Gaussian Mixture Model and EM Algorithm

Hints before Reading

In this note, we are going to discuss Gaussian mixture model and expectation-maximization (EM) algorithm. In order to understand the materials in this note, you will need to:

- understand multivariate Gaussian distribution, which is covered in Lecture 15, Part 9.
- understand the definition of mixture model, which is covered in Lecture 16, Part 3.

The theory behind Gaussian mixture and EM algorithm is not easy. To explain the theory clearly, it requires many derivations in math. As a result, this note will be dense in math and have many equations. To make the equations comprehensible, we show step-by-step derivations in this note. When you see an equation with multiple lines, don't be afraid of the length, because those lines consists of the step-by-step derivation details, which should be helpful to your understanding.

Material in this note can be challenging. If you want to understand everything in this note, you might need to read this note multiple times. Therefore, don't feel frustrated if you could not understand everything after the first time you go through the note.

Parameters in Gaussian Mixture Model

We assume that each instance of our data is a d-dimensional vector, which we denote as X, where $X \in \mathbb{R}^d$. We also assume that our data vector X is generated from a Gaussian Mixture Model (GMM) with K mixture components. This GMM is described by following parameters:

- K: the number of mixture components.
- $\mathbf{p} = \{p_1, p_2, ..., p_c, ..., p_K\}$: mixture weights, where $\sum_{c=1}^{K} p_c = 1$.
- A d-dimensional Gaussian PDF, $\mathcal{N}(\boldsymbol{\mu_c}, \sigma_c^2\mathbf{I})$ for every $c = 1, 2, ..., K$, where
  - $\boldsymbol{\mu_c} \in \mathbb{R}^d$ is the mean vector of the c^{th} Gaussian PDF;
  - $\sigma_c^2 \in \mathbb{R}$ is the variance of the c^{th} Gaussian PDF;
  - $\mathbf{I} \in \mathbb{R}^{d\times d}$ is a $d \times d$ identity matrix.

To compute the likelihood of data X under the c^{th} Gaussian PDF, we have:

$$\begin{aligned} P(X|\boldsymbol{\mu_c}, \sigma_c^2) &= \mathcal{N}(X;\, \boldsymbol{\mu_c}\,, \sigma_c^2\mathbf{I}) \\ &= \frac{1}{(2\pi)^{\frac{d}{2}}} \frac{1}{\det(\sigma_c^2\mathbf{I})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(X-\boldsymbol{\mu_c})^T(\sigma_c^2\mathbf{I})^{-1}(X-\boldsymbol{\mu_c})\right). \end{aligned}$$

Since X and $\boldsymbol{\mu_c}$ are d-dimensional vectors. We have $X = \{X_1, X_2, ..., X_i, ..., X_d\}$ and $\boldsymbol{\mu_c} = \{\mu_{c1},\ \mu_{c2},\ ...,\ \mu_{ci},\ ...,\ \mu_{cd}\}$, where X_i and μ_{ci} are the i^{th} entry of

X and $\boldsymbol{\mu_c}$ respectively. By using this notation to simplify the above equation, we have:

$$P(X|\boldsymbol{\mu_c}, \sigma_c^2) = (2\pi)^{-\frac{d}{2}} \det(\sigma_c^2\mathbf{I})^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{d}\frac{(X_i - \mu_{ci})^2}{\sigma_c^2}\right) \tag{1}$$

$$= (2\pi)^{-\frac{d}{2}} \left(\sigma_c^2\right)^{-\frac{d}{2}} \exp\left(-\frac{1}{2\sigma_c^2}\sum_{i=1}^{d}(X_i - \mu_{ci})^2\right). \tag{2}$$

Data Likelihood in Gaussian Mixture Model

To compute the likelihood of one data sample X under the GMM, we have:

$$P(X|\mathbf{p}, \mu, \boldsymbol{\sigma}^2) = \sum_{c=1}^{K} p_c \; P(X|\boldsymbol{\mu_c}, \sigma_c^2), \tag{3}$$

where $P(X|\boldsymbol{\mu_c}, \sigma_c^2)$ is defined in Equation 2 for $c = 1, 2, ..., K$.

Data Likelihood in Gaussian Mixture Model (Multiple Samples)

If we have a dataset D, which consists of N samples, we can represent this dataset by: $D = \{X_1, X_2, ..., X_j, ..., X_N\}$, where $X_j \in \mathbb{R}^d$ is the j^{th} sample in the dataset. In addition, we have $X_j = \{X_{j1}, X_{j2}, ..., X_{ji}, ..., X_{jd}\}$, where X_{ji} is the i^{th} entry in data X_j. The likelihood of dataset D under the GMM can be represented by:

$$P(D|\mathbf{p}, \mu, \boldsymbol{\sigma}^2) = P(X_1, X_2, ..., X_j, ..., X_N|\mathbf{p}, \mu, \boldsymbol{\sigma}^2)$$

We assume that samples in our data set D are independent and identically distributed. Based on this assumption, we have:

$$P(D|\mathbf{p}, \mu, \boldsymbol{\sigma}^2) = \prod_{j=1}^{N} P(X_j|\mathbf{p}, \mu, \boldsymbol{\sigma}^2)$$

According to on Equation 3, we have:

$$P(D|\mathbf{p}, \mu, \boldsymbol{\sigma}^2) = \prod_{j=1}^{N} P(X_j|\mathbf{p}, \mu, \boldsymbol{\sigma}^2) \tag{4}$$

$$= \prod_{j=1}^{N}\left(\sum_{c=1}^{K} p_c \; P(X_j|\boldsymbol{\mu_c}, \sigma_c^2)\right). \tag{5}$$

EM Algorithm

In this section, we will explain the theory behind the EM algorithm in detailed. Materials in this section are not included in Lecture 16, Part 5 for simplicity. Materials in this section are supplementary components for the lecture.

