EM Algorithm

Renhe W.

Contents

List of Figures

List of Tables

1 EM 算法

EM 算法的普遍应用主要归功于 DLR (Dempster et al., 1977)，他们在研究中还提供了许多其适用性示例，并在相当普遍的条件下确立了其收敛性和其他基本性质。对于一个具有隐藏状态的系统，我们假设观测序列为 $\mathbf{y}=\{y_1,y_2,\ldots,y_T\}$，隐藏状态为 $\mathbf{s}=\{s_1,s_2,\ldots,s_T\}$，其中 T 是观测序列时长. 定义非完全信息似然函数（incomplete-data likelihood function）为 $L(\mathbf{y}\mid\theta)$，其中 $\boldsymbol{\theta}$ 是模型参数，则有：

$$
\begin{aligned}
\log L(\mathbf{y}\mid\boldsymbol{\theta}) &= \log f(\mathbf{y}\mid\boldsymbol{\theta}) & \\
&= \log f(\mathbf{y}\mid\boldsymbol{\theta})\cdot\tfrac{f(\mathbf{y},\mathbf{s}\mid\boldsymbol{\theta})}{f(\mathbf{y},\mathbf{s}\mid\boldsymbol{\theta})} & \text{联系完全似然} \\
&= \log f(\mathbf{y},\mathbf{s}\mid\boldsymbol{\theta})\cdot\tfrac{f(\mathbf{y}\mid\boldsymbol{\theta})}{f(\mathbf{y},\mathbf{s}\mid\boldsymbol{\theta})} & \\
&= \log \underbrace{f(\mathbf{y},\mathbf{s}\mid\boldsymbol{\theta})}_{\text{完全信息似然函数}} - \log f(\mathbf{s}\mid\mathbf{y},\theta) & \text{条件概率}
\end{aligned}
\tag{1.1}
$$

同时对(1.1)式子两边关于 $\mathbf{y}$ 和 $\boldsymbol{\theta}'$ 取期望可以得到：

$$
\begin{aligned}
&\textstyle\sum_s \log L(\mathbf{y}\mid\boldsymbol{\theta})\cdot f(\mathbf{s}\mid\mathbf{y},\boldsymbol{\theta}') = \sum_s \log f(\mathbf{y},\mathbf{s}\mid\boldsymbol{\theta})\cdot f(\mathbf{s}\mid\mathbf{y},\boldsymbol{\theta}') - \sum_s \log f(\mathbf{s}\mid\mathbf{y},\boldsymbol{\theta})\cdot f(\mathbf{s}\mid\mathbf{y},\boldsymbol{\theta}') \\
\Rightarrow\quad &\log L(\mathbf{y}\mid\boldsymbol{\theta})\cdot\underbrace{\sum_s f(\mathbf{s}\mid\mathbf{y},\boldsymbol{\theta}')}_{=1} = E\left(\log f(\mathbf{y},\mathbf{s}\mid\boldsymbol{\theta})\mid\mathbf{y},\boldsymbol{\theta}'\right) - E\left(\log f(\mathbf{s}\mid\mathbf{y},\boldsymbol{\theta})\mid\mathbf{y},\boldsymbol{\theta}'\right) \\
\Rightarrow\quad &\log L(\mathbf{y}\mid\boldsymbol{\theta}) = Q(\boldsymbol{\theta},\boldsymbol{\theta}') - H(\boldsymbol{\theta},\boldsymbol{\theta}').
\end{aligned}
\tag{1.2}
$$

1.1 EM Algorithm

EM 算法（期望最大化算法）是一种用于含有隐变量的统计数据估计的迭代算法。它通过交替执行两个步骤：期望步骤（E-step）和最大化步骤（M-step）——来找到参数的最大似然估计或最大后验估计.EM 算法的目的是最大化非完全信息似然函数 $L(\mathbf{y}\mid\boldsymbol{\theta})$。在这种情况下，EM 算法的两个步骤可以这样表述：

EM 算法〈步骤〉

1. E-Step（期望步骤）：

在第 k 次迭代中，E-step 的目的是计算在当前参数估计 $\boldsymbol{\theta}^{(k)}$ 的条件下，完全数据对数似然函数 $\log L(\mathbf{y}, s \mid \boldsymbol{\theta})$ 的期望值。这一步骤可以表示为：

$$Q\left(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)}\right) = E\left(\log f(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\theta}) \mid \mathbf{y}, \boldsymbol{\theta}^{(k)}\right)$$

其中 $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ 是在给定观测数据 $\mathbf{y}$ 和当前参数估计 $\boldsymbol{\theta}^{(k)}$ 的情况下，关于隐状态的完全数据对数似然的期望.

2. M-Step（最大化步骤）：

在 M-step 中，目标是找到参数 $\boldsymbol{\theta}$ 的新估计值 $\boldsymbol{\theta}^{(k+1)}$，使得 $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$ 最大化. 数学上表示为：

$$\boldsymbol{\theta}^{(k+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(k)})$$

选择 $\boldsymbol{\theta}^{(k+1)}$ 作为参数 $\boldsymbol{\theta}$ 的新估计值，使得 Q 函数在这一点上达到最大.

通过交替执行这两个步骤，EM 算法在每次迭代中更新参数 $\boldsymbol{\theta}$ 的估计值，直到似然函数 $L(\boldsymbol{\theta})$ 的值收敛到一个固定值，或达到预定的迭代次数。这个过程保证了每次迭代后，不完全数据的对数似然函数 $L(\boldsymbol{\theta})$ 不会减少，从而实现对参数的有效估计.

