MS − E 2112 − Multivariate statistical analysis – Home exercise 8

Srikanth Gadicherla , 546195

March 11 *th* , 2016

Problem 1

Data: The data set consists of information of 48 athletes. We have been given total points and also individual points for 10 sports, they include, R100m, long jump, shot put, high jump, R400, hurdles, discus throw, pole vault, javelin throw, R1500. Then there is also information about height weight of the athletes.

a) How many pairs of canonical variables can we obtain?

In the first partition we have 2 variables, namely, height and weight. And the second partition has 10 variables, that is, the sports details of athletes. So, the pairs of canonical variables is minimum of variables in each partition. Therefore, there are 2 pairs of canonical variables.

b)After whitening the eigenvectors of M1 and M2 matrices, the canonical vectors are

alpha1 → (0.0574 0.1079)

beta1 →(-0.0037 -0.0007 0.0085 -0.0011 -0.0004 0.0030 0.0061 -0.0040 -0.0029 -0.0007)

alpha2 → (-0.3710 0.2573)

beta2 → (-0.0067 -0.0017 -0.0052 -0.0014 0.0112 0.011 0.0118 -0.0014 0.0019 0.0082)

The canonical variables are:

u1 → 0.0574*Height + 0.1079*Weight

v1 → -0.0037*R100m - 0.0007*LongJump + 0.0085*ShotPut - 0.0011*HighJump – 0.0004*R400m + 0.003*Hurdles + 0.0061*DiscusThrow -0.0040*PoleVault -0.0029*Javelin -0.0007*R1500

u2 → 0.3710*Height - 0.2573*Weight

v2 → - 0.0067*R100m - 0.0017*LongJump - 0.0052*ShotPut - 0.0014*HighJump + 0.0112*R400m + 0.0110*Hurdles + 0.0118*DiscusThrow - 0.0014*PoleVault + 0.0019*Javelin + 0.0082*R1500

