Approximate Bayesian Estimation: Lecture 2 Expectation Maximization

Lan Hua

Northwestern Polytechnical University
Key Laboratory of Information Fusion Technology

E-mail: lanhua@nwpu.edu.cn
Tel: 15399035213

October 23, 2019

Lecture Outline

What is the EM?

The expectation maximization (EM) algorithm is an iterative method to find ML or MAP estimates of parameters in statistical models, where the model depends on unobserved latent variables.

Two main applications

- Data has *missing values*, due to problems with or limitations of the observation process.
- *Optimizing*, the likelihood function is extremely hard, but the likelihood function can be simplified by assuming the existence of and values for additional missing or hidden parameters.

What is the EM?

Consider the following log-likelihood function

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \tag{1}$$

- $\mathbf{X}$ is the set of all observed data (measurements)
- $\boldsymbol{\theta}$ is the set of all model parameters (to be estimated)
- $\mathbf{Z}$ is the set of all latent variables (missing data, unobserved variables)

The ML estimate $\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} \ln p(\mathbf{X}|\boldsymbol{\theta})$ is intractable due to the summarization of latent variables $\mathbf{Z}$.

What is the EM?

Key idea of the EM algorithm

- Maximization of the *complete-data* log-likelihood function $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ is easier than $\ln p(\mathbf{X}|\boldsymbol{\theta})$.
- In practice, however, we are not given the complete data set $\{\mathbf{X}, \mathbf{Z}\}$, but only the incomplete data X. Knowledge of latent variable Z is given only by the posterior distribution $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$.
- Instead the (unknown) latent variable $\mathbf{Z}$ with its expected value $\mathbb{E}[\mathbf{Z}]$, which corresponds to the *E-Step* of the EM algorithm.
- In the subsequent *M-Step*, we maximize this expectation to obtain the ML estimate of parameter $\boldsymbol{\theta}$.

The EM algorithm iteratively estimate the latent variable $\mathbf{Z}$ in the E-Step and the parameters $\boldsymbol{\theta}$ in the M-step until convergence.

Derivation of the EM Algorithm

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

1. Choose an initial setting for the parameters $\boldsymbol{\theta}^{\text{old}}$.
2. **E-Step** Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$.
3. **M-Step** Evaluate $\boldsymbol{\theta}^{\text{new}}$ given by

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \tag{2}$$

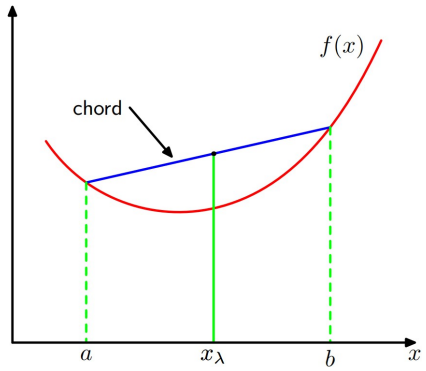
where the conditional expected log-likelihood (Q-function)

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \tag{3}$$

4. Check for convergence of either the log-likelihood or the parameter values. If the convergence criterion is not satisfied, then let $\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}}$ and return to step 2.

Derivation of the EM Algorithm

Interpretation of EM: (1) as a consequence of Jensen's inequality



- If a function is a convex, which implies

$$f(\lambda a + (1-\lambda)b) \leq \lambda f(a) + (1-\lambda) f(b)$$

Without proof, a convex function $f(x)$ satisfies

$$f\left(\sum_{i=1}^{M} \lambda_x x_i\right) \leq \sum_{i=1}^{M} \lambda_i f(x_i) \tag{4}$$

The result (4) is known as *Jensen's inequality*. If λ_i is the probability distribution over a discrete variable x, i.e., $\lambda_i = p(x_i)$, then

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)] \tag{5}$$

Derivation of the EM Algorithm

Define $L(\boldsymbol{\theta})$ as the incomplete-date log-likelihood, then

$$
\begin{aligned}
L(\theta) - L(\theta_n) &= \ln\left(\sum_Z p(X|Z,\theta)p(Z|\theta)\right) - \ln p(X|\theta_n) \qquad (6)\\
&= \ln\left(\sum_Z p(Z|X,\theta_n)\cdot\frac{p(X|Z,\theta)p(Z|\theta)}{p(Z|X,\theta_n)}\right) - \ln p(X|\theta_n)\\
&\geq \sum_Z p(Z|X,\theta_n)\ln\left(\frac{p(X|Z,\theta)p(Z|\theta)}{p(Z|X,\theta_n)}\right) - \ln p(X|\theta_n)\\
&= \sum_Z p(Z|X,\theta_n)\ln\left(\frac{p(X|Z,\theta)p(Z|\theta)}{p(Z|X,\theta_n)p(X|\theta_n)}\right)\\
&\triangleq \Delta(\theta|\theta_n)
\end{aligned}
$$

