Post-assignment

Boan Fan

10/24/2017

Introduction

Among the topics covered so far in this class, one particular method of analyzing data has caught my attention, the Principle Component Analysis (PCA). First introduced by Karl Pearson in 1904 and computationally implemented in 1960s, it is an way to condense information in variables while keeping as much variation in the data as possible. It is usually a hard task to rank different groups or to see which influencing factor plays the most important role in the data. PCA provides a way to agregate all the variables to one single numerical value and makes it much easier to compare data points.

Motivation

I was initially very confused with this method. What are the outputs like? What information can we collect from the data given the outputs of the prcomp function? And how can we use PCA on real examples. Therefore, I want to research deeper into PCA to be able to fully grasp its importance and functionality.

Examples

For demonstrational purposes, I will use the iris dataset. It gives the measurements in centimeters of the variables sepal length and width and petal length and width for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica. The data for the first 5 flowers gives a general idea what the data set looks like

```
data(iris)
head(iris, 5)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
```

Before the analysis, we have to perform a transformation on the data since skewness and magnitude have a great impact on the results of PCA. To remove the effect of skewness, we apply a log transformation; to remove the effect of magnitude, we standardize the values.

```
log.data <- log(iris[, 1:4])
```

After data transformation, we can apply the prcomp function, which has an output of class prcomp. Now, for any principle component analysis, we are expecting three main output: eigenvalues, scores and loadings. Scores (PCs) give coordinates to represents the objects. Loadings provide the weights of each contributing variable to every PC. And eigenvalues tell the amount of variability captured by each PC.

For function prcomp, eigenvalues are stored in sdev. The best way to visually present the variances and their respective proportion is through a table eigs.

```
pca <- prcomp(log.data)
eigs <- data.frame(
  eigenvalue = pca$sdev^2,
  prop = round((pca$sdev^2)/sum(pca$sdev^2), 4)
)
eigs
```

```
##    eigenvalue   prop
## 1 1.314598414 0.9702
## 2 0.019141453 0.0141
## 3 0.018294581 0.0135
## 4 0.002895452 0.0021
```

