**University of Illinois** Fall 2018

# ECE 544NA: Pattern Recognition
## Lecture 18: October 25

Lecturer: Alexander Schwing Scribe: Ranvir Rana

## 1 Overview

In this lecture we are going to discuss Maximum Likelihood estimation of parameters of complex statistical models using Expectation-Maximization(EM) algorithm. We describe the general EM algorithm to find parameters for any mixture models and see specific example of the EM algorithm by applying it for Gaussian mixture models. We also explore Concave-Convex procedure (CCCP) and show that CCCP is essentially similar to EM for a general distribution, moreover, we show similarities between inference and EM objectives.

### 1.1 Motivation for EM algorithm

In several practical use cases, the optimal solution of the likelihood is not directly solvable, thus we need some iterative algorithm to converge to an optimal solution. The complexity with solving the ML equation directly arises in case of mixture models with latent variables, for example GMMs. Typically to solve ML equation for mixture models, we have two options: (1)Gradient Descent or (2)Alternating maximization(similar to EM). The advantage of EM is that we have a closed form solution for each step of the iterative procedure, however in case of Gradient descent, we need don't have any closed form solution.

## 2 General EM algorithm

As discussed earlier, the EM algorithm is used to find Maximum likelihood solution for the statistical models that involve latent variables. In this section, we propose a general form of the EM algorithm and show that it converges to the maximum of the likelihood function.

Let $\mathbf{X}$ denote all the observed random variables and $\mathbf{Z}$ denote all hidden latent random variables. Our goal is to find the model parameters $\theta$ that maximize the likelihood of the observed random variables $\mathbf{X}$, i.e.

$$\max_{\theta} p(\mathbf{X}|\theta) \tag{1}$$

or alternatively maximize the log likelihood, i.e.

$$\max_{\theta} \ln p(\mathbf{X}|\theta) \tag{2}$$

Now since the observed random variables are part of the mixture models, they can be best described by the marginals over $\mathbf{Z}$ of the combined distribution over $\mathbf{X}$,$\mathbf{Z}$, i.e.

$$p(\mathbf{X}|\theta) = \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta) \tag{3}$$

In mixture models, the direct optimization of $p(\mathbf{X}|\theta)$ is difficult, however it is easier to optimize over $p(\mathbf{X}, \mathbf{Z}|\theta)$ . However, since we do not observe $\mathbf{Z}$, we cannot directly optimize over $p(\mathbf{X}, \mathbf{Z}|\theta)$. Thus, we define a distribution $q(\mathbf{Z})$ and decompose $p(\mathbf{X}|\theta)$ into two parts:

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(p(\mathbf{X}, \mathbf{Z}|\theta), q(\mathbf{Z})) + D_{KL}(q(\mathbf{Z}), p(\mathbf{Z}|\mathbf{X}, \theta)) \tag{4}$$

where

$$\mathcal{L}(p(\mathbf{X}, \mathbf{Z}|\theta), q(\mathbf{Z})) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \tag{5}$$

$$D_{KL}(q(\mathbf{Z}), p(\mathbf{Z}|\mathbf{X}, \theta)) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{X}, \theta)} \tag{6}$$

where $D_{KL}$ is a Kullback-Leibler divergence. Note that although the equations for $\mathcal{L}$ and $D_{KL}$ look very similar, $\mathcal{L}$ is not a KL divergence as the summation/integral happens over $\mathbf{Z}$ which is not the common domain of both the arguments of the $\mathcal{L}$ functional.

