**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：

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：

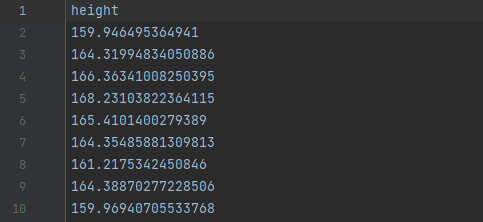


Figure 1：数据集前十的身高

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。

1. **初始化参数**：u1 = 172, u2 = 161, sigma1 = sigma2 = 10,

w1 = 0.8, w2 = 0.2

输出结果为：

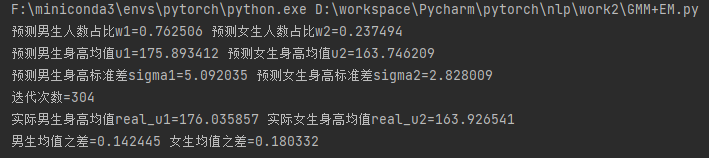


Figure 2：输出结果

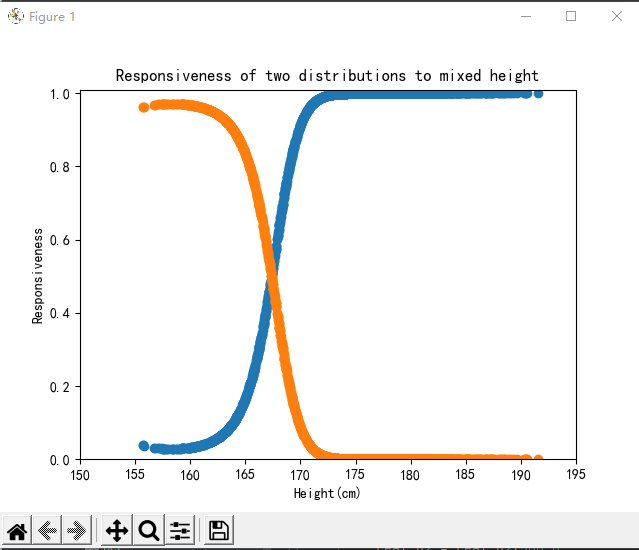


Figure 3：数据对两个分布的响应度

