数理统计 week 6

学业辅导中心

极大似然法的直观想法

为介绍极大似然法的想法, 我们可以考虑如下的例子:

离散分布

设 X 是从 $\{0,1,2\}$ 中根据 P_θ 取值的单一观测值, 其中 $\theta=\theta_0$ 或 θ_1, $P_{\theta_j}(\{i\})$ 的值由下表给出:

	$x=0$	$x=1$	$x=2$
$\theta=\theta_0$	0.8	0.1	0.1
$\theta=\theta_1$	0.2	0.3	0.5

如果观测到 $X=0$, 那么它来自 P_{θ_0} 是非常合理的, 因为 $P_{\theta_0}(\{0\})$ 比 $P_{\theta_1}(\{0\})$ 大得多. 所以我们用 θ_0 估计 θ. 另一方面, 如果 $X=1$ 或 2 , 那么它来自 P_{θ_1} 是比较合理的, 尽管这种情况下概率之间的差别不像 $X=0$ 的情况下那么大. 这就建议了下面这个 θ 的估计:

$$T(X)=\begin{cases}\theta_0, & X=0\\ \theta_1, & X\neq 0\end{cases}$$

极大似然法的直观想法

上面的想法可以很容易地推广到 P_θ 是离散分布的情况, $\theta \in \Theta \subset \mathcal{R}^k$.
如果观测到 $X = x, \theta_1$ 比 θ_2 更合理当且仅当 $P_{\theta_1}(\{x\}) > P_{\theta_2}(\{x\})$.
我们就用 $\theta \in \Theta$ 上使得 $P_\theta(\{x\})$ 最大的 $\hat{\theta}$ 来估计 θ(如果这样的 $\hat{\theta}$ 存在).
用"合理"这个词而不是用"可能"是因为 θ 是看作非随机的, P_θ 并不是指 θ 的分布, 而是关于 θ 的分布.
若 $P_{\theta_1}(\{x\}) > P_{\theta_2}(\{x\})$, 在用 $\{\theta_1, \ldots, \theta_m\}$ 上离散均匀分布作先验的 Bayes 方法下, $P_\theta(\{x\})$ 与后验概率成比例, 那我们可以说 θ_1 比 θ_2 更可能.

$$P_{\theta_1}(\{x\}) \propto P(\{x\}, \theta = \theta_1), \quad P_{\theta_2}(\{x\}) \propto P(\{x\}, \theta = \theta_2)$$

MLE

极大似然估计

设 $X \in \mathscr{X}$ 是来自关于σ 有限测度 ν 的 p.d.f. f_θ 的一个样本, 其中 $\theta \in \Theta \subset \mathcal{R}^k$.

1. 对于每个 $x \in \mathscr{X}$, $f_\theta(x)$ 作为 θ 的函数称为似然函数, 并记作 $L(\theta)$.
2. 设 $\bar{\Theta}$ 是 Θ 的闭包. 当 $X = x$ 固定时, 满足 $L(\hat{\theta}) = \max_{\theta \in \bar{\Theta}} L(\theta)$ 的 $\hat{\theta} \in \bar{\Theta}$ 称为 θ 的极大似然估计值 (MLE). 如果 $\hat{\theta}$ 是 X 的一个 Borel 函数 a.e. ν, 那么 $\hat{\theta}$ 称为 θ 的极大似然估计量 (MLE).
3. 设 g 是从 Θ 到 $\mathcal{R}^p$ 的 Borel 函数, $p \leqslant k$. 如果 $\hat{\theta}$ 是 θ 的一个 MLE, 那么 $\hat{\vartheta} = g(\hat{\theta})$ 定义为 $\vartheta = g(\theta)$ 的一个 MLE.

关于上述定义的注记

为理解上述的定义, 我们省略过多繁琐的证明, 只为理解定义的内容:

- 为什么需要 σ 有限?

定义 (σ 有限)

若存在一列 $\{A_i\}_{i=1}^{\infty} \subset \Omega$, 且满足

- $\nu(A_i) < \infty$
- $\bigcup A_i = \Omega$

则称 $(\Omega, \mathcal{F})$ 上的测度 ν 是 σ 有限的.

