Chapter 7

The Expectation-Maximisation Algorithm

7.1 The EM algorithm - a method for maximising the likelihood

Let us suppose that we observe $\underline{Y} = \{Y_i\}_{i=1}^n$. The joint density of $\underline{Y}$ is $f(\underline{Y}; \theta_0)$, and θ_0 is an unknown parameter. Our objective is to estimate θ_0. The log-likelihood of $\underline{Y}$ is

$$\mathcal{L}_n(\underline{Y}; \theta) = \log f(\underline{Y}; \theta),$$

Observe, that we have not specified that $\{Y_i\}$ are iid random variables. This is because the procedure that we will describe below is very general and the observations do not need to be either independent or identically distributed (indeed an interesting extension of this procedure, is to time series with missing data first proposed in Shumway and Stoffer (1982) and Engle and Watson (1982)). Our objective is to estimate θ_0, in the situation where either evaluating the log-likelihood $\mathcal{L}_n$ or maximising $\mathcal{L}_n$ is difficult. Hence an alternative means of maximising $\mathcal{L}_n$ is required. Often, there may exist unobserved data $\{\underline{U} = \{U_i\}_{i=1}^m\}$, where the likelihood of $(\underline{Y}, \underline{U})$ can be 'easily' evaluated. It is through these unobserved data that we find an alternative method for maximising $\mathcal{L}_n$.

The EM-algorithm was specified in its current form in Dempster, Laird and Rubin (1977)(`https://www.jstor.org/stable/pdf/2984875.pdf`) however it was applied previously to several specific models.

Example 7.1.1 *(i) Suppose that* $\{f_j(\cdot;\theta);\theta\}_{j=1}^m$ *are a sequence of densities from* m *exponential classes of densities. In Sections 1.6 and 1.6.5 we showed that it was straightforward to maximise each of these densities. However, let us suppose that each* $f_j(\cdot;\theta)$ *corresponds to one subpopulation. All the populations are pooled together and given an observation* X_i *it is unknown which population it comes from. Let* δ_i *denote the subpopulation the individual* X_i *comes from i.e.* $\delta_i \in \{1,\ldots,m\}$ *where* $P(\delta_i = j) = p_j$.

The density of all these mixtures of distribution is

$$f(x;\theta) = \sum_{j=1}^m f(X_i = x|\delta_i = j)P(\delta_i = j) = \sum_{j=1}^m p_j f_j(x;\theta)$$

where $\sum_{j=1}^m p_j = 1$. *Thus the log-likelihood of* $\{X_i\}$ *is*

$$\sum_{i=1}^n \log\left(\sum_{j=1}^m p_j f_j(X_i;\theta)\right).$$

Of course we require that $\sum_{j=1}^m p_j = 1$, *thus we include a lagrange multiplier to the likelihood to ensure this holds*

$$\sum_{i=1}^n \log\left(\sum_{j=1}^m p_j f_j(X_i;\theta)\right) + \lambda\left(\sum_{j=1}^m p_j - 1\right).$$

It is straightforward to maximise the likelihood for each individual subpopulation, however, it is extremely difficult to maximise the likelihood of this mixture of distributions.

The data $\{X_i\}$ *can be treated as missing, since the information* $\{\delta_i\}$ *about the which population each individual belongs to is not there. If* δ_i *were known the likelihood of* $\{X_i,\delta_i\}$ *is*

$$\prod_{j=1}^m\prod_{i=1}^n (p_j f_j(X_i;\theta))^{I(\delta_i=j)} = \prod_{i=1}^n p_{\delta_i} f_{\delta_i}(X_i;\theta)$$

which leads to the log-likelihood of $\{X_i,\delta_i\}$ *which is*

$$\sum_{i=1}^n \log p_{\delta_i} f_{\delta_i}(X_i;\theta) = \sum_{i=1}^n (\log p_{\delta_i} + \log f_{\delta_i}(X_i;\theta))$$

which is far easier to maximise. Again to ensure that $\sum_{j=1}^{m} p_j = 1$ *we include a Lagrange multiplier*

$$\sum_{i=1}^{n} \log p_{\delta_i} f_{\delta_i}(X_i;\theta) = \sum_{i=1}^{n} (\log p_{\delta_i} + \log f_{\delta_i}(X_i;\theta)) + \lambda \left(\sum_{j=1}^{m} p_j - 1 \right).$$

It is easy to show that $\widehat{p}_j = n^{-1} \sum_{i=1}^{n} I(\delta_i = j)$.

(ii) Let us suppose that $\{T_i\}_{i=1}^{n+m}$ *are iid survival times, with density* $f(x;\underline{\theta}_0)$. *Some of these times are censored and we observe* $\{Y_i\}_{i=1}^{n+m}$, *where* $Y_i = \min(T_i, c)$. *To simplify notation we will suppose that* $\{Y_i = T_i\}_{i=1}^{n}$, *hence the survival time for* $1 \leq i \leq n$, *is observed but* $Y_i = c$ *for* $n+1 \leq i \leq n+m$. *Using the results in Section the log-likelihood of* $\underline{Y}$ *is*

$$\mathcal{L}_n(\underline{Y};\theta) = \left(\sum_{i=1}^{n} \log f(Y_i;\theta) \right) + \left(\sum_{i=n+1}^{n+m} \log \mathcal{F}(Y_i;\theta) \right).$$

The observations $\{Y_i\}_{i=n+1}^{n+m}$ *can be treated as if they were missing. Define the 'complete' observations* $\underline{U} = \{T_i\}_{i=n+1}^{n+m}$, *hence* $\underline{U}$ *contains the unobserved survival times. Then the likelihood of* $(\underline{Y}, \underline{U})$ *is*

$$\mathcal{L}_n(\underline{Y}, \underline{U};\theta) = \sum_{i=1}^{n+m} \log f(T_i;\theta).$$

If no analytic express exists for the survival function $\mathcal{F}$, *it is easier to maximise* $\mathcal{L}_n(\underline{Y}, \underline{U})$ *than* $\mathcal{L}_n(\underline{Y})$.

We now formally describe the EM-algorithm. As mentioned in the discussion above it is often easier to maximise the joint likelihood of $(\underline{Y}, \underline{U})$ than with the likelihood of $\underline{Y}$ itself. the EM-algorithm is based on maximising an approximation of $(\underline{Y}, \underline{U})$ based on the data that is observed $\underline{Y}$.

Let us suppose that the joint likelihood of $(\underline{Y}, \underline{U})$ is

$$\mathcal{L}_n(\underline{Y}, \underline{U};\theta) = \log f(\underline{Y}, \underline{U};\theta).$$

This likelihood is often called the complete likelihood, we will assume that if $\underline{U}$ were known, then this likelihood would be easy to obtain and differentiate. We will assume

that the density $f(\underline{U}|\underline{Y};\theta)$ is also known and is easy to evaluate. By using Bayes theorem it is straightforward to show that

$$\begin{aligned} \log f(\underline{Y},\underline{U};\theta) &= \log f(\underline{Y};\theta) + \log f(\underline{U}|\underline{Y};\theta) \\ \Rightarrow \mathcal{L}_n(\underline{Y},\underline{U};\theta) &= \mathcal{L}_n(\underline{Y};\theta) + \log f(\underline{U}|\underline{Y};\theta). \end{aligned} \tag{7.1}$$

Of course, in reality $\log f(\underline{Y},\underline{U};\theta)$ is unknown, because $\underline{U}$ is unobserved. However, let us consider the expected value of $\log f(\underline{Y},\underline{U};\theta)$ given what we observe $\underline{Y}$. That is

$$Q(\theta_0,\theta) = \mathrm{E}\bigg(\log f(\underline{Y},\underline{U};\theta)\big|\underline{Y},\theta_0\bigg) = \int \bigg(\log f(\underline{Y},\underline{u};\theta)\bigg) f(\underline{u}|\underline{Y},\underline{\theta}_0)d\underline{u}, \tag{7.2}$$

where $f(\underline{u}|\underline{Y},\underline{\theta}_0)$ is the conditional distribution of $\underline{U}$ given $\underline{Y}$ and the unknown parameter θ_0. Hence if $f(\underline{u}|\underline{Y},\underline{\theta}_0)$ were known, then $Q(\theta_0,\theta)$ can be evaluated.

Remark 7.1.1 *It is worth noting that* $Q(\theta_0,\theta) = \mathrm{E}\big(\log f(\underline{Y},\underline{U};\theta)\big|\underline{Y},\theta_0\big)$ *can be viewed as the best predictor of the complete likelihood (involving both observed and unobserved data -* $(\underline{Y},\underline{U})$*) given what is observed* $\underline{Y}$*. We recall that the conditional expectation is the best predictor of* U *in terms of mean squared error, that is the function of* Y *which minimises the mean squared error:* $\mathrm{E}(U|Y) = \arg\min_g \mathrm{E}(U-g(Y))^2$.

The EM algorithm is based on iterating $Q(\cdot)$ in such a way that at each step we obtaining an estimator which gives a larger value of $Q(\cdot)$ (and as we will show later, this gives a larger $\mathcal{L}_n(\underline{Y};\theta)$). We describe the EM-algorithm below.

The EM-algorithm:

(i) Define an initial value $\theta_1 \in \Theta$. Let $\theta_* = \theta_1$.

(ii) **The expectation step (The (k+1)-step)**,

For a fixed θ_* evaluate

$$Q(\theta_*,\theta) = \mathrm{E}\bigg(\log f(\underline{Y},\underline{U};\theta)\bigg|\underline{Y},\theta_*\bigg) = \int \big(\log f(\underline{Y},\underline{u};\theta)\big) f(\underline{u}|\underline{Y},\underline{\theta}_*)du,$$

for all $\theta \in \Theta$.

(iii) **The maximisation step**

Evaluate $\theta_{k+1} = \arg\max_{\theta\in\Theta} Q(\theta_*,\theta)$.

We note that the maximisation can be done by finding the solution of

$$\mathrm{E}\left(\frac{\partial \log f(\underline{Y},\underline{U};\theta)}{\partial\theta}\Big|\underline{Y},\theta_*\right)=0.$$

(iv) If θ_k and θ_{k+1} are sufficiently close to each other stop the algorithm and set $\widehat{\theta}_n=\theta_{k+1}$. Else set $\theta_*=\theta_{k+1}$, go back and repeat steps (ii) and (iii) again.

