Statistical Computing
Chap. 2: Expectation–maximization Algorithm

LIU, Ran

Department of Statistics,
Beijing Normal University (Zhuhai Campus)

March 11, 2024

北京師範大學
BEIJING NORMAL UNIVERSITY

目录

介绍

频率学派通过极大化观测数据的对数似然函数，得到参数的估计。但是在有些情况下，得到的数据不是完整数据，还有一部分数据缺失。这导致无法直接极大化对数似然函数得到感兴趣的参数。

EM 算法（Expectation-Maximization Algorithm）就是用来解决隐含数据存在情况下的参数估计问题。该算法基础和收敛有效性等问题在 Dempster、Laird 和 Rubin 在 1977 年的文章“Maximum likelihood from incomplete data via the EM algorithm”中给出了详细的阐述。

它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.

基本思想

EM 算法的基本思想是：我们想要极大化观测数据的似然函数，但是很难求解。为此增设了一些辅助变量（或者这些变量原本就缺失了），这样的完整数据的似然函数更容易计算。

但完整数据中又存在未知的缺失变量，不能直接求导求最大值。所以 EM 基于观测数据和上一步估计出的参数值，根据设定的给定已知项下缺失项的条件分布，预测完整数据的对数似然函数，也就是 EM 算法的 E 步。

极大化预测的完整数据的对数似然函数，得到感兴趣参数的估计，此步称为 EM 算法的 M 步。E 步和 M 步反复迭代，直到估计参数基本无变化，算法收敛，得到最终的参数估计。

定义 $\boldsymbol{Y} = (\boldsymbol{X}, \boldsymbol{Z})$ 为完全数据，$\boldsymbol{X}$ 为观测数据，$\boldsymbol{Z}$ 为未观测到的缺失数据。

给定观测数据 x, 我们希望最大化观测数据的似然函数 $L(\boldsymbol{\theta}, x)$. 通常采用该似然函数会难以处理, 而采用 $\boldsymbol{Y}|\boldsymbol{\theta}$ 和 $\boldsymbol{Z}|(x, \boldsymbol{\theta})$ 的密度则较容易处理.

EM 算法通过采用这些较容易的密度避开了直接考虑 $L(\boldsymbol{\theta}|x)$.

但它本质是（后续会证明），**最大化观测数据的似然函数**。

目录

缺失数据，边际化和符号

完整数据为观测数据和缺失数据合起来，$\boldsymbol{Y} = (\boldsymbol{X}, \boldsymbol{Z})$.

设 $f_{\boldsymbol{X}}(x|\theta)$ 和 $f_{\boldsymbol{Y}}(y|\theta)$ 分别表示观测数据和完全数据的密度. 在给定观测数据下，缺失数据的条件密度为 $f_{\boldsymbol{Z}|\boldsymbol{X}}(z|x,\theta) = f_{\boldsymbol{Y}}(y|\theta)/f_{\boldsymbol{X}}(x|\theta)$.

Q 函数定义

设 $\theta^{(t)}$ 表示在迭代 t 时估计的最大值点，$t = 0, 1, ...$

定义 $\mathcal{Q}(\theta|\theta^{(t)})$ 为观测数据 $X = x$ 和当前迭代估计出的参数值 $\theta^{(t)}$ 条件下，完全数据的对数似然的对隐变量的条件期望. 即，

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= E\left\{\log L(\boldsymbol{\theta}|\boldsymbol{Y})|\boldsymbol{x}, \boldsymbol{\theta}^{(t)}\right\} \\ &= E\{\log f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta})|\boldsymbol{x}, \boldsymbol{\theta}^{(t)}\} \\ &= \int [\log f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta})] f_{\boldsymbol{Z}|\boldsymbol{X}}(z|\boldsymbol{x}, \boldsymbol{\theta}^{(t)}) \mathrm{d}z, \end{aligned}$$

其中，一旦给定 $X = x$, Y 的随机性就完全来自于 Z.

EM 算法

EM 从 θ^0 开始，然后在两步之间交替：E 表示期望，M 表示最大化. 该算法概括如下

1. 写出完全数据的似然函数.
2. **E 步**：计算条件期望 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E_z\{\log f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta})|\boldsymbol{x}, \boldsymbol{\theta}^{(t)}\}$.
3. **M 步**：寻求 $\boldsymbol{\theta}$ 来最大化 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. 设 $\boldsymbol{\theta}^{(t+1)}$ 为找到的点.
4. 返回 E 步，直到满足某停止规则为止.

在 EM 算法中，停止规则通常依赖于 $(\theta^{(t+1)} - \theta^{(t)})^{\mathrm{T}}(\theta^{(t+1)} - \theta^{(t)})$ 或 $|Q(\theta^{(t+1)}|\theta^{(t)}) - Q(\theta^{(t)}|\theta^{(t)})|$.

