# EM算法实现高斯混合聚类分析

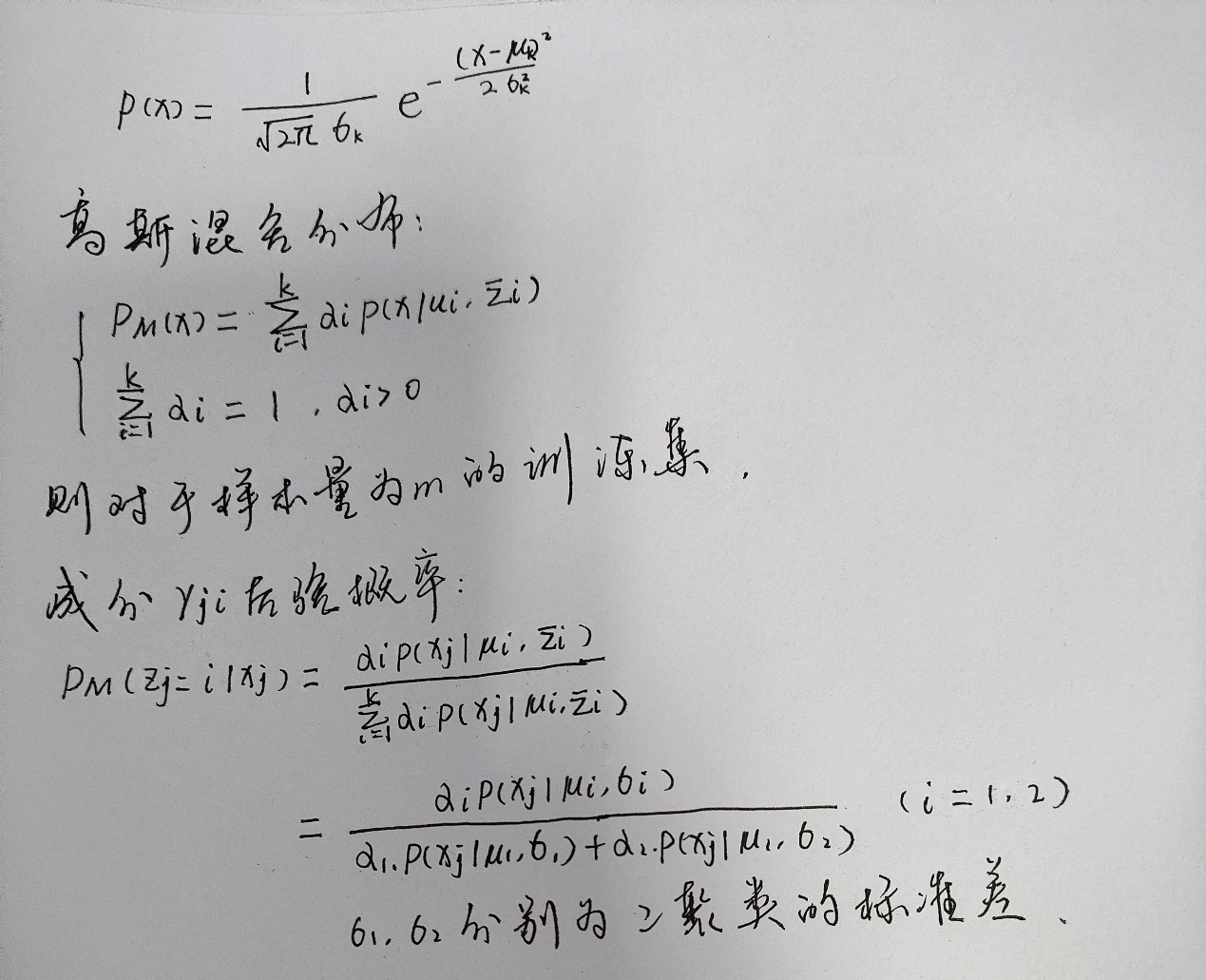
梁智博 ZY2303110

## 一、研究简述

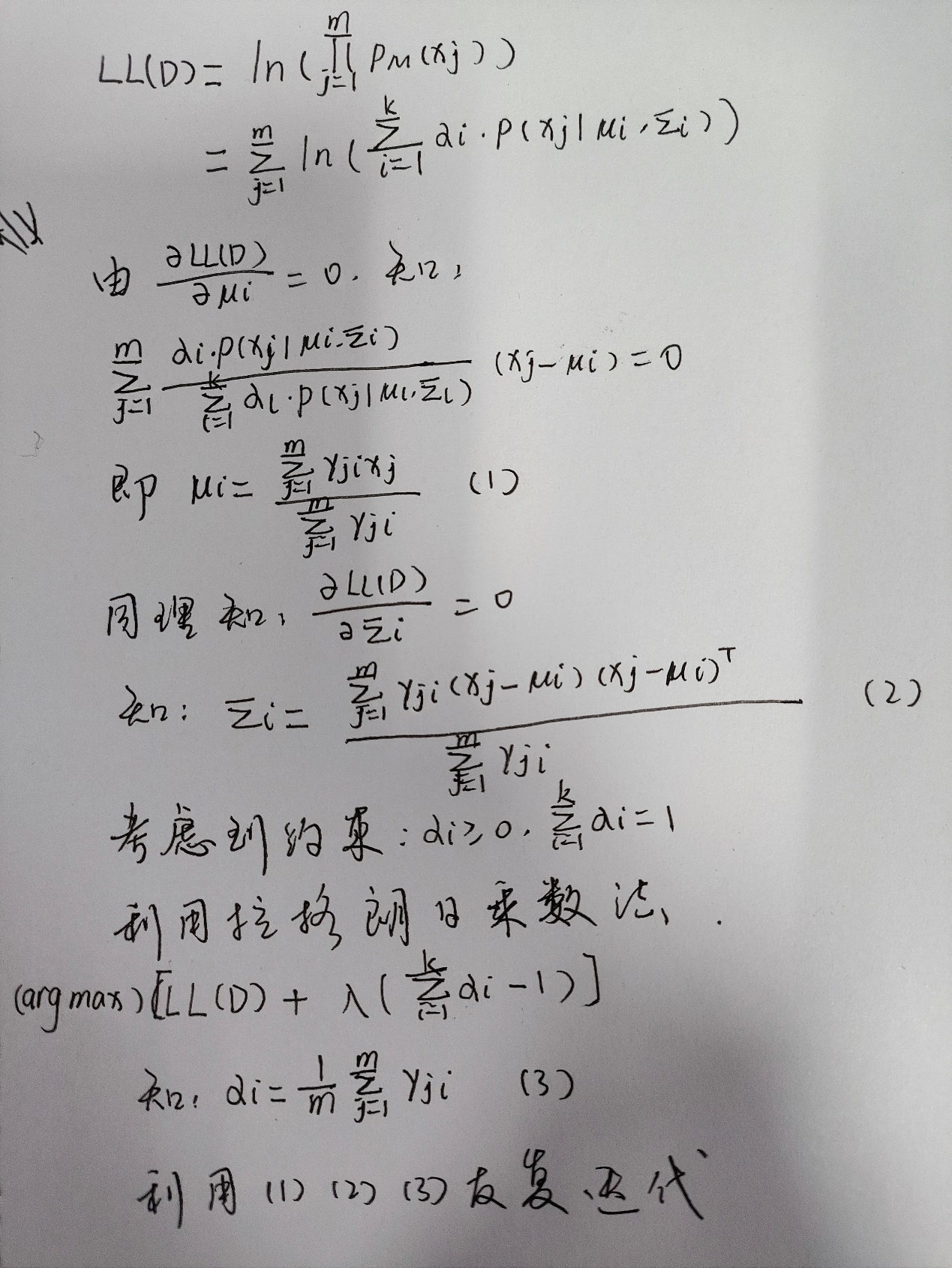
本次研究考虑m个样本，每个样本的属性是1维的身高(本次问题是无监督问题，只考虑聚类不考虑标签)，样本包含了男生和女生的身高数据。在样本足够大的前提下，身高数据是一种典型的符合高斯分布的变量。因此假设模型为高斯混合聚类模型，并利用EM算法对模型参数进行估计。

