蒙特卡洛随机模拟方法在估计高斯混合模型的复杂统计量的分布中的应用

SY2106346 郑耀彦

2021.10

摘要

蒙特卡洛随机方法是一种关于概率计算的经典实验方法，本文针对高斯混合模型的概率密度和其几种统计量的概率密度，使用了蒙特卡洛随机模拟方法，对这些统计量进行了随机抽样，并通过测定模拟生成的经验分布和准确分布之间的均方误差和 KL 散度两项指标，研究了模拟次数对于拟合精度的影响。本文使用 mathematica 软件计算统计量的精确概率密度函数，使用 python 进行蒙特卡洛随机模拟计算经验概率密度并计算均方误差和交叉熵等指标，同时借助 matplotlib 进行了概率密度的可视化。源代码位于 *https://github.com/BUAAHugeGun/Monte-Carlo*

关键词: 概率分布，蒙特卡洛，统计量，高斯混合模型

Abstract

Monte-carlo stochastic method is a classical experiment about probability calculation method, this paper, according to the probability density of gaussian mixture model(GMM) and its several statistic probability density, using the monte carlo stochastic simulation method, the random sampling of these statistics, and by measuring the experience and accurate distribution simulation to generate the mean square error (MSE) between the two indicators and cross entropy, The influence of simulation times on the fitting accuracy is studied. In this paper, mathematica is used to calculate the exact probability density function of statistics, Monte Carlo random simulation is used to calculate the empirical probability density and calculate the mean square error, cross entropy and other indicators, and matplotlib is used to visualize the probability density. Source code at *https://github.com/BUAAHugeGun/Monte-Carlo*

Keywords: Probability Distribution, Monte-Carlo, Statistics, Gaussian Mixture Model

目录

1 背景介绍

本文主要研究蒙特卡洛随机模拟方法在估计复杂概率分布的统计量概率密度中的作用。下面简单介绍本文使用的高斯混合模型、实验指标等。

1.1 高斯混合模型

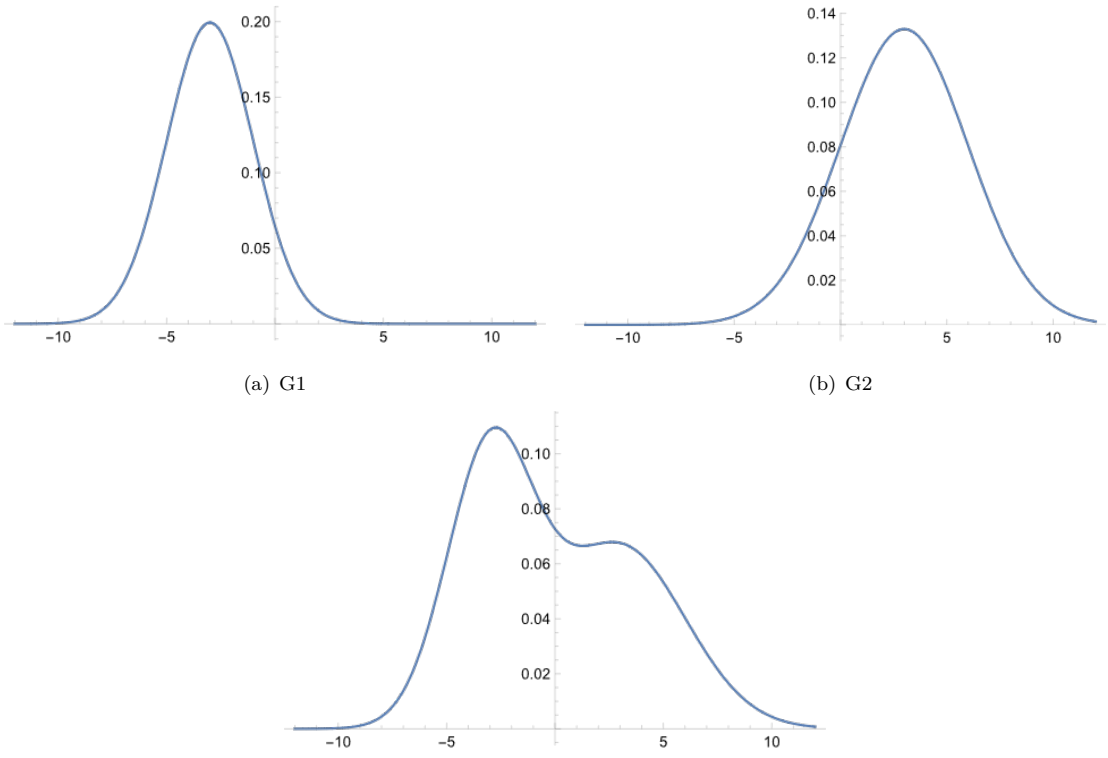
高斯混合模型 (GMM, Gaussian Mixture Model) 由多个高斯分布加权混合得到。高斯混合模型在图像处理、大数据、图像编码、图像生成等领域都有比较多的应用，而且通常高斯混合模型比较复杂，难以获得其统计量准确的分布函数和概率密度函数，因此本文选择高斯混合模型作为蒙特卡洛随机模拟方法的实验对象。本文使用一元高斯模型构成一元高斯混合模型，其概率密度函数可以表示为：

