Two EM Algorithm Examples, STAT 818M

Eric Slud 2/18/2019

As described in class, here are one discrete and one continuous example of EM algorithm. Both are small examples where a straightforward numerical maximization of the log observed-data likelihood would be possible and work just as well as EM.

I. A Contingency-Table Example.

This example is a 2×3 contingency-table setup, but the structure that makes it work is exactly the same as the one-way parameterized multinomial discussed in the handout `https://www.math.umd.edu/ slud/s705/LecNotes/Sec6NotF16.pdf` linked on the web-page. The point in that general example is that the cell-probabilities $\pi_j(\theta),\ j = 1, \ldots, C$, are pameterized together through a shared parameter θ, and then treat the cell-counts $Y_{K+1}, \ldots, Y_C$ is individually unobservable, while their sum $X_{K+1} = \sum_{j=K+1}^{C} Y_j$ along with the individual counts $Y_j,\ j = 1, \ldots, K$ *are* observable. The only role of the two-way contingency table here is to make the choice of the parameter θ look sensible.

So consider a 'complete' data situation where counts $\mathbf{X}_{com} = \{X_{ij}, \quad i = 1, 2, \quad j = 1, \ldots, 3\}$ arranged in a 2-way table can be viewed as multinomial with a fixed known number $N = \sum_{i=1}^{2} \sum_{j=1}^{3} X_{ij}$ of trials, and probabilities

$$\pi_{11} = \alpha \pi_1, \ \ \pi_{ij} = \gamma \, p_j \qquad \text{for} \quad (i, j) \neq (1, 1)$$

where the unknown parameter is $\theta = (\alpha, \gamma, p_1, p_2, p_3)$ which is effectively 3-dimensional because of the two constraints

$$p_1 + p_2 + p_3 = 1, \qquad \alpha p_1 + \gamma(p_1 + 2p_2 + 2p_3) = 2\gamma + (\alpha - \gamma) p_1 = 1 \tag{1}$$

In this setting with complete data, the multinomial has 5 degrees of freedom but the parameter dimension is 4, so the parameter remains identifiable when the observable Data are $\mathbf{X}_{obs}$ given by

X_{11}	X_{12}	X_{13}
X_{21}	$X_{22} + X_{23}$	

In the complete-data setting (with all X_{ij} observable), it is easy to check that the log-likelihood is

$$\log L_{com}(\theta) \ = \ X_{11} \log(\alpha/\gamma) \ + \ \sum_{j=1}^{3} X_{+j} \log p_j \ + \ N \log \gamma$$

and after maximizing the Lagrange-multiplier expression

$$\log L_{com}(\theta) - \lambda(\sum_{j=1}^{3} p_j - 1) - \mu(2\gamma + p_1(\alpha - \gamma))$$

the MLE $\hat{\theta}$ is given (after differentiating and using the constraints) by

$$\hat{\lambda} = N, \quad \hat{\mu} \;=\; 2\hat{\gamma}N, \quad \hat{\alpha}\hat{p}_1 \;=\; \frac{X_{11}}{N}, \quad \hat{\gamma}\hat{p}_1 \;=\; \frac{X_{21}}{N}$$

$$\hat{\gamma} \;=\; \frac{N - X_{11} + X_{21}}{2\,N}, \quad \hat{p}_1 = \frac{2\,X_{21}}{N - X_{11} + X_{21}}, \quad \hat{p}_j = \frac{X_{+j}}{N - X_{11} + X_{21}}, \qquad j = 1, 2 \tag{2}$$

These formulas provide an explicit function $\hat{\theta}_{com} = g(\mathbf{X}_{com})$. When the observed data are $\mathbf{X}_{obs}$ as above, the log-likelihood becomes

$$\log L_{obs,1}(\theta) \;=\; X_{11}\log(\alpha p_1) \;+\; X_{21}\log(p_1) \;+\; X_{12}\log(p_2) \;+\; X_{13}\log(p_3)$$
$$+\;(X_{22} + X_{23})\log(p_2 + p_3) \;+\; (2N - X_{11})\log\gamma$$

The EM algorithm says to replace $\log L_{obs}(\theta)$ by $E_{\theta_0}(\log L_{com}(\theta)\,|\,\mathbf{X}_{obs})$, which requires only replacing X_{2j} in $\log L_{com}$ for $j = 2, 3$ by

$$E_{\theta_0}(X_{2j} \mid X_{22} + X_{23}) \;=\; (X_{22} + X_{23})\,p_{j,0}/(p_{2,0} + p_{3,0})$$

In this particular example, where we set up a sequence of successive EM iterations , we can see that all of the MLEs for $\hat{\alpha}$, $\hat{\gamma}$, $\hat{p}_1$ are constant in all iterations, along with $\hat{p}_2 + \hat{p}_3 \;=\; 1 - \hat{p}_1$, but that the successive EM iterations map an initial guess $p_{2,0}$ to

$$\frac{1}{N - X_{11} + X_{21}}\left(X_{12} + \frac{p_{2,0}}{1 - \hat{p}_1}\,(X_{22} + X_{23})\right)$$

The unique fixed-point for this mapping is easy to write down explicitly.

Within this same complete-data setting, another possibility for observed data would be

<table>
<tr><td>X_{11}</td><td>X_{12}</td><td>X_{13}</td></tr>
<tr><td colspan="2">$X_{21} + X_{22}$</td><td>X_{23}</td></tr>
</table>

In this setting, the observed -data log–likelihood is

$$\log L_{obs,2}(\theta) \;=\; X_{11}\log(\alpha p_1) \;+\; (X_{21}+X_{22})\log(p_1+p_2) \;+\; X_{12}\log(p_2) \\ + X_{13}\log(p_3) \;+\; X_{23}\log(p_3) \;+\; (2N - X_{11})\log\gamma$$

Exercise I.(a) Verify the complete-data MLE formulas (2).

