PCA - Homework 4 - Lecture 6

Aleksey Kramer

November 14, 2015

After conduction Principle Component Analysis (as shown in the step-by-step manner below), components 1, 2, and 3 appear to explain 99% of the variance in the data set.

```
# Set working directory
setwd('C:\\Users\\Aleksey\\Documents\\School\\UW_Data_Science\\UW_Data_Science_450\\Week6\\homework')

# Read CSV file
data = read.csv("women_TnF_1984.csv")
```

Set working directory and load the dataset in the memory

```
# investigate data frame
str(data)
```

Look at the structure of the data

```
## 'data.frame':    55 obs. of  9 variables:
##  $ Obs     : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Country : Factor w/ 55 levels "Argentina","Australia",..: 1 2 3 4 5 6 7 8 9 10 ...
##  $ m100    : num  11.6 11.2 11.4 11.4 11.5 ...
##  $ m200    : num  22.9 22.4 23.1 23 23.1 ...
##  $ m400    : num  54.5 51.1 50.6 52 53.3 ...
##  $ m800    : num  2.15 1.98 1.99 2 2.16 2.1 2.18 2 2.05 2.08 ...
##  $ m1500   : num  4.43 4.13 4.22 4.14 4.58 4.49 4.45 4.06 4.23 4.33 ...
##  $ m3000   : num  9.79 9.08 9.34 8.88 9.81 9.77 9.51 8.81 9.37 9.31 ...
##  $ Marathon: num  179 152 159 158 170 ...
```

```
# look at the data
head(data)
```

Visually examine the dataset and remove unique identifier

```
##   Obs   Country  m100  m200  m400 m800 m1500 m3000 Marathon
## 1   1 Argentina 11.61 22.94 54.50 2.15  4.43  9.79   178.52
## 2   2 Australia 11.20 22.35 51.08 1.98  4.13  9.08   152.37
## 3   3   Austria 11.43 23.09 50.62 1.99  4.22  9.34   159.37
## 4   4   Belgium 11.41 23.04 52.00 2.00  4.14  8.88   157.85
## 5   5   Bermuda 11.46 23.05 53.30 2.16  4.58  9.81   169.98
## 6   6    Brazil 11.31 23.17 52.80 2.10  4.49  9.77   168.75
```

```
# Discovered unique identifier data$Obs
# Remove unique identifier
data$Obs = NULL
```

```
# display common stats
summary(data)
```

Visually examine the dataset again and convert minutes to seconds for fields m800, m1500, m3000, and Marathon

```
##       Country        m100            m200            m400
##  Argentina: 1   Min.   :10.79   Min.   :21.71   Min.   :47.99
##  Australia: 1   1st Qu.:11.27   1st Qu.:22.88   1st Qu.:51.55
##  Austria  : 1   Median :11.60   Median :23.54   Median :53.30
##  Belgium  : 1   Mean   :11.62   Mean   :23.64   Mean   :53.41
##  Bermuda  : 1   3rd Qu.:11.92   3rd Qu.:24.43   3rd Qu.:55.04
##  Brazil   : 1   Max.   :12.90   Max.   :27.10   Max.   :60.40
##  (Other)  :49
##       m800            m1500           m3000           Marathon
##  Min.   :1.890   Min.   :3.870   Min.   : 8.450   Min.   :142.7
##  1st Qu.:2.000   1st Qu.:4.115   1st Qu.: 8.860   1st Qu.:152.9
##  Median :2.050   Median :4.250   Median : 9.340   Median :164.7
##  Mean   :2.076   Mean   :4.325   Mean   : 9.448   Mean   :173.3
##  3rd Qu.:2.150   3rd Qu.:4.470   3rd Qu.: 9.840   3rd Qu.:181.6
##  Max.   :2.330   Max.   :5.810   Max.   :13.040   Max.   :306.0
##
```

```
# display first 10 rows
head(data)
```

```
##     Country  m100  m200  m400 m800 m1500 m3000 Marathon
## 1 Argentina 11.61 22.94 54.50 2.15  4.43  9.79   178.52
## 2 Australia 11.20 22.35 51.08 1.98  4.13  9.08   152.37
## 3   Austria 11.43 23.09 50.62 1.99  4.22  9.34   159.37
## 4   Belgium 11.41 23.04 52.00 2.00  4.14  8.88   157.85
## 5   Bermuda 11.46 23.05 53.30 2.16  4.58  9.81   169.98
## 6    Brazil 11.31 23.17 52.80 2.10  4.49  9.77   168.75
```

```
# Long way to convert data to seconds to be normalized
data$m800 = data$m800 * 60
data$m1500 = data$m1500 * 60
data$m3000 = data$m3000 * 60
data$Marathon = data$Marathon * 60
```

```
# display first 10 rows
head(data)
```

