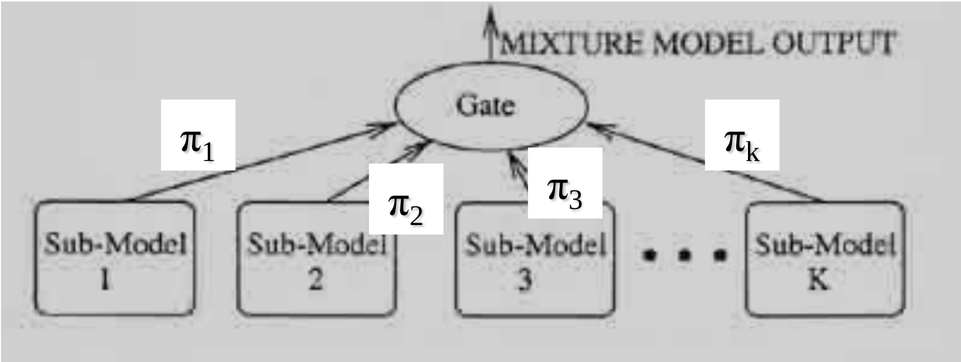
Adopted sildes from:
Prof. Bebis

Ch3-part5
EM-Mixture Model

- In a mixture model, there are many "sub-models", each of which has its own probability distribution which describes how it generates data when it is active.

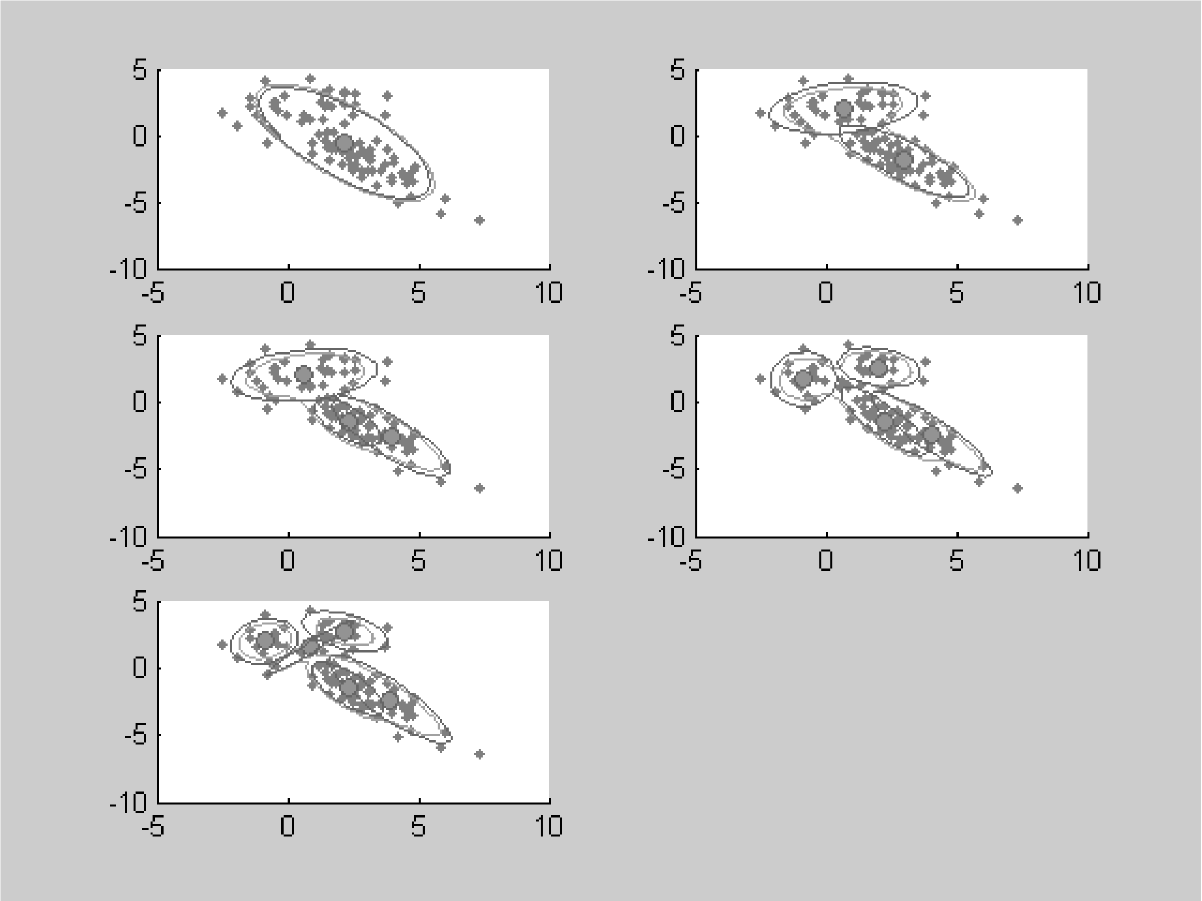
- There is also a "mixer" or "gate" which controls how often each sub-model is active.