一般需要先设置缺失数据的边际分布和观测数据基于缺失数据的条件分布才能得到完整数据的似然函数，也即完整数据的联合概率，然后计算 Q 函数。(当然若能直接设置完整数据的概率密度也行。)

混合高斯

假设我们从同一年级的学生总体中采样身高，男生的比例为 p，则在未知性别的情况下，身高服从混合高斯模型，它的似然函数是：

$$
\begin{aligned}
L(\theta \mid X_1, \cdots, X_n) &= L(\mu_1, \mu_2, \sigma_1, \sigma_2, p \mid X_1, \cdots, X_n) \\
&= \prod_{i=1}^{n} \left\{ p \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right) + (1-p) \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(x_i - \mu_2)^2}{2\sigma_2^2} \right) \right\}.
\end{aligned}
$$

如果我们假设男女性别为潜变量 ($Z_i = 0$ 为女性，$Z_i = 1$ 为男性)，那么我们有 $X_i \mid Z_i = 1 \sim N(\mu_1, \sigma_1^2)$ 和 $X_i \mid Z_i = 0 \sim N(\mu_2, \sigma_2^2)$. 所以 full data 的 density 是

$$
\begin{aligned}
&\mathbb{P}\left\{ X_i, Z_i \mid \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p \right\} \\
&= \mathbb{P}\left\{ Z_i \mid \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p \right\} \mathbb{P}\left\{ X_i \mid Z_i, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2, p \right\} \\
&= \left[p \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(x_i - \mu_1)^2}{2\sigma_1^2} \right) \right]^{Z_i} \left[(1-p) \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left(-\frac{(x_i - \mu_2)^2}{2\sigma_2^2} \right) \right]^{1-Z_i}
\end{aligned}
$$

E Step

根据定义得到 $\mathcal{Q}$ 函数为

$$
\begin{aligned}
& E\left\{\ell(\theta \mid X, Z) \mid X, \theta^{(t)}\right\} \\
&= \sum_{i=1}^{n} \mathbb{E}\left\{Z_i \mid X_i; \theta^{(t)}\right\}\left[\ln p - \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma_1^2 - \frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right] \\
&+ \sum_{i=1}^{n} \mathbb{E}\left\{1 - Z_i \mid X_i; \theta^{(t)}\right\}\left[\ln(1-p) - \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln \sigma_2^2 - \frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right],
\end{aligned}
$$

E Step

其中

$$\mathbb{E}\left\{Z_i \mid X_i;\theta^{(t)}\right\} = 0\cdot\mathbb{P}\left\{Z_i=0 \mid X_i;\theta^{(t)}\right\} + 1\cdot\mathbb{P}\left\{Z_i=1 \mid X_i;\theta^{(t)}\right\}$$

$$= \mathbb{P}\left\{Z_i=1 \mid X_i;\theta^{(t)}\right\}$$

$$= \frac{\mathbb{P}\left\{Z_i=1, X_i;\theta^{(t)}\right\}}{\mathbb{P}\left\{Z_i=1, X_i;\theta^{(t)}\right\} + \mathbb{P}\left\{Z_i=0, X_i;\theta^{(t)}\right\}}$$

$$= \frac{\mathbb{P}\left\{X_i \mid Z_i=1;\theta^{(t)}\right\}\mathbb{P}\left\{Z_i=1;\theta^{(t)}\right\}}{\mathbb{P}\left\{X_i \mid Z_i=1;\theta^{(t)}\right\}\mathbb{P}\left\{Z_i=1;\theta^{(t)}\right\} + \mathbb{P}\left\{X_i \mid Z_i=0;\theta^{(t)}\right\}\mathbb{P}\left\{Z_i=0;\theta^{(t)}\right\}}$$

$$= \frac{p^{(t)}\frac{1}{\sqrt{2\pi}\sigma_1^{(t)}}\exp\left(-\frac{(x_i-\mu_1^{(t)})^2}{2(\sigma_1^{(t)})^2}\right)}{p^{(t)}\frac{1}{\sqrt{2\pi}\sigma_1^{(t)}}\exp\left(-\frac{(x_i-\mu_1^{(t)})^2}{2(\sigma_1^{(t)})^2}\right) + (1-p^{(t)})\frac{1}{\sqrt{2\pi}\sigma_2^{(t)}}\exp\left(-\frac{(x_i-\mu_2^{(t)})^2}{2(\sigma_2^{(t)})^2}\right)}$$

$$:= \xi_i^{(t)}$$

M step

将 $\mathbb{E}\left\{Z_i \mid X_i; \theta^{(t)}\right\}$ 代入原式 $\mathcal{Q}(\theta|\theta^{(t)})$，则有

$$\begin{aligned}\mathcal{Q} = &\sum_{i=1}^{n} \xi_i^{(t)} \left[\ln p - \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma_1^2 - \frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right] \\ &+ \sum_{i=1}^{n}(1-\xi_i^{(t)}) \left[\ln(1-p) - \frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma_2^2 - \frac{(x_i-\mu_2)^2}{2\sigma_2^2}\right],\end{aligned}$$

我们令 $\mathcal{Q}(\theta|\theta^{(t)})$ 的一阶导数等于 0：

$$\frac{\partial Q}{\partial p} = \sum_{i=1}^{n}\left(\frac{\xi_i^{(t)}}{p} - \frac{1-\xi_i^{(t)}}{1-p}\right) = 0 \Rightarrow \widehat{p} = \frac{\sum_{i=1}^{n}\xi_i^{(t)}}{n},$$

$$\frac{\partial Q}{\partial \mu_1} = \sum_{i=1}^{n}\xi_i^{(t)}\frac{-2(x_i-\mu_1)}{2\sigma_1^2} = 0 \Rightarrow \widehat{\mu}_1 = \frac{\sum_{i=1}^{n}\xi_i^{(t)}X_i}{\sum_{i=1}^{n}\xi_i^{(t)}},$$

$$\frac{\partial Q}{\partial \sigma_1^2} = \sum_{i=1}^{n}\xi_i^{(t)}\left[-\frac{1}{2\sigma_1^2} + \frac{(x_i-\mu_1)^2}{2(\sigma_1^2)^2}\right] = 0 \Rightarrow \widehat{\sigma}_1^2 = \frac{\sum_{i=1}^{n}\xi_i^{(t)}(X_i-\widehat{\mu}_1)^2}{\sum_{i=1}^{n}\xi_i^{(t)}}.$$

