舍选法生成随机样本的 R 模拟

数 41 李博扬 2014012118

【模拟目的】

对于给定的目标概率密度函数，用 R 模拟通过舍选法生成若干以此函数为概率密度函数的随机样本，并探究几种不同情况下生成随机样本的可靠性和效率。

【模拟原理:舍选法】

设 f(x)为目标概率密度函数，随机变量 V 服从概率密度函数 g(v)，f 与 g 有相同的支撑集。

令$\mathrm{M} = sup_x(\mathrm{f(x)/g(x)}) < \infty$

第一步:生成独立随机变量 U、V,其中 U 服从(0,1)上均匀分布,V 服从 g(v)。

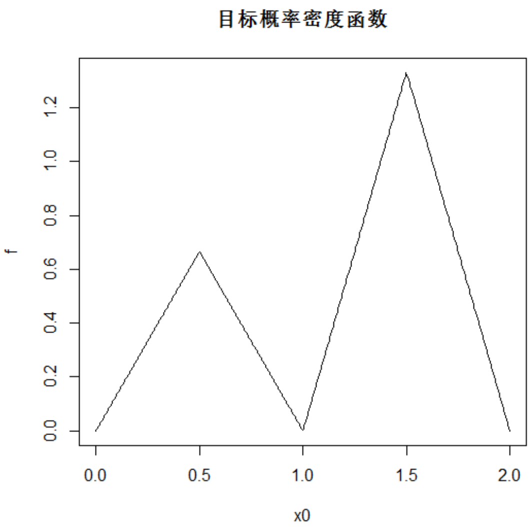
第二步:判断，若$U < \frac{1}{M} f(V)/g(V)$,则令随机变量 X=V。否则返回第一步。

上述过程在 R 模拟中重复 1000 次，得到 1000 个生成的满足概率密度函数 f(x)的随机样本 $X=(X_1, X_2, \cdots, X_{1000})$，并生成相应密度直方图与目标函数概率密度函数图对比。

【模拟过程】

目标函数为:（归一化后）

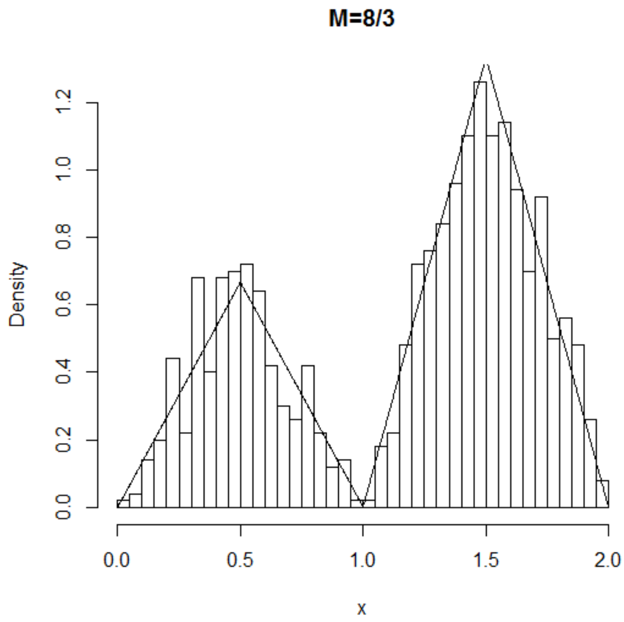
$$\mathrm{f(x)} = \begin{cases} \frac{4}{3}x & 0 \le x < 0.5 \\ -\frac{4}{3}x + \frac{4}{3} & 0.5 \le x < 1 \\ \frac{8}{3}x - \frac{8}{3} & 1 \le x < 1.5 \\ -\frac{8}{3}x + \frac{16}{3} & 1.5 \le x < 2 \end{cases}$$



