Statistical Computing

The Expectation-Maximization Algorithm II
The Mixture Model for 1-D Gaussians

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

Summary of the General EM scheme

1. **Initialize**: $\theta_t = 1^{\text{st}}$ guess for the parameter vector θ

2. **E-step**: calculate $P(Z = k \mid X = x_i, \theta_t)$ and then

$$Q_1(\theta, \theta_t) = \sum_{i=1}^{N} \sum_{k=1}^{K} P(Z = k | X = x_i, \theta_t) \cdot \log f_{X,Z}(x_i, Z = k | \theta)$$

- $X = \{x_1, x_2, \ldots, x_N\}$ are the genuine observations
- $Z = \{z_1, z_2, \ldots, z_K\}$ are the latent variables

3. M-step: update the estimate of the model parameters: $\theta_t \to \theta_{t+1}$

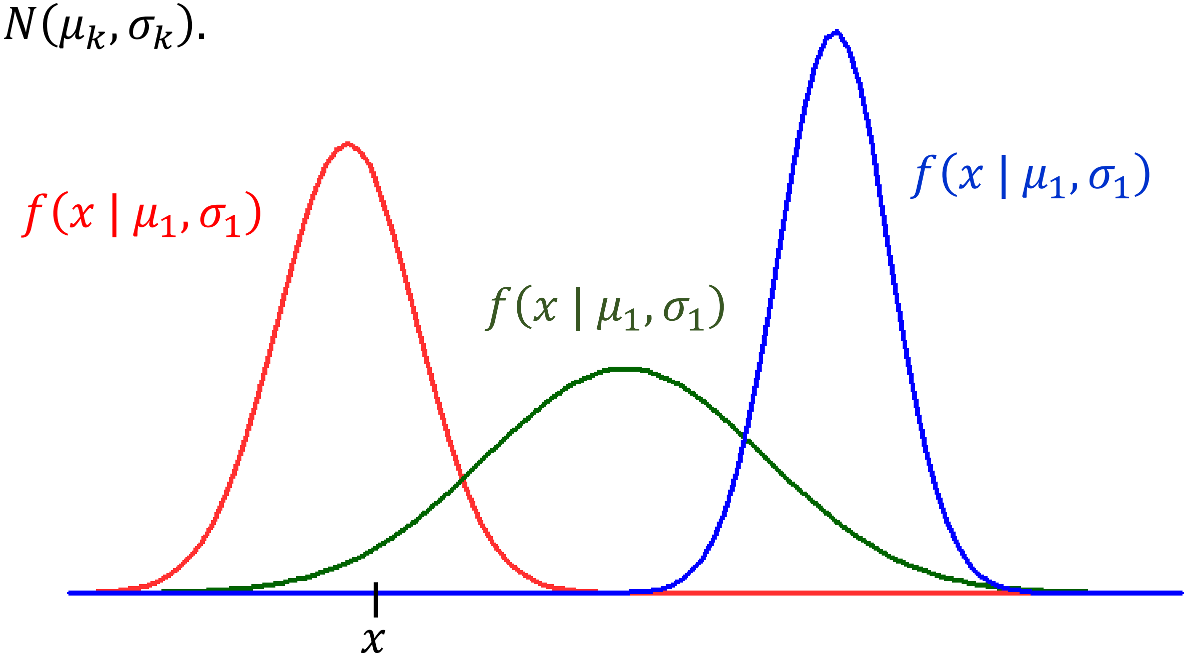
$$\theta_{t+1} = \underset{\theta}{\operatorname{argmax}} \, Q_1(\theta, \theta_t)$$

Iterate through steps 2 and 3 **until convergence**, i.e. until $\theta_{t+1} - \theta_t$ is small enough. The calculated θ_t is often a global maximum, but can also be a local maximum or a saddle point.