This section will be dense in math. When it comes to derivations in math, we will try our best to provide step-by-step derivation and we will provide comments in lightgray color to explain the reasons or mathematical theorems we use to go from one step to the next.

Maximum Likelihood Estimation is Intractable.

According to our earlier discussion, we have the data likelihood under GMM in Equation 5. If we apply log function on both sides Equation 5, we will get:

$$\log\left(P(D|\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)\right) = \sum_{j=1}^{N} \log\left(\sum_{c=1}^{K} p_c \; P(X_j|\boldsymbol{\mu}_c,\sigma_c^2)\right). \tag{6}$$

To find the optimal parameter in GMM, a straightforward approach is to apply maximum (log) likelihood estimation (MLE): given Equation 6, we take the partial derivative with respect to $\boldsymbol{\mu}_c, \sigma_c^2$ and p_c, set them to zero and solve for the optimal $\boldsymbol{\mu}_c, \sigma_c^2$ and p_c respectively for $c = 1, 2, ..., K$. Unfortunately, when we try to compute those partial derivatives based on Equation 6, we will encounter troubles: ideally, inside the log function, we would like to have product of terms, so that we can use the property of log function ($\log(ab) = \log a + \log b$) to separate different terms related to different c. However, in Equation 6, we have sum of terms inside the log function, rather than the product. This will make computation of MLE become difficult.

Approximation to Log Likelihood Function.

As we discussed from the above section, MLE computation is difficult. In this section, we are trying to come up with an approximation to the log likelihood function and hope this approximation can result in simple computation. Given the left hand side of Equation 6, we have:

$$\log\left(P(D|\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)\right) = \log\left(\sum_{y} P(D,y|\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)\right), \tag{7}$$

where y is an arbitrary random variable and $P(D,y|\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)$ is the joint distribution of our data D and the random variable y. The expression inside the log function on the right hand side is to compute the marginal distribution of our data D, $P(D|\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)$, from the joint distribution $P(D,y|\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)$. Given Equation 7, we have:

$$\log\left(P(D|\mathbf{p},\mu,\sigma^2)\right) = \log\left(\sum_y q(y|D)\frac{P(D,y|\mathbf{p},\mu,\sigma^2)}{q(y|D)}\right) \tag{8}$$

$$= \log\left(\mathbb{E}_{y\sim q(y|D)}\left[\frac{P(D,y|\mathbf{p},\mu,\sigma^2)}{q(y|D)}\right]\right), \tag{9}$$

where $q(y|D)$ is the conditional distribution distribution of y given the our data D.

Here is one question you might have: why do we want to introduce an extra random variable y in Equation 7 ? In fact, this is an important trick in EM algorithm. y is a random variable and there should be a probability distribution corresponding to y. Because y is an arbitrary random variable, we can put an arbitrary probability distribution on y, as long as it does not violate the definition of probability distribution. Here, we denote $q(y|D)$ in Equation 9 to be the probability distribution on y. Remember that our goal is to come up with an approximation to the log likelihood function. If we can choose a "nice" mathematical form for $q(y|D)$, it would bring convenience in computation when we try to approximate the log likelihood function. We will discuss how to choose a "nice" $q(y|D)$ in the later subsection of this note

We should continue our discussion based on Equation 9. According to Jensen's inequality, since log function is concave, we have $\log\left(\mathbb{E}[f(y)]\right) \geq \mathbb{E}[\log f(y)]$. By applying this relationship to Equation 9, we have:

$$\log\left(P(D|\mathbf{p},\mu,\sigma^2)\right) = \log\left(\mathbb{E}_{y\sim q(y|D)}\left[\frac{P(D,y|\mathbf{p},\mu,\sigma^2)}{q(y|D)}\right]\right) \tag{10}$$

$$\geq \mathbb{E}_{y\sim q(y|D)}\left[\log\left(\frac{P(D,y|\mathbf{p},\mu,\sigma^2)}{q(y|D)}\right)\right] \tag{11}$$

According to Equation 11, to maximize the log likelihood of our data, we need to maximize the lower bound $\mathbb{E}_{y\sim q(y|D)}\left[\log\left(\frac{P(D,y|\mathbf{p},\mu,\sigma^2)}{q(y|D)}\right)\right]$. In order to maximize the lower bound, we need to: (1) choose appropriate values for parameters $\mathbf{p},\mu,\sigma^2$; (2) choose "nice" function $q(y|D)$.

