

# Mixture Models and EM

Liang Sun  
sun.liang@asu.edu  
Arizona State University

August 27, 2007

- 1 K-means Clustering
- 2 Mixtures of Gaussians
  - Maximum Likelihood
  - EM for Gaussian Mixtures
- 3 An Alternative View of EM
  - Gaussian Mixtures Revisited
  - Relation to K-means
  - Mixtures of Bernoulli Distributions
- 4 The EM Algorithm in General
- 5 Summary

# Introduction to K-means – I

- Problem Statement

Given a set of data points  $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ , where  $\mathbf{x}_j \in \mathbb{R}^D$  and the cluster number  $K$ , the goal is:

- An assignment of data points to clusters
  - A set of vectors  $\{\mu_k\}$ , where  $\mu_k$  is the prototype of the  $k$ -th cluster, and the sum of the squares of the distances of each data point to its closest vector  $\mu_k$  is a minimum
- 1-of- $K$  coding scheme to describe assignment of data points to clusters

$$r_{nk} = \begin{cases} 1 & \text{if } \mathbf{x}_n \in \mathcal{C}_k \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

- Properties of 1-of- $K$  coding scheme

$$\sum_{k=1}^K r_{nk} = 1 \text{ and } r_{nk} \in \{0, 1\}$$

# Introduction to K-means – II

- The objective function in K-means, i.e., *distortion measure*:

$$J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \|\mathbf{x}_n - \mu_k\|^2 \quad (2)$$

- By minimizing the distortion measure, we can find the values of  $\{r_{nk}\}$  and  $\{\mu_k\}$
- Procedures of K-means
  - ① Choose some initial values for  $\mu_k$
  - ② Repeat until convergence
    - ① Keeping  $\{\mu_k\}$  fixed, and minimize  $J$  w.r.t.  $r_{nk}$
    - ② Keeping  $\{r_{nk}\}$  fixed, and minimize  $J$  w.r.t.  $\mu_k$

# Optimization Process

- When  $\{\mu_k\}$  is fixed,  $J$  is a linear function of  $r_{nk}$ , and  $J$  is minimized if

$$r_{nk} = \begin{cases} 1 & \text{if } k = \arg \min_j \|\mathbf{x}_n - \mu_j\|^2 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

- When  $\{r_{nk}\}$  is fixed,  $J$  is a quadratic function of  $\mu_k$

$$J = \sum_{n=1}^N \sum_{k=1}^K r_{nk} (\mu_k^T \mu_k - 2\mu_k^T \mathbf{x}_n + \mathbf{x}_n^T \mathbf{x}_n)$$

$$\Rightarrow \frac{\partial J}{\partial \mu_k} = 2 \sum_{n=1}^N r_{nk} (\mu_k - \mathbf{x}_n)$$

$$\text{Therefore, } \frac{\partial J}{\partial \mu_k} = 0 \Leftrightarrow \mu_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}} \quad (4)$$

- $\mu_k$  in (4) equals to the mean of all data points  $\mathbf{x}_n$  assigned to cluster  $k$ , and this is the reason why this algorithm is called **K-means** algorithm

# Online Stochastic Algorithm

- Two different kinds of algorithm
  - Batch algorithm
  - Online algorithm
- Key idea is to apply **Robbins-Monro** procedure to find the roots of the regression function (4), i.e., to find  $\mu_k$
- The updating scheme of  $\mu_k$  in online algorithm is

$$\mu_k^{\text{new}} = \mu_k^{\text{old}} + \eta_n(\mathbf{x}_n - \mu_k^{\text{old}}) \quad (5)$$

- $\eta_n$  is the learning rate parameter, and is typically made to decrease monotonically as more data points are considered

# K-medoids Algorithm – I

- Drawbacks of K-means algorithm
  - It cannot process categorical attributes
  - The determination of the cluster means is *nonrobust* to outliers
  - It may converge to a local rather than global minimum of  $J$
- K-medoids algorithm can overcome the first 2 drawbacks
- A more general dissimilarity measure  $\mathcal{V}(x, x')$  is used in K-medoids algorithm
- The new distortion measure is given

$$\tilde{J} = \sum_{n=1}^N \sum_{k=1}^K r_{nk} \mathcal{V}(\mathbf{x}_n, \mu_k) \quad (6)$$

# K-medoids Algorithm – II

- Procedures of K-medoids algorithm

- ① Choose some initial values for  $\mu_k$

- ② Repeat until convergence

- ① Keeping  $\{\mu_k\}$  fixed, assign each data point to the cluster for which the dissimilarity to the corresponding  $\mu_k$  is smallest

- ② Keeping  $\{r_{nk}\}$  fixed, and find the optimal  $\{\mu_k\}$

To simplify computation, we restrict each cluster prototype to be one of the data points

# Applications of K-means Algorithm – Image Segmentation

The goal of image segmentation is to partition an image into regions each of which has a reasonably homogeneous visual appearance or which corresponds to objects or parts of objects

$K = 2$



$K = 3$



$K = 10$



Original image



# Applications of K-means Algorithm – Vector Quantization

- *Vector Quantization* is used in *lossy data compression*
- For each of the  $N$  data points, we store only the identity  $k$  of the cluster to which it is assigned. We also store the values of the  $K$  cluster centres  $\mu_k$
- Each data point is then approximated by its nearest centre  $\mu_k$

- 1 K-means Clustering
- 2 Mixtures of Gaussians
  - Maximum Likelihood
  - EM for Gaussian Mixtures
- 3 An Alternative View of EM
  - Gaussian Mixtures Revisited
  - Relation to K-means
  - Mixtures of Bernoulli Distributions
- 4 The EM Algorithm in General
- 5 Summary

# Basis of Gaussian Distribution

- Gaussian distribution, or *normal distribution*, is a widely used model for the distribution of continuous variables
- Gaussian distribution in 1-dimensional space

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x - \mu)^2\right\} \quad (7)$$

where  $\mu$  is the mean, and  $\sigma^2$  is the variance

- Gaussian distribution in  $D$ -dimensional space

$$\mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right\} \quad (8)$$

where  $\mu$  is a  $D$ -dimensional mean vector,  $\Sigma$  is a  $D \times D$  covariance matrix, and  $|\Sigma|$  is the determinant of  $\Sigma$

# Mixture of Gaussian – I

- Gaussian Mixture Distribution is a linear superposition of Gaussians in the following form:

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \quad (9)$$

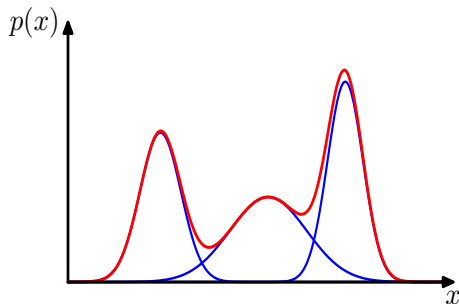
- Each Gaussian density  $\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)$  is called a *component* of the mixture
- $\pi_k$  is called *mixing coefficient*
- If we integrate both sides of (9) w.r.t.  $\mathbf{x}$ , we can get

$$\sum_{k=1}^K \pi_k = 1 \quad (10)$$

- $p(\mathbf{x}) \geq 0$ , and  $\mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \geq 0$  ensure that  $\pi_k \geq 0$  for all  $k$