Visually inspect the data again.

```
##     Country  m100  m200  m400  m800 m1500 m3000 Marathon
## 1 Argentina 11.61 22.94 54.50 129.0 265.8 587.4  10711.2
## 2 Australia 11.20 22.35 51.08 118.8 247.8 544.8   9142.2
## 3   Austria 11.43 23.09 50.62 119.4 253.2 560.4   9562.2
## 4   Belgium 11.41 23.04 52.00 120.0 248.4 532.8   9471.0
## 5   Bermuda 11.46 23.05 53.30 129.6 274.8 588.6  10198.8
## 6    Brazil 11.31 23.17 52.80 126.0 269.4 586.2  10125.0
```

```
# display common stats again on the transformed data
summary(data)
```

```
##       Country        m100            m200            m400      
##  Argentina: 1   Min.   :10.79   Min.   :21.71   Min.   :47.99  
##  Australia: 1   1st Qu.:11.27   1st Qu.:22.88   1st Qu.:51.55  
##  Austria  : 1   Median :11.60   Median :23.54   Median :53.30  
##  Belgium  : 1   Mean   :11.62   Mean   :23.64   Mean   :53.41  
##  Bermuda  : 1   3rd Qu.:11.92   3rd Qu.:24.43   3rd Qu.:55.04  
##  Brazil   : 1   Max.   :12.90   Max.   :27.10   Max.   :60.40  
##  (Other)  :49                                                  
##       m800            m1500           m3000          Marathon    
##  Min.   :113.4   Min.   :232.2   Min.   :507.0   Min.   : 8563  
##  1st Qu.:120.0   1st Qu.:246.9   1st Qu.:531.6   1st Qu.: 9177  
##  Median :123.0   Median :255.0   Median :560.4   Median : 9879  
##  Mean   :124.6   Mean   :259.5   Mean   :566.9   Mean   :10395  
##  3rd Qu.:129.0   3rd Qu.:268.2   3rd Qu.:590.4   3rd Qu.:10897  
##  Max.   :139.8   Max.   :348.6   Max.   :782.4   Max.   :18360  
##
```

```
rownames(data) = data[,1]
data[,1] = NULL
```

Naming rows in data to be country names from data[,1] and removing 'Country' field from the dataset

```
# display first 10 rows to make sure Country column is removed
head(data)
```

Visually inspect the datset again

```
##            m100  m200  m400  m800 m1500 m3000 Marathon
## Argentina 11.61 22.94 54.50 129.0 265.8 587.4  10711.2
## Australia 11.20 22.35 51.08 118.8 247.8 544.8   9142.2
## Austria   11.43 23.09 50.62 119.4 253.2 560.4   9562.2
## Belgium   11.41 23.04 52.00 120.0 248.4 532.8   9471.0
## Bermuda   11.46 23.05 53.30 129.6 274.8 588.6  10198.8
## Brazil    11.31 23.17 52.80 126.0 269.4 586.2  10125.0
```

```r
# Running PCA
fit <- princomp(data, cor=TRUE)

# Displaying summary of the fit
summary(fit)
```

Perform Principal Component Analysis (PCA) and display its summary

```
## Importance of components:
##                           Comp.1     Comp.2     Comp.3     Comp.4
## Standard deviation     2.4094991 0.80848347 0.54761522 0.35422802
## Proportion of Variance 0.8293837 0.09337793 0.04284035 0.01792536
## Cumulative Proportion  0.8293837 0.92276161 0.96560196 0.98352731
##                            Comp.5      Comp.6      Comp.7
## Standard deviation     0.231984732 0.197608919 0.149808546
## Proportion of Variance 0.007688131 0.005578469 0.003206086
## Cumulative Proportion  0.991215445 0.996793914 1.000000000
```

