Week8 Assignment

Mohit Mehndiratta

23/09/2021

# Principal Component Analysis

## Question 1

### 1. Load “iris” dataset from the ISLR Library.

library(ISLR)  
attach(iris)

### 2. Give a short description of the dataset.

head(iris)

## 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  
## 6 5.4 3.9 1.7 0.4 setosa

str(iris)

## 'data.frame': 150 obs. of 5 variables:  
## $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...  
## $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...  
## $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...  
## $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...  
## $ Species : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...

summary(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100   
## 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300   
## Median :5.800 Median :3.000 Median :4.350 Median :1.300   
## Mean :5.843 Mean :3.057 Mean :3.758 Mean :1.199   
## 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800   
## Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500   
## Species   
## setosa :50   
## versicolor:50   
## virginica :50   
##   
##   
##

dim(iris)

## [1] 150 5

cor(iris[,1:4])

## Sepal.Length Sepal.Width Petal.Length Petal.Width  
## Sepal.Length 1.0000000 -0.1175698 0.8717538 0.8179411  
## Sepal.Width -0.1175698 1.0000000 -0.4284401 -0.3661259  
## Petal.Length 0.8717538 -0.4284401 1.0000000 0.9628654  
## Petal.Width 0.8179411 -0.3661259 0.9628654 1.0000000

The iris data set has 150 observations of 5 variables. The variables are *“Sepal.Length”*, *“Sepal.Width”*, *“Petal.Length”*, *“Petal.Width”*, *“Species”*. All are numeric variables except “Species”.

### 3. Examine the mean of the variables in the dataset.

iris\_DS <- iris[,1:4]  
sapply(iris\_DS, mean)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## 5.843333 3.057333 3.758000 1.199333

The mean of the variables do not vary much except “Petal.Width”.

### 4. Examine the variance of the variables in the dataset.

sapply(iris\_DS, var)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## 0.6856935 0.1899794 3.1162779 0.5810063

The variance of the variables do vary from each other. High variance is observed in “Sepal.Length” & “Petal.Width”. While “Sepal.Width” & “Petal.Length” has much less variance compared to them.

### 5. Perform Principal Component Analysis for the iris dataset.

pr.out.iris <- prcomp(iris\_DS, scale. = TRUE)  
pr.out.iris

## Standard deviations (1, .., p=4):  
## [1] 1.7083611 0.9560494 0.3830886 0.1439265  
##   
## Rotation (n x k) = (4 x 4):  
## PC1 PC2 PC3 PC4  
## Sepal.Length 0.5210659 -0.37741762 0.7195664 0.2612863  
## Sepal.Width -0.2693474 -0.92329566 -0.2443818 -0.1235096  
## Petal.Length 0.5804131 -0.02449161 -0.1421264 -0.8014492  
## Petal.Width 0.5648565 -0.06694199 -0.6342727 0.5235971

For the 4 variables, we got 4 Principal component loading vectors.

cat("Standard Deviation of original variables are ", pr.out.iris$scale)

## Standard Deviation of original variables are 0.8280661 0.4358663 1.765298 0.7622377

cat("Standard Deviation of Principal Component vectors are ", pr.out.iris$sdev)

## Standard Deviation of Principal Component vectors are 1.708361 0.9560494 0.3830886 0.1439265

### 6. Display and explain the principal component loading vectors.

pr.out.iris$rotation

## PC1 PC2 PC3 PC4  
## Sepal.Length 0.5210659 -0.37741762 0.7195664 0.2612863  
## Sepal.Width -0.2693474 -0.92329566 -0.2443818 -0.1235096  
## Petal.Length 0.5804131 -0.02449161 -0.1421264 -0.8014492  
## Petal.Width 0.5648565 -0.06694199 -0.6342727 0.5235971

PCA for the iris data set gives 4 Principal Component Vectors. Those are PC1, PC2, PC3 & PC4.