1.2 How it works?

DLR（Dempster, Laird, and Rubin）在他们的工作中证明了，在一定条件下，EM 算法可以保证每次迭代后，不完全数据的对数似然函数 $L(\mathbf{y} \mid \boldsymbol{\theta})$ 不会减少. 这意味着，通过 EM 算法得到的参数估计序列将收敛到一个局部最大值.

命题 1.1 (单调不减（MONOTONICITY）). $L(\mathbf{y} \mid \boldsymbol{\theta}^{(k+1)}) \geq L(\mathbf{y} \mid \boldsymbol{\theta}^{(k)})$.

Proof. 根据(1.2)式，有

$$\begin{aligned}
&\log L\left(\mathbf{y} \mid \boldsymbol{\theta}^{(k+1)}\right) - \log L\left(y \mid \boldsymbol{\theta}^{(k)}\right) \\
=&\underbrace{\left\{Q\left(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}\right) - Q\left(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}\right)\right\}}_{\text{term1}} \underbrace{- \left\{H\left(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}\right) - H\left(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}\right)\right\}}_{\text{term2}} \\
=&\text{term1} + \text{term2},
\end{aligned}$$

其中 term1 中的 Q 函数每一步进行的过程中都有求导梯度更新,所以有 $Q\left(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}\right) \geq Q\left(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}\right)$，现在的主要工作为 term2，对于 $\boldsymbol{\theta} \in \Omega$，若 term2 每次更新都是负值，i.e. $H\left(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}\right) \leq H\left(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}\right)$ ，则命题1.1成立，下面给出其中一个证明：

对于任意参数 $\boldsymbol{\theta}$，有：

$$\begin{aligned}
&H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) - H(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}) \\
&\qquad = E\left(\log f(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta}) \mid \mathbf{y}, \boldsymbol{\theta}^{(k)}\right) - E\left(\log f(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta}^{(k)}) \mid \mathbf{y}, \boldsymbol{\theta}^{(k)}\right) \\
&\qquad = E\left(\log f(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta}) / f(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta}^{(k)}) \mid \mathbf{y}, \boldsymbol{\theta}^{(k)}\right) \\
&\qquad \leq \log[E\left(f(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta}) / f(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta}^{(k)}) \mid \mathbf{y}, \boldsymbol{\theta}^{(k)}\right)] \qquad \textit{Jensen inequality} \\
&\qquad = log \textstyle\sum_s \frac{f(\mathbf{s}|\mathbf{y},\boldsymbol{\theta})}{f(\mathbf{s}|\mathbf{y},\boldsymbol{\theta}^{(k)})} \cdot f(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta}^{(k)}) \\
&\qquad = log \textstyle\sum_s f(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta}) = \log 1 = 0.
\end{aligned}$$

综上，对数似然函数 $L(\mathbf{y} \mid \boldsymbol{\theta})$ 通过 EM 算法更新迭代一直单调不减. □

> **!** 对于 $H\left(\boldsymbol{\theta}^{(k+1)}, \boldsymbol{\theta}^{(k)}\right) \leq H\left(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)}\right)$，还可以运用 KL 散度进行证明，KL 散度是衡量两个概率分布间差异的度量，定义为
>
> $$D_{\mathrm{KL}}(P \| Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$$
>
> 其中，P 和 Q 是两个概率分布.

通过命题1.1以及相应的证明，对数似然函数 $L(\mathbf{y} \mid \boldsymbol{\theta})$ 通过 EM 算法更新迭代一直单调不减，下面给出算法收敛的证明.

Wu (1983)

Wu (1983) 为确保似然序列 $\{L(\mathbf{y} \mid \boldsymbol{\theta}^{(k)})\}$ 收敛到一个稳定值，给出以下几个条件：

- Ω 是 d 维欧几里得空间 $\mathbb{R}^d$ 的中的子集.
- 对于 $\forall L\left(\mathbf{y} \mid \boldsymbol{\theta}_{\mathbf{0}}\right) > -\infty$, $\Omega_{\boldsymbol{\theta}_{\mathbf{0}}} = \{\boldsymbol{\theta} \in \Omega : L(\mathbf{y} \mid \boldsymbol{\theta}) \geq L\left(\mathbf{y} \mid \boldsymbol{\theta}_{\mathbf{0}}\right)\}$ 是一个紧集.
- $L(\mathbf{y} \mid \boldsymbol{\theta})$ 在 Ω 中连续，在 Ω 上可微.

命题 1.2 (收敛性（CONVERGENCE）). *似然序列 $\{L(\mathbf{y} \mid \boldsymbol{\theta}^{(k)})\}$ 单调收敛到 $L^\star = L(\mathbf{y} \mid \boldsymbol{\theta}^\star)$.*

Proof. 似然序列 $\{L(\mathbf{y} \mid \boldsymbol{\theta}^{(k)})\}$ 单调收敛到 $L^\star = L(\mathbf{y} \mid \boldsymbol{\theta}^\star)$，即

$$\frac{\partial L(\mathbf{y} \mid \boldsymbol{\theta}^\star)}{\partial \boldsymbol{\theta}^\star} = 0.$$

即也表示为

$$\frac{\partial \log L(\mathbf{y} \mid \boldsymbol{\theta}^\star)}{\partial \boldsymbol{\theta}^\star} = 0.$$