# Mixture of Gaussian – II

- By using a sufficient number of Gaussians, and by adjusting their means and covariances as well as the coefficients in the linear combination, almost any continuous density can be approximated to arbitrary accuracy



# Mixture of Gaussian – From the Perspective of Latent Variables – I

- The mixtures of Gaussians can be described by introducing *latent variables*
- We also use 1-of- $K$  coding scheme to represent the latent variable  $z$ , and  $\mathbf{z}^T = [z_1, \dots, z_K]$

$$z_k \in \{0, 1\} \quad \text{and} \quad \sum_{k=1}^K z_k = 1 \quad (11)$$

- We can define the joint distribution  $p(\mathbf{x}, \mathbf{z})$  in terms of a marginal distribution  $p(\mathbf{z})$  and a conditional distribution  $p(\mathbf{x}|\mathbf{z})$

# Mixture of Gaussian – From the Perspective of Latent Variables – II

- The marginal distribution over  $\mathbf{z}$  is specified in terms of the mixing coefficients  $\pi_k$

$$p(z_k = 1) = \pi_k \quad (12)$$

- The parameters  $\{\pi_k\}$  must satisfy

$$0 \leq \pi_k \leq 1 \quad \text{and} \quad \sum_{k=1}^K \pi_k = 1 \quad (13)$$

- Like multinomial distribution,  $p(\mathbf{z})$  can be represented as:

$$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}, \quad \text{where } \mathbf{z}^T = [z_1, \dots, z_K] \quad (14)$$

# Mixture of Gaussian – From the Perspective of Latent Variables – III

- The conditional distribution of  $\mathbf{x}$  given a particular value for  $\mathbf{z}$  is a Gaussian

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k) \quad (15)$$

- This conditional distribution can be rewritten as

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)^{z_k} \quad (16)$$

- The joint distribution  $p(\mathbf{x}, \mathbf{z})$  is

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}) &= p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k} \prod_{k=1}^K \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k)^{z_k} \\ &= \prod_{k=1}^K (\pi_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k))^{z_k} \end{aligned}$$

# Mixture of Gaussian – From the Perspective of Latent Variables – IV

- The marginal distribution of  $\mathbf{x}$  is

$$\begin{aligned} p(\mathbf{x}) &= \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) \\ &= \sum_{\mathbf{z}} \prod_{k=1}^K (\pi_k \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k))^{z_k} \\ &= \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k) \end{aligned} \quad (17)$$

- The marginal distribution of  $\mathbf{x}$  is the same as the definition of Gaussian mixtures given by (9)
- If we have several observations  $\mathbf{x}_1, \dots, \mathbf{x}_N$ , it follows that for every observed data point  $\mathbf{x}_n$  there is a corresponding latent variable  $\mathbf{z}_n$  because we have represented the marginal distribution in the form  $p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$

# Mixture of Gaussian – From the Perspective of Latent Variables – V

- The conditional probability of  $\mathbf{z}$  given  $\mathbf{x}$  is denoted as  $\gamma(z_k)$

$$\begin{aligned}\gamma(z_k) &\equiv p(z_k = 1 | \mathbf{x}) \\ &= \frac{p(z_k = 1)p(\mathbf{x} | z_k = 1)}{\sum_{j=1}^K p(z_j = 1)p(\mathbf{x} | z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x} | \mu_j, \Sigma_j)}\end{aligned}$$

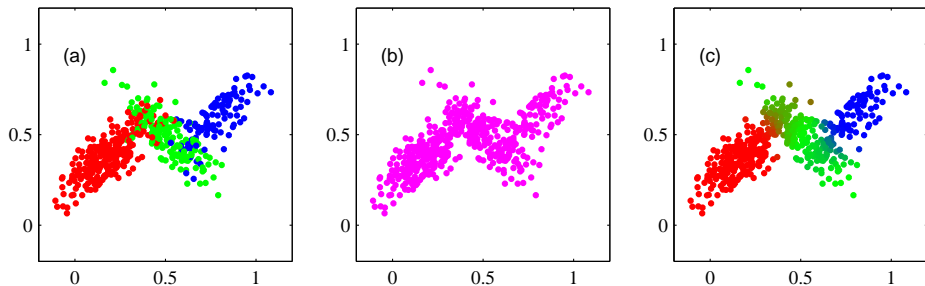
- $\pi_k$  can be considered as the prior probability of  $z_k = 1$ , and  $\gamma(z_k)$  can be considered as the posterior probability after observing  $\mathbf{x}$
- $\gamma(z_k)$  can also be viewed as the **responsibility** that component  $k$  takes for 'explaining' the observation  $\mathbf{x}$

# Mixture of Gaussian – From the Perspective of Latent Variables – VI – Sampling

- We can use the technique of ancestral sampling to generate random samples distributed according to the Gaussian mixture model
- Procedure of sampling
  - First generate a value for  $\mathbf{z}$  denote  $\hat{\mathbf{z}}$  from the marginal distribution  $p(\mathbf{z})$
  - Then generate a value for  $\mathbf{x}$  from the conditional distribution  $p(\mathbf{x}|\hat{\mathbf{z}})$
- Techniques for sampling will be discussed in Chapter 11 in detail

# Mixture of Gaussian – From the Perspective of Latent Variables – VII – Sampling

Example of 500 points drawn from the mixture of 3 Gaussians



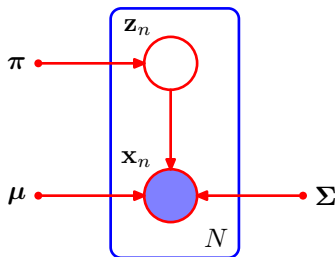
# Maximum Likelihood To Estimate Parameters of Gaussian Mixtures

- Given a data set of observations  $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ , and these  $N$  points are drawn independently from the distribution
- Here we wish to model this data using a mixture of Gaussian
- The question is how to estimate parameters  $\{\pi_k\}$ ,  $\{\mu_k\}$  and  $\{\Sigma_k\}$  for the data set
- The log of the likelihood function

$$\ln p(\mathbf{X}|\mu, \Sigma, \pi) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k) \right\} \quad (18)$$

# Graphical Representation of a Gaussian Mixture Model using Latent Variables

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  $n = 1, \dots, N$



# Limitations of ML – Singularities

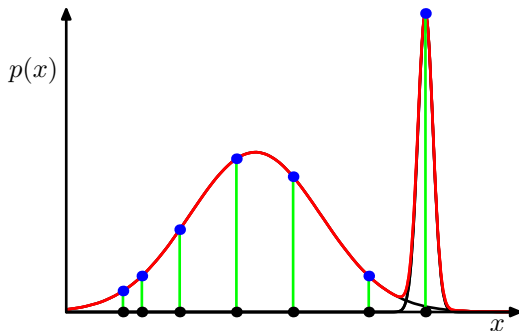
- For simplicity, suppose  $\Sigma_k = \sigma_k^2 \mathbf{I}$ , and  $\mu_j = \mathbf{x}_n$  for some value of  $n$
- Point  $\mathbf{x}_n$  will contribute a term in the likelihood function of the form

$$\mathcal{N}(\mathbf{x}_n | \mathbf{x}_n, \sigma_k^2 \mathbf{I}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{\sigma_j} \quad (19)$$

- $\sigma_j \rightarrow 0$ , then (19)  $\rightarrow +\infty$
- It means that maximizing the log likelihood function is an ill-posed problem
- The singularities will always be presented and will occur whenever one of the Gaussian components ‘collapses’ onto a specific data point
- But this problem does not arise in the case of a single Gaussian
  - $\sigma_j \rightarrow 0$ , then  $\mathcal{N}(\mathbf{x}_n | \mu, \Sigma) \rightarrow \infty$ , but  $\mathcal{N}(\mathbf{x}_j | \mu, \Sigma) \rightarrow 0$  for  $j \neq n$ , thus the log likelihood function will go to 0 rather than  $+\infty$

# Limitations of ML – Singularities

One of the components can have a finite variance and therefore assign finite probability to all of the data points while the other component can shrink onto one specific data point and thereby contribute an ever increasing additive value to the log likelihood.