当是多个类别的高斯混合模型，如何求解？

一般混合模型

如果一个数据集是由 K 个不同总体组成，密度函数具有如下混合密度形式

$$p(x;\theta) = \sum_{k=1}^{K} \pi_k p_k(x;\theta_j),$$

其中我们有 $\sum_{k=1}^{K} \pi_j = 1$, 以及模型参数包括 $(\mathbf{\Pi}, \mathbf{\Theta})$.

假定观测到的数据为 $\boldsymbol{X} = \{x_1, x_2, \cdots, x_n\}$, 未观测到的数据为 $\boldsymbol{z} = \{z_1, z_2, \cdots, z_n\}$, 其中 $z_i \in \{1, \cdots, K\}$ 代表数据的来源，也就是:

$$p(x_i \mid z_i = k; \mathbf{\Theta}) = p_k(x;\theta_k), \quad p(z_i = k) = \pi_j.$$

则完全数据的似然函数可写成

$$L(\mathbf{\Pi}, \mathbf{\Theta}|\boldsymbol{X}, \boldsymbol{Z}) = \prod_{i=1}^{n} \prod_{k=1}^{K} (\pi_k p_k(x_i; \theta_k))^{I(z_i=k)}$$

则对数似然函数为

$$\ell(\mathbf{\Pi}, \mathbf{\Theta}|\boldsymbol{X}, \boldsymbol{Z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} I(z_i = k)[\log \pi_k + \log(p_k(x_i; \theta_k)]$$

接下来再求 Q 函数

$$\begin{aligned} Q(\theta|\theta^{(t)}) &= E_Z \left\{ \ell(\mathbf{\Pi}, \mathbf{\Theta}|\boldsymbol{X}, \boldsymbol{Z}) \mid \boldsymbol{X}, \mathbf{\Pi}^{(t)}, \mathbf{\Theta}^{(t)} \right\} \\ &= \sum_{i=1}^{n} \sum_{k=1}^{K} E(I(z_i = k) \mid \boldsymbol{X}, \mathbf{\Pi}^{(t)}, \mathbf{\Theta}^{(t)})[\log \pi_k + \log(p_k(x_i; \theta_k))] \end{aligned}$$

其中针对条件期望，我们有

$$\begin{aligned} E(I(z_i = k) \mid \boldsymbol{X}, \boldsymbol{\Pi}^{(t)}, \boldsymbol{\Theta}^{(t)}) &= p(z_i = k \mid \boldsymbol{X}, \boldsymbol{\Pi}^{(t)}, \boldsymbol{\Theta}^{(t)}) \\ &= \frac{p(x_i, z_i = k|\theta_k^{(t)})}{\sum_{l=1}^K p(x_i, z_i = l|\theta_l^{(t)})} \\ &= \frac{\pi_k^{(t)} p_k(x_i|\theta_k^{(t)})}{\sum_{l=1}^K \pi_l^{(t)} p_l(x_i|\theta_l^{(t)})} \end{aligned}$$

代入到 Q 函数里面后，求导得到更新公式。

目录

Jensen's inequality

琴生不等式，它给出积分的凸函数值和凸函数的积分值间的关系，在此不等式最简单形式中，阐明了对一平均做凸函数变换，会小于等于先做凸函数变换再平均。

若将琴生不等式应用在二点上，就回到了凸函数的基本性质：过一个凸函数上任意两点所作割线一定在这两点间的函数图象的上方，即：

$$tf(x_1) + (1-t)f(x_2) \geq f\left(tx_1 + (1-t)x_2\right), 0 \leq t \leq 1.$$

如果 X 是随机变量且 ϕ 是凸函数，则

$$\mathrm{E}[\varphi(X)] \geq \varphi(\mathrm{E}[X]).$$

收敛性

为了观察 EM 算法的收敛性质, 我们通过说明每个最大化步提高了观测数据的对数似然 $l(\theta|x)$ 开始. 首先注意到观测数据密度的对数可重新表达为

$$\log f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) = \log f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}) - \log f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})$$

因此,

$$E\{\log f_{\boldsymbol{X}}(x|\boldsymbol{\theta})|\boldsymbol{x},\boldsymbol{\theta}^{(t)}\} = E\{\log f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta})|\boldsymbol{x},\boldsymbol{\theta}^{(t)}\} - E\{\log f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})|\boldsymbol{x},\boldsymbol{\theta}^{(t)}\},$$

其中期望是关于 $Z|(\boldsymbol{x},\boldsymbol{\theta}^{(t)})$ 求取的. 于是

$$\log f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}),$$

其中

$$H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathrm{E}\{\log f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{Z}|\boldsymbol{x},\boldsymbol{\theta})|\boldsymbol{x},\boldsymbol{\theta}^{(t)}\}.$$

我们将证明 $H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ 是不断减小的.