Optimal Choice for Parameters

In this step, we fix the function $q(y|D)$ and try to find optimal values of $\mathbf{p},\mu,\sigma^2$, which maximize the lower bound. We denote the optimal solution parameters as $\mathbf{p}_*,\mu_*,\sigma^2_*$. According to Equation 11, we have:

$$\mathbf{p}_*, \boldsymbol{\mu}_*, \boldsymbol{\sigma}^2_* = \arg\max_{\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2} \mathbb{E}_{y\sim q(y|D)}\left[\log\left(\frac{P(D,y|\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)}{q(y|D)}\right)\right] \tag{12}$$

// use the rule of log: $\log\frac{A}{B} = \log A - \log B$

// use the linearity of expectation

$$= \arg\max_{\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2}\left(\mathbb{E}_{y\sim q(y|D)}\left[\log P(D,y|\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)\right] - \mathbb{E}_{y\sim q(y|D)}\left[\log q(y|D)\right]\right) \tag{13}$$

Since the second term in the equation above is not related to $\mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2$, we have:

$$\mathbf{p}_*, \boldsymbol{\mu}_*, \boldsymbol{\sigma}^2_* = \arg\max_{\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2}\left(\mathbb{E}_{y\sim q(y|D)}\left[\log P(D,y|\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)\right]\right) \tag{14}$$

Optimal Choice for Function q

As mentioned above, we have the freedom to choose the function $q(y|D)$. Here, we would like to choose an appropriate form of $q(y|D)$, so that the lower bound in Equation 11 is maximized.

If we choose $q(y|D) = P(y|D, \mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, the lower bound would become:

$$\mathbb{E}_{y\sim q(y|D)}\left[\log\left(\frac{P(D,y|\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)}{q(y|D)}\right)\right] \tag{15}$$

// replace $q(y|D)$ with $P(y|D, \mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$

$$= \mathbb{E}_{y\sim P(y|D,\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)}\left[\log\left(\frac{P(D,y|\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)}{P(y|D,\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)}\right)\right] \tag{16}$$

// use the definition of expectation

$$= \sum_y P(y|D,\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)\log\left(\frac{P(D,y|\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)}{P(y|D,\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)}\right) \tag{17}$$

// use the definition of conditional probability on $P(D, y|\mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$

$$= \sum_y P(y|D,\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)\log\left(\frac{P(y|D,\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)P(D|\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)}{P(y|D,\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)}\right) \tag{18}$$

// simplify terms in log() by canceling out $P(y|D, \mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$

$$= \sum_y P(y|D,\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)\log\left(P(D|\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)\right) \tag{19}$$

// factor out the log() terms, since it does not related to y

// property of PDF: $\sum_y P(y|D, \mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = 1$

$$= \log\left(P(D|\mathbf{p},\boldsymbol{\mu},\boldsymbol{\sigma}^2)\right) \tag{20}$$

In Equation 11, we notice that $\log\left(P(D|\mathbf{p},\mu,\sigma^2)\right) \geq \mathbb{E}_{y\sim q(y|D)}\left[\log\left(\frac{P(D,y|\mathbf{p},\mu,\sigma^2)}{q(y|D)}\right)\right]$. If we choose $q(y|D) = P(y|D,\mathbf{p},\mu,\sigma^2)$, it brings the equal sign to above relationship, which maximize the lower bound. Therefore, $q(y|D) = P(y|D,\mathbf{p},\mu,\sigma^2)$ is the optimal choice for function $q(y|D)$.

EM Algorithm

Before introducing EM algorithm in details, we should summarize our analysis so far:

- Given data D, we want to fit a GMM model, but the straightforward MLE method is undesirable.
- Based on the log likelihood function of GMM, we come up with a lower bound, shown in Equation 11. We are going to find the GMM parameters by maximizing the lower bound, rather than using MLE.
- To maximize the lower bound, we need to choose an appropriate function $q(y|D)$ and compute the optimal parameters $\mathbf{p}_*, \mu_*, \sigma^2_*$.
- The optimal choice for $q(y|D)$ is to set $q(y|D) = P(y|D,\mathbf{p},\mu,\sigma^2)$, as our discussion in Section .
- Given $q(y|D)$ we can compute the optimal parameters $\mathbf{p}_*, \mu_*, \sigma^2_*$ by Equation 14. More specifically, since we know that $q(y|D) = P(y|D,\mathbf{p},\mu,\sigma^2)$, we have:

$$\begin{aligned}
\mathbf{p}_*, \mu_*, \sigma^2_* &= \underset{\mathbf{p},\mu,\sigma^2}{\arg\max}\left(\mathbb{E}_{y\sim q(y|D)}\left[\log P(D,y|\mathbf{p},\mu,\sigma^2)\right]\right) && (21)\\
&= \underset{\mathbf{p},\mu,\sigma^2}{\arg\max}\left(\mathbb{E}_{y\sim P(y|D,\mathbf{p},\mu,\sigma^2)}\left[\log P(D,y|\mathbf{p},\mu,\sigma^2)\right]\right) && (22)
\end{aligned}$$

There are several clarifications we need to address. Firstly, in Equation 7, we mentioned that y is an arbitrary random variable. In other words, we are free to choose the meaning of y based on the context. In GMM, we can choose y to be a vector consisting of "labels" as follows:

$$y = [y_1, y_2, ..., y_j, ..., y_N], \tag{23}$$

where $y_j \in \{1,2,...,K\}$ for $\forall j \in \{1,2,..,N\}$. When we have $y_j = c$, it means that our data sample X_j is assigned to cluster c.