# Limitations of ML

- In applying maximum likelihood to Gaussian mixture models we must take steps to avoid finding such pathological solutions caused by singularities and instead seek local maxima of the likelihood function that are well behaved
- Another limitation of ML: **Identifiability**
  - A  $K$ -component mixture will have a total of  $K!$  equivalent solutions corresponding to the  $K!$  ways of assigning  $K$  sets of parameters to  $K$  components
  - This is an important issue when we wish to interpret the parameter values discovered by a model
  - But if we are only interested in finding a good density model, it does not matter
- In practice, maximizing the log likelihood function is difficult because the presence of the summation over  $k$  that appears inside the log, so we cannot get the closed form solution

# EM for Gaussian Mixtures

- Expectation-Maximization algorithm can be used to find maximum likelihood solutions for models with latent variables
- To introduce EM algorithm, first we investigate what conditions must be satisfied at a maximum of the likelihood function
- In other words, we need to compute the derivatives of the log likelihood function w.r.t.  $\{\mu_k\}$ ,  $\{\Sigma_k\}$  and  $\{\pi_k\}$
- The log of the likelihood function

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

- And

$$\mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2} |\Sigma_k|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_n - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_n - \mu_k) \right\}$$

# Compute Derivative w.r.t. $\mu_k$

$$\begin{aligned} & \frac{\partial \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})}{\partial \mu_k} \\ = & \sum_{n=1}^N \frac{\frac{\partial(\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \boldsymbol{\Sigma}_k))}{\partial \mu_k}}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n|\mu_j, \boldsymbol{\Sigma}_j)} \\ = & \sum_{n=1}^N \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n|\mu_j, \boldsymbol{\Sigma}_j)} \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \mu_k) \\ = & \sum_{n=1}^N \gamma(z_{nk}) \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \mu_k) \end{aligned} \quad (20)$$

where 
$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n|\mu_j, \boldsymbol{\Sigma}_j)} \quad (21)$$

# Compute Derivative w.r.t. $\mu_k$

Setting the derivative of  $\ln p(\mathbf{X}|\mu, \Sigma, \pi)$  w.r.t.  $\mu_k$  to 0, we can get

$$\begin{aligned} & \sum_{n=1}^N \gamma(z_{nk}) \Sigma_k^{-1} (-\mathbf{x}_n + \mu_k) = 0 \\ \Leftrightarrow & \sum_{n=1}^N \gamma(z_{nk}) (-\mathbf{x}_n + \mu_k) = 0 \\ \Leftrightarrow & \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n = \sum_{n=1}^N \gamma(z_{nk}) \mu_k \\ \Leftrightarrow & \mu_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n \end{aligned} \quad (22)$$

$$\text{where } N_k = \sum_{n=1}^N \gamma(z_{nk}) \quad (23)$$

# Compute Derivative w.r.t. $\mu_k$

- $N_k$  can be considered as the **effective** number of points assigned to cluster  $k$
- $\mu_k$  is a weighted mean of all of the points in the data set, in which the weighting factor for data point  $\mathbf{x}_n$  is given by the posterior probability  $\gamma(z_{nk})$  that component  $k$  was responsible for generating  $\mathbf{x}_n$

# Compute Derivative w.r.t. $\Sigma_k^{-1}$

- First we compute the derivative of  $\pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)$  w.r.t.  $\Sigma_k^{-1}$
- Two useful formulas

$$\frac{d|M|}{dM} = |M|(M^{-1})^T$$
$$\frac{\partial(a^T M b)}{dM} = ab^T$$

- Denote  $c = \exp\left\{-\frac{1}{2}(\mathbf{x}_n - \mu_k)^T \Sigma_k^{-1}(\mathbf{x}_n - \mu_k)\right\}$

$$\frac{\partial c}{\partial \Sigma_k^{-1}} = c \left(-\frac{1}{2}\right) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

# Compute Derivative w.r.t. $\Sigma_k^{-1}$

$$\begin{aligned} & \frac{\partial(\pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k))}{\partial \Sigma_k^{-1}} \\ = & \frac{\partial\left(\frac{\pi_k |\Sigma_k^{-1}|^{1/2}}{(2\pi)^{D/2}} c\right)}{\partial \Sigma_k^{-1}} \\ = & \frac{\pi_k}{(2\pi)^{D/2}} \frac{c \partial(|\Sigma_k^{-1}|^{1/2})}{\partial \Sigma_k^{-1}} + \frac{\pi_k}{(2\pi)^{D/2}} \frac{|\Sigma_k^{-1}|^{1/2} \partial c}{\partial \Sigma_k^{-1}} \\ = & \frac{\pi_k c}{(2\pi)^{D/2}} \frac{1}{2} |\Sigma_k^{-1}|^{-\frac{1}{2}} |\Sigma_k^{-1}| \Sigma_k + \frac{\pi_k |\Sigma_k^{-1}|^{1/2}}{(2\pi)^{D/2}} c \left(-\frac{1}{2} (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T\right) \\ = & \frac{1}{2} \frac{\pi_k |\Sigma_k^{-1}|^{1/2} c}{(2\pi)^{D/2}} \left\{ \Sigma_k - (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \right\} \\ = & \frac{1}{2} \pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k) \left\{ \Sigma_k - (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \right\} \end{aligned} \quad (24)$$

# Compute Derivative w.r.t. $\Sigma_k^{-1}$

$$\begin{aligned} & \frac{\partial \ln p(\mathbf{X}|\mu, \Sigma, \pi)}{\partial \Sigma_k^{-1}} \\ &= \sum_{n=1}^N \frac{1}{\sum_{j=1}^K \pi_k \mathcal{N}(\mathbf{x}_n|\mu_j, \Sigma_j)} \frac{\partial(\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k))}{\partial \Sigma_k^{-1}} \\ &= \frac{1}{2} \sum_{n=1}^N \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k) \{ \Sigma_k - (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \}}{\sum_{j=1}^K \pi_k \mathcal{N}(\mathbf{x}_n|\mu_j, \Sigma_j)} \\ &= \frac{1}{2} \sum_{n=1}^N \gamma(z_{nk}) \{ \Sigma_k - (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \} \end{aligned} \quad (25)$$

Setting  $\frac{\partial \ln p(\mathbf{X}|\mu, \Sigma, \pi)}{\partial \Sigma_k^{-1}}$  to 0, we can get

