PCA Example

(Boston Data)

Intelligent Data Analysis 2020

Martin Russell

Data (Prototask.data)

Prototask - Notepad

File Edit Format View Help

```
0.10574 0.00 27.740 0 0.6090 5.9830 98.80 1.8681 4 711.0 20.10 390.11 18.07 13.60
7.75223 0.00 18.100 0 0.7130 6.3010 83.70 2.7831 24 666.0 20.20 272.21 16.23 14.90
0.02763 75.00 2.950 0 0.4280 6.5950 21.80 5.4011 3 252.0 18.30 395.63 4.32 30.80
0.09266 34.00 6.090 0 0.4330 6.4950 18.40 5.4917 7 329.0 16.10 383.61 8.67 26.40
15.17720 0.00 18.100 0 0.7400 6.1520 100.00 1.9142 24 666.0 20.20 9.32 26.45 8.70
0.37578 0.00 10.590 1 0.4890 5.4040 88.60 3.6650 4 277.0 18.60 395.24 23.98 19.30
0.21124 12.50 7.870 0 0.5240 5.6310 100.00 6.0821 5 311.0 15.20 386.63 29.93 16.50
1.19294 0.00 21.890 0 0.6240 6.3260 97.70 2.2710 4 437.0 21.20 396.90 12.26 19.60
0.13117 0.00 8.560 0 0.5200 6.1270 85.20 2.1224 5 384.0 20.90 387.69 14.09 20.40
0.06911 45.00 3.440 0 0.4370 6.7390 30.80 6.4798 5 398.0 15.20 389.71 4.69 30.50
4.75237 0.00 18.100 0 0.7130 6.5250 86.50 2.4358 24 666.0 20.20 50.92 18.13 14.10
0.08199 0.00 13.920 0 0.4370 6.0090 42.30 5.5027 4 289.0 16.00 396.90 10.40 21.70
0.34940 0.00 9.900 0 0.5440 5.9720 76.70 3.1025 4 304.0 18.40 396.24 9.97 20.30
10.06230 0.00 18.100 0 0.5840 6.8330 94.30 2.0882 24 666.0 20.20 81.33 19.69 14.10
0.13642 0.00 10.590 0 0.4890 5.8910 22.30 3.9454 4 277.0 18.60 396.90 10.87 22.60
0.06047 0.00 2.460 0 0.4880 6.1530 68.80 3.2797 3 193.0 17.80 387.11 13.15 29.60
4.89822 0.00 18.100 0 0.6310 4.9700 100.00 1.3325 24 666.0 20.20 375.52 3.26 50.00
6.65492 0.00 18.100 0 0.7130 6.3170 83.00 2.7344 24 666.0 20.20 396.90 13.99 19.50
0.14455 12.50 7.870 0 0.5240 6.1720 96.10 5.9505 5 311.0 15.20 396.90 19.15 27.10
0.02875 28.00 15.040 0 0.4640 6.2110 28.90 3.6659 4 270.0 18.20 396.33 6.21 25.00
0.03548 80.00 3.640 0 0.3920 5.8760 19.10 9.2203 1 315.0 16.40 395.18 9.25 20.90
1.20742 0.00 19.580 0 0.6050 5.8750 94.60 2.4259 5 403.0 14.70 292.29 14.43 17.40
0.08826 0.00 10.810 0 0.4130 6.4170 6.60 5.2873 4 305.0 19.20 383.73 6.72 24.20
0.18836 0.00 6.910 0 0.4480 5.7860 33.30 5.1004 3 233.0 17.90 396.90 14.15 20.00
0.08221 22.00 5.860 0 0.4310 6.9570 6.80 8.9067 7 330.0 19.10 386.09 3.53 29.60
0.15086 0.00 27.740 0 0.6090 5.4540 92.70 1.8209 4 711.0 20.10 395.09 18.06 15.20
0.80271 0.00 8.140 0 0.5380 5.4560 36.60 3.7965 4 307.0 21.00 288.99 11.69 20.20
0.10793 0.00 8.560 0 0.5200 6.1950 54.40 2.7778 5 384.0 20.90 393.49 13.00 21.70
```

