基于混合高斯模型的 EM 算法估计男女分布

刘张辰宇
ZY2203703

Abstract

由给定的 python 代码按照高斯分布随机生成男女样本分别为 1500、500 个，并将身高数据拼接得到混合身高的数据集，通过 csv 库导入数据集，在初始化参数之后，通过 em 算法迭代去估计混合数据集中的男女人数占比和两个高斯分布的参数,并与实际数据集进行对比验证。

Introduction

混合模型是一个可以用来表示在总体分布（distribution）中含有 K 个子分布的概率模型，换句话说，混合模型表示了观测数据在总体中的概率分布，它是一个由 K 个子分布组成的混合分布。混合模型不要求观测数据提供关于子分布的信息，来计算观测数据在总体分布中的概率。

高斯混合模型可以看作是由 K 个单高斯模型组合而成的模型,这 K 个子模型是混合模型的隐变量（Hidden variable）。一般来说，一个混合模型可以使用任何概率分布，这里使用高斯混合模型是因为高斯分布具备很好的数学性质以及良好的计算性能。

对于高斯混合模型，其似然函数是：

$$log^{L(\theta)} = \sum_{j=1}^{N} log^{P(x_j|\theta)} = \sum_{j=1}^{N} log^{(\sum_{k=1}^{K} \alpha_k \varphi(x|\theta_k))}$$

这里我们无法像单高斯模型那样使用最大似然法来求导求得使似然函数最大的参数,因为对于每个观测数据点来说，事先并不知道它是属于哪个子分布的（hidden variable），因此 log 里面还有求和，对于每个子模型都有未知的参数，直接求导无法计算。需要通过迭代的方法求解。

因此我们使用 EM 算法，进行迭代求解。EM 算法是一种迭代算法，1977 年由 Dempster 等人总结提出，用于含有隐变量（Hidden variable）的概率模型参数的最大似然估计。

每次迭代包含两个步骤：

1. E-Step：依据当前参数，计算每个数据 j 来自子模型 k 的可能性。
2. M-Step：根据 E 步算出的对每个分布的响应度，计算新一轮迭代的模型参数。

经过不断计算 E-Step 和 M-Step 直至参数收敛。

Methodology

本次实验已知男女混合身高数据集服从两个不同的高斯分布，由 EM 算法迭代估计两个高斯分布的参数及数据集中男女人数比。

M: GMM+EM

男女身高数据集服从高斯混合模型，假设男女身高分别服从两个高斯分布，即模型参数分别为男生身高均值 u1、标准差 sigma1，女生身高均值 u2、标准差 sigma2，以及数据集中男女人数占比 w1、w2。

对所有参数进行初始化之后，使用 EM 算法，其中 E 步根据身高数据集及所有参数，求出每个数据对两个模型的概率密度函数，再求出每个数据对两个模型的响应度 gama1、gama2；M 步则是由 E 步求得的响应度及数据集和模型均值更新所有的参数。

Experimental Studies

首先运行给定的 data.py 代码生成混合身高数据集，并保存到 height.csv 文件中，截取前 10 个身高数据，图如下：

	height
2	159.946495364941
3	164.31994834050886
4	166.36341008250395
5	168.23103822364115
6	165.4101400279389
7	164.35485881309813
8	161.2175342450846
9	164.38870277228506
10	159.96940705533768

Figure 1：数据集前十的身高

运行编写的 GMM+EM.py 代码，在初始化参数之后，对模型所有参数进行更新，并将每个身高数据对两个模型的响应度、混合身高的柱状图及估计得到的两个高斯分布的曲线、男女人数占比权重随迭代次数的变化都进行可视化，同时输出所有参数、迭代次数、真实身高均值。

A. **初始化参数**：u1 = 172, u2 = 161, sigma1 = sigma2 = 10, w1 = 0.8, w2 = 0.2

输出结果为：

```
F:\miniconda3\envs\pytorch\python.exe D:\workspace\Pycharm\pytorch\nlp\work2\GMM+EM.py
预测男生人数占比w1=0.762506 预测女生人数占比w2=0.237494
预测男生身高均值u1=175.893412 预测女生身高均值u2=163.746209
预测男生身高标准差sigma1=5.092035 预测女生身高标准差sigma2=2.828009
迭代次数=304
实际男生身高均值real_u1=176.035857 实际女生身高均值real_u2=163.926541
男生均值之差=0.142445 女生均值之差=0.180332
```

