Expectation-Maximization Clustering

Gaussian Mixture Method

$$f_i(\mathbf{x}) = f(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{(2\pi)^{\frac{d}{2}}|\boldsymbol{\Sigma}_i|^{\frac{1}{2}}} \exp\left\{-\frac{(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)}{2}\right\}$$

$$f(\mathbf{x}) = \sum_{i=1}^{k} f_i(\mathbf{x}) P(C_i) = \sum_{i=1}^{k} f(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) P(C_i)$$

$$\sum_{i=1}^{k} P(C_i) = 1$$

$$\boldsymbol{\theta} = \left\{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, P(C_i) \dots, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, P(C_k)\right\}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Maximum Likelihood Estimation

$$P(\mathbf{D}|\theta) = \prod_{j=1}^{n} f(\mathbf{x}_j)$$

$$\theta^* = \arg\max_{\theta}\{P(\mathbf{D}|\theta)\}$$

$$\theta^* = \arg\max_{\theta}\{\ln P(\mathbf{D}|\theta)\}$$

$$\ln P(\mathbf{D}|\theta) = \sum_{j=1}^{n} \ln f(\mathbf{x}_j) = \sum_{j=1}^{n} \ln\left(\sum_{i=1}^{k} f(\mathbf{x}_j|\mu_i, \mathbf{\Sigma}_i) P(C_i)\right)$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Expectation Step

$$P(C_i|\mathbf{x}_j) = \frac{P(C_i \text{ and } \mathbf{x}_j)}{P(\mathbf{x}_j)} = \frac{P(\mathbf{x}_j|C_i)P(C_i)}{\sum_{a=1}^{k} P(\mathbf{x}_j|C_a)P(C_a)}$$

$$P(\mathbf{x}_j|C_i) \simeq 2\epsilon \cdot f(\mathbf{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = 2\epsilon \cdot f_i(\mathbf{x}_j)$$

$$P(C_i|\mathbf{x}_j) = \frac{f_i(\mathbf{x}_j) \cdot P(C_i)}{\sum_{a=1}^{k} f_a(\mathbf{x}_j) \cdot P(C_a)}$$

$$f_i(x) = f(x|\mu_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left\{-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right\}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Maximization Step

$$\mu_i = \frac{\sum_{j=1}^{n} w_{ij} \cdot x_j}{\sum_{j=1}^{n} w_{ij}} \qquad \mu_i = \frac{\mathbf{w}_i^T X}{\mathbf{w}_i^T \mathbf{1}}$$

$$\sigma_i^2 = \frac{\sum_{j=1}^{n} w_{ij} (x_j - \mu_i)^2}{\sum_{j=1}^{n} w_{ij}} \qquad \sigma_i^2 = \frac{\mathbf{w}_i^T Z_i^s}{\mathbf{w}_i^T \mathbf{1}}$$

$$P(C_i) = \frac{\sum_{j=1}^{n} w_{ij}}{\sum_{a=1}^{k} \sum_{j=1}^{n} w_{aj}} = \frac{\sum_{j=1}^{n} w_{ij}}{\sum_{j=1}^{n} 1} = \frac{\sum_{j=1}^{n} w_{ij}}{n}$$

$$\sum_{i=1}^{k} w_{ij} = \sum_{i=1}^{k} P(C_i | x_j) = 1 \qquad P(C_i) = \frac{\mathbf{w}_i^T \mathbf{1}}{n}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

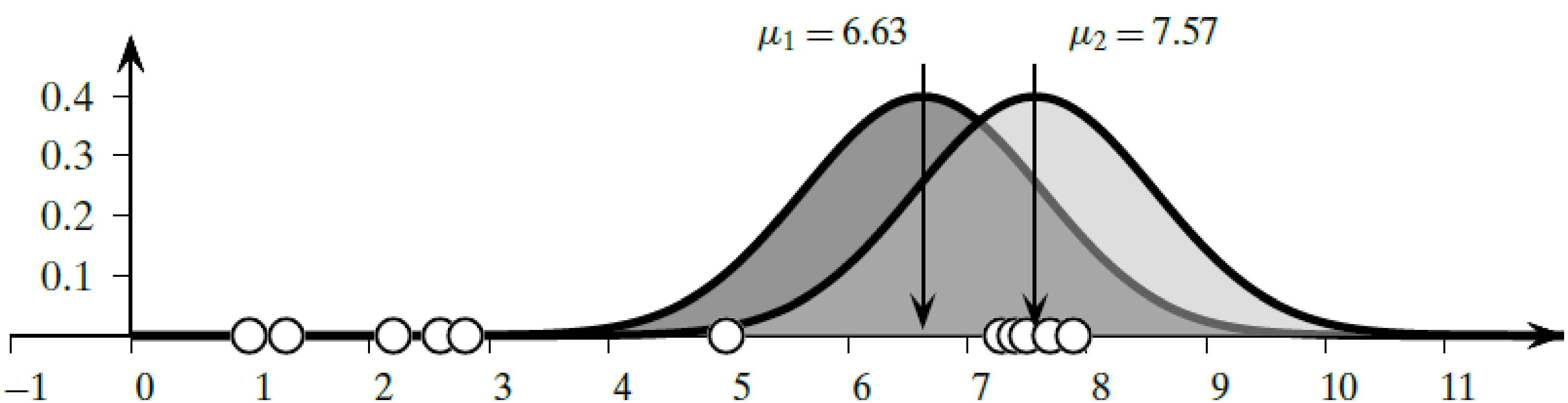
Example 13.4

$x_1 = 1.0 \quad x_2 = 1.3 \quad x_3 = 2.2 \quad x_4 = 2.6 \quad x_5 = 2.8$

$x_6 = 5.0 \quad x_7 = 7.3 \quad x_8 = 7.4 \quad x_9 = 7.5 \quad x_{10} = 7.7 \quad x_{11} = 7.9$