We use $\hat{\theta}_n$ as an estimator of θ_0. To understand why this iteration is connected to the maximising of $\mathcal{L}_n(\underline{Y};\theta)$ and, under certain conditions, gives a good estimator of θ_0 (in the sense that $\hat{\theta}_n$ is close to the parameter which maximises $\mathcal{L}_n$) let us return to (7.1). Taking the expectation of $\log f(\underline{Y},\underline{U};\theta)$, conditioned on $\underline{Y}$ we have

$$\begin{aligned} Q(\theta_*,\theta) &= \mathrm{E}\left(\log f(\underline{Y},\underline{U};\theta)\big|\underline{Y},\theta_*\right) \\ &= \mathrm{E}\left[\log f(\underline{Y};\theta)+\log f(\underline{U}|\underline{Y};\theta)\big|\underline{Y},\theta_*\right] \\ &= \log f(\underline{Y};\theta)+\mathrm{E}\left[\log f(\underline{U}|\underline{Y};\theta)\big|\underline{Y},\theta_*\right]. \end{aligned} \tag{7.3}$$

Define

$$D(\theta_*,\theta) = \mathrm{E}\left(\log f(\underline{U}|\underline{Y};\theta)\big|\underline{Y},\theta_*\right)=\int\left[\log f(\underline{u}|\underline{Y};\theta)\right]f(\underline{u}|\underline{Y},\theta_*)d\underline{u}.$$

Substituting $D(\theta_*,\theta)$ into (7.3) gives

$$Q(\theta_*,\theta)=\mathcal{L}_n(\theta)+D(\theta_*,\theta), \tag{7.4}$$

we use this in expression in the proof below to show that $\mathcal{L}_n(\theta_{k+1})>\mathcal{L}_n(\theta_k)$. First we that at the $(k+1)$th step iteration of the EM-algorithm, θ_{k+1} maximises $Q(\theta_k,\theta)$ over all $\theta\in\Theta$, hence $Q(\theta_k,\theta_{k+1})\geq Q(\theta_k,\theta_k)$ (which will also be used in the proof).

In the lemma below we show that $\mathcal{L}_n(\theta_{k+1})\geq\mathcal{L}_n(\theta_k)$, hence at each iteration of the EM-algorthm we are obtaining a θ_{k+1} which increases the likelihood over the previous iteration.

Lemma 7.1.1 *When running the EM-algorithm the inequality $\mathcal{L}_n(\theta_{k+1})\geq\mathcal{L}_n(\theta_k)$ always holds.*

Furthermore, if $\theta_k\to\widehat{\theta}$ and for every iteration $\frac{\partial Q(\theta_1,\theta_2)}{\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\theta_k,\theta_{k+1})}=0$, then $\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta=\widehat{\theta}}=0$ (this point can be a saddle point, a local maximum or the sought after global maximum).

PROOF. From (7.4) it is clear that

$$Q(\theta_k,\theta_{k+1})-Q(\theta_k,\theta_k)=\big[\mathcal{L}_n(\theta_{k+1})-\mathcal{L}_n(\theta_k)\big]+\big[D(\theta_k,\theta_{k+1})-D(\theta_k,\theta_k)\big], \qquad (7.5)$$

where we recall

$$D(\theta_1,\theta) \;=\; \mathrm{E}\bigg(\log f(\underline{U}|\underline{Y};\theta)\big|\underline{Y},\theta_1\bigg)=\int\big[\log f(\underline{u}|\underline{Y};\theta)\big]f(\underline{u}|\underline{Y},\theta_1)d\underline{u}.$$

We will show that $\big[D(\theta_k,\theta_{k+1})-D(\theta_k,\theta_k)\big]\leq 0$, the result follows from this. We observe that

$$\big[D(\theta_k,\theta_{k+1})-D(\theta_k,\theta_k)\big]=\int\log\frac{f(\underline{u}|\underline{Y},\theta_{k+1})}{f(\underline{u}|\underline{Y},\theta_k)}f(\underline{u}|\underline{Y},\theta_k)d\underline{u}.$$

By using the Jenson's inequality (which we have used several times previously)

$$\big[D(\theta_k,\theta_{k+1})-D(\theta_k,\theta_k)\big]\leq\log\int f(\underline{u}|\underline{Y},\theta_{k+1})d\underline{u}=0.$$

Therefore, $\big[D(\theta_k,\theta_{k+1})-D(\theta_k,\theta_k)\big]\leq 0$. Note that if θ uniquely identifies the distribution $f(\underline{u}|\underline{Y},\theta)$ then equality only happens when $\theta_{k+1}=\theta_k$. Since $\big[D(\theta_k,\theta_{k+1})-D(\theta_k,\theta_k)\big]\leq 0$ by (7.5) we have

$$\big[\mathcal{L}_n(\theta_{k+1})-\mathcal{L}_n(\theta_k)\big]\geq Q(\theta_k,\theta_{k+1})-Q(\theta_k,\theta_k)\geq 0.$$

and we obtain the desired result ($\mathcal{L}_n(\theta_{k+1})\geq\mathcal{L}_n(\theta_k)$).

To prove the second part of the result we will use that for all $\theta\in\Theta$

$$\frac{\partial D(\theta_1,\theta_2)}{\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)}=\int\frac{\partial\log f(\underline{u}|\underline{Y};\theta)}{\partial\theta}f(\underline{u}|\underline{Y},\theta)d\underline{u}=\frac{\partial}{\partial\theta}\int f(\underline{u}|\underline{Y},\theta)d\underline{u}=0. \quad (7.6)$$

We will to show that the derivative of the likelihood is zero at θ_* i.e. $\frac{\partial\mathcal{L}}{\partial\theta}\rfloor_{\theta=\theta_*}=0$. To show this we use the identity

$$\mathcal{L}_n(\theta_{k+1})=Q(\theta_k,\theta_{k+1})-D(\theta_k,\theta_{k+1}).$$

Taking derivatives with respect to θ_{k+1} gives

$$\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta=\theta_{k+1}}=\frac{\partial Q(\theta_1,\theta_2)}{\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\theta_k,\theta_{k+1})}-\frac{\partial D(\theta_1,\theta_2)}{\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\theta_k,\theta_{k+1})}.$$

By definition of θ_{k+1}, $\frac{\partial Q(\theta_1,\theta_2)}{\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\theta_k,\theta_{k+1})}=0$, thus we have

$$\frac{\partial\mathcal{L}_n(\theta)}{\partial\theta}\rfloor_{\theta=\theta_{k+1}}=-\frac{\partial D(\theta_1,\theta_2)}{\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\theta_k,\theta_{k+1})}.$$

Furthermore, since by assumption $\theta_k \to \widehat{\theta}$ this implies that as $k \to \infty$ we have

$$\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}\rfloor_{\theta=\widehat{\theta}} = -\frac{\partial D(\theta_1,\theta_2)}{\partial \theta_2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})} = 0,$$

which follows from (7.6), thus giving the required result. □

Further information on convergence can be found in Boyles (1983) (`http://www.jstor.org/stable/pdf/2345622.pdf?_=1460485744796`) and Wu (1983) (`https://www.jstor.org/stable/pdf/2240463.pdf?_=1460409579185`).

Remark 7.1.2 *Note that the EM algorithm will converge to a $\widehat{\theta}$ where $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}_{\theta=\widehat{\theta}} = 0$. The reason can be seen from the identity*

$$Q(\theta_*,\theta) = \mathcal{L}_n(\theta) + D(\theta_*,\theta).$$

The derivative of the above with respect to θ is

$$\frac{\partial Q(\theta_*,\theta)}{\partial \theta} = \frac{\partial \mathcal{L}_n(\theta)}{\partial \theta} + \frac{\partial D(\theta_*,\theta)}{\partial \theta}. \tag{7.7}$$

Observe that $D(\theta_*,\theta)$ is maximum only when $\theta = \theta_*$ (for all θ_*, this is clear from the proof above), thus $\frac{\partial D(\theta_*,\theta)}{\partial \theta}_{\theta=\theta_*}$ which for $\theta_* = \widehat{\theta}$ implies $\frac{\partial D(\widehat{\theta},\theta)}{\partial \theta}_{\theta=\widehat{\theta}} = 0$. Furthermore, by definition $\frac{\partial Q(\widehat{\theta},\theta)}{\partial \theta}_{\theta=\widehat{\theta}} = 0$.

Since $\frac{\partial Q(\widehat{\theta},\theta)}{\partial \theta}_{\theta=\widehat{\theta}} = 0$ and $\frac{\partial D(\widehat{\theta},\theta)}{\partial \theta}_{\theta=\widehat{\theta}} = 0$ by using (7.7) this implies $\frac{\partial \mathcal{L}_n(\theta)}{\partial \theta}_{\theta=\widehat{\theta}} = 0$.

In order to prove the results in the following section we use the following identities. Since

$$\begin{aligned}
& Q(\theta_1,\theta_2) = \mathcal{L}(\theta_2) + D(\theta_1,\theta_2) \\
\Rightarrow \quad & \frac{\partial^2 Q(\theta_1,\theta_2)}{\partial \theta_2^2} = \frac{\partial^2 \mathcal{L}(\theta_2)}{\partial \theta^2} + \frac{\partial^2 D(\theta_1,\theta_2)}{\partial \theta^2} \\
\Rightarrow \quad & \frac{\partial^2 Q(\theta_1,\theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)} = \frac{\partial^2 \mathcal{L}(\theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)} + \frac{\partial^2 D(\theta_1,\theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)} \\
\Rightarrow \quad & -\frac{\partial^2 \mathcal{L}(\theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)} = -\frac{\partial^2 Q(\theta_1,\theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)} + \frac{\partial^2 D(\theta_1,\theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)}
\end{aligned} \tag{7.8}$$

We observe that the LHS of the above is the observed Fisher information matrix $I(\theta|\underline{Y})$,

$$-\frac{\partial^2 Q(\theta_1,\theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)} = I_C(\theta|\underline{Y}) = -\int \frac{\partial^2 \log f(\underline{u},\underline{Y};\theta)}{\partial \theta^2} f(\underline{u}|\underline{Y},\theta)d\underline{u} \tag{7.9}$$

is the complete Fisher information *conditioned* on what is observed and

$$-\frac{\partial^2 D(\theta_1,\theta_2)}{\partial\theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\theta,\theta)} = I_M(\theta|\underline{Y}) = -\int \frac{\partial^2 \log f(\underline{u}|\underline{Y};\theta)}{\partial\theta^2} f(\underline{u}|\underline{Y},\theta)d\underline{u} \qquad (7.10)$$

is the Fisher information matrix of the unobserved data conditioned on what is observed. Thus

$$I(\theta|\underline{Y}) = I_C(\theta|\underline{Y}) - I_M(\theta|\underline{Y}).$$

7.1.1 Speed of convergence of θ_k to a stable point

When analyzing an algorithm it is instructive to understand how fast it takes to converge to the limiting point. In the case of the EM-algorithm, this means what factors determine the rate at which θ_k converges to a stable point $\widehat{\theta}$ (note this has *nothing* to do with the rate of convergence of an estimator to the true parameter, and it is important to understand this distinction).

