Department of Statistics and Data Science at SUSTech

MAT7035: Computational Statistics

Tutorial 5: Optimization (II): The EM Algorithm

D. The EM Algorithm

D.1 Summary of the EM Algorithm

(a) Augment the observed data Y_{obs} with latent variables $\mathbf{z}$;

(b) Find the complete-data log-likelihood function $\ell(\boldsymbol{\theta}|Y_{\text{obs}}, \mathbf{z})$;

(c) Find the complete-data MLE $\hat{\boldsymbol{\theta}}_{\text{com}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(\boldsymbol{\theta}|Y_{\text{obs}}, \boldsymbol{z})$. Suppose this expression contains some function of $\boldsymbol{z}$, say $g(\boldsymbol{z})$;

(d) Find the conditional predictive density $f(\boldsymbol{z}|Y_{\text{obs}}, \boldsymbol{\theta})$;

(e) E-step: Let $\boldsymbol{\theta}^{(t)}$ be the t-th approximate of the MLE $\hat{\boldsymbol{\theta}}$. Then compute

$$E\big[g(\mathbf{z})|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}\big] = \int_{\mathbb{Z}} g(\boldsymbol{z}) f\big(\boldsymbol{z}|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}\big)\,\mathrm{d}\boldsymbol{z};$$

(f) M-step: Replace $g(\boldsymbol{z})$ in the expression of $\hat{\boldsymbol{\theta}}_{\text{com}}$ by $E\big[g(\mathbf{z})|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}\big]$ and get the updated $\boldsymbol{\theta}^{(t+1)}$.

D.2 Remarks

(a) Usually, $E[g(\mathbf{z})]$ and $g(E[\mathbf{z}])$ are different.

(b) z_i^2 in the expression of $\hat{\boldsymbol{\theta}}_{\text{com}}$ should be replaced by $E[Z_i^2|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}]$.

E. The Ascent Property of the EM Algorithm

E.1 The Ascent Property of the EM Algorithm

(a) The observed-data log-likelihood and the complete-data log-likelihood have the following relationship:

$$\ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \ell(\boldsymbol{\theta}|Y_{\text{obs}}, \mathbf{z}) - \log[f(\mathbf{z}|Y_{\text{obs}}, \boldsymbol{\theta})]$$

(b) Take expectations of the both sides,

$$\begin{aligned} E_{\mathbf{z}}\big[\ell(\boldsymbol{\theta}|Y_{\text{obs}})\big|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}\big] &= E_{\mathbf{z}}\big[\ell(\boldsymbol{\theta}|Y_{\text{obs}}, \mathbf{z})\big|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}\big] \\ &- E_{\mathbf{z}}\big\{\log[f(\mathbf{z}|Y_{\text{obs}}, \boldsymbol{\theta})]\big|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}\big\}. \end{aligned}$$

(c) Define

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E_{\mathbf{z}}\big[\ell(\boldsymbol{\theta}|Y_{\text{obs}}, \mathbf{z})\big|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}\big], \\ H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E_{\mathbf{z}}\big\{\log[f(\mathbf{z}|Y_{\text{obs}}, \boldsymbol{\theta})]\big|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}\big\}. \end{aligned}$$

(d) Since $\ell(\boldsymbol{\theta}|Y_{\text{obs}}) = E_{\mathbf{z}}\big[\ell(\boldsymbol{\theta}|Y_{\text{obs}})\big|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}\big]$, we have

$$\ell(\boldsymbol{\theta}|Y_{\text{obs}}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}).$$

(e) **Given $\boldsymbol{\theta}^{(t)}$, $H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ is maximized at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$.**

Hint: Use Jensen's inequality to $-\log(u)$, which is strictly convex.

Solution:

$$\begin{aligned} & H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \\ =& E_{\mathbf{z}}\big\{\log[f(\mathbf{z}|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})]\big|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}\big\} - E_{\mathbf{z}}\big\{\log[f(\mathbf{z}|Y_{\text{obs}}, \boldsymbol{\theta})]\big|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}\big\} \\ =& E_{\mathbf{z}}\left\{-\log\left[\frac{f(\mathbf{z}|Y_{\text{obs}}, \boldsymbol{\theta})}{f(\mathbf{z}|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})}\right]\middle|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}\right\} \\ \geqslant& -\log\left\{E_{\mathbf{z}}\left[\frac{f(\mathbf{z}|Y_{\text{obs}}, \boldsymbol{\theta})}{f(\mathbf{z}|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})}\right]\middle|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}\right\} \quad \text{(by Jensen's inequality)} \end{aligned}$$