$$
\begin{aligned}
H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) &= \mathrm{E}\{\log f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{Z}|\boldsymbol{x},\boldsymbol{\theta}^{(t)}) - \log f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{Z}|\boldsymbol{x},\boldsymbol{\theta})|\boldsymbol{x},\boldsymbol{\theta}^{(t)}\} \\
&= \int -\log\left[\frac{f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})}{f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta}^{(t)})}\right] f_{\boldsymbol{Z}|\boldsymbol{X}}(z|\boldsymbol{x},\boldsymbol{\theta}^{(t)})\mathrm{d}z \\
&\geqslant -\log \int f_{\mathbf{Z}|X}(z|x,\theta)\mathrm{d}z \\
&= 0.
\end{aligned}
$$

其中的放缩是 Jensen 不等式的应用，因为 $-\log u$ 关于 u 是单调递减，严格凸的. 所以我们的每一步 $\mathcal{Q}$ 增大而 H 减小，

$$
\log f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}^{(t+1)}) - \log f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}^{(t)}) \geqslant 0.
$$

则观测数据的对数似然不断上升.

在每次迭代中选择 $\theta^{(t+1)}$ 来最大化 $\mathcal{Q}(\theta|\theta^{(t)})$ 构成了标准的 EM 算法. 如果取而代之的是只简单选取任一个使得 $\mathcal{Q}(\theta^{(t+1)}|\theta^{(t)}) > \mathcal{Q}(\theta^{(t)}|\theta^{(t)})$ 的 $\theta^{(t+1)}$，那么得到的算法称作广义 EM, 或者 GEM.

值得注意的是，若 $-\ell(\hat{\theta}|x)$ 是正定的，则 EM 算法有线性收敛，且收敛速度跟缺失信息比例有关。当缺失信息比例比较大，收敛会非常慢.

为进一步理解 EM 如何工作, 注意到

$$H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \geqslant 0$$

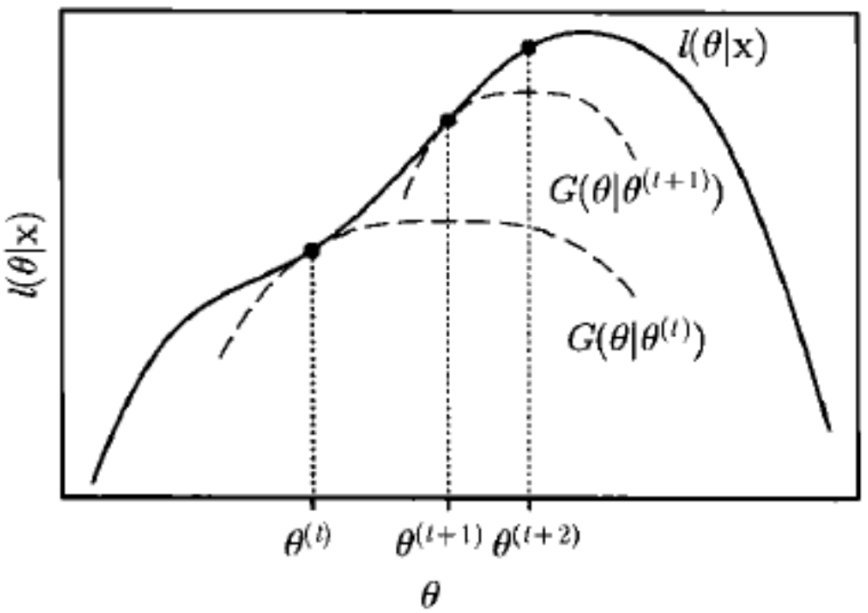
$$l(\boldsymbol{\theta}|\boldsymbol{x}) \geqslant Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + l(\boldsymbol{\theta}^{(t)}|\boldsymbol{x}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) = G(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}).$$

由于 $Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$ 和 $l(\boldsymbol{\theta}^{(t)}|\boldsymbol{x})$ 独立于 θ, 函数 Q 和 G 在相同的 θ 达到最大. 此外, G 在 $\theta^{(t)}$ 与 l 相切, 且在任一处低于 l. 我们说 G 是 l 的一个劣化函数 (minorizing function).

EM 策略将优化问题由观测数据的对数似然函数 l 转换到替代函数 G (有效地到 Q，因为后两项与 θ 无关，相当于常数), 这更便于最大化. G 的最大值点保证了在 l 值上的增加.

MM 算法 (Minorize-Maximization)

这个思想在下图中给出了解释. 每个 E 步等同于构造劣化函数 G, 而每个 M 步等同于最大化该函数以给出一个上升的路径.



Minorize-Maximization：每次迭代找到一个目标函数的下界函数（替代函数），求下界函数的最大值。

目录

方差估计

EM 的本质是极大化观测似然函数。因此, EM 参数估计后的协方差阵即观测数据的期望 Fisher 信息。一种方式是用观测数据的观测 Fisher 信息 $-\ell''(\hat{\theta}|x)$ 来近似, 其中 ℓ'' 是 $\log L(\theta|x)$ 的二阶导数，即 Hessian 矩阵.

在有些情形, 这个 Hessian 阵可以解析计算出来. 而在其他情形, 要得到或编码 Hessian 阵会很困难. 在这些场合, 可用多种其他方法来简化协方差阵的估计.

Louis' s Method

简单概括：用完全信息减去缺失信息来得到观测信息. 观测信息的逆即为想要的协方差.