The rate of convergence of an algorithm is usually measured by the ratio of the current iteration with the previous iteration:

$$R = \lim_{k\to\infty}\left(\frac{\theta_{k+1}-\widehat{\theta}}{\theta_k-\widehat{\theta}}\right),$$

if the algorithm converges to a limit in a finite number of iterations we place the above limit to zero. Thus the smaller R the faster the rate of convergence (for example if (i) $\theta_k - \widehat{\theta} = k^{-1}$ then $R = 1$ if (ii) $\theta_k - \widehat{\theta} = \rho^k$ then $R = \rho$, assuming $|\rho| < 1$). Note that since $(\theta_{k+1} - \widehat{\theta}) = \prod_{j=1}^{k}\left(\frac{\theta_{j+1}-\widehat{\theta}}{\theta_j-\widehat{\theta}}\right)$, then typically $|R| \leq 1$.

To obtain an approximation of R we will make a Taylor expansion of $\frac{\partial Q(\theta_1,\theta_2)}{\partial\theta_2}$ around the limit $(\theta_1,\theta_2) = (\widehat{\theta},\widehat{\theta})$. To do this we recall that for a bivariate function $f : \mathbb{R}^2 \to \mathbb{R}$ for (x_0, y_0) "close" to (x, y) we have the Taylor expansion

$$f(x,y) = f(x_0,y_0) + (x-x_0)\frac{\partial f(x,y)}{\partial x}\rfloor_{(x,y)=(x_0,y_0)} + (y-y_0)\frac{\partial f(x,y)}{\partial y}\rfloor_{(x,y)=(x_0,y_0)} + \text{lower order terms.}$$

Applying the above to $\frac{\partial Q(\theta_1,\theta_2)}{\partial\theta_2}$ gives

$$\begin{aligned}
&\frac{\partial Q(\theta_1,\theta_2)}{\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\theta_k,\theta_{k+1})} \\
\approx\ &\frac{\partial Q(\theta_1,\theta_2)}{\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})} + (\theta_{k+1}-\widehat{\theta})\frac{\partial^2 Q(\theta_1,\theta_2)}{\partial\theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})} + (\theta_k-\widehat{\theta})\frac{\partial^2 Q(\theta_1,\theta_2)}{\partial\theta_1\partial\theta_2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})}.
\end{aligned}$$

Since θ_{k+1} maximises $Q(\theta_k, \theta)$ and $\widehat{\theta}$ maximises $Q(\widehat{\theta}, \theta)$ within the interior of the parameter space then

$$\frac{\partial Q(\theta_1, \theta_2)}{\partial \theta_2}\rfloor_{(\theta_1,\theta_2)=(\theta_k,\theta_{k+1})} = 0 \text{ and } \frac{\partial Q(\theta_1, \theta_2)}{\partial \theta_2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})} = 0$$

This implies that

$$(\theta_{k+1} - \widehat{\theta})\frac{\partial^2 Q(\theta_1, \theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})} + (\theta_k - \widehat{\theta})\frac{\partial^2 Q(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})} = 0.$$

Thus

$$\lim_{k\to\infty}\left(\frac{\theta_{k+1} - \widehat{\theta}}{\theta_k - \widehat{\theta}}\right) = -\left(\frac{\partial^2 Q(\theta_1, \theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})}\right)^{-1}\frac{\partial^2 Q(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})}. \tag{7.11}$$

This result shows that the rate of convergence depends on the ratio of gradients of $Q(\theta_1, \theta_2)$ around $(\theta_1, \theta_2) = (\widehat{\theta}, \widehat{\theta})$. Some further simplifications can be made by noting that

$$Q(\theta_1, \theta_2) \;=\; \mathcal{L}_n(\theta_2) + D(\theta_1, \theta_2) \quad \Rightarrow \quad \frac{\partial^2 Q(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2} = \frac{\partial^2 D(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2}.$$

Substituting this into (7.11) gives

$$\lim_{k\to\infty}\left(\frac{\theta_{k+1} - \widehat{\theta}}{\theta_k - \widehat{\theta}}\right) = -\left(\frac{\partial^2 Q(\theta_1, \theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})}\right)^{-1}\frac{\partial^2 D(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})}. \tag{7.12}$$

To make one further simplification, we note that

$$\begin{aligned}
\frac{\partial^2 D(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})} &= \int \frac{1}{f(\underline{u}|\underline{Y}, \theta_2)}\frac{\partial f(\underline{u}|\underline{Y}, \theta_2)}{\partial \theta_2}\frac{\partial f(\underline{u}|\underline{Y}, \theta_1)}{\partial \theta_1}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})} d\underline{u} \\
&= \int \frac{1}{f(\underline{u}|\underline{Y}, \theta)}\left(\frac{\partial f(\underline{u}|\underline{Y}, \theta)}{\partial \theta}\right)^2\rfloor_{\theta=\widehat{\theta}} d\underline{u} \\
&= -\int \frac{\partial^2 \log f(\underline{u}|\underline{Y}, \theta)}{\partial \theta_2} f(\underline{u}|\underline{Y}, \theta)\rfloor_{\theta=\widehat{\theta}} d\underline{u}
\end{aligned} \tag{7.13}$$

were the last line of the above follows from the identity

$$\int \frac{1}{f(\underline{x}; \theta)}\left(\frac{\partial f(\underline{x}; \theta)}{\partial \theta}\right)^2 d\underline{x} + \int \frac{\partial^2 \log f(\underline{x}; \theta)}{\partial \theta_2} f(\underline{x}; \theta) d\underline{x} = 0$$

(see the proof of Corollary 1.3.1). Substituting (7.12) into (7.13) gives

$$\lim_{k\to\infty}\left(\frac{\theta_{k+1} - \widehat{\theta}}{\theta_k - \widehat{\theta}}\right) = \left(\frac{\partial^2 Q(\theta_1, \theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})}\right)^{-1}\frac{\partial^2 D(\theta_1, \theta_2)}{\partial \theta_2^2}\rfloor_{(\theta_1,\theta_2)=(\widehat{\theta},\widehat{\theta})}. \tag{7.14}$$

Substituting (7.9) and (7.10) into the above gives

$$\lim_{k\to\infty}\left(\frac{\theta_{k+1}-\widehat{\theta}}{\theta_k-\widehat{\theta}}\right)=I_C(\theta|\underline{Y})^{-1}I_M(\theta|\underline{Y}) \tag{7.15}$$

Hence the rate of convergence of the algorithm depends on on the ratio $I_C(\theta|\underline{Y})^{-1}I_M(\theta|\underline{Y})$. The closer the largest eigenvalue of $I_C(\theta|\underline{Y})^{-1}I_M(\theta|\underline{Y})$ to one, the slower the rate of convergence, and a larger number of iterations are required. The heuristic of this result is that if the missing information is a large proportion of the complete or total information than this ratio will be large.

Further details can be found in Dempster et. al. (1977) pages 9-10 and Meng and Rubin (1994) (`http://www.sciencedirect.com/science/article/pii/0024379594903638`).

7.2 Applications of the EM algorithm

7.2.1 Censored data

Let us return to the example at the start of this section, and construct the EM-algorithm for censored data. We recall that the log-likelihoods for censored data and complete data are

$$\mathcal{L}_n(\underline{Y};\theta)=\Big(\sum_{i=1}^{n}\log f(Y_i;\theta)\Big)+\Big(\sum_{i=n+1}^{n+m}\log\mathcal{F}(Y_i;\theta)\Big).$$

and

$$\mathcal{L}_n(\underline{Y},\underline{U};\theta)=\Big(\sum_{i=1}^{n}\log f(Y_i;\theta)\Big)+\Big(\sum_{i=n+1}^{n+m}\log f(T_i;\theta)\Big).$$

To implement the EM-algorithm we need to evaluate the expectation step $Q(\theta_*,\theta)$. It is easy to see that

$$Q(\theta_*,\theta)=\mathrm{E}\Big(\mathcal{L}_n(\underline{Y},\underline{U};\theta)\big|\underline{Y},\theta_*\Big) \;=\; \Big(\sum_{i=1}^{n}\log f(Y_i;\theta)\Big)+\Big(\sum_{i=n+1}^{n+m}\mathrm{E}\big(\log f(T_i;\theta)|\underline{Y},\theta_*\big)\Big).$$

To obtain $\mathrm{E}\big(\log f(T_i;\theta)|\underline{Y},\theta_*\big)$ $(i\geq n+1)$ we note that

$$\begin{aligned}\mathrm{E}\big(\log f(T_i;\theta)|\underline{Y},\theta_*\big) &= \mathrm{E}(\log f(T_i;\theta)|T_i\geq c)\\ &= \frac{1}{\mathcal{F}(c;\theta)}\int_c^{\infty}\big[\log f(T_i;\theta)\big]f(u;\theta_*)du.\end{aligned}$$

Therefore we have

$$Q(\theta_*,\theta) = \Big(\sum_{i=1}^{n}\log f(Y_i;\theta)\Big) + \frac{m}{\mathcal{F}(c;\theta_*)}\int_c^\infty \big[\log f(T_i;\theta)\big] f(u;\theta_*)du.$$

We also note that the derivative of $Q(\theta_*,\theta)$ with respect to θ is

$$\frac{\partial Q(\theta_*,\theta)}{\partial\theta} = \Big(\sum_{i=1}^{n}\frac{1}{f(Y_i;\theta)}\frac{\partial f(Y_i;\theta)}{\partial\theta}\Big) + \frac{m}{\mathcal{F}(c;\theta_*)}\int_c^\infty \frac{1}{f(u;\theta)}\frac{\partial f(u;\theta)}{\partial\theta} f(u;\theta_*)du.$$

Hence for this example, the EM-algorithm is

(i) Define an initial value $\theta_1 \in \Theta$. Let $\theta_* = \theta_1$.