- 概率测度是 σ 有限测度;
- Lebesgue 测度是 σ 有限测度;

Radon-Nikodym 定理

RN

设 ν 和 λ 是 $(\Omega, \mathcal{F})$ 上的两个测度, ν 是 σ 有限的. 如果 $\lambda \ll \nu$(λ 关于 ν 绝对连续), 即

$$\nu(A) = 0 \text{ 能推出} \lambda(A) = 0.$$

则存在 Ω 上的一个非负 Borel 函数 f, 使得

$$\lambda(A) = \int_A f d\nu, \quad A \in \mathcal{F}$$

此外, f 是唯一的 a.e. ν, 即, 如果对于任意 $A \in \mathcal{F}$ 有 $\lambda(A) = \int_A g d\nu$, 那么 $f = g$ a.e. ν.

上面的 f 被称为 RN 导数, 或者 λ 关于 ν 的密度.

对于 $f \geqslant 0$, 如果 $\int f d\nu = 1$ a.e. ν, 上述 λ 是一个概率测度, f 称为关于 ν 的概率密度函数 (p.d.f.).

极大似然估计

设 $X \in \mathscr{X}$ 是来自关于σ 有限测度 ν 的 p.d.f. f_θ 的一个样本, 其中 $\theta \in \Theta \subset \mathcal{R}^k$.

1. 对于每个 $x \in \mathscr{X}$, $f_\theta(x)$ 作为 θ 的函数称为似然函数, 并记作 $L(\theta)$.
2. 设 $\bar{\Theta}$ 是 Θ 的闭包. 当 $X = x$ 固定时, 满足 $L(\hat{\theta}) = \max_{\theta\in\bar{\Theta}} L(\theta)$ 的 $\hat{\theta} \in \bar{\Theta}$ 称为 θ 的极大似然估计值 (MLE). 如果 $\hat{\theta}$ 是 X 的一个 Borel 函数 a.e. ν, 那么 $\hat{\theta}$ 称为 θ 的极大似然估计量 (MLE).
3. 设 g 是从 Θ 到 $\mathcal{R}^p$ 的 Borel 函数, $p \leqslant k$. 如果 $\hat{\theta}$ 是 θ 的一个 MLE, 那么 $\hat{\vartheta} = g(\hat{\theta})$ 定义为 $\vartheta = g(\theta)$ 的一个 MLE.

为什么取闭包

注意到, MLE 的定义是利用 $\bar{\Theta}$ 而不是 Θ. 这是因为当 Θ 是开集时, $L(\theta)$ 的最大值可能不存在.
事实上, 这样的例子非常丰富,

Bernoulli 分布

设 $X_1, \ldots, X_n$ 是 i.i.d. 的 $0-1$ 随机变量, $P(X_1 = 1) = p \in \Theta = (0,1)$. 当观测到 $(X_1, \ldots, X_n) = (x_1, \ldots, x_n)$ 时, 似然函数是

$$L(p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{n\bar{x}}(1-p)^{n(1-\bar{x})},$$

其中 $\bar{x} = n^{-1}\sum_{i=1}^{n} x_i$. 注意到 $\bar{\Theta} = [0,1], \Theta^{\circ} = \Theta$. 似然方程变成

$$\frac{n\bar{x}}{p} - \frac{n(1-\bar{x})}{1-p} = 0.$$

为什么取闭包

Bernoulli 分布

如果 $0 < \bar{x} < 1$, 那么这个方程有唯一解 $\bar{x}$. $\log L(p)$ 的二阶导数是

$$-\frac{n\bar{x}}{p^2} - \frac{n(1-\bar{x})}{(1-p)^2},$$

它总是负的. 当 p 趋于 0 或 1(Θ 的边界) 时, $L(p) \to 0$. 因此, $\bar{x}$ 是 p 唯一的 MLE.

