Stat583

Piet Groeneboom

April 27, 1999

1 The EM algorithm and self-consistency

We recall the basic model and notation of Chapter 7, section 1, of Stat582, but use (for reasons that soon will become clear) the letters $\mathbf{Y}$, Y, $\mathcal{B}$ and Q_θ instead of $\mathbf{X}$, X, $\mathcal{A}$, and P_θ, respectively. So our model is a probability space $(\mathbf{Y}, \mathcal{B}, Q_\theta)$, where $\theta \in \Theta \subset \mathbb{R}^m$, and we assume that Y has the probability distribution Q_θ on this space, i.e.,

$$\mathbb{P}\{Y \in B\} = Q_\theta(B),\ B \in \mathcal{B}.$$

Also suppose, as in Chapter 7, section 1, of Stat582, that Y has a density q_θ w.r.t. some σ-finite measure ν, i.e., we can write

$$Q_\theta(B) = \int_B q_\theta(y)\,d\nu(y),\ B \in \mathcal{B}.$$

Then, for a given realization y of the random variable Y, we would compute the *maximum likelihood estimator* (MLE) $\hat\theta$ of θ by maximizing $q_\theta(y)$ as a function of θ.

But suppose that the maximization is, for some reason, difficult. The idea of the EM algorithm is then to construct a "hidden space" $(\mathbf{X}, \mathcal{A}, P_\theta)$, such that Y can be represented as $Y = T(X)$, where X has distribution P_θ and T is a measurable mapping from $(\mathbf{X}, \mathcal{A})$ to $(\mathbf{Y}, \mathcal{B})$, and such that the computation of the MLE of θ is easier on the space $(\mathbf{X}, \mathcal{A}, P_\theta)$. On the "hidden space" $(\mathbf{X}, \mathcal{A}, P_\theta)$ we again assume that X has a density p_θ w.r.t. some σ-finite measure μ, i.e., we can write

$$P_\theta(A) = \int_A p_\theta(x)\,d\mu(x),\ A \in \mathcal{A}.$$

Note that in this situation Q_θ can be represented as

$$Q_\theta = P_\theta T^{-1},$$

where the probability measure $P_\theta T^{-1}$ is defined by

$$P_\theta T^{-1}(B) = P_\theta\left(T^{-1}(B)\right),$$

for each $B \in \mathcal{B}$. Also note that we usually cannot observe the random variable X, and that the construction of the “hidden space” is just an artifice for computing the MLE of θ.

The EM algorithm now proceeds as follows for a given observation y.
Start with an initial estimate $\theta^{(0)}$ of θ. This yields an initial guess for the probability distribution $P_{\theta^{(0)}}$ (and hence also an initial guess for the probability distribution $Q_{\theta^{(0)}} = P_{\theta^{(0)}}T^{-1}$).
Do the *E-step*: compute, for $\theta \in \Theta$, the conditional expectation

$$\phi_0(\theta) \overset{\text{def}}{=} E_{P_{\theta^{(0)}}}\left\{\log p_\theta(X) \,\big|\, T(X) = y\right\}. \tag{1.1}$$

Then do the *M-step*: maximize

$$\phi_0(\theta), \tag{1.2}$$

as a function of θ.
Suppose that $\theta^{(1)}$ maximizes (1.2).
Next start with $P_{\theta^{(1)}}$ instead of $P_{\theta^{(0)}}$, and compute in the E-step

$$\phi_1(\theta) \overset{\text{def}}{=} E_{P_{\theta^{(1)}}}\left\{\log p_\theta(X) \,\big|\, T(X) = y\right\}.$$

Then, in the M-step, we maximize $\phi_1(\theta)$ as a function of θ.
Generally, in the mth step, we first compute the conditional expectation

$$\phi_m(\theta) \overset{\text{def}}{=} E_{P_{\theta^{(m)}}}\left\{\log p_\theta(X) \,\big|\, T(X) = y\right\}. \tag{1.3}$$

and then maximize

$$\phi_m(\theta), \tag{1.4}$$

as a function of θ.
Repeat these E- and M-steps until $\theta^{(m)}$ does not change in, say, the 10^{th} decimal (or until some other criterion is met) at, say, the mth iteration step. Then we take $\theta^{(m)}$ as our estimate of the MLE.

Will this work? Sometimes it will and sometimes it won’t! We now first give an argument, explaining why the EM algorithm might work, and, after that, some examples of situations where it indeed works.

Suppose that, for given y, the real MLE is given by $\hat{\theta}$, where $\hat{\theta}$ is an interior point of Θ, and that the function

$$\theta \mapsto q_\theta(y),\ \theta \in \Theta,$$

is differentiable on the interior of Θ. Then Rolle’s theorem tells us that we must have:

$$\frac{\partial}{\partial\theta}\, q_\theta(y)\Big|_{\theta=\hat{\theta}} = 0. \tag{1.5}$$

But if the EM algorithm converges to an interior point $\theta^{(\infty)} \in \Theta$, then $\theta^{(\infty)}$ maximizes the function

$$\theta \mapsto E_{P_{\theta^{(\infty)}}}\left\{\log p_\theta(X) \,\big|\, T(X) = y\right\}, \tag{1.6}$$

see (1.3) and (1.4). But this implies, assuming that (1.6) is differentiable at interior points $\theta \in \Theta$ and that we may interchange expectation and differentiation:

$$\frac{\partial}{\partial\theta} E_{P_{\theta^{(\infty)}}} \left\{ \log p_\theta(X) \,\Big|\, T(X) = y \right\} = E_{P_{\theta^{(\infty)}}} \left\{ \frac{\partial}{\partial\theta} \log p_\theta(X) \,\Big|\, T(X) = y \right\} = 0. \tag{1.7}$$

at $\theta = \theta^{(\infty)}$.

Let, as in Chapter 7, section 1, of Stat582, $\dot{l}_\theta$ be defined by

$$\dot{l}_\theta(x) = \frac{\partial}{\partial\theta} \log p_\theta(x).$$

Then, using properties of conditional expectations, and assuming that certain interchanges of differentiation and integration (or summation) are allowed (homework assignment!), it is seen that:

$$E_{P_\theta} \left\{ \dot{l}_\theta(X) \,\Big|\, T(X) = y \right\} q_\theta(y) = \frac{\partial}{\partial\theta} q_\theta(y).$$

Hence (1.7) would imply, for $\theta = \theta^{(\infty)}$,

$$\frac{\partial}{\partial\theta} q_\theta(y) = 0,$$

at a value y such that $q_\theta(y) > 0$. Or, written differently, we would have, for $\theta = \theta^{(\infty)}$,

$$\frac{\partial}{\partial\theta} \log q_\theta(y) = \frac{\frac{\partial}{\partial\theta} q_\theta(y)}{q_\theta(y)} = 0, \tag{1.8}$$

if (1.7) is satisfied and $q_\theta(y) > 0$. So (1.5) would be satisfied at $\theta = \theta^{(\infty)}$, and hence, if there is only one θ for which this "score equation" is zero, $\theta^{(\infty)}$ would be the MLE!

The equation

$$E_{P_\theta} \left\{ \dot{l}_\theta(X) \,\Big|\, T(X) = y \right\} = 0, \tag{1.9}$$

that is satisfied at $\theta = \theta^{(\infty)}$, is called the *self-consistency equation* (the reason for this name will become clearer in the sequel). So the reason for believing that the EM algorithm might work is the fact that (1.9) implies (1.8) for $\theta = \theta^{(\infty)}$, if $q_{\theta^{(\infty)}}(y) > 0$. So, if the likelihood function $\theta \mapsto \log q_\theta(y)$ is only maximized at a value θ where the derivative w.r.t. θ is zero, then a stationary point $\theta^{(\infty)}$ of the EM algorithm would give the MLE.

This argument also points to potential difficulties with the EM algorithm: it might not work if the maximum is not attained at an interior point, or if the likelihood function is not differentiable at the MLE, or if the score equation (1.8) has multiple roots, some (or all) of which do not maximize the likelihood. Indeed all these situation can occur.

Example 1.1 (from DEMPSTER, LAIRD AND RUBIN (1977)) Suppose that $Y = (Y_1, \ldots, Y_4) \sim Q_\theta$, where Q_θ is the multinomial $\text{Mult}_4\left(n, \underline{q}(\theta)\right)$–distribution, with

$$\underline{q}(\theta) = \left(\tfrac{1}{2} + \tfrac{1}{4}\theta, \tfrac{1}{4}(1-\theta), \tfrac{1}{4}(1-\theta), \tfrac{1}{4}\theta\right), \; \theta \in (0,1).$$

Then Q_θ has the density

$$q_\theta(y_1,\ldots,y_4) = \frac{n!}{y_1!\ldots y_4!}\left(\tfrac{1}{2}+\tfrac{1}{4}\theta\right)^{y_1}\left(\tfrac{1}{4}(1-\theta)\right)^{y_2}\left(\tfrac{1}{4}(1-\theta)\right)^{y_3}\left(\tfrac{1}{4}\theta\right)^{y_4}$$

w.r.t. counting measure ν on $\mathbb{N}^4$, where $\mathbb{N}$ is the set of natural numbers with 0 included. So our "observation space" is $(\mathbf{Y},\mathcal{B},Q_\theta)$, where $\mathbf{Y}=\mathbb{N}^4$ and $\mathcal{B}$ is the set of subsets of $\mathbf{Y}$.

To compute the MLE of θ with the EM algorithm, we introduce the "hidden space" $(\mathbf{X},\mathcal{A},P_\theta)$, where $\mathbf{X}=\mathbb{N}^5$. $\mathcal{A}$ is the set of subsets of $\mathbf{X}$, and P_θ is the multinomial $\text{Mult}_5\left(n,\underline{p}(\theta)\right)$–distribution, with

$$\underline{p}(\theta)=\left(\tfrac{1}{2},\tfrac{1}{4}\theta,\tfrac{1}{4}(1-\theta),\tfrac{1}{4}(1-\theta),\tfrac{1}{4}\theta\right),\ \theta\in(0,1),$$

and again $n=197$. Then P_θ has the density

$$p_\theta(x_1,\ldots,x_5)=\frac{n!}{x_1!\ldots x_5!}\left(\tfrac{1}{2}\right)^{x_1}\left(\tfrac{1}{4}\theta\right)^{x_2}\left(\tfrac{1}{4}(1-\theta)\right)^{x_3}\left(\tfrac{1}{4}(1-\theta)\right)^{x_4}\left(\tfrac{1}{4}\theta\right)^{x_5}$$

w.r.t. counting measure μ on $\mathbb{N}^5$, and we introduce a random variable $X=(X_1,\ldots,X_5)$ with distribution P_θ on $\mathbf{X}$. The mapping T, taking X to Y is given by

$$T(x)=(x_1+x_2,x_3,x_4,x_5),\ x\in\mathbb{N}^5.$$

It is easily verified that if X is distributed according to P_θ, then $Y=T(X)$ is distributed according to Q_θ, and hence $Q_\theta=P_\theta T^{-1}$, as is required for the application of the EM algorithm (see above).

