0.001274631
aai       0.080591892  0.044072893  0.02015585 -0.052790749  0.011499737
fev       0.187360175  0.364181511  0.08970387  0.043428568  0.061224563
dsst      0.093851306  0.056383103 -0.07861906 -0.024805326  0.052820262
atrophy   0.024497447  0.075383306 -0.05142430  0.009867980  0.058280743
                crt           plt           sbp         aai         fev
weight    0.102458395  0.008118457  0.027333323  0.08059189  0.18736018
height    0.165083863 -0.077065852 -0.033429683  0.04407289  0.36418151
physact   0.010784881 -0.018121619  0.055014394  0.02015585  0.08970387
ldl      -0.022917696  0.083593055 -0.043139441 -0.05279075  0.04342857
alb       0.056908832 -0.067389756  0.001274631  0.01149974  0.06122456
crt       0.793611036 -0.008966337 -0.110924212  0.08250748  0.13413640
plt      -0.008966337  0.835784106 -0.009502598 -0.04718415 -0.01556534
sbp      -0.110924212 -0.009502598  0.713096669 -0.02479916  0.03734034
aai       0.082507476 -0.047184149 -0.024799162  0.67480185  0.08013754
fev       0.134136395 -0.015565337  0.037340339  0.08013754  0.79343476
dsst     -0.008149417 -0.002376835  0.063210054 -0.03480217  0.08420792
atrophy   0.050412261 -0.032009621 -0.078916828  0.06984529  0.01739542
                dsst     atrophy
weight    0.093851306  0.02449745
height    0.056383103  0.07538331
physact  -0.078619059 -0.05142430
ldl      -0.024805326  0.00986798
alb       0.052820262  0.05828074
crt      -0.008149417  0.05041226
plt      -0.002376835 -0.03200962
sbp       0.063210054 -0.07891683
aai      -0.034802170  0.06984529
fev       0.084207924  0.01739542
dsst      0.806301754  0.13211295
atrophy   0.132112950  0.91665263
```

```
print(residualMatrix3)
```

```
              weight       height     physact           ldl          alb
weight    0.83477484  0.338797243  0.06575430  0.082646259  0.01828434
```

```
height   0.33879724  0.710808544  0.07362961 -0.003621288  0.09549848
physact  0.06575430  0.073629608  0.45916396  0.160590481  0.22683661
ldl      0.08264626 -0.003621288  0.16059048  0.844380583  0.03913048
alb      0.01828434  0.095498481  0.22683661  0.039130485  0.91055988
crt      0.09837148  0.165378350  0.03760371 -0.033131727  0.04596731
plt      0.01172413 -0.077325662 -0.04178245  0.092604351 -0.05773664
sbp      0.05100433 -0.035135319 -0.10031756  0.016019196  0.06464681
aai      0.08519520  0.043741197 -0.01005160 -0.041286152  0.02382374
fev      0.19764978  0.363440084  0.02218225  0.069144375  0.08877196
dsst     0.06448318  0.058499251  0.11409805 -0.098202210 -0.02580427
atrophy -0.02052445  0.078627400  0.24401469 -0.102650836 -0.06225216
                crt           plt          sbp         aai         fev
weight   0.098371479  0.011724126  0.051004333  0.08519520  0.19764978
height   0.165378350 -0.077325662 -0.035135319  0.04374120  0.36344008
physact  0.037603709 -0.041782445 -0.100317562 -0.01005160  0.02218225
ldl     -0.033131727  0.092604351  0.016019196 -0.04128615  0.06914437
alb      0.045967313 -0.057736636  0.064646805  0.02382374  0.08877196
crt      0.792252357 -0.007767647 -0.103054881  0.08403783  0.13755713
plt     -0.007767647  0.834726565 -0.016445291 -0.04853430 -0.01858327
sbp     -0.103054881 -0.016445291  0.667518294 -0.03366280  0.01752777
aai      0.084037827 -0.048534297 -0.033662802  0.67307813  0.07628458
fev      0.137557133 -0.018583273  0.017527768  0.07628458  0.78482239
dsst    -0.017912732  0.006236819  0.119758191 -0.02380524  0.10878897
atrophy  0.035444914 -0.018804727  0.007772539  0.08670380  0.05507863
                dsst       atrophy
weight   0.064483179 -0.020524449
height   0.058499251  0.078627400
physact  0.114098052  0.244014686
ldl     -0.098202210 -0.102650836
alb     -0.025804270 -0.062252156
crt     -0.017912732  0.035444914
plt      0.006236819 -0.018804727
sbp      0.119758191  0.007772539
aai     -0.023805237  0.086703799
fev      0.108788967  0.055078627
dsst     0.736143662  0.024559264
atrophy  0.024559264  0.751770792
```

Part C

In Model 1, the first factor is shaped by variables such as weight, height, physical activity physact, fev, and dsst. This factor indicates a combination of physical activity levels, and cognitive and respiratory elements. The second factor, in contrast, is influenced by sbp, aai, and atrophy. It places emphasis on physiological aspects such as blood pressure, arterial health, and the presence of atrophy in the studied subjects.

Moving on to Model 2, the first factor shares similarities with Model 1, being driven by weight, height, physical activity ,fev, and the dsst. This factor continues to represent a blend of body measurements, physical activity, and cognitive and respiratory factors. The second factor, however, is influenced by additional variables including ldl, crt, plt, sbp, aai, and atrophy. In other words, this factor focuses on physiological factors such as lipid levels, blood pressure, arterial health, and the presence of atrophy. Lastly, the third factor features crt, plt, fev, and dsst, respectively.

Together, these models from the MLFA provide the health indicators and factors within the studied population.