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \quad (26)$$

# Compute Derivative w.r.t. $\pi_k$

$$\begin{aligned} & \frac{\partial \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})}{\partial \pi_k} \\ = & \frac{\partial \left\{ \sum_{n=1}^N \ln \left\{ \sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \right\} \right\}}{\partial \pi_k} \\ = & \sum_{n=1}^N \frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \\ = & \sum_{n=1}^N \frac{\gamma(z_{nk})}{\pi_k} \end{aligned}$$

# Compute Derivative w.r.t. $\pi_k$

- For  $\{\pi_k\}$ , we need to consider the constraint  $\sum_{k=1}^K \pi_k = 1$
- We can form the Lagrangian:

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

- Setting  $\frac{\partial L}{\partial \pi_k}$  to 0, we can get

$$\begin{aligned} \frac{\partial L}{\partial \pi_k} &= 0 \\ \Leftrightarrow \sum_{n=1}^N \frac{\gamma(z_{nk})}{\pi_k} - \lambda &= 0 \\ \Leftrightarrow \sum_{n=1}^N \gamma(z_{nk}) &= \lambda \pi_k \end{aligned}$$

# Compute Derivative w.r.t. $\pi_k$

$$\sum_{n=1}^N \gamma(z_{nk}) = \lambda \pi_k$$

$$\Rightarrow \sum_{k=1}^K \sum_{n=1}^N \gamma(z_{nk}) = \sum_{k=1}^K (\lambda \pi_k)$$

$$\Leftrightarrow \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) = \lambda \sum_{k=1}^K \pi_k$$

$$\Leftrightarrow \sum_{n=1}^N 1 = \lambda$$

$$\Leftrightarrow \lambda = N$$

# Compute Derivative w.r.t. $\pi_k$

$$\sum_{n=1}^N \gamma(z_{nk}) = \lambda \pi_k$$

$$\Leftrightarrow N_k = N \pi_k$$

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

# Summary of Derivatives

- If we set the derivatives of the log of the likelihood function w.r.t.  $\{\mu_k\}$ ,  $\{\Sigma_k^{-1}\}$  and  $\{\pi_k\}$  to 0, we can get

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \quad (27)$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \quad (28)$$

$$\pi_k = \frac{N_k}{N} \quad (29)$$

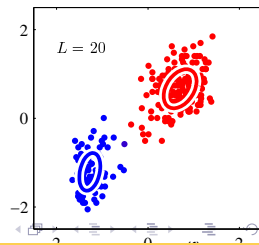
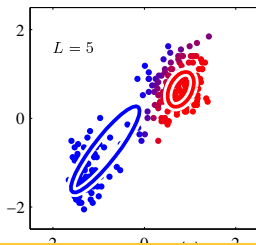
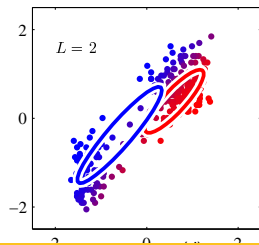
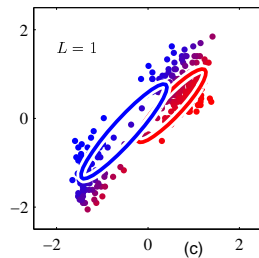
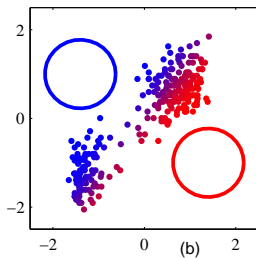
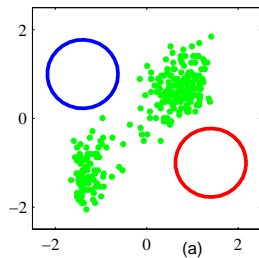
where 
$$N_k = \sum_{n=1}^N \gamma(z_{nk})$$

- Note that these are not closed form solution for parameters  $\{\mu_k\}$ ,  $\{\Sigma_k\}$  and  $\{\pi_k\}$  because  $\gamma(z_{nk})$  are dependent on these parameters

# EM Algorithm for Gaussian Mixtures

- First choose some initial values for the means  $\{\mu_k\}$ ,  $\{\Sigma_k\}$ , and mixing coefficients  $\{\pi_k\}$
- In the expectation step, or E step, we use the current values for the parameters to evaluate the posterior probabilities, or responsibilities  $\{\gamma(z_{nk})\}$
- In the maximization step, or M step, we use  $\{\gamma(z_{nk})\}$  to re-estimate the means  $\{\mu_k\}$ , covariances  $\{\Sigma_k\}$ , and mixing coefficients  $\{\pi_k\}$
- E step and M step are repeated until it is converged

# Illustration of EM Algorithm



# Drawbacks of EM Algorithm

- Compared with the K-means algorithm, EM algorithm takes many more iterations to reach (approximate) convergence
- Each cycle requires significantly more computation in EM algorithm
- EM is not guaranteed to find the largest of these maxima

# EM for Gaussian Mixtures – Detailed Procedures

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)

- Initialize the means  $\{\mu_k\}$ , covariances  $\{\Sigma_k\}$  and mixing coefficients  $\{\pi_k\}$ , and evaluate the initial value of the log likelihood.
- **(E Step)** Evaluate the responsibilities using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \mu_j, \Sigma_j)} \quad (30)$$

- **(M Step)** Re-estimate the parameters using the current responsibilities

$$\mu_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n \quad (31)$$

$$\Sigma_k^{\text{new}} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \mu_k^{\text{new}})(\mathbf{x}_n - \mu_k^{\text{new}})^T \quad (32)$$

# EM for Gaussian Mixtures – Detailed Procedures – Continued

- **M Step (continued)** Re-estimate the parameters using the current responsibilities

$$\pi_k^{\text{new}} = \frac{N_k}{N} \quad (33)$$

where

$$N_k = \sum_{n=1}^N \gamma(z_{nk}) \quad (34)$$

- Evaluate the log likelihood

$$\ln p(\mathbf{X}|\mu, \Sigma, \pi) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k) \right\} \quad (35)$$

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

- 1 K-means Clustering
- 2 Mixtures of Gaussians
  - Maximum Likelihood
  - EM for Gaussian Mixtures
- 3 An Alternative View of EM
  - Gaussian Mixtures Revisited
  - Relation to K-means
  - Mixtures of Bernoulli Distributions
- 4 The EM Algorithm in General
- 5 Summary

# An Alternative View of EM: From the Perspective of Latent Variables

- This framework considers the mixture model by using *latent variables*
- The goal of the EM algorithm is to find maximum likelihood solutions for models having latent variables
- Notations
  - The set of all observed data is denoted as  $X$
  - The set of all latent variables is denoted as  $Z$
  - The set of all model parameters is denoted as  $\theta$
- The log likelihood function is

$$\ln p(X|\theta) = \ln \left\{ \sum_Z p(X, Z|\theta) \right\} \quad (36)$$

- The presence of the sum prevents the logarithm from acting directly on the joint distribution  $p(X, Z|\theta)$ , resulting in complicated expressions for the maximum likelihood solution

# Assumptions

- For each observation  $\mathbf{x}_n$  in  $X$ , the corresponding value of the latent variable  $\mathbf{z}_n$  in  $Z$   
 $\{X, Z\}$  is called the *complete* data set, and we refer to the actual observed data  $X$  as incomplete
- The maximization of the complete-data log likelihood function  $\ln p(X, Z|\theta)$  w.r.t.  $\theta$  is straightforward

