正交三角分解

$\beta_1' = \alpha_1 \quad \beta_1 = \beta_1'/\|\beta_1'\|_2$

$\beta_2' = \alpha_2 - (\alpha_2, \beta_1)\beta_1 \quad \beta_2 = \beta_2'/\|\beta_2'\|_2$

$\beta_3' = \alpha_3 - (\alpha_3, \beta_1)\beta_1 - (\alpha_3, \beta_2)\beta_2 \quad \beta_3 = \beta_3'/\|\beta_3'\|_2$

$$A = [\beta_1, \beta_2 \cdots \beta_n] \begin{bmatrix} \|\beta_1'\|_2 & (\alpha_2, \beta_1) & (\alpha_3, \beta_1) & \cdots \\ & \|\beta_2'\|_2 & (\alpha_3, \beta_2) & \cdots \\ & & \|\beta_3'\|_2 & \end{bmatrix}$$

特征值分解

$A = Q\Sigma Q^{-1}$ Q为特征向量矩阵，Σ为特征值对角阵

奇异值分解

$A = U\Sigma_A V^T$

$U = (u_1, u_2 \cdots u_m)$

$V = (v_1, v_2 \cdots v_n)$

$(A^TA)v_i = \lambda_i v_i$

$\sigma_i = \sqrt{\lambda_i}$

$u_i = \frac{1}{\sigma_i} A v_i$

$$\Sigma_A = \begin{bmatrix} \sigma_1 & & & 0 & 0 \\ & \ddots & & \vdots & \vdots \\ & & \sigma_r & 0 & 0 \\ 0 & \cdots & & 0 & 0 \end{bmatrix}$$

非0奇异值从大到小↗

线性同余随机数发生器

$X_n = (aX_{n-1} + c) \pmod M \qquad R_n = \frac{X_n}{M}$

满周期：

1. c与M互素
2. 对M的任一素因子p，a-1被p整除
3. 如果4是M的因子，则a-1被4整除

阶数(乘同余法)

$a^v \equiv 1 \pmod M$ a与M互素，v为a对M的阶数

K-S检验法

$F_n(X_k)$ 实际累计分布 $= \frac{\text{累计点数}}{\text{总点数}}$ $\qquad F_0(X_k)$ = 待验证的理论分布

$d_{k1} = |F_n(X_k) - F_0(X_k)| \quad d_{k2} = |F_n(X_k) - F_0(X_{k-1})| \quad \delta_k = \max(d_{k1}, d_{k2})$

$D = \max(\delta_k)$

逆变换法

$Y=F^{-1}(U)$

$Y=a_i$ 当且仅当 $F(a_{i-1})<U\le F(a_i)$

利用变换生成随机数

6.3 X有密度 $p(x)$，y为x的函数 $g(x)$，$g(x)$ 反函数为 $h(x)$，则Y的密度函数为

$$f(y)=p(h(y))\cdot|h'(y)|$$

6.4 随机向量(X,Y)有联合密度 $p(x,y)$，有 $\begin{cases}u=g_1(x,y)\\v=g_2(x,y)\end{cases}$

$$\Rightarrow\begin{cases}x=h_1(u,v)\\y=h_2(u,v)\end{cases}\quad J=\begin{vmatrix}\frac{\partial x}{\partial u}&\frac{\partial x}{\partial v}\\\frac{\partial y}{\partial u}&\frac{\partial y}{\partial v}\end{vmatrix}$$

则(U,V)的联合密度为：

$$f(u,v)=p(h_1(u,v),h_2(u,v))\cdot|J|$$

6.5 U_1,U_2 独立且同服从 $U(0,1)$

$$\begin{cases}X=\sqrt{-2\ln U_1}\cos(2\pi U_2)\\Y=\sqrt{-2\ln U_1}\sin(2\pi U_2)\end{cases}$$