$$P(x|\theta) = \sum_{i=1}^{k} \alpha_i p(x|\theta_i)$$

$$p(x|\theta_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} exp(-\frac{(x-\mu_i)^2}{2\sigma_i^2})$$

$$\theta = (\theta_1, \theta_2, ..., \theta_k), \quad \theta_i = (\mu_i, \sigma_i)$$

其中 p 表示一元高斯概率密度函数，一共有 k 个，α_i 表示第 i 个高斯分布在混合模型中的权重。图 1所示高斯混合模型就是由两个均值和方差分别为 $\theta_1 = (-3, 2)$、$\theta_2 = (3, 3)$ 的两个高斯模型经过权重 $\alpha = (0.5, 0.5)$ 混合得到。



(a) G1

(b) G2

(c) GMM

图 1: 高斯混合模型

1.2 蒙特卡洛随机模拟

蒙特卡洛随机方法，即统计模拟方法，是一类以概率统计理论为指导的数值计算方法。本质上是用部分估计整体，采样越多，则越近似最优解。

比如想要估计一个图形的面积，蒙特卡洛方法就是随机在包含这个图形的全集内随机撒 n 个点，最后统计有 m 个点落在图形内部，只要我们知道全集的面积大小 S，就能通过在图形内部的点的比例近似求出图形的面积 s，而随机撒点的点数趋于无穷大时，其估算误差也会趋于 0，其计算公式为:

$$s = \frac{m}{n} S$$

在本文的实验中，我们针对了几种高斯混合模型的简单样本的统计量，使用了蒙特卡洛随机模拟方法对统计量的概率分布做了估计，得出经验分布。并同时使用均方误差和 KL 散度两种指标对随机次数与经验分布的误差的关系做了分析。

2 实验过程

本文使用 mathematica 软件对统计量的精确分布进行计算并化简。蒙特卡洛随机模拟和可视化以及实验结果指标的测定使用 python 完成，下面介绍各个部分的结构。

2.1 高斯混合模型的构建

基本概率模型 base model 和高斯混合模型在 model.py 中实现，基本概率模型必须有 cdf、ppf 和 plot 函数，表示概率模型的概率密度函数、累计分布函数和可视化函数。