Next, we prove that KL divergence is always non negative. To prove this, we need to use the Jensen's inequality for concave functions $f$ given by:

$$\sum_{\mathbf{Z}} q(\mathbf{Z}) f(g(\mathbf{Z})) \geq f\left(\sum_{\mathbf{Z}} q(\mathbf{Z}) g(\mathbf{Z})\right) \tag{7}$$

from the RHS of the $D_{KL}$ equation, we can see that $g(Z) = \frac{p(\mathbf{Z}|\mathbf{X},\theta)}{q(\mathbf{Z})}$ and $f() = \ln()$ which is a concave function, thus applying Jensen's inequality to the $D_{KL}$ equation, we can see that:

$$D_{KL}(q(\mathbf{Z}), p(\mathbf{Z}|\mathbf{X}, \theta)) = \sum_{\mathbf{Z}} -q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} \tag{8}$$

$$\geq \ln \sum_{\mathbf{Z}} -q(\mathbf{Z}) \frac{p(\mathbf{Z}|\mathbf{X}, \theta)}{q(\mathbf{Z})} = \ln \sum_{\mathcal{Z}} p(\mathbf{Z}|\mathbf{X}, \theta) = 0 \tag{9}$$

Thus:

$$D_{KL} \geq 0 \tag{10}$$

which means that:

$$\ln p(\mathbf{X}|\theta) = \mathcal{L}(p(\mathbf{X}, \mathbf{Z}|\theta), q(\mathbf{Z})) + D_{KL}(q(\mathbf{Z}), p(\mathbf{Z}|\mathbf{X}, \theta)) \geq \mathcal{L}(p(\mathbf{X}, \mathbf{Z}|\theta), q(\mathbf{Z})) \tag{11}$$

thus, $\mathcal{L}(p(\mathbf{X}, \mathbf{Z}|\theta), q(\mathbf{Z}))$ is the lower bound on $\ln p(\mathbf{X}|\theta)$. This decomposition is shown in figure 1

Maximization of log likelihood is difficult, thus we try to maximize the lower bound on the log likelihood, i.e. our new optimization problem is:

$$\max_{q,\theta} \mathcal{L}(p(\mathbf{X}, \mathbf{Z}|\theta), q(\mathbf{Z})) \tag{12}$$

Now, since this optimization involves maximizing over both $q$ and $\theta$, we follow an alternate maximization approach, the two steps of which, will be termed as Estimation and Maximization.

In the first step, we maximize w.r.t. $q$, this is called the Estimation step, since $\mathbf{Z}$ is not observed and we use $q(\mathbf{Z})$ to estimate it. Please refer to figure 1, we see that $\ln p(\mathbf{X}|\theta)$ does not vary with $q(\mathbf{Z})$, thus changing $q(\mathbf{Z})$ changes $\mathcal{L}(p(\mathbf{X}, \mathbf{Z}|\theta), q(\mathbf{Z}))$ and $D_{KL}(q(\mathbf{Z}), p(\mathbf{Z}|\mathbf{X}, \theta))$, thus if we want

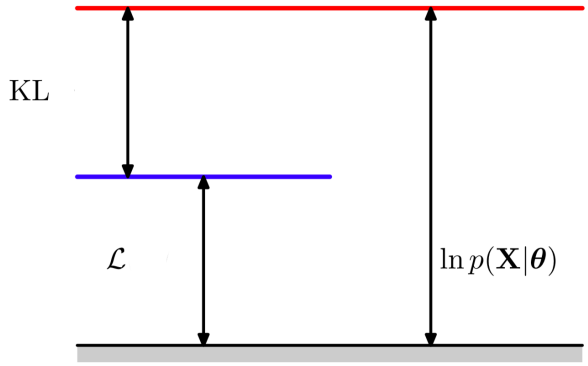


Figure 1: Decomposition of $\ln p(\mathbf{X}|\theta)$ into $\mathcal{L}$ and $D_{KL}$

to maximize $\mathcal{L}$ we want to minimize $D_{KL}$. Now we know that the minimum value of any KL divergence, which is 0, is obtained when the two distributions are that same, i.e. :

$$q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta) \tag{13}$$

which results in $D_{KL}(q(\mathbf{Z}), p(\mathbf{Z}|\mathbf{X}, \theta)) = 0$. This step is shown in figure 2.

