关于混合高斯分布的讨论

刘发中

November 2022

1 背景知识

1.1 混合模型

混合模型是一个可以用来表示在总体分布中含有 K 个子分布的概率模型，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。

1.2 高斯模型

当样本数据 X 是一维数据时，高斯分布遵从下方概率密度函数:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

理论上的正态分布曲线是一条中间高，两端逐渐下降且完全对称的钟形曲线。

1.3 混合高斯模型

$$X \sim N(\mu_1, \sigma_1^2)$$

$$Y \sim N(\mu_2, \sigma_2^2).$$

$$Z = X + \eta y$$

Z 服从的分布称为混合高斯分布

1.4 中心极限定理

当随机变量序列 $X_1, X_2, ...X_n, ...$ 相互独立同分布，且有期望与方差 $E(X_k) = \mu, D(X_k) = \sigma^2 > 0, k = 1, 2...$，则对任意实数 x，有

$$\lim_{p\to\infty} P\left(\frac{\sum_{k=1}^{n} X_k - n\mu}{\sqrt{n}\sigma} \leq x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

2 任务一

2.1 任务概述

1. 设定不同的参数:μ_1,σ_1^2,μ_2,σ_2^2,p
2. 通过软件生成 N（默认为 5000）个混合高斯分布的随机数
3. 画出其频率分布直方图并讨论不同参数对其分布“峰”的影响。

2.2 随机数生成

初始参数设置如下表:

μ_1	σ_1^2	μ_2	σ_2^2	p
0.113	0.263	3.141	1.020	0.387

通过 python 生成 5000 个随机数，统计随机数分布。

2.3 频率分布直方图

约定组数 $m = 500$，绘制频率分布直方图如下:

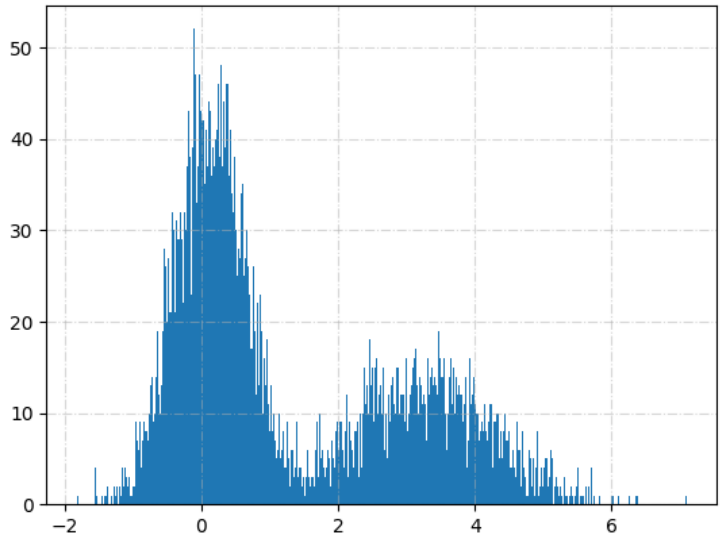

图 1: 频率分布直方图 1

2.4 有关参数讨论

正态分布表达式中有两个参数，即期望（均数）μ 和标准差 σ，σ^2 为方差。正态分布具有两个参数 μ 和数 σ^2 的连续型随机变量的分布，第一参数 μ 是服从正态分布的随机变量的均值，第二个参数 σ^2 是此随机变量的方差，所以正态分布记作 N(μ,σ^2)。

μ 是正态分布的位置参数，描述正态分布的集中趋势位置。概率规律为取与 μ 邻近的值的概率大，而取离 μ 越远的值的概率越小。正态分布以 X=μ 为对称轴，左右完全对称。

σ^2 描述正态分布资料数据分布的离散程度，σ^2 越大，数据分布越分散，σ^2 越小，数据分布越集中。也称为是正态分布的形状参数，σ^2 越大，曲线越扁平，反之，σ^2 越小，曲线越瘦高。

服从混合高斯分布随机数的频率分布直方图呈现“双峰”图像。其中，参数 μ_1 与参数 μ_2 决定两个峰的分布中心位置，由混合高斯分布定义可知，峰 1 的分布中心为 μ_1，峰 2 的分布中心为 $\mu_1 + \mu_2$，参数 σ_1^2 与参数 σ_1^2 决定两个峰的分布宽度（集中程度），参数 p 决定两个峰对应的样本数量大小(峰的高度)。

