SIMULATION METHODS FOR STOCHASTIC SYSTEMS

MIXTURE MODELS SIMULATIONS-PROJECT REPORT

Harish Settikere Prabhakara

CONTENTS

SUMMARY

Problem-1 [EM]

Steps:

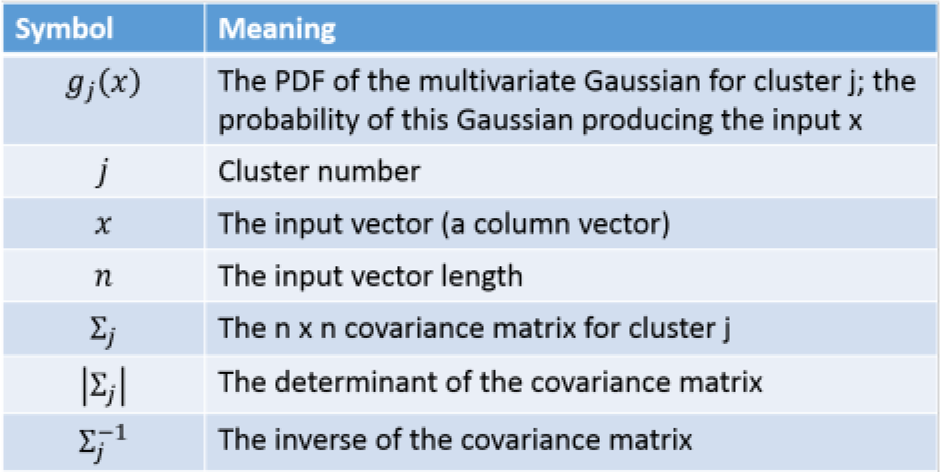
- Initially, a 2-dimensional Random number generator for a Gaussian mixture model (GMM) pdf with 2 sub-populations is implemented and displayed.
- Given the GMMs, we will find the clusters using a technique called "Expectation Maximization.
    - In the "Expectation" step, we will calculate the probability that each data point belongs to each cluster (using our current estimated mean vectors and covariance matrices).
    - In the "Maximization" step, we'll re-calculate the cluster means and co-variances based on the probabilities calculated in the expectation step.

- Detailed explanation about the steps for Expectation Maximization algorithm:

I. **Initialization**: We randomly select data points to use as the initial means, and set the covariance matrix for each cluster to be equal to the covariance of the full training set. Also, we give each cluster equal "prior probability". A cluster's "prior probability" is just the fraction of the dataset that belongs to each cluster. We'll start by assuming the dataset is equally divided between the clusters.

II. **Expectation:** In the “Expectation” step, we calculate the probability that each data point belongs to each cluster.

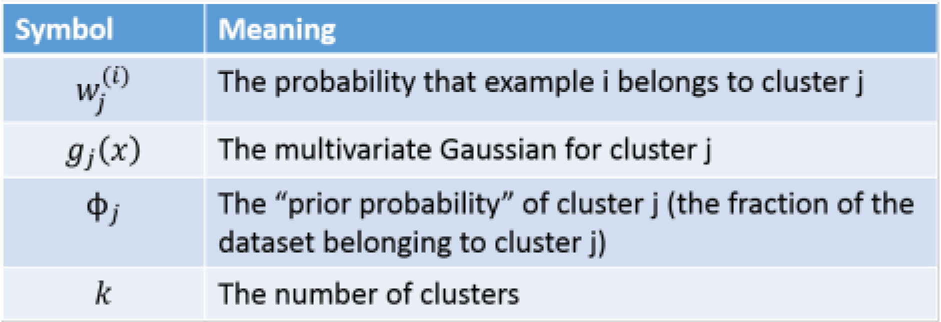
$$g_j(x) = \frac{1}{\sqrt{(2\pi)^n|\Sigma_j|}} e^{-\frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j)}$$



Symbol	Meaning
$g_j(x)$	The PDF of the multivariate Gaussian for cluster j; the probability of this Gaussian producing the input x
j	Cluster number
x	The input vector (a column vector)
n	The input vector length
Σ_j	The n x n covariance matrix for cluster j
$\|\Sigma_j\|$	The determinant of the covariance matrix
Σ_j^{-1}	The inverse of the covariance matrix

The probability that example point i belongs to cluster j can be calculated using the following:

$$w_j^{(i)} = \frac{g_j(x)\phi_j}{\sum_{l=1}^{k} g_l(x)\phi_l}$$



Symbol	Meaning
$w_j^{(i)}$	The probability that example i belongs to cluster j
$g_j(x)$	The multivariate Gaussian for cluster j
ϕ_j	The “prior probability” of cluster j (the fraction of the dataset belonging to cluster j)
k	The number of clusters

We’ll apply this equation to every example and every cluster, giving us a matrix with one row per example and one column per cluster.