这里假设 $L(y \mid \theta)$ 是单峰函数（在 Ω 中，并可微），对(1.2)式两边求导，有

$$\frac{\partial \log L(\mathbf{y} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \frac{\partial Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}} - \frac{\partial H\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}}, \tag{1.3}$$

由命题1.1的证明，有 $H(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}) \leq H(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}^{(k)})$ ，此时可以理解为到达了平稳点. 则对于所有 $\boldsymbol{\theta} \in \Omega$，有：

$$\left.\frac{\partial H\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(k)}} = 0. \tag{1.4}$$

令 $\boldsymbol{\theta}_{\mathbf{0}}$ 是 $\boldsymbol{\theta}$ 的任一值，将 $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}_{\mathbf{0}}$ 放入(1.3)中，又根据(1.4)，可以得到

$$\left.\frac{\partial \log L(\mathbf{y} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathbf{0}}} = \left.\frac{\partial Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\mathbf{0}}} \tag{1.5}$$

假设 $\boldsymbol{\theta} = \boldsymbol{\theta}^\star$ 时，$\boldsymbol{\theta}^\star$ 是 $\log L(\mathbf{y} \mid \boldsymbol{\theta})$ 的一个平稳点，由(1.5)得

$$\left.\frac{\partial \log L(\mathbf{y} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star} = \left.\frac{\partial Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)}\right)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star} \tag{1.6}$$

则若 $Q\left(\boldsymbol{\theta}, \boldsymbol{\theta}^\star\right)$ 在 $\boldsymbol{\theta}^\star \in \Omega$ 全局最优，则 EM 算法可以收敛到鞍点 $\boldsymbol{\theta}^\star$. □

1.3 Score Statistic

Score Statistic

对数似然函数的梯度向量:

$$S(\boldsymbol{\theta}) = \frac{\partial \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

完全对数似然函数的梯度向量:

$$S_c(\boldsymbol{\theta}) = \frac{\partial \log f(\mathbf{y}, \mathbf{s}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

其中 $S(\boldsymbol{\theta})$ 可以通过 $S(\boldsymbol{\theta})$ 表示:

$$\begin{aligned}
S(\boldsymbol{\theta}) &= \tfrac{\partial \log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \tfrac{\partial f(\mathbf{y}|\boldsymbol{\theta})/\partial \boldsymbol{\theta}}{f(\mathbf{y}|\boldsymbol{\theta})} \\
&= \sum \tfrac{\partial f(\mathbf{y},\mathbf{s}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} / f(\mathbf{y} \mid \boldsymbol{\theta}) \\
&= \sum \tfrac{\partial \log f(\mathbf{y},\mathbf{s}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \cdot \tfrac{f(\mathbf{y},\mathbf{s}|\boldsymbol{\theta})}{f(\mathbf{y}|\boldsymbol{\theta})} \qquad \text{加一个 } \mathit{log} \text{ 变换} \\
&= \sum \tfrac{\partial \log f(\mathbf{y},\mathbf{s}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \cdot f(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta}) \\
&= E\{\tfrac{\partial \log f(\mathbf{y},\mathbf{s}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \mid \mathbf{y}, \boldsymbol{\theta}\} \\
&= E\{S_c(\boldsymbol{\theta}) \mid \mathbf{y}, \boldsymbol{\theta}\}.
\end{aligned} \tag{1.7}$$

i.e. $S(\boldsymbol{\theta}) = \frac{\partial Q(\boldsymbol{\theta}_0, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}_0}\Big|_{\boldsymbol{\theta}_0 = \boldsymbol{\theta}}$.

1.4 Missing Information

最大似然估计的渐进方差由费希尔（Fisher）信息量决定，根据第1.3节的定义，费希尔信息量为：

$$\mathcal{F} = E\{S(\boldsymbol{\theta})S(\boldsymbol{\theta})^T \mid \mathbf{y}, \boldsymbol{\theta}\} = E\{J(\boldsymbol{\theta}) \mid \mathbf{y}, \boldsymbol{\theta}\} \tag{1.8}$$

其中 $J(\boldsymbol{\theta}) = -\frac{\partial^2 log f(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$，令 $J_c(\boldsymbol{\theta}) = -\frac{\partial^2 log f(\mathbf{y},\mathbf{s}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$，$\mathcal{F}_c = E\{J_c \boldsymbol{\theta}) \mid \mathbf{y}, \boldsymbol{\theta}\}$. 由(1.1)得：

$$\log f(\mathbf{y} \mid \boldsymbol{\theta}) = \log f(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\theta}) - \log f(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta}),$$

对上面等式关于参数 $\boldsymbol{\theta}$ 同时求二次导，有：

$$J(\boldsymbol{\theta}) = J_c(\boldsymbol{\theta}) + \frac{\partial^2 log f(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T},$$

两边求条件期望得：

$$E\{J(\boldsymbol{\theta}) \mid \mathbf{y}, \boldsymbol{\theta}\} = E\{J_c(\boldsymbol{\theta}) \mid \mathbf{y}, \boldsymbol{\theta}\} + E\{\frac{\partial^2 log f(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \mid \mathbf{y}, \boldsymbol{\theta}\}, \tag{1.9}$$

观察等式左边，有

$$\begin{aligned} E\{J(\boldsymbol{\theta}) \mid \mathbf{y}, \boldsymbol{\theta}\} &= \sum_s J(\boldsymbol{\theta}) \cdot f(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta}) \\ &= \sum_s -\frac{\partial^2 log f(\mathbf{y} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \cdot f(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta}) \\ &= J(\boldsymbol{\theta}) \cdot \sum_s f(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta}) \\ &= J(\boldsymbol{\theta}) \cdot 1 \\ &= J(\boldsymbol{\theta}) \end{aligned}$$

