DATA MINING
THE EM ALGORITHM

Maximum Likelihood Estimation

MIXTURE MODELS AND THE EM ALGORITHM

Model-based clustering

- In order to understand our data, we will assume that there is a generative process (a model) that creates/describes the data, and we will try to find the model that best fits the data.
  - Models of different complexity can be defined, but we will assume that our model is a distribution from which data points are sampled
  - Example: the data is the height of all adults in Greece

- In most cases, a single distribution is not good enough to describe all data points: different parts of the data follow a different distribution
  - Example: the data is the height of all adults and children in Greece
  - We need a mixture model
  - Different distributions correspond to different clusters in the data.

Gaussian Distribution

- Example: the data is the height of all adults in Greece
  - Experience has shown that this data follows a Gaussian (Normal) distribution
  - Reminder: Normal distribution:

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  - μ = mean, σ = standard deviation

Gaussian Model

- What is a model?
  - A Gaussian distribution is fully defined by the mean μ and the standard deviation σ
  - We define our model as the pair of parameters $\theta = (\mu, \sigma)$

- This is a general principle: a model is defined as a vector of parameters θ

Fitting the model

- We want to find the normal distribution that best fits our data
  - Find the best values for μ and σ
  - But what does best fit mean?

Maximum Likelihood Estimation (MLE)

- Find the most likely parameters given the data. Given the data observations X, find θ that maximizes $P(\theta|X)$
  - Problem: We do not know how to compute $P(\theta|X)$

- Using Bayes Rule:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

- If we have no prior information about θ, or X, we can assume uniform. Maximizing $P(\theta|X)$ is now the same as maximizing $P(X|\theta)$

Maximum Likelihood Estimation (MLE)

- We have a vector $X = (x_1, \dots, x_n)$ of values and we want to fit a Gaussian $N(\mu, \sigma)$ model to the data
  - Our parameter set is $\theta = (\mu, \sigma)$
- Probability of observing point x_i given the parameters θ

$$P(x_i|\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

We cheated a little here.
More accurately we look at:
$P(x_i \le x \le x_i + dx)$

- Probability of observing all points (assume independence)

$$P(X|\theta) = \prod_{i=1}^{n} P(x_i|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- We want to find the parameters $\theta = (\mu, \sigma)$ that maximize the probability $P(X|\theta)$

Maximum Likelihood Estimation (MLE)

- The probability $P(X|\theta)$ as a function of θ is called the Likelihood function

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}$$

- It is usually easier to work with the Log-Likelihood function

$$LL(\theta) = -\sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} - \frac{1}{2} n \log 2\pi - n \log \sigma$$

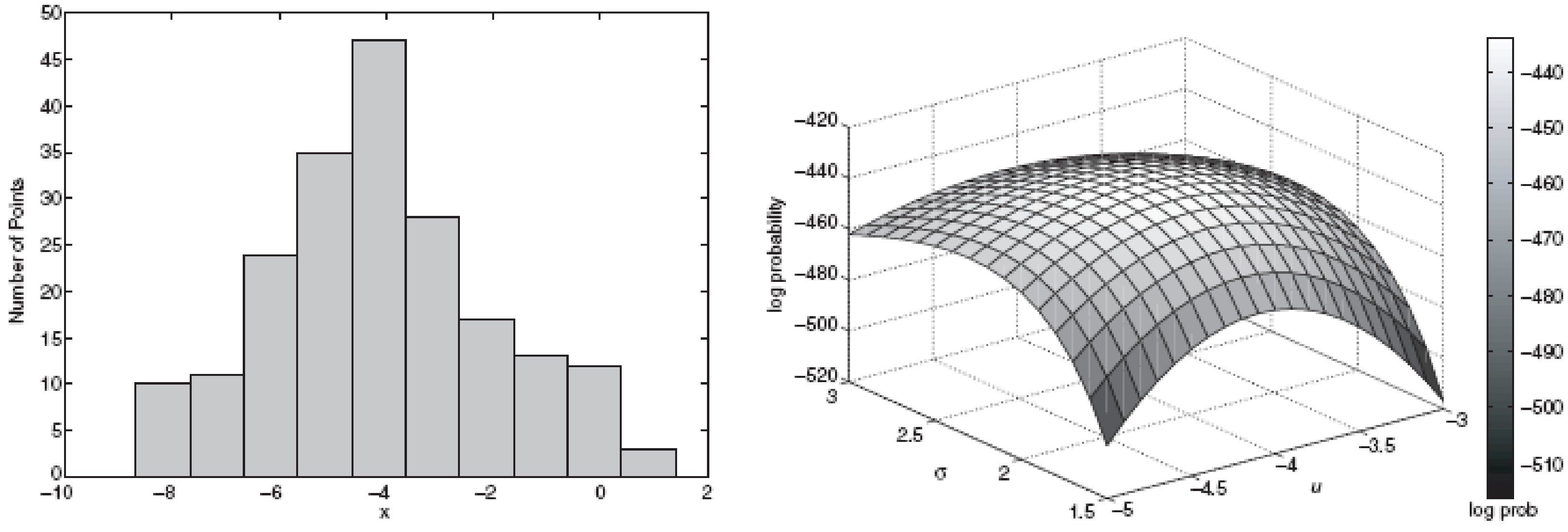
- Maximum Likelihood Estimation
  - Find parameters μ, σ that maximize $LL(\theta)$

