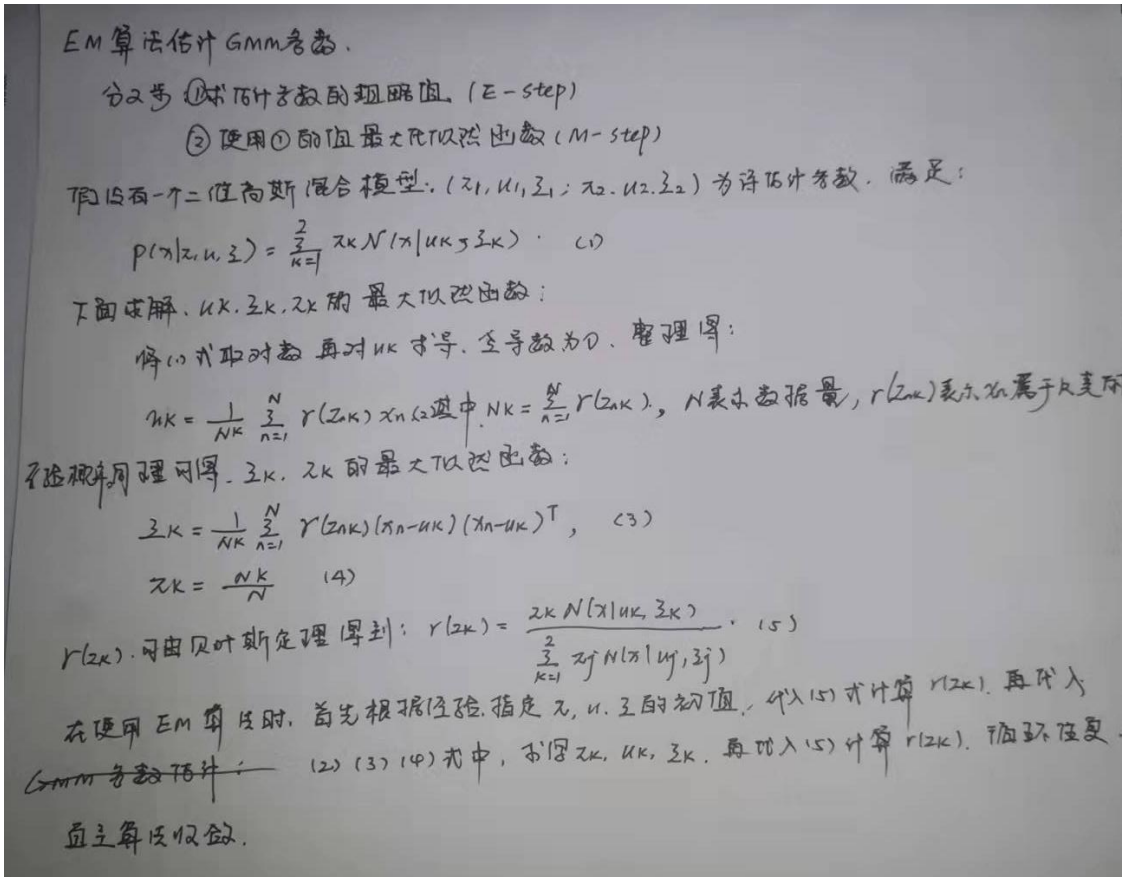
作业 2 基于 EM 算法的高斯混合模型

一、实验内容

给定男女生身高均值和方差，通过 EM 算法和高斯混合模型得到全班男女生的平均身高、方差和男女比例。

二、实验原理

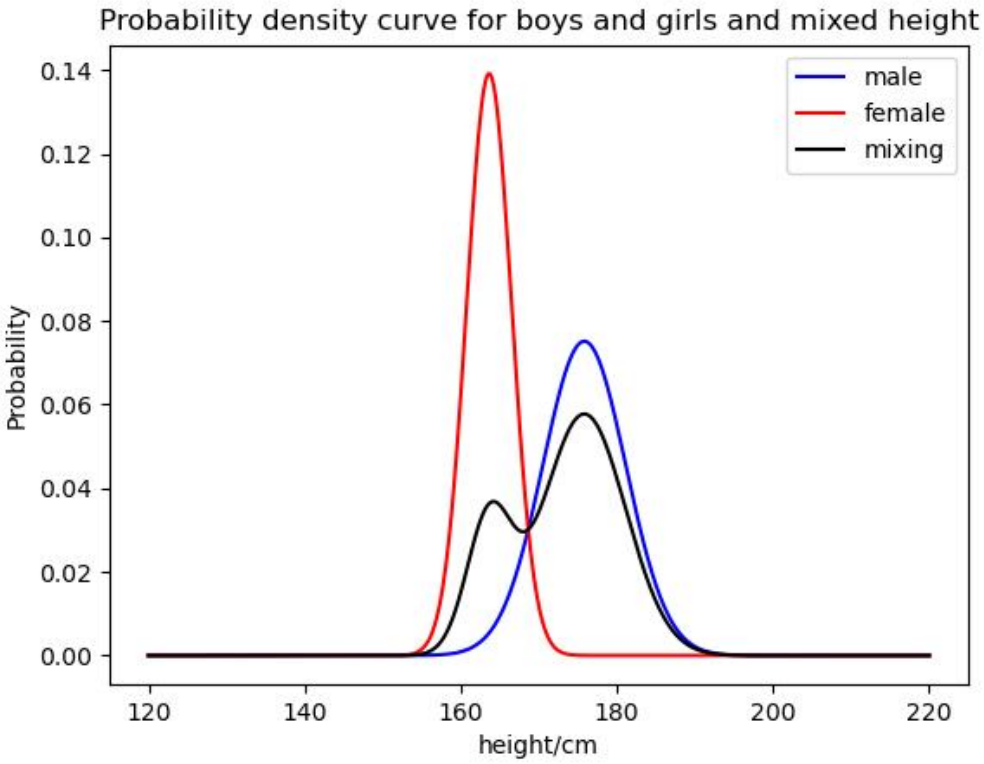


三、实验结果与分析

采用数据：男生平均身高 176，标准差 5，样本数 1500.女生平均身高 164，标准差 3，样本数 500.可以得到男女占比分别为 0.75、0.25.

设定初始条件：男生平均身高 170，标准差 10，比例 0.7.女生平均身高 160，标准差 10，比例 0.3.

经过 EM 算法的 100 次迭代，得到男生平均身高 175.82，标准差 5.309，占比 0.769，女生平均身高 163.63，标准差 2.866，占比 0.231.改变初始条件和增加 EM 算法的迭代轮数，结果无变化。得到男女生身高的混合概率密度曲线如下图：



EM 算法的预测结果与真实数据总体比较接近，但有一定误差。

分析误差原因，可能是样本数量太少、采样数据并不严格满足正态分布。为了验证这一猜想，本人将男女生的样本量增加到 15000、5000 重新进行实验。经过 EM 算法的 100 次迭代，得到男生平均身高 176.00，标准差 5.047，占比 0.752，女生平均身高 163.92，标准差 2.918，占比 0.248.课件，在增大样本量后，EM 算法的预测结果更接近真实值。