Data processing with Principal Components Analysis in R

Principal component analysis is a method of transforming multiple variables into a few principal components through dimensionality reduction technology. These principal components can reflect most of the information of the original variables, and they are usually expressed as linear combinations of the original variables. In statistics, principal component analysis is a technique to simplify data sets, while maintaining the feature that contributes the most to the variance. In R language, the PCA analysis function prcomp is built-in. Calling this function directly can quickly perform PCA analysis on a set of data. With ggplot2 and other drawing packages, we can easily generate PCA analysis visualization results. In the following analysis, we use the mtcars data set as an example.

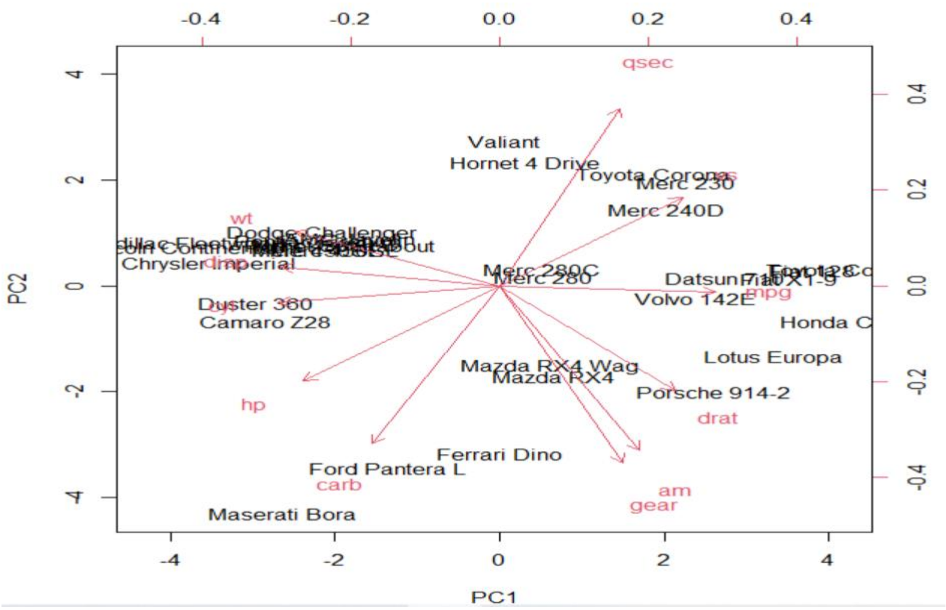
1. Calculate the principal components

```
> data("mtcars")
> head(mtcars)
                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
> dim(mtcars)
[1] 32 11
>
> res <- prcomp(mtcars, scale = TRUE)
> names(res)
[1] "sdev"     "rotation" "center"   "scale"    "x"
> res$rotation <- -1*res$rotation
> res$rotation
            PC1         PC2         PC3          PC4         PC5         PC6          PC7
mpg   0.3625305 -0.01612440  0.22574419  0.022540255 -0.10284468  0.10879743 -0.367723810
cyl  -0.3739160 -0.04374371  0.17531118  0.002591838 -0.05848381 -0.16855369 -0.057277736
disp -0.3681852  0.04932413  0.06148414 -0.256607885 -0.39399530  0.33616451 -0.214303077
hp   -0.3300569 -0.24878402 -0.14001476  0.067676157 -0.54004744 -0.07143563  0.001495989
drat  0.2941514 -0.27469408 -0.16118879 -0.854828743 -0.07732727 -0.24449705 -0.021119857
wt   -0.3461033  0.14303825 -0.34181851 -0.245899314  0.07502912  0.46493964  0.020668302
qsec  0.2004563  0.46337482 -0.40316904 -0.068076532  0.16466591  0.33048032 -0.050010522
vs    0.3065113  0.23164699 -0.42881517  0.214848616 -0.59953955 -0.19401702  0.265780836
am    0.2349429 -0.42941765  0.20576657  0.030462908 -0.08978128  0.57081745  0.587305101
gear  0.2069162 -0.46234863 -0.28977993  0.264690521 -0.04832960  0.24356284 -0.605097617
carb -0.2140177 -0.41357106 -0.52854459  0.126789179  0.36131875 -0.18352168  0.174603192
              PC8          PC9        PC10         PC11
mpg   0.754091423 -0.235701617 -0.13928524  0.124895628
cyl   0.230824925 -0.054035270  0.84641949  0.140695441
disp -0.001142134 -0.198427848 -0.04937979 -0.660606481
hp    0.222358441  0.575830072 -0.24782351  0.256492062
drat -0.032193501  0.046901228  0.10149369  0.039530246
wt    0.008571929 -0.359498251 -0.09439426  0.567448697
qsec  0.231840021  0.528377185  0.27067295 -0.181361780
vs   -0.025935128 -0.358582624  0.15903909 -0.008414634
am    0.059746952  0.047403982  0.17778541 -0.029823537
gear -0.336150240  0.001735039  0.21382515  0.053507085
carb  0.395629107 -0.170640677 -0.07225950 -0.319594676
```

There are eleven principal components. The first principal component (PC1) has high values for mpg, vs and drat, which indicates that this principal component describes the most variation in these variables. We can also see that the second principal component (PC2) has a high value for qsec. We can see this situation more clearly through the picture below.

