第九次作业

PB18000029 舒文炫

目录

8.4 题

(1)

证明: 由 BIC 的定义

$$BIC_j = -2l_j(\hat{p}_j) + d_j lnn$$

其中 l_j 为第 j 个模型下的对数似然，$\hat{p}_j$ 为第 j 个模型下的极大似然估计，d_j 为参数的维数，这里维数为 1.

这里 p 分布的先验为 $Be(\alpha,\beta)$, 在给定 p 的条件下, 其中 $t = \sum_i x_i/n$

$$f(x|p) = \binom{n}{t} p^t (1-p)^{n-t}$$
$$lnf(x|p) = ln\binom{n}{t} + tlnp + (n-t)ln(1-p)$$

对上面的对数似然函数求导

$$\frac{\partial lnf}{\partial p} = \frac{t}{p} - \frac{n-t}{1-p}$$

从而 p 的极大似然估计为 $\hat{p}_{MLE} = \frac{t}{n}$ 将其代入对数似然函数

$$l(x|\hat{p}_{MLE}) = ln\binom{n}{t} + tln\hat{p}_{MLE} + (n-t)(1-\hat{p}_{MLE})$$

从而有

$$-\frac{1}{2}BIC = ln\binom{n}{t} + tln\hat{p}_{MLE} + (n-t)(1-\hat{p}_{MLE}) - \frac{1}{2}lnn$$

证毕

(2)

令 $\alpha = 2, beta = 4, n = 10$, 则有

$$p(x_n) = \binom{n}{t}\frac{5!}{1!3!}\frac{(\alpha+t-1)!(n+\beta-t-1)!}{(n+\alpha+\beta-1)!}$$

下面用 R 代码来计算

```
set.seed(1)## 固定随机数，好说结果
sum1<-0
n<-10000
ru<-runif(n)

## 先产生 10000 个 (0,1) 随机数
for (i in c(1:n)){## 这里我需要的其实只是 xn 的和，这里就用 sum1 表示了
  pi<-rbeta(1,2,4)## 每次抽取时先从 beta(2,4) 中产生一个新的 p，以这个 p 为伯努利分布的概率
  if(ru[i]<pi){
    sum1<-sum1+1
  }
}
print(sum1)
```

```
## [1] 3348
```

```
## 计算 P(xn)
Pxn<-choose(n,sum1)*factorial(5)/factorial(3)*factorial(1+sum1)*factorial(n+3-sum1)/fac
print(Pxn)
```

```
## [1] NaN
```

```r
bic<-log(choose(n,sum1))+sum1*log(sum1/n)+(n-sum1)*log(1-sum1/n)-0.5*log(n)
ebic<-exp(bic)
print(ebic)
```

```
## [1] Inf
```

这里发现抽 10000 次样，这个样本量太大，导致算出来的这些阶乘太大，最后直接溢出了，那这里阶乘显然不能直接运算，我将其分子分母上下约去可得到这样的结果

$$p(x_n) = \frac{5!}{3!}\frac{(t+1)(n-t+1)(n-t+2)(n-t+3)}{\prod_{i=1}^{5}(n+i)}$$

然后后面 bic 的计算可将乘法换成加法，因为有这个取 ln 的运算, 这样精度也就控制住了

下面是 r 代码

```r
set.seed(1)## 固定随机数，好说结果
n<-10000

x<-c(rep(0,times=n))
pi<-rbeta(n,2,4)## 先抽出 n 个 p 备用
for (i in 1:n){## 再对每个 p，对 x 进行抽样
 x[i]<-rbinom(1,1,pi[i])
}
t<-sum(x)
Pxn<-factorial(5)/factorial(3)*(t+1)/(n+1)*(n-t+1)/(n+2)*(n-t+2)/(n+3)*(n-t+3)/(n+4)/(n
print(Pxn)
```

```
## [1] 0.0001983287
```

```r
sum1<-0##sum1 用来保存前面阶乘取对数的结果
for(j in 1:t){
  sum1<-sum1+log(n-t+j)-log(j)
}
bic<-sum1+t*log(t/n)+(n-t)*log(1-t/n)-0.5*log(n)
```

```
ebic<-exp(bic)
print(ebic)
```

```
## [1] 8.48144e-05
```

```
print(abs(Pxn-ebic))
```

```
## [1] 0.0001135143
```