III. **Maximization:** To find the average value of a set of *m*values, where you have a weight *w* defined for each of the values, you can use the following equation:

$$\bar{y} = \frac{\sum_{i=1}^{m}(w_i y_i)}{\sum_{i=1}^{m} w_i}$$

The update rules for the maximization step are given below:

$$\phi_j := \frac{1}{m}\sum_{i=1}^{m} w_j^{(i)},$$

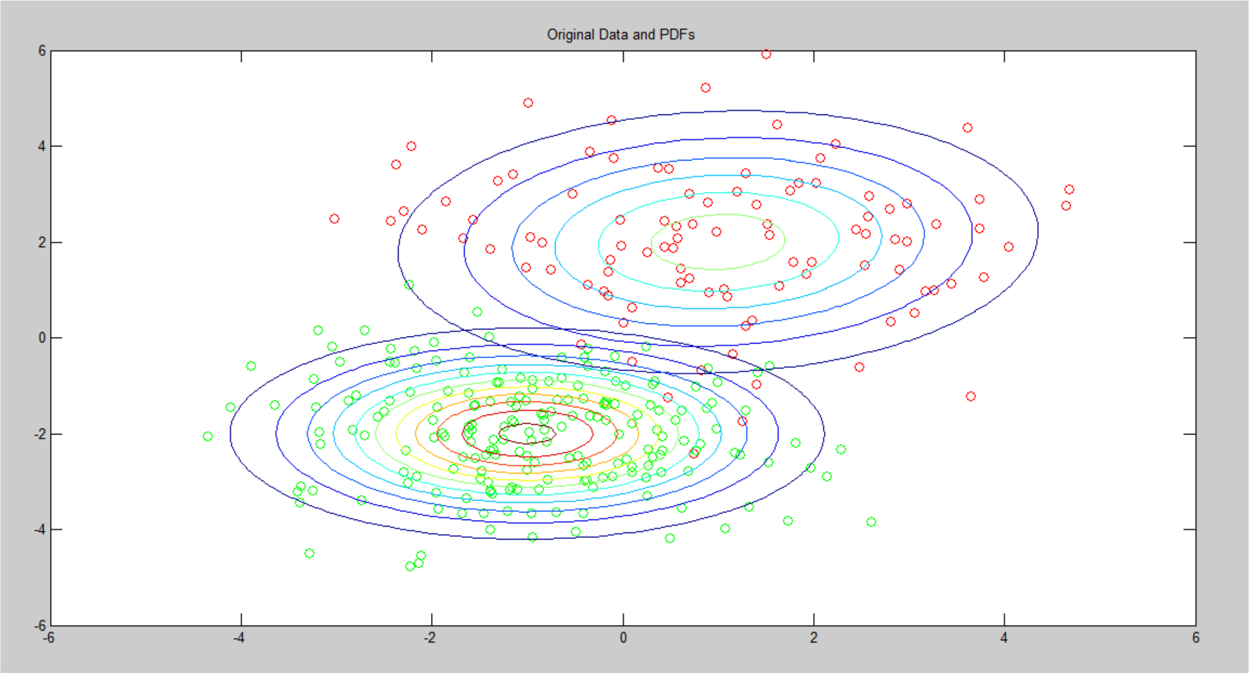
$$\mu_j := \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}},$$

$$\Sigma_j := \frac{\sum_{i=1}^{m} w_j^{(i)} (x^{(i)} - \mu_j)(x^{(i)} - \mu_j)^T}{\sum_{i=1}^{m} w_j^{(i)}}$$