则我们可以简化(1.9)表示为：

$$\underbrace{J(\boldsymbol{\theta})}_{\text{观测信息}} = \underbrace{E\{J_c(\boldsymbol{\theta}) \mid \mathbf{y}, \boldsymbol{\theta}\}}_{\text{条件期望完整信息}} - \underbrace{J_m(\boldsymbol{\theta})}_{\text{缺失信息}} \tag{1.10}$$

其中 $J_m(\boldsymbol{\theta}) = E\{\frac{\partial^2 log f(\mathbf{s}|\mathbf{y}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \mid \mathbf{y}, \boldsymbol{\theta}\}$，可以发现存在以下关系：

$$\begin{aligned} -J_m(\boldsymbol{\theta}) &= cov\{S_c(\boldsymbol{\theta}) \mid \mathbf{y}, \boldsymbol{\theta}\} \\ &= E\{S_c(\boldsymbol{\theta}) S_c(\boldsymbol{\theta})^T \mid \mathbf{y}, \boldsymbol{\theta}\} - (E\{S_c(\boldsymbol{\theta}) \mid \mathbf{y}, \boldsymbol{\theta}\})^2 \\ &= E\{S_c(\boldsymbol{\theta}) S_c(\boldsymbol{\theta})^T \mid \mathbf{y}, \boldsymbol{\theta}\} - S(\boldsymbol{\theta}) S(\boldsymbol{\theta})^T \end{aligned}$$

根据(1.7)

则(1.9)可以表示为：

$$\begin{aligned} J(\boldsymbol{\theta}) &= E\{J_c(\boldsymbol{\theta}) \mid \mathbf{y}, \boldsymbol{\theta}\} - E\{S_c(\boldsymbol{\theta}) S_c(\boldsymbol{\theta})^T \mid \mathbf{y}, \boldsymbol{\theta}\} + S(\boldsymbol{\theta}) S(\boldsymbol{\theta})^T \\ &= E\{-\frac{\partial^2 log f(\mathbf{y}, \mathbf{s}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} \mid \mathbf{y}, \boldsymbol{\theta}\} - E\{S_c(\boldsymbol{\theta}) S_c(\boldsymbol{\theta})^T \mid \mathbf{y}, \boldsymbol{\theta}\} + S(\boldsymbol{\theta}) S(\boldsymbol{\theta})^T \\ &= -\frac{\partial^2 Q(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} - E\{S_c(\boldsymbol{\theta}) S_c(\boldsymbol{\theta})^T \mid \mathbf{y}, \boldsymbol{\theta}\} + \underbrace{S(\boldsymbol{\theta}) S(\boldsymbol{\theta})^T}_{\text{最优时，一般约等于 } 0} \end{aligned} \tag{1.11}$$

求和再求导等价于
求导再求和

1.5 Toy Example: 求解混合分布参数

Toy Example〈求解混合分布参数〉

如下数据：

$$3.54, 3.90, 3.93, 5.19, 3.58, 4.60, 3.85, 4.69, 4.29, 4.067,$$
$$3.77, 3.45, 5.36, 2.62, 4.80, 4.65, 3.65, 3.67, 6.23, 3.35,$$
$$1.58, 0.19, -1.89, 0.08, 0.34, 0.90, -0.03, 0.55, -0.57, -1.20$$

可能来自于正态分布 $N(0,1)$ 与 $N(\mu,1)$ 的混合，混合比为 $1-p$ 与 p，且 $0<p<1$。求出 p 与 μ 的极大似然估计.

首先给出混合密度：

$$f(y;p,\mu) = p\phi(y-\mu) + (1-p)\phi(y)$$

其中 $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$，设从混合分布中抽取样本 $Y=(Y_1,Y_2,\cdots,Y_n)$，得到其似然函数：

$$L(\mu,p,Y) = \prod_{i=1}^{n}(p\phi(Y_i-\mu) + (1-p)\phi(Y_i)).$$

对数后

$$l(\mu,p;Y) = \sum_{i=1}^{n} log(p\phi(Y_i-\mu) + (1-p)\phi(Y_i))$$

运用 EM 算法求解：引入潜在变量 $Z=(z_1,z_2,\cdots,z_n)$，且 $z_1,z_2,\cdots,z_n$ 相互独立，其中：

$$Z_i = \begin{cases} 1 & N(\mu,1) \\ 0 & N(0,1) \end{cases}$$

以及 $P(Z_i=1)=p, i=1,2,\cdots,n$，有 $Y_i|Z_i=1 \sim N(\mu,1)$，$Y_i|Z_i=0 \sim N(0,1)$，则 (Z_i,Y_i) 的似然函数为：

$$L(\mu,p;Y,Z) = \prod_{i=1}^{n} p^{Z_i}\phi(Y_i-\mu)^{Z_i}\phi(Y_i)^{1-Z_i}$$

对上述似然函数取对数并去掉与 p、μ 无关的量得:

$$l_1(\mu, p; Y, Z) = \sum_{i=1}^{n} Z_i logp - \frac{1}{2}\sum_{i=1}^{n} Z_i (Y_i - \mu)^2 + (n - \sum_{i=1}^{n} Z_i) log(1-p).$$

假设在第 k 步迭代中，有估计值 μ^k、p^k，通过 E 步和 M 步得到 μ、p 的新的估计值 $\mu^{(k+1)}$、$p^{(k+1)}$.