- Rows correspond to Boston neighbourhoods
- 506 neighbourhoods in total

Column labels

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

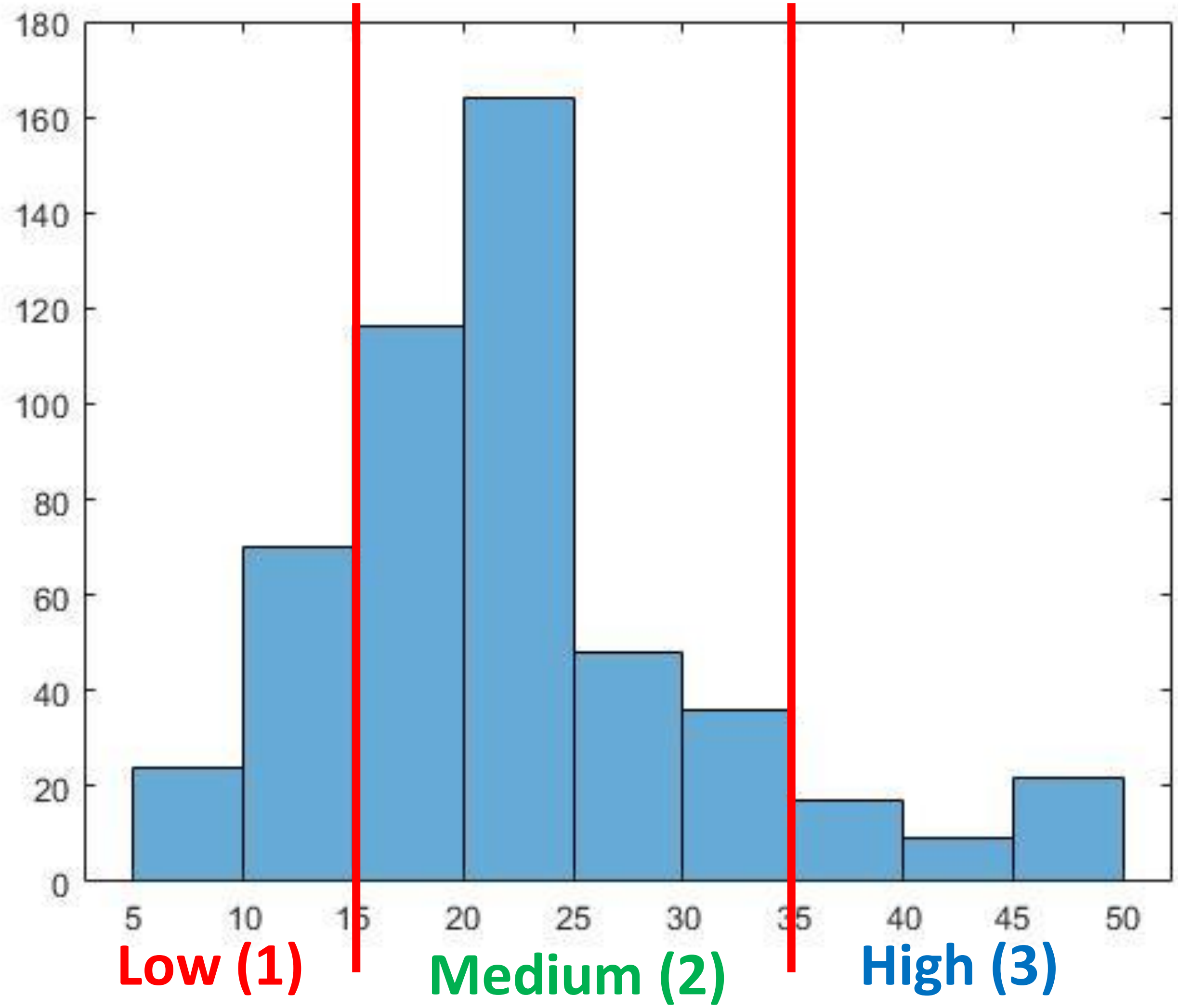
Task

- Two tasks associated with this data set:
  1. **price** - can the median value of a home be predicted from the other 13 variables?
  2. **nox** – can the nitrous oxide level be predicted from the other 13 variables?
- Use PCA to try to answer these questions

