Lecture 8

15. The EM-algorithm

The EM-algorithm (Expectation-Maximization algorithm) is an iterative procedure for computing the maximum likelihood estimator when only a subset of the data is available. The first proper theoretical study of the algorithm was done by Dempster, Laird, and Rubin (1977). The EM algorithm is extensively used throughout the statistics literature. We will give an overview on how it works.

Let $X = (X_1, \dots, X_n)$ be a sample with conditional density $f_{X|\Theta}(x \mid \theta)$ given $\Theta = \theta$. We will write

$$l(\theta; X) = \log f_{X|\Theta}(X \mid \theta)$$

for the log-likelihood function. The EM-algorithm is used when we do not have a complete data set of observations from X. We will assume that the data X consists of observed variables $Y = (Y_1, \dots, Y_k)$ and unobserved (missing or latent variables) $Z = (Z_1, \dots, Z_{n-k})$. We write $X = (Y, Z)$. With this notion the log-likelihood function for the observed data Y is

$$l_{obs}(\theta; Y) = \log \int f_{X|\Theta}(Y, z \mid \theta)\nu_z(dz).$$

The problem here is that when maximizing the likelihood we have to compute the integral and this can be difficult. We might not be able to find a closed form expression for $l_{obs}(\theta; Y)$. To maximize $l_{obs}(\theta; Y)$ with respect to θ the idea is to do an iterative procedure where each iteration has two steps, called the E-step and the M-step. Let $\theta^{(i)}$ denote the estimate of Θ after the ith step. Then the two steps in the $(i+1)$th iteration are

E-step: Compute $Q(\theta \mid \theta^{(i)}) = E_{\theta^{(i)}}[l(\theta; X) \mid Y]$.
M-step: Maximize $Q(\theta \mid \theta^{(i)})$ with respect to θ and put $\theta^{(i+1)} = \operatorname{argmax} Q(\theta \mid \theta^{(i)})$.

This procedure is iterated until it converges.

15.1. **A Multinomial example.** This is one of the original illustrating examples for the use of the EM algorithm. One considers data in which 197 animals are distributed multinomially into four categories with cell-probabilities $(1/2+\theta/4, (1-\theta)/4, (1-\theta)/4, \theta/4)$ for some unknown $\theta \in [0,1]$. The observed number in each cell was $Y = y = (125, 18, 20, 34)$. The density of the observed data is

$$f_{Y|\Theta}(y \mid \theta) = \frac{n!}{y_1!y_2!y_3!y_4!}\left(\frac{1}{2}+\frac{\theta}{4}\right)^{y_1}\left(\frac{1}{4}-\frac{\theta}{4}\right)^{y_2}\left(\frac{1}{4}-\frac{\theta}{4}\right)^{y_3}\left(\frac{\theta}{4}\right)^{y_4}.$$

The log-likelihood function is then

$$l(\theta, y) = c + y_1 \log(2+\theta) + (y_2+y_3)\log(1-\theta) + y_4 \log\theta.$$

Differentiating w.r.t. θ we get that the score is

There are some typos in this paper.

$$\boxed{\partial_\theta l(\theta, y) = \frac{y_1}{1-\theta} - \frac{y_2+y_3}{1-\theta} + \frac{y_4}{\theta}}$$

and the Fisher information is

$$\mathcal{I}(\theta) = -\partial_\theta^2 l(\theta, y) = \frac{y_1}{(2+\theta)^2} + \frac{y_2+y_3}{(1-\theta)^2} + \frac{y_4}{\theta^2}.$$

Although the log-likelihood can be maximized explicitly we use the example to illustrate the EM algorithm. To view the problem as an unobserved data problem we would think of it as a multinomial experiment with five categories with observations $x = (y_{11}, y_{12}, y_2, y_3, y_4, y_5)$, each with cell probability $(1/2, \theta/4, (1-\theta)/4, (1-\theta)/4, \theta/4)$. That is, we split the first category into two, and we can only observe the sum $y_1 = y_{11} + y_{12}$. Then y_{11} and y_{12} are considered as the unobservable variables. The complete likelihood for the data is then

$$f_{X|\Theta}(y_{11}, y_{12}, y_2, y_3, y_4 \mid \theta) = \frac{n!}{y_{11}!y_{12}!y_2!y_3!y_4!}\Big(\frac{1}{2}\Big)^{y_{11}}\Big(\frac{\theta}{4}\Big)^{y_{12}}\Big(\frac{1-\theta}{4}\Big)^{y_2}\Big(\frac{1-\theta}{4}\Big)^{y_3}\Big(\frac{\theta}{4}\Big)^{y_4}$$

and the log-likelihood is (apart from a term not involving θ)

$$l(\theta, y_{11}, y_{12}, y_2, y_3, y_4) = (y_{12} + y_4)\log\theta + (y_2 + y_3)\log(1-\theta).$$

Since y_{12} is unobservable we cannot maximize this directly. This obstacle is overcome by the E-step.

