Intelligent Data Analysis

Demonstration of PCA and SOM

Peter Tiňo
School of Computer Science
University of Birmingham

Boston Housing Dataset

Information collected by the U.S Census Service concerning housing in the area of Boston Mass.

Obtained from the StatLib archive
`http://lib.stat.cmu.edu/datasets/boston`.

506 cases.

Two prototasks associated with this data set:
For a given neighborhood, predict

1. nitrous oxide level
2. median value of a home

Boston Housing Dataset - attributes

There are 14 attributes in each case of the dataset:

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 USD 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 USD 1000's

How well-posed is the median house price prediction?

Gain more insight about this data set.

One may ask, for example, how well-posed is the task No. 2 of predicting the median value of a home based on the remaining 13 attributes (features) that vaguely characterise the neighborhood.

Prepare the data

From the original data set construct two data sets:
column No. 14 (house prices) only
the remaining columns No. 1-13.

View histogram of possible prices to see what the price distribution looks like.

It makes sense to discretize the house prices into:
"Low" - 1
"Medium" - 2
"High" - 3
"Very High" - 4

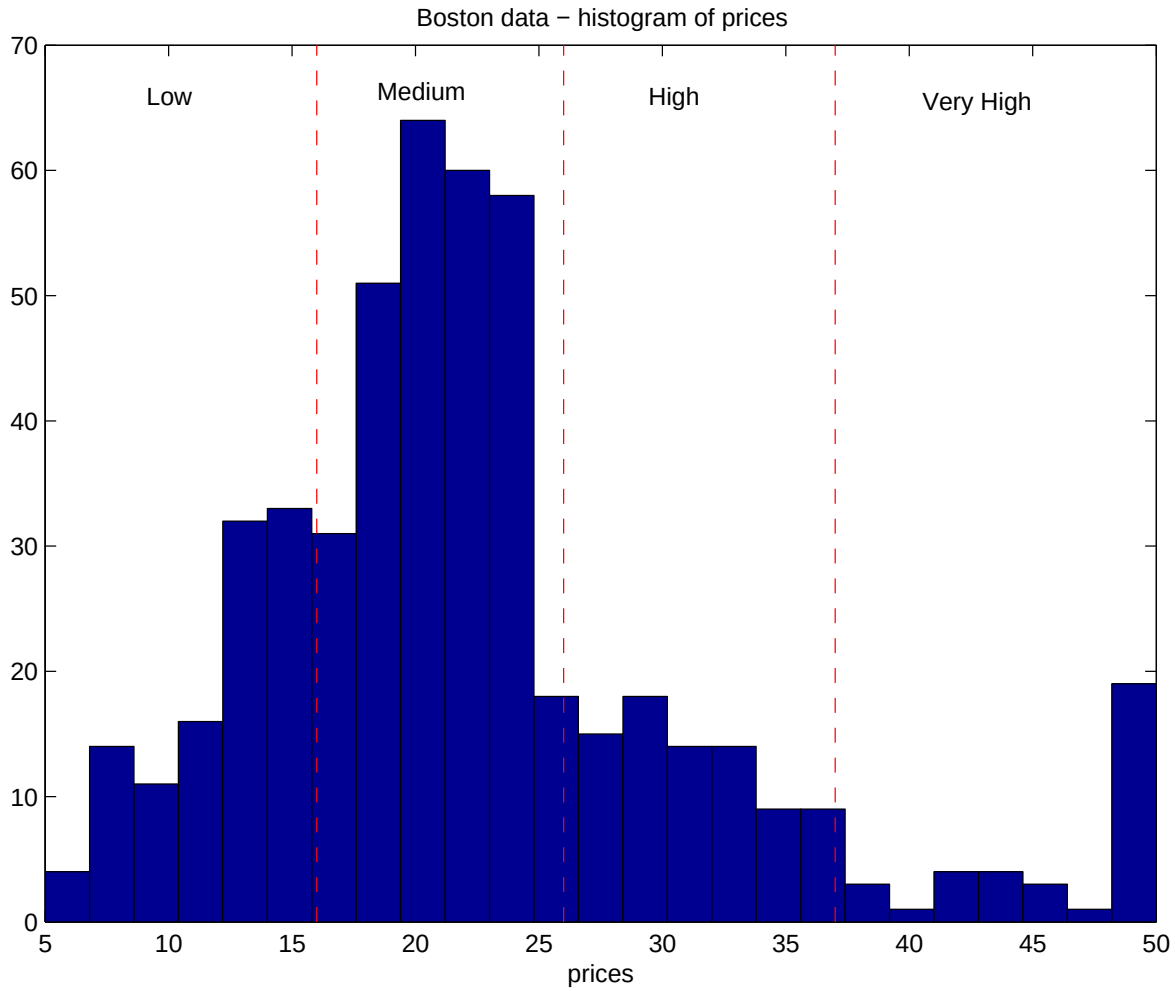
Most prices are in the Medium range, there are few extremely expensive houses.

