第 4 章

EM 优化算法

杜江

北京工业大学

2021 年 3 月 14 日

目录

前言

期望最大化（EM）算法是一种迭代优化策略，它是受**缺失思想**以及考虑给定已知项下**缺失项的条件分布**而激发产生的。EM 算法的普及源自于它能非常简单地执行并且能通过稳定、上升的步骤非常可靠地找到全局最优值。

考虑由随机变量 X 生成的观测数据和来自随机变量 Z 的缺失或未观测数据。设完全数据 Y=（X，Z），给定观测数据 x，若想求极大似然估计，直接使用似然函数 $L(\theta|x)$ 会难以处理。EM 算法通过采用 $Y|\theta$ 和 $Z|(x,\theta)$ 这些较容易的密度避开直接考虑 $L(\theta|x)$。

4.1 缺失数据、边际化和符号

缺失数据可能不是真的缺少，它们可能仅是一个简化问题概念上的策略。这种情况下，Z 通常称为**潜数据**。

无论认为 Z 是潜在或缺失的，都可认为它是从完整的 Y 中删除了。这样潜在或缺失数据的假设等价于一个边际化模型，从模型中可以观察到：

$$f_X(x|\theta) = \int_{\{y:M(y)=x\}} f_Y(y|\theta)dy$$

以及

$$f_{Z|X}(z|x,\theta) = f_Y(y|\theta)/f_X(x|\theta)$$

在 θ 后验密度的贝叶斯应用中，以下两种方式下可以考虑用后验表示一个更宽泛问题的边际化：

1. 把似然函数 $L(\theta|x)$ 看作完全数据 Y 的似然函数 $L(\theta|y) = L(\theta|x, z)$ 的一个边际化。这时缺失数据是 z，采用上面相同的一类符号。
2. 可以考虑有缺失参数 ϕ 来简化贝叶斯计算。在采用频率论者的记号时，可以用后验和 ϕ 代替似然函数和 Z 来考虑贝叶斯学派的观点。

4.2 EM 算法

定义 $Q(\theta|\theta^{(t)})$ 为观测数据 $X = x$ 条件下完全数据的联合对数似然的期望：

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E\{logL(\theta|Y)|x, \theta^{(t)}\} \\ &= E\{logf_Y(y|\theta)|x, \theta^{(t)}\} \\ &= \int [logf_Y(y|\theta)] f_{Z|X}(z|x, \theta^{(t)}) dz \end{aligned}$$

EM 从 $\theta^{(0)}$ 开始，在两步之间交换：E 表示期望，M 表示最大化。该算法可概括为以下三步：

1. E 步：计算 $Q(\theta|\theta^{(t)})$。
2. M 步：关于 θ 最大化 $Q(\theta|\theta^{(t)})$。设 $\theta^{(t+1)}$ 为 Q 的最大值点。
3. 返回 E 步，知道满足某停止规则为止。

停止规则通常依赖 $(\theta^{(t+1)} - \theta^{(t)})^T(\theta^{(t+1)} - \theta^{(t)})$ 或 $|Q((\theta^{(t+1)}|\theta^{(t)}) - Q((\theta^{(t)}|\theta^{(t)})|$。

例 4.1(简单的指数密度)

考虑到 $Y_1, Y_2 \sim i.i.d.Exp(\theta)$。假定 $y_1 = 5$ 是观测到的，y_2 值是缺失的。则完全数据的对数似然函数：

$$logL(\theta|y) = logf_Y(y|\theta) = 2log\{\theta\} - \theta y_1 - \theta y_2$$

取其条件期望：

$$Q(\theta|\theta^{(t)}) = 2log\{\theta\} - 5\theta - \theta/\theta^{(t)}$$

对 $Q(\theta|\theta^{(t)})$ 取极大值点得到关于 θ 的更新方程为：

$$\theta^{(t+1)} = 2\theta^{(t)}/(5\theta^{(t)} + 1)$$

取某初值后可知估计收敛到 $\hat{\theta} = 0.2$。

例 4.2(椒花蛾)

示例介绍：椒花蛾又叫桦尺蛾，这些蛾子的色彩已确认由某单个基因决定，该基因具有三个可能的等位基因，我们记为 C，I 和 T。三者之中，C 对 I 是显性的，而 T 对 I 是隐性的。因此基因型 CC，CI 和 CT 导致黑化 (carbonaria) 表型，它呈现纯黑色。基因型 TT 导致典型 (typica) 表型，它呈现浅色图案的翅膀。基因型 II 和 IT 产生一个称之为岛屿 (insularia) 的中间表型，它在外观上变化很广泛，但通常以中间色彩杂色而成。这样，有六种可能的基因型，但只有三种基因型在田间工作中是可测的。

如果种群中等位基因的频率为 p_C, p_I 和 p_T，那么基因型 CC，CI，CT，II，IT 和 TT 的频率应分别为 $p_C^2, 2p_Cp_I, 2p_Cp_T, p_I^2, 2p_Ip_T$ 和 p_T^2 且满足 $p_C + p_I + p_T = 1$。