And scores are stored in x.

```
x <- pca$x
x
```

```
##                PC1          PC2           PC3           PC4
##   [1,] -1.6736850  0.020556829 -6.042345e-02  0.0152177314
##   [2,] -1.6688436 -0.067970291  6.921821e-02  0.0415883446
##   [3,] -1.7142099  0.020062520  7.515646e-02  0.0022096017
##   [4,] -1.6422467 -0.096482818  4.725452e-02 -0.0468650526
##   [5,] -1.6773055  0.036758216 -7.085250e-02 -0.0130909947
##   [6,] -0.9833353  0.257579515 -6.701669e-02 -0.0026976389
##   [7,] -1.3357117  0.184338124  1.169346e-01 -0.0429463541
##   [8,] -1.6391536 -0.043484509 -5.949231e-02 -0.0111992590
##   [9,] -1.6777510 -0.087355218  1.460744e-01 -0.0382345131
##  [10,] -2.2285854 -0.404841756 -1.651759e-01 -0.0258972307
##  [11,] -1.6362581  0.005028942 -1.575174e-01  0.0219985030
##  [12,] -1.6106631 -0.087793683 -6.707655e-02 -0.0659635400
```

```
##  [13,] -2.2636375 -0.376257686 -1.022216e-01 -0.0100414377
##  [14,] -2.3965886 -0.210462789  5.454271e-02 -0.0328523738
##  [15,] -1.7462857  0.203078282 -1.528773e-01  0.1201121080
##  [16,] -1.0480685  0.412834306 -1.261652e-01  0.0340066577
##  [17,] -1.1188812  0.441902070  4.751005e-02  0.0779355848
##  [18,] -1.3269699  0.200903692  4.576239e-02  0.0349698666
##  [19,] -1.2223823  0.114652014 -1.502561e-01  0.0404123014
##  [20,] -1.2968462  0.200745808 -4.241478e-02 -0.0183259507
##  [21,] -1.5681470 -0.129549049 -1.505728e-01  0.0178542532
##  [22,] -1.0493120  0.313382741  5.196798e-02  0.0062464619
##  [23,] -1.8557282  0.268017433  1.135804e-01  0.0158071617
##  [24,] -0.7886704  0.260905932  1.386676e-01  0.0247914862
##  [25,] -1.5238323 -0.205870950 -1.404425e-01 -0.1176172875
##  [26,] -1.5993356 -0.159736147  2.328927e-03  0.0189546426
##  [27,] -1.0138307  0.220476465  9.448092e-02  0.0031686271
##  [28,] -1.6368657 -0.026864332 -9.937627e-02  0.0113028937
##  [29,] -1.6700562  0.003886589 -4.922315e-02  0.0435166328
##  [30,] -1.6092959 -0.122605226 -1.348853e-02 -0.0602014269
##  [31,] -1.6053431 -0.140863179 -1.116530e-03 -0.0293926286
##  [32,] -1.0386747  0.264754785  8.438744e-02  0.0892415586
##  [33,] -2.2386917 -0.244267888 -3.938835e-01 -0.0851050260
##  [34,] -1.6765667  0.125238119 -2.275470e-01  0.0084511134
##  [35,] -1.6358720 -0.096536757  1.635003e-02  0.0078692675
##  [36,] -1.7484098  0.075006553  7.906127e-02  0.0798734213
##  [37,] -1.7035108  0.071411502 -6.572057e-02  0.1029077793
##  [38,] -2.2720574 -0.271529536 -2.424961e-01 -0.0643598777
##  [39,] -1.7171481 -0.016959264  1.535047e-01 -0.0293813108
##  [40,] -1.6371555 -0.043501415 -6.917893e-02  0.0059565211
##  [41,] -1.3664125  0.251839736  8.708704e-02  0.0400890284
##  [42,] -1.3528628  0.010719001  4.384255e-01  0.1150327130
##  [43,] -1.7208650  0.020118832  1.074204e-01 -0.0549323307
##  [44,] -0.6687851  0.417476990  1.799679e-01  0.0114444980
##  [45,] -0.9314076  0.166282651 -6.799337e-02 -0.0753639472
##  [46,] -1.3242089  0.112394177  1.854901e-01  0.0434771957
##  [47,] -1.6109519 -0.023945081 -1.761533e-01 -0.0574766980
##  [48,] -1.6789353 -0.030838256  5.403837e-02 -0.0386969955
##  [49,] -1.6381441  0.005044900 -1.483740e-01  0.0058047870
##  [50,] -1.6722943 -0.013230784 -8.721225e-03  0.0213571053
##  [51,]  0.6393130  0.002192084 -1.587653e-01  0.0558106888
##  [52,]  0.6672955  0.062834292 -7.829792e-02 -0.0053923095
##  [53,]  0.7197418 -0.013981321 -1.287790e-01  0.0467498232
##  [54,]  0.4891452 -0.109485660  2.444536e-01  0.0224893870
##  [55,]  0.6876556 -0.028995782  8.421056e-05  0.0542988062
##  [56,]  0.5409323 -0.077431919  3.623532e-02 -0.0598474716
##  [57,]  0.7410932  0.079354135 -9.423007e-02 -0.0411448457
##  [58,]  0.1534901 -0.069455538  2.839850e-01 -0.0493923712
##  [59,]  0.5648089 -0.072498302 -6.991746e-02  0.0466617104
##  [60,]  0.5248289  0.033039059  1.876126e-01 -0.0783632513
##  [61,]  0.1957593 -0.214647491  3.791710e-01  0.0226058036
##  [62,]  0.6279446  0.073230228  3.703161e-02 -0.0295766811
##  [63,]  0.2761359 -0.251795002  1.649228e-01  0.1026882920
##  [64,]  0.6310964 -0.054245334 -2.115465e-02 -0.0244432614
##  [65,]  0.4243776  0.096064123  1.150999e-01 -0.0220028844
##  [66,]  0.6033952  0.029308821 -8.650982e-02  0.0502574653
##  [67,]  0.6575391  0.025870199  3.310440e-02 -0.0955246873
##  [68,]  0.2733969 -0.151075543  2.472898e-02 -0.0151826958
##  [69,]  0.6856718 -0.152404141  2.047855e-01  0.1154449784
##  [70,]  0.3305200 -0.118505675  1.431597e-01  0.0045603306
##  [71,]  0.8476011  0.099654538 -1.831212e-02 -0.0863820369
##  [72,]  0.4882634  0.003438186  5.334304e-02  0.0343123981
##  [73,]  0.7229516 -0.137487525  6.932283e-02  0.0531007880
##  [74,]  0.5013025 -0.142970382 -3.646735e-02 -0.0180599019
##  [75,]  0.5276280 -0.026133537 -2.607335e-02  0.0402741418
##  [76,]  0.6037663  0.010483543 -5.574004e-02  0.0502111743
##  [77,]  0.6547163 -0.088964102 -5.822460e-02  0.0772350885
##  [78,]  0.8358978  0.008995843 -6.682348e-02  0.0342739224
##  [79,]  0.6664530  0.006334479  2.356368e-02 -0.0223317611
##  [80,]  0.1938698 -0.064028179  1.277338e-01  0.0322497388
##  [81,]  0.3179284 -0.124095381  1.922124e-01  0.0129193305
##  [82,]  0.2229535 -0.148164814  1.786371e-01  0.0162921041
##  [83,]  0.4040322 -0.035618865  9.382689e-02  0.0087308954
##  [84,]  0.7889970 -0.092012227  3.805698e-02 -0.0285177430
##  [85,]  0.6538696  0.025901248  5.089394e-02 -0.1270313806
##  [86,]  0.7124792  0.126425004 -7.311616e-02 -0.0821620359
##  [87,]  0.6957179  0.014676847 -9.660016e-02  0.0337931754
##  [88,]  0.5510050 -0.175088646  1.373355e-01  0.1114917242
##  [89,]  0.4881371  0.026182226  3.537021e-02 -0.0745152343
##  [90,]  0.4843430 -0.061582002  1.849142e-01 -0.0105216754
##  [91,]  0.4617967 -0.140138516  9.525657e-02 -0.0585963280
##  [92,]  0.6182775 -0.019991734 -3.618097e-02 -0.0314007913
##  [93,]  0.4189981 -0.074696813  1.099671e-01  0.0160625406
##  [94,]  0.1579797 -0.093923756  3.044928e-01 -0.0150404903
##  [95,]  0.5063809 -0.050905840  1.003162e-01 -0.0400457495
##  [96,]  0.4336539 -0.025992370 -4.537474e-03 -0.0703237704
##  [97,]  0.5040513 -0.009866984  4.063235e-02 -0.0530027696
```

```
##  [98,]  0.5244245 -0.026106431 -1.054321e-02  0.0127690195
##  [99,]  0.1885183  0.061843065  3.009168e-01  0.0023950175
## [100,]  0.4938965 -0.013470055  7.597731e-02 -0.0318668954
## [101,]  1.2461009  0.110071969 -8.160560e-02 -0.0928033743
## [102,]  0.9325263 -0.015545992  9.964546e-02 -0.0495163213
## [103,]  1.1060697 -0.010789437 -1.105105e-01  0.0450550287
## [104,]  0.9377771 -0.062872374 -4.591725e-02 -0.0369135325
## [105,]  1.1283031  0.021723066 -4.784058e-02 -0.0240319072
## [106,]  1.1695858 -0.087882652 -1.916643e-01  0.0703128064
## [107,]  0.7615938 -0.023090227  2.613893e-01 -0.1329290684
## [108,]  1.0121545 -0.143926161 -1.682661e-01  0.0553170148
## [109,]  0.9702669 -0.172304899  1.497075e-02  0.0646288858
## [110,]  1.2629147  0.148589784 -2.161115e-01 -0.0165363577
## [111,]  0.9780994  0.104780357 -6.397627e-02 -0.0155668787
## [112,]  0.9618948 -0.042059983  3.507073e-02  0.0242041418
## [113,]  1.0662415  0.037484341 -5.942104e-02  0.0287548968
## [114,]  0.9690593 -0.023325101  1.849946e-01 -0.0256634226
## [115,]  1.1301969  0.109257194  1.348574e-01 -0.0525339040
## [116,]  1.1154820  0.140528373 -3.621245e-02 -0.0337522545
## [117,]  0.9298738 -0.031041833 -7.771998e-02 -0.0178440905
## [118,]  1.2046681  0.058273368 -3.610912e-01 -0.0142032517
## [119,]  1.2793969 -0.160186023 -9.102916e-02  0.1293623824
## [120,]  0.7355988 -0.224768717  1.758446e-01  0.0553692407
## [121,]  1.1598352  0.090471641 -1.040709e-01  0.0095471156
## [122,]  0.9505386  0.055679753  1.213540e-01 -0.0797920491
## [123,]  1.1407558 -0.159564797 -1.679911e-01  0.1020553689
## [124,]  0.8744232 -0.012177732  6.211565e-02  0.0315134010
## [125,]  1.0773047  0.067712066 -1.354799e-01 -0.0325495406
## [126,]  0.9804411 -0.053836047 -2.109816e-01  0.0190596310
## [127,]  0.8602960  0.022896930  5.277644e-02  0.0094512391
## [128,]  0.8651001  0.048380594  2.662668e-03 -0.0381480149
## [129,]  1.0732021 -0.014481476  1.180657e-02 -0.0018679371
## [130,]  0.8663118 -0.120009319 -1.812699e-01  0.0490628092
## [131,]  1.0454810 -0.117883322 -1.219319e-01  0.0933275423
## [132,]  1.1026091  0.047333906 -3.790380e-01  0.0171379258
## [133,]  1.1129816  0.006210166  2.398954e-02  0.0003982743
## [134,]  0.7366382 -0.099866362 -2.867956e-02 -0.0037909937
## [135,]  0.7259110 -0.237363203 -1.797767e-02 -0.0338727132
## [136,]  1.2088895  0.006699396 -1.406014e-01  0.1097488199
## [137,]  1.1746144  0.156470192 -8.415881e-02 -0.0858734228
## [138,]  0.9264210 -0.012190480 -9.354985e-02 -0.0442575678
## [139,]  0.8530139  0.062562093  1.955086e-02 -0.0462703620
## [140,]  1.0565547  0.068917575 -8.214246e-02  0.0339360748
## [141,]  1.1861457  0.103348141 -4.831034e-02  0.0040274665
## [142,]  1.1054647  0.148654120 -3.391622e-02  0.0555480855
## [143,]  0.9325263 -0.015545992  9.964546e-02 -0.0495163213
## [144,]  1.1757870  0.066788862 -1.116525e-01 -0.0134660412
## [145,]  1.2263951  0.145262725 -8.981912e-02 -0.0240559573
## [146,]  1.1141967  0.116499067 -4.404227e-03  0.0372107293
## [147,]  0.9352966 -0.046225296  1.226050e-01  0.0585439986
## [148,]  0.9916278  0.054360212 -2.618188e-02  0.0041475663
## [149,]  1.1182316  0.162541744 -7.195189e-02 -0.0908772430
## [150,]  0.8819499  0.020921633  1.890457e-03 -0.0790532121
```