可以看到这里的误差值为 0.0001135143, 几乎可以忽略不计了，说明这个量在样本量足够大时，是一个很好的逼近

8.5 题

(1)

要精确计算贝叶斯因子 BF_{01}, 需要算出样本在模型 M_0, M_1 下的边际密度 $P(x_n|M_0), P(x_n|M_1)$ x_n 服从分布 $N(\mu, 1)$

$M_0 : \mu$ 服从分布 $N(1,1)$ 则边际密度计算如下

$$
\begin{aligned}
P(X_n|M_0) &= \int_{-\infty}^{+\infty} f(X_n|\mu)\pi(\mu|M_0)d\mu \\
&= \int_{-\infty}^{+\infty} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{(\mu-1)^2}{2}} d\mu \\
&= \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{n+1}} e^{-\frac{-\frac{(n\bar{x}+1)^2}{n+1}+\sum_{i=1}^{n} x_i^2+1}{2}}
\end{aligned}
$$

$M_1 : \mu$ 服从分布 $U(-1,1)$ 则边际密度计算如下

$$
\begin{aligned}
P(X_n|M_1) &= \int_{-\infty}^{+\infty} f(X_n|\mu)\pi(\mu|M_1)d\mu \\
&= \int_{-1}^{1} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2}} \frac{1}{2} d\mu \\
&= \frac{1}{(2\pi)^{\frac{n}{2}} 2} \int_{-1}^{1} e^{-\frac{n(\mu-\bar{x})^2+\sum_{i=1}^{n} x_i^2-n\bar{x}^2}{2}} d\mu
\end{aligned}
$$

下面使用 R 代码进行精确计算, 这边两个式子里面的 π 项可以约掉

```r
set.seed(2)## 先固定一组随机数
n<-30
xn<-rnorm(n,1,1)## 从正态分布 N(1,1) 中抽样
meanx<-mean(xn)
vsumx<-sum(xn^2)
P0<-exp(-(-(n*meanx+1)^2/(n+1)+vsumx+1)/2)/sqrt(n+1)
myfunc<-function(x){
  return(exp(-(n*(x-meanx)^2+vsumx-n*meanx^2)/2))
}
P1<-integrate(myfunc,-1,1)$value/2
BF01=P0/P1
print(BF01)
```

```
## [1] 7.27497
```

这表明模型 0 为模型 1 可能性的 7.27 倍，根据前面贝叶斯因子的解释，在该样本下有较强的证据支持模型 0

(2)

重要性抽样方法

对 M_0 考虑重要性抽样密度为 $N(1,1)$, 对 M_1 考虑重要性抽样密度为 $U(-1,1)$, 这样的话，这两者的 $w_r(M_i)$ 的形式都是一样的约去常数都为 $e^{-\frac{n(\mu-\bar{x})^2+\sum_{i=1}^n x_i^2-n\bar{x}^2}{2}}$, 这样计算起来就很方便计算结果用 R 代码表示

```r
set.seed(2)
n<-30## 样本个数
nsample<-1000## 重要性抽样次数
xn<-rnorm(n,1,1)
meanx<-mean(xn)
vsumx<-sum(xn^2)
myfunc<-function(x){
  return(exp(-(n*(x-meanx)^2+vsumx-n*meanx^2)/2))
```

```
}
q1<-rnorm(nsample,1,1)
q2<-runif(nsample,-1,1)
impBF01<-sum(myfunc(q1))/sum(myfunc(q2))
print(impBF01)
```

```
## [1] 7.319779
```