$\mu_1 = 6.63 \quad \sigma_1^2 = 1 \quad P(C_2) = 0.5$

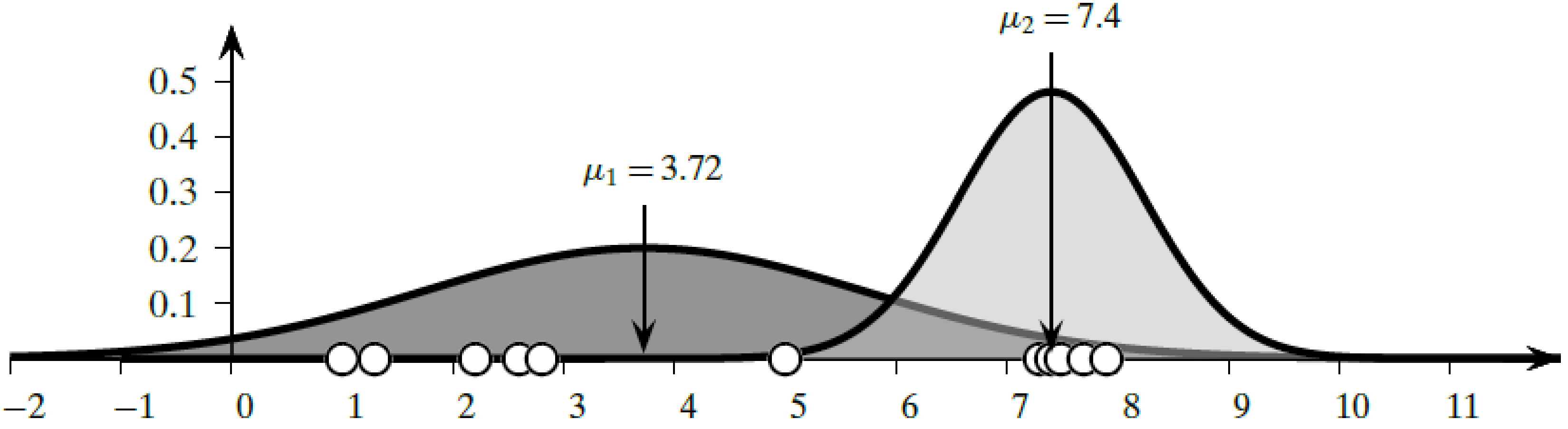
$\mu_2 = 7.57 \quad \sigma_2^2 = 1 \quad P(C_2) = 0.5$



Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Example 13.4 continued

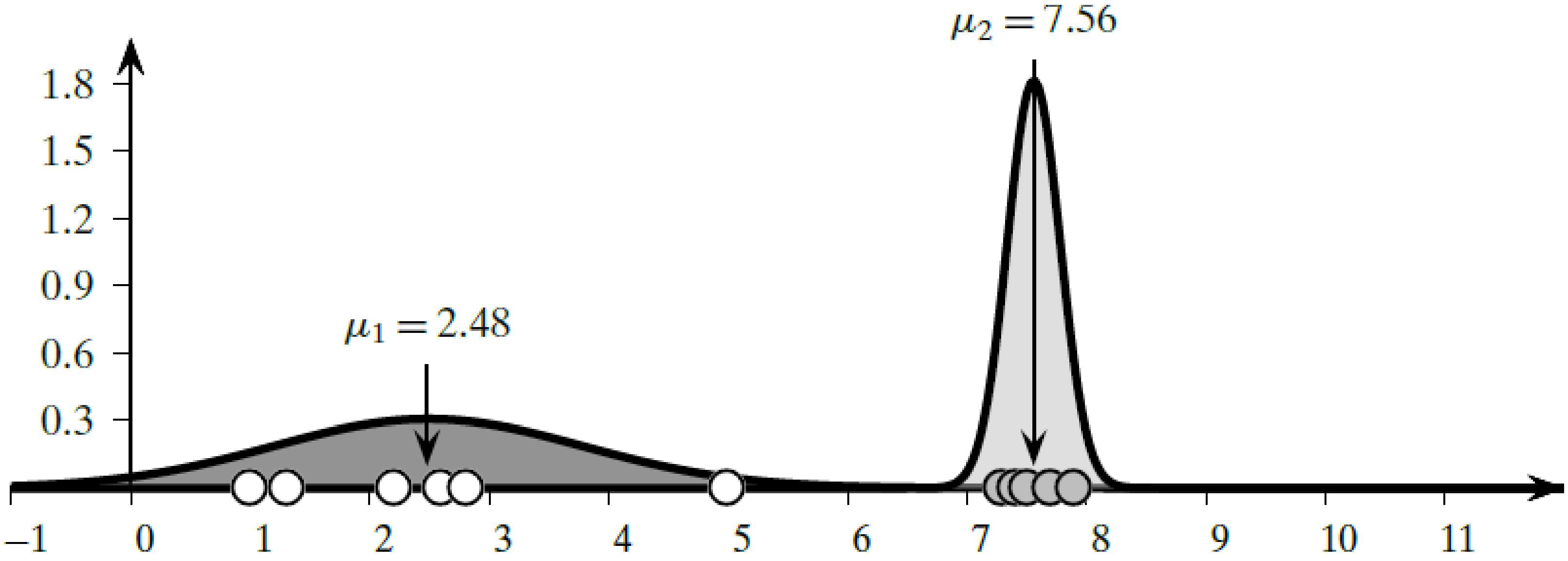
$\mu_1 = 3.72$	$\sigma_1^2 = 6.13$	$P(C_1) = 0.71$
$\mu_2 = 7.4$	$\sigma_2^2 = 0.69$	$P(C_2) = 0.29$



Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Example 13.4 continued

$\mu_1 = 2.48$	$\sigma_1^2 = 1.69$	$P(C_1) = 0.55$
$\mu_2 = 7.56$	$\sigma_2^2 = 0.05$	$P(C_2) = 0.45$



Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Expectation-Maximization Clustering