And loadings are stored in rotation.

```
pca$rotation
```

```
##                     PC1           PC2        PC3         PC4
## Sepal.Length  0.10090019 -0.0008537483 -0.4891583  0.86633858
## Sepal.Width  -0.05759298  0.5745110809 -0.7140592 -0.39590340
## Petal.Length  0.50527032 -0.6870939247 -0.4269180 -0.30057416
## Petal.Width   0.85510473  0.4447900940  0.2618865  0.04871476
```

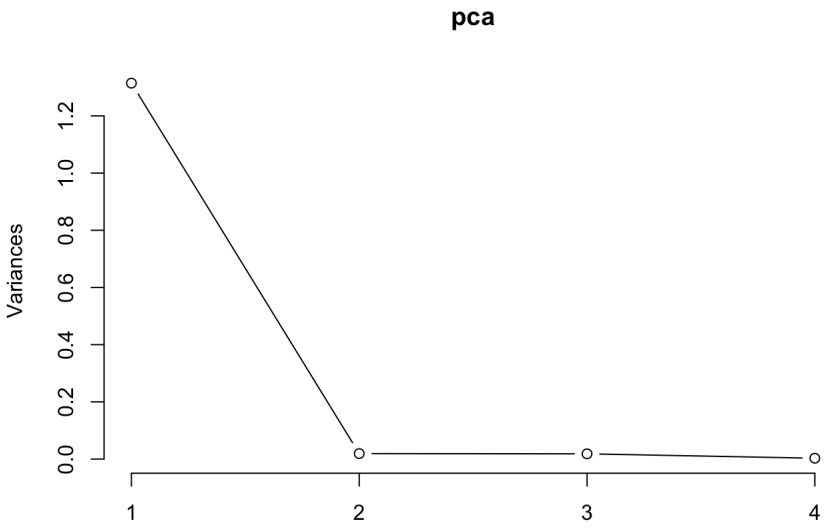
The third line of the summary function gives the cummulative proportion of variances given by each PC. We can decide on which PCs to keep for further analysis

```
summary(pca)
```

```
## Importance of components%s:
##                           PC1     PC2    PC3     PC4
## Standard deviation     1.1466 0.13835 0.1353 0.05381
## Proportion of Variance 0.9702 0.01413 0.0135 0.00214
## Cumulative Proportion  0.9702 0.98436 0.9979 1.00000
```

To see how many PCs are actually useful for further analysis, we can plot the correlation between the variances and PCs.

```
plot(pca, type = "l")
```



from the plot, we can see that PC1 captures most of the variance, which means it is the most useful PC to differentiate and rank the groups.

During my research, a very good way to visually show the relations between PC1 and PC2 is through a function ggbiplot.

```
library(devtools)
install_github("ggbiplot", "vqv")
```

```
## Warning: Username parameter is deprecated. Please use vqv/ggbiplot
```

```
## Skipping install of 'ggbiplot' from a github remote, the SHA1 (7325e880) has not changed since last install.
##   Use `force = TRUE` to force installation
```

```
library(ggbiplot)
```

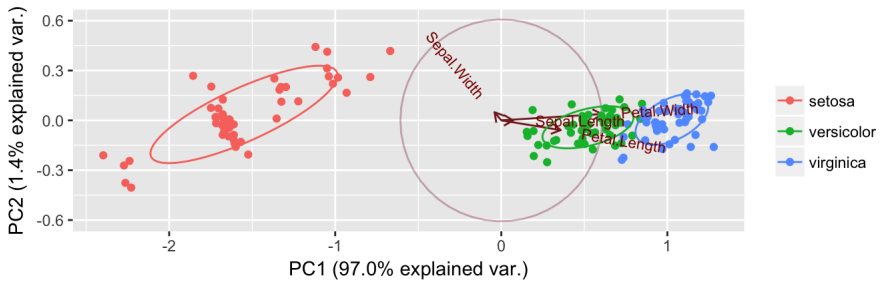
```
## Loading required package: ggplot2
```

```
## Loading required package: plyr
```

```
## Loading required package: scales
```

```
## Loading required package: grid
```

```
species <- iris[, 5]
g <- ggbiplot(pca, obs.scale = 1, var.scale = 1, groups = species, ellipse = TRUE, circle = TRUE)
g <- g + scale_color_discrete(name = '')
g
```



Each point is colored based on the

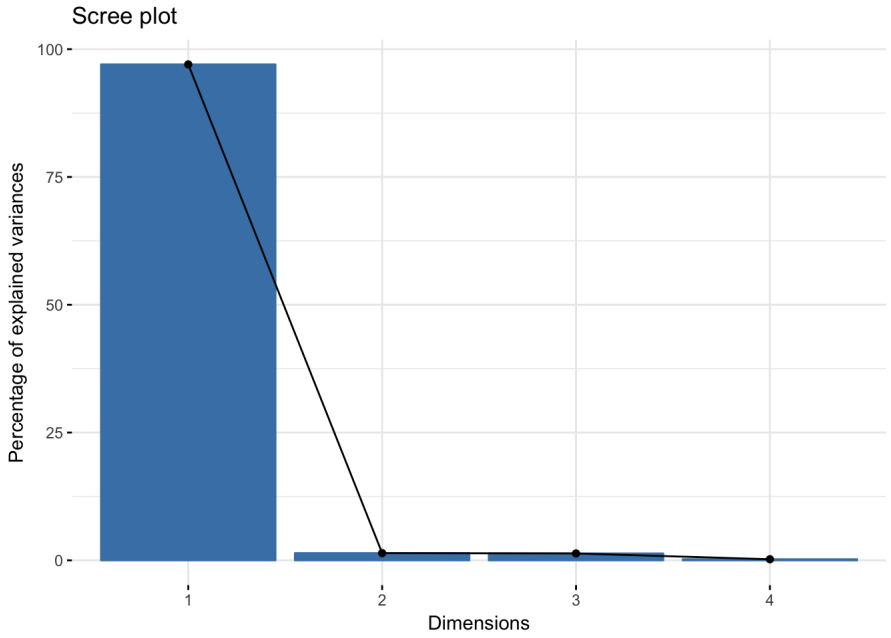
species of the flower and contour line is drawn to capture 68% probablity for each group. In the homework assignment, we firstly computed the sum of each variable fro all the players of a NBA team and then perform the PCA on the team data. We can also compare the teams without losing the information for individual players by performing PCA on all players, and use ggbiplot to see where most players of a team lie on a plot. This method will greatly reduce the effect of an outlier (if one player is very successful while all other players are lame, then the team data will be very good, and the ranking will not be very reliable).