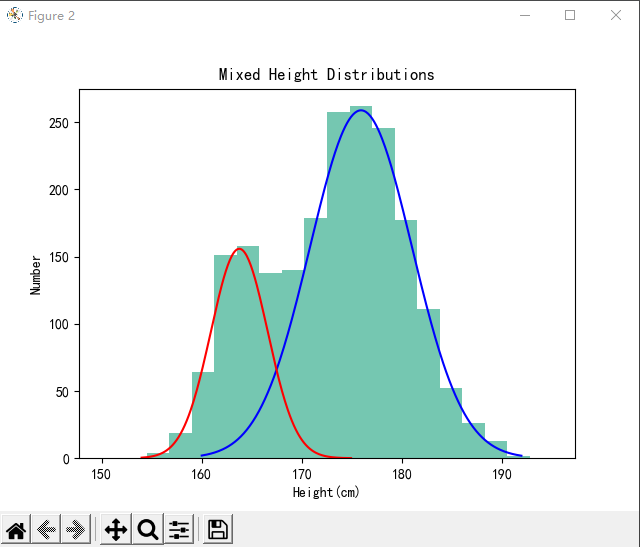


Figure 4：身高柱状图及估计的高斯分布曲线图

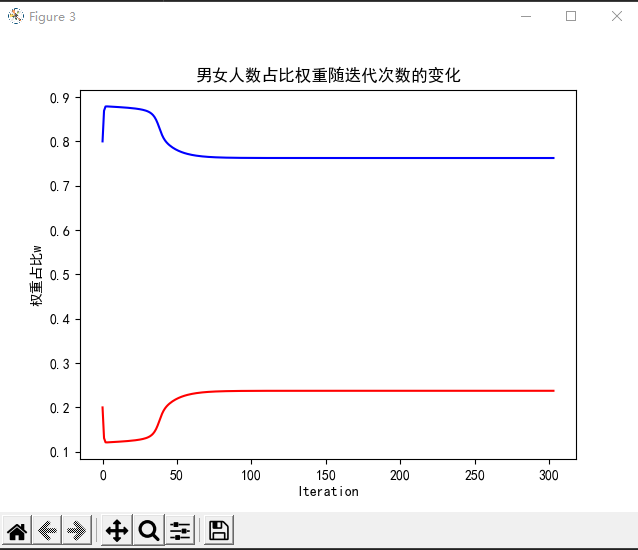


Figure 5：男女人数占比权重变化图

1. **初始化参数**：u1 = 170.1, u2 = 170, sigma1 = sigma2 = 10,

w1 = 0.2, w2 = 0.8

输出结果为：

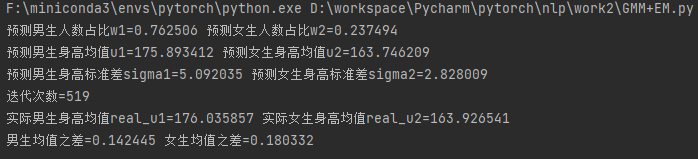


Figure 6：输出结果

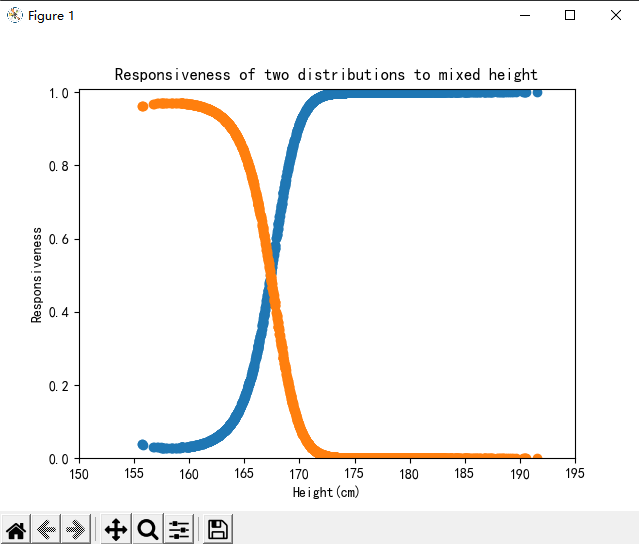


Figure 7：数据对两个分布的响应度

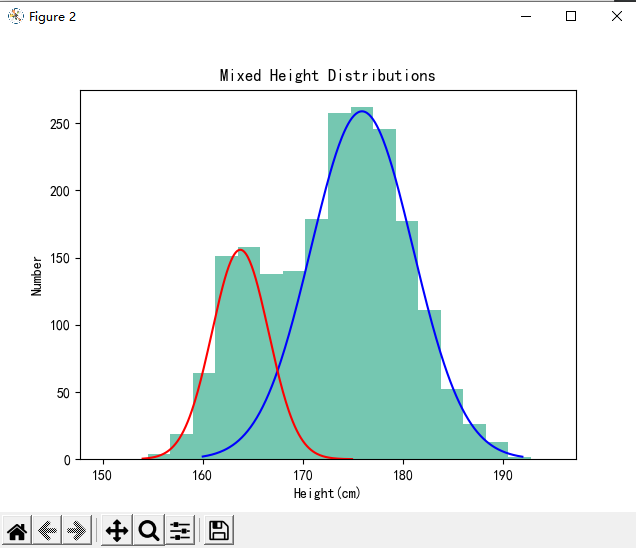


Figure 8：身高柱状图及估计的高斯分布曲线图

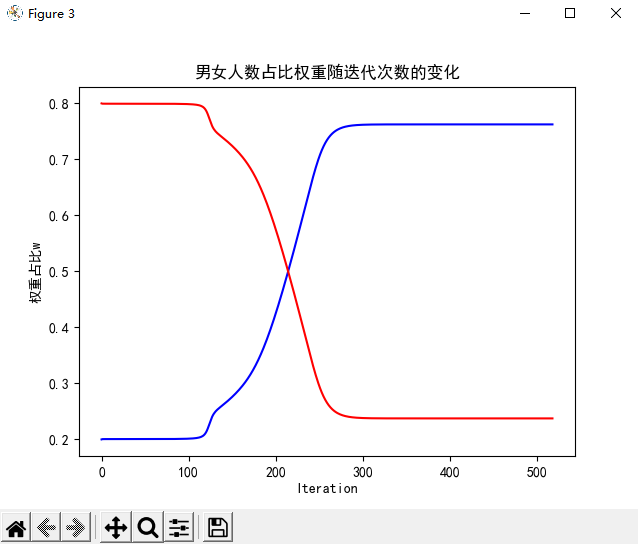