which implies that

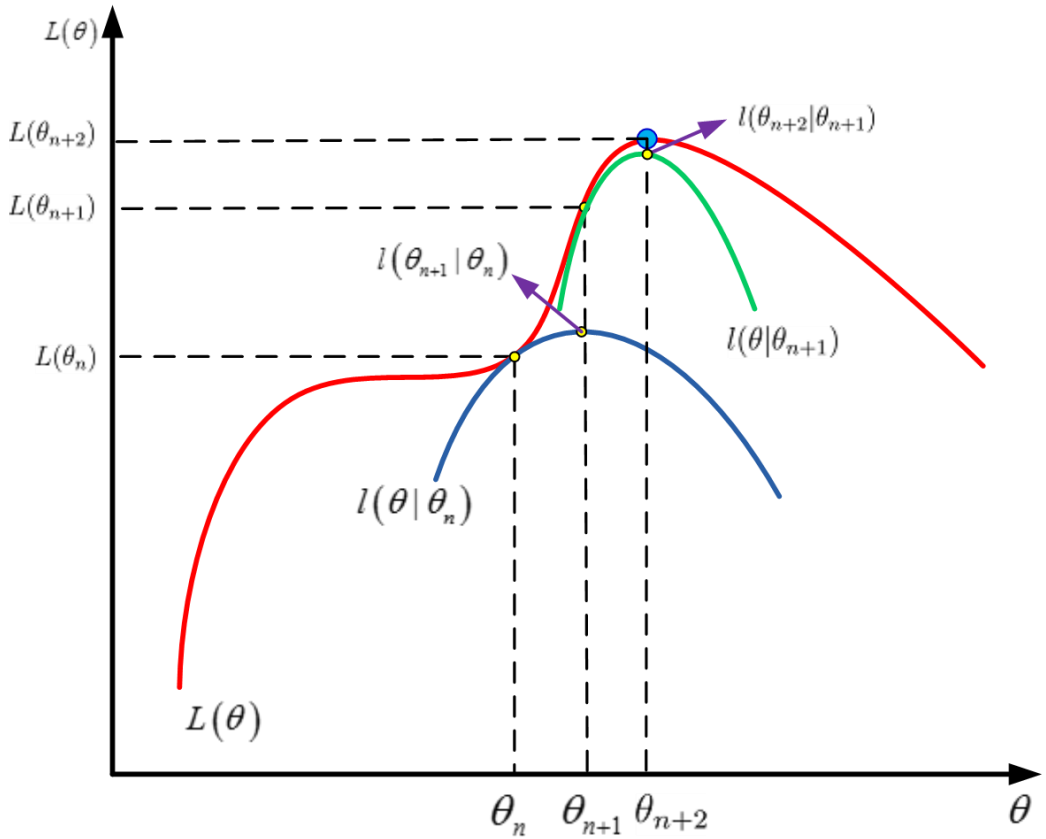
$$
L(\theta) \geq \underbrace{L(\theta_n) + \Delta(\theta|\theta_n)}_{l(\theta|\theta_n)\rightarrow\text{lower bound}} \qquad (7)
$$

Derivation of the EM Algorithm

$$
\begin{aligned}
\theta_{n+1} &= \arg\max_{\theta}\{l(\theta|\theta_n)\} \\
&= \arg\max_{\theta}\left\{L(\theta_n) + \sum_{Z} p(Z|X,\theta_n)\ln\frac{p(X|Z,\theta)p(Z|\theta)}{p(X|\theta_n)p(Z|X,\theta_n)}\right\} \\
&= \arg\max_{\theta}\underbrace{\sum_{Z} p(Z|X,\theta_n)\ln p(X,Z|\theta)}_{\mathcal{Q}(\theta,\theta_n)} \\
&= \arg\max_{\theta}\left\{E_{Z|X,\theta_n}[\ln p(X,Z|\theta)]\right\}
\end{aligned}
\tag{8}
$$

Maximize the complete data log-likelihood $\ln p(X,Z|\theta)$, where the unknown latent variable Z is instead with its expectation.

