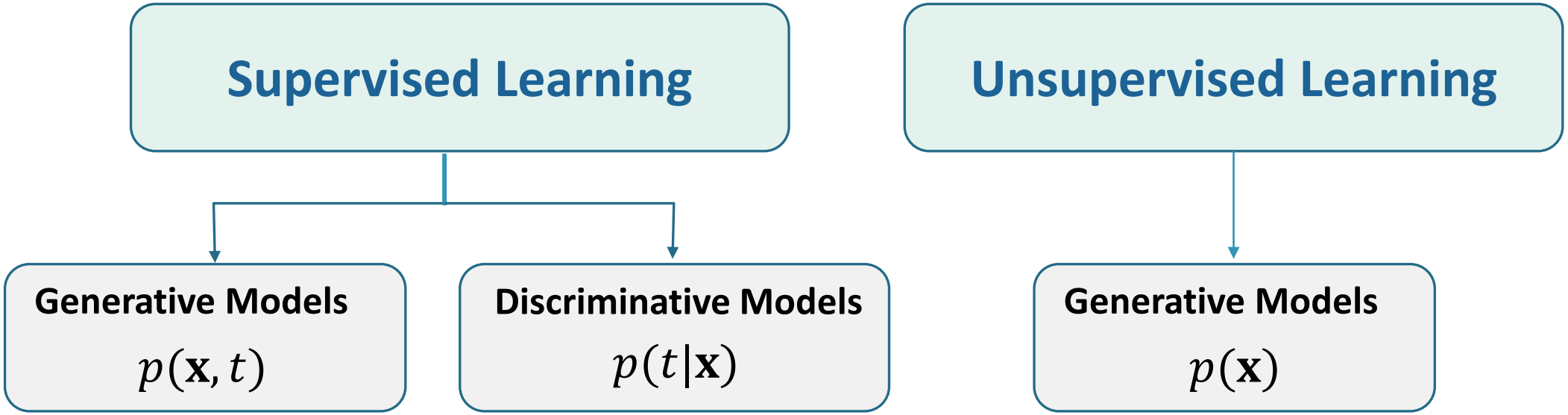
Introduction to Machine Learning

Mixture Models & EM

SHENG-JYH WANG

NATIONAL YANG MING CHIAO TUNG UNIVERSITY, TAIWAN

SPRING, 2024



Remark: We call it a classification/regression problem if ***t*** is discrete/continuous.

K-means Clustering (1/7)

Given a data set $\{x_1, \ldots, x_N\}$ and the cluster number K, we want to partition the data set into K clusters.

We define μ_k, $k = 1, \ldots, K$, as the centers of the clusters.

$\Rightarrow$ Aim to find an assignment of data points to clusters, as well as a set of vectors $\{\mu_k\}$, such that the sum of the squares of the distances of each data point to its closest vector μ_k is a minimum.

K-means Clustering (2/7)

1-of-K coding scheme

For each data point x_n, we introduce a corresponding set of binary indicator variables $r_{nk} \in \{0, 1\}$, where $k = 1, \ldots, K$.
If the data point x_n is assigned to the kth cluster, then $r_{nk} = 1$ and $r_{nj} = 0$ for $j \neq k$.

Distortion Measure:

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

K-means Clustering (3/7)

The K-means Algorithm

(1)Choose some initial values for μ_k.

(2)Two-stage process

- ✓ Keep μ_k fixed and minimize J with respect to r_{nk}. (E step)
- ✓ Keep r_{nk} fixed and minimize J with respect to μ_k. (M step)

(3)Repeat (2) until convergence.

Remark: Convergence of the K-means algorithm is assured. However, it may converge to a local minimum of J.

K-means Clustering (4/7)

Determination of r_{nk}.

Since J is a linear function of r_{nk}, we can optimize for each n separately.

