Chapter 7: Principal Component Analysis

Principal Component Analysis

Assumes that $X = (x_{ij})$ is normalised with $\frac{1}{n}\sum_{i=1}^n x_{ij} = 0$ and $\frac{1}{n}\sum_{i=1}^n x_{ij}^2 = 1$. We look for the 1st PC that ensures linear combination of the sample feature values $z_{i1} = \phi_{11}x_{i1} + \cdots + \phi_{p1}x_{ip}$ has largest sample variance. In other words, we seek for $\phi_{j1}, j = 1, \ldots, p$:

$$\min_{\phi_{11},\ldots,\phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \phi_{j1} x_{ij} \right)^2 \right\} \ s.t. \sum_{j=1}^p \phi_{j1}^2 = 1,$$

Then, step by step, look for the 2nd, 3rd, ..., p-th PC that also has the largest sample variance and linear independent with all the previous PC's:

$$\min_{\phi_{12},\ldots,\phi_{p2}} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \phi_{j2} x_{ij} \right)^2 \right\} \ s.t. \sum_{j=1}^p \phi_{j2}^2 = 1, \sum_{j=1}^p \phi_{j2}\phi_{j1} = 0,$$

$$\min_{\phi_{13},\ldots,\phi_{p3}} \left\{ \frac{1}{n} \sum_{i=1}^n \left(\sum_{j=1}^p \phi_{j3} x_{ij} \right)^2 \right\} \ s.t. \sum_{j=1}^p \phi_{j3}^2 = 1, \sum_{j=1}^p \phi_{j3}\phi_{j1} = 0, \sum_{j=1}^p \phi_{j3}\phi_{j2} = 0,$$

et al.

Principal Component Analysis

We perform PCA on the "USArrests" data set. The row of the data set contain 50 states, in alphabetical order.

```
> states <-row.names(USArrests)
> states
```

The column of the data contain the four variables.

```
> names(USArrests)
[1] "Murder"   "Assault"  "UrbanPop" "Rape"
> apply(USArrests, 2, var)
    Murder    Assault   UrbanPop       Rape
  18.97047 6945.16571  209.51878   87.72916
```

4 variables have vastly different means and variance, hence it is important to standardise the variables to have mean zero and std one before performing PCA.

Principal Component Analysis

prcomp() function is used to perform PCA. By default, the prcomp() function centres the variables to have mean zero. By using the option scale=TRUE, we scale the variables to have deviation one.

```
> pr_out <-prcomp(USArrests, scale = TRUE)
> names(pr_out)
[1] "sdev"     "rotation" "center"   "scale"    "x"
```

The center and scale components correspond to the means and standard deviations of the variables that were used for scaling prior to implementing PCA.

```
> pr_out$center
 Murder  Assault UrbanPop     Rape
  7.788  170.760   65.540   21.232
> pr_out$scale
   Murder   Assault  UrbanPop      Rape
 4.355510 83.337661 14.474763  9.366385
```

Principal Component Analysis

The rotation matrix provides the principal component loadings, each column contains the corresponding PCA loading vector.

```
> 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
```

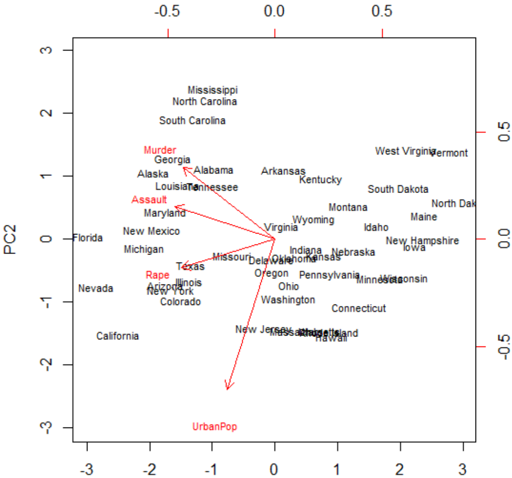
The 50 by 4 matrix x has its columns the principal component score vectors, i.e. the k-th column is the k-th PC score vector.

```
> dim(pr_out$x)
[1] 50  4
> head(pr_out$x, 5)
                  PC1        PC2         PC3        PC4
Alabama    -0.9756604  1.1220012 -0.43980366  0.1546966
Alaska     -1.9305379  1.0624269  2.01950027 -0.4341755
Arizona    -1.7454429 -0.7384595  0.05423025 -0.8262642
Arkansas    0.1399989  1.1085423  0.11342217 -0.1809736
California -2.4986128 -1.5274267  0.59254100 -0.3385592
```

Principal Component Analysis

We can plot the first two principal components as follows. The scale=0 argument to biplot() ensures that the arrows are scaled to represent the loadings.

```
> biplot(pr_out, scale=0, cex=0.7)
```



Principal Component Analysis

The summary() function for prcomp() outputs show the standard deviation, proportion of variance explained (PVE) of each principal component, and the cumulative PVE.

```
> summary(pr_out)
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
```

Principal Component Analysis

We can plot the PVE by each component, as well as the cumulative PVE, as follows.

```
par(mfrow = c(1, 2))
pr_imp <- summary(pr_out)$importance
plot(pr_imp[2,], xlab="Principal Component",
     ylab="PVE", ylim=c(0,1), type='b')
plot(pr_imp[3,], xlab="Principal Component",
     ylab="Cumulative PVE", ylim=c(0,1), type='b')
par(mfrow = c(1, 1))
```