Derivation of the EM Algorithm



- At each iteration, the log-likelihood $L(\theta)$ increase.
- At each iteration, the estimated parameter θ maximize $l(\theta|\theta_n)$.

Derivation of the EM Algorithm

Interpretation of EM: (2) from the perspective of variational Bayes

For any introduced distribution $q(\mathbf{Z})$, the following decomposition holds

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q||p) \tag{9}$$

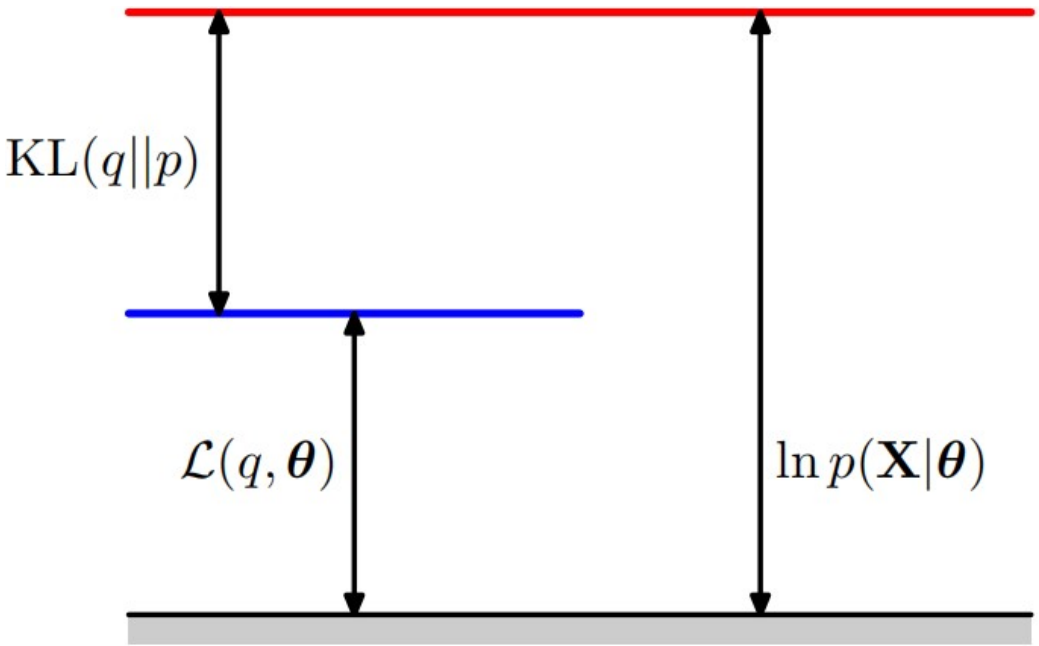
where

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \tag{10}$$

$$\mathrm{KL}(q||p) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \left\{ \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \right\} \tag{11}$$

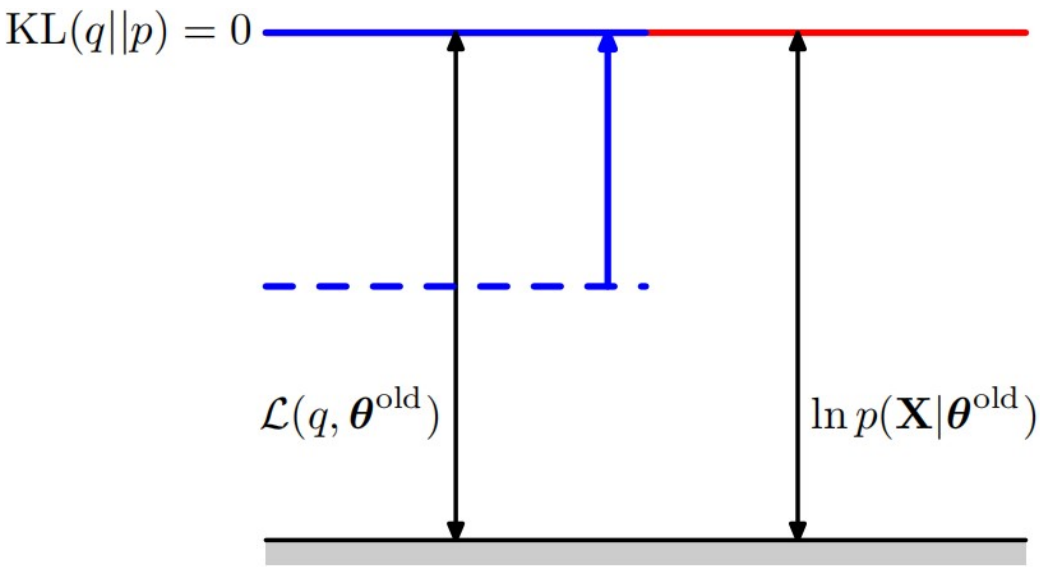
- $\mathcal{L}(q, \boldsymbol{\theta})$ is a functional of $q(\mathbf{Z})$ and a function of the parameters $\boldsymbol{\theta}$.
- $\mathrm{KL}(q||p) \geq 0$ is the Kullback-Leibler divergence from q to p.