In addition, $P(X_j|y_j = c,\mathbf{p},\mu,\sigma^2)$ represents the likelihood of observing data

X_j under cluster c:

$$
\begin{aligned}
P(X_j|y_j = c, \mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = P(X_j|\boldsymbol{\mu_c}, \sigma_c^2)& \\
&// \text{ c is the cluster assignment indicator. Once c is given,} \\
&// \text{ we can directly determine which cluster does data } X_j \text{ come from,} \\
&// \text{ without relying on } \mathbf{p}. \\
&// \text{ In other words, } X_j \text{ and } \mathbf{p} \text{ are conditional independent given c.} \\
&// \text{ If c is not given, we need to rely on } \mathbf{p} \text{ to compute the probability} \\
&// \text{ of assigning data } X_j \text{ to cluster c.} \\
= \mathcal{N}(X_j \ ; \ \boldsymbol{\mu_c}, \sigma_c^2 \mathbf{I}).& \qquad (24)
\end{aligned}
$$

$P(y_j = c| \ \mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ represents the probability of observing cluster c in the GMM:

$$
P(y_j = c| \ \mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = p_c \tag{25}
$$

Secondly, Equation 14 is derived based on the fact that $q(y|D)$ is not related to parameters $\mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2$. On the other hand, we also say that $q(y|D) = P(y|D, \mathbf{p}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ is the optimal choice for $q(y|D)$, which looks contradictory. In fact, EM algorithm is an iterative algorithm. In iteration $t+1$, we evaluate $q(y|D)$ using parameters from previous iteration, which means $q(y|D) = P(y|D, \ {}^{t}\mathbf{p}, \ {}^{t}\boldsymbol{\mu}, \ {}^{t}\boldsymbol{\sigma}^2)$. When we compute the optimal parameters, we use:

$$
\begin{aligned}
&{}^{t+1}\mathbf{p}_*, \ {}^{t+1}\boldsymbol{\mu}_*, \ {}^{t+1}\boldsymbol{\sigma}^2_* \\
&= \underset{{}^{t+1}\mathbf{p}, \ {}^{t+1}\boldsymbol{\mu}, \ {}^{t+1}\boldsymbol{\sigma}^2}{\arg\max} \left(\mathbb{E}_{y \sim P(y|D, \ {}^{t}\mathbf{p}, \ {}^{t}\boldsymbol{\mu}, \ {}^{t}\boldsymbol{\sigma}^2)} \left[\log P(D, y| \ {}^{t+1}\mathbf{p}, \ {}^{t+1}\boldsymbol{\mu}, \ {}^{t+1}\boldsymbol{\sigma}^2) \right] \right).
\end{aligned}
\tag{26}
$$

Since $q(y|D) = P(y|D, \ {}^{t}\mathbf{p}, \ {}^{t}\boldsymbol{\mu}, \ {}^{t}\boldsymbol{\sigma}^2)$ and it is not related to ${}^{t+1}\mathbf{p}, \ {}^{t+1}\boldsymbol{\mu}, \ {}^{t+1}\boldsymbol{\sigma}^2$, therefore, it is consistent with the result in Equation 14.

E-Step in EM Algorithm

In E-Step of the EM Algorithm, parameters ${}^{t}\mathbf{p}, \ {}^{t}\boldsymbol{\mu}, \ {}^{t}\boldsymbol{\sigma}^2$ are given. We are going to compute $P(y|D, \ {}^{t}\mathbf{p}, \ {}^{t}\boldsymbol{\mu}, \ {}^{t}\boldsymbol{\sigma}^2)$. Given that $y = [y_1, ..., y_j, ..., y_N]$ and $D = \{X_1, X_2, ..., X_j, ..., X_N\}$, if we expend $P(y|D, \ {}^{t}\mathbf{p}, \ {}^{t}\boldsymbol{\mu}, \ {}^{t}\boldsymbol{\sigma}^2)$, we will get:

$$
\begin{aligned}
&P(y|D, \ {}^{t}\mathbf{p}, \ {}^{t}\boldsymbol{\mu}, \ {}^{t}\boldsymbol{\sigma}^2) \\
&= P(y_1, ..., y_j, ..., y_N|X_1, ..., X_j, ..., X_N, \ {}^{t}\mathbf{p}, \ {}^{t}\boldsymbol{\mu}, \ {}^{t}\boldsymbol{\sigma}^2).
\end{aligned}
$$

Given parameter ${}^{t}\mathbf{p}$, y_j and $y_{j'}$ are independent with each other, for $j \neq j'$; y_j

and $X_{j'}$ are independent with each other, for $j \neq j'$. Therefore, we have:

$$P(y|D,\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2) = \prod_{j=1}^{N} P(y_j|X_j,\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2).$$

Remember from Equation 23, $y_j \in \{1, 2, ..., K\}$. We need to compute $P(y_j = c|X_j,\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2)$ for $\forall c = \{1, 2, ..., K\}$. By applying Bayes' rule and use information from Equation 24 and 25, we can get:

$$\begin{aligned}
&P(y_j = c|X_j,\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2) \\
&\qquad\qquad \text{// Apply Bayes' Rule.} \\
&= \frac{P(X_j|y_j = c,\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2)P(y_j = c|\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2)}{\sum_{b=1}^{K} P(X_j|y_j = b,\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2)P(y_j = b|\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2)} \\
&\qquad\qquad \text{// Refer to Equation 24 for } P(X_j|y_j = c,\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2) \\
&\qquad\qquad \text{// Refer to Equation 25 for } P(y_j = c|\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2) \\
&= \frac{\mathcal{N}(X_j\ ;\ {}^t\boldsymbol{\mu}_c,\ {}^t\sigma_c^2\mathbf{I}) \cdot\ {}^tp_c}{\sum_{b=1}^{K}\ \mathcal{N}(X_j\ ;\ {}^t\boldsymbol{\mu}_b,\ {}^t\sigma_b^2\mathbf{I}) \cdot\ {}^tp_b} \quad \text{for } \forall c \in \{1, 2, ..., K\} \text{ and } \forall j \in \{1, 2, ..., N\}.
\end{aligned}$$