Another way to visualize PCA in R is through the factoextra package. We can directly visualize eigenvalues and see the percentage explained by each principal component.

```
library(factoextra)
```

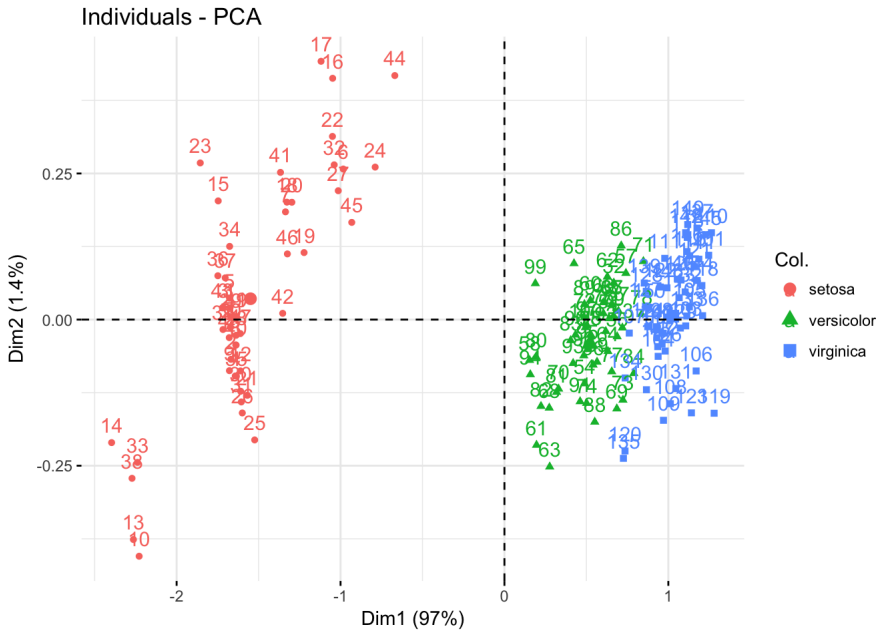
```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

```
fviz_eig(pca)
```



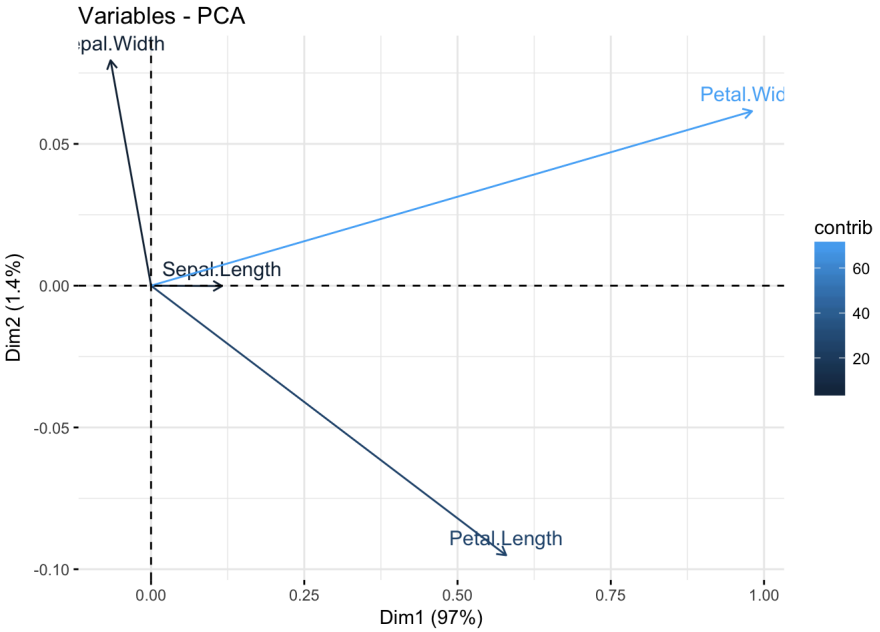
We can also get a graph of individual data points. Individuals with a similar profile are grouped together.

```
fviz_pca_ind(pca, col.ind = species)
```



From the graph of individual data points, we can see that even within the same species, there are patterns of certain data points closer than others. I think this suggests the existence of sub-species.

```
fviz_pca_var(pca, col.var = "contrib")
```



We can also get a graph of all the variables. The longer a line is, the more contribution the variable has on the PC; the shorter a line is, the less contribution the variable has on the PC. Also, positive correlated variables point to the same side of the plot, negative correlated variables point to opposite sides of the graph.

Another built-in function in R that performs PCA is princomp. It adopts a different approach from prcomp. PCs from prcomp are calculated based on the covariances and correlations between the individual data points while PCs from princomp are calculated based on covariances and correlations between the variables.