$$= -\log\left[\int_{\mathbb{S}(\mathbf{z})} \frac{f(\boldsymbol{z}|Y_{\text{obs}},\boldsymbol{\theta})}{f(\boldsymbol{z}|Y_{\text{obs}},\boldsymbol{\theta}^{(t)})} f(\boldsymbol{z}|Y_{\text{obs}},\boldsymbol{\theta}^{(t)})\,\mathrm{d}\boldsymbol{z}\right]$$

$$= -\log\left[\int_{\mathbb{S}(\mathbf{z})} f(\boldsymbol{z}|Y_{\text{obs}},\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{z}\right]$$

$$= -\log 1 = 0. \qquad \|$$

(f) For all $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{(t)}$, we have

$$\ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) \leqslant \ell(\boldsymbol{\theta}|Y_{\text{obs}}) - Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}),$$

indicating that $\ell(\boldsymbol{\theta}|Y_{\text{obs}}) - Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ attains its minimum at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)} \in \boldsymbol{\Theta}$.

(g) <u>The ascent property of the EM algorithm</u>:

- Increasing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ causes an increase in $\ell(\boldsymbol{\theta}|Y_{\text{obs}})$.

—We choose $\boldsymbol{\theta}^{(t+1)}$ to maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$;

—Then

$$\ell(\boldsymbol{\theta}^{(t+1)}|Y_{\text{obs}}) - Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geqslant \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$$

$$\Rightarrow \quad \ell(\boldsymbol{\theta}^{(t+1)}|Y_{\text{obs}}) - \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}) \geqslant Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) \geqslant 0.$$

E.2 Remarks

(a) <u>Jensen's Inequality</u>:

- $\varphi(\cdot)$ is a convex function, and X is a r.v. taking values in the domain of $\varphi(\cdot)$.
- $\varphi[E(X)] \leqslant E[\varphi(X)]$ provided that both $E(X)$ and $E[\varphi(X)]$ exist.

(b) <u>Generalized EM alorithm (GEM)</u>:

- Standard EM algorithm: choosing $\boldsymbol{\theta}^{(t+1)}$ to maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.
- Generalized EM (GEM): only select a $\boldsymbol{\theta}^{(t+1)}$ with $Q(\boldsymbol{\theta}^{(t+1)}|\boldsymbol{\theta}^{(t)}) \geqslant Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$, but **not necessarily the maximum**.
- A step that increases Q function also increases ℓ, has the ascent property.

(c) Expectation/Conditional Maximization algorithm (ECM):

- **Replaces each M-step of the EM by a sequence of conditional maximization steps, i.e. CM-steps.**

— Suppose $\boldsymbol{\theta}$ could be divided into two parts $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$;

— Maximizing $Q(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2^{(t)}|\boldsymbol{\theta}^{(t)})$ over $\boldsymbol{\theta}_1$ to get $\boldsymbol{\theta}_1^{(t+1)}$;

— Maximizing $Q(\boldsymbol{\theta}_1^{(t+1)}, \boldsymbol{\theta}_2|\boldsymbol{\theta}^{(t)})$ over $\boldsymbol{\theta}_2$ to get $\boldsymbol{\theta}_2^{(t+1)}$.

Example T5.1 (Right censored regression model). Consider the linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{I}_m),$$

where $\boldsymbol{y} = (y_1, \ldots, y_m)^\top$ is the response vector, $\boldsymbol{X} = (\boldsymbol{x}_{(1)}, \ldots, \boldsymbol{x}_{(m)})^\top$ the covariate matrix, $\boldsymbol{\beta}$ and σ^2 the unknown parameters. Suppose that the first r components of $\boldsymbol{y}$ are uncensored and the remaining $m-r$ ones are right censored (c_i denotes a censored time). We augment the observed data $Y_{\text{obs}} = \{y_1, \ldots, y_r; c_{r+1}, \ldots, c_m\}$ with the unobserved uncensored times $\mathbf{z} = (Z_{r+1}, \ldots, Z_m)^\top$. If we had observed the value of $\mathbf{z}$, say $\boldsymbol{z} = (z_{r+1}, \ldots, z_m)^\top \equiv (y_{r+1}, \ldots, y_m)^\top$ with $z_i > c_i (i = r+1, \ldots, m)$, we could have the complete-data likelihood

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2|Y_{\text{obs}}, \boldsymbol{z}) &= \prod_{i=1}^m \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(y - \boldsymbol{x}_{(i)}^\top\boldsymbol{\beta})^2}{2\sigma^2} \right\} \\ &= (2\pi\sigma^2)^{-\frac{m}{2}} \exp\left\{ -\frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2} \right\}. \end{aligned}$$

Use EM algorithm to find the MLEs of $\boldsymbol{\beta}$ and σ^2.