Here we can observe, majority and almost equal contributions are given by : For, PC1 - Sepal.Length, Petal.Length & Petal.Width. PC2 - Sepal.Width. PC3 - Sepal.Length & Petal.Width. PC4 - Petal.Length & Petal.Width.

### 7. Give the first two Principal components using loading parameters in part 6.

pc1.iris.vector <- pr.out.iris$rotation[,1]  
PC1.iris <- pc1.iris.vector[[1]] \* Sepal.Length + pc1.iris.vector[[2]]\* Sepal.Width +  
 pc1.iris.vector[[3]] \* Petal.Length + pc1.iris.vector[[4]]\* Petal.Width  
  
pc2.iris.vector <- pr.out.iris$rotation[,2]  
PC2.iris <- pc2.iris.vector[[1]] \* Sepal.Length + pc2.iris.vector[[2]]\* Sepal.Width +  
 pc2.iris.vector[[3]] \* Petal.Length + pc2.iris.vector[[4]]\* Petal.Width  
PC1.iris; PC2.iris

## [1] 2.640270 2.670730 2.454606 2.545517 2.561228 2.975946 2.463157 2.673139  
## [9] 2.437132 2.645351 2.800761 2.626967 2.562138 2.127481 2.754260 2.881509  
## [17] 2.743781 2.696755 3.102715 2.673992 2.997648 2.757413 2.120637 3.037720  
## [25] 2.801091 2.838920 2.844152 2.750418 2.719311 2.628730 2.707772 2.994537  
## [33] 2.532324 2.660153 2.701837 2.552885 2.790655 2.452636 2.352156 2.725246  
## [41] 2.586608 2.649291 2.298287 2.930188 2.962643 2.675109 2.675548 2.460541  
## [49] 2.748655 2.642033 6.304290 5.932054 6.451687 5.302329 6.149941 5.562075  
## [57] 6.025581 4.387009 6.062141 5.036715 4.662937 5.551266 5.420340 5.916135  
## [65] 4.960662 6.000781 5.569071 5.239494 6.097188 5.129554 6.015102 5.480295  
## [73] 6.300656 5.830099 5.783804 5.975609 6.365857 6.545421 5.804432 4.866075  
## [81] 5.046341 4.931814 5.236383 6.263035 5.464857 5.726244 6.231391 5.951347  
## [89] 5.223934 5.248460 5.397205 5.831159 5.321359 4.466050 5.362780 5.277596  
## [97] 5.361017 5.679591 4.346649 5.329910 7.288489 6.328278 7.502162 6.768663  
## [105] 7.187966 8.168984 5.451969 7.696018 7.200911 7.734685 6.614836 6.757001  
## [113] 7.113677 6.328486 6.583772 6.848269 6.787900 8.120139 8.615925 6.283181  
## [121] 7.340968 6.137534 8.276516 6.416243 7.096848 7.388983 6.279160 6.231226  
## [129] 7.017161 7.213799 7.715462 7.937257 7.073647 6.335934 6.519311 8.043855  
## [137] 6.972903 6.708859 6.121078 7.080807 7.262134 7.019655 6.328278 7.404944  
## [145] 7.322791 7.000417 6.584640 6.726747 6.748228 6.243095

