An example of EM algorithm for estimating the parameters of a mixture of two multivariate Gaussian distributions

Made by Jukka Talvitie
jukka.talvitie@tut.fi
Department of Electronics and Communications engineering
Tampere University of Technology

Considered system model

Suppose we take samples $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \ldots, \mathbf{x}_N)$ from a mixture of two multivariate normal distributions of dimension 2 (i.e., sample $\mathbf{x}_i$ is a 2x1 vector including the sample coordinates). Let $\mathbf{z} = (z_1, z_2, z_3, \ldots, z_N)$ be hidden variable which determines from which of the two distributions the sample is originated. If z_i=1, then the i^{th} sample is from the distribution #1, and if z_i=2, then the i^{th} sample is from the distribution #2. Hence, the i^{th} sample has the following probability distribution:

$$x_i \sim \begin{cases} \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), & \text{when } z_i = 1 \\ \mathrm{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), & \text{when } z_i = 2 \end{cases}.$$

The probabilities of having the sample from the distribution #1 and distribution #2 are given as

$$\begin{aligned} P(z_i = 1) &= \tau_1 \\ P(z_i = 2) &= \tau_2 = 1 - \tau_1 \end{aligned}.$$

Now, the objective is to estimate (in ML sense) the mean and the covariance of both distributions and the sampling probability $\boldsymbol{\tau} = [\tau_1 \, \tau_2]^T$:

$$\theta = (\boldsymbol{\tau}, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)$$

Finding out the likelihood function

The E-step of the EM algorithm, which we will further take, requires the likelihood function $L(\theta \mid \mathbf{x}, \mathbf{z}) = P(\mathbf{x}, \mathbf{z} \mid \theta)$ of the estimated parameter θ.

To begin with, let us consider the likelihood (and probability) for the i^{th} sample:

$$L(\theta \mid \mathbf{x}_i, z_i) = P(\mathbf{x}_i, z_i \mid \theta) = \begin{cases} \tau_1 \cdot f(\mathbf{x}_i \mid \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), & \text{when } z_i = 1 \\ \tau_2 \cdot f(\mathbf{x}_i \mid \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) & \text{when } z_i = 2 \end{cases}$$

where $f(\mathbf{x}_i|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$ is the probability density function (pdf) of a multivariate normal distribution (i.e., the value of the pdf at $\mathbf{x}_i$, with the distribution mean and covariance as $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$, respectively) defined as

$$f(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right),$$

where $|\boldsymbol{\Sigma}|$ is determinant of $\boldsymbol{\Sigma}$.

The above likelihood $L(\theta \mid \mathbf{x}, \mathbf{z})$ can be presented in separate sum terms by using so called indicator function:

$$L(\theta \mid \mathbf{x}_i, z_i) = P(\mathbf{x}_i, z_i \mid \theta) = \sum_{j=1}^{2} \mathrm{I}_j(z_i) \cdot \tau_j \cdot f(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

where, $\mathrm{I}_j(z_i) = \begin{cases} 1, \text{ when } z_i = j \\ 0, \text{ when } z_i \neq j \end{cases}$.

Excursion for the indicator function (this is a side topic; not related to this example):

Indicator function is actually quite useful in many problems in statistics. It is especially convenient in problems, where some function is defined piecewise for different input values. Let us define *A* to be a subset of *Ω*. For an element *α* taken from the set *Ω*, indicator function is defined as

$$\mathrm{I}_A(\alpha) = \begin{cases} 1, \text{ when } \alpha \in A \\ 0, \text{ when } \alpha \notin A \end{cases}.$$

So, the indicator function returns value 1, if *α* belongs to the set *A*, and 0, if *α* does not belong to the set *A*. For example, suppose a random variable *X* whose pdf is defined as

$$p(x) = \begin{cases} 0, \text{ when } x < 0 \\ 2x, \text{ when } 0 \leq x < 5 \\ 20 - 2x, \text{ when } 5 \leq x \leq 10 \\ 0, \text{ when } x > 10 \end{cases}$$

Now, exploiting the indicator function, this can be conveniently written as

$$p(x) = 2x\mathrm{I}_{[0,5)}(x) + (20 - 2x)\mathrm{I}_{[5,10]}(x),$$

and this is actually the format that we are using in our likelihood function.