1、探究不同的 M 值对生成的随机样本 X 的影响（V 为(0,2)上的均匀分布时）

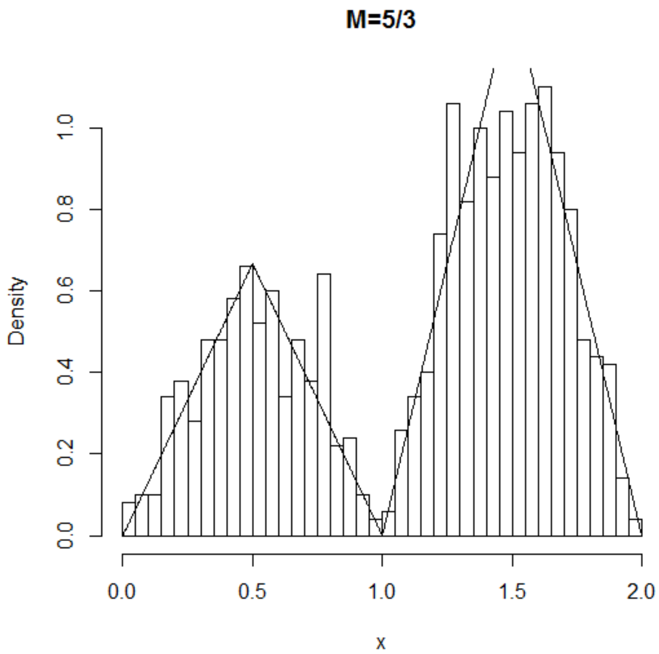
（1）M=8/3 时（此时的 M 即为$sup_x(\mathrm{f(x)/g(x)})$