## 二、方法描述

高斯混合模型采用概率模型来表达聚类原型，利用后验概率确定原型的分类。其生成后验概率的步骤如下：

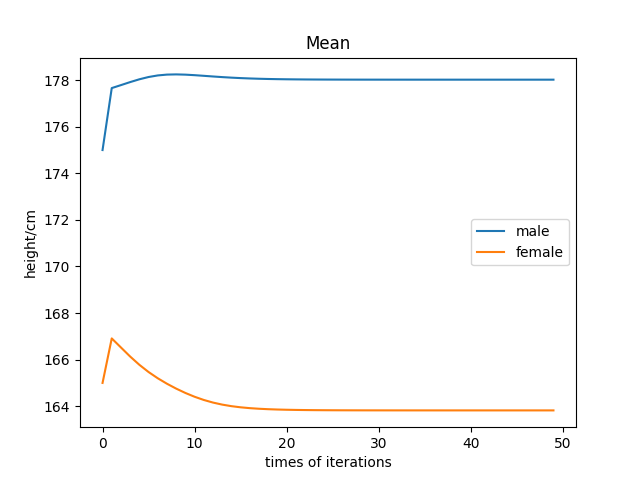


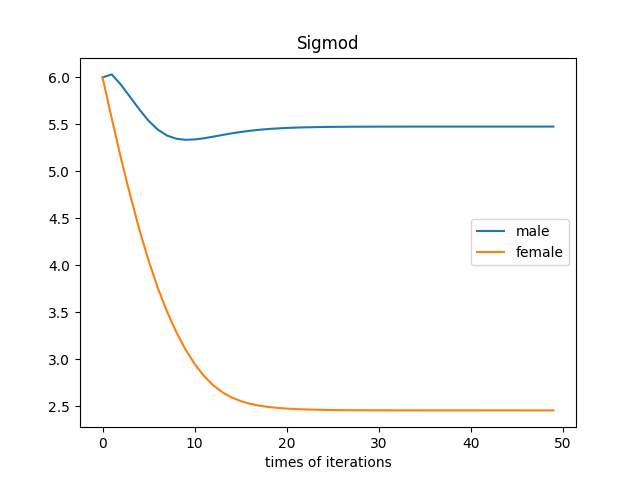
借助生成的后验概率，我们便可以进行簇分类。因此如何确定混合高斯分布的参数，是我们要研究的核心问题。在此我们利用极大似然估计的方法，并利用EM算法进行迭代优化：