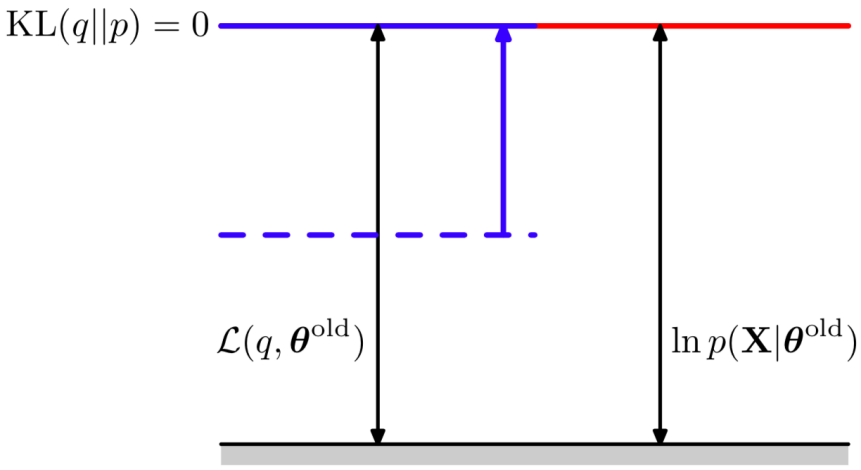


Figure 2: Illustration of E step

Now, the second step is called maximization step, this involves maximization of $\mathcal{L}$ over $\theta$. We can write $\mathcal{L}$ as:

$$\begin{aligned}\mathcal{L}(p(\mathbf{X}, \mathbf{Z}|\theta), q(\mathbf{Z})) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(\mathbf{X}, \mathbf{Z}, \theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln q(\mathbf{Z}) \\ &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \ln p(\mathbf{X}, \mathbf{Z}, \theta) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \ln p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \\ &= \mathcal{Q}(\theta, \theta^{old}) + H(q(\mathbf{Z}))\end{aligned} \tag{14}$$

where

$$\mathcal{Q}(\theta, \theta^{old}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \ln p(\mathbf{X}, \mathbf{Z}, \theta) \tag{15}$$

The second term is independent of $\theta$ thus, is constant w.r.t. this optimization. We now maximize $\mathcal{Q}(\theta, \theta^{old})$, we observe that this term is essentially maximizing the likelihood of the joint distribution $p(\mathbf{X}, \mathbf{Z}|\theta)$. Thus the name of this step is Maximization.

Note that since, we optimize $\mathcal{L}$, $\mathcal{L}(p(\mathbf{X}, \mathbf{Z}|\theta^{new})$ will be more than $\mathcal{L}(p(\mathbf{X}, \mathbf{Z}|\theta^{old})$. Moreover $D_{KL}$ will no longer be zero as $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta^{old}) \neq p(\mathbf{Z}|\mathbf{X}, \theta^{new})$. This step is illustrated in figure 3.

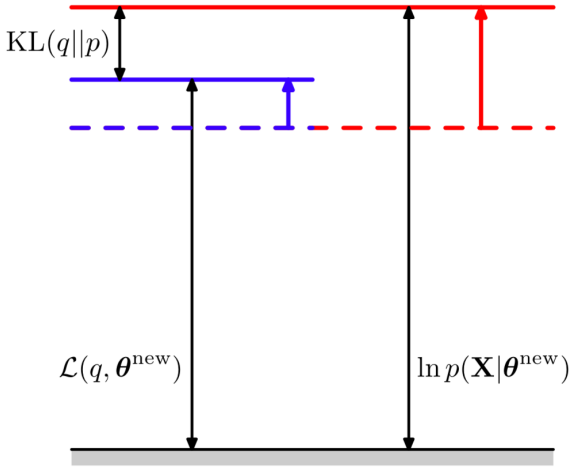


Figure 3: Illustration of M step

Thus, this step will increase the value of both the components $\mathcal{L}$ and $D_{KL}$ of $\ln p(\mathbf{X}|\theta)$, thus increasing the value of $\ln p(\mathbf{X}|\theta)$, unless it is already at its maximum. Thus showing convergence of the EM algorithm.