## [1] -5.204041 -4.666910 -4.773636 -4.648463 -5.258629 -5.707321 -4.929697  
## [8] -5.076419 -4.385872 -4.754994 -5.504375 -5.003385 -4.622474 -4.426418  
## [15] -5.924983 -6.277296 -5.697524 -5.210735 -5.721522 -5.490173 -5.232284  
## [22] -5.404538 -5.097865 -5.046812 -5.010732 -4.709550 -5.092257 -5.244232  
## [29] -5.149453 -4.780984 -4.726396 -5.240775 -5.791515 -6.001315 -4.761689  
## [36] -4.884413 -5.352559 -5.214193 -4.475752 -5.114161 -5.170545 -3.873881  
## [43] -4.660411 -5.197975 -5.506664 -4.635862 -5.485928 -4.738344 -5.466633  
## [50] -4.981640 -5.805299 -5.580644 -5.686820 -4.384368 -5.251517 -4.933745  
## [57] -5.646824 -4.213020 -5.368200 -4.644706 -3.886342 -5.199929 -4.460665  
## [64] -5.188634 -4.966290 -5.592396 -5.094051 -4.849278 -4.581865 -4.590931  
## [71] -5.419365 -5.072466 -4.906392 -5.082916 -5.285369 -5.462325 -5.362946  
## [78] -5.534844 -5.152688 -4.704512 -4.458411 -4.449267 -4.857768 -4.989418  
## [85] -5.018567 -5.621030 -5.606438 -4.696099 -5.070866 -4.569027 -4.664459  
## [92] -5.278515 -4.767888 -4.158432 -4.796326 -5.104363 -5.018727 -5.209885  
## [99] -4.380180 -4.923948 -5.738911 -4.934017 -5.734631 -5.312937 -5.512425  
## [106] -5.940484 -4.381599 -5.707499 -5.099484 -6.358025 -5.666552 -5.165366  
## [113] -5.611609 -4.715862 -5.059818 -5.653791 -5.478301 -6.726005 -5.629643  
## [120] -4.518627 -5.852296 -4.952659 -5.789321 -5.111134 -5.855754 -5.939398  
## [127] -5.163272 -5.312639 -5.278432 -5.736452 -5.654707 -6.780753 -5.285126  
## [134] -5.188279 -4.933688 -5.979368 -5.814750 -5.532889 -5.272448 -5.739231  
## [141] -5.688728 -5.745272 -4.934017 -5.819453 -5.882531 -5.579908 -4.935618  
## [148] -5.484342 -5.765416 -5.242054

### 8. Explain the proportion of variance explained by each PC using graphs and summarise your results.

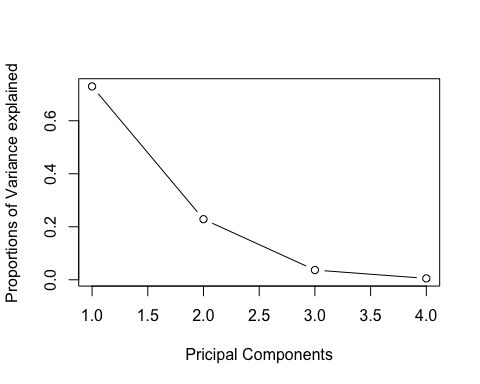
summary(pr.out.iris)

## Importance of components:  
## PC1 PC2 PC3 PC4  
## Standard deviation 1.7084 0.9560 0.38309 0.14393  
## Proportion of Variance 0.7296 0.2285 0.03669 0.00518  
## Cumulative Proportion 0.7296 0.9581 0.99482 1.00000

pr.out.iris$var <- pr.out.iris$sdev ^ 2  
PVE.iris <- pr.out.iris$var / sum(pr.out.iris$var)  
PVE.iris

## [1] 0.729624454 0.228507618 0.036689219 0.005178709

plot(PVE.iris, type = "b", xlab = "Pricipal Components", ylab = "Proportions of Variance explained")

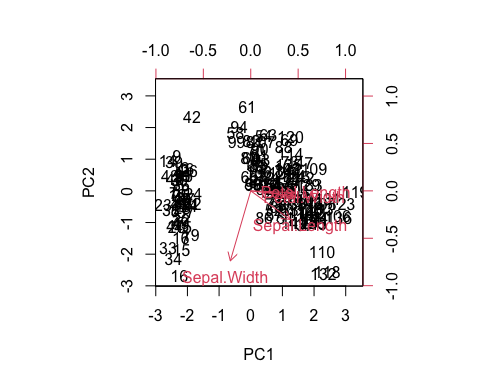


From the results, we can see that maximum proportion is explained by PC1 which is 72.9% followed by PC2 which explains 22.8% proportions of variance. While PC3 and PC4 explained 3.6% and 0.5% respectively.