Derivation of the EM Algorithm



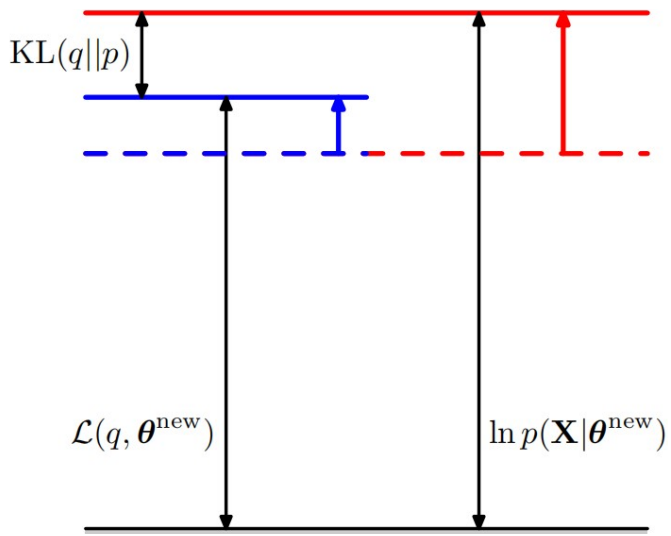
- Illustration of the decomposition given by (9).
- Because the KL divergence satisfies $\mathrm{KL}(q||p) \geq 0$, the quantity $\mathcal{L}(q, \boldsymbol{\theta})$ is a lower bound on the log-likelihood function $\ln p(\mathbf{X}|\boldsymbol{\theta}.)$

Derivation of the EM Algorithm



- In the E-Step, the lower bound $\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})$ is maximized with respect to $q(\mathbf{Z})$ while holding $\boldsymbol{\theta}^{\mathrm{old}}$ fixed.
- Since $\ln p(\mathbf{X}|\boldsymbol{\theta}^{\mathrm{old}})$ does not depend on $q(Z)$, the largest value of $\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})$ will occur when the KL divergence vanishes.
- We obtain $q(\mathbf{Z}) = \ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\mathrm{old}})$.

Derivation of the EM Algorithm



- In the M-Step, $q(\mathbf{Z})$ is held fixed and the lower bound $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ is maximized with respect to $\boldsymbol{\theta}$ to give some new value $\boldsymbol{\theta}^{\text{new}}$.
- The lower bound $\mathcal{L}$ will increase, which will necessarily cause the corresponding log-likelihood function to increase.
- Because q is determined using the old parameter values rather than the new values, it will not equal the new posterior distribution $p(Z|X, \boldsymbol{\theta}^{\text{new}})$, and hence there will be a nonzero KL-divergence.
- The increase in $\ln p(\mathbf{X}|\boldsymbol{\theta})$ is therefore greater than the increase in $\mathcal{L}$.

Properties of the EM Algorithm

Advantages:

- Under fairly general conditions, the convergence of EM algorithm is robust to initialization.
- The EM algorithm is generally easy to program, and requires small storage space.
- Many extension of the EM algorithm in literature deal with the complex situations where a closed form formulation does not exist.
- The EM algorithm is numerically stable, with each EM iteration increasing the likelihood.

Properties of the EM Algorithm

Disadvantage:

- The EM iteration does not automatically provide an estimate of the covariance matrix of the parameter estimates.
- The EM algorithm may converge slowly for some nonlinear systems with large incomplete uncertainties, and may be analytically intractable, especially when the latent variable is of high-dimension.
- Many extension of the EM algorithm in literature deal with the complex situations where a closed form formulation does not exist.
- The EM algorithm does guarantee convergence to the global maximum when multiple maxima exist.

Extension of the EM Algorithm

Generalized EM (GEM) addresses the problem of an intractable M-step. Instead of aiming to maximize $Q-$function $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ with respect to $\boldsymbol{\theta}$.