Gaussian mixture

Assume we have K normal distributions (Gaussians) with densities $f_k(x \mid \mu_k, \sigma_k)$, $k = (1, 2, \dots, K)$. The parameters μ_k and σ_k are the mean and the standard deviation of the k^th Gaussian. We carry out a two-step experiment:

1. Choose a Gaussian f_k randomly with some probability α_k. This can be described by a multinomially distributed random variable $Z \sim Mult(\alpha_{1,} \alpha_2, \dots, \alpha_K)$ with sample space $\Omega_Z = \{1, 2, \dots, K\}$ and probability mass function $P(Z = k) = \alpha_k$. We have $\sum \alpha_k = 1$ and $\alpha_k > 0$ for all k.
2. Generate a sample x from the above chosen distribution f_k. Thus, x is an observation of a normally distributed random variable X with parameters μ_k and σ_k, i.e. $X \sim N(\mu_k, \sigma_k)$.



Maximum Likelihood for a Gaussian mixture

The experiment includes a discrete (Z) and a continuous (X) random variable. The (mixed) **joint density of X and Z** can be written:

$$f_{X,Z}(x, Z = k) = P(Z = k) \cdot f_{X|Z}(x|Z = k) \qquad f_{X|Z}\text{: conditional probability}$$

$$= \alpha_k \cdot f_k(x \mid \mu_k, \sigma_k)$$

where f_k is a Gaussian: $\quad f_k(x \mid \mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi} \cdot \sigma_k} \exp\left[-\frac{(x - \mu_k)^2}{2\sigma_k^2}\right]$

In practice, we often only have the observation x, without knowing from which Gaussian x was emitted, i.e. the variable Z is not observed (hidden, latent). The Maximum-Likelihood (ML) method must maximize with respect to the real observations, i.e. we have to find parameters $\boldsymbol{\theta}$ that maximize $f_X(x \mid \boldsymbol{\theta})$, not $f_{X,Z}(x, Z \mid \boldsymbol{\theta})$. Here, the vector $\boldsymbol{\theta}$ represents all parameters of the model: $\boldsymbol{\theta} = \{\alpha_k, \mu_k, \sigma_k\}$. An expression for the density $f_X(x \mid \boldsymbol{\theta})$ that incorporates the latent variables Z can be obtained by applying the **law of total probability:**

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

Maximum Likelihood for a Gaussian mixture

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

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

If we have multiple independent observations $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$, the likelihood is the product of the density for the individual observations:

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

Note that the likelihood is considered as a function of the vector $\boldsymbol{\theta}$. The task of ML is to calculate the $\boldsymbol{\theta}$ that maximizes $L(\boldsymbol{\theta})$:

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

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 $L(\boldsymbol{\theta})$:

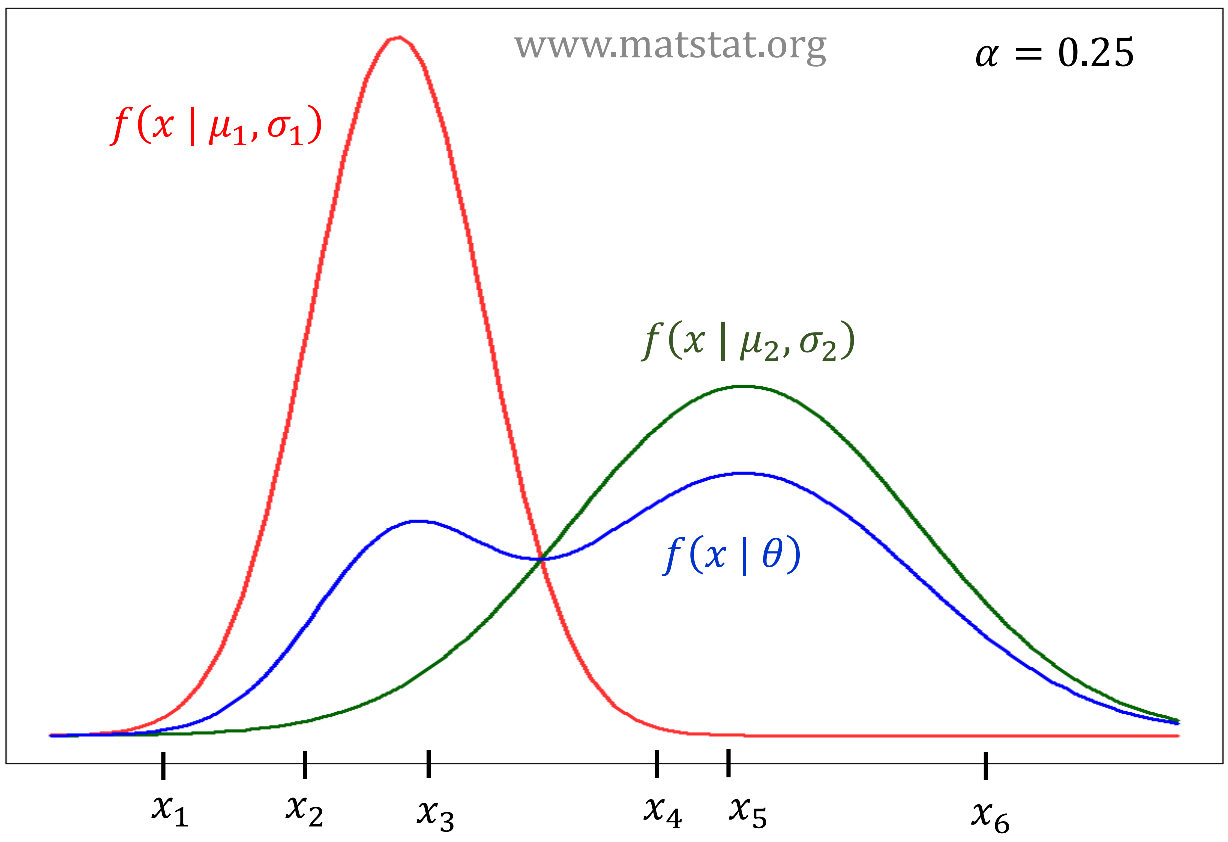
$$l\left(\boldsymbol{\theta}\right) = \log L(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log \left[\sum_{k=1}^{K} \alpha_k \cdot f(x_i \mid \mu_k, \sigma_k) \right]$$

Mixture of Two Gaussians

Let's first look at a mixture of two Gaussians:

$$f(x|\boldsymbol{\theta}) = \alpha \cdot f(x|\mu_1, \sigma_1) + (1-\alpha) \cdot f(x|\mu_2, \sigma_2)$$

We have 5 parameters to estimate: $\boldsymbol{\theta} = (\alpha, \mu_1, \sigma_1, \mu_2, \sigma_2)$ because $\alpha_1 + \alpha_2 = 1$



Mixture of Two Gaussians

Having a mixture of two Gaussians, the likelihood and log-likelihood in the presence of N independent observations for ***X*** read:

$$L\left(\alpha, \mu_1, \sigma_1, \mu_2, \sigma_2\right) = \prod_{i=1}^{N} \left\{\alpha \cdot f\left(x_i | \mu_1, \sigma_1\right) + (1-\alpha) \cdot f\left(x_i | \mu_2, \sigma_2\right)\right\}$$

$$l\left(\alpha, \mu_1, \sigma_1, \mu_2, \sigma_2\right) = \sum_{i=1}^{N} \log\left\{\alpha \cdot f\left(x_i | \mu_1, \sigma_1\right) + (1-\alpha) \cdot f\left(x_i | \mu_2, \sigma_2\right)\right\}$$

