Hoework 4

Elena Volpi

11/21/2021

```r
PhysioData <-  read.delim("~/Desktop/Multivariate/PhysioData.txt")
#PhysioData2 <- read_csv("~/Desktop/Multivariate/PhysioData.csv")

#physio loaded in funky, making matrix to compensate
physio_cor <- as.matrix(PhysioData[,2:13])
colnames(physio_cor) <- c('weight', 'height', 'physact', 'ldl', 'alb', 'crt', 'plt', 'sbp', 'aai','fec'
```

a) Perform a principal components factor analysis based on the given correlation matrix, for m = 2 and m = 3 factors. Describe how you might interpret the resulting factors for each model: can you describe the underlying latent variables for these two models? Which variables contribute most to each factor?

```r
#given the correlaation matrix PhysioData, perform PCA with m=2 and m=3
#eigenvalues of cvariance matrix
eigen_1 <- eigen(physio_cor)
#eigen_2 <- eigen(PhysioData2[,-1])

load_pcfa <- eigen_1$vectors %*% diag(sqrt(eigen_1$val))

load_pcfa[,1:2]  #m=2
```

```
##              [,1]        [,2]
##  [1,] -0.67032225 -0.11199841
##  [2,] -0.85586395 -0.12731545
##  [3,] -0.08117641  0.14480727
##  [4,]  0.23417307  0.12765583
##  [5,] -0.14462115  0.07513461
##  [6,] -0.52025732 -0.40987168
##  [7,]  0.45812656  0.10680428
##  [8,]  0.17581320 -0.61654067
##  [9,] -0.27740006  0.68320497
## [10,] -0.73431536  0.18605277
## [11,] -0.12659588  0.56614787
## [12,] -0.15419489 -0.34196705
```

For factor 1, we see the strong contributors are weight, height and for expiratory volume. With factor 2, the strong contributors are sytemic blood pressure and the ratio of systemic blood pressure between arm and ankle. Blood pressure would be a latent variable.

```r
load_pcfa[,1:3]    #m=3
```

```
##               [,1]        [,2]         [,3]
##  [1,] -0.67032225 -0.11199841  0.233316706
##  [2,] -0.85586395 -0.12731545  0.038000375
##  [3,] -0.08117641  0.14480727 -0.322368206
##  [4,]  0.23417307  0.12765583  0.694537196
##  [5,] -0.14462115  0.07513461  0.556830421
##  [6,] -0.52025732 -0.40987168 -0.006760628
##  [7,]  0.45812656  0.10680428  0.274416227
##  [8,]  0.17581320 -0.61654067  0.021936623
##  [9,] -0.27740006  0.68320497 -0.143342355
## [10,] -0.73431536  0.18605277  0.024790535
## [11,] -0.12659588  0.56614787  0.219687698
## [12,] -0.15419489 -0.34196705  0.299617821
```

We see the strong contributors to factor three to be cholestrol (ldl) and albumin levels (alb).

