Mixture Models and EM

Chapter 9

Mixture Models

- Can be used to build more complex probability distribution from simple ones.
- Advantageous for clustering.
- Latent variables can be cased to the mixture models.
- Gaussian mixtures models are widely used in data mining, pattern recognition, machine learning and statistical analysis.

K-means Clustering

- Uses to identify clusters in multidimensional space.

- Aim: Partition the data set into some number K of clusters, where K is given.

- Note: A cluster comprises of a group of data points whose inter-point distances are small compared with the distances to points outside of the cluster. Each cluster is represented by a vector representing the centeroid of each cluster.

K-means Clustering

- Define an Objective function or Distortion Measure

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2$$

K-means Clustering

- Choose some initial values of $\boldsymbol{\mu}_k$
- First Phase:
  - Minimize J with respect to r_{nk} keeping the $\boldsymbol{\mu}_k$ fixed.
- Second Phase:
  - Minimize J with respect to $\boldsymbol{\mu}_k$ keeping the r_{nk} fixed.
- Iterate until convergence

K-means Clustering

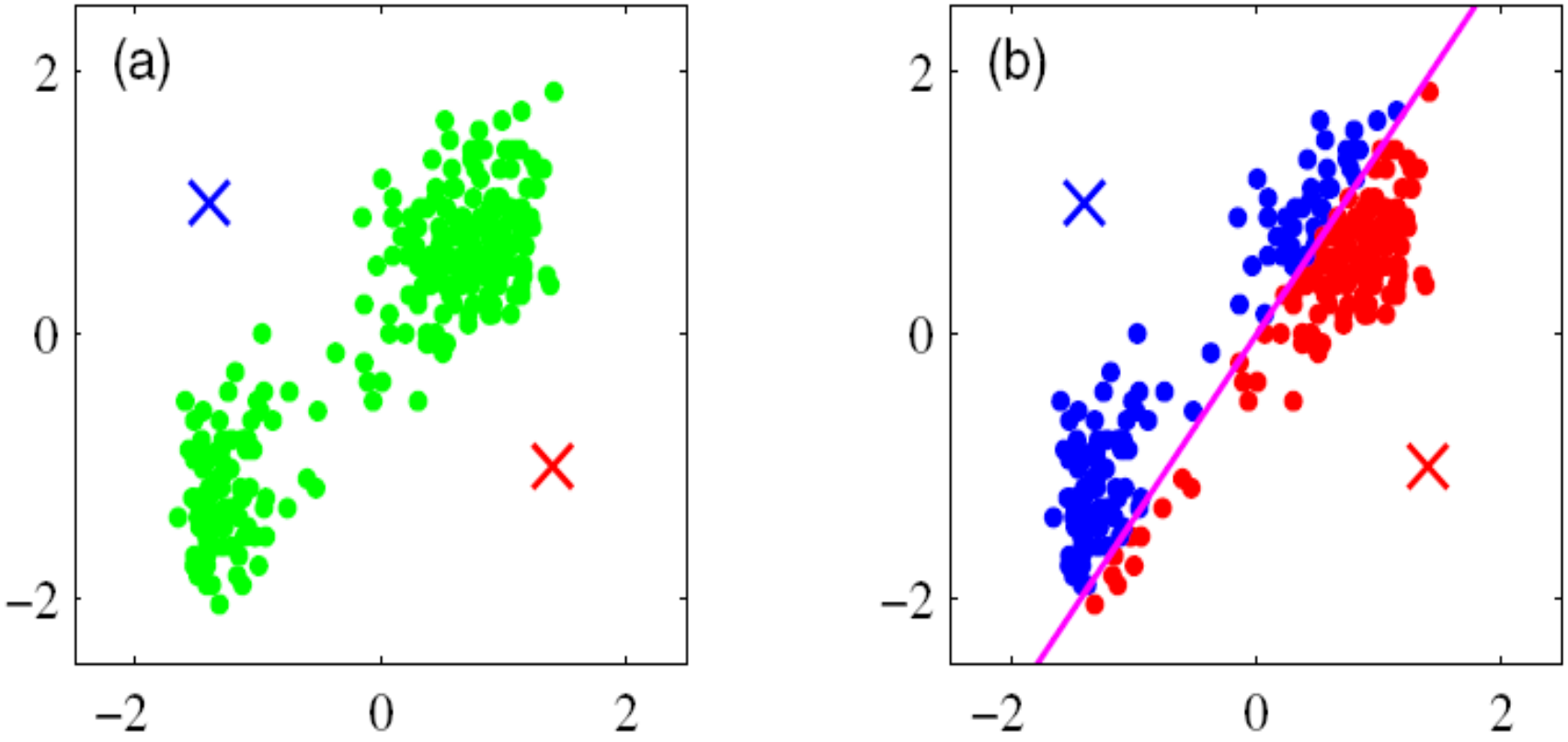
- While optimizing the above equations, we will get

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg\min_j \|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2 \\ 0 & \text{otherwise.} \end{cases}$$