Hint: (Truncated normal distribution, see Exercise 2.3). Let $X \sim \mathrm{TN}(\mu, \sigma^2; a, b)$, then

$$\begin{aligned} E(X) &= \mu + \sigma\frac{\phi(a_1) - \phi(b_1)}{\Phi(b_1) - \Phi(a_1)}, \\ \mathrm{Var}(X) &= \sigma^2\left[1 + \frac{a_1\phi(a_1) - b_1\phi(b_1)}{\Phi(b_1) - \Phi(a_1)}\right] - [E(X) - \mu]^2, \end{aligned}$$

where $a_1 = \frac{a-\mu}{\sigma}$ and $b_1 = \frac{b-\mu}{\sigma}$, ϕ and Φ are the pdf and cdf of $N(0,1)$, respectively.

Solution: The complete-data log-likelihood is

$$\ell(\boldsymbol{\beta}, \sigma^2 | Y_{\text{obs}}, \boldsymbol{z}) = -\frac{m}{2}\log(\sigma^2) - \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}.$$

To find the complete-data MLEs of $\boldsymbol{\beta}$ and σ^2, set

$$\frac{\partial \ell(\boldsymbol{\beta}, \sigma^2 | Y_{\text{obs}}, \boldsymbol{z})}{\partial \boldsymbol{\beta}} = 0 \quad \text{and} \quad \frac{\partial \ell(\boldsymbol{\beta}, \sigma^2 | Y_{\text{obs}}, \boldsymbol{z})}{\partial \sigma^2} = 0.$$

Since

$$\frac{\partial(\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = (\boldsymbol{y}^\top \boldsymbol{X})^\top \quad \text{and} \quad \frac{\partial(\boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta},$$

we obtain:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \boldsymbol{\beta}}\left[-\frac{m}{2}\log\sigma^2 - \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}\right] \\
&= -\frac{1}{2\sigma^2}\frac{\partial}{\partial \boldsymbol{\beta}}\Big(\boldsymbol{y}^\top \boldsymbol{y} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}\Big) \\
&= -\frac{1}{2\sigma^2}\Big[-2(\boldsymbol{y}^\top \boldsymbol{X})^\top + 2(\boldsymbol{X}^\top \boldsymbol{X})\boldsymbol{\beta}\Big],
\end{aligned}$$

which results in $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{y}$, where $\boldsymbol{y} = (y_1, \ldots, y_r; z_{r+1}, \ldots, z_m)^\top$.

On the other hand, from

$$\begin{aligned}
0 &= \frac{\partial}{\partial \sigma^2}\left[-\frac{m}{2}\log\sigma^2 - \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}\right] \\
&= -\frac{m}{2\sigma^2} + \frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2(\sigma^2)^2},
\end{aligned}$$

we obtain $\hat{\sigma}^2 = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/m$.

The conditional predictive density is the product of $(m - r)$ independent truncated normal densities:

$$f(\boldsymbol{z} | Y_{\text{obs}}, \boldsymbol{\beta}, \sigma^2) = \prod_{i=r+1}^{m} \text{TN}(z_i | \boldsymbol{x}_{(i)}^\top \boldsymbol{\beta}, \sigma^2; c_i, +\infty).$$

The E-step requires to calculate

$$\begin{aligned}
E[Z_i|Y_{\rm obs},\boldsymbol{\beta},\sigma^2] &= \boldsymbol{x}_{(i)}^\top\boldsymbol{\beta} + \sqrt{\sigma^2}\,\frac{\phi\Big(\frac{c_i-\boldsymbol{x}_{(i)}^\top\boldsymbol{\beta}}{\sqrt{\sigma^2}}\Big) - \phi\Big(\frac{+\infty-\boldsymbol{x}_{(i)}^\top\boldsymbol{\beta}}{\sqrt{\sigma^2}}\Big)}{\Phi\Big(\frac{+\infty-\boldsymbol{x}_{(i)}^\top\boldsymbol{\beta}}{\sqrt{\sigma^2}}\Big) - \Phi\Big(\frac{c_i-\boldsymbol{x}_{(i)}^\top\boldsymbol{\beta}}{\sqrt{\sigma^2}}\Big)} \\
&= \boldsymbol{x}_{(i)}^\top\boldsymbol{\beta} + \sqrt{\sigma^2}\,\frac{\phi\Big(\frac{c_i-\boldsymbol{x}_{(i)}^\top\boldsymbol{\beta}}{\sqrt{\sigma^2}}\Big)}{1-\Phi\Big(\frac{c_i-\boldsymbol{x}_{(i)}^\top\boldsymbol{\beta}}{\sqrt{\sigma^2}}\Big)}
\end{aligned}$$

