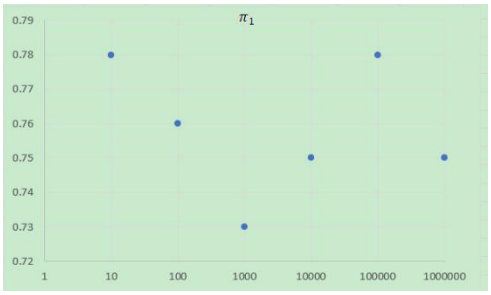
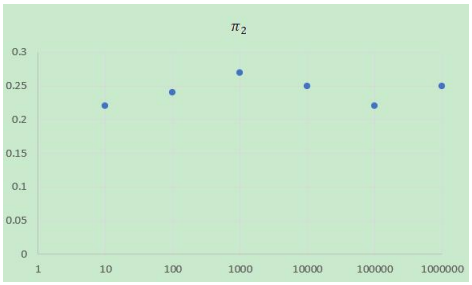
# EM 算法

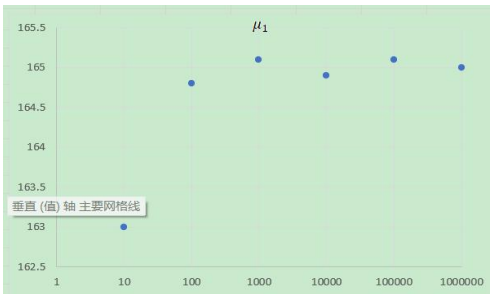
## 一、 数据集大小对算法结果影响

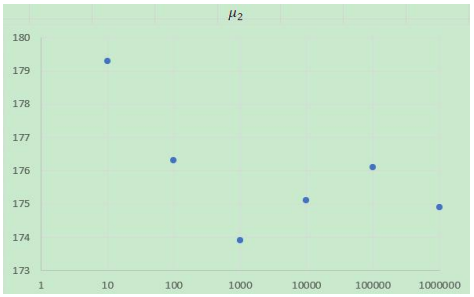
设定参数$\pi_1$= 0.75，$\pi_2$=0.25;$\mu_1$= 165，$\mu_2$= 175;$\sigma_1$= 5，$\sigma_2$= 8 生成数据集;把生成数据集的参数作为初始值，改变生成的数据集的大小进而观测其对算法结果影响

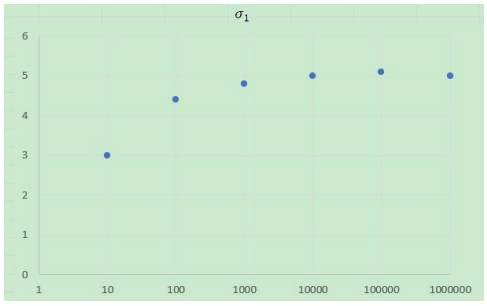
| datasize | $\pi_1$ | $\pi_2$ | $\mu_1$ | $\mu_2$ | $\sigma_1$ | $\sigma_2$ |
|---|---|---|---|---|---|---|
| $10^1$ | 0.78 | 0.22 | 163.0 | 179.3 | 3.0 | 2.4 |
| $10^2$ | 0.76 | 0.24 | 164.8 | 176.3 | 4.4 | 4.3 |
| $10^3$ | 0.73 | 0.27 | 165.1 | 173.9 | 4.8 | 7.6 |
| $10^4$ | 0.75 | 0.25 | 164.9 | 175.1 | 5.0 | 8.0 |
| $10^5$ | 0.78 | 0.22 | 165.1 | 176.1 | 5.1 | 7.6 |
| $10^6$ | 0.75 | 0.25 | 165.0 | 174.9 | 5.0 | 8.1 |