2.4.1 参数 μ_1 的讨论

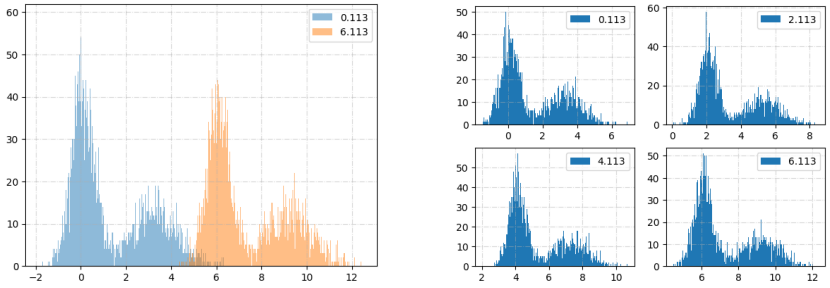

图 2: 参数 μ_1

通过观察多个直方图可知，随着 μ_1 的增加，峰 1 和峰 2 的位置都会逐渐右移。

2.4.2 参数 σ_1^2 的讨论

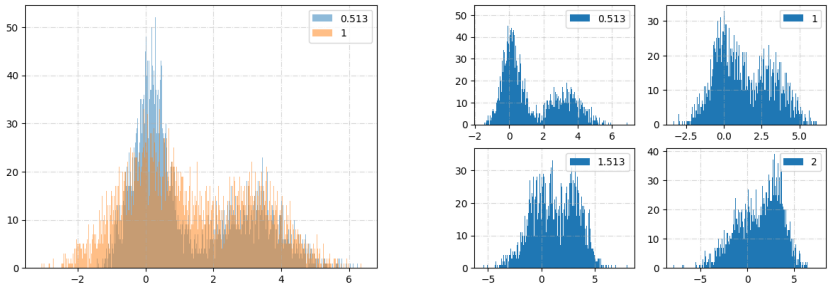

图 3: 参数 σ_1^2

通过观察多个直方图可知，随着 σ_1^2 的增加，峰 1 的离散程度逐渐增加，峰 2 由于受到峰 1 的扩散的影响也会发生形态变化。

2.4.3 参数 μ_2 的讨论

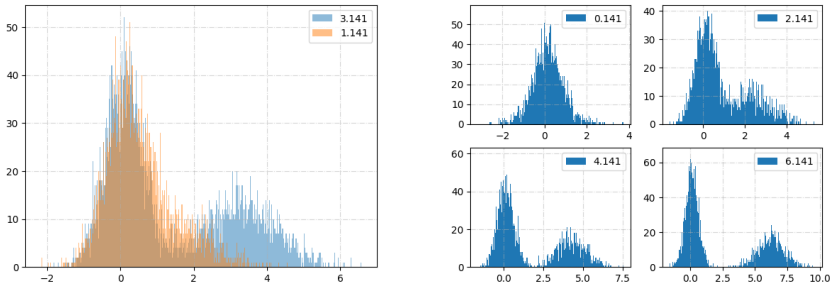

图 4: 参数 μ_2

通过观察多个直方图可知，随着 μ_2 的增加，峰 2 的位置会逐渐右移，而峰 1 基本不受影响。

2.4.4 参数 σ_2^2 的讨论

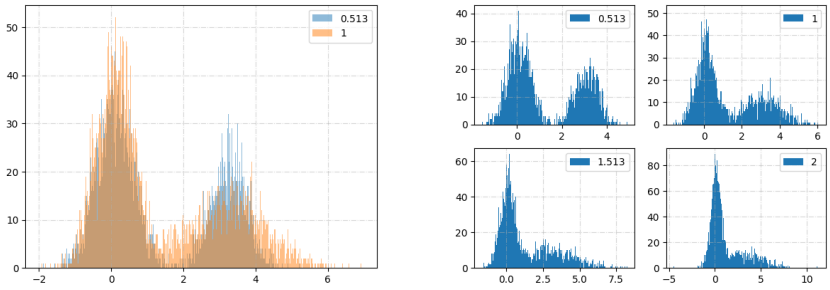

图 5: 参数 σ_2^2

通过观察多个直方图可知，随着 σ_2^2 的增加，峰 2 的离散程度逐渐增加，峰 1 由于受到峰 2 的扩散的影响，会提高峰值。

2.4.5 参数 p 的讨论

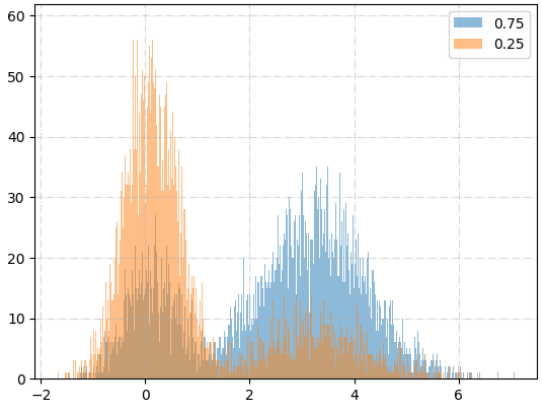

图 6: 频率分布直方图

参数 p 主要影响两个峰对应的样本数量。通过观察直方图可知，当 p 增大时，组成峰 2 的样本数量变多，组成峰 1 的样本数量减小，反映到图上即峰 2 更突出，峰 1 被削弱。

