In All Likelihood: Statistical Modelling and Inference Using Likelihood – By Yudi Pawitan

Chapter 12: Expectation–Maximization (**EM Algorithm)**

- Motivation (What is the missing and complete data?)
- General Specification (What do we mean by *E-step* and *M-step*?)
- Exponential family model
- General Properties
- Examples
  - Mixture models
  - Robust estimations

12.1 Motivation

Finding maximum likelihood estimates usually requires a numerical method, which has been motivated from calculus, but a statistically motivated EM (Expectation-Maximization) algorithm appears naturally in problems where

- some parts of the data are missing, and analysis of the *incomplete data* is somewhat complicated or nonlinear;
- it is possible to 'fill in' the missing data, and analysis of the *complete data* is relatively simple.

Example 12.1: Consider a two-way table of y_{ij} for $i = 1, 2$ and $j = 1, 2, 3$ with one missing cell y_{23}:

10	15	17
22	23	-

Suppose we consider a linear model

$$y_{ij} = \mu + \alpha_i + \beta_j + e_{ij}$$

where $\sum_i \alpha_i = \sum_j \beta_j = 0$, and e_{ij}'s are an iid sample from $N(0, \sigma^2)$. The MLEs of μ, α_i and β_j are the minimizer of the sum of square

$$\sum_{ij} (y_{ij} - \mu - \alpha_i - \beta_j)^2$$

subject to the constraints. The solution can be obtained from the least square estimate $(X'X)^{-1}X'y$, where X is the design matrix:

$$X = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & -1 & 0 & 1 \\ 1 & 1 & -1 & -1 \end{pmatrix}.$$

We will get $\hat{\mu} = 19, \widehat{\alpha_1} = -5, \widehat{\beta_1} = -3$ and $\widehat{\beta_2} = 0$.

Had there been no missing data there is a simple closed form solution:

$$\hat{\mu} = \bar{y}, \qquad \hat{\alpha}_i = \bar{y}_{i\cdot} - \bar{y}, \qquad \hat{\beta}_j = \bar{y}_{\cdot j} - \bar{y}$$

The question is, how can we use this result for the 'complete date' to help us estimate the parameters from the incomplete data?

One way to do this is first to 'fill in' the missing data y_{23}, by the average of the available data y, then compute the parameter estimates according to the complete data formulae. This constitutes one cycle of an iteration. The iteration continues by recomputing the missing data

$$\widehat{y}_{23} = \widehat{\mu} + \widehat{\alpha}_2 + \widehat{\beta}_3$$

and the parameter estimates until convergence. In this example, starting with $\hat{y}_{23} = 17.4$ we obtain:

Iteration	$\hat{\mu}$	$\hat{\alpha}_1$	$\hat{\beta}_1$	$\hat{\beta}_2$
1	17.400	-3.400	-1.400	1.600
2	17.933	-3.933	-1.933	1.067
3	18.289	-4.289	-2.289	0.711
10	18.958	-4.958	-2.958	0.042
15	18.995	-4.995	-2.995	0.005
21	19.000	-5.000	-3.000	0.000

Thus the algorithm arrives at the solution without inverting the matrix. □

12.2 General Specification

In Problems where EM is relevant, we will denote the available data set y as 'incomplete data' and x as 'complete data'. In general $y = h(x)$ for some array-valued function $h(\cdot)$; this means that y is completely determined by x, but not vice versa. For example, if $x = (x_1, x_2, x_3)$ then $y = (x_1, x_2 + 2x_3)$ is a form of incomplete data. **The key idea is that some information is lost by going from x to y.**

We need to estimate θ from the likelihood based on y, $L(\theta; y) = p_\theta(y)$. The dependence on y is made explicit, so we can distinguish from $L(\theta; x) = p_\theta(x)$, the likelihood based on x.