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

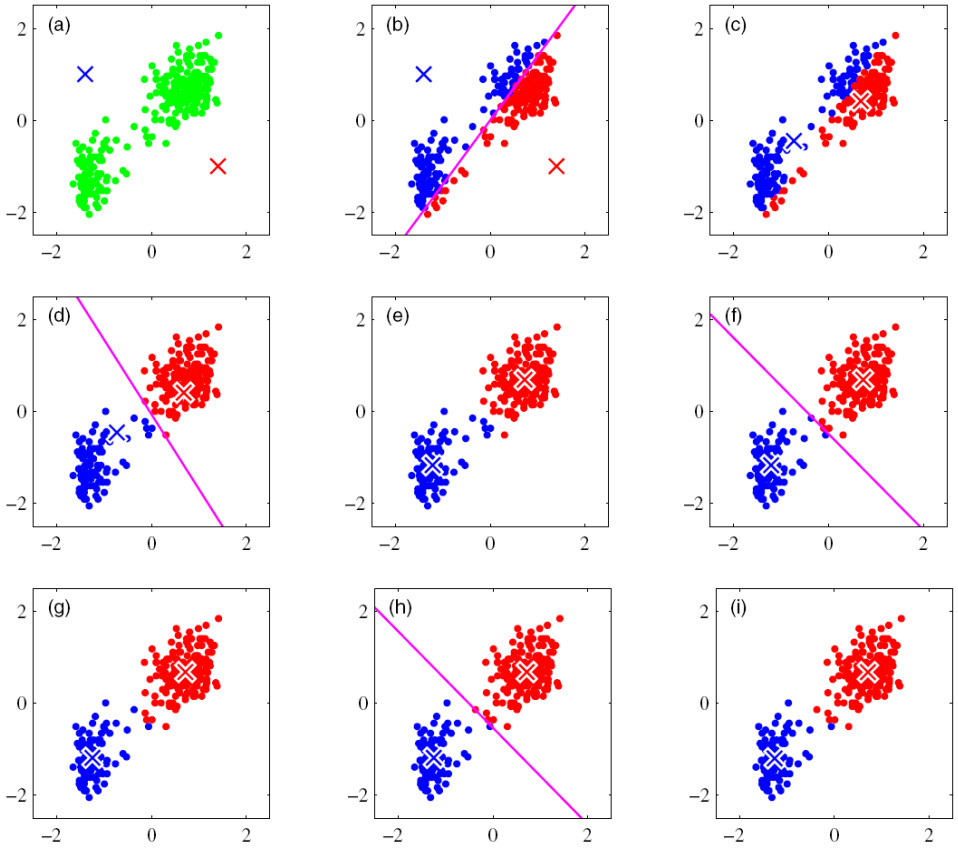
Determination of μ_k

$$\frac{\partial J}{\partial \boldsymbol{\mu}_k} = 0$$

$$2\sum_{n=1}^{N} r_{nk}(\mathbf{x}_n - \boldsymbol{\mu}_k) = 0$$

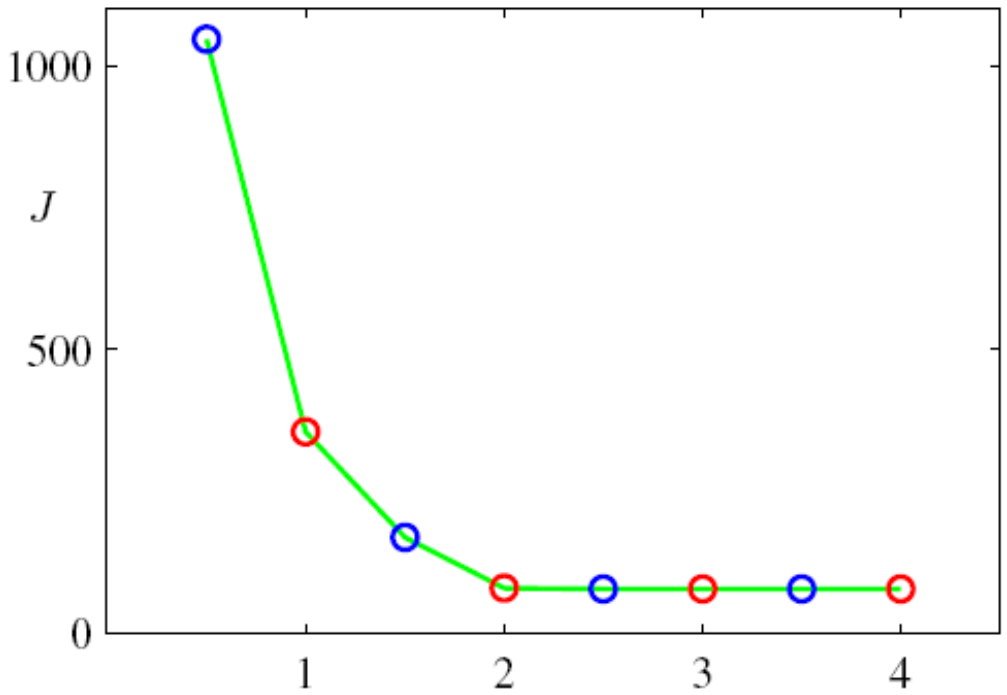
$$\boldsymbol{\mu}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

K-means Clustering (5/7)



K-means Clustering (6/7)

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 (7/7)

Example of K-means Algorithm: Image Segmentation

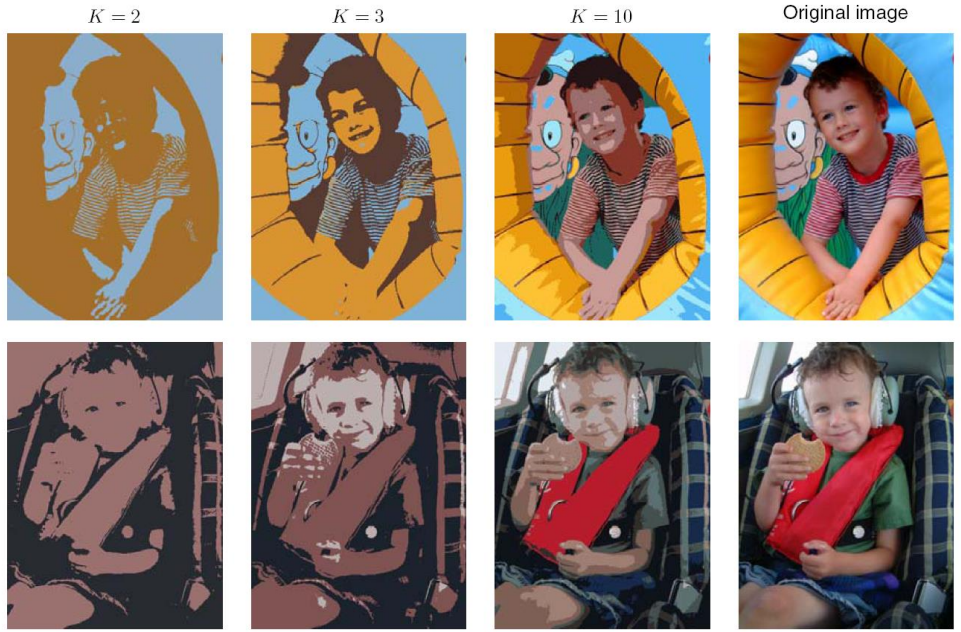


Figure 9.3 Two examples of the application of the K-means clustering algorithm to image segmentation showing the initial images together with their K-means segmentations obtained using various values of K. This also illustrates of the use of vector quantization for data compression, in which smaller values of K give higher compression at the expense of poorer image quality.

Mixtures of Gaussians (1/12)

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

The mixtures of Gaussians can be described with the introduction of the ***latent variables* z**.

z: *K*-dimensional binary random variable with a 1-of-*K* representation.

a particular element z_k is equal to 1 while all other elements are 0.

