Rで学ぶベイズ統計学入門

5章　ベイズ計算入門

5.1　はじめに

5.2　積分を計算する

5.3　Rで問題を設定する

5.4　過分散に対するベータ・二項モデル

5.5　事後モードに基づく近似

茨城大学工学部情報工学科

菊池　裕紀

5.1　はじめに

ベイズ計算問題を取り上げ、より洗練された計算方法を説明する。

・モード周辺での事後分布の挙動に基づく方法
・要約の計算にシミュレーションを利用する方法
・モンテカルロ法
・棄却サンプリング
・重点サンプリング
・SIR法

など…

5.2　積分を計算する(1)

サンプリング密度f(y|θ)からデータyを観測するとき…

θ：パラメータのベクトル

g(θ)：事後分布

すると…θの事後分布密度は次に比例する。

$$g(\theta \mid y) \propto g(\theta) f(y \mid \theta)$$

問題：θの関数を推論するため、多変量確率分布を要約する点

事後要約の多くは積分で表現できる！！

5.2　積分を計算する(2)

例えば…

関数h(θ)の事後平均は次の積分の比で表される。

$$E(h(\theta)\mid y)=\frac{\int h(\theta)g(\theta)f(y\mid\theta)d\theta}{\int g(\theta)f(y\mid\theta)d\theta}$$

ここでh(θ)がある集合Aに入る事後確率を求めたいときは…

$$P(h(\theta)\in A\mid y)=\frac{\int_{h(\theta)\in A} g(\theta)f(y\mid\theta)d\theta}{\int g(\theta)f(y\mid\theta)d\theta}$$

を計算する。

5.2　積分を計算する(3)

関心のあるパラメータの周辺密度を計算する場合

例えば…

$\theta=(\theta_1、\theta_2)$というパラメータがある。

θ_1：関心のあるパラメータ

θ_2：局外母数

θ_1の周辺事後密度は、結合事後密度から局外母数を積分消去

$$g(\theta_1 \mid y) \propto \int g(\theta_1,\theta_2 \mid y)d\theta_2$$

積分を数値的に表意化する方法に求積法が利用できる。

本章では、高次元ベイズ問題に適用可能な積分計算に焦点を…

5.3　Rで問題を設定する(1)

Rでベイズ問題を設定する方法

事後密度がパラメータを変換して表現されたとき

密度要約をする最初のステップは…

結合事後密度の対数を定義するR関数を書くこと！

- mylogposterior <- function(theta, data)
- {
- [対数密度を計算する式をここに定義する]
- return(val)
- }

theta:$\theta=(\theta_1、\ldots、\theta_k)$

data:観測値ベクトル or 事前密度の超パラメータのようなモデルを特徴づける値からなるリスト

5.3　Rで問題を設定する(2)

サンプリング密度f(y|θ)からランダム標本(y_1,…y_n)を観測
θに事前密度g(θ)を割り当てる場合…

θの事前密度対数は、 $\log g(\theta \mid y) = \log g(\theta) + \sum_{i=1}^{n} \log f(y_i \mid \theta)$

平均μで分散σの正規分布からサンプリングすると仮定…

パラメータベクトル：θ=(μ、log σ)
μ：N(10,20)の正規事前分布
log σ：平坦な事前分布　　を当てはめる。

すると、対数事後分布は以下になる。

$$\log g(\theta \mid y) = \log \phi(\mu;10,20) + \sum_{i=1}^{n} \log \phi(y_i;\mu,\sigma)$$

5.3　Rで問題を設定する(3)

ここで、$\phi(y;\mu,\sigma)$は平均μで標準偏差σの正規密度

この関数をプログラム化するために…

- logf <- function(y, mu, sigma)
- dnorm(y, mean=mu, sd=sigma, log=TRUE)

引数dataが観測値$y_1,\ldots,y_n$のベクトルとすると…

$\sum_{i=1}^{n}\log\phi(y_i;\mu,\sigma)$は関数sum()を使って計算できる。

- sum(logf(data, mu, sigma))

5.3　Rで問題を設定する(4)

対数事後密度を定義するmylogposterior()関数

```
−mylogposterior <- function(theta, data)
−{
−n <- length(data)
−mu <- theta[1]; sigma<- exp(theta[2])
−logf <- function(y, mu, sigma)
−dnorm(y, mean=10, sd=20, log=TRUE)
−val <- dnorm(mu, mean=10, ssd=20, log=TRUE) +
−          sum(logf(data, mu, sigma))
−return(val)
−}
```