- Formally, a mixture is defined as a weighted sum of K components where each component is a parametric density function $p(x/\theta_k)$:

$$p(x/\theta) = \sum_{k=1}^{K} p(x/\theta_k)\pi_k$$

Mixture of Gaussian Data - Example



Mixture Parameters

- The parameters θ to estimate are:

  * the values of π_k
  * the parameters θ_k of $p(x/\theta_k)$

- The component densities $p(x/\theta_k)$ may be of different parametric forms and are specified using knowledge of the data generation process, if available.

- The weights π_k are the *mixing parameters* and they sum to unity:

$$\sum_{k=1}^{K} \pi_k = 1$$

Fitting a Mixture Model to a set of Observations D_x

- Estimate the mixture parameters that best describe the data D_x (i.e., ML problem).

- Two fundamental issues
  - Estimate mixture parameters
  - Estimate number of mixture components

Mixtures of Gaussians

$$p(x/\theta) = \sum_{k=1}^{K} p(x/\theta_k)\pi_k$$

- Each $p(x/\theta_k)$ is a multivariate Gaussian.
- The parameters θ_k are (μ_k, Σ_k)

Mixtures of Gaussians (cont'd)

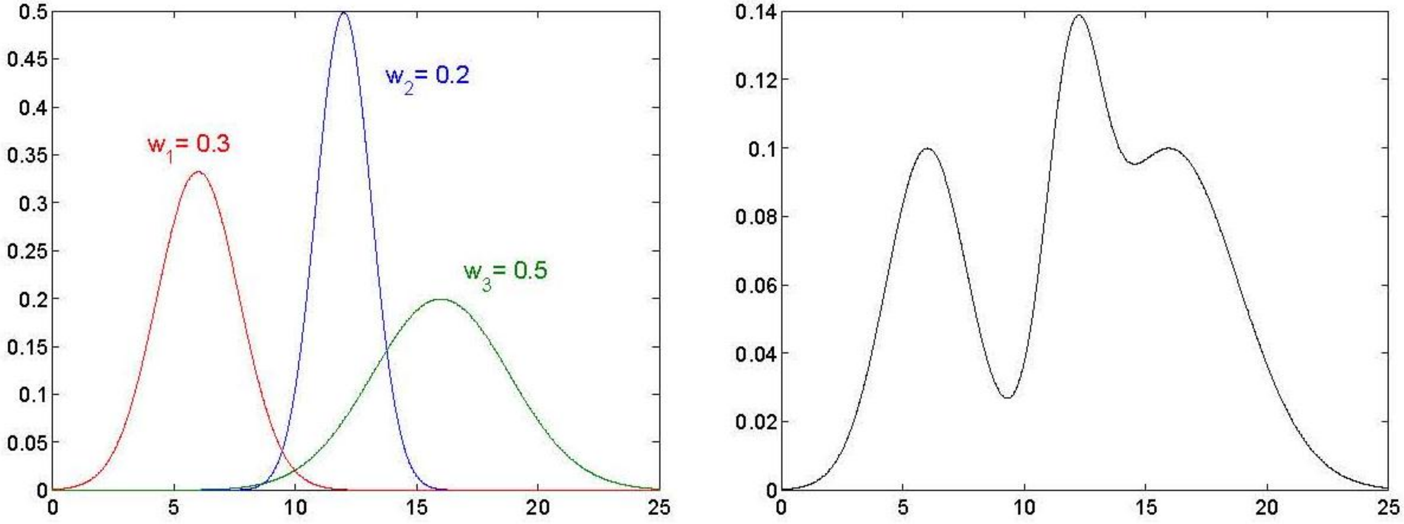


Figure 1: One dimensional Gaussian mixture pdf, consisting of 3 single Gaussians

Data Generation Process Using Mixtures of Gaussians

- Each instance is generated using a two-step process:

  (1) One of the K Gaussians is selected at random, with probabilities $\pi_1, \pi_2, \ldots, \pi_K$.