If we could observe a realization of the vector X (instead of Y), we could easily compute the MLE of θ. This is seen as follows. For a given $x=(x_1,\ldots,x_5)$, the log likelihood for θ (or, equivalently p_θ) is

$$\log p_\theta(x)=c+(x_2+x_5)\log\left(\tfrac{1}{4}\theta\right)+(x_3+x_4)\log\left(\tfrac{1}{4}(1-\theta)\right). \tag{1.10}$$

where c is a part of the log likelihood not depending on θ. Setting the derivative w.r.t. θ equal to zero yields

$$\frac{x_2+x_5}{\theta}-\frac{x_3+x_4}{1-\theta}=0.$$

Hence the MLE of θ would be

$$\hat\theta=\frac{x_2+x_5}{x_2+x_3+x_4+x_5}\,, \tag{1.11}$$

since we can verify that the stationary point indeed corresponds to a maximum.

However, we do not observe X, but instead a realization of the random variable Y which is distributed as $T(X)$. So we want to apply the EM algorithm by updating our estimates of the “hidden” X in the E-step and subsequently maximizing over θ in the M-step. For the E-step we have to compute

$$\phi_m(\theta) \stackrel{\text{def}}{=} E_{P_{\theta^{(m)}}}\left\{\log p_\theta(X) \,\middle|\, T(X)=y\right\},$$

see (1.3). But by (1.10) is is sufficient to compute the conditional expectations

$$E_{P_{\theta^{(m)}}}\left\{X_2+X_5 \,\middle|\, T(X)=y\right\}$$

and

$$E_{P_{\theta^{(m)}}}\left\{X_3+X_4 \,\middle|\, T(X)=y\right\},$$

since these are the only ingredients that are needed in the M-step. We have:

$$E_{P_{\theta^{(m)}}}\left\{X_1 \,\middle|\, T(X)=y\right\} = y_1\frac{1/2}{1/2+\theta^{(m)}/4}, \tag{1.12}$$

$$E_{P_{\theta^{(m)}}}\left\{X_2 \,\middle|\, T(X)=y\right\} = y_1\frac{\theta^{(m)}/4}{1/2+\theta^{(m)}/4}, \tag{1.13}$$

and

$$E_{P_{\theta^{(m)}}}\left\{X_k \,\middle|\, T(X)=y\right\} = y_{k-1},\ k=3,4,5. \tag{1.14}$$

So we get

$$\begin{aligned}\phi_m(\theta) &= E_{P_{\theta^{(m)}}}\left\{\log p_\theta(X) \,\middle|\, T(X)=y\right\}\\ &= \left(y_1\frac{\theta^{(m)}/4}{1/2+\theta^{(m)}/4}+y_4\right)\log\left(\tfrac{1}{4}\theta\right)+(y_2+y_3)\log\left(\tfrac{1}{4}(1-\theta)\right)+g(y),\end{aligned}$$

where $g(y)$ only depends on y (and not on θ), and where $y_1\left(\theta^{(m)}/4\right)\big/\left\{1/2+\theta^{(m)}/4\right\}$ is the updated estimate of X_2 at the mth iteration. Note that this updated estimate of X_2 will in general not be an integer!

Writing

$$x_2^{(m)} = y_1\frac{\theta^{(m)}/4}{1/2+\theta^{(m)}/4}$$

for this updated estimate of X_2, we find by (1.11) that

$$\theta^{(m+1)} = \frac{x_2^{(m)}+y_4}{x_2^{(m)}+y_2+y_3+y_4}$$

maximizes $\phi_m(\theta)$ at the mth iteration step.

A simple C program that implements the E- and M-steps above, and started with $\theta^{(0)} = 0.5$, produced the following output for the observation vector $y = (125, 18, 20, 34)$ (hence $n = 197$):

Iteration	$\theta^{(m)}$	$\theta^{(m)} - \hat{\theta}$	log likelihood	$\frac{\theta^{(m+1)}-\hat{\theta}}{\theta^{(m)}-\hat{\theta}}$
0	0.500000000000	-0.126821497871	67.320170488171	0.146458412039
1	0.608247422680	-0.018574075191	67.382924965794	0.134620296088
2	0.624321050369	-0.002500447502	67.384081218564	0.133023705192
3	0.626488879080	-0.000332618791	67.384101726379	0.132811268713
4	0.626777322347	-0.000044175524	67.384102088226	0.132783053828
5	0.626815632110	-0.000005865761	67.384102094606	0.132779307312
6	0.626820719019	-0.000000778852	67.384102094718	0.132778809278
7	0.626821394456	-0.000000103415	67.384102094720	0.132778740769
8	0.626821484140	-0.000000013731	67.384102094720	0.132778707205
9	0.626821496048	-0.000000001823	67.384102094720	0.132778513910
10	0.626821497629	-0.000000000242	67.384102094720	-
11	0.626821497839	-0.000000000032	67.384102094720	-
12	0.626821497867	-0.000000000004	67.384102094720	-
13	0.626821497870	-0.000000000001	67.384102094720	-
14	0.626821497871	0.000000000000	67.384102094720	-
15	0.626821497871	0.000000000000	67.384102094720	-

Note that in Dempster, Laird and Rubin (1977) the minus signs are missing in the second column (a mistake that has subsequently been copied by many authors referring to this example!). Here the "log likelihood" is not really the log likelihood $\log q_\theta(y)$, but a version that only differs from $\log q_\theta(y)$ by constant. I took:

$$\tilde{l}(\theta) \overset{\text{def}}{=} y_1 \log(2+\theta) + (y_2 + y_3)\log(1-\theta) + y_4 \log(\theta). \tag{1.15}$$

Note that $\tilde{l}(\theta)$ increases in the first 8 iterations, but after that doesn't change any more, using 12 decimals. The real MLE $\hat{\theta}$ is approximately (in 15 decimals):

$$\hat{\theta} \approx 0.626821497870982.$$

The column headed by $\frac{\theta^{(m+1)}-\hat{\theta}}{\theta^{(m)}-\hat{\theta}}$ provides an estimate of the "rate of convergence" of the EM algorithm to the stationary point for this particular example. In this example it is suggested that this (linear) rate of convergence is approximately 0.132778. This means that we have:

$$\lim_{m\to\infty} \frac{\theta^{(m+1)} - \hat{\theta}}{\theta^{(m)} - \hat{\theta}} \approx 0.132778.$$

In practice this rate of convergence is often estimated by the ratio

$$\frac{\theta^{(m+1)} - \theta^{(m)}}{\theta^{(m)} - \theta^{(m-1)}}$$

for large m (but not so large that the denominator is too close to zero). As an example, in the present situation we get

$$\frac{\theta^{(m+1)} - \theta^{(m)}}{\theta^{(m)} - \theta^{(m-1)}} = 0.132778715949,$$

for $m = 10$.

A linear rate of convergence means that we have

$$\left|\theta^{(m+1)} - \hat{\theta}\right| \leq c\left|\theta^{(m)} - \hat{\theta}\right|,$$

for some $c \in (0, 1)$. In this case this is apparently satisfied for $c \approx 0.132778$ (and m sufficiently large). There are other methods that have *superlinear* convergence in this case, meaning that

$$\left|\theta^{(m+1)} - \hat{\theta}\right| \leq c\left|\theta^{(m)} - \hat{\theta}\right|^{\alpha},$$

for some $\alpha > 1$ and m sufficiently large For example, with Newton's method we would get a relation of this type with $\alpha = 2$, so-called *quadratic* convergence. Using (1.15) (the likelihood, modulo a constant not involving θ), the Newton method is based on the iterations

$$\theta^{(m+1)} = \theta^{(m)} + I\left(\theta^{(m)}\right)^{-1} \frac{\partial}{\partial\theta}\tilde{l}(\theta)\Big|_{\theta=\theta^{(m)}},$$

where

$$\frac{\partial}{\partial\theta}\tilde{l}(\theta) = \frac{y_1}{2+\theta} - \frac{y_2+y_3}{1-\theta} + \frac{y_4}{\theta}$$

and

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

The corresponding table for Newton's method is, for the present example:

Iteration	$\theta^{(m)}$	$\theta^{(m)} - \hat{\theta}$	log likelihood	$\frac{\theta^{(m+1)}-\hat{\theta}}{\theta^{(m)}-\hat{\theta}}$
0	0.500000000000000	-0.126821497870982	64.629744483953	-0.075240701717
1	0.636363636363636	0.009542138492654	67.366740797467	0.015423623821
2	0.626968672225539	0.000147174354557	67.384098005534	0.000228693022
3	0.626821531528730	0.000000033657748	67.384102094720	0.000000065971
4	0.626821497870984	0.000000000000002	67.384102094720	-
5	0.626821497870982	0.000000000000000	67.384102094720	-
6	0.626821497870982	0.000000000000000	67.384102094720	-

So in this case the method has fully converged at the $5th$ iteration step, using the same starting value. Also notice the rapid decrease of the numbers in the last column, indicating the quadratic convergence.

Example 1.2 (The exponential mixtures model) Let $Y = (Y_1 \ldots, Y_n)$, where the Y_i are i.i.d. with density

$$g_{(p,\lambda,\mu)}(y) = \Big\{ p\lambda e^{-\lambda y} + (1-p)\mu e^{-\mu y} \Big\}\, 1_{(0,\infty)}(y)$$

w.r.t. Lebesgue measure on $\mathbb{R}$, where $p \in (0,1)$ and $\lambda, \mu > 0$. This means: Y is distributed according to the probability distribution Q_θ, where $\theta = (p, \lambda, \mu)$, and where Q_θ has density

$$q_\theta(y_1, \ldots, y_n) = g_\theta(y_1) \ldots g_\theta(y_n),$$

w.r.t. Lebesgue measure on $\mathbb{R}^n_+$. So our observation space is $(\mathbf{Y}, \mathcal{B}, Q_\theta)$, where

$$\mathbf{Y} = \mathbb{R}^n_+,$$

and $\mathcal{B}$ is (as usual) the collection of Borel sets on $\mathbf{Y}$.