c) The score vectors corresponding to canonical variables is given in the below figures.

```
> t(fii1)
        [,1]     [,2]     [,3]     [,4]     [,5]    [,6]    [,7]    [,8]     [,9]    [,10]
[1,] 19.3104 20.91357 20.68383 20.09367 22.70487 17.2755 20.6264 19.3104 20.46103 20.07287
        [,11]    [,12]   [,13]   [,14]   [,15]    [,16]   [,17]   [,18]    [,19]  [,20]
[1,] 20.79177 20.9502 19.3609 20.3026 20.3026 19.79957 19.6342 21.9929 21.28787 19.756
        [,21]    [,22]   [,23]    [,24]    [,25]    [,26]   [,27]    [,28]    [,29]    [,30]
[1,] 18.87173 19.87087 21.0512 21.11557 19.96493 19.19553 20.0798 19.41833 17.67754 19.41833
        [,31]    [,32]    [,33]    [,34]    [,35]    [,36]    [,37]    [,38]   [,39]   [,40]
[1,] 19.86394 18.54794 20.51847 20.41053 19.85007 20.74127 20.19467 19.62726 18.3687 19.8065
      [,41]    [,42]    [,43]   [,44]    [,45]    [,46]    [,47]    [,48]
[1,] 19.958 19.69857 20.29567 18.4697 19.14503 19.42527 20.29567 21.17993
```

```
> t(Re(eta1))
         [,1]     [,2]     [,3]     [,4]     [,5]        [,6]     [,7]     [,8]     [,9]
[1,] 1.854322 3.321614 2.765372 2.618159 4.798891 -0.01425697 3.144691 2.826071 3.200653
        [,10]    [,11]    [,12]     [,13]    [,14]    [,15]    [,16]    [,17]    [,18]
[1,] 2.798003 3.106082 4.298037 0.3587265 3.385253 2.701827 2.564909 3.142365 3.277108
        [,19]    [,20]    [,21]    [,22]    [,23]    [,24]    [,25]    [,26]    [,27]    [,28]
[1,] 2.840761 1.436025 1.269002 2.312152 2.759078 2.190071 1.906144 2.045877 3.085346 2.668214
        [,29]     [,30]    [,31]    [,32]    [,33]   [,34]    [,35]    [,36]    [,37]    [,38]
[1,] 0.523169 0.8968909 2.710556 1.501191 2.597447 2.68002 2.181798 3.295735 3.108272 2.133732
         [,39]    [,40]    [,41]    [,42]    [,43]    [,44]    [,45]    [,46]    [,47]
[1,] 0.4661145 2.647806 2.784648 1.977997 2.545014 2.493123 0.638667 1.737009 1.335766
        [,48]
[1,] 3.555324
```

```
> t(fii2)
          [,1]      [,2]     [,3]      [,4]      [,5]      [,6]     [,7]      [,8]      [,9]
[1,] -47.41774 -49.18243 -47.6984 -49.61366 -46.43482 -46.68255 -47.3274 -47.41774 -47.21377
        [,10]     [,11]     [,12]     [,13]     [,14]     [,15]     [,16]     [,17]     [,18]
[1,] -46.6155 -47.44103 -46.55527 -46.78936 -48.09952 -48.09952 -46.75924 -46.64561 -46.60868
         [,19]     [,20]     [,21]     [,22]     [,23]    [,24]     [,25]     [,26]     [,27]
[1,] -47.78192 -48.38701 -47.44785 -49.12902 -45.29851 -46.6689 -46.87287 -46.67572 -47.61489
         [,28]     [,29]     [,30]     [,31]     [,32]     [,33]     [,34]     [,35]     [,36]
[1,] -47.16036 -49.27959 -47.16036 -48.12964 -48.21998 -47.58477 -47.84215 -46.13086 -48.06941
        [,37]     [,38]     [,39]     [,40]     [,41]   [,42]     [,43]     [,44]     [,45]
[1,] -48.3569 -45.64622 -46.10757 -47.75863 -45.87349 -48.016 -47.10013 -44.85081 -47.30411
         [,46]     [,47]    [,48]
[1,] -48.15975 -47.10013 -48.0393
```

```
> t(Re(eta2))
        [,1]     [,2]     [,3]    [,4]     [,5]     [,6]     [,7]     [,8]     [,9]    [,10]
[1,] 20.7142 21.25526 19.37122 20.4409 19.62487 21.19916 20.63428 21.34279 21.21518 19.50683
        [,11]    [,12]    [,13]    [,14]    [,15]    [,16]    [,17]    [,18]    [,19]    [,20]
[1,] 19.40545 21.02671 20.20108 22.26779 20.95471 20.88271 19.26651 19.60636 20.63885 21.17877
      [,21]    [,22]    [,23]    [,24]    [,25]    [,26]    [,27]    [,28]    [,29]    [,30]
[1,] 20.795 21.73332 19.71401 19.83112 19.92603 17.15143 20.84521 19.66547 21.23949 21.12165
        [,31]    [,32]    [,33]    [,34]    [,35]    [,36]    [,37]    [,38]    [,39]   [,40]
[1,] 20.44969 19.60545 19.60174 19.07439 19.44748 19.51605 20.42804 19.32046 19.29933 19.9842
        [,41]    [,42]    [,43]    [,44]    [,45]    [,46]    [,47]    [,48]
[1,] 18.02678 19.80772 21.36931 18.43188 18.65028 20.05659 20.46978 19.91923
```

Correlation Structure

The correlation structure of the canonical variables is gvien below. It can be seen that (u1,v1) are fairly correlated with 0.78 magnitude but (u2,v2) less correlated with 0.49 magnitude

```
     [,1] [,2] [,3] [,4]
[1,] 1.00 0.00 0.78 0.00
[2,] 0.00 1.00 0.00 0.49
[3,] 0.78 0.00 1.00 0.00
[4,] 0.00 0.49 0.00 1.00
```

d) The first pair of canonical variables (u1,v1) has been plotted. (equation is again given below for convenience)