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i = \mu_X$$

Sample Mean

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 = \sigma_X^2$$

Sample Variance



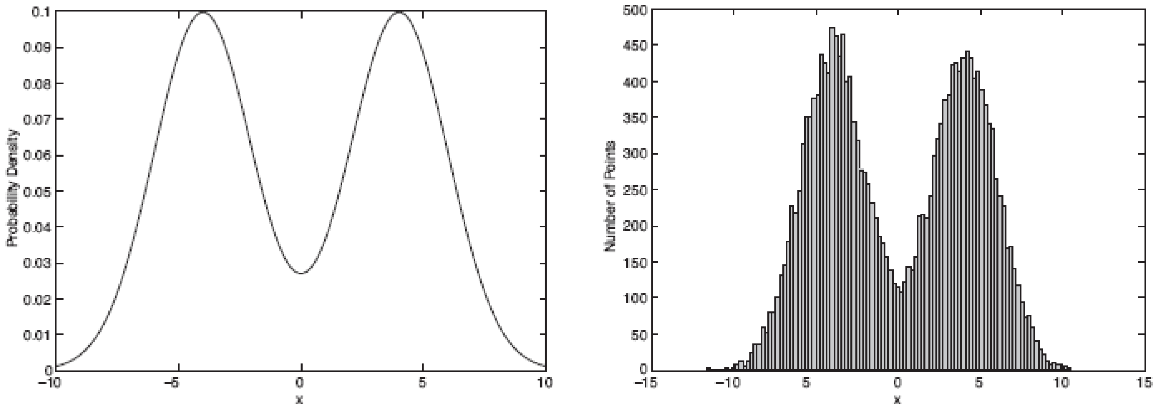
(a) Histogram of 200 points from a Gaussian distribution.

(b) Log likelihood plot of the 200 points for different values of the mean and standard deviation.

Figure 9.3. 200 points from a Gaussian distribution and their log probability for different parameter values.

Mixture of Gaussians

- Suppose that you have the heights of adults and children, and the distribution looks like the figure below



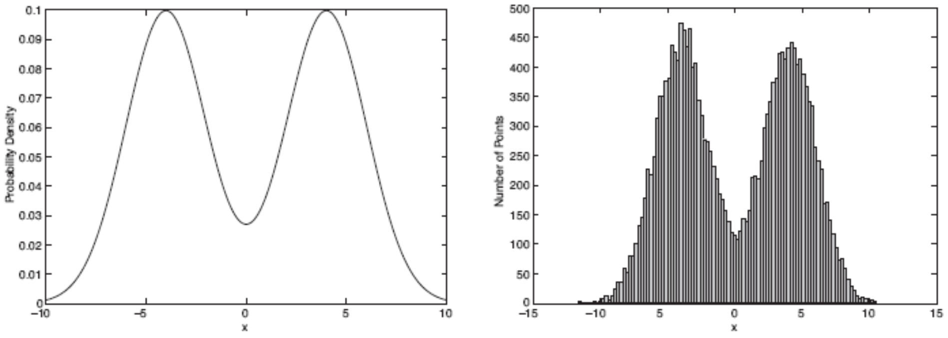
(a) Probability density function for the mixture model.