Let $y = (y_1, \ldots, y_n)$ be a (sample) realization of the random vector $Y = (Y_1, \ldots, Y_n)$. Then the log likelihood for $\theta = (p, \lambda, \mu)$ is given by

$$l(\theta|y) = \log g_\theta(y_1) + \ldots + \log g_\theta(y_n),$$

and the partial derivatives w.r.t. p, λ and μ are given by

$$\dot{l}_p(\theta|y) = \sum_{i=1}^n \frac{\lambda e^{-\lambda y_i} - \mu e^{-\mu y_i}}{g_\theta(y_i)}\,,$$

$$\dot{l}_\lambda(\theta|y) = \sum_{i=1}^n \frac{p e^{-\lambda y_i}\{1 - \lambda y_i\}}{g_\theta(y_i)}\,,$$

and

$$\dot{l}_\mu(\theta|y) = \sum_{i=1}^n \frac{(1-p) e^{-\mu y_i}\{1 - \mu y_i\}}{g_\theta(y_i)}\,,$$

respectively. The MLE $\hat{\theta} = (\hat{p}, \hat{\lambda}, \hat{m})$ of $\theta = (p, \lambda, \mu)$ is found by solving the score equation

$$\Big(\dot{l}_p(\theta|y), \dot{l}_\lambda(\theta|y), \dot{l}_\mu(\theta|y)\Big) = (0, 0, 0).$$

in (p, λ, μ). This is a complicated set of equations (in fact a lot more complicated than the corresponding 1-dimensional score equation in Example 1.1!).

Now we introduce the ”hidden space” $(\mathbf{X}, \mathcal{A}, P_\theta)$, where

$$\mathbf{X} = (\mathbb{R}_+ \times \{0, 1\})^n$$

and P_θ is specified by the density

$$p_\theta\left((y_1, \delta_1), \ldots, (y_n, \delta_n)\right) = f_\theta(y_1, \delta_1) \ldots f_\theta(y_n, \delta_n),$$

where

$$f_\theta(y,\delta) = \left(p\lambda e^{-\lambda y}\right)^\delta \left((1-p)\mu e^{-\mu y}\right)^{1-\delta},\ \theta = (p,\lambda,\mu). \tag{1.16}$$

The idea is that on the hidden space we introduce an extra experiment, with outcome 1 or 0, telling us whether we get the exponential distribution with parameter λ (this happens when the outcome is 1) or the exponential distribution with parameter μ (when the outcome is 0). The outcome is represented by the value of the indicator δ_i. If $\delta_i = 1$ (resp. $\delta_i = 0$) the outcome of the ith variable of our vector of n variables is coming from the exponential distribution with parameter λ (resp. μ).

The mapping going from the hidden space to the observation space is simply:

$$T(x) = (y_1,\dots,y_n)\,,\ \text{if } x = ((y_1,\delta_1),\dots,(y_n,\delta_n))\,. \tag{1.17}$$

For the E-step we have to compute:

$$\begin{aligned}
&E_{P_{\theta^{(m)}}}\left\{\log p_\theta(X)\;\middle|\;T(X)=y\right\}\\
&= E_{P_{\theta^{(m)}}}\left\{\sum_{i=1}^n (\Delta_i\log(p\lambda) - \Delta_i\lambda Y_i\right.\\
&\qquad\qquad\left. +(1-\Delta_i)\log((1-p)\mu) - (1-\Delta_i)\mu Y_i)\;\middle|\;T(X)=(y_1,\dots,y_n)\right\}\\
&= \sum_{i=1}^n E_{P_{\theta^{(m)}}}\left\{\Delta_i\log(p\lambda) - \Delta_i\lambda y_i\right.\\
&\qquad\qquad\left. +(1-\Delta_i)\log((1-p)\mu) - (1-\Delta_i)\mu y_i\;\middle|\;T(X)=(y_1,\dots,y_n)\right\},
\end{aligned}$$

where

$$X = ((Y_1,\Delta_1),\dots,(Y_1,\Delta_1))\,,$$

and where we switched to capitals to indicate that the mapping (1.17) now has a random argument. But

$$E_{P_{\theta^{(m)}}}\left\{\Delta_i\;\middle|\;T(X)=(y_1,\dots,y_n)\right\} = \frac{f_{\theta^{(m)}}(y_i,1)}{g_{\theta^{(m)}}(y_i)}\,, \tag{1.18}$$

see (1.16). Define

$$\delta_i^{(m)} = \frac{f_{\theta^{(m)}}(y_i,1)}{g_{\theta^{(m)}}(y_i)},$$

i.e., $\delta_i^{(m)}$ is the updated conditional expectation of Δ_i at the mth iteration. Then we get, using (1.18),

$$\begin{aligned}
&E_{P_{\theta^{(m)}}}\left\{\log p_\theta(X)\;\middle|\;T(X)=(y_1,\dots,y_n)\right\}\\
&= \sum_{i=1}^n\left\{\delta_i^{(m)}\log p + \left(1-\delta_i^{(m)}\right)\log(1-p)\right.\\
&\qquad\qquad\left. +\delta_i^{(m)}(\log\lambda - \lambda y_i) + \left(1-\delta_i^{(m)}\right)(\log\mu - \mu y_i)\right\}.
\end{aligned} \tag{1.19}$$

For the M-step we have to maximize (1.19) over $\theta = (p, \lambda, \mu)$. Setting the partial derivative w.r.t. p equal to zero yields:

$$\frac{\sum_{i=1}^n \delta_i^{(m)}}{p} - \frac{\sum_{i=1}^n \left(1-\delta_i^{(m)}\right)}{1-p} = 0,$$

and hence

$$p^{(m+1)} = \frac{1}{n}\sum_{i=1}^n \delta_i^{(m)}.$$

Similarly we get:

$$1/\lambda^{(m+1)} = \frac{\sum_{i=1}^n \delta_i^{(m)} y_i}{\sum_{i=1}^n \delta_i^{(m)}},$$

and

$$1/\mu^{(m+1)} = \frac{\sum_{i=1}^n \left(1-\delta_i^{(m)}\right) y_i}{\sum_{i=1}^n \left(1-\delta_i^{(m)}\right)}.$$

It can be verified (look at the diagonal matrix of the second derivatives!) that $\theta^{(m+1)} = \left(p^{(m+1)}, \lambda^{(m+1)}, \mu^{(m+1)}\right)$ indeed maximizes (1.19) as a function of θ. So we have specified the E-step and M-step, and writing a simple C program, implementing these two step is no more difficult than in Example 1.1.

Monotonicity of EM

Let, as before,

$$l_\theta(y) = \log q_\theta(y).$$

We want to show

$$l_{\theta^{(m+1)}}(y) \geq l_{\theta^{(m)}}(y),$$

This is what we call *"monotonicity of EM"*: at each step of the EM algorithm we will get a likelihood that is at least as big as the likelihood at the preceding step.

Suppose T is, as before, the mapping from the hidden space $\mathbf{X}$ to the observation space $\mathbf{Y}$, that $Q_\theta = P_\theta T^{-1}$, that Q_θ has a density q_θ w.r.t. a σ-finite measure ν and P_θ has a density p_θ w.r.t. a σ-finite measure μ. Let $k_\theta(x|y)$ be the conditional density of X, given $T(X) = y$. Then:

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

for values of x such that $T(x) = y$. If $T(x) \neq y$, we put $k_\theta(x|y) = 0$. We can now write:

$$l_\theta(y) = \log p_\theta(x) - \log k_\theta(x|y), \text{ if } T(x) = y,\ p_\theta(x) > 0 \text{ and } q_\theta(y) > 0.$$

Suppose that $q_\theta(y) > 0$ (note that any "candidate MLE" θ will at least need to have this property, since otherwise the log likelihood is $-\infty$ for this candidate MLE). Then we get, for each m (m being the index of the mth iteration of the EM algorithm):

$$\begin{aligned}\log q_\theta(y) &= E_{P_\theta^{(m)}}\left\{\log q_\theta(y) \bigm| T(X) = y\right\}\\ &= E_{P_\theta^{(m)}}\left\{\log p_\theta(X) - \log k_\theta(X|y) \bigm| T(X) = y\right\}\\ &= E_{P_\theta^{(m)}}\left\{\log p_\theta(X) \bigm| T(X) = y\right\} - E_{P_\theta^{(m)}}\left\{\log k_\theta(X|y) \bigm| T(X) = y\right\}.\end{aligned} \tag{1.20}$$

(to think about what happens if $p_\theta(X) = 0$ will be part of the homework!)

Now we look separately at the two terms in the last line of (1.20), and compare the expressions we get, by replacing θ by $\theta^{(m)}$ and $\theta^{(m+1)}$, respectively. Note that we keep the distribution $P_\theta^{(m)}$, determining the distribution of the conditional expectation, fixed!

First of all, we get:

$$E_{P_\theta^{(m)}}\left\{\log p_{\theta^{(m+1)}}(X) \bigm| T(X) = y\right\} - E_{P_\theta^{(m)}}\left\{\log p_{\theta^{(m)}}(X) \bigm| T(X) = y\right\} \ge 0, \tag{1.21}$$

since $\theta^{(m+1)}$ maximizes the conditional expectation

$$\phi_m(\theta) \stackrel{\text{def}}{=} E_{P_\theta^{(m)}}\left\{\log p_\theta(X) \bigm| T(X) = y\right\}$$

over all θ, so the value we get by plugging in $\theta^{(m+1)}$ will certainly be at least as big as the value we get by plugging in $\theta^{(m)}$!