Let $\theta^{(0)}$ be an initial guess for θ. The E-step requires computation of

$$\begin{aligned} Q(\theta \mid \theta^{(0)}) &= E_{\theta^{(0)}}[l(\theta, Y_{11}, Y_{12}, Y_2, Y_3, Y_4) \mid Y_1, \dots, Y_4] \\ &= E_{\theta^{(0)}}[(Y_{12} + Y_4)\log\theta + (Y_2 + Y_3)\log(1-\theta) \mid Y_1, \dots, Y_4] \\ &= (E_{\theta^{(0)}}[Y_{12} \mid Y_1] + Y_4)\log\theta + (Y_2 + Y_3)\log(1-\theta). \end{aligned}$$

Thus, we need to compute the conditional expectation of Y_{12} given Y_{11} given $\Theta = \theta^{(0)}$. But this is a Binomial distribution with sample size Y_1 and parameter $p = (\theta^{(0)}/4)/(1/2 + \theta^{(0)}/4)$. Hence, the expected value is

$$E_{\theta^{(0)}}[Y_{12} \mid Y_1] = \frac{Y_1\theta^{(0)}}{2 + \theta^{(0)}} := y_{12}^{(0)}$$

and the expression for $Q(\theta \mid \theta^{(0)})$ is

$$Q(\theta \mid \theta^{(0)}) = (y_{12}^{(0)} + y_4)\log\theta + (y_2 + y_3)\log(1-\theta).$$

In the M-step we maximize this with respec to θ to get

$$\theta^{(1)} = \frac{y_{12}^{(0)} + y_4}{y_{12}^{(0)} + y_2 + y_3 + y_4}.$$

Then, iterating this gives us finally the estimate for θ. Summarizing, we get the iterations

$$\theta^{(i+1)} = \frac{y_{12}^{(i)} + y_4}{y_{12}^{(i)} + y_2 + y_3 + y_4}$$

where

$$y_{12}^{(i)} = \frac{y_1\theta^{(i)}}{2 + \theta^{(i)}}.$$

15.2. **Exponential families.** The EM-algorithm is particularly easy for regular exponential families.

If $f_{X|\Theta}(x \mid \theta)$ form a regular exponential family with natural parameter Θ then

$$f_{X|\Theta}(x \mid \theta) = c(\theta)h(x)\exp\Big\{\sum_{j=1}^{k}\theta_j t_j(x)\Big\}$$

Recall that

$$E_\theta[t_j(X)] = -\partial_{\theta_j} \log c(\theta),$$
$$\mathrm{cov}_\theta[t_j(X)t_k(X)] = (\mathcal{I}_X(\theta))_{jk} = -\partial_{\theta_j}\partial_{\theta_k} \log c(\theta).$$

The log-likelihood of X is then (ignoring terms that do not depend on θ)

$$l(\theta, X) = \log c(\theta) + \sum_{i=1}^{k} \theta_j t_j(X).$$

Then, observing Y and with Z as unobserved such that $X = (Y, Z)$ we get

$$Q(\theta \mid \theta^{(i)}) = \log c(\theta) + \sum_{j=1}^{k} \theta_j t_j^{(i)}$$

with

$$t_j^{(i)}(Y) = E_{\theta^{(i)}}[t_j(Y, Z) \mid Y].$$

Maximizing $Q(\theta \mid \theta^{(i)})$ with respect to θ leads, by differentiation, to taking $\theta^{(i+1)}$ by solving

$$0 = \partial_{\theta_j} Q(\theta \mid \theta^{(i)}) = \partial_{\theta_j} \log c(\theta) + t_j^{(i)}(Y) \quad j = 1, \dots, k.$$

That is, find $\theta^{(i+1)}$ such that

$$E_{\theta^{(i+1)}}[t_j(X)] = t_j^{(i)} \quad j = 1, \dots, k.$$

15.3. **Why it works.** To give a heuristic explanation why the algorithm works one can argue as follows. First note that what we want originally is to maximize the log-likelihood $l_{obs}(\theta, Y)$ of the observations. For this we want to find a $\hat{\theta}$ which solves $\partial_{\theta_j} l_{obs}(\theta, Y)|_{\theta=\hat{\theta}} = 0$ $j = 1, \dots, k$. We will show that if $\theta^{(i)}$ coming from the algorithm converges to some θ^* then this θ^* satisfies $\partial_{\theta_j} l_{obs}(\theta, Y)|_{\theta=\theta^*} = 0$ $j = 1, \dots, k$ which is exactly what we want.

