Consider the `USArrests` data set from *R*. For each of the 50 states in the United States, the data set contains the number of arrests per 100,000 residents for each of three crimes: `Assault, Murder,` and `Rape`. It also records `UrbanPop` (the percent of the population in each state living in urban areas). Use the function `prcomp` to examine differences between the states via the two largest principal components.

a) Compare eigenvalues and the variances of all variables in `USArrests`.

b) Use function `prcomp()` to find principal components (scale the data set first), and the loading and score vectors. Call it `m1` object.

c) Use function `eigen()` to find eigenvalues and eigenvectors of the covariance matrix of the scaled data set. Call it `m2` object.

d) Use eigenvectors to define the PC variables.

e) Verify that the variance of the PCs are the eigenvalues

f) Find the proportion of variance explained (PVE) by each principal component.

g) Plot the PVE explained by each component (individual and cumulative).

h) Use a biplot display to plot the original data on PC1 and PC2 axes.

i) Interpret the principal components.

```
d1=USArrests
dim(d1)            # [1] 50  4
head(d1)

# a) Compare eigenvalues & variances
#==============================================================
summary(d1)
#      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

# covariance matrix
var(d1)
#              Murder   Assault   UrbanPop      Rape
# Murder    18.970465  291.0624   4.386204  22.99141
# Assault  291.062367 6945.1657 312.275102 519.26906
# UrbanPop   4.386204  312.2751 209.518776  55.76808
# Rape      22.991412  519.2691  55.768082  87.72916
apply(d1,2,var)
#     Murder    Assault   UrbanPop       Rape
#   18.97047 6945.16571  209.51878   87.72916
apply(d1,2,sd)
#    Murder   Assault  UrbanPop      Rape
#  4.355510 83.337661 14.474763  9.366385

eigen(var(d1))
# $values
# [1] 7011.114851  201.992366   42.112651    6.164246
# $vectors
#             [,1]        [,2]        [,3]        [,4]
# [1,] -0.04170432  0.04482166  0.07989066  0.99492173
# [2,] -0.99522128  0.05876003 -0.06756974 -0.03893830
# [3,] -0.04633575 -0.97685748 -0.20054629  0.05816914
# [4,] -0.07515550 -0.20071807  0.97408059 -0.07232502

sum(eigen(var(d1))$values)  #[1] 7261.384
sum(diag(var(d1)))          #[1] 7261.384
# sum of eigenvalues = sum variances

# means and variances very differents, need to standardize (scale)
```

```
# b) Find principal components (scaled data)
#==============================================================

m1=prcomp(d1, scale=T)
names(m1)
# [1] "sdev"     "rotation" "center"   "scale"    "x"

# mean and sd of d1 -unscaled-

m1$center
#  Murder  Assault UrbanPop     Rape
#   7.788  170.760   65.540   21.232
m1$scale
#   Murder   Assault  UrbanPop      Rape
# 4.355510 83.337661 14.474763  9.366385

# loading (eigen) vectors in the rotation matrix

m1$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

# sqrt(eigenvalues)
m1$sdev
# 1.5748783 0.9948694 0.5971291 0.4164494

# transformed data on PC axes
d2 = m1$x

head(d2)
#                 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

apply(d2,2,sd)
#       PC1       PC2       PC3       PC4
# 1.5748783 0.9948694 0.5971291 0.4164494
```

```
# eigen() function
#=============================================================

cova=var(scale(d1))
#              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

m2 = eigen(cova)
#$values
#[1] 2.4802416 0.9897652 0.3565632 0.1734301
#$vectors
#          [,1]       [,2]       [,3]        [,4]
#[1,] 0.5358995  0.4181809 -0.3412327  0.64922780
#[2,] 0.5831836  0.1879856 -0.2681484 -0.74340748
#[3,] 0.2781909 -0.8728062 -0.3780158  0.13387773
#[4,] 0.5434321 -0.1673186  0.8177779  0.08902432

# covariance matrix of transformed data
var(d2)
#               PC1           PC2           PC3           PC4
# PC1  2.480242e+00  6.706371e-17  4.573978e-17 -3.198568e-16
# PC2  6.706371e-17  9.897652e-01 -9.581526e-17 -1.516830e-16
# PC3  4.573978e-17 -9.581526e-17  3.565632e-01  5.281033e-17
# PC4 -3.198568e-16 -1.516830e-16  5.281033e-17  1.734301e-01

round(var(d2),5)
#       PC1     PC2     PC3     PC4
#PC1 2.48024 0.00000 0.00000 0.00000
#PC2 0.00000 0.98977 0.00000 0.00000
#PC3 0.00000 0.00000 0.35656 0.00000
#PC4 0.00000 0.00000 0.00000 0.17343

# This is Big lambda diagonal matrix (eigenvalues on main diagonal)
sum(diag(var(d2)))      # 4

# covariances (off diagonal) all equal to 0  (PCs uncorrelated)
# PC1 with largest variance across states
```

```
# Use eigenvectors to define the PC variables.
#======================================================================

m1$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