#Question 1, b

```r
#compute uniquesness

#residual matrix for pfc model with m<p
res_matrix_pfc <- function(m){
  uni_pcfa_m <- diag(physio_cor -load_pcfa[,1:m] %*% t(load_pcfa[,1:m]))
  fit_pcfa_m <- load_pcfa[,1:m]%*% t(load_pcfa[,1:m]) +diag(uni_pcfa_m)
  res_pcfa_m <- physio_cor-fit_pcfa_m
  return(res_pcfa_m)
}

m<- 3
res_matrix_pfc(m)
```

```
##              weight       height      physact          ldl         alb
##  [1,]  0.000000000 -0.049071083  0.008980265  0.012792307 -0.17170757
##  [2,] -0.049071083  0.000000000  0.026145439  0.033730359 -0.04717074
##  [3,]  0.008980265  0.026145439  0.000000000  0.192663856  0.17164826
##  [4,]  0.012792307  0.033730359  0.192663856  0.000000000 -0.23793103
##  [5,] -0.171707566 -0.047170736  0.171648256 -0.237931032  0.00000000
##  [6,] -0.139076024 -0.132338421 -0.011385879  0.047493230  0.00360448
##  [7,]  0.105581673  0.100502962  0.101178182 -0.114722825 -0.15886857
##  [8,]  0.054142113 -0.003798259  0.112805812 -0.007920818  0.04282627
##  [9,]  0.011401649 -0.071989068 -0.090137682  0.018771238  0.01939325
## [10,] -0.140915920 -0.028699844  0.023602061  0.067138604 -0.06850872
## [11,] -0.015434904 -0.023635419 -0.053196873 -0.194401576 -0.11440597
## [12,] -0.146680797 -0.065362200  0.051291396 -0.147055232 -0.11566849
##                crt         plt          sbp         aai         fec        dsst
##  [1,] -0.13907602  0.10558167  0.054142113  0.01140165 -0.14091592 -0.01543490
##  [2,] -0.13233842  0.10050296 -0.003798259 -0.07198907 -0.02869984 -0.02363542
##  [3,] -0.01138588  0.10117818  0.112805812 -0.09013768  0.02360206 -0.05319687
##  [4,]  0.04749323 -0.11472283 -0.007920818  0.01877124  0.06713860 -0.19440158
##  [5,]  0.00360448 -0.15886857  0.042826273  0.01939325 -0.06850872 -0.11440597
##  [6,]  0.00000000  0.12915795 -0.162114901  0.08880215 -0.08279136  0.02128937
```

```
##  [7,]  0.12915795  0.00000000  0.010295178  0.01457247  0.13000876 -0.04607817
##  [8,] -0.16211490  0.01029518  0.000000000  0.14327258  0.13167768  0.20413176
##  [9,]  0.08880215  0.01457247  0.143272578  0.00000000 -0.09904170 -0.18171784
## [10,] -0.08279136  0.13000876  0.131677675 -0.09904170  0.00000000 -0.04841282
## [11,]  0.02128937 -0.04607817  0.204131759 -0.18171784 -0.04841282  0.00000000
## [12,] -0.06399230 -0.03699742 -0.129737653  0.15177944 -0.06722203  0.11477297
##           atrophy
##  [1,] -0.14668080
##  [2,] -0.06536220
##  [3,]  0.05129140
##  [4,] -0.14705523
##  [5,] -0.11566849
##  [6,] -0.06399230
##  [7,] -0.03699742
##  [8,] -0.12973765
##  [9,]  0.15177944
## [10,] -0.06722203
## [11,]  0.11477297
## [12,]  0.00000000
```

```
m <- 2
res_matrix_pfc(2)
```

```
##            weight       height      physact          ldl           alb
##  [1,]  0.00000000 -0.040204960 -0.066233623  0.174839438 -0.0417897267
##  [2,] -0.04020496  0.000000000  0.013895326  0.060123033 -0.0260109710
##  [3,] -0.06623362  0.013895326  0.000000000 -0.031232854 -0.0078561683
##  [4,]  0.17483944  0.060123033 -0.031232854  0.000000000  0.1488084075
##  [5,] -0.04178973 -0.026010971 -0.007856168  0.148808408  0.0000000000
##  [6,] -0.14065339 -0.132595327 -0.009206468  0.042797723 -0.0001600439
##  [7,]  0.16960756  0.110930881  0.012715115  0.075869452 -0.0060652708
##  [8,]  0.05926029 -0.002964659  0.105734142  0.007314982  0.0550412519
##  [9,] -0.02204252 -0.077436131 -0.043928664 -0.080785359 -0.0604241380
## [10,] -0.13513187 -0.027757795  0.015610381  0.084356553 -0.0547045993
## [11,]  0.03582191 -0.015287204 -0.124017202 -0.041820298  0.0079228219
## [12,] -0.07677495 -0.053976610 -0.045295864  0.061040490  0.0511678281
##                 crt          plt          sbp         aai         fec
##  [1,] -0.1406533915  0.169607563  0.059260293 -0.02204252 -0.13513187
##  [2,] -0.1325953275  0.110930881 -0.002964659 -0.07743613 -0.02775779
##  [3,] -0.0092064679  0.012715115  0.105734142 -0.04392866  0.01561038
##  [4,]  0.0427977226  0.075869452  0.007314982 -0.08078536  0.08435655
##  [5,] -0.0001600439 -0.006065271  0.055041252 -0.06042414 -0.05470460
##  [6,]  0.0000000000  0.127302720 -0.162263207  0.08977124 -0.08295896
##  [7,]  0.1273027199  0.000000000  0.016314943 -0.02476300  0.13681169
##  [8,] -0.1622632068  0.016314943  0.000000000  0.14012813  0.13222150
##  [9,]  0.0897712366 -0.024762997  0.140128131  0.00000000 -0.10259523
## [10,] -0.0829589569  0.136811685  0.132221496 -0.10259523  0.00000000
## [11,]  0.0198041422  0.014207696  0.208950965 -0.21320839 -0.04296664
## [12,] -0.0660179040  0.045222570 -0.123165050  0.10883152 -0.05979435
##               dsst     atrophy
##  [1,]  0.035821906 -0.07677495
##  [2,] -0.015287204 -0.05397661
##  [3,] -0.124017202 -0.04529586
##  [4,] -0.041820298  0.06104049
```

```
##  [5,]  0.007922822  0.05116783
##  [6,]  0.019804142 -0.06601790
##  [7,]  0.014207696  0.04522257
##  [8,]  0.208950965 -0.12316505
##  [9,] -0.213208390  0.10883152
## [10,] -0.042966644 -0.05979435
## [11,]  0.000000000  0.18059532
## [12,]  0.180595322  0.00000000
```