Combining the likelihoods over all observations (i.e. multiplying individual observation likelihoods) we get

$$\begin{aligned} L(\theta \mid \mathbf{x}, \mathbf{z}) &= \prod_{i=1}^{N} L(\theta \mid \mathbf{x}_i, z_i) = \prod_{i=1}^{N} \sum_{j=1}^{2} \mathrm{I}_j(z_i) \cdot \tau_j \cdot f(\mathbf{x}_i \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \\ &= \prod_{i=1}^{N} \sum_{j=1}^{2} \mathrm{I}_j(z_i) \tau_j \frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}_j|}} \exp\left(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_j) \right) \end{aligned}$$

and the corresponding log-likelihood required for the E-step finally as

$$\log L(\theta|\mathbf{x},\mathbf{z}) = \sum_{i=1}^{N}\sum_{j=1}^{2} \mathrm{I}_j(z_i)\left[\log\tau_j - \log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_j)\right].$$

E-step

The target is to calculate the expected value of the above defined likelihood with respect to hidden/latent variable **z**, given the observations **x** and the current estimate of the parameter set $\theta^{(t)}$. Here **z** is the "complete data", i.e., if we would know this, we would be able to calculate the maximum likelihood (ML) estimate directly without the EM algorithm (find from literature: "ML estimate of parameters of multivariate normal distribution"). Here the superscript (t) refers to the t^{th} iteration, and thus, $\theta^{(t)}$ is the value of the parameter estimate at t^{th} iteration. So, at this point the variable **z** inside the indicator function is the only "unknown" variable which needs to be made "known" before the likelihood can be computed. For this purpose we use the expectation, for which we need to define the probability of **z**, given the observations and the current parameter estimate $\theta^{(t)}$:

$$T_{j,i}^{(t)} := P(z_i = j \mid x_i, \theta^{(t)}) = \frac{\tau_j^{(t)} f(\mathbf{x}_i \mid \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}{\tau_1^{(t)} f(\mathbf{x}_i \mid \boldsymbol{\mu}_1^{(t)}, \boldsymbol{\Sigma}_1^{(t)}) + \tau_2^{(t)} f(\mathbf{x}_i \mid \boldsymbol{\mu}_2^{(t)}, \boldsymbol{\Sigma}_2^{(t)})}$$

Here $T_{j,i}^{(t)}$ gives the probability of i^{th} sample to belong into j^{th} distribution (given the parameter estimates at t^{th} iteration). Notice that for each observation we must always have $\sum_{j=1}^{2} T_{j,i}^{(t)} = 1$, i.e. the probability of sample belonging in any of the defined distributions equals one. For example, digital transmission system using blind channel estimation (estimation without knowledge of transmitted symbols), $T_{j,i}^{(t)}$ could model the unknown symbols. In this case, assuming samples taken at symbol-rate, $T_{j,i}^{(t)}$ would give the likelihood of i^{th} symbol to originate from j^{th} symbol in the constellation. In other words, we would get soft symbol estimates as a side product of the EM algorithm while estimating the channel.

The E-step can now be expressed as

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= \mathrm{E}\left\{\log L(\theta|\mathbf{x},\mathbf{z})\right\} = \mathrm{E}\left\{\sum_{i=1}^{N}\log L(\theta|\mathbf{x}_i,\mathbf{z}_i)\right\} \\ &= \sum_{i=1}^{N}\mathrm{E}\left\{\log L(\theta|\mathbf{x}_i,\mathbf{z}_i)\right\} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{2} T_{j,i}^{(t)}\left[\log\tau_j - \log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_j)^T\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_j)\right] \end{aligned}$$

M-step

In M-step, to obtain updated parameter estimates, we maximize the function provided in the E-step with respect to the estimated parameters:

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta \,|\, \theta^{(t)})$$
$$= \arg\max_{\theta} \sum_{i=1}^{N}\sum_{j=1}^{2} T_{j,i}^{(t)} \left[\log\tau_j - \log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_j| - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_j) \right]$$

Notice that here the estimated parameters $\boldsymbol{\tau}$, ($\boldsymbol{\mu}_1$,$\boldsymbol{\Sigma}_1$), ($\boldsymbol{\mu}_2$,$\boldsymbol{\Sigma}_2$) appear in separate terms, so we can maximize them independently (i.e., parameters do not exist in partial derivatives of other parameters). We start by maximizing the likelihood with respect to $\boldsymbol{\tau}$, and then, we continue by maximizing with respect to ($\boldsymbol{\mu}_1$,$\boldsymbol{\Sigma}_1$). Here, because of the symmetry of the likelihood function, ($\boldsymbol{\mu}_2$,$\boldsymbol{\Sigma}_2$) will have similar solution with the latter case. Nevertheless, for each estimated parameter, the function is maximized by differentiating the likelihood function and setting the derivative to zero.

Estimate update for τ:

When estimating $\boldsymbol{\tau}$, we must also remember the constraint $\tau_1+\tau_2=1$. Therefore, the updated estimate is

$$\boldsymbol{\tau}^{(t+1)} = \arg\max_{\boldsymbol{\tau}} Q(\theta \,|\, \theta^{(t)})$$
$$= \arg\max_{\boldsymbol{\tau}} \sum_{i=1}^{N}\sum_{j=1}^{2} T_{j,i}^{(t)} \log\tau_j = \arg\max_{\boldsymbol{\tau}} \left\{ \sum_{i=1}^{N} T_{1,i}^{(t)} \log\tau_1 + \sum_{i=1}^{N} T_{2,i}^{(t)} \log\tau_2 \right\}$$
$$= \arg\max_{\boldsymbol{\tau}} \left\{ \log\tau_1 \sum_{i=1}^{N} T_{1,i}^{(t)} + \log\tau_2 \sum_{i=1}^{N} T_{2,i}^{(t)} \right\}$$

with constraint $\tau_1+\tau_2=1$.

The method of Lagrangian multipliers can be used to solve constrained maximization (or minimization) problems. The Lagrange function is defined as

$$\Phi(\tau_1, \tau_2, \lambda) = f(\tau_1, \tau_2) + \lambda\left(g(\tau_1, \tau_2) - c\right),$$

where $f(\cdot)$ is the function being maximized ($\log\tau_1 \sum_{i=1}^{N} T_{1,i}^{(t)} + \log\tau_2 \sum_{i=1}^{N} T_{2,i}^{(t)}$) and $g(\cdot)$ is the constraint function ($\tau_1+\tau_2=1$). Now, the partial derivatives of the Lagrange function are computed, and their values are set to zero:

$$\frac{\partial\Phi(\tau_1, \tau_2, \lambda)}{\partial\tau_1} = \frac{1}{\tau_1}\sum_{i=1}^{N} T_{1,i}^{(t)} + \lambda = 0$$
$$\frac{\partial\Phi(\tau_1, \tau_2, \lambda)}{\partial\tau_2} = \frac{1}{\tau_2}\sum_{i=1}^{N} T_{2,i}^{(t)} + \lambda = 0$$
$$\frac{\partial\Phi(\tau_1, \tau_2, \lambda)}{\partial\lambda} = \tau_1 + \tau_2 - 1 = 0$$

From this set of equations we can easily solve the new estimates for $\boldsymbol{\tau}$ as

$$\left.\begin{aligned}\tau_1^{(t+1)} &= \frac{\sum_{i=1}^{N} T_{1,i}^{(t)}}{\sum_{i=1}^{N} T_{1,i}^{(t)} + \sum_{i=1}^{N} T_{2,i}^{(t)}} = \frac{1}{N}\sum_{i=1}^{N} T_{1,i}^{(t)} \\ \tau_2^{(t+1)} &= \frac{\sum_{i=1}^{N} T_{1,i}^{(t)}}{\sum_{i=1}^{N} T_{1,i}^{(t)} + \sum_{i=1}^{N} T_{2,i}^{(t)}} = \frac{1}{N}\sum_{i=1}^{N} T_{2,i}^{(t)}\end{aligned}\right\} \Rightarrow \tau_j^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} T_{j,i}^{(t)}$$