```
pca2 <- princomp(log.data)
```

The scores on stored in the element of output scores

```
pca2$scores
```

```
##            Comp.1       Comp.2        Comp.3        Comp.4
##   [1,] -1.6736850  0.020556829 -6.042345e-02  0.0152177314
##   [2,] -1.6688436 -0.067970291  6.921821e-02  0.0415883446
##   [3,] -1.7142099  0.020062520  7.515646e-02  0.0022096017
##   [4,] -1.6422467 -0.096482818  4.725452e-02 -0.0468650526
##   [5,] -1.6773055  0.036758216 -7.085250e-02 -0.0130909947
##   [6,] -0.9833353  0.257579515 -6.701669e-02 -0.0026976389
##   [7,] -1.3357117  0.184338124  1.169346e-01 -0.0429463541
##   [8,] -1.6391536 -0.043484509 -5.949231e-02 -0.0111992590
##   [9,] -1.6777510 -0.087355218  1.460744e-01 -0.0382345131
##  [10,] -2.2285854 -0.404841756 -1.651759e-01 -0.0258972307
##  [11,] -1.6362581  0.005028942 -1.575174e-01  0.0219985030
##  [12,] -1.6106631 -0.087793683 -6.707655e-02 -0.0659635400
##  [13,] -2.2636375 -0.376257686 -1.022216e-01 -0.0100414377
##  [14,] -2.3965886 -0.210462789  5.454271e-02 -0.0328523738
##  [15,] -1.7462857  0.203078282 -1.528773e-01  0.1201121080
##  [16,] -1.0480685  0.412834306 -1.261652e-01  0.0340066577
##  [17,] -1.1188812  0.441902070  4.751005e-02  0.0779355848
##  [18,] -1.3269699  0.200903692  4.576239e-02  0.0349698666
##  [19,] -1.2223823  0.114652014 -1.502561e-01  0.0404123014
##  [20,] -1.2968462  0.200745808 -4.241478e-02 -0.0183259507
##  [21,] -1.5681470 -0.129549049 -1.505728e-01  0.0178542532
##  [22,] -1.0493120  0.313382741  5.196798e-02  0.0062464619
##  [23,] -1.8557282  0.268017433  1.135804e-01  0.0158071617
##  [24,] -0.7886704  0.260905932  1.386676e-01  0.0247914862
##  [25,] -1.5238323 -0.205870950 -1.404425e-01 -0.1176172875
##  [26,] -1.5993356 -0.159736147  2.328927e-03  0.0189546426
##  [27,] -1.0138307  0.220476465  9.448092e-02  0.0031686271
##  [28,] -1.6368657 -0.026864332 -9.937627e-02  0.0113028937
##  [29,] -1.6700562  0.003886589 -4.922315e-02  0.0435166328
##  [30,] -1.6092959 -0.122605226 -1.348853e-02 -0.0602014269
##  [31,] -1.6053431 -0.140863179 -1.116530e-03 -0.0293926286
##  [32,] -1.0386747  0.264754785  8.438744e-02  0.0892415586
##  [33,] -2.2386917 -0.244267888 -3.938835e-01 -0.0851050260
##  [34,] -1.6765667  0.125238119 -2.275470e-01  0.0084511134
##  [35,] -1.6358720 -0.096536757  1.635003e-02  0.0078692675
##  [36,] -1.7484098  0.075006553  7.906127e-02  0.0798734213
##  [37,] -1.7035108  0.071411502 -6.572057e-02  0.1029077793
##  [38,] -2.2720574 -0.271529536 -2.424961e-01 -0.0643598777
##  [39,] -1.7171481 -0.016959264  1.535047e-01 -0.0293813108
##  [40,] -1.6371555 -0.043501415 -6.917893e-02  0.0059565211
##  [41,] -1.3664125  0.251839736  8.708704e-02  0.0400890284
##  [42,] -1.3528628  0.010719001  4.384255e-01  0.1150327130
##  [43,] -1.7208650  0.020118832  1.074204e-01 -0.0549323307
##  [44,] -0.6687851  0.417476990  1.799679e-01  0.0114444980
```