The EM algorithm obtains the MLE $\hat{\theta}$ by the following iteration:

- Start with initial value θ^0
- *E-step:* compute the conditional expected value

$$Q(\theta) = Q(\theta|\theta^0) \equiv E\{\log L(\theta; x)|y, \theta^0\}$$

- *M-step:* maximize $Q(\theta)$ to give an updated value θ^1, the go to the E-step using the updated value, and iterate until convergence.

Example 12.2: The famous genetic example from Rao (1973, page 369) assume that the phenotype data

$$y = (y_1, y_2, y_3, y_4) = (125, 18, 20, 34)$$

is distributed according to multinomial distribution with probabilities

$$\left\{\frac{1}{2} + \frac{\theta}{4}, \frac{(1-\theta)}{4}, \frac{(1-\theta)}{4}, \frac{\theta}{4}\right\}.$$

The log-likelihood based on y is

$$\log L(\theta; y) = y_1 \log(2+\theta) + (y_2 + y_3)\log(1-\theta) + y_4 \log(\theta), \tag{1}$$

which does not yield a closed form estimate of θ. Now lets treat y as an incomplete data from $x = (x_2, x_1, x_3, x_4, x_5)$ with multinomial probabilities

$$\left\{\frac{1}{2}, \frac{\theta}{4}, \frac{(1-\theta)}{4}, \frac{(1-\theta)}{4}, \frac{\theta}{4}\right\}.$$

Here $y_1 = x_1 + x_2$. The log-likelihood based on x is

$$\log L(\theta; x) = (x_2 + x_5)\log(\theta) + (x_3 + x_4)\log(1-\theta), \tag{2}$$

which readily yields

$$\hat{\theta} = \frac{x_2 + x_5}{x_2 + x_3 + x_4 + x_5},$$

here is the case where the 'complete data' x is simpler than y.

In this example the E-step is to find

$$\begin{aligned} Q(\theta) &= E(x_2 + x_5|y, \theta^0)\log(\theta) + E(x_3 + x_4|y, \theta^0)\log(1-\theta) \\ &= \{E(x_2|y, \theta^0) + x_5\}\log(\theta) + (x_3 + x_4)\log(1-\theta), \end{aligned}$$

so we only need to compute

$$\widehat{x}_2 = E(x_2|y, \theta^0).$$

Since $x_1 + x_2 = y_1$, the conditional distribution of $x_2|y_1 \sim$ Binomial$(y_1 = 125, p^0 = \frac{\theta^0/4}{1/2+\theta^0/4})$, so

$$\widehat{x}_2 = y_1 \frac{\theta^0/4}{1/2 + \theta^0/4}. \tag{3}$$

The M-step yields an update

$$\theta_1 = \frac{\widehat{x}_2 + x_5}{\widehat{x}_2 + x_3 + x_4 + x_5}, \tag{4}$$

The algorithm iterates between (3) and (4). From the last category of y we may obtain a starting value: $\theta^0/4 = 34/197$. The first five iterates are 0.690, 0.635, 0.627, 0.627, giving the MLE $\hat{\theta} = 0.627$.□

It's easy to see that, in the EM algorithm, the objective function $\log L(\theta; y)$ is approximated by $Q(\theta)$, then by using an initial estimate θ^0, we try to find θ^1 as the maximizer of $Q(\theta)$.

12.3 Exponential family model

In general, there is no guarantee that a particular EM algorithm exists for a particular incomplete data, but the algorithm is simple and theoretically illuminating, if the complete data x is in the full exponential family:

$$\log L(\theta; x) = \theta' T - A(\theta)$$

where $T \equiv T(x)$ is a p-vector of sufficient statistics. At the n'th iteration the E-step is to find

$$\begin{aligned} Q(\theta|\theta^n) &= E\{\log L(\theta; x)|y, \theta^n\} \\ &= \theta' E(T|y, \theta^n) - A(\theta), \end{aligned}$$

which reduces to finding the conditional expected value $\hat{T} = E\{T|y, \theta^n\}$.