Secondly, we get:

$$\begin{aligned}&E_{P_{\theta^{(m)}}}\left\{\log k_{\theta^{(m+1)}}(X|y) \bigm| T(X) = y\right\} - E_{P_{\theta^{(m)}}}\left\{\log k_{\theta^{(m)}}(X|y) \bigm| T(X) = y\right\}\\ &E_{P_{\theta^{(m)}}}\left\{\log \frac{k_{\theta^{(m+1)}}(X|y)}{k_{\theta^{(m)}}(X|y)} \bigm| T(X) = y\right\}\\ &\le \log E_{P_{\theta^{(m)}}}\left\{\frac{k_{\theta^{(m+1)}}(X|y)}{k_{\theta^{(m)}}(X|y)} \bigm| T(X) = y\right\},\end{aligned} \tag{1.22}$$

where (the conditional form of) Jensen's inequality is used in the last step. But we have:

$$E_{P_{\theta^{(m)}}}\left\{\frac{k_{\theta^{(m+1)}}(X|y)}{k_{\theta^{(m)}}(X|y)} \bigm| T(X) = y\right\} = 1, \text{ a.e. } [Q_{\theta^{(m)}}]. \tag{1.23}$$

This is seen in the following way. Let the function $g : \mathbf{Y} \to \mathbb{R}$ represent the left-hand side of (1.23):

$$g(y) = E_{P_{\theta^{(m)}}}\left\{\frac{k_{\theta^{(m+1)}}(X|y)}{k_{\theta^{(m)}}(X|y)} \bigm| T(X) = y\right\}. \tag{1.24}$$

Then g is a $\mathcal{B}$-measurable function that is defined by the following relation:

$$\int_B g(y)\, dQ_{\theta^{(m)}}(y) = \int_{T^{-1}(B)} \frac{k_{\theta^{(m+1)}}(x|T(x))}{k_{\theta^{(m)}}(x|T(x))}\, dP_{\theta^{(m)}}(x),\ \forall B \in \mathcal{B}, \tag{1.25}$$

(using the general definition of conditional expectations). But since we can write

$$\frac{k_{\theta^{(m+1)}}(x|T(x))}{k_{\theta^{(m)}}(x|T(x))}=\frac{p_{\theta^{(m+1)}}(x)}{q_{\theta^{(m+1)}}(T(x))}\cdot\frac{q_{\theta^{(m)}}(T(x))}{p_{\theta^{(m)}}(x)}\,,$$

the right-hand side of (1.25) can be written:

$$\begin{aligned}&\int_{T^{-1}(B)}\frac{p_{\theta^{(m+1)}}(x)}{q_{\theta^{(m+1)}}(T(x))}\cdot q_{\theta^{(m)}}(T(x))\,d\mu(x)=\int_{T^{-1}(B)}\frac{q_{\theta^{(m)}}(T(x))}{q_{\theta^{(m+1)}}(T(x))}\,dP_{\theta^{(m+1)}}(x)\\&=\int_B\frac{q_{\theta^{(m)}}(y)}{q_{\theta^{(m+1)}}(y)}\,dQ_{\theta^{(m+1)}}(y)=\int_B q_{\theta^{(m)}}(y)\,d\nu(y)=\int_B 1\,dQ_{\theta^{(m)}}(y).\end{aligned}\tag{1.26}$$

implying, using (1.25) and (1.26),

$$g(y)=E_{P_\theta^{(m)}}\left\{\frac{k_{\theta^{(m+1)}}(X|y)}{k_{\theta^{(m)}}(X|y)}\mid T(X)=y\right\}=1\text{ a.e. }\left[Q_{\theta^{(m)}}\right].$$

So (neglecting things happening on sets of $Q_{\theta^{(m)}}$-measure zero) we get that the last expression in (1.22) is equal to zero!

The preceding somewhat elaborate argument is needed, since in general the conditional density $k(x|y)$ will be singular w.r.t. the measure μ, implying that the simple argument

$$\begin{aligned}&E_{P_\theta^{(m)}}\left\{\frac{k_{\theta^{(m+1)}}(X|y)}{k_{\theta^{(m)}}(X|y)}\mid T(X)=y\right\}\\&=\int_{T^{-1}(y)}\frac{k_{\theta^{(m+1)}}(x|y)}{k_{\theta^{(m)}}(x|y)}\,k_{\theta^{(m)}}(x|y)\,d\mu(x)\\&=\int_{T^{-1}(y)}k_{\theta^{(m+1)}}(x|y)\,d\mu(x)=1\end{aligned}$$

does not work, since in general, in dealing with absolutely continuous distributions, we will get

$$\int_{T^{-1}(y)}k_{\theta^{(m+1)}}(x|y)\,d\mu(x)=0.$$

As an example, take μ to be Lebesgue measure on the unit square, and consider the mapping $T(x,y)=y$. Then a conditional density of the type $k_{\theta^{(m)}}(x|y)$ will be concentrated on a line segment and the corresponding measure is singular w.r.t. μ. This difficulty is completely glossed over in, e.g., Dempster, Laird and Rubin (1977); the difficulty is already there in their very first formula (1.1) (and returns, for example, in the first line of the proof of Lemma 2). If one writes dx everywhere (as they do), it is of course generally unclear what really is going on, and how discrete distributions are covered. Similar difficulties occur in the book McLachlan and Krishnan (1997), completely devoted to the EM algorithm (see, e.g., (3.10) on p. 83 of their book).

Now, by combining (1.20), (1.21) and (1.22), we get

$$l_{\theta^{(m+1)}}(y) - l_{\theta^{(m)}}(y) \geq 0,$$

i.e., the likelihood for the parameter θ in the observation space is increased (at least "nondecreased") at each step of the EM algorithm.

For the general theory on change of variables in integrals w.r.t. measures, see, e.g., Billingsley (1995), third edition (the first edition of this book contained an incorrect result of this type!).

2 Nonparametric maximum likelihood estimators

Suppose that $X_1, \ldots, X_n$ is a sample of 1-dimensional random variables, generated by a distribution with distribution function (df) F_0, and that we want to estimate F_0 by maximizing a likelihood. If we want to do this "nonparametrically", i.e., without making any parametric assumptions on F_0, like F_0 is a normal distribution function with location parameter μ and variance σ^2, we are in trouble, because there is no dominating measure that allows us to specify a density. In the case that we want to estimate a finite-dimensional parameter, like in the case where assume that F_0 is the distribution function of a normal distribution with location parameter μ and variance σ^2, the likelihood of a realization $(x_1, \ldots, x_n)$ of our random vector $(X_1, \ldots, X_n)$ would be

$$\prod_{i=1}^{n} \frac{1}{\sigma} \phi\left(\frac{x_i - \mu}{\sigma}\right), \text{ where } \phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\tfrac{1}{2} x^2\right\},$$

w.r.t. Lebesgue measure on $\mathbb{R}^n$. Similarly, if we assume that our sample vector $(x_1, \ldots, x_n)$ is generated by a Binomial $\text{Bin}(n, p)$ distribution, the likelihood would be

$$\prod_{i=1}^{n} \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i},$$

w.r.t. counting measure on $\{0, \ldots, n\}^n$. So in these cases there is a fixed dominating measure, respectively Lebesgue measure and counting measure. But in the case of nonparametric maximum likelihood such a fixed dominating measure is often not available.

One way out of the difficulty is to restrict the set of distributions over which we are going the maximize to a set for which we can specify a dominating measure. As an example, let us take the counting measure μ on the set $\{x_1, \ldots, x_n\}$, where, for simplicity, the x_i's are assumed to be different. Then the likelihood of the sample would be

$$\prod_{i=1}^{n} p_F(x_i)$$

w.r.t. counting measure on $\{x_1, \dots, x_n\}^n$, where $p_F(x_i)$ is the probability that $X_i = x_i$, if the underlying distribution function is F. This means that we restrict the maximization problem to the set of discrete dfs F, corresponding to a probability distribution that is concentrated on the finite set $\{x_1, \dots, x_n\}$.

Now the maximization problem becomes simply: maximize

$$\sum_{i=1}^{n} \log p_i$$

over the vector $(p_1, \dots, p_n)$, under the restrictions $\sum_{i=1}^n p_i = 1$ and $p_i \ge 0$, $i = 1, \dots, n$.

One way to do this is to use a Lagrange multiplier. So we consider the problem of minimizing the function

$$\phi_\lambda(p) = -\sum_{i=1}^{n} \log p_i + \lambda \left\{ \sum_{i=1}^{n} p_i - 1 \right\},$$

as a function of $p = (p_1, \dots, p_n)$, for a suitably chosen Lagrange multiplier λ. The reason for going from a maximization problem to a minimization problem is that we want to put the problem into the general framework of minimization of convex functions under side constraints, which will be useful later, if we meet more difficult problems of this type. Setting the partial derivatives w.r.t. p_i equal to zero gives us the equations

$$-\frac{1}{p_i} + \lambda = 0, \; i = 1, \dots, n.$$

Since $\sum_{i=1}^n p_i = 1$, we must have:

$$\sum_{i=1}^{n} p_i = \frac{n}{\lambda} = 1,$$

and hence $\lambda = n$ and $p_i = 1/n$. We now get for any vector $q = (q_1, \dots, q_n)$ with nonnegative components and such that $\sum_{i=1}^n q_i = 1$:

$$-\sum_{i=1}^{n} \log q_i = \phi_n(q) \ge \phi_n(p) = -\sum_{i=1}^{n} \log p_i,$$

where the inequality holds since p minimizes the function ϕ_n over all vectors $r = (r_1, \dots, r_n)$ with nonnegative components (where the components r_i do not necessarily sum to 1).

So the distribution function maximizing the likelihood for all distributions which have as support the finite set of points $\{x_, \dots, x_n\}$ is just the empirical distribution function $\mathbb{F}_n$. So in this sense the empirical distribution function is the *nonparametric maximum likelihood estimator* (NPMLE) of F_0. Note, however, that the dominating measure with respect to which we maximized the likelihood depended on the sample, and hence this dominating measure will change from sample to sample.

So the dominating measure is in fact itself random! But this is not necessarily a bad thing. For example, we know that, by Donsker's theorem

$$\sqrt{n}\left\{\mathbb{F}_n - F_0\right\} \xrightarrow{\mathcal{D}} B \circ F_0,$$

where $\xrightarrow{\mathcal{D}}$ means convergence in distribution and B is the Brownian bridge. This means that the "distance" between $\mathbb{F}_n$ and F_0 is of order $n^{-1/2}$, so the NPMLE stabilizes and will be closer and closer to the real distribution function, as the sample size increases. Also, the Glivenko-Cantelli theorem tells us that

$$\sup_{x \in I\!\!R} |\mathbb{F}_n(x) - F_0(x)| \to 0,$$

with probability one (this is sometimes called "the fundamental theorem of statistics"), which also points to this stabilizing phenomenon. The following example describes a situation where we do not have the problem of specifying the dominating measure.

Example 2.1 (Current status data) Let $(X_1, U_1), \ldots, (X_n, U_n)$ be a sample of random variables in $I\!\!R_+^2$, where X_i and U_i are independent (non-negative) random variables with distribution functions F_0 and G, respectively. The only observations that will be available are U_i ("observation time") and $\Delta_i = \{X_i \le U_i\}$. Here and (sometimes) in the sequel I will denote the indicator of an event A (such as $\{X_i \le U_i\}$) just by A, instead of 1_A. The (marginal) log likelihood for F_0 at a realization $((u_1, \delta_1), \ldots, (u_n, \delta_n))$ of $((U_1, \Delta_1), \ldots, (U_n, \Delta_n))$ is given by the function

$$F \mapsto \sum_{i=1}^{n} \left\{\delta_i \log F(u_i) + (1-\delta_i)\log\bigl(1 - F(u_i)\bigr)\right\}, \tag{2.1}$$

where F is a right-continuous distribution function. This is the simplest case of *interval censoring* and often called the "current status" model. A *nonparametric maximum likelihood estimator* (NPMLE) $\hat{F}_n$ of F_0 is a (right-continuous) distribution function F, maximizing (2.1).