The complete EM algorithm can be summarized schematically as shown in figure 4. We start with some initial parameter $\theta^{old}$, we perform E step to find $q(\mathbf{Z})$, optimize the $\mathcal{L}(p(\mathbf{X}, \mathbf{Z}|\theta)$ to move from $\theta^{old}$ to $\theta^{new}$, as shown by the blue curve, then we repeat the E and M steps again till convergence. Note that at every step, the $\mathcal{L}$ curve is tangential to the curve of $\ln p(\mathbf{X}|\theta)$.

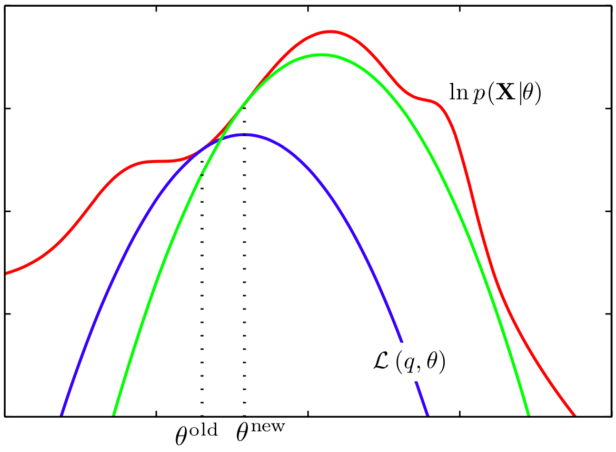


Figure 4: Summary of EM over parameter space

# 3 Similarity of EM with different approaches

## 3.1 Similarity between EM and Concave-Convex procedure

CCCP [2] is used to find out optima of non convex optimization objectives by splitting the concave part of the objective function into convex and concave parts. Before we explain CCCP in detail, let us recap the lower bound optimization used in EM algorithm for distributions in the general form:

$$p(x^{(i)}, \mathbf{Z}|\theta) = \frac{1}{Z(\theta)} \exp F(x^{(i)}, \mathbf{Z}, \theta) \tag{16}$$

where $x^{(i)}$ is a single observed random variable and $Z(\theta)$ is the partition function. We see that the negative of the lower bound on log likelihood $\mathcal{L}$ is given by :

$$-\mathcal{L}(p(x^{(i)}, \mathbf{Z}|\theta), q(\mathbf{Z})) = -\sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(x^{(i)}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \tag{17}$$

$$= \ln(Z(\theta)) - \sum_{\mathbf{Z}} q(\mathbf{Z}) F(x^{(i)}, \mathbf{Z}, \theta) - H(q(\mathbf{Z})) \tag{18}$$

where $H(q(\mathbf{Z}))$ is the entropy of the $q$ distribution. Keeping equation 18 in mind, we proceed to define the convex concave procedure.

We use the same joint distribution model as before, and the same objective to obtain the maximum likelihood of the marginal distribution of $\mathbf{X}$ as shown below:

$$\min_{\theta} - \ln \sum_{\mathbf{Z}} \frac{\exp F(x^{(i)}, \mathbf{Z}, \theta)}{Z(\theta)} \tag{19}$$

$$\min_{\theta} \quad \ln Z(\theta) \quad - \ln \sum_{\mathbf{Z}} \exp F(x^{(i)}, \mathbf{Z}, \theta) \tag{20}$$

We observe that the first term $\ln Z(\theta)$ is convex if $F$ is linear in $\theta$, and the second term $-\ln \sum_{\mathbf{Z}} \exp F(x^{(i)}, \mathbf{Z}, \theta)$ is concave if $F$ is linear in $\theta$.