$z_k \in \{0, 1\}$

$$\sum_k z_k = 1$$

$\mathbf{z} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ or $\begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}$ or $\cdots$

The marginal distribution over **z** is specified in terms of the mixing coefficients π_k

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

where $0 \leqslant \pi_k \leqslant 1$ $\quad \sum_{k=1}^{K} \pi_k = 1$

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

Mixtures of Gaussians (2/12)

The conditional distribution of **x** given **z** is

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

第一群

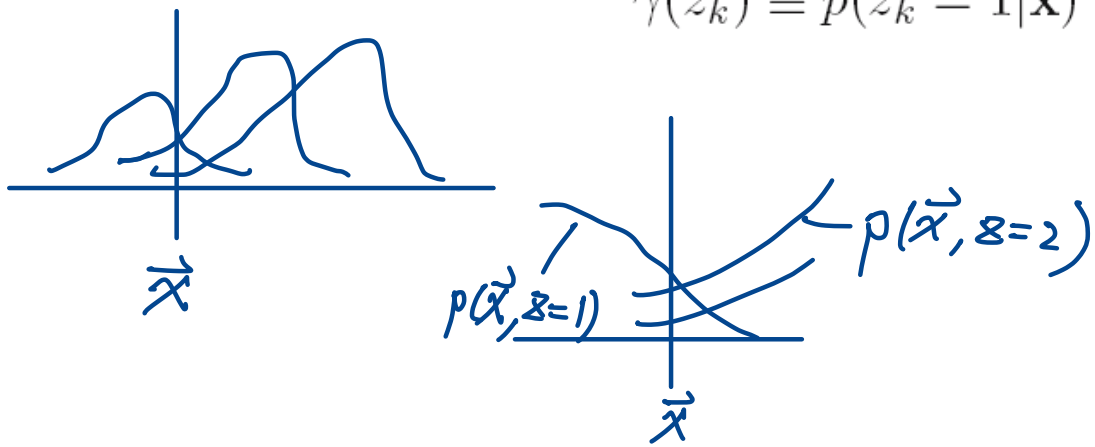
$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

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

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

Remark: For every observed data point $\mathbf{x}_n$, there is a corresponding latent variable $\mathbf{z}_n$.

Mixtures of Gaussians (3/12)



$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) = \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(\mathbf{x}|z_j = 1)}$$

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

We can use the *ancestral sampling* technique to generate random samples distributed according to the Gaussian mixture model.

(1) Generate a value for **z** from the marginal distribution $p(\mathbf{z})$.

(2) Generate a value for **x** from the conditional distribution $p(\mathbf{x}|\mathbf{z})$.

Mixtures of Gaussians (4/12)

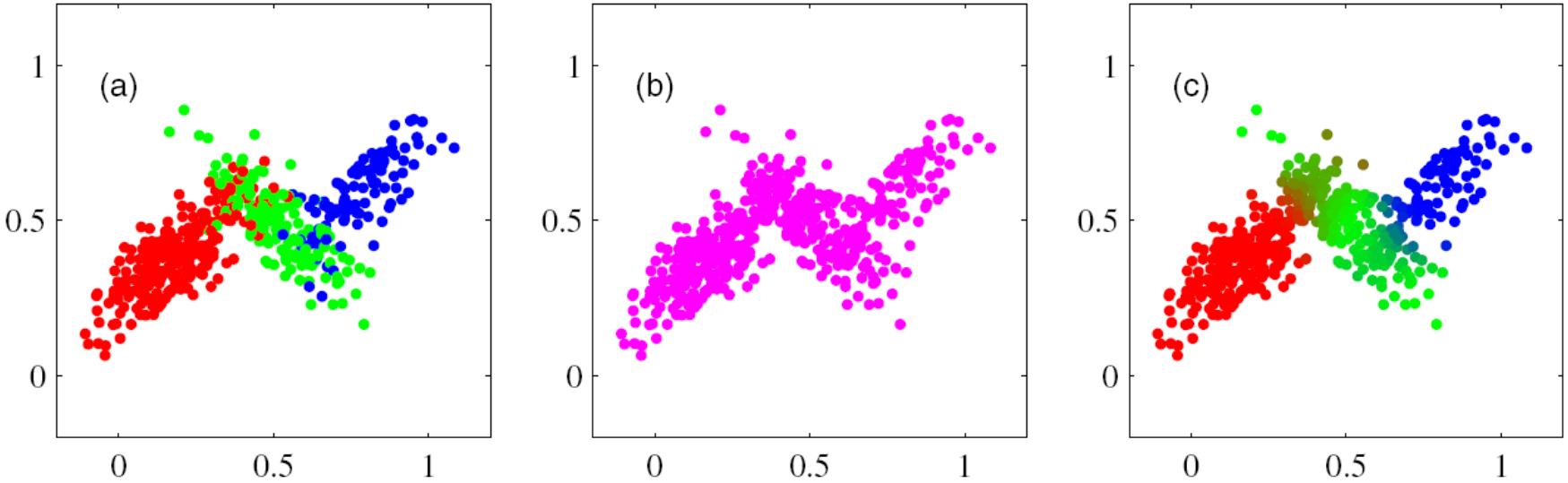


Figure 9.5 Example of 500 points drawn from the mixture of 3 Gaussians shown in Figure 2.23. (a) Samples from the joint distribution $p(\mathbf{z})p(\mathbf{x}|\mathbf{z})$ in which the three states of $\mathbf{z}$, corresponding to the three components of the mixture, are depicted in red, green, and 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

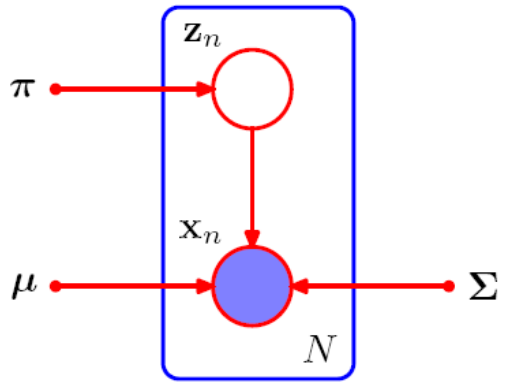
Mixtures of Gaussians (5/12)

Maximum Likelihood Estimate of Model Parameters

Suppose we have a data set of observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, and we wish to model this data using a mixture of Gaussians.

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

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 $n = 1, \ldots, N$.