X,Y独立且服从标准正态分布（Box-Muller）

舍选法

until $(Y\le p(X))$ {

从U(0,1)中取 U_1，U_2

取 $X\leftarrow a+(b-a)*U_1$，$Y\leftarrow M*U_2$

}

输出 $Z\leftarrow X$

复合法

① $P(I=i)=\alpha_i$，$z_1,z_2\dots z_m$ 都为离散型随机变量，

$$X=\begin{cases}z_1 & I=1\\z_2 & I=2\\\dots\\z_m & I=m\end{cases}$$

则 $P(X=x)=\sum_{i=1}^{m}\alpha_i P(z_i=x)$

② $z_1,z_2\dots z_m$ 都为连续型随机变量，密度函数分别为 $p_1(z)$，$p_2(z)\dots p_m(z)$

则有X的分布函数表为

$$F(x)=P(X\le x)=\sum_{i=1}^{m}\alpha_i P(z_i\le x)$$

X的密度函数为

$$p(x)=F'(x)=\sum_{i=1}^{m}\alpha_i p_i(x)$$

随机模拟法（蒙特卡洛法）

求解概率模型，产生大量随机数，分概率模型两个求解问题

随机投点法

1> 独立地产生 $U(0,1)$ 随机数：u_i，v_i $(i=1,\dots,N)$

2> 计算 $x_i=a+u_i(b-a)$，$y_i=Mv_i$，$f(x_i)$

3> 统计 $f(x_i)\geq y_i$ 的个数，得到 $\hat{p}$

4> $\hat{I}=\hat{p}M(b-a)$

平均值法

1> 独立地产生 N 个 $U(0,1)$ 随机数 u_i

2> 计算 $x_i=a+u_i(b-a)$ 和 $h(x_i)$

3> $\hat{I}=\frac{b-a}{N}\sum_{i=1}^{N}h(x_i)$

多重积分的随机投点法

1. 赋初值：试验次数 $n=0$，成功次数 $m=0$，规定随机投点试验总次数 N

2. 向 $k+1$ 维立方体 $[0\leq x_i\leq 1\ (i=1,\dots,k),\ 0\leq y\leq 1]$ 内随机投点，即产生 $k+1$ 个相互独立的均匀随机数 $\xi=(\xi_1,\dots,\xi_k,\eta)$，置 $n=n+1$

3. 判断 $n\leq N$ 是否成立，成立则转 4，否则停止模拟实验，转 5

4. 检验点 ξ 是否落入 V 中，即检验条件 $\eta\leq f(\xi_1,\dots,\xi_k)$ 是否成立，成立则置 $m=m+1$ 后转 2，否则直接转 2

5. 计算 $\theta_1=m/N$，其中 m 是 N 次试验中成功的总次数，则 $I=\theta_1$

多重积分的平均值法

1. 赋初值：ξ 落入 D 的次数 $m=0$，试验次数 $n=0$，试验总次数为 N

2. 产生 k 个相互独立服从 $[a,b]$ 区间上的均匀随机数 $\xi=(\xi_1,\dots,\xi_k)$，置 $n=n+1$

3. 判断 $n\leq N$ 是否成立，成立则转 3，否则停止抽样，转 4

4. 检验 k 维空间的点 ξ 是否落入积分区域 D，若 $\xi\in D$，置 $m=m+1$，并令 $\eta_m=\xi$，计算 $f(\eta_m)$，转 1；否则舍去，转 1，重新产生 k 维均匀随机数

5. 计算 $V_D\approx\frac{m}{N}(b-a)^k$，$E[f(\xi)\mid\xi\in D]\approx\frac{1}{m}\sum_{i=1}^{m}f(\eta_i)$

则 $I_k\approx\frac{1}{N}(b-a)^k\sum_{i=1}^{m}f(\eta_i)$

重要抽样法

$g(x)$：试投密度或重要抽样密度

令 $y_i=\frac{h(x_i)}{g(x_i)}$，则 $E(y)=\int_c\frac{h(x)}{g(x)}g(x)\,dx=I$

从而可用 y_i 的样本均值来估计 I（y_i 服从概率密度 $g(x)$）

$$I=\frac{1}{N}\sum_{i=1}^{N}\frac{h(x_i)}{g(x_i)}$$

重要性权重

$Y\sim f(y)$，求 $E(h(Y))=\int h(y)f(y)\,dy$，则有 $\hat{I}=\frac{1}{N}\sum_{i=1}^{N}h(x_i)\frac{f(x_i)}{g(x_i)}$