  (2) A single random instance x_i is generated according to this selected distribution.

- This process is repeated to generate a set of data points D.

Estimating Mixture Parameters Using ML – difficult!

- As we have seen, given a set of data $D=(x_1, x_2, ..., x_n)$, ML seeks the value of θ that maximizes the following probability:

$$p(D/\theta) = \prod_{i=1}^{n} p(x_i/\theta)$$

- Since $p(x_i/\theta)$ is modeled as a mixture (i.e., $p(x_i/\theta) = \sum_{k=1}^{K} p(x_i/\theta_k)\pi_k$) the above expression can be written as:

$$p(D/\theta) = \prod_{i=1}^{n} \sum_{k=1}^{K} p(x_i/\theta_k)\pi_k$$

- In general, it is not possible to solve $\frac{\partial p(D/\theta)}{\partial \theta} = 0$ explicitly for the parameters and iterative schemes must be employed.

Estimating Mixture Parameters Using EM: Case of Means

- Assumptions

(1) $\pi_1 = \pi_2 = \cdots = \pi_K$ (uniform distribution)

(2) Each Gaussian has the same variance σ^2 which is known.

- The problem is to estimate the means of the Gaussians $\theta = (\mu_1, \mu_2, \ldots, \mu_K)$

Note: if we knew which Gaussian generated each datapoint, then it would be easy to find the parameters for each Gaussian using ML.

Estimating Mixture Parameters Using EM: Case of Means (cont'd)

- Introducing <u>hidden</u> or <u>unobserved</u> variables

- We can think of the full description of each instance x_i as

$$y_i = (x_i, z_i) = (x_i, z_{i1}, z_{i2}, \ldots, z_{iK})$$

where z_i is a class indicator vector (hidden variable):

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \text{ was generated by } j-th \text{ component} \\ 0 & \text{otherwise} \end{cases}$$

- In this case, x_i are observable and z_i non-observable.

Estimating Mixture Parameters Using EM: Case of Means (cont'd)

- Main steps using EM

- The EM algorithm searches for a ML hypothesis through the following iterative scheme:

(1) Initialize the hypothesis $\theta^0=(\mu_1^0, \mu_2^0, \ldots, \mu_K^0)$

(2) Estimate the expected values of the hidden variables z_{ij} using the current hypothesis $\theta^t=(\mu_1^t, \mu_2^t, \ldots, \mu_K^t)$

(3) Update the hypothesis $\theta^{t+1}=(\mu_1^{t+1}, \mu_2^{t+1}, \ldots, \mu_K^{t+1})$ using the expected values of the hidden variables from step 2.

- Repeat steps (2)-(3) until convergence.

Estimating Mixture Parameters Using EM: Case of Means (cont'd)

- Derivation of the Expectation Step

- We must derive an expression for $Q(\theta; \theta^t) = E_{z_i}(ln\ p(D_y/\theta) / D_x, \theta^t)$

don't get confused by notation

(1) Derive the form of ln $p(D_y/\theta)$:

$$p(D_y/\theta) = \prod_{i=1}^{n} p(y_i/\theta)$$

- We can write $p(y_i/\theta)$ as follows:

$$p(y_i/\theta) = p(x_i, z_i/\theta) = p(x_i/z_i, \theta)p(z_i/\theta) = p(x_i/\theta_j)\pi_j$$

(assuming z_{ij}=1 and z_{ik}=0 for $k \neq j$)

- We can rewrite $p(x_i/\theta_j)\pi_j$ as follows:

$$p(y_i/\theta) = \prod_{k=1}^{K} [p(x_i/\theta_k)\pi_k]^{z_{ik}}$$

Estimating Mixture Parameters Using EM: Case of Means (cont'd)