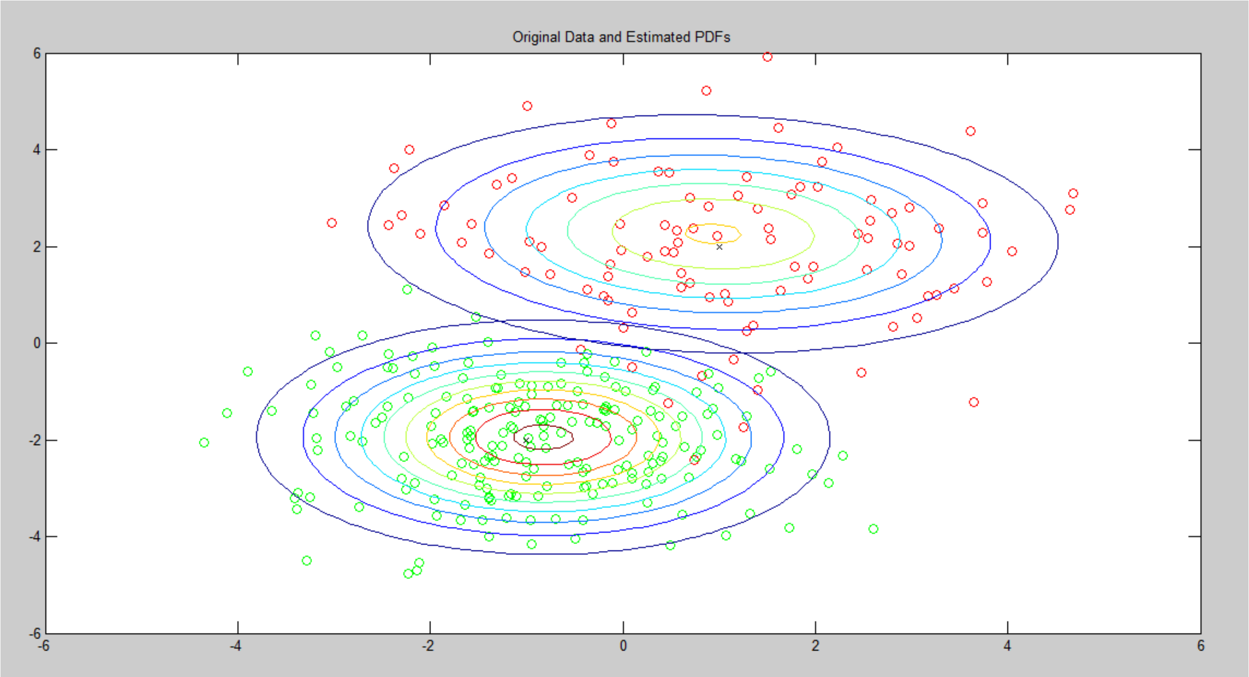
The equation for mean (mu) of cluster j is just the average of all data points in the training set, with each example weighted by its probability of belonging to cluster j. Similarly, the equation for the covariance matrix is the same as the equation you would use to estimate the covariance of a dataset, except that the contribution of each example is again weighted by the probability that it belongs to cluster j. The prior probability of cluster j, denoted as phi, is calculated as the average probability that a data point belongs to cluster j.

Graphs

Original Data and PDFs



Original Data and Estimated PDFs



Problem-2 [Testing Faith]

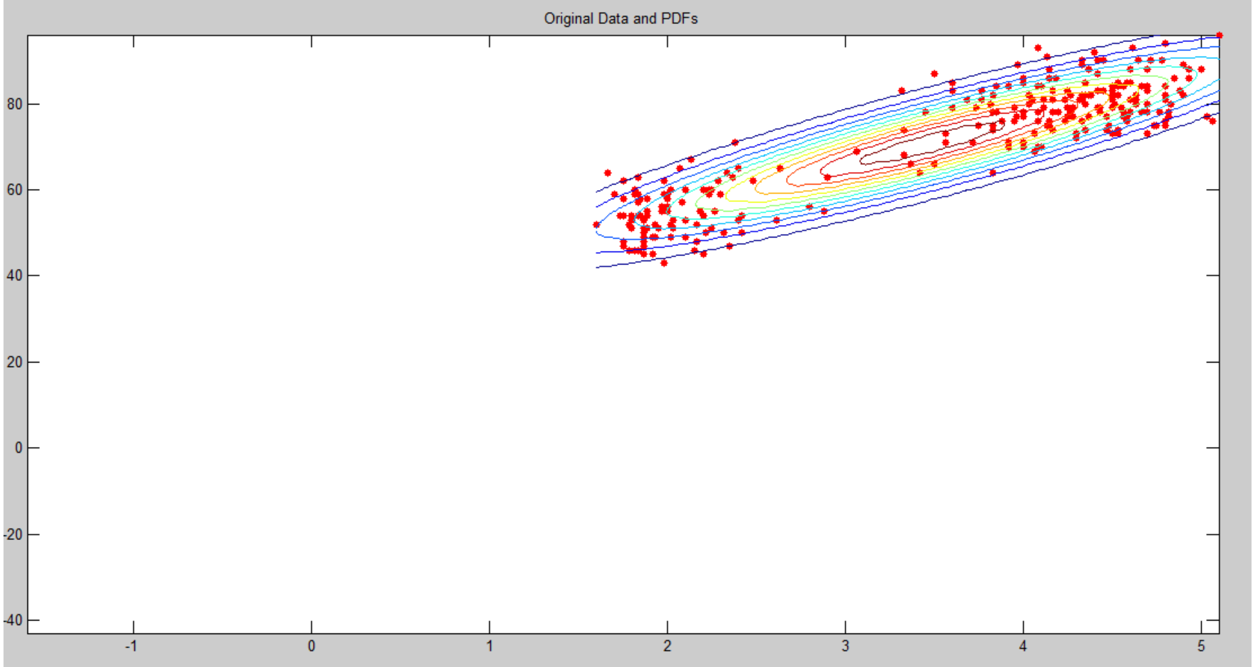
- The “old faithful” data set which has samples of a 2-D random variable, the first dimension being the duration of the geyser eruption, and the second being waiting time for the next eruption is downloaded from blackboard and saved as “**old.txt**”. The required data from the file is fetched appropriately.
- GMM-EM algorithm is applied to fit the data to a GMM pdf.
- A contour plot of the final GMM pdf is plotted. Also, the contour plot is overlaid with a scatterplot of the data set.