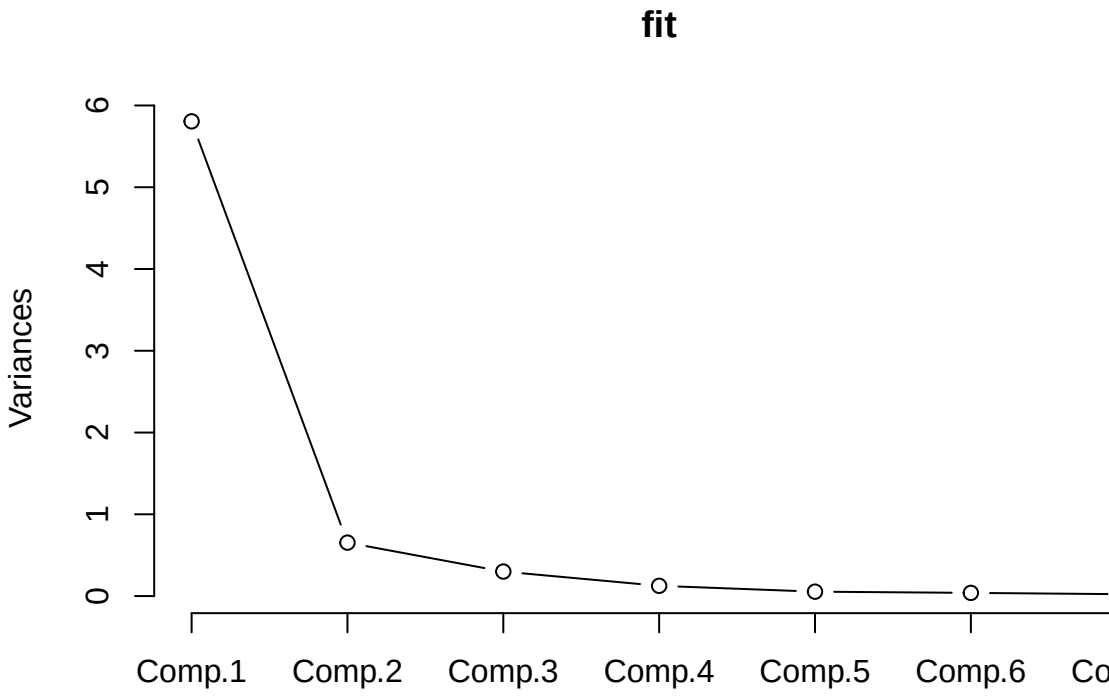
```r
# Display loadings
loadings(fit)
```

Display Loadings

```
##
## Loadings:
##          Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## m100      0.368  0.490  0.286 -0.319 -0.231  0.620
## m200      0.365  0.537  0.230                -0.711 -0.109
## m400      0.382  0.247 -0.515  0.347  0.572  0.191  0.208
## m800      0.385 -0.155 -0.585        -0.620        -0.315
## m1500     0.389 -0.360        -0.430        -0.231  0.693
## m3000     0.389 -0.348  0.153 -0.363  0.463        -0.598
## Marathon  0.367 -0.369  0.484  0.672 -0.131  0.142
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.143  0.143  0.143  0.143  0.143  0.143  0.143
## Cumulative Var  0.143  0.286  0.429  0.571  0.714  0.857  1.000
```

```r
# Plot the fit
plot(fit, type="l")
```

Plot the 'fit' obtained by running PCA. Plot clealy shows componnents 1, 2, and 3 to be the principal components derived (explain more than 99% of the variance in the dataset as seen in