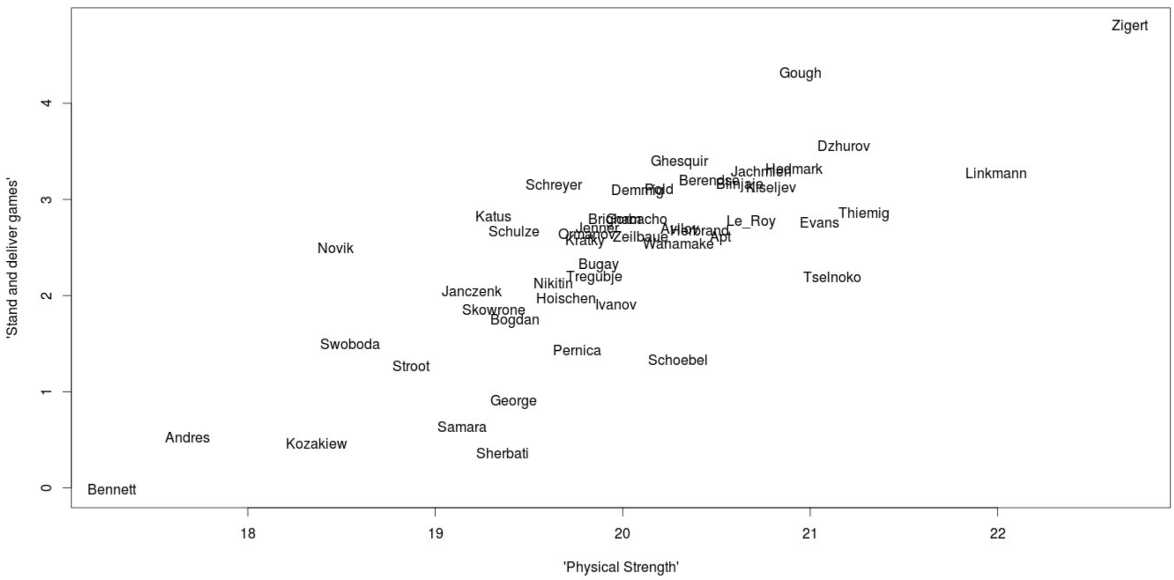
```
u1 →  0.0574*Height + 0.1079*Weight


v1 → -0.0037*R100m - 0.0007*LongJump + 0.0085*ShotPut - 0.0011*HighJump – 0.0004*R400m
+ 0.003*Hurdles + 0.0061*DiscusThrow -0.0040*PoleVault -0.0029*Javelin -0.0007*R1500
```

It can be seen that the physical attributes like height and weight are positively related to games like shot put, hurdles and discus throw. Except for hurdles the shot put and discus throw can be categorized as, what I cal, 'Stand and Deliver Games'. These games don't require much of stamina but need more physical strength of which main attributes are height and weight. So, athletes with more height and weight would score maximum in this analysis.

From the plot, it can be seen that Zigert is the tallest and heaviest with 198 cms and 105 kgs. So naturally, he scores maximum and he is on the top of the plot. It can also seen from the data the he has a good score of 924 in shot put.

It can also been seen that Bennett and Andres are at the left corner of the plot. They score badly in the stand and deliver games as they both are lightest and shortest athletes in the data.



e) The canonical variable (u2,v2) has been plotted. (equation is again given below for convenience)

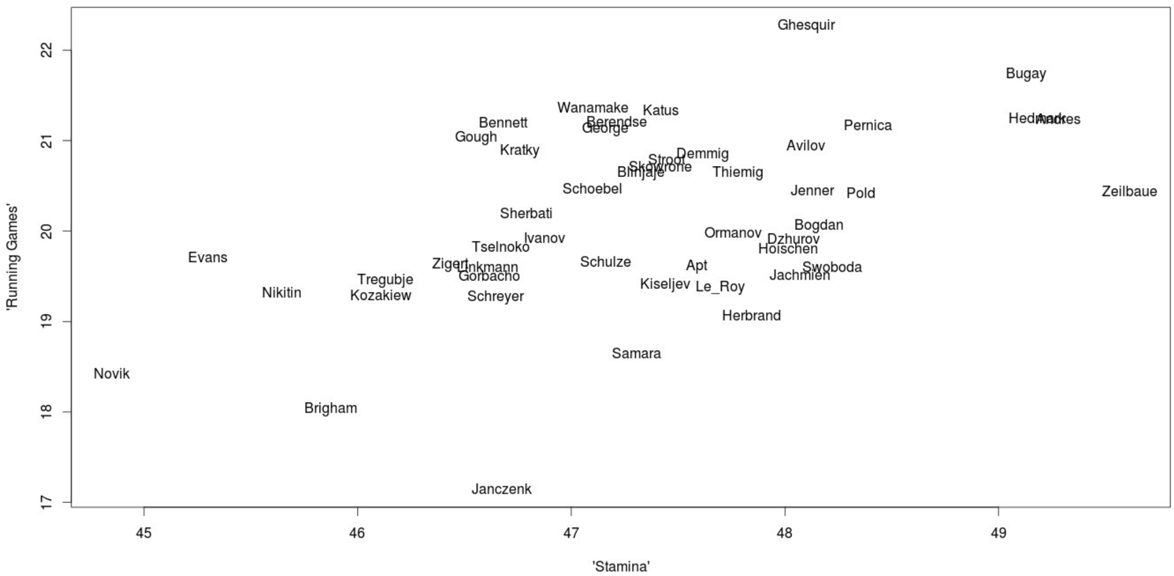
```
u2 → 0.3710*Height - 0.2573*Weight

v2 → - 0.0067*R100m - 0.0017*LongJump - 0.0052*ShotPut - 0.0014*HighJump + 0.0112*R400m + 0.0110*Hurdles + 0.0118*DiscusThrow - 0.0014*PoleVault + 0.0019*Javelin + 0.0082*R1500
```

The second pair of canonical variables can be seen as 'Stamina vs Running Games'. From the equations it can be seen that height is positively related to games like R400, hurdles, discus throw and R1500. Except for discus throw, the other games are basically running games which requires a lot of stamina and for running one needs better height and less weight (moslty dependent on height)

This can be verified by athletes like Bugay, who has a good height of 190 cms and less weight 83 kgs, so going by above interpretation he has good stamina and aslo performs well in R1500 game with 655 points (696 maximum).