5.4　過分散に対するベータ・二項モデル(1)

n_j:ある都市でのリスク男性の数
y_j：癌での死亡数

最初に考えられるモデル…
$\{y_j\}$は標本サイズ$\{n_j\}$で共通の死亡確率pの独立した二項標本

But…

過分散が認められてしまう。

頻度$\{y_j\}$の分散が、確率定数pの二項モデルのもとで予測したものよりも大きくなる。

y_jが平均ηで精度パラメータKのベータ・二項モデルに従うとするモデルを考える。

5.4　過分散に対するベータ･二項モデル(2)

$$f(y_i \mid \eta, K) = \binom{n_j}{y_j} \frac{B(K\eta + y_j, K(1-\eta) + n_j - y_j}{B(K\eta, K(1-\eta))}$$

パラメータに次に比例する曖昧事前分布を割り当てると...

$$g(\eta, K) \propto \frac{1}{\eta(1-\eta)} \frac{1}{(1+K)^2}$$

(η,K)の事後密度は、比例定数を除いて以下のように与えられる

$$g(\eta, K \mid data) \propto \frac{1}{\eta(1-\eta)} \frac{1}{(1+K)^2} \prod_{j=1}^{20} \frac{B(K\eta + y_j, K(1-\eta) + n_j - y_j)}{B(K\eta, K(1-\eta))}$$

5.4　過分散に対するベータ･二項モデル(3)

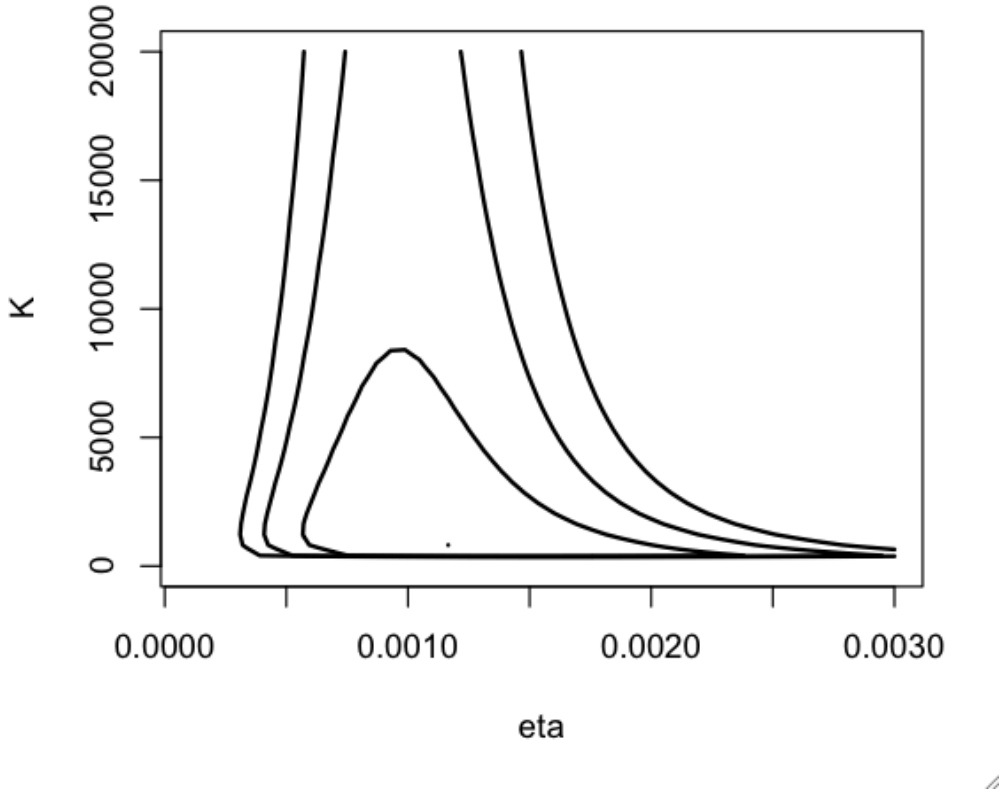
関数betabinexchを作成

```
➢betabinexch <- function(theta, data)
+ {
+ eta <- theta[1]
+ K <- theta[2]
+ y <- data[,1]
+ n <- data[,2]
+ N <- length(y)
+ logf <- function(y, n, K, eta) lbeta(K*eta+y, K*(1-eta)+n-y)
lbeta(K*eta,K*(1-eta))
+ val <- sum(logf(y,n,K,eta))
+ val <- val-2*log(1+K)-log(eta)-log(1-eta)
+ return(val)+ }
```

5.4　過分散に対するベータ･二項モデル(4)

Cancermortalityデータセットを読み込んで投稿線図を得た。

```
>data(cancermortality)
>mycontour(betabinexch,c(.0001,.003,1,20000),cancermortality,xlab="eta",ylab="K")
```



5.4　過分散に対するベータ･二項モデル(5)

それぞれのパラメータを再定式化によって実数直線上の値に変換

$$\theta_1 = \log it(\eta) = \log\left(\frac{\eta}{1-\eta}\right) \quad 、\ \theta_2 = \log(K)$$

