一．摘要

1.1 EM 算法（Expectation-Maximum）

EM 算法中文名为期望最大化算法，是一种用于估计概率分布中未知参数的统计方法，特别有效于处理不完整的残缺数据。EM 算法的基础和收敛有效性等问题在 Dempster、Laird 和 Rubin 三人于 1977 年的文章 *Maximum likelihood from incomplete data via the EM algorithm* 中给出了详尽的阐释。EM 算法在两个步骤之间交替进行：期望步骤（E 步）和最大化步骤（M 步）。

在 E 步中，算法根据当前参数估计计算不完整数据的期望值。接着该步骤使用当前参数估计值来估计缺失数据。

在 M 步中，算法最大化似然函数，似然函数时观察到的数据在给定参数下的概率。该步骤使用 E 步中估计的缺失数据来更新参数估计值。

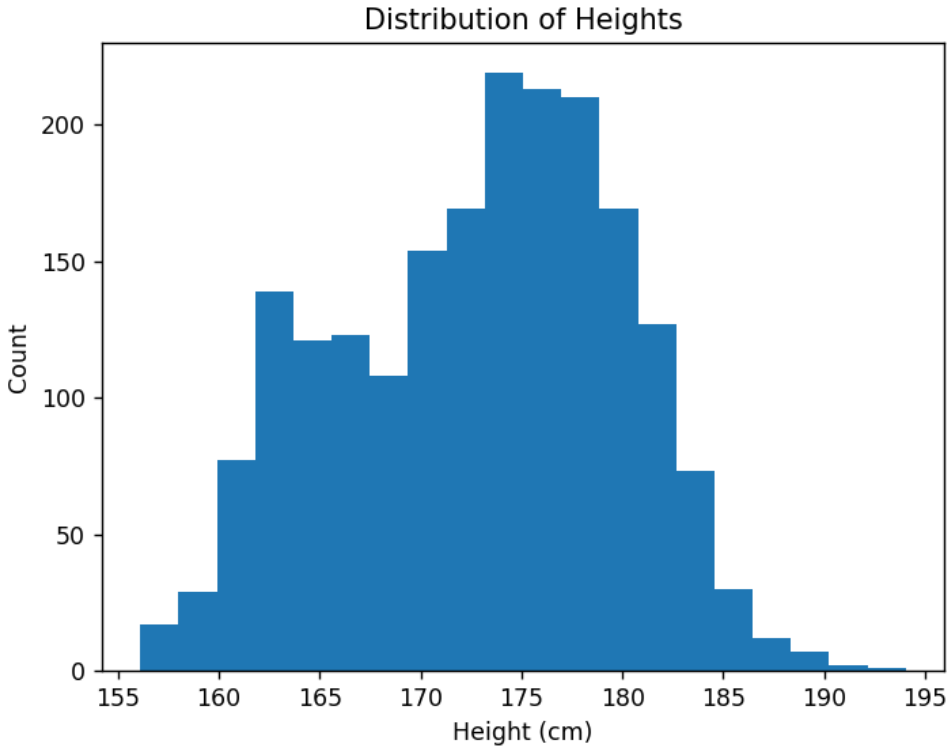
算法重复进行这两个步骤，反复迭代，直到参数的估计值收敛。

1.2 混合高斯模型（GMM）

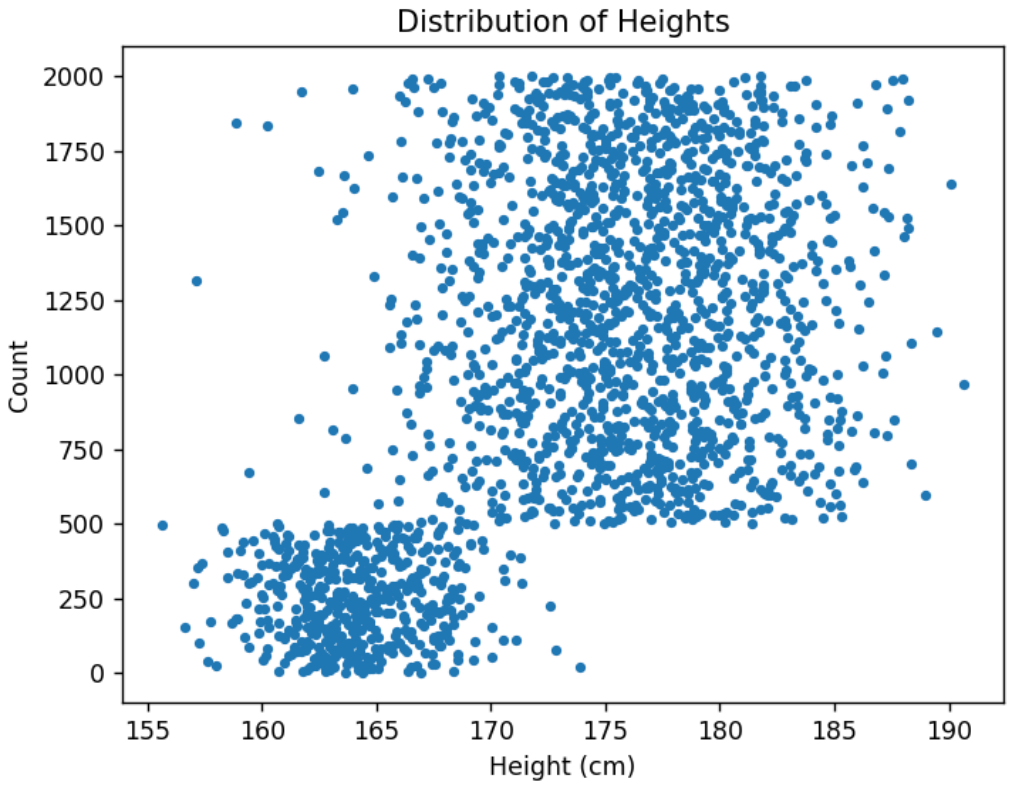
二．实验过程及结果

2.1 原始数据

运行 student_height.py 程序，获得学生身高原始数据，将身高数据绘制为如下所示的直方图：



为了对身高数据的分布有更为清晰的了解，使用 matplotlib 库的 scatter 函数绘制身高数据的散点图如下：



可以明显发现，数据来源可以大致分为两个类别。

2.1 程序设计

使用 Matlab 编程语言，在程序伊始，我们对高斯混合模型中的若干参数进行了值初始化，它们分别是男女生身高各自满足的正态分布参数（$N_1(\mu_1,\sigma_1)$，$N_2(\mu_2,\sigma_2)$）以及男生人数占比 π（即高斯分布权重）。有身高变量的概率密度：

$$p(h|\theta) = \pi N_1(\mu_1,\sigma_1) + (1-\pi)N_2(\mu_2,\sigma_2))$$

接着读入 student_height.py 生成的身高数据 csv 文件。根据如下基于 EM 算法的公式设计循环，迭代次数为 200：

E-step：

$$R_i^{(t)} = \frac{\pi^{(t)}N({\mu_1}^{(t)},{\sigma_1}^{(t)})}{\pi^{(t)}N({\mu_1}^{(t)},{\sigma_1}^{(t)}) + (1-\pi)^{(t)}N({\mu_2}^{(t)},{\sigma_2}^{(t)})}$$