# Main Idea of EM in the Framework of Latent Variables

- Since latent variable  $Z$  is not observed, we cannot use the complete-data log likelihood. Instead, we use the expectation of complete-data log likelihood under the posterior distribution of the latent variable to approximate  $\ln p(X|\theta)$   
Only the incomplete data  $X$  is observed in practice. Our knowledge of the values of the latent variables in  $Z$  is given only by the posterior distribution  $p(Z|X, \theta)$
- By maximizing the approximated  $\ln p(X|\theta)$ , we can get updated parameter  $\theta^{\text{new}}$
- Like K-means, we can repeat the process until it converges

# EM in the Framework of Latent Variables

- Initially, the parameter  $\theta$  is set as  $\theta_0$
- In E step
  - We use the current parameter values  $\theta^{\text{old}}$  to find the posterior distribution of the latent variables given by  $p(Z|X, \theta^{\text{old}})$
  - Use  $p(Z|X, \theta^{\text{old}})$  to compute the expectation of the complete-data log likelihood  $\ln p(X, Z|\theta)$  under  $p(Z|X, \theta^{\text{old}})$

$$Q(\theta, \theta^{\text{old}}) = \sum_Z p(Z|X, \theta^{\text{old}}) \ln p(X, Z|\theta) \quad (37)$$

- In M step, we need to compute  $\theta^{\text{new}}$  which maximizes  $Q(\theta, \theta^{\text{old}})$

$$\theta^{\text{new}} = \arg \max_{\theta} Q(\theta, \theta^{\text{old}}) \quad (38)$$

- Note that in this framework log function acts directly on  $p(X, Z|\theta)$ , and by assumption this is tractable

# The General EM Algorithm

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

- Choose an initial setting for the parameters  $\theta^{\text{old}}$
- **E step** Evaluate  $p(Z|X, \theta^{\text{old}})$
- **M step** Evaluate  $\theta^{\text{new}}$  given by

$$\theta^{\text{new}} = \arg \max_{\theta} Q(\theta, \theta^{\text{old}}) \quad (39)$$

where

$$Q(\theta, \theta^{\text{old}}) = \sum_Z p(Z|X, \theta^{\text{old}}) \ln p(X, Z|\theta) \quad (40)$$

- Check for convergence of either the log likelihood or the parameter values. If the convergence criterion is not satisfied, then let  $\theta^{\text{old}} \leftarrow \theta^{\text{new}}$  and return to E step

# Gaussian Mixtures Revisited – I

- We apply the general EM algorithm to Gaussian mixtures from the perspective of latent variables
- The complete-data likelihood function is

$$\begin{aligned} p(X, Z | \mu, \Sigma, \pi) &= \prod_{n=1}^N p(\mathbf{x}_n, \mathbf{z}_n | \mu, \Sigma, \pi) \\ &= \prod_{n=1}^N \prod_{k=1}^K \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)^{z_{nk}} \end{aligned}$$

- The log likelihood function is given as follows

$$\ln p(X, Z | \mu, \Sigma, \pi) = \sum_{n=1}^N \sum_{k=1}^K z_{nk} \{ \ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k) \} \quad (41)$$

- This complete-data log likelihood function can be maximized w.r.t.  $\mu_k, \Sigma_k, \pi_k$  in closed form

## Gaussian Mixtures Revisited – II

To apply the general framework of EM algorithm, we first need to compute the posterior distribution  $p(Z|X, \mu, \Sigma, \pi)$

$$\begin{aligned} p(Z|X, \mu, \Sigma, \pi) &\propto p(Z|\mu, \Sigma, \pi)p(X|Z, \mu, \Sigma, \pi) \\ &= \prod_{n=1}^N p(\mathbf{z}_n|\mu, \Sigma, \pi)p(\mathbf{x}_n|\mathbf{z}_n, \mu, \Sigma, \pi) \\ &= \prod_{n=1}^N \prod_{k=1}^K [\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)]^{z_{nk}} \end{aligned}$$

## Gaussian Mixtures Revisited – III

The expectation of  $z_{nk}$  under the posterior distribution  $p(Z|X, \mu, \Sigma, \pi)$  is

$$\begin{aligned}\mathbb{E}[z_{nk}] &= 1 \times p(z_{nk} = 1|X, \mu, \Sigma, \pi) + 0 \times p(z_{nk} = 0|X, \mu, \Sigma, \pi) \\ &= p(z_{nk} = 1|X, \mu, \Sigma, \pi) \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \mu_j, \Sigma_j)} \\ &= \gamma(z_{nk})\end{aligned}$$

# Gaussian Mixtures Revisited – IV

- The expectation  $Q$  of the complete-data log likelihood function is

$$\begin{aligned} Q &= \mathbb{E}[\ln p(\mathbf{X}, Z | \mu, \Sigma, \pi)] \\ &= \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) (\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \mu_k, \Sigma_k)) \\ &= \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) \left( \ln \pi_k + \frac{1}{2} \ln |\Sigma_k^{-1}| - \frac{1}{2} (\mathbf{x}_n - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_n - \mu_k) \right. \\ &\quad \left. - \frac{D}{2} \ln(2\pi) \right) \end{aligned}$$

- By setting the derivatives of  $Q$  w.r.t  $\mu_k, \Sigma_k, \pi_k$ , we can get  $\mu_k^{\text{new}}, \Sigma_k^{\text{new}}, \pi_k^{\text{new}}$
- Note that  $\gamma(z_{nk})$  are considered as fixed values when computing derivatives

# Gaussian Mixtures Revisited – V

$$\frac{\partial Q}{\partial \mu_k} = \sum_{n=1}^N \gamma(z_{nk}) (-\Sigma_k^{-1} \mu_k + \Sigma_k^{-1} \mathbf{x}_n)$$

By setting  $\frac{\partial Q}{\partial \mu_k}$  to 0, we can get

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n$$

# Gaussian Mixtures Revisited – VI

$$\begin{aligned}\frac{\partial Q}{\partial \Sigma_k^{-1}} &= \sum_{n=1}^N \gamma(z_{nk}) \left( -\frac{1}{2} (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T + \frac{1}{2} |\Sigma_k| |\Sigma_k^{-1}| \Sigma_k \right) \\ &= \frac{1}{2} \sum_{n=1}^N \gamma(z_{nk}) (\Sigma_k - (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T)\end{aligned}$$

Setting  $\frac{\partial Q}{\partial \Sigma_k^{-1}}$  to 0, we can get

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

## Gaussian Mixtures Revisited – VII

For  $\pi_k$ , we need to consider the constraint  $\sum_{k=1}^K \pi_k = 1$ , thus we can construct the Lagrangian

$$L = \mathcal{Q} + \lambda \left( \sum_{k=1}^K \pi_k - 1 \right)$$

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^N \frac{\gamma(z_{nk})}{\pi_k} - \lambda$$

By setting  $\frac{\partial L}{\partial \pi_k}$  to 0, we can get

$$\frac{\partial L}{\partial \pi_k} = 0$$

$$\Rightarrow \lambda = N$$

$$\Rightarrow \sum_{n=1}^N \gamma(z_{nk}) = N\pi_k$$

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

# Comparison of K-means and EM Algorithm – I

- Comparison of K-means and EM Algorithm
  - K-means gives a *hard* assignment of data points to clusters
  - EM algorithm makes a *soft* assignment based on the posterior probabilities
- K-means algorithm can be considered as a particular limit of EM for Gaussian mixtures