the display of loadings)

```
# Display scores
fit$scores
```

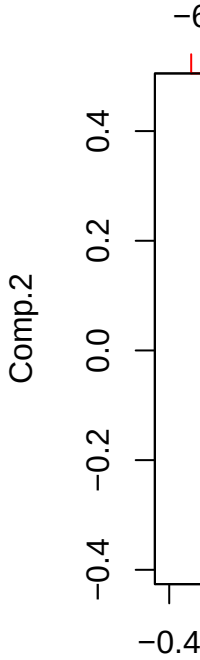
Display the scores for each record

```
##                Comp.1       Comp.2      Comp.3      Comp.4       Comp.5
## Argentina    0.53212195 -0.680937797 -0.62126301 -0.04594894  0.0025811586
## Australia   -2.11284700 -0.537709378  0.04357281 -0.18910495  0.2203680482
## Austria     -1.39316656 -0.277530069  0.53721926 -0.43016371  0.0257389828
## Belgium     -1.52390695  0.091802341  0.08422604  0.03978605  0.0306058833
## Bermuda      0.39139200 -0.985408838 -0.65485261 -0.47882949 -0.2062048001
## Brazil      -0.11948856 -0.919916756 -0.32511042 -0.34410818  0.0968457310
## Burma        1.69754142  0.591594341 -0.07652477  0.14645483 -0.6089939909
## Canada      -2.63217070 -0.701696215 -0.11055186 -0.03359162 -0.1423824032
## Chile        0.55287936  1.182727959  0.23952060  0.09701175  0.2176191364
## China        0.64718368  0.989080326 -0.05420239 -0.02315029  0.0584398887
## Columbia     0.14287696  0.155892166 -0.21654575 -0.17777045 -0.1912689370
## CookIs       6.13329130  1.412402548  0.21479144  0.28344586  0.0534064942
## Costa        2.64336942  0.308776354 -1.10468957 -0.41583379  0.5901095778
## Czech       -3.08194526 -1.022064547  1.05845762 -0.37661064  0.0261044669
## Denmark     -1.12666477  0.545109728 -0.41477324  0.17754024  0.1050907283
## Domrep       2.31659294 -0.621775472 -0.65228802  0.26485362 -0.4000628361
## Finland     -2.20194950 -0.676426245 -0.08161977 -0.08786917 -0.3329832852
```

```
## France      -1.90960964 -0.447922751 -0.14598782  0.05371977  0.1913722401
## GDR         -3.53833101 -1.213583461  0.53088333  0.12545146 -0.0892198174
## FRG         -2.95274367 -0.441166540  0.20306008  0.02608293 -0.0484772004
## Gbni        -2.80880784 -0.583680262 -0.13427581  0.13144851 -0.0421251378
## Greece       0.82175014  0.235868928  0.16138683  0.08774033  0.4590729787
## Guatemala    3.25704356 -0.927528451 -0.44712955  0.09157478 -0.3578559637
## Hungary     -1.49082853  0.059931588  0.15144440 -0.12620125 -0.0932958595
## India        1.02388750  0.255942992  0.51541826 -0.05323981 -0.0513764306
## Indonesia    2.13183387 -0.381756355 -0.33979611  0.18942353 -0.3786096894
## Ireland     -1.12764937  0.517771847 -0.45125716  0.08878343  0.0257994289
## Israel      -0.14428519  0.157303606 -0.75829151  0.16758828  0.2932604006
## Italy       -2.15926044  0.355236135  0.07802937  0.29320309  0.2265506376
## Japan       -0.05977861  0.664038312 -0.40411743 -0.43394737 -0.1242097243
## Kenya       -0.43486576  0.485040815  0.76003818  0.32902859  0.0649744043
## Korea        1.24523372  0.821904700 -0.56153092 -0.24180716 -0.0131019597
## DprKorea     0.46655772  1.733718848  1.93732968 -0.34498766 -0.3406608161
## Luxembourg   1.31374285  1.193070361  0.10608922  0.03180666  0.4536981596
## Malaysia     2.36210756 -0.001271429 -0.11928549 -0.85435662 -0.1289014697
## Mauritius    4.27286963 -1.191080643  0.08853806  1.40696396  0.1655729653
## Mexico      -0.06406697  0.574213788  0.09124219 -0.45631268  0.2989152081
## Netherlands -1.81096545 -0.047519331 -0.14402369  0.10900426  0.3659311379
## NewZealand  -1.52518788  0.381404100 -0.06168639 -0.37241743 -0.2651386357
## Norway      -1.49667849  0.912528976 -0.39098664  0.14663839 -0.1323825474
## PNG          4.01755107 -0.343380195  0.29100630  0.29669938 -0.1610292597
## Philippines  1.65530683 -0.884117095 -0.22419099  