```r
# Maximum Likelihood Factor Analysis

# Factors = 2 & 3
mlfaResult2 <- factanal(covmat = correlationMatrix, factors = 2, method = "ml")
mlfaResult3 <- factanal(covmat = correlationMatrix, factors = 3, method = "ml")

# Loadings for factors = 2 & 3
loadings2 <- mlfaResult2$loadings
loadings3 <- mlfaResult3$loadings
print(loadings2)
```

```
Loadings:
        Factor1 Factor2
weight   0.569
height   0.956
physact
ldl     -0.159
alb
crt      0.395  -0.126
plt     -0.308
sbp             -0.443
aai              0.686
fev      0.592   0.283
dsst             0.342
atrophy  0.132  -0.151

               Factor1 Factor2
SS loadings      1.900   0.918
Proportion Var   0.158   0.077
Cumulative Var   0.158   0.235
```

```r
print(loadings3)
```

```
Loadings:
        Factor1 Factor2 Factor3
weight   0.563   0.144
height   0.945
physact
ldl     -0.235   0.967
alb              0.151
crt      0.407          -0.108
plt     -0.316   0.122
sbp                     -0.441
aai                      0.695
fev      0.579           0.304
dsst                     0.339
atrophy  0.139          -0.145

               Factor1 Factor2 Factor3
SS loadings      1.895   1.016   0.945
Proportion Var   0.158   0.085   0.079
Cumulative Var   0.158   0.243   0.321
```

Part D

After some side research, it was found that the residual values are zero (NULL), which means our factors and their connections perfectly match what we observe in the study. Again, this is not something that happens a lot in real-world situations. Getting a perfect match like this is like finding a needle in a haystack.

If this is not the correct method to get the residuals, then I must please leave me feedback with the correct R code. But I am certain this is the method.

```
# Residual Matrix for factors = 2 & 3
residualMatrix2 <- residuals(mlfaResult2)
residualMatrix3 <- residuals(mlfaResult3)
print(residualMatrix2)
```

```
NULL
```

```
print(residualMatrix3)
```

```
NULL
```

Part E

I would opt for using Principal Component Analysis for this dataset since I prioritize interpretability and simplicity. PCA's ability to capture maximum variance is a very useful tool. However, the choice is influenced by the challenge of understanding the factors from Maximum Likelihood Factor Analysis due to limited information on loadings and residuals. This lack of clarity led me to favor PCA for its transparency and ease of use in this context. Plus, I've used Principle Component Analysis before on a few machine learning projects so it's more easier for me to work with.

Part F

For models with m = 2 factors, both PCA and MLFA yield similar factors, emphasizing anthropometric, physiological, cognitive, and respiratory aspects. Likewise the m = 3 models, yielded similar results with PCA's PC3 aligning with MLFA's third factor, introducing additional variation related to cognitive, respiratory, and physical health factors.

Therefore, both methods are effective, and improved interpretability depends on one's familiarity with each tool.

Question 2

The first canonical variate, the weights for glucose intolerance, insulin response to oral glucose, and insulin resistance are approximately -528.870, -2174.966, and -2383.596, just to highlight.

Now, here is my best interpretation of the canonical variables (still rather new for me to grasp). In the context of this study, the results should provide insights into how the variables of the primary variables, glucose and insulin, and the secondary variables, weight and glucose levels, are correlated in non-diabetic patients. The canonical variables and correlations help summarize and quantify these relationships in a way that maximizes the shared information between the two sets of variables.

I think that is how it is? Truthfully, I do not fully understand how to interpret the results of the canonical correlations, but it's still a wonderful tool that I may touch upon in the near future.