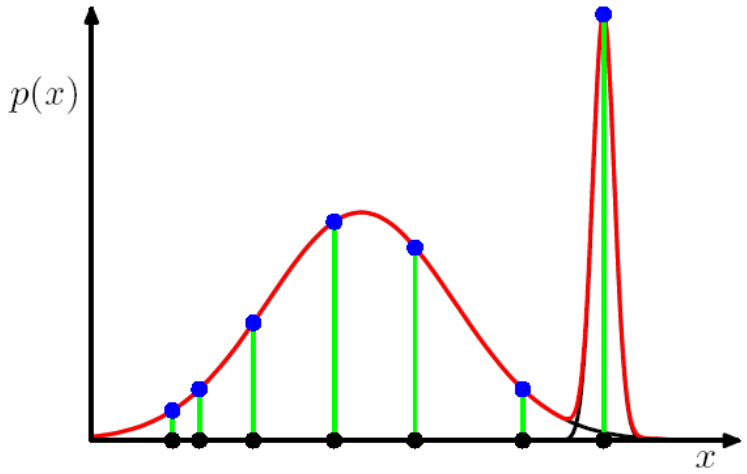


Mixtures of Gaussians (6/12)

However, maximizing the log likelihood function is an ill-posed problem.

Example: Suppose $\boldsymbol{\Sigma}_k = \sigma_k^2 \mathbf{I}$ and $\mu_j = \mathbf{x}_n$ for some value of n.

$\Rightarrow \sigma_j \to 0$



Mixtures of Gaussians (7/12)

In applying maximum likelihood to Gaussian mixture models, we must take steps to avoid finding such pathological solutions and instead seek local maxima of the likelihood function that are well behaved.

Moreover, for any given maximum likelihood solution, 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.

Mixtures of Gaussians (8/12)

EM (Expectation-Maximization) algorithm for Gaussian Mixtures

An elegant and powerful method for finding maximum likelihood solutions for models with latent variables.

The conditions that must be satisfied at a maximum of the likelihood function:

$$\frac{\partial \ln p(\mathbf{X} \mid \pi, \mu, \Sigma)}{\partial \mu_k} = 0$$

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

Mixtures of Gaussians (9/12)

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

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

N_k: the effective number of points assigned to cluster k.

$\boldsymbol{\mu}_k$: a weighted mean of all of the points in the data set.

Similarly, we have

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

Mixtures of Gaussians (10/12)

To maximize ln $p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\mu},\boldsymbol{\Sigma})$ with respect to π_k, we define the Lagrangian function as

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

$$\Rightarrow \quad 0 = \sum_{n=1}^{N}\frac{\mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)} + \lambda$$

$$\Rightarrow \quad \lambda = N$$

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

Mixtures of Gaussians (11/12)

EM for Gaussian Mixtures

1. Initialize the means μ_k, covariances $\mathbf{\Sigma}_k$ and mixing coefficients π_k.
2. **E step**: Evaluate the responsibilities using the current parameter values

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

3. **M step**: Re-estimate the parameters using the current $\gamma(z_{nk})$.

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

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

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

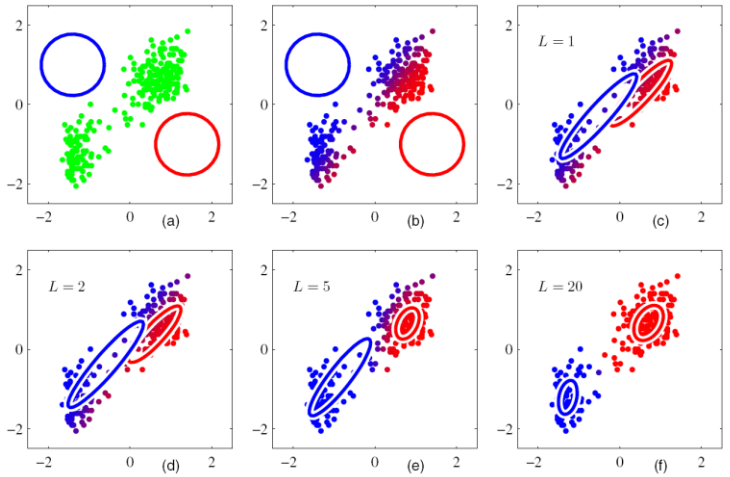
where $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$

Mixtures of Gaussians (12/12)

4. Evaluate the log likelihood

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

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



An Alternative View of EM (1/10)

X: all observed data, in which the *n*th row represents $\mathbf{x}_n^T$

Z: the set of all latent variables, in which the *n*th row represents $\mathbf{z}_n^T$

θ: the set of all model parameters

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \left\{ \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \right\}$$

The presence of the sum within logarithm results in complicated expressions for the maximum likelihood solution.

An Alternative View of EM (2/10)

{**X**,**Z**}: the *complete* data set

In practice, we are not given the complete data set {**X**,**Z**}, but only the incomplete data **X**.

- ✓ Maximization of this complete-data log likelihood function is straightforward.
- ✓ Our state of knowledge of the values of the latent variables in **Z** is given only by the posterior distribution $p(\mathbf{Z}|\mathbf{X},\theta)$.

Since we cannot use the complete-data log likelihood, we consider instead its expected value under the posterior distribution of the latent variable (the E step) and then maximize this expectation (the M step).

An Alternative View of EM (3/10)

The General EM Algorithm

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 $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$.

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

3. **M step** Evaluate θ^{new} given by

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

4. 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 step 2.

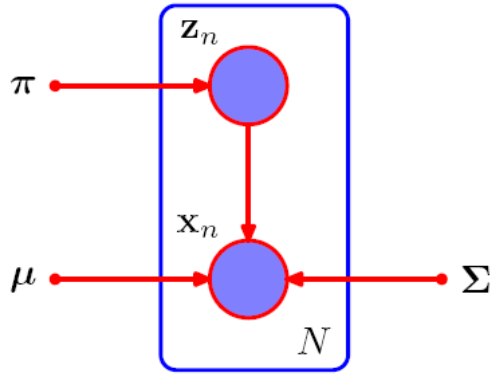
An Alternative View of EM (4/10)

Remarks:

1. In the definition of $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}})$, the logarithm acts directly on the joint distribution $p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$, and so the corresponding M-step maximization will be tractable.
2. The EM algorithm can also be used to find MAP (maximum posterior) solutions. In this case, the E step remains the same as in the maximum likelihood case, while in the M step the quantity to be maximized is given by $Q(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}) + \ln p(\boldsymbol{\theta})$.

