

## Cost Function to be Minimized

$$X_{1,1}, X_{1,2}, \dots, X_{1,n_1} \quad \mu_1$$

$$X_{2,1}, X_{2,2}, \dots, X_{2,n_2} \quad \mu_2$$

$$X_{k,1}, X_{k,2}, \dots, X_{k,n_k} \quad \mu_k$$

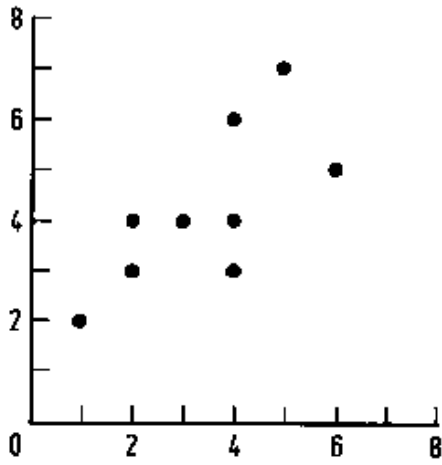
$\mu = \text{global mean vector}$

$$\begin{aligned} \text{Total Sum of Squares} &= \sum_{i=1}^k \sum_{j=1}^{n_i} \|X_{i,j} - \mu\|^2 \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} \|X_{i,j} - \mu_i\|^2 + \sum_{i=1}^k n_i \|\mu_i - \mu\|^2 \end{aligned}$$

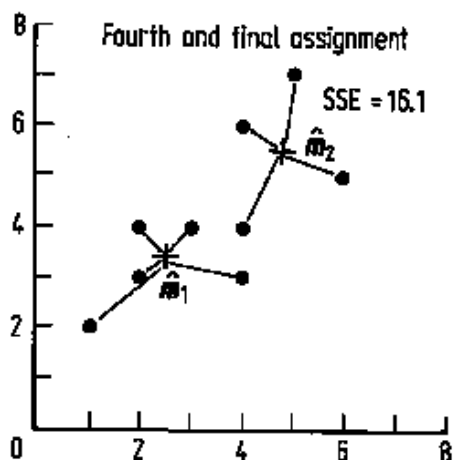
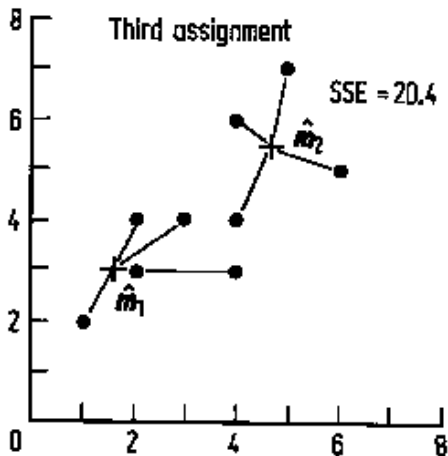
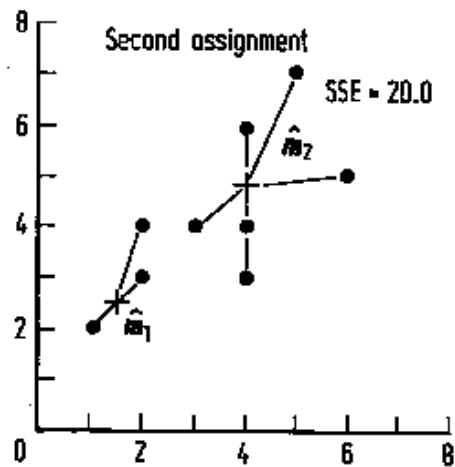
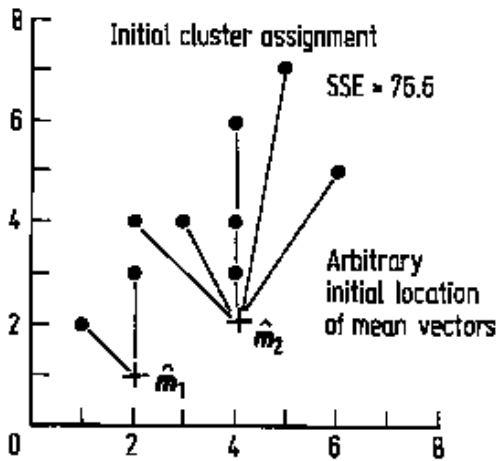
= Within Sum of Squares + Between Sum of Squares