Notice from the earlier discussion that the suppression of the above denominator term is based on the fact that $\sum_{i=1}^{N} T_{1,i}^{(t)} + \sum_{i=1}^{N} T_{2,i}^{(t)} = \sum_{i=1}^{N} T_{2,i}^{(t)} + T_{1,i}^{(t)} = \sum_{i=1}^{N} 1 = N$.

Estimate update for $\boldsymbol{\mu}_1$ (and $\boldsymbol{\mu}_2$):

Then the final task is to find updated estimates for ($\boldsymbol{\mu}_1$,$\boldsymbol{\Sigma}_1$) and ($\boldsymbol{\mu}_2$,$\boldsymbol{\Sigma}_2$). Since they appear symmetrically in the likelihood function, we concentrate only on ($\boldsymbol{\mu}_1$,$\boldsymbol{\Sigma}_1$) and later deduce ($\boldsymbol{\mu}_2$,$\boldsymbol{\Sigma}_2$) directly from this.

We start by maximizing with respect to $\boldsymbol{\mu}_1$. However, since the covariance matrix $\boldsymbol{\Sigma}_1$ appears in the same term with $\boldsymbol{\mu}_1$, it must be included in the maximization process (at least for now). So, we define the estimate updates together for $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$ as

$$\begin{aligned}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)^{(t+1)} &= \arg\max_{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1} Q(\theta \,|\, \theta^{(t)}) \\ &= \arg\max_{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1} \sum_{i=1}^{N} T_{1,i}^{(t)} \left[-\frac{1}{2}\log|\boldsymbol{\Sigma}_1| - \frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_1) \right] \\ &= \arg\max_{\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1} \sum_{i=1}^{N} \frac{1}{2} T_{1,i}^{(t)} \left[-\log|\boldsymbol{\Sigma}_1| - \mathbf{x}_i^T \boldsymbol{\Sigma}_1^{-1}\mathbf{x}_i + \mathbf{x}_i^T \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1}\mathbf{x}_i - \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 \right]\end{aligned}$$

where again we have dropped all the terms from Q(·) that are not dependent on $\boldsymbol{\mu}_1$ or $\boldsymbol{\Sigma}_1$. We begin by differentiating the above equation with respect to $\boldsymbol{\mu}_1$ and set its value to zero:

$$\begin{aligned}\frac{\partial Q(\theta \,|\, \theta^{(t)})}{\partial \boldsymbol{\mu}_1} &= \sum_{i=1}^{N} \frac{1}{2} T_{1,i}^{(t)} \left[\mathbf{x}_i^T \boldsymbol{\Sigma}_1^{-1} + \left(\boldsymbol{\Sigma}_1^{-1}\mathbf{x}_i\right)^T - 2\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} \right] \\ &= \sum_{i=1}^{N} \frac{1}{2} T_{1,i}^{(t)} \left[\mathbf{x}_i^T \boldsymbol{\Sigma}_1^{-1} + \mathbf{x}_i^T \boldsymbol{\Sigma}_1^{-1} - 2\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} \right] \\ &= \sum_{i=1}^{N} \frac{1}{2} T_{1,i}^{(t)} \left[2\mathbf{x}_i^T \boldsymbol{\Sigma}_1^{-1} - 2\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} \right] \\ &= \sum_{i=1}^{N} T_{1,i}^{(t)} \mathbf{x}_i^T \boldsymbol{\Sigma}_1^{-1} - \sum_{i=1}^{N} T_{1,i}^{(t)} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} = 0\end{aligned}$$

