Report6

PB20020480

王润泽

1 Question

对两个函数线型(Gauss 分布和类 Lorentz 型分布)，设其一为 p(x)，另一为 F(x)，其中常数 a≠b≠1，用舍选法对 p(x)抽样。将计算得到的归一化频数分布直方图 与理论曲线 p(x)进行比较，讨论差异，讨论抽样效率。

$$Gaussian: \exp(-ax^2)$$
$$Lorantzianlike: \frac{1}{1+bx^4}$$

2.1取合适的参数

取Gauss分布为

$$p(x) = G(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

取类Lorentz分布为

$$F(x) = L(x) = \frac{1}{\sqrt{2\pi}}\frac{1.01}{1+0.25x^4}$$

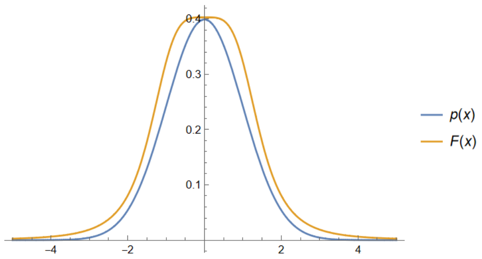

图1：比较函数与待抽样函数

可以找到$F(x) > p(x)$始终成立，取Gauss分布为 $p(x)$，而Lorentz分布为 $F(x)$

2.2 抽样比较函数F(x)

为了抽样关于 $F(x)$的分布函数，取 ξ为 $(-1,1)$的均匀分布函数

$$\xi = \frac{\int_{-\infty}^{x} F(t)dt}{\int_{-\infty}^{+\infty} F(t)dt} = \int_{-\infty}^{x} f(t)dt$$
$$=0.31831\left(0.5\tan^{-1}(x+1)\right) - 0.5\tan^{-1}(1-x) + 0.25\log\left(x^2+2x+2\right) - 0.25\log\left(-x^2+2x-2\right)\Big)$$

$$\xi = \frac{\int_{-\infty}^{x} F(t)dt}{\int_{-\infty}^{+\infty} F(t)dt} = \int_{-\infty}^{x} f(t)dt$$
$$=0.31831\left(0.5\tan^{-1}(x+1)\right) - 0.5\,\mathrm{ta}$$

几乎无法求解反函数，故采用乘积舍选法进行抽样 $f(x)$ 。以下有形式：

$$F(x) = q(x)h(x)$$
$$q(x) = \frac{1}{\pi(1+x^2)}$$
$$h(x) = f(x)/q(x)$$

1. 产生分布 $q(x)$的随机抽样 ξ_x，取 ξ_1为 $(0,1)$的均匀分布函数

$$\xi_1 = \int_{-\infty}^{\xi_x} q(x)dx = \frac{\arctan x}{\pi} + \frac{1}{2}$$
$$\xi_x = \tan(\pi\xi - \pi/2) = -\cot(\pi\xi_1) \iff \tan(\pi\xi_1)\ \ \xi_1 \in (-1/2, 1/2)\text{均匀分布}$$

2. 另外再产生一个[0, 1]区间中均匀分布的随机抽样值 ξ_2 ，判断条件 $M\xi_2 \le h(\xi_x)$ 是否成立。（其中M取 $h(x)$的一个上界2.05）
3. 是，则取 $x_1 = \xi_x$；否，则舍去

2.3 抽样原分布p(x)

1. 根据已抽样得到的分布 x_1，再取 $y_1 = \xi_3 F(x_1)$，ξ_3为 $(0,1)$的均匀分布函数
2. 那么根据舍选法：若 $y_1 < p(x_1)$, 则取 $x = x_1$；否，则舍去。

这样就得到一个关于标准正态高斯分布的抽样

3. Experiment

3.0 区间截断

考虑到对正态分布的抽样是在实数域全集上，而计算机浮点表示数的范围是有限的，所以，在实际抽样时，采取了将5σ之外的点全部替换为 5σ处的值的策略。这样做的依据是正态分布在 5σ之外点的概率非常低，几乎不可能，从之后实际抽样结果可以看出，确实如此，所以次截断是合理的。

```
    # from -infinity to +infinity, actually the infinity is 1.63e+16
    xi_1 = np.tan(xi_1)

    #为了防止范围过大，且|x|>5之后概率密度非常小，故将超过这部分的值进行替代
    xi_1 = np.clip(xi_1,-5,5)
```

3.1 抽样结果

实验中得到标准高斯分布的统计图如下所示。

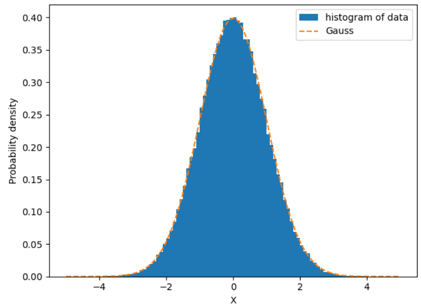

图2：抽样直方图与正态分布比较

可见归一化直方统计结果与正态分布密度函数拟合得十分的好。

3.2 抽样效率

再统计一下两次抽样的抽样效率，在程序中输出如下：

```
First rate of sampling:  0.62618
Second rate of sampling:  0.7791369893640806
```

两次抽样后，总抽样效率只有

$$r = 0.62618 \times 0.77914 \approx 48.79\%$$

可见效率并不是太好,但也有一般的抽样效率。

分析一下可以看到，在第一次对F(x)进行舍选法时，选择函数h(x)与 M 对比如下

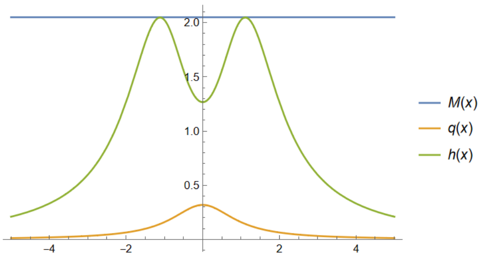

图3：第一次抽样图像

由图可见，第一次抽样时，q(x)较为平缓，有较多区域被舍去，所以导致了舍选效率偏低。

但从下图可以看到第二次对高斯函数抽样效率明显较高，是由于选择的比较函数效果较好。

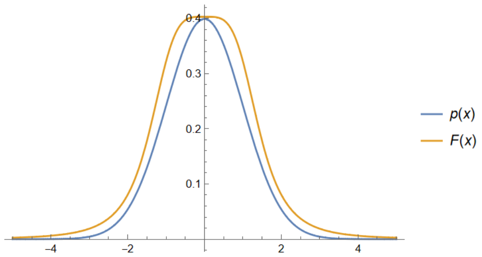

图4：第二次抽样图像

4. Summary

本次实验采取了乘分布与比较法，进行舍选抽样，最终成功得到满足高斯分布的抽样样本。

实验过程中，由于分布的区间是实数集，采取了用$\pm 5\sigma$处的值替代 5σ之外的值进行截断处理。

在实验中也看到抽样效率只有大约 50%左右，主要是由于第一次乘分布选择的函数分布平缓，导致的抽样效率偏低，或许可以微调一些参数，使得乘分布法抽样效率可以提高。

从比较法抽样中可以看到，选择一个合适的比较函数，可以大大提高抽样的最终效率。