(ii) **The expectation step**:

For a fixed θ_* evaluate

$$\frac{\partial Q(\theta_*,\theta)}{\partial\theta} = \Big(\sum_{i=1}^{n}\frac{1}{f(Y_i;\theta)}\frac{\partial f(Y_i;\theta)}{\partial\theta}\Big) + \frac{m}{\mathcal{F}(c;\theta_*)}\int_c^\infty \frac{1}{f(u;\theta)}\frac{\partial f(u;\theta)}{\partial\theta} f(u;\theta_*)du.$$

(iii) **The maximisation step**:

Solve for $\frac{\partial Q(\theta_*,\theta)}{\partial\theta}$. Let θ_{k+1} be such that $\frac{\partial Q(\theta_*,\theta)}{\partial\theta}\rfloor_{\theta=\theta_k} = 0$.

(iv) If θ_k and θ_{k+1} are sufficiently close to each other stop the algorithm and set $\hat{\theta}_n = \theta_{k+1}$. Else set $\theta_* = \theta_{k+1}$, go back and repeat steps (ii) and (iii) again.

7.2.2 Mixture distributions

We now consider a useful application of the EM-algorithm, to the estimation of parameters in mixture distributions. Let us suppose that $\{Y_i\}_{i=1}^n$ are iid random variables with density

$$f(y;\theta) = pf_1(y;\theta_1) + (1-p)f_2(y;\theta_2),$$

where $\theta = (p,\theta_1,\theta_2)$ are unknown parameters. For the purpose of identifiability we will suppose that $\theta_1 \neq \theta_2$, $p \neq 1$ and $p \neq 0$. The log-likelihood of $\{Y_i\}$ is

$$\mathcal{L}_n(\underline{Y};\theta) = \sum_{i=1}^{n}\log\big(pf_1(Y_i;\theta_1) + (1-p)f_2(Y_i;\theta_2)\big). \tag{7.16}$$

Now maximising the above can be extremely difficult. As an illustration consider the example below.

Example 7.2.1 *Let us suppose that $f_1(y;\theta_1)$ and $f_2(y;\theta_1)$ are normal densities, then the log likelihood is*

$$\mathcal{L}_n(\underline{Y};\theta) = \sum_{i=1}^{n} \log\left(p\frac{1}{\sqrt{2\pi\sigma_1^2}}\exp(-\frac{1}{2\sigma_1^2}(Y_i-\mu_1)^2) + (1-p)\frac{1}{\sqrt{2\pi\sigma_2^2}}\exp(-\frac{1}{2\sigma_2^2}(Y_i-\mu_2)^2)\right).$$

We observe this is extremely difficult to maximise. On the other hand if Y_i were simply normally distributed then the log-likelihood is extremely simple

$$\mathcal{L}_n(\underline{Y};\theta) \propto -\sum_{i=1}^{n}\left(\log\sigma_1^2 + \frac{1}{2\sigma_1^2}(Y_i-\mu_1)^2)\right). \tag{7.17}$$

In other words, the simplicity of maximising the log-likelihood of the exponential family of distributions (see Section 1.6) is lost for mixtures of distributions.

We use the EM-algorithm as an indirect but simple method of maximising (7.17). In this example, it is not clear what observations are missing. However, let us consider one possible intepretation of the mixture distribution. Let us define the random variables δ_i and Y_i, where $\delta_i \in \{1,2\}$,

$$P(\delta_i = 1) = p \text{ and } P(\delta_i = 2) = (1-p)$$

and the density of $Y_i|\delta_i = 1$ is f_1 and the density of $Y_i|\delta_i = 1$ is f_2. Based on this definition, it is clear from the above that the density of Y_i is

$$f(y;\theta) = f(y|\delta=1,\theta)P(\delta=1) + f(y|\delta=2,\theta)P(\delta=2) = pf_1(y;\theta_1) + (1-p)f_2(y;\theta_2).$$

Hence, one interpretation of the mixture model is that there is a hidden unobserved random variable which determines the state or distribution of Y_i. A simple example, is that Y_i is the height of an individual and δ_i is the gender. However, δ_i is unobserved and only the height is observed. Often a mixture distribution has a physical interpretation, similar to the height example, but sometimes it can be used to parametrically model a wide class of densities.

Based on the discussion above, $\underline{U} = \{\delta_i\}$ can be treated as the missing observations. The likelihood of (Y_i, U_i) is

$$\big\{p_1 f_1(Y_i;\theta_1)\big\}^{I(\delta_i=1)}\big\{p_2 f_2(Y_i;\theta_2)\big\}^{I(\delta_i=2)} = p_{\delta_i} f_{\delta_i}(Y_i;\theta_{\delta_i}).$$

where we set $p_2 = 1 - p$. Therefore the log likelihood of $\{(Y_i, \delta_i)\}$ is

$$\mathcal{L}_n(\underline{Y}, \underline{U}; \theta) = \sum_{i=1}^{n} \left(\log p_{\delta_i} + \log f_{\delta_i}(Y_i; \theta_{\delta_i})\right).$$

We now need to evaluate

$$Q(\theta_*, \theta) = \mathrm{E}\big(\mathcal{L}_n(\underline{Y}, \underline{U}; \theta)|\underline{Y}, \theta_*\big) = \sum_{i=1}^{n} \left[\mathrm{E}\big(\log p_{\delta_i}|Y_i, \theta_*\big) + \mathrm{E}\big(\log f_{\delta_i}(Y_i; \theta_{\delta_i})|Y_i, \theta_*\big)\right].$$

We see that the above expectation is taken with respect the distribution of δ_i conditioned on Y_i and the parameter θ_* Thus, in general,

$$\mathrm{E}\left(A(Y, \delta)|Y, \theta^*\right) = \sum_j A(Y, \delta = j) P(\delta = j|Y_i, \theta^*),$$

which we apply to $Q(\theta_*, \theta)$ to give

$$Q(\theta_*, \theta) = \sum_j \sum_{i=1}^{n} \left[\log p_{\delta_i = j} + \log f_{\delta_i = j}(Y_i; \theta)\right] P(\delta_i = j|Y_i, \theta^*).$$

Therefore we need to obtain $P(\delta_i = j|Y_i, \theta^*)$. By using conditioning arguments it is easy to see that [1]

$$\begin{aligned}
P(\delta_i = 1|Y_i = y, \theta_*) &= \frac{P(\delta_i = 1, Y_i = y; \theta_*)}{P(Y_i = y; \theta_*)} = \frac{p_* f_1(y, \theta_{1,*})}{p_* f_1(y, \theta_{1,*}) + (1 - p_*) f_2(y, \theta_{2,*})} \\
&:= w_1(\theta_*, y) \\
P(\delta_i = 2|Y_i = y, \theta_*) &= \frac{p_* f_2(y, \theta_{2,*})}{p_* f_1(y, \theta_{1,*}) + (1 - p_*) f_2(y, \theta_{2,*})} \\
&:= w_2(\theta_*, y) = 1 - w_1(\theta_*, y).
\end{aligned}$$

Therefore

$$Q(\theta_*, \theta) = \sum_{i=1}^{n} \Big(\log p + \log f_1(Y_i; \theta_1)\Big) w_1(\theta_*, Y_i) + \sum_{i=1}^{n} \Big(\log(1 - p) + \log f_2(Y_i; \theta_2)\Big) w_2(\theta_*, Y_i).$$

Now maximising the above with respect to p, θ_1 and θ_2 in general will be much easier than maximising $\mathcal{L}_n(\underline{Y}; \theta)$. For this example the EM algorithm is

(i) Define an initial value $\theta_1 \in \Theta$. Let $\theta_* = \theta_1$.

[1] To see why note that $P(\delta_i = 1 \text{ and } Y_i \in [y - h/2, y + h/2]|\theta^*) = h p_* f_1(y)$ and $P(Y_i \in [y - h/2, y + h/2]|\theta^*) = h\left(p_* f_1(y) + (1 - p_*) f_2(y)\right)$.

(ii) **The expectation step**:

For a fixed θ_* evaluate

$$Q(\theta_*,\theta) = \sum_{i=1}^{n}\Big(\log p + \log f_1(Y_i;\theta_1)\Big)w_1(\theta_*,Y_i) + \sum_{i=1}^{n}\Big(\log(1-p) + \log f_2(Y_i;\theta_2)\Big)w_2(\theta_*,Y_i).$$

(iii) **The maximisation step**:

Evaluate $\theta_{k+1} = \arg\max_{\theta\in\Theta} Q(\theta_*,\theta)$ by differentiating $Q(\theta_*,\theta)$ wrt to θ and equating to zero. Since the parameters p and θ_1,θ_2 are in separate subfunctions, they can be maximised separately.

(iv) If θ_k and θ_{k+1} are sufficiently close to each other stop the algorithm and set $\hat{\theta}_n = \theta_{k+1}$. Else set $\theta_* = \theta_{k+1}$, go back and repeat steps (ii) and (iii) again.

