Statistical Computing

The Expectation-Maximization Algorithm III
Mixture Model for Multi-dimensional Gaussians

Uwe Menzel, 2018

uwe.menzel@matstat.de

www.matstat.org

Multi-dimensional Gaussians

- Very similar to the 1-D case, for details see part II
- d = dimension of the sample space
- $\boldsymbol{\mu_k}$: vector of means (of length d)
- $\boldsymbol{x_i}$: one observation, vector of length d
- $\boldsymbol{\Sigma_k}$: **covariance matrix** ($d \times d$ matrix):

$$\boldsymbol{\Sigma_k} = \begin{pmatrix} Cov(\boldsymbol{x_1}, \boldsymbol{x_1}) & Cov(\boldsymbol{x_1}, \boldsymbol{x_2}) & \dots & Cov(\boldsymbol{x_1}, \boldsymbol{x_d}) \\ Cov(\boldsymbol{x_2}, \boldsymbol{x_1}) & Cov(\boldsymbol{x_2}, \boldsymbol{x_2}) & \dots & Cov(\boldsymbol{x_2}, \boldsymbol{x_d}) \\ \dots & \dots & \dots & \dots \\ Cov(\boldsymbol{x_d}, \boldsymbol{x_1}) & Cov(\boldsymbol{x_d}, \boldsymbol{x_2}) & \dots & Cov(\boldsymbol{x_d}, \boldsymbol{x_d}) \end{pmatrix}$$

In two dimensions, this translates to:

$$\boldsymbol{\Sigma_k} = \begin{pmatrix} Cov(\boldsymbol{x_1}, \boldsymbol{x_1}) & Cov(\boldsymbol{x_1}, \boldsymbol{x_2}) \\ Cov(\boldsymbol{x_2}, \boldsymbol{x_1}) & Cov(\boldsymbol{x_2}, \boldsymbol{x_2}) \end{pmatrix} = \begin{pmatrix} \sigma_x^2 & \rho \cdot \sigma_x \cdot \sigma_y \\ \rho \cdot \sigma_x \cdot \sigma_y & \sigma_y^2 \end{pmatrix}$$

$$Cov(\boldsymbol{x}, \boldsymbol{x}) = \sigma_x^2 \qquad \rho = \rho(\boldsymbol{x}, \boldsymbol{y}) = \frac{Cov(\boldsymbol{x}, \boldsymbol{y})}{\sigma_x \sigma_y}$$ correlation coefficient

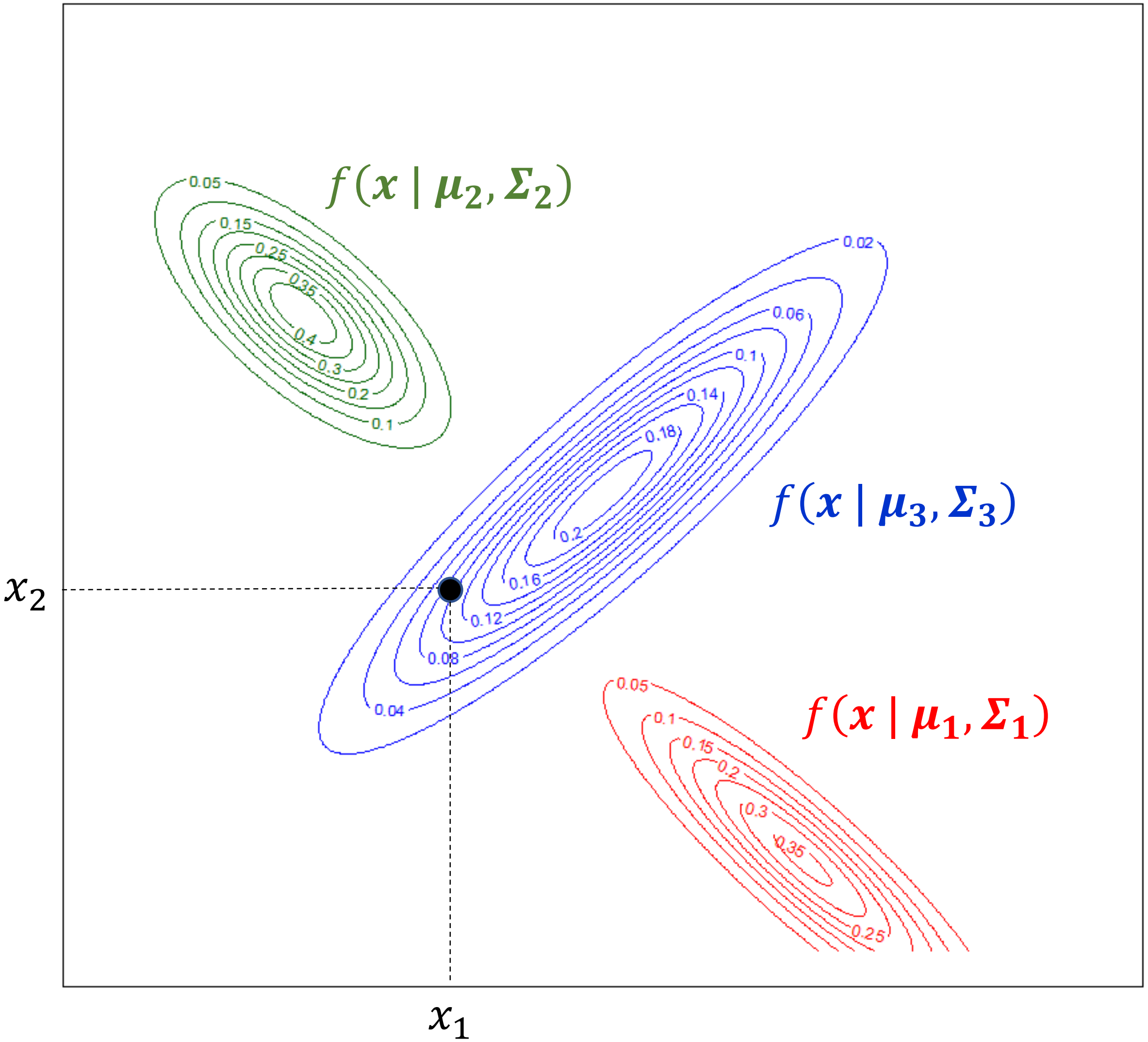
Two-step experiment

Assume we have K d-dimensional Gaussians with densities $f_k(\boldsymbol{x} \mid \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$, where $k = (1, 2, \ldots, K)$. We carry out the same two-step experiment as we did in the one-dimensional case (see part II):

1. Choose a Gaussian f_k randomly with some probability α_k. This can be described by a multinomially distributed random variable Z with sample space $\Omega_Z = \{1, 2, \ldots, K\}$ and probability mass function $P(Z = k) = \alpha_k$. The α_k are constrained by $\sum \alpha_k = 1$ and $\alpha_k > 0$ for all k.
2. Generate a sample $\boldsymbol{x}$ from the above chosen distribution f_k. The vector $\boldsymbol{x}$ is an observation of a d-dimensional normally distributed random variable $\boldsymbol{X}$ with parameters $\boldsymbol{\mu_k}$ and $\boldsymbol{\Sigma_k}$, i.e. $\boldsymbol{X} \sim MVN(\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$.