$$T_i^{(t)} = \frac{(1-\pi)^{(t)}N({\mu_2}^{(t)},{\sigma_2}^{(t)})}{\pi^{(t)}N({\mu_1}^{(t)},{\sigma_1}^{(t)}) + (1-\pi)^{(t)}N({\mu_2}^{(t)},{\sigma_2}^{(t)})}$$

M-step:

$$\pi^{(t+1)} = \frac{\sum_{i=1}^{2000} R_i^{(t)}}{N}$$

$$\mu_1{}^{(t+1)} = \frac{\sum_{i=1}^{2000} h_i R_i^{(t)}}{\sum_{i=1}^{2000} R_i^{(t)}}$$

$$\sigma_1{}^{(t+1)} = \frac{\sum_{i=1}^{2000} R_i^{(t)} (h_i - \mu_1{}^{(t+1)})^2}{\sum_{i=1}^{2000} R_i^{(t)}}$$

$$(1-\pi)^{(t+1)} = \frac{\sum_{i=1}^{2000} T_i^{(t)}}{N}$$

$$\mu_2{}^{(t+1)} = \frac{\sum_{i=1}^{2000} h_i T_i^{(t)}}{\sum_{i=1}^{2000} T_i^{(t)}}$$

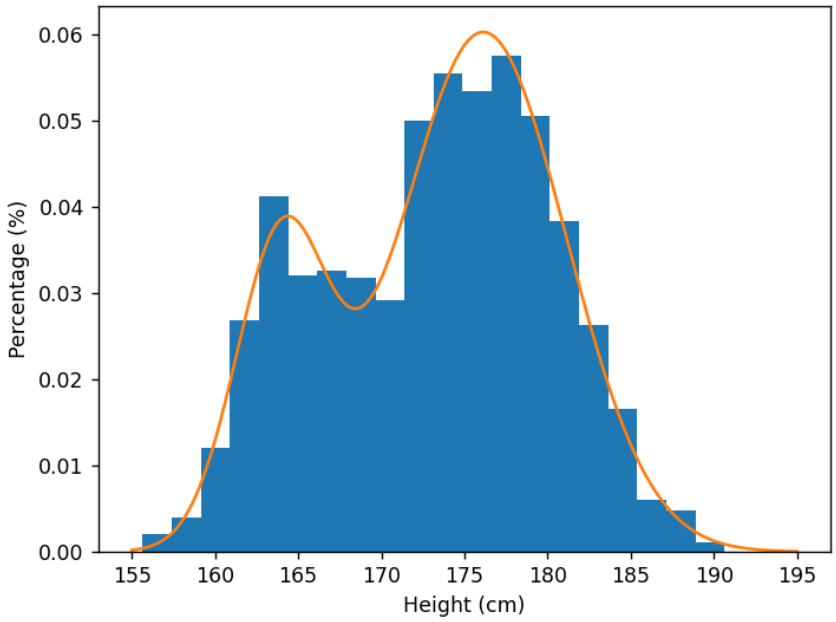
$$\sigma_2{}^{(t+1)} = \frac{\sum_{i=1}^{2000} T_i^{(t)} (h_i - \mu_2{}^{(t+1)})^2}{\sum_{i=1}^{2000} T_i^{(t)}}$$

其中，N为数据量，即为 2000，h_i为第i个身高数据，R和T分别为身高数据来自男生和女生的概率。

最终估计结果如下:

$$\hat{\pi} = 0.7539$$
$$\widehat{\mu_1} = 176.1271$$
$$\widehat{\sigma_1} = 4.9889$$
$$\widehat{1-\pi} = 0.2461$$
$$\widehat{\mu_2} = 163.9943$$
$$\widehat{\sigma_2} = 2.7715$$

将估计的结果和对人数进行归一化后的身高数据绘制在同一张图内如下:



三．结论

本实验基于 EM 算法完成了对混合高斯模型中两个高斯分布的包括均值方差和权重在内的 3 个参数估计，估计的结果较为准确，性能比较理想。但注意到，对女生身高分布的估计结果相较于对男生估计的结果较差，这可能是因为女生的数据量较小导致 EM 算法估计的结果偏差较大。