舍选抽样与采样重要性重抽样算法的比较

王丙参[1],魏艳华[1],孙永辉[2]
(1.天水师范学院 数学与统计学院,甘肃 天水 741001; 2.河海大学 能源与电气学院,南京 210098)

摘 要:文章比较研究了舍选法和重要性重抽样(SIR)算法生成随机数的理论基础,给出了二者的区别与联系,特别讨论了压挤舍选抽样和自适应舍选抽样,并给出了包络函数和重要性抽样函数的选择标准,探讨二者对随机数生成速度和质量的影响。

关键词:舍选法;重要性重抽样;接受概率;包络函数
中图分类号:O212 **文献标识码**:A **文章编号**:1002-6487(2014)21-0009-05

0 引言

蒙特卡洛方法(M-C)是工程、科学、金融等领域中常用的数值方法,在解决实际问题时,首先要产生目标分布的随机数,然而真随机数由于受到技术、成本、可重复性的约束往往不可取。科学计算界广为大家接受的替代方法就是利用随机数生成器产生伪随机数,虽然它并不是真正的随机数,但其统计性质与真随机数相互几乎没有区别,且伪随机数成本低、方便、可重复[1~3]。舍选抽样和采样重要性重抽样[4,5]是非常有用的随机数生成方法,二者具有联系和区别,虽然目前对二者研究的文献很多[6,7],但几乎很少有学者对二者进行比较研究。二者在理论上具有相通之处,比如包络函数的选择,但在很多方面存在区别,遗憾的是很多读者二种方法中的某些概念常常混淆。鉴于此,本文比较研究了舍选法和重要性重抽样(SIR)算法生成随机数的理论基础,给出了二者的区别与联系,特别讨论了压挤舍选抽样和自适应舍选抽样,并给出了包络函数和重要性抽样函数的选择标准,探讨二者对随机数生成速度和质量的影响,最后结合实例验证了结论。

1 舍选抽样法算法

基金项目:国家自然科学研究基金资助项目(61104045);甘肃省自然科学研究基金计划(096RJZE106)
作者简介:王丙参(1983-),男,河南南阳人,硕士,讲师,研究方向:随机过程和统计计算。

经济总收入Y可采用一般的不放回不等概率抽样中的Horvitz-Thompson估计量:$\hat{Y}_{HT}=\sum_{i=1}^{n}\frac{y_i}{\pi_i}$;$\hat{Y}_{HT}$是全部城乡居民总体的非正规经济总收入Y的无偏估计,其方差为:$V(\hat{Y}_{HT})=\sum_{i=1}^{N}\frac{1-\pi_i}{\pi_i}Y_i^2+2\sum_{i=1}^{N}\sum_{j>i}^{N}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}Y_iY_j$;当n固定时,其又可表示为:$V(\hat{Y}_{HT})=\sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i\pi_j-\pi_{ij})(\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j})^2$。

由于在πps系统抽样中,并不总能保证所有的$\pi_{ij}>0$,所以,不一定能用样本进行简单估计上述方差,其两种方差估计都不适用于系统样本:

$$v_1(\hat{Y}_{HT})=\sum_{i=1}^{n}\frac{1-\pi_i}{\pi_i^2}y_i^2+2\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j\pi_{ij}}y_iy_j;$$

$$v_2(\hat{Y}_{HT})=\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{(\pi_i\pi_j-\pi_{ij})}{\pi_{ij}}(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j})^2$$

当我们把不放回的πps系统抽样样本作为放回的pps抽样样本处理时,可以得到以下方差的估计形式:

$$v_3=\frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{y_i}{z_i}-\hat{Y}_{HT})^2=\frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{ny_i}{\pi_i}-\hat{Y}_{HT})^2$$

但这样处理会"高估"方差,因此,可将其乘以有限总体修正系数$1-f$。考虑到这里的单元抽取概率不相等,f不能简单地取为$\frac{n}{N}$,一般情况下,可采用f的一个简单估计$\hat{f}=\frac{1}{n}\sum_{i=1}^{n}\pi_i$,有方差估计量的另一种表达式:

$$v_4=\left(1-\hat{f}\right)v_3=\frac{1-\sum_{i=1}^{n}\frac{\pi_i}{n}}{n(n-1)}\sum_{i=1}^{n}(\frac{ny_i}{\pi_i}-\hat{Y}_{HT})^2$$