Example 7.2.2 (Normal mixtures and mixtures from the exponential family) *(i)* *We briefly outline the algorithm in the case of a mixture two normal distributions. In this case*

$$Q(\theta_*,\theta) = -\frac{1}{2}\sum_{j=1}^{2}\sum_{i=1}^{n} w_j(\theta_*,Y_i)\left(\sigma_j^{-2}(Y_i-\mu_j)^2 + \log\sigma_j^2\right) + \sum_{i=1}^{n} w_j(\theta_*,Y_i)\left(\log p + \log(1-p)\right).$$

By differentiating the above wrt to μ_j, σ_j^2 (for $j=1$ and 2) and p it is straightforward to see that the μ_j, σ_j^2 and p which maximises the above is

$$\widehat{\mu}_j = \frac{\sum_{i=1}^{n} w_j(\theta_*,Y_i)Y_i}{\sum_{i=1}^{n} w_j(\theta_*,Y_i)} \text{ and } \widehat{\sigma}_j^2 = \frac{\sum_{i=1}^{n} w_j(\theta_*,Y_i)(Y_i-\widehat{\mu}_j)^2}{\sum_{i=1}^{n} w_j(\theta_*,Y_i)}$$

and

$$\widehat{p} = \frac{\sum_{i=1}^{n} w_1(\theta_*,Y_i)}{n}.$$

Once these estimators are obtained we let $\theta_* = (\widehat{\mu}_1,\widehat{\mu}_2,\widehat{\sigma}_1^2,\widehat{\sigma}_2^2,\widehat{p})$. The quantities $w_j(\theta_*,Y_i)$ are re-evaluated and $Q(\theta_*,\theta)$ maximised with respect to the new weights.

(ii) In general if Y is a mixture from the exponential family with density

$$f(y;\theta) = \sum_{j=1}^{m} p_j \exp\left(y\theta_j - \kappa_j(\theta_j) + c_j(y)\right)$$

the corresponding $Q(\theta_*, \theta)$ *is*

$$Q(\theta_*, \theta) = \sum_{j=1}^{m} \sum_{i=1}^{n} w_j(\theta_*, Y_i) \left[Y_i\theta_j - \kappa_j(\theta_j) + c_j(Y_i) + \log p_j \right],$$

where

$$w_j(\theta_*, Y_i) = \frac{p_j^* \exp\left(Y_i\theta_j^* - \kappa_j(\theta_j^*) + c_j(Y_i)\right)}{\sum_{k=1}^{m} p_k^* \exp\left(Y_i\theta_k^* - \kappa_k(\theta_k^*) + c_k(Y_i)\right)}$$

subject to the constraint that $\sum_{j=1}^{m} p_j = 1$. *Thus for* $1 \leq j \leq m$, $Q(\theta_*, \theta)$ *is maximised for*

$$\widehat{\theta}_j = \mu_j^{-1} \left(\frac{\sum_{i=1}^{n} w_j(\theta_*, Y_i) Y_i}{\sum_{i=1}^{n} w_j(\theta_*, Y_i)} \right)$$

where $\mu_j = \kappa_j'$ *(we assume all parameter for each exponential mixture is open) and*

$$\widehat{p}_j = \frac{\sum_{i=1}^{n} w_j(\theta_*, Y_i)}{n}.$$

Thus we set $\theta_* = (\{\widehat{\theta}_j, \widehat{p}_j\}_{j=1}^{m})$ *and re-evaluate the weights.*

Remark 7.2.1 *Once the algorithm is terminated, we can calculate the chance that any given observation* Y_i *is in subpopulation* j *since*

$$\widehat{P}(\delta_i = j | Y_i) = \frac{\widehat{p}_j f_j(Y; \widehat{\theta})}{\sum_{j=1}^{2} \widehat{p}_j f_j(Y; \widehat{\theta})}.$$

This allows us to obtain a classifier for each observation Y_i.

It is straightforward to see that the arguments above can be generalised to the case that the density of Y_i is a mixture of m different densities. However, we observe that the selection of m can be quite adhoc. There are methods for choosing m, these include the reversible jump MCMC methods.

7.2.3 Problems

Example 7.2.3 *Question: Suppose that the regressors* x_t *are believed to influence the response variable* Y_t. *The distribution of* Y_t *is*

$$P(Y_t = y) = p\frac{\lambda_{t1}^y \exp(-\lambda_{t1} y)}{y!} + (1-p)\frac{\lambda_{t2}^y \exp(-\lambda_{t2} y)}{y!},$$

where $\lambda_{t1} = \exp(\beta_1' x_t)$ *and* $\lambda_{t2} = \exp(\beta_2' x_t)$.

(i) State minimum conditions on the parameters, for the above model to be identifiable?

(ii) Carefully explain (giving details of $Q(\theta^*, \theta)$ and the EM stages) how the EM-algorithm can be used to obtain estimators of β_1, β_2 and p.

(iii) Derive the derivative of $Q(\theta^*, \theta)$, and explain how the derivative may be useful in the maximisation stage of the EM-algorithm.

(iv) Given an initial value, will the EM-algorithm always find the maximum of the likelihood?

Explain how one can check whether the parameter which maximises the EM-algorithm, maximises the likelihood.

Solution

(i) $0 < p < 1$ and $\beta_1 \neq \beta_2$ (these are minimum assumptions, there could be more which is hard to account for given the regressors x_t).

(ii) We first observe that $P(Y_t = y)$ is a mixture of two Poisson distributions where each has the canonical link function. Define the unobserved variables, $\{U_t\}$, which are iid and where $P(U_t = 1) = p$ and $P(U_t = 2) = (1-p)$ and $P(Y = y|U_i = 1) = \frac{\lambda_{t1}^y \exp(-\lambda_{t1} y)}{y!}$ and $P(Y = y)|U_i = 2) = \frac{\lambda_{t2}^y \exp(-\lambda_{t2} y)}{y!}$. Therefore, we have

$$\log f(Y_t, U_t, \theta) = \bigg(Y_t \beta'_{u_t} x_t - \exp(\beta'_{u_t} x_t) + \log Y_t! + \log p \bigg),$$

where $\theta = (\beta_1, \beta_2, p)$. Thus, $E(\log f(Y_t, U_t, \theta)|Y_t, \theta_*)$ is

$$\begin{aligned} E(\log f(Y_t, U_t, \theta)|Y_t, \theta_*) &= \bigg(Y_t \beta'_1 x_t - \exp(\beta'_1 x_t) + \log Y_t! + \log p \bigg) \pi(\theta_*, Y_t) \\ &\quad + \bigg(Y_t \beta'_2 x_t - \exp(\beta'_2 x_t) + \log Y_t! + \log p \bigg) (1 - \pi(\theta_*, Y_t)). \end{aligned}$$

where $P(U_i|Y_t, \theta^*)$ is evaluated as

$$P(U_i = 1|Y_t, \theta^*) = \pi(\theta_*, Y_t) = \frac{p f_1(Y_t, \theta_*)}{p f_1(Y_t, \theta_*) + (1-p) f_2(Y_t, \theta_*)},$$

with

$$f_1(Y_t, \theta_*) = \frac{\exp(\beta'_{*1} x_t Y_t) \exp(-Y_t \exp(\beta'_{*1} x_t))}{Y_t!} \quad f_1(Y_t, \theta_*) = \frac{\exp(\beta'_{*1} x_t Y_t) \exp(-Y_t \exp(\beta'_{*1} x_t)}{Y_t!}.$$

Thus $Q(\theta_*, \theta)$ is

$$\begin{aligned} Q(\theta_*, \theta) &= \sum_{t=1}^{T} \Big(Y_t \beta_1' x_t - \exp(\beta_1' x_t) + \log Y_t! + \log p \Big) \pi(\theta_*, Y_t) \\ &\quad + \Big(Y_t \beta_2' x_t - \exp(\beta_2' x_t) + \log Y_t! + \log(1-p) \Big) (1 - \pi(\theta_*, Y_t)). \end{aligned}$$

Using the above, the EM algorithm is the following:

(a) Start with an initial value which is an estimator of β_1, β_2 and p, denote this as θ_*.

(b) For every θ evaluate $Q(\theta_*, \theta)$.

(c) Evaluate $\arg\max_\theta Q(\theta_*, \theta)$. Denote the maximum as θ_* and return to step (b).

(d) Keep iterating until the maximums are sufficiently close.

(iii) The derivative of $Q(\theta_*, \theta)$ is

$$\begin{aligned} \frac{\partial Q(\theta_*, \theta)}{\partial \beta_1} &= \sum_{t=1}^{T} \Big(Y_t - \exp(\beta_1' x_t) \Big) x_t \pi(\theta_*, Y_t) \\ \frac{\partial Q(\theta_*, \theta)}{\partial \beta_2} &= \sum_{t=1}^{T} \Big(Y_t - \exp(\beta_2' x_t) \Big) x_t (1 - \pi(\theta_*, Y_t)) \\ \frac{\partial Q(\theta_*, \theta)}{\partial p} &= \sum_{t=1}^{T} \Big(\frac{1}{p} \pi(\theta_*, Y_t) - \frac{1}{1-p} (1 - \pi(\theta_*, Y_t) \Big). \end{aligned}$$

Thus maximisation of $Q(\theta_*, \theta)$ can be achieved by solving for the above equations using iterative weighted least squares.

(iv) Depending on the initial value, the EM-algorithm may only locate a local maximum.

To check whether we have found the global maximum, we can start the EM-algorithm with several different initial values and check where they converge.

Example 7.2.4 *Question*

(2) Let us suppose that $\mathcal{F}_1(t)$ and $\mathcal{F}_2(t)$ are two survival functions. Let x denote a univariate regressor.

(i) Show that $\mathcal{F}(t; x) = p\mathcal{F}_1(t)^{\exp(\beta_1 x)} + (1-p)\mathcal{F}_2(t)^{\exp(\beta_2 x)}$ is a valid survival function and obtain the corresponding density function.

(ii) Suppose that T_i are survival times and x_i is a univariate regressor which exerts an influence an T_i. Let $Y_i = \min(T_i, c)$, where c is a common censoring time. $\{T_i\}$ are independent random variables with survival function $\mathcal{F}(t; x_i) = p\mathcal{F}_1(t)^{\exp(\beta_1 x_i)} + (1-p)\mathcal{F}_2(t)^{\exp(\beta_2 x_i)}$, where both $\mathcal{F}_1$ and $\mathcal{F}_2$ are known, but p, β_1 and β_2 are unknown.

State the censored likelihood and show that the EM-algorithm together with iterative least squares in the maximisation step can be used to maximise this likelihood (sufficient details need to be given such that your algorithm can be easily coded).

