Expectation–Maximisation (EM) Algorithm Derivations Specifically for the Multivariate Gaussian Mixture Model (GMM)

Contents

1 Introduction

The Gaussian Mixture Model (GMM) is a special case of the Expectation–Maximisation (EM) algorithm, where each mixture is a Gaussian distribution. The following shows the derivations for the GMM in the multivariate case.

Let,
- n be the number of observations
- K be the number of mixtures
- X_i be an observation
- Z_i be the mixture of an observation
- θ_i be a mixture's parameters
- π_i be the weight of a mixture

2 Derivation

2.1 Complete likelihood

$$P(X, Z|\theta) = P(X_1, \ldots, X_n, Z_1, \ldots, Z_n|\theta)$$

As each observation is independent

$$\begin{aligned} P(X, Z|\theta) &= \prod_{i=1}^{n} P(X_i, Z_i|\theta) \\ &= \prod_{i=1}^{n} P(X_i|Z_i, \theta)P(Z_i|\theta) \end{aligned}$$

As $Z_i \sim$ categorical and $P(Z_i) = \prod_{k=1}^{K} P(Z_i = k)^{I(Z_i=k)}$, where $I(A)$ is an indicator function

$$P(X, Z|\theta) = \prod_{i=1}^{n} \prod_{j=1}^{K} (P(X_i|Z_i = j, \theta_j)P(Z_i = j|\theta_j))^{I(Z_i=j)}$$

2.2 Complete log-likelihood

$$\begin{aligned}
\log P(X,Z|\theta) &= \log\left(\prod_{i=1}^{n}\prod_{j=1}^{K}(P(X_i|Z_i=j,\theta_j)P(Z_i=j|\theta_j))^{I(Z_i=j)}\right)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{K} I(Z_i=j)(\log P(X_i|Z_i=j,\theta_j)+\log P(Z_i=j|\theta_j))
\end{aligned}$$

As $P(X_i|Z_i=j,\theta_j)$ is the PDF, $f(x_i;\theta_j)$, and $P(Z_i=j|\theta_j)$ is the weight, π_j

$$\log P(X,Z|\theta) = \sum_{i=1}^{n}\sum_{j=1}^{K} I(Z_i=j)(\log\pi_j+\log f(x_i;\theta_j))$$

2.3 Expectation of the complete log-likelihood

Let t be the current iteration of parameters

$$\begin{aligned}
Q\left(\theta^{t+1},\theta^t\right) &= E_{Z|X,\theta^t}\left[\log P\left(X,Z|\theta^{t+1}\right)\right]\\
&= E_{Z|X,\theta^t}\left[\sum_{i=1}^{n}\sum_{j=1}^{K} I(Z_i=j)\left(\log\pi_j^{t+1}+\log f\left(x_i;\theta_j^{t+1}\right)\right)\right]\\
&= \sum_{i=1}^{n}\sum_{j=1}^{K} E_{Z|X,\theta^t}[I(Z_i=j)]\left(\log\pi_j^{t+1}+\log f\left(x_i;\theta_j^{t+1}\right)\right)
\end{aligned}$$

Aside,

As $I(Z_i=j)$ has 2 values, 0 and 1, that are dependent on the mixtures

$$\begin{aligned}
E_{Z|X,\theta^t}[I(Z_i=j)] &= 0P_{Z|X,\theta^t}(Z_i\neq j)+1P_{Z|X,\theta^t}(Z_i=j)\\
&= P_{Z|X,\theta^t}(Z_i=j)\\
&= P\left(Z_i=j|X,\theta^t\right)
\end{aligned}$$

Let,

$$P\left(Z_i=j|X,\theta^t\right) = A_{ij}^t$$

So,

$$Q\left(\theta^{t+1},\theta^t\right) = \sum_{i=1}^{n}\sum_{j=1}^{K} A_{ij}^t\left(\log\pi_j^{t+1}+\log f\left(x_i;\theta_j^{t+1}\right)\right)$$