## Comparison of K-means and EM Algorithm – II

- Consider a Gaussian mixture model in which the covariance matrices of the mixture components are given by  $\epsilon I$ , i.e.,  $\Sigma_k = \epsilon I$ , where  $\epsilon \in \mathbb{R}$  is a fixed constant, and it need not to be estimated
- Each component is given by

$$p(\mathbf{x}|\mu_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}\epsilon^{1/2}} \exp\left\{-\frac{1}{2\epsilon}\|\mathbf{x}_n - \mu_k\|^2\right\} \quad (42)$$

- The posterior probability of  $\mathbf{z}_k$  given observation  $\mathbf{x}_n$  is

$$\gamma(\mathbf{z}_{nk}) = \frac{\pi_k \exp(-\|\mathbf{x}_n - \mu_k\|^2/2\epsilon)}{\sum_j \pi_j \exp(-\|\mathbf{x}_n - \mu_j\|^2/2\epsilon)} \quad (43)$$

- Suppose  $j = \arg \min_k \|\mathbf{x}_n - \mu_k\|^2$ , if  $\epsilon \rightarrow 0$  and  $\pi_k \neq 0$  for all  $k$ , then

$$\gamma(\mathbf{z}_{nk}) \rightarrow \begin{cases} 1 & \text{if } j = k \\ 0 & \text{otherwise} \end{cases} \quad (44)$$

## Comparison of K-means and EM Algorithm – III

- (44) implies that  $\gamma(z_{nk}) \rightarrow r_{nk}$  when  $\epsilon \rightarrow 0$
- For parameter  $\{\mu_k\}$ , the formula in EM is

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n \rightarrow \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}} \quad (45)$$

- For parameter  $\{\pi_k\}$ , the formula in EM is

$$\pi_k = \frac{N_k}{N} \rightarrow \frac{\sum_{n=1}^N r_{nk}}{N} \quad (46)$$

- It means that the limit of the EM algorithm for this particular Gaussian mixture is the exact K-means

# Mixtures of Bernoulli Distributions – I

- In this part we focus on mixtures of discrete binary variables described by Bernoulli distributions, or **latent class analysis**
- Bernoulli distribution, namely, binary distribution

$$\text{Bern}(x|\mu) = \mu^x(1 - \mu)^{1-x}, \text{ and } x \in \{0, 1\} \quad (47)$$

- Consider a set of  $D$  binary variables  $x_i$ , where  $i = 1, \dots, D$ , each of which is governed by a Bernoulli distribution with parameter  $\mu_{i0}$ , so that

$$p(\mathbf{x}|\mu) = \prod_{i=1}^D \mu_{i0}^{x_i} (1 - \mu_{i0})^{1-x_i} \quad (48)$$

where  $\mathbf{x} = (x_1, \dots, x_D)^T$ ,  $\mu = (\mu_{10}, \dots, \mu_{D0})^T$

# Mixtures of Bernoulli Distributions – II

- The mixture model is given by

$$p(\mathbf{x}|\mu, \pi) = \sum_{k=1}^K \pi_k p(\mathbf{x}|\mu_k) \quad (49)$$

where  $\mu = \{\mu_1, \dots, \mu_K\}$

- Given a data set  $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ , the log likelihood function for this model is given by

$$\ln p(X|\mu, \pi) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k p(\mathbf{x}_n|\mu_k) \right\} \quad (50)$$

- Still, the summation appears inside the logarithm, which makes the maximum likelihood solution to  $\ln p(X|\mu, \pi)$  difficult

# Mixtures of Bernoulli Distributions – III

- To derive the EM algorithm for the mixtures of Bernoulli distributions, the latent variables  $\{\mathbf{z}_n\}$  are introduced
- For each  $\mathbf{x}_n$ , a latent variable  $\mathbf{z}_n = (z_{n1}, \dots, z_{nK})^T$  is introduced, and

$$\sum_{k=1}^K z_{nk} = 1, \text{ and } z_{nk} \in \{0, 1\} \quad (51)$$

- The conditional distribution of  $\mathbf{x}_n$  given  $\mathbf{z}_n$  is given by

$$p(\mathbf{x}_n | \mathbf{z}_n, \mu) = \prod_{k=1}^K p(\mathbf{x}_n | \mu_k)^{z_{nk}} \quad (52)$$

- The prior distribution of latent variables is

$$p(\mathbf{z}_n | \pi) = \prod_{k=1}^K \pi_k^{z_{nk}} \quad (53)$$

## Mixtures of Bernoulli Distributions – IV

- The joint distribution of  $\mathbf{x}_n$  and  $\mathbf{z}_n$  is

$$p(\mathbf{x}_n, \mathbf{z}_n | \mu, \pi) = \prod_{k=1}^K (\pi_k p(\mathbf{x}_n | \mu_k))^{z_{nk}}$$

- The complete-data log likelihood function is given by

$$\begin{aligned} \ln p(X, Z | \mu, \pi) &= \sum_{n=1}^N \ln p(\mathbf{x}_n, \mathbf{z}_n | \mu, \pi) \\ &= \sum_{n=1}^N \ln \left\{ \prod_{k=1}^K (\pi_k p(\mathbf{x}_n | \mu_k))^{z_{nk}} \right\} \\ &= \sum_{n=1}^N \sum_{k=1}^K z_{nk} \left\{ \ln \pi_k + \sum_{i=1}^D [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})] \right\} \end{aligned}$$

- Like Gaussian mixtures, to compute the expectation of  $\ln p(X, Z | \mu, \pi)$ , we only need to compute the expectation of  $z_{nk}$

# Mixtures of Bernoulli Distributions – V

- Like Gaussian mixtures, the expectation of  $z_{nk}$  is

$$\mathbb{E}[z_{nk}] = \frac{\pi_k p(\mathbf{x}_n | \mu_k)}{\sum_{j=1}^K \pi_j p(\mathbf{x}_n | \mu_j)} = \gamma(z_{nk})$$

- The expectation of the complete-data log likelihood function is given by

$$\begin{aligned} Q &= \mathbb{E}[\ln p(X, Z | \mu, \pi)] \\ &= \sum_{n=1}^N \sum_{k=1}^K \gamma(z_{nk}) \left\{ \ln \pi_k + \sum_{i=1}^D [x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki})] \right\} \end{aligned}$$

- In M step, we need to maximize the expectation of the log of complete-data likelihood function
- The method to compute optimal  $\{\mu_k\}$  and  $\{\pi_k\}$  in M step is similar to that of Gaussian mixture models, i.e., by setting the derivative of  $Q = \mathbb{E}[\ln p(X, Z | \mu, \pi)]$  w.r.t.  $\{\mu_k\}$  and  $\{\pi_k\}$  to 0, and compute  $\{\mu_k\}$  and  $\{\pi_k\}$

## Mixtures of Bernoulli Distributions – VI

$$\begin{aligned}\frac{\partial Q}{\partial \mu_{ki}} &= \sum_{n=1}^N \gamma(z_{nk}) \left( \frac{x_{ni}}{\mu_{ki}} - \frac{1 - x_{ni}}{1 - \mu_{ki}} \right) \\ &= \sum_{n=1}^N \gamma(z_{nk}) \frac{x_{ni} - \mu_{ki}}{\mu_{ki}(1 - \mu_{ki})}\end{aligned}$$