0.02248064 -0.2075468646
## Poland      -2.69672473 -0.708796713  0.60146348  0.02405997 -0.0367527726
## Portugal    -0.22635133  1.264207647 -0.46643810  0.02370870 -0.2406682777
## Rumania     -2.04853470  0.623743207  0.84617721  0.47391000  0.0746945195
## Singapore    1.98828979  0.960756220  0.39288468 -0.40700939 -0.0739539451
## Spain       -0.35893084  0.934008388  0.05056545  0.10437043 -0.0935688291
## Sweden      -1.84460096 -0.257169493 -0.25034454  0.15039551  0.0006341491
## Switzerland -1.35906564  0.518920084 -0.24744025  0.22622341  0.0866356489
## Taipei      -0.50472889 -1.246032831 -0.31447126  0.04825215 -0.0094885051
## Thailand     1.97117932 -0.141163113 -0.82110657 -0.66155209  0.1590843033
## Turkey       1.62302658  0.599820487 -0.16253455  0.82274936 -0.1491062390
## USA         -3.36655736 -0.691419036 -0.38475273  0.27307377  0.1972839807
## USSR        -3.49662048 -0.247337282  0.65233230  0.20934990  0.0832371859
## WSamoa       8.40968394 -2.348426494  1.50639215 -0.40801087  0.3457387522
##                   Comp.6       Comp.7
## Argentina    0.462128370 -0.080699873
## Australia    0.139238539 -0.009905621
## Austria     -0.082435556 -0.107151129
## Belgium      0.063456859  0.139767314
## Bermuda     -0.049881900  0.048270306
## Brazil      -0.303217238 -0.006786235
## Burma        0.279882154  0.056340098
## Canada      -0.117509178 -0.118324358
## Chile        0.130070468  0.004004527
## China        0.047048439  0.173928433
## Columbia    -0.327553026 -0.123571003
## CookIs      -0.056219489 -0.291941223
## Costa       -0.066261725 -0.045099878
## Czech        0.047845977  0.189965751
## Denmark     -0.181623092  0.325385154
```

```
## Domrep      -0.006839652  0.324435318
## Finland     -0.031928343 -0.184419113
## France      -0.036137721  0.053714341
## GDR         -0.047587303 -0.177690502
## FRG         -0.193244641  0.087528634
## Gbni         0.012482219  0.060527247
## Greece       0.094738351 -0.122222962
## Guatemala   -0.282280845 -0.030793981
## Hungary      0.062273088 -0.002906953
## India        0.135730171 -0.446306741
## Indonesia   -0.013558797 -0.058987131
## Ireland     -0.150791833 -0.044039586
## Israel      -0.091782995 -0.095867007
## Italy        0.012160659  0.080935631
## Japan       -0.183706704  0.143305470
## Kenya        0.120250076 -0.011156425
## Korea       -0.028441879 -0.027955187
## DprKorea    -0.567062179 -0.072739937
## Luxembourg  -0.116390132  0.124542185
## Malaysia     0.377469165 -0.141452453
## Mauritius   -0.324641495 -0.210982892
## Mexico       0.406538778 -0.128453534
## Netherlands  0.052043082 -0.072629607
## NewZealand   0.078855177  0.200383029
## Norway       0.228463484  0.087093179
## PNG          0.112715839  0.039720948
## Philippines  0.270288501 -0.095656782
## Poland       0.146256176 -0.250884729
## Portugal    -0.041462984  0.041425643
## Rumania     -0.051264108  0.169370959
## Singapore    0.092279870  0.018173587
## Spain        0.125567520 -0.021728569
## Sweden      -0.161316105  0.036668857
## Switzerland  0.047766703  0.069231063
## Taipei       0.024460467 -0.093939922
## Thailand    -0.486457551 -0.032495552
## Turkey       0.289750729  0.193033266
## USA         -0.047483131  0.040740404
## USSR         0.098863036  0.017920516
## WSamoa       0.088455707  0.380377027
```

```r
# Display biplot
biplot(fit)
```



Create a biplot to visualize the directions of the most varying data (in 2D)

End of PCA assignment