(θ_1、θ_2)の事後密度は以下で与えられる。

ヤコビアン項

$$g_1(\theta_1,\theta_2 \mid data) = g\left(\frac{e^{\theta_1}}{1+e^{\theta_1}}\right)\frac{e^{\theta_1+\theta_2}}{(1+e^{\theta_1})^2}$$

5.4　過分散に対するベータ･二項モデル(6)

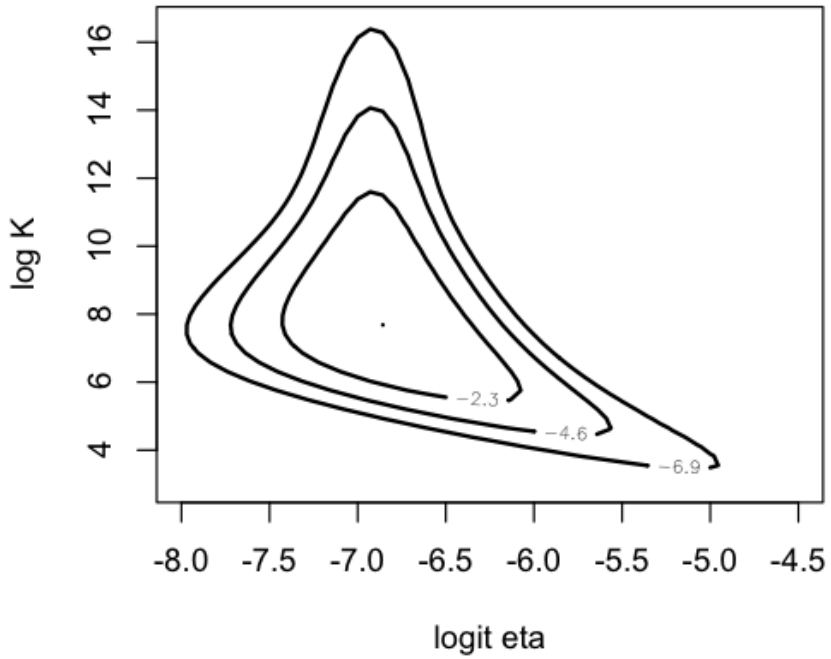
変換後の対数事後密度をbetabinexch0()にコーディングする。

```
➢betabinexch0 <- function(theta,data)
+ {
+ eta <- exp(theta[1])/(1+exp(theta[1]))
+ K <- exp(theta[2])
+ y <- data[,1]
+ n <- data[,2]
+ N <- length(y)
+ logf <- function(y,n,K,eta) lbeta(K*eta+y,K*(1-eta)+n-y)-
lbeta(K*eta,K*(1-eta))
+ val<-sum(logf(y,n,K,eta))
+ val<-val+theta[2]-2*log(1+exp(theta[2]))
+ return(val)+ }
```

5.4　過分散に対するベータ･二項モデル(7)

$(\theta_1、\theta_2)$の事後密度の等高線図を示す。

- mycontour(betabinexch0,c(-8,-4.5,3,16.5),cancermortality,
+ xlab="logit eta",ylab="log K")



事後モードに基づく近似(1)

モード周辺の密度の挙動を調べる。
–多変量事後分布の要約方法の一つ

θ：事前密度g(θ)のパラメータを表すベクトルとする。

サンプリング密度f(y|θ)のデータyを観察
θとyの密度を検討する。

$$h(\theta, y) = \log(g(\theta) f(y \mid \theta))$$

以降では、対数密度をh(θ)と示す。

事後モードに基づく近似(2)

$\hat{\theta}$:θの事後モード

対数密度を $\hat{\theta}$ 周辺で2次のテイラー級数に展開。
近似は以下で示される。

$$h(\theta) \approx h(\hat{\theta}) + (\theta - \hat{\theta})' h''(\hat{\theta})(\theta - \hat{\theta})/2$$

事後密度は、平均が $\hat{\theta}$ で分散共分散行列が以下で
与えられる多変量正規密度で近似される。

$$V = (-h''(\hat{\theta}))^{-1}$$

事後モードに基づく近似(3)

前ページでの近似では、結合密度からθを積分消去できる。

$$f(y) \approx (2\pi)^{d/2} g(\hat{\theta}) f(y \mid \hat{\theta}) \mid -h''(\hat{\theta}) \mid^{-1/2}$$

※dはθの次元である。

ニュートン法によりθの事後密度のモードを求める。
例えば…
確定値のある事後モードθ^{0}があるとする。
t-1回目の反復でのモードの推定値がθ^{t-1}とすると…

続く反復は…

$$\theta^t = \theta^{t-1} - [h''(\theta^{t-1})]^{-1} h'(\theta^{t-1})$$

事後モードに基づく近似(4)

・事後モードを求めるアルゴリズム

–ネルダー–ミード法アルゴリズム

Rの基本パッケージのoptim()関数での方法

初期値の影響を受けにくいのでニュートン法より望ましい。