Data Mining and Machine Learning in Bioinformatics

Exercise Series 6

Group members (Name, Student ID, E-Mail):

- Baldomero Valdez, Valenzuela, 2905175, baldmer.w@gmail.com
- Omar Trinidad Gutierrez Mendez, 2850441, omar.vpa@gmail.com
- Shinho Kang, 2890169, wis.shinho.kang@gmail.com

Task 1:

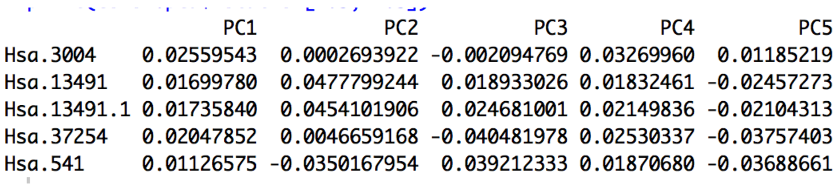
A) Load `colonCA` dataset and use the function prcomp to calculate the PCA

```
library(colonCA)
data(colonCA)

colon.ds = log(exprs(colonCA))
colon.ds = t(colon.ds)

colon.pca = prcomp(colon.ds,
                   center = TRUE,
                   scale. = TRUE
                   )
print(colon.pca$rotation[1:5, 1:5])

> print(colon.pca$rotation[1:5, 1:5])
```

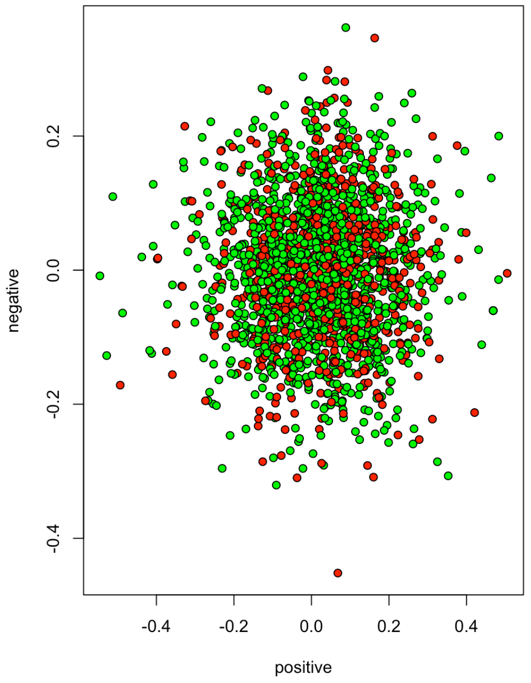


```
                  PC1           PC2          PC3        PC4         PC5
Hsa.3004   0.02559543  0.0002693922 -0.002094769 0.03269960  0.01185219
Hsa.13491  0.01699780  0.0477799244  0.018933026 0.01832461 -0.02457273
Hsa.13491.1 0.01735840  0.0454101906  0.024681001 0.02149836 -0.02104313
Hsa.37254  0.02047852  0.0046659168 -0.040481978 0.02530337 -0.03757403
Hsa.541    0.01126575 -0.0350167954  0.039212333 0.01870680 -0.03688661
```

B) 2D PCA plot from normal patients and patients with cancer

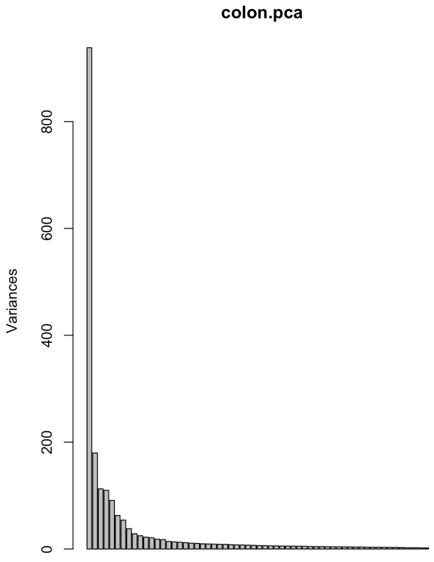
```
negative = apply(colon.pca$rotation[, colonCA$class == 'n'], 1, sum)
positive = apply(colon.pca$rotation[, colonCA$class == 't'], 1, sum)

plot(negative ~ positive,
     pch = 21,
     bg = c('red', 'green')[unclass(colonCA$class)])
```