标准化重要抽样法

$\tilde{f}(x)=cf(x)$，$W_i=\frac{\tilde{f}(x_i)}{g(x_i)}$

$\hat{I}=\frac{\sum_{i=1}^{N}W_i h(x_i)}{\sum_{i=1}^{N}W_i}$

分层抽样法

$I=\int_C h(x)dx$ 分解为 m 个不相交子集 G_j 上的积分，在 G_j 投 n_j 个随机点，I 的 m 个部分分别用平均值法估计，得 I 的分层估计：

$$\hat{I}=\sum_{j=1}^{m}\frac{V(G_j)}{n_j}\sum_{i=1}^{n_j}h(x_{ji})$$

控制变量法

要估计随机变量 X 的期望 $\theta=E(X)$，另有随机变量 Y 满足 $E(Y)=0$，$Cov(X,Y)<0$，则 $Z=X+Y$，$E(Z)=\theta$；也可令 $Z=X+bY$，$b=-Cov(X,Y)/Var(Y)=-\rho_{X,Y}\sqrt{Var(X)/Var(Y)}$（使 $Var(Y)+2Cov(X,Y)<0$）

对立变量法

14.1 设 g 为单调函数，$U\sim U(0,1)$，则 $Cov(g(U),g(1-U))\le 0$

14.2 设 $h(X_1,X_2,\dots,X_n)$ 是关于每个自变量单调的函数，$U_1,U_2,\dots U_n$ 相互独立，则有 $Cov(h(U_1,U_2,\dots,U_n),h(1-U_1,1-U_2,\dots,1-U_n))\le 0$

$U\sim U(0,1)$，$X=F^{-1}(U)$，$Y=F^{-1}(1-U)$，$Z=\frac{X+Y}{2}$，为估计 $I=E(X)$，用：

$$Z_i=\frac{1}{2}(F^{-1}(U_i)+F^{-1}(1-U_i))$$

条件期望法

对 $Z=E(Y|X)$ 抽样，用 Z 的样本均值来估计 $I=E(Y)$.

①估计量的标准误差的 bootstrap 估计

标准误差：$X\sim F(x,\phi)$，ϕ 是一个参数，$\hat{\phi}$ 是 ϕ 的估计量，称 $SE=\sqrt{Var(\hat{\phi})}$ 为 $\hat{\phi}$ 的标准误差.

在原始样本中作有放回抽取，抽 B 个独立样本 $Y^{(b)}$，每个 $Y^{(b)}$ 样本量为 n，称 $Y^{(b)}$ 为 bootstrap 样本.

从每个 $Y^{(b)}$ 可估计得到 $\hat{\phi}^{(b)}=g(Y^{(b)})$

- 求 SE 的 bootstrap 估计步骤

1. 自原始样本 $X=(X_1,X_2,\dots,X_n)$ 按放回抽取抽得容量为 n 的样本 $Y=(Y_1,Y_2,\dots Y_n)$
2. 独立地求出 B 个（$B\ge 1000$）容量为 n 的 bootstrap 样本，$Y^{(i)}=(Y_1^{(i)},Y_2^{(i)}\dots Y_n^{(i)})$，计算每个 bootstrap 样本的 $\hat{\phi}^{(i)}$
3. 计算 $SE=\sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\phi}^{(i)}-\bar{\phi})^2}$，其中 $\bar{\phi}=\frac{1}{B}\sum_{i=1}^{B}\hat{\phi}^{(i)}$

②估计量的均方误差及偏差的 bootstrap 估计

均方误差 $MSE=\sum_{i=1}^{n}\frac{1}{n}(M-\theta)^2$

偏差 $b=\sum\frac{1}{n}(M-\theta)$

③ bootstrap置信区间——分位数法

求θ的置信水平为1-α的置信区间

对样本 $x=(x_1, x_2, \dots x_n)$ 中抽出的容量为n的 bootstrap 样本.

