

Project Two: Clustering with Gaussian Mixture Models (Due Friday, April 11, 2003)

The data set, called project 2 data, can be downloaded from the course website. There are 10 independent variables denoted by x_1, x_2, \dots, x_{10} , and 500 observations. It is suspected that the data are clustered. The goal of the project is to identify these clusters using finite Gaussian mixture models. You are required to use EM algorithm to compute the estimates. The following are the tasks you need accomplish in this project.

- a) Generalize the one-dimensional two-component Gaussian mixture model discussed in class to multi-component and multi-dimensional Gaussian mixture model. Work out the details of EM algorithm for estimating the MLEs of the general model.
- b) Explore the data. Is there any evidence that the data are clustered?
- c) Identify these clusters and derive the location, dispersion (variance-covariance matrix), and proportion of each cluster. Since the number of clusters is unknown, you may need try out different possible numbers. You can assume that the number is less than 10.
- d) How to determine the number of clusters?
- e) Use any other clustering methods you know such as hierarchical clustering, K-means and Self-Organized Mapping etc. to analyze the data, and compare the results with those from Gaussian mixture modeling.

The report of the project should be written in the format of a journal paper. It is due to my office on Friday, April 11.