Label the 13-dimensional data points (original data without the price attribute) based on where the corresponding house price falls.

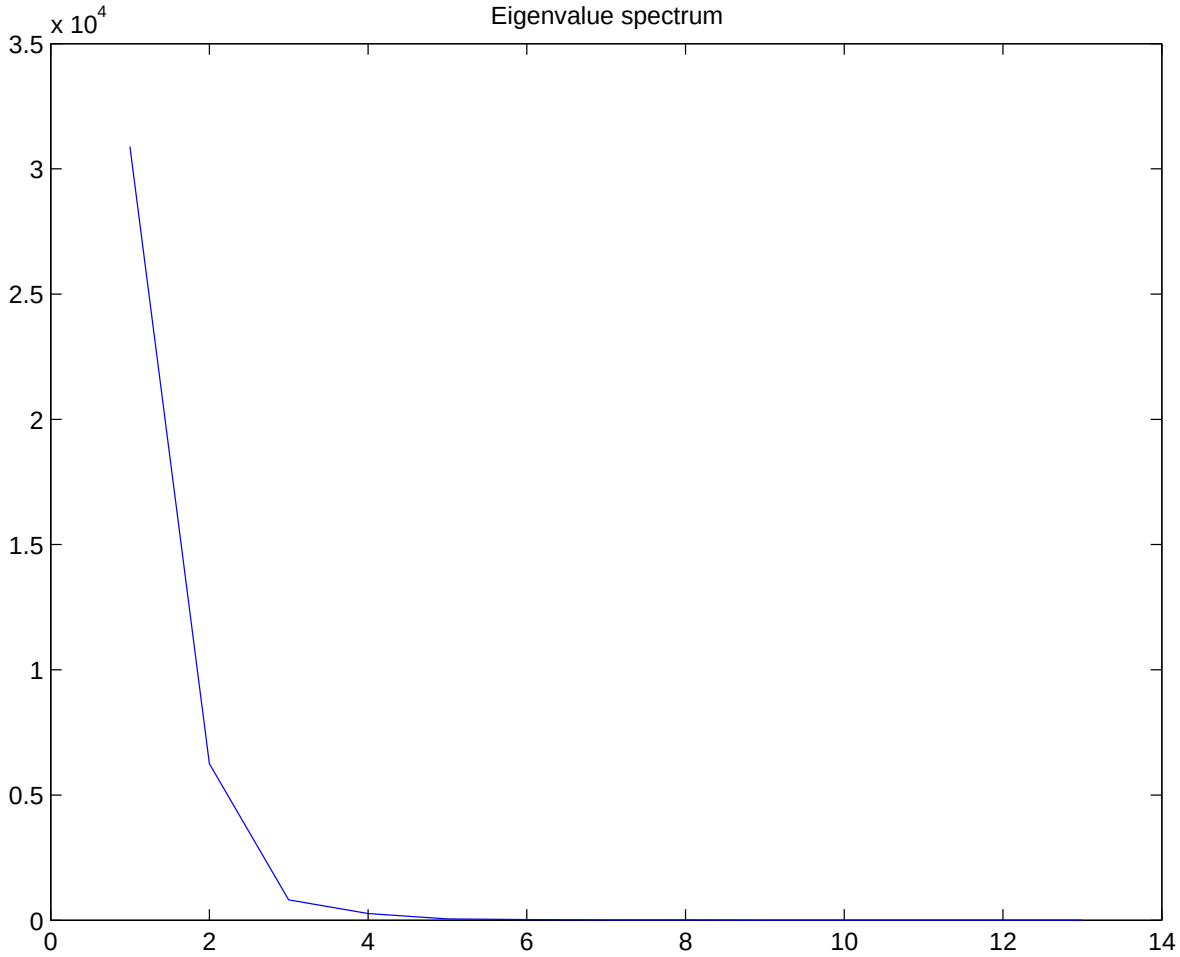
"Low" - black star
"Medium" - blue circle
"High" - green cross
"Very High" - red square