根据有关模拟研究表明,对于随机排列的总体,v_4是一个较好的方差估计量。

参考文献:
[1]European Economic Community Commission; International Monetary Fund; Organization for Economic Co-operation and Development; United Nations; World Bank. System of National Accounts 2008 [EB/OL]. P472,http://unstats.un.org/unsd/nationalaccount/sna, 2008.asp (2009-06-01)[2009-08-25].
[2]杜子芳.抽样技术及其应用[M].北京:清华大学出版社,2005.
[3]金勇进等.抽样技术[M].北京:中国人民大学出版社,2002.

(责任编辑/亦 民)

舍选抽样法至少在理论上可从任意维数的给定概率分布中抽样，它不是对所产生的随机数都录用，而是建立一个检验条件，利用这一检验条件进行舍选得到所需的随机数。由于舍选抽样灵活、计算简单、使用方便而得到了较为广泛的应用[6,7]。

如果定义在任意空间 Ω 上的 f 是我们感兴趣的目标密度函数（也常称为目标分布），由于 $f(x)=\int_0^{f(x)}\mathrm{d}u$，则 f 可以作为联合分布 $(X,U)\sim U\{(x,u):0<u<f(x)\}$ 的对于随机变量 X 的边缘密度。由于 U 和原问题没有直接关系，故称为辅助变量。我们可以通过集合 $\{(x,u):0<u<f(x)\}$ 上的均匀随机数生成联合分布 (X,U) 随机数，由于边缘分布 X 具有目标分布 f，故我们已经产生了 f 随机数。

简言之，模拟 $X\sim f(x)$ 等价于模拟 $(X,U)\sim U\{(x,u):0<u<f(x)\}$。

例如，在1维场合，假定 $\int_a^b f(x)\mathrm{d}x=1$ 且 $f\le M_1$，我们可以通过模拟 $Y\sim U(a,b)$ 和 $U|Y=y\sim U(0,M_1)$ 生成随机对 $(Y,U)\sim U(0<u<M_1)$，如果 $0<u<f(y)$，则接受 Y 的取值，称为 X 随机数，因为

$$P(X\le x)=P(Y\le x|U<f(Y))=\frac{\int_a^x\int_0^{f(y)}\mathrm{d}u\mathrm{d}y}{\int_a^b\int_0^{f(y)}\mathrm{d}u\mathrm{d}y}=\int_0^x f(y)\mathrm{d}y$$

这就是说，如果 $A\subset B$ 且我们生成了 B 上随机数，则通过舍选抽样就可得到 A 上随机数。我们也很容易计算出接受概率

$$p=P(U<f(Y))=\frac{1}{M_1}\int_0^1\int_0^{f(y)}\mathrm{d}u\mathrm{d}y=\frac{1}{M_1}$$

假定备选区域 $\mathcal{F}=\{(y,u):0<u<m(y)\}$，$f(x)\le m(x)$，且很容易生成 $\mathcal{F}$ 上的均匀随机数，显然 $\mathcal{F}$ 的测度有限，否则不存在 $\mathcal{F}$ 上的均匀分布，$m(x)$ 不是密度函数，但我们可写成

$m(x)=Mg(x)$，且 $\int_X m(x)\mathrm{d}x=\int_X Mg(x)\mathrm{d}x=M$

我们首先生成 $Y\sim g$，$U|Y=y\sim U(0,Mg(y))$，当 $u<f(y)$ 时，我们接受 y，则 $y\sim f(x)$。事实上，对于任意可测集 A，我们有

$$P(X\in A)=P(Y\in A|U<f(Y))$$

$$=\frac{\int_A\int_0^{f(y)}\frac{1}{Mg(u)}\mathrm{d}u g(y)\mathrm{d}y}{\int_A\int_0^{+\infty}\frac{1}{Mg(u)}\mathrm{d}u g(y)\mathrm{d}y}=\int_A f(y)\mathrm{d}y$$