Data pre-processing – Task 1

- **Price** - median value of a home – is column 14
- Create two data sets – **X** and **L**
  - **X** consists of columns 1 to 13 of the original data
    - This is the data we will apply PCA to
  - **L** consists of column 14 of the original data
    - This is the 'price' data. We will use this to colour-code (label) the points in the PCA plot
    - First we need to divide the elements of **L** into categories
    - Look at a histogram of the values that **L** takes

Histogram of the ‘price’ values



Apply PCA to the data (‘price’ excluded)

1. Calculate the sample mean vector
2. Subtract the mean from each data point

   ```
   b = mean(X,1);
   ```

3. Calculate the covariance matrix of X

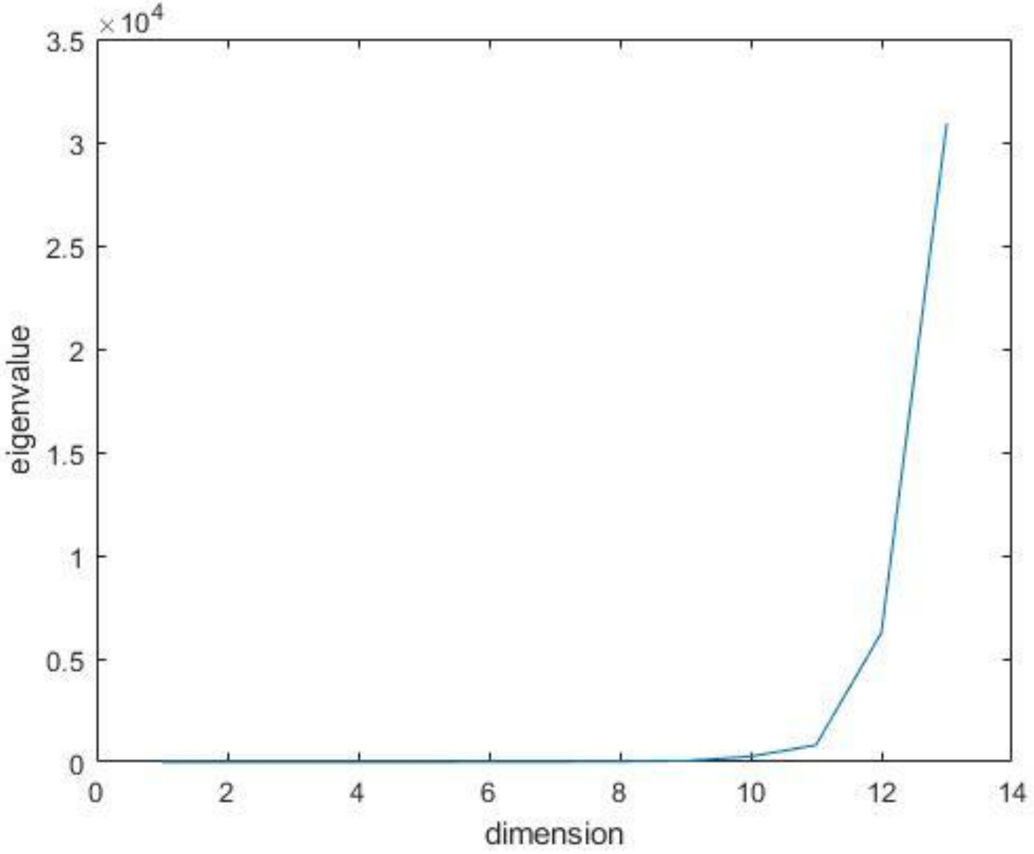
   ```
   C = X’*X/(N-1);
   ```
   (N = number of data points = 506)

4. Apply eigenvalue decomposition

   ```
   [U,D] = eig(C);
   ```

5. Find the two biggest eigenvalues

Eigenvalues



- Biggest eigenvalues are numbers 13 and 12
- Corresponding eigenvalues are 13th and 12th columns of U