By setting  $\frac{\partial Q}{\partial \mu_{ki}}$  to 0, we can get

$$\mu_{ki} = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) x_{ni} \quad (54)$$

Or equivalently,

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n = \bar{\mathbf{x}}_n \quad (55)$$

# Mixtures of Bernoulli Distributions – VII

To compute  $\{\pi_k\}$ , we need to construct the Lagrangian

$$L = \mathcal{Q} - \lambda \left( \sum_{k=1}^K \pi_k - 1 \right)$$

therefore, 
$$\frac{\partial L}{\partial \pi_k} = \sum_{i=1}^N \gamma(z_{nk}) \frac{1}{\pi_k} - \lambda$$

$$\Rightarrow \lambda = N$$

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

- 1 K-means Clustering
- 2 Mixtures of Gaussians
  - Maximum Likelihood
  - EM for Gaussian Mixtures
- 3 An Alternative View of EM
  - Gaussian Mixtures Revisited
  - Relation to K-means
  - Mixtures of Bernoulli Distributions
- 4 The EM Algorithm in General
- 5 Summary

# The EM Algorithm in General – I

- The EM algorithm, is a general technique for finding **maximum likelihood solutions** for probabilistic models having **latent variables**
- In this part we consider a general probabilistic model
- Problem statement in general discussion
  - All of the observed variables are denoted by  $X$
  - All of the hidden variables are denoted by  $Z$
  - The joint distribution  $p(X, Z|\theta)$  is governed by a set of parameters denoted by  $\theta$
  - The goal is to maximize the likelihood function given by

$$p(X|\theta) = \sum_Z p(X, Z|\theta) \quad (56)$$

- In our discussion,  $Z$  is assumed as discrete variables, but the conclusion still holds for continuous variables and the combination of discrete and continuous variables

# The EM Algorithm in General – II

- Assumptions
  - The direct optimization of  $p(X|\theta)$  is difficult
  - The optimization of the complete-data likelihood function  $p(X, Z|\theta)$  is much easier
- For latent variables  $Z$ , the distribution is denoted as  $q(Z)$
- The decomposition of  $p(X|\theta)$  for any choice of  $q(Z)$

$$\ln p(X|\theta) = \mathcal{L}(q, \theta) + \text{KL}(q||p) \quad (57)$$

$$\text{where } \mathcal{L}(q, \theta) = \sum_Z q(Z) \ln \left\{ \frac{p(X, Z|\theta)}{q(Z)} \right\} \quad (58)$$

$$\text{KL}(q||p) = - \sum_Z q(Z) \ln \left\{ \frac{p(Z|X, \theta)}{q(Z)} \right\} \quad (59)$$

- $\mathcal{L}(q, \theta)$  is a functional of the distribution  $q(Z)$ , and a function of the parameters  $\theta$

# The EM Algorithm in General – III

## Definition and Properties of K-L divergence $KL(q||p)$

- For probability distributions  $q$  and  $p$  of a discrete random variable  $i$ , the Kullback-Leibler divergence (also information divergence, information gain, or relative entropy) of  $p$  from  $q$  is defined as

$$KL(q||p) = \sum_i q(i) \ln \frac{q(i)}{p(i)} \quad (60)$$

- The K-L divergence is a measure of the difference between two probability distributions: from a “true” probability distribution  $q$  to an arbitrary probability distribution  $p$
- $KL(q||p) \geq 0$  for any distribution  $p$  and  $q$
- $KL(q||p) = 0$  iff  $q = p$

# The EM Algorithm in General – IV

$$\begin{aligned}\ln p(X, Z|\theta) &= \ln p(Z|X, \theta) + \ln p(X|\theta) \\ \Rightarrow \mathcal{L}(q, \theta) &= \sum_Z q(Z) \ln \left\{ \frac{p(X, Z|\theta)}{q(Z)} \right\} \\ &= \sum_Z q(Z) (\ln p(X, Z|\theta) - \ln q(Z)) \\ &= \sum_Z q(Z) (\ln p(Z|X, \theta) + \ln p(X|\theta) - \ln q(Z)) \\ &= \sum_Z q(Z) \ln \left\{ \frac{p(Z|X, \theta)}{q(Z)} \right\} + \sum_Z q(Z) \ln p(X|\theta) \\ &= -\text{KL}(q\|p) + \ln p(X|\theta) \\ \Rightarrow \ln p(X|\theta) &= \mathcal{L}(q, \theta) + \text{KL}(q\|p)\end{aligned}$$

# The EM Algorithm in General – V

- Since  $\text{KL}(q||p) \geq 0$ ,  $\mathcal{L}(q, \theta)$  is a lower bound on  $\ln p(\mathbf{X}|\theta)$

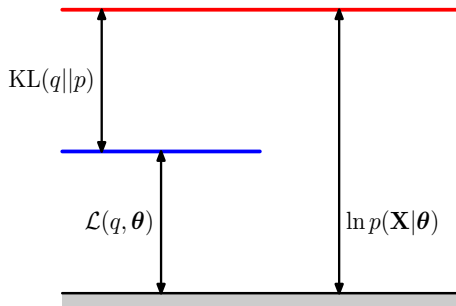


Figure: Illustration of the Decomposition

# The EM Algorithm in General – VI – E Step

- Suppose that the current value of the parameter vector  $\theta$  is  $\theta^{\text{old}}$
- In the E step, the lower bound  $\mathcal{L}(q, \theta)$  is maximized with respect to  $q(Z)$  while holding  $\theta^{\text{old}}$  fixed
- Note that

$$\mathcal{L}(q, \theta^{\text{old}}) = -\text{KL}(q||p) + \ln p(X|\theta^{\text{old}}) \quad (61)$$

- When  $\text{KL}(q||p) = 0$  or  $q(Z) = p(Z|X, \theta^{\text{old}})$ ,  $\mathcal{L}(q, \theta^{\text{old}})$  is maximized

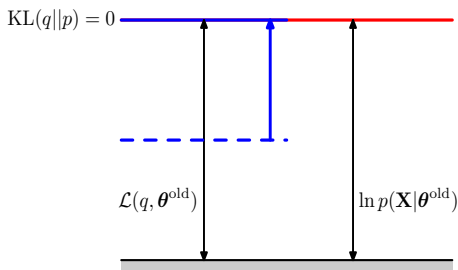


Figure: Illustration of the E Step

# The EM Algorithm in General – VII – M Step

- In the M step, the distribution  $q(Z)$  is held fixed and the lower bound  $\mathcal{L}(q, \theta)$  is maximized with respect to  $\theta$  to give some new value  $\theta^{\text{new}}$
- In E step,  $q(Z)$  is fixed as  $q(Z) = p(Z|X, \theta^{\text{old}})$

$$\mathcal{L}(q, \theta) = \sum_z p(Z|X, \theta^{\text{old}}) \ln p(X, Z|\theta) \quad (62)$$

$$- \sum_z p(Z|X, \theta^{\text{old}}) \ln p(Z, X|\theta^{\text{old}})$$

$$= Q(\theta, \theta^{\text{old}}) + \text{const} \quad (63)$$

$$\text{where } Q(\theta, \theta^{\text{old}}) = \sum_z p(Z|X, \theta^{\text{old}}) \ln p(X, Z|\theta) \quad (64)$$