c)

```
#MLE factor analysis for m=2
mlfal_2 <- factanal(covmat = physio_cor, factors=2, rotation="none")
mlfal_2
```

```
##
## Call:
## factanal(factors = 2, covmat = physio_cor, rotation = "none")
##
## Uniquenesses:
##  weight  height physact     ldl     alb     crt     plt     sbp     aai     fec
##   0.675   0.084   0.988   0.974   0.990   0.828   0.903   0.801   0.526   0.569
##    dsst atrophy
##   0.883   0.960
##
## Loadings:
##       Factor1 Factor2
##  [1,]  0.570
##  [2,]  0.956
##  [3,]
##  [4,] -0.160
##  [5,]
##  [6,]  0.385  -0.154
##  [7,] -0.310
##  [8,]         -0.438
##  [9,]  0.105   0.681
## [10,]  0.610   0.241
## [11,]          0.340
## [12,]  0.121  -0.160
##
##                Factor1 Factor2
## SS loadings      1.930   0.889
## Proportion Var   0.161   0.074
## Cumulative Var   0.161   0.235
##
## The degrees of freedom for the model is 43 and the fit was 0.1927
```

The variables that are large contributors to factor 1 are again height, weight, and forced expiratory volume. A latent variable could be gender. The variables that are large contributors to factor 2 is the ratio of systemic blood pressure between arm and ankle. A possible latent variable here could be blood pressure.

```r
#MLE factor analysis for m=3
mlfal_3 <- factanal(covmat = physio_cor, factors=3, rotation="none")
mlfal_3
```

```
##
## Call:
## factanal(factors = 3, covmat = physio_cor, rotation = "none")
##
## Uniquenesses:
##  weight  height physact     ldl     alb     crt     plt     sbp     aai     fec
##   0.659   0.097   0.988   0.005   0.969   0.821   0.881   0.800   0.517   0.564
##    dsst atrophy
##   0.884   0.960
##
## Loadings:
##       Factor1 Factor2 Factor3
##  [1,]  0.584
##  [2,]  0.935  -0.165
##  [3,]
##  [4,]          0.997
##  [5,]  0.119   0.124
##  [6,]  0.369  -0.135  -0.158
##  [7,] -0.281   0.200
##  [8,]                 -0.437
##  [9,]  0.102           0.685
## [10,]  0.613           0.235
## [11,]                  0.337
## [12,]  0.119          -0.160
##
##                Factor1 Factor2 Factor3
## SS loadings      1.858   1.106   0.891
## Proportion Var   0.155   0.092   0.074
## Cumulative Var   0.155   0.247   0.321
##
## The degrees of freedom for the model is 33 and the fit was 0.1187
```