For a multivariate normal, where d is its dimension

$$\begin{aligned}
Q\left(\theta^{t+1},\theta^t\right) &= \sum_{i=1}^{n}\sum_{j=1}^{K} A_{ij}^t\left(\log\pi_j^{t+1}+\log\left(\frac{1}{\sqrt{(2\pi)^d|\Sigma_j^{t+1}|}}e^{-\frac{1}{2}\left(\vec{x}_i-\vec{\mu}_j^{t+1}\right)^T\left(\Sigma_j^{t+1}\right)^{-1}\left(\vec{x}_i-\vec{\mu}_j^{t+1}\right)}\right)\right)\\
&= \sum_{i=1}^{n}\sum_{j=1}^{K} A_{ij}^t\left(\log\pi_j^{t+1}-\frac{d}{2}\log 2\pi-\frac{1}{2}\log\left|\Sigma_j^{t+1}\right|-\frac{1}{2}\left(\vec{x}_i-\vec{\mu}_j^{t+1}\right)^T\left(\Sigma_j^{t+1}\right)^{-1}\left(\vec{x}_i-\vec{\mu}_j^{t+1}\right)\right)
\end{aligned}$$

2.4 Posterior distribution of the latent variables (E-Step)

$$
\begin{aligned}
A_{ij}^t = P\big(Z_i = j | X, \theta^t\big) &= \frac{P(Z_i = j, X|\theta^t)}{P(X|\theta^t)} \\
&= \frac{P(X|Z_i = j, \theta^t)P(Z_i = j|\theta^t)}{\sum_{k=1}^{K} P(X|Z_i = k, \theta^t)P(Z_i = k|\theta^t)}
\end{aligned}
$$

As $P(X_i|Z_i = j, \theta^t)$ is the PDF, $f\big(x_i; \theta_j^t\big)$, and $P(Z_i = j|\theta^t)$ is the weight, π_j^t

$$
A_{ij}^t = \frac{f\big(x_i; \theta_j^t\big)\pi_j^t}{\sum_{k=1}^{K} f(x_i; \theta_k^t)\pi_k^t}
$$

2.5 Maximisation of the latent variables (M-Step)

2.5.1 Update equation for $\vec{\mu}$

As Σ is symmetric and using the matrix derivative rule $\frac{\partial}{\partial \vec{a}}\left[\left(\vec{b} - \vec{a}\right)^T X \left(\vec{b} - \vec{a}\right)\right] = -2X\left(\vec{b} - \vec{a}\right)$, where X is symmetric

$$
\begin{aligned}
\frac{\partial Q}{\partial \vec{\mu}_j^{t+1}} = 0 &= \sum_{i=1}^{n} A_{ij}^t \left(-\frac{1}{2}\left(-2\big(\Sigma_j^{t+1}\big)^{-1}\big(\vec{x}_i - \vec{\mu}_j^{t+1}\big)\right)\right) \\
&= \sum_{i=1}^{n} A_{ij}^t \vec{x}_i - \vec{\mu}_j^{t+1} \sum_{i=i}^{n} A_{ij}^t \\
\vec{\mu}_j^{t+1} &= \frac{1}{\sum_{i=1}^{n} A_{ij}^t} \sum_{i=1}^{n} A_{ij}^t \vec{x}_i
\end{aligned}
$$

2.5.2 Update equation for Σ

Using the matrix derivative rules $\frac{\partial}{\partial X}[\log|X|] = (X^T)^{-1}$ and $\frac{\partial}{\partial X}\left[\vec{a}^T X^{-1} \vec{b}\right] = -\big(X^T\big)^{-1} \vec{a}\vec{b}^T \big(X^T\big)^{-1}$

$$
\frac{\partial Q}{\partial \Sigma_j^{t+1}} = 0 = \sum_{i=1}^{n} A_{ij}^t \left(-\frac{1}{2}\left(\big(\Sigma_j^{t+1}\big)^T\right)^{-1} - \frac{1}{2}\left(-\left(\big(\Sigma_j^{t+1}\big)^T\right)^{-1}\big(\vec{x}_i - \vec{\mu}_j^{t+1}\big)\big(\vec{x}_i - \vec{\mu}_j^{t+1}\big)^T \left(\big(\Sigma_j^{t+1}\big)^T\right)^{-1}\right)\right)
$$

As Σ is symmetric then $\Sigma^T = \Sigma$

$$
\begin{aligned}
0 &= \big(\Sigma_j^{t+1}\big)^{-1} \sum_{i=1}^{n} A_{ij}^t - \big(\Sigma_j^{t+1}\big)^{-1} \left(\sum_{i=1}^{n} A_{ij}^t \big(\vec{x}_i - \vec{\mu}_j^{t+1}\big)\big(\vec{x}_i - \vec{\mu}_j^{t+1}\big)^T\right)\big(\Sigma_j^{t+1}\big)^{-1} \\
\sum_{i=1}^{n} A_{ij}^t I &= \left(\sum_{i=1}^{n} A_{ij}^t \big(\vec{x}_i - \vec{\mu}_j^{t+1}\big)\big(\vec{x}_i - \vec{\mu}_j^{t+1}\big)^T\right)\big(\Sigma_j^{t+1}\big)^{-1} \\
\Sigma_j^{t+1} &= \frac{1}{\sum_{i=1}^{n} A_{ij}^t} \sum_{i=1}^{n} A_{ij}^t \big(\vec{x}_i - \vec{\mu}_j^{t+1}\big)\big(\vec{x}_i - \vec{\mu}_j^{t+1}\big)^T
\end{aligned}
$$