The End

EM in d Dimensions

EM in d Dimensions

$$\mu_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{id})^T$$

$$\mathbf{\Sigma}_i = \begin{pmatrix} (\sigma_1^i)^2 & \sigma_{12}^i & \ldots & \sigma_{1d}^i \\ \sigma_{21}^i & (\sigma_2^i)^2 & \ldots & \sigma_{2d}^i \\ \vdots & \vdots & \ddots & \\ \sigma_{d1}^i & \sigma_{d2}^i & \ldots & (\sigma_d^i)^2 \end{pmatrix}$$

$$\mathbf{\Sigma}_i = \begin{pmatrix} (\sigma_1^i)^2 & 0 & \ldots & 0 \\ 0 & (\sigma_2^i)^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \ldots & (\sigma_d^i)^2 \end{pmatrix}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Initialization

- Randomly select the mean for each cluster across each dimension uniformly
- Initialize the covariance matrix, d x d identity matrix, $\mathbf{\Sigma}_i = \mathbf{I}.$
- Initialize the prior probability equally across all clusters, $P(C_i) = \frac{1}{k}$

Expectation

- Compute the posterior probability for each cluster
- Remember the posterior probability $P(C_i|\mathbf{x}_j)$ can be considered a weight for each point in the cluster
- Use the notation $\mathbf{w}_i = (w_{i1}, w_{i2}, \ldots, w_{in})^T$ across all points to denote the weight vector for the cluster

Maximization

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{n} w_{ij} \cdot \mathbf{x}_j}{\sum_{j=1}^{n} w_{ij}} \qquad\qquad \boldsymbol{\mu}_i = \frac{\mathbf{D}^T \mathbf{w}_i}{\mathbf{w}_i^T \mathbf{1}}$$

$$\boldsymbol{\Sigma}_i = \frac{\sum_{j=1}^{n} w_{ij} \mathbf{z}_{ji} \mathbf{z}_{ji}^T}{\mathbf{w}_i^T \mathbf{1}}$$

$$\sigma_{ab}^i = \frac{\sum_{j=1}^{n} w_{ij} (x_{ja} - \mu_{ia})(x_{jb} - \mu_{ib})}{\sum_{j=1}^{n} w_{ij}}$$

$$P(C_i) = \frac{\sum_{j=1}^{n} w_{ij}}{n} = \frac{\mathbf{w}_i^T \mathbf{1}}{n}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Expectation-Maximization Algorithm

EXPECTATION-MAXIMIZATION $(\mathbf{D}, k, \epsilon)$:

$t \leftarrow 0$

`// Initialization`

Randomly initialize $\boldsymbol{\mu}_1^t, \ldots, \boldsymbol{\mu}_k^t$

$\boldsymbol{\Sigma}_i^t \leftarrow \mathbf{I},\ \forall i = 1, \ldots, k$

$P^t(C_i) \leftarrow \frac{1}{k},\ \forall i = 1, \ldots, k$

repeat

  $t \leftarrow t + 1$

  `// Expectation Step`

  for $i = 1, \ldots, k$ *and* $j = 1, \ldots, n$ **do**

    $w_{ij} \leftarrow \frac{f(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \cdot P(C_i)}{\sum_{a=1}^{k} f(\mathbf{x}_j | \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a) \cdot P(C_a)}$ `// posterior probability` $P^t(C_i | \mathbf{x}_j)$

  `// Maximization Step`

  for $i = 1, \ldots, k$ **do**

    $\boldsymbol{\mu}_i^t \leftarrow \frac{\sum_{j=1}^{n} w_{ij} \cdot \mathbf{x}_j}{\sum_{j=1}^{n} w_{ij}}$ `// re-estimate mean`

    $\boldsymbol{\Sigma}_i^t \leftarrow \frac{\sum_{j=1}^{n} w_{ij} (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T}{\sum_{j=1}^{n} w_{ij}}$ `// re-estimate covariance matrix`

    $P^t(C_i) \leftarrow \frac{\sum_{j=1}^{n} w_{ij}}{n}$ `// re-estimate priors`

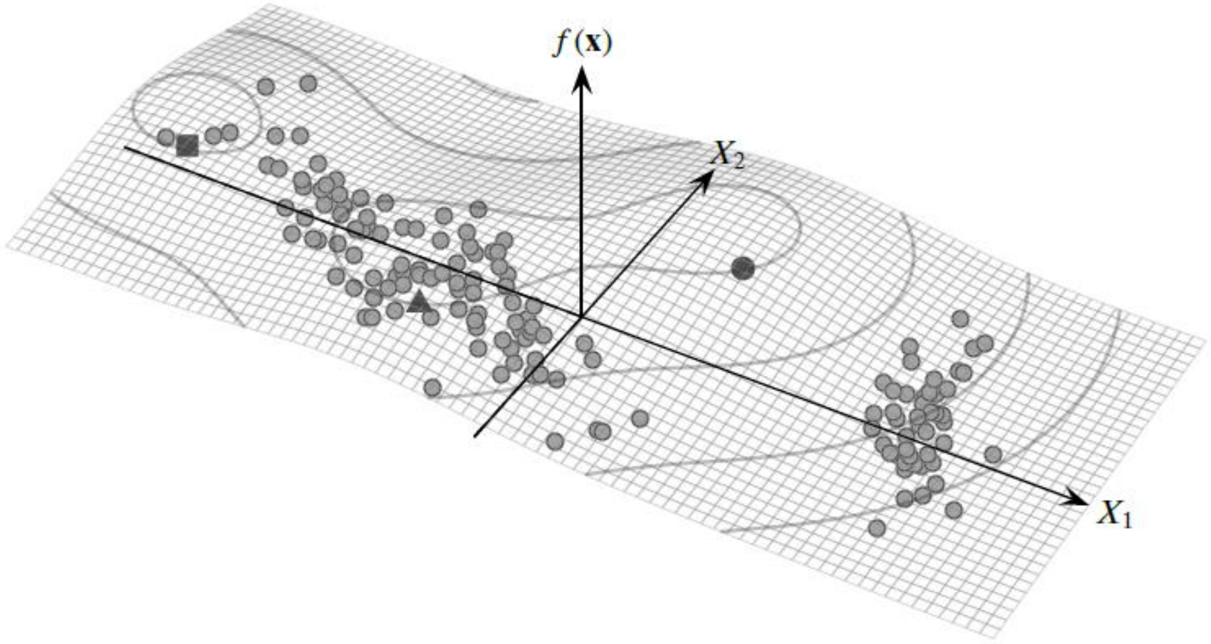
until $\sum_{i=1}^{k} \left\| \boldsymbol{\mu}_i^t - \boldsymbol{\mu}_i^{t-1} \right\|^2 \leq \epsilon$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Example 13.5