Results:

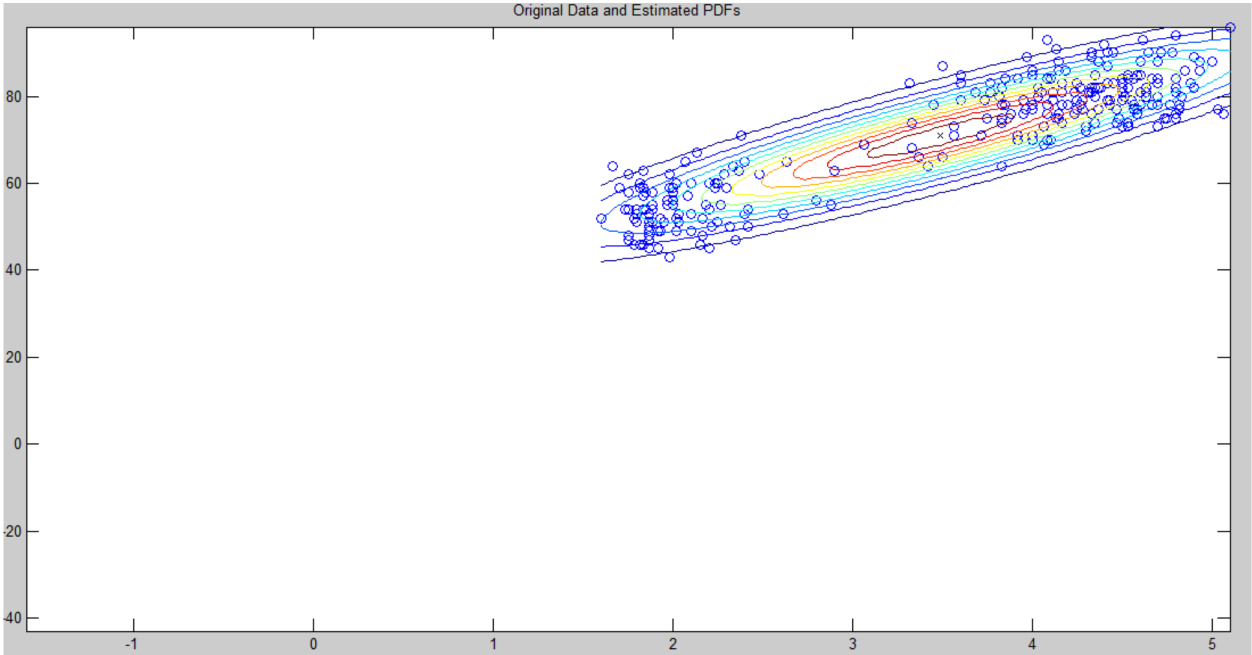
We can observe that 2 EM iterations are needed for convergence.

Graphs

Original Data and PDFs



Original Data and Estimated PDFs


Original Data and Estimated PDFs