To see this note that the complete likelihood is, $x = (y, z)$,

$$l(\theta, x) = \log f_{X|\Theta}(x \mid \theta) = \log f_{Z|Y,\Theta}(z \mid y, \theta) + \log f_{Y|\Theta}(y \mid \theta).$$

Then

$$Q(\theta \mid \theta^{(i)}) = \int \log f_{Z|Y,\Theta}(z \mid y, \theta) f_{Z|Y,\Theta}(z \mid y, \theta^{(i)}) \nu_z(dz) + l_{obs}(\theta, Y).$$

Differentiating w.r.t. θ_j and putting equal to 0 in order to maximize Q gives

$$0 = \partial_{\theta_j} Q(\theta \mid \theta^{(i)}) = \int \frac{\partial_{\theta_j} f_{Z|Y,\Theta}(z \mid y, \theta)}{f_{Z|Y,\Theta}(z \mid y, \theta)} f_{Z|Y,\Theta}(z \mid y, \theta^{(i)}) \nu_z(dz) + \partial_{\theta_j} l_{obs}(\theta, Y).$$

If $\theta^{(i)} \to \theta^*$ then we have for θ^* that (with $j = 1, \dots, k$)

$$\begin{aligned}
0 &= \partial_{\theta_j} Q(\theta^* \mid \theta^*) \\
&= \int \frac{\partial_{\theta_j} f_{Z|Y,\Theta}(z \mid y, \theta^*)}{f_{Z|Y,\Theta}(z \mid y, \theta^*)} f_{Z|Y,\Theta}(z \mid y, \theta^*) \nu_z(dz) + \partial_{\theta_j} l_{obs}(\theta^*, Y) \\
&= \partial_{\theta_j} \int f_{Z|Y,\Theta}(z \mid y, \theta^*) \nu_z(dz) + \partial_{\theta_j} l_{obs}(\theta^*, Y) \\
&= \partial_{\theta_j} l_{obs}(\theta^*, Y).
\end{aligned}$$

Hence, θ^* satisfies the desired equation. To actually prove that the EM-algorithm works we need to show that the sequence $\theta^{(i)}$ actually converges to some θ^*. This is usually done it two steps.

In the first step we show that the sequence $l_{obs}(\theta^{(i)}, Y)$ is non-decreasing. If it is also bounded, then it converges. It should be noted that it may converge to a local maximum, and not necessarily to a global maximum. In practice one would typically choose different initial values $\theta^{(0)}$ to see if there seems to be several local maximima. In the second step we show that $l_{obs}(\theta^{(i)}, Y) \to l^*$ implies $\theta^{(i)} \to \theta^*$. This second step requires stronger assumoptions than the first step.

Proposition 3. *The sequence $l_{obs}(\theta^{(i)}, Y)$ in the EM-algorithm is nondecreasing.*

Proof. We write $X = (Y, Z)$ for the complete data, with Y the observed data and Y the unobserved data. Then

$$f_{Z|Y,\Theta}(Z \mid Y, \theta) = \frac{f_{X|\Theta}((Y, Z) \mid \theta)}{f_{Y|\Theta}(Y \mid \theta)}.$$

Hence,

$$l_{obs}(\theta, Y) = \log f_{Y|\Theta}(Y \mid \theta) = \log f_{X|\Theta}((Y, Z) \mid \theta) - \log f_{Z|Y,\Theta}(Z \mid Y, \theta).$$

Taking conditional expectation given Y and $\Theta = \theta^{(i)}$ on both sides yields

$$\begin{aligned} l_{obs}(\theta, Y) &= E_{\theta^{(i)}}[l_{obs}(\theta, Y) \mid Y] \\ &= E_{\theta^{(i)}}[\log f_{X|\Theta}((Y, Z) \mid \theta) \mid Y] - E_{\theta^{(i)}}[\log f_{Z|Y,\Theta}(Z \mid Y, \theta) \mid Y] \\ &= Q(\theta \mid \theta^{(i)}) - H(\theta \mid \theta^{(i)}), \end{aligned}$$

with

$$H(\theta \mid \theta^{(i)}) = E_{\theta^{(i)}}[\log f_{Z|Y,\Theta}(Z \mid Y, \theta) \mid Y].$$