(b) 20,000 points generated from the mixture model.

Figure 9.2. Mixture model consisting of two normal distributions with means of -4 and 4, respectively. Both distributions have a standard deviation of 2.

Mixture of Gaussians

- In this case the data is the result of the mixture of two Gaussians
  - One for Adults, and one for Children
  - Identifying for each value which Gaussian is most likely to have generated it will give us a clustering.



(a) Probability density function for the mixture model.

(b) 20,000 points generated from the mixture model.

Figure 9.2. Mixture model consisting of two normal distributions with means of -4 and 4, respectively. Both distributions have a standard deviation of 2.

Mixture model

- A value x_i is generated according to the following process:
  - First select the age group
    - With probability π_A select Adult, with probability π_C select Child $(\pi_G + \pi_C = 1)$

We can also think of this as a Hidden Variable Z that takes two values: Adult and Child

$$\pi_A = P(Z = \text{Adult}), \pi_C = P(Z = \text{Child})$$

  - Given the nationality, generate the point from the corresponding Gaussian
    - $P(x_i|\theta_A) \sim N(\mu_A, \sigma_A)$ if Adult
    - $P(x_i|\theta_C) \sim N(\mu_C, \sigma_C)$ if Child

θ_G: parameters of the Adult distribution
θ_C: parameters of the Child distribution

Using the Hidden Variable Z:

$$P(x_i|Z = \text{Adult}) = P(x_i|\theta_A) \sim N(\mu_A, \sigma_A)$$
$$P(x_i|Z = \text{Child}) = P(x_i|\theta_C) \sim N(\mu_C, \sigma_C)$$

Mixture Model

- Our model has the following parameters

$$\Theta = (\pi_A, \pi_C, \mu_A, \sigma_A, \mu_C, \sigma_C)$$

Mixture probabilities

θ_A: parameters of the Adult distribution

θ_C: parameters of the Child distribution

Mixture Model

- Our model has the following parameters

$$\Theta = (\pi_A, \pi_C, \mu_A, \sigma_A, \mu_C, \sigma_C)$$

Mixture probabilities | Distribution Parameters

- For value x_i, we have:

$$P(x_i|\Theta) = \pi_A P(x_i|\theta_A) + \pi_C P(x_i|\theta_C)$$

- For all values $X = (x_1, \ldots, x_n)$

$$P(X|\Theta) = \prod_{i=1}^{n} P(x_i|\Theta)$$

- We want to estimate the parameters that maximize the Likelihood of the data

Mixture Models

- Once we have the parameters $\Theta = (\pi_A, \pi_C, \mu_A, \mu_C, \sigma_A, \sigma_C)$ we can estimate the membership probabilities $P(A|x_i)$ and $P(C|x_i)$ for each point x_i:
  - This is the probability that point x_i belongs to the Adult or the Child population (cluster)
  - Using Bayes Rule:

Given from the Gaussian distribution $N(\mu_G, \sigma_G)$ for Greek

$$P(A|x_i) = \frac{P(x_i|A)P(A)}{P(x_i|A)P(A) + P(x_i|C)P(C)}$$

$$= \frac{P(x_i|\theta_A)\pi_A}{P(x_i|\theta_A)\pi_A + P(x_i|\theta_C)\pi_C}$$

EM (Expectation Maximization) Algorithm

- Initialize the values of the parameters in Θ to some random values
- Repeat until convergence
  - E-Step: Given the parameters Θ estimate the membership probabilities $P(A|x_i)$ and $P(C|x_i)$
  - M-Step: Compute the parameter values that (in expectation) maximize the data likelihood $LL(\Theta) = \sum_{x_i} \log\big(\pi_C P(x_i|\theta_C) + \pi_A P(x_i|\theta_A)\big)$

$$\pi_C = \frac{1}{n}\sum_{i=1}^{n} P(C|x_i) \qquad \pi_A = \frac{1}{n}\sum_{i=1}^{n} P(A|x_i)$$