We will use the label information to set markers for data projections on the visualization plots.

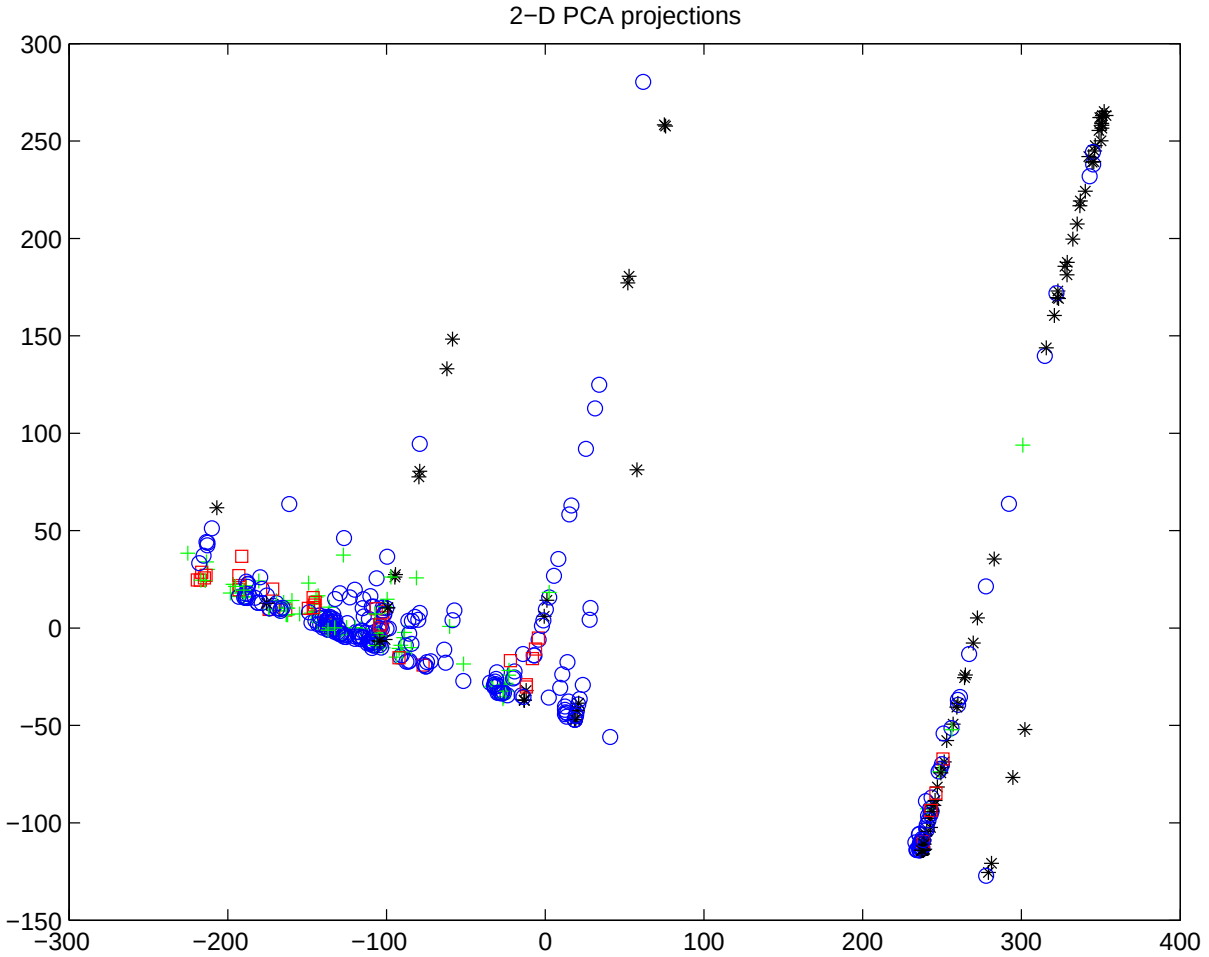
Histogram of house prices