定理 设 $f(x),g(x)$ 为pdf，$h(x)$ 为给定的函数，不一定是pdf，如果按下法进行舍选抽样：

（1）生成 $X\sim f(x)$，$Y\sim g(x)$，且 X,Y 相互独立；

（2）若 $Y\le h(X)$，令 $Z=X$，则 Z 的pdf为 $p(z)=\frac{f(z)G(h(z))}{\int_{-\infty}^{+\infty}f(y)G(h(y))\mathrm{d}y}$，其中 $G(y)=\int_{-\infty}^{y}g(x)\mathrm{d}x$

证明：

$$P(Z\le z)=P(X\le z|Y\le h(x))$$

$$=\frac{P(X\le z,Y\le h(x))}{P(Y\le h(x))}=\frac{\int_{-\infty}^{z}\int_{-\infty}^{h(x)}f(x)g(y)\mathrm{d}y\mathrm{d}x}{\int_{-\infty}^{+\infty}\int_{-\infty}^{h(x)}f(x)g(y)\mathrm{d}y\mathrm{d}x}=\frac{\int_{-\infty}^{z}f(x)G(h(x))dx}{\int_{-\infty}^{+\infty}f(x)G(h(x))dx},$$

求导可得结论成立。

推论1 设 Z 的pdf $p(z)\le M(z),\forall z\in R$，令 $C=\int_{-\infty}^{+\infty}M(x)\mathrm{d}x$，$f(x)=\frac{M(x)}{C}$，$h(x)=\frac{p(x)}{M(x)}$，如果按下法进行舍选抽样：

（1）生成 $X\sim f(x)$，$U\sim U[0,1]$，且 X,U 相互独立；

（2）若 $U\le h(X)$，令 $Z=X$，则 $Z\sim p(z)$。

推论2 设r.v Z 的pdf $p(z)=Lh(z)f(z)$，其中 $L=(\int_{-\infty}^{+\infty}f(x)h(x)\mathrm{d}x)^{-1}>1$，$0\le h(z)\le 1$，$f(z)$ 是r.v X 的pdf，如果按下法进行舍选抽样：

（1）生成 $X\sim f(x)$，$U\sim U[0,1]$，且 X,U 相互独立；

（2）若 $U\le h(X)$，令 $Z=X$，则 $Z\sim p(z)$。

推论3 设r.v Z 的pdf $p(z)=L\int_{-\infty}^{h(z)}g(z,y)dy$，其中 $g(z,y)$ 为随机向量 (X,Y) 的联合pdf，$h(z)$ 在 Y 的定义域上取值，L 为规格化常量，如果按下法进行舍选抽样：

（1）生成 $(X,Y)\sim g(x,y)$；（2）若 $Y\le h(X)$，令 $Z=X$，则 $Z\sim p(z)$。

注：设 $X,Y\sim U(0,1)$ 相互独立，即 $g(x,y)=f(x)\varphi(y)$，则 $p(z)=Lh(z)f(z)$，此时正是推论2；若 $Y\sim U(0,1)$，$X\sim f(x)=M(x)/L$，其中 $M(x)$ 是 $p(x)$ 的上界函数，则 $p(x)=M(x)h(x)$，此时正是推论1。可见推论1、2是推论3的特例。

产生一对随机数 (X,U) 称作一次试验，一次试验不能保证产生一个随机数 $Z\sim p(z)$，一次试验产生随机数 Z 的概率叫做舍选法的接受概率[5]，记作 p_0，即 $p(z)$ 随机数 Z 在取舍原则中被选中的概率（舍选抽样法的效率），经取舍原则首次接受时已取舍的次数记为 N，则

$$p_0=P(U\le p(X)/M(X))=\int_a^b\frac{p(x)}{M(x)}f(x)dx=\frac{1}{C},$$

$N\sim Ge(p_0)$，$EN=C$，可见 C 越小，取舍的效率越高。$M(z)$ 叫做函数 $p(z)$ 的优函数，不附加任何条件的优函数容易找到，好的优函数应该可快速产生 $X\sim M(x)/C$ 和高的接受概率，但两者往往相互制约，实用的优函数是两者的合理妥协，因此舍选法的关键是找出满足下述条件的优函数：