On the other hand athletes like Janczenk, who is at the bottom of the plot, has less height 182 cms and he fairs badly in R1500 game 378 points which is almost 50% of the maximum.



Appendix

The code for the problem solved above.

%%%

```
setwd("~/Documents/OneDrive/Aalto/Sem2/MSA/Ex 8")


dec_data <- read.table("DECATHLON.txt",header=T,sep="\t")


Y <- dec_data[,c(3,4,5,6,7,8,9,10,11,12)] # X = (Price,Value)
X <- dec_data[,c(13,14)] # Y = (Economy,Service, Desing,Sport,Safety,
#                               Easy.h)
# (a)
#How many pairs of canonical variables can we obtain?
# Two pairs, since we need orthogonal u_i and v_i vectors
# and X has two components (2<6)
XY <- as.matrix(cbind(X,Y))


R <- cov(XY)
R11 <- R[1:2,1:2]
R22 <- R[3:12,3:12]
R21 <- R[3:12,1:2]
R12 <- R[1:2,3:12]
R11.inv <- solve(R11)
R22.inv <- solve(R22)


#Non-zero eigenvalues of M1 and M2 are the same
M1 <- R11.inv %*% R12 %*% R22.inv %*% R21
M2 <- R22.inv %*% R21 %*% R11.inv %*% R12




# Note that the eigenvectors are unique up to sings
# (sometimes R calculates the eigenvectors multiplied with -1)
va1 <- eigen(M1)$vectors[,1]
va2 <- eigen(M1)$vectors[,2]


vb1 <- eigen(M2)$vectors[,1]
```

```
vb2 <- eigen(M2)$vectors[,2]

# a1 = alpha1 , b1 = beta1

# Correct scaling: Whitening of data

a1 <- -va1/sqrt(va1%*%R11%*%va1)

a2 <- -va2/sqrt(va2%*%R11%*%va2)

b1 <- vb1/sqrt(vb1%*%R22%*%vb1)

b2 <- vb2/sqrt(vb2%*%R22%*%vb2)


fii1 <- XY[,1:2]%*%a1

fii2 <- XY[,1:2]%*%a2

eta1 <- XY[,3:12]%*%b1

eta2 <- XY[,3:12]%*%b2


round(cor(cbind(fii1,fii2,Re(eta1),Re(eta2))),2)


#canonical correlations

# g1 = 0.78

# g2 = 0.49


# Value = "Loss of value" = "How fast the value goes down"??


# u1 <- 0.0574*Height + 0.1079*Weight

# v1 <- -0.0037*R100m - 0.0007*Long Jump + 0.0085*Shot Put - 0.0011*High Jump -
0.0004*R400m

#        + 0.003*Hurdles + 0.0061*Discus Throw -0.0040*Pole Vault -0.0029*Javelin
-0.0007*R1500


## Stand and deliver games(like shot put and discus throw) requires more physical
strength which are mainly

## contributed by physical characteristics


plot(fii1,Re(eta1),xlab="'Physical Strength'",ylab="'Stand and deliver
games'",pch="")

text(fii1,Re(eta1),labels=dec_data$NAME)
```

```
# u2 <- -0.3710*Height + 0.2573*Weight

# v2 <- -0.0067*R100m -0.0017*LongJump - 0.0052*Shot Put - 0.0014*High Jump + 0.0112*R400m

#     + 0.0110*Hurdles + 0.0118*DiscusThrow - 0.0014*Pole Vault + 0.0019*Javelin + 0.0082*R1500


# The second pair of canonical variables provides more insight into the relation ship between the two sets

# of variables. u2 has low negative values for cars with good marks both in price and value, e.g., VW and Opel.

# On the right hand side, we should see cars with bad marks in these two variables such as Ferrari and Wartburg.

# The canonical variable v2 consists mainly of variables economy, service and easy handling. The position

# of cars is displayed in the plot below


plot(fii2,Re(eta2),xlab="'Stamina'",ylab="'Running Games'",pch="")

text(fii2,Re(eta2),labels=dec_data$NAME)


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```