EM Algorithm

19374128 白梓彤

April 2022

1 引言

在统计学中，期望最大化（EM）算法是一种迭代方法，用于寻找统计模型中参数的（局部）最大似然或最大后验（MAP）估计，其中模型取决于未观察到的隐变量。EM 迭代交替进行期望（E）步骤和最大化（M）步骤，前者为使用当前参数估计值评估的对数似然的期望值创建一个函数，后者计算参数，使在 E 步骤中找到的期望对数似然最大化。然后，这些参数估计值被用来确定下一个 E 步骤中潜在变量的分布。[2]

在方程不能直接求解的情况下，EM 算法被用来寻找统计模型的（局部）最大似然参数。通常这些模型除了未知的参数和已知的数据观测外，还涉及隐变量。寻找最大似然解通常需要对所有的未知值、参数和隐变量取似然函数的导数，并同时求解所得的方程。而对于含有隐变量的统计模型而言，这种方法很难实现。因为在这种情况下，计算的结果通常是一组交错的方程，其中参数的解需要隐变量的值，而隐变量的解又需要参数的值，但将一组方程代入另一组方程会产生一个无法解决的方程。

EM 算法的出发点是，通过数值方法求解参数和隐变量。首先从为参数和隐变量这两组未知集合中的一组设置一个初值，用它们来估计第二组，然后用这些新的值来寻找第一组的更好的估计值，然后在这两组之间不断交替，直到得到的值都收敛。此外，可以证明似然的导数在该点为零（或是近似为 0），这又意味着该点要么是局部最大值，要么是鞍点。一般来说，可能会出现多个最大值，不能保证找到全局最大值 [3]。有些可能性中也有奇异点，即无意义的最大值。例如，在一个混合模型中，EM 可能找到的解决方案之一是将其中一个成分的方差设为零，同一成分的平均参数等于其中一个数据点。

2 EM 算法

2.1 算法描述

假设统计模型有一组已知的观测数据 $\boldsymbol{X}$，一组无法观测到的隐变量 $\boldsymbol{Z}$，以及一组未知的参数 $\boldsymbol{\theta}$，似然函数为 $L(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}) = p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$，未知参数的最大似然估计（MLE）是通过最大化观测数据的边际似然来确定的：

$$L(\boldsymbol{\theta}; \mathbf{X}) = p(\mathbf{X} \mid \boldsymbol{\theta}) = \int p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\theta}) \mathrm{d}\mathbf{Z} = \int p(\mathbf{X} \mid \mathbf{Z}, \boldsymbol{\theta}) p(\mathbf{Z} \mid \boldsymbol{\theta}) \mathrm{d}\mathbf{Z} \tag{1}$$

然而，由于隐变量 $\boldsymbol{Z}$ 的存在，MLE 很难直接实现。

EM 算法则是通过迭代以下两个步骤来找到边际似然的最大似然估计（MLE）：

- E 步（Expectation step）：定义 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ 为关于 $\boldsymbol{\theta}$ 的对数似然函数的期望值，这个期望值与给定 $\boldsymbol{X}$ 以及当前估计的参数 $\boldsymbol{\theta}^{(\boldsymbol{t})}$ 时 $\boldsymbol{Z}$ 的条件概率分布有关：

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{(t)}}\left[\ln L(\boldsymbol{\theta};\boldsymbol{X},\boldsymbol{Z})\right] = \int \ln L(\boldsymbol{\theta};\boldsymbol{X},\boldsymbol{Z}) p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{(t)}) \mathrm{d}\boldsymbol{Z}$$

- M 步（Maximization step）：计算能够最大化 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ 的参数值：