回顾定义，我们有以下式子

$$\log f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) = \log f_{\boldsymbol{Y}}(\boldsymbol{y}|\boldsymbol{\theta}) - \log f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})$$

两边对隐变量，在观测数据和参数 $\boldsymbol{\omega} = \boldsymbol{\theta}$ 给定下，求条件期望后可得：

$$\log f_{\boldsymbol{X}}(\boldsymbol{x}|\boldsymbol{\theta}) = Q(\boldsymbol{\theta}|\boldsymbol{\omega})|_{\boldsymbol{\omega}=\boldsymbol{\theta}} - H(\boldsymbol{\theta}|\boldsymbol{\omega})|_{\boldsymbol{\omega}=\boldsymbol{\theta}}$$

我们再对第一个 θ 求二阶偏导数，则有

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

$$\hat{\boldsymbol{i}}_{\boldsymbol{X}}(\boldsymbol{\theta}) = \hat{\boldsymbol{i}}_{\boldsymbol{Y}}(\boldsymbol{\theta}) - \hat{\boldsymbol{i}}_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{\theta}),$$

其中 $\hat{\boldsymbol{i}}_{\boldsymbol{X}}(\boldsymbol{\theta})$ 为观测信息 (观测似然函数)，而 $\hat{\boldsymbol{i}}_{\boldsymbol{Y}}(\boldsymbol{\theta})$ 和 $\hat{\boldsymbol{i}}_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{\theta})$ 分别称作完全信息和缺失信息.

我们有

$$\hat{\boldsymbol{i}}_{\boldsymbol{Y}}(\boldsymbol{\theta}) = -\boldsymbol{Q}''(\boldsymbol{\theta}|\boldsymbol{\omega})|_{\boldsymbol{\omega}=\boldsymbol{\theta}} = -\mathrm{E}\{\boldsymbol{l}''(\boldsymbol{\theta}|\boldsymbol{Y})|\boldsymbol{x}, \boldsymbol{\theta}\}$$

并且我们可以证明

$$\hat{i}_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{\theta}) = \mathrm{var}\left\{\frac{\mathrm{d}\log f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{Z}|\boldsymbol{x}, \boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\right\},$$

其中方差是关于 $f_{Z|X}$ 求的. 进一步，因为得分期望为 0，故有

$$\hat{\boldsymbol{i}}_{\boldsymbol{Z}|\boldsymbol{X}}(\hat{\boldsymbol{\theta}}) = \int S_{\boldsymbol{Z}|\boldsymbol{X}}(\hat{\boldsymbol{\theta}})\boldsymbol{S}_{\boldsymbol{Z}|\boldsymbol{X}}(\hat{\boldsymbol{\theta}})^{\mathrm{T}} f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{X}, \hat{\boldsymbol{\theta}})\mathrm{d}\boldsymbol{z},$$

其中 $S_{\boldsymbol{Z}|\boldsymbol{X}}(\theta) = \frac{\mathrm{d}\log f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}$.

注意完全信息的定义其实跟之前信息定义不太一样，区别在于在 x 已经给定.

参考：https://bookdown.org/rdpeng/advstatcomp/missing-information-principle.html

观测信息等于完全信息减去缺失信息，该结果称为缺失信息法则。使得我们能够用完全数据似然和给定观测数据下缺失数据的条件密度来表达 $\hat{\boldsymbol{i}}_{\boldsymbol{X}}(\boldsymbol{\theta})$, 而且可以避免包括观测数据的可能复杂的边际似然的计算. 在某些情况下该方法可较容易得到并编码, 但它并不总比直接计算 $-\ell''(\theta|x)$ 容易.

如果 $\hat{\boldsymbol{i}}_{\boldsymbol{Y}}(\boldsymbol{\theta})$ 或者 $\hat{\boldsymbol{i}}_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{\theta})$ 难于解析计算, 可以通过 Monte Carlo 方法来估计. 例如,$\hat{\boldsymbol{i}}_{\boldsymbol{Y}}(\boldsymbol{\theta})$ 的最简单的 Monte Carlo 估计为

$$\frac{1}{m}\sum_{i=1}^{m}-\frac{\mathrm{d}^2\log f_{\boldsymbol{Y}}(\boldsymbol{y}_i|\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}\cdot\mathrm{d}\boldsymbol{\theta}},$$

其中对 $i=1,...,m$, $y_i=(x,z_i)$ 是模拟的完全数据集, 它是由观测数据和从 $f_{Z|X}$ 抽取的独立同分布假设下的缺失数据值 z 构成的. $\hat{\boldsymbol{i}}_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{\theta})$ 类似.

删失的指数数据

假定我们试图在模型 $Y_1, \cdots, Y_n \sim \text{i.i.d. Exp}(\lambda)$ 下观测到完全数据，但有些情形是右删失的（超过某个值就不变了）. 观测数据是 $\boldsymbol{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$, 其中 $\boldsymbol{x}_i = (\min(y_i, c_i), \delta_i)$,$c_i$ 是删失水平，如果 $y_i \leqslant c_i, \delta_i = 1$, 否则 $\delta_i = 0$.