- Derivation of the Expectation Step (cont'd)

- Thus, $p(D_y/\theta)$ can be written as follows (π_k's are all equal):

$$p(D_y/\theta) = \prod_{i=1}^{n}\prod_{k=1}^{K}[p(x_i/\theta_k)]^{z_{ik}}$$

- We have assumed the form of $p(x_i/\theta_k)$ to be Gaussian:

$$p(x_i/\theta_k) = \frac{1}{\sigma\sqrt{2\pi}} exp[-\frac{(x_i-\mu_k)^2}{2\sigma^2}], \quad \text{thus}$$

$$\prod_{k=1}^{K}[p(x_i/\theta_k)]^{z_{ik}} = \frac{1}{\sigma\sqrt{2\pi}} exp[-\frac{1}{2\sigma^2}\sum_{k=1}^{K} z_{ik}(x_i-\mu_k)^2]$$

which leads to the following form for $p(D_y/\theta)$:

$$p(D_y/\theta) = \prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}} exp[-\frac{1}{2\sigma^2}\sum_{k=1}^{K} z_{ik}(x_i-\mu_k)^2]$$

Estimating Mixture Parameters Using EM: Case of Means (cont'd)

- Derivation of the Expectation Step (cont'd)

- Let's compute now $\ln\ p(D_y/\theta)$:

$$\ln\ p(D_y/\theta) = \sum_{i=1}^{n}(ln\ \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2}\sum_{k=1}^{K} z_{ik}(x_i - \mu_k)^2)$$

(2) Take the expected value of $\ln\ p(D_y/\theta)$:

$$E_{z_i}(ln\ p(D_y/\theta)/D_x, \theta^t) = E(\sum_{i=1}^{n}(ln\ \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2}\sum_{k=1}^{K} z_{ik}(x_i - \mu_k^t)^2))) =$$

$$\sum_{i=1}^{n}(ln\ \frac{1}{\sigma\sqrt{2\pi}} - \frac{1}{2\sigma^2}\sum_{k=1}^{K} E(z_{ik})(x_i - \mu_k^t)^2)$$

Estimating Mixture Parameters Using EM: Case of Means (cont'd)

- Derivation of the Expectation Step (cont'd)

- $E(z_{ik})$ is just the probability that the instance x_i was generated by the k-th component (i.e., $E(z_{ik}) = \sum_j z_{ij} P(z_{ij}) = P(z_{ik}) = P(k/x_i)$):

$$E(z_{ik}) = \frac{exp[-\frac{(x_i - \mu_k^t)^2}{2\sigma^2}]}{\sum_{j=1}^{K} exp[-\frac{(x_i - \mu_j^t)^2}{2\sigma^2}]}$$

Estimating Mixture Parameters Using EM: Case of Means (cont'd)

- Derivation of the Maximization Step

- Maximize $Q(\theta; \theta^t) = E_{z_i}(ln\ p(D_y/\theta) \ / \ D_x, \theta^t)$

$$\frac{\partial Q}{\partial \mu_k} = 0 \quad \text{or} \quad \boxed{\mu_k^{t+1} = \frac{\sum_{i=1}^{n} E(z_{ik}) x_i}{\sum_{i=1}^{n} E(z_{ik})}}$$

Estimating Mixture Parameters Using EM: Case of Means (cont'd)

- Summary of steps

Initialization step

$$\theta_k^0 = \mu_k^0$$

Expectation step

$$E(z_{ik}) = \frac{exp[-\frac{(x_i - \mu_k^t)^2}{2\sigma^2}]}{\sum_{j=1}^{K} exp[-\frac{(x_i - \mu_j^t)^2}{2\sigma^2}]}$$

Estimating Mixture Parameters Using EM: Case of Means (cont'd)

- Summary of steps (cont'd)

Maximization step

$$\mu_k^{t+1} = \frac{\sum_{i=1}^{n} E(z_{ik}) x_i}{\sum_{i=1}^{n} E(z_{ik})}$$

(4) If $\|\theta^{t+1} - \theta^t\| < \varepsilon$, stop; otherwise, go to step 2.