Visualise

6. Identify the two principal components

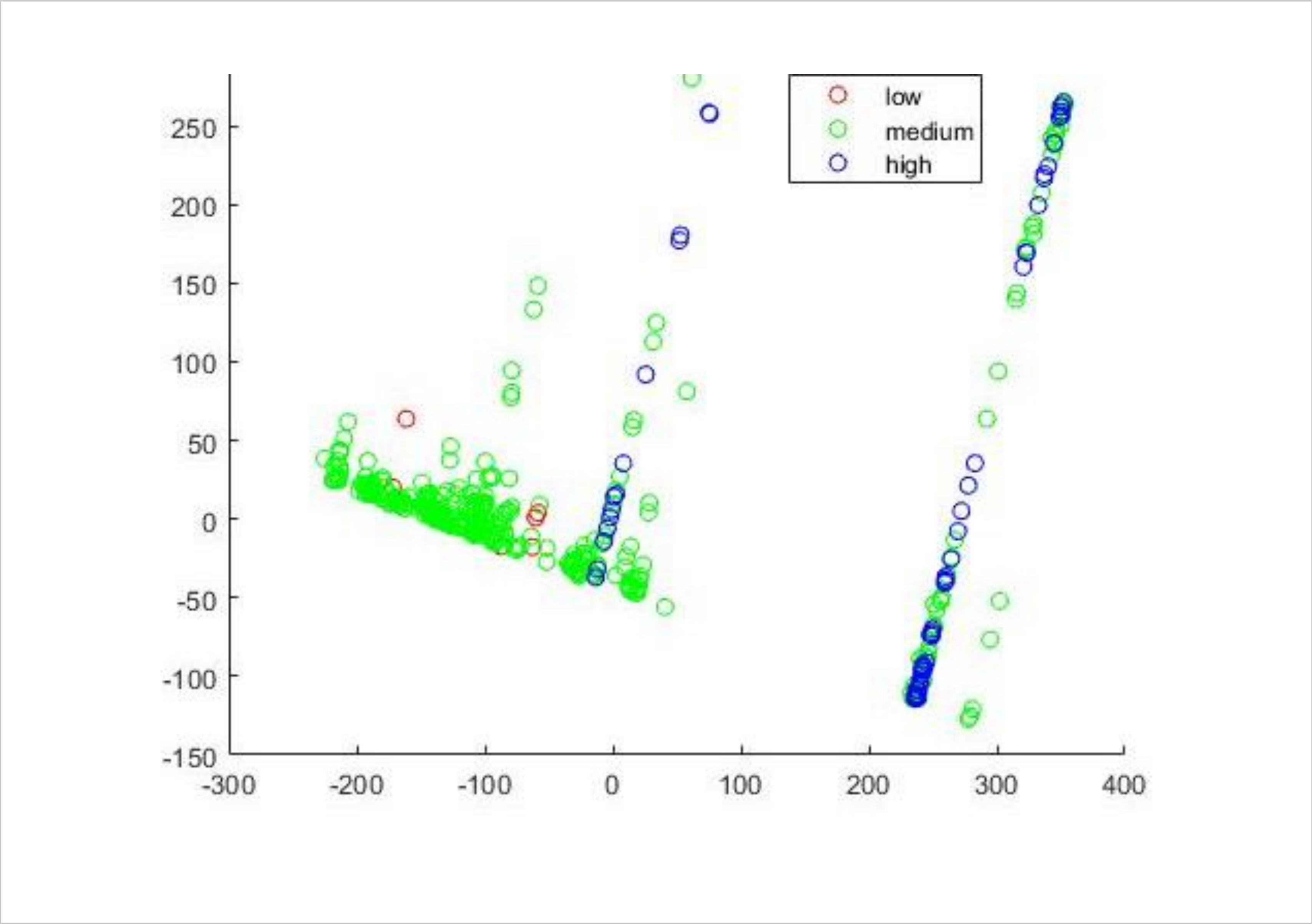
```
e1 = U(:,13);
e2 = U(:,12);
```

7. Project the data points onto the plane

```
x1 = X*e1;
x2 = X*e2;
```

8. Plot the data

```
scatter(x1(L1==1),x2(L1==1),'r');
scatter(x1(L1==2),x2(L1==2),'g');
scatter(x1(L1==3),x2(L1==3),'b');
```