If we consider the cumulative proportions, then we can see that 95% of the total proportions of variance are explained by PC1 and PC2.

### 9. Draw the biplot and interpret it.

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



From the above plot, we can see that Petal.Length and Petal.Width are very close to horizontal axis(from (0,0)). This means that they have more weightage toward PC1.

Sepal.Length is somewhat closed to PC1 as well as PC2, but it gives more weightage toward PC1.

While, Sepal.Width is very close to vertical axis, thus more contributions towards PC2.

To check which of the variables are associated with each other or correlated with each other, we can focus on angle between the variables.

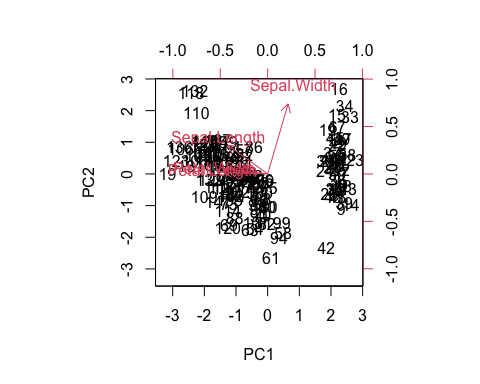
we can see that angle of Petal.Length, Petal.Width are very close to each other and in same direction that means they are highly positive correlated with each other.

On the other hand, angles of Petal.Length, Petal.Width are nearly 90 degrees with respect to Sepal.Width therefore, they seem uncorrelated towards them.

Sepal.Length is somewhat positively correlated with Petal.Length and Petal.Width.

### 10. Rotate the graph and compare the results.

pr.out.iris$rotation <- -pr.out.iris$rotation  
pr.out.iris$x <- -pr.out.iris$x  
biplot(pr.out.iris, scale = 0)



With the rotations, we can see the clear difference toward magnitude/direction in each of the variable as they are opposite as compared towards previous graph, but the interpretations are similar about correlations.

## Question 2

### 1. Load the “USArrests” dataset built in R.

arrest <- USArrests  
attach(arrest)

### 2. Give a short description of the dataset.

head(arrest)

## 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

dim(arrest)

## [1] 50 4

str(arrest)

## 'data.frame': 50 obs. of 4 variables:  
## $ Murder : num 13.2 10 8.1 8.8 9 7.9 3.3 5.9 15.4 17.4 ...  
## $ Assault : int 236 263 294 190 276 204 110 238 335 211 ...  
## $ UrbanPop: int 58 48 80 50 91 78 77 72 80 60 ...  
## $ Rape : num 21.2 44.5 31 19.5 40.6 38.7 11.1 15.8 31.9 25.8 ...

summary(arrest)

## Murder Assault UrbanPop Rape   
## Min. : 0.800 Min. : 45.0 Min. :32.00 Min. : 7.30   
## 1st Qu.: 4.075 1st Qu.:109.0 1st Qu.:54.50 1st Qu.:15.07   
## Median : 7.250 Median :159.0 Median :66.00 Median :20.10   
## Mean : 7.788 Mean :170.8 Mean :65.54 Mean :21.23   
## 3rd Qu.:11.250 3rd Qu.:249.0 3rd Qu.:77.75 3rd Qu.:26.18   
## Max. :17.400 Max. :337.0 Max. :91.00 Max. :46.00

cor(arrest)

## Murder Assault UrbanPop Rape  
## Murder 1.00000000 0.8018733 0.06957262 0.5635788  
## Assault 0.80187331 1.0000000 0.25887170 0.6652412  
## UrbanPop 0.06957262 0.2588717 1.00000000 0.4113412  
## Rape 0.56357883 0.6652412 0.41134124 1.0000000

The USArrest data set has 50 observations of 4 variables. The variables are *“Murder”*, *“Assault”*, *“UrbanPop”* and *“Rape”*. All are numeric variables.