The figure on the next page illustrates this experiment: In step 1, cluster 3 was chosen. Then the vector $\boldsymbol{x} = (x_1, x_2)$ is generated from the distribution $\boldsymbol{X} \sim MVN(\boldsymbol{\mu_3}, \boldsymbol{\Sigma_3})$.

Contour plot for 2-D Gaussian mixture



Uwe Menzel, 2018

Probability Density Function for multidimensional Gaussian

The experiment includes a discrete (Z) and a d-dimensional continuous ($\boldsymbol{X}$) random variable. The (mixed) **joint density of $\boldsymbol{X}$ and $\boldsymbol{Z}$** can be written:

$$f_{\boldsymbol{X},Z}(\boldsymbol{x}, Z=k) = P(Z=k) \cdot f_{\boldsymbol{X}|Z}(\boldsymbol{x}|Z=k)$$
$$= \alpha_k \cdot f_k(\boldsymbol{x} \mid \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$$

$f_{X|Z}$: conditional probability

where f_k is the probability density function for a d-dimensional Gaussian:

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2} \cdot \sqrt{|\boldsymbol{\Sigma_k}|}} \cdot \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_k})^T \cdot (\boldsymbol{\Sigma_k})^{-1} \cdot (\boldsymbol{x}-\boldsymbol{\mu_k})\right\}$$

Here, $|\boldsymbol{\Sigma_k}|$ is the determinant of the covariance matrix of the k^{th} component (cluster).

Probability Density Function for two-dimensional Gaussian

In the 2-D case, the joint density of $\mathbf{x} = (x, y)$ simplifies to

$$f_k(\boldsymbol{x}) = f_k(x, y) =$$

$$= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \cdot \exp\left\{-\frac{(x-\mu_x)^2}{2\sigma_x^2} + \rho \cdot \frac{(y-\mu_y)(x-\mu_x)}{\sigma_x\sigma_y} - \frac{(y-\mu_y)^2}{2\sigma_y^2}\right\}$$

because $\left|\boldsymbol{\Sigma_k}\right| = \sigma_x^2\sigma_y^2 \cdot \left(1-\rho^2\right)$ and

$$(\boldsymbol{x} - \boldsymbol{\mu_k})^T \cdot (\boldsymbol{\Sigma_k})^{-1} \cdot (\boldsymbol{x} - \boldsymbol{\mu_k}) =$$

$$= \frac{1}{(1-\rho^2)} \cdot \left\{\frac{(x-\mu_x)^2}{\sigma_x^2} - 2 \cdot \rho \frac{(y-\mu_y)(x-\mu_x)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right\}$$

See the appendix for a derivation of the two-dimensional expressions.

Gaussian mixture

We have to find parameters $\boldsymbol{\theta}$ that maximize $f_X(\boldsymbol{x} \mid \boldsymbol{\theta})$. As in the 1-D case, an expression for the density $f_X(\boldsymbol{x} \mid \boldsymbol{\theta})$ that incorporates the latent variables Z can be found by using the **law of total probability:**

$$f_X(\boldsymbol{x} \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \underbrace{f_{\boldsymbol{X}|Z=k}(\boldsymbol{x} \mid Z = k)}_{f_k(\boldsymbol{x} \mid \mu_k, \Sigma_k)} \cdot \underbrace{P(Z = k)}_{\alpha_k} = \sum_{k=1}^{K} \alpha_k \cdot f_k(\boldsymbol{x} \mid \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$$

where K is the number of clusters.

The density f_X can be seen as a superposition of multiple probability density functions (Gaussians):

$$f_X(\boldsymbol{x} \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \alpha_k \cdot f_k(\boldsymbol{x} \mid \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$$

Maximum Likelihood for a Gaussian mixture

If we have multiple independent observations $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N)^T$, the likelihood is the product of the density for the individual observations:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{N} f_X(\boldsymbol{x_i} \mid \boldsymbol{\theta}) = \prod_{i=1}^{N} \sum_{k=1}^{K} \alpha_k \cdot f_k(\boldsymbol{x_i} \mid \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$$

Note that each $\boldsymbol{x_i}$ is a d-dimensional vector here. We aim at calculating the $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta})$:

$$\boldsymbol{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\, L(\boldsymbol{\theta})$$

i.e. we search for the parameter (vector) $\boldsymbol{\theta}$ that makes the observed data most likely. Often, it is more convenient to maximize the logarithm of $L(\boldsymbol{\theta})$:

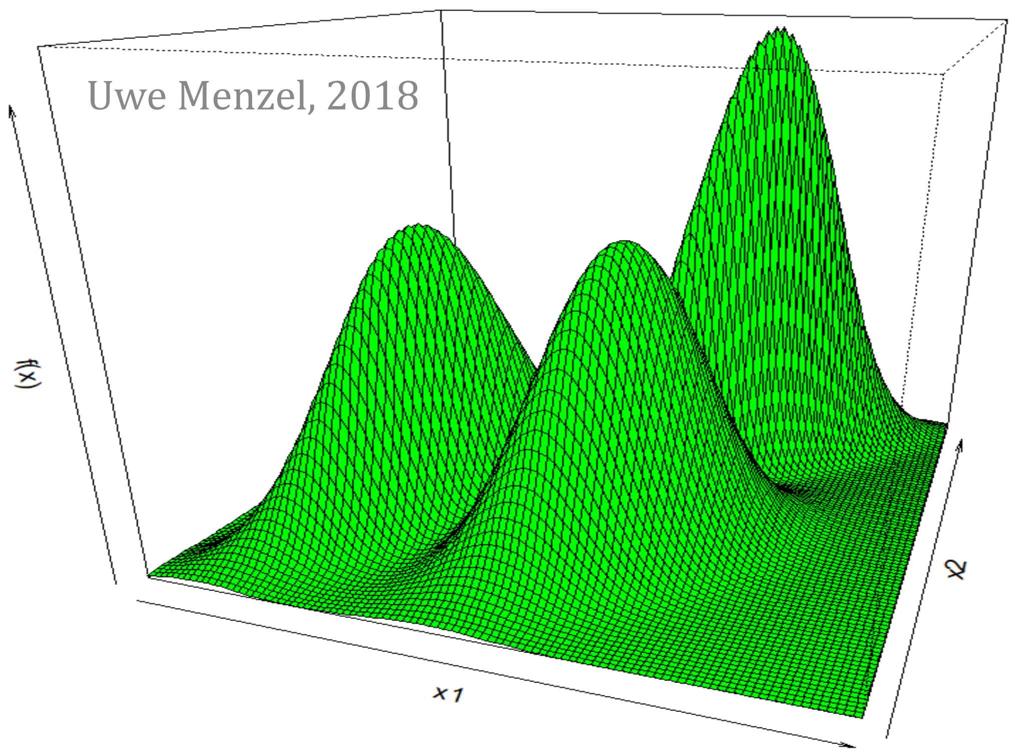
$$l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} \alpha_k \cdot f(\boldsymbol{x_i} \mid \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \right]$$