假定捕获到 n 只蛾子，其中黑化、岛屿和典型表型的分别有 n_C, n_I 和 n_T 只。于是 $n_C + n_I + n_T = n$。因为每只蛾子在讨论的基因上有两个等位基因，样本中一共有 2n 个等位基因。如果我们知道每只蛾子的基因型而不仅仅是它的表型，就能生成基因型数 $n_{CC}, n_{CI}, n_{CT}, n_{II}, n_{IT}$ 和 n_{TT}，由它们较容易列出等位基因的频率。例如，有基因型 CI 的每只蛾子贡献一个 C 等位基因和一个 I 等位基因，而一个 II 型的蛾子贡献两个 I 等位基因。这样的等位基因数会立刻提供 p_C, p_I 和 p_T 的估计。仅由表型个数如何估计等位基因频率还是远不明朗的。

例 4.2(椒花蛾)

在 EM 符号下，椒花蛾示例观测数据为 $x=(n_C,n_I,n_T)$，完全数据为 $y=(n_{CC},n_{CI},n_{CT},n_{II},n_{IT},n_{TT})$。完全数据的对数似然函数为：

$$\begin{aligned} logf_Y(y|p) =& n_{CC}log\{p_C^2\}+n_{CI}log\{2p_Cp_I\}+n_{CT}log\{2p_Cp_T\} \\ &+n_{II}log\{p_I^2\}+n_{IT}log\{2p_Ip_T\}+n_{TT}log\{p_T^2\} \\ &+log\left(\begin{array}{cccccc} & & n & & & \\ n_{CC} & n_{CI} & n_{CT} & n_{II} & n_{IT} & n_{TT} \end{array}\right) \end{aligned}$$

完全数据不是都能观测到的。设 $Y=(N_{CC},N_{CI},N_{CT},N_{II},N_{IT},n_{TT})$，已知 $N_{TT}=n_{TT}$，但其他频率不是直接观测到的。

为计算 $Q(p|p^{(t)})$，在条件 n_C 和 $p^{(t)}=(p_C^{(T)},p_I^{(T)})$ 下，三种黑化基因型的潜在数目有一个三元多项式分布，两个岛屿单元也有类似结果。

例 4.2(椒花蛾)

似然函数中前五个随机部分的期望值为:

$$E\{N_{CC}|n_C, n_I, n_T, p^{(t)}\} = n_{CC}^{(t)} = \frac{n_C(p_C^{(t)})^2}{(p_C^{(t)})^2 + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}}$$

$$E\{N_{CI}|n_C, n_I, n_T, p^{(t)}\} = n_{CI}^{(t)} = \frac{2n_C p_C^{(t)} p_I^{(t)}}{(p_C^{(t)})^2 + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}}$$

$$E\{N_{CT}|n_C, n_I, n_T, p^{(t)}\} = n_{CT}^{(t)} = \frac{2n_C p_C^{(t)} p_T^{(t)}}{(p_C^{(t)})^2 + 2p_C^{(t)}p_I^{(t)} + 2p_C^{(t)}p_T^{(t)}}$$

$$E\{N_{II}|n_C, n_I, n_T, p^{(t)}\} = n_{II}^{(t)} = \frac{n_I(p_I^{(t)})^2}{(p_I^{(t)})^2 + 2p_I^{(t)}p_T^{(t)}}$$

$$E\{N_{IT}|n_C, n_I, n_T, p^{(t)}\} = n_{IT}^{(t)} = \frac{2n_I p_I^{(t)} p_T^{(t)}}{(p_I^{(t)})^2 + 2p_I^{(t)}p_T^{(t)}}$$

例 4.2(椒花蛾)

似然函数中的多项式系数有一个条件期望，比如 $k(n_C, n_I, n_T, p^{(t)})$，它不依赖于 p。于是有：

$$\begin{aligned} Q(p|p^{(t)}) =& n_{CC}^{(t)} log\{p_C^2\} + n_{CI}^{(t)} log\{2p_C p_I\} \\ &+ n_{CT}^{(t)} log\{2p_C p_T\} + n_{II}^{(t)} log\{p_I^2\} \\ &+ n_{IT}^{(t)} log\{2p_I p_T\} + n_{TT} log\{p_T^2\} + k(n_C, n_I, n_T, p^{(t)}) \end{aligned}$$

关于 p_C 和 p_I 求极大似然估计完成 M 步，得到：

$$p_C^{(t+1)} = \frac{2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}}{2n}$$

$$p_I^{(t+1)} = \frac{2n_{II}^{(t)} + n_{IT}^{(t)} + n_{CI}^{(t)}}{2n}$$

$$p_T^{(t+1)} = \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{2n}$$

例 4.2(椒花蛾)