可以看到重要性抽样方法得到的结果为 7.32，与精确计算的结果相近

MCMC 抽样方法

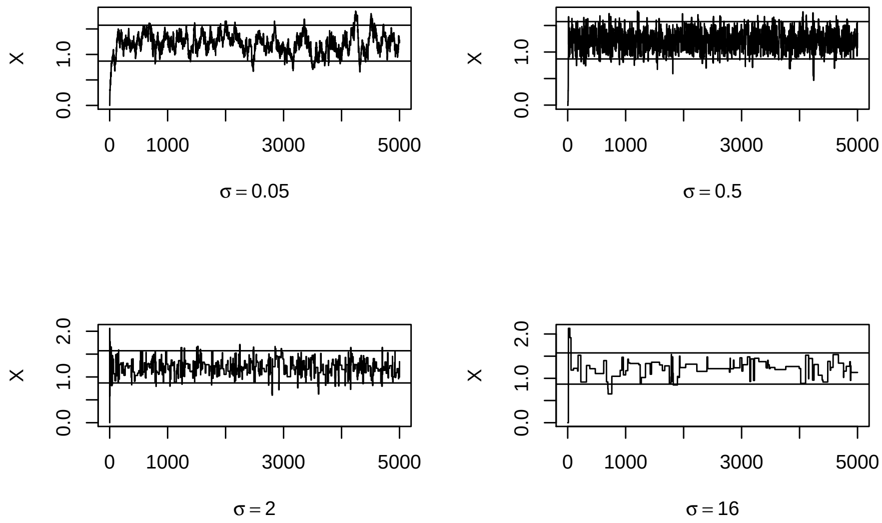
这里同样的我们考虑 $q_i = \pi_i$

对 M_0 进行 MCMC 抽样, 这里取对称的提议分布 $N(X_t, \sigma^2)$, 后验分布服从正态分布 $N(\frac{n\bar{x}+1}{n+1}, \frac{1}{n+1})$

```
set.seed(2)
##sigma 为提议分布标准差
##x0 为初始值
##N 为链跑的次数
rwMetro<-function(n,barx,sigmax,sigma,x0,N){

  x<-numeric(N)
  x[1]<-x0
  u<-runif(N)
  k<-0
  for(i in 2:N){
    y<-rnorm(1,x[i-1],sigma)
      if(u[i]<=exp(-((y-(n*barx+1)/(n+1))^2/2)*sigmax)/exp(-((x[i-1]-(n*barx+1)/(n+1))^
        x[i]<-y
      else{
        x[i]=x[i-1]
        k=k+1
      }
  }
```

```
  return(list(x=x,k=k))
}
n<-30
xn<-rnorm(n,1,1)
barx<-mean(xn)
sigmax<-n+1
sigma<-c(0.05,0.5,2,16)
N<-5000
x0<-0
mynorm1<-rwMetro(n,barx,sigmax,sigma[1],x0,N)
mynorm2<-rwMetro(n,barx,sigmax,sigma[2],x0,N)
mynorm3<-rwMetro(n,barx,sigmax,sigma[3],x0,N)
mynorm4<-rwMetro(n,barx,sigmax,sigma[4],x0,N)
mynorm<-cbind(mynorm1$x,mynorm2$x,mynorm3$x,mynorm4$x)
par(mfrow=c(2,2))## 方便比较四个图以 2*2 的形式排列
refline<-qnorm(c(0.025,0.975),(n*barx+1)/(n+1),1/sqrt(n+1))
for(j in 1:4){
  plot(mynorm[,j],type="l",xlab=bquote(sigma==.(round(sigma[j],3))),ylab="X",ylim=range
  abline(h=refline)
}
```




```
par(mfrow=c(1,1))
```

这里可以看到 $\sigma = 2$ 时收敛的比较好，所以我取 $sigma = 2$ 时的那条链

```
burn<-1001## 预烧期
vx<-sum(xn^2)
myfunc1<-function(x){
  return(exp(-(n*(x-barx)^2+vx-n*meanx^2)/2))
}
y<-mynorm3$x[burn:N]

m1<-sum(1/myfunc1(y))
print(1/m1)
```