2.5.3 Update equation for π

As $\pi_K = 1 - \sum_{i=1}^{K-1} \pi_i$ the π_K part is also included in the derivative of π_i

$$\begin{aligned}
\frac{\partial Q}{\partial \pi_j^{t+1}} = 0 &= \sum_{i=1}^{n} \left[\frac{A_{ij}^t}{\pi_j^{t+1}} - \frac{A_{iK}^t}{\pi_K^{t+1}} \right] \\
&= \frac{1}{\pi_j^{t+1}} \sum_{i=1}^{n} A_{ij}^t - \frac{1}{\pi_K^{t+1}} \sum_{i=1}^{n} A_{iK}^t \\
&= \frac{\pi_K^{t+1} \sum_{i=1}^{n} A_{ij}^t - \pi_j^{t+1} \sum_{i=1}^{n} A_{iK}^t}{\pi_j^{t+1} \pi_K^{t+1}} \\
0 &= \pi_K^{t+1} \sum_{i=1}^{n} A_{ij}^t - \pi_j^{t+1} \sum_{i=1}^{n} A_{iK}^t \\
\pi_K^{t+1} \sum_{i=1}^{n} A_{ij}^t &= \pi_j^{t+1} \sum_{i=1}^{n} A_{iK}^t \qquad \text{(i)}
\end{aligned}$$

Aside,

By summing all (i) together for $j = 1, \ldots, K-1$

$$\begin{aligned}
\pi_K^{t+1} \sum_{i=1}^{n} A_{i1}^t + \cdots + \pi_K^{t+1} \sum_{i=1}^{n} A_{i(K-1)}^t &= \pi_1^{t+1} \sum_{i=1}^{n} A_{iK}^t + \cdots + \pi_{K-1}^{t+1} \sum_{i=1}^{n} A_{iK}^t \\
&= \sum_{k=1}^{K-1} \pi_k^{t+1} \sum_{i=1}^{n} A_{iK}^t
\end{aligned}$$

As $\sum_{k=1}^{K-1} \pi_k = 1 - \left(1 - \sum_{k=1}^{K-1} \pi_k\right) = 1 - \pi_K$

$$\begin{aligned}
\pi_K^{t+1} \sum_{i=1}^{n} A_{i1}^t + \cdots + \pi_K^{t+1} \sum_{i=1}^{n} A_{i(K-1)}^t &= \left(1 - \pi_K^{t+1}\right) \sum_{i=1}^{n} A_{iK}^t \\
&= \sum_{i=1}^{n} A_{iK}^t - \pi_K^{t+1} \sum_{i=1}^{n} A_{iK}^t \\
\pi_K^{t+1} \sum_{i=1}^{n} \sum_{j=1}^{K} A_{iK}^t &= \sum_{i=1}^{n} A_{iK}^t
\end{aligned}$$

As $\sum_{i=1}^{n} A_{iK}^t = 1$ because it is the sum of all probabilities

$$\begin{aligned}
\pi_K^{t+1} \sum_{i=1}^{n} 1 &= \sum_{i=1}^{n} A_{iK}^t \\
\pi_K^{t+1} &= \frac{1}{n} \sum_{i=1}^{n} A_{iK}^t \qquad \text{(ii)}
\end{aligned}$$

So,

From (i)

$$\pi_j^{t+1} = \frac{1}{\sum_{i=1}^{n} A_{iK}^t} \pi_K^{t+1} \sum_{i=1}^{n} A_{ij}^t$$

Substituting in (ii)

$$\begin{aligned}
\pi_j^{t+1} &= \frac{1}{\cancel{\sum_{i=1}^{n} A_{iK}^t}} \left(\frac{1}{n} \cancel{\sum_{i=1}^{n} A_{iK}^t} \right) \sum_{i=1}^{n} A_{ij}^t \\
\pi_j^{t+1} &= \frac{1}{n} \sum_{i=1}^{n} A_{ij}^t
\end{aligned}$$