An Alternative View of EM (5/10)

Example: ***Mixtures of Gaussian Distributions***



The complete-data likelihood function is

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{nk}}$$

z_{nk}: the kth component of $\mathbf{z}_n$.

An Alternative View of EM (6/10)

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

The complete-data log likelihood function can be maximized in closed form. However, since we do not have values for the latent variables, we consider the expectation, with respect to the posterior distribution of the latent variables, of the complete-data log likelihood.

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_{nk}}$$

An Alternative View of EM (7/10)

$$E[z_{nk}] = \frac{\sum_{z_n} z_{nk} \prod_{k'} [\pi_{k'} N(\mathbf{x}_n \mid \mu_{k'}, \mathbf{\Sigma}_{k'})]^{z_{nk'}}}{\sum_{z_n} \prod_{j} [\pi_j N(x_n \mid \mu_j, \Sigma_j)]^{z_{nj}}} = \frac{\pi_k N(\mathbf{x}_n \mid \mu_k, \mathbf{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j N(\mathbf{x}_n \mid \mu_j, \mathbf{\Sigma}_j)} = \gamma(z_{nk})$$

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

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

First we choose some initial values for the parameters μ^{old}, Σ^{old}, and π^{old}.

Keep the responsibilities $\gamma(z_{nk})$ fixed and maximize

$$\mathbb{E}_{\mathbf{Z}}[\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\}$$

with respect to μ_k, Σ_k, and π_k.

An Alternative View of EM (8/10)

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

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

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

Relation to K-means

The *K*-means algorithm performs a *hard* assignment of data points to clusters, in which each data point is associated uniquely with one cluster.

The EM algorithm makes a *soft* assignment based on the posterior probabilities.

An Alternative View of EM (9/10)

Consider a Gaussian mixture model in which the covariance matrices of the mixture components are given by $\varepsilon\mathbf{I}$.
Each component is given by

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

The posterior probabilities, or responsibilities, for a particular data point $\mathbf{x}_n$, are given by

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

An Alternative View of EM (10/10)

If we consider the limit $\varepsilon \to 0$, the responsibilities $\gamma(z_{nk})$ for the data point $\mathbf{x}_n$ all go to zero except for term j, for which the responsibility $\gamma(z_{nj})$ will go to unity.

$\Rightarrow$ We obtain a hard assignment of data points to clusters, just as in the K-means algorithm.

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

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

The EM Algorithm in General (1/10)

- ✓ The EM algorithm is a general technique for finding maximum likelihood solutions for probabilistic models having latent variables.
- ✓ 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.

X : all of the observed variables

Z : all of the hidden variables

Our goal is to maximize the likelihood function

$$p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$$

The EM Algorithm in General (2/10)

We introduce a distribution $q(\mathbf{Z})$ defined over the latent variables.
For any choice of $q(\mathbf{Z})$, the following decomposition holds

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p)$$

where

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

$$\mathrm{KL}(q\|p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\}$$

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

The EM Algorithm in General (3/10)

Remarks:

1. $L(q,\boldsymbol{\theta})$ contains the joint distribution of $\mathbf{X}$ and $\mathbf{Z}$.

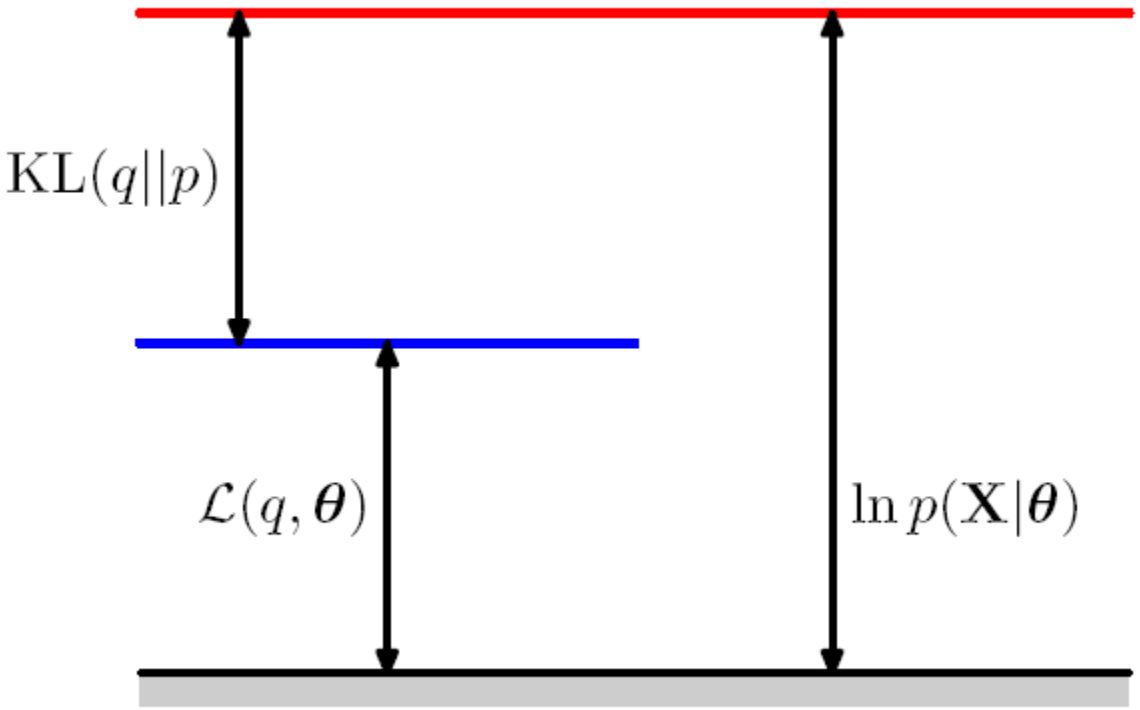
   $\mathrm{KL}(q||p)$ contains the conditional distribution of $\mathbf{Z}$ given $\mathbf{X}$.

