Statistical Computing

The Expectation-Maximization Algorithm I
The General scheme of EM

Uwe Menzel, 2018

uwe.menzel@matstat.de

www.matstat.org

Expectation–Maximization

- The **E**xpectation–**M**aximization (**EM**) algorithm is an iterative scheme for calculation of maximum likelihood (ML) or maximum a posteriori (MAP) estimates of parameters in statistical models.
- The EM scheme includes so-called latent (hidden) variables, which might be missing values or data that cannot be observed on principle. Very often, the latent variables are just artificially incorporated into the model in order to facilitate an EM scheme.
- The EM iteratetively alternates between an **E**xpectation step and a **M**aximization step.
- The EM algorithm is widely applied in different fields of statistical modeling.
- Here, the general scheme of EM is introduced. In following lectures, specific algorithms utilizing this general scheme will be presented.

Maximum Likelihood

Here, we develop the EM scheme for the case that the observations derive from a continuous random variable $\boldsymbol{X}$, and the latent variables $\boldsymbol{Z}$ are discrete with an integer sample space. This setup occurs frequently, e. g. when dealing with Gaussian mixtures. The N independent observations from the variable X can be denoted by $\boldsymbol{x} = \{x_1, x_2, \ldots, x_N\}$. The statistical model is established in form of a probability density function (PDF) $f_X(\boldsymbol{x} \mid \boldsymbol{\theta})$ for $\boldsymbol{X}$ which is parameterized on a set of parameters condensed in the vector $\boldsymbol{\theta}$. The likelihood is then

$$L(\theta) = p(X|\theta) = \prod_i^N f_X(x_i|\theta) \qquad (1)$$

To obtain a Maximum Likelihood (ML) estimation for θ, we are looking for the θ that makes the observed data most likely, i. e. we attempt to calculate

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}}\, L(\theta) \qquad (2)$$

Instead of maximizing $L(\theta)$, we can also maximize $l(\theta) = \log L(\theta)$, since both functions have their maxima at the same place:

$$l(\theta) = \sum_i^N \log f_X(x_i|\theta) \qquad (3)$$

Introduction of latent variables

We can introduce into the model latent variables **Z** by taking advantage of the marginal rule:

$$f_X(x_i|\theta) = \sum_{k=1}^{K} f_{X,Z}\left(x_i, Z = k|\theta\right) \qquad (4)$$

where $f_{X,Z}$ denotes the joint distribution of X and Z. As already noted above, we have assumed that Z is a discrete random variable with a sample space that includes the integers from 1 to K, i.e. $Z \in \{1, 2, \ldots, K\}$. If Z is continuous, the sum has to be replaced by an integral.

It seems as if we have made the problem even more complicated by introducing unknown latent variables Z, especially in view of the fact that we have to maxmize $p(X \mid \theta)$, not $p(X, Z \mid \theta)$, i.e. we maximize only with respect to the observed data. However, later we'll see by means of examples that it makes sense to include the latent variables into the model if they are chosen in a clever way.

By using the above expression (eqn. 4), the equation for the log-likelihood (eqn. 3) transformes to:

$$l(\theta) = \sum_{i=1}^{N} \log \left\{ \sum_{k=1}^{K} f_{X,Z}\left(x_i, Z = k|\theta\right) \right\} \qquad (5)$$

Conditional probability for the latent variables

EM attempts to find a ML estimation for θ in an iterative manner. Let θ_t be a first guess of the parameters (or the estimates obtained in the precedent iteration step) so that θ_t can be regarded as known. We can rewrite eqn. (5) by expanding each term within the k-sum with the conditional probability of the latent variables Z, given the observations x_i and the θ_t, i.e. we can expand with $P(Z = k \mid X = x_i, \theta_t)$ to obtain:

$$l(\theta) = \sum_{i=1}^{N} \log \left\{ \sum_{k=1}^{K} P(Z = k|X = x_i, \theta_t) \cdot \frac{f_{X,Z}(x_i, Z = k|\theta)}{P(Z = k|X = x_i, \theta_t)} \right\} \quad (6)$$

Remark regarding the notation used:

Because X is a continuous variable, we write $f_{X,Z}(x_i, Z = k)$ to make clear that this is a probability density.
On the other hand, we use the notation $P(Z = k \mid X = x_i)$ because Z is discrete. Depending on the nature of the observed and latent variables, the notation should be adapated.