表 4.1 说明了 EM 算法如何收敛到极大似然估计。并且后三列给出了收敛性的诊断。

```
> b4_1
  p[C]^(t) p[I]^(t)      R^(t) D[C]^(t) D[I]^(t)
0 0.333333 0.333333         NA        NA        NA
1 0.081994 0.237406 0.5706852    0.0425     0.337
2 0.071249 0.197870 0.1631212    0.0369     0.188
3 0.070852 0.190360 0.0357560    0.0367     0.178
4 0.070837 0.189023 0.0065860    0.0367     0.176
5 0.070837 0.188787 0.0011683    0.0367     0.176
6 0.070837 0.188745 0.0002058    0.0367     0.176
7 0.070837 0.188738 0.0000362    0.0367     0.176
8 0.070837 0.188737 0.0000064    0.0367     0.176
```

例 4.2(椒花蛾)

假定观测到的基因型数目为 $n_C = 85, n_I = 196$ 及 $n_T = 341$。表 4.1 说明了 EM 算法如何收敛到极大似然估计，大约为 $\hat{p}_C = 0.07804, \hat{p}_I = 0.18874$ 及 $\hat{p}_T = 0.74043$。

相对收敛准则

$$R^{(t)} = \frac{||p^{(t)} - p^{(t-1)}||}{||p^{(t-1)}||}$$

概括了由下一次迭代到下一次迭代在 $p^{(t)}$ 上相对改变的总量。为了说明，还给出了 $D_C^{(t)} = (p_C^{(t)} - \hat{p}_C)/(p_C^{(t-1)} - \hat{p}_C)$ 和类似的变量 $D_I^{(t)}$。

例 4.2(贝叶斯后验众数)

考虑一个具有似然函数、先验函数以及缺失数据或缺失参数的贝叶斯问题。为了找到后验函数，E 步需要：

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E\{log\{L(\theta|Y)f(\theta)k(Y)\}|x,\theta^{(t)}\} \\ &= E\{log\{L(\theta|Y)\}|x,\theta^{(t)}\} + logf(\theta) + E\{logk(Y)|x,\theta^{(t)}\} \end{aligned}$$

上式的最后一项是一个可以忽略的归一化常数，因为 Q 是要关于 θ 的最大化。

4.2 EM 算法
4.2.1 收敛性

注意到观测数据密度的对数可重新表示为：

$$logf_X(x|\theta) = logf_Y(y|\theta) - logf_{Z|X}(z|x,\theta)$$

那么：

$$logf_X(x|\theta) = Q(\theta|\theta^{(t)}) - H(\theta|\theta^{(t)})$$

其中 $H(\theta|\theta^{(t)})$ 是 $logf_{Z|X}(z|x,\theta)$ 关于 $Z|(x,\theta^{(t)})$ 的期望。

可以看到，当 $\theta = \theta^{(t)}$ 时 $H(\theta|\theta^{(t)})$ 关于 θ 取得最大值。用 $\theta^{(t+1)}$ 来关于 θ 最大化 $Q(\theta|\theta^{(t)})$ 构成标准 EM 算法，得到该结果后，考虑收敛阶数：

定义映射 $\theta^{(t+1)} = \Psi(\theta^{(t)})$。当 EM 收敛时，如果收敛到该映射的不动点，那么 $\hat{\theta} = \Psi(\hat{\theta})$，设 $\Psi'(\theta)$ 表示 Jacobi 矩阵，则 Ψ 的 Taylor 级数展开为：

$$\theta^{(t+1)} - \hat{\theta} \approx \Psi'(\theta^{(t)})(\theta^{(t)} - \hat{\theta})$$

可以看到，当 $p = 1$ 时 EM 算法线性收敛。$p > 1$ 时，若 $-l''(\hat{\theta}|x)$ 是正定的，仍线性收敛。

为进一步理解 EM 如何工作，注意到：

$$l(\theta|x) \geq Q(\theta|\theta^{(t)}) + l(\theta^{(t)}|x) - Q(\theta^{(t)}|\theta^{(t)}) = G(\theta|\theta^{(t)})$$

其中 G 在 $\theta^{(t)}$ 处与 l 相切，图 4.1 给出了 EM 算法的工作思想。

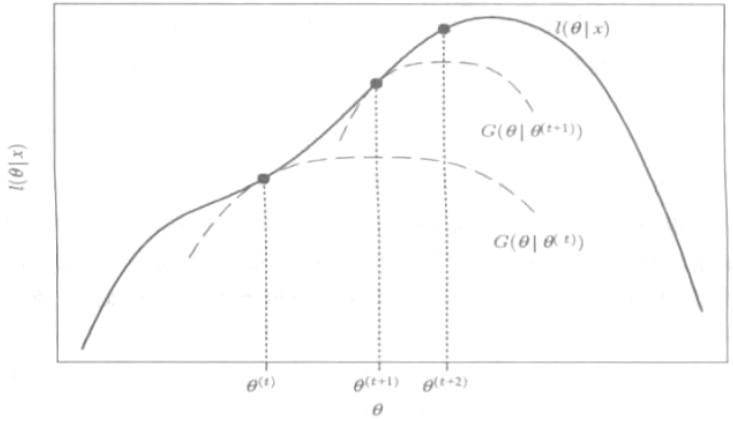


图 4.1　作为一种劣化或优化转换策略的 EM 算法的一维图示

4.2 EM 算法
4.2.2 在指数族中的应用

当完全数据被建模为具有指数分布时，E 步得出：

$$Q(\theta|\theta^{(t)}) = k + logc_2(\theta) + \int \theta^T s(y) f_{Z|X}(z|x, \theta^{(t)}) dz$$

