HDDA Tutorial: PCA : Solutions

Department of Econometrics and Business Statistics, Monash University

Tutorial 6

The data for todays tutorial are a subset of the well-known Boston Housing dataset created by Harrison and Rubinfeld and used in their 1978 paper, Hedonic prices and the demand for clean air, J. Environ. Economics & Management, vol.5, 81-102. The data were obtained from the UCI Machine Learning Repository. In this dataset each observation corresponds to a town (or suburb) in or around Boston. The towns are numbered rather than named and are stored in the variable `Town`. Excluding `Town` are 14 variables which are summarised below:

- CRIM: per capita crime rate by town
- ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
- INDUS: proportion of non-retail business acres per town
- CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
- NOX: nitric oxides concentration (parts per 10 million)
- RM: average number of rooms per dwelling
- AGE: proportion of owner-occupied units built prior to 1940
- DIS: weighted distances to five Boston employment centres
- RAD: index of accessibility to radial highways
- TAX: full-value property-tax rate per $10,000
- PTRATIO: pupil-teacher ratio by town
- B: 1000(Bk 0.63) 2 where Bk is the proportion of African Americans by town
- LSTAT: % lower status of the population
- MEDV: Median value of owner-occupied homes in $1000s

Answer the following questions.

1. Should the data be standardised prior to carrying out Principal Components?

The data are measured in different units so the data should be standardised before carrying out PCA. This ensures that results are not sensitive to the units of measurement.

2. Carry out Principal Components Analysis on this data.

```
#First load required packages
library(tidyverse)
Boston<-readRDS('Boston.rds')
Boston%>%
  column_to_rownames('Town')%>% #see comment below
  prcomp(scale.=TRUE)->pcaout

#This sets the column Town to be the row name.  There is a
#debate around whether dataframes should have rownames but the
#existing prcomp function works better when observation IDs are
#rownames rather than a separate column
```

3. What proportion of total variance is explained by the first four principal components together?
4. What proportion of total variance is explained by the second principal component (on its own)?
5. How many PCAs would be selected using Kaisers Rule?

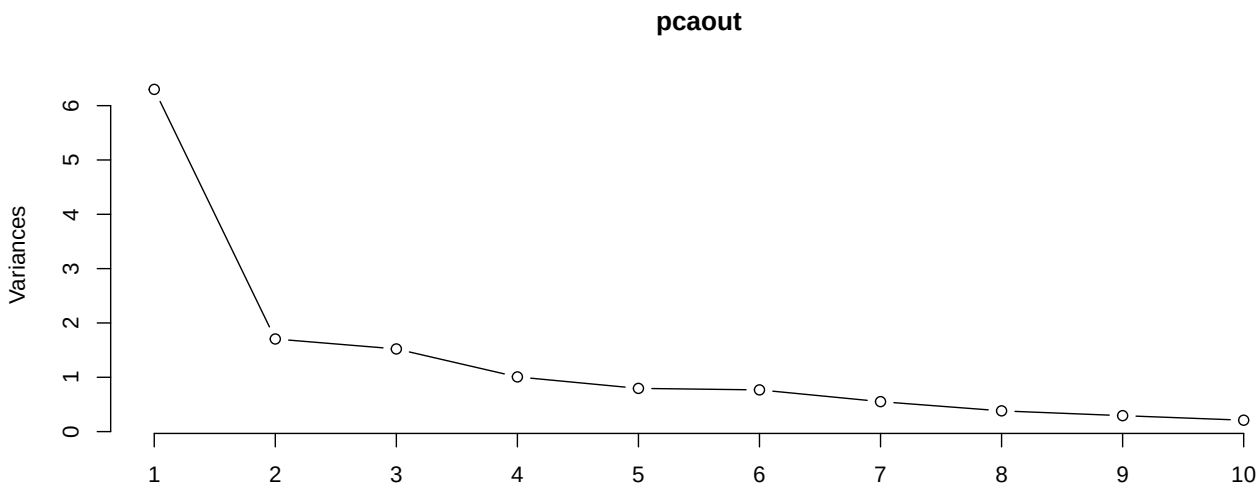
```
#Questions 4-5 can be answered using the summary function
summary(pcaout)
```

```
## Importance of components:
##                            PC1    PC2    PC3    PC4     PC5     PC6    PC7
## Standard deviation     2.5098 1.3058 1.2335 1.0033 0.89191 0.87632 0.7427
## Proportion of Variance 0.4499 0.1218 0.1087 0.0719 0.05682 0.05485 0.0394
## Cumulative Proportion  0.4499 0.5717 0.6804 0.7523 0.80913 0.86399 0.9034
##                            PC8     PC9   PC10    PC11    PC12    PC13
## Standard deviation     0.61795 0.54176 0.4582 0.42482 0.37896 0.32909
## Proportion of Variance 0.02728 0.02096 0.0150 0.01289 0.01026 0.00774
## Cumulative Proportion  0.93066 0.95162 0.9666 0.97951 0.98977 0.99750
##                           PC14
## Standard deviation     0.1870
## Proportion of Variance 0.0025
## Cumulative Proportion  1.0000
```

```
#Proportion of variance explained by the first four PCs together is 75.23%
#Proportion of variance explained by the second PC alone is 12.18%
#By Kaisers rule select 4 PCs (variance and standard deviation greater than 1.)
```