Figure 9：男女人数占比权重变化图

1. **初始化参数**：u1 = 169.9, u2 = 170, sigma1 = sigma2 = 10,

w1 = 0.8, w2 = 0.2

输出结果为：

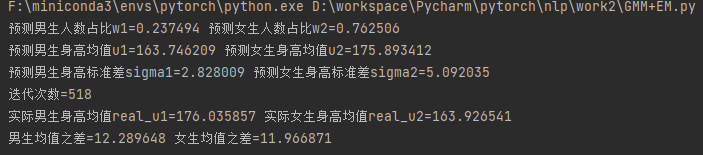


Figure 10：输出结果

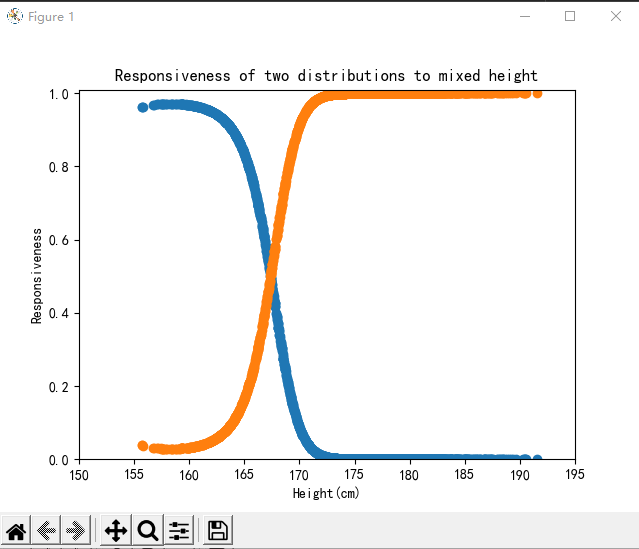


Figure 11：数据对两个分布的响应度

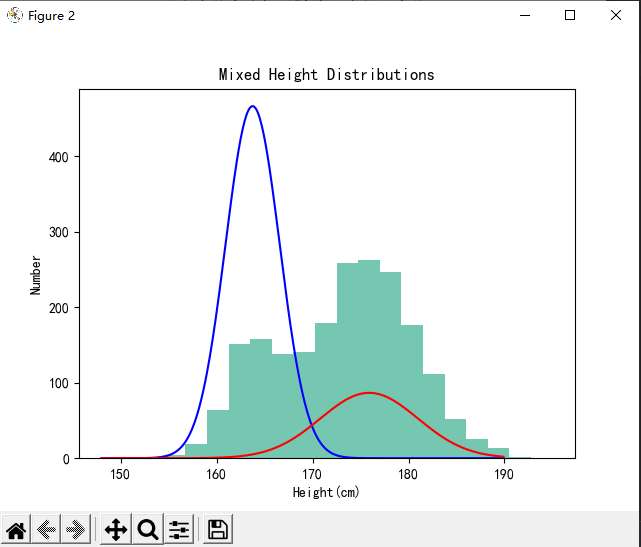


Figure 12：身高柱状图及估计的高斯分布曲线图

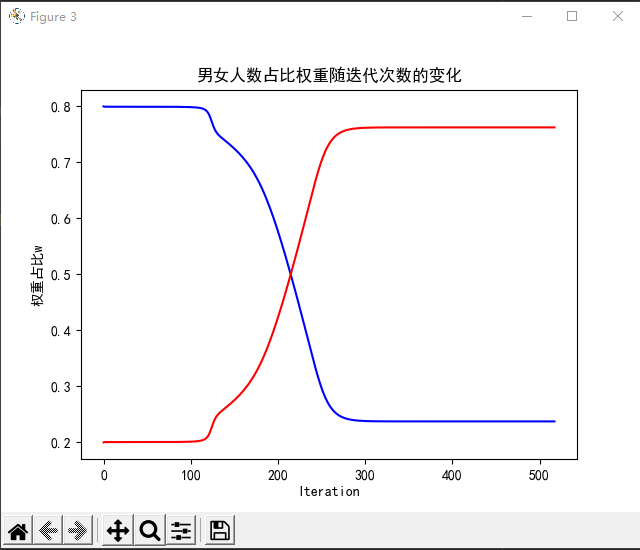


Figure 13：男女人数占比权重变化图

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.