From here we can solve $\boldsymbol{\mu}_1$, and thus, get the updated estimate $\boldsymbol{\mu}^{(t+1)}$. Also, we see that the covariance $\boldsymbol{\Sigma}_1$ does not affect the maximization with respect to $\boldsymbol{\mu}_1$. Finally, we get the estimate update as

$$\sum_{i=1}^{N} T_{1,i}^{(t)} \mathbf{x}_i^T \boldsymbol{\Sigma}_1^{-1} = \sum_{i=1}^{N} T_{1,i}^{(t)} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} \Leftrightarrow$$

$$\boldsymbol{\Sigma}_1^{-1} \sum_{i=1}^{N} T_{1,i}^{(t)} \mathbf{x}_i^T = \boldsymbol{\Sigma}_1^{-1} \sum_{i=1}^{N} T_{1,i}^{(t)} \boldsymbol{\mu}_1^T$$

$$\Rightarrow \boldsymbol{\mu}_1^{(t+1)} = \frac{\sum_{i=1}^{N} T_{1,i}^{(t)} \mathbf{x}_i^T}{\sum_{i=1}^{N} T_{1,i}^{(t)}}$$

Estimate update for $\boldsymbol{\Sigma}_1$ (and $\boldsymbol{\Sigma}_2$):

Now, we should do a similar type of procedure for the covariance $\boldsymbol{\Sigma}_1$ (notice that we can now use knowledge of $\boldsymbol{\mu}^{(t+1)}$ here). However, differentiating the likelihood with respect to $\boldsymbol{\Sigma}_1$ is a little bit trickier. Before we are able to use basic matrix calculus rules for the differentiation (e.g., Wikipedia for "Matrix calculus"), we must do a small trick using trace function $tr(\cdot)$. Now, the starting point is

$$\boldsymbol{\Sigma}_1^{(t+1)} = \arg\max_{\boldsymbol{\Sigma}_1} \sum_{i=1}^{N} T_{1,i}^{(t)} \left[-\frac{1}{2} \log|\boldsymbol{\Sigma}_1| - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})^T \boldsymbol{\Sigma}_1^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)}) \right].$$

Notice that the quadratic term inside the sum ($(\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})^T \boldsymbol{\Sigma}_1^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})$), is a scalar, i.e., 1x1 matrix. For this reason, we are able to take trace of the term and maintain the equality:

$$\begin{aligned}
\boldsymbol{\Sigma}_1^{(t+1)} &= \arg\max_{\boldsymbol{\Sigma}_1} \sum_{i=1}^{N} T_{1,i}^{(t)} \left[-\frac{1}{2} \log|\boldsymbol{\Sigma}_1| - \frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})^T \boldsymbol{\Sigma}_1^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)}) \right] \\
&= \arg\max_{\boldsymbol{\Sigma}_1} \sum_{i=1}^{N} T_{1,i}^{(t)} \left[-\frac{1}{2} \log|\boldsymbol{\Sigma}_1| - tr\left(\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})^T \boldsymbol{\Sigma}_1^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)}) \right) \right] \\
&= \arg\max_{\boldsymbol{\Sigma}_1} \sum_{i=1}^{N} -\frac{1}{2} T_{1,i}^{(t)} \log|\boldsymbol{\Sigma}_1| - \frac{1}{2} \sum_{i=1}^{N} T_{1,i}^{(t)} tr\left((\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})^T \boldsymbol{\Sigma}_1^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)}) \right)
\end{aligned}$$