Solution

i) Since $\mathcal{F}_1$ and $\mathcal{F}_2$ are monotonically decreasing positive functions where $\mathcal{F}_1(0) = \mathcal{F}_2(0) = 1$ and $\mathcal{F}_1(\infty) = \mathcal{F}_2(\infty) = 0$, then it immediately follows that

$$\mathcal{F}(t, x) = p\mathcal{F}_1(t)^{e^{\beta_1 x}} + (1-p)\mathcal{F}_2(t)^{e^{\beta_2 x}}$$

satisfies the same conditions. To obtain the density we differential wrt x

$$\begin{aligned} \frac{\partial \mathcal{F}(t,x)}{\partial t} &= -pe^{\beta_1 x} f_1(t)\mathcal{F}_1(t)^{e^{\beta_1 x}-1} - (1-p)e^{\beta_2 x} f_2(t)\mathcal{F}_2(t)^{e^{\beta_2 x}-1} \\ \Rightarrow f(t;x) &= pe^{\beta_1 x} f_1(t)\mathcal{F}_1(t)^{e^{\beta_1 x-1}} + (1-p)e^{\beta_2 x} f_2(t)\mathcal{F}_2(t)^{e^{\beta_2 x-1}}, \end{aligned}$$

where we use that $\frac{dF_(t)}{dt} = -f_(t)$.

ii) The censored log likelihood is

$$\mathcal{L}_n(\beta_1, \beta_2, p) = \sum_{i=1}^{n} [\delta_i \log f(Y_i; \beta_1, \beta_2, p) + (1-\delta_i) \log \mathcal{F}(Y_i; \beta_1, \beta_2, p)].$$

Clearly, directly maximizing the above is extremely difficult. Thus we look for an alternative method via the EM algorithm.

We first define the indicator variable (which corresponds to the missing variables) which denotes the state 1 or 2

$$I_i = \begin{cases} 1 & \text{with } P(I_i = 1) = p = p_1 \\ 2 & \text{with } P(I_i = 2) = (1-p) = p_2. \end{cases}$$

Then the joint density of (Y_i, δ_i, I_i) is

$$p_{I_i}\left(e^{\beta_{I_i}x_i}f_{I_i}(t)\mathcal{F}_{I_i}(t)^{e^{\beta_{I_i}x_i}-1}\right)\left(\mathcal{F}_{I_i}(t)^{e^{\beta_{I_i}x_i}}\right)^{1-\delta_i}$$

which gives the log-density

$$\delta_i\left\{\log p_{I_i}+\beta_{I_i}x_i+\log f_{I_i}(Y_i)+(e^{\beta_{I_i}x_i}-1)\log F_{I_i}(Y_i)\right\}+(1-\delta_i)\left\{\log p_{I_i}+(e^{\beta_{I_i}x_i})\log F_{I_i}(Y_i)\right\}.$$

Thus the complete log likelihood of (Y_i, δ_i, I_i) is

$$\begin{aligned}\mathcal{L}_n(\underline{Y},\underline{\delta},I_i;\beta_1,\beta_2,p)=\sum_{i=1}^{n}\{\delta_i[&\log p_{I_i}+\beta_{I_i}x_i+\log f_{I_i}(Y_i)+(e^{\beta_{I_i}x_i}-1)\log\mathcal{F}_{I_i}(Y_i)\\&+(1-\delta_i)[\log p_{I_i}+(e^{\beta_{I_i}x_i})\log\mathcal{F}_{I_i}(Y_i)]\}\end{aligned}$$

Next we need to calculate $P(I_i = 1|Y_i, \delta_i, \theta^*)$ and $P(I_i = 2|Y_i, \delta_i, \theta^*)$;

$$\begin{aligned}\omega_i^{\delta_i=1}(1) &= P(I_i=1|Y_i,\delta_i=1,p^*,\beta_1^*,\beta_2^*)\\ &= \frac{p^*e^{\beta_1^*x_i}f_1(Y_i)\mathcal{F}_1(Y_i)^{e^{\beta_1^*x_i}-1}}{p^*e^{\beta_1^*x_i}f_1(Y_i)\mathcal{F}_1(Y_i)^{e^{\beta_1^*x_i}-1}+(1-p^*)e^{\beta_2^*x_i}f_2(Y_i)\mathcal{F}_2(Y_i)^{e^{\beta_2^*x_i}-1}}\\ \omega_i^{\delta_i=0}(1) &= P(I_i=1|Y_i,\delta_i=0,p^*,\beta_1^*,\beta_2^*)\\ &= \frac{p^*\mathcal{F}_1(Y_i)^{e^{\beta_1^*x_i}}}{p^*\mathcal{F}_1(Y_i)^{e^{\beta_1^*x_i}}+(1-p^*)\mathcal{F}_2(Y_i)^{e^{\beta_2^*x_i}}}\end{aligned}$$

and $\omega_i^{\delta_i=1}(2) = 1-\omega_i^{\delta_i=1}(1)$ and $\omega_i^{\delta_i=0}(2) = 1-\omega_i^{\delta_i=0}(1)$. Let $p_1 = p$ and $p_2 = 1-p$. Therefore the complete likelihood conditioned on what we observe is

$$\begin{aligned}Q(\theta_*,\theta) &= \sum_{s=1}^{2}\sum_{i=1}^{n}\{\delta_i\omega_i^{\delta_i=1}(s)[\log p_s+\beta_1x_i+\log f_s(Y_i)+(e^{\beta_sx_i}-1)\log\mathcal{F}_s(Y_i)]\\ &\quad+(1-\delta_i)\omega_i^{\delta_i=0}(s)\left[\log p_s+e^{\beta_sx_i}\log\mathcal{F}_s(Y_i)\right]\}\\ &= \sum_{s=1}^{2}\sum_{i=1}^{n}\Big\{\{\delta_i\omega_i^{\delta_i=1}(s)[\beta_1x_i+\log f_s(Y_i)+(e^{\beta_sx_i}-1)\log\mathcal{F}_s(Y_i)]\\ &\quad+e^{\beta_sx_i}\log\mathcal{F}_s(Y_i)\Big\}\\ &\quad+\sum_{s=1}^{2}\sum_{i=1}^{n}\left\{\delta_i\omega_i^{\delta_i=1}(s)\log p_s+(1-\delta_i)\omega_i^{\delta_i=0}(s)\log p_s\right\}\\ &= Q(\theta_*,\beta_1)+Q(\theta_*,\beta_2)+Q(\theta_*,p_1,p_2)\end{aligned}$$

The conditional likelihood, above, looks unwieldy. However, the parameter estimators can to be separated. First, differentiating with respect to p gives

$$\begin{aligned}\frac{\partial Q}{\partial p} &= \frac{\partial Q(\theta_*, p, 1-p)}{\partial p} \\ &= \sum_{i=1}^{n} \delta_i \omega_i^{\delta_i=1}(1)\frac{1}{p} + \sum_{i=1}^{n} \omega_i^{\delta_i=0}(1)(1-\delta_i)\frac{1}{p} - \\ &\quad \sum_{i=1}^{n} \delta_i \omega_i^{\delta_i=1}(2)\frac{1}{1-p} - \sum_{i=1}^{n} \omega_i^{\delta_i=0}(2)(1-\delta_i)\frac{1}{1-p}.\end{aligned}$$

Equating the above to zero we have the estimator $\hat{p} = \frac{a}{a+b}$, where

$$\begin{aligned}a &= \sum_{i=1}^{n} \delta_i \omega_i^{\delta_i=1}(1) + \sum_{i=1}^{n} \omega_i^{\delta_i=0}(1)(1-\delta_i) \\ b &= \sum_{i=1}^{n} \delta_i \omega_i^{\delta_i=1}(2) + \sum_{i=1}^{n} \omega_i^{\delta_i=0}(2)(1-\delta_i).\end{aligned}$$

Next we consider the estimates of β_1 and β_2 at the ith iteration step. Differentiating Q wrt to β_1 and β_2 gives for $s = 1, 2$

$$\begin{aligned}\frac{\partial Q}{\partial \beta_s} &= \frac{\partial Q_s(\theta_*, \beta_s)}{\partial \beta_s} \\ &= \sum_{i=1}^{n} \left\{ \delta_i \omega_i^{\delta_i=1}(s) \left[1 + e^{\beta_s x_i} \log F_s(Y_i)\right] + (1-\delta_i)\omega_i^{\delta_i=0}(s) e^{\beta_s x_i} \log F_s(Y_i) \right\} x_i \\ \frac{\partial^2 Q(\theta_*, \theta)}{\partial \beta_s^2} &= \frac{\partial^2 Q_s(\theta^*, \beta_s)}{\partial \beta_s} \\ &= \sum_{i=1}^{n} \left\{ \delta_i \omega_i^{\delta_i=1}(s) e^{\beta_s x_i} \log F_s(Y_i) + (1-\delta_i)\omega_i^{\delta_i=0}(s) e^{\beta_s x_i} \log F_s(Y_i) \right\} x_i^2 \\ \frac{\partial^2 Q(\theta_*; \theta)}{\partial \beta_1 \partial \beta_2} &= 0.\end{aligned}$$

Observe that setting the first derivative to zero, we cannot obtain an explicit expression for the estimators at each iteration. Thus we need to use the Newton-Rapshon scheme but in a very simply set-up. To estimate (β_1, β_2) at the jth iteration we use

$$\begin{bmatrix} \beta_1^{(j)} \\ \beta_2^{(j)} \end{bmatrix} = \begin{bmatrix} \beta_1^{(j-1)} \\ \beta_2^{(j-1)} \end{bmatrix} + \begin{bmatrix} \frac{\partial^2 Q}{\partial \beta_1^2} & 0 \\ 0 & \frac{\partial^2 Q}{\partial \beta_1^2} \end{bmatrix}^{-1}_{\underline{\beta}^{(j-1)}} \begin{bmatrix} \frac{\partial Q}{\partial \beta_1} \\ \frac{\partial Q}{\partial \beta_2} \end{bmatrix}_{\underline{\beta}^{(j-1)}}$$

Thus for $s = 1, 2$ we have $\beta_s^{(j)} = \beta_s^{(j-1)} + \left(\frac{\partial^2 Q}{\partial \beta_s^2}\right)^{-1} \frac{\partial Q}{\partial \beta_s} \rfloor_{\underline{\beta}^{(j-1)}}$.