指数分布的概率分布为 $f(y|\lambda) = \lambda e^{-\lambda x}, \quad x > 0.$

完全数据对数似然为 $l(\lambda|y_1, \cdots, y_n) = n \log \lambda - \lambda \sum_{i=1}^n y_i$. 这样

$$
\begin{aligned}
Q(\lambda|\lambda^{(t)}) &= \mathrm{E}(l(\lambda|Y_1, \cdots, Y_n)|\boldsymbol{x}, \lambda^{(t)}) \\
&= n \log \lambda - \lambda \sum_{i=1}^n \mathrm{E}\{Y_i|\boldsymbol{x}_i, \lambda^{(\boldsymbol{t})}\} \\
&= n \log \lambda - \lambda \sum_{i=1}^n \left[x_i \delta_i + (c_i + 1/\lambda^{(t)})(1 - \delta_i)\right] \\
&= n \log \lambda - \lambda \sum_{i=1}^n \left[x_i \delta_i + c_i(1 - \delta_i)\right] - C\lambda/\lambda^{(t)}
\end{aligned}
$$

其中 $C = \sum_{i=1}^{n}(1-\delta_i)$ 表示删失事件的个数. 则我们有 $-Q''(\lambda|\lambda^{(t)}) = n/\lambda^2$.

未观测到的变量 Z_i，也就是 y_i 超过 c_i 的部分（条件概率），有密度 $f_{Z_i|X}(z_i|x,\lambda) = \lambda\exp\{-\lambda(z_i-c_i)\}1_{\{z_i>c_i\}}$，则

$$\frac{\mathrm{d}\log f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{Z}|\boldsymbol{x},\lambda)}{\mathrm{d}\lambda} = C/\lambda - \sum_{\{i:\delta_i=0\}}(Z_i - c_i)$$

所以缺失信息是

$$\hat{i}_{\boldsymbol{Z}|\boldsymbol{X}}(\lambda) = -E\frac{\mathrm{d}^2\log f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{Z}|\boldsymbol{x},\lambda)}{\mathrm{d}\lambda^2} = \mathrm{var}\,\frac{\mathrm{d}\log f_{\boldsymbol{Z}|\boldsymbol{X}}(\boldsymbol{Z}|\boldsymbol{x},\lambda)}{\mathrm{d}\lambda} = C/\lambda^2.$$

应用 Louis 方法，

$$\hat{i}_{\boldsymbol{X}}(\lambda) = n/\lambda^2 - C/\lambda^2 = U/\lambda^2,$$

其中 $U = \sum_{i=1}^{n}\delta_i$ 表示未删失事件的个数. 对这个基本的例子，通过直接分析容易验证 $-l''(\lambda|\boldsymbol{x}) = U/\lambda^2$.

上述定义有个问题，x 和 z 到底是什么？指代不明。

$$Y_i = \begin{cases} X_i & Y_i \leq c_i \\ Z_i & Y_i > c_i \end{cases}$$

也就是我们的观测值是 $(x_1, x_2, c_3, c_4, ..., x_n)$，则我们有隐变量 (z_3, z_4).

则观测值的边际概率分布为:

$$p(x_i \leq c_i) = \int_0^{x_i} \lambda \exp\{-\lambda x\} dx = 1 - e^{-\lambda x_i}$$

$$p(x_i = c_i) = \int_{c_i}^{\infty} \lambda \exp\{-\lambda x\} dx = e^{-\lambda c_i}$$

而隐变量的条件概率为（因为已知 $x_i = c_i$ 了）:

$$p(z_i | x_i, \lambda) = p(y_i | y_i > c) I(y_i = z_i)$$

Y 并不必须为 X 和 Z 的并集，只需要 Y 的随机性与 (X, Z) 的随机性相互确定就行。换句话说，用 Y 能确定性地变换得到 (X, Z)；同样，用 (X, Z) 能确定性地变换得到 Y.

我们可检查概率分解是否成立

$$p(y|\lambda) = p(x|\lambda)p(z|x, \lambda)$$

在此例子中，删失标签 δ_i 已知。所以理应在观测变量和完全变量的集合中，都加入进去，但因为是固定的，计算过程中可以省去。

$$
\begin{aligned}
&\mathrm{E}\{Y_i|\boldsymbol{X}, \delta_i, c_i, \lambda^{(\boldsymbol{t})}\} \\
&= \delta_i \mathrm{E}\{Y_i|\boldsymbol{X}, \delta_i = 1, c_i, \lambda^{(\boldsymbol{t})}\} + (1-\delta_i)\mathrm{E}\{Y_i|\boldsymbol{X}, \delta_i = 0, c_i, \lambda^{(\boldsymbol{t})}\} \\
&= \delta_i x_i + (1-\delta_i)\mathrm{E}\{z_i|\boldsymbol{X}, \delta_i = 0, c_i, \lambda^{(\boldsymbol{t})}\} \\
&= \delta_i x_i + (1-\delta_i)(c_i + 1/\lambda^{(t)})
\end{aligned}
$$

构建完统计模型之后，再想一遍，这些观测值是怎样由模型产生的。(有时即使写出似然函数，回想数据生成过程还是会有问题)

比如此例中，观测值是怎么产生的？

自助法 (Bootstrap)

Bootstrap 是非常经典的求估计量方差的方法，不仅仅适用于 EM 算法的估计量. 对独立同分布的观测数据 $x_1, ..., x_n$ 来说:

1. 有放回地从 $x_1, ..., x_n$ 抽取 $x_1^*, ..., x_n^*$
2. 通过 EM 算法在这组新的数据上得到新的估计 $\hat{\theta}_b$
3. 重复上述过程，直到采样出 B 组参数 $\{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_B\}$.

我们假定 $\{\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_B\}$ 是从原本 EM 估计量 $\hat{\theta}$ 的分布中采样出来的样本，则可以方便地画出估计量的分布图以及计算方差或分位数.

若是使用参数自助法 (Parametric bootstrap), 则第一步改成用第二步估计出来的参数生成一组数据，而不是 resampling.