We see the same variables are still the largest contributors to factor 1. For factor 2, we now see the greatest contributor to be cholesterol, which is nearly 1, latent factor may be inactivity. Factor 3 highest contributor is aai, the ratio of sbp between arm and ankle. d) Residual matrix for mlfa

```r
#m=3
fit_mlf_3 <- mlfal_3$load %*% t(mlfal_3$load) + diag(mlfal_3$uni)
res_mlf_3 <- physio_cor - fit_mlf_3
res_mlf_3
```

```
##             weight        height       physact           ldl           alb
##  [1,]  1.769071e-06  1.715552e-03 -6.367751e-02  4.930839e-05 -2.203323e-02
##  [2,]  1.715552e-03 -1.434591e-07  5.151534e-03 -8.841224e-06 -7.031566e-04
##  [3,] -6.367751e-02  5.151534e-03 -9.850531e-07  1.740750e-05  8.126801e-03
##  [4,]  4.930839e-05 -8.841224e-06  1.740750e-05 -1.488487e-07  7.889531e-05
##  [5,] -2.203323e-02 -7.031566e-04  8.126801e-03  7.889531e-05  2.104348e-07
##  [6,]  3.741820e-02 -8.562872e-03 -3.980536e-02 -3.531039e-05  2.320952e-02
##  [7,]  1.466025e-02 -1.919502e-04  1.700254e-02  5.173403e-06 -5.476550e-02
```

```
##  [8,]  6.251393e-02 -9.205251e-03  4.505925e-02 -9.551266e-05  1.464274e-02
##  [9,]  3.219461e-02 -4.912088e-03  9.644994e-03 -4.221579e-06  3.195010e-04
## [10,] -2.015677e-02  1.870694e-03  4.145989e-02  3.735397e-05 -7.565595e-03
## [11,]  3.581256e-02 -3.353576e-03 -6.361512e-02 -1.814167e-04  5.108954e-02
## [12,] -5.758711e-03  4.482888e-04 -7.637260e-02  4.381138e-05  4.217393e-02
##                 crt           plt           sbp           aai           fec
##  [1,]  3.741820e-02  1.466025e-02  6.251393e-02  3.219461e-02 -2.015677e-02
##  [2,] -8.562872e-03 -1.919502e-04 -9.205251e-03 -4.912088e-03  1.870694e-03
##  [3,] -3.980536e-02  1.700254e-02  4.505925e-02  9.644994e-03  4.145989e-02
##  [4,] -3.531039e-05  5.173403e-06 -9.551266e-05 -4.221579e-06  3.735397e-05
##  [5,]  2.320952e-02 -5.476550e-02  1.464274e-02  3.195010e-04 -7.565595e-03
##  [6,]  2.870878e-08 -2.811321e-02 -3.922948e-02  1.661071e-02  2.440522e-02
##  [7,] -2.811321e-02  2.484734e-07 -2.902823e-04 -2.148498e-02  1.214759e-02
##  [8,] -3.922948e-02 -2.902823e-04  4.039692e-07 -2.284557e-02  4.679824e-02
##  [9,]  1.661071e-02 -2.148498e-02 -2.284557e-02  1.167642e-06  3.360116e-04
## [10,]  2.440522e-02  1.214759e-02  4.679824e-02  3.360116e-04  1.511709e-06
## [11,] -1.079476e-01  3.622663e-02 -1.118506e-02 -2.660727e-02  5.113672e-02
## [12,]  8.238260e-02 -2.842938e-02  1.192972e-03  1.448303e-02 -4.693377e-02
##                dsst       atrophy
##  [1,]  3.581256e-02 -5.758711e-03
##  [2,] -3.353576e-03  4.482888e-04
##  [3,] -6.361512e-02 -7.637260e-02
##  [4,] -1.814167e-04  4.381138e-05
##  [5,]  5.108954e-02  4.217393e-02
##  [6,] -1.079476e-01  8.238260e-02
##  [7,]  3.622663e-02 -2.842938e-02
##  [8,] -1.118506e-02  1.192972e-03
##  [9,] -2.660727e-02  1.448303e-02
## [10,]  5.113672e-02 -4.693377e-02
## [11,] -7.112001e-07  5.579803e-02
## [12,]  5.579803e-02 -2.512108e-06
```