Lower bound for the log-likelihood

Note that the $P(Z = k \mid X = x_i, \theta_t)$ sum up to 1 when summed over k, and are all non-negative, so that we can apply **Jensen's inequality** to the log of the k-sum. The log is a concave function, so that Jensen's inequality reads:

$$\log\left\{\sum_k a_k B_k\right\} \geq \sum_k a_k \log B_k \quad \text{if} \quad \sum_k a_k = 1 \quad a_k \geq 0$$

(see a little bit more about Jensen's inequality in the appendix)

This can be used to find a lower limit of $l(\theta)$:

$$l(\theta) = \sum_{i=1}^{N} \log\left\{\sum_{k=1}^{K} \underbrace{P(Z = k|X = x_i, \theta_t)}_{a_k} \cdot \underbrace{\frac{f_{X,Z}(x_i, Z = k|\theta)}{P(Z = k|X = x_i, \theta_t)}}_{B_k}\right\} \quad (7)$$

$$\geq \sum_{i=1}^{N} \left\{\sum_{k=1}^{K} \underbrace{P(Z = k|X = x_i, \theta_t)}_{a_k} \cdot \log \underbrace{\frac{f_{X,Z}(x_i, Z = k|\theta)}{P(Z = k|X = x_i, \theta_t)}}_{B_k}\right\} \quad (8)$$

Lower bound for the log-likelihood

By using Jensen's inequality, we have identified a lower bound for $l(\theta)$, which is often called $Q(\theta, \theta_t)$:

$$l(\theta) \geq Q(\theta, \theta_t) \qquad (9)$$

where

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

The **E-step** involves the calculation of the quantities $P(Z = k \mid X = x_i, \theta_t)$ and $Q(\theta, \theta_t)$. In the **M-step**, the quantity Q is maximized w.r.t. θ, in order to find an update of θ:

$$\theta_{t+1} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta_t) \qquad (10)$$

Eqn. (9) is valid for all θ, i.e. eqn (9) implies:

$$l(\theta_{t+1}) \geq Q(\theta_{t+1}, \theta_t) \qquad (11)$$

$Q(\theta_t, \theta_t) = l(\theta_t)$

It is important to know what $Q(\theta, \theta_t)$ becomes when $\theta = \theta_t$:

$$Q(\theta_t, \theta_t) = \sum_{i=1}^{N} \sum_{k=1}^{K} P(Z = k|X = x_i, \theta_t) \cdot \log \underbrace{\frac{f_{X,Z}(x_i, Z = k|\theta_t)}{P(Z = k|X = x_i, \theta_t)}}_{\text{see below}}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{K} P(Z = k|X = x_i, \theta_t) \cdot \underbrace{\log f_X(x_i|\theta_t)}_{\text{independent of k}}$$

$$= \sum_{i=1}^{N} \log f_X(x_i|\theta_t) \underbrace{\sum_{k=1}^{K} P(Z = k|X = x_i, \theta_t)}_{= 1}$$

$$= l(\theta_t) \quad \text{compare eqn. (3)} \qquad (12)$$

Above, we have used the general relation (Z discrete, X continuous):

$$f_{X,Z}(x, z) = P(Z = z) \cdot f_{X|Z}(x|z) = f_X(x) \cdot P(Z = z|X = x)$$

leading to $$\frac{f_{X,Z}(x, z)}{P(Z = z|X = x)} = f_X(x)$$

Simplifying the M-step

The M-step can be simplified by splitting the log of the ratio:

$$
\begin{aligned}
Q(\theta, \theta_t) &= \sum_{i=1}^{N} \sum_{k=1}^{K} P(Z = k | X = x_i, \theta_t) \cdot \log \frac{f_{X,Z}(x_i, Z = k|\theta)}{P(Z = k | X = x_i, \theta_t)} \\
&= \sum_{i=1}^{N} \sum_{k=1}^{K} P(Z = k | X = x_i, \theta_t) \cdot \log f_{X,Z}(x_i, Z = k|\theta) \\
&- \sum_{i=1}^{N} \sum_{k=1}^{K} P(Z = k | X = x_i, \theta_t) \cdot \log P(Z = k | X = x_i, \theta_t)
\end{aligned}
$$