(b). Give the explicit EM-iteration limit in the first observed-data setting $L_{obs,1}(\theta)$, and show directly that is the maximizer of $\log L_{obs,1}(\theta)$.

(c). For the specific observed-data table

<table>
<tr><td>30</td><td>25</td><td>45</td></tr>
<tr><td>50</td><td colspan="2">58</td></tr>
</table>

give the estimated Fisher information two ways: using the Louis (1982) formula (Thm 2.7 in the Kim-Shao book) and using the observed information $I_{obs}(\hat{\theta})$ from $L_{obs,1}(\theta)$.

We will show separate R-code for the EM-algorithm in the second observed-data setting $L_{obs,2}(\theta)$, where it turns out that there is no explicit observed-data MLE.

II. An Unbalanced ANOVA Example.

The second example presented in class is 2-way unbalanced ANOVA viewed as a missing-data problem. Define

$$X_{ij} \;=\; \mu \,+\, \alpha_j \,+\, \epsilon_{ij}, \quad j = 1, \ldots, m, \quad i = 1, \ldots, n_j$$

where $\alpha_j \sim \mathcal{N}(0, \sigma_a^2$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_e^2)$ are all jointly independent. Let $N = \sum_{j=1}^m n_j$ and $\theta = (\mu,\, \sigma_a^2,\, \sigma_e^2)$. The observed data in the example are $\mathbf{Y}_{obs} \;=\; \{X_{ij},\; 1 \le j \le m,\;\; 1 \le i \le n_j\}$; the complete or augmented data are $\mathbf{Y}_{com} \;=\; (\mathbf{Y}_{obs},\; \{\alpha_j\}_{j=1}^m)$; and the unknown parameters to be estimated are $\theta = (\mu, \sigma_a^2, \sigma_e^2)$.

The complete-data log-likelihood is $\log Lik_{com}(\theta) \;=$

$$-\frac{N+m}{2}\log(2\pi) \;-\; \frac{m}{2}\log(\sigma_a^2) \;-\; \frac{N}{2}\log(\sigma_e^2) \;-\; \frac{1}{2\sigma_a^2}\sum_{j=1}^m \alpha_j^2 \;-\; \frac{1}{2\sigma_e^2}\sum_{i,j}(X_{ij} - \alpha_j - \mu)^2$$

It is fairly straightforward to check that the complete-data MLEs are given by the formulas

$$\hat{\mu} \;=\; \frac{1}{N}\sum_{i,j}(X_{ij}-\alpha_j) \;=\; \frac{1}{N}\sum_{j=1}^{m} n_j\,(\bar{X}_{\cdot j}-\alpha_j), \quad \hat{\sigma}^2_a \;=\; \frac{1}{m}\,\alpha_j^2, \quad \hat{\sigma}^2_e \;=\; \frac{1}{N}\sum_{i,j}(X_{ij}-\hat{\mu}-\alpha_j)^2 \tag{3}$$

where

$$\bar{X}_{\cdot j} \;=\; \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij} \qquad \text{for} \qquad j=1,\ldots,m$$

To compute the conditional expected $\log Lik_{com}(\theta)$ under the model with parameters $\theta_0 = (\mu_0,\, \sigma^2_{a,0},\, (\sigma^2_{e,0})$ (the E-step) and maximize it over θ to define θ_1 (the M-step), we make the remark that conditionally given $\mathbf{Y}_{obs}$,

$$\alpha_j \;\sim\; \mathcal{N}\Big(\gamma_j^0(\bar{X}_{\cdot j}-\mu_0),\, \frac{\gamma_j^0}{n_j}\,\sigma^2_{e,0}\Big)\;, \quad \text{where} \qquad \gamma_j^0 \;\equiv\; \frac{\sigma^2_{a,0}}{(\sigma^2_{a,0}+\sigma^2_{e,0}/n_j)}$$

so that

$$E_{\theta_0}(\alpha_j^2\,|\,\mathbf{Y}_{obs}) \;=\; (\gamma_j^0)^2\,(\bar{X}_{\cdot j}-\mu_0)^2 \;+\; \frac{\gamma_j^0}{n_j}\,\sigma^2_{e,0}$$

$$E_{\theta_0}((X_{ij}-\mu-\alpha_j)^2\,|\,\mathbf{Y}_{obs}) \;=\; (X_{ij}-\mu-\gamma_j^0\,(\bar{X}_{\cdot j}-\mu_0))^2 \;+\; \frac{\gamma_j^0}{n_j}\,\sigma^2_{e,0}$$

It follows, after substituting these formulas into $E_{\theta_0}(\log L_{com}(\theta)\,|\,\mathbf{Y}_{obs})$, that the M-step equations are:

$$\mu_1 \;=\; \frac{1}{N}\sum_{j=1}^{m} n_j((1-\gamma_j^0)\bar{X}_{\cdot j}+\gamma_j^0\mu_0), \quad \sigma^2_{a,1} \;=\; \frac{1}{m}\sum_{j=1}^{m}\Big\{(\gamma_j^0)^2\,(\bar{X}_{\cdot j}-\mu_0)^2 \;+\; \frac{\gamma_j^0}{n_j}\,\sigma^2_{e,0}\Big\}$$

$$\sigma^2_{e,1} \;=\; \frac{1}{N}\sum_{j=1}^{m}\sum_{i=1}^{n_j}\Big\{(X_{ij}-\mu_1-\gamma_j^0(\bar{X}_{\cdot j}-\mu_0))^2 \;+\; \frac{\gamma_j^0}{n_j}\,\sigma^2_{e,0}\Big\}$$