（1）$M(x)$ 应从上方尽量接近pdf $p(x)$；

（2）容易生成 $X\sim M(x)/C$。

常数优函数产生 $X\sim M(x)/C$ 最快、最简单，但往往接受概率太低。遗憾的是使用舍选法通常难解决高维蒙特卡罗模拟问题。

综上所述，舍选抽样法的基本思想是：按照给定的pdf $p(x)$，对易生成的随机数列 $\{r_i\}$ 进行舍选。舍选的原则是：在 $p(x)$ 大的地方，保留较多随机数 r_i，在 $p(x)$ 小的地方，保留较少随机数 r_i，，使得到的子样本中 r_i 的分布满

足pdf $p(x)$的要求。

一般的舍选抽样需要对每个备选抽样 Z 有一个 $p(x)$ 值，在 $p(x)$ 求值昂贵但舍选法却吸引人的时候，压挤舍选抽样可以提高模拟速度。若非负函数 $s(x)$ 在 $p(x)$ 的支撑上处处不超过 $p(x)$，则可选 $s(x)$ 作为压挤函数，像一般舍选法，也要用到包络 $M(x)\geq g(x)$，由推论1知算法如下：

(1)生成 $X\sim f(x)$，$U\sim U[0,1]$，且 X, U 相互独立；

(2)若 $U\leq \frac{s(X)}{M(X)}$，令 $Z=X$，则 $Z\sim p(z)$，然后转到(5)；

(3)否则，确定是否有 $U\leq h(X)$，如果不等式成立，令 $Z=X$，则 $Z\sim p(z)$，然后转到(4)；

如果 X 仍未保留，拒绝其成为目标随机样本之一；

返回(1)，直到达到所需的样本量。

可见，压挤舍选抽样的总接受概率仍为 $1/C$，步骤(2)基于 $s(x)$ 而非 $p(x)$ 决定时候保留 X。当 $s(x)$ 紧紧靠在 $p(x)$ 的下方时，且 $s(x)$ 容易计算，则可大大减少计算量，避免计算 $p(x)$ 的比例为 $\int s(x)\mathrm{d}x \Big/ \int M(x)\mathrm{d}x$。

舍选抽样中的关键是构造合适的包络，Gilks和Wild提出了一种针对支撑连通区域上连续、可导、对数凹密度的自动包络生成方法，也成为自适应舍选抽样。

令 $l(x)=\log p(x)$，假设在某实区间上 $p(x)>0$，可能取值无穷，$p(x)$ 是凹的，满足支撑区域内任意三点 $a<b<c$ 有 $l(a)-2l(b)+l(c)<0$。$l'(x)$ 存在且随 x 的增大单调递减，但可能有间断点。在点 $x_1<x_2<\cdots<x_k$ 处计算 $l(x), l'(x)$，如果 $p(x)$ 的支撑延伸到 $-\infty$，选择 x_1 s.t. $l'(x_1)>0$，如果 $p(x)$ 的支撑延伸到 ∞，选择 x_k s.t. $l'(x_k)<0$。令 $T_k=\{x_1,\cdots,x_k\}$，T_k 上拒绝包络为 l 在 T_k 内各点处切线组成的分段线性上覆盖指数。$l(x)$ 在 x_i 的切线公式由点斜式可得 $l(x_i)+(x-x_i)l'(x_i)$，在 x_{i+1} 处得切线为 $l(x_{i+1})+(x-x_{i+1})l'(x_{i+1})$，两切线在点 $z_i=\frac{l(x_{i+1})-l(x_i)-x_{i+1}l'(x_{i+1})+x_il'(x_i)}{l'(x_{i+1})-l'(x_i)}$ 处相交，因此 l 的上覆盖为 $m_k^*=l(x_i)+(x-x_i)l'(x_i), x\in[z_{i-1},z_i]$，且 $i=1,\cdots,k$，z_0, z_k 分别为 $p(x)$ 支撑区域的下界和上届，可能取无穷大。综上所述，拒绝包络 $M_k(x)=\exp\{m_k^*(x)\}$。