根据上面的结论, 当 $\bar{x} = 0$ 时, $L(p) = (1-p)^n$ 是 p 的严格递减函数, 因此它唯一的最大值点是 0 . 类似地, 当 $\bar{x} = 1$ 时, MLE 是 1 . 与之前的结论相结合, 我们可知 p 的 MLE 是 $\bar{x}$.
但是, 当 $\bar{x} = 0$ 或 1 时, $L(p)$ 的最大值在 $\Theta = (0,1)$ 上不存在, 尽管 $\sup_{p\in(0,1)} L(p) = 1$; MLE 在 Θ 外取值, 因此, 不是一个合理的估计. 然而, 如果 $p \in (0,1)$, 当 $n \to \infty$ 时, $\bar{x} = 0$ 或 1 的概率很快趋于 0 .

极大似然估计

设 $X \in \mathscr{X}$ 是来自关于σ 有限测度 ν 的 p.d.f. f_θ 的一个样本, 其中 $\theta \in \Theta \subset \mathcal{R}^k$.

1. 对于每个 $x \in \mathscr{X}$, $f_\theta(x)$ 作为 θ 的函数称为似然函数, 并记作 $L(\theta)$.
2. 设 $\bar{\Theta}$ 是 Θ 的闭包. 当 $X = x$ 固定时, 满足 $L(\hat{\theta}) = \max_{\theta \in \bar{\Theta}} L(\theta)$ 的 $\hat{\theta} \in \bar{\Theta}$ 称为 θ 的极大似然估计值 (MLE). 如果 $\hat{\theta}$ 是 X 的一个 Borel 函数 a.e. ν, 那么 $\hat{\theta}$ 称为 θ 的极大似然估计量 (MLE).
3. 设 g 是从 Θ 到 $\mathcal{R}^p$ 的 Borel 函数, $p \leqslant k$. 如果 $\hat{\theta}$ 是 θ 的一个 MLE, 那么 $\hat{\vartheta} = g(\hat{\theta})$ 定义为 $\vartheta = g(\theta)$ 的一个 MLE.

不变性性质

Invariance Property

设 $\{f_\theta : \theta \in \Theta\}$ 是关于 σ 有限测度的一族 p.d.f., 其中 $\Theta \subset \mathcal{R}^k$; h 是 Θ 到 $\Lambda \subset \mathcal{R}^p$ 的一个 Borel 函数, $1 \leqslant p \leqslant k$; 设 $\tilde{L}(\lambda) = \sup_{\theta : h(\theta) = \lambda} L(\theta)$ 是变换过的参数 λ 的似然函数. 证明, 如果 $\hat{\theta} \in \Theta$ 是 θ 的 MLE, 那么 $\hat{\lambda} = h(\hat{\theta})$ 使 $\tilde{L}(\lambda)$ 达到最大.

证明.

$$\tilde{L}(\lambda) = \sup_{\theta : h(\theta) = \lambda} L(\theta) \leq \sup_{\theta \in \Theta} L(\theta) = L(\hat{\theta}) \leq \tilde{L}(\hat{\lambda})$$

对上述式子左边的 λ 取值 $\hat{\lambda}$, 于是 $\hat{\lambda} = h(\hat{\theta})$ 是 λ 的极大似然估计. $\square$

MLE 的具体求法

如果参数空间 Θ 包含有限多个点, 那么 $\bar{\Theta} = \Theta$ 且 MLE 总是可以通过比较有限多个值 $L(\theta), \theta \in \Theta$, 而得到. 如果 $L(\theta)$ 在 Θ 的内点集 Θ° 上可微, 那么 MLE 可能的 $\theta \in \Theta^\circ$ 值必需满足

$$\frac{\partial L(\theta)}{\partial \theta} = 0,$$

这称为似然方程. 注意到, 满足 (4.50) 的 θ 可能是局部或全局最小、局部或全局最大或者就是简单的稳定点. 另外, 极值可能出现在 Θ 的边界或者当 $\|\theta\| \to \infty$ 时. 此外, 如果 $L(\theta)$ 不总是可微的, 那么极值可能出现在 $L(\theta)$ 的不可微或不连续点上. 因此, 分析整个似然函数来找到其最大值是很重要的.