高斯混合模型支持 k 个高斯模型的混合，需要给定 k 和各个高斯模型的 (μ, σ) 以及每个模型所占的权重，它们的权重和必须为 1.0。

2.2 统计量的随机模拟

统计量的随机模拟和测试等环节在 main.py 中实现，本文主要实现了两种统计量的蒙特卡洛模拟。对于每个统计量，我们实现了随机数据的存取和不同随机次数的模拟。而对于统计量 T 在 $t = x$ 的经验概率密度，我们使用频率替换估计法，用了二分法求出了 $[x - 0.5, x + 0.5]$ 内的数据个数占比 (即频率) 作为概率。

2.3 统计量精确分布的计算

我们使用 mathematica 软件构造了本文使用的高斯混合模型，并通过 TransformedDistribution 求出了两种统计量的精确分布并化简。

得到了精确的概率密度函数表达式后，我们在 python 中实现了两种统计量的概率密度函数用于之后的指标测试和可视化。

3 实验结果

3.1 指标介绍

3.1.1 均方误差

本文使用了均方误差 (MSE, Mean Square Error) 衡量两个概率分布之间的相似性，当且仅当两个概率密度相等时，它们的均方误差为 0。其计算公式为：

$$MSE = \frac{1}{r-l}\int_{l}^{r}[f(t)-\hat{f}(t)]^2dt$$

其中 f 和 f' 分别为统计量服从的真实分布的概率密度和蒙特卡洛随机模拟得到的经验分布的概率密度，l, r 为蒙特卡洛模拟得到的统计量样本的最小值和最大值。在本文的实验中，f 是一个连续可导的函数，而 $\hat{f}$ 是不连续的。所以我们使用频率分布直方图来表示概率密度，将均方误差公式改写为：

$$MSE = \frac{1}{m}\sum_{i=1}^{m}[\hat{f}_i - \int_{l_i}^{r_i} f(t)dt]^2$$

其中 m 为频率分布直方图的组数，$\hat{f}_i$ 为直方图第 i 组的频率，l_i, r_i 表示直方图第 i 组横坐标的左右边界。

3.1.2 KL 散度

KL 散度 (Kullback-Leibler divergence) 可以从相对熵的角度衡量两个概率分布之间的相似性或者区别。KL 散度等于交叉熵减去熵，而分布的熵很容易计算：$D_{KL}(f||\hat{f}) = H(f,\hat{f}) - S_f$

$$S_f = \int_{-\infty}^{\infty} f(t)\log_2 f(t)$$

$$S_{\hat{f}} = \sum_{i=1}^{n}\hat{f}(t_i)\log_2\hat{f}(t_i)$$

其中 n 为统计量 t 的样本容量,H 是相对熵。一般来说 KL 散度计算 $D_{KL}(f||\hat{f}$，因为 f 是确定的，KL 散度的值一定会大于 S_f，是一个定值，比较起来比较方便。但是相对熵 $H(f,\hat{f})$ 不好算，因为 $\hat{f}$ 不连续。但是 f 是连续的，所以我们计算 f 相对 $\hat{f}$ 的 KL 散度：

$$D_{KL}(\hat{f}||f) = \sum_{i=1}^{n}\hat{f}(t_i)\log_2 f(t_i) - S_{\hat{f}} = \sum_{i=1}^{n}\hat{f}(t_i)\log_2\frac{f(t_i)}{\hat{f}(t_i)}$$

3.2 实验设置

对于高斯混合模型，本文使用两个参数分别为 $(\mu=-4,\sigma=2)$ 和 $(\mu=4,\sigma=2)$ 的高斯模型进行 1:1 的混合，其概率密度和蒙特卡洛模拟的频率分布直方图如图 2所示：