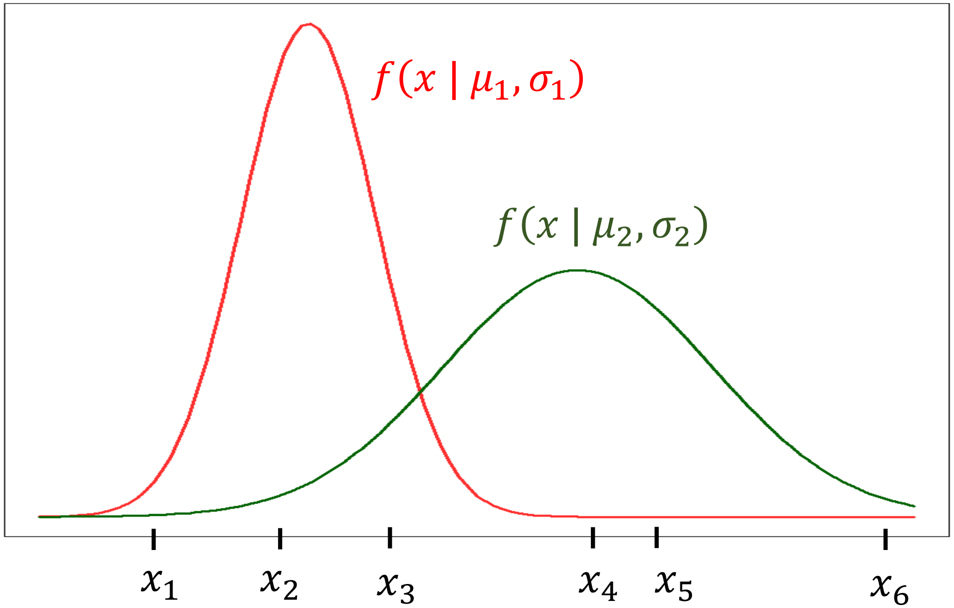
In order to maximize the log-likelihood, we have to solve the equations

$$\frac{\partial l}{\partial \mu_1} = 0 \; ; \quad \frac{\partial l}{\partial \mu_2} = 0 \; ; \quad \frac{\partial l}{\partial \sigma_1} = 0 \; ; \quad \frac{\partial l}{\partial \sigma_2} = 0 \; ; \quad \frac{\partial l}{\partial \alpha} = 0$$

Solving these equations causes problems because the log of a sum is inconvenient to handle numerically.
(With just one component, there was no problem → see appendix)

Introduction of latent variables

- If it was known for each observation x_i from which Gaussian it was emitted, we could solve the problem easily by just estimating the μ_k and σ_k for each component separately (as shown in the appendix).
- Therefore, if the parent components of the observations are unobserved, it seems convenient to artificially introduce a latent variable Z_i for each x_i, so that Z_i assigns x_i to one of the components. Hence, the sample space of each Z_i is $\Omega_{Z_i} = \{1, 2, \dots, K\}$.
- The introduction of the Z_i enables the EM scheme for parameter estimation →



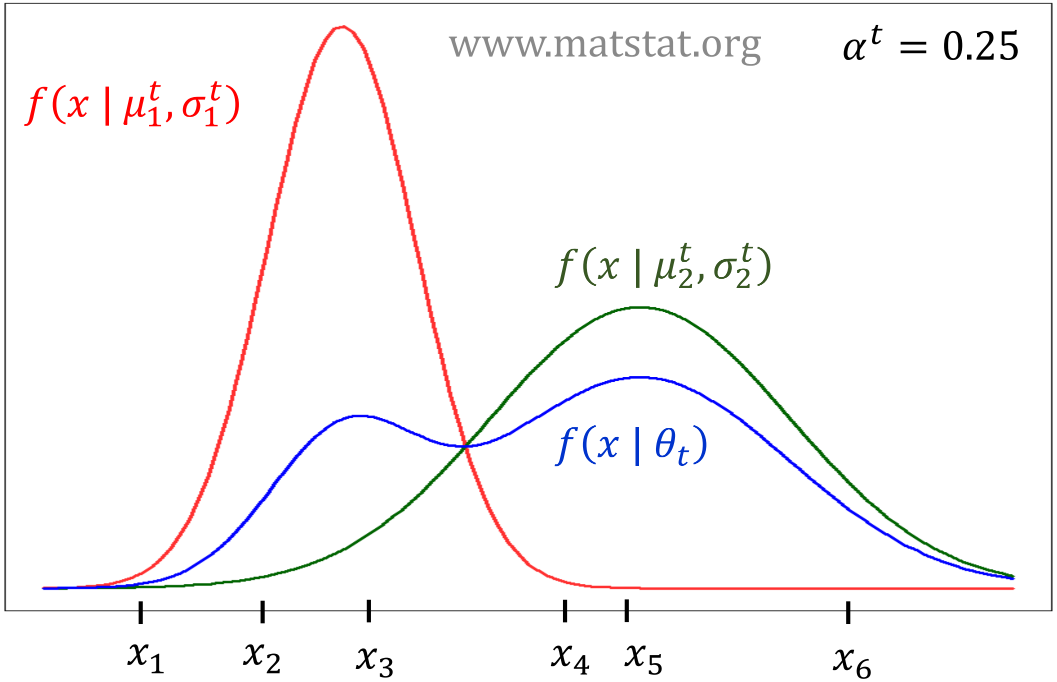
For example, if we knew that x_1, x_2, x_3 belonged to $f(x \mid \mu_1, \sigma_1)$, we could estimate μ_1 and σ_1 on the basis of these 3 points, which is easy to achieve.