2. Draw the biplot

```
biplot(res, scale = 0)
```



From the above we can see that each of the 32 cars represented in a simple two-dimensional space. The cars that are close to each other on the plot have similar data patterns regards to the variables in the original dataset.

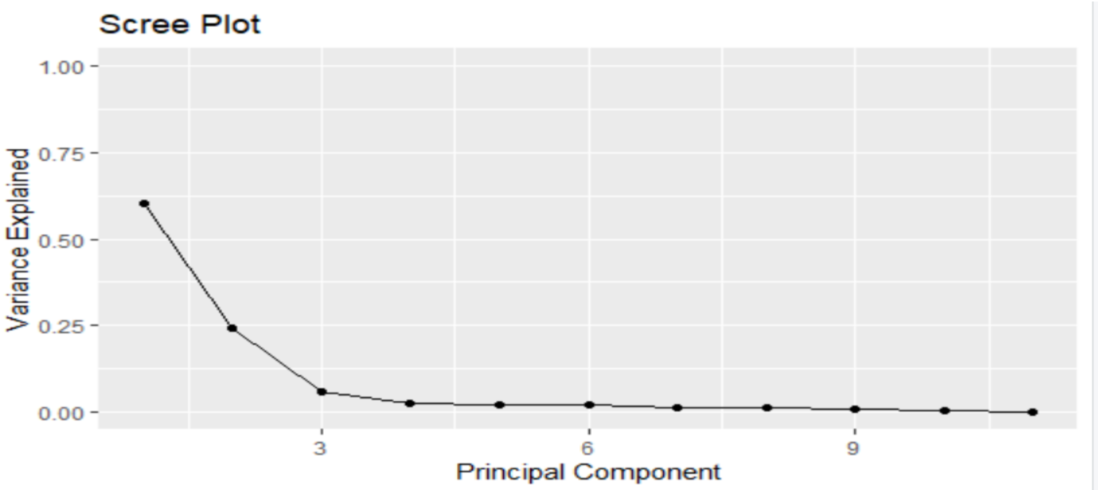
3. Calculate proportion of variance

```
> var_explained = res$sdev^2 / sum(res$sdev^2)
> var_explained
 [1] 0.600763659 0.240951627 0.057017934 0.024508858 0.020313737 0.019236011 0.012296544
 [8] 0.011172858 0.007004241 0.004730495 0.002004037
```

Form the results, we can get the proportion variance of the first two principal component are about 0.60, 0.24 and 0.05. We need to calculate the cumulative value of the proportion of Variance which generally reaching about 80% can represent the data. Here the value is approximately 80%. It is a good signal. We can also use the scree chart to observe the slope of variance.

4. Make scree plot

```
> library(ggplot2)
> qplot(c(1:11), var_explained) +
+   geom_line() +
+   xlab("Principal Component") +
+   ylab("Variance Explained") +
+   ggtitle("Scree Plot") +
+   ylim(0, 1)
```



When the variance change reaches the third component, the change is no longer obvious, so we choose the first three principal component.

5. Get new data

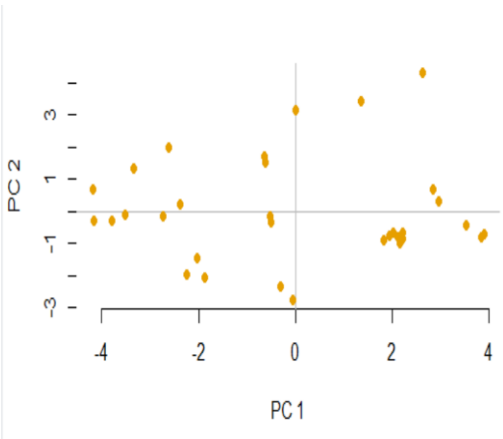
```
> new.data<-as.data.frame(predict(res)[,1:3])
> head(new.data)
                          PC1        PC2        PC3
Mazda RX4           0.64686274 -1.7081142  0.5917309
Mazda RX4 Wag       0.61948315 -1.5256219  0.3763013
Datsun 710          2.73562427  0.1441501  0.2374391
Hornet 4 Drive      0.30686063  2.3258038  0.1336213
Hornet Sportabout  -1.94339268  0.7425211  1.1165366
Valiant             0.05525342  2.7421229 -0.1612456
```

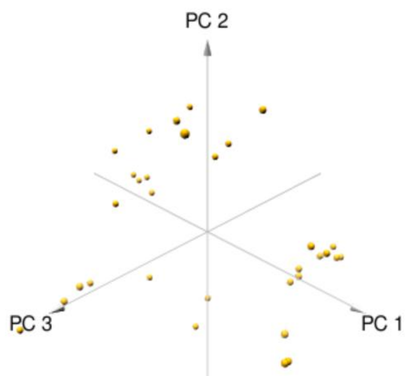
The above is to extract the data after dimensionality reduction.

