PCA_USArrests

Ye

8/6/2017

Overview of the Data

```
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```
USArrests.means<-apply(USArrests , 2, mean)
USArrests.vars<-apply(USArrests , 2, var)
rbind(USArrests.means,USArrests.vars)
```

```
##                   Murder  Assault UrbanPop     Rape
## USArrests.means  7.78800  170.760  65.5400 21.23200
## USArrests.vars  18.97047 6945.166 209.5188 87.72916
```

Apply PCA

```
pr.out=prcomp(USArrests , scale=TRUE)
names(pr.out)
```

```
## [1] "sdev"     "rotation" "center"   "scale"    "x"
```

```
#standardize the dataset to mean = 0 and std = 1
print(pr.out$center)
```

```
##   Murder  Assault UrbanPop     Rape
##    7.788  170.760   65.540   21.232
```

```
print(pr.out$scale)
```

```
##    Murder   Assault  UrbanPop      Rape
##  4.355510 83.337661 14.474763  9.366385
```

```
#score vectors
head(pr.out$x)
```

```
##                   PC1        PC2         PC3          PC4
## Alabama    -0.9756604  1.1220012 -0.43980366  0.154696581
## Alaska     -1.9305379  1.0624269  2.01950027 -0.434175454
## Arizona    -1.7454429 -0.7384595  0.05423025 -0.826264240
## Arkansas    0.1399989  1.1085423  0.11342217 -0.180973554
## California -2.4986128 -1.5274267  0.59254100 -0.338559240
```

```
## Colorado   -1.4993407 -0.9776297  1.08400162  0.001450164
```

```
smry<-summary(pr.out)
print(smry$importance)
```

```
##                             PC1       PC2       PC3       PC4
## Standard deviation     1.574878 0.9948694 0.5971291 0.4164494
## Proportion of Variance 0.620060 0.2474400 0.0891400 0.0433600
## Cumulative Proportion  0.620060 0.8675000 0.9566400 1.0000000
```

```
#Same as
#pr.var=pr.out$sdev ^2
#pve=pr.var/sum(pr.var)
```

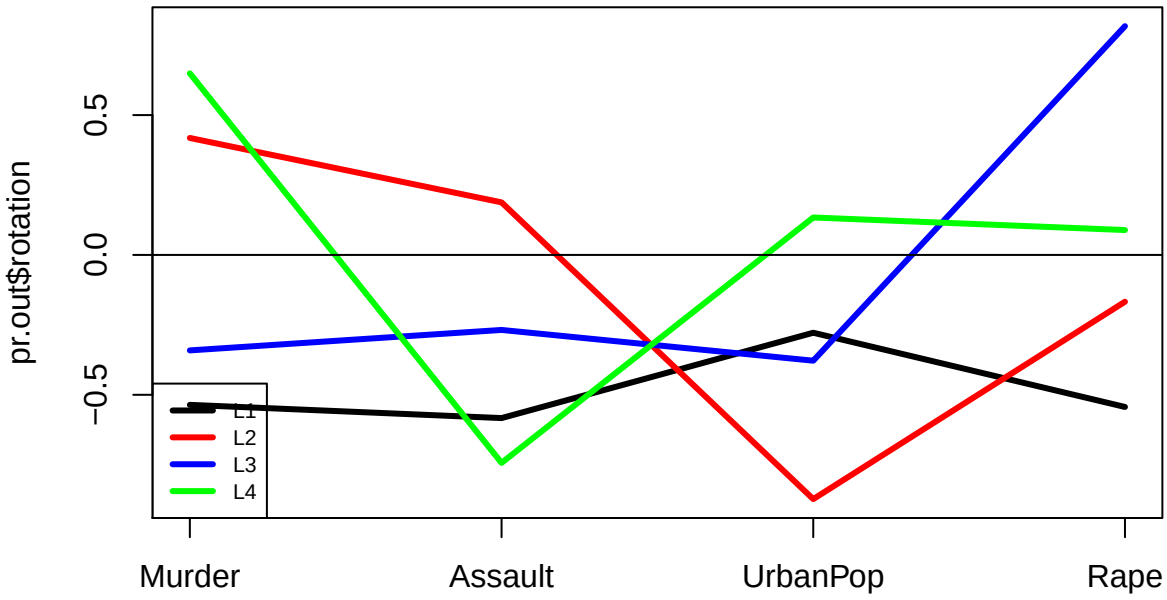
The first principal component corresponds to the direction with the highest variation. Here in this case, the first principal component explains 62% of the variation.