Now, using two properties of the trace function, $tr(\mathbf{AB}) = tr(\mathbf{BA})$ and $tr(\mathbf{A}) + tr(\mathbf{B}) = tr(\mathbf{A}+\mathbf{B})$ (these are valid as long as the dimensions match), we can process the equation further as

$$
\begin{aligned}
\boldsymbol{\Sigma}_1^{(t+1)} &= \arg\max_{\boldsymbol{\Sigma}_1} \sum_{i=1}^{N} -\frac{1}{2} T_{1,i}^{(t)} \log\left|\boldsymbol{\Sigma}_1\right| - \frac{1}{2} \sum_{i=1}^{N} T_{1,i}^{(t)} tr\left((\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})^T \boldsymbol{\Sigma}_1^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})\right) \\
&= \arg\max_{\boldsymbol{\Sigma}_1} \frac{1}{2} \sum_{i=1}^{N} -T_{1,i}^{(t)} \log\left|\boldsymbol{\Sigma}_1\right| - \frac{1}{2} \sum_{i=1}^{N} T_{1,i}^{(t)} tr\left(\boldsymbol{\Sigma}_1^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})^T\right) \\
&= \arg\max_{\boldsymbol{\Sigma}_1} \frac{1}{2} \sum_{i=1}^{N} -T_{1,i}^{(t)} \log\left|\boldsymbol{\Sigma}_1\right| - \frac{1}{2} tr\left(\sum_{i=1}^{N} T_{1,i}^{(t)} \boldsymbol{\Sigma}_1^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})^T\right) \\
&= \arg\max_{\boldsymbol{\Sigma}_1} \frac{1}{2} \sum_{i=1}^{N} -T_{1,i}^{(t)} \log\left|\boldsymbol{\Sigma}_1\right| - \frac{1}{2} tr\left(\boldsymbol{\Sigma}_1^{-1} \sum_{i=1}^{N} T_{1,i}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})^T\right) \\
&= \arg\max_{\boldsymbol{\Sigma}_1} \frac{1}{2} \sum_{i=1}^{N} \left\{-T_{1,i}^{(t)} \log\left|\boldsymbol{\Sigma}_1\right|\right\} - \frac{1}{2} tr\left(\boldsymbol{\Sigma}_1^{-1} \mathbf{S}\right)
\end{aligned}
$$

where $\mathbf{S} = \sum_{i=1}^{N} T_{1,i}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})^T$. Finally, the above format of the function can be easily differentiated using the basic matrix calculus rules:

$$
\frac{\partial tr(\mathbf{X}^{-1}\mathbf{A})}{\partial \mathbf{X}} = -\left(\mathbf{X}^{-1}\right)^T \mathbf{A} \left(\mathbf{X}^{-1}\right)^T
$$

$$
\frac{\partial \log(\left|\mathbf{X}\right|)}{\partial \mathbf{X}} = \mathbf{X}^{-1}
$$

Differentiating the equation and setting its derivative to zero we get

$$
\frac{\partial \left\{ \frac{1}{2} \sum_{i=1}^{N} \left\{-T_{1,i}^{(t)} \log\left|\boldsymbol{\Sigma}_1\right|\right\} - \frac{1}{2} tr\left(\boldsymbol{\Sigma}_1^{-1} \mathbf{S}\right) \right\}}{\partial \boldsymbol{\Sigma}_1}
$$

$$
= -\frac{1}{2} \sum_{i=1}^{N} T_{1,i}^{(t)} \boldsymbol{\Sigma}_1^{-1} + \frac{1}{2} \left(\boldsymbol{\Sigma}_1^{-1}\right)^T \mathbf{S} \left(\boldsymbol{\Sigma}_1^{-1}\right)^T \qquad \left| \text{ notice that } \boldsymbol{\Sigma}_1^{-1} \text{ is symmetric} \right.
$$

$$
= -\frac{1}{2} \sum_{i=1}^{N} T_{1,i}^{(t)} \boldsymbol{\Sigma}_1^{-1} + \frac{1}{2} \boldsymbol{\Sigma}_1^{-1} \mathbf{S} \boldsymbol{\Sigma}_1^{-1} = 0
$$