目录

EM 变型

EM 有时候表达式太过复杂，难以精确地计算期望或者求出最值，在这种情况下，我们选择牺牲一些计算速度或精确性做一些近似计算.

改进 E 步 (MCEM)

E 步需要找到在观测数据条件下完全数据的期望对数似然. 我们已经用 $Q(\theta|\theta^{(t)})$ 表示该期望，当该期望难以解析计算时，可以用 Monte Carlo 方法来近似.

Monte Carlo EM

Wei and Tanner 提出第 t 个 E 步可以用下面的两步替代

1. 从 $f_{\boldsymbol{Z}|\boldsymbol{X}}(z|\boldsymbol{x},\boldsymbol{\theta}^{(t)})$ 中抽取独立同分布的缺失数据集 $Z_1^{(t)},\cdots,Z_{m^{(t)}}^{(t)}$. 每个 $\boldsymbol{Z}_j^{(t)}$ 是用来补齐观测数据集的所有缺失值的一个向量，这样 $\mathbf{Y}_j=(\mathbf{x},\mathbf{Z}_j)$ 表示一个补齐的数据集，其中缺失值由 $\mathbf{z}_j$ 代替.
2. 计算 $\hat{Q}^{(t+1)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})=\frac{1}{m^{(t)}}\sum_{j=1}^{m^{(t)}}\log f_{\boldsymbol{Y}}(\boldsymbol{Y}_j^{(t)}|\boldsymbol{\theta})$.

那么 $\hat{Q}^{(t+1)}(\theta|\boldsymbol{\theta}^{(t)})$ 就是 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ 的 Monte Carlo 估计. M 步改为最大化 $\hat{Q}^{(t+1)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

删失的指数数据

假定我们试图在模型 $Y_1, \cdots, Y_n \sim \text{i.i.d. Exp}(\lambda)$ 下观测到完全数据，但有些情形是右删失的. 这样，观测数据是 $\boldsymbol{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$，其中 $\boldsymbol{x}_i = (\min(y_i, c_i), \delta_i)$,$c_i$ 是删失水平，如果 $y_i \leqslant c_i, \delta_i = 1$，否则 $\delta_i = 0$.

完全数据对数似然为 $l(\lambda|y_1, \cdots, y_n) = n \log \lambda - \lambda \sum_{i=1}^n y_i$. 且 Q 函数为

$$Q(\lambda|\lambda^{(t)}) = n \log \lambda - \lambda \sum_{i=1}^{n} [x_i \delta_i + c_i(1 - \delta_i)] - C\lambda/\lambda^{(t)}$$

所以我们的标准 EM 更新为

$$\lambda^{(t+1)} = \frac{n}{\sum_{i=1}^n x_i \delta_i + \sum_{i=1}^n c_i(1 - \delta_i) + C/\lambda^{(t)}}.$$

而 MCEM，将用 MC 的方法估计 Q 函数:

$$\hat{Q}^{(t+1)}(\lambda|\lambda^{(t)}) = n \log \lambda - \frac{\lambda}{m^{(t)}} \sum_{j=1}^{m^{(t)}} \boldsymbol{Y}_j^{\mathrm{T}} \mathbf{1},$$

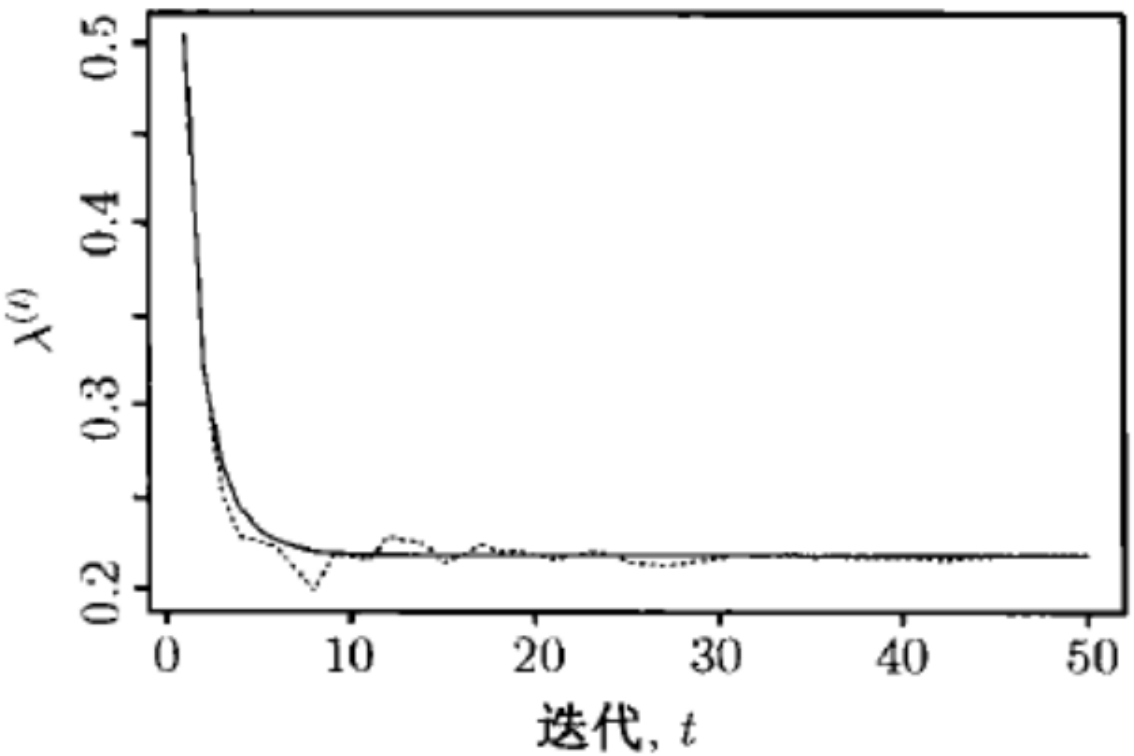
其中 $\mathbf{1}$ 是所有元素均为 1 的向量，Y_j 是包含未删失数据和模拟数据 $Z_j = (Z_{j1}, \cdots, Z_{jC})$ 的第 j 个补齐的数据集，
$Z_{jk} - c_k \sim$i.i.d.Exp$(\lambda^{(t)}), \quad k = 1, \cdots, C,$ 是用来代替删失值的. 令 $\hat{Q}'(\lambda|\lambda^{(t)}) = 0$ 且对 λ 求解得到

