Lab10_Answer

Anyi Guo

04/12/2018

```
states<-row.names(USArrests)
states[1:10]
```

```
##  [1] "Alabama"     "Alaska"      "Arizona"     "Arkansas"    "California"
##  [6] "Colorado"    "Connecticut" "Delaware"    "Florida"     "Georgia"
```

```
names(USArrests)
```

```
## [1] "Murder"   "Assault"  "UrbanPop" "Rape"
```

1) Calculate the mean and variance of each column, by using apply() function. Hint: `apply(dataset, 1, func)` is to apply the func to each row of dataset, and `apply(dataset, 2, func)` is to apply the func to each column of dataset.

```
# mean
print(apply(USArrests,2,mean))
```

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

```
# variance
print(apply(USArrests,2,var))
```

```
##     Murder    Assault   UrbanPop       Rape
##   18.97047 6945.16571  209.51878   87.72916
```

2) What conclusions can you draw from 1)? And consequently what transformation would you do to your dataset? Assault has very high variance compared ot the other variables - we should scale the variables.

3) Perform principal component analysis using the prcomp() function.

```
pr.arrest<-prcomp(USArrests,scale=TRUE)
```

4) Check the results, report the number of PCs and their center, scale, and rotation. There are 4 PCs. Center

```
pr.arrest$center
```

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

Scale

```
pr.arrest$scale
```

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

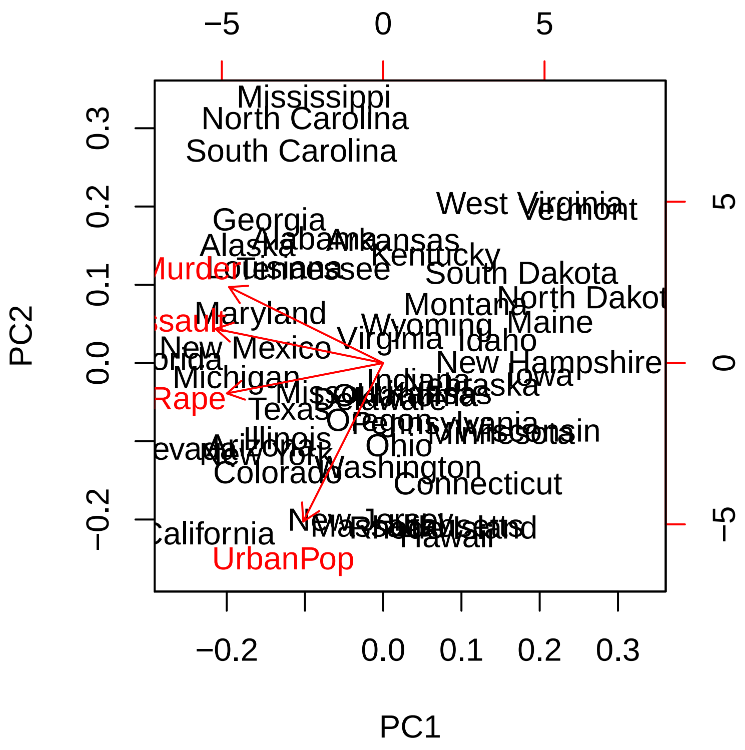
Rotation:

```
pr.arrest$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
```

5) Plot the first two PCs.

```
biplot(pr.arrest,scale=TRUE)
```



6) What are the standard deviation of each principal component? Based on this result, calculate the variance explained by each PC and the proportion of variance explained by each PC.

```r
# standard deviation
pr.arrest$sdev
```

```
## [1] 1.5748783 0.9948694 0.5971291 0.4164494
```

```r
# variance
pr.arrest$sdev^2
```

```
## [1] 2.4802416 0.9897652 0.3565632 0.1734301
```

```r
# proportion of variance explained
pr.arrest.var<-pr.arrest$sdev^2
pve<-pr.arrest.var/sum(pr.arrest.var)
pve
```