为了实现 M 步，设 Q 关于 θ 的梯度等于 0，重新整理后得到：

$$\frac{-c_2'(\theta)}{c_2(\theta)} = \int s(y) f_{Z|X}(z|x, \theta^{(t)}) dz$$

上式意味着 M 步是通过设 $\theta^{(t+1)}$ 等于求解

$$E\{s(Y)|\theta\} = \int s(y) f_{Z|X}(z|x, \theta^{(t)}) dz$$

得到的 θ 而完成。

总的来看，指数族的 EM 算法由下面步骤组成：

1. E 步：给定观测数据并利用现有参数估计 $\theta^{(t)}$，计算完全数据充分统计量的期望。
2. M 步：通过求解 $E\{s(Y)|\theta\} = s^{(t)}$ 得到 $\theta^{(t+1)}$。
3. 返回 E 步，直到满足某收敛准则为止。

4.2 EM 算法
4.2.3 方差估计

EM 算法用来寻找一个极大似然估计，但并不自动产生协方差阵的估计。

那么如何估计协方差阵？

- Louis 方法
- SEM 算法（推荐）
- 自助法 (Bootstrapping)（推荐）
- 经验信息
- 数值微分

4.2.3.1 Louis 方法

考虑等式：

$$-l''(\theta|x) = -Q''(\theta|\omega)|_{\omega=\theta} + H''(\theta|\omega)|_{\omega=\theta}$$

其中 Q'' 和 H'' 表示 Q 和 H 关于 θ 的二阶导。该等式可写成：

$$\hat{i}_X(\theta) = \hat{i}_Y(\theta) - \hat{i}_{Z|X}(\theta)$$

上式从左到右三项分别为观测信息、完全信息和缺失信息。该式引入缺失信息法则：观测信息等于完全信息减去缺失信息。

该法则可用来得到 $\hat{\theta}$ 的协方差阵的一个估计。

4.2.3.1 Louis 方法

例 4.5（删失的指数数据）：　假定在 $Exp(\lambda)$ 下观测到完全数据，但有些是右删失的。那么观测数据 $x=(x_1,...,x_n)$，其中 $x_i=(min(y_i,c_i),\delta_i)$, c_i 是删失水平，如果 $y_i \le c_i$，则 $\delta_i=1$，否则 $\delta_i=0$。

得到：

$$
\begin{aligned}
Q(\lambda|\lambda^{(t)}) &= E(l(\lambda|Y_1,...,Y_n)|x,\lambda^{(t)}) \\
&= nlog\lambda - \lambda\sum_{i=1}^{n} E\{Y_i|x_i,\lambda^{(t)}\} \\
&= nlog\lambda - \lambda\sum_{i=1}^{n}[y_i\delta_i + (c_i+\frac{1}{\lambda^{(t)}})(1-\delta_i)] \\
&= nlog\lambda - \lambda\sum_{i=1}^{n}[y_i\delta_i + c_i(1-\delta_i)] - \frac{C\lambda}{\lambda^{(t)}}
\end{aligned}
$$

4.2.3.1 Louis 方法

Z_i 有密度 $f_{Z_i|X}(z_i|x,\lambda) = \lambda exp\{-\lambda(z_i - c_i)\}1_{\{z_i > c_i\}}$。利用等式

$$\hat{i}_{Z|X}(\theta) = var\{\frac{dlogf_{Z|X}(Z|x,\theta)}{d\theta}\}$$

计算缺失信息：

$$\hat{i}_{Z|X}(\lambda) = \sum_{\{i:\delta_i=0\}} var\{Z_i - c_i\} = \frac{C}{\lambda^2}$$

这样通过 Louis 方法：

$$\hat{i}_X(\lambda) = \frac{n}{\lambda^2} - \frac{C}{\lambda^2} = \frac{U}{\lambda^2}$$

其中 $U = \Sigma_{i=1}^{n}\delta_i$ 表示未删失事件的个数。容易验证 $-l''(\lambda|x) = U/\lambda^2$。

4.2.3.2 SEM 算法

用 Ψ 表示 EM 的映射，Jacobi 矩阵 $\Psi'(\theta)$，有等式：

$$\Psi'(\hat{\theta})^T = \hat{i}_{Z|X}(\hat{\theta})\hat{i}_Y(\hat{\theta})^{-1}$$

将缺失信息法换一种表达方式，将上式带入并对 $\hat{i}_X(\hat{\theta})$ 求逆得到估计：

$$\widehat{var}\{\hat{\theta}\} = \hat{i}_Y(\hat{\theta})^{-1}(I + \Psi'(\hat{\theta})^T[I - \Psi'(\hat{\theta})^T]^{-1})$$