$$\lambda^{(t+1)} = \frac{n}{\sum_{j=1}^{(t)} \boldsymbol{Y}_j^{\mathrm{T}} \mathbf{1}/m^{(t)}}$$

作为 MCEM 的更新.

MCEM 和 EM 对比

模拟 $n = 30$ 个观测, 且包括 $C = 17$ 个删失观测. 用 MCEM 和 EM 估计 λ. 对 MCEM, 我们用 $m^{(t)} = 5^{1+\lfloor t/10 \rfloor}$, 其中 $\lfloor z \rfloor$ 表示 z 的整数部分. 两种算法的初始值均为 $\lambda^{(0)} = 0.5042$, 它是无视删失的所有 30 个数据值的均值. (也可以只用完全的数据均值)



改进 M 步

EM 算法的吸引力之一在于 $Q(\theta|\theta^{(t)})$ 的求导和最大化通常比不完全数据极大似然的计算简单，这是因为 $Q(\theta|\theta^{(t)})$ 与完全数据似然有关.

然而，在某些情况下，即使导出 $Q(\theta|\theta^{(t)})$ 的 E 步是直接了当的，M 步也不容易实施. 为此人们提出了多种策略以便于 M 步的实施.

ECM 算法

Meng 和 Rubin 的 ECM 算法是用一系列计算较简单的条件极大化 (conditional maximization) 步骤代替 M 步. 每次条件极大化均被设计为一个简单的优化问题, 该优化问题把 θ 限制在某特殊子空间, 使得可以得到解析解或非常初等的数值解.

我们称第 t 个 E 步后的所有 CM 步所形成的集合为一个 CM 循环. 因此, ECM 的第 t 次迭代包括第 t 个 E 步和第 t 次 CM 循环. 令 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 $Q(\theta|\theta^{(t)})$, 其中 $\theta^{(t+(s-1)/S)}$ 是在当前循环的第 $(s-1)$ 个 CM 步中求得的极大值点. 当 S 个 CM 步的整个循环完成时, 我们令 $\theta^{(t+1)} = \theta^{(t+S/S)}$ 并进行第 $(t+1)$ 次迭代的 E 步.

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 $Q(\theta|\theta^{(t)})$.

这定义看起来很复杂，但其实很好理解，原理就是用之前迭代得到的参数来约束这一步参数的搜索空间，以方便搜索。上式也可写成：

$$\max_{\theta} Q(\theta|\theta^{(t)}), \quad \text{when} \quad g_s(\theta, \theta^{(t+(s-1)/S)}) = 0.$$

构造有效 ECM 算法的技巧在于巧妙地选择约束条件. 通常，可自然地把 θ 分成 S 个子向量 $\theta = (\theta_1, \cdots, \theta_S)$. 然后在第 s 个 CM 步中，我们可以固定 θ 其余的元素而关于 θ_s 寻求最大化 Q. 这等同于用函数 $g_s(\theta) = (\theta_1, \cdots, \theta_{s-1}, \theta_{s+1}, \cdots, \theta_S)$ 导出的约束条件. 类似坐标下降法 (Coordinate Descent).

另外，第 s 个 CM 步也可以在固定 θ_s 下关于 θ 的其他元素最大化 Q. 在这种情况下，$g_s(\theta) = \theta_s$. 也可根据特定的问题背景想象其他的约束体系. ECM 的一种变型是在每两个 CM 步之间插入一个 E 步，由此在 CM 循环的每一个阶段均更新了 Q.

EM 梯度算法

如果最大化不能用解析的方法来实现，那么可以考虑采用迭代优化方法来实施每个 M 步. 这将会产生一个有嵌套迭代循环的算法. ECM 算法在 EM 算法的每次迭代中搬入 S 个条件最大化步骤，这也会产生嵌套迭代.

为避免嵌套循环的计算负担，Lange 提出用单步 Newton 法替代 M 步，从而可近似取得最大值而不用真正地精确求解. M 步是用由

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

给出的更新替代的，其中 $l'(\boldsymbol{\theta}^{(t)}|\boldsymbol{x})$ 是当前迭代得分函数的估值. 注意第二个等式是由 $\theta^{(t)}$ 最大化 $Q(\theta|\theta^{(t)}) - l(\theta|x)$ 的结论得来的. (一般来说 $l'(\theta|x)$ 的表达式要简单些)

当然我们也可以用最速梯度法代替 Newton 法，这样避免了计算 Q 函数矩阵的逆矩阵.

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