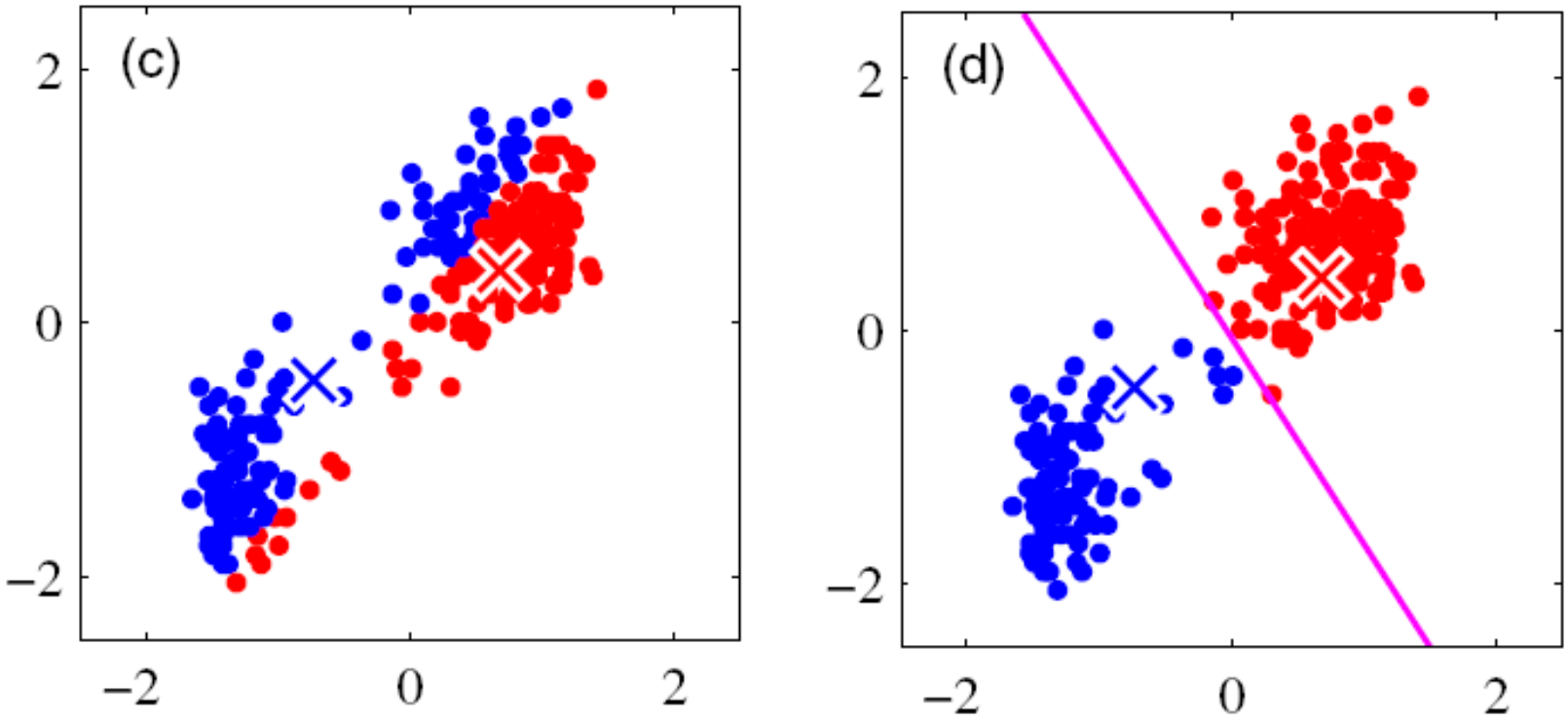
$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}.$$

- Note that the denominator of $\boldsymbol{\mu}_k$ is equal to the number of points assigned to cluster k.