$n = 150$ $k = 3$, $\Sigma_i = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ $P(C_i) = 1/3$

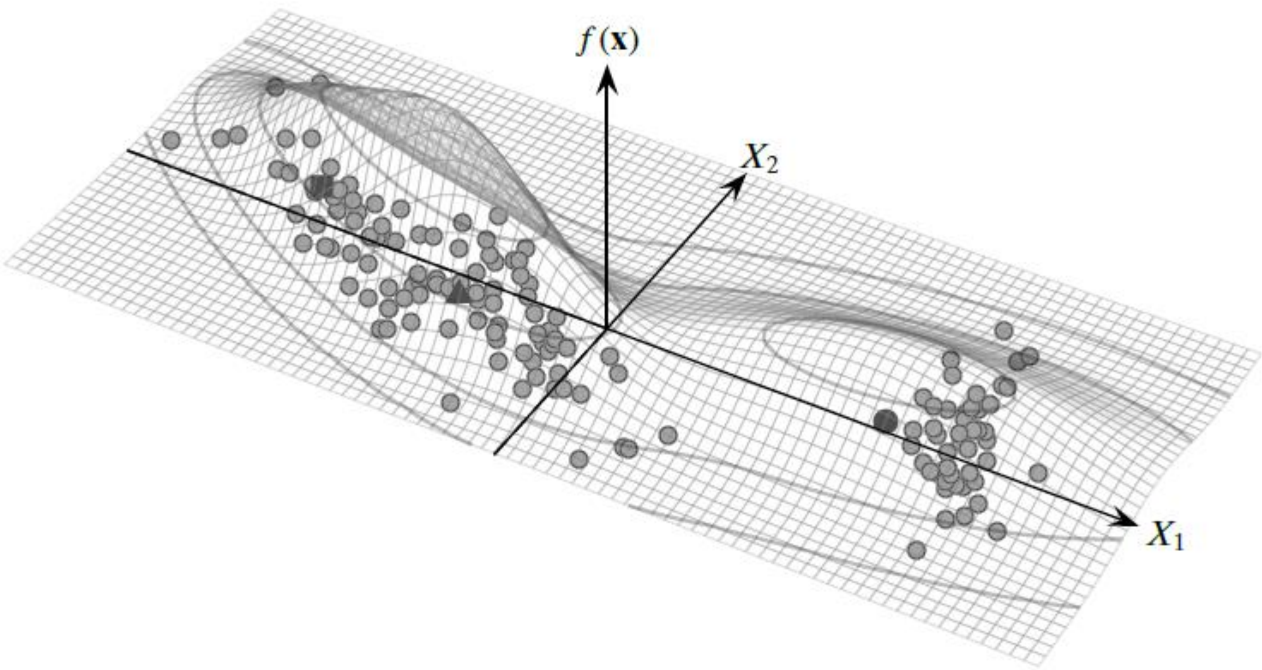
$\mu_1 = (-3.59, 0.25)^T$ $\mu_2 = (-1.09, -0.46)^T$ $\mu_3 = (0.75, 1.07)^T$



Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Example 13.5 continued

$\epsilon = 0.001$ $t = 1$



Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Example 13.5 continued

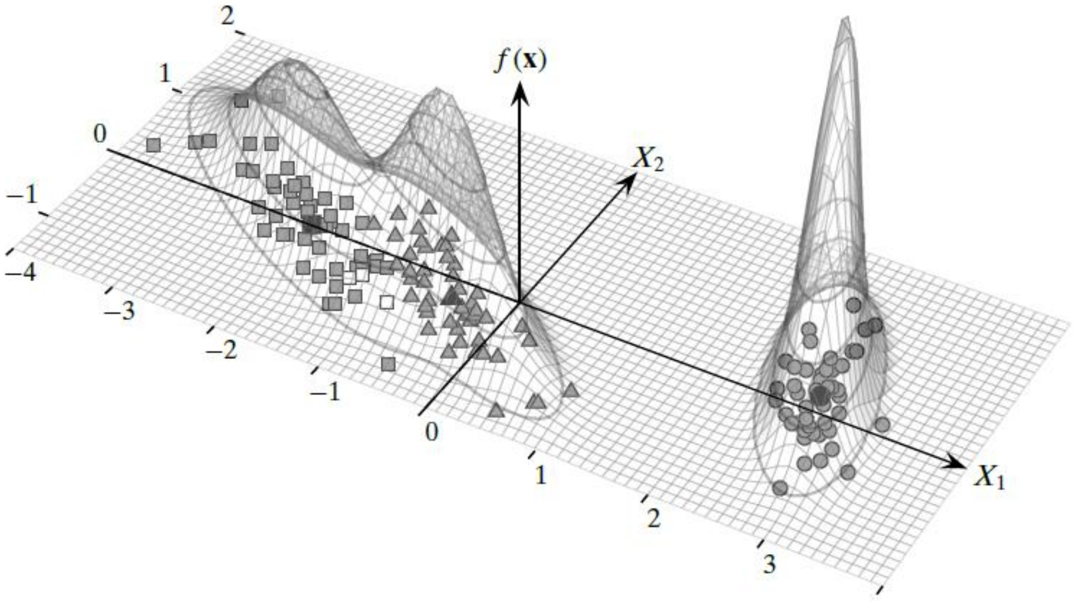
$\mu_1 = (-2.02, 0.017)^T$ $\mu_2 = (-0.51, -0.23)^T$ $\mu_3 = (2.64, 0.19)^T$