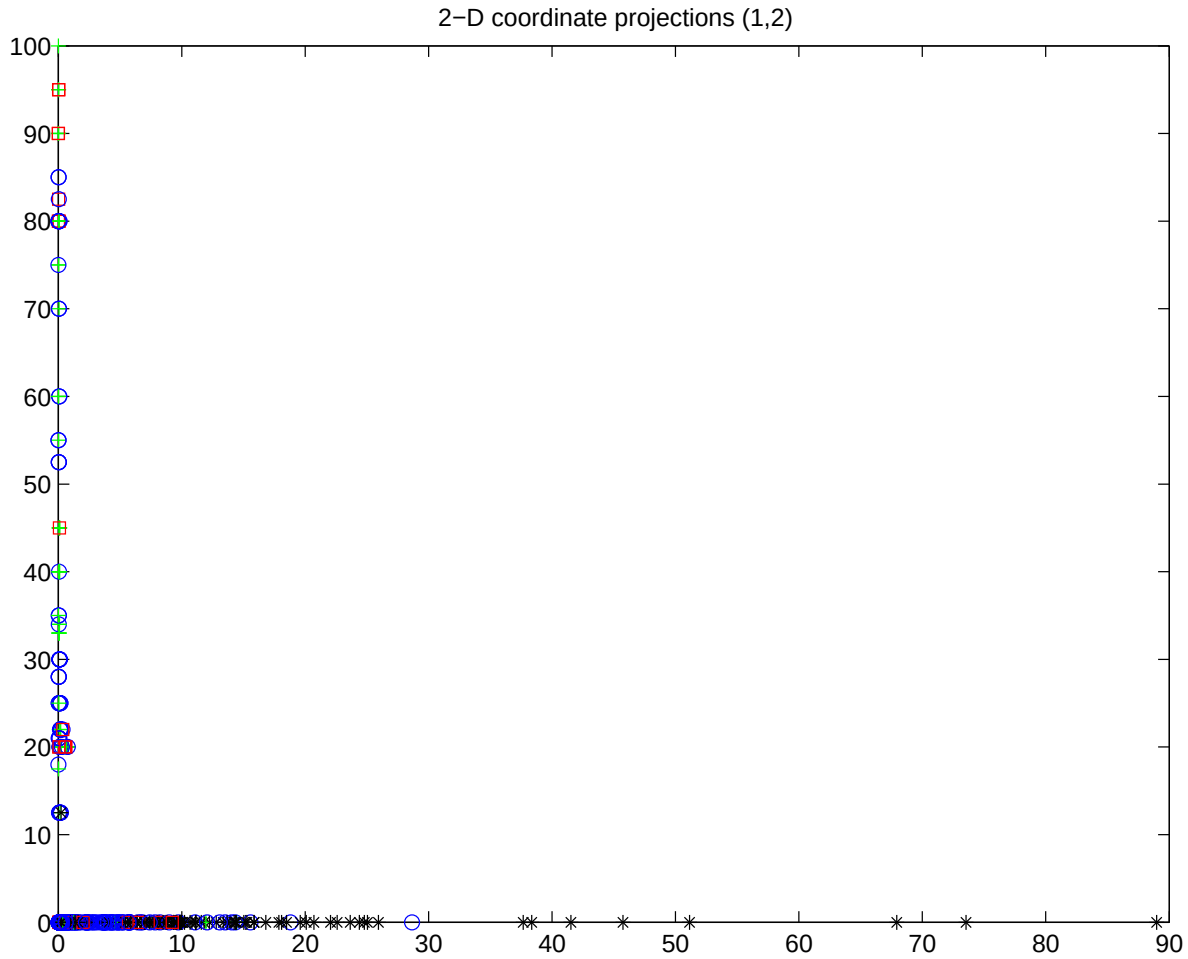
Eigenvalues of covariance matrix



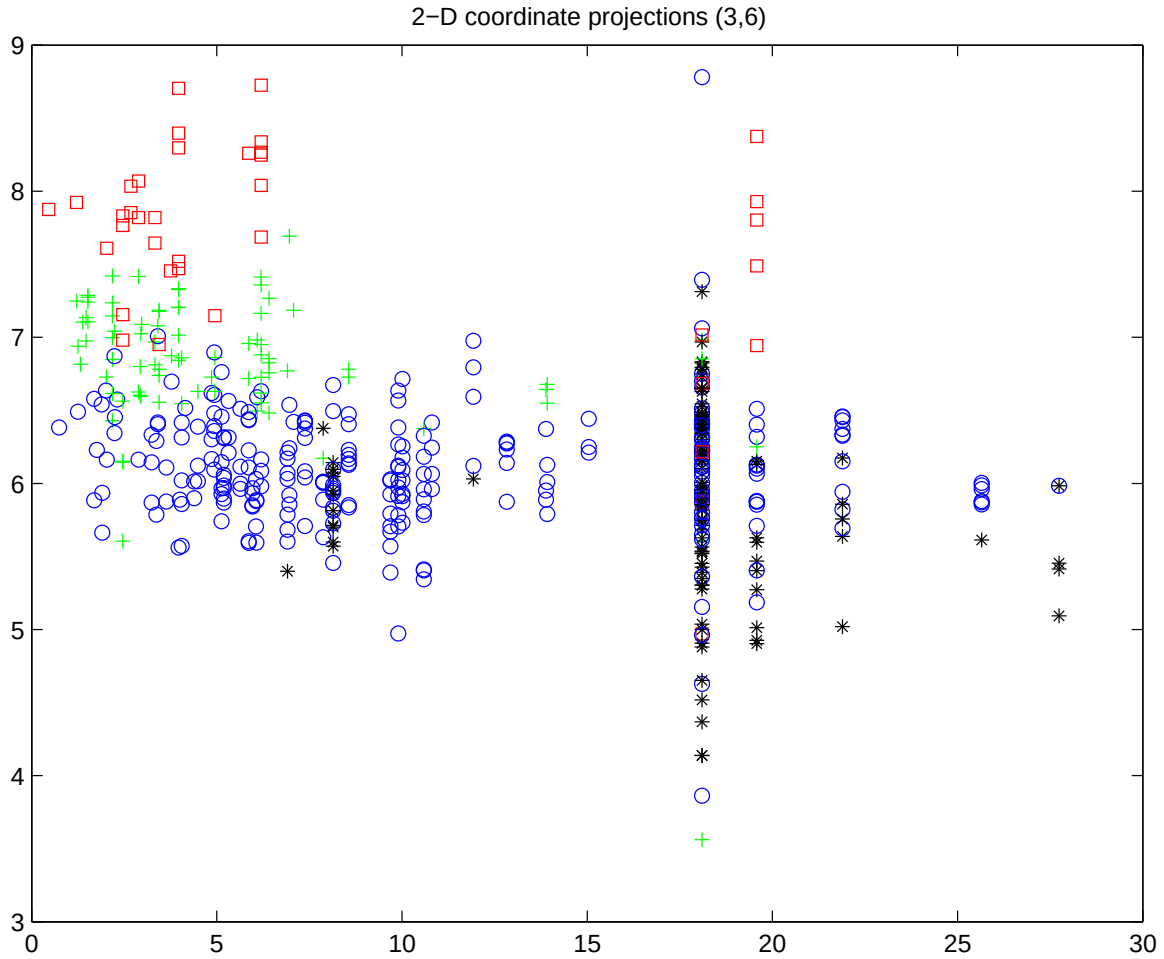
