> ####PCA

> #Carry out Principal Component Analysis (PCA) on the basis of covariance matrix

> D=read.table(file.choose(),header=TRUE)

> n1= names(D)

> n1

[1] "LENGTH" "LEFTHEIGHT" "RIGHTHEIGHT" "LOWERMARGIN" "UPPERMARGIN" "DIAGONAL"

> d=dim(D)

> d

[1] 200 6

> n=dim(D)[1]

> n

[1] 200

> p=dim(D)[2]

> p

[1] 6

> M=apply(D,2,mean)

> M

LENGTH LEFTHEIGHT RIGHTHEIGHT LOWERMARGIN UPPERMARGIN DIAGONAL

214.8960 130.1215 129.9565 9.4175 10.6505 140.4835

> D=as.matrix(D)

> S=(t(D)%\*%D)-(n\*M%\*%t(M))

> V=S/n

> V

LENGTH LEFTHEIGHT RIGHTHEIGHT LOWERMARGIN UPPERMARGIN DIAGONAL

LENGTH 0.141084 0.0312860 0.0229760 -0.1027300 -0.0184480 0.0838840

LEFTHEIGHT 0.031286 0.1296877 0.1078853 0.2147237 0.1045143 -0.2082952

RIGHTHEIGHT 0.022976 0.1078853 0.1624577 0.2827112 0.1293468 -0.2392677

LOWERMARGIN -0.102730 0.2147237 0.2827112 2.0764438 0.1637162 -1.0318113

UPPERMARGIN -0.018448 0.1045143 0.1293468 0.1637162 0.6414998 -0.5468668

DIAGONAL 0.083884 -0.2082952 -0.2392677 -1.0318113 -0.5468668 1.3210777

> eval=eigen(V)$values

> evec=eigen(V)$vectors

> eval

[1] 2.98530335 0.93094242 0.24219664 0.19368545 0.08478579 0.03533710

#**Correlation shows you how the two variables are related**.

> c=cor(D)

> c

LENGTH LEFTHEIGHT RIGHTHEIGHT LOWERMARGIN UPPERMARGIN DIAGONAL

LENGTH 1.00000000 0.2312926 0.1517628 -0.1898009 -0.06132141 0.1943015

LEFTHEIGHT 0.23129257 1.0000000 0.7432628 0.4137810 0.36234960 -0.5032290

RIGHTHEIGHT 0.15176280 0.7432628 1.0000000 0.4867577 0.40067021 -0.5164755

LOWERMARGIN -0.18980092 0.4137810 0.4867577 1.0000000 0.14185134 -0.6229827

UPPERMARGIN -0.06132141 0.3623496 0.4006702 0.1418513 1.00000000 -0.5940446

DIAGONAL 0.19430146 -0.5032290 -0.5164755 -0.6229827 -0.59404464 1.0000000

> evec

[,1] [,2] [,3] [,4] [,5] [,6]

[1,] -0.04377427 0.01070966 -0.3263165 0.5616918 0.75257278 0.09809807

[2,] 0.11216159 0.07144697 -0.2589614 0.4554588 -0.34680082 -0.76651197

[3,] 0.13919062 0.06628208 -0.3447327 0.4153296 -0.53465173 0.63169678

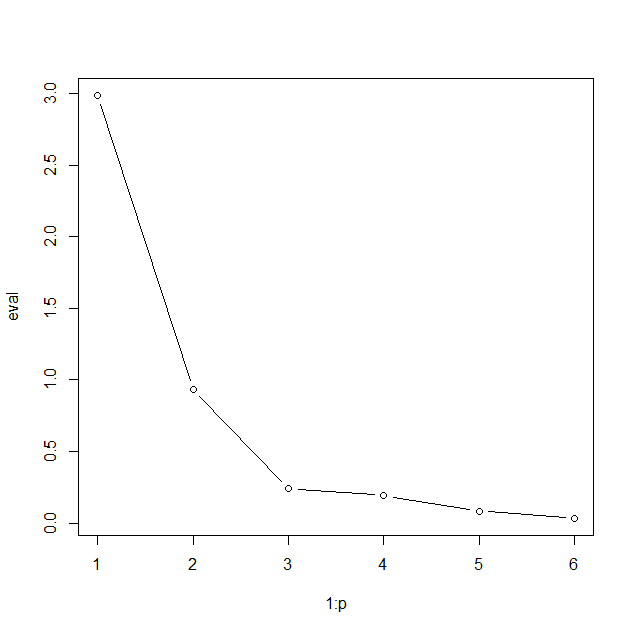
[4,] 0.76830499 -0.56307225 -0.2180222 -0.1861082 0.09996771 -0.02221711