Then we have

$$l_{obs}(\theta^{(i+1)}, Y) - l_{obs}(\theta^{(i)}, Y) = Q(\theta^{(i+1)} \mid \theta^{(i)}) - Q(\theta^{(i)} \mid \theta^{(i)}) - [H(\theta^{(i+1)} \mid \theta^{(i)}) - H(\theta^{(i)} \mid \theta^{(i)})].$$

Since $\theta^{(i+1)}$ maximizes $Q(\theta \mid \theta^{(i)})$ the first term is nonnegative. For the second term we have, for any θ, by Jensen's inequality for concave functions (log is concave)

$$\begin{aligned} H(\theta \mid \theta^{(i)}) - H(\theta^{(i)} \mid \theta^{(i)}) &= E_{\theta^{(i)}}\Big[\log\Big(\frac{f_{Z|Y,\Theta}(Z \mid Y, \theta)}{f_{Z|Y,\Theta}(Z \mid Y, \theta^{(i)})}\Big) \Bigm| Y\Big] \\ &\le \log E_{\theta^{(i)}}\Big[\frac{f_{Z|Y,\Theta}(Z \mid Y, \theta)}{f_{Z|Y,\Theta}(Z \mid Y, \theta^{(i)})} \Bigm| Y\Big] \\ &= \log \int \frac{f_{Z|Y,\Theta}(z \mid Y, \theta)}{f_{Z|Y,\Theta}(z \mid Y, \theta^{(i)})} f_{Z|Y,\Theta}(z \mid Y, \theta^{(i)}) \nu_z(dz) \\ &= \log 1 = 0. \end{aligned}$$

This shows $l_{obs}(\theta^{(i+1)}, Y) - l_{obs}(\theta^{(i)}, Y) \ge 0$. □

This proposition is the main ingredient in the following theorem due to Wu (1983). We also need the following assumptions:

- Ω is a subset of $\mathbb{R}^k$.
- $\Omega_{\theta_o} = \{\theta \in \Omega : l((\theta, y) \ge l(\theta_o, y)\}$ is compact for any $l(\theta_o, y) > -\infty$.
- $l(\theta_o, x)$ is continuous and differentiable in the interior of Ω.

Theorem 22. *Suppose that $Q(\theta \mid \phi)$ is continuous in both θ and ϕ. Then all limit points of any instance $\{\theta^{(i)}\}$ of the EM algorithm are stationary points, i.e. $\theta^* = argmax Q(\theta \mid \theta^*)$, and $l(\theta^{(i)}, y)$ converges monotonically to some value $l^* = l(\theta^*, y)$ for some stationary point θ^*.*

To show that $\theta^{(i)}$ converges to some θ^* we can use the next theorem.

Theorem 23. *Assume the hypothesis of Theorem 22. Suppose in addition that $\partial_\theta Q(\theta \mid \phi)$ is continuous in θ and ϕ. Then $\theta^{(i)}$ converges to a stationary point θ^* with $l(\theta^*, y) = l^*$, the limit of $l(\theta^{(i)})$ if either*

$$\{\theta : l(\theta, y) = l^*\} = \{\theta^*\}$$

or $|\theta^{(i+1)} - \theta^{(i)}| \to 0$ and $\{\theta : l(\theta, y) = l^*\}$ is discrete.

In the case the likelihood is unimodal everything simplifies.

Corollary 1. *Suppose $l(\theta, y)$ is unimodal in Ω with θ^* being the only stationary point and $\partial_\theta Q(\theta \mid \phi)$ is continuous in θ and ϕ. Then $\theta^{(i)}$ converges to the unique maximizer θ^* of $l(\theta, y)$ which is the MLE.*

15.4. **EM algorithm in Bayesian estimation.** In this section we show how the EM algorithm can be used to produce a maximum a posteriori (MAP) estimate in a Bayesian framework.

Suppose we have a prior density $f_\Theta(\theta)$ for Θ and we write $f_{\Theta|Y}(\theta \mid y)$ and $f_{\Theta|X}(\theta \mid x)$ for the posterior of the observed and complete data, respectively. Then the MAP estimate is the value of θ that maximizes

$$\log f_{\Theta|Y}(\theta \mid y) = \log f_{Y|\Theta}(y \mid \theta) + \log f_\Theta(\theta).$$

The EM algorithm can then be implemented as follows.

E-step: Compute (ignoring terms that does not depend on the parameter)

$$E_{\theta^{(i)}}[\log f_{\Theta|X}(\theta \mid X) \mid Y = y] = Q(\theta \mid \theta^{(i)}) + \log f_\Theta(\theta).$$

M-step: Choose $\theta^{(i+1)}$ to maximize this expression among θ.