T_k 上压挤函数为 l 在 T_k 内各相邻点的弦组成的分段线性下覆盖指数。$l(x)$ 的下覆盖由 $s_k^*(x)=\frac{(x_{i+1}-x_i)l(x_i)+(x-x_i)l(x_{i+1})}{x_{i+1}-x_i}, x\in[x_i,x_{i+1}]$，$i=1,\cdots,k$ 给出。当 $x<x_1$ 或 $x>x_k$ 时，令 $s_k^*(x)=-\infty$。这样压挤函数为 $s_k(x)=\exp\{s_k^*(x)\}$。

自适应舍选抽样通过选择一个适合的 k 和相应的网格 T_k 来初始化。第一次迭代像压挤舍选抽样一样进行，分别用 $M_k(x), s_k(x)$ 作为包络及压挤函数。当一个备选抽样被接受时，如果满足压挤条件，就不用计算 $l(x), l'(x)$ 即可直接接受。不过，它也可能在第二阶段被接受，这是需要计算备选抽样出的 $l(x), l'(x)$，同时接受点加到 T_k 中，得到 T_{k+1}，并计算函数 $M_k(x), s_k(x)$。迭代继续。如果一个新的接受点与 T_k 中的点重合，则不必更新 T_k 与 $M_k(x), s_k(x)$。

若对点集 T_k，定义 $L_i(x)$ 为连接 $(x_i, l(x_i))$ 和 $(x_{i+1}, l(x_{i+1}))$ 的直线函数，其中 $i=1,\cdots,k-1$，则包络函数 $m_k^*(x)=\begin{cases}\min\{L_{i-1}(x), L_{i+1}(x)\}, x\in[x_i,x_{i+1}]\\ L_1(x), x<x_1\\ L_{k-1}(x), x>x_k\end{cases}$，并约定 $L_0(x)=L_k(x)=\infty$，则 $m_k^*(x)$ 是 $l(x)$ 的上覆盖，$M_k(x)=\exp\{m_k^*(x)\}$ 为 $p(x)$ 的包络函数。这样生成包络函数可以避免计算 $l'(x)$。我们希望在 $p(x)$ 取最大值的附近网格点最密集，幸运的是，这将自动发生。因为这样的点在迭代中最可能保留，从而被更新到 T_k 中。

总之，不加约束的包络函数 $M(x)$ 很容易选择，比如常数包络函数，但接受效率太低，因此它选择标准是：(1)容易生成 $M(x)/C$ 随机数，且易计算 $M(x)$，从而提高运算速度；(2) $M(x)$ 应从上方尽可能接近 $p(x)$，从而提高接受效率。有时候pdf $p(x)$ 计算复杂，为较少它的计算次数，我们可引进压挤函数 $s(x)$ 来提高计算速度，它的选择原则是：$s(x)$ 应从下方尽可能接近 $p(x)$，且容易计算。

2 采样重要性重抽样(SIR)算法

SIR算法仿真了近似目标分布，简单说，就是通过一个重要性抽样函数 $g(x)$ 中抽取一个样本来进行，在非正式场合，我们称 $g(x)$ 为包络，样本中的每个点通过加权修正抽样概率从而近似目标分布 $f(x)$。从严格意义上说，此处的包络 $g(x)$ 不同于舍选抽样中的包络，舍选抽样中的包络在每点处函数值都大于目标分布 $f(x)$，不是密度函数，而在SIR算法中 $g(x)$ 是密度函数，是重要性抽样函数，它不可能每点函数值都大于 $f(x)$，因为 $g(x)$，$f(x)$ 都是密度函数，与 x 轴围成的面积都是1。

$w(x_i)=\frac{w(x_i)}{\sum_{i=i}^m w(x_i)}$ 称为标准化权重，其中 $w(x)=f(x)/g(x)$，$x_1,\cdots,x_m$ 是来自包络 $g(x)$ 的i.i.d样本。它在 f 仅差一个比例常数下使用，比如在贝叶斯分析中，f 是后验密度。