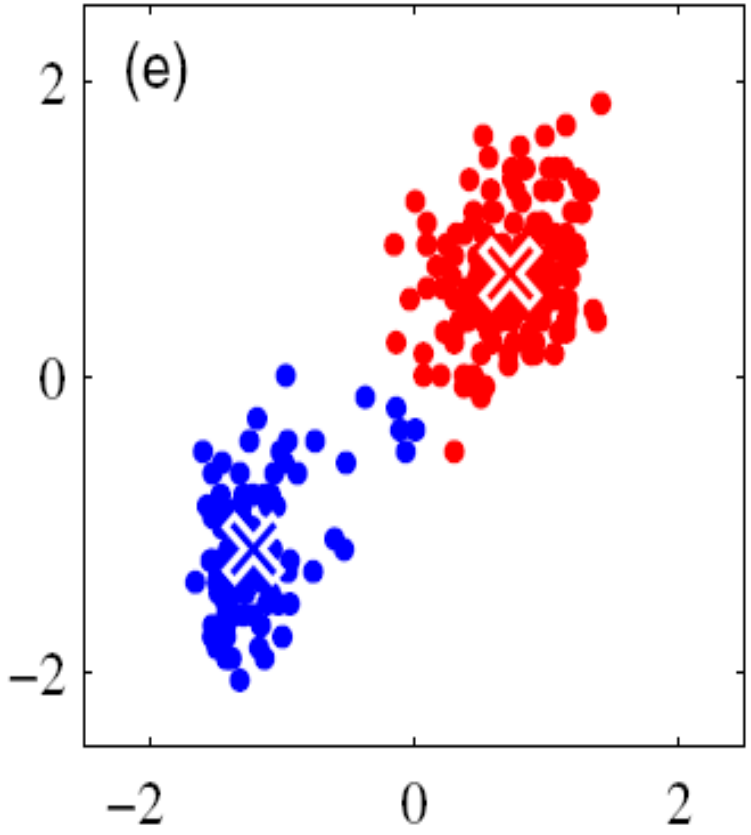
K-means Clustering

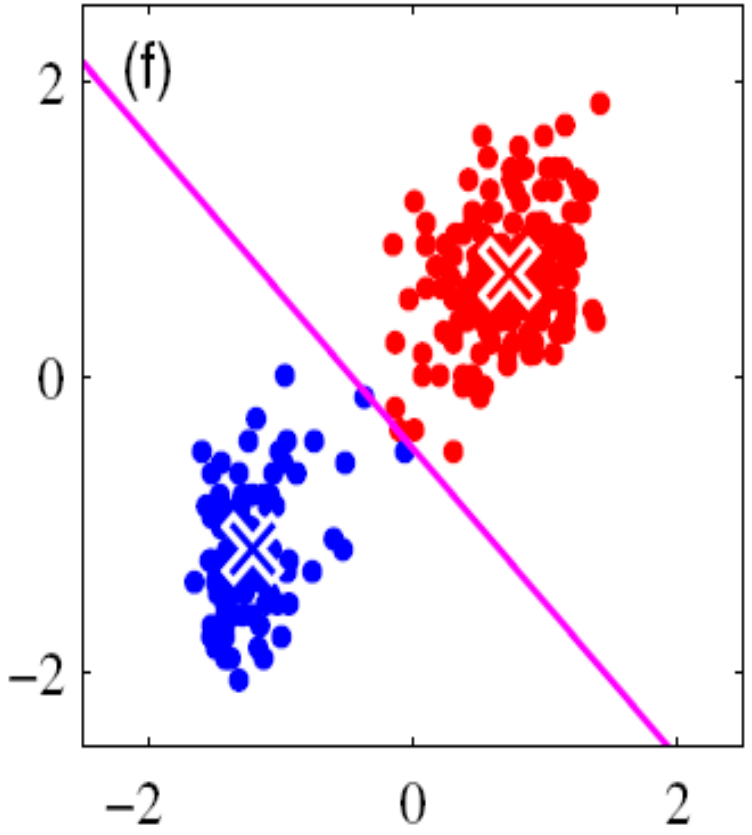


K-means Clustering

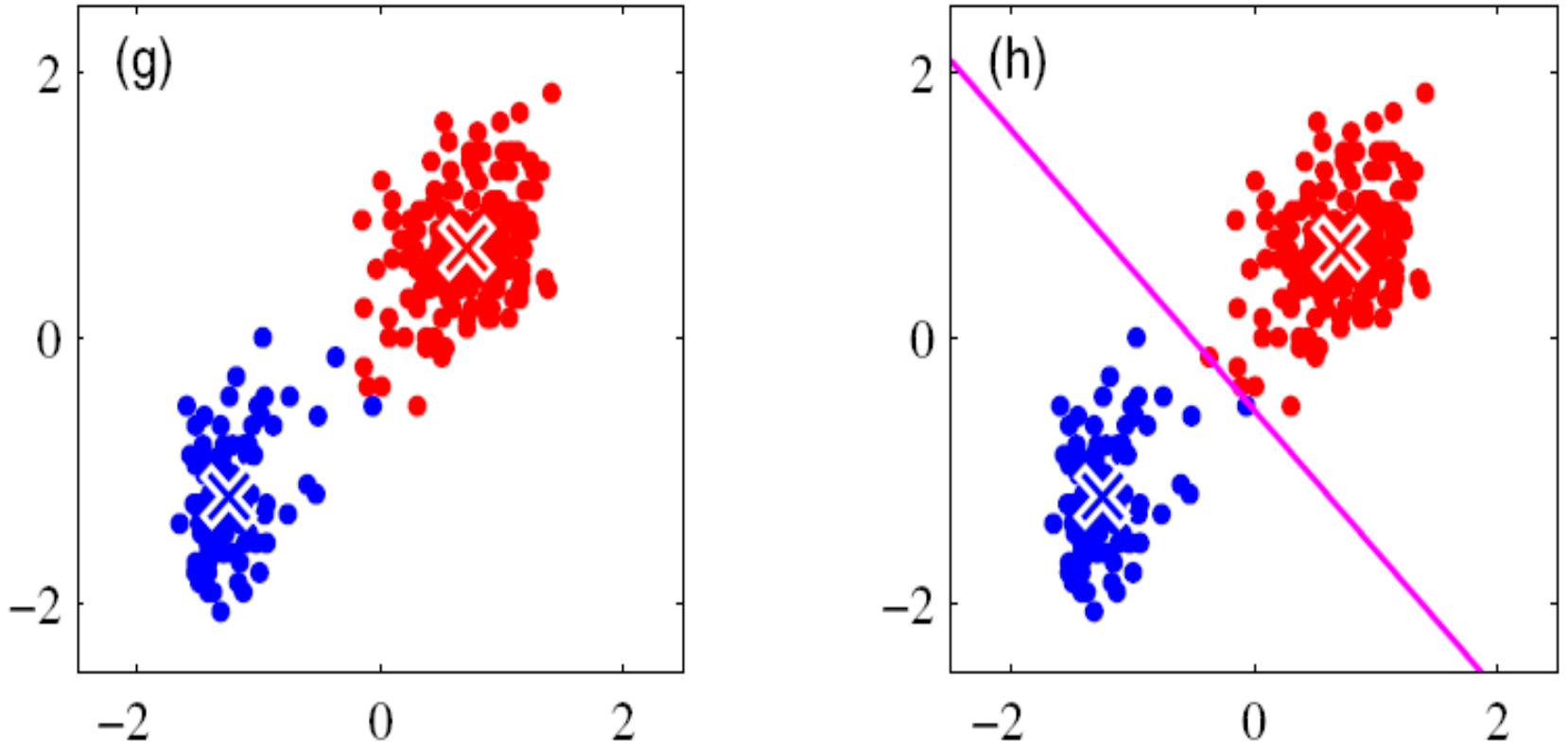


K-means Clustering





K-means Clustering



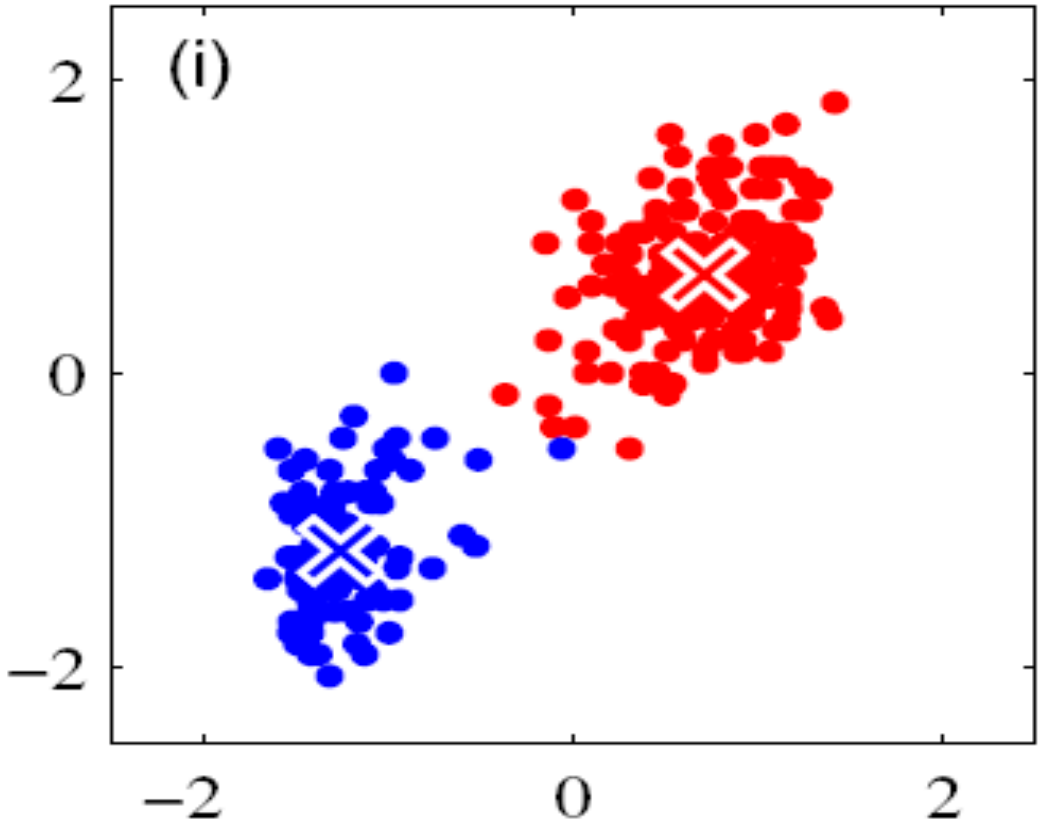
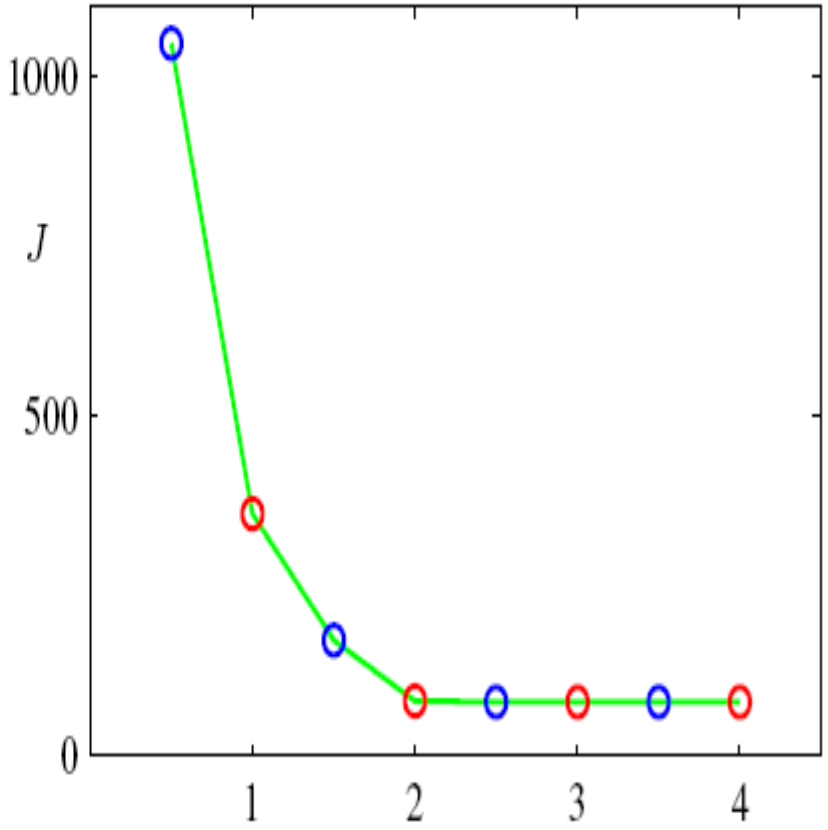


Figure 9.1 Illustration of the K-means algorithm using the re-scaled Old Faithful data set. (a) Green points denote the data set in a two-dimensional Euclidean space. The initial choices for centres μ_1 and μ_2 are shown by the red and blue crosses, respectively. (b) In the initial E step, each data point is assigned either to the red cluster or to the blue cluster, according to which cluster centre is nearer. This is equivalent to classifying the points according to which side of the perpendicular bisector of the two cluster centres, shown by the magenta line, they lie on. (c) In the subsequent M step, each cluster centre is re-computed to be the mean of the points assigned to the corresponding cluster. (d)–(i) show successive E and M steps through to final convergence of the algorithm.

K-means Clustering

Figure 9.2 Plot of the cost function J given by (9.1) after each E step (blue points) and M step (red points) of the K-means algorithm for the example shown in Figure 9.1. The algorithm has converged after the third M step, and the final EM cycle produces no changes in either the assignments or the prototype vectors.



K-means Clustering

- Note that the previous algorithm is a batch version of K-means and there is also a sequential version.

- The dissimilarity measure used previously is Euclidean distance and it has some limitations. Hence there are some modification of the dissimilarity measure.

- K-means is a hard clustering algorithm in which a point is assigned to one and only one cluster and it has limitations for points lying at equidistant between two clusters.

Image Segmentation and Compression

- Goal: Partition an image into regions each of which has a reasonably homogenous visual appearance or which corresponds to objects or parts of objects.

- Treat each pixel as a 3-D data point of the RGB intensities and apply K-means clustering. Finally replace each pixel with the corresponding RGB intensity of the cluster it is assigned to.

- Lossy Image compression can be performed by storing indexes of cluster centres and the cluster centres.