The loading vectors

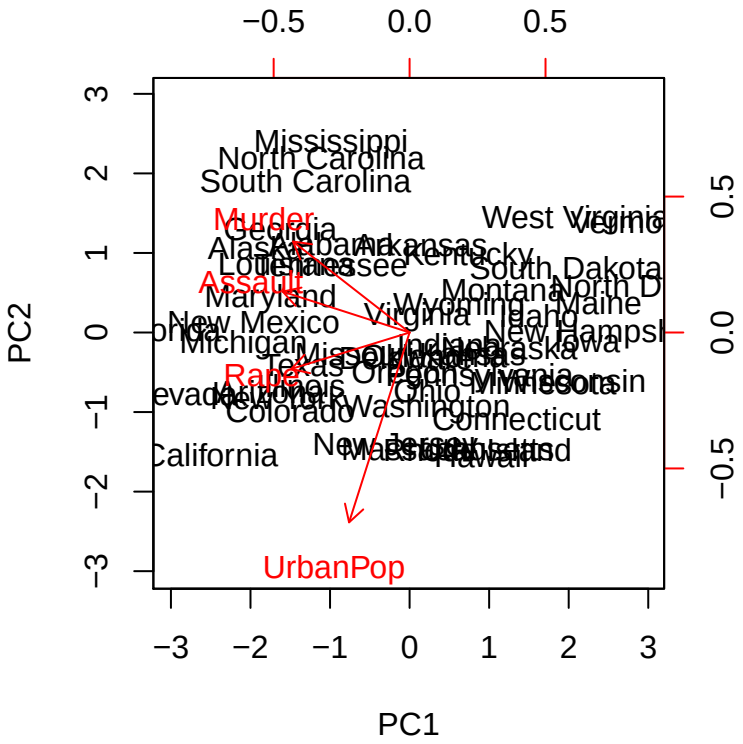
```
#loading vectors
print(pr.out$rotation)
```

```
##                 PC1        PC2        PC3         PC4
## Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
## Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
## UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
## Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

```
matplot(pr.out$rotation,type="l",lty=1,lwd=3,col=c("black","red","blue","green"),xaxt = "n")
abline(h=0)
legend("bottomleft",
       legend=c("L1","L2","L3","L4"),
       lty=1,lwd=3,cex=.7,col=c("black","red","blue","green"))
axis(1,1:4,rownames(pr.out$rotation))
```