Mixture of Gaussians: E-step

1. **Initialize**: $\boldsymbol{\theta}_t = 1^{\text{st}}$ guess for the set of parameters

2. **E-step**: calculate $P(Z_i = k \mid \boldsymbol{X_i} = \boldsymbol{x_i}, \boldsymbol{\theta}_t)$.

This is the probability that Z_i is equal to k, i.e. the probability that an observation $\boldsymbol{x_i}$ originates from the k^{th} Gaussian, given that observation ($\boldsymbol{x_i}$) and all parameters $\boldsymbol{\theta_t} = \{\alpha_k^t, \boldsymbol{\mu}_k^t, \boldsymbol{\Sigma}_k^t\}$. Since the parameters can be considered as given (we made a 1^{st} guess) , we know the exact positions and shapes of the d-dimensional Gaussians and it's superposition, as shown in the figure for $d = 2$.



Mixture of Gaussians: E-step

2. **E-step**: calculate $P(Z_i = k \mid \boldsymbol{X} = \boldsymbol{x_i}, \boldsymbol{\theta}_t)$.

Knowledge of $\boldsymbol{x_i}$ and $\boldsymbol{\theta_t}$ enables us to calculate the conditional probability using **Bayes theorem** :

$$P\left(Z_i = k \mid \boldsymbol{X} = \boldsymbol{x_i}, \boldsymbol{\theta_t}\right) = \frac{f_{\boldsymbol{X}|Z}\left(\boldsymbol{x_i} | Z_i = k, \boldsymbol{\theta_t}\right) \cdot P\left(Z_i = k | \boldsymbol{\theta_t}\right)}{f_X\left(\boldsymbol{x_i} | \boldsymbol{\theta_t}\right)}$$

$$= \frac{f_k\left(\boldsymbol{x_i} \mid \boldsymbol{\mu_k^t}, \boldsymbol{\Sigma_k^t}\right) \cdot \alpha_k^t}{\sum_k \alpha_k^t \cdot f_k\left(\boldsymbol{x_i} \mid \boldsymbol{\mu_k^t}, \boldsymbol{\Sigma_k^t}\right)} = \omega_{ik}$$

The last ratio is labelled ω_{ik} and often named ”degree of membership” (of observation x_i to component k). The ω_{ik} are known numbers since they are calculated based on the known $\boldsymbol{\theta_t} = \{\alpha_k^t, \mu_k^t, \sigma_k^t\}$.

$$\omega_{ik} = \frac{\alpha_k^t \cdot f_k\left(\boldsymbol{x_i} \mid \boldsymbol{\mu_k^t}, \boldsymbol{\Sigma_k^t}\right)}{\sum_k \alpha_k^t \cdot f_k\left(\boldsymbol{x_i} \mid \boldsymbol{\mu_k^t}, \boldsymbol{\Sigma_k^t}\right)} \qquad \sum_{k=1}^{K} \omega_{ik} = 1$$

Degree of membership: ω_{ik}

It is not easy to illustrate the geometric meaning of the ω_{ik} in the multi-dimesional case. In two dimensions, having two clusters, we might see the figure below as a cross section perpendicular to the $x - y$ plane.

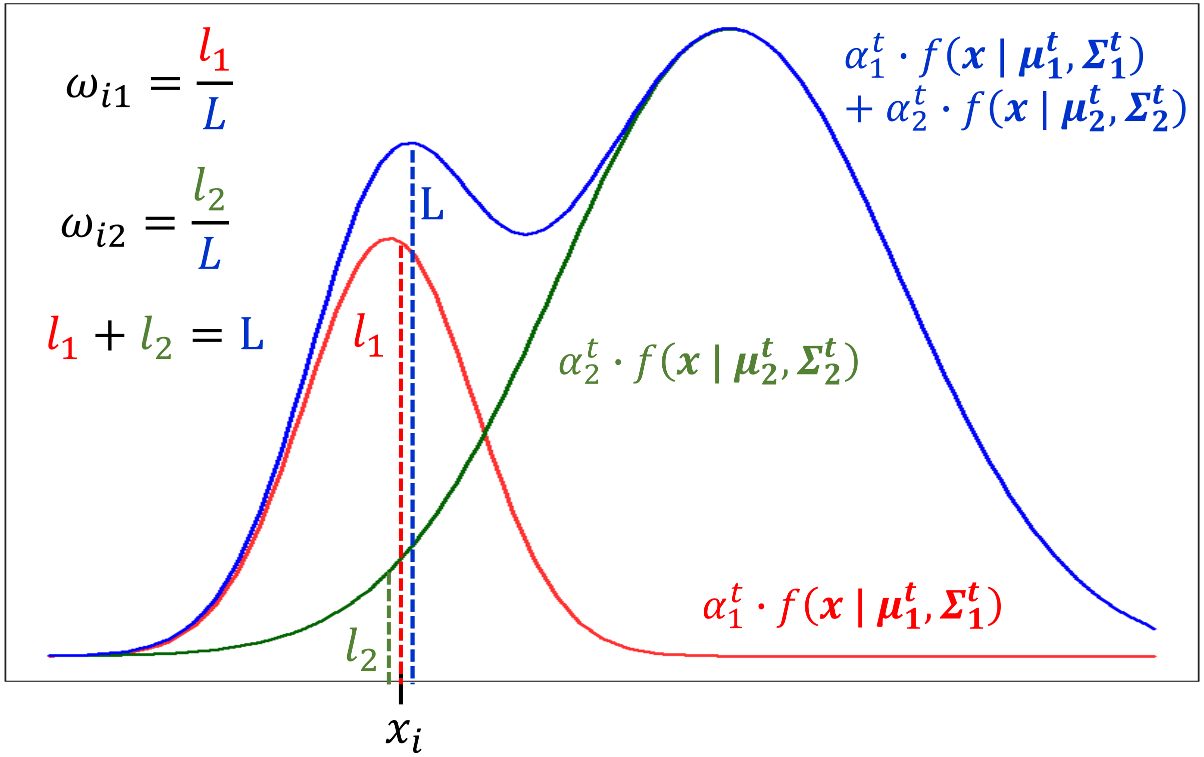


Illustration of the ω_{ik} (dashed lines jittered around x_i for better visibility)