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$

2.2 算法证明

将对数似然函数进行如下的变形：

$$\begin{aligned}
\ln L(\boldsymbol{\theta};\mathbf{X}) = \ln p(\mathbf{X} \mid \boldsymbol{\theta}) &= \ln \int p(\mathbf{X},\mathbf{Z} \mid \boldsymbol{\theta}) \mathrm{d}\mathbf{Z} \\
&= \ln \int \frac{p(\mathbf{X},\mathbf{Z} \mid \boldsymbol{\theta})}{p(\boldsymbol{Z} \mid \boldsymbol{X},\boldsymbol{\theta}^{(t)})} p(\boldsymbol{Z} \mid \boldsymbol{X},\boldsymbol{\theta}^{(t)}) \mathrm{d}\mathbf{Z} \\
&= \ln E\left(\frac{p(\mathbf{X},\mathbf{Z} \mid \boldsymbol{\theta})}{p(\boldsymbol{Z} \mid \boldsymbol{X},\boldsymbol{\theta}^{(t)})} \mid \boldsymbol{X},\boldsymbol{\theta}^{(t)}\right) \\
&\geq E\left(\ln \frac{p(\mathbf{X},\mathbf{Z} \mid \boldsymbol{\theta})}{p(\boldsymbol{Z} \mid \boldsymbol{X},\boldsymbol{\theta}^{(t)})} \mid \boldsymbol{X},\boldsymbol{\theta}^{(t)}\right) \\
&= E(\ln p(\mathbf{X},\mathbf{Z} \mid \boldsymbol{\theta}) \mid \boldsymbol{X},\boldsymbol{\theta}^{(t)}) - E(\ln p(\boldsymbol{Z} \mid \boldsymbol{X},\boldsymbol{\theta}^{(t)}) \mid \boldsymbol{X},\boldsymbol{\theta}^{(t)}) \\
&= Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - E(\ln p(\boldsymbol{Z} \mid \boldsymbol{X},\boldsymbol{\theta}^{(t)}) \mid \boldsymbol{X},\boldsymbol{\theta}^{(t)})
\end{aligned}$$

在上面的推导中，第三行到第四行是根据 Jensen 不等式得出，因为对数函数是一个凹函数，当 $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ 时该不等式取等号。令 $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - E(\ln p(\boldsymbol{Z} \mid \boldsymbol{X},\boldsymbol{\theta}^{(t)}) \mid \boldsymbol{X},\boldsymbol{\theta}^{(t)})$，可以得到下面的不等式：

$$\ln L(\boldsymbol{\theta};\mathbf{X}) \geq g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \tag{2}$$

这个不等式对于所有的 $\boldsymbol{\theta}$ 成立，并且当 $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$ 时取等号。因此，所有能够使得 $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ 增大的 $\boldsymbol{\theta}$ 值都会使得 $\ln L(\boldsymbol{\theta};\mathbf{X})$ 增大，也就是会使得 $\ln L(\boldsymbol{\theta}^{(t+1)};\mathbf{X}) \geq \ln L(\boldsymbol{\theta}^{(t)};\mathbf{X})$。EM 算法中的 M 步使通过最大化 $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ 可以等价地最大化 $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$，进而可以最大化似然函数。

3 EM 算法在 GMM 模型中的应用

EM 算法的一个常见应用是估计混合高斯模型的最大似然。

3.1 GMM 模型

假设有 n 个观测样本 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n\}$。为了使论证更加简洁，这里只讨论一元随机变量，因此写成 $\boldsymbol{X}=\{x_1,x_2,\ldots,x_n\}$。他们相互独立且服从相同的混合分布：

$$f(x)=\sum_{j=1}^{m}p_j\cdot f_j(x)=\sum_{j=1}^{m}p_j\cdot\frac{1}{\sqrt{2\pi\sigma_j^2}}e^{-\frac{(x-\mu_j)^2}{2\sigma_j^2}} \tag{3}$$

其中 $f_j(x)$ 是参数为 μ_j,σ_j 的正态分布的概率密度函数；$p_j\geq 0$ 且有 $\sum p_j=1$。因此 GMM 模型中的参数为：$\mu_j,\sigma_j,p_j,j=1,\ldots,m$。

3.2 应用 EM 算法

这里使用 EM 算法对这些参数进行估计，而不是直接使用数值优化的方法去估计最大似然。因此，定义一个额外的随机变量 y，满足 $P(y=j)=p_j,j=1,\ldots,m$。y 的取值 j 表示相应的样本值是由混合分布中第 j 个分布产生的。每一个样本都有一个这样的随机变量，因此得到集合 $\boldsymbol{Y}=\{y_1,y_2,\ldots,y_n\}$。因此可以得到：

$$f_{x|y}(x_i\mid y_i=j,\boldsymbol{\theta})=\frac{1}{\sqrt{2\pi\sigma_j^2}}e^{-(x_i-\mu_j)/2\sigma_j^2} \tag{4}$$