这个结果很有意义，因为它把协方差阵表示为完全数据协方差阵加一个考虑缺失数据的不确定性的增量矩阵。

用 SEM 算法得到 $\hat{\Psi}'(\hat{\theta})$ 的估计有如下步骤：

1. 运行 EM 算法直至收敛，找到极大值点 $\hat{\theta}$
2. 从 $\theta^{(0)}$ 重新开始运行算法。最好选择靠近 $\hat{\theta}$ 的 $\theta^{(0)}$。

4.2.3.3 自助法

用自助法对独立同分布的观测数据 $x_1, ..., x_n$ 得到协方差阵的估计需进行如下步骤：

1. 对于 $x_1, ..., x_n$ 选取一个合适 EM 方法计算 $\hat{\theta}_{EM}$。令 $j = 1$ 且设 $\hat{\theta}_j = \hat{\theta}_{EM}$。
2. 增加 j。从原观测中抽取伪数据 $X_1^*, ..., X_n^*$。
3. 对抽取的数据用相同方法计算 $\hat{\theta}_j$。
4. 如果 j 足够大，停止；否则返回第 2 步。

通过自助法得到一组参数估计，这组估计的样本方差就是 $\hat{\theta}$ 的估计方差。$\hat{\theta}$ 样本分布的其他特征也可以用自助法得到的数据得到。

4.2.3.4 经验信息

当数据独立同分布时，经验信息定义为：

$$\frac{1}{n}\sum_{i=1}^{n} l'(\theta|x_i)l'(\theta|x_i)^T - \frac{1}{n^2}l'(\theta|x_i)l'(\theta|x_i)^T$$

该方法的引人之处在于上式中所有项都是 M 步的副产品，不需要额外分析。注意到 $\theta^{(t)}$ 关于 θ 最大化 $Q(\theta|\theta^{(t)}) - l(\theta|x)$。关于 θ 取导数得：

$$Q'(\theta|\theta^{(t)})|_{\theta=\theta^{(t)}} = l'(\theta|x)|_{\theta=\theta^{(t)}}$$

由于 Q' 通常在每个 M 步计算，该方法的单个项是可以得到的。

4.2.3.5 数值微分

为估计 Hessian 阵，可以考虑用下式：

$$\frac{df(x)}{dx_i} \approx \frac{f(x+\epsilon_i e_i) - f(x-\epsilon_i e_i)}{2\epsilon_i}$$

计算 l' 在 $\hat{\theta}$ 处的数值导数，每次一个坐标。

Hessian 各行需要通过向 $\hat{\theta}$ 的坐标加一个小的扰动得到，扰动的大小都会影响估计的准确性，需要慎重处理。

R 语言示例

```r
rm(list=ls())
set.seed(1234567)
Sim=2000
n=500
lambda=1/200
Sim=1000
SD=matrix(0,Sim,4)
for(sim in 1:Sim){

  Y=rexp(n,lambda)
  lambda.gold=mean(Y)
  #C=runif(n,150,180)
  C=rexp(n,lambda/2)
```

R 语言示例

```
delta=1*(Y<C)
mean(delta)
X=pmin(Y,C)
lambda.hat=1/(sum(X)/sum(delta))
LB=NULL
B <- 1000
for(b in 1:B){
  index=sample.int(n,n,replace = T)
  X.new=X[index]
  delta.new=delta[index]
  lambda.b=1/mean(X.new)
  for(it in 1:20)
    lambda.b=n/(sum(X.new)+(n-sum(delta.new))/ lambda.b)
```

R 语言示例

```
    sd1=sqrt(1/(sum(delta)/lambda.hat^2))
    sd2=sd(LB)
    Low= lambda.hat-1.96*sd1
    Up= lambda.hat+1.96*sd1
    ecp1=1*(lambda<Up)*(lambda>Low)#落在置信区间的概率
    Low= lambda.hat-1.96*sd2
    Up= lambda.hat+1.96*sd2
    ecp2=1*(lambda<Up)*(lambda>Low)#落在置信区间的概率
    SD[sim,]=c(sd1,ecp1,sd2,ecp2)
}
  out=round(rbind(apply(SD,2,mean),apply(SD,2,sd))*100,2)
  print(out)
```

R 语言示例

以上 R 代码展示了 Louis 方法和自助法在存在删失的指数数据的方差估计上的一些对比。输出结果如下：

```
> print(out)
     [,1] [,2] [,3]  [,4]
[1,] 0.03 94.4 0.03 94.10
[2,] 0.00 23.0 0.00 23.57
```

4.3 EM 变形

4.3.1 改进 E 步

4.3.1.1 Monte Carlo EM

如果 E 步所求期望难以计算，可以用 Monte Carlo 方法近似:

1. 从 $f_{Z|X}(z|x,\theta^{(t)})$ 中抽取缺失数据集 $Z_1^{(t)},...,Z_{m^{(t)}}^{(t)}$。用 $Y_j=(x,Z_j)$ 表示一个补齐的数据集，缺失值由 Z_j 代替。
2. 计算 $\hat{Q}^{(t+1)}(\theta|\theta^{(t)})=(1/m^{(t)})\Sigma_{j=1}^{m^{(t)}}logf_Y(Y_j^{(t)}|\theta)$。

那么 $\hat{Q}^{(t+1)}(\theta|\theta^{(t)})$ 就是 $Q(\theta|\theta^{(t)})$ 的 MC 估计。M 步修改为最大化 $\hat{Q}^{(t+1)}(\theta|\theta^{(t)})$。

例 4.7(删失的指数数据，续): 例 4.5 给出了普通 EM 算法，MCEM 算法给出:

$$\hat{Q}^{(t+1)}(\lambda|\lambda^{(t)}) = nlog\lambda - \frac{\lambda}{m^{(t)}} \sum_{j=1}^{m^{(t)}} Y_j^T 1$$

令 $\hat{Q}'(\lambda|\lambda^{(t)}) = 0$ 且对 λ 求解得到:

$$\lambda^{(t+1)} = \frac{n}{\sum_{j=1}^{m^{(t)}} Y_j^T 1/m^{(t)}}$$

作为 MCEM 更新。

R 语言示例

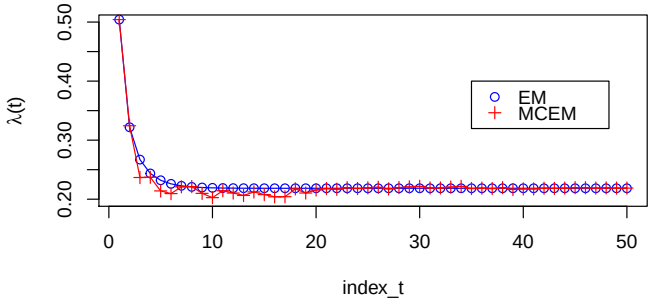
```
DATA<-read.table("censor.txt",header = T)
Y <- DATA[,1]
C <- DATA[,2]
X <- DATA[,3]
delta <- DATA[,4]
lambda.new=1/mean(X)
lambda.new1=1/mean(X)
S <- matrix(lambda.new1,50,2)
n=30
for(sim in 2:50){
  lambda.new<-n/(sum(X)+sum(1-delta)/lambda.new)#普通EM更新
  S[sim,1]<-lambda.new
  m=5^(1+floor(sim/10))
```

R 语言示例

```
  K=0
  for(j in 1:m){
    Z.new=rexp(sum(1-delta),lambda.new1)
    K=K+sum(Z.new)+sum(X)}
  lambda.new1=n*m/K#MCEM更新
  S[sim,2]<-lambda.new1}
print(S)
plot(S[,1],xlab="index_t",ylab=expression(paste(lambda,"(t)
  ")),type="o",col=4, ylim=c(0.2,0.5))
points(S[,2],type="o",pch=3,col=2,ylim=c(0.2,0.5))
legend(35,0.4,c("EM","MCEM"), col=c(4,2),pch=c(1,3))
```

R 语言示例

以上 R 代码是例 4.7 的绘图过程，图象提供 4.7 中数据的 EM 和 MCEM 的迭代比较。图象如下：



4.3 EM 变形
4.3.2 改进 M 步

通常 EM 算法的 E 步比较直接了当，M 步相对不容易实施。

更加方便的实施 M 步需要什么策略？

1. ECM 算法
2. EM 梯度算法

4.3.2.1 ECM 算法

ECM 算法用一系列计算较简单的条件极大化 (CM) 步骤代替 M 步。

称 t 个 E 步后的较简单 CM 步的集合称为一个 CM 循环。所以 ECM 算法的第 t 次迭代包括第 t 个 E 步和第 t 次 CM 循环。S 表示每个 CM 循环里 CM 步的数目，第 t 次循环里第 s 个 CM 步需要在约束

$$g_s(\theta) = g_s(\theta^{(t+(s-1)/S)})$$

下最大化 $Q(\theta|\theta^{(t)})$，当 S 个 CM 步完成时，令 $\theta^{(t+1)} = \theta^{(t+S/S)}$ 并进行第 t+1 次迭代的 E 步。

例 4.8(带缺失值的多元回归): 设 $U_1, ..., U_n$ 是从 d 维正态模型

$$U_i \sim N_d(\mu_i, \Sigma)$$

观测的 n 个独立 d 维向量。其中 $\mu_i = V_i\beta$。假定一些 U_i 的某些元素是缺失的。

将 U_i 和 μ_i 的元素以及 V_i 的行进行重新排序，使得 U_i 中观测元素在前缺失元素在后。记 $U_i = (U_{obs,i}, U_{miss,i}), \mu_i = (\mu_{obs,i}, \mu_{miss,i})$ 及

$$\Sigma_i = \begin{pmatrix} \Sigma_{obs,i} & \Sigma_{cross,i} \\ \Sigma_{cross,i}^T & \Sigma_{miss,i} \end{pmatrix}$$