### 3. Examine the mean of the variables in the dataset.

sapply(arrest, mean)

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

Here, we can see that mean of the variables vary alot. Means are given as : 1. Murder - 7.788 2. Assault - 170.76 3. UrbanPop - 65.54 4. Rape - 21.232

### 4. Examine the variance of the variables in the dataset.

sapply(arrest, var)

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

The variance of the variables do vary from each other. High variance is observed in “Assault” & “UrbanPop”. While “Rape” has much less variance compared to them. Murder has the lowest variance out of all.

### 5. Perform Principal Component Analysis for USArrests.

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

## Standard deviations (1, .., p=4):  
## [1] 1.5748783 0.9948694 0.5971291 0.4164494  
##   
## Rotation (n x k) = (4 x 4):  
## 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

For the 4 variables, we got 4 Principal component loading vectors.

cat("Standard Deviation of original variables are ", pr.out.arrest$scale)

## Standard Deviation of original variables are 4.35551 83.33766 14.47476 9.366385

cat("Standard Deviation of Principal Component vectors are ", pr.out.arrest$sdev)

## Standard Deviation of Principal Component vectors are 1.574878 0.9948694 0.5971291 0.4164494

### 6. Display and explain the principal component loading vectors.

pr.out.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

PCA for the USArrest data set gives 4 Principal Component Vectors. Those are PC1, PC2, PC3 & PC4.

Here we can observe, majority and almost equal contributions are given by : For, PC1 - Murder, Assault & Rape. PC2 - UrbanPop. PC3 - Rape. PC4 - Murder & Assault

### 7. Give the first two Principal components using loading parameters in part 6.

pc1.arrest.vector <- pr.out.arrest$rotation[,1]  
PC1.arrest <- pc1.arrest.vector[[1]]\*Murder + pc1.arrest.vector[[2]]\*Assault +   
 pc1.arrest.vector[[3]]\*UrbanPop + pc1.arrest.vector[[4]]\*Rape  
pc2.arrest.vector <- pr.out.arrest$rotation[,2]  
PC2.arrest <- pc2.arrest.vector[[1]]\*Murder + pc2.arrest.vector[[2]]\*Assault +   
 pc2.arrest.vector[[3]]\*UrbanPop + pc2.arrest.vector[[4]]\*Rape  
PC1.arrest; PC2.arrest

## [1] -172.36104 -196.27218 -214.89844 -140.02728 -213.16049 -165.93278  
## [7] -93.37146 -170.57548 -243.21012 -163.08840 -63.73389 -94.11442  
## [13] -186.91829 -99.25271 -55.83492 -98.42389 -92.08911 -183.89037  
## [19] -67.95614 -214.75696 -121.75649 -194.85680 -69.89389 -181.20563  
## [25] -143.42793 -90.43882 -88.00356 -201.03159 -55.10815 -131.66737  
## [31] -209.23412 -192.18512 -224.76742 -42.87944 -106.38786 -121.38329  
## [37] -129.91345 -93.32051 -132.00910 -196.00557 -71.66473 -147.74398  
## [43] -160.13862 -106.39678 -44.16034 -124.30686 -121.25108 -66.19586  
## [49] -56.53174 -122.70567

## [1] -4.285324 4.281646 -16.356340 -7.505767 -30.570846 -32.901422  
## [7] -47.004900 -18.277839 -5.746797 -9.743883 -64.959054 -25.861916  
## [13] -25.301065 -35.992818 -40.193462 -36.489515 -23.566431 -8.071282  
## [19] -29.337216 -2.008348 -47.065969 -17.464225 -45.434205 14.156362  
## [25] -28.589754 -26.003238 -35.902032 -25.919780 -38.873315 -47.841092  
## [31] -8.124203 -27.038198 26.817391 -30.831002 -43.430090 -31.551374  
## [37] -31.441654 -42.774080 -43.191573 12.810421 -23.662108 -15.135156  
## [43] -30.995117 -49.759641 -19.860460 -25.569994 -39.167964 -17.985040  
## [49] -48.361743 -21.869230