Mixture of Gaussians: E-step

1. **Initialize**: $\theta_t = 1^{st}$ guess for the set of parameters

2. **E-step**: calculate $P(Z_i = k \mid X_i = x_i, \theta_t)$.

This is the probability that Z_i is equal to k, i.e. the probability that x_i originates from the kth Gaussian, given the observation x_i and all parameters $\boldsymbol{\theta_t} = \{\alpha_k^t, \mu_k^t, \sigma_k^t\}$. Since the parameters can be considered given, we know the exact positions and shapes of the Gaussians and it's superposition, as shown in the figure.



Mixture of Gaussians: E-step

2. **E-step**: calculate $P(Z_i = k \mid X = x_i, \theta_t)$.

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

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

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

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 $\boldsymbol{\theta_t} = \{\alpha_k^t, \mu_k^t, \sigma_k^t\}$.

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

Degree of membership: ω_{ik}

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

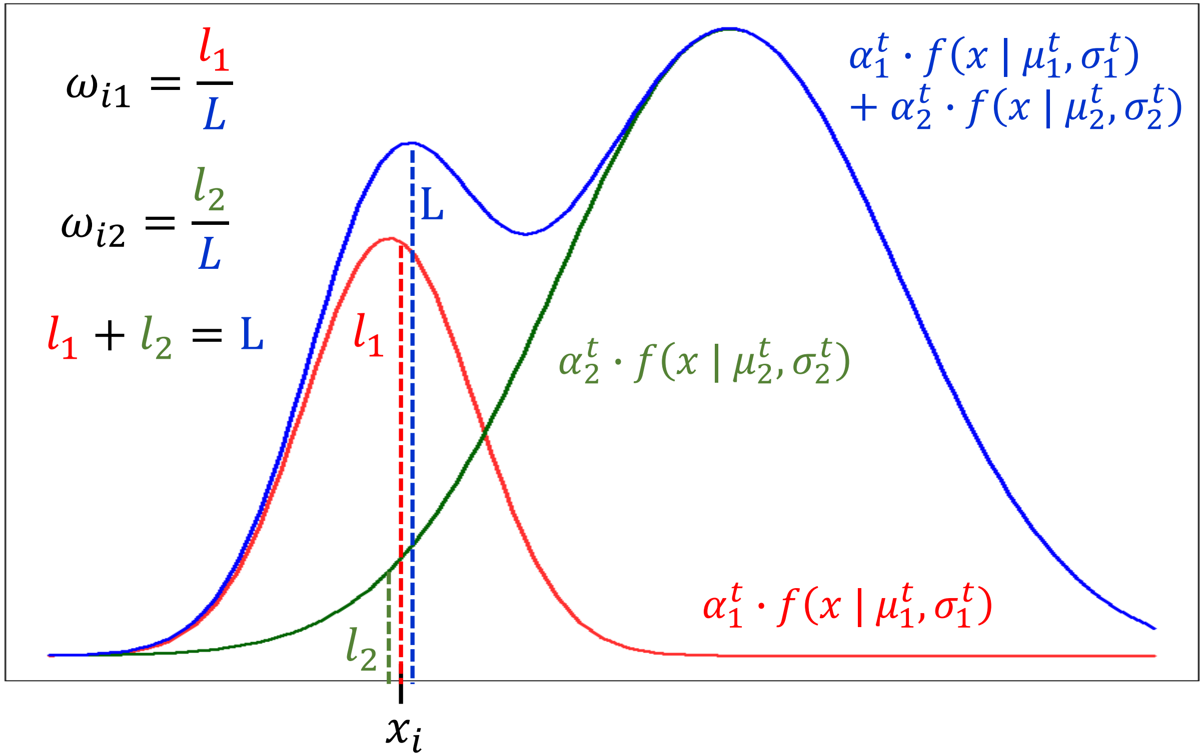


Illustration of the ω_{ik} (dashed lines jittered around x_i for better visibility)

Uwe Menzel, 2018

Completion of the E-step

It remains to calculate:

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

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

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

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 $\alpha_k, \mu_k, \sigma_k \rightarrow$ **M-step**.

Mixture of Gaussians: M-step

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

Q_1 has to be maximized for the unknown $\alpha_k, \mu_k, \sigma_k$

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

$$f_k\left(x_i\mid\mu_k,\sigma_k\right)=\frac{1}{\sqrt{2\pi}\cdot\sigma_k}\exp\left[-\frac{\left(x_i-\mu_k\right)^2}{2\sigma_k^2}\right]$$

