HDDA Tutorial: PCA

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?
2. Carry out Principal Components Analysis on this data.
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?
6. Produce and Interpret the Scree Plot. How many PCs should be used?

All remaining quesions should be answered using a biplot

7. Name two variables that have a strong positive association with one another.
8. Name two variables that have a strong negative association with one another.
9. Name two variables that are only weakly associated with another.
10. Name two towns that are similar to one another.
11. Name two towns that are different to one another.
12. Describe the characteristics of town 354.
13. Name a town that has a high level of crime and a high pupil-teacher ratio.