I will show later that there exists a 1-step algorithm for computing the NPMLE in this model. But, since this is a clear case of a situation where we have a (real!) hidden space, it is tempting to put this into the framework of the EM algorithm. The random variables (X_i, U_i) are living on the hidden space $\mathbf{X}$, the random variables (U_i, Δ_i) on the observation space $\mathbf{Y}$, and our mapping T is given by:

$$T\left((x_1, u_1), \ldots, (x_n, u_n)\right) = \left((u_1, \delta_1), \ldots, (u_n, \delta_n)\right), \text{ where } \delta_i = \{x_i \le u_i\}. \tag{2.2}$$

We now proceed from here as in section 3.1 of Groeneboom and Wellner (1992) (things are done a bit differently in Groeneboom (1996), where also a more general discussion of the merits of the EM algorithm versus other algorithms is presented).

Suppose that, in our search for an NPMLE of F_0, we restrict attention to the class $\mathcal{F}$ of purely discrete distribution functions F of distributions with mass concentrated on the set of observation points $u_{(i)}$, $i = 1, \ldots, n$, where $u_{(i)}$ is the ith order statistics of the set $\{u_1, \ldots, u_n\}$, with an arbitrary additional point $u_{(n+1)} > u_{(n)}$ for "remaining mass". The latter point is needed, since we may have evidence that the distribution, corresponding to F_0, has mass beyond the largest observation point $u_{(n)}$, and, at first sight perhaps somewhat artificially, we take an arbitrary point $u_{(n+1)} > u_{(n)}$ for the location of this mass (but it will be clear from the sequel that we can indeed do this without limiting the generality of our approach).

By restricting ourselves to the set $\mathcal{F}$ of purely discrete distribution functions F of distributions with mass concentrated on the set of observation points $u_{(i)}$, $i = 1, \ldots, n$ and the additional point $u_{(n+1)}$, we have in fact reduced the problem to a finite-dimensional maximization problem: the distribution $F \in \mathcal{F}$ is completely specified by the parameters

$$p_i = P_F\left\{X = u_{(i)}\right\},\ 1 \le i \le n+1, \tag{2.3}$$

where X has the probability distribution, specified by the distribution function F, and where the parameters p_i satisfy $\sum_{i=1}^{n+1} p_i = 1$ and $p_i \ge 0$, $1 \le i \le n+1$.

We now denote the probability density p_F in (2.3) by

$$f(x) = P_F\left\{X = x\right\},\ x \in \left\{u_{(1)}, \ldots, u_{(n+1)}\right\},$$

and take as our starting point of the EM algorithm for F the discrete uniform distribution function on the points $u_{(1)}, \ldots, u_{(n+1)}$. Note that the X_i's are supposed to be identically distributed so that

$$P_F\left\{X_i = x\right\} = f(x),\ i = 1, \ldots, n.$$

In the E-step we have to compute

$$E^{(0)}\left\{\sum_{i=1}^n \log f(X_i) \;\Big|\; T\left((X_1, U_1), \ldots, (X_n, U_n)\right) = \left((u_1, \delta_1), \ldots, (u_n, \delta_n)\right)\right\}, \tag{2.4}$$

where $E^{(0)}$ denotes the expectation at the first step of the EM algorithm (where we have the uniform distribution on the points $u_{(1)}, \ldots, u_{(n+1)}$). We will denote the probability measure, corresponding to $E^{(0)}$ by $P^{(0)}$. Note that in taking the conditional expectation, the distribution of the U_i's does not have to specified, since we can immediately reduce (2.4) to:

$$E^{(0)}\left\{\sum_{i=1}^n \log f(X_i) \;\Big|\; T\left((X_1, u_1), \ldots, (X_n, u_n)\right) = \left((u_1, \delta_1), \ldots, (u_n, \delta_n)\right)\right\}, \tag{2.5}$$

implying that we only have to specify the distribution of the X_i's to compute this conditional expectation.

We can rewrite (2.4) in the following way:

$$\sum_{i=1}^n E^{(0)}\left\{\log f(X_i) \;\Big|\; T\left((X_1, u_1), \ldots, (X_n, u_n)\right) = \left((u_1, \delta_1), \ldots, (u_n, \delta_n)\right)\right\}$$

$$\sum_{i=1}^{n}\left\{\sum_{k=1}^{n+1}\log f\left(u_{(k)}\right)P^{(0)}\left\{X_i=u_{(k)}\bigm|\Delta_i=\delta_i\right\}\right\}$$
$$\sum_{k=1}^{n+1}\log f\left(u_{(k)}\right)\sum_{i=1}^{n}P^{(0)}\left\{X_i=u_{(k)}\bigm|\Delta_i=\delta_i\right\}.$$

So, if we denote $f\left(u_{(k)}\right)$ by p_k, the M-step of the first iteration consists of maximizing

$$\sum_{k=1}^{n+1}\log p_k\sum_{i=1}^{n}P^{(0)}\left\{X_i=u_{(k)}\bigm|\Delta_i=\delta_i\right\}. \tag{2.6}$$

over the set of parameters $p=(p_1,\dots,p_{n+1})$ such that $\sum_{i=1}^{n+1}p_i=1$ and $p_i\ge 0$, $1\le i\le n+1$. It is easily shown (homework!) that (2.6) is maximized over this set by taking

$$p_k=\frac{1}{n}\sum_{i=1}^{n}P^{(0)}\left\{X_i=u_{(k)}\bigm|\Delta_i=\delta_i\right\},\ 1\le k\le n+1.$$

Hence the combined E- and M-step yield at the end of the first iteration:

$$p_k^{(1)}=\frac{1}{n}\sum_{i=1}^{n}P^{(0)}\left\{X_i=u_{(k)}\bigm|\Delta_i=\delta_i\right\},\ 1\le k\le n+1.$$

So the estimate of the distribution function F_0 that we obtain at the end of the first E- and M-step is given by

$$F^{(1)}(t)=\sum_{\{k:u_{(k)}\le t\}}p_k^{(1)}, \tag{2.7}$$

where we put $F^{(1)}(t)=0$, if $t<u_{(1)}$ (i.e., when the set of indices over which we sum in (2.7) is empty). Generally we get, as the result of the E- and M-step at the mth iteration:

$$p_k^{(m+1)}=\frac{1}{n}\sum_{i=1}^{n}P^{(m)}\left\{X_i=u_{(k)}\bigm|\Delta_i=\delta_i\right\},\ 1\le k\le n+1, \tag{2.8}$$

where $P^{(m)}$ denotes the probability distribution at the mth iteration step. The corresponding relation for the distribution function, obtained at the end of the mth iteration step is:

$$F^{(m+1)}(t)=\frac{1}{n}\sum_{i=1}^{n}P^{(m)}\left\{X_i\le t\bigm|\Delta_i=\delta_i\right\}. \tag{2.9}$$

But the right-hand side of (2.9) can be written

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}E^{(m)}\left\{1_{\{X_i\le t\}}\bigm|\Delta_i=\delta_i\right\}\\&=E^{(m)}\left\{\frac{1}{n}\sum_{i=1}^{n}1_{\{X_i\le t\}}\bigm|T\left((X_1,U_1),\dots,(X_n,U_n)\right)=((u_1,\delta_1),\dots,(u_n,\delta_n))\right\}\\&=E^{(m)}\left\{\mathbb{F}_n(t)\bigm|T\left((X_1,U_1),\dots,(X_n,U_n)\right)=((u_1,\delta_1),\dots,(u_n,\delta_n))\right\},\end{aligned} \tag{2.10}$$

where $E^{(m)}$ is the expectation under $P^{(m)}$, and $\mathbb{F}_n$ is the (unobservable!) empirical distribution function of the X_i. So, combining (2.9) and (2.10), we get

$$F^{(m+1)}(t) = E^{(m)}\left\{\mathbb{F}_n(t) \;\middle|\; T\left((X_1,U_1),\ldots,(X_n,U_n)\right) = ((u_1,\delta_1),\ldots,(u_n,\delta_n))\right\}. \tag{2.11}$$

Hence, if the EM algorithm converges to a limit distribution $F^{(\infty)}$ with corresponding expectation $E^{(\infty)}$, the corresponding relation would be:

$$F^{(\infty)}(t) = E^{(\infty)}\left\{\mathbb{F}_n(t) \;\middle|\; T\left((X_1,U_1),\ldots,(X_n,U_n)\right) = ((u_1,\delta_1),\ldots,(u_n,\delta_n))\right\}. \tag{2.12}$$

This is an equation that again is called a *self-consistency equation.* It tells us that, if $F^{(\infty)}$ really is the NPMLE of F_0, the conditional expectation of the NPMLE in the hidden space is equal to the NPMLE in the observation space (since the empirical distribution function $\mathbb{F}_n$ is the NPMLE in the hidden space). Note that the expectation $E^{(\infty)}$ in principle also involves the distribution of the U_i's, but that, in taking the conditional expectation, we only have to specify the distribution of the X_i (see (2.5)).

We can write (2.8) in the following explicit form:

$$p_k^{(m+1)} = \frac{1}{n}\sum_{i=1}^n \left\{\frac{\delta_i}{F^{(m)}(u_i)}\{u_i \ge u_{(k)}\} + \frac{1-\delta_i}{1-F^{(m)}(u_i)}\{u_i < u_{(k)}\}\right\} p_k^{(m)},\ 1 \le k \le n+1,$$

since

$$P^{(m)}\left\{X_i = u_{(k)} \;\middle|\; \Delta_i = \delta_i\right\} = p_k^{(m)}\left\{\frac{\delta_i}{F^{(m)}(u_i)}\{u_i \ge u_{(k)}\} + \frac{1-\delta_i}{1-F^{(m)}(u_i)}\{u_i < u_{(k)}\}\right\}. \tag{2.13}$$

This means that if the EM algorithm converges, we get for the probability masses $p_k^{(\infty)}$ in the limit

$$p_k^{(\infty)} = \frac{1}{n}\sum_{i=1}^n \left\{\frac{\delta_i}{F^{(\infty)}(u_i)}\{u_i \ge u_{(k)}\} + \frac{1-\delta_i}{1-F^{(\infty)}(u_i)}\{u_i < u_{(k)}\}\right\} p_k^{(\infty)},\ 1 \le k \le n+1. \tag{2.14}$$

Hence, if $p_k^{(\infty)} > 0$, we get:

$$1 = \frac{1}{n}\sum_{i=1}^n \left\{\frac{\delta_i}{F^{(\infty)}(u_i)}\{u_i \ge u_{(k)}\} + \frac{1-\delta_i}{1-F^{(\infty)}(u_i)}\{u_i < u_{(k)}\}\right\}, \tag{2.15}$$

and if $p_k^{(\infty)} = 0$:

$$1 \ge \frac{1}{n}\sum_{i=1}^n \left\{\frac{\delta_i}{F^{(\infty)}(u_i)}\{u_i \ge u_{(k)}\} + \frac{1-\delta_i}{1-F^{(\infty)}(u_i)}\{u_i < u_{(k)}\}\right\}. \tag{2.16}$$

We are now going to show that, in the previous example, instead of the EM algorithm, we can use a 1-step algorithm for computing the NPMLE. For this we will need to develop a little bit of convex duality theory (which will also be useful for other purposes). For this I will use some material that is also included in the notes Groeneboom (1999).