The 2nd term in the difference does not depend on θ. Therefore, it is sufficient to maximize

$$
Q_1(\theta, \theta_t) = \sum_{i=1}^{N} \sum_{k=1}^{K} P(Z = k | X = x_i, \theta_t) \cdot \log f_{X,Z}(x_i, Z = k|\theta) \qquad (13)
$$

This can be interpreted as the **conditional expectation** $E_{Z|X=x_i}$ of the log-likelihood of the complete data (X, Z).

Each iteration improves the log-likelihood

We have now obtained several important equations:

$$l(\theta) \geq Q(\theta, \theta_t) \qquad (9)$$ which implies

$$l(\theta_{t+1}) \geq Q(\theta_{t+1}, \theta_t) \qquad (11)$$

$$Q(\theta_t, \theta_t) = l(\theta_t) \qquad (12)$$

Moreover, it is

$$Q(\theta_{t+1}, \theta_t) \geq Q(\theta_t, \theta_t) \qquad (14)$$

because θ_{t+1} maximizes the quantity $Q(\theta, \theta_t)$:

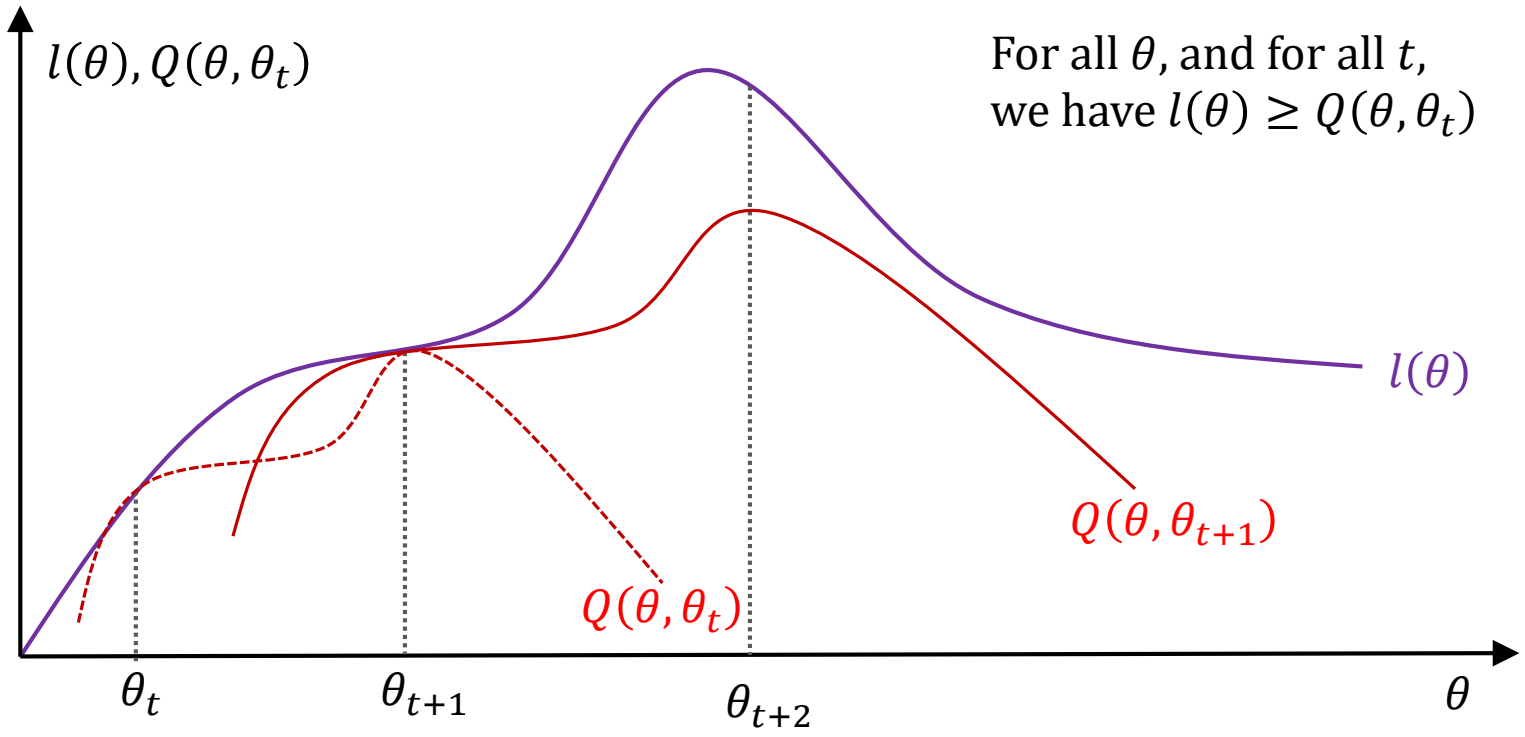
$$\theta_{t+1} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta_t) \quad \text{(M-step)}$$