Since the final result is not explicitly related to y_i, we can rewrite the equation above for simplicity:

$$P(c|X_j,\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2) = \frac{\mathcal{N}(X_j\ ;\ {}^t\boldsymbol{\mu}_c,\ {}^t\sigma_c^2\mathbf{I}) \cdot\ {}^tp_c}{\sum_{b=1}^{K}\ \mathcal{N}(X_j\ ;\ {}^t\boldsymbol{\mu}_b,\ {}^t\sigma_b^2\mathbf{I}) \cdot\ {}^tp_b} \tag{27}$$
$$\text{for } \forall c \in \{1, 2, ..., K\} \text{ and } \forall j \in \{1, 2, ..., N\}.$$

M-Step in EM Algorithm

In M-Step of the EM Algorithm, $P(y|D,\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2)$ is given. We need to compute optimal solutions for ${}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^2$, which maximize the object function shown in Equation 26. Here, we need to analyze the objective function in details:

$$\begin{aligned}
&\mathbb{E}_{y\sim P(y|D,\,{}^{t}\mathbf{p},\,{}^{t}\boldsymbol{\mu},\,{}^{t}\boldsymbol{\sigma}^2)}\left[\log P(D,y|\,{}^{t+1}\mathbf{p},\,{}^{t+1}\boldsymbol{\mu},\,{}^{t+1}\boldsymbol{\sigma}^2)\right]\\
&\qquad\qquad // \; D=\{X_1,...,X_j,...,X_N\}\\
&\qquad\qquad // \; y=[y_1,...,y_j,...,y_N]\\
&=\mathbb{E}_{y\sim P(y|D,\,{}^{t}\mathbf{p},\,{}^{t}\boldsymbol{\mu},\,{}^{t}\boldsymbol{\sigma}^2)}\left[\log P(X_1,...,X_j,...,X_N,y_1,...,y_j,...y_N|\,{}^{t+1}\mathbf{p},\,{}^{t+1}\boldsymbol{\mu},\,{}^{t+1}\boldsymbol{\sigma}^2)\right]\\
&\qquad\qquad // \text{ Use the independent assumption:}\\
&\qquad\qquad // \text{ A data sample and its label } (X_j,y_j)\\
&\qquad\qquad // \text{ are independent of other } (X_{j'},y_{j'}) \text{ for } j\neq j^{'}\\
&=\mathbb{E}_{y\sim P(y|D,\,{}^{t}\mathbf{p},\,{}^{t}\boldsymbol{\mu},\,{}^{t}\boldsymbol{\sigma}^2)}\left[\log\left(\prod_{j=1}^{N}P(X_j,y_j|\,{}^{t+1}\mathbf{p},\,{}^{t+1}\boldsymbol{\mu},\,{}^{t+1}\boldsymbol{\sigma}^2)\right)\right]\\
&\qquad\qquad // \text{ Use the property of log: } \log(\prod_j a_j)=\sum_j \log(a_j)\\
&=\mathbb{E}_{y\sim P(y|D,\,{}^{t}\mathbf{p},\,{}^{t}\boldsymbol{\mu},\,{}^{t}\boldsymbol{\sigma}^2)}\left[\sum_{j=1}^{N}\log P(X_j,y_j|\,{}^{t+1}\mathbf{p},\,{}^{t+1}\boldsymbol{\mu},\,{}^{t+1}\boldsymbol{\sigma}^2)\right]\\
&\qquad\qquad // \text{ Use the linearity of expectation to put}\\
&\qquad\qquad // \text{ the summation outside the expectation.}\\
&=\sum_{j=1}^{N}\mathbb{E}_{y\sim P(y|D,\,{}^{t}\mathbf{p},\,{}^{t}\boldsymbol{\mu},\,{}^{t}\boldsymbol{\sigma}^2)}\left[\log P(X_j,y_j|\,{}^{t+1}\mathbf{p},\,{}^{t+1}\boldsymbol{\mu},\,{}^{t+1}\boldsymbol{\sigma}^2)\right]\\
&\qquad\qquad // \text{ Use the definition of expectation.}\\
&=\sum_{j=1}^{N}\left[\sum_{y}P(y|D,\,{}^{t}\mathbf{p},\,{}^{t}\boldsymbol{\mu},\,{}^{t}\boldsymbol{\sigma}^2)\,\log P(X_j,y_j|\,{}^{t+1}\mathbf{p},\,{}^{t+1}\boldsymbol{\mu},\,{}^{t+1}\boldsymbol{\sigma}^2)\right] \qquad (28)
\end{aligned}$$

The first term inside the square bracket can be factored as follows:

$$\sum_{y} P(y|D,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2})$$

// $D = \{X_1, ..., X_j, ..., X_N\}$

// $y = [y_1, ..., y_j, ..., y_N]$

$$= \left(\sum_{y_1=1}^{K} ... \sum_{y_j=1}^{K} ... \sum_{y_N=1}^{K} P(y_1, ..., y_j, ..., y_N | X_1, ..., X_j, ..., X_N,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2}) \right)$$

// based on conditional independence: given X_j,

// y_j is independent of $y_{j'}$ and $X_{j'}$ for $\forall j' \neq j$

$$= \left(\sum_{y_1=1}^{K} ... \sum_{y_j=1}^{K} ... \sum_{y_N=1}^{K} P(y_1|X_1,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2}) ... P(y_j|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2}) ... P(y_N|X_N,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2}) \right)$$

// Factor out different terms: you can get the result below by induction.