$$\log f_k(x_i|\mu_k,\sigma_k)=-\log\sqrt{2\pi}-\log\sigma_k-\frac{\left(x_i-\mu_k\right)^2}{2\sigma_k^2}$$

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

Mixture of Gaussians: M-step

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

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

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

$$\frac{\partial Q_1}{\partial \mu_m} = \sum_{i=1}^{N} \omega_{im} \cdot \frac{(x_i - \mu_m)}{\sigma_m^2} = 0$$

$$\sum_{i=1}^{N} \omega_{im} \cdot (x_i - \mu_m) = 0 \quad \Rightarrow \quad \boxed{\mu_m = \frac{\sum_i^N \omega_{im} \cdot x_i}{\sum_i^N \omega_{im}}} \quad \text{update of } \mu_m$$

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: $\mu = \frac{1}{N} \cdot \sum x_i$

Mixture of Gaussians: M-step

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

Maximization of Q_1 with respect to σ_m:

$$\frac{\partial Q_1}{\partial \sigma_m} = -\frac{\partial}{\partial \sigma_m} \sum_{i=1}^{N} \sum_{k=1}^{K} \omega_{ik} \cdot \left\{ \log \sigma_k + \frac{(x_i - \mu_k)^2}{2\sigma_k^2} \right\}$$

$$\frac{\partial Q_1}{\partial \sigma_m} = -\sum_{i=1}^{N} \omega_{im} \cdot \left\{ \frac{1}{\sigma_m} - \frac{(x_i - \mu_m)^2}{\sigma_m^3} \right\} = 0$$

$$\sum_{i=1}^{N} \omega_{im} \cdot \frac{1}{\sigma_m} = \sum_{i=1}^{N} \omega_{im} \cdot \frac{(x_i - \mu_m)^2}{\sigma_m^3} \Rightarrow \boxed{\sigma_m^2 = \frac{\sum_i \omega_{im} \cdot (x_i - \mu_m)^2}{\sum_i \omega_{im}}}$$

The new estimates for the variances σ_m^2 are weighted means of the squared distances between datapoints and mean μ_m (weighted with the degree of membership of each datapoint). This can be compared with the ML estimation of the variance for a single Gaussian: $\sigma^2 = \frac{1}{N}\sum(x_i - \mu)^2$ (if not corrected for bias).