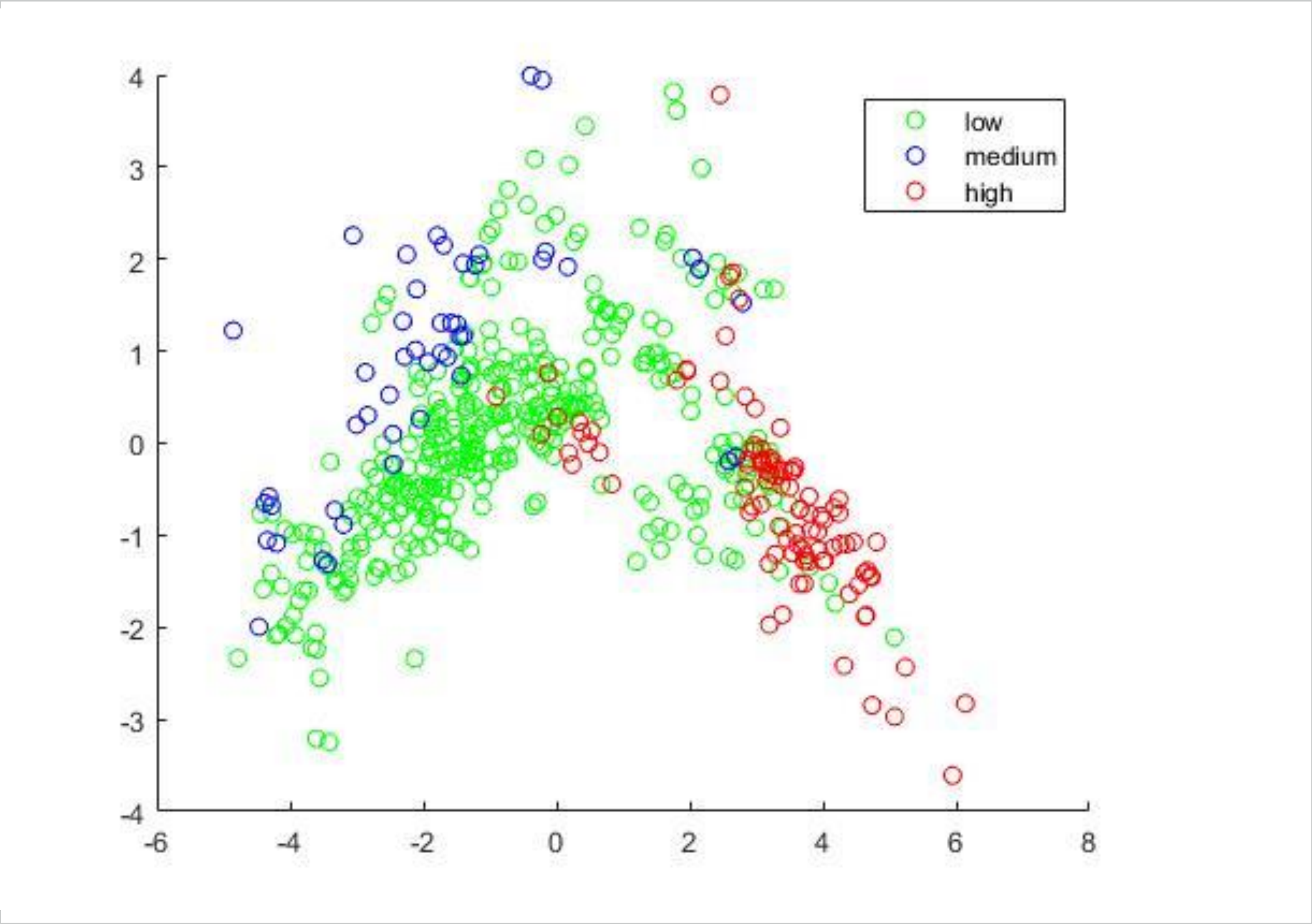
Basic PCA plot (‘price’ colour coded)

Comments on Basic PCA plot

- Not very informative
- May be due to the difference in the dynamic ranges of the numbers in the different columns
- Normalise each column (for example to have standard deviation 1)

```
m = mean(X,1);
v = var(X,1);
X=(X-m)./sqrt(v);
```