Expectation conditional maximization (ECM) replaces the complicated M-step of EM by a number of computationally simpler conditional maximization (CM)-steps because the Q-function is difficult to be optimized in the whole parameter space.

Expectation conditional maximization either (ECME) is an extension of ECM, which partitions the CM-steps into two groups ϕ_Q and ϕ_L. While the CM-steps index by $s \in \phi_Q$ remain the same with ECM, and the CM-steps index by $s \in \phi_L$ maximize incomplete-data likelihood function.

Extension of the EM Algorithm

Monte Carlo EM (MCEM) addresses the problem of an intractable E-Step. Instead of computing the intractable analytical solution of Q-function,, the MCEM algorithm approximates it via the Monte Carlo method.

Parameter expanded EM (PX-EM) use a "covariance adjustment" to correct the analysis of the M-step by expanding the complete-data model $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ to a larger model $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}, \kappa)$, and κ is an auxiliary parameter.

Accelerated EM (AEM) is developed by appending a line search to each iteration. Formally, given initial value $\boldsymbol{\theta}^{\mathbf{0}}$, in the $(n+1)$ iteration, the new estimate θ^{n+1} is computed by $\theta^{n+1} = \theta^n + \varsigma^n d^n$.

EM for Gaussian Mixture Model

Gaussian Mixture Model can be written as a linear superposition of Gaussian as

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k}) \tag{12}$$

- K is the number of Gaussian
- π_k is mixing coefficient: weight for each Gaussian distribution, which satisfies $0 \leq \pi_k \leq 1$ and $\sum_{i=1}^{K} \pi_k = 1$.

EM for Gaussian Mixture Model

Let us introduce a K-dimensional binary random variable $\mathbf{z}$ having a 1-of-K representation in which a particular element z_k is equal to 1 and all other elements are equal to 0. The marginal distribution over z_k is

$$p(z_k = 1) = \pi_k \tag{13}$$

and the joint distribution

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{14}$$

The conditional distribution of $\mathbf{x}$ given z_i is a Gaussian

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k}) \tag{15}$$

which can also be written in the form

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k})_k^z \tag{16}$$

EM for Gaussian Mixture Model

The marginal distribution of $\mathbf{x}$ is then obtained by summing the joint distribution over all possible states of $\mathbf{z}$

$$p(\mathbf{x}) = \sum_{\mathbf{x}} p(\mathbf{z})p(\mathbf{x}|\mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_\mathbf{k}, \boldsymbol{\Sigma}_\mathbf{k}) \tag{17}$$

By introducing latent variable $\mathbf{z}$ to derive the Gaussian mixture, we gain

- Work with the joint distribution $p(\mathbf{x}, \mathbf{z})$ instead of the marginal distribution $p(\mathbf{x})$.
- The conditional probability of $\mathbf{z}$ given $\mathbf{x}$, i.e, $\gamma(z_k) = p(z_k = 1|\mathbf{x})$ will play an important role in the later EM algorithm.
- $\gamma(z_k)$ can be viewed as the *responsibility* that component k takes for 'explaining' the observation $\mathbf{x}$.

Maximum likelihood

Suppose we have a i.i.d data set of observations $\{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$, and we wish to model this data using a mixture of Gaussians. Denote the data set as $\mathbf{X} \in \mathcal{R}^{N \times D}$, and the corresponding latent variables $\mathbf{Z} \in \mathcal{R}^{N \times K}$. From (17) the log-likelihood function is given by

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

Maximizing the log-likelihood function (18) is intractable, the main difficulty arises from the presence of the summation over k that appear insider the logarithm in (18). If we set the derivatives of the log-likelihood to zero, we will no longer obtain a closed form solution.

EM for Gaussian Mixture Model

Now consider the problem of maximizing the likelihood for the complete data set $\{\mathbf{X}, \mathbf{Z}\}$, which takes the form

$$p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu}, \boldsymbol{\Sigma_k})^{z_{nk}} \tag{19}$$

Taking the logarithm, we obtain

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \{\ln \pi_k + \ln \mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu}, \boldsymbol{\Sigma_k})\} \tag{20}$$

Compared with the log-likelihood function for incomplete data, i.e., (18), the summation over k and the logarithm have been interchanged in (20). The logarithm now acts directly on the Gaussian distribution.