在 E 步中，令:

$$\begin{aligned}
Q(\mu, p|\mu^{(k)}, p^{(k)}, Y) &= E_Z[l_1(\mu, p; Y, Z)|\mu^{(k)}, p^{(k)}, Y] \\
&= \sum_{i=1}^{n} E_Z[Z_i|\mu^{(k)}, p^{(k)}, Y] logp \\
&- \frac{1}{2}\sum_{i=1}^{n} E_Z[Z_i|\mu^{(k)}, p^{(k)}, Y](Y_i - \mu)^2 \\
&+ (n - \sum_{i=1}^{n} E_Z[Z_i|\mu^{(k)}, p^{(k)}, Y] log(1-p))
\end{aligned}$$

易知:

$$Z_i^{(K+1)} = E_Z[Z_i|\mu^{(k)}, p^{(k)}, Y] = \frac{p^{(k)}\phi(Y_i - \mu^{(k)})}{p^{(k)}\phi(Y_i - \mu^{(k)}) + (1-p^{(k)})\phi(Y_i)}$$

根据贝叶斯定理，我们有:

$$P(Z_i = 1|Y_i, \mu^{(k)}, p^{(k)}) = \frac{P(Y_i|Z_i = 1, \mu^{(k)})P(Z_i = 1|\mu^{(k)}, p^{(k)})}{P(Y_i|\mu^{(k)}, p^{(k)})}$$

这里:

- $P(Y_i|Z_i = 1, \mu^{(k)})$ 是在 $Z_i = 1$ 条件下的 Y_i 的概率密度，即 $\phi(Y_i - \mu^{(k)})$.
- $P(Z_i = 1|\mu^{(k)}, p^{(k)})$ 是 $Z_i = 1$ 的先验概率，即 $p^{(k)}$.
- $P(Y_i|\mu^{(k)}, p^{(k)})$ 是 Y_i 的总概率，它等于两种情况的加权和，即 $p^{(k)}\phi(Y_i - \mu^{(k)}) + (1-p^{(k)})\phi(Y_i)$.

因此，

$$Z_i^{(k+1)} = E[Z_i|\mu^{(k)}, p^{(k)}, Y_i] = \frac{p^{(k)}\phi(Y_i - \mu^{(k)})}{p^{(k)}\phi(Y_i - \mu^{(k)}) + (1 - p^{(k)})\phi(Y_i)}$$

在 M 步中，解：

$$\begin{cases} \frac{\partial Q}{\partial \mu} = 0 \\ \frac{\partial Q}{\partial p} = 0 \end{cases}$$

求得：

$$\mu^{(k+1)} = \frac{\sum_{i=1}^n \frac{\phi(Y_i - \mu^{(k)})Y_i}{p^{(k)}\phi(Y_i - \mu^{(k)}) + (1 - p^{(k)})\phi(Y_i)}}{\sum_{i=1}^n \frac{\phi(Y_i - \mu^{(k)})}{p^{(k)}\phi(Y_i - \mu^{(k)}) + (1 - p^{(k)})\phi(Y_i)}};$$

$$p^{(k+1)} = \frac{1}{n}\sum_{i=1}^n \frac{p^{(k)}\phi(Y_i - \mu^{(k)})}{p^{(k)}\phi(Y_i - \mu^{(k)}) + (1 - p^{(k)})\phi(Y_i)}.$$

2 Example: MULTINOMIAL WITH COMPLEX CELL STRUCTURE

假设我们有 $n = 435$ 次观测，观测对象是一个有四个遗传性状的多项式分布，这些性状的概率结构如表1所示. 表中还给出了这些性状观测到的频率. （其中基因为 O、A 以及 B，三个基因相互组合，通过观测值求解每个基因对应的概率 r、A 和 B.）

2.1 MLE Method

Table 1: 观测细胞数据

类别（细胞）	细胞概率	观测频率
O	r^2	$n_O = 176$
A	$p^2 + 2pr$	$n_A = 182$
B	$q^2 + 2qr$	$n_B = 60$
AB	$2pq$	$n_{AB} = 17$

因此，观测数据由性状频率的向量给出：

$$\boldsymbol{y} = (n_O, n_A, n_B, n_{AB})^T .$$

未知参数的向量为：

$$\Psi = (p, q)^T,$$

因为 $r = 1 - p - q$. 目标是基于 $\boldsymbol{y}$ 找到 Ψ 的最大似然估计（MLE）. 这是遗传学中基因频率估计的一个著名问题，很多研究都有讨论.

参数 $\boldsymbol{\Psi}$ 的对数似然函数（除了一个加性常数之外）为

$$\log L(\Psi) = 2n_0 \log \underbrace{r}_{\pi_1} + n_A \log \underbrace{\left(p^2 + 2pr\right)}_{\pi_2} + n_B \log \underbrace{\left(q^2 + 2qr\right)}_{\pi_3} + n_{AB} \log \underbrace{(2pq)}_{\pi_4},$$

它没有一个封闭形式的解决方案来获得 $\hat{\boldsymbol{\Psi}}$，即 Ψ 的最大似然估计（MLE）.