因为 $\log x$ 是严格递增的, 且不失一般性可以假定 $L(\theta)$ 总是正的, 故 $\hat{\theta}$ 是 MLE 当且仅当它使得对数似然函数 $\log L(\theta)$ 达到最大值. 通常, 处理 $\log L(\theta)$ 以及与 (4.50) 类似的下式 (称为对数似然方程或简称似然方程) 更方便:

$$\frac{\partial \log L(\theta)}{\partial \theta} = 0$$

均匀分布长度的 MLE

例

设 $X_1, \ldots, X_n$ i.i.d. 服从区间 $\mathcal{I}_\theta$ 上的均匀分布, θ 未知. 考虑 $\mathcal{I}_\theta = (0, \theta)$ 和 $\theta > 0$ 的情况.

似然函数是 $L(\theta) = \theta^{-n} I_{\left(x_{(n)}, \infty\right)}(\theta)$, 它并不总是可微的. 在这个情况下, $\Theta^\circ = \left(0, x_{(n)}\right) \cup \left(x_{(n)}, \infty\right)$. 但是, 在 $\left(0, x_{(n)}\right)$ 上, $L \equiv 0$; 在 $\left(x_{(n)}, \infty\right)$ 上, $L'(\theta) = -n\theta^{n-1} < 0$ 对于所有的 θ 成立. 因此, 似然方程的办法在这个问题不适用. 因为 $L(\theta)$ 在 $\left(x_{(n)}, \infty\right)$ 上严格递减, 且在 $\left(0, x_{(n)}\right)$ 上为 0 , 所以 $L(\theta)$ 的唯一最大值点是 $x_{(n)}$, 是 $L(\theta)$ 的不连续点. 这说明 θ 的 MLE 是最大次序统计量 $X_{(n)}$.

均匀分布中点的 MLE

例

设 $X_1, \ldots, X_n$ i.i.d. 服从区间 $\mathcal{I}_\theta$ 上的均匀分布, θ 未知. 考虑 $\mathcal{I}_\theta = \left(\theta - \dfrac{1}{2}, \theta + \dfrac{1}{2}\right)$ 的情况, $\theta \in \mathcal{R}$.

似然函数是 $L(\theta) = I_{\left(x_{(n)}-\frac{1}{2}, x_{(1)}+\frac{1}{2}\right)}(\theta)$. 同样, 似然方程的方法不适用. 然而, 从定义可知, 任意满足 $x_{(n)} - \dfrac{1}{2} \leqslant T(x) \leqslant x_{(1)} + \dfrac{1}{2}$ 的统计量 $T(X)$ 是 θ 的 MLE. 这个例子说明 MLE 可能不是唯一的, 也可能是不合理的.

例题

练习

令 $X=(X_1,\cdots,X_n)$ 来源于密度 f_θ 的简单随机样本, 求下面 f_θ 中 θ 的极大似然估计.

1. $f_\theta(x)=\theta(1-x)^{\theta-1}I_{(0,1)}(x),\theta>1.$
2. $f_\theta(x)=\theta^x(1-\theta)^{1-x}I_{\{0,1\}}(x),\theta\in\left[\frac{1}{2},\frac{3}{4}\right].$
3. $f_\theta(x)=\frac{1}{\sqrt{2\pi}\sigma x}e^{-(\log x-\mu)^2/\left(2\sigma^2\right)}I_{(0,\infty)}(x),\theta=\left(\mu,\sigma^2\right)\in\mathcal{R}\times(0,\infty).$
4. $f_\theta(x)=\begin{pmatrix}\theta\\x\end{pmatrix}p^x(1-p)^{\theta-x}I_{\{0,1,\ldots,\theta\}}(x),\theta=1,2,\ldots,$ 这里 p 是已知的.