and

$$\begin{aligned}
E[Z_i^2|Y_{\rm obs},\boldsymbol{\beta},\sigma^2] &= {\rm Var}[Z_i|Y_{\rm obs},\boldsymbol{\beta},\sigma^2] + (E[Z_i|Y_{\rm obs},\boldsymbol{\beta},\sigma^2])^2 \\
&= (\boldsymbol{x}_{(i)}^\top\boldsymbol{\beta})^2 + \sigma^2 + \sqrt{\sigma^2}(c_i+\boldsymbol{x}_{(i)}^\top\boldsymbol{\beta})\frac{\phi\Big(\frac{c_i-\boldsymbol{x}_{(i)}^\top\boldsymbol{\beta}}{\sqrt{\sigma^2}}\Big)}{1-\Phi\Big(\frac{c_i-\boldsymbol{x}_{(i)}^\top\boldsymbol{\beta}}{\sqrt{\sigma^2}}\Big)}
\end{aligned}$$

for $i = r+1,\ldots,m$, where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the pdf and cdf of $N(0,1)$, respectively.

For the M-step,

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y} \quad \text{and} \quad \hat{\sigma}^2 = \frac{(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})^\top(\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})}{m}$$

are updated by replacing $\{z_i\}_{i=r+1}^m$ in $\boldsymbol{y} = (y_1,\ldots,y_r; z_{r+1},\ldots,z_m)^\top$ with $E[Z_i|Y_{\rm obs},\boldsymbol{\beta},\sigma^2]$ and $\{z_i^2\}_{i=r+1}^m$ in $\boldsymbol{y}^\top\boldsymbol{y} = \sum_{i=1}^r y_i^2 + \sum_{i=r+1}^m z_i^2$ with $E[Z_i^2|Y_{\rm obs},\boldsymbol{\beta},\sigma^2]$. $\|$

Example T5.2 (Cell probability vector). Let $Y_{\rm obs} = \{n_1,n_2,n_3,n_4;m_1,m_2\}$ denote the observed frequencies and $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_4)^\top$ be the cell probability vector satisfying $\theta_i \geqslant 0, \theta_1+\cdots+\theta_4 = 1$. Suppose that the observed-data likelihood function of $\boldsymbol{\theta}$ is given by

$$L(\boldsymbol{\theta}|Y_{\rm obs}) \propto \left(\prod_{i=1}^4 \theta_i^{n_i}\right)(\theta_1+\theta_2)^{m_1}(\theta_1+\theta_2+\theta_3)^{m_2}.$$

Use the EM algorithm to find the maximum likelihood estimator of $\boldsymbol{\theta}$.

Solution: To split the term $(\theta_1+\theta_2)^{m_1}$, we introduce a latent variable W following the conditional predictive distribution

$$W|(Y_{\rm obs},\boldsymbol{\theta}) \sim \text{Binomial}\left(m_1, \frac{\theta_1}{\theta_1+\theta_2}\right).$$

To split the term $(\theta_1+\theta_2+\theta_3)^{m_2}$, we introduce a latent vector $\mathbf{z}=(Z_1,Z_2,Z_3)^\top$ following the conditional predictive distribution

$$\mathbf{z}|(Y_{\text{obs}},\boldsymbol{\theta}) \sim \text{Multinomial}_3\left(m_2; \frac{\theta_1}{\theta_1+\theta_2+\theta_3}, \frac{\theta_2}{\theta_1+\theta_2+\theta_3}, \frac{\theta_3}{\theta_1+\theta_2+\theta_3}\right).$$