Image Segmentation and Compression

$K = 2$

$K = 3$

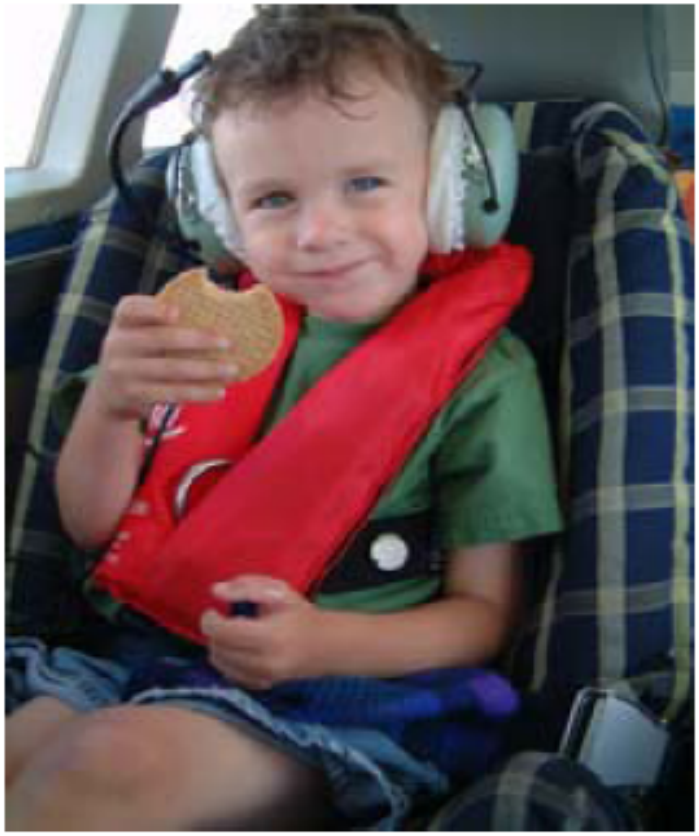
Image Segmentation and Compression

$K = 10$

Original image

Image Segmentation and Compression

Image Segmentation and Compression



Mixtures of Gaussian

- Gaussian Mixture Distribution can be given as

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

$$p(z_k = 1) = \pi_k$$

$$0 \leqslant \pi_k \leqslant 1$$

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

Mixtures of Gaussian

- Using 1-of-K coding scheme for z

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}.$$

- Define conditional distribution of $\mathbf{x}$ given a particular $\mathbf{z}$

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}.$$

- Define Joint Distribution over all possible states of $\mathbf{z}$

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Figure 9.4 Graphical representation of a mixture model, in which the joint distribution is expressed in the form $p(\mathbf{x}, \mathbf{z}) = p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$.

Mixtures of Gaussian

- Define the conditional probability of *z* given *x*.

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)} = \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}.$$

- Note that π_k is the prior probability of $z_k = 1$ and the quantity $\gamma(z_k)$ as the corresponding posterior probability once we observed *x*.

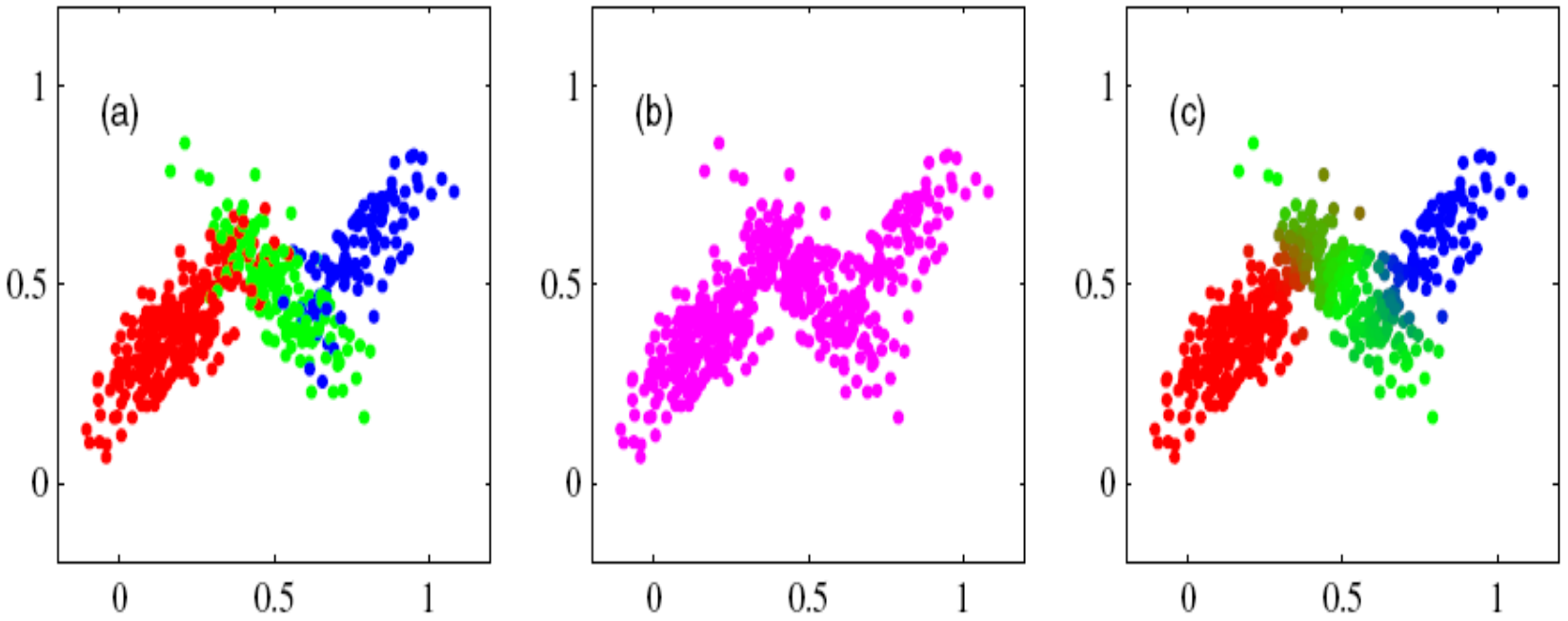


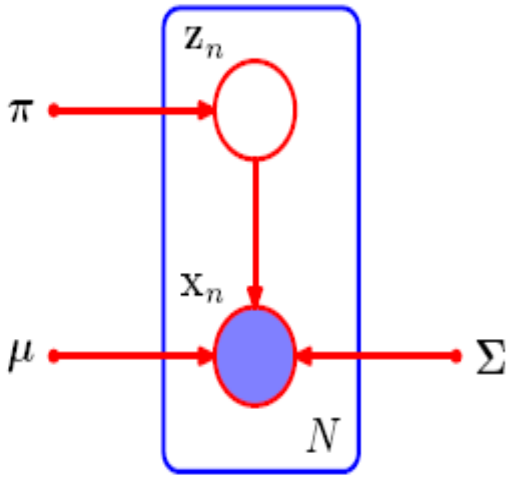
Figure 9.5 Example of 500 points drawn from the mixture of 3 Gaussians shown in Figure 2.23. (a) Samples from the joint distribution $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ in which the three states of $\mathbf{z}$, corresponding to the three components of the mixture, are depicted in red, green, and blue, and (b) the corresponding samples from the marginal distribution $p(\mathbf{x})$, which is obtained by simply ignoring the values of $\mathbf{z}$ and just plotting the $\mathbf{x}$ values. The data set in (a) is said to be *complete*, whereas that in (b) is *incomplete*. (c) The same samples in which the colours represent the value of the responsibilities $\gamma(z_{nk})$ associated with data point $\mathbf{x}_n$, obtained by plotting the corresponding point using proportions of red, blue, and green ink given by $\gamma(z_{nk})$ for $k = 1, 2, 3$, respectively

Mixtures of Gaussian:
Maximum Likelihood

- Taking data points drawn independently from a mixture of Gaussians, the log likelihood is given by

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}.$$

Figure 9.6 Graphical representation of a Gaussian mixture model for a set of N i.i.d. data points $\{\mathbf{x}_n\}$, with corresponding latent points $\{\mathbf{z}_n\}$, where $n = 1, \ldots, N$.