Lagrange Optimization

- Suppose we want to maximize $f(x)$ subject to some constraint expressed in the form:

$$g(x) = 0$$

- To find the maximum, first we form the Lagrangian function:

$$L(x, \lambda) = f(x) + \lambda g(x)$$

(λ is called the Lagrange undetermined multiplier)

- Take the derivative and set it equal to zero:

$$\frac{\partial L(x, \lambda)}{\partial x} = \frac{\partial f(x)}{\partial x} + \lambda \frac{\partial g(x)}{\partial x} = 0$$

- Solve the resulting equation for λ and the value x that maximizes $f(x)$

Lagrange Optimization (cont'd)

- If x is d-dimensional, we have $d+1$ equations and $d+1$ unknowns!
- Example: find the stationary point of $f(x_1,x_2)=x_1x_2$
  subject to the constraint $g(x_1,x_2)=x_1+x_2-1=0$

$$\frac{\partial L(x_1,x_2,\lambda)}{\partial x_1}=x_2+\lambda=0$$

$$L(x_1,x_2,\lambda)=f(x_1,x_2)+\lambda g(x_1,x_2)$$

$$\frac{\partial L(x_1,x_2,\lambda)}{\partial x_2}=x_1+\lambda=0$$

$$\frac{\partial L(x_1,x_2,\lambda)}{\partial \lambda}=x_1+x_2-1=0$$

Estimating Mixture Parameters Using EM: General Case

- If we knew which sub-model was responsible for generating each datapoint, then it would be easy to find the ML parameters for each sub-model.

(1) Use EM to estimate which sub-model was responsible for generating each datapoint.

(2) Find the ML parameters based on these estimates.

(3) Use the new ML parameters to re-estimate the responsibilities and iterate.

Estimating Mixture Parameters Using EM: General Case (cont'd)

- Involving hidden variables

- We do not know which instance x_i was generated by which component (i.e., the missing data are the labels showing which sub-model generated each datapoint).

- Augment each instance x_i by the missing information:

$$y_i = (x_i, z_i)$$

where z_i is a class indicator vector $z_i = (z_{1i}, z_{2i}, \ldots, z_{Ki})$:

$$z_{ij} = \begin{cases} 1 & \text{if } x_i \textit{ generated by } j-th \textit{ component} \\ 0 & \textit{otherwise} \end{cases}$$

(x_i are observable and z_i non-observable)

Estimating Mixture Parameters Using EM: General Case (cont'd)

- Derivation of the Expectation Step

- We must derive an expression for $Q(\theta; \theta^t) = E_{z_i}(ln\ p(D_y/\theta) \ / \ D_x, \theta^t)$

(1) Derive the form of ln $p(D_y/\theta)$:

$$p(D_y/\theta) = \prod_{i=1}^{n} p(y_i/\theta)$$

- We can write $p(y_i/\theta)$ as follows:

$$p(y_i/\theta) = p(x_i, z_i/\theta) = p(x_i/z_i, \theta)p(z_i/\theta) = p(x_i/\theta_j)\pi_j$$

(assuming z_{ij}=1 and z_{ik}=0 for $k \neq j$)

- We can rewrite the above expression as follows:

$$p(y_i/\theta) = \prod_{k=1}^{K}[p(x_i/\theta_k)\pi_k]^{z_{ik}}$$

Estimating Mixture Parameters Using EM: General Case (cont'd)

- Derivation of the Expectation Step (cont'd)

  - Thus, $p(D_y/\theta)$ can be written as follows:

$$p(D_y/\theta) = \prod_{i=1}^{n} \prod_{k=1}^{K} [p(x_i/\theta_k)\pi_k]^{z_{ik}}$$

  - We can now compute $\ln p(D_y/\theta)$

$$\ln p(D_y/\theta) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln(p(x_i/\theta_k)\pi_k) =$$

$$\sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln(p(x_i/\theta_k)) + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \ln(\pi_k)$$

Estimating Mixture Parameters Using EM: General Case (cont'd)

- Derivation of the Expectation Step (cont'd)

(2) Take the expected value of ln $p(D_y/\theta)$:

$$E(ln\ p(D_y/\theta)/D_x, \theta^t) = \sum_{i=1}^{n}\sum_{k=1}^{K} E(z_{ik}) ln\,(p(x_i/\theta_k^t)) + \sum_{i=1}^{n}\sum_{k=1}^{K} E(z_{ik}) ln\,(\pi_k^t)$$