2. $\mathrm{KL}(q||p)$ is the Kullback-Leibler divergence between $q(\mathbf{Z})$ and the posterior distribution $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$.

3. $\mathrm{KL}(q||p) \geq 0$, with equality if, and only if, $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta})$.

4. $L(q,\boldsymbol{\theta}) \leq \ln p(\mathbf{X}|\boldsymbol{\theta})$. That is, $L(q,\boldsymbol{\theta})$ is a lower bound on $\ln p(\mathbf{X}|\boldsymbol{\theta})$.

The EM Algorithm in General (4/10)

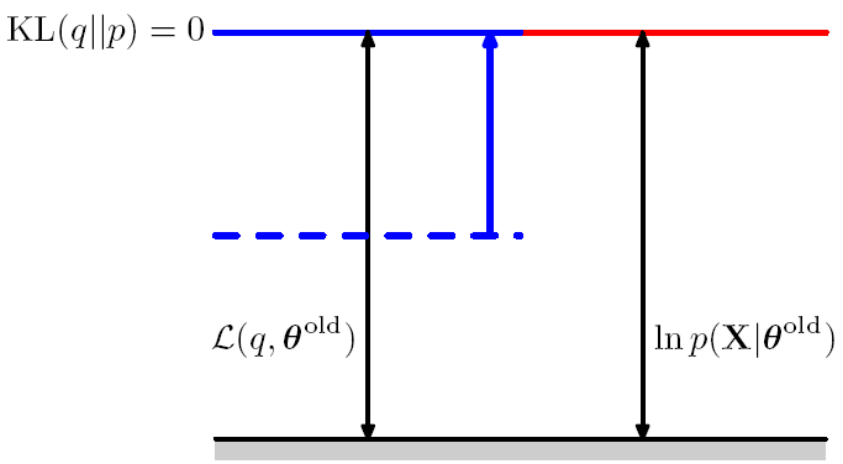


The EM Algorithm in General (5/10)

Suppose that the current value of the parameter vector is $\boldsymbol{\theta}^{\text{old}}$.

E step: The lower bound $L(q,\boldsymbol{\theta}^{\text{old}})$ is maximized with respect to $q(\mathbf{Z})$ while holding $\boldsymbol{\theta}^{\text{old}}$ fixed.

Since the value of $\ln p(\mathbf{X}|\boldsymbol{\theta}^{\text{old}})$ does not depend on $q(\mathbf{Z})$, the largest value of $L(q,\boldsymbol{\theta}^{\text{old}})$ will occur when $q(\mathbf{Z})$ is equal to the posterior distribution $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}})$.



The EM Algorithm in General (6/10)

M step: the distribution $q(\mathbf{Z})$ is held fixed and the lower bound $L(q,\boldsymbol{\theta})$ is maximized with respect to $\boldsymbol{\theta}$ to give some new value $\boldsymbol{\theta}^{\text{new}}$.

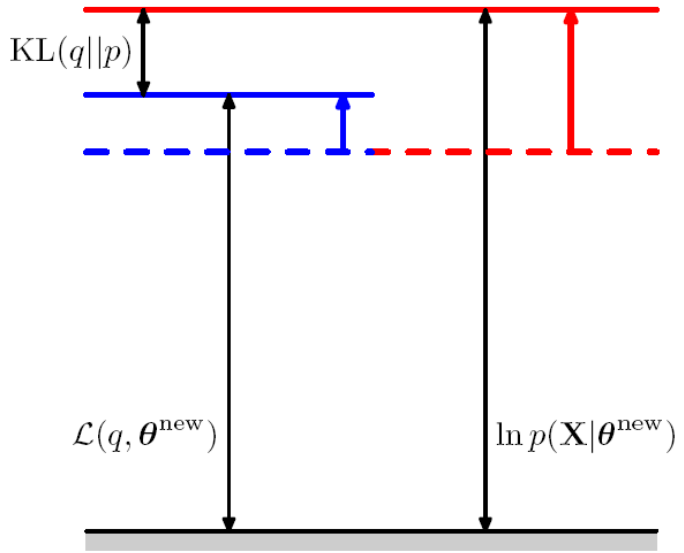
This will cause the lower bound L to increase (unless it is already at a maximum), which will necessarily cause the corresponding log likelihood function to increase.

Because the distribution q is determined using the old parameter values rather than the new values and is held fixed during the M step, it will not equal the new posterior distribution $p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{new}})$. $\Rightarrow$ There will be a nonzero KL divergence.

$\Rightarrow$ Both KL$(q||p)$ and $L(q,\boldsymbol{\theta})$ are increased in the M step.

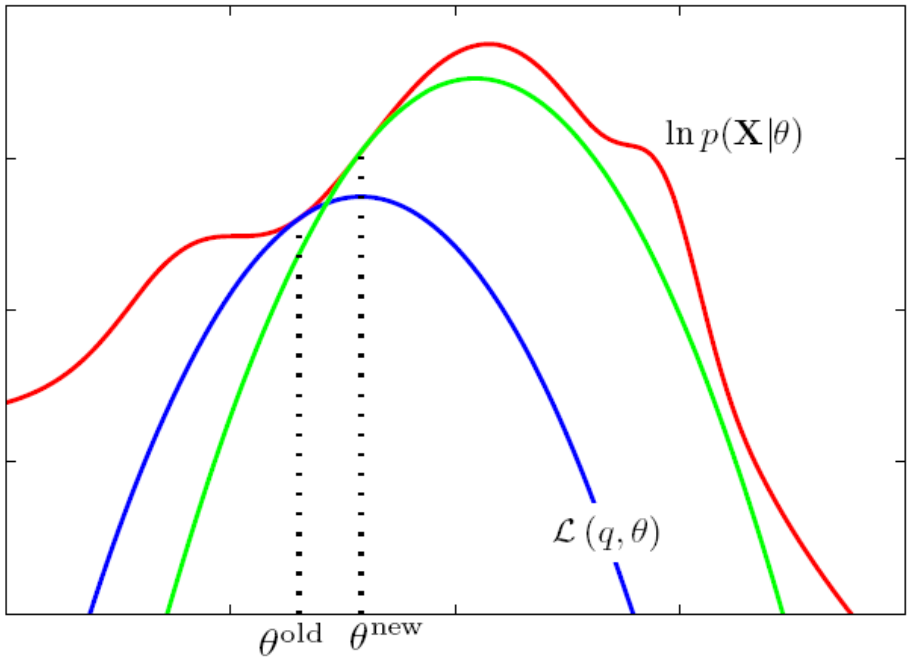
The EM Algorithm in General (7/10)