$\Sigma_1 = \begin{pmatrix} 0.56 & -0.29 \\ -0.29 & 0.23 \end{pmatrix}$ $\Sigma_2 = \begin{pmatrix} 0.36 & -0.22 \\ -0.22 & 0.19 \end{pmatrix}$ $\Sigma_3 = \begin{pmatrix} 0.05 & -0.06 \\ -0.06 & 0.21 \end{pmatrix}$

$P(C_1) = 0.36$

$P(C_2) = 0.31$

$P(C_3) = 0.33$



Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Example 13.5 continued

$$\mu_1 = (-2.02, 0.017)^T \qquad \mu_2 = (-0.51, -0.23)^T \qquad \mu_3 = (2.64, 0.19)^T$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.56 & -0.29 \\ -0.29 & 0.23 \end{pmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.36 & -0.22 \\ -0.22 & 0.19 \end{pmatrix} \qquad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 0.05 & -0.06 \\ -0.06 & 0.21 \end{pmatrix}$$

$$P(C_1) = 0.36 \qquad P(C_2) = 0.31 \qquad P(C_3) = 0.33$$

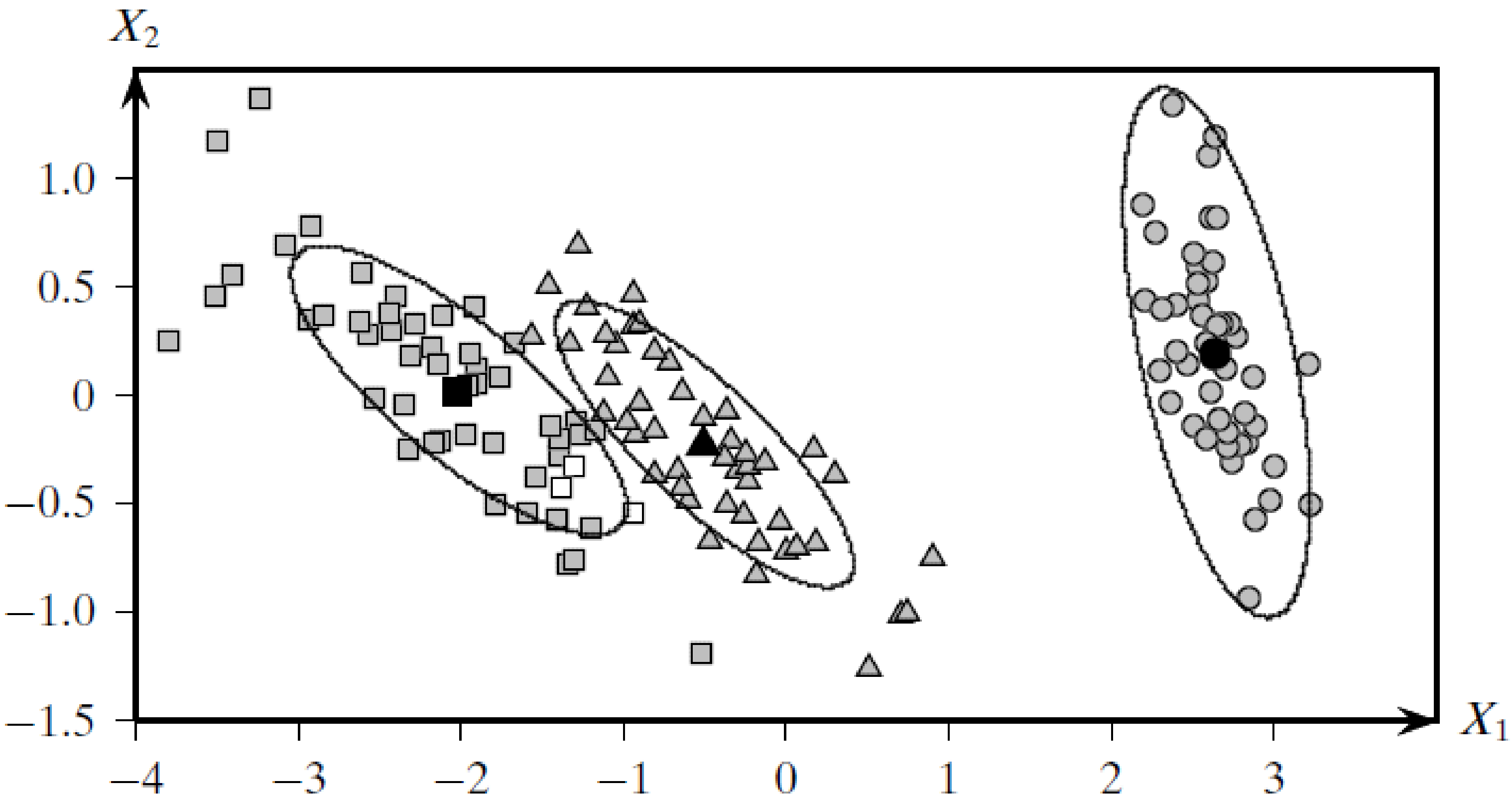
$$\mu_1 = (-2.1, 0.28)^T \qquad \mu_2 = (-0.67, -0.40)^T \qquad \mu_3 = (2.64, 0.19)^T$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 0.59 & 0 \\ 0 & 0.11 \end{pmatrix} \qquad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.49 & 0 \\ 0 & 0.11 \end{pmatrix} \qquad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.21 \end{pmatrix}$$