- Now repeat the process on slides 7, 8, 9 and 10



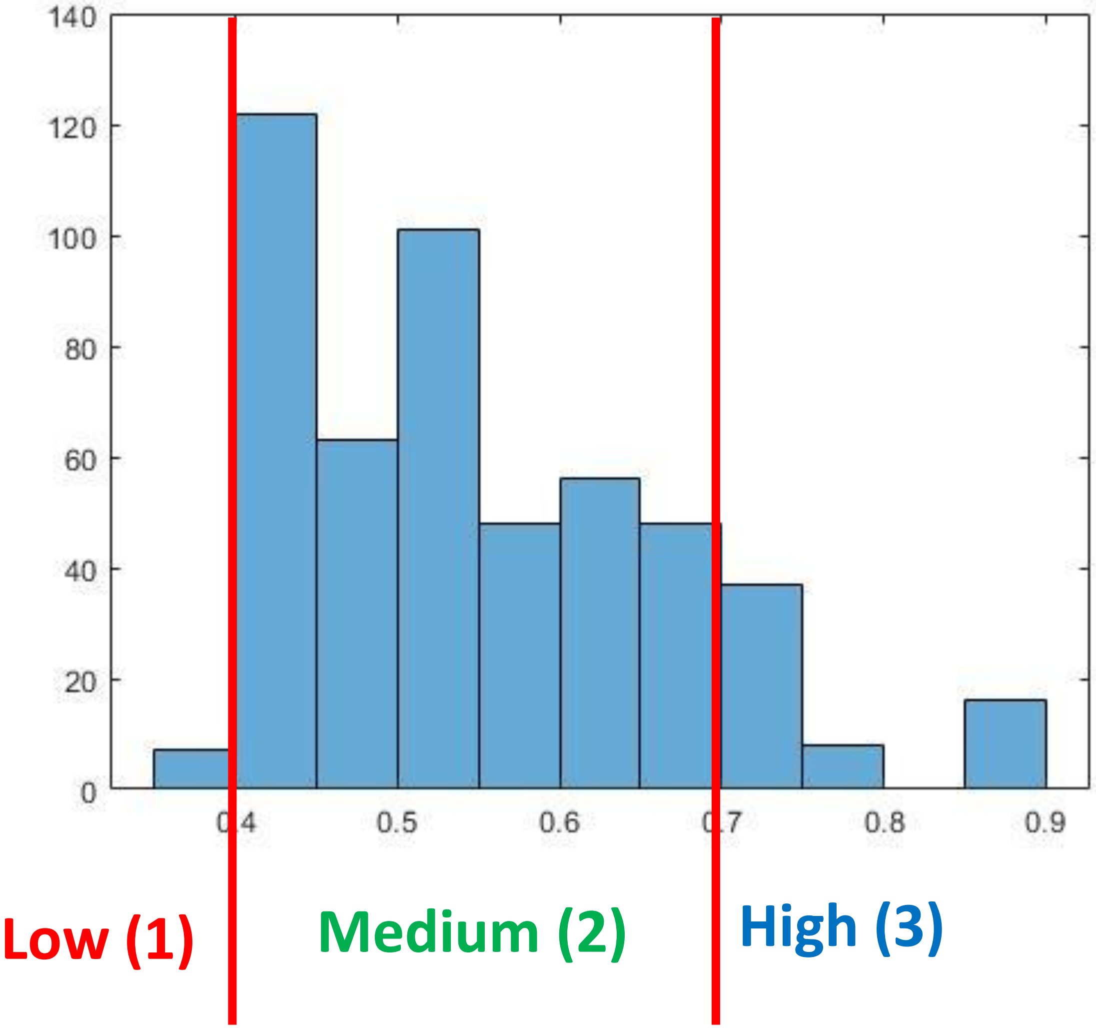
PCA plot – normalised data ('price' colour coded)