Mixture of Gaussians: M-step

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

Maximization of Q_1 with respect to α_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 $\sum_k \alpha_k = 1$

$$\Lambda = \sum_{i=1}^{N} \sum_{k=1}^{K} \omega_{ik} \cdot \log \alpha_k - \lambda \cdot \left(\sum_k \alpha_k - 1 \right)$$

λ : Lagrange multiplier

(only the part of Q_1 depending on α_k was included, other terms disappear by derivation)

$$\frac{\partial \Lambda}{\partial \alpha_m} = \sum_{i=1}^{N} \omega_{im} \cdot \frac{1}{\alpha_m} - \lambda = 0 \quad \Rightarrow \quad \alpha_m = \frac{1}{\lambda} \cdot \sum_{i=1}^{N} \omega_{im}$$

Mixture of Gaussians: M-step

$$\alpha_m = \frac{1}{\lambda} \cdot \sum_{i=1}^{N} \omega_{im}$$ it remains to find λ

The Lagrange multiplier λ can be determined using:

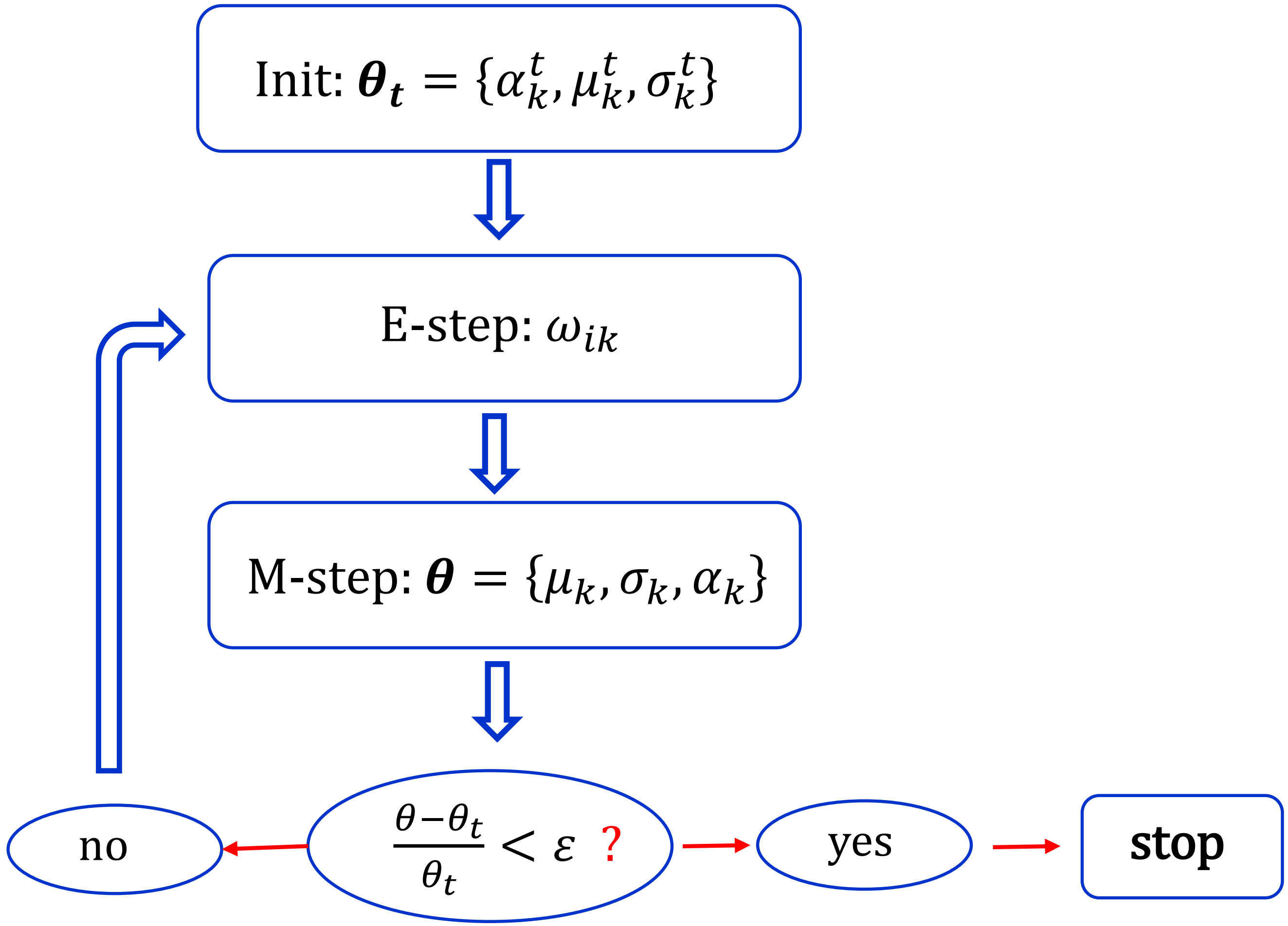
$$1 = \sum_{m=1}^{K} \alpha_m = \sum_{m=1}^{K} \frac{1}{\lambda} \cdot \sum_{i=1}^{N} \omega_{im} = \frac{1}{\lambda} \sum_{i=1}^{N} \underbrace{\sum_{m=1}^{K} \omega_{im}}_{=1} = \frac{1}{\lambda} \sum_{i=1}^{N} 1 = \frac{N}{\lambda}$$