For the M-step, taking the derivative of $Q(\theta|\theta^n)$ with respect to θ, and setting it to zero, we solve the equation

$$\frac{\partial}{\partial\theta} A(\theta) = \hat{T}$$

to get an update θ^{n+1}.

Recall that for a full exponential family

$$\frac{\partial}{\partial\theta} A(\theta) = E(T|\theta),$$

so the updating equation for θ satisfies

$$E(T|\theta^{n+1}) = E(T|y, \theta^n), \tag{5}$$

and at convergence we have the MLE $\hat{\theta}$ satisfying

$$E(T|\hat{\theta}) = E(T|y, \hat{\theta}). \tag{6}$$

This means that $\theta = \hat{\theta}$ is the value that makes T and y uncorrelated. Now assume $h_1(\theta) \equiv E(T|y, \theta)$ and $h_2(\theta) \equiv E(T|\theta)$. The conditional density of x given y is

$$p_\theta(x|y) = \frac{p_\theta(x)}{p_\theta(y)}$$

since y is completely determined by x. So, with obvious notations,

$$\begin{aligned} \log L(\theta; x|y) &= \log L(\theta; x) - \log L(\theta; y) \\ &= \theta T - A(\theta) - \log L(\theta; y), \end{aligned} \tag{7}$$

which is also in the exponential family. Taking the derivative with respect to θ, and taking conditional expectation, we obtain

$$E(T|y, \theta) - A'(\theta) - S(\theta; y) = 0$$

so

$$h_1(\theta) = E(T|y, \theta) = A'(\theta) + S(\theta; y)$$

and

$$h_2(\theta) \equiv E(T|\theta) = A'(\theta)$$

The slops of these functions are

$$\begin{aligned} h_1'(\theta) &= A''(\theta) - I(\theta; y) \\ h_2'(\theta) &= A''(\theta), \end{aligned}$$

where $I(\theta; y)$ is the Fisher information based on y. Since

$$\begin{aligned} var(T) &= A''(\theta) > 0 \\ var(T|y) &= A''(\theta) - I(\theta; y) > 0 \end{aligned}$$

both $h_1(\theta)$ and $h_1(\theta)$ are increasing functions of θ. Furthermore $I(\theta; y) > 0$ for θ near $\hat{\theta}$, so $h_2(\theta)$ has a steeper slope.
Taking the conditional expectation of (7) given y yields

$$E\{\log L(\theta; x|y)|y, \theta^0\} = Q(\theta|\theta^0) - \log L(\theta; y) \tag{8}$$

Derivatives of $E\{\log L(\theta; x|y)|y, \theta^0\}$ behave like the expected score and Fisher information; for example,

$$\begin{aligned} I(\theta; x|y) &\equiv -\partial^2 E\{\log L(\theta; x|y)|y, \theta^0\}/\partial\theta^2 \\ &= -E\{\partial^2 \log L(\theta; x|y)/\partial\theta^2|y, \theta^0\}. \end{aligned}$$

Defining

$$I(\theta; x) \equiv -\partial^2 Q(\theta|\theta^0)/\partial\theta^2,$$

we have from (8)

$$I(\theta; x|y) = I(\theta; x) - I(\theta; y)$$

or

$$I(\theta; y) = I(\theta; x) - I(\theta; x|y) \tag{9}$$

This intuitively means that **the information in the incomplete data y is equal to the information in the complete data x minus the extra information in x which is not in y**. This is a form of the so-called 'missing information principle'. Near the solution $\hat{\theta}$ we have

$$\begin{aligned} E(T|\theta) &\approx E(T|\hat{\theta}) - I(\hat{\theta}; x)(\theta - \hat{\theta}) \\ E(T|y, \theta) &\approx E(T|y, \hat{\theta}) - I(\hat{\theta}; x|y)(\theta - \hat{\theta}) \end{aligned}$$

Assuming $\theta^n \to \hat{\theta}$ as $n \to \infty$, and in view of (5) and (6). we have

$$\frac{\theta^{n+1} - \hat{\theta}}{\theta^n - \hat{\theta}} \approx \frac{I(\hat{\theta}; x|y)}{I(\hat{\theta}; x)}.$$

Smaller $I(\hat{\theta}; x|y)$, meaning less missing information in y relative to x, implies, faster convergence. This is a more precise expression of the previous notion that the speed of convergence is determined by how close $Q(\theta|\theta^0)$ is to $\log L(\theta; y)$.