// When you sum over y_N, the only term related to y_N is $P(y_N|X_N,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2})$,

// other terms from $P(y_1|X_1,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2})$ to $P(y_{N-1}|X_{N-1},\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2})$ can be factored out.

// When you sum over y_{N-1}, the only term related to y_{N-1} is $P(y_{N-1}|X_{N-1},\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2})$,

// other terms from $P(y_1|X_1,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2})$ to $P(y_{N-2}|X_{N-2},\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2})$ can be factored out.

...

$$= \left(\sum_{y_1=1}^{K} P(y_1|X_1,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2}) \right) ... \left(\sum_{y_j=1}^{K} P(y_j|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2}) \right) ... \left(\sum_{y_N=1}^{K} P(y_N|X_N,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2}) \right).$$

Since the log term in Equation 28 is related to the j^{th} term above, we can assign the log term to the j^{th} term, and we can get:

$$\sum_{y} P(y|D,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2})\ \log P(X_j, y_j|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^{2})$$

// expand $\sum_{y} P(y|D,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2})$ based on the analysis above and

// assign $\log P(X_j, y_j|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^{2})$ to the summation term related to y_j

$$= \left(\sum_{y_j=1}^{K} P(y_j|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2})\ \log P(X_j, y_j|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^{2}) \right) \times$$
$$\left(\sum_{y_1=1}^{K} P(y_1|X_1,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2}) \right) \times \ldots \times \left(\sum_{y_{j-1}=1}^{K} P(y_{j-1}|X_{j-1},\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2}) \right) \times$$
$$\left(\sum_{y_{j+1}=1}^{K} P(y_{j+1}|X_{j+1},\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2}) \right) \times \ldots \times \left(\sum_{y_N=1}^{K} P(y_N|X_N,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2}) \right)$$

// property of PDF: $\sum_{y_{j'}=1}^{K} P(y_{j'}|X_{j'},\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2}) = 1$

$$= \sum_{y_j=1}^{K} P(y_j|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2})\ \log P(X_j, y_j|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^{2}).$$

Based on this result, Equation 28 becomes:

$$\mathbb{E}_{y\sim P(y|D,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2})} \left[\log P(D, y|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^{2}) \right]$$
$$= \sum_{j=1}^{N} \left[\sum_{y_j=1}^{K} P(y_j|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2})\ \log P(X_j, y_j|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^{2}) \right]$$

// apply the definition of conditional probability $P(A, B|C) = P(A|B, C)P(B|C)$

// on the term $P(X_j, y_j|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^{2})$

$$= \sum_{j=1}^{N} \left[\sum_{y_j=1}^{K} P(y_j|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2})\ \log P(X_j|y_j,\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^{2}) P(y_j|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^{2}) \right]$$

// use the property of log: $\log AB = \log A + \log B$

// on term $\log P(X_j|y_j,\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^{2}) P(y_j|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^{2})$

$$= \sum_{j=1}^{N} \left[\sum_{y_j=1}^{K} P(y_j|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2})\ \log P(X_j|y_j,\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^{2}) \right]$$
$$+ \sum_{j=1}^{N} \left[\sum_{y_j=1}^{K} P(y_j|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^{2})\ \log P(y_j|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^{2}) \right].$$

In the above equation, y_i is a cluster index, which has the same function as parameter c in Equation 5. Therefore, to make the notations consistent, we can reparameterize the above equation by replacing y_i with c:

$$
\begin{aligned}
&\mathbb{E}_{y\sim P(y|D,\ {}^{t}\mathbf{p},\ {}^{t}\mu,\ {}^{t}\sigma^2)}\left[\log P(D,y|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\mu,\ {}^{t+1}\sigma^2)\right] \\
&= \sum_{j=1}^{N}\left[\sum_{c=1}^{K} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\mu,\ {}^{t}\sigma^2)\log P(X_j|c,\ {}^{t+1}\mathbf{p},\ {}^{t+1}\mu,\ {}^{t+1}\sigma^2)\right] \\
&\quad+\sum_{j=1}^{N}\left[\sum_{c=1}^{K} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\mu,\ {}^{t}\sigma^2)\log P(c|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\mu,\ {}^{t+1}\sigma^2)\right]. \qquad (29)
\end{aligned}
$$

In Equation 29, the term $P(X_j|c,\ {}^{t+1}\mathbf{p},\ {}^{t+1}\mu,\ {}^{t+1}\sigma^2)$ is the likelihood of data X_j under the c^{th} Gaussian. Therefore, we have:

$$
\begin{aligned}
&\log P(X_j|c,\ {}^{t+1}\mathbf{p},\ {}^{t+1}\mu,\ {}^{t+1}\sigma^2) \\
&\qquad\qquad //\ \text{According to Equation 24} \\
&= \log P(X_j|c,\ {}^{t+1}\mu,\ {}^{t+1}\sigma^2) \\
&\qquad\qquad //\ \text{According to Equation 24} \\
&= \log P(X_j|\ {}^{t+1}\mu_c,\ {}^{t+1}\sigma_c^2) \\
&\qquad\qquad //\ \text{According to Equation 2} \\
&= \log\left[(2\pi)^{-\frac{d}{2}}\ ({}^{t+1}\sigma_c^2)^{-\frac{d}{2}}\ \exp\left(-\frac{1}{2\left({}^{t+1}\sigma_c^2\right)}\sum_{i=1}^{d}(X_{ji}-\ {}^{t+1}\mu_{ci})^2\right)\right] \\
&\qquad\qquad //\ \text{use the property of log: } \log ABC = \log A + \log B + \log C \\
&= -\frac{d}{2}\log 2\pi - \frac{d}{2}\log\left({}^{t+1}\sigma_c^2\right) - \frac{1}{2\left({}^{t+1}\sigma_c^2\right)}\sum_{i=1}^{d}(X_{ji}-\ {}^{t+1}\mu_{ci})^2. \qquad (30)
\end{aligned}
$$