$$\begin{aligned}\mathcal{L}(q,\boldsymbol{\theta}) &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \\ &= \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}}) + \text{const}\end{aligned}$$



The EM Algorithm in General (8/10)

The EM algorithm involves alternately computing a lower bound on the log likelihood for the current parameter values and then maximizing this bound to obtain the new parameter values. See the text for a full discussion.



The EM Algorithm in General (9/10)

We can also use the EM algorithm to maximize the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$ for models in which we have introduced a prior $p(\boldsymbol{\theta})$ over the parameters.

$$\ln p(\boldsymbol{\theta}|\mathbf{X}) = \ln p(\boldsymbol{\theta}, \mathbf{X}) - \ln p(\mathbf{X})$$

$$\begin{aligned} \ln p(\boldsymbol{\theta}|\mathbf{X}) &= \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q\|p) + \ln p(\boldsymbol{\theta}) - \ln p(\mathbf{X}) \\ &\geqslant \mathcal{L}(q, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - \ln p(\mathbf{X}). \end{aligned}$$

Nevertheless, for complex models it may be the case that either the E step or the M step, or indeed both, remain intractable.

$\Rightarrow$ Extensions of the EM algorithm

The EM Algorithm in General (10/10)

- ***The generalized EM (GEM) algorithm***:
  Instead of aiming to maximize $L(q,\theta)$ with respect to θ, it seeks to change the parameters in such a way as to increase its value.
  - One way is to use one of the nonlinear optimization strategies, such as the conjugate gradients algorithm, during the M step.
  - One way is to make several constrained optimizations within each M step (*expectation conditional maximization*, or ECM).

    Example: The parameters are 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.
- For the E step, we can perform a partial, rather than complete, optimization of $L(q,\theta)$ with respect to $q(\mathbf{Z})$.

Appendix

Calculus of Variations

- Function: return an output value for an input value.

$$x \xrightarrow{f(x)} y$$

- Functional: return an output value for an input function.

$$y(x) \xrightarrow{F(y)} F$$

function value

Example: Entropy

$$H[p] = -\int p(x) \ln p(x)\, dx$$

- In Calculus

find a value of x that maximizes (or minimizes) a function y(x).

- In Calculus of Variations

find a function y(x) that maximizes (or minimizes) a functional F[y].

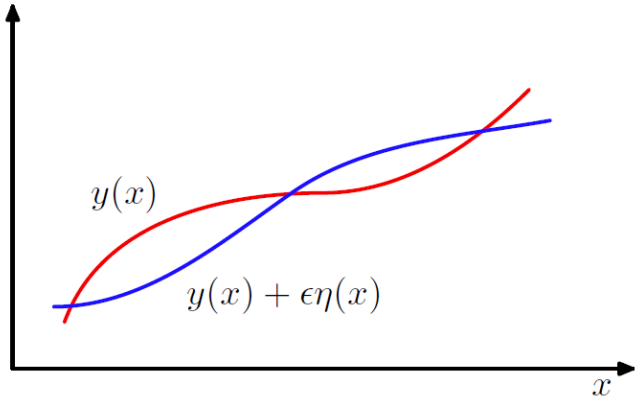
- In Calculus

$$y(x+\epsilon) = y(x) + \frac{\mathrm{d}y}{\mathrm{d}x}\epsilon + O(\epsilon^2)$$

$$y(x_1+\epsilon_1, \ldots, x_D+\epsilon_D) = y(x_1, \ldots, x_D) + \sum_{i=1}^{D} \frac{\partial y}{\partial x_i}\epsilon_i + O(\epsilon^2)$$

- In Calculus of Variations

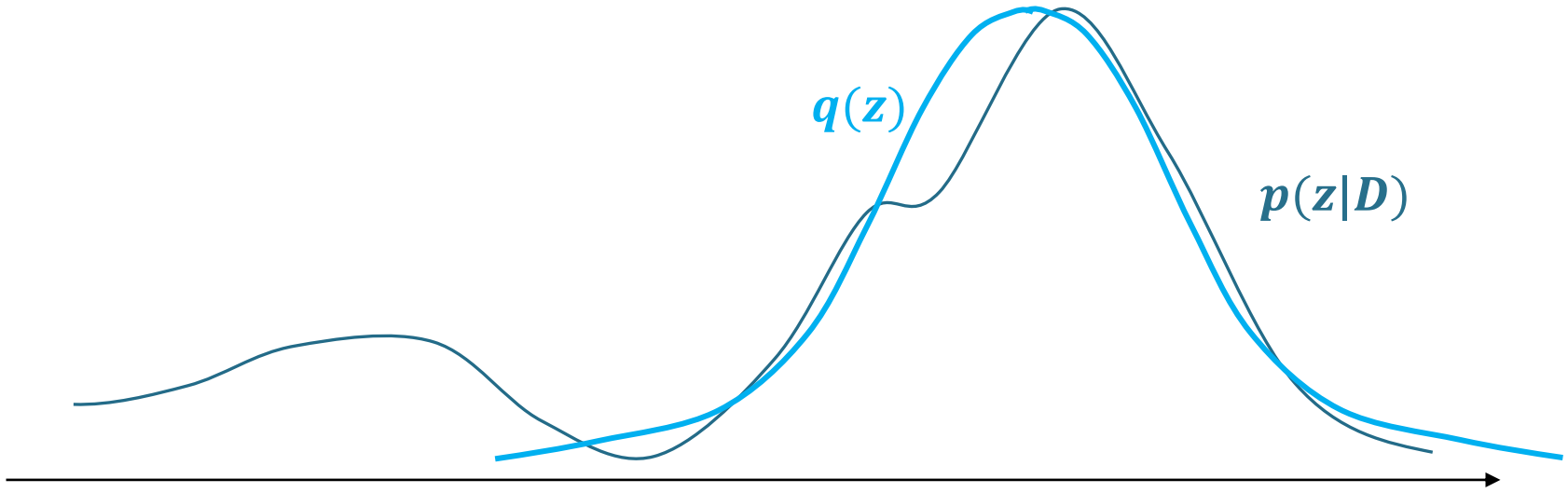
$$F[y(x)+\epsilon\eta(x)] = F[y(x)] + \epsilon \int \frac{\delta F}{\delta y(x)}\eta(x)\,\mathrm{d}x + O(\epsilon^2)$$