$$P(C_1) = 0.30 \qquad P(C_2) = 0.37 \qquad P(C_3) = 0.33$$

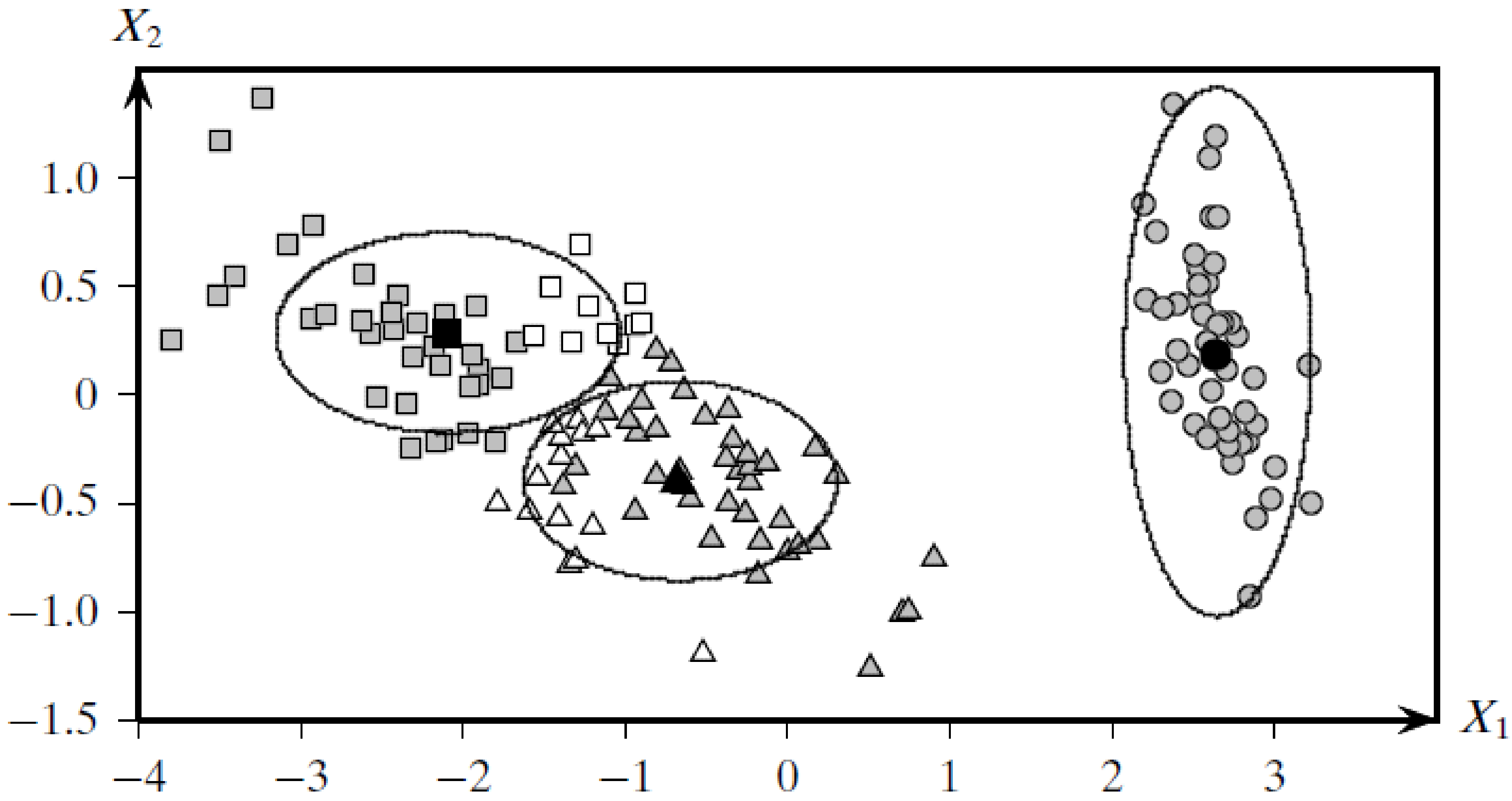
Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Example 13.5 continued



Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Example 13.5 continued



Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Computational Complexity

- The posterior probability for the expectation step takes $O(kd^3)$
- Evaluating the density takes $O(knd^2)$
- In the maximization step the time is dominated by updating Σ_i , $O(knd^2)$
- For a total of $O(t(kd^3 + nkd^2))$ where t = number of iterations

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Computational Complexity

- For the diagonal covariance method the density computation for the expectation step takes $O(knd)$.
- The same is true for the maximization step
- For a total time of $O(tnkd)$
- The I/O time takes $O(t)$ because the entire database of points must be read for each iteration

EM in d Dimensions

The End

K-means as Specialization of EM

K-means as Specialization of EM

$$P(\mathbf{x}_j|C_i) = \begin{cases} 1 & \text{if } C_i = \arg\min_{C_a} \left\{ \left\| \mathbf{x}_j - \mu_a \right\|^2 \right\} \\ 0 & \text{otherwise} \end{cases}$$

$$P(C_i|\mathbf{x}_j) = \frac{P(\mathbf{x}_j|C_i)P(C_i)}{\sum_{a=1}^{k} P(\mathbf{x}_j|C_a)P(C_a)}$$

$$P(C_i|\mathbf{x}_j) = \begin{cases} 1 & \text{if } \mathbf{x}_j \in C_i, \text{i.e., if } C_i = \arg\min_{C_a} \left\{ \left\| \mathbf{x}_j - \mu_a \right\|^2 \right\} \\ 0 & \text{otherwise} \end{cases}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Maximum Likelihood Estimation

$$\frac{\partial}{\partial \boldsymbol{\theta}_i} \ln\Big(P(\mathbf{D}|\theta)\Big) = \frac{\partial}{\partial \boldsymbol{\theta}_i}\left(\sum_{j=1}^{n} \ln f(\mathbf{x}_j)\right)$$

$$= \sum_{j=1}^{n}\left(\frac{1}{f(\mathbf{x}_j)} \cdot \frac{\partial f(\mathbf{x}_j)}{\partial \boldsymbol{\theta}_i}\right)$$

$$= \sum_{j=1}^{n}\left(\frac{1}{f(\mathbf{x}_j)} \sum_{a=1}^{k} \frac{\partial}{\partial \boldsymbol{\theta}_i}\Big(f(\mathbf{x}_j|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_a)P(C_a)\Big)\right)$$

$$= \sum_{j=1}^{n}\left(\frac{1}{f(\mathbf{x}_j)} \cdot \frac{\partial}{\partial \boldsymbol{\theta}_i}\Big(f(\mathbf{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)P(C_i)\Big)\right)$$

