Data mining 2nd Report  
(Student Number: 1651046, Name: Daiki Kubo)

1. Explain Principal Component Analysis in maximum two pages.

Answer:  
　Principal Component Analysis (PCA) is a statistical technique for displaying the feature quantity of the pair that does not have a linear correlation, orthogonally transforming a set of observed values.

In principal component analysis, a feature value which is obtained by orthogonal transformation of the observed values called principal component (PC).  
　Furthermore, Conversion to make principal component maximizes the variance of the first principal component. Following principal component is chosen so as to maximize the dispersion under constraint condition that is perpendicular to the main component was determined previously.  
　Maximizing the variance of the principal components is to have the principal component described capability to changes in the observed values. The selected principal components are orthogonal to each other, so it can represent a set of given observations as a linear combination.

　Therefore, Principal components eigenvalue decomposition of the covariance matrix, or may be obtained by singular value decomposition of the corresponding deviation matrix to the correlation coefficient matrix. Principal component analysis is a major tool in exploratory data analysis, it is also used in the prediction model building.

　Principal component analysis eigenvalue decomposition, or with respect to the observed value of covariance matrix or correlation matrix is performed by singular value decomposition of the (usually normalized) data matrix.

　The results of principal component analysis is evaluated by principal component score and the factor loadings.

1. Apply PCA to the data of the file "PCATutorial.txt" and try to plot Comp1 vs. Comp2 and Comp1 vs. Comp3.  
   Answer:

> data<-read.table("PCATutorial.txt", header=T)

> data.pca <- prcomp(data[,2:11],center=T,scale.=T)

> data.pca

Standard deviations:

[1] 1.9691433 1.3178499 1.0692950 0.9021047 0.8263766 0.7586129 0.6530684

[8] 0.5282379 0.4959506 0.4676208

Rotation:

PC1 PC2 PC3 PC4 PC5 PC6

S1 -0.3440561 0.35546953 0.04317156 -0.1464110 0.0005650767 -0.06374166

S2 -0.2689869 -0.15770988 0.23031671 -0.7942485 0.3033979955 -0.01952665

S3 -0.4039714 0.18719487 0.20595069 -0.1011077 -0.4002112772 -0.02213714

S4 -0.3844764 0.31249999 0.19424935 0.1397021 0.1096796403 -0.08207743

S5 -0.3812649 0.22311880 -0.24898032 0.2000384 -0.3563886772 -0.02590634

S6 -0.3424906 0.07128363 -0.23129501 0.2909822 0.6998242734 0.16991610

S7 -0.2401414 -0.31469277 -0.51289330 -0.2307860 -0.2773255574 0.54122201

S8 -0.3143276 -0.44645802 -0.21340932 0.1673518 0.1348923980 -0.20643118

S9 -0.2378182 -0.52482080 0.13470537 0.1345532 -0.1478625884 -0.54998817

S10 -0.1502216 -0.29565911 0.65015747 0.3064870 -0.0448884762 0.56636053

PC7 PC8 PC9 PC10

S1 -0.813808119 0.1771715 0.18439752 -0.01509072

S2 0.209319671 0.1728061 -0.14679669 -0.17674720

S3 0.145650451 -0.2707449 -0.29436154 0.63820221

S4 0.287149785 -0.5045879 0.43733673 -0.38752862

S5 0.268143504 0.5218662 -0.28208064 -0.38903702

S6 0.141396894 0.2134922 0.01645686 0.39367813

S7 -0.008189715 -0.1284809 0.37341318 0.03370899

S8 -0.287918670 -0.4226750 -0.49888454 -0.23890660

S9 0.054585730 0.2788106 0.44127111 0.17880170

S10 -0.111331547 0.1414797 -0.06043556 -0.12057045

> summary(data.pca)

Importance of components:

PC1 PC2 PC3 PC4 PC5 PC6 PC7

Standard deviation 1.9691 1.3178 1.0693 0.90210 0.82638 0.75861 0.65307

Proportion of Variance 0.3877 0.1737 0.1143 0.08138 0.06829 0.05755 0.04265

Cumulative Proportion 0.3877 0.5614 0.6758 0.75714 0.82543 0.88298 0.92563

PC8 PC9 PC10

Standard deviation 0.5282 0.4960 0.46762

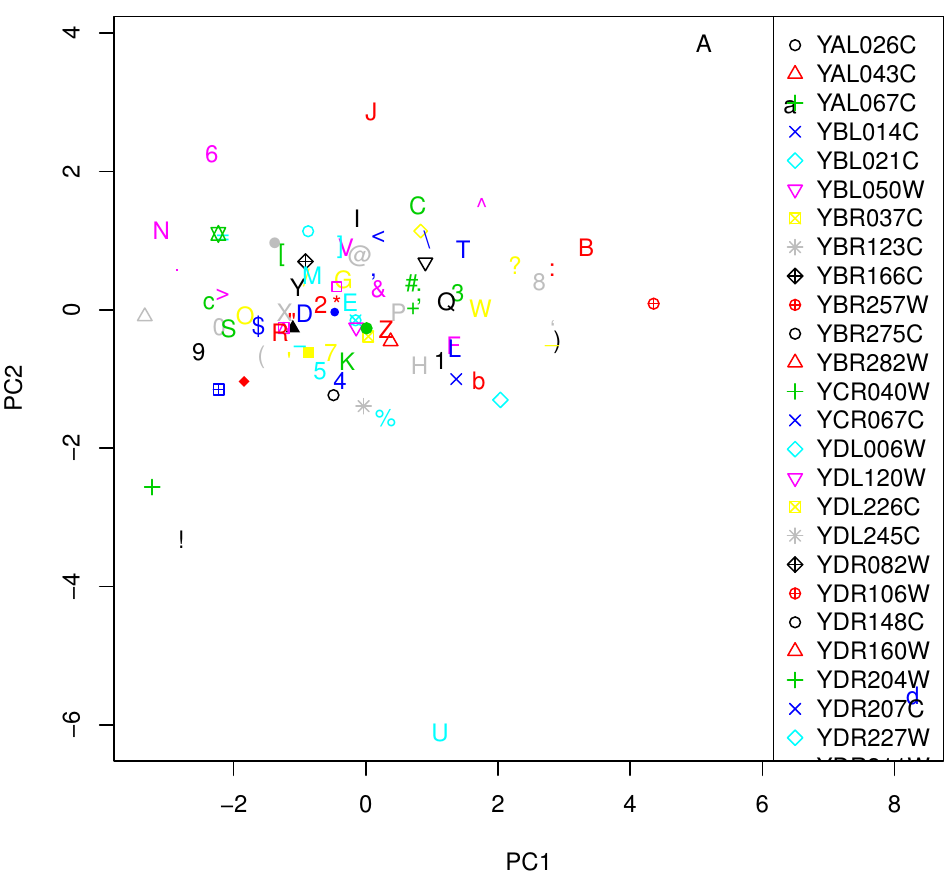
Proportion of Variance 0.0279 0.0246 0.02187

Cumulative Proportion 0.9535 0.9781 1.00000

>plot(data.pca$x[,1:2],col=as.integer(data[,1])  
 ,pch=as.integer(data[,1]))

> par(new=T)

> legend("topright",legend=levels(data[,1]),col=1:10,pch=1:10)



1. Apply PLS to the data of the file "PLSTutorial.txt" by considering MEDV as response variable and others as predictor variables. Plot measured data vs. predicted results.  
   Answer:  
   > library(pls)  
   > data<-read.table("PLSTutorial.txt",header=T)  
   > data.plsr<-plsr(MEDV~CRIM+ZN+INDUS+CHAS+NOX+RM+  
    AGE+DIS+RAD+TAX+PTRATIO+B+LSTAT,  
    ncomp=5,data=data,validation="LOO")  
   > summary(data.plsr)  
    Data: X dimension: 506 13   
    Y dimension: 506 1  
   Fit method: kernelpls  
   Number of components considered: 5  
     
   VALIDATION: RMSEP  
   Cross-validated using 506 leave-one-out segments.  
      
   (Intercept) 1 comps 2 comps 3 comps 4 comps 5 comps  
   CV 9.206 8.032 7.9 7.634 6.536 5.907  
   adjCV 9.206 8.032 7.9 7.634 6.536 5.907  
     
   TRAINING: % variance explained  
   1 comps 2 comps 3 comps 4 comps 5 comps  
   X 80.51 94.45 98.97 99.34 99.80  
   MEDV 24.23 26.94 32.05 51.05 60.08

> plot(data.plsr,ncomp=5,asp=1,line=TRUE)  