Variational Lower Bound

$p(z|D)$: the posterior probability of the latent variable z given the observed data D.

- The posterior probability $p(z|D)$ could be very complicated and hard to describe and compute.
  ⇒ find a surrogate posterior $q(z)$, which is easier to work with, to approximate $p(z|D)$.



Variational Lower Bound

$$p(\mathbf{z}|\mathbf{x}=\mathrm{D}) = \frac{p(\mathbf{x}=\mathrm{D}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x}=\mathrm{D})}$$

Typically, the bottleneck in finding $p(\mathbf{z}|\mathbf{x})$ is the computation of the marginal probability $p(\mathbf{x})$, which can be expressed as

$$p(\mathbf{x}=\mathrm{D}) = \int \cdots \int p(\mathbf{x}=\mathrm{D},\mathbf{z}) d\mathbf{z}_1 d\mathbf{z}_2 \cdots d\mathbf{z}_D$$

The computation is usually intractable!

⇒ Find the optimal surrogate posterior $q(\mathbf{z})$ in Q that is most similar to $p(\mathbf{z}|\boldsymbol{x}=\mathrm{D})$

$$q^*(z) = \arg\min_{q(z)\in Q}(\ \mathrm{KL}(\ q(\mathbf{z})\ ||\ p(\mathbf{z}|\mathbf{x}=\mathrm{D})\ )$$

K-L Divergence

Variational Lower Bound

$$\mathrm{KL}(\, q(\mathbf{z}) \,||\, p(\mathbf{z}|\mathbf{x}) \,) = E_{\mathbf{z}\sim q(\mathbf{z})}[\log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})}] = \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})} d\mathbf{z}$$

However, we don't have the posterior probability function $p(\mathbf{z}|\mathrm{D})$!

$$\mathrm{KL}(\, q(\mathbf{z}) \,||\, p(\mathbf{z}|\mathbf{x}) \,) = \int q(\mathbf{z}) \log \frac{q(\mathbf{z})p(\mathbf{x})}{p(\mathbf{z},\mathbf{x})} d\mathbf{z}$$

$$= \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z},\mathbf{x})} d\mathbf{z} + \int q(\mathbf{z}) \log(p(\mathbf{x})) \, d\mathbf{z}$$

$$= \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z},\mathbf{x})} d\mathbf{z} + \log(p(\mathbf{x})) = -E_{z\sim q(\mathbf{z})}[\log \frac{p(\mathbf{z},\mathbf{x})}{q(\mathbf{z})}] + \log(p(\mathbf{x}))$$

Known

$L(q(\mathbf{z}))$

Variational Lower Bound

$$\mathrm{KL}(\, q(\mathbf{z}) \,||\, p(\mathbf{z}|\mathbf{x}=\mathrm{D})\,) = -L(q(\mathbf{z})) + \underbrace{\log(p(\mathbf{x}=\mathrm{D}))}_{\text{Fixed}}$$

where $L\big(q(\mathbf{z})\big) = E_{z\sim q(\mathbf{z})}\left[\log \frac{p(\mathbf{z},\mathbf{x})}{q(\mathbf{z})}\right]$

$$\mathrm{KL}(\, q(\mathbf{z}) \,||\, p(\mathbf{z}|\mathbf{x})\,) = \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z}|\mathbf{x})}\, d\mathbf{z} = -\int q(\mathbf{z}) \log \frac{p(\mathbf{z}|\mathbf{x})}{q(\mathbf{z})}\, d\mathbf{z}$$

$\log(p(\mathbf{x}=\mathbf{D}))$: Evidence

$L\big(q(\mathbf{z})\big) = \log\big(p(\mathbf{x}=\mathrm{D})\big) - \mathrm{KL}(\, q(\mathbf{z}) \,||\, p(\mathbf{z}|\mathbf{x}=\mathrm{D})\,) \le \log(p(\mathbf{x}=\mathrm{D}))$

Evidence Lower Bound (ELBO)

$L\big(q(\mathbf{z})\big) = \log(p(\mathbf{x}=\mathrm{D}))$ if and only if $\mathrm{KL}(\, q(\mathbf{z}) \,||\, p(\mathbf{z}|\mathbf{x})\,) = 0$

Variational Lower Bound

Summary

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \mathrm{KL}(q\|p)$$

where

$$\mathcal{L}(q) = \int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

$$\mathrm{KL}(q\|p) = -\int q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X})}{q(\mathbf{Z})} \right\} \mathrm{d}\mathbf{Z}$$

- We maximize the evidence lower bound L(q) by optimization with respect to the distribution q(Z).
- The maximum of L(q) occurs when KL(q||p) vanishes. That is, when q(Z) = p(Z|X).
- When p(Z|X) is intractable, we consider a restricted family of distributions q(Z) and minimize the KL divergence.