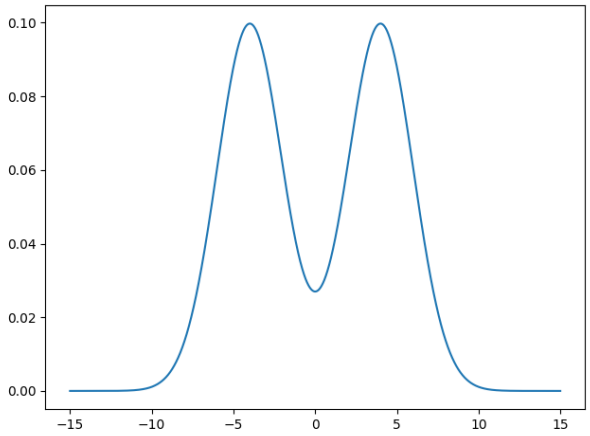


图 2: 本文使用的高斯混合模型 GMM

对于统计量，我们选用 $T_1 = x_1 + x_2$ 和 $T_2 = max\{x_1, x_2\}$ 作为我们研究的统计量，其中 x_1, x_2 都服从上述的高斯混合模型。

对于蒙特卡洛随机模拟的随机次数，我们分别进行了 $n = 100, 1000, 5000, 10000, 50000, 100000$ 的实验，通过测定其指标研究了模拟次数和拟合精度的关系。

3.3 实验结果

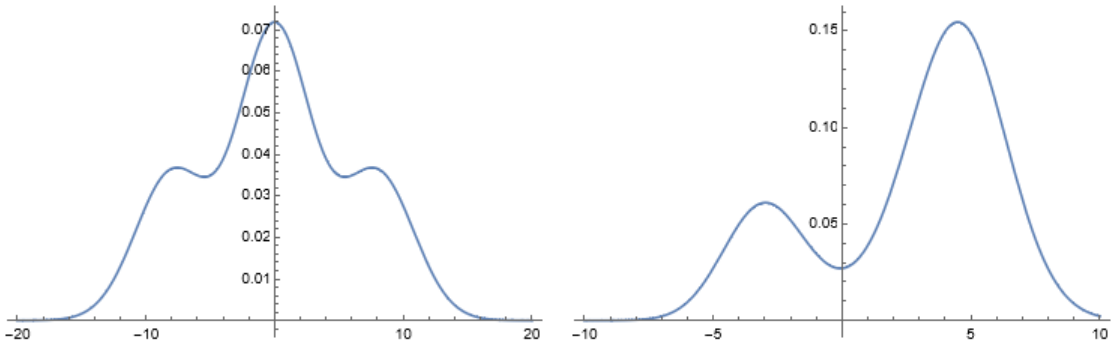
3.3.1 统计量服从的精确分布

我们使用 mathematica 计算出了 $T1$、$T2$ 真实概率密度函数的表达式：

$$f_{T_1} = \frac{1}{8\sqrt{\pi}} e^{-4-\frac{x^2}{16}} [e^4 + cosh(x)]$$

$$f_{T_2} = \frac{1}{8\sqrt{2\pi}} e^{-\frac{1}{8}(4+x)^2} [1 + e^{2x}][2 + erf(\frac{-4+x}{2\sqrt{2}}) + erf(\frac{4+x}{2\sqrt{2}})]$$