Figure 2：输出结果

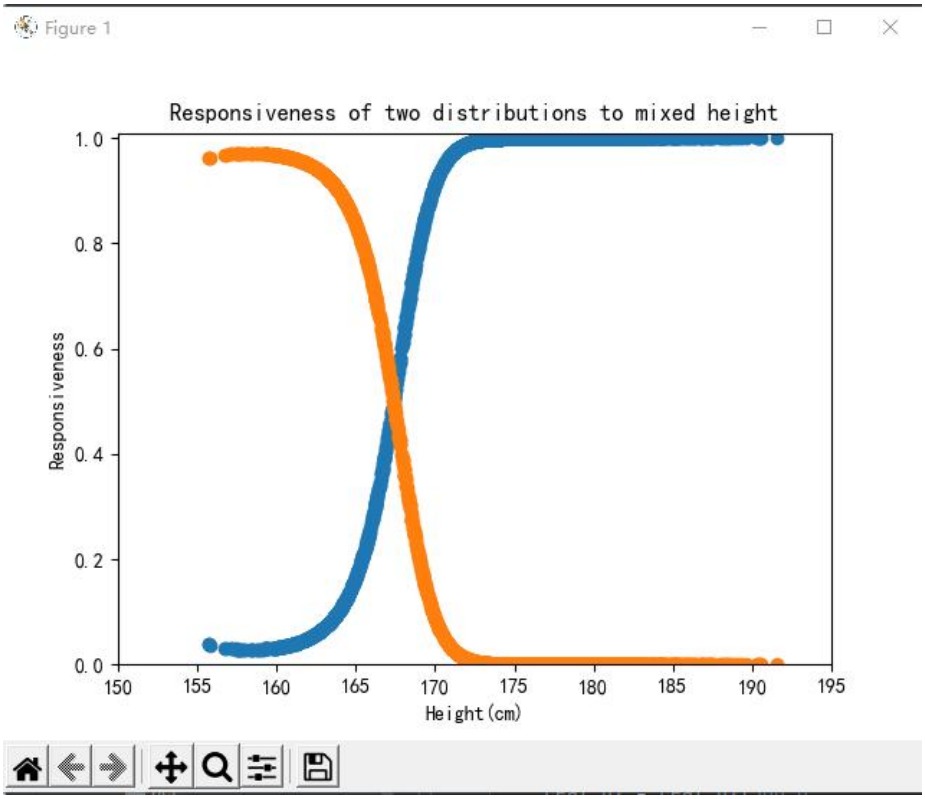

Figure 3：数据对两个分布的响应度

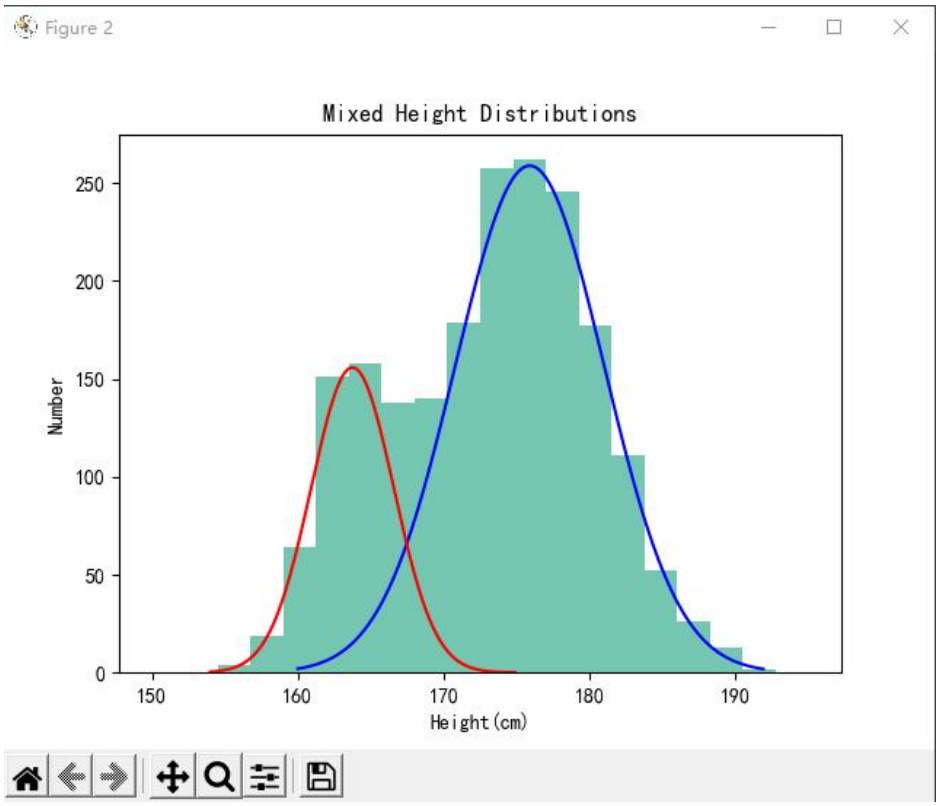

Figure 4：身高柱状图及估计的高斯分布曲线图

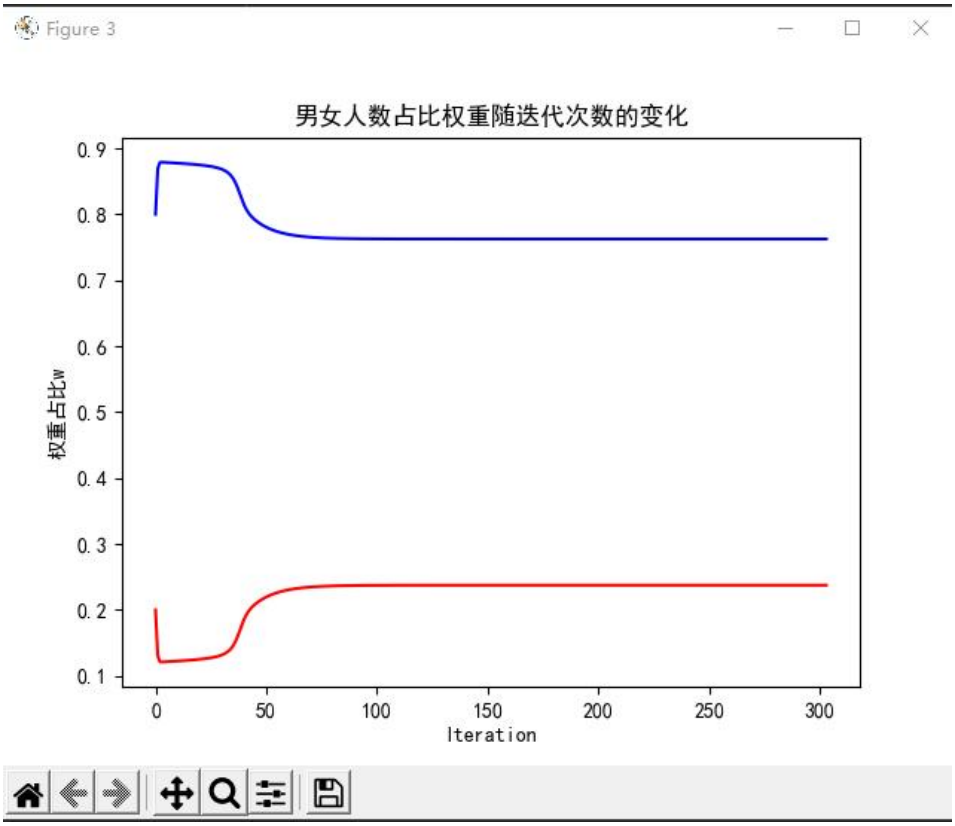

Figure 5：男女人数占比权重变化图

B. **初始化参数**：u1 = 170.1, u2 = 170, sigma1 = sigma2 = 10, w1 = 0.2, w2 = 0.8

输出结果为：

```
F:\miniconda3\envs\pytorch\python.exe D:\workspace\Pycharm\pytorch\nlp\work2\GMM+EM.py
预测男生人数占比w1=0.762506 预测女生人数占比w2=0.237494
预测男生身高均值u1=175.893412 预测女生身高均值u2=163.746209
预测男生身高标准差sigma1=5.092035 预测女生身高标准差sigma2=2.828009
迭代次数=519
实际男生身高均值real_u1=176.035857 实际女生身高均值real_u2=163.926541
男生均值之差=0.142445 女生均值之差=0.180332
```