$$\Rightarrow \quad \lambda = \frac{1}{N} \quad \Rightarrow \quad \boxed{\alpha_m = \frac{1}{N} \cdot \sum_{i=1}^{N} \omega_{im}}$$

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

Iteration

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



Summary of EM for 1-D Gaussian mixture

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

E-step:

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

M-step:

$$\mu_m = \frac{\sum_i^N \omega_{im} \cdot x_i}{\sum_i^N \omega_{im}}$$

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

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

$$\boxed{\sum_{k=1}^{K} \omega_{ik} = 1 \qquad \sum_{i=1}^{N} \sum_{m=1}^{K} \omega_{im} = N}$$

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

Alternatively, check convergence of the likelihood.

Appendix

The Expectation-Maximization algorithm II

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

ML for a single Gaussian

Maximum-likelihood estimate for μ, σ for a single Gaussian:

$$L(\mu, \sigma) = \prod_{i=1}^{N} f(x_i|\mu, \sigma)$$

Likelihood for N independent samples

$$l(\mu, \sigma) = \sum_{i=1}^{N} \log f(x_i|\mu, \sigma)$$

Log-likelihood

$$f(x_i|\mu, \sigma) = \frac{1}{\sqrt{2\pi} \cdot \sigma} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

$$\log f(x_i|\mu, \sigma) = -\log\sqrt{2\pi} - \log\sigma - \frac{(x_i - \mu)^2}{2\sigma^2}$$

$$l(\mu, \sigma) = \sum_{i=1}^{N} \left[-\log\sqrt{2\pi} - \log\sigma - \frac{(x_i - \mu)^2}{2\sigma^2}\right]$$

ML for a single Gaussian

$$l(\mu,\sigma) = \sum_{i=1}^{N}\left[-\log\sqrt{2\pi} - \log\sigma - \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$\frac{\partial l}{\partial \mu} = -\frac{\partial}{\partial \mu}\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2} = \sum_{i=1}^{N}\frac{2(x_i-\mu)}{2\sigma^2} = \sum_{i=1}^{N}\frac{(x_i-\mu)}{\sigma^2} = 0$$

$$\sum_{i=1}^{N}(x_i-\mu) = 0 \quad \Rightarrow \quad \sum_{i=1}^{N}x_i - N\cdot\mu = 0 \quad \Rightarrow \quad \hat{\mu} = \frac{1}{N}\cdot\sum_{i=1}^{N}x_i$$

$$\frac{\partial l}{\partial \sigma} = \frac{\partial}{\partial \sigma}(-N\cdot\log\sigma) - \frac{\partial}{\partial \sigma}\sum_{i=1}^{N}\frac{(x_i-\mu)^2}{2\sigma^2} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i-\mu)^2 = 0$$

$$\frac{N}{\sigma} = \frac{1}{\sigma^3}\sum_{i=1}^{N}(x_i-\mu)^2 \quad \Rightarrow \quad \hat{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$$

(We know that we have to replace N by $N-1$ in the last expression in order to get an unbiased estimation of the variance)