The CCCP can be summarized as shown by algorithm 1.

**Algorithm 1** CCCP

```
1: procedure CCCP
2:     Initialize θ
3:     while growth rate ≥ threshold do
4:         Decompose concave part into 'convex+concave' parts at current θ
5:         Solve the convex program
```

Now, as we showed for the negative log likelihood minimization, we have two parts of the equation 18: convex and concave. We try to decompose the concave part of the equation by introducing a latent variable distribution $q(\mathbf{Z})$ as follows:

$$\ln \sum_{\mathbf{Z}} \exp F(x^{(i)}, \mathbf{Z}, \theta) = \ln \sum_{\mathbf{Z}} \frac{\exp F(x^{(i)}, \mathbf{Z}, \theta)}{q(\mathbf{Z})} \tag{21}$$

Now, we apply Jensen's inequality to the RHS of equation 21, we see that

$$\ln \sum_{\mathbf{Z}} \frac{\exp F(x^{(i)}, \mathbf{Z}, \theta)}{q(\mathbf{Z})} \geq \sum_{\mathbf{Z}} q(\mathbf{Z}) F(x^{(i)}, \mathbf{Z}, \theta) + H(q(\mathbf{Z})) \tag{22}$$

now, we can reduce the inequality to an equality in equation 22 and combine it with equation 21 to obtain:

$$\ln \sum_{\mathbf{Z}} \exp F(x^{(i)}, \mathbf{Z}, \theta) = \max_{q(\mathbf{Z})} \left( \sum_{\mathbf{Z}} q(\mathbf{Z}) F(x^{(i)}, \mathbf{Z}, \theta) + H(q(\mathbf{Z})) \right) \tag{23}$$

If we combine this update in equation 23 to the minimization in 20, we observe that the updated minimization in the CCCP repeat loop is given by:

$$\min_{\theta, q} \ln Z(\theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) F(x^{(i)}, \mathbf{Z}, \theta) - H(q(\mathbf{Z})) \tag{24}$$

we observe that this optimization objective is the same objective obtained by the EM algorithm as shown in 18. Thus, we have shown that CCCP is similar to EM for these set of general distributions and ML estimation.

### 3.2 Similarity with variational inference

Let us take a closer look at the general inference problem, where we need to infer the value of the random variable $\mathbf{y}$ under a model with parameters $\mathbf{w}$, our objective to find $\mathbf{y}$ which optimizes $F()$ is given by:

$$\mathsf{arg}\ \max_{\hat{y}} F(x^{(i)}, \hat{\mathbf{y}}, \mathbf{w}) \tag{25}$$

now, we split into $r$ different components, making the objective function

$$\mathsf{arg}\ \max_{\hat{y}} \sum_{r} f_r(x^{(i)}, \hat{\mathbf{y}}_\mathbf{r}, \mathbf{w}) \tag{26}$$

Finding appropriate $\hat{\mathbf{y}}$ is similar to finding a distribution $p(\hat{\mathbf{y}})$ on $\hat{\mathbf{y}}$ to optimize the objective function.

Thus our updated objective function takes the form:

$$\mathsf{arg}\ \max_{\hat{y}} \sum_{\hat{\mathbf{y}}} p(\hat{\mathbf{y}}) \sum_{r} f_r(x^{(i)}, \hat{\mathbf{y}}_\mathbf{r}, \mathbf{w}) \tag{27}$$

marginalizing the above equation and taking into account various marginalization constraints, we get the objective function:

$$\mathsf{arg}\ \max_{\hat{y}} \sum_{r} \sum_{\hat{\mathbf{y}}} p(\hat{\mathbf{y}}) f_r(x^{(i)}, \hat{\mathbf{y}}_\mathbf{r}, \mathbf{w}) \tag{28}$$

We can see that the above equation takes a form very similar to the objective function used in the Expectation Maximization algorithm.