16. Monte Carlo EM algorithm

In some applications of the EM algorithm the E-step is complex and does not admit a closed form solution. That is, the $Q(\theta \mid \theta^{(i)})$ function cannot be computed explicitly. A solution is provided by evaluating the $Q(\theta \mid \theta^{(i)})$ function by Monte Carlo methods. This will then be called the MCEM algorithm. Note that we can write

$$\begin{aligned} Q(\theta \mid \theta^{(i)}) &= E_{\theta^{(i)}}[\log f_{X|\Theta}(X \mid \theta) \mid Y] \\ &= E_{\theta^{(i)}}[\log f_{X|\Theta}((Y, Z) \mid \theta) \mid Y]. \end{aligned}$$

The MCEM algorithm consists of the following steps. Suppose we observe $Y = y$.

- (MC E-step) On the $(i+1)$th iteration, draw $z^1, \dots, z^{(M)}$ from $f_{Z|Y,\Theta}(\cdot \mid y, \theta^{(i)})$. Approximate the Q-function as

$$Q(\theta \mid \theta^{(i)}) = \frac{1}{M} \sum_{m=1}^{M} \log f_{X|\Theta}((y, z^m) \mid \theta).$$

- (M-step) Maximize the approximate $Q(\theta \mid \theta^{(i)})$ and put $\theta^{(i+1)}$ as the maximizer.

Example 25. In the multinomial example it would look as follows (although it is not needed there because explicit computations are possible).

The missing data $Z = Y_{12}$ which, conditionally on $Y_1 = y_1$ and $\Theta = \theta^{(i)}$ has Binomial distribution $\text{Bin}(y_1, \theta^{(i)}/(2+\theta^{(i)}))$. In the MC E-step we would generate $z^{(1)}, \dots, z^{(M)}$ from this Binomial distribution. The approximate Q function is then given by

$$\begin{aligned} Q(\theta \mid \theta^{(i)}) &= \frac{1}{M} \sum_{m=1}^{M} \log f_{X|\Theta}((y, z^m) \mid \theta) \\ &= (\bar{z}_M + y_4) \log \theta + (y_2 + y_3) \log(1-\theta), \end{aligned}$$

where $\bar{z}_M = M^{-1} \sum_{i=1}^{M} z^{(i)}$. Thus, the updated $\theta^{(i+1)}$ becomes

$$\theta^{(i+1)} = \frac{\bar{z}_M + y_4}{\bar{z}_M + y_4 + y_2 + y_3}.$$

In the Bayesian setting we want to compute

$$E_{\theta^{(i)}}[\log f_{\Theta|X}(\theta \mid X) \mid Y] = \int \log f_{\Theta|X}(\theta \mid (Y, z)) f_{Z|Y,\Theta}(z \mid Y, \theta^{(i)}) \nu_z(dz), \quad (16.1)$$

in the E-step.

The MCEM algorithm then consists of the following steps. Suppose we observe $Y = y$.

- (MC E-step) On the $(i+1)$th iteration, draw $z^1, \dots, z^{(M)}$ from $f_{Z|Y,\Theta}(\cdot \mid y, \theta^{(i)})$. Approximate the function (16.1) by

$$\frac{1}{M} \sum_{m=1}^{M} \log f_{\Theta|X}(\theta \mid z^m, y)$$

- (M-step) Maximize this function and put $\theta^{(i+1)}$ as the maximizer.

For the MCEM algorithm monotonicity propeties are lost, but there are results that say that in certain cases the algorithm gets close to a maximizer with high probability.

16.1. **Stochastic EM algorithm.** A particular instance of the MCEM algorithm has recieved the name stochastic EM algorithm. It is actually just the MCEM with $M = 1$ (the number of samples).

Thus, the stochastic EM algorithm then consists of the following steps. Suppose we observe $Y = y$.

- (MC E-step) On the $(i+1)$th iteration, draw z from $f_{Z|Y,\Theta}(\cdot \mid y, \theta^{(i)})$. Plug it in as the true value of z so that the likelihood function is $l(\theta, (y, z))$.
- (M-step) Maximize this function and put $\theta^{(i+1)}$ as the maximizer.

The stochastic EM algorithm prevents the sequence from staying near an unstable stationary point of the likelihood function. The sequence of estimates actually form an ergodic Markov chain and converges weakly to a stationary distribution.

References

[1] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm, *Annals of Statistics* 39, 1 – 38.

[2] MacLachlan, G.J. and Krishnan, T. (1997) *The EM algorithm and extensions*, John Wiley & Sons ISBN 0-471-12358-7.