From this we can calculate the updated estimate for $\boldsymbol{\Sigma}_1$ as

$$-\frac{1}{2}\sum_{i=1}^{N} T_{1,i}^{(t)}\boldsymbol{\Sigma}_1^{-1} + \frac{1}{2}\boldsymbol{\Sigma}_1^{-1}\mathbf{S}\boldsymbol{\Sigma}_1^{-1} = 0 \Leftrightarrow$$

$$\sum_{i=1}^{N} T_{1,i}^{(t)} = \mathbf{S}\boldsymbol{\Sigma}_1^{-1} \Leftrightarrow$$

$$\boldsymbol{\Sigma}_1 \sum_{i=1}^{N} T_{1,i}^{(t)} = \mathbf{S} \Leftrightarrow$$

$$\boldsymbol{\Sigma}_1 = \frac{\mathbf{S}}{\sum_{i=1}^{N} T_{1,i}^{(t)}}$$

and by substituting the definition of **S**, the updated estimate becomes

$$\boldsymbol{\Sigma}_1^{(t+1)} = \frac{\sum_{i=1}^{N} T_{1,i}^{(t)}(\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})^T}{\sum_{i=1}^{N} T_{1,i}^{(t)}}$$

Now, similarly for the other distribution ($\boldsymbol{\mu}_2$,$\boldsymbol{\Sigma}_2$) we get

$$\boldsymbol{\mu}_2^{(t+1)} = \frac{\sum_{i=1}^{N} T_{2,i}^{(t)}\mathbf{x}_i^T}{\sum_{i=1}^{N} T_{2,i}^{(t)}}$$

$$\boldsymbol{\Sigma}_2^{(t+1)} = \frac{\sum_{i=1}^{N} T_{2,i}^{(t)}(\mathbf{x}_i - \boldsymbol{\mu}_2^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_2^{(t+1)})^T}{\sum_{i=1}^{N} T_{2,i}^{(t)}}$$

With these new estimates we are again ready to take the E-step to get higher likelihood for the estimated parameters, and so on…

In practice the iterations are run until the updated estimates converge, i.e. the new estimates equal the old estimates with some predefined accuracy.

SUMMA SUMMARUM – PSEUDOCODE FOR THIS EXAMPLE

Initialize $\theta^{(0)} = (\boldsymbol{\tau}^{(0)}, \boldsymbol{\mu}_1^{(0)}, \boldsymbol{\mu}_2^{(0)}, \boldsymbol{\Sigma}_1^{(0)}, \boldsymbol{\Sigma}_2^{(0)})$

For each iteration index (from t=0 until convergence or when the maximum number of iterations is reached)

1. (E-step) Compute: $T_{j,i}^{(t)} = \dfrac{\tau_j^{(t)} f(\mathbf{x}_i \mid \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}{\tau_1^{(t)} f(\mathbf{x}_i \mid \boldsymbol{\mu}_1^{(t)}, \boldsymbol{\Sigma}_1^{(t)}) + \tau_2^{(t)} f(\mathbf{x}_i \mid \boldsymbol{\mu}_2^{(t)}, \boldsymbol{\Sigma}_2^{(t)})}$ for j=1, 2

2. (M-step) Define updated estimates for the next iteration round as

$$\tau_1^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} T_{1,i}^{(t)} \text{ and } \tau_2^{(t+1)} = \frac{1}{N}\sum_{i=1}^{N} T_{2,i}^{(t)}$$

$$\boldsymbol{\mu}_1^{(t+1)} = \frac{\sum_{i=1}^{N} T_{1,i}^{(t)} \mathbf{x}_i^T}{\sum_{i=1}^{N} T_{1,i}^{(t)}} \text{ and } \boldsymbol{\Sigma}_1^{(t+1)} = \frac{\sum_{i=1}^{N} T_{1,i}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_1^{(t+1)})^T}{\sum_{i=1}^{N} T_{1,i}^{(t)}}$$

$$\boldsymbol{\mu}_2^{(t+1)} = \frac{\sum_{i=1}^{N} T_{2,i}^{(t)} \mathbf{x}_i^T}{\sum_{i=1}^{N} T_{2,i}^{(t)}} \text{ and } \boldsymbol{\Sigma}_2^{(t+1)} = \frac{\sum_{i=1}^{N} T_{2,i}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_2^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_2^{(t+1)})^T}{\sum_{i=1}^{N} T_{2,i}^{(t)}}$$

end