图 3为他们的概率密度函数图像。

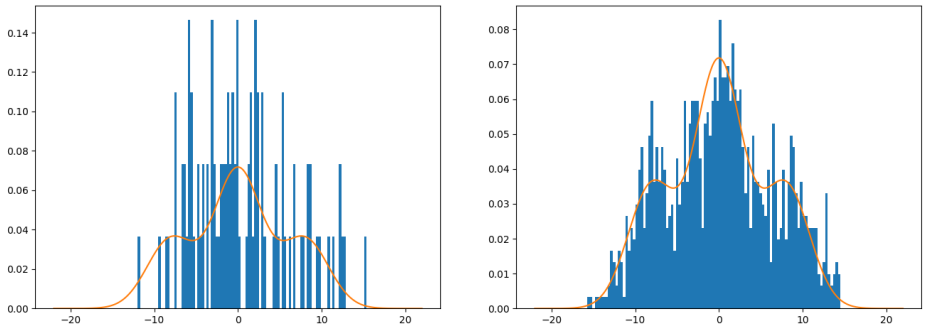


(a) T_1 的概率密度函数　　(b) T_2 的概率密度函数

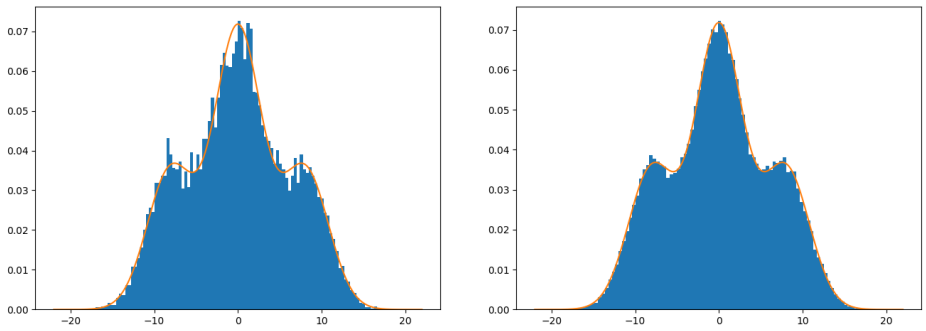
图 3: 统计量的概率密度函数

3.3.2 蒙特卡洛随机模拟

如图 4和图 5所示，蒙特卡洛模拟对统计量的分布有比较好的估计效果。

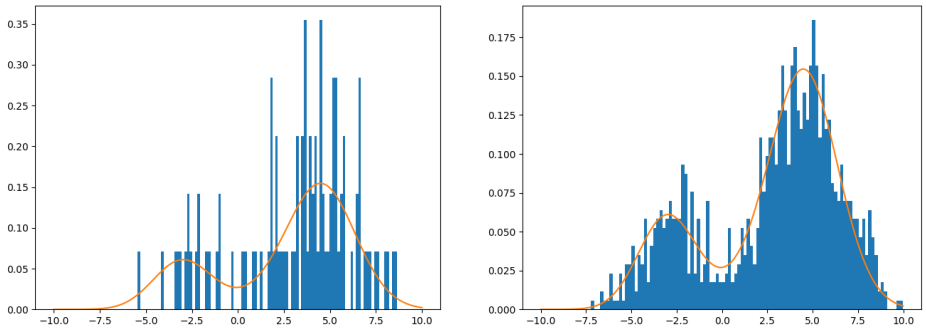


(a) n=100 (b) n=1000

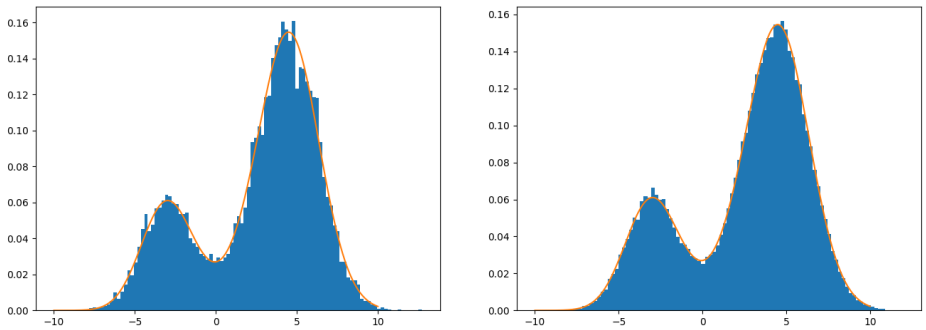


(c) n=10000 (d) n=100000