In Equation 29, the term $P(c|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\mu,\ {}^{t+1}\sigma^2)$ is the probability of observing cluster c among all the K clusters. As a result, by referring Equation 25, we have:

$$
\begin{aligned}
&\log P(c|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\mu,\ {}^{t+1}\sigma^2) \\
&= \log P(c|\ {}^{t+1}\mathbf{p}) \\
&= \log({}^{t+1}p_c) \qquad (31)
\end{aligned}
$$

Based on the analysis above, Equation 29 becomes:

$$\begin{aligned}
&\mathbb{E}_{y\sim P(y|D,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)}\left[\log P(D,y|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^2)\right]\\
&= -\frac{Nd}{2}\log 2\pi - \frac{d}{2}\sum_{j=1}^{N}\sum_{c=1}^{K} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)\log\left({}^{t+1}\sigma_c^2\right) -\\
&\quad\sum_{j=1}^{N}\sum_{c=1}^{K} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)\frac{1}{2\left({}^{t+1}\sigma_c^2\right)}\sum_{i=1}^{d}(X_{ji} -\ {}^{t+1}\mu_{ci})^2 -\\
&\quad\sum_{j=1}^{N}\sum_{c=1}^{K} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)\log({}^{t+1}p_c). \qquad (32)
\end{aligned}$$

Remember that in Section "Parameters in Gaussian Mixture Mode", we have a constraint on parameter ${}^{t+1}\mathbf{p}$, which is $\sum_{c=1}^{K}\ {}^{t+1}p_c = 1$. Therefore, our problem becomes an optimization problem, which can be concluded as follows:

$$\begin{aligned}
&\max\ \mathbb{E}_{y\sim P(y|D,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)}\left[\log P(D,y|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^2)\right],\\
&\text{subject to } \sum_{c=1}^{K}\ {}^{t+1}p_c = 1.
\end{aligned}$$

By using the Lagrange multiplier, we can absorb the constraint into the objective function. The new objective function augmented by Lagrange multiplier can be written as follows:

$$\begin{aligned}
&J(\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^2)\\
&= \mathbb{E}_{y\sim P(y|D,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)}\left[\log P(D,y|\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^2)\right] + \lambda\left(1 - \sum_{c=1}^{K}\ {}^{t+1}p_c\right)\\
&\qquad\qquad \text{// Use Equation 32 to expand the expectation term.}\\
&= -\frac{Nd}{2}\log 2\pi - \frac{d}{2}\sum_{j=1}^{N}\sum_{c=1}^{K} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)\log\left({}^{t+1}\sigma_c^2\right) -\\
&\quad\sum_{j=1}^{N}\sum_{c=1}^{K} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)\frac{1}{2\left({}^{t+1}\sigma_c^2\right)}\sum_{i=1}^{d}(X_{ji} -\ {}^{t+1}\mu_{ci})^2 -\\
&\quad\sum_{j=1}^{N}\sum_{c=1}^{K} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)\log({}^{t+1}p_c) + \lambda\left(1 - \sum_{c=1}^{K}\ {}^{t+1}p_c\right). \qquad (33)
\end{aligned}$$

To compute the optimal solution for ${}^{t+1}\boldsymbol{\mu}$, we carry computation as follows:

$$\frac{\partial\, J(\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^2)}{\partial\ {}^{t+1}\mu_{ci}} = 0$$

// $J(\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^2)$ is shown in Equation 33

$$\sum_{j=1}^{N} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)\frac{1}{2\left({}^{t+1}\sigma_c^2\right)}\ 2(X_{ji} -\ {}^{t+1}\mu_{ci}) = 0$$

$${}^{t+1}\mu_{ci} = \frac{\sum_{j=1}^{N} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)X_{ji}}{\sum_{j=1}^{N} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)}$$

for $\forall\ c = \{1, 2, ..., K\}$ and $i = \{1, 2, ..., d\}$.

Since ${}^{t+1}\boldsymbol{\mu_c} = [\ {}^{t+1}\mu_{c1}, ...,\ {}^{t+1}\mu_{ci}, ...,\ {}^{t+1}\mu_{cd}]$ and $X_j = [X_{j1}, ..., X_{ji}, ..., X_{jd}]$, we have:

$${}^{t+1}\boldsymbol{\mu_c} = \frac{\sum_{j=1}^{N} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)\ X_j}{\sum_{j=1}^{N} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)} \text{ for } \forall\ c = \{1, 2, ..., K\}. \tag{34}$$

To compute the optimal solution for ${}^{t+1}\boldsymbol{\sigma^2}$, we carry computation as follows:

$$\frac{\partial\, J(\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^2)}{\partial\ {}^{t+1}\sigma_c^2} = 0$$