对每个bootstrap样本求θ的bootstrap估计:

$$\hat{\theta}_1^*, \hat{\theta}_2^*, \dots, \hat{\theta}_B^*$$

将它们从小到大排序得

$$\hat{\theta}_{(1)}^*, \hat{\theta}_{(2)}^*, \dots, \hat{\theta}_{(B)}^*$$

用 $\hat{\theta}^*$ 的分布作θ的近似分布，求 $\hat{\theta}^*$ 的分布的近似分位数 $\hat{\theta}_{\frac{\alpha}{2}}^*$ 和 $\hat{\theta}_{1-\frac{\alpha}{2}}^*$，使:

$$P\{\hat{\theta}_{\frac{\alpha}{2}}^* < \hat{\theta}^* < \hat{\theta}_{1-\frac{\alpha}{2}}^*\} = 1-\alpha$$

记 $k_1 = B\times\frac{\alpha}{2}$，$k_2 = B\times(1-\frac{\alpha}{2})$，从 $\hat{\theta}_{(k_1)}^*$ 和 $\hat{\theta}_{(k_2)}^*$ 分别作为分位数 $\hat{\theta}_{\frac{\alpha}{2}}^*$ 和 $\hat{\theta}_{1-\frac{\alpha}{2}}^*$ 的估计

则得到θ置信水平为1-α的近似置信区间 $(\hat{\theta}_{(k_1)}^*, \hat{\theta}_{(k_2)}^*)$

④ bootstrap置信区间 — bootstrap方法

设φ是总体F的一个参数(如期望)

$X=(x_1, x_2, \dots x_n)$ 为来自总体的样本，容量为n。总体的均值与方差均为未知参数，我们要利用样本值来估计总体的期望φ

枢轴量 $g(X)=\frac{\bar{x}-\mu}{S/\sqrt{n}} \sim t(n-1)$

$$W^* = \frac{\bar{X}^* - \bar{X}}{S^*/\sqrt{n}}$$

用 W* 的分布近似 g(X) 的分布，求出 W* 的近似分位数为 $W^*_{\frac{\alpha}{2}}$ 和 $W^*_{1-\frac{\alpha}{2}}$.

$$P\left\{W^*_{\frac{\alpha}{2}} < \frac{\bar{X}^*-\bar{x}}{S^*/\sqrt{n}} < W^*_{\frac{1-\alpha}{2}}\right\} = 1-\alpha$$

$$P\left\{W^*_{\frac{\alpha}{2}} < \frac{\bar{x}-\mu}{S/\sqrt{n}} < W^*_{\frac{1-\alpha}{2}}\right\} = 1-\alpha$$

$$P\left\{\bar{x} - W^*_{\frac{1-\alpha}{2}}\frac{S}{\sqrt{n}} < \phi < \bar{x} - W^*_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}\right\} = 1-\alpha$$

将 W* 的B个 bootstrap 值自小到大排序

$$W^*_{(1)} \le W^*_{(2)} \le \dots \le W^*_{(B)}$$

记 $k_1=[B\times\frac{\alpha}{2}]$，$k_2=[B\times\frac{1-\alpha}{2}]$，作为分位数自的估计，得到φ置信水平为1-α自的 bootstrap 置信区间:

$$\left(\bar{x} - W_{(k_2)}\frac{S}{\sqrt{n}},\ \bar{x} - W_{(k_1)}\frac{S}{\sqrt{n}}\right)$$

马氏链

平稳分布 $\pi(j) = \sum_{i=0}\pi(i)P_{ij}$

$\pi_i P_{ij} = \pi_j P_{ji}$

Metropolis抽样

1、输入任意状态转移矩阵Q，平稳分布$\pi(x)$，设定状态转移次数阈值n_1，需要样本个数n_2

2、从任意简单概率分布抽样得到初始状态值x_0。

3. for t=0 to n_1+n_2-1:

a> 从条件概率分布$Q(x|x_t)$中抽样得到样本x_*

b> 从均匀分布抽样 $u\sim U[0,1]$

c> 如果 $u<\alpha(x_t,x_*)=\pi(x_*)Q(x_*,x_t)$，则接受转移 $x_t\to x_*$，即 $x_{t+1}=x_*$

否则不接受转移，即 $x_{t+1}=x_t$

↓改为 $\min\left\{\frac{\pi(j)Q(j,i)}{\pi(i)Q(i,j)},1\right\}$ 即为Metropolis-Hastings抽样

样本集$(x_{n_1},x_{n_1+1},\dots x_{n_1+n_2+1})$即为我们所需平稳分布对应样本集.