- $E(z_{ik})$ is just the probability that instance x_i was generated by the k-th component (i.e., $E(z_{ik}) = \sum_j z_{ij} P(z_{ij}) = P(z_{ik}) = P(k/x_i)$:

$$\boxed{E(z_{ik}) = \frac{p(x_i/\theta_k^t)\pi_k^t}{\sum_{j=1}^{K} p(x_i/\theta_j^t)\pi_j^t}}$$

Estimating Mixture Parameters Using EM: General Case (cont'd)

- Derivation of the Maximization Step

- Maximize $Q(\theta; \theta^t)$ subject to the constraint $\sum_{k=1}^{K} \pi_k = 1$:

$$Q'(\theta; \theta^t) = \sum_{i=1}^{n} \sum_{k=1}^{K} E(z_{ik}) ln\,(p(x_i/\theta_k)) + \sum_{i=1}^{n} \sum_{k=1}^{K} E(z_{ik}) ln\,(\pi_k) + \lambda(1 - \sum_{k=1}^{K} \pi_k)$$

where λ is the Langrange multiplier.

$$\frac{\partial Q'}{\partial \pi_k} = 0 \quad \text{or} \quad \sum_{i=1}^{n} E(z_{ik}) \frac{1}{\pi_k} - \lambda = 0 \quad \text{or} \quad \boxed{\pi_k^{t+1} = \frac{1}{n} \sum_{i=1}^{n} E(z_{ik})}$$

(the constraint $\sum_{k=1}^{K} \pi_k = 1$ gives $\sum_{k=1}^{K} \sum_{i=1}^{n} E(z_{ik}) = \lambda$)

Estimating Mixture Parameters Using EM: General Case (cont'd)

- Derivation of the Maximization Step (cont'd)

$$\frac{\partial Q'}{\partial \mu_k} = 0 \quad \text{or} \quad \boxed{\mu_k^{t+1} = \frac{1}{n\pi_k^{t+1}} \sum_{i=1}^{n} E(z_{ik}) x_i}$$

$$\frac{\partial Q'}{\partial \Sigma_k} = 0 \quad \text{or} \quad \boxed{\Sigma_k^{t+1} = \frac{1}{n\pi_k^{t+1}} \sum_{i=1}^{n} E(z_{ik})(x_i - \mu_k^{t+1})(x_i - \mu_k^{t+1})^T}$$

Estimating Mixture Parameters Using EM: General Case (cont'd)

- Summary of Steps

Initialization step

$$\theta_k^0 = (\pi_k^0, \mu_k^0, \Sigma_k^0)$$

Expectation step

$$E(z_{ik}) = \frac{p(x_i/\theta_k^t)\pi_k^t}{\sum_{j=1}^{K} p(x_i/\theta_j^t)\pi_j^t}$$

Estimating Mixture Parameters Using EM: General Case (cont'd)

- Summary of Steps (cont'd)

Maximization step

$$\pi_k^{t+1} = \frac{1}{n}\sum_{i=1}^{n} E(z_{ik})$$

$$\mu_k^{t+1} = \frac{1}{n\pi_k^{t+1}}\sum_{i=1}^{n} E(z_{ik})x_i$$

$$\Sigma_k^{t+1} = \frac{1}{n\pi_k^{t+1}}\sum_{i=1}^{n} E(z_{ik})(x_i - \mu_k^{t+1})(x_i - \mu_k^{t+1})^T$$

(4) If $\|\theta^{t+1} - \theta^t\| < \varepsilon$, stop; otherwise, go to step 2.

Estimating the Number of Components K

- Use EM to obtain a sequence of parameter estimates for a range of values K

$$\{\Theta_{(K)}, K=K_{\min},\ldots,K_{\max}\}$$

- The estimate of K is then defined as a minimizer of some cost function:

$$\hat{K} = arg\ \min_K(C(\Theta_{(K)}, K), K=K_{\min},\ldots,K_{\max}$$

- Most often, the cost function includes ln $p(D_y/\theta)$ and an additional term whose role is to penalize large values of K.

- Several criteria have been used, e.g., Minimum description length (MDL)

Estimating the Number of Components K (cont'd)