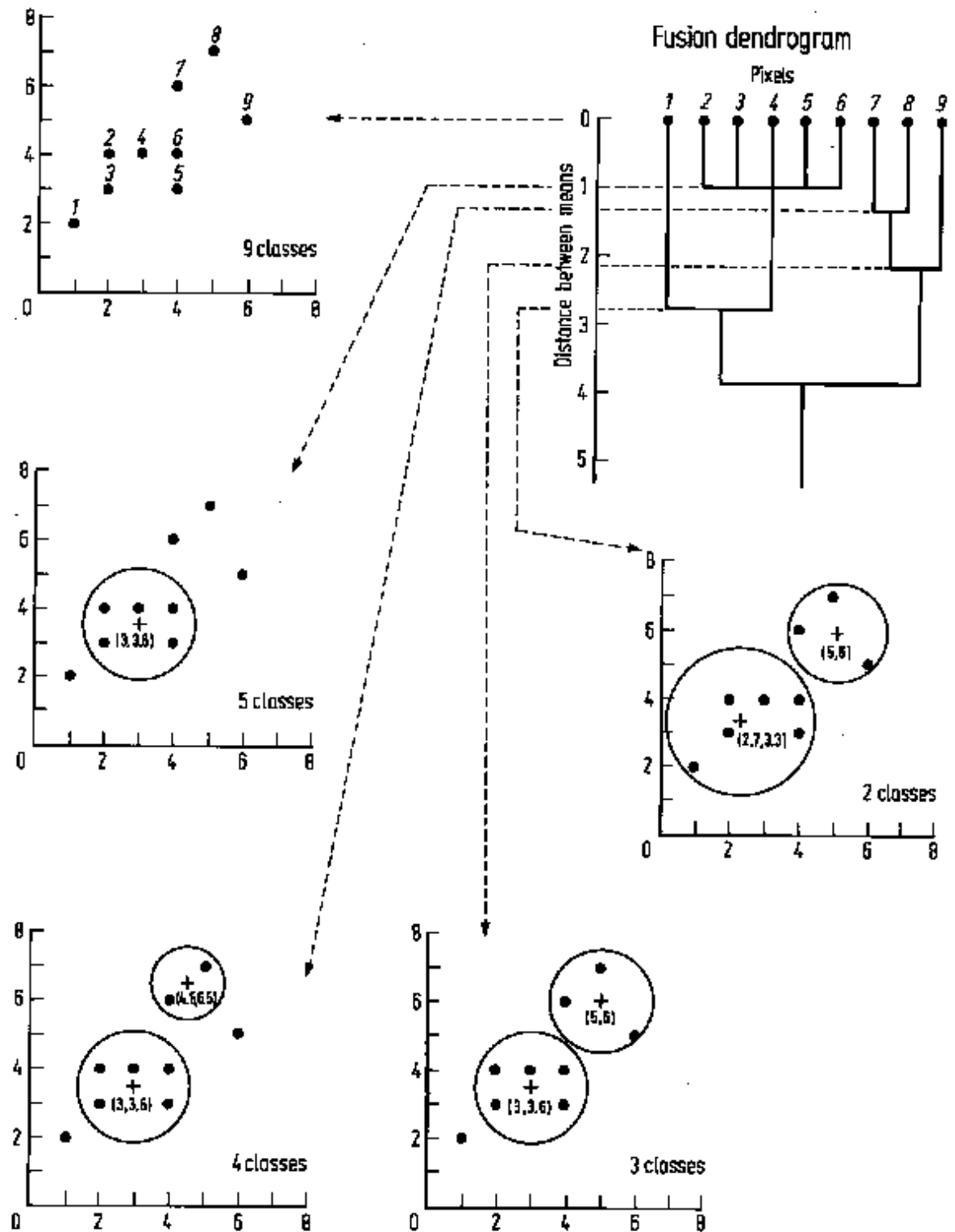
A clustering method aims at partitioning the input set so that Within Sum of Squares is minimized (or equivalently, Between Sum of Squares is maximized).

An illustration of clustering by K-means algorithm which leads to a progressive reduction in SSE (Within Sum of Squares).



Two clusters (spectral classes) to be found





An illustration of agglomerative hierarchical clustering using Euclidean distance.

## Expectation Maximization (EM) Algorithm Mixture Models

$$g(X) = P_1 f_1(X) + P_2 f_2(X) + \dots + P_k f_k(X)$$

is a mixture of  $k$  individual probability density functions where  $0 < P_i < 1$  and  $\sum P_i = 1$ . An element in the input set is assumed to be generated from the mixture probability density function  $g(X)$ .

The task of clustering can be achieved by estimating the unknown parameters of  $g(X)$  which are  $P_i$  and the parameters of  $g(X)$ .

For example, if  $f_i$  is a multivariate normal distribution, its unknown parameters are the mean vector and covariance matrix, ie,  $\mu_i$  and  $\Sigma_i$ .

For the input set

$$X_1, X_2, \dots, X_n$$

the **likelihood function** is defined as  $g(X_1) * g(X_2) * \dots * g(X_n)$ .

The maximum likelihood estimates of  $P_i, \mu_i, \Sigma_i$  are those that maximize

$$h(P_i, \mu_i, \Sigma_i) = \sum_i \log\{g(X_i)\} - \lambda(\sum P_i - 1)$$

**Expectation (E-step):** Compute the posterior probability estimates  $q_{ji}$  the probability that

the  $j$ -th sample belongs to the  $i$ -th component as

$$q_{ji} = \frac{P_i f_i(X)}{P_1 f_1(X) + P_2 f_2(X) + \dots + P_k f_k(X)}$$

**Maximization (M-step):** Re-estimate the model parameters,  $(P_i, \mu_i, \Sigma_i)$  from the probabilistically re-labeled data.

$$P_i = (\sum_j q_{ji})/n$$

$$\mu_i = (\sum_j q_{ji} X_j) / (\sum_j q_{ji})$$

$$\delta_{rs} = (\sum_j q_{ji} (X_{jr} - \mu_{ir})(X_{jr} - \mu_{ir})) / (\sum_j q_{ji})$$

# Classification

## Bayes Rule

**Box 1:** 9 red balls + 1 blue ball

**Box 2 :** 1 red ball + 9 blue balls

Prob(Box 1| red ball) = ?