12.4 General properties

One of the important properties of the EM algorithm is that its step always increases the likelihood

$$L(\theta^{n+1}; y) \geq L(\theta^n; y) \tag{10}$$

This makes EM a numerically stable procedure as it climbs the likelihood surface; in contrast, we don't have this property in Newton-Raphson algorithm. However, the likelihood-climbing property doesn't guarantee the convergence. **Another practical advantage of the EM algorithm is that it usually handles parameter constrains automatically**. This is because each M-step produces an MLE-type estimate.

The main disadvantage of the EM algorithm compared with the competing Newton-Raphson algorithm are

- **The convergence can be very slow**. As discussed above, the speed of convergence is determined by the amount of missing information in y relative to x. There is no general technique to manipulate the complete data x to minimize the amount of missing information.
- **there are no immediate standard errors for the estimators**. If there is an explicit log-likelihood function $\log L(\theta; y)$, then one can easily find the observed information $I(\hat{\theta}; y)$ numerically, and find the standard errors from the inverse of $I(\hat{\theta}; y)$.

To prove (10), recall from (8) that

$$E\{\log L(\theta; x|y)|y, \theta^0\} \quad = \quad Q(\theta|\theta^0) - \log L(\theta; y) \tag{11}$$

Let $h(\theta|\theta^0) \equiv E\{\log L(\theta; x|y)|y, \theta^0\}$. From the information inequality (Theorem 9.5), for any two densities $f(x) \neq g(x)$ we have

$$E_g \log f(X) \leq E_g \log g(X).$$

Applying this to the conditional density of $x|y$,

$$h(\theta|\theta^0) \leq h(\theta^0|\theta^0)$$

and at the next iterate θ^1 we have

$$Q(\theta^1|\theta^0) - \log L(\theta^1; y) \leq Q(\theta^0|\theta^0) - \log L(\theta^0; y),$$

or

$$\log L(\theta^1; y) - \log L(\theta^0; y) \geq Q(\theta^1|\theta^0) - Q(\theta^0|\theta^0) \geq 0$$

The right-hand side is positive by definition of θ^1 as the maximizer of $Q(\theta|\theta^0)$. In fact, the monotone likelihood-climbing property is satisfies as long as we choose the next iterate that satisfies $Q(\theta^1|\theta^0) - Q(\theta^0|\theta^0) \geq 0$.

In particular, the EM algorithm can get trapped in a local maximum or saddle points. If the likelihood surface is unimodal and $Q(\theta|\theta^0)$ is continuous in both θ and θ^0, then the EM algorithm is convergent.

12.5 Mixture Models

Let $y = (y_1, \cdots, y_n)$ be an iid sample from a mixture model with density

$$p_\theta(u) = \sum_{j=1}^{J} \pi_j p_j(u|\theta_j) \tag{12}$$

where π_j's are unknown mixing probabilities, such that $\sum_j \pi_j = 1$, $p_j(u|\theta_j)$'s are probability models, and θ_j's are unknown parameters. Hence θ is the collection of π_j's and θ_j's. The log-likelihood based on the observed data y is

$$\log L(\theta; y) = \sum_i \log\left\{\sum_{j=1}^{J} \pi_j p_j(u|\theta_j)\right\}.$$

Because of the constrains on π_j's, a simplistic application of the Newton-Raphson algorithm is prone to failure.