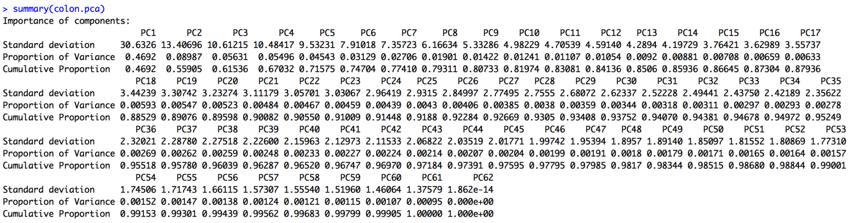
C) screeplot of eigenvalues

```
# colon.pca contains a `sdev` component
screeplot(colon.pca, npcs = 62)
```



D) principal components analysis

```
summary(colon.pca)
```



```
> summary(colon.pca)
Importance of components:
                           PC1      PC2      PC3      PC4     PC5     PC6     PC7     PC8     PC9    PC10    PC11    PC12   PC13    PC14    PC15    PC16    PC17
Standard deviation     30.6326 13.40696 10.61215 10.48417 9.53231 7.91018 7.35723 6.16634 5.33286 4.98229 4.70539 4.59140 4.2894 4.19729 3.76421 3.62989 3.55737
Proportion of Variance  0.4692  0.08987  0.05631  0.05496 0.04543 0.03129 0.02706 0.01901 0.01422 0.01241 0.01107 0.01054 0.0092 0.00881 0.00708 0.00659 0.00633
Cumulative Proportion   0.4692  0.55905  0.61536  0.67032 0.71575 0.74704 0.77410 0.79311 0.80733 0.81974 0.83081 0.84136 0.8506 0.85936 0.86645 0.87304 0.87936
                          PC18    PC19    PC20    PC21    PC22    PC23    PC24   PC25    PC26    PC27   PC28    PC29    PC30    PC31    PC32    PC33    PC34    PC35
Standard deviation     3.44239 3.30742 3.23274 3.11179 3.05701 3.03067 2.96419 2.9315 2.84997 2.77495 2.7555 2.68072 2.62337 2.52228 2.49441 2.43750 2.42189 2.35622
Proportion of Variance 0.00593 0.00547 0.00523 0.00484 0.00467 0.00459 0.00439 0.0043 0.00406 0.00385 0.0038 0.00359 0.00344 0.00318 0.00311 0.00297 0.00293 0.00278
Cumulative Proportion  0.88529 0.89076 0.89598 0.90082 0.90550 0.91009 0.91448 0.9188 0.92284 0.92669 0.9305 0.93408 0.93752 0.94070 0.94381 0.94678 0.94972 0.95249
                          PC36    PC37    PC38    PC39    PC40    PC41    PC42    PC43    PC44    PC45    PC46    PC47   PC48    PC49    PC50    PC51    PC52    PC53
Standard deviation     2.32021 2.28780 2.27518 2.22600 2.15963 2.12973 2.11533 2.06822 2.03519 2.01771 1.99742 1.95394 1.8957 1.89140 1.85097 1.81552 1.80869 1.77310
Proportion of Variance 0.00269 0.00262 0.00259 0.00248 0.00233 0.00227 0.00224 0.00214 0.00207 0.00204 0.00199 0.00191 0.0018 0.00179 0.00171 0.00165 0.00164 0.00157
Cumulative Proportion  0.95518 0.95780 0.96039 0.96287 0.96520 0.96747 0.96970 0.97184 0.97391 0.97595 0.97795 0.97985 0.9817 0.98344 0.98515 0.98680 0.98844 0.99001
                          PC54    PC55    PC56    PC57    PC58    PC59    PC60    PC61      PC62
Standard deviation     1.74506 1.71743 1.66115 1.57307 1.55540 1.51960 1.46064 1.37579 1.862e-14
Proportion of Variance 0.00152 0.00147 0.00138 0.00124 0.00121 0.00115 0.00107 0.00095 0.000e+00
Cumulative Proportion  0.99153 0.99301 0.99439 0.99562 0.99683 0.99799 0.99905 1.00000 1.000e+00
```

- Which proportion of the overall variance do the first 2 principal components explain? => 0.55905
- How many principal components would you need to explain 90% and 95% of the overall variance?
  - To explain 90% we need the first 21 PCA whose cumulative proportion is 0.90082
  - To explain 95% we need the first 34 PCA whose cumulative proportion is 0.94972