其中 $\boldsymbol{\theta}$ 为待求的模型参数，$\boldsymbol{\theta}=\{p_1,\ldots,p_m,\mu_1,\ldots,\mu_m,\sigma_1,\ldots,\sigma_m\}$。似然函数如下：

$$L(\theta;\boldsymbol{X},\boldsymbol{Y})=\prod_{i=1}^{n}p_{y_i}\frac{1}{\sqrt{2\pi\sigma_{y_i}^2}}e^{-(x_i-\mu_{y_i})/2\sigma_{y_i}^2} \tag{5}$$

EM 算法从一个初始值开始，依次进行 E 步和 M 步的迭代，直至收敛。

- E 步：首先计算 $Q(\boldsymbol{\theta}\mid\boldsymbol{\theta}^{(t)})=E_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta}^{(t)}}\left[\ln L(\boldsymbol{\theta};\boldsymbol{X},\boldsymbol{Y})\right]$。对其变形后可以得到：

$$\begin{aligned}Q(\boldsymbol{\theta}\mid\boldsymbol{\theta}^{(t)})&=E_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta}^{(t)}}\left[\ln L(\boldsymbol{\theta};\boldsymbol{X},\boldsymbol{Y})\right]\\&=E_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta}^{(t)}}\left[\sum_{i=1}^{n}\ln\left(p_{y_i}f_{x|y}(x_i\mid y_i,\boldsymbol{\theta})\right)\right]\\&=\sum_{i=1}^{n}\sum_{j=1}^{m}P(y_i=j\mid x_i,\boldsymbol{\theta}^{(t)})\ln\left(P(y_i=j\mid\boldsymbol{\theta})f_{x|y}(x_i\mid y_i=j,\boldsymbol{\theta})\right)\end{aligned} \tag{6}$$

上式中的 $P(y_i=j\mid x_i,\boldsymbol{\theta}^{(t)})$ 可以根据贝叶斯公式进行计算：

$$P(y_i=j\mid x_i,\boldsymbol{\theta}^{(t)})=\frac{P(y_i=j,x_i\mid\boldsymbol{\theta}^{(t)})}{P(x_i\mid\boldsymbol{\theta}^{(t)})}=\frac{f_{x|y}(x_i\mid y_i=j,\boldsymbol{\theta}^{(t)})P(y_i=j\mid\boldsymbol{\theta}^{(t)})}{\sum_{k=1}^{m}f_{x|y}(x_i\mid y_i=k,\boldsymbol{\theta}^{(t)})P(y_i=k\mid\boldsymbol{\theta}^{(t)})}$$

因此，式6可以计算。

- M 步：可以通过解析的方法最大化 $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(t)})$。令 $\partial Q / \partial \boldsymbol{\theta} = 0$，可以得到：

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^{n} x_i P(y_i = j \mid x_i, \boldsymbol{\theta}^{(t)})}{\sum_{i=1}^{n} P(y_i = j \mid x_i, \boldsymbol{\theta}^{(t)})} \tag{7}$$

$$(\sigma_j^{(t+1)})^2 = \frac{\sum_{i=1}^{n} (x_i - \mu_j)^2 P(y_i = j \mid x_i, \boldsymbol{\theta}^{(t)})}{P(y_i = j \mid x_i, \boldsymbol{\theta}^{(t)})} \tag{8}$$

$$p_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} P(y_i = j \mid x_i, \boldsymbol{\theta}^{(t)}) \tag{9}$$

用这些新的参数代替旧的参数，不断迭代，直至参数收敛。

3.3 身高问题

3.3.1 问题描述

假设人类的身高近似服从正态分布，且相同年龄段的男性和女性身高所服从的正态分布的参数有所不同 [1]。先有相同年龄段的男女生身高数据，试使用 EM 算法，估计混合高斯模型的参数：男生身高的均值和方差、女生身高的均值和方差、男女生的比例。