二维Gibbs抽样步骤

1、输入平稳分布$\pi(x_1,x_2)$，设定状态转移次数阈值n_1，需要的样本个数为n_2

2、随机初始化初始状态值$x_1^{(0)}$和$x_2^{(0)}$

3. for t=0 to n_1+n_2-1:

a> 从条件概率分布$P(x_1|x_2^{(t)})$中抽样得到样本x_1^{t+1}

b> 从条件概率分布$P(x_2|x_1^{(t+1)})$中抽样得到样本x_2^{t+1}

样本集$\{(x_1^{(n_1)},x_2^{(n_1)}),(x_1^{(n_1+1)},x_2^{(n_1+1)}),\dots,(x_1^{(n_1+n_2+1)},x_2^{(n_1+n_2+1)})\}$即为平稳分布对应的样本集.

多维Gibbs抽样步骤

1、输入平稳分布$\pi(x_1,x_2,\dots,x_n)$或者对应的所有特征的条件概率分布，设定状态转移次数阈值n_1，需要的样本个数n_2

2、随机初始化初始状态值$(x_1^{(0)},x_2^{(0)},\dots,x_n^{(0)})$

3. for t=0 to n_1+n_2-1:

a> 从条件概率分布$P(x_1|x_2^{(t)},x_3^{(t)},\dots,x_n^{(t)})$中抽样得到样本$x_1^{t+1}$

b> 从条件概率分布$P(x_2|x_1^{(t+1)},x_3^{(t)},\dots,x_n^{(t)})$中抽样得到样本$x_2^{t+1}$

c>...

d> 从条件概率分布$P(x_j|x_1^{(t+1)},x_2^{(t+1)},\dots x_{j-1}^{(t+1)},x_{j+1}^{(t)},\dots x_n^{(t)})$中抽样得到样本$x_j^{t+1}$

e>...

f> 从条件概率分布$P(x_n|x_1^{(t+1)},x_2^{(t+1)},\dots,x_{n-1}^{(t+1)})$中抽样得到样本$x_n^{t+1}$

样本集$\{(x_1^{(n_1)},x_2^{(n_1)},\dots,x_n^{(n_1)}),\dots(x_1^{(n_1+n_2-1)},x_2^{(n_1+n_2-1)},\dots,x_n^{(n_1+n_2-1)})\}$即为平稳分布对应的样本集

最大似然估计

求能使抽样结果出现概率最大的θ

1> 写出似然函数 $L(\theta)=\prod_{i=1}^{n}P(x_i|\theta)$

2> 取对数并整理 $\sum_{i=1}^{n}\log P(x_i|\theta)=f(\theta)$

3> 求导

4> 解似然方程

EM方法

输入：观测变量数据Y，隐变量数据Z，联合分布$P(Y,Z|\theta)$，条件分布$P(Z|Y,\theta)$

输出：模型参数θ

1) 选择参数的初值$\theta^{(0)}$，开始迭代

2) E步：记$\theta^{(i)}$为第i次迭代参数的估计值，在第$i+1$次迭代的E步，计算：

$$Q(\theta,\theta^{(i)}) = E_Z[\log P(Y,Z|\theta)|Y,\theta^{(i)}]$$
$$= \sum_Z \log P(Y,Z|\theta) P(Z|Y,\theta^{(i)})$$

3) M步：求使$Q(\theta,\theta^{(i)})$极大化的θ，确定第$i+1$次迭代的参数估计值$\theta^{(i+1)}$

4) 停止迭代条件：

$$\|\theta^{(i+1)}-\theta^{(i)}\|<\varepsilon_1 \text{ 或 } \|Q(\theta^{(i+1)},\theta^{(i)})-Q(\theta^{(i)},\theta^{(i)})\|<\varepsilon_2$$

Q函数：

完全数据的对数似然函数$\log P(Y,Z|\theta)$关于在给定观测数据Y和当前参数$\theta^{(i)}$下，对未观测数据Z的条件概率分布的期望.