PCA projection



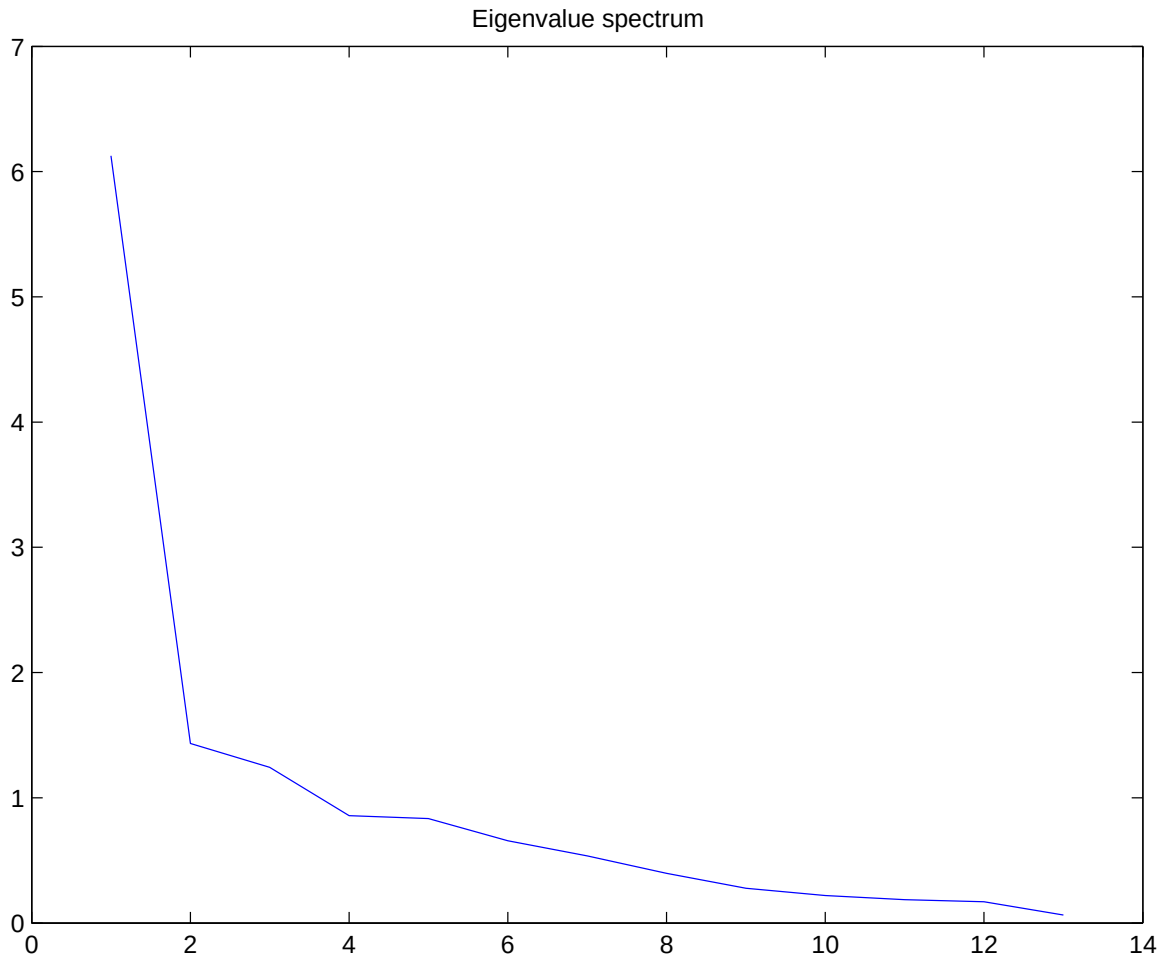
Coordinate projections