Figure 6：输出结果

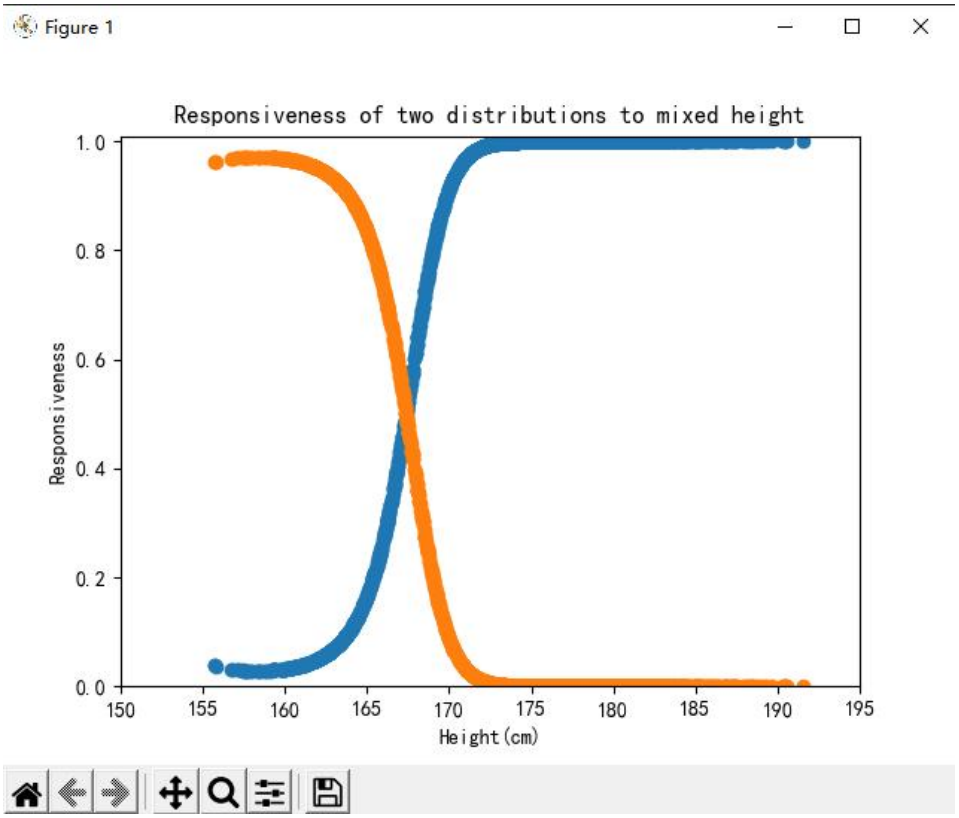

Figure 7：数据对两个分布的响应度

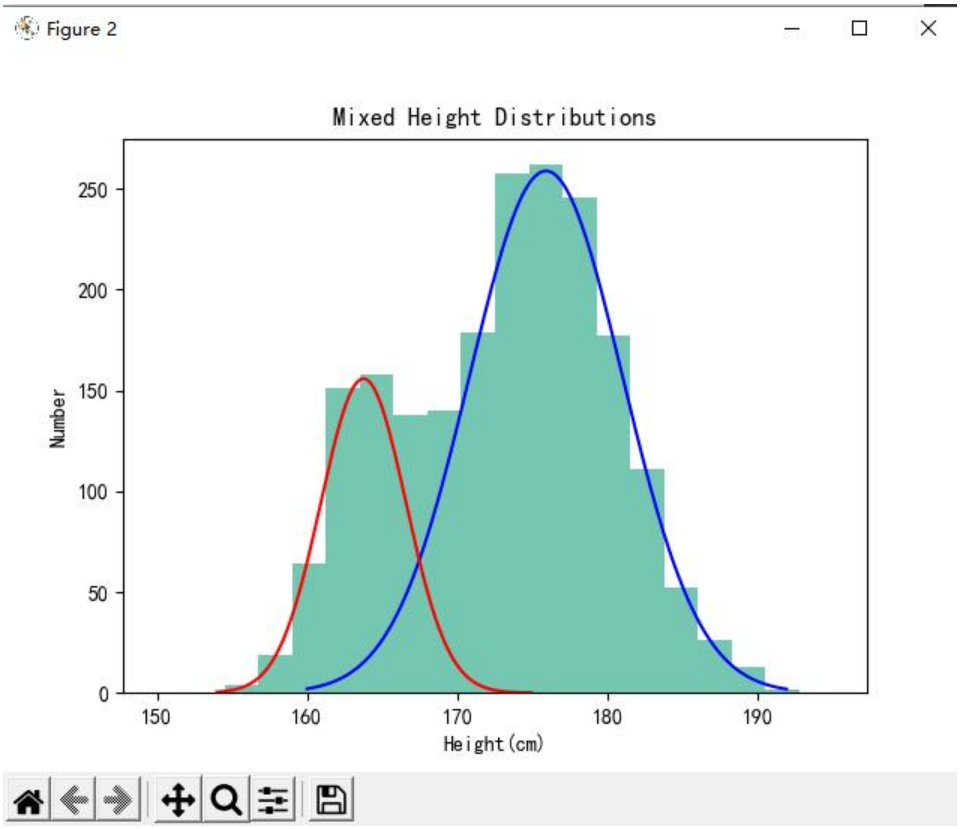

Figure 8：身高柱状图及估计的高斯分布曲线图

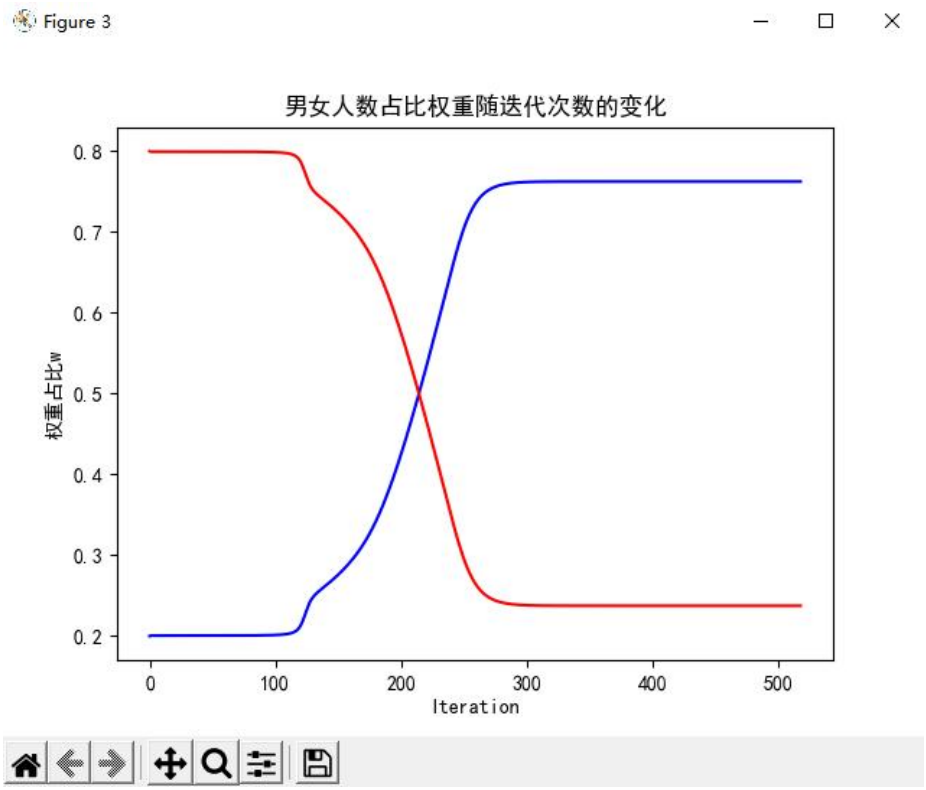

Figure 9：男女人数占比权重变化图

C. **初始化参数**：u1 = 169.9, u2 = 170, sigma1 = sigma2 = 10, w1 = 0.8, w2 = 0.2

输出结果为：

```
F:\miniconda3\envs\pytorch\python.exe D:\workspace\Pycharm\pytorch\nlp\work2\GMM+EM.py
预测男生人数占比w1=0.237494 预测女生人数占比w2=0.762506
预测男生身高均值u1=163.746209 预测女生身高均值u2=175.893412
预测男生身高标准差sigma1=2.828009 预测女生身高标准差sigma2=5.092035
迭代次数=518
实际男生身高均值real_u1=176.035857 实际女生身高均值real_u2=163.926541
男生均值之差=12.289648 女生均值之差=11.966871
```