### 8. Explain the proportion of variance explained by each PC using graphs and summarise your results.

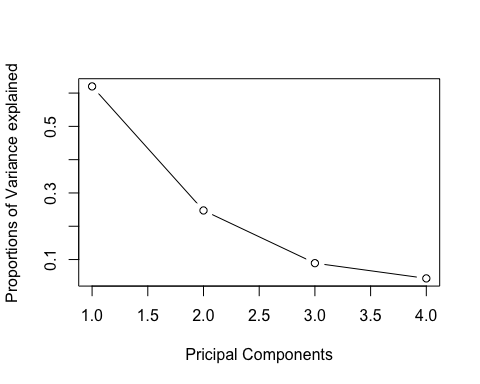
summary(pr.out.arrest)

## Importance of components:  
## PC1 PC2 PC3 PC4  
## Standard deviation 1.5749 0.9949 0.59713 0.41645  
## Proportion of Variance 0.6201 0.2474 0.08914 0.04336  
## Cumulative Proportion 0.6201 0.8675 0.95664 1.00000

pr.out.arrest$var <- pr.out.arrest$sdev ^ 2  
PVE.arrest <- pr.out.arrest$var / sum(pr.out.arrest$var)  
PVE.arrest

## [1] 0.62006039 0.24744129 0.08914080 0.04335752

plot(PVE.arrest, type = "b", xlab = "Pricipal Components", ylab = "Proportions of Variance explained")

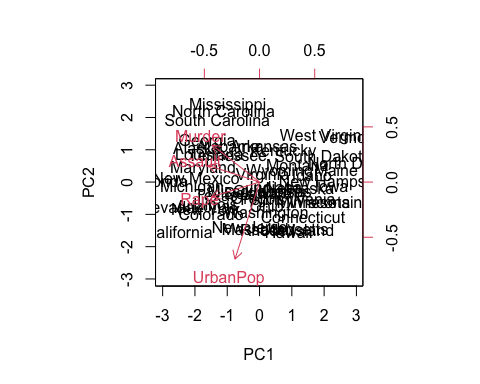


From the results, we can see that maximum proportion is explained by PC1 which is 62.01% followed by PC2 which explains 24.74% proportions of variance. While PC3 and PC4 explained 8.9% and 4.3% respectively.

If we consider the cumulative proportions, then we can see that 95.66% of the total proportions of variance are explained by PC1,PC2 and PC3.

### 9. Draw the biplot and interpret it.

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



From the above plot, we can see that Murder, Assault and Rape are very close to horizontal axis(from (0,0)). This means that they have more weightage toward PC1.

While, UrbanPop is very close to vertical axis, thus more contributions towards PC2.

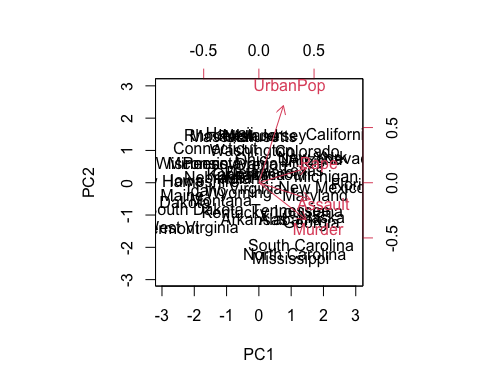
To check which of the variables are associated with each other or correlated with each other, we can focus on angle between the variables.

we can see that angle of Murder, Assault and Rape are close to each other and in same direction that means they are highly positive correlated with each other.

On the other hand, angles of Murder, Assault and Rape are nearly 90 degrees with respect to UrbanPop therefore, they seem uncorrelated towards them.

### 10. Rotate the graph and compare the results.

pr.out.arrest$rotation <- -pr.out.arrest$rotation  
pr.out.arrest$x <- -pr.out.arrest$x  
biplot(pr.out.arrest, scale = 0)



With the rotations, we can see the clear difference toward magnitude/direction in each of the variable as they are opposite as compared towards previous graph, but the interpretations are similar about correlations.