图 4: 蒙特卡洛随机模拟估计 T_1 的频率分布直方图



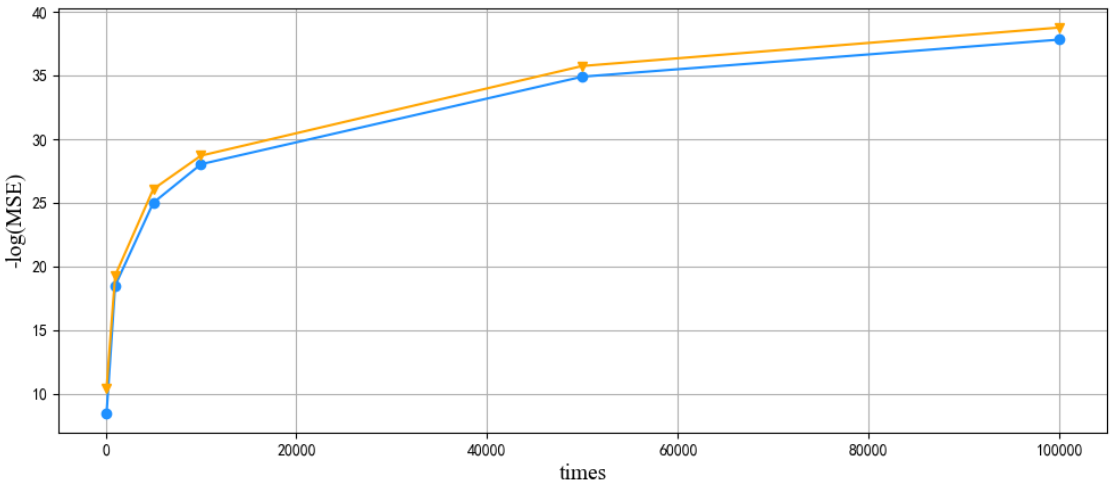
(a) n=100 (b) n=1000



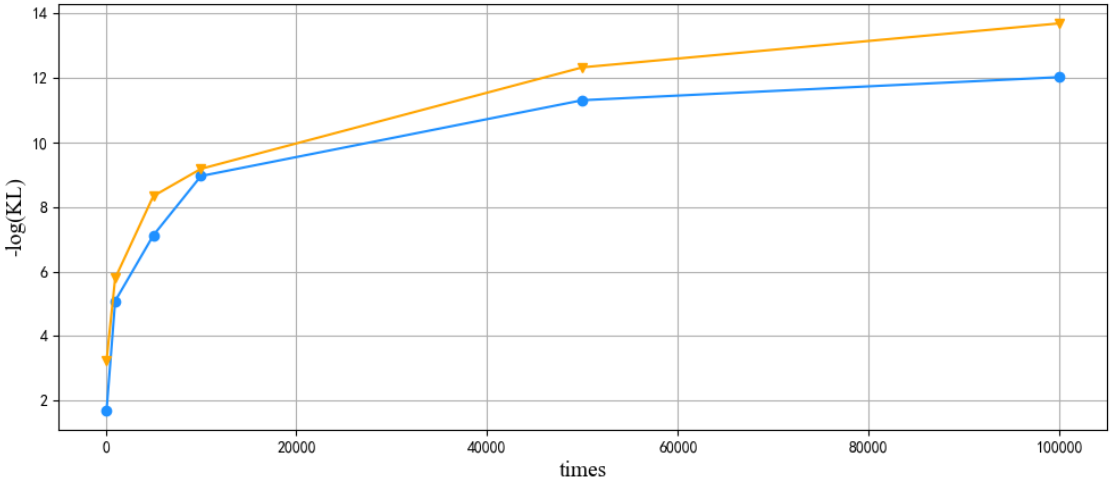
(c) n=10000 (d) n=100000

图 5: 蒙特卡洛随机模拟估计 T_2 的频率分布直方图

如图 6所示，随着随机次数的增加，均方误差和 KL 散度的指标都有明显的提升并且提升速度会越来越慢。由于原指标值太小，本文对其取了对数和相反数后进行了可视化。



(a) MSE



(b) KL 散度

图 6: 指标随机模拟次数的变化