- Thus in M step, we only need to maximize  $Q(\theta, \theta^{\text{old}})$  w.r.t.  $\theta$
- $Q(\theta, \theta^{\text{old}})$  is the expectation of the complete-data log likelihood

# The EM Algorithm in General – VIII – M Step

- By maximizing  $Q(\theta, \theta^{\text{old}})$  w.r.t.  $\theta$  we can get  $\theta^{\text{new}}$
- $\text{KL}(q||p) = \text{KL}(p(Z|X, \theta^{\text{old}})||p(Z|X, \theta^{\text{new}})) > 0$  in most cases
- Both  $\mathcal{L}(q, \theta)$  and  $\text{KL}(q||p)$  are increased in the M step

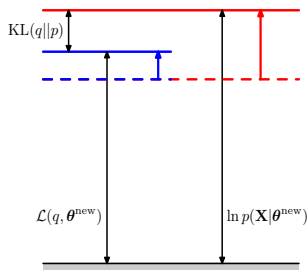


Figure: Illustration of the M Step

# The EM Algorithm in General – IX

- Red curve is  $\ln p(X|\theta)$
- Blue curve is the lower bound  $\mathcal{L}(q, \theta^{\text{old}})$  given  $\theta^{\text{old}}$
- Both curves have the same gradient and same value for the point  $\theta^{\text{old}}$
- Since  $\mathcal{L}(q, \theta^{\text{old}})$  is a convex function having a unique maximum for mixture components from the exponential family, we can get  $\theta^{\text{new}}$  in M step

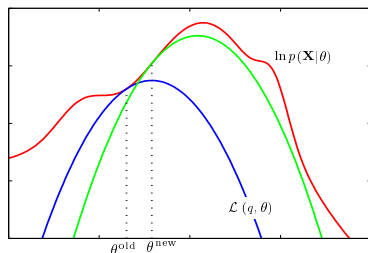


Figure: Illustration of EM Algorithm

# The EM Algorithm in General – IX

- In each cycle of E step and M step,  $\ln p(\mathbf{x}|\theta)$  is guaranteed to be increased
- For any optimization problem, if the maximum is not  $+\infty$ , any algorithm which increases the objective function in every step or every  $t(t > 1)$  steps, is guaranteed to find a local maximum
- For the general EM algorithm, a local solution is guaranteed

# Extensions of EM Algorithm

- The EM algorithm breaks down the potentially difficult problem of maximizing the likelihood function into two stages, the E step and the M step, each of which will often prove simpler to implement
- For complex models it may be the case that either the E step or the M step, or indeed both, remain intractable
- This leads to two possible extensions of the EM algorithm
  - The *generalized EM*, or GEM, addresses the problem of an intractable M step
  - For complex E step, we can perform a partial, rather than complete, optimization of  $\mathcal{L}(q, \theta^{\text{old}})$  with respect to  $q(Z)$

# Generalized EM Algorithm

- GEM addresses the problem of an intractable M step
- GEM seeks to change the parameters in such a way as to increase the value of  $\mathcal{L}(q, \theta)$
- Different forms of GEM
  - Use one of the nonlinear optimization strategies, such as the conjugate gradients algorithm, during the M step
  - *Expectation conditional maximization*, or ECM, involves making several constrained optimizations within each M step
    - Example: Parameters might be partitioned into groups, and the M step is broken down into multiple steps each of which involves optimizing one of the subset with the remainder held fixed

# Incremental EM Algorithm

- This algorithm can be utilized to process independent data points  $\mathbf{x}_1, \dots, \mathbf{x}_N$  with corresponding latent variables  $\mathbf{z}_1, \dots, \mathbf{z}_N$
- The joint distribution  $p(X, Z|\theta)$  can factorize over the data points, i.e.

$$p(X, Z|\theta) = \prod_{n=1}^N p(\mathbf{x}_n, \mathbf{z}_n|\theta) \quad (65)$$

- In incremental EM algorithm, at each EM cycle only one data point is processed
  - In E step, the responsibility of only one point is re-evaluated
  - In M step, we can use some simple statistics to update parameter  $\theta$  if the mixture components are members of the exponential family

# Incremental EM Algorithm for Gaussian Mixture Model

- Suppose we perform an update for data point  $m$ , and the old and new values of the responsibilities are denoted as  $\gamma^{\text{old}}(z_{mk})$  and  $\gamma^{\text{new}}(z_{mk})$
- Before updating data point  $m$

$$N_k^{\text{old}} = \sum_{n=1}^N \gamma^{\text{old}}(z_{nk})$$

- After updating data point  $m$

$$\begin{aligned} N_k^{\text{new}} &= \sum_{n \neq m} \gamma^{\text{old}}(z_{nk}) + \gamma^{\text{new}}(z_{mk}) \\ &= N_k^{\text{old}} - \gamma^{\text{old}}(z_{mk}) + \gamma^{\text{new}}(z_{mk}) \end{aligned}$$

# Incremental EM Algorithm for Gaussian Mixture Model – Estimate $\mu_k^{\text{new}}$ in M Step

$$\begin{aligned}\mu_k^{\text{new}} &= \frac{1}{N_k^{\text{new}}} \sum_{n=1}^N \gamma^{\text{new}}(z_{nk}) \mathbf{x}_n \\ &= \frac{1}{N_k^{\text{new}}} \left( \sum_{n \neq m} \gamma^{\text{old}}(z_{nk}) \mathbf{x}_n + \gamma^{\text{new}}(z_{mk}) \mathbf{x}_m \right) \\ &= \frac{1}{N_k^{\text{new}}} \left( N_k^{\text{old}} \mu_k^{\text{old}} - \gamma^{\text{old}}(z_{mk}) \mathbf{x}_m + \gamma^{\text{new}}(z_{mk}) \mathbf{x}_m \right) \\ &= \frac{1}{N_k^{\text{new}}} \left( \left( N_k^{\text{new}} + \gamma^{\text{old}}(z_{mk}) - \gamma^{\text{new}}(z_{mk}) \right) \mu_k^{\text{old}} \right. \\ &\quad \left. - \gamma^{\text{old}}(z_{mk}) \mathbf{x}_m + \gamma^{\text{new}}(z_{mk}) \mathbf{x}_m \right) \\ &= \mu_k^{\text{old}} + \left( \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} \right) (\mathbf{x}_m - \mu_k^{\text{old}})\end{aligned}$$

# Incremental EM Algorithm

- Advantages of incremental EM algorithm
  - Both the E step and the M step take fixed time that is independent of the total number of data points in each iteration
  - The convergence rate is faster than the batch version of EM algorithm

# Summary

- The K-means Algorithm
- The EM algorithm
  - The EM algorithm for mixtures of continuous variables
    - Mixtures of Gaussian Distributions
    - Two different methods to derive the EM algorithm for mixtures of Gaussian distributions
      - From the perspective of maximum likelihood
      - From the perspective of latent variables
  - The EM algorithm for mixtures of discrete variables
    - Mixtures of Bernoulli Distributions
  - The EM algorithm in general
    - Extensions of EM algorithm