Completion of the E-step

It remains to calculate:

$$Q_1(\boldsymbol{\theta}, \boldsymbol{\theta_t}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \underbrace{P(Z_i = k \mid \boldsymbol{X} = \boldsymbol{x_i}, \boldsymbol{\theta_t})}_{\omega_{ik}} \cdot \log f_{\boldsymbol{X},Z}(\boldsymbol{x_i}, Z_i = k \mid \boldsymbol{\theta})$$

$$f_{\boldsymbol{X},Z}(\boldsymbol{x}, Z) = \alpha_k \cdot f_k(\boldsymbol{x} \mid \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})$$ (mixed) joint probability distribution

$$Q_1(\boldsymbol{\theta}, \boldsymbol{\theta_t}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \omega_{ik} \cdot \log \underbrace{\{\alpha_k \cdot f_k(\boldsymbol{x_i} \mid \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})\}}_{\text{unknown parameters (depending on } \boldsymbol{\theta})}$$

unknown parameters (depending on $\boldsymbol{\theta}$)

ω_{ik} known (depending on $\boldsymbol{\theta}_t$)

The expression Q_1 has to be maximized for the unknown $\alpha_k, \mu_k, \sigma_k$ → **M-step**.

Mixture of multivariate Gaussians: M-step

$$Q_1(\boldsymbol{\theta}, \boldsymbol{\theta_t}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \omega_{ik} \cdot \log \left\{ \alpha_k \cdot f_k(\boldsymbol{x_i} \mid \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \right\}$$

Q_1 has to be maximized for the unknown $\alpha_k, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}$

$$Q_1(\boldsymbol{\theta}, \boldsymbol{\theta_t}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \omega_{ik} \cdot \left\{ \log \alpha_k + \log f_k(\boldsymbol{x_i} \mid \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \right\}$$

$$\log f_k(\boldsymbol{x_i} \mid \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) = -\frac{d}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma_k}| - \frac{1}{2} (\boldsymbol{x_i} - \boldsymbol{\mu_k})^T \cdot (\boldsymbol{\Sigma_k})^{-1} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_k})$$

$$Q_1(\boldsymbol{\theta}, \boldsymbol{\theta_t}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \omega_{ik} \cdot \left\{ \log \alpha_k - \frac{d}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma_k}| - \frac{1}{2} (\boldsymbol{x_i} - \boldsymbol{\mu_k})^T \cdot (\boldsymbol{\Sigma_k})^{-1} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_k}) \right\}$$

The parameters $\boldsymbol{\Sigma_k}, \boldsymbol{\mu_k}$ and α_k in the curly brackets are the actual variables Q_1 must be maximized for, while the ω_{ik} contain only variables that have been specified by first guess, so that ω_{ik} can be treated as a constant in this expression.

Mixture of Gaussians: M-step

$$Q_1(\boldsymbol{\theta}, \boldsymbol{\theta_t}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \omega_{ik} \cdot \left\{ \log \alpha_k - \frac{1}{2} \log |\boldsymbol{\Sigma_k}| - \frac{1}{2} (\boldsymbol{x_i} - \boldsymbol{\mu_k})^T \cdot (\boldsymbol{\Sigma_k})^{-1} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_k}) \right\}$$

(here, we have ignored the term $-\frac{d}{2} \cdot \log 2\pi$ since it doesn't depend on any parameter)

Maximization of $\boldsymbol{Q_1}$ with regard to $\boldsymbol{\mu_m}$:

$$\frac{\partial Q_1}{\partial \boldsymbol{\mu_m}} = -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\mu_m}} \sum_{i=1}^{N} \sum_{k=1}^{K} \omega_{ik} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_k})^T \cdot (\boldsymbol{\Sigma_k})^{-1} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_k})$$

$$= -\frac{1}{2} \sum_{i=1}^{N} \omega_{im} \cdot \frac{\partial}{\partial \boldsymbol{\mu_m}} (\boldsymbol{x_i} - \boldsymbol{\mu_m})^T \cdot (\boldsymbol{\Sigma_m})^{-1} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_m})$$

$$= \sum_{i=1}^{N} \omega_{im} \cdot (\boldsymbol{\Sigma_m})^{-1} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_m}) = 0$$

(see Appendix for the last step)

$$\Longrightarrow \quad (\boldsymbol{\Sigma_m})^{-1} \sum_{i=1}^{N} \omega_{im} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_m}) = 0$$

Mixture of Gaussians: M-step

$$(\boldsymbol{\Sigma_m})^{-1}\sum_{i=1}^{N}\omega_{im}\cdot(\boldsymbol{x_i}-\boldsymbol{\mu_m})=0$$

$\Rightarrow$ $$\sum_{i=1}^{N}\omega_{im}\cdot(\boldsymbol{x_i}-\boldsymbol{\mu_m})=0$$ (Σ is positive definite, covariance)

$\Rightarrow$ $$\boldsymbol{\mu_m}=\frac{\sum_{i=1}^{N}\omega_{im}\cdot\boldsymbol{x_i}}{\sum_{i=1}^{N}\omega_{im}}$$

The new means are weighted means of the x_i (weighted with the degree of membership of each datapoint). This can be compared with the ML estimation of the mean for a single Gaussian: $\mu=\frac{1}{N}\cdot\sum x_i$

This is similar to the 1-D case where we had $\mu_m=\frac{\sum_i^N\omega_{im}\cdot x_i}{\sum_i^N\omega_{im}}$

Mixture of Gaussians: M-step

$$Q_1(\boldsymbol{\theta}, \boldsymbol{\theta_t}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \omega_{ik} \cdot \left\{ \log \alpha_k - \frac{1}{2} \log |\boldsymbol{\Sigma_k}| - \frac{1}{2} (\boldsymbol{x_i} - \boldsymbol{\mu_k})^T \cdot (\boldsymbol{\Sigma_k})^{-1} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_k}) \right\}$$

Maximization of $\boldsymbol{Q_1}$ with respect to $\boldsymbol{\Sigma_m}$:

Instead of deriving for $\boldsymbol{\Sigma_m}$, we derive for $(\boldsymbol{\Sigma_m})^{-1}$. This will automatically lead to an expression for $\boldsymbol{\Sigma_m}$. This strategy was already presented by Xavier Bourret Sicotte[1].

$$\frac{\partial Q_1}{\partial (\boldsymbol{\Sigma_m})^{-1}} = -\frac{1}{2} \sum_{i=1}^{N} \omega_{im} \cdot \frac{\partial}{\partial (\boldsymbol{\Sigma_m})^{-1}} \left\{ \log |\boldsymbol{\Sigma_m}| + (\boldsymbol{x_i} - \boldsymbol{\mu_m})^T \cdot (\boldsymbol{\Sigma_m})^{-1} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_m}) \right\}$$

We obtain:

$$\frac{\partial Q_1}{\partial (\boldsymbol{\Sigma_m})^{-1}} = \frac{1}{2} \sum_{i=1}^{N} \omega_{im} \cdot \left\{ \boldsymbol{\Sigma_m} - (\boldsymbol{x_i} - \boldsymbol{\mu_m}) \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_m})^T \right\}$$

see Appendix

[1] https://stats.stackexchange.com/questions/351549/maximum-likelihood-estimators-multivariate-gaussian

Mixture of Gaussians: M-step

$$\frac{\partial Q_1}{\partial\left(\boldsymbol{\Sigma_m}\right)^{-1}} = \frac{1}{2}\sum_{i=1}^{N}\omega_{im}\cdot\left\{\boldsymbol{\Sigma_m} - \left(\boldsymbol{x_i} - \boldsymbol{\mu_m}\right)\cdot\left(\boldsymbol{x_i} - \boldsymbol{\mu_m}\right)^T\right\}$$

$$\Rightarrow \quad \frac{1}{2}\sum_{i=1}^{N}\omega_{im}\cdot\left\{\boldsymbol{\Sigma_m} - \left(\boldsymbol{x_i} - \boldsymbol{\mu_m}\right)\cdot\left(\boldsymbol{x_i} - \boldsymbol{\mu_m}\right)^T\right\} = 0$$

$$\Rightarrow \quad \boldsymbol{\Sigma_m} = \frac{\sum_{i=1}^{N}\omega_{im}\cdot\left(\boldsymbol{x_i} - \boldsymbol{\mu_m}\right)\cdot\left(\boldsymbol{x_i} - \boldsymbol{\mu_m}\right)^T}{\sum_{i=1}^{N}\omega_{im}}$$

The new estimate for the covariance matrix $\boldsymbol{\Sigma_m}$ is a matrix where each element is a weighted mean of the squared distances between datapoints and mean $\boldsymbol{\mu_m}$ (weighted with the degree of membership of each datapoint). Thsi can be compared with the one-dimensional case, where we obtained for the variance:

$$\sigma_m^2 = \frac{\sum_i \omega_{im}\cdot\left(x_i - \mu_m\right)^2}{\sum_i \omega_{im}}$$

Mixture of Gaussians: M-step

$$Q_1(\boldsymbol{\theta}, \boldsymbol{\theta_t}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \omega_{ik} \cdot \left\{ \log \alpha_k - \frac{1}{2} \log |\boldsymbol{\Sigma_k}| - \frac{1}{2} (\boldsymbol{x_i} - \boldsymbol{\mu_k})^T \cdot (\boldsymbol{\Sigma_k})^{-1} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_k}) \right\}$$

Maximization of $\boldsymbol{Q_1}$ with respect to $\boldsymbol{\alpha_m}$:

$$\frac{\partial Q_1}{\partial \alpha_m} = -\frac{\partial}{\partial \alpha_m} \sum_{i=1}^{N} \sum_{k=1}^{K} \omega_{ik} \cdot \log \alpha_k$$

This has to be solved with the constraint $\sum_k \alpha_k = 1$

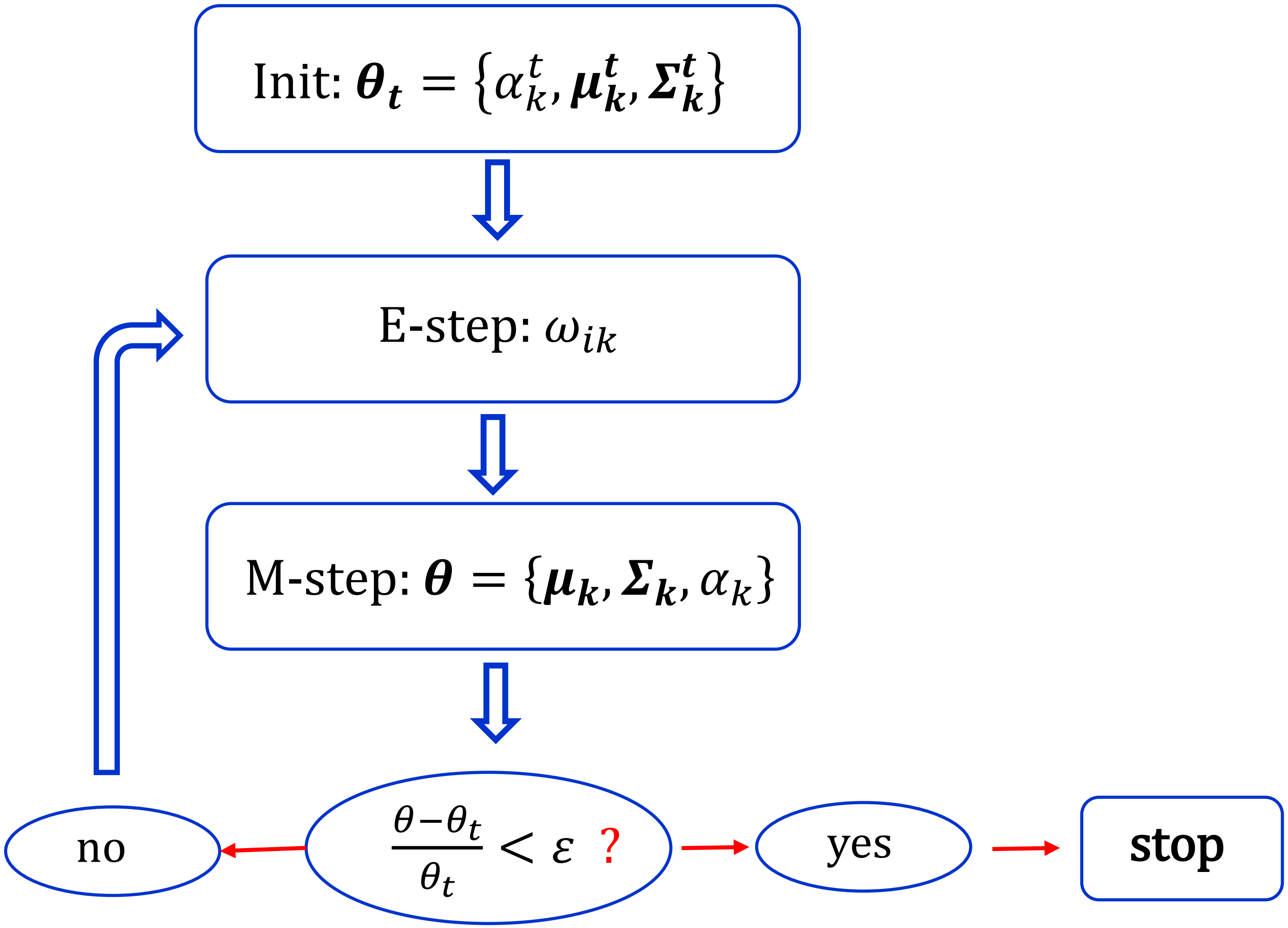
This expression is the same as for the one-dimensional case, so that we can directly adopt the solution (see part II):