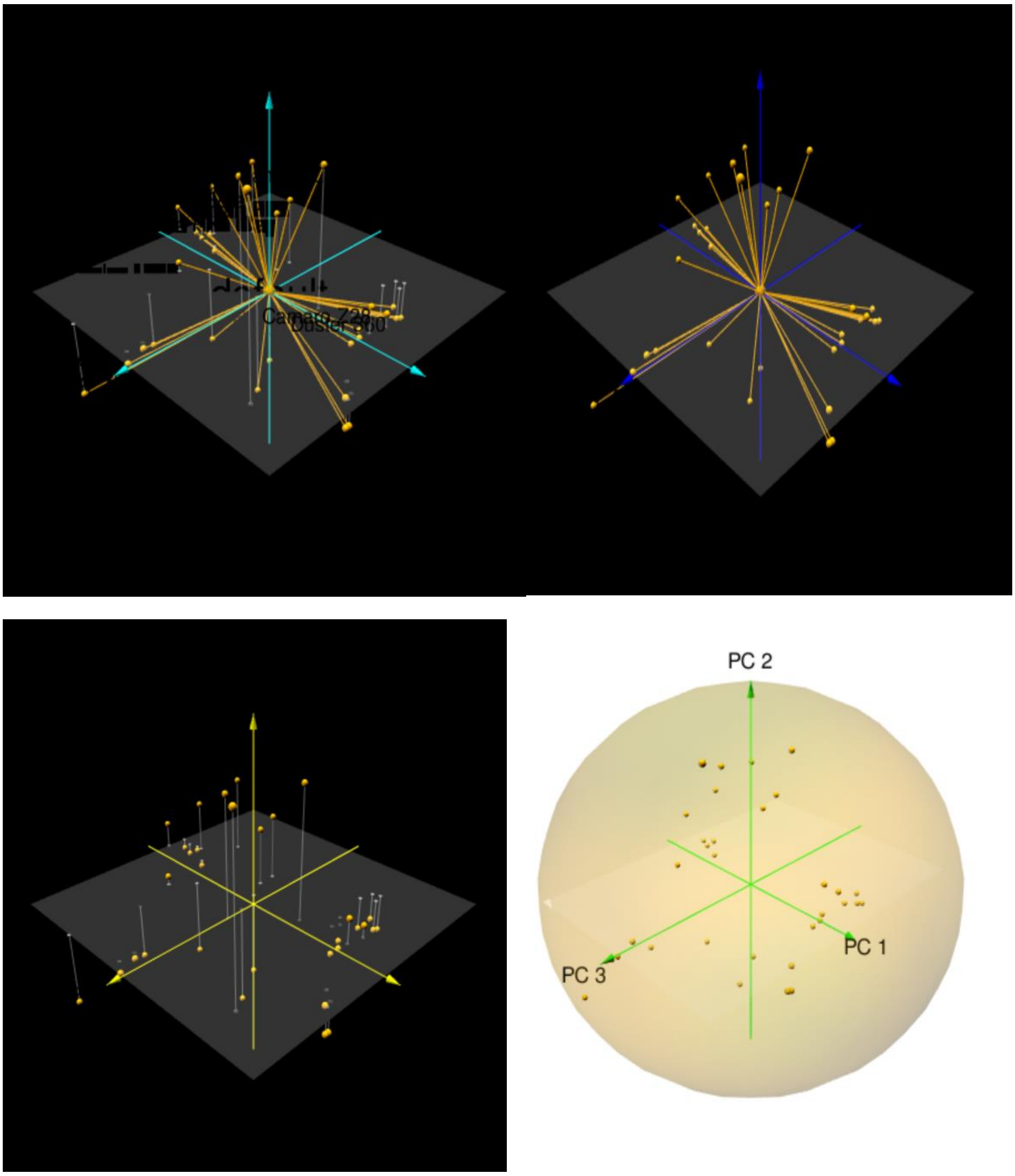
We can use some functions of R to visualize PCA.

```
library(pca3d)
data("mtcars")
pca <- prcomp(mtcars, scale = TRUE)
#2D
pca2d(pca)
#3D
pca3d(pca)
#3D+
pca3d(pca,fancy=TRUE, bg= "black",axes.color= "cyan",new=TRUE)
#3D++
pca3d(pca, fancy=FALSE, bg= "black", axes.color= "blue", new=TRUE,show.centroids=TRUE)
#3D+++
pca3d(pca, fancy=FALSE, bg= "black", axes.color= "yellow", new=TRUE,show.shadows=TRUE)

pca3d(pca, fancy=FALSE, bg= "white", axes.color= "green", new=TRUE,show.ellipses=TRUE)
```







References:

Francis, H. (2016). Principal Components Analysis using R. Retrieved from *http://faculty.missouri.edu/huangf/data/mvnotes/Documents/pca_in_r_2.pdf*

Gregory, B. (2013). Principal Components Analysis in R. Retrieved from *https://www.ime.usp.br/~pavan/pdf/PCA-R-2013*