```
#m=2
fit_mlf_2 <- mlfal_2$load %*% t(mlfal_2$load) + diag(mlfal_2$uni)
res_mlf_2 <- physio_cor - fit_mlf_2
res_mlf_2
```

```
##              weight        height       physact           ldl           alb
##  [1,] -2.295501e-06  2.316174e-03 -6.600738e-02  9.474229e-02 -6.364201e-03
##  [2,]  2.316174e-03  5.795176e-08  4.071739e-03 -3.665135e-03  4.081529e-04
##  [3,] -6.600738e-02  4.071739e-03 -4.414328e-07 -2.110953e-02  5.157061e-03
##  [4,]  9.474229e-02 -3.665135e-03 -2.110953e-02  1.515341e-07  1.394198e-01
##  [5,] -6.364201e-03  4.081529e-04  5.157061e-03  1.394198e-01 -3.679141e-07
##  [6,]  3.444910e-02 -8.229091e-03 -3.857461e-02 -6.982532e-02  1.442844e-02
##  [7,]  2.744664e-02  1.002223e-03  1.400327e-02  1.472004e-01 -3.435965e-02
##  [8,]  5.876081e-02 -8.040394e-03  4.604408e-02 -4.377930e-02  8.276838e-03
##  [9,]  2.805056e-02 -4.595808e-03  1.118192e-02 -4.186461e-02 -5.241403e-03
## [10,] -1.168937e-02  1.341863e-03  4.041874e-02  3.369120e-02 -7.839538e-04
## [11,]  3.688455e-02 -2.267372e-03 -6.381315e-02  6.538310e-03  5.216409e-02
## [12,] -4.485594e-03  7.105639e-05 -7.647922e-02  7.147180e-04  4.267490e-02
##                 crt           plt           sbp           aai           fec
##  [1,]  3.444910e-02  2.744664e-02  5.876081e-02  2.805056e-02 -1.168937e-02
##  [2,] -8.229091e-03  1.002223e-03 -8.040394e-03 -4.595808e-03  1.341863e-03
##  [3,] -3.857461e-02  1.400327e-02  4.604408e-02  1.118192e-02  4.041874e-02
```

```
## [4,] -6.982532e-02  1.472004e-01 -4.377930e-02 -4.186461e-02  3.369120e-02
## [5,]  1.442844e-02 -3.435965e-02  8.276838e-03 -5.241403e-03 -7.839538e-04
## [6,]  2.907132e-06 -3.956903e-02 -3.556760e-02  1.837698e-02  2.507144e-02
## [7,] -3.956903e-02  2.774056e-06 -7.107719e-03 -2.818814e-02  1.605121e-02
## [8,] -3.556760e-02 -7.107719e-03  3.996594e-07 -2.291624e-02  4.608291e-02
## [9,]  1.837698e-02 -2.818814e-02 -2.291624e-02  1.046105e-06 -3.191512e-05
## [10,]  2.507144e-02  1.605121e-02  4.608291e-02 -3.191512e-05  6.081541e-06
## [11,] -1.080193e-01  3.696668e-02 -1.042492e-02 -2.647690e-02  5.121941e-02
## [12,]  8.309776e-02 -2.852986e-02  1.028431e-03  1.386842e-02 -4.580250e-02
##               dsst       atrophy
## [1,]  3.688455e-02 -4.485594e-03
## [2,] -2.267372e-03  7.105639e-05
## [3,] -6.381315e-02 -7.647922e-02
## [4,]  6.538310e-03  7.147180e-04
## [5,]  5.216409e-02  4.267490e-02
## [6,] -1.080193e-01  8.309776e-02
## [7,]  3.696668e-02 -2.852986e-02
## [8,] -1.042492e-02  1.028431e-03
## [9,] -2.647690e-02  1.386842e-02
## [10,]  5.121941e-02 -4.580250e-02
## [11,] -3.194236e-07  5.637991e-02
## [12,]  5.637991e-02 -9.733176e-07
```