```
##  [45,] -0.9314076  0.166282651 -6.799337e-02 -0.0753639472
##  [46,] -1.3242089  0.112394177  1.854901e-01  0.0434771957
##  [47,] -1.6109519 -0.023945081 -1.761533e-01 -0.0574766980
##  [48,] -1.6789353 -0.030838256  5.403837e-02 -0.0386969955
##  [49,] -1.6381441  0.005044900 -1.483740e-01  0.0058047870
##  [50,] -1.6722943 -0.013230784 -8.721225e-03  0.0213571053
##  [51,]  0.6393130  0.002192084 -1.587653e-01  0.0558106888
##  [52,]  0.6672955  0.062834292 -7.829792e-02 -0.0053923095
##  [53,]  0.7197418 -0.013981321 -1.287790e-01  0.0467498232
##  [54,]  0.4891452 -0.109485660  2.444536e-01  0.0224893870
##  [55,]  0.6876556 -0.028995782  8.421056e-05  0.0542988062
##  [56,]  0.5409323 -0.077431919  3.623532e-02 -0.0598474716
##  [57,]  0.7410932  0.079354135 -9.423007e-02 -0.0411448457
##  [58,]  0.1534901 -0.069455538  2.839850e-01 -0.0493923712
##  [59,]  0.5648089 -0.072498302 -6.991746e-02  0.0466617104
##  [60,]  0.5248289  0.033039059  1.876126e-01 -0.0783632513
##  [61,]  0.1957593 -0.214647491  3.791710e-01  0.0226058036
##  [62,]  0.6279446  0.073230228  3.703161e-02 -0.0295766811
##  [63,]  0.2761359 -0.251795002  1.649228e-01  0.1026882920
##  [64,]  0.6310964 -0.054245334 -2.115465e-02 -0.0244432614
##  [65,]  0.4243776  0.096064123  1.150999e-01 -0.0220028844
##  [66,]  0.6033952  0.029308821 -8.650982e-02  0.0502574653
##  [67,]  0.6575391  0.025870199  3.310440e-02 -0.0955246873
##  [68,]  0.2733969 -0.151075543  2.472898e-02 -0.0151826958
##  [69,]  0.6856718 -0.152404141  2.047855e-01  0.1154449784
##  [70,]  0.3305200 -0.118505675  1.431597e-01  0.0045603306
##  [71,]  0.8476011  0.099654538 -1.831212e-02 -0.0863820369
##  [72,]  0.4882634  0.003438186  5.334304e-02  0.0343123981
##  [73,]  0.7229516 -0.137487525  6.932283e-02  0.0531007880
##  [74,]  0.5013025 -0.142970382 -3.646735e-02 -0.0180599019
##  [75,]  0.5276280 -0.026133537 -2.607335e-02  0.0402741418
##  [76,]  0.6037663  0.010483543 -5.574004e-02  0.0502111743
##  [77,]  0.6547163 -0.088964102 -5.822460e-02  0.0772350885
##  [78,]  0.8358978  0.008995843 -6.682348e-02  0.0342739224
##  [79,]  0.6664530  0.006334479  2.356368e-02 -0.0223317611
##  [80,]  0.1938698 -0.064028179  1.277338e-01  0.0322497388
##  [81,]  0.3179284 -0.124095381  1.922124e-01  0.0129193305
##  [82,]  0.2229535 -0.148164814  1.786371e-01  0.0162921041
##  [83,]  0.4040322 -0.035618865  9.382689e-02  0.0087308954
##  [84,]  0.7889970 -0.092012227  3.805698e-02 -0.0285177430
##  [85,]  0.6538696  0.025901248  5.089394e-02 -0.1270313806
##  [86,]  0.7124792  0.126425004 -7.311616e-02 -0.0821620359
##  [87,]  0.6957179  0.014676847 -9.660016e-02  0.0337931754
##  [88,]  0.5510050 -0.175088646  1.373355e-01  0.1114917242
##  [89,]  0.4881371  0.026182226  3.537021e-02 -0.0745152343
##  [90,]  0.4843430 -0.061582002  1.849142e-01 -0.0105216754
##  [91,]  0.4617967 -0.140138516  9.525657e-02 -0.0585963280
##  [92,]  0.6182775 -0.019991734 -3.618097e-02 -0.0314007913
##  [93,]  0.4189981 -0.074696813  1.099671e-01  0.0160625406
##  [94,]  0.1579797 -0.093923756  3.044928e-01 -0.0150404903
##  [95,]  0.5063809 -0.050905840  1.003162e-01 -0.0400457495
##  [96,]  0.4336539 -0.025992370 -4.537474e-03 -0.0703237704
##  [97,]  0.5040513 -0.009866984  4.063235e-02 -0.0530027696
##  [98,]  0.5244245 -0.026106431 -1.054321e-02  0.0127690195
##  [99,]  0.1885183  0.061843065  3.009168e-01  0.0023950175
## [100,]  0.4938965 -0.013470055  7.597731e-02 -0.0318668954
## [101,]  1.2461009  0.110071969 -8.160560e-02 -0.0928033743
## [102,]  0.9325263 -0.015545992  9.964546e-02 -0.0495163213
## [103,]  1.1060697 -0.010789437 -1.105105e-01  0.0450550287
## [104,]  0.9377771 -0.062872374 -4.591725e-02 -0.0369135325
## [105,]  1.1283031  0.021723066 -4.784058e-02 -0.0240319072
## [106,]  1.1695858 -0.087882652 -1.916643e-01  0.0703128064
## [107,]  0.7615938 -0.023090227  2.613893e-01 -0.1329290684
## [108,]  1.0121545 -0.143926161 -1.682661e-01  0.0553170148
## [109,]  0.9702669 -0.172304899  1.497075e-02  0.0646288858
## [110,]  1.2629147  0.148589784 -2.161115e-01 -0.0165363577
## [111,]  0.9780994  0.104780357 -6.397627e-02 -0.0155668787
## [112,]  0.9618948 -0.042059983  3.507073e-02  0.0242041418
## [113,]  1.0662415  0.037484341 -5.942104e-02  0.0287548968
## [114,]  0.9690593 -0.023325101  1.849946e-01 -0.0256634226
## [115,]  1.1301969  0.109257194  1.348574e-01 -0.0525339040
## [116,]  1.1154820  0.140528373 -3.621245e-02 -0.0337522545
## [117,]  0.9298738 -0.031041833 -7.771998e-02 -0.0178440905
## [118,]  1.2046681  0.058273368 -3.610912e-01 -0.0142032517
## [119,]  1.2793969 -0.160186023 -9.102916e-02  0.1293623824
## [120,]  0.7355988 -0.224768717  1.758446e-01  0.0553692407
## [121,]  1.1598352  0.090471641 -1.040709e-01  0.0095471156
## [122,]  0.9505386  0.055679753  1.213540e-01 -0.0797920491
## [123,]  1.1407558 -0.159564797 -1.679911e-01  0.1020553689
## [124,]  0.8744232 -0.012177732  6.211565e-02  0.0315134010
## [125,]  1.0773047  0.067712066 -1.354799e-01 -0.0325495406
## [126,]  0.9804411 -0.053836047 -2.109816e-01  0.0190596310
## [127,]  0.8602960  0.022896930  5.277644e-02  0.0094512391
## [128,]  0.8651001  0.048380594  2.662668e-03 -0.0381480149
## [129,]  1.0732021 -0.014481476  1.180657e-02 -0.0018679371
```