观测数据全集表示成 $U_{obs} = (U_{obs,1}, ..., U_{obs,n})$。

4.3.2.1 ECM 算法

直接处理似然函数较为麻烦，但本例中 E 步等价于在观测数据和当前参数 $\beta^{(t)}$ 和 $\Sigma^{(t)}$ 条件下求完全数据充分统计量的期望。对于 $j = 1, ..., d$:

$$E\{\sum_{i=1}^{n} U_{ij} | u_{obs}, \beta^{(t)}, \Sigma^{(t)}\} = \sum_{i=1}^{n} a_{ij}^{(t)}$$

类似地，对于 $j, k = 1, ..., d$:

$$E\{\sum_{i=1}^{n} U_{ij}U_{ik} | u_{obs}, \beta^{(t)}, \Sigma^{(t)}\} = \sum_{i=1}^{n} (a_{ij}^{(t)} a_{ik}^{(t)} + b_{ijk}^{(t)})$$

其中:

$$a_{ij}^{(t)} = \begin{cases} E\{U_{ij} | u_{obs,i}, \beta_i^{(t)}, \Sigma_i^{(t)}\}, & \text{如果} U_{ij} \text{缺失} \\ u_{ij}, & \text{如果观察到} U_{ij} = u_{ij} \end{cases}$$

4.3.2.1 ECM 算法

接上例，式中：

$$b_{ijk}^{(t)} = \begin{cases} cov\{U_{ij}, U_{ik}|u_{obs,i}, \beta_i^{(t)}, \Sigma_i^{(t)}\}, & \text{如果}U_{ij}\text{和}U_{ik}\text{都缺失} \\ u_{ij}, & \text{其他} \end{cases}$$

可以通过缺失数据 $U_{miss,i}|(u_{obs,i}, \beta_i^{(t)}, \Sigma_i^{(t)})$ 的条件分布得到均值向量和协方差阵。知道这些后就可通过指数族 E 步求出 $Q(\beta, \Sigma|\beta^{(t)}, \Sigma^{(t)})$。　　以上就完成了 E 步，M 步可通过 ECM 策略。先加入约束 $\Sigma = \Sigma^{(t)}$，利用加权最小二乘估计：

$$\beta^{(t+1/2)} = (\sum_{i=1}^{n} V_i^T (\Sigma_i^{(t)})^{-1} V_i)^{-1} (\sum_{i=1}^{n} V_i^T (\Sigma_i^{(t)})^{-1} a_i^{(t)})$$

关于 β 最大化 $Q(\beta, \Sigma|\beta^{(t)}, \Sigma^{(t)})$，这构成了两个 CM 步的第一步。

4.3.2.1 ECM 算法

第二个 CM 步依据

$$E\{\frac{1}{n}\sum_{i=1}^{n}(U_i - V_i\beta^{(t+1/2)})(U_i - V_i\beta^{(t+1/2)})^T|u_{obs}, \beta^{(t+1/2)}, \Sigma^{(t)}\}$$

在约束 $\beta = \beta^{t+1/2}$ 下关于 Σ 最大化 $Q(\beta, \Sigma|\beta^{(t)}, \Sigma^{(t)})$。将这两个 CM 步结合起来得到:

$$(\beta^{(t+1)}, \Sigma^{(t+1)}) = (\beta^{(t+1/2)}, \Sigma^{(t+2/2)})$$

且保证在 Q 函数上有一个增量。

总结: ECM 算法在下面两步之间交替进行:

1. 创建更新了的完全数据集
2. 用当前的完全数据轮流固定 β 和 Σ 中的某一个值来估计另一个参数。

4.3.2.2 EM 梯度算法

用单步 Newton 法替代 M 步，可以近似最大值而不用真正地精确求解。M 步是由：

$$\begin{aligned}\theta^{(t+1)} &= \theta^{(t)} - Q''(\theta|\theta^{(t)})^{-1}|\theta=\theta^{(t)} Q'(\theta|\theta^{(t)})^{-1}|\theta=\theta^{(t)} \\ &= \theta^{(t)} - Q''(\theta|\theta^{(t)})^{-1}|\theta=\theta^{(t)} l'(\theta^{(t)}|x)\end{aligned}$$

给出的更新替代，其中 $l'(\theta^{(t)}|x)$ 是当前迭代得分函数的估计。这种 EM 梯度算法和完全 EM 算法对 $\hat{\theta}$ 有相同的收敛速度。

例 4.9(椒花蛾，续) 对椒花蛾数据应用 EM 梯度算法，得到：

$$\frac{d^2Q(p|p^{(t)})}{dp_C^2} = -\frac{2n_{CC}^{(t)} + n_{CI}^{(t)} + n_{CT}^{(t)}}{p_C^2} - \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{(1-p_C-p_I)^2}$$