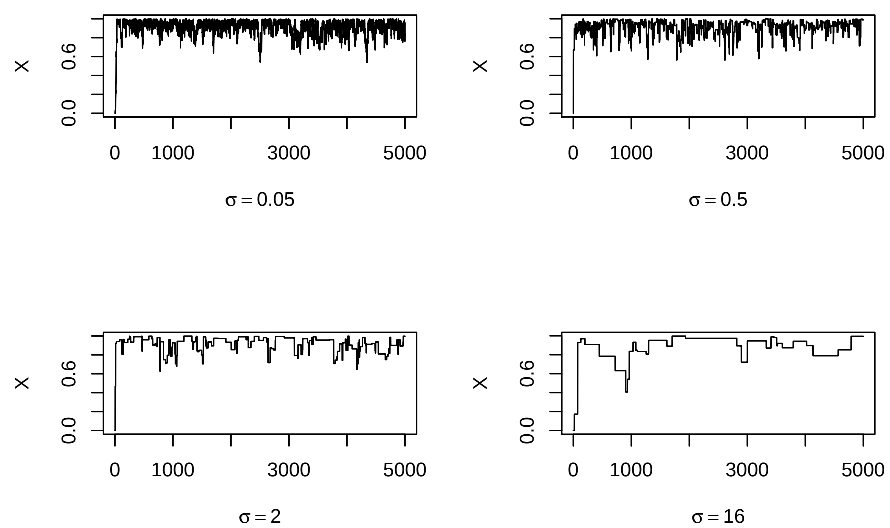
```
## [1] 1.70276e-13
```

对 M_1 进行 MCMC 抽样, 这里取对称的提议分布 $N(X_t, \sigma^2)$, 后验分布正比于 $e^{-\frac{n(\mu-\bar{x})^2}{2}} I_{-1<\mu<1}$, 所以如果提议分布中抽出了大于 1 或小于-1 的数，这个直接拒绝

```
set.seed(2)
##sigma 为提议分布标准差
##x0 为初始值
##N 为链跑的次数
rwMetro1<-function(n,barx,sigma,x0,N){

  x<-numeric(N)
  x[1]<-x0
  u<-runif(N)
  k<-0
  for(i in 2:N){
    y<-rnorm(1,x[i-1],sigma)
    if(y<=-1 || y>=1){
      x[i]=x[i-1]
      k=k+1
    }
    else{
      if(u[i]<=exp(-((y-barx)^2/2)*n)/exp(-((x[i-1]-barx)^2/2)*n))
        x[i]<-y
      else{
        x[i]=x[i-1]
        k=k+1
      }
  }}
  return(list(x=x,k=k))
}
sigma<-c(0.05,0.5,2,16)
N<-5000
x0<-0
my1norm1<-rwMetro1(n,barx,sigma[1],x0,N)
my1norm2<-rwMetro1(n,barx,sigma[2],x0,N)
my1norm3<-rwMetro1(n,barx,sigma[3],x0,N)
my1norm4<-rwMetro1(n,barx,sigma[4],x0,N)
```

```
my1norm<-cbind(my1norm1$x,my1norm2$x,my1norm3$x,my1norm4$x)
par(mfrow=c(2,2))## 方便比较四个图以 2*2 的形式排列
for(j in 1:4){
  plot(my1norm[,j],type="l",xlab=bquote(sigma==.(round(sigma[j],3))),ylab="X",ylim=rang
}
```




```
par(mfrow=c(1,1))
```

选取 $\sigma = 0.05$ 的链

```
burn<-1001## 预烧期

y<-my1norm1$x[burn:N]
m2<-sum(1/myfunc1(y))
print(1/m2)
```

```
## [1] 4.108017e-14
```

```
print(m2/m1)
```

```
## [1] 4.144968
```

最后我们给出结果 $BF_{01} = 4.14$，说明在该样本下，有较强的证据支持模型 M_0