Combining these equations we can conclude that

$$l(\theta_{t+1}) \geq Q(\theta_{t+1}, \theta_t) \geq Q(\theta_t, \theta_t) = l(\theta_t)$$

i.e. the EM scheme improves the ML on each iteration step (at least, the log-likelihood cannot decrease in any iteration). That makes EM appealing for practical applications.

Each iteration improves the log-likelihood



For $\theta = \theta_t$, we have $l(\theta_t) = Q(\theta_t, \theta_t)$. Then, θ_{t+1} is calculated to be the maximum of $Q(\theta, \theta_t)$. The new $Q(\theta, \theta_{t+1})$ is also a lower bound of $l(\theta)$, and the equation $l(\theta_{t+1}) = Q(\theta_{t+1}, \theta_{t+1})$ applies. Next, the maximum of $Q(\theta_t, \theta_{t+1})$ is calculated in order to get θ_{t+2}, etc. EM ensures that the log-likelihood $l(\theta)$ cannot decrease on any iteration. Iterations continue until convergence is reached, i.e. until $\theta_{t+1} - \theta_t$ is small enough.

Summary of the EM meta-algorithm

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

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

$$Q_1(\theta, \theta_t) = \sum_{i=1}^{N} \sum_{k=1}^{K} P(Z = k | X = x_i, \theta_t) \cdot \log f_{X,Z}(x_i, Z = k | \theta) \qquad (13)$$

   - $X = \{x_1, x_2, \ldots, x_N\}$ 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: $\theta_t \rightarrow \theta_{t+1}$

$$\theta_{t+1} = \underset{\theta}{\operatorname{argmax}}\, Q_1(\theta, \theta_t) \qquad (10)$$

Iterate through steps 2 and 3 until convergence, i.e. until $\theta_{t+1} - \theta_t$ is small (or the difference of the likelihoods is small).

Convergence of the EM method

- We have seen that the (log-) likelihood cannot decrease at any iteration step of the EM scheme.
- The scheme guarantees that a stationary point of the likelihood is found.
- This point is not nessecarily is a global maximum of $L(\theta)$, but can also be a local maximum or a saddle point.

What have we won by using EM?

Instead of calculating the θ that maximizes $l(\theta) = \sum_i^N \log P(X = x_i|\theta)$

we have to find the θ which maximizes the expression $Q_1(\theta, \theta_t)$ on each iteration:

$$\theta_{t+1} = \underset{\theta}{\operatorname{argmax}}\, Q_1(\theta, \theta_t) \qquad \text{see eqn. (10)}$$

At a first glance, this doesn't seem to be much of an advantage. However, as we'll see in the examples, solving eqn. (10) can be much easier than maximizing $l(\theta)$ directly, if the latent variables Z are chosen in a clever manner.