# 4 EM for GMM's

In this section we take a look at an example of Expectation Maximization algorithm applied to Gaussian mixture models. The mixture model consists of observed random variables $\mathbf{X}$ and latent random variables $\mathbf{Z}$, the joint distribution is given by:

$$p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi) = \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{n,k}} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{n,k}} \tag{29}$$

where $K$ is the total number of components of $\mathbf{z}_n$, $N$ is the total number of observed data points, $z_{n,k}$ is the $k^{th}$ component of $\mathbf{z}^n$, $x_n$ is the $n^{th}$ data point, $\pi$ is the mixture distribution, and $\mu$ and $\Sigma$ are the mean and covarience matrices, with $\mu_k$ and $\Sigma_k$ representing the $k^{th}$ gaussian distribution. A graphical representation of this model is given in figure 5.

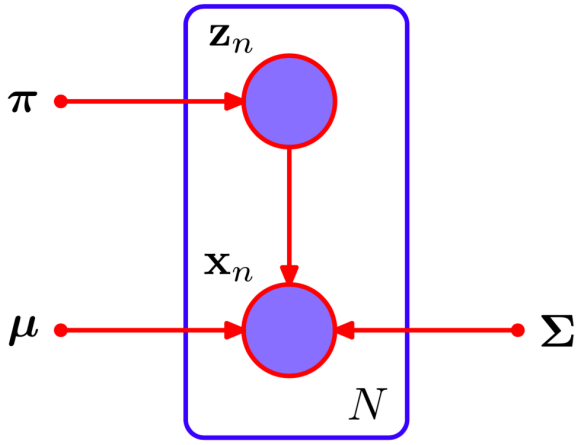


Figure 5: Graphical representation of GMM

Taking the logarithm, we get

$$\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} (\ln \pi_k + \ln \mathcal{N}(x_n|\mu_k, \Sigma_k)) \tag{30}$$

We can see from equation 30 that the logarithm is inside the summation and hence is easy to optimize, however, we don't have the value for the latent variables, hence the expectation step is needed. More concretely, let us define the lower bound $\mathcal{L}$ defined in the generalized EM earlier.

$$\mathcal{L}(p(x_n, \mathbf{Z}|\theta), q(\mathbf{Z})) = \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln p(x_n, \mathbf{Z}, \theta) - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln q(\mathbf{Z}) \tag{31}$$

$$= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \prod_{k=1}^{K} \pi_k^{z_{n,k}} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{n,k}} + H(q(\mathbf{Z})) \tag{32}$$

$$= \sum_{\mathbf{Z}} \sum_{k=1}^{K} q(\mathbf{Z}) \ln \pi_k^{z_{n,k}} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{n,k}} + H(q(\mathbf{Z})) \tag{33}$$

we can see from the above equation 33 is easy to optimize once we calculate the estimation of $z_{n,k}$ by setting

$$q(\mathbf{Z} = p(\mathbf{Z}|\mathbf{X}, \mu, \Sigma, \pi) \tag{34}$$

now, we can derive $p(\mathbf{Z}|\mathbf{X}, \mu, \Sigma, \pi)$ as follows:

$$p(\mathbf{Z}|\mathbf{X}, \mu, \Sigma, \pi) = \frac{p(\mathbf{Z}, \mathbf{X}|\mu, \Sigma, \pi)}{p(\mathbf{X}|\mu, \Sigma, \pi)} \tag{35}$$

note that our aim is to only calculate $\mathbb{E}[z_{nk}]$ which is an expectation of an indicator random variable $z_{nk}$, thus we need not calculate $p(\mathbf{X}|\mu, \Sigma, \pi)$, thus, it suffices to say that:

$$p(\mathbf{Z}|\mathbf{X}, \mu, \Sigma, \pi) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{n,k}} \mathcal{N}(x_n|\mu_k, \Sigma_k)^{z_{n,k}} \tag{36}$$