Figure 10：输出结果

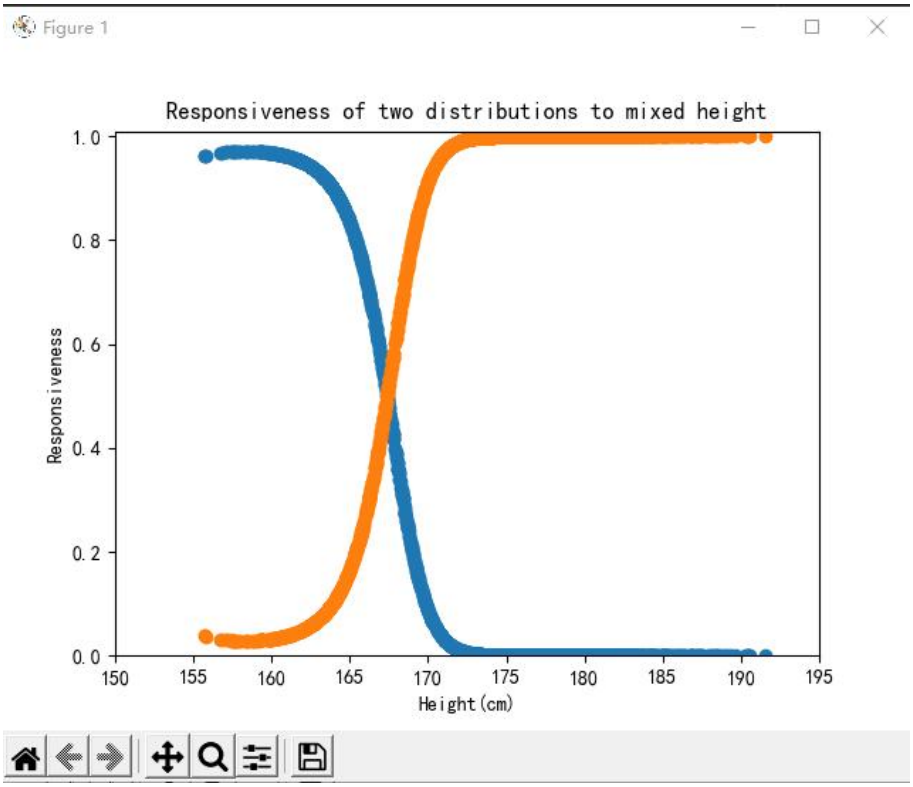

Figure 11：数据对两个分布的响应度

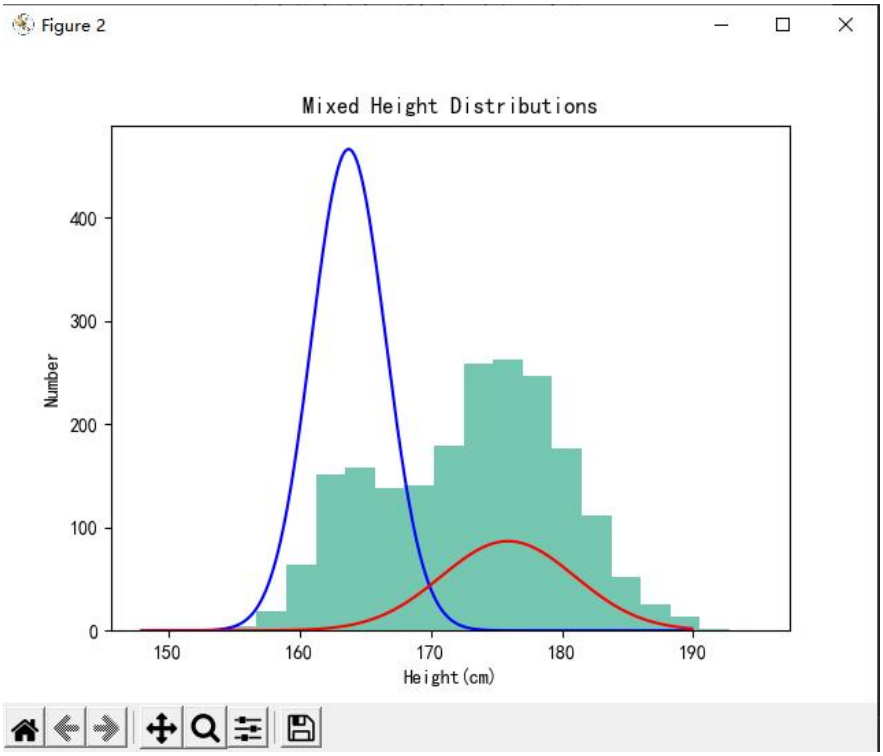

Figure 12：身高柱状图及估计的高斯分布曲线图

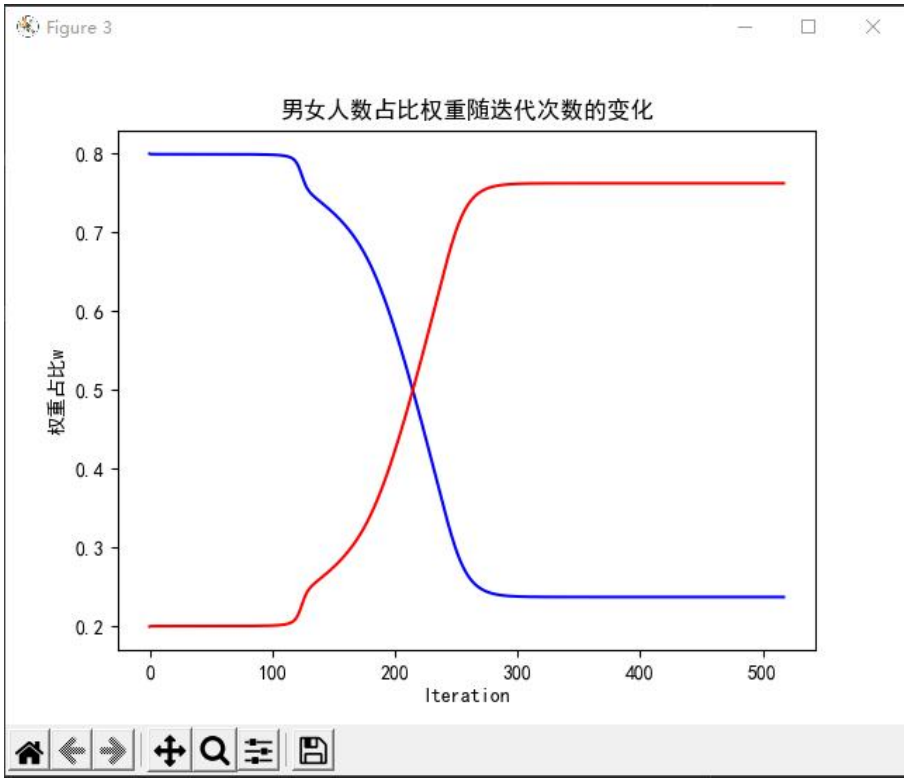

Figure 13：男女人数占比权重变化图

对于本问题，在进行 A 参数初始化时，选择了较为不错的参数，男生身高均值 u1=172 远大于女生身高均值 u2=161，占比权重 w1=0.8 也接近真实权重(0.75)，故迭代 304 次即参数收敛，最终的参数也都比较接近真实值。

而在进行 B 参数初始化时，选择的参数 u1=170.1，u2=170，w1=0.2，是一组比较差的参数，但 u1 仍大于 u2，故最终的参数也能收敛至 A 的结果，但迭代次数增加到 519 次。

但在进行 C 参数初始化时，选择的参数 u1=169.9，u2=170，w1=0.8，是一组很差的参数，即使 w1 比较接近真实值，但 u1 此时小于 u2，导致最终两个分布的参数交换，且迭代次数也达到 518 次。

Conclusions

对于混合高斯模型这种无法直接对似然函数求导得模型参数的问题，使用 EM 算法迭代可以较为准确地获得模型的参数，从而可以预测数据属于哪个模型。但在参数进行初始化时需要一定的知识。

对于本问题，进行参数初始化时，男生身高均值 u1 必须大于女生身高均值 u2(若等于，则 EM 算法失效)，否则结果出错；w1 和 w2 的取值好坏影响收敛速度即反映到迭代次数上，但不影响最终参数结果。

References

[1]McLachlan G J, Krishnan T. The EM algorithm and extensions[M]. John Wiley & Sons, 2007.