Mixtures of Gaussian:
Maximum Likelihood

- Consider a diagonal covariance matrices for simplicity and consider the case $\boldsymbol{\mu}_j = \mathbf{x}_n$

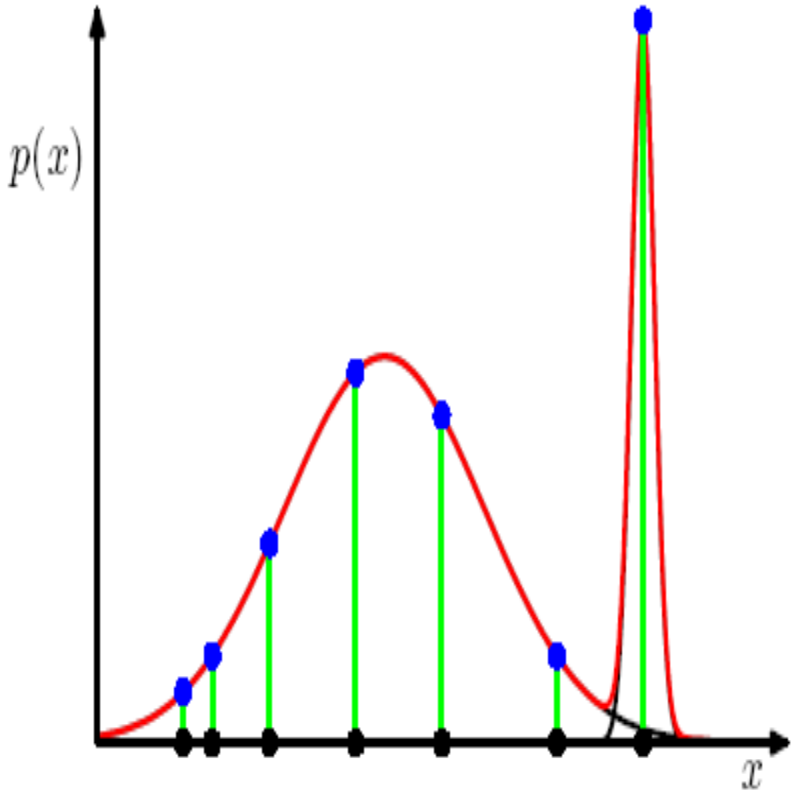
$$\mathcal{N}(\mathbf{x}_n|\mathbf{x}_n, \sigma_j^2\mathbf{I}) = \frac{1}{(2\pi)^{1/2}}\frac{1}{\sigma_j}$$

- If $\sigma_j \to 0$ then $\mathcal{N}(\mathbf{x}_n|\mathbf{x}_n, \sigma_j^2\mathbf{I}) \to \infty$ and $\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \to \infty$
- Note that this problem does not occur in the case of single Gaussian function where the overall likelihood goes to zero rather than infinity.

- Maximum likelihood also gives many identical distribution.

Mixtures of Gaussian:
Maximum Likelihood

Figure 9.7 Illustration of how singularities in the likelihood function arise with mixtures of Gaussians. This should be compared with the case of a single Gaussian shown in Figure 1.14 for which no singularities arise.



EM for Gaussian Mixtures

- Expectation-Maximization: powerful method for finding maximum likelihood models with latent variables.

- Differentiate with respect to $\boldsymbol{\mu}_k$

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}}_{\gamma(z_{nk})} \boldsymbol{\Sigma}_k(\mathbf{x}_n - \boldsymbol{\mu}_k)$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum_{n=1}^{N} \gamma(z_{nk})\mathbf{x}_n \qquad N_k = \sum_{n=1}^{N} \gamma(z_{nk}).$$