SIR算法其实是将每个观测点 x_i 有概率 $w(x_i)$ 的离散分布来近似目标分布 $f(x)$，具体算法如下：

生成i.i.d的备选样本 $y_1,\cdots,y_n\sim g$。

计算标准化权重 $w(y_i)=\frac{w(y_i)}{\sum_{i=i}^m w(y_i)}$，其中 $w(y)=f(y)/g(y)$。

以概率 $w(y_i), i=1,\cdots,m$ 从 $y_i, i=1,\cdots,m$ 中有放回地重新抽取样本 $x_1,\cdots,x_n$。

当 $m\to\infty$ 时，样本 $X_1,\cdots,X_n$ 的分布收敛到 $f(x)$。我

们可简要证明如下：

设 $Y_1,\cdots,Y_m$ i.i.d 于 g，给定集合 A，则 $P(X\in A|Y_1,\cdots,Y_m)=\frac{\sum_{i=1}^m I_{\{Y_i\in A\}}w(Y_i)}{\sum_{i=1}^m w(Y_i)}$。由强大数定律可得，当 $m\to\infty$ 时，

$\frac{1}{m}\sum_{i=1}^m w(Y_i)\to 1$，

$\frac{1}{m}\sum_{i=1}^m I_{\{Y_i\in A\}}w(Y_i)\to E[I_{\{Y_i\in A\}}w(Y_i)]=\int_A w(y)g(y)\mathrm{d}y=\int_A f(y)\mathrm{d}y$，

即 $P(X\in A|Y_1,\cdots,Y_m)\to\int_A f(y)\mathrm{d}y$。最后，由控制收敛定理可得

$P(X\in A)=E[P(X\in A|Y_1,\cdots,Y_m)]\to\int_A f(y)\mathrm{d}y$。

虽然舍选抽样和SIR都依赖目标密度函数 f 与包络 g 的比例，但在某种重要程度上它们是不一样的。舍选抽样仿真了精确分布 f，但SIR算法仿真了近似目标分布 f，但舍选抽样生成容量为 n 的一个样本所需要的随机数个数是随机的，依赖于接受概率，而SIR利用确定个数个随机数生成容量为 n 的一个样本，不过它的样本点对 f 有个随机的近似程度，即样本 $X_1,\cdots,X_n$ 的分布收敛到 $f(x)$。

在利用SIR算法时，需要考虑初始抽样和重抽样的相对大小，即需要考虑 m,n，理论上，样本依分布收敛需要 $\frac{n}{m}\to 0$。当 $n\to\infty$ 时，意味着 $m\to\infty$ 的速度更快。当 n 固定时，只要 $m\to\infty$ 就会出现样本依分布收敛。在实际中，我们会选择尽可能大的 m，但为了提高精度，我们也会选择尽可能大的 n，可见，$\frac{n}{m}$ 的最大容量取决于包络的质量。在一般情况下，当 $\frac{n}{m}\le\frac{1}{10}$ 时就可以，只要生成的样本相对初始样本没有过多重复就行。

由于利用来自 g 的重置样本近似来自 f 的样本，故包络 g 的支撑一定要包含目标分布 f 的支撑且 g 应该比 f 更厚的尾，在这点上，拒绝抽样和SIR对包络的要求是一致的。我们应当选择 g 以保证 $\frac{f(x)}{g(x)}$ 的增长不要过快，如果 $g(x)$ 很多点处几乎为0，而 $f(x)>0$，尽管这样的样本很难出现，但一旦出现，将获得很大的权重，以至于在SIR重抽样中经常出现，导致近似效果变差，但如果在舍选抽样中出现这样情况，将会导致抽样效率大大降低，而不会导致仿真出现偏差，即样本仍然是精确分布 f。如果在SIR中有几个标准化重要权重远远大于其他权重，就会导致二次抽样几乎是几个样本的重复值，当问题不是特别严重时，我们建议采用无放回的再抽样策略，它渐进等价于有放回抽样，但可避免过度重复样本的出现，遗憾的是它在最后抽样中又导致了近似，当问题严重时，建议更改包络函数 g。