第一题

似然函数,

$$L(\theta) = \theta^n \prod_{i=1}^{n} (1 - X_i)^{\theta-1} I_{(0,1)}(X_i)$$

以及

$$\frac{\partial \log L(\theta)}{\partial \theta} = \frac{n}{\theta} + \sum_{i=1}^{n} \log(1 - X_i) \quad , \quad \frac{\partial^2 \log L(\theta)}{\partial \theta^2} = -\frac{n}{\theta^2} < 0$$

因此似然方程

$$\frac{\partial \log L(\theta)}{\partial \theta} = 0$$

有唯一解 $\hat{\theta} = -n / \sum_{i=1}^{n} \log(1 - X_i)$

当 $\hat{\theta} > 1$ 时, $\hat{\theta}$ 最大化似然函数, 当 $\hat{\theta} \leq 1$ 时, $L(\theta)$ 在 $(0, 1)$ 上单调减. 综上 MLE 是 $\max\{1, \hat{\theta}\}$.

第二题

似然函数,

$$L(\theta)=\theta^{\sum_{i=1}^{n}X_i}(1-\theta)^{n-\sum_{i=1}^{n}X_i}I_{(0,1)}(X_1,\cdots,X_n)$$

以及

$$\frac{\partial\log L(\theta)}{\partial\theta}=\frac{n\bar{X}}{\theta}-\frac{n-n\bar{X}}{1-\theta}$$

类似前面的讨论,

$$\hat{\theta}=\begin{cases}\frac{1}{2} & \text{if } \bar{X}\in\left[0,\frac{1}{2}\right)\\ \bar{X} & \text{if } \bar{X}\in\left[\frac{1}{2},\frac{3}{4}\right)\\ \frac{3}{4} & \text{if } \bar{X}\in\left(\frac{3}{4},1\right]\end{cases}$$

第三题

适当做变换, 令 $Y_i = \log X_i$, 于是

$$L(\theta) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2 - \sum_{i=1}^{n} Y_i\right\}$$

求解似然方程得到 μ 的极大似然估计,

$$\bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$$

σ^2 的极大似然估计 (要带入 $\mu = \bar{Y}$),

$$n^{-1}\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$$

第四题

似然函数是

$$L(\theta)=\prod_{i=1}^{n}\binom{\theta}{X_i}p^T(1-p)^{n\theta-T}I_{\{X_{(n)},X_{(n)}+1,\ldots\}}(\theta).$$

其中 $T=\sum_{i=1}^{n}X_i$, 现在对上述的离散的似然函数考察

$$\frac{L(\theta+1)}{L(\theta)}=(1-p)^n\prod_{i=1}^{n}\frac{\theta+1}{\theta+1-X_i}$$

由于 $(\theta+1)/(\theta+1-X_i)$ 关于 θ 是递减的, 于是 $\dfrac{L(\theta+1)}{L(\theta)}$ 关于 θ 是递减的, 再根据

$$\lim_{\theta\to\infty}L(\theta+1)/L(\theta)=(1-p)^n<1$$

因此, MLE 是

$$\max\left\{\theta:\theta\geq X_{(n)},L(\theta+1)/L(\theta)\geq 1,\theta\in\mathbb{N}\right\}$$

MLE 的数值计算

在应用中, 多数情形下 MLE 没有解析形式, 不得不用一些数值方法来计算 MLE. 普遍运用的数值方法是 Newton-Raphson 迭代法, 就是重复计算

$$\hat{\theta}^{(t+1)} = \hat{\theta}^{(t)} - \left[\left. \frac{\partial^2 \log L(\theta)}{\partial\theta\partial\theta^\top} \right|_{\theta=\hat{\theta}^{(t)}} \right]^{-1} \left. \frac{\partial \log L(\theta)}{\partial\theta} \right|_{\theta=\hat{\theta}^{(t)}},$$

$t = 0, 1, \ldots,$ 其中 $\hat{\theta}^{(0)}$ 是初始值, $\partial^2 \log L(\theta)/\partial\theta\partial\theta^\tau$ 对于每个 $\theta \in \Theta$ 假定满秩. 在每次迭代中, 如果我们用其期望值 $E\left[\partial^2 \log L(\theta)/\partial\theta\partial\theta^\top\right]$ 代替上式中的 $\partial^2 \log L(\theta)/\partial\theta\partial\theta^\top$, 其中期望是关于 P_θ 的, 那么这个方法就是 Fisher-scoring (得分) 法. 如果迭代收玫, 那么 $\hat{\theta}^{(\infty)}$ 或 t 充分大时的 $\hat{\theta}^{(t)}$ 是似然方程解的数值近似.

