Cluster Analysis:

In Classification, we *knew* that there were J classes and our goal was to classify new obserations into one of the classes. This is done in practice using a learning sample, $L = (\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ on n cases where $\mathbf{x}_i \in \chi$, $y_i \in \{1, ..., J\}$, for $i = 1..., n$. Note that for each observation $\mathbf{x}_i$ we know the value of y_i, i.e., into which class the ith observation falls.

Often, there is no training sample from which to form the classification rule $\delta : \chi \to C$. Still, we want to classify observations. This is Cluster analysis. Two components are unknown:

A) The number of classes (or clusters), J.

B) Parameters in each class.

E.G.

Have n observations from $N_p(\mu_1, \mathbf{\Sigma}_1), ..., N_p(\mu_J, \mathbf{\Sigma}_J)$.

The goal is to determine J and $(\mu_1, \mathbf{\Sigma}_1, ..., \mu_J, \mathbf{\Sigma}_J)$

Two different philosophies:

1) Nonparametric model, a rough description of observations.

2) Parametric Model, formal clustering and parameter estimation.

Probabilistic Formulation:

Assume we have n independent $p-$vectors, $\mathbf{x}_i, ..., \mathbf{x}_n$. Each observation comes from one of the J populations with p.d.f. $f(\mathbf{x}, \theta_j)$, $j = 1, ..., J$. For now, we assume that J is known. Note that this is still not the usual discrimination problem as we don't know from which $f(\mathbf{x}, \theta_j)$ each $\mathbf{x}_i$ comes.

Let $\gamma = (\gamma_1, ..., \gamma_n)$ be labels so that $\gamma_i = j \iff \mathbf{x}_i$ has distribution $f(\mathbf{x}, \theta_j)$ for $i = 1, ..., n$ and $j = 1, ..., J$.

Let C_j be the set of all $\mathbf{x}_i$ assigned to the jth class by γ. Then the likelihood function is given by:

$$L(\gamma, \theta_1, ..., \theta_J) = \prod_{\mathbf{x} \in C_1} f(\mathbf{x}, \theta_1) ... \prod_{\mathbf{x} \in C_J} f(\mathbf{x}, \theta_J).$$

Let $\hat{\gamma}$, $\hat{\theta}_1$,...,$\hat{\theta}_J$ denote the parameter values that maximize $L(\gamma, \theta_1, ..., \theta_J)$ Under, $\hat{\gamma}$ we have the induced partition: $\hat{C}_1, ..., \hat{C}_J$.

Notice that moving one obervation from class j to class l, i.e., from $\hat{C}_j$ to $\hat{C}_l$ changes the Likelihood function

to: $L(\hat{\gamma}, \hat{\theta}_1, ..., \hat{\theta}_J)=$

$$\prod_{\mathbf{x}\in\hat{C}_1} f(\mathbf{x}, \hat{\theta}_1)... \prod_{\mathbf{x}\in\hat{C}_J} f(\mathbf{x}, \hat{\theta}_J)[f(\mathbf{x}, \hat{\theta}_l)/f(\mathbf{x}, \hat{\theta}_j)],$$

and that this is no larger than $L(\hat{\gamma}, \hat{\theta}_1, ..., \hat{\theta}_J) =$

$$\prod_{\mathbf{x}\in C_1} f(\mathbf{x}, \hat{\theta}_1)... \prod_{\mathbf{x}\in C_J} f(\mathbf{x}, \hat{\theta}_J).$$

So we have:

$$f(\mathbf{x}, \hat{\theta}_l) \leq f(\mathbf{x}, \hat{\theta}_j), \text{ for } \mathbf{x} \in \hat{C}_j,\ l \neq j.$$

Recall, this is the *classification rule* under uniform prior class probabilities: $\pi(j) = 1/J,\ j = 1, ..., J.$

E.G.

Let $f(\mathbf{x}, \theta_j)$ denote the $N_p(\mu_j, \mathbf{\Sigma}_j)$ p.d.f., $j = 1, ..., J.$

Then $L(\gamma, \theta)(2\pi)^{np/2} =$

$$\prod_{j=1}^{J} \prod_{\mathbf{x}_i\in C_j} |\mathbf{\Sigma}_j|^{-1/2} exp[(-1/2)(\mathbf{x}_i - \mu_j)'\mathbf{\Sigma}_j^{-1}(\mathbf{x}_i - \mu_j)].$$

with associated log likelihood $l(\gamma, \theta) + (np/2)log(2\pi) =$

$$-(1/2)\sum_{j=1}^{J} n_j log(|\mathbf{\Sigma}_j|) - (1/2)\sum_{j=1}^{J}\sum_{\mathbf{x}_i\in C_j} (\mathbf{x}_i - \mu_j)'\mathbf{\Sigma}_j^{-1}(\mathbf{x}_i - \mu_j).$$