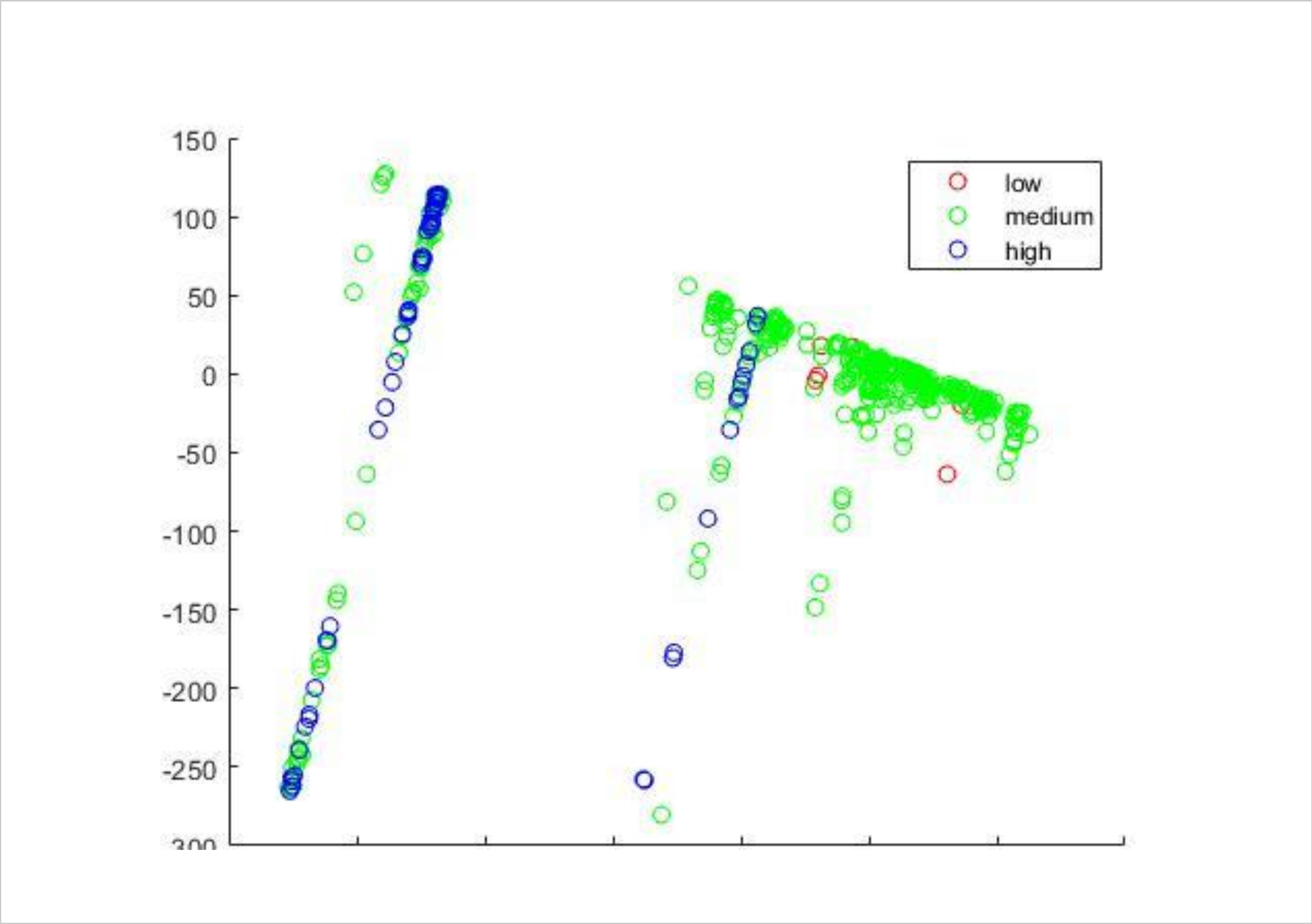
1^{st} Principal Component

$$e_1 = \begin{bmatrix} 0.25 \\ -0.26 \\ 0.35 \\ 0 \\ 0.34 \\ -0.19 \\ 0.31 \\ -0.32 \\ 0.32 \\ 0.34 \\ 0.2 \\ -0.2 \\ 0.31 \end{bmatrix}$$