$$\alpha_m = \frac{1}{N} \cdot \sum_{i=1}^{N} \omega_{im}$$

Again, it is $\sum_{m=1}^{K} \alpha_m = 1$ and $\sum_{i=1}^{N} \sum_{m=1}^{K} \omega_{im} = N$

Iteration

Now, we set $\boldsymbol{\theta}_t = \boldsymbol{\theta}$ and repeat the E step. This continues until convergence is reached:



Summary of EM for multidimensional Gaussian mixture

Initialization: 1st guess $\boldsymbol{\theta_t} = \{\alpha_k^t, \boldsymbol{\mu_k^t}, \boldsymbol{\Sigma_k^t}\}$

E-step:

$$\omega_{ik} = \frac{\alpha_k^t \cdot f_k\left(\boldsymbol{x_i} \mid \boldsymbol{\mu_k^t}, \boldsymbol{\Sigma_k^t}\right)}{\sum_k \alpha_k^t \cdot f_k\left(\boldsymbol{x_i} \mid \boldsymbol{\mu_k^t}, \boldsymbol{\Sigma_k^t}\right)}$$

M-step:

$$\boldsymbol{\mu_m} = \frac{\sum_{i=1}^N \omega_{im} \cdot \boldsymbol{x_i}}{\sum_{i=1}^N \omega_{im}}$$

$$\boldsymbol{\Sigma_m} = \frac{\sum_{i=1}^N \omega_{im} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_m}) \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_m})^T}{\sum_{i=1}^N \omega_{im}}$$

$$\alpha_m = \frac{1}{N} \cdot \sum_{i=1}^N \omega_{im}$$

$$\sum_{k=1}^K \omega_{ik} = 1$$

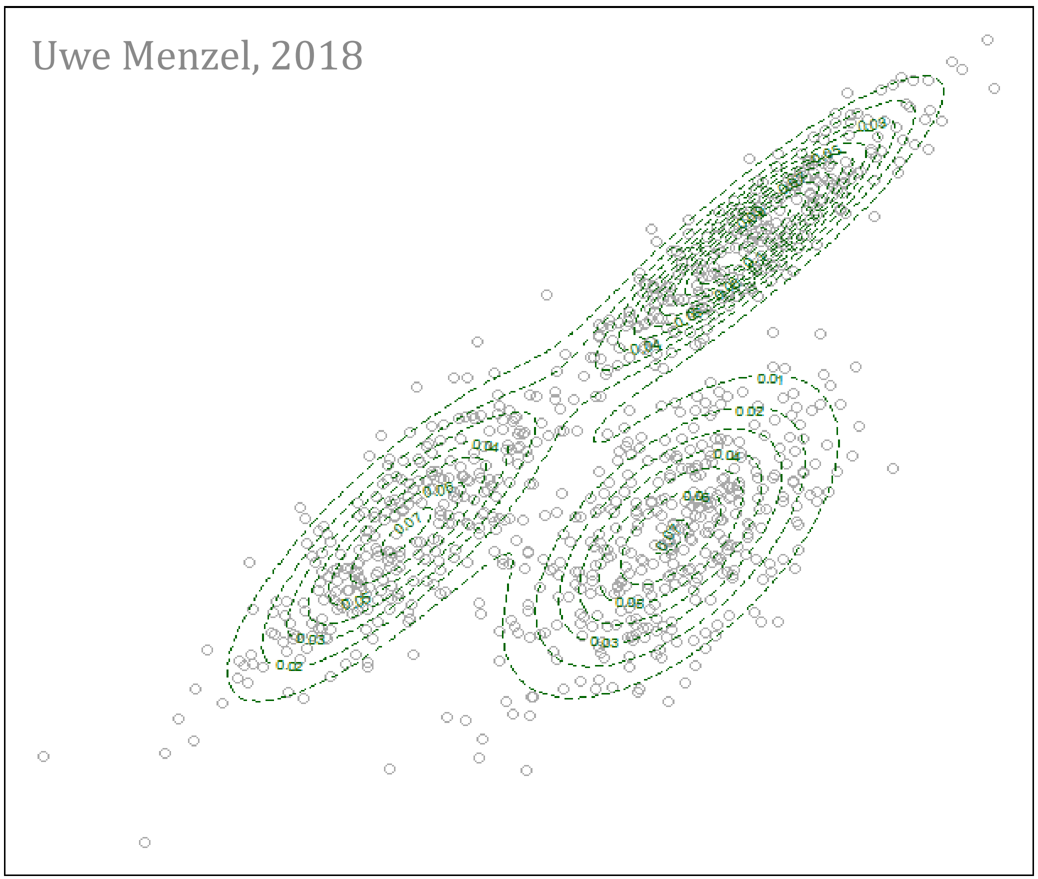
$$\sum_{i=1}^N \sum_{m=1}^K \omega_{im} = N$$

Iterate between E- and M-step until convergence, i.e. until $\frac{\theta - \theta_t}{\theta_t} < \varepsilon$.

Alternatively, check convergence of the likelihood.

R-script, EM for 2-D Gaussian mixture

- In the R-code provided, we generate sample points for a two-dimensional Gaussian mixture by modelling the two-step experiment described above (grey points in the figure below)
- Starting with these data, we calculate the means, covariance matrices, and mixture weights using the EM algorithm presented here. The solution is indicated by the green contour lines.