在很多场合，最初也许只能找到一个很差的包络，例如，当目标函数是多维分布或定义域为曲面时，由于有未被分析员充分了解的变量相依性导致了这种情况，这就需要调整重要性包络。同拒绝抽样类似，也可采用自适应重要性抽样。我们从初始包络 e_1 中抽取 m_1 个初始样本，将样本加权重以得到感兴趣的初始估计或者 f 本身初始观测值，基于此改进包络产生 e_2，重复进行，直到满足要求。在参数的自适应SIR中，一般假定包络属于某个低维分布族，参数的最优选择都在每次迭代中进行估计直到该参数的估计稳定为止。在非参自适应SIR中，通常假定包络为混合分布，重要性抽样由包络更新、加、减及混合成分更新交替进行。

有时候，我们需要对多个相关问题构造重要性抽样，但没有包络适合所有感兴趣的目标分布。在贝叶斯统计中，通常对估计一对密度的归一化常数的比率感兴趣。例如，如果 $\pi_i(\theta|x)=c_i q_i(\theta|x)$，$i=1,2$ 表示两个竞争模型的后验密度，其中 c_i 未知，q_i 已知，则 $r=\frac{c_2}{c_1}$ 表示第1个模型与第2个模型的后验胜算比。一般情况下，很难对目标分布 π_1,π_2 都找到较好的重要性抽样包络，当 π_2 支撑包含 π_1 的支撑且 $r=E[\frac{q_1(\theta|x)}{q_2(\theta|x)}]$，我们可采用单个包络来估计后验胜算比 r。如果 π_1 与 π_2 差别大，由于没有一个包络能充分提供 c_1,c_2 的信息，故这样的方法很差，但我们可以用一个未归一化的密度 q_b，即介于 q_1,q_2 的密度。由于 $r=\frac{E_{\pi_2}[q_b(\theta|x)/q_1(\theta|x)]}{E_{\pi_1}[q_b(\theta|x)/q_1(\theta|x)]}$，我们看利用SIR方法估计分母和分子，这可简化计算，因为 q_b 与每个 q_i 之间的距离比两个 q_i 之间的距离更近。

3 实例分析

例1 假定从混合泊松总体 $X|\lambda\sim P(\lambda)$ 中随机抽取10个样本，观测值为(6 2 7 8 1 7 2 3 4 3)，λ 的先验分布为对数正态分布：$\log(\lambda)\sim N(1,0.5^2)$，密度函数为 $f(\lambda)$，记似然函数为 $L(\lambda|x)$，由于似然估计值使得样本出现的概率最大，故 $\hat{\lambda}=\bar{x}=4.3$ 使得似然函数为 $L(\lambda|x)$ 关于 λ 最大，进而有未归一化后验密度 $g(\lambda|x)=f(\lambda)L(\lambda|x)$ 被 $e(\lambda)=L(4.3|\lambda)f(\lambda)$ 覆盖，即 $e(\lambda)$ 为后验密度的包络[7-8]。综上所述，抽取后验分布随机数的舍选算法如下：

独立抽取对数正态随机数 $\log(x)\sim N(1,0.5^2)$ 和均匀随机数 $u\sim U(0,1)$。

如果 $u<\frac{g(\lambda|x)}{e(\lambda)}=\frac{L(\lambda|x)}{L(4.3|x)}$，令 $\lambda=x$，则 λ 为贝叶斯后验分布随机数。

我们利用此算法生成10000个贝叶斯后验分布随机数，其频率直方图如图1，其中实线为先验分布的密度函数。后验分布的均值为4.1377，标准差为0.6132，即 λ 的贝叶斯估计为4.1377，比经典估计4.3偏小。值得注意的是接受概率与先验分布的关系密切，在一定范围内，对数正态的均值越大，尤其远远大于样本观测值的均值4.3时，接受概率越低。当先验分布的均值接近4.3时，接受概率还是可以的，并且该方法简单准确。