e) Which method do you prefer

I think it is preferable to use the method that produced smaller residuals which was the maximum likelihood approach.

f) are they similiar for m=2 and m=3? For m=2, we had very similiar results and the most important contributors for each factor were the same. For m=3, however, the most important variables for factor 2 was ytemic blood pressure and the ratio of systemic blood pressure between arm and ankle (aai) factor three were cholestrol (ldl) and albumin levels (alb) for the principal components approach. Wheareas, for maximum likelihood approach it was choesterol (ldl) for factor 2 and aai for factor 3.

#Problem 2

```
S <- matrix(c(1106, 396.7, 108.4, 0.79, 26.23,
                 396.7, 2382, 1143, -.21, -23.96,
                 108.4, 1143, 2136, 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)

s11 <- as.matrix(S[1:3,1:3])
s12 <- as.matrix(S[1:3,4:5])
s21 <- as.matrix(S[4:5,1:3])
s22 <- as.matrix(S[4:5,4:5])

sig11.eig <- eigen(s11)
sig11.5 <- sig11.eig$vec %*% diag(sqrt(sig11.eig$val)) %*%
  t(sig11.eig$vec)

sig22.eig <- eigen(s22)
```

```r
sig22.5 <- sig22.eig$vec %*% diag(sqrt(sig22.eig$val)) %*%
  t(sig22.eig$vec)

A1 <- solve(sig11.5) %*% s12 %*% solve(s22) %*% t(s12) %*% solve(sig11.5)
A2 <- solve(sig22.5) %*% t(s12) %*% solve(s11) %*% s12 %*% solve(sig22.5)

A1.eig <- eigen(A1)
A2.eig <- eigen(A2)
# First canonical variates loadings:
e1 <- A1.eig$vec[,1]
f1 <- A2.eig$vec[,1]

(a1 <- e1 %*% solve(sig11.5))
```

```
##          [,1]        [,2]       [,3]
## [1,] 0.013188 -0.01443349 0.02337164
```

```r
(b1 <- f1 %*% solve(sig22.5))
```

```
##           [,1]       [,2]
## [1,] -7.185484 0.01611295
```

```r
#Second canonical variate lodings
e2 <- A1.eig$vec[,2]
f2 <- A2.eig$vec[,2]

(a2 <- e2 %*% solve(sig11.5))
```

```
##            [,1]         [,2]         [,3]
## [1,] 0.02471366 -0.009285288 -0.008727547
```

```r
(b2 <- f2 %*% solve(sig22.5))
```

```
##         [,1]       [,2]
## [1,] 0.38023 -0.1200668
```

```r
sqrt(A1.eig$val[2])
```

```
## [1] 0.1254845
```

```r
sqrt(A2.eig$val[2])
```

```
## [1] 0.1254845
```

```r
sqrt(A1.eig$val[1])
```

```
## [1] 0.4611246
```

```r
sqrt(A2.eig$val[1])
```

```
## [1] 0.4611246
```

From our canonical variate loadings, we see that the weighted difference between glucose intolerance, insulin response to oral glucose, and insulin resistance is most highly correlated with the weighted difference between relative weight and fasting plasma glucose.

From canonical variate loading 1, we see $U_1 = 0.013188(X_1^1) - 0.01443349(X_2^1) + 0.02337164(X_3^1)$ is most highly correlated with the variable $V_1 = -7.185484(X_1^2) + 0.01611295(X_2^2)$.