比较两个迭代法

考察正态分布

$$\log L(\theta) = -\frac{1}{2\sigma^2}\sum^{n}(x_i-\mu)^2 - \frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi).$$

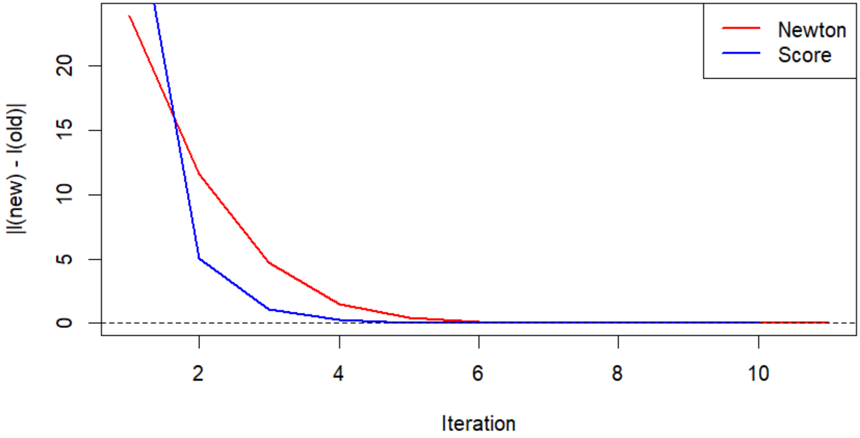
梯度

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0 \quad 和 \quad \frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2-\frac{n}{\sigma^2}=0.$$

Hessian

$$\frac{\partial^2\log L(\theta)}{\partial\theta\partial\theta^\top} = -\begin{pmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu) \\ \frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i-\mu) & \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i-\mu)^2-\frac{n}{2\sigma^4} \end{pmatrix}$$

比较两个迭代法



EM 算法: 背景

EM 算法最初用于缺失数据模型参数估计，现在已经用在许多优化问题中。设模型中包含 $\boldsymbol{X}_{\text{obs}}$ 和 $\boldsymbol{X}_{\text{mis}}$ 两个随机成分，有联合密度函数或概率函数 $f(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \theta), \theta$ 为未知参数。称 $f(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \theta)$ 为完全数据的密度，一般具有简单的形式。实际上我们只有 $\boldsymbol{X}_{\text{obs}}$ 的观测数据 $\boldsymbol{X}_{\text{obs}} = \boldsymbol{x}_{\text{obs}} \boldsymbol{X}_{\text{mis}}$ 不能观测得到，这一部分可能是缺失观测数据，也可能是潜在影响因素。所以实际的似然函数为

$$L(\theta) = f(\boldsymbol{x}_{\text{obs}} \mid \theta) = \int f(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \theta)\, d\boldsymbol{x}_{\text{mis}},$$

这个似然函数通常比完全数据的似然函数复杂得多，所以很难直接从 $L(\theta)$ 求最大似然估计。

EM 算法

EM 算法的想法是，已经有了参数的近似估计值 $\theta^{(t)}$ 后，假设 $(\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}})$ 近似服从完全密度 $f\left(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \theta^{(t)}\right)$, 这里 $\boldsymbol{X}_{\text{obs}} = \boldsymbol{x}_{\text{obs}}$ 已知，所以认为 $\boldsymbol{X}_{\text{mis}}$ 近似服从由 $f\left(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \theta^{(t)}\right)$ 导出的条件分布

$$f\left(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \theta^{(t)}\right) = \frac{f\left(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \theta^{(t)}\right)}{f\left(\boldsymbol{x}_{\text{obs}} \mid \theta^{(t)}\right)},$$