Appendix

The Expectation-Maximization algorithm III

Uwe Menzel, 2018
uwe.menzel@slu.se ; uwe.menzel@matstat.de
www.matstat.org

Probability density for the 2-D Gaussian

Probability density function for the d-dimensional case:

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2} \cdot \sqrt{|\boldsymbol{\Sigma_k}|}} \cdot \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu_k})^T \cdot (\boldsymbol{\Sigma_k})^{-1} \cdot (\boldsymbol{x} - \boldsymbol{\mu_k}) \right\}$$

Probability density function for the 2-dimensional case:

$$f_k(\boldsymbol{x}) = f_k(x, y) =$$
$$= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \cdot \exp\left\{ -\frac{(x-\mu_x)^2}{2\sigma_x^2} + \rho \cdot \frac{(y-\mu_y)(x-\mu_x)}{\sigma_x\sigma_y} - \frac{(y-\mu_y)^2}{2\sigma_y^2} \right\}$$

Determinat of the covariance matrix for the 2-D case:

$$|\boldsymbol{\Sigma_k}| = \det \begin{pmatrix} \sigma_x^2 & \rho \cdot \sigma_x\sigma_y \\ \rho \cdot \sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} = \sigma_x^2\sigma_y^2 - \rho^2 \cdot \sigma_x^2\sigma_y^2 = \sigma_x^2\sigma_y^2 \cdot (1 - \rho^2)$$

Inverse of the 2D covariance matrix

A **general expression** for the inverse of a 2-D matrix is:

$$\boldsymbol{A}^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{\det \boldsymbol{A}} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Hence, the **inverse of the 2-D covariance matrix** is:

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} \sigma_x^2 & \rho \cdot \sigma_x \sigma_y \\ \rho \cdot \sigma_x \sigma_y & \sigma_y^2 \end{pmatrix}^{-1} = \frac{1}{\sigma_x^2 \sigma_y^2 \cdot (1 - \rho^2)} \cdot \begin{pmatrix} \sigma_y^2 & -\rho \cdot \sigma_x \sigma_y \\ -\rho \cdot \sigma_x \sigma_y & \sigma_x^2 \end{pmatrix}$$

We still need the exponent $(\boldsymbol{x} - \boldsymbol{\mu})^T \cdot (\boldsymbol{\Sigma})^{-1} \cdot (\boldsymbol{x} - \boldsymbol{\mu})$ ⇨

(index ***k*** suppressd to simplify notation)

Note:

$(\boldsymbol{x} - \boldsymbol{\mu})$ is a column vector: 2x1

$$(\boldsymbol{x} - \boldsymbol{\mu}) = \begin{pmatrix} x - \mu_x \\ y - \mu_y \end{pmatrix}$$

$(\boldsymbol{x} - \boldsymbol{\mu})^T$ is a row vector: 1x2

$$(\boldsymbol{x} - \boldsymbol{\mu})^T = (x - \mu_x \, , y - \mu_y)$$

Exponent of the 2D probability density function

$$(\boldsymbol{\Sigma})^{-1}\cdot(\boldsymbol{x}-\boldsymbol{\mu}) = \frac{1}{\sigma_x^2\sigma_y^2\cdot(1-\rho^2)}\cdot\begin{pmatrix}\sigma_y^2 & -\rho\cdot\sigma_x\sigma_y\\ -\rho\cdot\sigma_x\sigma_y & \sigma_x^2\end{pmatrix}\cdot\begin{pmatrix}x-\mu_x\\ y-\mu_y\end{pmatrix}$$

$$= \frac{1}{\sigma_x^2\sigma_y^2\cdot(1-\rho^2)}\cdot\begin{pmatrix}\sigma_y^2(x-\mu_x)-\rho\cdot\sigma_x\sigma_y(y-\mu_y)\\ -\rho\cdot\sigma_x\sigma_y(x-\mu_x)+\sigma_x^2(y-\mu_y)\end{pmatrix}$$

It remains to multiply this with $(\boldsymbol{x}-\boldsymbol{\mu})^T$ from the left side:

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\cdot(\boldsymbol{\Sigma})^{-1}\cdot(\boldsymbol{x}-\boldsymbol{\mu}) =$$

$$= \frac{1}{\sigma_x^2\sigma_y^2\cdot(1-\rho^2)}\begin{pmatrix}x-\mu_x\\ y-\mu_y\end{pmatrix}^T\cdot\begin{pmatrix}\sigma_y^2(x-\mu_x)-\rho\cdot\sigma_x\sigma_y(y-\mu_y)\\ -\rho\cdot\sigma_x\sigma_y(x-\mu_x)+\sigma_x^2(y-\mu_y)\end{pmatrix}$$

$$= \frac{1}{\sigma_x^2\sigma_y^2\cdot(1-\rho^2)}\cdot\left\{\sigma_y^2(x-\mu_x)^2-2\cdot\rho\cdot\sigma_x\sigma_y(y-\mu_y)(x-\mu_x)+\sigma_x^2(y-\mu_y)^2\right\}$$

$$= \frac{1}{(1-\rho^2)}\cdot\left\{\frac{(x-\mu_x)^2}{\sigma_x^2}-2\cdot\rho\frac{(y-\mu_y)(x-\mu_x)}{\sigma_x\sigma_y}+\frac{(y-\mu_y)^2}{\sigma_y^2}\right\}$$

Derivative of the quadratic form $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}$

If the matrix $\boldsymbol{A}$ is symmetric ($\boldsymbol{A}^T = \boldsymbol{A}$) and independent of the vector $\boldsymbol{a}$, the following is generally valid:

$$\frac{\partial}{\partial \boldsymbol{a}}\left(\boldsymbol{a}^T \boldsymbol{A}\boldsymbol{a}\right) = 2\boldsymbol{A}\boldsymbol{a}$$

Therefore:

$$\frac{\partial}{\partial \underbrace{(\boldsymbol{x_i} - \boldsymbol{\mu_m})}_{\boldsymbol{a}}}\left[\underbrace{(\boldsymbol{x_i} - \boldsymbol{\mu_m})^T}_{\boldsymbol{a}^T} \cdot \underbrace{(\boldsymbol{\Sigma_m})^{-1}}_{\boldsymbol{A}} \cdot \underbrace{(\boldsymbol{x_i} - \boldsymbol{\mu_m})}_{\boldsymbol{a}}\right] = 2 \cdot \underbrace{(\boldsymbol{\Sigma_m})^{-1}}_{\boldsymbol{A}} \underbrace{(\boldsymbol{x_i} - \boldsymbol{\mu_m})}_{\boldsymbol{a}}$$