Fraction of population in G,C

$$\mu_C = \frac{1}{n \cdot \pi_C}\sum_{i=1}^{n} P(C|x_i)x_i \qquad \mu_G = \frac{1}{n \cdot \pi_A}\sum_{i=1}^{n} P(A|x_i)x_i$$

$$\sigma_C^2 = \frac{1}{n \cdot \pi_C}\sum_{i=1}^{n} P(C|x_i)(x_i - \mu_C)^2 \qquad \sigma_G^2 = \frac{1}{n \cdot \pi_A}\sum_{i=1}^{n} P(A|x_i)(x_i - \mu_A)^2$$

MLE Estimates if π's were fixed

Relationship to K-means

- E-Step: Assignment of points to clusters
  - K-means: hard assignment, EM: soft assignment
- M-Step: Computation of centroids
  - K-means assumes common fixed variance (spherical clusters)
  - EM: can change the variance for different clusters or different dimensions (ellipsoid clusters)
- If the variance is fixed then both minimize the same error function

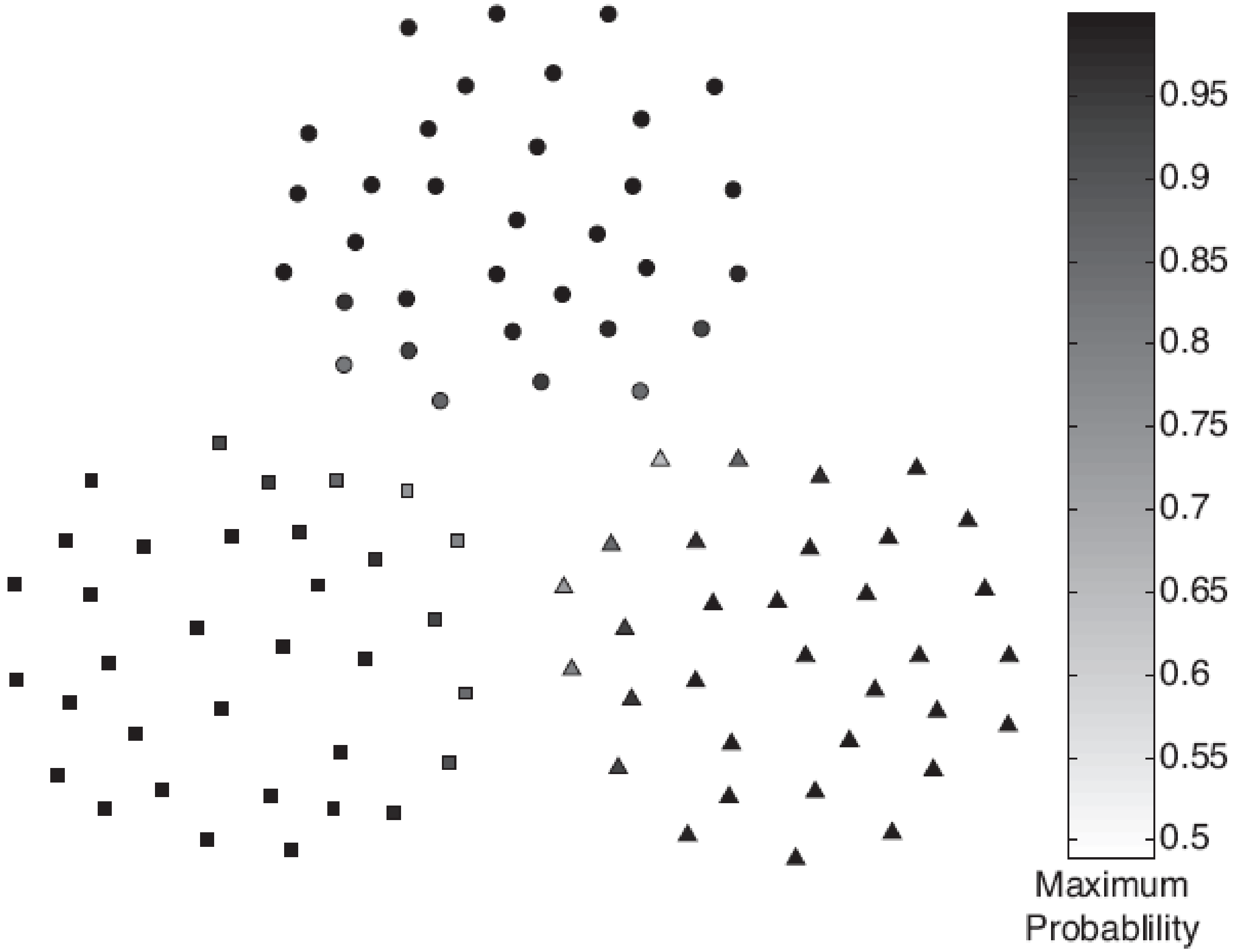


Figure 9.4. EM clustering of a two-dimensional point set with three clusters.