6. Produce and Interpret the Scree Plot. How many PCs should be used?

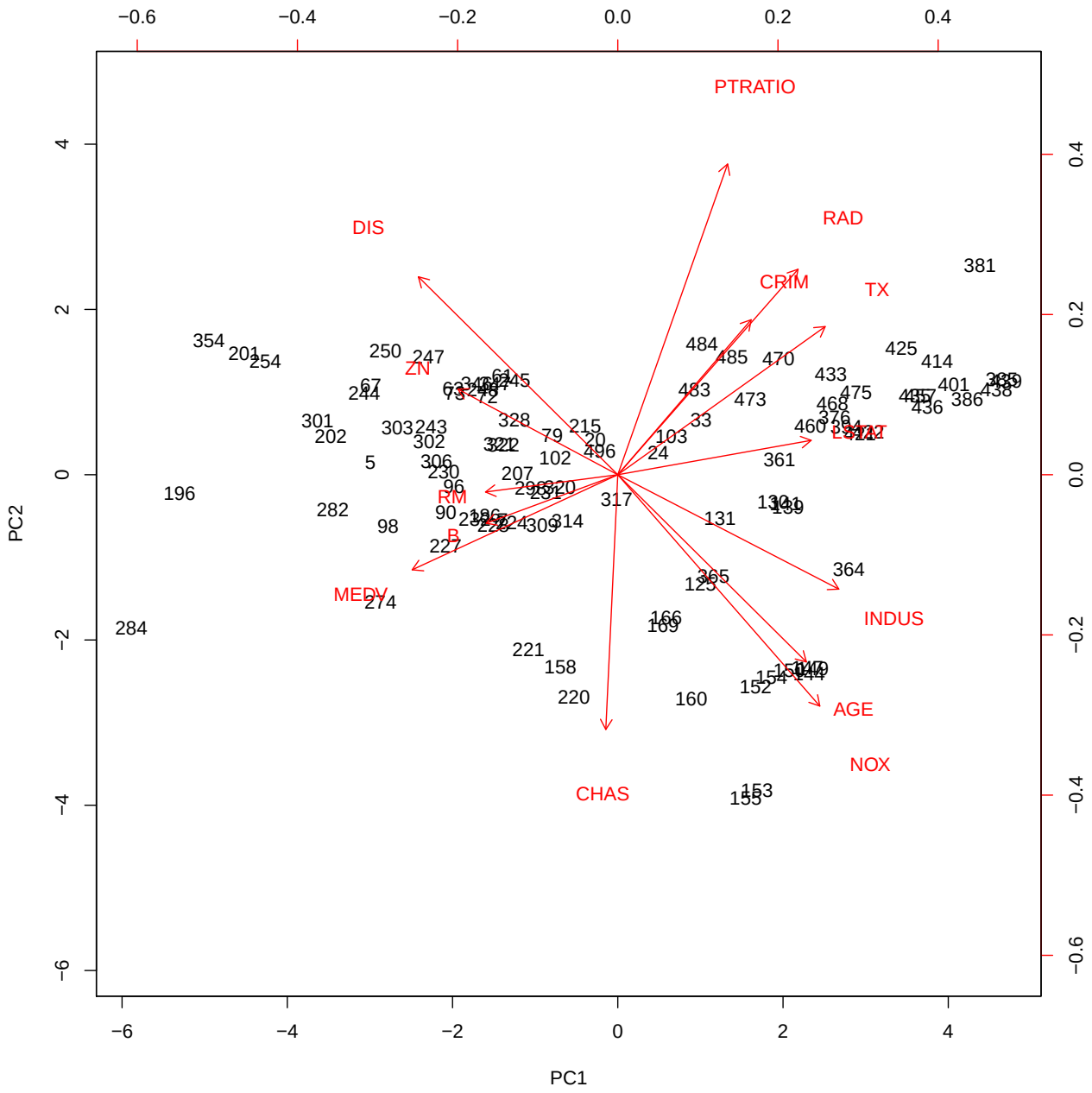
```
screeplot(pcaout,type = 'l')
```



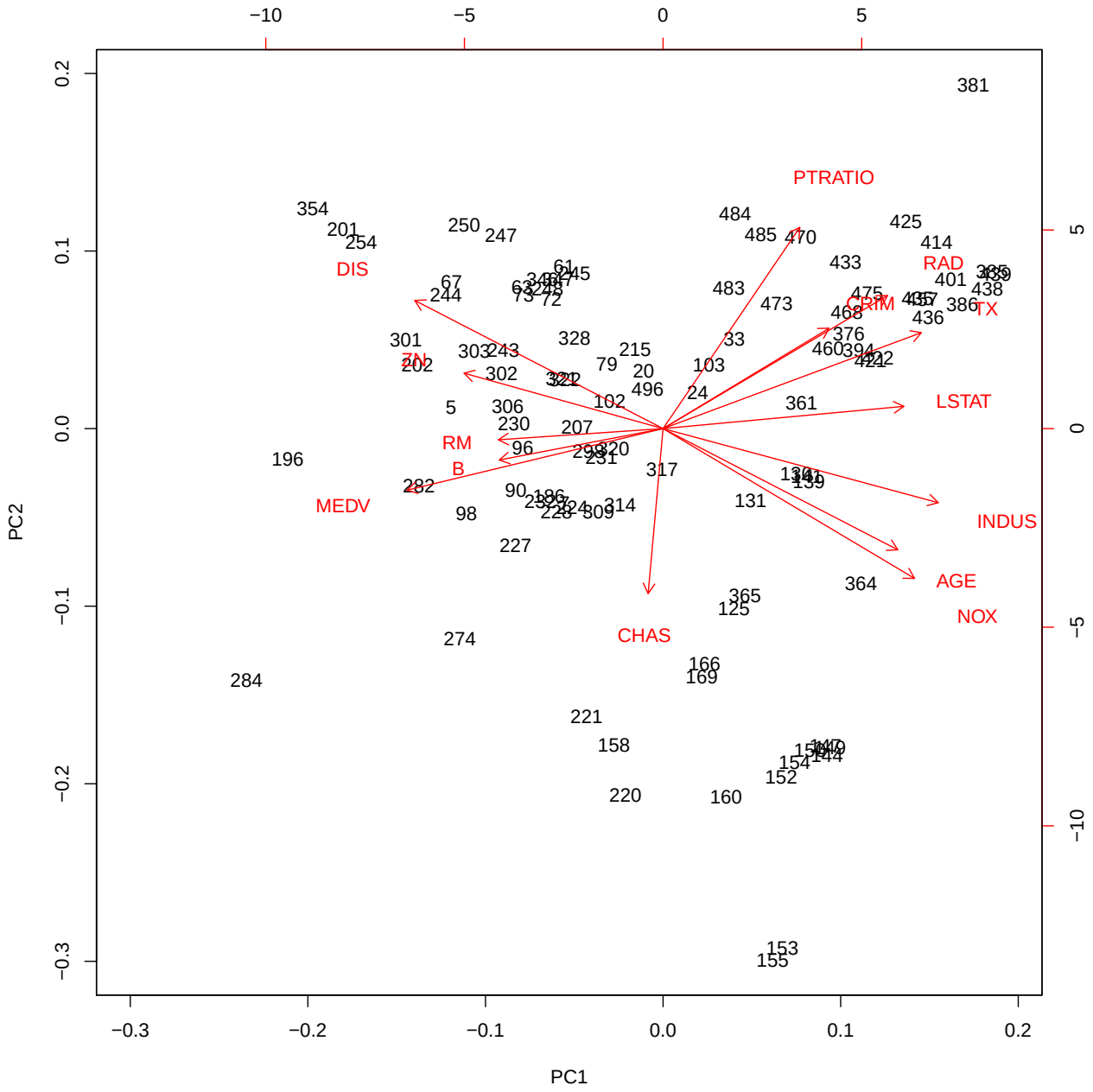

```
# The elbow appears at the second PC, therefore 2 PCs should be used.
# Note that this is different to Kaiser's rule.
```

All remaining quesions should be answered using a biplot

```
#Correlation biplot used to answer Q7, Q8 and Q9
biplot(pcaout,scale=0)
```




```
#Distance biplot used to answer Q10, Q11
biplot(pcaout)
```




```
#Either biplot can be used to answer Q12 and Q13
```

7. Name two variables that have a strong positive association with one another.

Any two arrows pointing in the same direction for example AGE and NOX.

8. Name two variables that have a strong negative association with one another.

Any two arrows pointing in the opposite direction for example AGE and DIS.

9. Name two variables that are only weakly associated with another.

Any two arrows at a 90 degree angle for example NOX and PTRATIO.

10. Name two towns that are similar to one another.

Any two points that are close, e.g 153 and 155.

11. Name two towns that are different to one another.

Any points that are far apart, e.g. 284 and 381.

12. Describe the characteristics of town 354.

Town 354 is characterised by an association with high values of DIS meaning it is far away from employment centres and high levels of ZN meaning that it is mostly a residential area. It also is characterised by a negative association with AGE and INDUS. The value of MEDV is also likely to be high for town 354. Overall Town 301 is probably a new outlying suburb and a result of urban sprawl.

13. Name a town that has a high level of crime and a high pupil-teacher ratio.

Town 381. To see this, imagine the CRIM and PTRATIO arrow extending out. Then draw a line from point 381 to those lines crossing at right angles. The point where this line crosses will be far away from the center of the plot. Remember that this is only an approximation.