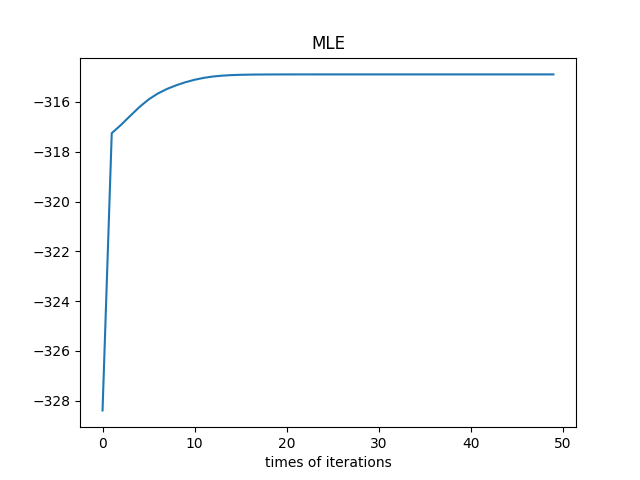


## 三、实践验证

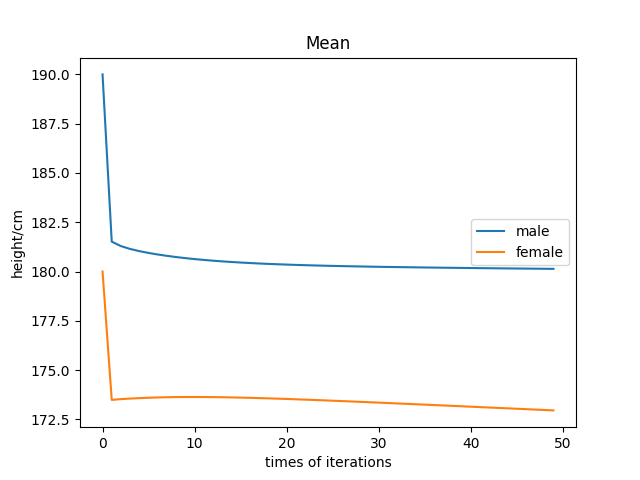
本次研究考虑的高斯混合聚类模型混合成分个数k=2，本质上是借助EM算法把男生和女生两个聚类进行区分。实验时发现，当选取不同的初始值时，产生的结果差异很大，有时候甚至会出现不收敛的情况。因此利用一些先验知识(男生的身高初始均值取为170，女生取为160)选取了合适的初值后，得到的实验结果如下：

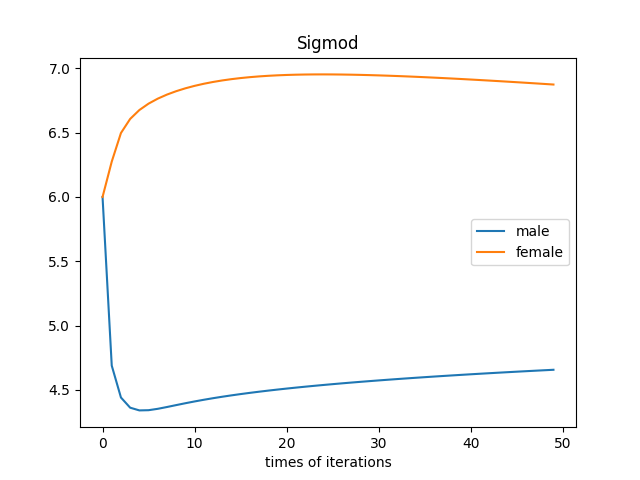


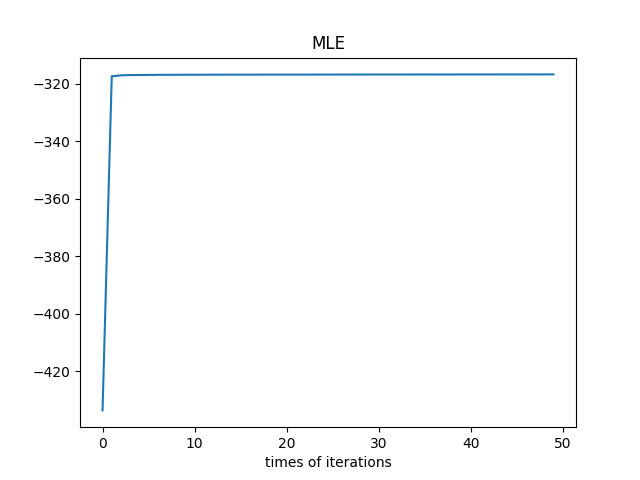




分别选取了男女身高均值（Mean）、标准差（Sigmod）和似然函数（MLE）随着迭代次数的变化来分析学习过程。

 当选取的初始值不合适时，实验结果如下：





实际样本中，男生样本均值为178cm，标准差为5.5；女生样本均值为167cm，标准差为5。可以很明显看出，当初始值合适时，对男生样本的估计精确度很高，但对女生的估计偏差较大。当初始值不合适时，对两种样本的估计效果都不好。

## 四、总结与展望

本次实验利用EM算法迭代来估计高斯混合模型的参数，当初值合适时，对男生样本的估计较准确，对女生估计偏差比较大，原因应该是女生样本较少，频率转换为概率的误差较大。当初值不合适时，由于似然函数的极值不唯一，会收敛到局部最优值，造成结果不准确。

通过以上分析，可以看出该算法的局限性，首先是这种估计方法并不能保证是无偏估计，因此会出现估计出来的期望值有偏差。其次，这种算法很依赖初值的选取（这就需要先验知识的引入，局限性比较大），可以引入正则化来惩罚与合理参数偏差较大的估计值，这样能有效避免收敛到局部最优值的问题。

# 附录：代码