我们将单元频率表示为 $\pi_j(j = 1, 2, 3, 4)$。那么它们关于 Ψ 的一阶和二阶导数

如下：

$$\frac{\partial \pi_1(\mathbf{\Psi})}{\partial \mathbf{\Psi}} = \begin{pmatrix} -2r \\ -2r \end{pmatrix}; \qquad \frac{\partial \pi_2(\mathbf{\Psi})}{\partial \mathbf{\Psi}} = \begin{pmatrix} 2p+2r \\ -2p \end{pmatrix}$$

$$\frac{\partial \pi_3(\mathbf{\Psi})}{\partial \mathbf{\Psi}} = \begin{pmatrix} 2q+2r \\ -2q \end{pmatrix}; \qquad \frac{\partial \pi_4(\mathbf{\Psi})}{\partial \mathbf{\Psi}} = \begin{pmatrix} 2q \\ 2p \end{pmatrix}$$

$$\frac{\partial^2 \pi_1(\mathbf{\Psi})}{\partial \mathbf{\Psi} \partial \mathbf{\Psi}^T} = \begin{pmatrix} -2 & -2 \\ -2 & -2 \end{pmatrix}; \quad \frac{\partial^2 \pi_2(\mathbf{\Psi})}{\partial \mathbf{\Psi} \partial \mathbf{\Psi}^T} = \begin{pmatrix} 0 & -2 \\ -2 & 0 \end{pmatrix}$$

$$\frac{\partial^2 \pi_3(\mathbf{\Psi})}{\partial \mathbf{\Psi} \partial \mathbf{\Psi}^T} = \begin{pmatrix} -2 & 0 \\ 2 & -2 \end{pmatrix}; \quad \frac{\partial^2 \pi_4(\mathbf{\Psi})}{\partial \mathbf{\Psi} \partial \mathbf{\Psi}^T} = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}.$$

这导致了如下的似然方程：

$$\partial \log L(\mathbf{\Psi})/\partial \mathbf{\Psi} = \sum_{j=1}^{4} \left(\frac{n_j}{\pi_j}\right) \frac{\partial \pi_j(\mathbf{\Psi})}{\partial \mathbf{\Psi}} = \mathbf{0},$$

以及对数似然的 Hessian 矩阵：

$$\partial^2 \log L(\mathbf{\Psi})/\partial \mathbf{\Psi} \partial \mathbf{\Psi}^T = \sum_{j=1}^{4} n_j \left\{ \left(\frac{1}{\pi_j}\right) \frac{\partial^2 \pi_j(\mathbf{\Psi})}{\partial \mathbf{\Psi} \mathbf{\Psi}^T} - \left(\frac{1}{\pi_j^2}\right) \frac{\partial \pi_j(\mathbf{\Psi})}{\partial \mathbf{\Psi}} \left(\frac{\partial \pi_j(\mathbf{\Psi})}{\partial \mathbf{\Psi}^T}\right) \right\}.$$

费舍尔（预期）信息矩阵由以下公式给出：

$$\begin{aligned} \mathcal{I}(\mathbf{\Psi}) &= E\left\{-\partial^2 \log L(\mathbf{\Psi})/\partial \mathbf{\Psi} \partial \mathbf{\Psi}^T\right\} \\ &= n \left\{ \sum_{j=1}^{4} \left(\frac{1}{\pi_j}\right) \frac{\partial \pi_j(\mathbf{\Psi})}{\partial \mathbf{\Psi}} \left(\frac{\partial \pi_j(\mathbf{\Psi})}{\partial \mathbf{\Psi}^T}\right) \right\}, \end{aligned}$$

当条件设置为 $\mathbf{\Psi} = \hat{\mathbf{\Psi}}$ 时，所得到的协方差矩阵为：

$$\begin{pmatrix} 0.000011008 & -0.000103688 \\ -0.000103688 & 0.000040212 \end{pmatrix};$$

参见Monahan (2011) 对这个例子中牛顿方法、评分方法和 EM 算法的有趣讨论.

2.2 EM Method

现在让我们讨论将 EM 算法应用于这个问题. 在将 EM 算法应用于这个问题时，一个自然的选择是完整数据向量为：

$$\boldsymbol{x} = \left(n_O, \boldsymbol{z}^T\right)^T,$$

其中

$$\boldsymbol{z} = (n_{AA}, n_{AO}, n_{BB}, n_{BO})^T$$

表示不可观测或“缺失”的数据。这些数据被认为是频率 n_{AA}, n_{AO}, n_{BB}（因为这些基因显现出来是 A，实则可能是 AO 或者 AA），和 n_{BO}，对应于表格 2.7 中的中间单元格。值得注意的是，由于总频率 n 是固定的，因此变量 $\boldsymbol{x}$ 中的五个单元格频率足以代表完整数据. 如果我们认为 $\boldsymbol{x}$ 的分布是关于表格 2中指定的六个单元格概率的 n 次抽取的多项式分布，那么很明显，观察到的频率向量 y 具有所需的多项式分布，如表1所指定.

Table 2: Complete-Data Structure for Example

Category (Cell)	Cell Probability	Notation for Frequency
O	r^2	n_O
AA	p^2	n_{AA}
AO	$2pr$	n_{AO}
BB	q^2	n_{BB}
BO	$2qr$	n_{BO}
AB	$2pq$	n_{AB}

对于 $\boldsymbol{\Psi}$，完整数据的对数似然函数可以写成（除了一个加法常数）如下形式：

$$\log L_c(\boldsymbol{\Psi}) = 2n_A^+ \log p + 2n_B^+ \log q + 2n_O^+ \log r, \tag{2.1}$$

其中

$$n_A^+ = n_{AA} + \frac{1}{2}n_{AO} + \frac{1}{2}n_{AB},$$
$$n_B^+ = n_{BB} + \frac{1}{2}n_{BO} + \frac{1}{2}n_{AB},$$

和

$$n_O^+ = n_O + \frac{1}{2}n_{AO} + \frac{1}{2}n_{BO}.$$

在这里，$\log L_c(\boldsymbol{\Psi})$ 表示的是完整数据的对数似然函数，它是用于估计统计模型参数的一个关键函数.