We can rewrite the above Newton Raphson scheme as something that resembles weighted least squares. We recall the weighted least squares estimator are the parameters α which minimise the weighted least squares criterion

$$\sum_{i=1}^{n} W_{ii} \left(Y_i - \underline{x}_i' \underline{\alpha}\right)^2.$$

The α which minimises the above is

$$\widehat{\underline{\alpha}} = (\underline{X}' W \underline{X})^{-1} \underline{X}' W \underline{Y}.$$

The Newtons-Raphson scheme can be written as

$$\begin{aligned} \beta_s^{(j)} &= \beta_s^{(j-1)} - \left(\frac{\partial Q^2}{\partial \beta_s^2}\right)^{-1} \frac{\partial Q}{\partial \beta_s} \rfloor_{\underline{\beta}^{(j-1)}} \\ &= \beta_s^{(j-1)} - (\underline{X}' W_s^{(j-1)} \underline{X})^{-1} \underline{X} \underline{S}_s^{(j-1)} \end{aligned}$$

where

$$\begin{aligned} \underline{X}' &= (x_1, x_2, \ldots, x_n), \\ W_s^{(j-1)} &= \mathrm{diag}[\omega_1^{(j-1)}(s), \ldots, \omega_n^{(j-1)}(s)], \\ \underline{S}_s^{(j-1)} &= \begin{bmatrix} S_{s1}^{(j-1)} \\ \vdots \\ S_{sn}^{(j-1)} \end{bmatrix}, \end{aligned}$$

where the elements of the above are

$$\begin{aligned} \omega_{si}^{(j-1)} &= \delta_i \omega_i^{\delta_i=1} e^{\beta_s^{(j-1)}} \log \mathcal{F}_s(Y_i) + (1-\delta_i) \omega_i^{\delta_i=0} e^{\beta_s^{(j-1)} x_i} \log \mathcal{F}_s(Y_i) \\ S_{si}^{(j-1)} &= \delta_i \omega_i^{\delta_i=1} [1 + e^{\beta_s^{(j-1)} x_i} \log \mathcal{F}_s(Y_i)] + (1-\delta_i) \omega_i^{\delta_i=0} e^{\beta_s^{(j-1)} x_i} \log \mathcal{F}_s(Y_i)]. \end{aligned}$$

By using algebraic manipulations we can rewrite the iteration as an iterated weighted least squared algorithm

$$\begin{aligned} \beta_s^{(j)} &= \beta_s^{(j-1)} - \left(\frac{\partial Q^2}{\partial \beta_s^2}\right)^{-1} \frac{\partial Q}{\partial \beta_s} \rfloor_{\underline{\beta}^{(j-1)}} \\ &= \beta_s^{(j-1)} - (\underline{X}' \omega_s^{(j-1)} \underline{X})^{-1} \underline{X}' \underline{S}_s^{(j-1)} \\ &= (\underline{X}' W_s^{(j-1)} \underline{X})^{-1} (\underline{X}' W_s^{(j-1)} \underline{X}) \beta_s^{(j-1)} - (\underline{X}' W_s^{(j-1)} \underline{X})^{-1} \underline{X} \underline{S}_s^{(j-1)} \\ &= (\underline{X}' W_s^{(j-1)} \underline{X})^{-1} \underline{X}' W_s^{(j-1)} \underline{X} \beta_s^{(j-1)} - (\underline{X}' W_s^{(j-1)} \underline{X})^{-1} \underline{X} W_s^{(j-1)} [W_s^{(j-1)}]^{-1} \underline{S}_s^{(j-1)} \end{aligned}$$

Now we rewrite the above in weighted least squares form. Define

$$\underline{Z}_s^{(j-1)} = \underline{X}\beta_s^{(j-1)} - [W_s^{(j-1)}]^{-1}\underline{S}_s^{(j-1)}$$

this "acts" as our pseudo y-variable. Using this notation we have

$$\beta_s^{(j)} \;=\; (\underline{X}'W_s^{(j-1)}\underline{X})^{-1}\underline{X}'W_s^{(j-1)}\underline{Z}_s^{(j-1)}.$$

Thus at each step of the Newton-Raphson iteration we minimise the weighted least equation

$$\sum_{i=1}^{n}\omega_{si}^{(j-1)}\left(Z_s^{(j-1)} - \beta x_i\right)^2 \text{ for } s = 1, 2.$$

Thus altogether in the EM-algorithm we have:
Start with initial value $\beta_1^0, \beta_2^0, p^0$

Step 1 Set $(\beta_{1,r-1}, \beta_{2,r-1}, p_{r-1}) = (\beta_1^*, \beta_2^*, p^*)$. Evaluate $\omega_i^{\delta_i}$ and $\omega_i^{1-\delta_i}$ (these probabilies/weights stay the same throughout the iterative least squares).

Step 2 Maximize $Q(\theta_*, \theta)$ by using the algorithm $p_r = \frac{a_r}{a_r+b_r}$ where a_r, b_r are defined previously. Now evaluate for $s = 1, 2$

$$\beta_s^{(j)} = (\underline{X}'W_s^{(j-1)}\underline{X})^{-1}\underline{X}'W_s^{(j-1)}Z_s^{(j-1)}.$$

Iterate until convergence of the parameters.

Step 3 Go back to step 1 until convergence of the EM algorithm.

7.2.4 Exercises

Exercise 7.1 *Consider the linear regression model*

$$Y_i = \underline{\alpha}'\underline{x}_i + \sigma_i\varepsilon_i$$

where ε_i follows a standard normal distribution (mean zero and variance 1) and σ_i^2 follows a Gamma distribution

$$f(\sigma^2; \lambda) = \frac{\sigma^{2(\kappa-1)}\lambda^\kappa \exp(-\lambda\sigma^2)}{\Gamma(\kappa)}, \quad \sigma^2 \geq 0,$$

with $\kappa > 0$.

Let us suppose that $\underline{\alpha}$ and λ are unknown parameters but κ is a known parameter. We showed in Exercise 1.1 that directly maximising the log-likelihood was extremely difficult.

Derive the EM-algorithm for model. In your derivation explain what quantities will have to be evaluated numerically.

Exercise 7.2 *Consider the following shifted exponential mixture distribution*

$$f(x;\lambda_1,\lambda_2,p,a) = p\frac{1}{\lambda_1}\exp(-x/\lambda_1)I(x \geq 0) + (1-p)\frac{1}{\lambda_2}\exp(-(x-a)/\lambda_2)I(x \geq a),$$

where p, λ_1, λ_2 and a are unknown.

(i) Make a plot of the above mixture density.

Considering the cases $x \geq a$ and $x < a$ separately, calculate the probability of belonging to each of the mixtures, given the observation X_i (i.e. Define the variable δ_i, where $P(\delta_i = 1) = p$ and $f(x|\delta_i = 1) = \frac{1}{\lambda_1}\exp(-x/\lambda_1)$ and calculate $P(\delta_i|X_i)$).

(ii) Show how the EM-algorithm can be used to estimate $a, p, \lambda_1, \lambda_2$. At each iteration you should be able to obtain explicit solutions for most of the parameters, give as many details as you can.

Hint: It may be beneficial for you to use profiling too.

(iii) From your knowledge of estimation of these parameters, what do you conjecture the rates of convergence to be? Will they all be the same, or possibly different?

Exercise 7.3 *Suppose $\{Z_i\}_{i=1}^n$ are independent random variables, where Z_i has the density*

$$f_Z(z;\beta_0,\beta_1,\mu,\alpha,u_i) = ph(z;\beta_0,\beta_1,u_i) + (1-p)g(z;\alpha,\mu),$$

$g(x;\alpha,\mu) = (\frac{\alpha}{\mu})(\frac{x}{\mu})^{\alpha-1}\exp(-(x/\mu)^\alpha)I_{(0,\infty)}(x)$ (the Weibull distribution) and $h(x;\beta_0,\beta_1,u_i) = \frac{1}{\lambda_i}\exp(-x/\lambda_i)I_{(0,\infty)}(x)$ (the exponential distribution), with $\lambda_i = \beta_0\exp(\beta_1 u_i)$ and $\{u_i\}_{i=1}^n$ are observed regressors.

The parameters p, β_0, β_1, μ and α are unknown and our objective in this question is to estimate them.

(a) What is the log-likelihood of $\{Z_i\}$? (Assume we also observe the deterministic regressors $\{u_i\}$.)

(b) By defining the correct dummy variable δ_i derive the steps of the EM-algorithm to estimate the parameters $p, \beta_0, \beta_1, \mu, \alpha$ (using the method of profiling if necessary).

7.3 Hidden Markov Models

Finally, we consider applications of the EM-algorithm to parameter estimation in Hidden Markov Models (HMM). This is a model where the EM-algorithm pretty much surpasses any other likelihood maximisation methodology. It is worth mentioning that the EM-algorithm in this setting is often called the *Baum-Welch algorithm*.

Hidden Markov models are a generalisation of mixture distributions, however unlike mixture distibutions it is difficult to derive an explicit expression for the likelihood of a Hidden Markov Models. HMM are a general class of models which are widely used in several applications (including speech recongition), and can easily be generalised to the Bayesian set-up. A nice description of them can be found on Wikipedia.

In this section we will only briefly cover how the EM-algorithm can be used for HMM. We do not attempt to address any of the issues surrounding how the maximisation is done, interested readers should refer to the extensive literature on the subject.

The general HMM is described as follows. Let us suppose that we observe $\{Y_t\}$, where the rvs Y_t satisfy the Markov property $P(Y_t|Y_{t-1}, Y_{t-1}, \ldots) = P(Y_t|Y_{t-1})$. In addition to $\{Y_t\}$ there exists a 'hidden' unobserved discrete random variables $\{U_t\}$, where $\{U_t\}$ satisfies the Markov property $P(U_t|U_{t-1}, U_{t-2}, \ldots) = P(U_t|U_{t-1})$ and 'drives' the dependence in $\{Y_t\}$. In other words $P(Y_t|U_t, Y_{t-1}, U_{t-1}, \ldots) = P(Y_t|U_t)$. To summarise, the HMM is described by the following properties:

(i) We observe $\{Y_t\}$ (which can be either continuous or discrete random variables) but do not observe the hidden discrete random variables $\{U_t\}$.

(ii) Both $\{Y_t\}$ and $\{U_t\}$ are time-homogenuous Markov random variables that is $P(Y_t|Y_{t-1}, Y_{t-1}, \ldots) = P(Y_t|Y_{t-1})$ and $P(U_t|U_{t-1}, U_{t-1}, \ldots) = P(U_t|U_{t-1})$. The distributions of $P(Y_t)$, $P(Y_t|Y_{t-1})$, $P(U_t)$ and $P(U_t|U_{t-1})$ do not depend on t.