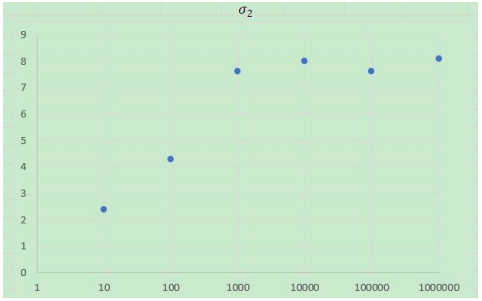


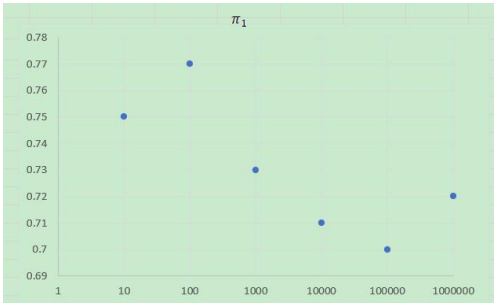



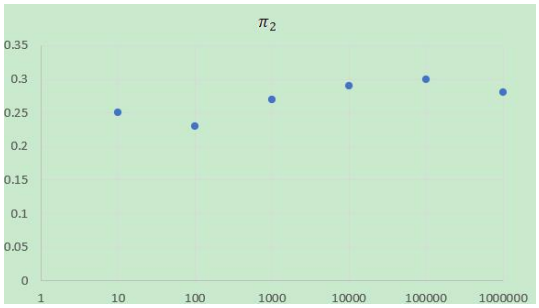

通过以上数据可知，方差的估计对数据量的大小比较“敏感”，当数据量较小时，方差估计偏差较大，这可能是因为方差表征了数据的离散程度，而这在数据量较小时难以充分预估;对于均值$\mu_1$的估计总体要好于$\mu_2$，无论是在数据量较大或者说数据量较小的情况。这可能是因为$\mu_1$所在的 Gauss 分布所占比率$\pi_1$要远高于$\mu_2$所在的 Gauss 分布所占比率$\pi_2$。

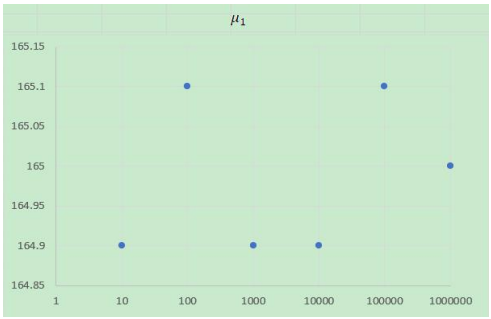
## 二、 迭代次数对算法结果影响

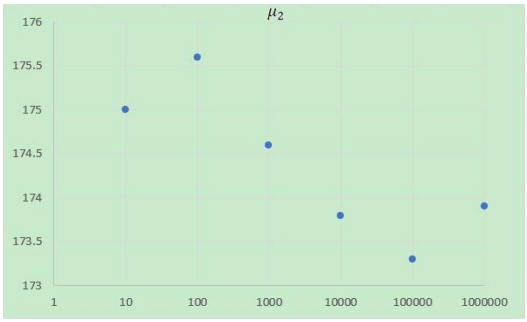
设定参数$\pi_1$= 0.75，$\pi_2$=0.25;$\mu_1$= 165，$\mu_2$= 175;$\sigma_1$= 5，$\sigma_2$= 8 生成数据集;把生成数据集的参数作为初始值，改变迭代次数的大小进而观测其对算法结果影响