Then the complete-data likelihood function is given by

$$L(\boldsymbol{\theta}|Y_{\text{obs}},W,\mathbf{z}) \propto \theta_1^{n_1+W+Z_1}\theta_2^{n_2+m_1-W+Z_2}\theta_3^{n_3+m_2-Z_1-Z_2}(1-\theta_1-\theta_2-\theta_3)^{n_4}.$$

The complete-data log-likelihood function without the constant term is

$$\begin{aligned}\ell(\boldsymbol{\theta}|Y_{\text{obs}},W,\mathbf{z}) &= (n_1+W+Z_1)\log\theta_1+(n_2+m_1-W+Z_2)\log\theta_2 \\ &\quad +(n_3+m_2-Z_1-Z_2)\log\theta_3+n_4\log(1-\theta_1-\theta_2-\theta_3).\end{aligned}$$

With $\nabla\ell(\boldsymbol{\theta}|Y_{\text{obs}},W,Z)=0$, we have the complete data MLE of $\boldsymbol{\theta}$:

$$\begin{aligned}\hat{\theta}_1 &= \frac{n_1+W+Z_1}{\sum_{i=1}^4 n_i+m_1+m_2}, \\ \hat{\theta}_2 &= \frac{n_2+m_1-W+Z_2}{\sum_{i=1}^4 n_i+m_1+m_2}, \\ \hat{\theta}_3 &= \frac{n_3+m_2-Z_1-Z_2}{\sum_{i=1}^4 n_i+m_1+m_2}, \\ \hat{\theta}_4 &= 1-\hat{\theta}_1-\hat{\theta}_2-\hat{\theta}_3=\frac{n_4}{\sum_{i=1}^4 n_i+m_1+m_2}.\end{aligned}$$

E-step:

$$\begin{aligned}E(W|Y_{\text{obs}},\boldsymbol{\theta}) &= \frac{m_1\theta_1}{\theta_1+\theta_2}, \\ E(Z_1|Y_{\text{obs}},\boldsymbol{\theta}) &= \frac{m_2\theta_1}{\theta_1+\theta_2+\theta_3}, \\ E(Z_2|Y_{\text{obs}},\boldsymbol{\theta}) &= \frac{m_2\theta_2}{\theta_1+\theta_2+\theta_3}.\end{aligned}$$

M-step:

$$\theta_1^{(t+1)} = \frac{n_1 + E(W|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}) + E(Z_1|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})}{\sum_{i=1}^4 n_i + m_1 + m_2},$$

$$\theta_2^{(t+1)} = \frac{n_2 + m_1 - E(W|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}) + E(Z_2|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})}{\sum_{i=1}^4 n_i + m_1 + m_2},$$

$$\theta_3^{(t+1)} = \frac{n_3 + m_2 - E(Z_1|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)}) - E(Z_2|Y_{\text{obs}}, \boldsymbol{\theta}^{(t)})}{\sum_{i=1}^4 n_i + m_1 + m_2},$$

$$\theta_4^{(t+1)} = \frac{n_4}{\sum_{i=1}^4 n_i + m_1 + m_2}. \qquad \|$$

Example T5.3 (Possion additive model). Let $Y_i \stackrel{\text{ind}}{\sim} \text{Poisson}(\boldsymbol{a}_i^\top \boldsymbol{\theta}), 1 \leqslant i \leqslant n$, where $\boldsymbol{a}_i = (a_{i1}, \dots, a_{ip})^\top$ is known vector and each element is nonnegative. The aim is to estimate the unknown parameter vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^\top$ based on the observed data $Y_{\text{obs}} = \{y_i\}_{i=1}^n$.

(a) Find the log-likelihood function, the score vector and the observed information matrix. Then, write down the iteration formula of Newton–Raphson Method.

(b) For any i, we introduce a latent vector $\mathbf{z}_i = (Z_{i1}, Z_{i2}, \dots, Z_{ip})^\top$ by splitting Y_i and $\boldsymbol{a}_i^\top \boldsymbol{\theta}$ as follows:

$$\begin{aligned} Y_i &= Z_{i1} + \cdots + Z_{ij} + \cdots + Z_{ip}, \\ \boldsymbol{a}_i^\top \boldsymbol{\theta} &= a_{i1}\theta_1 + \cdots + a_{ij}\theta_j + \cdots + a_{ip}\theta_p. \end{aligned}$$

where $Z_{ij} \stackrel{\text{ind}}{\sim} \text{Poisson}(a_{ij}\theta_j)$. Use EM algorithm to find the maximum likelihood estimator of $\boldsymbol{\theta}$.