// $J(\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^2)$ is shown in Equation 33

$$\frac{1}{2}\frac{1}{({}^{t+1}\sigma_c^2)^2}\sum_{j=1}^{N} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)\sum_{i=1}^{d} (X_{ji} -\ {}^{t+1}\mu_{ci})^2 = \frac{d}{2}\frac{\sum_{j=1}^{N} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)}{{}^{t+1}\sigma_c^2}$$

$${}^{t+1}\sigma_c^2 = \frac{\sum_{j=1}^{N} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)\sum_{i=1}^{d} (X_{ji} -\ {}^{t+1}\mu_{ci})^2}{d\ \sum_{j=1}^{N} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)}$$

$${}^{t+1}\sigma_c^2 = \frac{\sum_{j=1}^{N} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)\ ||\ X_j -\ {}^{t+1}\boldsymbol{\mu_c}||^2}{d\ \sum_{j=1}^{N} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)} \text{ for } \forall\ c = \{1, 2, ..., K\}. \tag{35}$$

To compute the optimal solution for ${}^{t+1}\mathbf{p}$, we carry computation as follows:

$$\begin{aligned}
\frac{\partial\, J(\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^2)}{\partial\ {}^{t+1}p_c} &= 0 \\
\text{// } J(\ {}^{t+1}\mathbf{p},\ {}^{t+1}\boldsymbol{\mu},\ {}^{t+1}\boldsymbol{\sigma}^2) \text{ is shown in Equation 33} \\
\frac{\sum_{j=1}^{N} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)}{{}^{t+1}p_c} &= \lambda \\
{}^{t+1}p_c = \frac{\sum_{j=1}^{N} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)}{\lambda} & \text{ for } \forall\ c = \{1, 2, ..., K\}.
\end{aligned}$$

According to the constraint:

$$\begin{aligned}
\sum_{c=1}^{K}\ {}^{t+1}p_c &= 1 \\
\frac{1}{\lambda}\sum_{c=1}^{K}\sum_{j=1}^{N} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2) &= 1 \\
\lambda = \sum_{j=1}^{N}\sum_{c=1}^{K} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)& \\
\lambda &= N.
\end{aligned}$$

Therefore, the optimal solution for ${}^{t+1}\mathbf{p}$ is:

$$ {}^{t+1}p_c = \frac{\sum_{j=1}^{N} P(c|X_j,\ {}^{t}\mathbf{p},\ {}^{t}\boldsymbol{\mu},\ {}^{t}\boldsymbol{\sigma}^2)}{N} \text{ for } \forall\ c = \{1, 2, ..., K\}. \tag{36}$$

EM Algorithm Summary

- Given Data $D = \{X_1, X_2, ..., X_j, ..., X_N\}$, where $X_j \in \mathbb{R}^d$ for $\forall j \in \{1, 2, .., N\}$.
- Given the number of mixture component $K \in \{1, 2, 3, ...\}$.
- Given a $d \times d$ identity matrix I.
- Randomly initialize ${}^0\mathbf{p} = [{}^0p_1, ..., {}^0 p_c, ..., {}^0 p_K]$, where ${}^0p_c \in \mathbb{R}$ for $\forall c \in \{1, 2, ..., K\}$, and $\sum_{c=1}^{K} {}^0p_c = 1$.
- Randomly initialize ${}^0\boldsymbol{\mu} = [{}^0\boldsymbol{\mu}_1, ..., {}^0 \boldsymbol{\mu}_c, ..., {}^0 \boldsymbol{\mu}_K]$, where ${}^0\boldsymbol{\mu}_c \in \mathbb{R}^d$ for $\forall c \in \{1, 2, ..., K\}$.
- Randomly initialize ${}^0\boldsymbol{\sigma}^2 = [{}^0\sigma_1^2, ..., {}^0 \sigma_c^2, ..., {}^0 \sigma_K^2]$, where ${}^0\sigma_c^2 \in \mathbb{R}$ for $\forall c \in \{1, 2, ..., K\}$.

- for t = $\{0, 1, 2, 3, ..., T\}$
  - Compute E-Step update: (Equation 27)

$$P(c|X_j,\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2) = \frac{\mathcal{N}(X_j\ ;\ {}^t\boldsymbol{\mu}_c,\ {}^t\sigma_c^2\mathbf{I}) \cdot\ {}^tp_c}{\sum_{b=1}^{K}\ \mathcal{N}(X_j\ ;\ {}^t\boldsymbol{\mu}_b,\ {}^t\sigma_b^2\mathbf{I}) \cdot\ {}^tp_b}$$

for $\forall c \in \{1, 2, ..., K\}$ and $\forall j \in \{1, 2, ..., N\}$.

  - Compute M-Step update: (Equation 34, 35, 36)

$${}^{t+1}\boldsymbol{\mu}_c = \frac{\sum_{j=1}^{N} P(c|X_j,\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2)\ X_j}{\sum_{j=1}^{N} P(c|X_j,\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2)} \text{ for } \forall\ c = \{1, 2, ..., K\}.$$

$${}^{t+1}\sigma_c^2 = \frac{\sum_{j=1}^{N} P(c|X_j,\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2)\ ||\ X_j -\ {}^{t+1}\boldsymbol{\mu}_c||^2}{d\ \sum_{j=1}^{N} P(c|X_j,\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2)} \text{ for } \forall\ c = \{1, 2, ..., K\}.$$

$${}^{t+1}p_c = \frac{\sum_{j=1}^{N} P(c|X_j,\ {}^t\mathbf{p},\ {}^t\boldsymbol{\mu},\ {}^t\boldsymbol{\sigma}^2)}{N} \text{ for } \forall\ c = \{1, 2, ..., K\}.$$