得到生成的随机样本 X 的密度直方图如下



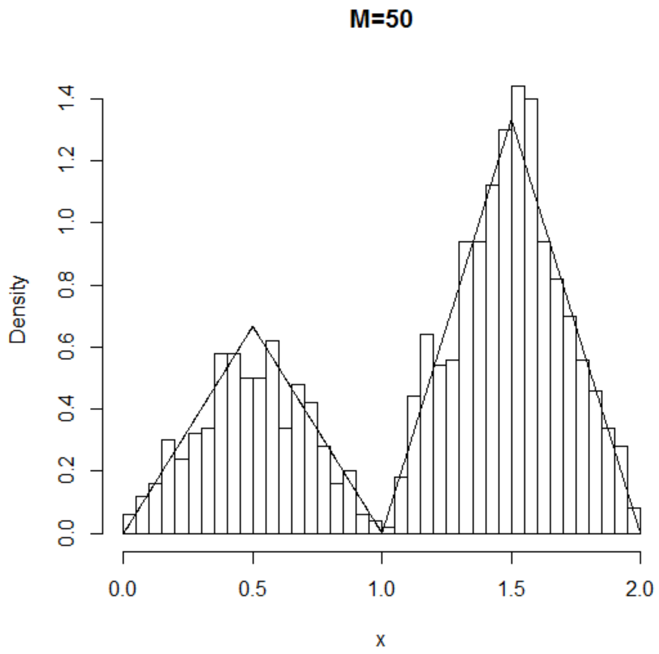
（2）M=5/3 时

得到生成的随机样本 X 的密度直方图如下



（3）M=50 时

得到生成的随机样本 X 的密度直方图如下



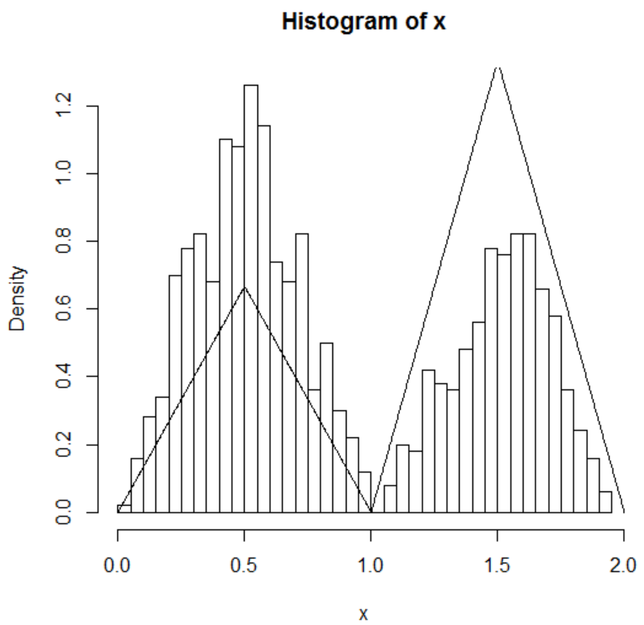
生成一个随机样本 X 平均需要的计算次数如下：

M=5/3	1.836
M=8/3	2.819
M=50	48.333

2、探究不同概率密度函数 g 对生成的随机样本 X 的影响

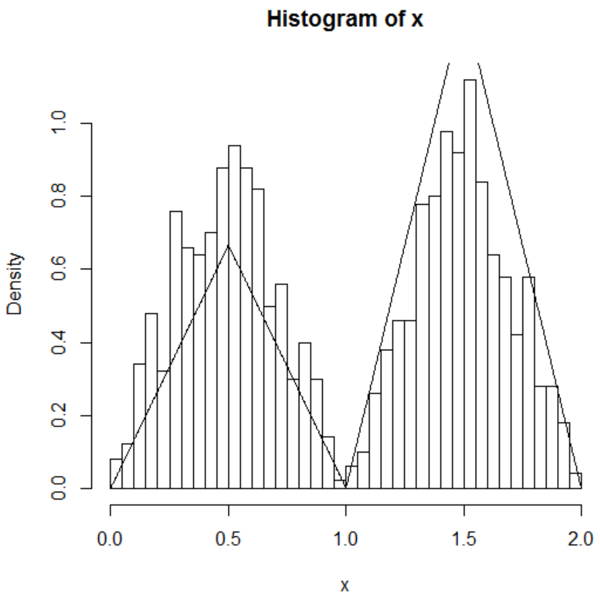
（1）取 g(v)= $\begin{cases} \frac{1}{4} & 0 \le v < 1 \\ \frac{3}{4} & 1 \le v \le 2 \end{cases}$ 此时 M= $sup_x(f(x)/g(x))=\frac{8}{3}$

得到生成的随机样本 X 密度直方图如下



（2）取 $g(v)=\begin{cases}\frac{1}{3} & 0\le v<1\\ \frac{2}{3} & 1\le v\le 2\end{cases}$　　此时 M=$sup_x(\mathrm{f(x)/g(x)})$=2

得到生成的随机样本 X 密度直方图如下



生成一个随机样本 X 平均需要的计算次数如下：

g 为（0,2）均匀分布	2.819
$g(v)=\begin{cases}\frac{1}{3} & 0\le v<1\\ \frac{2}{3} & 1\le v\le 2\end{cases}$	1.951
$g(v)=\begin{cases}\frac{1}{4} & 0\le v<1\\ \frac{3}{4} & 1\le v\le 2\end{cases}$	2.478

【模拟结论】

1、当固定概率密度函数 g 时，M 取不同值（分别取了$sup_x(\mathrm{f(x)/g(x)})$以及比它大、小的值），可以从生成的随机样本 X 的密度直方图看出，当 M 恰好取$sup_x(\mathrm{f(x)/g(x)})$时，密度直方图与密度函数曲线符合得最好；当 M 取小于$sup_x(\mathrm{f(x)/g(x)})$时，密度直方图反映出生成的随机样本 X 取值出现明显偏差，且对 M 的变化十分敏感，虽然平均生成随机样本效率高，但不可靠；当 M 取大于$sup_x(\mathrm{f(x)/g(x)})$时，从图上看几乎对随机样本 X 的取值无影响，但生成随机样本 X 的效率随 M 增大越来越低。

事实上，从原理上来看，当 M>$sup_x(\mathrm{f(x)/g(x)})$时，舍选法的第二步：$U<\frac{1}{M}f(V)/g(V)<\frac{1}{sup_x(\mathrm{f(x)/g(x)})}f(V)/g(V)$，于是选出的 X 就等价于 M= $sup_x(\mathrm{f(x)/g(x)})$时选出的 X，这也就解释了当 M 取大于$sup_x(\mathrm{f(x)/g(x)})$时为何对选出的随机样本 X 无影响。

2、当固定 M 取值为$sup_x(\mathrm{f(x)/g(x)})$时，概率密度函数变为

g(v)=$\begin{cases}\frac{1}{3} & 0 \le v < 1 \\ \frac{2}{3} & 1 \le v \le 2\end{cases}$和 g(v)=$\begin{cases}\frac{1}{4} & 0 \le v < 1 \\ \frac{3}{4} & 1 \le v \le 2\end{cases}$时，生成的随机样本 X 的可靠性明显降低，虽然效率较高，但不可取。