(iii) The dependence between $\{Y_t\}$ is driven by $\{U_t\}$, that is $P(Y_t|U_t, Y_{t-1}, U_{t-1}, \ldots) = P(Y_t|U_t)$.

There are several examples of HMM, but to have a clear intepretation of them, in this section we shall only consider one classical example of a HMM. Let us suppose that the hidden random variable U_t can take N possible values $\{1, \ldots, N\}$ and let $p_i = P(U_t = i)$ and $p_{ij} = P(U_t = i|U_{t-1} = j)$. Moreover, let us suppose that Y_t are continuous random variables where $(Y_t|U_t = i) \sim \mathcal{N}(\mu_i, \sigma_i^2)$ and the conditional random variables $Y_t|U_t$ and $Y_\tau|U_\tau$ are independent of each other. Our objective is to estimate the parameters $\theta = \{p_i, p_{ij}, \mu_i, \sigma_i^2\}$ given $\{Y_i\}$. Let $f_i(\cdot;\theta)$ denote the normal distribution $\mathcal{N}(\mu_i, \sigma_i^2)$.

Remark 7.3.1 (HMM and mixture models) *Mixture models (described in the above section) are a particular example of HMM. In this case the unobserved variables $\{U_t\}$ are iid, where $p_i = P(U_t = i|U_{t-1} = j) = P(U_t = i)$ for all i and j.*

Let us denote the log-likelihood of $\{Y_t\}$ as $\mathcal{L}_T(\underline{Y};\theta)$ (this is the observed likelihood). It is clear that constructing an explicit expression for $\mathcal{L}_T$ is difficult, thus maximising the likelihood is near impossible. In the remark below we derive the observed likelihood.

Remark 7.3.2 *The likelihood of $\underline{Y} = (Y_1, \ldots, Y_T)$ is*

$$\begin{aligned} L_T(\underline{Y};\theta) &= f(Y_T|Y_{T-1}, Y_{T-2}, \ldots;\theta) \ldots f(Y_2|Y_1;\theta)P(Y_1;\theta) \\ &= f(Y_T|Y_{T-1};\theta) \ldots f(Y_2|Y_1;\theta)f(Y_1;\theta). \end{aligned}$$

Thus the log-likelihood is

$$\mathcal{L}_T(\underline{Y};\theta) = \sum_{t=2}^{T} \log f(Y_t|Y_{t-1};\theta) + f(Y_1;\theta).$$

The distribution of $f(Y_1;\theta)$ is simply the mixture distribution

$$f(Y_1;\theta) = p_1 f(Y_1;\theta_1) + \ldots + p_N f(Y_1;\theta_N),$$

where $p_i = P(U_t = i)$. The conditional $f(Y_t|Y_{t-1})$ is more tricky. We start with

$$f(Y_t|Y_{t-1};\theta) = \frac{f(Y_t, Y_{t-1};\theta)}{f(Y_{t-1};\theta)}.$$

An expression for $f(Y_t;\theta)$ is given above. To evaluate $f(Y_t, Y_{t-1};\theta)$ we condition on

U_t, U_{t-1} to give (using the Markov and conditional independent propery)

$$\begin{aligned} f(Y_t, Y_{t-1}; \theta) &= \sum_{i,j} f(Y_t, Y_{t-1}|U_t = i, U_{t-1} = j) P(U_t = i, U_{t-1} = j) \\ &= \sum_{i,j} f(Y_t|U_t = i) P(Y_{t-1}|U_{t-1} = j) P(U_t = i|U_{t-1} = j) P(U_{t-1} = i) \\ &= \sum_{i,j} f_i(Y_t; \theta_i) f_j(Y_{t-1}; \theta_j) p_{ij} p_i. \end{aligned}$$

Thus we have

$$f(Y_t|Y_{t-1}; \theta) = \frac{\sum_{i,j} f_i(Y_t; \theta_i) f_j(Y_{t-1}; \theta_j) p_{ij} p_i}{\sum_i p_i f(Y_{t-1}; \theta_i)}.$$

We substitute the above into $\mathcal{L}_T(\underline{Y}; \theta)$ to give the expression

$$\mathcal{L}_T(\underline{Y}; \theta) = \sum_{t=2}^{T} \log \left(\frac{\sum_{i,j} f_i(Y_t; \theta_i) f_j(Y_{t-1}; \theta_j) p_{ij} p_i}{\sum_i p_i f(Y_{t-1}; \theta_i)} \right) + \log \left(\sum_{i=1}^{N} p_i f(Y_1; \theta_i) \right).$$

Clearly, this is extremely difficult to maximise.

Instead we seek an indirect method for maximising the likelihood. By using the EM algorithm we can maximise a likelihood which is a lot easier to evaluate. Let us suppose that we observe $\{Y_t, U_t\}$. Since $P(\underline{Y}|\underline{U}) = P(Y_T|Y_{T-1}, \ldots, Y_1, \underline{U}) P(Y_{T-1}|Y_{T-2}, \ldots, Y_1, \underline{U}) \ldots P(Y_1|\underline{U}) = \prod_{t=1}^{T} P(Y_t|U_t)$, and the distribution of $Y_t|U_t$ is $\mathcal{N}(\mu_{U_t}, \sigma^2_{U_t})$, then the complete likelihood of $\{Y_t, U_t\}$ is

$$\left(\prod_{t=1}^{T} f(Y_t|U_t; \theta) \right) p_{U_1} \prod_{t=2}^{T} p_{U_t|U_{t-1}}.$$

Thus the log-likelihood of the complete observations $\{Y_t, U_t\}$ is

$$\mathcal{L}_T(\underline{Y}, \underline{U}; \theta) = \sum_{t=1}^{T} \log f(Y_t|U_t; \theta) + \sum_{t=2}^{T} \log p_{U_t|U_{t-1}} + \log p_{U_1}.$$

Of course, we do not observe the complete likelihood, but the above can be used in order to define the function $Q(\theta_*, \theta)$ which is maximised in the EM-algorithm. It is worth mentioning that given the transition probabilities of a discrete Markov chain (that is $\{p_{i,j}\}_{ij}$) the marginal/stationary probabilities $\{p_i\}$ can be obtained by solving $\pi = \pi P$, where P is the transition matrix. Thus it is not necessary to estimate the marginal

probabilities $\{p_i\}$ (note that the exclusion of $\{p_i\}$ in the log-likelihood, above, gives the conditional complete log-likelihood).

We recall that to maximise the observed likelihood $\mathcal{L}_T(\underline{Y};\theta)$ using the EM algorithm involves evaluating $Q(\theta_*,\theta)$, where

$$\begin{aligned} Q(\theta_*,\theta) &= \mathrm{E}\Big(\sum_{t=1}^{T}\log f(Y_t|U_t;\theta)+\sum_{t=2}^{T}\log p_{U_t|U_{t-1}}+\log p_{U_1}\Big|\underline{Y},\theta_*\Big) \\ &= \sum_{\underline{U}\in\{1,\ldots,N\}^T}\Big(\sum_{t=1}^{T}\log f(Y_t|U_t;\theta)+\sum_{t=2}^{T}\log p_{U_t|U_{t-1}}+\log p_{U_1}\Big)p(\underline{U}|\underline{Y},\theta_*). \end{aligned}$$

Note that each step in the algorithm the probability $p(\underline{U}|\underline{Y},\theta_*)$ needs to be evaluated. This is done by using conditioning

$$\begin{aligned} p(\underline{U}|\underline{Y},\theta_*) &= p(U_1|\underline{Y},\theta_*)\prod_{t=2}^{T}P(U_t|U_{t-1},\ldots,U_1\underline{Y};\theta_*) \\ &= p(U_1|\underline{Y},\theta_*)\prod_{t=2}^{T}P(U_t|U_{t-1},\underline{Y};\theta_*) \text{ using the Markov property.} \end{aligned}$$

Evaluation of the above is not simple (mainly because one is estimating the probability of being in state U_t based on U_{t-1} *and* the observation information Y_t in the past, present and future). This is usually done using the so called forward backward algorithm (and is related to the idea of Kalman filtering).

For this example the EM algorithm is

(i) Define an initial value $\theta_1\in\Theta$. Let $\theta_*=\theta_1$.

(ii) **The expectation step**,

For a fixed θ_* evaluate $P(U_t,\underline{Y},\theta_*)$, $P(U_t|U_{t-1},\underline{Y},\theta_*)$ and $Q(\theta_*,\theta)$.

(iii) **The maximisation step**

Evaluate $\theta_{k+1}=\arg\max_{\theta\in\Theta}Q(\theta_*,\theta)$ by differentiating $Q(\theta_*,\theta)$ wrt to θ and equating to zero.

(iv) If θ_k and θ_{k+1} are sufficiently close to each other stop the algorithm and set $\hat{\theta}_n=\theta_{k+1}$. Else set $\theta_*=\theta_{k+1}$, go back and repeat steps (ii) and (iii) again.

Since $P(U_1|\underline{Y},\theta_*) = P(U_1,\underline{Y},\theta_*)/P(\underline{Y},\theta_*)$ and $P(U_t,U_{t-1}|\underline{Y},\theta_*) = P(U_t,U_{t-1},\underline{Y},\theta_*)/P(\underline{Y},\theta_*)$; $P(\underline{Y},\theta_*)$ is common to all $\underline{U}$ in $\{1,\ldots,N\}^T$ and is independent of θ_*, Thus rather than maximising $Q(\theta_*,\theta)$ one can equvalently maximise

$$\widetilde{Q}(\theta_*,\theta) = \sum_{\underline{U}\in\{1,\ldots,N\}^T} \left(\sum_{t=1}^{T} \log f(Y_t|U_t;\theta) + \sum_{t=2}^{T} \log p_{U_t|U_{t-1}} + \log p_{U_1} \right) p(\underline{U},\underline{Y},\theta_*),$$

noting that $\widetilde{Q}(\theta_*,\theta) \propto Q(\theta_*,\theta)$ and the maximum of $\tilde{Q}(\theta_*,\theta)$ with respect to θ is the same as the maximum of $Q(\theta_*,\theta)$.