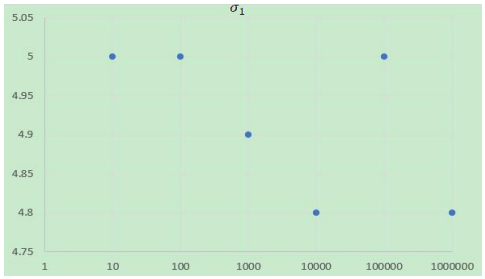
| 迭代次数 | $\pi_1$ | $\pi_2$ | $\mu_1$ | $\mu_2$ | $\sigma_1$ | $\sigma_2$ |
|---|---|---|---|---|---|---|
| $10^1$ | 0.75 | 0.25 | 164.9 | 175.0 | 5.0 | 8.1 |
| $10^2$ | 0.77 | 0.23 | 165.1 | 175.6 | 5.0 | 7.7 |
| $10^3$ | 0.73 | 0.27 | 164.9 | 174.6 | 4.9 | 8.1 |
| $10^4$ | 0.71 | 0.29 | 164.9 | 173.8 | 4.8 | 8.3 |
| $10^5$ | 0.70 | 0.30 | 165.1 | 173.3 | 5.0 | 8.5 |
| $10^6$ | 0.72 | 0.28 | 165.0 | 173.9 | 4.8 | 8.2 |

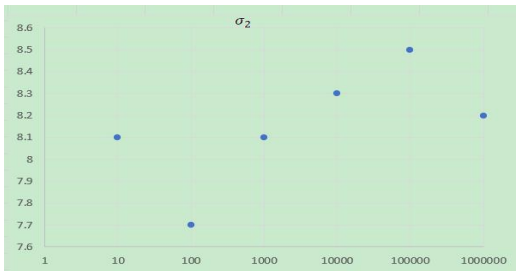



通过以上数据可知，通过增加迭代次数，并不一定能改善参数的估计结果，甚至在迭代次数高的情况下，参数的估计误差反而越大。此外，对于 Gauss 分布 1 的均值$\mu_1$和方差$\sigma_1$的估计，总体要好于 Gauss 分布 2 的均值$\mu_2$和方差$\sigma_2$的估计。这可能是因为 Gauss 分布 1 所占比率$\pi_1$要远高于 Gauss 分布 2 所占比率$\pi_2$。

## 三、 初始值对算法结果影响

设定参数$\pi_1$= 0.75，$\pi_2$=0.25;$\mu_1$= 165, $\mu_2$= 175;$\sigma_1$= 5，$\sigma_2$= 8 生成数据集;改变初始值的大小进而观测其对算法结果影响

| 初始值 | $\pi_1$ | $\pi_2$ | $\mu_1$ | $\mu_2$ | $\sigma_1$ | $\sigma_2$ |
|---|---|---|---|---|---|---|
| $\pi_1$= 0.5<br>$\pi_2$=0.5 | 0.59 | 0.41 | 165.2 | 176.8 | 5.1 | 7.2 |
| $\mu_1$=175<br>$\mu_2$=165 | 0.73 | 0.27 | 175.1 | 165.5 | 4.9 | 7.9 |
| $\sigma_1$=8<br>$\sigma_2$=5 | 0.69 | 0.31 | 164.3 | 174.8 | 7.7 | 5.3 |

如果我们改变初始值，设定$\pi_1$=$\pi_2$=0.5,结果发现$\pi_1$会有增大趋势，$\pi_2$会有减小趋势，但离准确值还是存在一定误差。此外，尝试把两个分布的均值和方差分别对调，会发现其估计值离准确值有误差。所以可见，EM 算法对初始值的依赖度较高。初始值的准确设定与否对结果的估计误差有很大影响。

## 总结

本篇文章主要研究了数据集大小，迭代次数和初始值对算法结果的影响。结果发现：方差的估计对数据量的大小比较“敏感”，当数据量较小时，方差估计偏差较大，这可能是因为方差表征了数据的离散程度。此外，对于 Gauss 分布 1 的参数估计，总体要好于 Gauss 分布 2 的参数估计，无论是在数据集大小还是在迭代次数的测试中，这可能是因为 Gauss 分布 1 所占比率$\pi_1$要远高于 Gauss 分布 2 所占比率$\pi_2$。同时，EM 算法对初始值的依赖度较高。初始值的准确设定与否对结果的估计误差有很大影响。