Solution: (a) The observed-data likelihood function of $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}|Y_{\text{obs}}) = \prod_{i=1}^n \frac{(\boldsymbol{a}_i^\top \boldsymbol{\theta})^{y_i} \exp(-\boldsymbol{a}_i^\top \boldsymbol{\theta})}{y_i!}.$$

The observed-data log-likelihood function without the constant term is

$$\ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \sum_{i=1}^{n} [y_i \log(\boldsymbol{a}_i^\top \boldsymbol{\theta}) - \boldsymbol{a}_i^\top \boldsymbol{\theta}].$$

The score vector is

$$\nabla\ell(\boldsymbol{\theta}|Y_{\text{obs}}) = \sum_{i=1}^{n} \left(y_i \frac{\boldsymbol{a}_i}{\boldsymbol{a}_i^\top \boldsymbol{\theta}} - \boldsymbol{a}_i \right) = \sum_{i=1}^{n} \left[\left(y_i \frac{1}{\boldsymbol{a}_i^\top \boldsymbol{\theta}} - 1 \right) \boldsymbol{a}_i \right].$$

The observed information matrix is

$$-\nabla^2 \ell(\boldsymbol{\theta}|Y_{\text{obs}}) = -\frac{\partial \nabla \ell(\boldsymbol{\theta}|Y_{\text{obs}})}{\partial \boldsymbol{\theta}^\top} = \sum_{i=1}^{n} \left[\frac{y_i}{(\boldsymbol{a}_i^\top \boldsymbol{\theta})^2} \boldsymbol{a}_i \boldsymbol{a}_i^\top \right].$$

Thus the Newton–Raphson algorithm is defined by

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + [-\nabla^2 \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}})]^{-1} \nabla \ell(\boldsymbol{\theta}^{(t)}|Y_{\text{obs}}).$$

(b) Let $Y_{\text{mis}} = \{\mathbf{z}_i\}_{i=1}^n$ and $Y_{\text{com}} = \{Y_{\text{obs}}, Y_{\text{mis}}\} = Y_{\text{mis}}$. The complete-data likelihood function is

$$L(\boldsymbol{\theta}|Y_{\text{com}}) = \prod_{i=1}^{n} \prod_{j=1}^{p} \frac{(a_{ij}\theta_j)^{Z_{ij}} \exp(-a_{ij}\theta_j)}{Z_{ij}!} \propto \prod_{j=1}^{p} \left[\theta_j^{\sum_{i=1}^n Z_{ij}} \exp\left(-\theta_j \sum_{i=1}^{n} a_{ij} \right) \right].$$

The complete-data log-likelihood function without the constant term is

$$\ell(\boldsymbol{\theta}|Y_{\text{com}}) = \sum_{j=1}^{p} \left(\sum_{i=1}^{n} Z_{ij} \right) \log(\theta_j) - \sum_{j=1}^{p} \theta_j \left(\sum_{i=1}^{n} a_{ij} \right).$$

Setting $\nabla \ell(\boldsymbol{\theta}|Y_{\text{com}}) = 0$, we have the complete data MLE of $\boldsymbol{\theta}$,

$$\hat{\theta}_j = \frac{\sum_{i=1}^n Z_{ij}}{\sum_{i=1}^n a_{ij}},$$

where $1 \leqslant j \leqslant p$. The conditional predictive distribution is

$$\mathbf{z}_i | (Y_{\text{obs}}, \boldsymbol{\theta}) \sim \text{Multinomial}_p \Big(y_i; \frac{a_{i1}\theta_1}{\boldsymbol{a}_i^\top \boldsymbol{\theta}}, \ldots, \frac{a_{ip}\theta_p}{\boldsymbol{a}_i^\top \boldsymbol{\theta}} \Big).$$

where $i = 1, \ldots, n$ and $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are independent.

$$E(Z_{ij}|Y_{\text{obs}},\boldsymbol{\theta}) = \frac{y_i a_{ij}\theta_j}{\boldsymbol{a}_i^\top\boldsymbol{\theta}}, \quad 1 \leqslant i \leqslant n,\ 1 \leqslant j \leqslant p.$$

M-step:

$$\theta_j^{(t+1)} = \theta_j^{(t)} \frac{\sum_{i=1}^n \left[y_i a_{ij}/(\boldsymbol{a}_i^\top\boldsymbol{\theta}^{(t)})\right]}{\sum_{i=1}^n a_{ij}}, \quad 1 \leqslant j \leqslant p.$$

‖