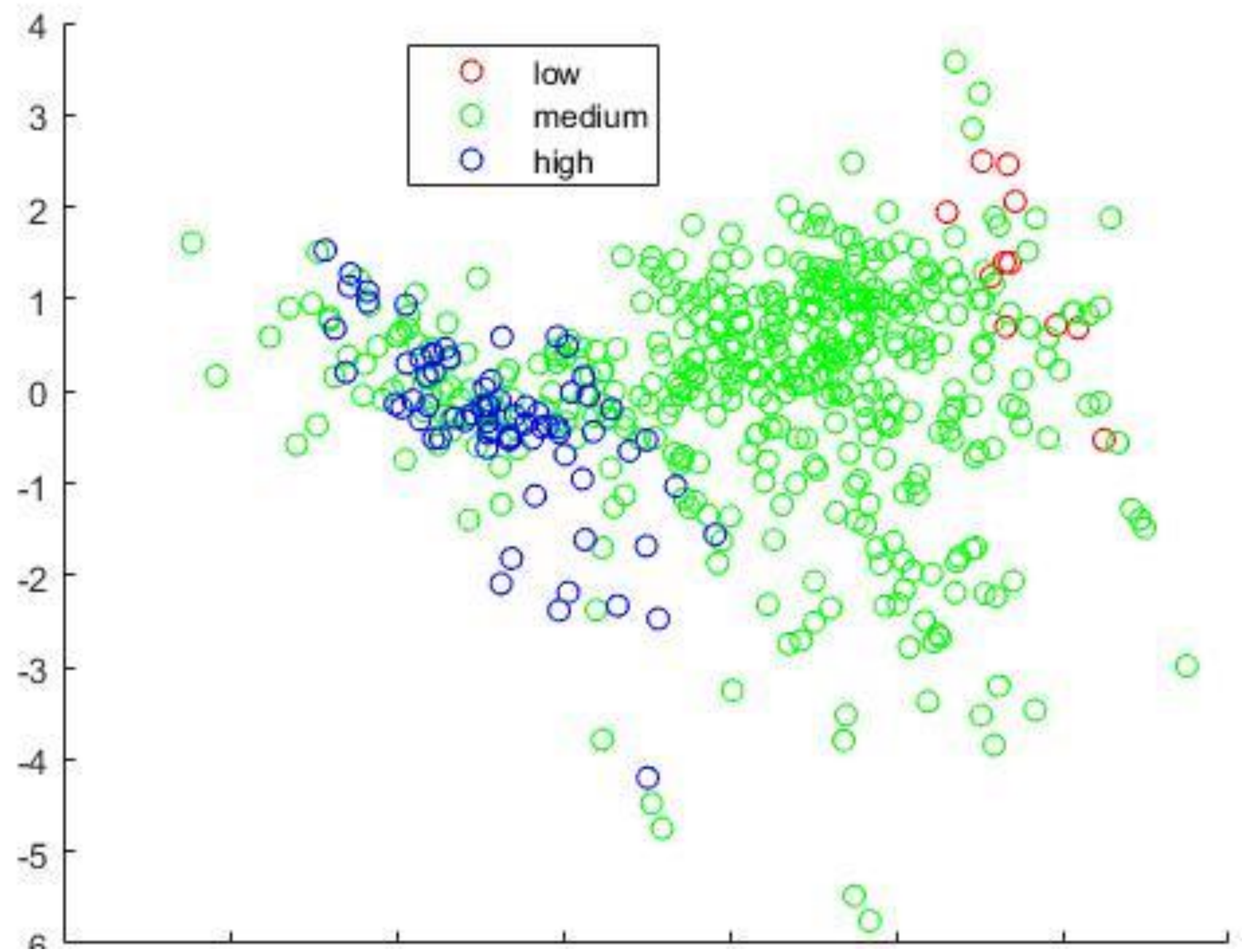
1. CRIM - per capita crime rate by town
2. ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS - proportion of non-retail business acres per town
4. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
5. NOX - nitric oxides concentration (parts per 10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

Now repeat for Task 2 – 'nox'





PCA plot – 'nox' – un-normalised



PCA Plot – ‘nox’ – variance normalised

Homework

- Work through the PCA analysis of the boston data with respect to the 'nox' data
- Try to obtain the same figures