EM for Gaussian Mixtures

- Differentiate with respect to $\boldsymbol{\Sigma}_k$

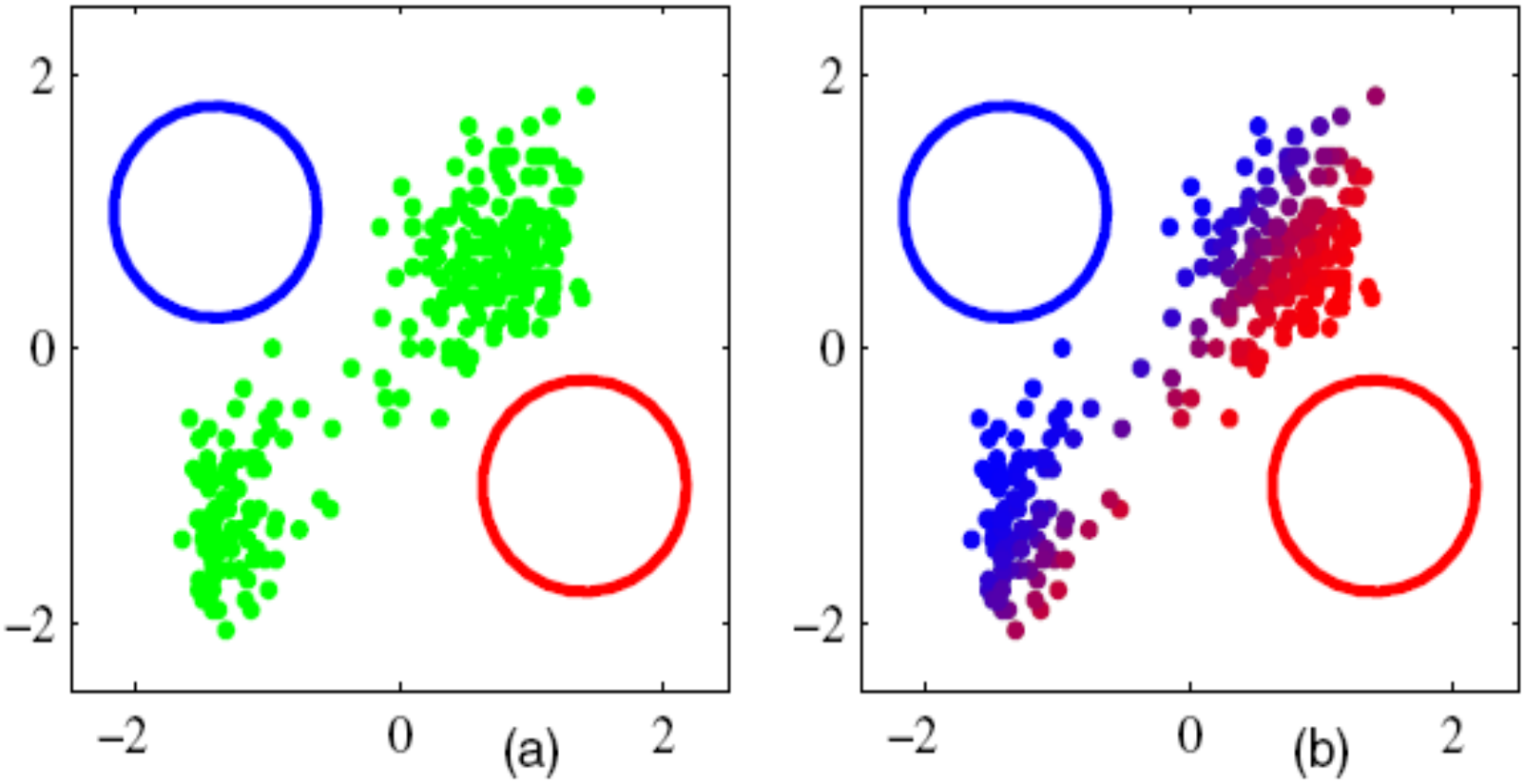
$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^{\mathrm{T}}$$

- Use Lagrange multiplier in order to differentiate with respect to π_k

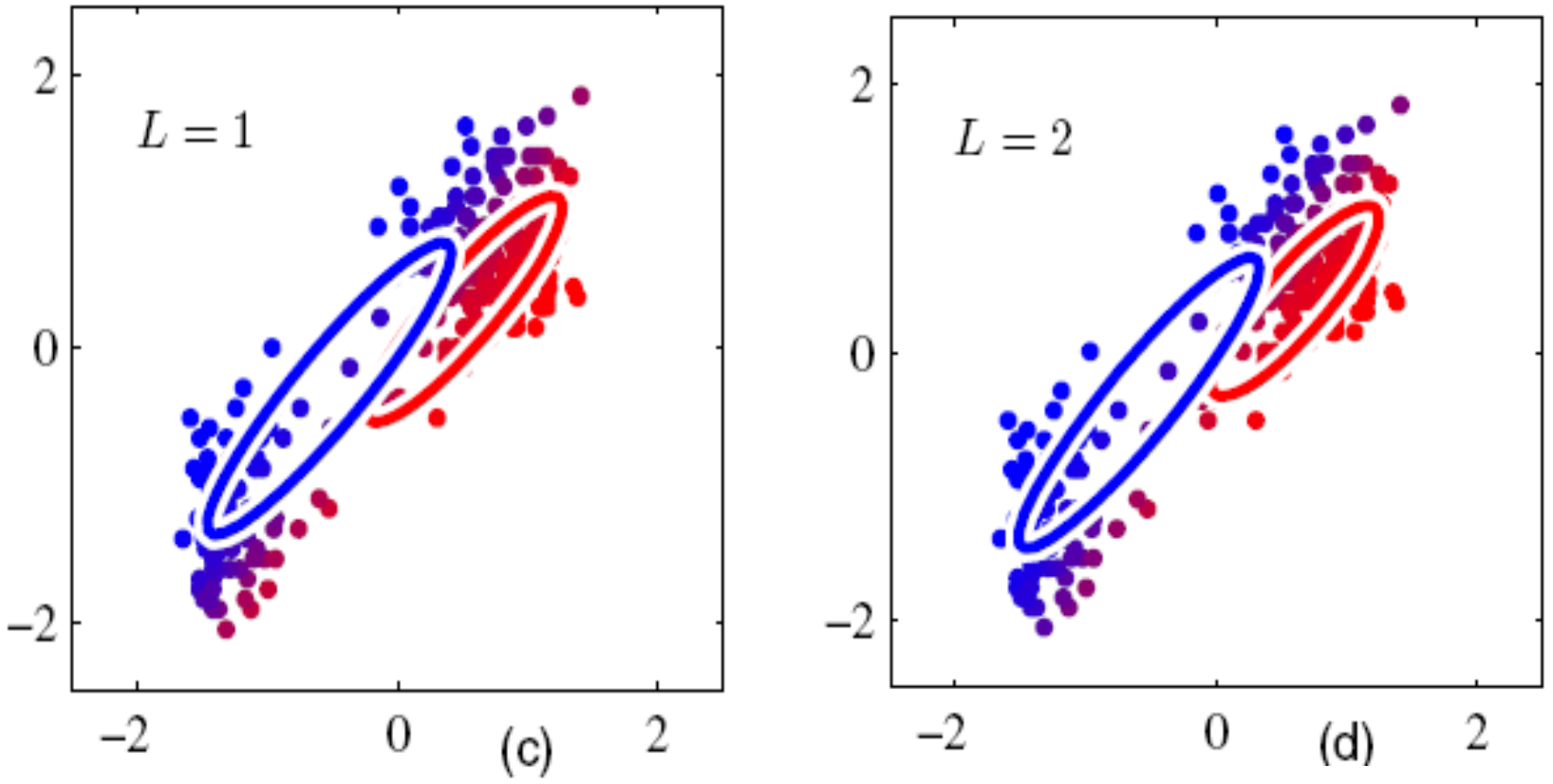
$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) + \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right)$$