3.3.2 数据预处理与符号

男女生身高的频数直方图以及男女生比例如下所示。具体使用的数据见附录。

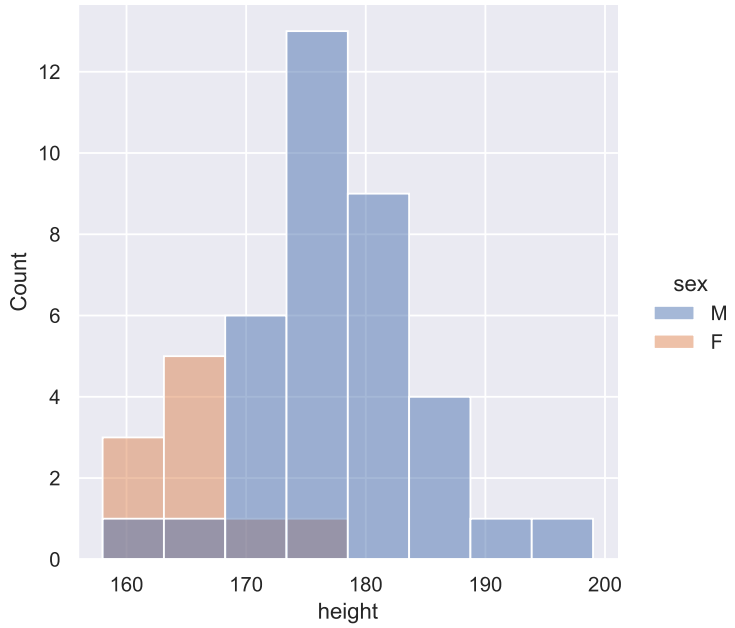

图 1: 男女生身高的频数直方图

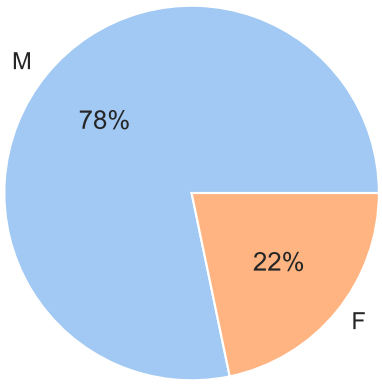

图 2: 男女生比例

3.3.3 计算与结果

根据3.2中描述的算法，在 Python 中进行编程计算，使用不同的初始值，可以得到如表1和图3、4、5、6、8、7所示。代码已上传至 gitee 仓库：`https://gitee.com/buaaxiaobaige/intelligent-system.git`

初值						结果					
μ_1	μ_2	σ_1	σ_2	p_1	p_2	μ_1	μ_2	σ_1	σ_2	p_1	p_2
180	165	5	5	0.7	0.3	177.284279	164.115894	7.319264	0.942801	0.861201	0.138799
170	160	5	5	0.5	0.5	177.284279	164.115894	7.319264	0.942801	0.861201	0.138799
170	160	10	10	0.5	0.5	176.74956	172.39867	8.232876	7.197745	0.702811	0.297189