Example 12.3: The waiting time (in minutes) of $N = 299$ consecutive eruptions of the Old Faithful geyser in Yellowstone National Park. The bimodal nature of the distribution, as shown in Figure 1, suggests a mixture of two processes. We model the data as coming from a normal mixture

$$\pi_1 N(\mu_1, \sigma_1^2) + \pi_2 N(\mu_2, \sigma_2^2).$$

Here $\pi_2 = 1 - \pi_1$, so $\theta = (\pi_1, \mu_1, \sigma_1, \mu_2, \sigma_2)$. The log-likelihood function is

$$\log L(\theta; y) = \sum_i \log\{\pi_1 \phi(y_i, \mu_1, \sigma_1^2) + (1 - \pi_1)\phi(y_i, \mu_2, \sigma_2^2)\},$$

where $\phi(y, \mu, \sigma^2)$ is the density of $N(\mu, \sigma^2)$.□

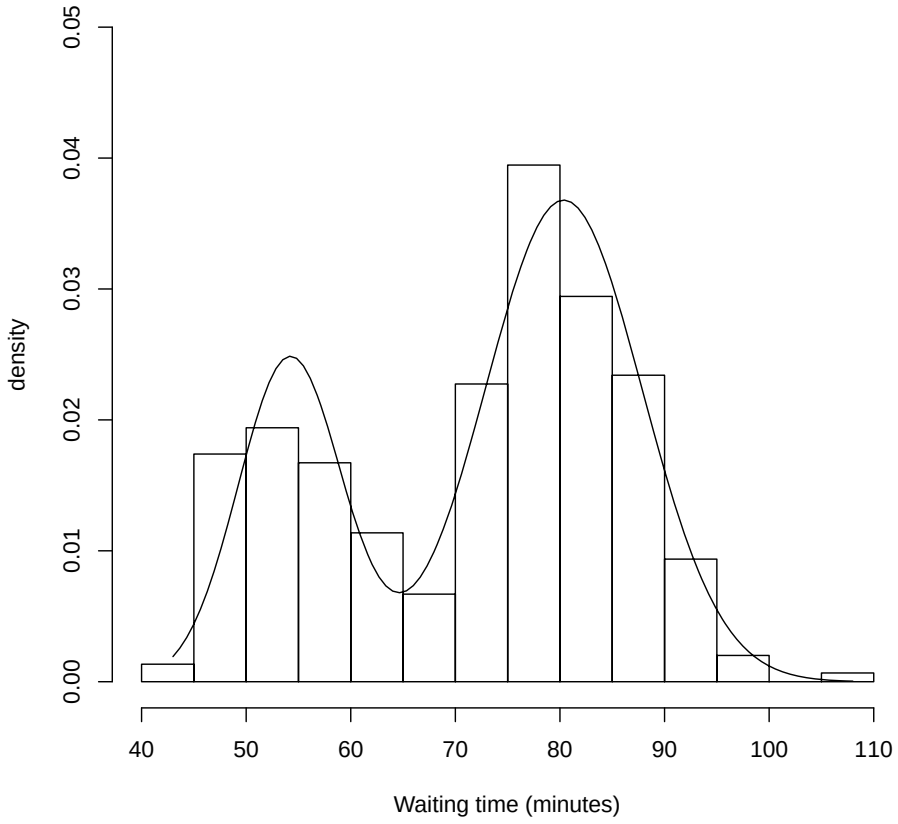

Figure 1: The histogram of the geyser waiting time and the parametric density estimate (solid line) based on mixture of two normals.

One interpretation of the mixture model (12) is that y_i comes from one of the J populations, but we do not know which one. Had we observed the indicator z_i of where y_i is coming from, each θ_j could be estimated separately, and the estimation of θ would become trivial.