Using the chain rule, we get:

$$\frac{\partial}{\partial \boldsymbol{\mu_m}} = \frac{\partial (\boldsymbol{x}_i - \boldsymbol{\mu_m})}{\partial \boldsymbol{\mu_m}} \cdot \frac{\partial}{\partial (\boldsymbol{x}_i - \boldsymbol{\mu_m})} = (-\boldsymbol{I}) \cdot \frac{\partial}{\partial (\boldsymbol{x}_i - \boldsymbol{\mu_m})}$$

$\boldsymbol{I}$ is the identity matrix

Finally, we get:

$$\frac{\partial}{\partial \boldsymbol{\mu_m}}\left[(\boldsymbol{x_i} - \boldsymbol{\mu_m})^T \cdot (\boldsymbol{\Sigma_m})^{-1} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_m})\right] = -2 \cdot (\boldsymbol{\Sigma_m})^{-1} (\boldsymbol{x_i} - \boldsymbol{\mu_m})$$

Derivation of Q_1 for $(\Sigma_m)^{-1}$

Instead of deriving w.r.t. Σ_m

$$\frac{\partial Q_1}{\partial \boldsymbol{\Sigma_m}} = -\frac{1}{2}\sum_{i=1}^{N} \omega_{im} \cdot \frac{\partial}{\partial \boldsymbol{\Sigma_m}} \left\{ \log|\boldsymbol{\Sigma_m}| + (\boldsymbol{x_i} - \boldsymbol{\mu_m})^T \cdot (\boldsymbol{\Sigma_m})^{-1} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_m}) \right\}$$

we derive w.r.t. $(\Sigma_m)^{-1}$, so that we have to calculate :

$$\frac{\partial Q_1}{\partial (\boldsymbol{\Sigma_m})^{-1}} = -\frac{1}{2}\sum_{i=1}^{N} \omega_{im} \cdot \frac{\partial}{\partial (\boldsymbol{\Sigma_m})^{-1}} \left\{ \log|\boldsymbol{\Sigma_m}| + (\boldsymbol{x_i} - \boldsymbol{\mu_m})^T \cdot (\boldsymbol{\Sigma_m})^{-1} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_m}) \right\}$$

Derivation of Q_1 for $(\Sigma_m)^{-1}$

$$\frac{\partial Q_1}{\partial (\boldsymbol{\Sigma_m})^{-1}} = -\frac{1}{2}\sum_{i=1}^{N} \omega_{im} \cdot \frac{\partial}{\partial (\boldsymbol{\Sigma_m})^{-1}} \left\{ \log|\boldsymbol{\Sigma_m}| + (\boldsymbol{x_i} - \boldsymbol{\mu_m})^T \cdot (\boldsymbol{\Sigma_m})^{-1} \cdot (\boldsymbol{x_i} - \boldsymbol{\mu_m}) \right\}$$

Calculation of $$\frac{\partial}{\partial (\boldsymbol{\Sigma_m})^{-1}} \log|\boldsymbol{\Sigma_m}|$$

$|\boldsymbol{\Sigma_m}|$ is the determinat of $\boldsymbol{\Sigma_m}$

Since $|\boldsymbol{A}^{-1}| = 1/|\boldsymbol{A}|$ we can write

$$\frac{\partial}{\partial (\boldsymbol{\Sigma_m})^{-1}} \log|\boldsymbol{\Sigma_m}| = -\frac{\partial}{\partial (\boldsymbol{\Sigma_m})^{-1}} \log|(\boldsymbol{\Sigma_m})^{-1}|$$

Furthermore, if **A** is symmetric, we have: $$\frac{\partial}{\partial \boldsymbol{A}} \log|\boldsymbol{A}| = \boldsymbol{A}^{-1} \qquad (*)$$

so that $$\frac{\partial}{\partial (\boldsymbol{\Sigma_m})^{-1}} \log|\boldsymbol{\Sigma_m}| = -\boldsymbol{\Sigma_m}$$

(∗): see ”The Matrix Cookbook” by Petersen & Pedersen
https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf)

Derivation of Q_1 for $(\Sigma_m)^{-1}$

$$\frac{\partial Q_1}{\partial\left(\boldsymbol{\Sigma_m}\right)^{-1}} = -\frac{1}{2}\sum_{i=1}^{N}\omega_{im}\cdot\frac{\partial}{\partial\left(\boldsymbol{\Sigma_m}\right)^{-1}}\left\{\log\left|\boldsymbol{\Sigma_m}\right| + \left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)^T\cdot\left(\boldsymbol{\Sigma_m}\right)^{-1}\cdot\left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)\right\}$$

Calculation of $\frac{\partial}{\partial\left(\boldsymbol{\Sigma_m}\right)^{-1}}\left\{\left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)^T\cdot\left(\boldsymbol{\Sigma_m}\right)^{-1}\cdot\left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)\right\}$

In general we have : $\frac{\partial \boldsymbol{a}^T\boldsymbol{X}\boldsymbol{a}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{a}^T \quad (*)$

Using this, we get:

$$\frac{\partial}{\partial\left(\boldsymbol{\Sigma_m}\right)^{-1}}\left\{\left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)^T\cdot\left(\boldsymbol{\Sigma_m}\right)^{-1}\cdot\left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)\right\} = \left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)\cdot\left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)^T$$

$(*)$ see “The Matrix Cookbook” by Petersen and Pedersen, eqn. (72)
https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

Derivation of Q_1 for $(\Sigma_m)^{-1}$

In **summary**, we obtained

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

$$\frac{\partial}{\partial\left(\boldsymbol{\Sigma_m}\right)^{-1}}\left\{\left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)^T\cdot\left(\boldsymbol{\Sigma_m}\right)^{-1}\cdot\left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)\right\} = \left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)\cdot\left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)^T$$

so that the final result becomes

$$\frac{\partial Q_1}{\partial\left(\boldsymbol{\Sigma_m}\right)^{-1}} = -\frac{1}{2}\sum_{i=1}^{N}\omega_{im}\cdot\frac{\partial}{\partial\left(\boldsymbol{\Sigma_m}\right)^{-1}}\left\{\log\left|\boldsymbol{\Sigma_m}\right| + \left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)^T\cdot\left(\boldsymbol{\Sigma_m}\right)^{-1}\cdot\left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)\right\}$$

$$= \frac{1}{2}\sum_{i=1}^{N}\omega_{im}\cdot\left\{\boldsymbol{\Sigma_m}-\left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)\cdot\left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)^T\right\}$$

$$\boxed{\frac{\partial Q_1}{\partial\left(\boldsymbol{\Sigma_m}\right)^{-1}} = \frac{1}{2}\sum_{i=1}^{N}\omega_{im}\cdot\left\{\boldsymbol{\Sigma_m}-\left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)\cdot\left(\boldsymbol{x_i}-\boldsymbol{\mu_m}\right)^T\right\}}$$