表 1: 初值与计算结果

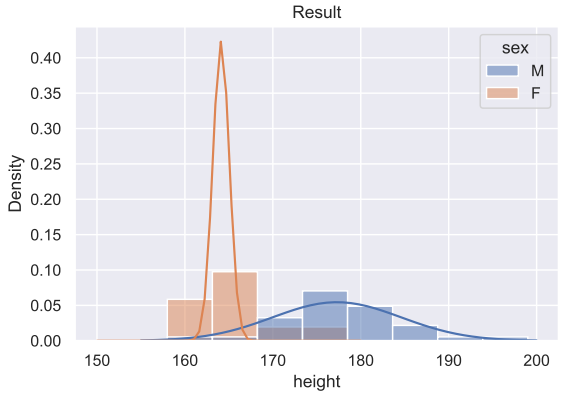

图 3: 第一组的结果

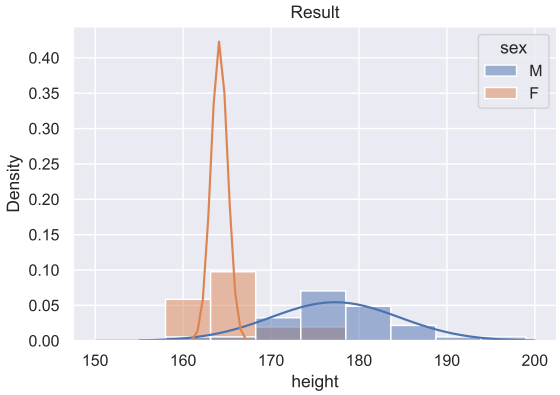

图 4: 第二组的结果

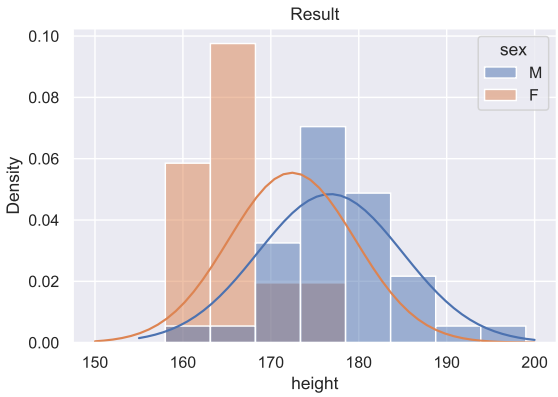

图 5: 第三组的结果

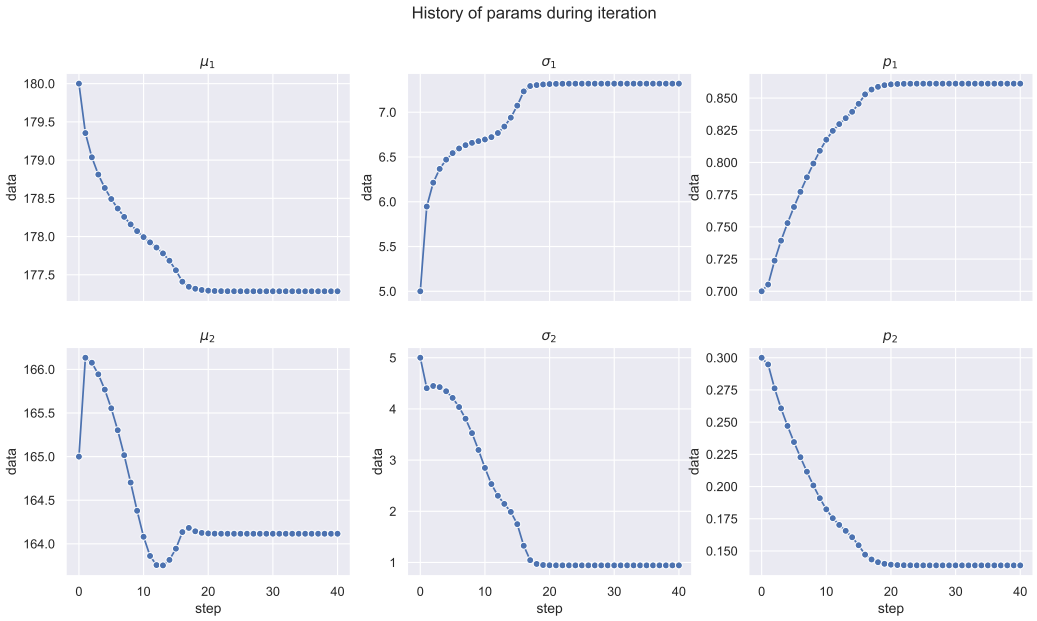

图 6: 第一组参数的变化

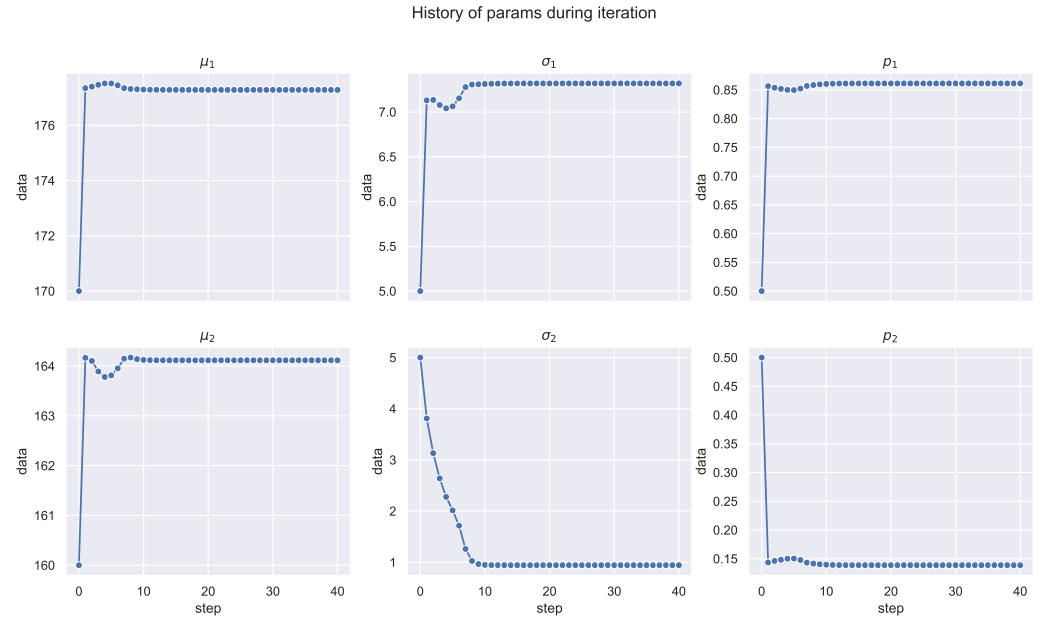

图 7: 第二组参数的变化

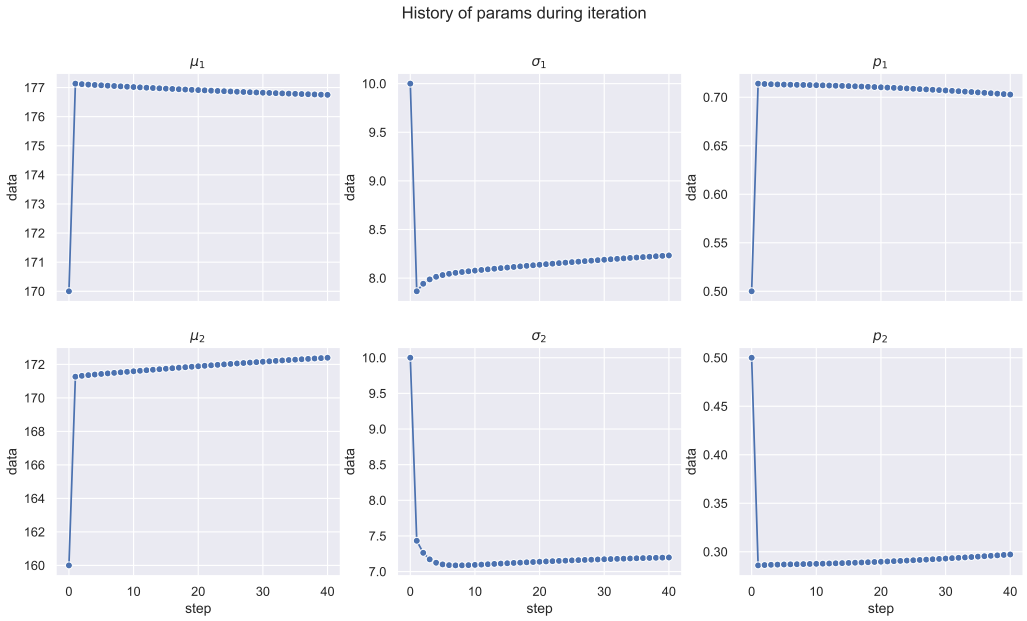

图 8: 第三组参数的变化

由计算结果可以看出，使用 EM 算法对 GMM 模型的参数进行估计，可以取得与真实值相近的结果。但是结果受初值的影响较大，因此在使用时需要有一个较为贴近真实值的先验估计。

4 总结

本文从 EM 算法的原理出发，给出了 EM 算法的描述以及推导过程，并将其应用在一元混合高斯分布模型上对男女生身高分布的参数进行估计，效果较好。

References

[1] Mark F Schilling, Ann E Watkins, and William Watkins. “Is human height bimodal?” In: *The American Statistician* 56.3 (2002), pp. 223–229.

[2] Wikipedia contributors. *Expectation–maximization algorithm — Wikipedia, The Free Encyclopedia*. [Online; accessed 5-April-2022]. 2022. URL: `https://en.wikipedia.org/w/index.php?title=Expectation%E2%80%93maximization_algorithm&oldid=1078223536`.

[3] CF Jeff Wu. “On the convergence properties of the EM algorithm”. In: *The Annals of statistics* (1983), pp. 95–103.

A 数据

性别	身高（cm）	性别	身高（cm）	性别	身高（cm）
M	175	M	173	M	185
F	163	M	172	M	180
M	182	M	186	M	177
M	173	F	163	F	165
F	165	M	158	M	175
M	178	F	163	F	165
M	176	M	171	M	178
M	172	M	179	M	181
M	199	F	167	F	174
M	183	M	182	F	165
M	178	M	182	M	182
M	186	M	190	M	176
M	178	M	171	M	175
F	170	M	174	M	175
M	178	M	187		
M	180	M	164		

表 2: 男女生身高数据