$$f(\mathbf{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = (2\pi)^{-\frac{d}{2}}\ |\boldsymbol{\Sigma}_i^{-1}|^{\frac{1}{2}}\ \exp\Big\{g(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\Big\}$$

$$g(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = -\frac{1}{2}(\mathbf{x}_j - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_j - \boldsymbol{\mu}_i)$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Maximum Likelihood Estimation

$$\frac{\partial}{\partial \boldsymbol{\theta}_i} \ln\Big(P(\mathbf{D}|\theta)\Big) =$$

$$\sum_{j=1}^{n}\left(\frac{1}{f(\mathbf{x}_j)} \cdot \frac{\partial}{\partial \boldsymbol{\theta}_i}\Big((2\pi)^{-\frac{d}{2}}\,|\boldsymbol{\Sigma}_i^{-1}|^{\frac{1}{2}} \exp\big\{g(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\big\} P(C_i)\Big)\right)$$

$$\frac{\partial}{\partial \boldsymbol{\theta}_i} \exp\big\{g(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\big\} = \exp\big\{g(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\big\} \cdot \frac{\partial}{\partial \boldsymbol{\theta}_i} g(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Estimation of Mean

$$\frac{\partial}{\partial \boldsymbol{\mu}_i} g(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_j - \boldsymbol{\mu}_i)$$

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}_i} \ln(P(\mathbf{D}|\boldsymbol{\theta})) &= \sum_{j=1}^{n} \left(\frac{1}{f(\mathbf{x}_j)} (2\pi)^{-\frac{d}{2}} \, |\boldsymbol{\Sigma}_i^{-1}|^{\frac{1}{2}} \exp\left\{ g(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \right\} P(C_i) \, \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_j - \boldsymbol{\mu}_i) \right) \\
&= \sum_{j=1}^{n} \left(\frac{f(\mathbf{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) P(C_i)}{f(\mathbf{x}_j)} \cdot \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_j - \boldsymbol{\mu}_i) \right) \\
&= \sum_{j=1}^{n} w_{ij} \, \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_j - \boldsymbol{\mu}_i)
\end{aligned}$$

$$w_{ij} = P(C_i|\mathbf{x}_j) = \frac{f(\mathbf{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) P(C_i)}{f(\mathbf{x}_j)}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Estimation of Mean

$$\sum_{j=1}^{n} w_{ij}(\mathbf{x}_j - \boldsymbol{\mu}_i) = \mathbf{0}, \text{ which implies that}$$

$$\sum_{j=1}^{n} w_{ij}\mathbf{x}_j = \boldsymbol{\mu}_i \sum_{j=1} w_{ij}, \text{ and therefore}$$

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{n} w_{ij}\mathbf{x}_j}{\sum_{j=1}^{n} w_{ij}}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Estimation of Covariance Matrix

$$\frac{\partial |\mathbf{\Sigma}_i^{-1}|^{\frac{1}{2}}}{\partial \mathbf{\Sigma}_i^{-1}} = \frac{1}{2} \cdot |\mathbf{\Sigma}_i^{-1}|^{-\frac{1}{2}} \cdot |\mathbf{\Sigma}_i^{-1}| \cdot \mathbf{\Sigma}_i = \frac{1}{2} \cdot |\mathbf{\Sigma}_i^{-1}|^{\frac{1}{2}} \cdot \mathbf{\Sigma}_i$$

$$\frac{\partial}{\partial \mathbf{\Sigma}_i^{-1}} \exp\left\{g(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)\right\} = -\frac{1}{2} \exp\left\{g(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)\right\} (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T$$

$$\begin{aligned}
&\frac{\partial}{\partial \mathbf{\Sigma}_i^{-1}} |\mathbf{\Sigma}_i^{-1}|^{\frac{1}{2}} \exp\left\{g(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)\right\} \\
&\quad = \frac{1}{2} |\mathbf{\Sigma}_i^{-1}|^{\frac{1}{2}} \mathbf{\Sigma}_i \exp\left\{g(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)\right\} - \frac{1}{2} |\mathbf{\Sigma}_i^{-1}|^{\frac{1}{2}} \exp\left\{g(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)\right\} (\mathbf{x}_j - \boldsymbol{\mu}_i)\ (\mathbf{x}_j - \boldsymbol{\mu}_i)^T \\
&\quad = \frac{1}{2} \cdot |\mathbf{\Sigma}_i^{-1}|^{\frac{1}{2}} \cdot \exp\left\{g(\boldsymbol{\mu}_i, \mathbf{\Sigma}_i)\right\} \left(\mathbf{\Sigma}_i - (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T\right)
\end{aligned}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Estimation of Covariance Matrix

$$= \frac{1}{2}\sum_{j=1}^{n}\frac{f(\mathbf{x}_j|\boldsymbol{\mu}_i,\boldsymbol{\Sigma}_i)P(C_i)}{f(\mathbf{x}_j)}\cdot\Big(\boldsymbol{\Sigma}_i-(\mathbf{x}_j-\boldsymbol{\mu}_i)(\mathbf{x}_j-\boldsymbol{\mu}_i)^T\Big)$$

$$= \frac{1}{2}\sum_{j=1}^{n} w_{ij}\Big(\boldsymbol{\Sigma}_i-(\mathbf{x}_j-\boldsymbol{\mu}_i)(\mathbf{x}_j-\boldsymbol{\mu}_i)^T\Big)$$

$$\sum_{j=1}^{n} w_{ij}\big(\boldsymbol{\Sigma}_i-(\mathbf{x}_j-\boldsymbol{\mu}_i)(\mathbf{x}_j-\boldsymbol{\mu}_i)^T\big)=\mathbf{0}_{d\times d}, \text{which implies that}$$

$$\boldsymbol{\Sigma}_i=\frac{\sum_{j=1}^{n} w_{ij}(\mathbf{x}_j-\boldsymbol{\mu}_i)(\mathbf{x}_j-\boldsymbol{\mu}_i)^T}{\sum_{j=1}^{n} w_{ij}}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