其中 $f\left(\boldsymbol{x}_{\text{obs}} \mid \theta^{(t)}\right)$ 是由 $f\left(\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}} \mid \theta^{(t)}\right)$ 决定的边缘密度。据此近似条件分布，在完全数据对数似然函数 $\log f\left(\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}_{\text{mis}} \mid \theta\right)$ 中，把 $\boldsymbol{X}_{\text{obs}} = \boldsymbol{x}_{\text{obs}}$ 看成已知，关于未知部分 $\boldsymbol{X}_{\text{mis}}$ 按密度 $f\left(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \theta^{(t)}\right)$ 求期望，得到 θ 的函数 $Q_t(\theta)$，再求 $Q_t(\theta)$ 的最大值点作为下一个 $\theta^{(t+1)}$ 。

EM 算法每次迭代有如下的 *E* 步 (期望步) 和 M 步 (最大化步)：

- **E 步**: 计算完全数据对数似然函数的期望 (Q 函数) $Q_t(\boldsymbol{\theta}) = E\{\log f(\boldsymbol{x}_{\text{obs}}, \boldsymbol{X}_{\text{mis}} \mid \boldsymbol{\theta})\}$, 其中期望针对随机变量 $\boldsymbol{X}_{\text{mis}}$ ，求期望时假定 $\boldsymbol{X}_{\text{mis}}$ 服从条件密度 $f\left(\boldsymbol{x}_{\text{mis}} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{\theta}^{(t)}\right)$ 决定的分布。
- **M 步**: 求 $Q_t(\boldsymbol{\theta})$ 的最大值点，记为 $\boldsymbol{\theta}^{(t+1)}$ ，迭代进入下一步。

EM 算法的有效性

EM 算法得到的估计序列 $\boldsymbol{\theta}^{(t)}$ 使得似然函数值 $L\left(\boldsymbol{\theta}^{(t)}\right)$ 单调不减。

EM 算法的注记

在适当正则性条件下，EM 算法的迭代序列 $\theta^{(t)}$ 依概率收敛到 $L(\theta)$ 的最大值点 $\hat{\theta}$。但是，上述定理仅保证 EM 算法最终能收敛，但不能保证 EM 算法会收敛到似然函数的全局最大值点，算法也可能收敛到局部极大值点或者鞍点。

在实际问题中，往往 E 步和 M 步都比较简单，有时 E 步和 M 步都有解析表达式，这时 EM 算法实现很简单。EM 算法优点是计算稳定，可以保持原有的参数约束，缺点是收敛可能很慢，尤其是接近最大值点时可能收敛更慢。如果似然函数不是凸函数，算法可能收敛不到全局最大值点，遇到这样的问题可以多取不同初值比较，用矩估计等合适的近似值作为初值。

混合分布

EM 算法可以用来估计混合分布的参数。设随机变量 $Y_1 \sim \mathrm{N}(\mu_1, \delta_1)$, $Y_2 \sim \mathrm{N}(\mu_2, \delta_2)$, Y_1, Y_2 独立。记 $N(\mu, \delta)$ 的密度为 $f(x \mid \mu, \delta)$ 。设随机变量 $W \sim \mathrm{b}(1, \lambda), 0 < \lambda < 1, W$ 与 Y_1, Y_2 独立，令

$$X = (1 - W)Y_1 + WY_2,$$

则 $W = 0$ 条件下 $X \sim \mathrm{N}(\mu_1, \delta_1)$, $W = 1$ 条件下 $X \sim \mathrm{N}(\mu_2, \delta_2)$, 但 X 的边缘密度为

$$f(x \mid \boldsymbol{\theta}) = (1 - \lambda) f(x \mid \mu_1, \delta_1) + \lambda f(x \mid \mu_2, \delta_2),$$

其中 $\boldsymbol{\theta} = (\mu_1, \delta_1, \mu_2, \delta_2, \lambda)$ 。

混合分布

设 X 有样本 $\boldsymbol{X}=(X_1,\ldots,X_n)$, 样本值为 $\boldsymbol{x}$ ， 实际观测数据的似然函数为

$$L(\boldsymbol{\theta})=\prod_{i=1}^{n} f(x_i \mid \boldsymbol{\theta})$$

这个函数是光滑函数但是形状很复杂，直接求极值很容易停留在局部极值点。用 EM 算法，以 $\boldsymbol{W}=(W_1,\ldots,W_n)$ 为没有观测到的部分, 完全数据的似然函数和对数似然函数为