# Score vectors are PC1, PC2, defined as follows

# PC1 = 0.536 Murder + 0.58Assault + 0.28 UrbanPop + 0.543 Rape
# A weighted average of crime rates (almost exclude UrbanPop)

# PC2 = 0.4 Murder - 0.87 UrbanPop
# Weighted average of Urban Pop and Murder

# transformed variables in the principal components space.
#===============================================================
# eigenvectors span a new p-dimensional space
# score vectors are the transformed observations in this new space
d2 = m1$x
head(d2)
#                   PC1        PC2         PC3          PC4
# Alabama    -0.9756604  1.1220012 -0.43980366  0.154696581
# Alaska     -1.9305379  1.0624269  2.01950027 -0.434175454
# Arizona    -1.7454429 -0.7384595  0.05423025 -0.826264240
# Arkansas    0.1399989  1.1085423  0.11342217 -0.180973554
# California -2.4986128 -1.5274267  0.59254100 -0.338559240
# Colorado   -1.4993407 -0.9776297  1.08400162  0.001450164
tail(m1$x)
#                      PC1        PC2         PC3        PC4
# Vermont       2.7732561  1.3881944  0.83280797 -0.1434337
# Virginia      0.0953667  0.1977278  0.01159482  0.2092464
# Washington    0.2147234 -0.9603739  0.61859067 -0.2186282
# West Virginia 2.0873931  1.4105263  0.10372163  0.1305831
# Wisconsin     2.0588120 -0.6051251 -0.13746933  0.1822534
# Wyoming       0.6231006  0.3177866 -0.23824049 -0.1649769

# Variance of the PCs are the eigenvalues
#=================================================================
apply(d2,2,var)
#       PC1       PC2       PC3       PC4
# 2.4802416 0.9897652 0.3565632 0.1734301

m2$values
# 2.4802416 0.9897652 0.3565632 0.1734301
```

```
# proportion of variance explained (PVE) by each PC
#==============================================================

# variance of PCs
aux=m1$sdev^2
#  2.4802416 0.9897652 0.3565632 0.1734301

sum(aux)   # 4
pve=aux/sum(aux)
# [1] 0.62006039 0.24744129 0.08914080 0.04335752

m2$values/4
# [1] 0.62006039 0.24744129 0.08914080 0.04335752
# each eigenvalue divided by 4

cumsum(pve) # [1] 0.6200604 0.8675017 0.9566425 1.0000000
#  87% variability in the dataset explained by PC1 & PC2

# plots
plot(pve, xlab="PC", ylab="% of Variance Explained", ylim=c(0,1),type='l')
grid()

plot(cumsum(pve), xlab="PC", ylab="Cumulative % of Variance Explained", ylim=c(0,1),type='l')
grid()
```

```
# biplots
#==============================================================

biplot(m1, scale=0)
biplot(m1, scale=0,cex=0.6)
grid()

head(d2)
#                  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

# rowname is State name, located at (PC1,PC2) coordinates

# mirror image (main diagonal line)
m1$rotation=-m1$rotation
m1$x=-m1$x
biplot(m1, scale=0,cex=0.6)
grid()

rot=m1$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

# Murder axis
slope1=rot[1,2]/rot[1,1]
slope1   # -0.7803345
abline(0,slope1)

# interpret the PCs
#===================================================================

# states are the observations

# states with large values in PC1 have high crime rates
#       (PC1 weights -col1- in rotation are 0.5359, 0.5831, 0.5434)
# California, Nevada, Florida   vs   North Dakota, Vermont

# states with large values in PC2 have large urban areas
#       (PC2 largest weight -col2- in rotation is 0.8728)
# California  vs   Mississippi
```

```
# original vs transformed values
#===========================================================

d3=data.frame(d1,d2)
head(d3)
#            Murder Assault UrbanPop Rape        PC1        PC2         PC3          PC4
# Alabama      13.2     236       58 21.2 -0.9756604  1.1220012 -0.43980366  0.154696581
# Alaska       10.0     263       48 44.5 -1.9305379  1.0624269  2.01950027 -0.434175454
# Arizona       8.1     294       80 31.0 -1.7454429 -0.7384595  0.05423025 -0.826264240
# Arkansas      8.8     190       50 19.5  0.1399989  1.1085423  0.11342217 -0.180973554
# California    9.0     276       91 40.6 -2.4986128 -1.5274267  0.59254100 -0.338559240
# Colorado      7.9     204       78 38.7 -1.4993407 -0.9776297  1.08400162  0.001450164
tail(d3)
#               Murder Assault UrbanPop Rape       PC1        PC2         PC3        PC4
# Vermont          2.2      48       32 11.2 2.7732561  1.3881944  0.83280797 -0.1434337
# Virginia         8.5     156       63 20.7 0.0953667  0.1977278  0.01159482  0.2092464
# Washington       4.0     145       73 26.2 0.2147234 -0.9603739  0.61859067 -0.2186282
# West Virginia    5.7      81       39  9.3 2.0873931  1.4105263  0.10372163  0.1305831
# Wisconsin        2.6      53       66 10.8 2.0588120 -0.6051251 -0.13746933  0.1822534
# Wyoming          6.8     161       60 15.6 0.6231006  0.3177866 -0.23824049 -0.1649769
```