[5,] 0.20176610 0.65928988 -0.5566857 -0.4506985 0.10190229 -0.03485874

[6,] -0.57890193 -0.48854255 -0.5917628 -0.2584483 -0.08445895 -0.04567946

> #(b) Draw the scree plot.

> plot(1:p,eval,type="b")



#As it can be seen from the scree plot, three PCs should be appropriate.

#(c) How many PCs are required to explain 90% variation

> prop=cumsum(eval)/sum(eval)

> prop

[1] 0.6675170 0.8756767 0.9298321 0.9731404 0.9920986 1.0000000

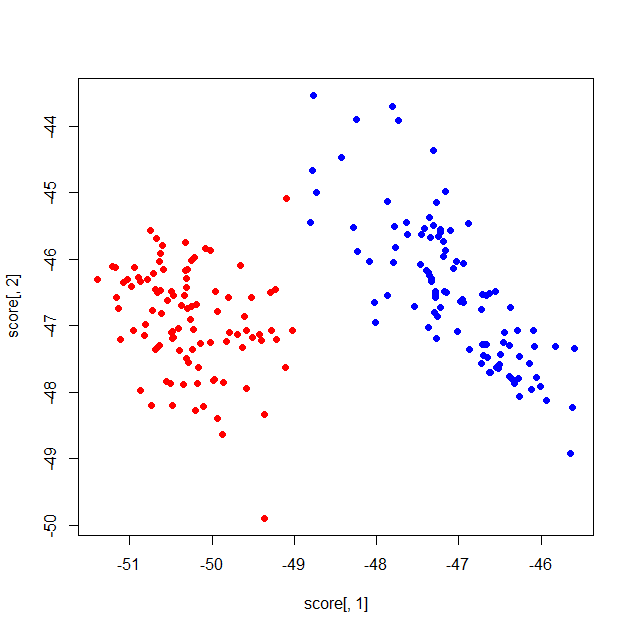
#Just as in the case of scree plot, we can see that three PCs will be able to explain 90% of the variation in the data.

##d)

#Will Draw the score plot for the first two PCs and mark the observations of the two groups (genuine and forged) by two different colors. & will see are the two groups clearly distinguishable?

> score=D%\*%evec

> plot(score[,1],score[,2],col=rep(c("red","blue"),each=100),pch=16)



> #(e) Now will see On the basis of first two PCs, which variables are important?

> #Q1e

> evec

[,1] [,2] [,3] [,4] [,5] [,6]

[1,] -0.04377427 0.01070966 -0.3263165 0.5616918 0.75257278 0.09809807

[2,] 0.11216159 0.07144697 -0.2589614 0.4554588 -0.34680082 -0.76651197

[3,] 0.13919062 0.06628208 -0.3447327 0.4153296 -0.53465173 0.63169678

[4,] 0.76830499 -0.56307225 -0.2180222 -0.1861082 0.09996771 -0.02221711

[5,] 0.20176610 0.65928988 -0.5566857 -0.4506985 0.10190229 -0.03485874

[6,] -0.57890193 -0.48854255 -0.5917628 -0.2584483 -0.08445895 -0.04567946

> #On the basis of the first PC, the Lower Margin variable is the most important.

> #On the basis of the second PC, the Upper Margin variable is the most important.