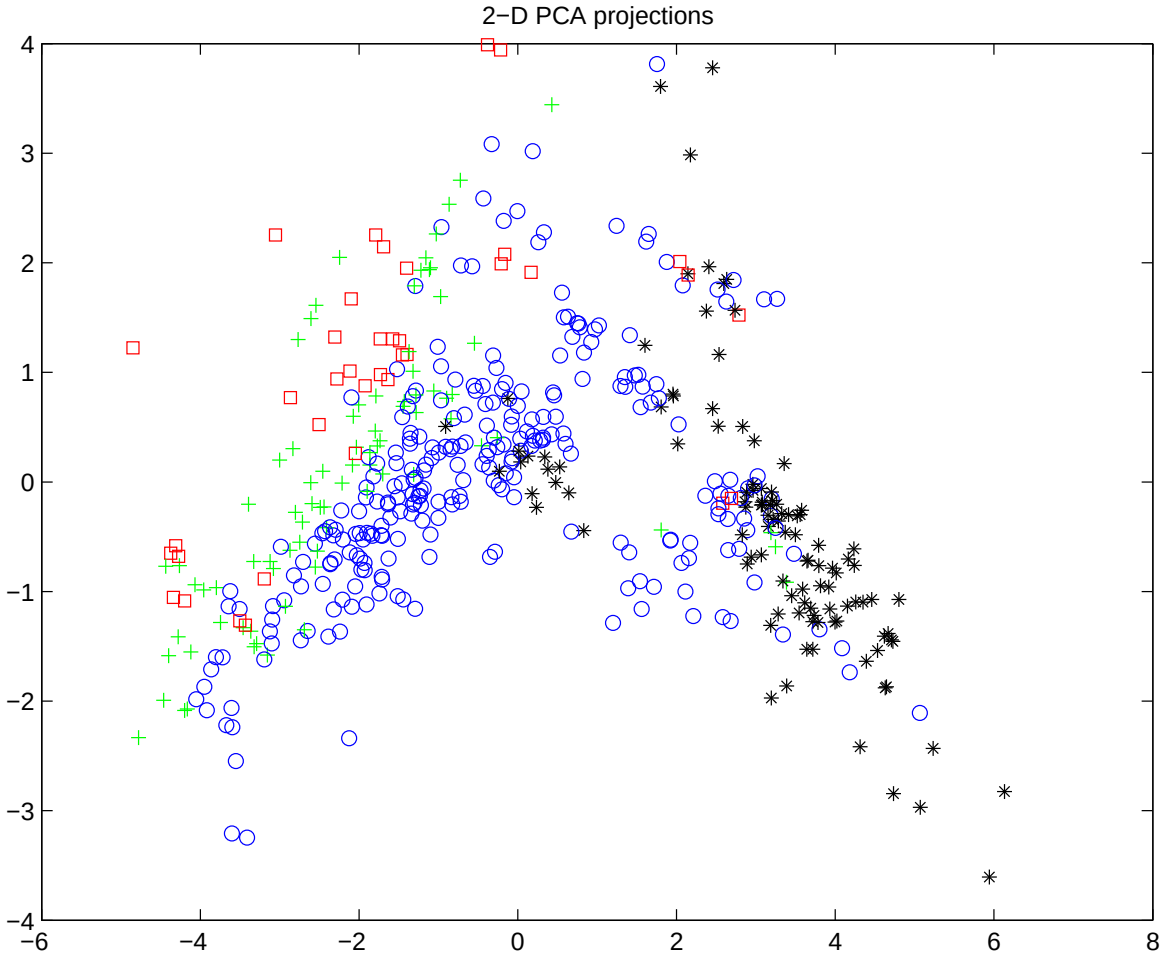
Coordinate projections - after considerable search