```
biplot(pr.out , scale=0)
```



From the figures, we may tell the first principal component mainly describe the overall level of crime, while the second principal component is more responseible to the Urban population. From the biplot figure, we may tell that California is a state with generally higher urban population and high level of crime rate.

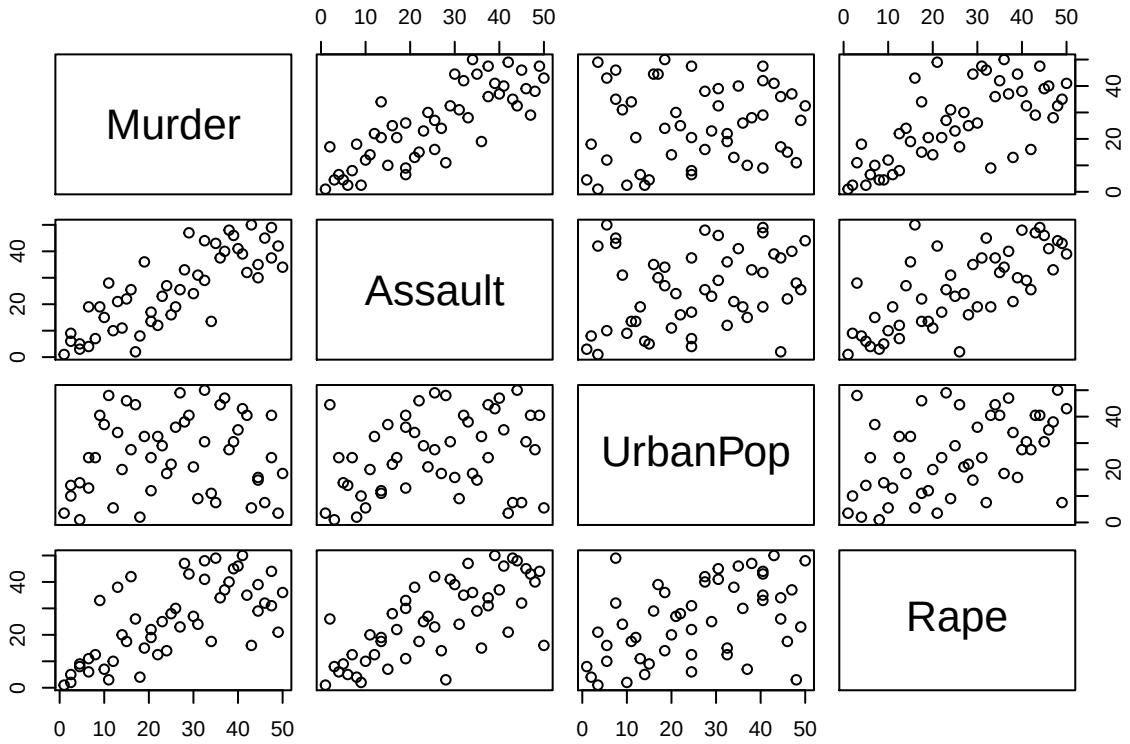
Dimension Reduction

Here, we perform dimension reduction on the first two principal components and then compare the pair plot before and after the dimension reduciton.

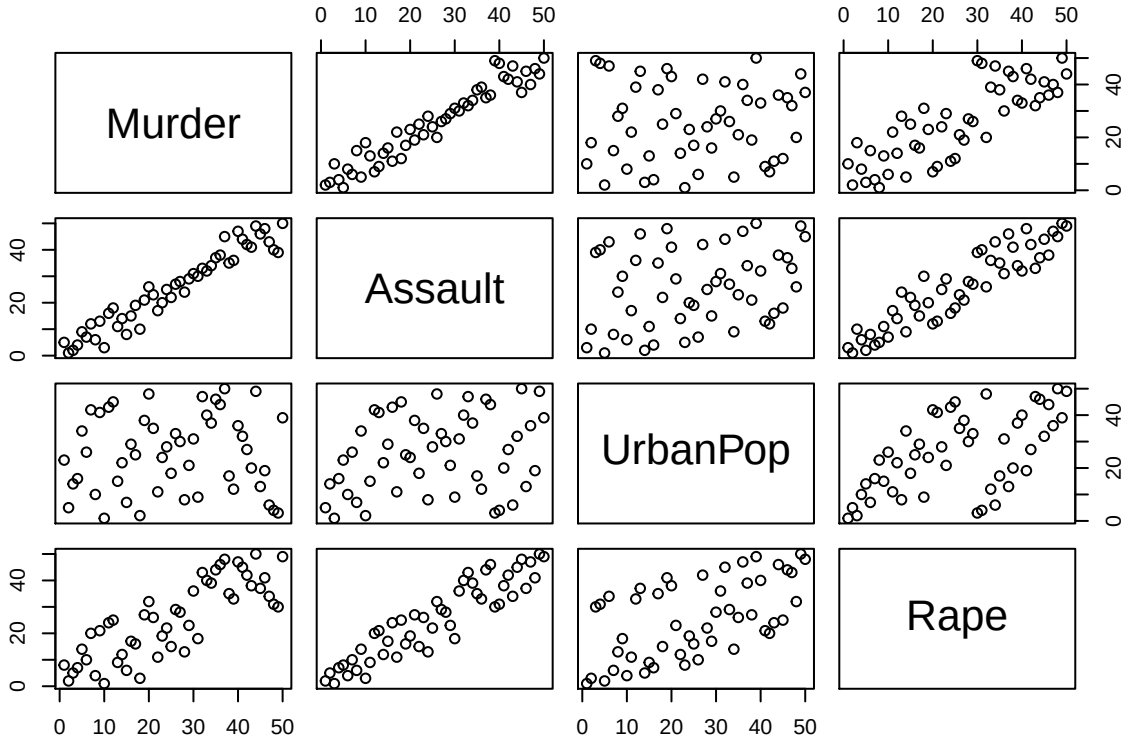
```
USArrests.approximations<-pr.out$x[,1:2]%*%t(pr.out$rotation[,1:2])
head(USArrests.approximations)
```