K-means as Specialization of EM

The End

Estimating the Prior Probability: Mixture Parameters

Mixture Parameters

$$\frac{\partial}{\partial P(C_i)}\left(\ln(P(\mathbf{D}|\theta)) + \alpha\left(\sum_{a=1}^{k} P(C_a) - 1\right)\right)$$

$$\frac{\partial}{\partial P(C_i)}\ln(P(\mathbf{D}|\theta)) = \sum_{j=1}^{n}\frac{f(\mathbf{x}_j|\mu_i,\mathbf{\Sigma}_i)}{f(\mathbf{x}_j)}$$

$$\left(\sum_{j=1}^{n}\frac{f(\mathbf{x}_j|\mu_i,\mathbf{\Sigma}_i)}{f(\mathbf{x}_j)}\right) + \alpha$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Mixture Parameters

$$\sum_{j=1}^{n} \frac{f(\mathbf{x}_j|\mu_i, \Sigma_i) P(C_i)}{f(\mathbf{x}_j)} = -\alpha P(C_i)$$

$$\sum_{j=1}^{n} w_{ij} = -\alpha P(C_i)$$

$$\sum_{i=1}^{k} \sum_{j=1}^{n} w_{ij} = -\alpha \sum_{i=1}^{k} P(C_i) \quad \text{or } n = -\alpha$$

$$P(C_i) = \frac{\sum_{j=1}^{n} w_{ij}}{n}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

EM Approach

$$P(\mathbf{C} = \mathbf{c}_j) = \prod_{i=1}^{k} P(C_i)^{c_{ji}} \qquad f(\mathbf{x}_j|\mathbf{c}_j) = \prod_{i=1}^{k} f(\mathbf{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)^{c_{ji}}$$

$$f(\mathbf{x}_j \text{ and } \mathbf{c}_j) = f(\mathbf{x}_j|\mathbf{c}_j)P(\mathbf{c}_j) = \prod_{i=1}^{k} \Big(f(\mathbf{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)P(C_i) \Big)^{c_{ji}}$$

$$\ln P(\mathbf{D}|\boldsymbol{\theta}) = \ln \prod_{j=1}^{n} f(\mathbf{x}_j \text{ and } \mathbf{c}_j|\boldsymbol{\theta})$$

$$= \sum_{j=1}^{n} \sum_{i=1}^{k} c_{ji} \Big(\ln f(\mathbf{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) + \ln P(C_i) \Big)$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Expectation Step

$$E[\ln P(\mathbf{D}|\theta)] = \sum_{j=1}^{n}\sum_{i=1}^{k} E[c_{ji}]\Big(\ln f(\mathbf{x}_j|\mu_i, \mathbf{\Sigma}_i) + \ln P(C_i)\Big)$$

$$\begin{aligned}
E[c_{ji}] &= 1 \cdot P(c_{ji}=1|\mathbf{x}_j) + 0 \cdot P(c_{ji}=0|\mathbf{x}_j) = P(c_{ji}=1|\mathbf{x}_j) = P(C_i|\mathbf{x}_j) \\
&= \frac{P(\mathbf{x}_j|C_i)P(C_i)}{P(\mathbf{x}_j)} = \frac{f(\mathbf{x}_j|\mu_i, \mathbf{\Sigma}_i)P(C_i)}{f(\mathbf{x}_j)} \\
&= w_{ij}
\end{aligned}$$

$$E[\ln P(\mathbf{D}|\theta)] = \sum_{j=1}^{n}\sum_{i=1}^{k} w_{ij}\Big(\ln f(\mathbf{x}_j|\mu_i, \mathbf{\Sigma}_i) + \ln P(C_i)\Big)$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Maximization Step

$$\frac{\partial}{\partial \boldsymbol{\mu}_i} \ln E[P(\mathbf{D}|\theta)] = \frac{\partial}{\partial \boldsymbol{\mu}_i} \sum_{j=1}^{n} w_{ij} \ln f(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

$$= \sum_{j=1}^{n} w_{ij} \, \boldsymbol{\Sigma}_i^{-1} (\mathbf{x}_j - \boldsymbol{\mu}_i)$$

$$\frac{\partial}{\partial \boldsymbol{\mu}_i} f(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = f(\mathbf{x}_j | \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \, \boldsymbol{\Sigma}_i^{-1} \, (\mathbf{x}_j - \boldsymbol{\mu}_i)$$

$$\boldsymbol{\mu}_i = \frac{\sum_{j=1}^{n} w_{ij} \mathbf{x}_j}{\sum_{j=1}^{n} w_{ij}}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Maximization Step

$$\frac{\partial}{\partial \boldsymbol{\Sigma}_i^{-1}} \ln E[P(\mathbf{D}|\theta)]$$

$$= \sum_{j=1}^{n} w_{ij} \cdot \frac{1}{f(\mathbf{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)} \cdot \frac{1}{2} f(\mathbf{x}_j|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\big(\boldsymbol{\Sigma}_i - (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T\big)$$

$$= \frac{1}{2} \sum_{j=1}^{n} w_{ij} \cdot \big(\boldsymbol{\Sigma}_i - (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T\big)$$

$$\boldsymbol{\Sigma}_i = \frac{\sum_{j=1}^{n} w_{ij} (\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^T}{\sum_{j=1}^{n} w_{ij}}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Maximization Step

$$\frac{\partial}{\partial P(C_i)}\left(\ln E[P(\mathbf{D}|\theta)] + \alpha\left(\sum_{i=1}^{k} P(C_i) - 1\right)\right) = \frac{\partial}{\partial P(C_i)}\left(w_{ij} \ln P(C_i) + \alpha P(C_i)\right)$$

$$= \left(\sum_{j=1}^{n} w_{ij} \cdot \frac{1}{P(C_i)}\right) + \alpha$$

$$\sum_{j=1}^{n} w_{ij} = -\alpha \cdot P(C_i)$$

$$P(C_i) = \frac{\sum_{j=1}^{n} w_{ij}}{n}$$

Zaki, A., Meira Jr., W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.

Estimating the Prior Probability: Mixture Parameters

The End