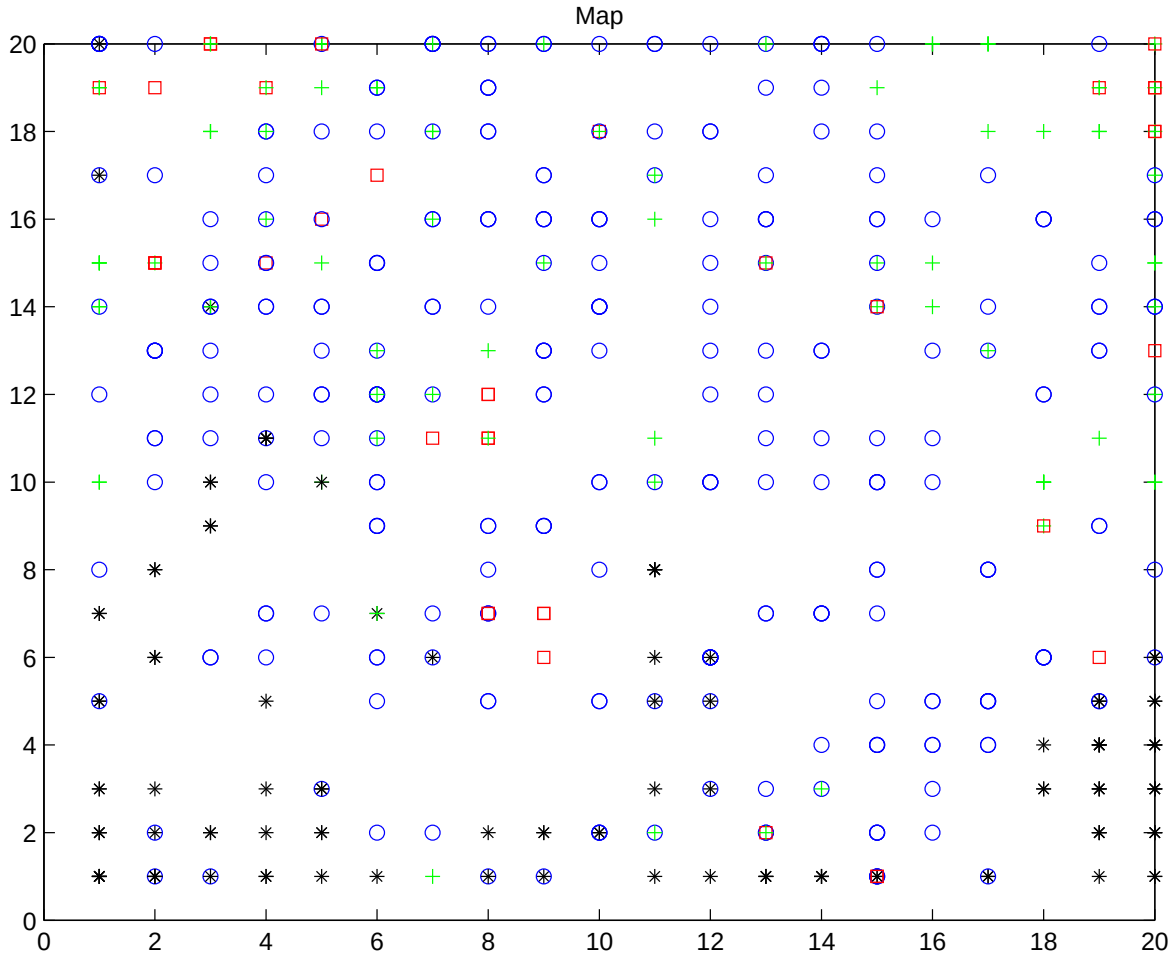
Normalize the data ($\mu = 0$ and $\sigma = 1$)!



PCA projection



SOM - original data



SOM - normalized data