```
## [130,]  0.8663118 -0.120009319 -1.812699e-01  0.0490628092
## [131,]  1.0454810 -0.117883322 -1.219319e-01  0.0933275423
## [132,]  1.1026091  0.047333906 -3.790380e-01  0.0171379258
## [133,]  1.1129816  0.006210166  2.398954e-02  0.0003982743
## [134,]  0.7366382 -0.099866362 -2.867956e-02 -0.0037909937
## [135,]  0.7259110 -0.237363203 -1.797767e-02 -0.0338727132
## [136,]  1.2088895  0.006699396 -1.406014e-01  0.1097488199
## [137,]  1.1746144  0.156470192 -8.415881e-02 -0.0858734228
## [138,]  0.9264210 -0.012190480 -9.354985e-02 -0.0442575678
## [139,]  0.8530139  0.062562093  1.955086e-02 -0.0462703620
## [140,]  1.0565547  0.068917575 -8.214246e-02  0.0339360748
## [141,]  1.1861457  0.103348141 -4.831034e-02  0.0040274665
## [142,]  1.1054647  0.148654120 -3.391622e-02  0.0555480855
## [143,]  0.9325263 -0.015545992  9.964546e-02 -0.0495163213
## [144,]  1.1757870  0.066788862 -1.116525e-01 -0.0134660412
## [145,]  1.2263951  0.145262725 -8.981912e-02 -0.0240559573
## [146,]  1.1141967  0.116499067 -4.404227e-03  0.0372107293
## [147,]  0.9352966 -0.046225296  1.226050e-01  0.0585439986
## [148,]  0.9916278  0.054360212 -2.618188e-02  0.0041475663
## [149,]  1.1182316  0.162541744 -7.195189e-02 -0.0908772430
## [150,]  0.8819499  0.020921633  1.890457e-03 -0.0790532121
```

Another use of PCA is prediction. If we observe new data and want to predict their PC values, we can just use the predict function. Assume we just got the data for the last four rows of iris dataset.

```
predict(pca, newdata = tail(log.data, 4))
```

```
##           PC1          PC2           PC3          PC4
## 147 0.9352966 -0.04622530  0.122604961  0.058543999
## 148 0.9916278  0.05436021 -0.026181878  0.004147566
## 149 1.1182316  0.16254174 -0.071951892 -0.090877243
## 150 0.8819499  0.02092163  0.001890457 -0.079053212
```

Finally, there is a function that can automatically correct for skewness, directly center and scale each variable and perform PCA in one step. We can use the preProcess function in the caret package.

```
library(caret)
```

```
## Loading required package: lattice
```

```
trans <- preProcess(iris[, 1:4], method = c("scale", "pca", "center"))
```

This function only keeps the PCs that cummutively explain 95% of the variability in the data.

Conclusion

PCA is an extremely efficient way to reduce dimensions of the dataset and visually present the aggregate value of each data point. The outputs of the PCA functions give the standard deviations of the principal components, the eigenvectors, the standard deviation of variables, and the coordinates of the individuals on the PCs. However, the outputs the functions do not give much information. To be able to compare the data points and variable, we can take advantage of the various packages available and generate graphs using factoextra or ggbiplot.

Take-home message

After we get the outputs of the PCA function, we need to first use the summary function to see the cumulative proportions of the variability captured by each PC. Usually only the ones that capture 95% of the variability are useful for further analysis. Then, we should generate a plot of all the variables to see how much contribution each variable makes to the PCs. We can tell which variables are more decisive compared to others. Thirdly, we should generate a scatterplot between PC1 and PC2(the two most useful PCs) to see where each individual data point lies. We can tell that data points closer to each other belong to the same group. Those three steps are essential for PC analysis.

References

http://www.sthda.com/english/wiki/fviz-pca-quick-principal-component-analysis-data-visualization-r-software-and-data-mining

https://www.statmethods.net/advstats/factor.html

http://www.gastonsanchez.com/visually-enforced/how-to/2012/06/17/PCA-in-R/

https://tgmstat.wordpress.com/2013/11/21/introduction-to-principal-component-analysis-pca/

https://www.r-bloggers.com/computing-and-visualizing-pca-in-r/

http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/

https://www.analyticsvidhya.com/blog/2016/03/practical-guide-principal-component-analysis-python/