```
##               Murder   Assault   UrbanPop        Rape
## Alabama    0.9920554 0.7799093 -0.7078698  0.3424735
## Alaska     1.4788608 1.3255791 -0.3902348  0.8713524
## Arizona    0.6265723 0.8790939  1.1300983  1.0720877
## Arkansas   0.3885458 0.1267449 -1.0064890 -0.2615597
## California 0.7002647 1.1700159  2.0282388  1.6133934
## Colorado   0.3946699 0.6906107  1.2703841  0.9783655
```

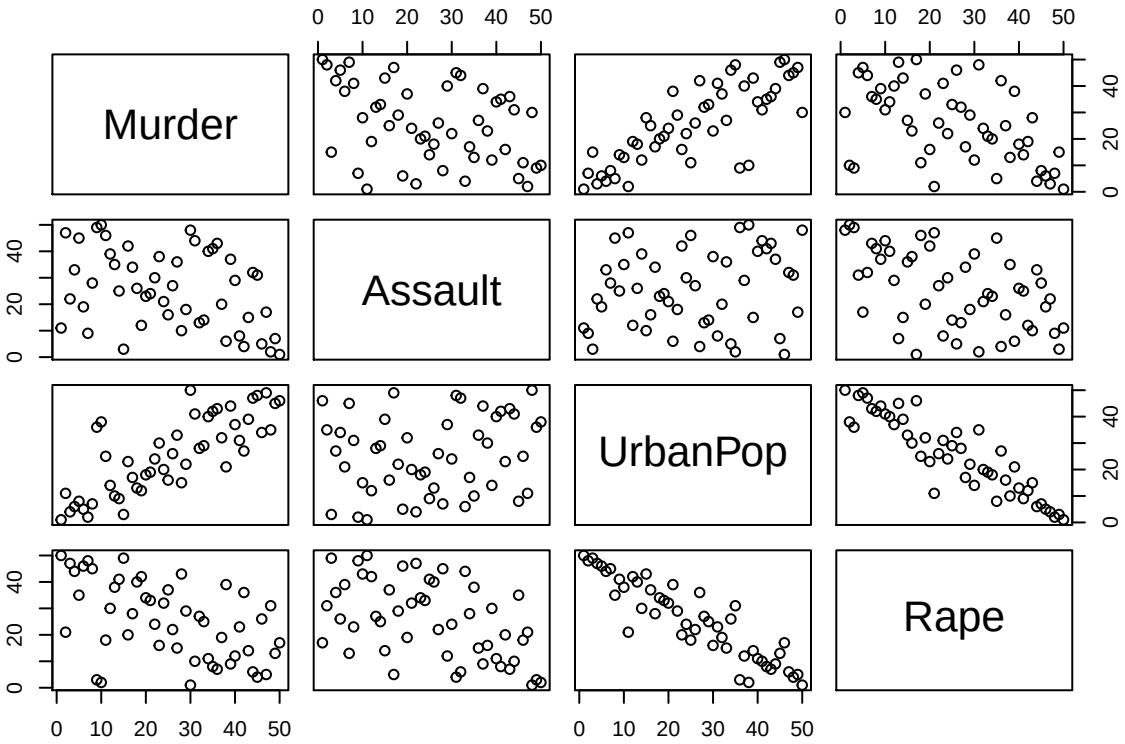
```
USArrests.scaled<-apply(USArrests,2,scale)
pairs(apply(USArrests.scaled,2,rank))
```




```
pairs(apply(USArrests.approximations,2,rank))
```




```
pr.out.residuals<-USArrests.scaled-USArrests.approximations
pairs(apply(pr.out.residuals,2,rank))
```



Clearly, we may tell that PCA exaggerates the correlation between variables.