We define the 'complete data' $x = (x_1, \cdots, x_n)$, where $x_i = (y_i, z_i)$. The marginal probability of Z_i is $P(Z_i = j) = \pi_j$; conditional on $z_i = j$, assume y_i has density $p_j(u|\theta_j)$. Now let

$$\log L(\theta; x) = \sum_i \log L(\theta, x_i)$$

where the contribution of x_i to the log-likelihood is

$$\begin{aligned} \log L(\theta; x_i) &= \log p_{z_i}(y_i|\theta_{z_i}) + \log \pi_{z_i} \\ &= \sum_{j=1}^{J} \{I(z_i = j) \log p_j(y_i|\theta_j) + I(z_i = j) \log \pi_j\}, \end{aligned} \tag{13}$$

and $I(z_i = j) = 1$ if $z_i = j$ and zero otherwise. So, with starting value θ^0, the E-step consists of finding the

conditional probabilities

$$\begin{aligned}\widehat{p}_{ij} &= E\{I(Z_i = j)|y_i, \theta^0\} \\ &= P(Z_i = j|y_i, \theta^0) \\ &= \frac{\pi_j^0 p_j(y_i|\theta_j^0)}{p_{\theta^0}(y_i)} \\ &= \frac{\pi_j^0 p_j(y_i|\theta_j^0)}{\sum_k \pi_k^0 p_k(y_i|\theta_k^0)}.\end{aligned}$$

This is the estimated probability of y_i coming from population j; in clustering problems it is the quantity of interest. It is immediate that $\sum_j \widehat{p}_{ij} = 1$ for each i.

From (13) the M-step update of each θ_j is based on *separately* maximizing the weighted log-likelihood

$$\sum_i \widehat{p}_{ij} \log p_j(y_i|\theta_j).$$

While there is no guarantee of a closed-form update, this is a major simplification of the problem. Explicit formulae are available, for example, for the normal model. We can also show the update formula

$$\pi_j^1 = \frac{\sum_i \widehat{p}_{ij}}{n}$$

for the mixing probabilities.

Example 12.3: continued. For the Old Faithful data the weighted likelihood of the j'th parameter $\theta_j = (\mu_j, \sigma_j)$ is

$$-\frac{1}{2}\sum_i \widehat{p}_{ij}\left\{\log \sigma_j^2 + \frac{(y_i - \mu_j)^2}{\sigma_j^2}\right\},$$

which yields the following weighted averages as updates:

$$\begin{aligned}\mu_j^1 &= \frac{\sum_i \widehat{p}_{ij} y_i}{\sum_i \widehat{p}_{ij}} \\ \sigma_j^{2(1)} &= \frac{\sum_i \widehat{p}_{ij}(y_i - \mu_j^1)^2}{\sum_i \widehat{p}_{ij}}.\end{aligned}$$

Starting with $(\pi_1^0 = 0.3, \mu_1^0 = 55, \sigma_1^0 = 4, \mu_2^0 = 80, \sigma_2^0 = 7)$ we obtain the following iterations:

Iteration	π_1	μ_1	σ_1	μ_2	σ_2
1	0.306	54.092	4.813	80.339	7.494
2	0.306	54.136	4.891	80.317	7.542
3	0.306	54.154	4.913	80.323	7.541
5	0.307	54.175	4.930	80.338	7.528
10	0.307	54.195	4.946	80.355	7.513
15	0.308	54.201	4.951	80.359	7.509
25	0.308	54.203	4.952	80.360	7.508

Hence we obtain $(\hat{\pi}_1 = 0.308, \hat{\mu}_1 = 54.203, \hat{\sigma}_1 = 4.952, \hat{\mu}_2 = 80.360, \hat{\sigma}_2 = 7.508)$, giving the density estimate

$$\hat{p}(u) = \hat{\pi}_1 \phi(y_i, \hat{\mu}_1, \hat{\sigma}_1^2) + (1 - \hat{\pi}_1)\phi(y_i, \hat{\mu}_2, \hat{\sigma}_2^2).$$

Figure 1 compares this parametric density with the histogram.□

12.6 Robust estimation

As described in Section 6.9 we can perform a robust regression analysis by assuming a heavy-tailed error distribution. The EM algorithm applied to this problem becomes as IWLS algorithm.