Appendix

The Expectation Maximization Algorithm

Uwe Menzel, 2018

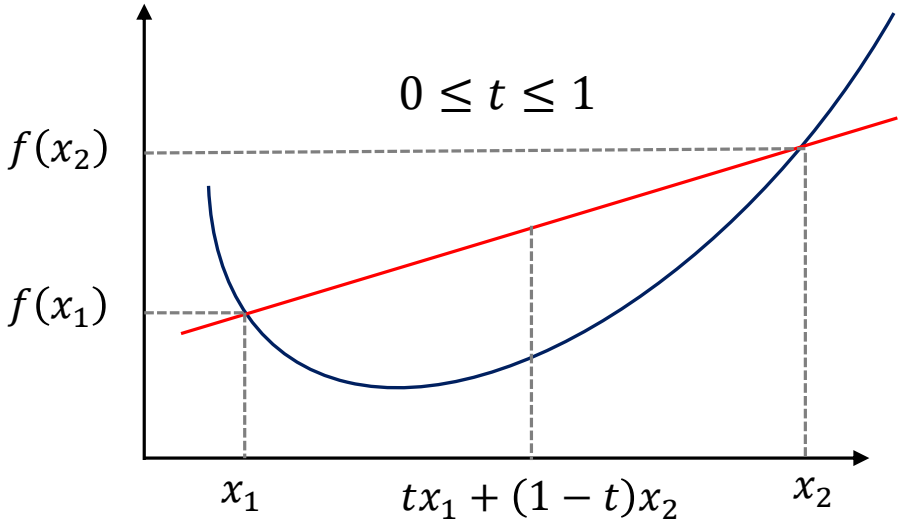
uwe.menzel@matstat.de

www.matstat.org

The Jensen inequality

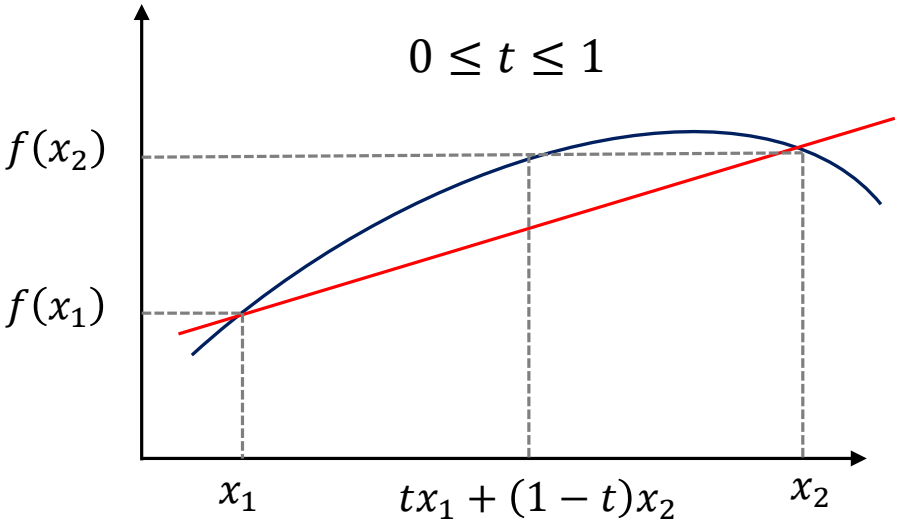
Convex functions:

$$\sum_i a_i f(x_i) \geq f\left(\sum_i a_i x_i\right)$$
$$\sum_i a_i = 1 \quad a_i > 0$$



Concave functions:

$$f\left(\sum_i a_i x_i\right) \geq \sum_i a_i f(x_i)$$
$$\sum_i a_i = 1 \quad a_i > 0$$



The Jensen inequality

Convex functions:

$$\sum_i a_i f(x_i) \geq f\left(\sum_i a_i x_i\right)$$
$$\sum_i a_i = 1 \quad a_i > 0$$

Concave functions:

$$f\left(\sum_i a_i x_i\right) \geq \sum_i a_i f(x_i)$$
$$\sum_i a_i = 1 \quad a_i > 0$$

Convex functions:

$$\int \Phi\left[f(x)\right] p(x)dx \geq \Phi\left[\int f(x)p(x)dx\right]$$
$$\int p(x) = 1 \quad p(x) > 0$$

Concave functions:

$$\Phi\left[\int f(x)p(x)dx\right] \geq \int \Phi\left[f(x)\right] p(x)dx$$
$$\int p(x) = 1 \quad p(x) > 0$$

This is not more than a hint. An excellent presentation of Jensen's (and other) inequalities is by Dragos Hrimiuc (University of Alberta) and is linked here: https://www.math.ualberta.ca/pi/issue4/ ("Pi in the sky" December 2001 issue).