This gives us $\mathbb{E}[z_{nk}]$ as follows:

$$\mathbb{E}[z_{nk}] = \frac{\sum_{z_{nk}} z_{nk}[\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)]^{z_{nk}}}{\sum_{z_{nj}} [\pi_j \mathcal{N}(\mathbf{x}_n|\mu_j, \Sigma_j)]^{z_{nj}}} \tag{37}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\mu_j, \Sigma_j)} = \gamma(z_{nk}) \tag{38}$$

We can use equation 38 to calculate the expectation of the first term $\mathcal{L}$ functional which the expected value of the log likelihood of the joint distribution given by:

$$\mathbb{E}[\ln p(\mathbf{X}, \mathbf{Z}|\mu, \Sigma, \pi)] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk})(\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)) \tag{39}$$

We can calculate the derivative of this equation to calculate the optimal parameters for the given estimate of $z_{nk}$, updating $\mu^{old}$, $\Sigma^{old}$ and $\pi^{old}$, to the new parameters $\mu^{new}$, $\Sigma^{new}$ and $\pi^{new}$ using equations 41, 42, 43 respectively.

Let us define $N_k$ as follows:

$$N_k = \sum_{n=1}^{N} \gamma(z_{nk}) \tag{40}$$

then,

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n \tag{41}$$

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T \tag{42}$$

$$\pi_k = \frac{N_k}{N} \tag{43}$$

The iterations are terminated when the growth rate of the log likelihood defined in equation 44 falls below a threshold.

$$\ln p(\mathbf{X}|\mu, \Sigma, \pi) = \sum_{n=1}^{N} \ln \left( \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k) \right) \tag{44}$$

The implementation of EM algorithm for GMM on Old faithful dataset is illustrated in figure 6.

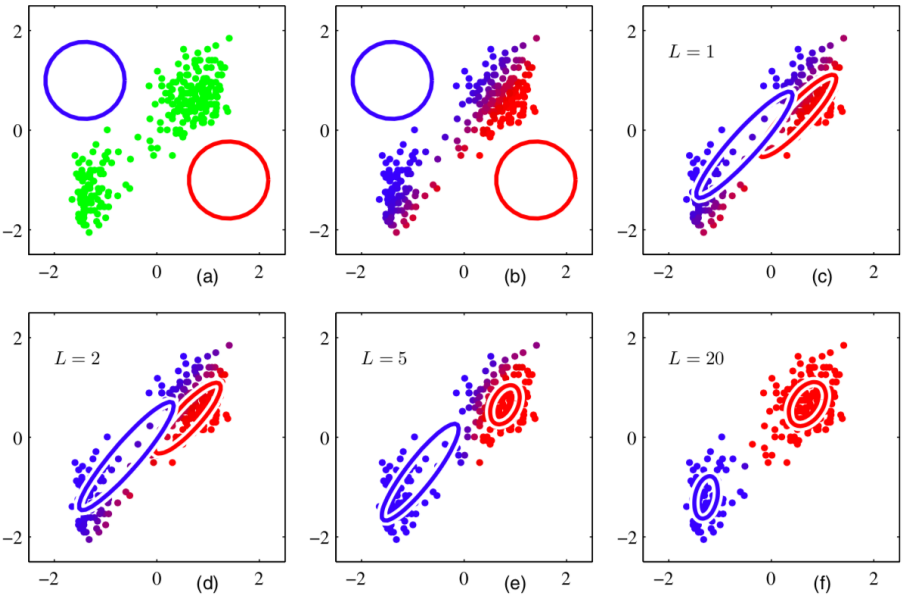


Figure 6: EM algorithm on Old Faithful dataset

For further reference to EM algorithms, please refer to Chapter 9 of [1]

## References


[1] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[2] A. L. Yuille and A. Rangarajan. The concave-convex procedure. *Neural computation*, 15(4):915–936, 2003.