Suppose $y_1, \cdots, y_n$ are independent with locations $\mu_1, \cdots, \mu_n$ and a common scale parameter, such that

$$\mu_i = x_i'\beta,$$

or we write it as a regression model

$$y_i = x_i'\beta + e_i,$$

where the e_i has a t_k-distribution with unknown scale σ and degrees of freedom k. Hence the total parameter is $\theta = (\beta, \sigma, k)$. From t_k-density function, the contribution of y_i to the observed-data likelihood is

$$\log L(\theta; y_i) = \log\Gamma(k/2+1/2) - \log\Gamma(k/2) - \frac{1}{2}\log k - \frac{1}{2}\log\sigma^2 - \frac{k+1}{2}\log\left\{k + \frac{(y_i-\mu_i)^2}{\sigma^2}\right\}. \quad (14)$$

The way to proceed with EM algorithm may not be immediately obvious here, but recall that we may write

$$e_i = \sigma z_i/\sqrt{w_i},$$

where z_i is $N(0,1)$, and w_i is χ^2_k/k independent of z_i. So, if we knew w_i the regression problem would reduce to a normal-based regression problem.

Defining the 'complete data' as $x_i = (y_i, w_i)$, for $i = 1, \cdots, n$, the contribution of x_i to the complete data likelihood is

$$L(\theta; x_i) = p(y_i|w_i)p(w_i).$$

The conditional distribution $y_i|w_i$ is normal with mean μ_i and variance σ^2/w_i; the density of w_i is

$$p(w) = \frac{1}{2^{k/2}\Gamma(k/2)} w^{k/2-1} e^{-kw/2}.$$

Hence

$$\log L(\theta; x_i) = v(k) + \frac{k-1}{2}\log w_i - \frac{kw_i}{2} - \frac{1}{2}\log\sigma^2 - \frac{w_i(y_i-\mu_i)^2}{\sigma^2},$$

where $v(k)$ is a function of involving k only.

The E-step consists of finding $E(\log w_i|y_i, \theta^0)$ and $E(w_i|y_i, \theta^0)$. There is no closed form result for the former, thus raising a question regarding the practically of the algorithm. Note, however, that $E(\log w_i|y_i, \theta^0)$ is only needed for updating k, while updating β and σ^2 only requires $E(w_i|y_i, \theta^0)$.

What we can do instead is to consider the estimation of β and σ^2 at each fixed k, such that we get a profile likelihood of k from $\log L(\theta; y)$ in (14). The MLE of k is then readily available from the profile likelihood. Thus the EM algorithm can be performed at each k. The E-step reduces to finding

$$\hat{w}_i \equiv E(w_i|y_i, \beta^0, \sigma^{2(0)}).$$

We can show that the conditional distribution of $w_i|y_i$ is is $\chi^2_{k+1}/(k+d_i^2)$, where

$$d_i^2 = \frac{(y_i-\mu_i^0)^2}{\sigma_i^{2(0)}},$$

so

$$\hat{w}_i = \frac{k+1}{k+d_i^2}.$$

For β and σ^2 only, the relevant term of the E-step function is

$$Q = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_i \hat{w}_i(y_i - \mu_i)^2.$$

This is the usual likelihood associated with weighted least squared, so the update of β is

$$\beta^1 = (X'WX)^{-1}X'Wy,$$

where the weight matrix W is diagonal matrix of $\hat{w}_i$, and the update of σ^2 is

$$\sigma^{2(1)} = \frac{1}{n}\sum_i \hat{w}_i(y_i - \mu_i)^2,$$

thus completing the M-step.

For one sample problems, where $y_1, \cdots, y_n$ have a common location μ, the update formula for μ is simply a weighted average

$$\mu_1 = \frac{\sum_i \hat{w}_i y_i}{\sum_i \hat{w}_i}.$$

The weight $\hat{w}_i$ is small if y_i is far from μ^0, so outlying observation are downweighted. This is the source of the robustness in the estimation.