公式(2.1)呈现的是关于频率 $2n_A^+, 2n_B^+$ 和 $2n_O^+$ 的多项式对数似然函数，对应于三个单元格的概率 p, q 和 r. 因此，通过最大化(2.1) 得到这些概率的完整数据最大似然估计（MLE）如下：

$$\hat{p} = \frac{n_A^+}{n}; \quad \hat{q} = \frac{n_B^+}{n}. \tag{2.2}$$

当完整数据似然函数属于规则指数族时，E 步骤和 M 步骤会简化，就像这个例子一样. E 步骤仅需要计算当前条件下 $\boldsymbol{\Psi}$ 的充分统计量的期望值，这里是 $\left(n_A^+, n_B^+\right)^T$. M 步骤随后通过解由等同于这个期望的方程得到 $\boldsymbol{\Psi}^{(k+1)}$. 对于这个问题，实际上 $\boldsymbol{\Psi}^{(k+1)}$ 是通过用观察到的数据给定的 n_A^+ 和 n_B^+ 的当前条件期望来替换(2.2)右侧的值来得到的.

为了计算 n_A^+ 和 n_B^+（即 E 步）的这些条件期望，我们需要计算这个问题中不可观察数据 z 的条件期望. 考虑 z 的第一个元素，从变量 $\boldsymbol{x}$ 可知是 n_{AA}.

首先，可以验证，在给定 $\boldsymbol{y}$ 的条件下，n_A, n_{AA} 实际上具有二项分布，样本大小为 n_A，概率参数为

$$p^{(k)^2} / \left(p^{(k)^2} + 2p^{(k)}r^{(k)}\right),$$

Why 服从二项分布

在迭代的第 k 步，根据贝叶斯定理，有

$$p\left(\boldsymbol{z}^{(k+1)} = AA \mid \mathbf{y} = A, \boldsymbol{\Psi}^{(k)}\right) = \frac{p\left(\mathbf{y} = A \mid \boldsymbol{z}^{(k)} = AA, \boldsymbol{\Psi}^{(k)}\right) \cdot P\left(\boldsymbol{z}^{(k)} = AA \mid \boldsymbol{\Psi}^{(k)}\right)}{p\left(\mathbf{y} = A \mid \boldsymbol{z}^{(k)} = AA, \boldsymbol{\Psi}^{(k)}\right)}$$

$$= \frac{p\left(\mathbf{y} = A \mid \boldsymbol{z}^{(k)} = AA, \boldsymbol{\Psi}^{(k)}\right) \cdot P\left(\boldsymbol{z}^{(k)} = AA \mid \boldsymbol{\Psi}^{(k)}\right)}{p\left(\mathbf{y} = A, \boldsymbol{z}^{(k)} = AA \mid \boldsymbol{\Psi}^{(k)}\right) + p\left(\mathbf{y} = A, \boldsymbol{z}^{(k)} = AO \mid \boldsymbol{\Psi}^{(k)}\right)}$$

$$= \frac{p^{(k)^2} \cdot p\left(\mathbf{y} = A \mid \boldsymbol{z}^{(k)} = AA, \boldsymbol{\Psi}^{(k)}\right)}{(p^{(k)^2} + 2p^{(k)}r^{(k)}) \cdot p\left(\mathbf{y} = A \mid \boldsymbol{z}^{(k)} = AA, \boldsymbol{\Psi}^{(k)}\right)}$$

$$= \frac{p^{(k)^2}}{p^{(k)^2} + 2p^{(k)}r^{(k)}}.$$

这里 $\boldsymbol{\Psi}^{(k)}$ 代替了未知参数向量 $\boldsymbol{\Psi}$ 在第 $k+1$ 次迭代中的使用. 因此，给定 y 的 n_{AA} 的当前条件期望可以通过

$$E_{\boldsymbol{\Psi}^{(k)}}\left(n_{AA}\right) = n_{AA}^{(k)},$$

得到，其中

$$n_{AA}^{(k)} = n_A p^{(k)^2} / \left(p^{(k)^2} + 2p^{(k)}r^{(k)}\right). \tag{2.3}$$

同样，给定 $\boldsymbol{y}$ 的 n_{AO}, n_{BB} 和 n_{BO} 的当前条件期望也可以计算出来. M 步骤的执行给出

$$p^{(k+1)} = \left(n_{AA}^{(k)} + \frac{1}{2}n_{AO}^{(k)} + \frac{1}{2}n_{AB}\right)/n$$

和

$$q^{(k+1)} = \left(n_{BB}^{(k)} + \frac{1}{2}n_{BO}^{(k)} + \frac{1}{2}n_{AB}\right)/n.$$

这个问题的 EM 算法结果在表3中给出。可以将 $\boldsymbol{\Psi}$ 的最大似然估计（MLE）视为第 $k=4$ 次迭代时 $\Psi^{(k)}$ 的值.

Table 3: Results of the EM Algorithm for Example.