3 任务二

3.1 任务概述

1. 画出 U_i 的频率分布直方图（i=1,2,...,1000）
2. 讨论 n=2,3,4,5,10,20,50,1000,5000 对频率分布直方图“峰”的影响
3. 分析并得出结论

3.2 频率分布直方图

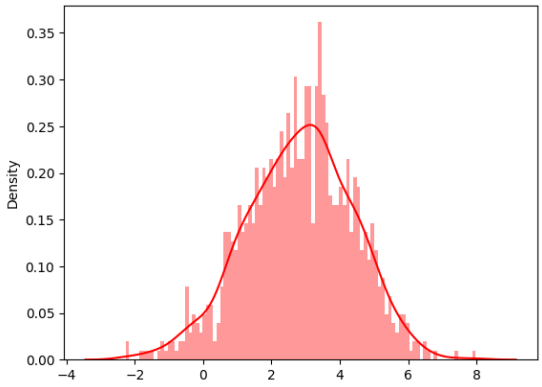

图 7: 频率分布直方图

3.3 有关参数讨论

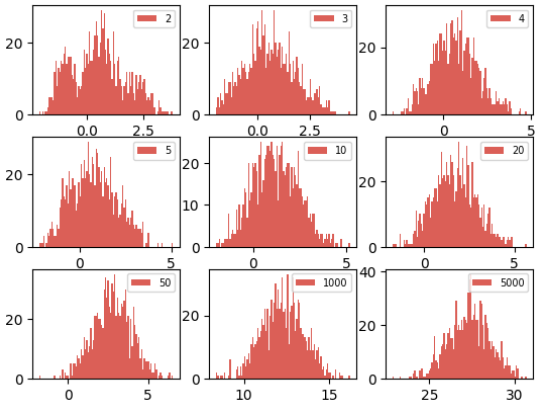

图 8: 频率分布直方图

在生成不同 n 对应的频率分布直方图后，通过 python 对直方图进行函数拟合。

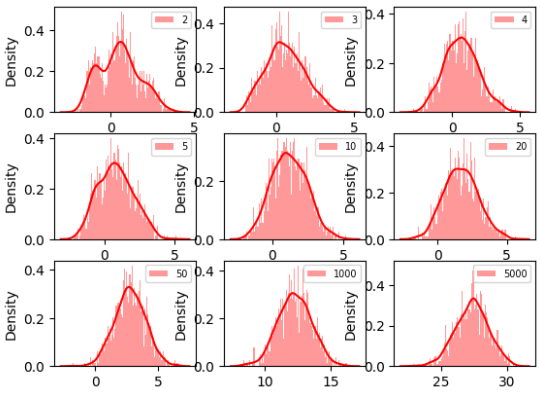

图 9: 频率分布直方图

观察图像可发现，当 n 较小时，直方图形状与“峰”的样态不规则，当 n 较大时，直方图呈现较明显的正态分布样态，为单峰函数。总结规律后，大概在 n>20 后频率分布直方图稳定为正态函数图像。

3.4 结论

由 Linderberg-Levy 中心极限定理可知，随机变量序列 $Z_1, Z_2, ...Z_n, ...$ 相互独立同分布，且有期望与方差 $E(Z_k) = \mu, D(Z_k) = \sigma^2 > 0, k = 1, 2...$，则对于任意实数 z，应有:

$$\lim_{p\to\infty} P\left(\frac{\sum_{j=1}^{n} Z_j - nE(Z)}{\sqrt{nDZ}} \leq z\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt$$

因此在 n 趋向于无穷（即 n 较大时），频率分布直方图应为正态分布图像，而当 n 有限（即 n 较小时），不满足中心极限定理的条件，因此不呈现明显的正态分布图像，而是不规则的图像。

通过计算机生成随机数，再次验证了中心极限定理的正确性。当所取的 n 越大时，频率分布直方图应越趋近于正态分布图像。

4 结语与感想

这次大作业的完成让我收获良多，通过 python 编程来解决实际数学问题的经历很宝贵，我不仅对正态分布、混合高斯分布、中心极限定理等数学知识有了更深刻的理解，也在过程中锻炼了信息检索、编程、论文写作等能力。

在完成论文的过程中由于能力有限，遇到了许多困难（悲）。首先是由于对混合高斯分布的公式理解不足，导致花费大量时间搜索如何用 python 实现错误的算法，在迷惑于“多维混合高斯分布”“机器学习 DMM”等等问题后，才发现混合高斯分布的实现只需最基础的正态分布函数。后来又遭遇到如参数与参数名称不对应导致图像混乱，缺乏绘制直方图经验导致画出来的图像难以辨识规律，在完成论文写作后才发现最基本的画图函数写错了结果还要全部重来等等问题，不胜枚举。经历重重艰难险阻，最终得出还可以的结论，不胜感激。

感谢熊德文老师以及各位助教的辛勤付出。