```
# Given covariance matrix
S <- matrix(c(1106.00, 396.70, 108.40, 0.79, 26.23,
              396.70, 2382.00, 1143.00, -0.21, -23.96,
              108.40, 1143.00, 2136.00, 2.19, -20.84,
              0.79, -0.21, 2.19, 0.02, 0.22,
              26.23, -23.96, -20.84, 0.22, 70.56), nrow = 5, byrow = TRUE)

# Split the covariance matrix into S11, S12, S21, and S22
S11 <- S[1:3, 1:3]
S12 <- S[1:3, 4:5]
S21 <- S[4:5, 1:3]
S22 <- S[4:5, 4:5]

# Find canonical variable
A1 <- S11^(-1/2) %*% S12 %*% S22^(-1) %*%  S21 %*% S11^(-1/2)
A2 <- S22^(-1/2) %*% S21 %*% S11^(-1) %*%  S12 %*% S22^(-1/2)

# Compute canonical variables and correlations for all canonical variates
num_canonical_vars <- min(dim(S11)[1], dim(S22)[1])
canonical_variables <- matrix(NA, nrow = dim(S11)[1], ncol = num_canonical_vars)
canonical_correlations <- numeric(num_canonical_vars)

for (i in 1:num_canonical_vars) {
  eigenvectors_A1 <- eigen(A1)$vectors
  eigenvectors_A2 <- eigen(A2)$vectors

  a <- eigenvectors_A1[, i]
  b <- eigenvectors_A2[, i]

  U <- Re(a %*% t(S11))
  V <- Re(b %*% t(S22))

  lambda <- eigen(A1)$values[i]
  r <- sqrt(lambda)

  canonical_variables[, i] <- U
  canonical_correlations[i] <- r
}
```

```r
# Print canonical variables and correlations for all canonical variates
print("Canonical Variables:")
```

```
[1] "Canonical Variables:"
```

```r
print(canonical_variables)
```

```
          [,1]        [,2]
[1,]  -528.870  -117.57781
[2,] -2174.966 -1542.42981
[3,] -2383.596   -62.88942
```

```r
print("Canonical Correlations:")
```

```
[1] "Canonical Correlations:"
```

```r
print(canonical_correlations)
```

```
[1] 1.798948e+00 7.300048e-08
```

Question 3

Part A

```
library(MASS)
library(caret)

crudeOil <- read.csv("CrudeOilData.csv")
colnames(crudeOil) <- c("Population",
                        "Vanadium",
                        "Iron",
                        "Beryllium",
                        "SaturatedHydrocarbons",
                        "AromaticHydrocarbons")
crudeOil$Population <- as.factor(crudeOil$Population)
head(crudeOil)
```

	Population	Vanadium	Iron	Beryllium	SaturatedHydrocarbons	AromaticHydrocarbons
1	1	5.0	47	0.07	7.06	6.10
2	1	3.4	32	0.20	5.82	4.69
3	1	1.2	12	0.00	5.54	3.15
4	1	8.4	17	0.07	6.31	4.55
5	1	4.2	36	0.50	9.25	4.95
6	1	4.2	35	0.50	5.69	2.22

```
ldaModel <- lda(Population ~ ., data = crudeOil)
x0 <- data.frame(Vanadium = 1.0,
                 Iron = 30.0, Beryllium = 0.07,
                 SaturatedHydrocarbons = 8.34,
                 AromaticHydrocarbons = 9.59)
classificationX0 <- predict(ldaModel, newdata = x0)$class
print(paste("Classification for x0: ", classificationX0))
```

```
[1] "Classification for x0:  1"
```

Part B

```r
set.seed(2392)
trainIndex <- createDataPartition(crudeOil$Population, p = 0.8, list = FALSE)
trainData <- crudeOil[trainIndex, ]
testData <- crudeOil[-trainIndex, ]

ldaModel <- lda(Population ~ ., data = trainData)
predictions <- predict(ldaModel, newdata = testData)$class
confMatrix <- confusionMatrix(predictions, testData$Population)
APER <- 1 - confMatrix$overall["Accuracy"]

print("Confusion Matrix:")
```

```
[1] "Confusion Matrix:"
```

```r
print(confMatrix)
```

```
Confusion Matrix and Statistics

          Reference
Prediction 1 2
         1 1 0
         2 1 7

               Accuracy : 0.8889
                 95% CI : (0.5175, 0.9972)
    No Information Rate : 0.7778
    P-Value [Acc > NIR] : 0.372

                  Kappa : 0.6087

 Mcnemar's Test P-Value : 1.000

            Sensitivity : 0.5000
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.8750
             Prevalence : 0.2222
         Detection Rate : 0.1111
   Detection Prevalence : 0.1111
      Balanced Accuracy : 0.7500

       'Positive' Class : 1
```

```r
print(paste("Apparent Error Rate: ", APER))
```

```
[1] "Apparent Error Rate:  0.111111111111111"
```

Part C

```
ldaModel <- lda(Population ~ ., data = crudeOil)
newObs <- data.frame(Vanadium = 4.0,
                             Iron = 17.0,
                             Beryllium = 0.50,
                             SaturatedHydrocarbons = 5.54,
                             AromaticHydrocarbons = 3.51)
classificationX <- predict(ldaModel, newdata = newObs)$class
print(paste("Classification for x: ", classificationX))
```

```
[1] "Classification for x:  2"
```