Let ϕ be a smooth convex function defined on $I\!R^n$. The following lemma gives necessary and sufficient conditions for a vector $\hat{x}$ to be the minimizer of ϕ over a convex cone $\mathcal{K}$ in $I\!R^n$, where a cone in $I\!R^n$ is a subset $\mathcal{K}$ of $I\!R^n$, satisfying

$$x \in \mathcal{K} \Longrightarrow c \cdot x \in \mathcal{K}, \text{ for all } c \ge 0.$$

The elementary proof of the following lemma is based on the proof of Lemma 2.1 in Jongbloed (1995). It is a special case of Fenchel's duality theorem (see Rockafellar (1970), Theorem 31.4) and it is also used at several places in Groeneboom and Wellner (1992), see, e.g., the proof of Proposition 1.1 on p. 39.

We write $\nabla\phi$ for the vector of partial derivatives of ϕ,

$$\nabla\phi(x) = \left(\frac{\partial}{\partial x_1}\phi(x), \cdots, \frac{\partial}{\partial x_n}\phi(x)\right),$$

and $\langle\cdot,\cdot\rangle$ for the usual inner product in $I\!R^n$.

Lemma 2.1 *Let $\phi : I\!R^n \to I\!R \cup \{\infty\}$ be a continuous convex function. Let $\mathcal{K} \subset I\!R^n$ be a convex cone and let $\mathcal{K}_0 = \mathcal{K} \cap \phi^{-1}(I\!R)$. Moreover, suppose that $\mathcal{K}_0$ is non-empty, and that ϕ is differentiable on $\mathcal{K}_0$. Then $\hat{x} \in \mathcal{K}_0$ satisfies*

$$\phi(\hat{x}) = \min_{x\in\mathcal{K}} \phi(x), \tag{2.17}$$

if and only if $\hat{x}$ satisfies

$$\forall x \in \mathcal{K} : \langle x, \nabla\phi(\hat{x})\rangle \ge 0, \tag{2.18}$$

and

$$\langle \hat{x}, \nabla\phi(\hat{x})\rangle = 0. \tag{2.19}$$

Proof: We first prove the if-part. Let $x \in \mathcal{K}$ be arbitrary and let $\hat{x} \in \mathcal{K}_0$ satisfy (2.18) and (2.19). Then we get, using the convexity of ϕ,

$$\phi(x) - \phi(\hat{x}) \ge \langle x - \hat{x}, \nabla\phi(\hat{x})\rangle \ge 0,$$

implying $\phi(\hat{x}) = \min_{x\in\mathcal{K}} \phi(x)$. Note that the inequality is trivially satisfied if $x \in \mathcal{K} \setminus \mathcal{K}_0$.

Conversely, let $\hat{x}$ satisfy (2.17), and first suppose that (2.18) is not satisfied. Then there exists an $x \in \mathcal{K}$ such that $\langle x, \nabla\phi(\hat{x})\rangle < 0$. Since, for each $\epsilon \ge 0$,

$$\hat{x} + \epsilon x = (1+\epsilon)\left\{\frac{1}{1+\epsilon}\hat{x} + \left(1 - \frac{1}{1+\epsilon}\right)x\right\} \in \mathcal{K},$$

we have, for $\epsilon \downarrow 0$, using the continuity of ϕ and the assumption that ϕ is differentiable on $\mathcal{K}_0$,

$$\phi(\hat{x}+\epsilon x)-\phi(\hat{x})=\epsilon\,\langle x,\nabla\phi(\hat{x})\rangle+o(\epsilon).$$

This shows that for ϵ sufficiently small, $\phi(\hat{x}+\epsilon x)<\phi(\hat{x})$, contradicting the assumption that $\hat{x}$ minimizes ϕ over $\mathcal{K}$.

Now suppose (2.19) is not satisfied. Then $\hat{x}\neq 0$ and, for $|\epsilon|\leq 1$, $(1+\epsilon)\hat{x}\in\mathcal{K}$. Taking the sign of ϵ opposite to that of $\langle\hat{x},\nabla\phi(\hat{x})\rangle$, we get for $\epsilon\to 0$

$$\phi((1+\epsilon)\hat{x})-\phi(\hat{x})=\epsilon\,\langle\hat{x},\nabla\phi(\hat{x})\rangle+o(\epsilon),$$

and hence the left-hand side will be negative for $|\epsilon|$ sufficiently small, contradicting again the assumption that $\hat{x}$ minimizes ϕ over $\mathcal{K}$. □

We say that a cone $\mathcal{K}$ is *finitely generated* if there are finitely many vectors $z^{(1)},\ldots,z^{(k)}\in\mathcal{K}$ such that

$$x\in\mathcal{K}\iff\exists\alpha_1,\alpha_2,\ldots,\alpha_k\geq 0\text{ such that }x=\sum_{i=1}^{k}\alpha_i z^{(i)}.$$

For finitely generated convex cones we have the following corollary of Lemma 2.1.

Corollary 2.1 *Let ϕ satisfy the conditions of lemma 2.1 and let the convex cone $\mathcal{K}$ be generated by the vectors $z^{(1)},z^{(2)},\ldots,z^{(k)}$. Then $\hat{x}\in\mathcal{K}_0=\mathcal{K}\cap\phi^{-1}(\mathbb{R})$ satisfies*

$$\phi(\hat{x})=\min_{x\in\mathcal{K}}\phi(x),$$

if and only if

$$\langle z^{(i)},\nabla\phi(\hat{x})\rangle\geq 0,\ \textit{for } 1\leq i\leq k, \tag{2.20}$$

$$\langle z^{(i)},\nabla\phi(\hat{x})\rangle=0,\ \textit{if } \hat{\alpha}_i>0, \tag{2.21}$$

where the nonnegative numbers $\hat{\alpha}_1,\hat{\alpha}_2,\ldots,\hat{\alpha}_k$ satisfy

$$\hat{x}=\sum_{i=1}^{k}\hat{\alpha}_i z^{(i)}.$$

Proof: If $x\in\mathcal{K}$, then

$$x=\sum_{i=1}^{k}\alpha_i z^{(i)},$$

where the α_i are nonnegative. Hence we can write

$$\langle x,\nabla\phi(\hat{x})\rangle=\sum_{i=1}^{k}\alpha_i\,\langle z^{(i)},\nabla\phi(\hat{x})\rangle. \tag{2.22}$$

If (2.20) and (2.21) hold, then (2.18) follows, since all terms in the sum on the right-hand side of (2.22) are nonnegative. If $x = \hat{x}$, (2.19) follows since in that case all terms on the right-hand side of (2.22) are zero.

Suppose (2.18) and (2.19) hold. Then (2.20) follows trivially. Taking $x = \hat{x}$ in (2.22) and observing that all terms in the sum on the right-hand side of (2.22) are nonnegative, it follows that (2.19) can only hold if (2.21) holds. □

We now show how Corollary 2.1 leads to a one-step algorithm for computing the NPMLE for the distribution function F_0 in Example 2.1 (current status data). It follows from Theorem 1.5.1 in Robertson *et. al.* (1988), applied to the convex function Φ, defined by

$$\Phi(x) = x \log x + (1-x)\log(1-x),\ x \in (0,1),$$

extended to $[0,1]$ by defining $\Phi(0) = \Phi(1) = 0$, that *maximizing* the function

$$x \mapsto \sum_{i=1}^n \{\delta_{(i)} \log x_i + (1-\delta_{(i)})\log(1-x_i)\},\ x = (x_1, \dots, x_n) \in [0,1]^n,$$

over all vectors $x \in [0,1]^n$ with ordered components $x_1 \le \ldots \le x_n$, is equivalent to *minimizing* the convex function

$$\phi : x \mapsto \sum_{i=1}^n \{x_i - \delta_{(i)}\}^2,\ x = (x_1, \dots, x_n) \in [0,1]^n, \tag{2.23}$$

over all such vectors. Here $\delta_{(i)}$ is a realization of an indicator Δ_j, corresponding to the ith order statistic $u_{(i)}$ of the (realized) observation times $u_1, \dots, u_n$, and is equal to zero or one. We can extend ϕ to $I\!R^n$ by defining

$$\phi(x) = \sum_{i=1}^n \{x_i - \delta_{(i)}\}^2,\ x = (x_1, \dots, x_n) \in I\!R^n,$$

and for this extended function the conditions of Lemma 2.1 are satisfied.

Let $\mathcal{K}$ be the convex cone

$$\mathcal{K} = \{x \in I\!R^n \,:\, x = (x_1, \dots, x_n),\ x_1 \le \cdots \le x_n\}. \tag{2.24}$$

Then $\mathcal{K}$ is finitely generated by the vectors $z^{(i)} = \sum_{j=i}^n e_j$, $1 \le i \le n$, where the e_j are the unit vectors in $I\!R_n$, and the vector $z^{(0)} = -z^{(1)}$. This means that any $x \in \mathcal{K}$ can be written in the form

$$x = \sum_{i=1}^n \alpha_i z^{(i)},\ \alpha_1 \in I\!R,\ \alpha_i \ge 0,\ i = 2, \dots, n.$$

By Corollary 2.1, $\hat{x} = \sum_{i=1}^n \hat{\alpha}_i z^{(i)}$ minimizes $\phi(x)$ over $\mathcal{K}$, if and only if

$$\langle z^{(i)}, \nabla\phi(\hat{x})\rangle \begin{cases} \ge 0 & \text{for } 1 \le i \le n, \\ = 0 & \text{if } \hat{\alpha}_i > 0 \text{ or } i = 1. \end{cases} \tag{2.25}$$

The condition $\langle z^{(1)}, \nabla\phi(\hat{x})\rangle = 0$ arises from the fact that the inner product of $\nabla\phi(\hat{x})$ with both $z^{(1)}$ and $-z^{(1)}$ has to be non-negative.