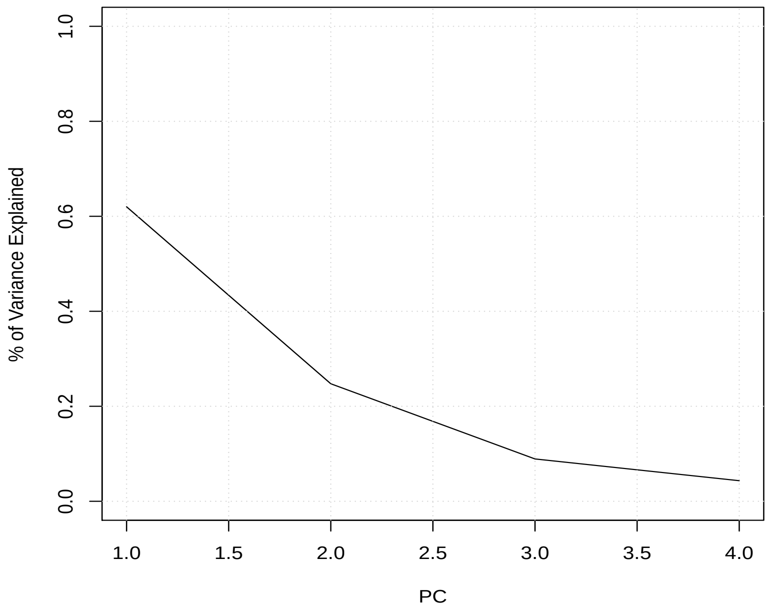

Figure 1: PVE by each pricipal component

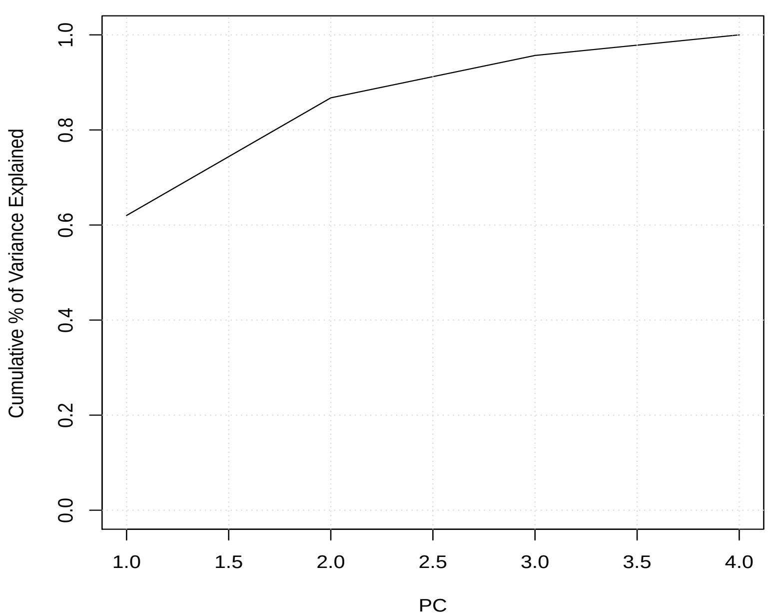

Figure 2: Cumulative PVE

Figure 3: Scatterplot on first two PC axes

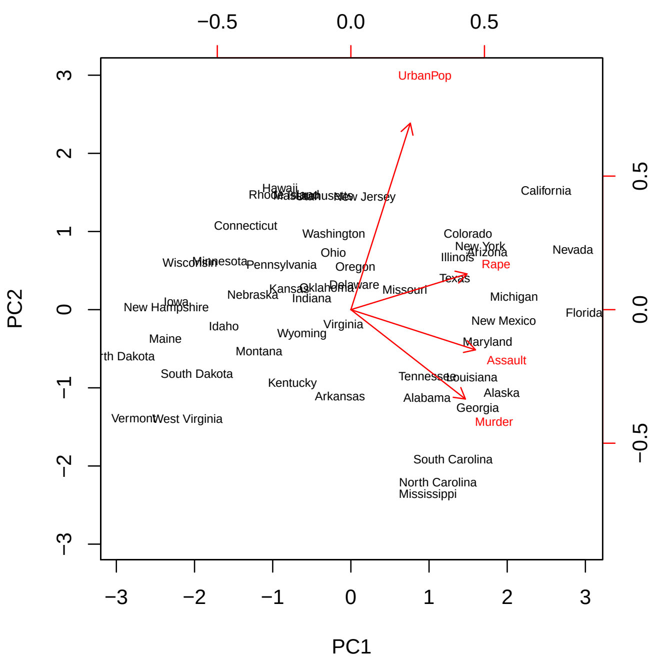

Figure 4: Reversed biplot