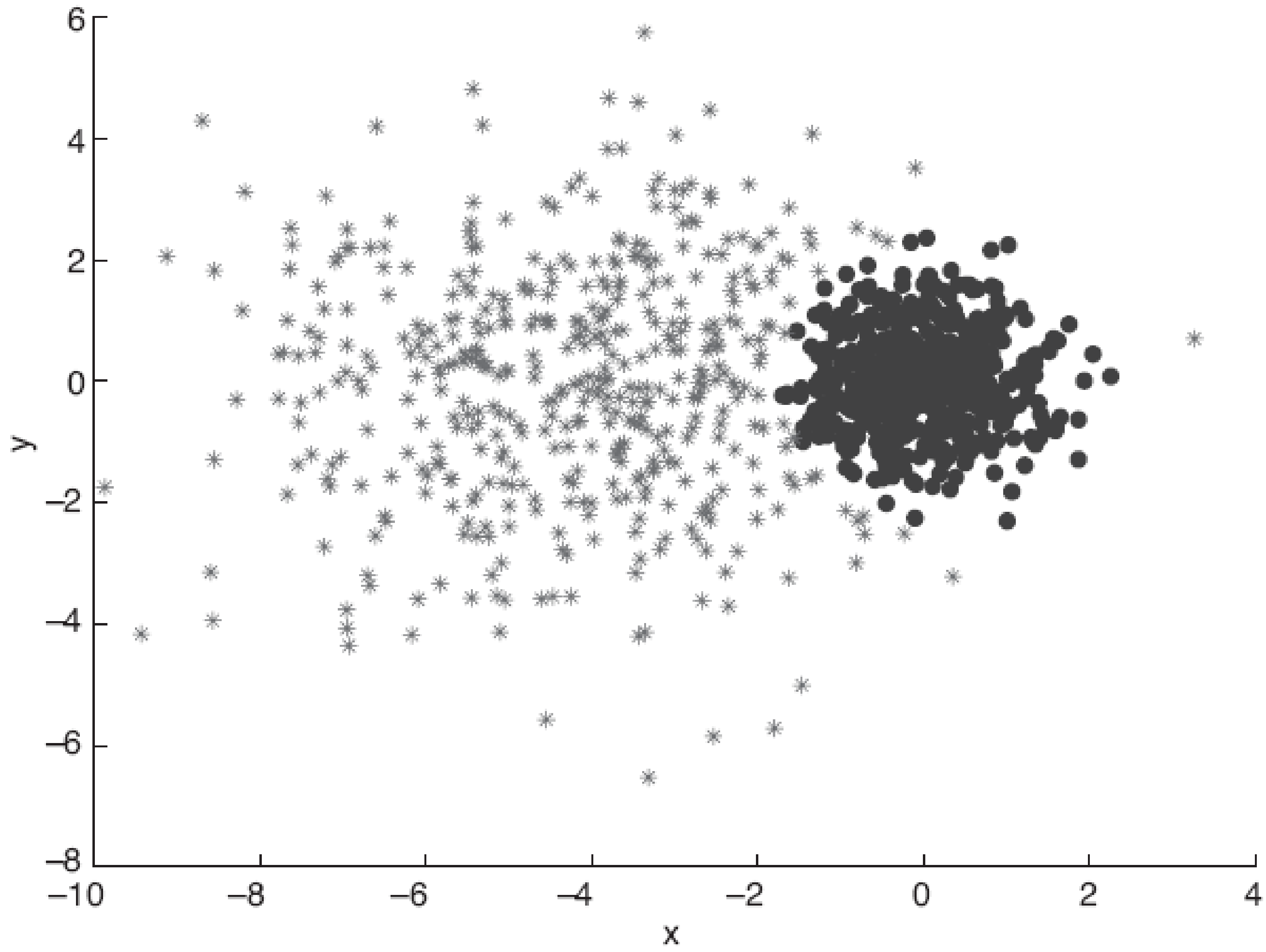
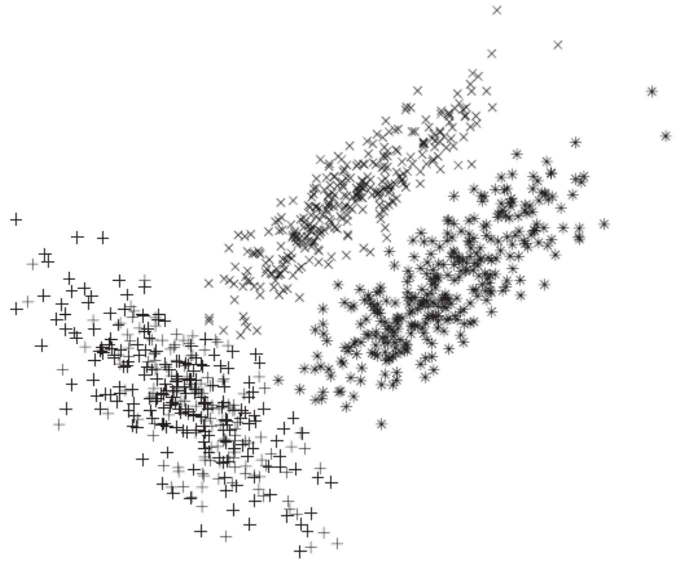
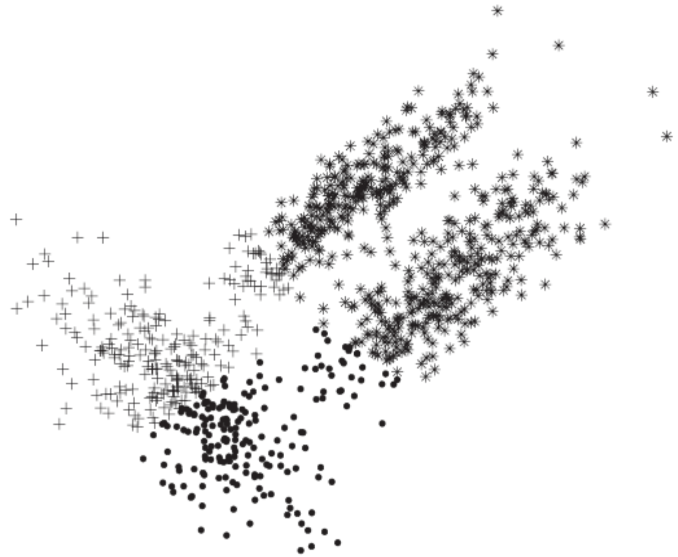


Figure 9.5. EM clustering of a two-dimensional point set with two clusters of differing density.



(a) Clusters produced by mixture model clustering.



(b) Clusters produced by K-means clustering.

Figure 9.6. Mixture model and K-means clustering of a set of two-dimensional points.