Since $z^{(i)} = \sum_{j=i}^n e_j$, $i \geq 1$, this gives, by (2.23), that $\hat{x} = \sum_{i=1}^n \hat{\alpha}_i z^{(i)}$ has to satisfy

$$\sum_{j=i}^n \hat{x}_j \geq \sum_{j=i}^n \delta_{(j)},\ i = 1, \ldots, n, \text{ and } \sum_{j=i}^n \hat{x}_j = \sum_{j=i}^n \delta_{(j)},\ \text{if } \hat{\alpha}_i > 0 \text{ or } i = 1. \tag{2.26}$$

Let $P_0 = (0, 0)$ and $P_i = (i, \sum_{j=1}^i \delta_{(j)})$, $i = 1, \ldots, n$. Furthermore, let $C : [0, n] \to I\!R$ be the biggest convex function on $[0, n]$, lying below (or touching) the points P_i. The set of points P_i is usually denoted as the *cumulative sum diagram* (or just *cusum diagram*) and the function C as the *(greatest) convex minorant* of this cusum diagram. Then, defining $\hat{x}_i$ as the left derivative of the convex minorant C at i, it is easily verified that the $\hat{x}_i$'s satisfy (2.26).

In fact, the (greatest) convex minorant has to touch the cusum diagram at P_0 and P_n, which means that we will have

$$\sum_{j=1}^n \hat{x}_j = \sum_{j=1}^n \delta_{(j)}. \tag{2.27}$$

Furthermore, since the convex minorant lies below the points P_i, we must have

$$\sum_{j=1}^i \hat{x}_j \leq \sum_{j=1}^i \delta_{(j)},\ i = 1, \ldots, n. \tag{2.28}$$

By (2.27) and (2.28) we now get

$$\sum_{j=i}^n \hat{x}_j \geq \sum_{j=i}^n \delta_{(j)},\ i = 1, \ldots, n,$$

Defining the $\hat{\alpha}_i$'s by $\hat{x} = \sum_{i=1}^n \hat{\alpha}_i z^{(i)}$, where $\hat{x} = (\hat{x}_1, \ldots, \hat{x}_n)$, it is seen that $\hat{\alpha}_i > 0$, $i > 1$, means that the slope of C changes at $i-1$, and this means that C touches the cusum diagram at P_{i-1}. Since $\hat{x}_i$ is the left-continuous slope of C, we get from this

$$\sum_{j=i}^n \hat{x}_j = \sum_{j=i}^n \delta_{(j)}.$$

It now follows that $\hat{x}$ satisfies (2.26), and since it is also easily seen that $0 \leq \hat{x}_i \leq 1$, for $1 \leq i \leq n$, we get that $\hat{x}$ actually minimizes $\phi(x)$ over all vectors $x \in [0, 1]^n$ with ordered components.

Hence we have the following one-step algorithm for computing the NPMLE: construct the cusum diagram, consisting of the points P_i, and construct its convex minorant. Then the value $\hat{F}_n(u_{(i)})$ of the NPMLE $\hat{F}_n$ at the ith ordered observation time $u_{(i)}$ is given by the left-continuous slope of the convex minorant at i.

Example 2.2 Suppose that $\delta_{(1)} = \delta_{(4)} = \delta_{(6)} = 1$ and $\delta_{(2)} = \delta_{(3)} = \delta_{(5)} = 0$. Then the cusum diagram consists of the points $(0,0)$, $(1,1)$, $(2,1)$, $(3,1)$, $(4,2)$, $(5,2)$ and $(6,3)$. A picture of the cusum diagram and the convex minorant for this situation is shown below.

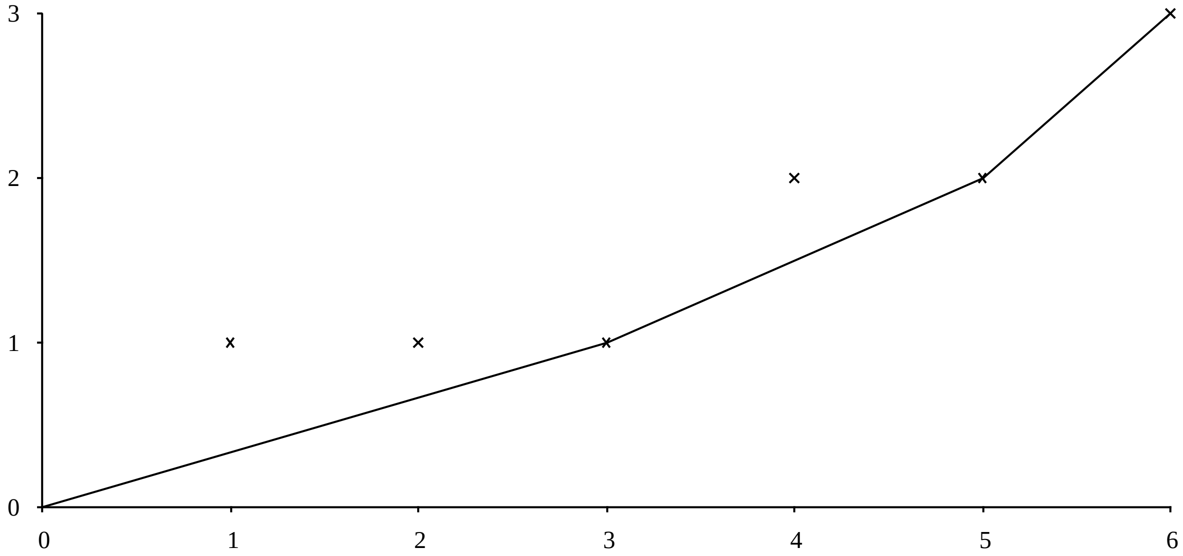

Figure 1: Cusum diagram

From this diagram we can see that $\hat{x}_1 = \hat{x}_2 = \hat{x}_3 = 1/3$, $\hat{x}_4 = \hat{x}_5 = 1/2$, and $\hat{x}_6 = 1$, since these are the left-continuous slopes of the convex minorant at the points $1, \ldots, 6$. So, if $u_{(1)} < \ldots, u_{(6)}$ are our (strictly) ordered observation points, the NPMLE $\hat{F}_6$ is given by

$$\hat{F}_6\left(u_{(i)}\right) = \hat{x}_i,\ i = 1, \ldots, 6.$$

Note that the exact location of the points $u_{(i)}$ does not matter in the computation of the NPMLE.

It was shown in exercises 3 and 4 of the second homework assignment that, if $\delta_{(1)} = 1$ and $\delta_{(n)} = 0$, the relations (2.15) and (2.16), characterizing the NPMLE, are equivalent to the following two relations:

$$\sum_{i=k}^{n} \left\{ \frac{\delta_{(i)}}{F\left(u_{(i)}\right)} - \frac{1 - \delta_{(i)}}{1 - F\left(u_{(i)}\right)} \right\} \leq 0,\ k = 1, \ldots, n, \tag{2.29}$$

and

$$\sum_{i=1}^{n}\left\{\frac{\delta_{(i)}}{F\left(u_{(i)}\right)}-\frac{1-\delta_{(i)}}{1-F\left(u_{(i)}\right)}\right\}F\left(u_{(i)}\right)=0. \tag{2.30}$$

The last equation is the equivalent of a "score equation" in the case of finite-dimensional maximum likelihood estimation. But because we are essentially dealing with an infinite dimensional set this time (the set of all distribution functions on $I\!R$ or on $[0,\infty)$), the situation is more complicated now; we also need the inequalities (2.29).

If we make the identification

$$\hat{x}\overset{\mathrm{def}}{=}(\hat{x}_1,\dots,\hat{x}_n)=\left(\hat{F}_n\left(u_{(1)}\right),\dots,\hat{F}_n\left(u_{(n)}\right)\right),$$

and define the cone $\mathcal{K}$ as in (2.24), then (2.29) corresponds to (2.20) and (2.30) corresponds to (2.21). Note that the set of vectors $\left(F\left(u_{(1)}\right),\dots,F\left(u_{(n)}\right)\right)$ is a subset of the cone $\mathcal{K}$, but that the cone $\mathcal{K}$ itself contains many more vectors, since components bigger than 1 and smaller than 0 are allowed. But if $\delta_{(1)}=1$ and $\delta_{(n)}=0$, we know that the function

$$-\sum_{i=1}^{n}\left\{\delta_{(i)}\log F(u_{(i)})+(1-\delta_{(i)})\left(1-\log F(u_{(i)})\right)\right\}, \tag{2.31}$$

(the log likelihood with a minus sign in front), is infinite if $F(u_{(1)})=0$ or $F(u_{(n)})=1$. In such a case one says that the constraints that the values of the distribution function have to be between zero and one *are not active*, and that we can reduce the problem of minimizing (2.31) to a minimization problem over the cone $\mathcal{K}$, defining

$$-\log x=\infty,\ \text{if } x\le 0.$$

This is the reason that (2.30) and (2.29) characterize the NPMLE in this case.

Now, exactly as in the usual finite-dimensional maximum likelihood problems, we can try to get the solution by taking partial derivatives of the log likelihood. The big difficulty here, though, is that we can only take partial derivatives in certain directions, because we have to stay inside the parameter space. For example, one could wonder whether (2.12) is, in some sense, really the same as (1.9), since both equations are called "self-consistency equations".

To make the connection, we start by assuming that the probability mass in the hidden space is concentrated on the set of points $\{u_{(k_1)},\dots,u_{(k_{m+1})}\}$ which is a subset of the set of points $\{u_{(1)},\dots,u_{(n+1)}\}$, and we introduce the parameter vector

$$\theta=\left(p_{k_1},\dots,p_{k_m}\right),\ 1\le k_1<\dots<k_m\le n,$$

representing the probability masses at the points $u_{(k_j)}$. For similar reasons as in the case of a multinomial distribution, we express $p_{k_{m+1}}$ in terms of the p_{k_i}, $i\le m$, and do not include this parameter in our parameter vector θ.