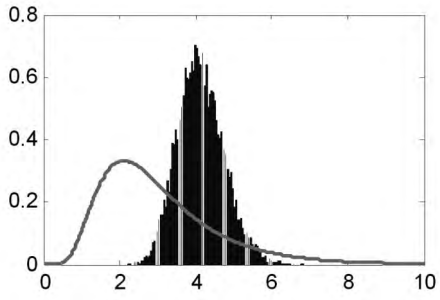

图1 舍选抽样法生产贝叶斯后验随机数的直方图

例2 如果 $X\sim N(0,1)$，$U\sim U(0,1)$，则 $Y=\frac{X}{U}$ 服从斜线分布。

当 $y\neq 0$ 时，$f_Y(y)=\int_{-\infty}^{+\infty} f_X(uy)f_U(u)|u|\mathrm{d}u$ $=\int_0^1 \frac{1}{\sqrt{2\pi}}\exp(-\frac{y^2u^2}{2})u\mathrm{d}u=\frac{1-\exp(-y^2/2)}{y^2\sqrt{2\pi}}, y\neq 0$。由密度函数的连续性可得斜线分布的 pdf 为 $f(y)=\begin{cases}\frac{1}{2\sqrt{2\pi}}, & y=0\\ \frac{1-\exp(-y^2/2)}{y^2\sqrt{2\pi}}, & y\neq 0\end{cases}$。

下面考虑用斜线分布作为 $N(0,1)$ 的重要性抽样函数来生成 $N(0,1)$ 随机数，如图2左，并反过来利用正态分布作为重要性抽样函数生成斜线随机数，如图3.2右，其中 $m=100000$，$n=5000$，实线表示目标密度。尽管我们可以利用标准算法模拟斜线分布和正态分布，SIR算法并不是必须的，但这具有启发性。

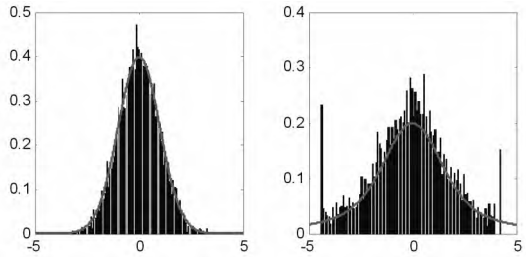

图2 SIR近似抽样的直方图与目标密度

由于 $f(y)$ 具有厚尾，故它是一个很好的重要性抽样函数，模拟结果显示样本的频率直方图非常接近标准正态分布的密度曲线，效果很好。当利用正态密度作为斜线分布的重要抽样性函数时，由于正态密度的尾部远远轻于目标密度的尾部，尽管斜线分布在远离原点10个单位处的概率是可估的，但由于正态密度的备选样本很少有超过原点5个单位的地方，故斜线分布的密度函数的模拟尾部被截去了，并且在接近 ±5 的样本点重要性比例很高，这导致了再抽样时产生了大量重复，最终模拟效果不尽人意。由此例也验证了前面的理论分析，即重要抽样性函数 g 的支撑一定要包含目标分布 f 的支撑且 g 应该比 f 更厚的尾，应当选择 g 以保证不会产生某些样本点获得很大的权重。

参考文献：

[1]刘军著,唐年胜,周勇,徐亮译.科学计算中的蒙特卡罗策略[M].北京:北京大学出版社,2009.

[2]魏艳华,王丙参,何万生.利用样本分位数求逆威布尔分布参数的渐近估计[J].统计与决策,2011,27(16).

[3]魏艳华,王丙参编著.概率论与数理统计[M].成都:西南交通大学出版社,2013.

[4]孙文彩,杨自春,李昆锋.结构混合可靠度计算的自适应重要性重要性抽样方法[J].华中科技大学学报(自然科学版), 2012, 40 (10).

[5]苏兵,高理峰.非线性贝叶斯动态模型的重要性再抽样[J].数学杂志, 2012, 32 (2).

[6]Givens G H,Hoeting J A 著,王兆军,刘民千,邹长亮等译.计算统计[M].北京:人民邮电出版社,2009.

[7]Robert C P,Casella G.Monte Carlo Statistical Methods[M].北京:世界图书出版社,2009.

（责任编辑/浩　天）