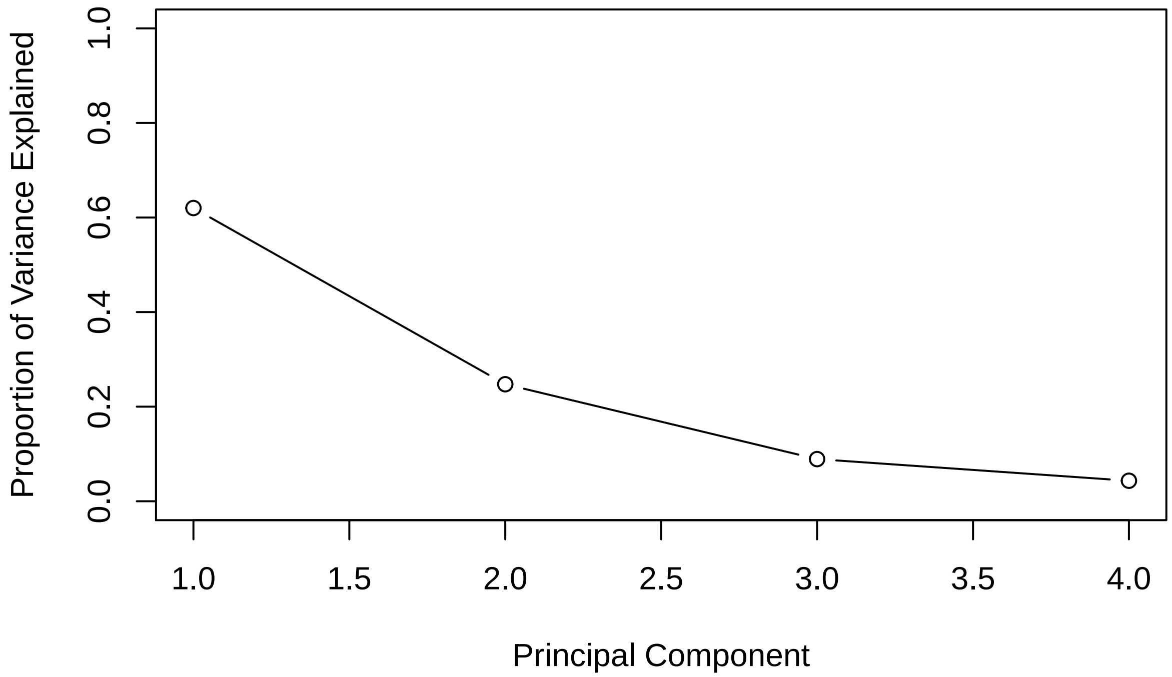
```
## [1] 0.62006039 0.24744129 0.08914080 0.04335752
```

First PC: 62.0% Second PC: 24.7% Third PC:8.9% Fourth PC: 4.3%

7) Plot the PVE explained by each component as well as the cumulative PVE. Hint: the cumulative PVE can be obtained by the cumsum() function.

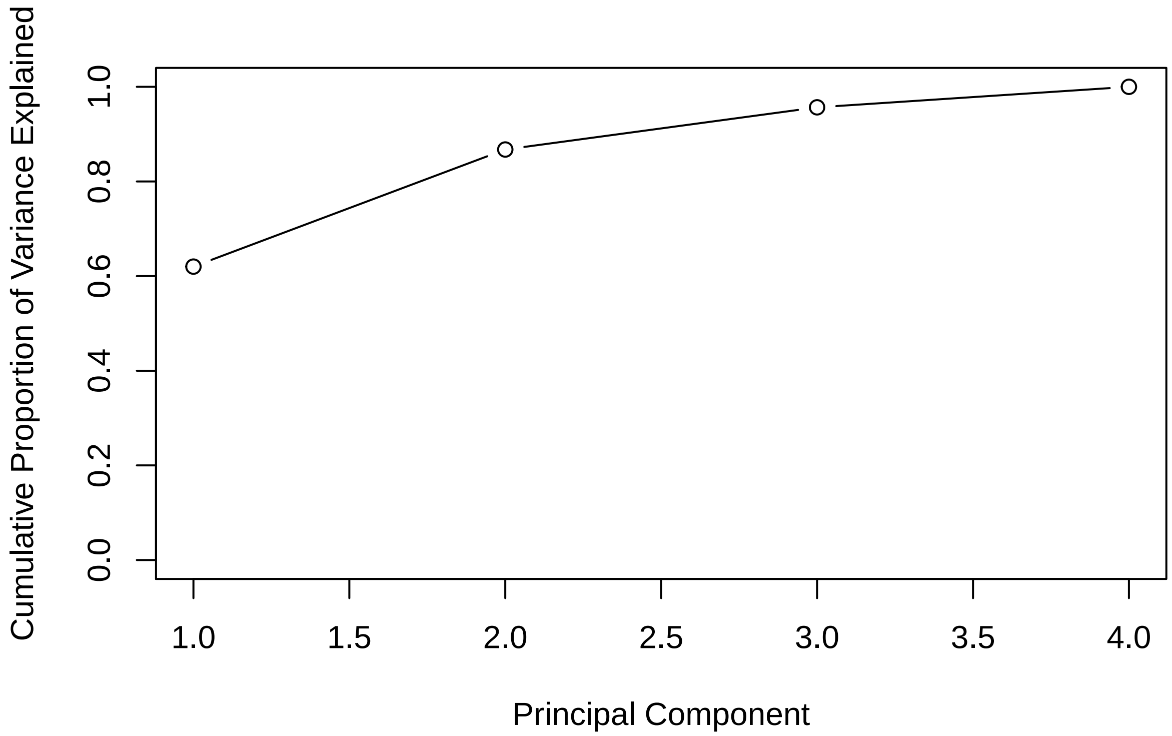
For each component:

```r
plot(pve,xlab="Principal Component", ylab="Proportion of Variance Explained", type="b",ylim=c(0,1))
```



Cumulative PVE:

```r
plot(cumsum(pve),xlab="Principal Component", ylab="Cumulative Proportion of Variance Explained", type="
```


Cumulative Proportion of Variance Explained
Principal Component