$$\tilde{L}(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{W})=\prod_{W_i=0} f(x_i \mid \mu_1, \delta_1) \prod_{W_i=1} f(x_i \mid \mu_2, \delta_2)\, \lambda^{\sum_{i=1}^n W_i}(1-\lambda)^{n-\sum_{i=1}^n W_i},$$

$$\begin{aligned}\tilde{l}(\boldsymbol{\theta} \mid \boldsymbol{x}, \boldsymbol{W})=&\sum_{i=1}^{n}\left[(1-W_i)\log f(x_i \mid \mu_1, \delta_1)+W_i \log f(x_i \mid \mu_2, \delta_2)\right]\\&+\left(\sum_{i=1}^{n} W_i\right)\log\lambda+\left(n-\sum_{i=1}^{n} W_i\right)\log(1-\lambda).\end{aligned}$$

混合分布:E step

在 E 步，设已有 θ 的近似值 $\boldsymbol{\theta}^{(t)}=\left(\mu_1^{(t)},\delta_1^{(t)},\mu_2^{(t)},\delta_2^{(t)},\lambda^{(t)}\right)$, 以 $\boldsymbol{\theta}^{(t)}$ 为分布参数，在 $\boldsymbol{X}=\boldsymbol{x}$ 条件下，W_i 的条件分布为

$$\gamma_i^{(t)} \triangleq P\left(W_i=1 \mid \boldsymbol{x},\boldsymbol{\theta}^{(t)}\right)=P\left(W_i=1 \mid X_i=x_i,\boldsymbol{\theta}^{(t)}\right)$$

$$=\frac{\lambda^{(t)} f\left(x_i \mid \mu_2^{(t)},\delta_2^{(t)}\right)}{\left(1-\lambda^{(t)}\right) f\left(x_i \mid \mu_1^{(t)},\delta_1^{(t)}\right)+\lambda^{(t)} f\left(x_i \mid \mu_2^{(t)},\delta_2^{(t)}\right)}.$$

这里的推导类似于逆概率公式。利用 W_i 的条件分布求完全数据对数似然的期望，得

$$Q_t(\boldsymbol{\theta})=\sum_{i=1}^n\left[\left(1-\gamma_i^{(t)}\right)\log f\left(x_i \mid \mu_1,\delta_1\right)+\gamma_i^{(t)}\log f\left(x_i \mid \mu_2,\delta_2\right)\right]$$

$$+\left(\sum_{i=1}^n \gamma_i^{(t)}\right)\log\lambda+\left(n-\sum_{i=1}^n \gamma_i^{(t)}\right)\log(1-\lambda)$$

混合分布: M step

令 $\nabla Q_t(\boldsymbol{\theta}) = \mathbf{0}$, 求得 $Q_t(\boldsymbol{\theta})$ 的最大值点 $\boldsymbol{\theta}^{(t+1)}$ 为

$$
\begin{cases}
\mu_1^{(t+1)} = \dfrac{\sum_{i=1}^n \left(1-\gamma_i^{(t)}\right) x_i}{\sum_{i=1}^n \left(1-\gamma_i^{(t)}\right)} \\
\delta_1^{(t+1)} = \dfrac{\sum_{i=1}^n \left(1-\gamma_i^{(t)}\right)\left(x_i-\mu_1^{(t+1)}\right)^2}{\sum_{i=1}^n \left(1-\gamma_i^{(t)}\right)} \\
\mu_2^{(t+1)} = \dfrac{\sum_{i=1}^n \gamma_i^{(t)} x_i}{\sum_{i=1}^n \gamma_i^{(t)}} \\
\delta_2^{(t+1)} = \dfrac{\sum_{i=1}^n \gamma_i^{(t)} \left(x_i-\mu_2^{(t+1)}\right)^2}{\sum_{i=1}^n \gamma_i^{(t)}} \\
\lambda^{(t+1)} = \dfrac{1}{n} \sum\limits_{i=1}^n \gamma_i^{(t)}
\end{cases}
$$