The log likelihood in the "hidden space" is now given by:

$$\sum_{i=1}^{n} \log f_\theta(x_i),$$

where

$$f_\theta\left(u_{(k_i)}\right) = p_{k_i},\ i = 1, \dots, m+1,$$

The score function $\dot{l}_\theta$ in the hidden space is now given by

$$\dot{l}_\theta(x_1, \dots, x_n) = \left(\frac{n_1}{p_{k_1}} - \frac{n_{m+1}}{p_{k_{m+1}}}, \dots, \frac{n_m}{p_{k_m}} - \frac{n_{m+1}}{p_{k_{m+1}}}\right), \tag{2.32}$$

where n_j is the number of x_i's that is equal to $u_{(k_j)}$. So, if we want (1.9) to be true for this parametrization, we must have for the corresponding random variables N_j:

$$E_\theta\left\{\frac{N_j}{p_{k_j}} - \frac{N_{m+1}}{p_{k_{m+1}}} \,\middle|\, T\left((X_1,U_1),\dots,(X_n,U_n)\right) = ((u_1,\delta_1),\dots,(u_n,\delta_n))\right\} = 0,\ j = 1,\dots,m. \tag{2.33}$$

where the mapping T is defined as in (2.2). But since

$$E_\theta\left\{N_j \mid T\left((X_1,U_1),\dots,(X_n,U_n)\right) = ((u_1,\delta_1),\dots,(u_n,\delta_n))\right\} = \sum_{i=1}^{n} P_\theta\left\{X_i = u_{(k_j)} \mid \Delta_i = \delta_i\right\}, \tag{2.34}$$

for $j = 1, \dots, m+1$, (2.33) can be written:

$$\frac{\sum_{i=1}^n P_\theta\left\{X_i = u_{(k_1)} \mid \Delta_i = \delta_i\right\}}{p_{k_1}} = \dots = \frac{\sum_{i=1}^n P_\theta\left\{X_i = u_{(k_{m+1})} \mid \Delta_i = \delta_i\right\}}{p_{k_{m+1}}}. \tag{2.35}$$

However, (2.35) implies

$$p_{k_j} = \frac{1}{n}\sum_{i=1}^{n} P_\theta\left\{X_i = u_{(k_j)} \mid \Delta_i = \delta_i\right\},\ j = 1, \dots, m+1. \tag{2.36}$$

Now, if the distribution function F_θ is defined by

$$F_\theta(t) = \sum_{j: u_{(k_j)} \le t} p_{k_j},$$

we get from (2.36):

$$F_\theta(t) = E_\theta\left\{\mathbb{F}_n(t) \mid T\left((X_1,U_1),\dots,(X_n,U_n)\right) = ((u_1,\delta_1),\dots,(u_n,\delta_n))\right\}, \tag{2.37}$$

where $\mathbb{F}_n$ denotes the empirical distribution function in the hidden space (see the transition from (2.8) to (2.11)). So indeed (1.9) leads us to (2.12). *Note however that the biggest problem* *in the maximum likelihood procedure, i.e., finding the points with strictly positive masses, is not addressed by this approach; we chose these points in advance.* This means that we have absolutely no guarantee that the distribution function found in this way is really the NPMLE! For checking that we really found the NPMLE, we also need *inequalities* like (2.29).

Example 2.3 (Right-censoring and the Kaplan-Meier estimator) In accordance with our approach so far, we will discuss the right-censoring model in the context of a mapping from a hidden space to an observation space. Let $(X_1, U_1), \dots, (X_n, U_n)$ be a sample of random variables in $\mathbb{R}_+^2$, where X_i and U_i are independent (non-negative) random variables with distribution functions F_0 and G, respectively. The observations available to us will be:

$$Y_i = X_i \wedge U_i \text{ and } \Delta_i = \{X_i \le U_i\},$$

where $X_i \wedge U_i = \min\{X_i, U_i\}$. The random variables (X_i, U_i) are living on the hidden space $\mathbf{X}$, the random variables (Y_i, Δ_i) on the observation space $\mathbf{Y}$, and our mapping T is given by:

$$T\left((x_1, u_1), \dots, (x_n, u_n)\right) = \left((x_1 \wedge u_1, \delta_1), \dots, (x_n \wedge u_n, \delta_n)\right), \text{ where } \delta_i = \{x_i \le u_i\}. \tag{2.38}$$

Very often the X_i have the interpretation of *survival times* and the U_i the interpretation of *censoring times*. As in the case of the preceding examples, we are going to construct an NPMLE of F_0 by restricting the allowed distribution functions to distribution functions with mass concentrated on a finite set, in this case $\{y_1, \dots, y_n\} \cup \{y_{(n+1)}\}$, where $y_i = x_i \wedge u_i$, $i = 1, \dots, n$, and where $y_{(n+1)}$ is an extra point to the right of all y_i's, $i \le n$, for extra mass (Pollard (1984) uses on p. 183 the catchy description "we dump the remaining mass all down on a fictitious supersurvivor at ∞", taking $y_{(n+1)} = \infty$). For simplicity we will assume that the y_i's are different.

The (part of the) likelihood (involving the distribution of the X_i) in the observation space is then given by

$$\prod_{i=1}^n f(y_i)^{\delta_i} \{1 - F(y_i)\}^{1-\delta_i}. \tag{2.39}$$

Note that the big difference with the likelihood in the current status model is the presence of the density f in (2.39): the factor $F(u_i)^{\delta_i}$ is now replaced by $f(y_i)^{\delta_i}$ (and of course the u_i have a different interpretation than the y_i, but this in itself has no effect on the maximization procedure).

Switching, as before, to the order statistics $y_{(i)}$, the log likelihood becomes:

$$\sum_{i=1}^n \Big\{ \delta_{(i)} \log f\left(y_{(i)}\right) + \{1 - \delta_{(i)}\} \log\left\{1 - F\left(y_{(i)}\right)\right\} \Big\}, \tag{2.40}$$

where $\delta_{(i)} = \delta_j$ if $y_{(i)} = y_j$. Contrary to the situation just discussed for the current status model, we now know exactly where to put the probability mass: it is concentrated on the set of points $\left\{y_{(k_1)},\dots,y_{(k_{m+1})}\right\}$, where $y_{(k_{m+1})}$ is the extra point $y_{(n+1)}$ if the last observation is censored ($\delta_{(n)}=0$), and where $y_{(k_{m+1})}=y_{(n)}$, otherwise, and where the other points $y_{(k_i)}$ are points corresponding to noncensored observations ($\delta_{(k_i)}=1$). So we take

$$\theta=(p_{k_1},\dots,p_{k_m})\,,\ \text{where } p_i=f\left(y_{(i)}\right),$$

as the parameter we want to estimate by maximum likelihood, and again express $p_{k_{m+1}}$ in terms of the p_{k_i}, $i\le m$, and do not include this parameter in our parameter vector θ.

The log likelihood in the "hidden space" is again given by:

$$\sum_{i=1}^{n}\log f_\theta(x_i),$$

where

$$f_\theta\left(y_{(k_i)}\right)=p_{k_i},\ i=1,\dots,m+1,$$

and, likewise, the score function $\dot{l}_\theta$ in the hidden space is given by

$$\dot{l}_\theta(x_1,\dots,x_n)=\left(\frac{n_1}{p_{k_1}}-\frac{n_{m+1}}{p_{k_{m+1}}},\dots,\frac{n_m}{p_{k_m}}-\frac{n_{m+1}}{p_{k_{m+1}}}\right), \tag{2.41}$$

where n_j is the number of x_i's that is equal to $y_{(k_j)}$. This leads to the equations

$$p_{k_j}=\frac{1}{n}\sum_{i=1}^{n}P_\theta\left\{X_i=y_{(k_j)}\;\middle|\;\Delta_i=\delta_i\right\},\ j=1,\dots,m+1. \tag{2.42}$$

(compare with (2.36)), and, if the distribution function F_θ is defined by

$$F_\theta(t)=\sum_{j:y_{(k_j)}\le t}p_{k_j},$$

we get from (2.42):

$$F_\theta(t)=E_\theta\left\{\mathbb{F}_n(t)\;\middle|\;T\left((X_1,U_1),\dots,(X_n,U_n)\right)=\left((y_1,\delta_1),\dots,(y_n,\delta_n)\right)\right\}, \tag{2.43}$$

But now it is very easy to solve these equations. The solution vector $\hat\theta=(\hat p_{k_1},\dots,\hat p_{k_m})$ satisfies:

$$\frac{\hat p_{k_i}}{\sum_{j=i}^{m+1}\hat p_{k_j}}=\frac{1}{n-k_i+1}\,,i=1,\dots,m. \tag{2.44}$$

The number $n-k_i+1$ is called "the size of the population at risk" just before time $y_{(k_i)}$ (in the literature on the right-censoring model). This leads to the *Kaplan-Meier estimator* $\hat F_n$ (sse Kaplan and Meier (1958)), defined by

$$1-\hat F_n(t)=\prod_{i:y_{(k_i)}\le t}\left(1-\frac{1}{n-k_i+1}\right). \tag{2.45}$$

In the present set-up, the Kaplan-Meier estimator is the NPMLE in the right-censoring model, and can also be defined by

$$\hat{F}_n(t) = \sum_{i: y_{(k_i)} \le t} \hat{p}_{k_i}, \tag{2.46}$$

where the $\hat{p}_{k_i}$ are defined by the equations (2.44). Hence the Kaplan-Meier estimator is another example of an estimator satisfying the self-consistency equations (in this case given by (2.43)).

References

Billingsley, P. (1995). Probability and Measure, 3rd edition, Wiley.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). *Maximum likelihood from incomplete data via the EM algorithm*, J. of the Royal Statistical Society B, vol. 39, 1-38.

Groeneboom, P. and Wellner, J.A. (1992). *Information bounds and nonparametric maximum likelihood estimation*, Birkhäuser Verlag.

Groeneboom, P. (1996). *Lectures on inverse problems*, in: Lectures on probability theory, Ecole d'Eté de Probabilités de Saint-Flour XXIV-1994, Editor: P. Bernard. Springer Verlag, Berlin.

Groeneboom, P. (1996). *Special topics course on inverse problems, 593C*, `http://www.stat.washington.edu/tech.reports`.

Kaplan, E.L. and Meier, P. (1958). *Nonparametric estimation from incomplete observations*, J. Amer. Statist. Assoc., vol. 53, 457-481.

Jongbloed, G. (1995). *Three statistical inverse problems*, Ph. D. dissertation, Delft University.

McLachlan, G.J. and Krishnan, T. (1997). The EM algorithm and Extensions. Wiley.

Pollard, D. (1984). Convergence of stochastic processes. Springer-Verlag.

Robertson,T., Wright, F.T. and Dykstra, R.L. (1988), *Order Restricted Statistical Inference*. Wiley, New York.

Rockafellar, R.T. (1970). Convex Analysis. Princeton University Press.