Iteration	$p^{(k)}$	$q^{(k)}$	$r^{(k)}$	$-\log L\left(\boldsymbol{\Psi}^{(k)}\right)$
0	0.26399	0.09299	0.64302	2.5619001
1	0.26436	0.09316	0.64248	2.5577875
2	0.26443	0.09317	0.64240	2.5577729
3	0.26444	0.09317	0.64239	2.5577726
4	0.26444	0.09317	0.64239	2.5577726

3 Monte Carlo Versions of the EM Algorithm

3.1 MONTE CARLO EM

在 EM 算法中，E 步骤可能难以实施，因为难以计算对数似然的期望值.Wei and Tanner (1990a,b) 建议采用蒙特卡洛方法，通过在第 $(k+1)$ 次迭代的 E 步骤中从条件分布 $f(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta})$ 模拟缺失数据 $\mathbf{s}$，然后最大化完全数据对数似然的近似条件期望：

$$\hat{Q}\left(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}\right) = \frac{1}{m} \sum_{i=1}^{m} \log\left[f\left(\mathbf{y}, \mathbf{s}^{(i)} \mid \boldsymbol{\theta}\right)\right], \tag{3.1}$$

当 $m \to \infty$ 时，这个公式的极限形式就是实际的 $Q\left(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}\right)$. 这正是蒙特卡洛积分的核心思想. 尽管最大化公式(3.1)通常可能很困难，但有时，在指数族情形下，最大化问题可以有解析形式的解.

! 在 MCEM（Monte Carlo Expectation-Maximization）中，蒙特卡洛误差在 E 步骤引入，丧失了单调性属性. 但在某些情况下，该算法以很高的概率接近一个极大化值.

3.2 Estimation of Standard Error with MCEM

最大似然估计（MLE）$\hat{\boldsymbol{\theta}}$ 的协方差矩阵估计由观测信息矩阵 $J(\hat{\boldsymbol{\theta}})$ 的逆给出，根据公式(1.10)，

$$J(\boldsymbol{\theta}) = E\{J_c(\boldsymbol{\theta}) \mid \mathbf{y}, \boldsymbol{\theta}\} - J_m(\boldsymbol{\theta}), \tag{3.2}$$

其中

$$J_c(\boldsymbol{\theta}) = -\frac{\partial^2 log f(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T},$$

现在考虑单一未知参数 $\boldsymbol{\theta}$ 的情况，并将完全数据对数似然函数 $\log f(\mathbf{y}, \mathbf{s} \mid \boldsymbol{\theta})$ 写为 $\log f(\mathbf{y}, \mathbf{s}^{(i)} \mid \boldsymbol{\theta})$. 为了通过蒙特卡洛评估计算(3.2)中的期望值，我们可以将 $J(\boldsymbol{\theta})$ 表示为以下形式：

$$\begin{aligned} J(\boldsymbol{\theta}) &\approx \frac{1}{m} \sum_{j=1}^{m} -\partial^2 \log f(\mathbf{y}, \mathbf{s}^{(i)} \mid \boldsymbol{\theta}) / \partial \boldsymbol{\theta}^2 \\ &+ \frac{1}{m} \sum_{j=1}^{m} \left\{ \partial \log f(\mathbf{y}, \mathbf{s}^{(i)} \mid \boldsymbol{\theta}) / \partial \boldsymbol{\theta} - \frac{1}{m} \sum_{j=1}^{m} \partial \log f(\mathbf{y}, \mathbf{s}^{(i)} \mid \boldsymbol{\theta}) / \partial \boldsymbol{\theta} \right\}^2 . \end{aligned}$$

其中 $\mathbf{s}^{(j)} (j = 1, \ldots, m)$ 是从缺失数据分布生成的，使用 MCEM 估计的 $\boldsymbol{\theta}$.

3.3 Continue to Section 2

在例子 Section 2 中，如果我们采用蒙特卡罗（MC）期望步骤（E-step），我们可以分别从两个独立的二项分布中抽取 $z_{11}, \ldots, z_{1m}$ 和 $z_{21}, \ldots, z_{2m}$，其中第一个二项分布的样本大小为 n_A，概率参数为

$$p^{(k)^2} / \left(p^{(k)^2} + 2p^{(k)} r^{(k)} \right),$$

而第二个二项分布的样本大小为 n_B，概率参数为

$$q^{(k)^2} / \left(q^{(k)^2} + 2q^{(k)} r^{(k)} \right),$$

在第 $k+1$ 次迭代中，用 $\boldsymbol{\Psi}^{(k)}$ 替代未知的参数向量 $\boldsymbol{\Psi}$. 然后，这些抽取的值可以代替方程(2.3)使用，如下所示：

$$n_{AA}^{(k)} = \bar{z}_{1m} = \frac{1}{m} \sum_{j=1}^{m} z_{1i}, \quad n_{BB}^{(k)} = \bar{z}_{2m} = \frac{1}{m} \sum_{j=1}^{m} z_{2i}.$$

在这个例子中，通过使用蒙特卡罗方法模拟缺失数据 z 的可能值，我们可以获得对 n_{AA} 和 n_{BB} 的估计。这是通过在每个迭代中从相应的二项分布中抽取样本来实现的，然后计算这些抽取值的平均数来估计 n_{AA} 和 n_{BB}. 这种方法允许我们在存在缺失或不完整数据时，使用模拟数据来估计缺失值，从而在 EM 算法中实施期望步骤.

References

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

John F Monahan. *Numerical methods of statistics*. Cambridge University Press, 2011.

Greg CG Wei and Martin A Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American statistical Association*, 85(411):699–704, 1990a.

Greg CG Wei and Martin A Tanner. Posterior computations for censored regression data. *Journal of the American Statistical Association*, 85(411):829–839, 1990b.

CF Jeff Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.