EM for Gaussian Mixture Model

E-step: Evaluate the *responsibilities* using the current parameter values

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x_n}|\boldsymbol{\mu_j}, \boldsymbol{\Sigma_j})} \tag{21}$$

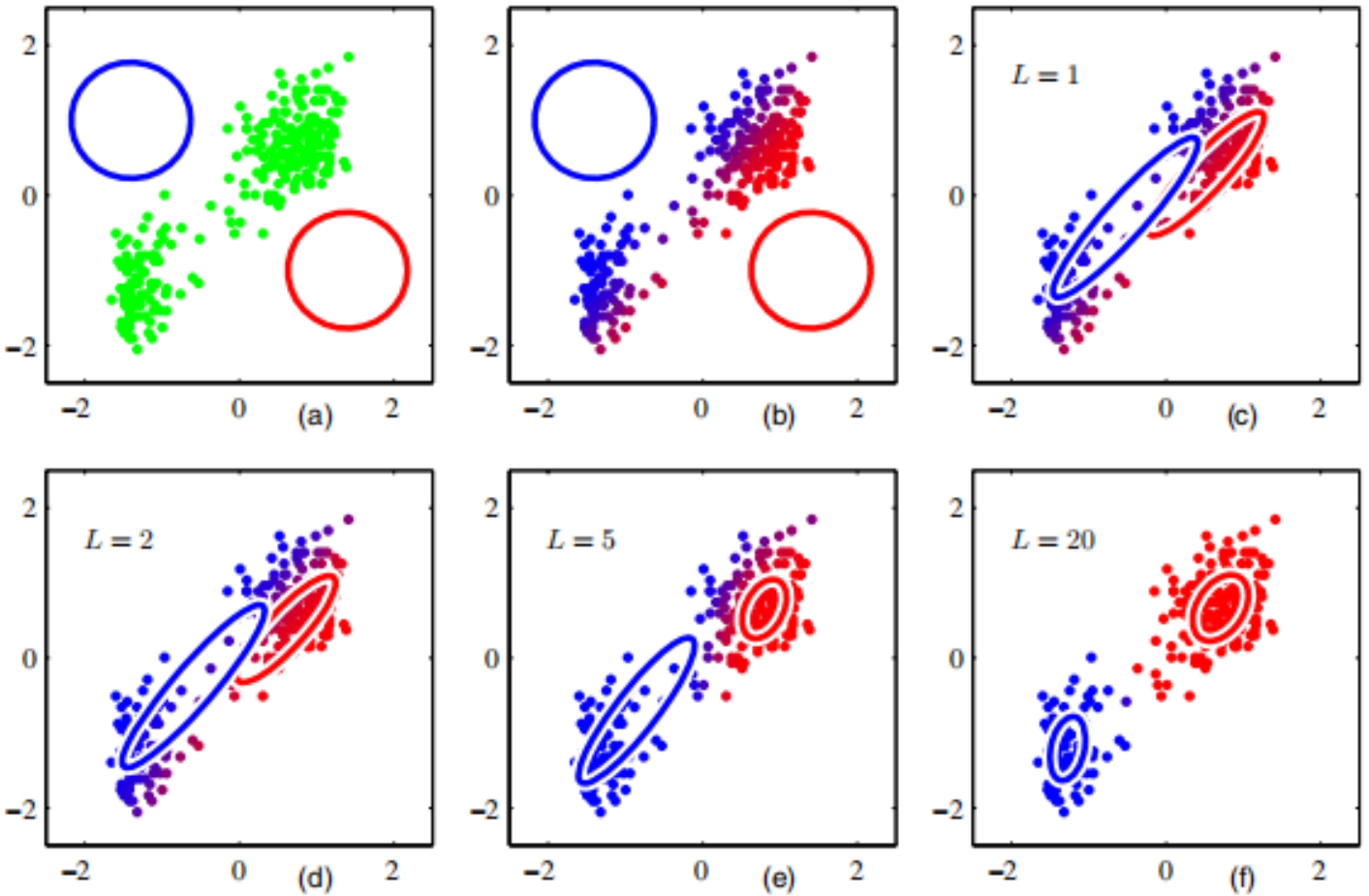
M-step: Re-estimate the parameters using the current responsibilities

$$\mu_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x_n}}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{22}$$

$$\Sigma_k = \frac{\sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x_n} - \mu_k)(\mathbf{x_n} - \mu_k)^T}{\sum_{n=1}^{N} \gamma(z_{nk})} \tag{23}$$

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma(z_{nk}) \tag{24}$$

EM for Gaussian Mixture Model



References


Christopher M.Bishop, Pattern Recognition and Machine Learning, *Spring*, 2006.


Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 1977, 39(1): 1-22.

The End