From canonical variate loading 2, $U_2 = 0.02471366(X_1^1) - 0.009285288(X_2^1) - 0.008727547(X_3^1)$ and $V_2 = 0.38023(X_1^2) - 0.1200668(X_2^2)$ are the most highly correlated combinations that are uncorrelated with the first conanical variates U_1 and V_1.

#Problem 3

```r
CrudeOilData <- read_csv("CrudeOilData.csv", show_col_types = FALSE)
```

a) Obtain the Fisher's (Linear) Discriminant Function rule for this data: what are the values for a new observation x_0 for which the observation would be classified as coming from π_1?

```r
(oil_lda <- lda(Population~., data=CrudeOilData))
```

```
## Call:
## lda(Population ~ ., data = CrudeOilData)
##
## Prior probabilities of groups:
##         1         2
## 0.2244898 0.7755102
##
## Group means:
##   Vanadium     Iron Beryllium SatHydroCarb AroHydroCarb
## 1 4.445455 33.09091 0.1709091     6.560909     5.483636
## 2 7.226316 22.25263 0.4321053     4.658158     5.767895
##
## Coefficients of linear discriminants:
##                       LD1
## Vanadium      0.210812184
## Iron         -0.037069792
## Beryllium     2.960016454
## SatHydroCarb -0.846215662
## AroHydroCarb -0.001726671
```

```r
pop1 <- CrudeOilData[CrudeOilData$Population == 1, -1]
pop2 <- CrudeOilData[CrudeOilData$Population == 2, -1]
n1 <- nrow(pop1)
n2 <- nrow(pop2)
pop1.xbar <- apply(pop1, 2, mean)
pop2.xbar <- apply(pop2, 2, mean)
pop1.cov <- cov(pop1)
pop2.cov <- cov(pop2)
```

```r
oil.Sp <- ((n1-1)*pop1.cov + (n2-1)*pop2.cov)/(n1+n2-2)

a_T <- t(pop1.xbar - pop2.xbar)%*% solve(oil.Sp)
a_T
```

```
##        Vanadium     Iron Beryllium SatHydroCarb AroHydroCarb
## [1,] -0.7106066 0.124955 -9.977636     2.852427  0.005820271
```

A new observation, X_0 will be classified as belonging to population one if $a^T X_0 \geq \frac{a^T \bar{X}_1 + a^T \bar{X}_2}{2}$.

#b) Construct the confusion matrix for the given data, comparing true population membership to the predicted classification based on Fisher's (Linear) Discriminant Function. What is the apparent error rate (APER) for this classifier, based on the given data?

```r
oil.ldaPred <- predict(oil_lda, newdata=CrudeOilData[,-1])
actual <- as.factor(CrudeOilData$Population)

confusionMatrix(actual, oil.ldaPred$class)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  1  2
##          1  9  2
##          2  1 37
##
##                Accuracy : 0.9388
##                  95% CI : (0.8313, 0.9872)
##     No Information Rate : 0.7959
##     P-Value [Acc > NIR] : 0.005575
##
##                   Kappa : 0.8183
##
##  Mcnemar's Test P-Value : 1.000000
##
##             Sensitivity : 0.9000
##             Specificity : 0.9487
##          Pos Pred Value : 0.8182
##          Neg Pred Value : 0.9737
##              Prevalence : 0.2041
##          Detection Rate : 0.1837
##    Detection Prevalence : 0.2245
##       Balanced Accuracy : 0.9244
##
##        'Positive' Class : 1
##
```

```r
error <- (1+2)/nrow(CrudeOilData)
error
```

```
## [1] 0.06122449
```

Our error is 0.06. d)

```
oil.newObs <- data.frame("Vanadium"=4, "Iron"=17, "Beryllium"=0.5, "SatHydroCarb"=5.54, "AroHydroCarb" =

oil.newObs.ldaPred <- predict(oil_lda, newdata=oil.newObs)
oil.newObs.ldaPred
```

```
## $class
## [1] 2
## Levels: 1 2
##
## $posterior
##            1         2
## 1 0.03047993 0.9695201
##
## $x
##         LD1
## 1 -0.270082
```

This observation would be assigned to population 2.