$$\pi_k = \frac{N_k}{N}$$

EM for Gaussian Mixtures



EM for Gaussian Mixtures



EM for Gaussian Mixtures

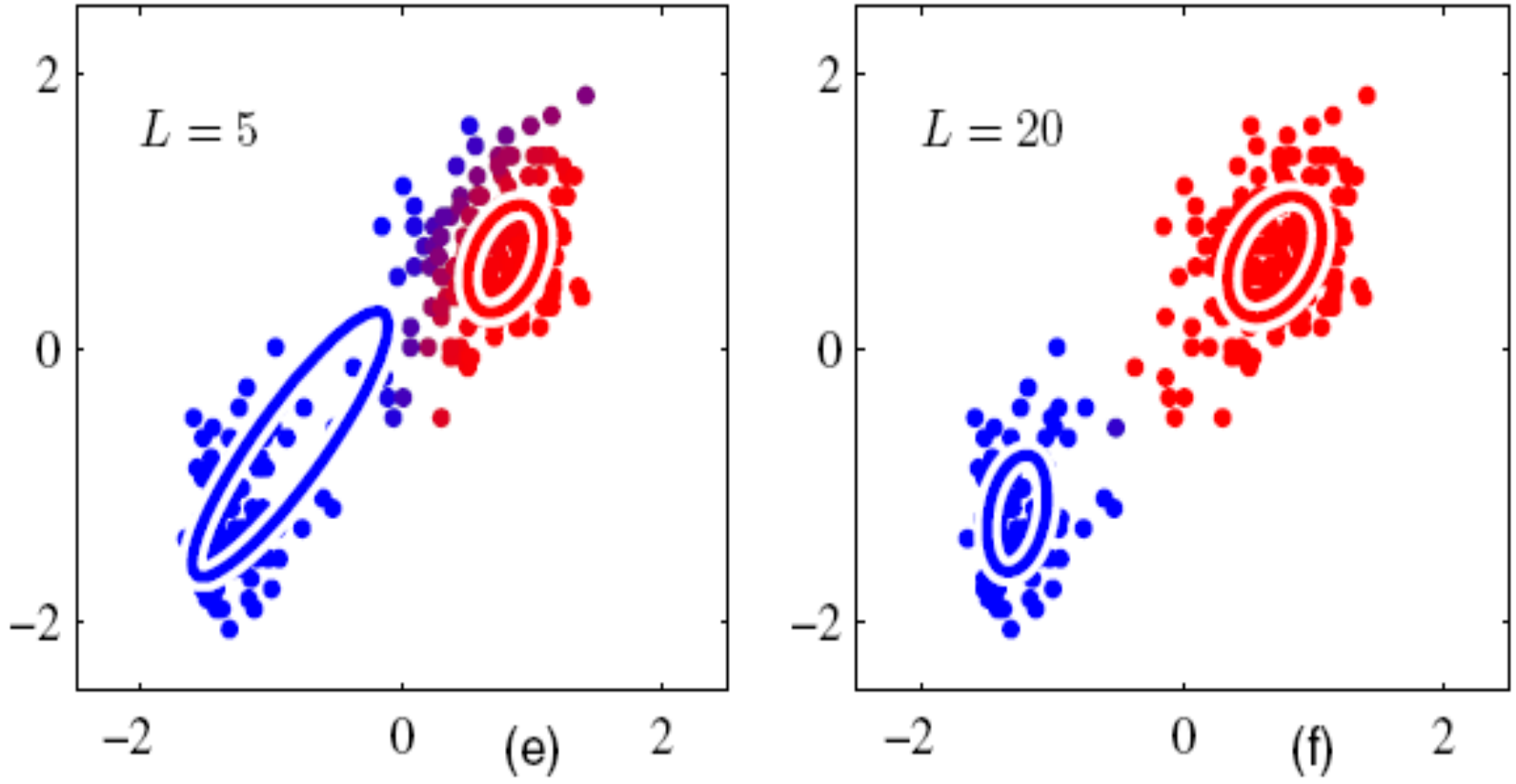


Figure 9.8 Illustration of the EM algorithm using the Old Faithful set as used for the illustration of the K-means algorithm in Figure 9.1. See the text for details.

EM for Gaussian Mixtures

EM for Gaussian Mixtures

Given a Gaussian mixture model, the goal is to maximize the likelihood function with respect to the parameters (comprising the means and covariances of the components and the mixing coefficients).

1. Initialize the means $\boldsymbol{\mu}_k$, covariances $\boldsymbol{\Sigma}_k$ and mixing coefficients π_k, and evaluate the initial value of the log likelihood.
2. **E step**. Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}. \tag{9.23}$$

EM for Gaussian Mixtures

3. **M step**. Re-estimate the parameters using the current responsibilities

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\mathbf{x}_n \tag{9.24}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})\,(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}})\,(\mathbf{x}_n-\boldsymbol{\mu}_k^{\text{new}})^{\mathrm{T}} \tag{9.25}$$

$$\pi_k^{\text{new}} = \frac{N_k}{N} \tag{9.26}$$

where

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk}). \tag{9.27}$$

4. Evaluate the log likelihood

$$\ln p(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\pi}) = \sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right\} \tag{9.28}$$

and check for convergence of either the parameters or the log likelihood. If the convergence criterion is not satisfied return to step 2.

General EM Algorithm

The General EM Algorithm

Given a joint distribution $p(\mathbf{X}, \mathbf{Z}|\theta)$ over observed variables $\mathbf{X}$ and latent variables $\mathbf{Z}$, governed by parameters θ, the goal is to maximize the likelihood function $p(\mathbf{X}|\theta)$ with respect to θ.

1. Choose an initial setting for the parameters θ^{old}.
2. **E step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$.

General EM Algorithm

3. **M step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \tag{9.32}$$

where

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}). \tag{9.33}$$

4. Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}} \tag{9.34}$$

and return to step 2.

Relation to K-means

- For identical covariance

$$p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi\epsilon)^{1/2}} \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x} - \boldsymbol{\mu}_k\|^2\right\}.$$

$$\gamma(z_{nk}) = \frac{\pi_k \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2/2\epsilon\right\}}{\sum_j \pi_j \exp\left\{-\|\mathbf{x}_n - \boldsymbol{\mu}_j\|^2/2\epsilon\right\}}.$$

- Expected complete-data log likelihood

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] \rightarrow -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk}\|\mathbf{x}_n - \boldsymbol{\mu}_k\|^2 + \text{const.}$$