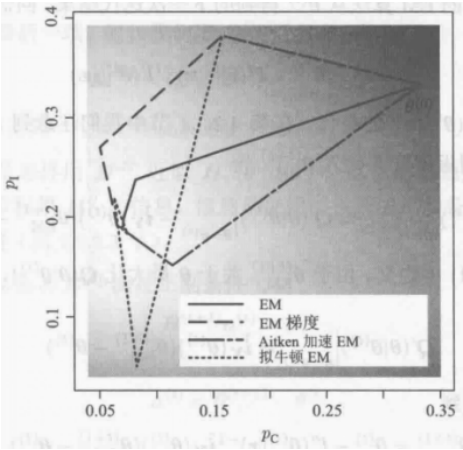
$$\frac{d^2Q(p|p^{(t)})}{dp_I^2} = -\frac{2n_{II}^{(t)} + n_{IT}^{(t)} + n_{CI}^{(t)}}{p_I^2} - \frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{(1-p_C-p_I)^2}$$

和

$$\frac{d^2Q(p|p^{(t)})}{dp_Cdp_I} = -\frac{2n_{TT}^{(t)} + n_{CT}^{(t)} + n_{IT}^{(t)}}{(1-p_C-p_I)^2}$$

4.3.2.2 EM 梯度算法

下图显示了从 $p_C = p_I = p_T = 1/3$ 开始的 EM 梯度算法步骤。也给出普通 EM 算法作对比。



4.3 EM 变形
4.3.3 加速方法

EM 方法收敛较慢是一个明显缺点。如何加速 EM 方法？

1. Aitken 加速
2. 拟 Newton 加速

4.3.3.1 Aitken 加速

设 $\theta_{EM}^{(t+1)}$ 是由标准 EM 算法从 $\theta^{(t)}$ 得到的下一次迭代结果。在牛顿更新中，用 $Q'(\theta|\theta^{(t)})$ 代替 $l'(\theta^{(t)}|x)$ 并将 Q' 在 $\theta^{(t)}$ 附近展开并带入 $\theta_{EM}^{(t+1)}$ 得：

$$Q'(\theta|\theta^{(t)})|_{\theta=\theta_{EM}^{(t+1)}} \approx Q'(\theta|\theta^{(t)})|_{\theta=\theta^{(t)}} - \hat{i}(\theta^{(t)})(\theta_{EM}^{(t+1)} - \theta^{(t)})$$

由于 $\theta_{EM}^{(t+1)}$ 关于 θ 最大化 $Q(\theta|\theta^{(t)})$，所以上式左边为 0，再由最大化对数似然牛顿更新得到：

$$\theta^{(t+1)} = \theta^{(t)} - l''(\theta^{(t)}|x)^{-1}\hat{i}_Y(\theta^{(t)})(\theta_{EM}^{(t+1)} - \theta^{(t)})$$

这种更新是 Aitken 加速的一般策略的一个例子。EM 的 Aitken 加速等同于用 Newton-Raphson 方法求 $\Psi(\theta) - \theta$ 的一个零点。

4.3.3.2 拟 Newton 加速

拟 Newton 优化方法依据的是用 $M^{(t)}$ 代替 $l''(\theta^{(t)}|x)$，得到：

$$\theta^{(t+1)} = \theta^{(t)} - (M^{(t)})^{-1} l'(\theta^{(t)}|x)$$

在 EM 框架下，可以把 $l''(\theta^{(t)}|x)$ 分解成一个在 EM 期间计算的部分和一个余项。用 $B^{(t)}$ 近似余项得到关于 $M^{(t)}$ 的等式：

$$M^{(t)} = Q''(\theta|\theta^{(t)})|_{\theta=\theta^{(t)}} - B^{(t)}$$

将其带入拟 Newton 法更新方程中就得到了一个拟 Newton 加速。

这个方法关键在于怎样用 $B^{(t)}$ 近似余项，这里的思想是以 $B^{(0)} = 0$ 为初始值，随后随着迭代逐步累积余项的信息。

4.3.3.2 拟 Newton 加速

特别地，可以要求：

$$B^{(t+1)}a^{(t)} = b^{(t)}$$

其中

$$a^{(t)} = \theta^{(t+1)} - \theta^{(t)}$$

且

$$b^{(t)} = H'(\theta|\theta^{(t+1)})|_{\theta=\theta^{(t+1)}} - H'(\theta|\theta^{(t+1)})|_{\theta=\theta^{(t)}}$$

根据拟 Newton 法的更新方程，可以设：

$$B^{(t+1)} = B^{(t)} + c^{(t)}v^{(t)}(v^{(t)})^T$$

其中 $v^{(t)} = b^{(t} - B^{(t)}a^{(t)}$ 且 $c^{(t)} = 1/[(v^{(t)})^T a^{(t)}]$。

实际上，EM 梯度算法恰是最大化 $Q(\theta|\theta^{(t)})$ 的 Newton-Raphson 算法，这里描述的方法成为最大化 $l(\theta|x)$ 的近似 Newton-Raphson 算法。

谢谢观看！