

# Rで学ぶベイズ統計学入門

## 第10章 ギブスサンプリング

### 10.1 はじめに

### 10.2 頑健なモデル化

茨城大学工学部

佐々木稔

# サンプリング手法（復習）

## モンテカルロ法

棄却サンプリング

重点サンプリング

Sampling Importance  
resampling (SIR)

## マルコフ連鎖 モンテカルロ法

メトロポリスーヘイス  
ティングアルゴリズム

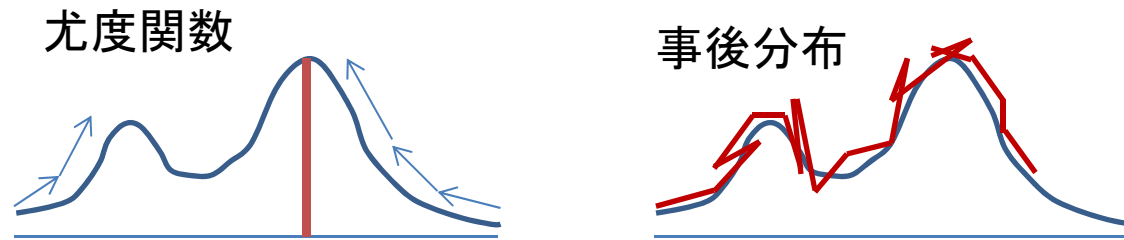
ギブスサンプリング

# はじめに

- 6.4節でのギブスサンプリング
  - 1個のパラメータを推定
    - $(p - 1)$  個のパラメータとデータを使う
- 推定するパラメータを拡張
  - 1個のパラメータ集合の各値を推定
    - $(p - 1)$  個のパラメータ集合とデータを使う
- ギブスサンプラーによる標本
  - 一般正則化条件のもとで、目標の結合事後分布  $(\theta_1, \dots, \theta_p)$  に収束

# はじめに

- ギブスサンプリング
  - すべての条件付き事後分布が計算できる



- ギブスサンプリングの性質
  - 調整のための定数や提案密度の設定が不要
  - 複雑なMCMCアルゴリズムを使うための基本
  - Rでプログラムが書きやすい

# はじめに

- ギブスサンプリングの実現方法を説明
  - tサンプリングモデル(10.2節)
  - プロビットリンクを使った2値反応モデル(10.3節)
  - 2元分割表へギブスサンプリングを適用(10.4節)

# 裾の重い分布のモデル化(1/3)

- 外れ値を持つ可能性があるデータ
  - 正規分布よりも重い裾を持つ分布
  - 自由度の小さい t 分布族
- 標本  $y_1, \dots, y_n$ 
  - 位置パラメータ  $\mu$ , 尺度パラメータ  $\sigma$ , 自由度  $\nu$  の t 分布からの標本と仮定
  - $(\mu, \sigma)$  は無情報事前分布  $g(\mu, \sigma) \propto \frac{1}{\sigma}$
  - 事後密度分布

$$g(\mu, \sigma | y) \propto \frac{1}{\sigma} \prod_{i=1}^n \frac{1}{\sigma} \left( 1 + \frac{(y_i - \mu)^2}{\sigma^2} \right)^{-(\nu+1)/2}$$

## 裾の重い分布のモデル化(2/3)

- 各観測値  $y_i$  を以下の混合分布で仮定

$$y_i | \lambda_i \approx N\left(\mu, \sigma / \sqrt{\lambda_i}\right)$$

$$\lambda_i \approx \text{gamma}(\nu/2, \nu/2)$$

$$(\mu, \sigma) \approx g(\mu, \sigma) \propto \frac{1}{\sigma}$$

- このときの結合密度

$$\frac{1}{\sigma^2} \prod_{i=1}^n \left( \frac{\lambda_i^{1/2}}{\sigma} \exp\left(-\frac{\lambda_i}{2\sigma^2} (y_i - \mu)^2\right) \right) \prod_{i=1}^n \left( \lambda_i^{\nu/2-1} \exp\left(-\frac{\nu\lambda_i}{2}\right) \right)$$

# 裾の重い分布のモデル化(3/3)

- 各パラメータの推定方法
  - ギブスサンプリングにより推定
  - 結合密度の  $i$  番目をまとめると導出可能

$$\lambda_i \approx \text{gamma}\left(\frac{\nu + 1}{2}, \frac{(y_i - \mu)^2}{2\sigma^2} + \frac{\nu}{2}\right)$$

$$\mu \approx N\left(\frac{\sum_{i=1}^n \lambda_i y_i}{\sum_{i=1}^n \lambda_i}, \frac{\sigma}{\sqrt{\sum_{i=1}^n \lambda_i}}\right)$$

$$\sigma^2 \approx \text{inverse gamma}\left(\frac{n}{2}, \frac{\sum_{i=1}^n \lambda_i (y_i - \mu)^2}{2}\right)$$

# Rによる実装

- $\lambda_i$

- ガンマ分布の乱数を生成

- ```
lam <- rgamma(n, shape=(v+1)/2, rate=v/2+(y-mu)^2/2/sig2)
```

- $\mu$

- ガウス分布の乱数を生成

- ```
mu <- rnorm(1, mean=sum(y*lam)/sum(lam), sd=sqrt(sig2/sum(lam)))
```

- $\sigma^2$

- 逆ガンマ密度分布の乱数を生成

- ```
sig2 <- rigamma(1, n/2, sum(lam*(y-mu)^2/2)
```

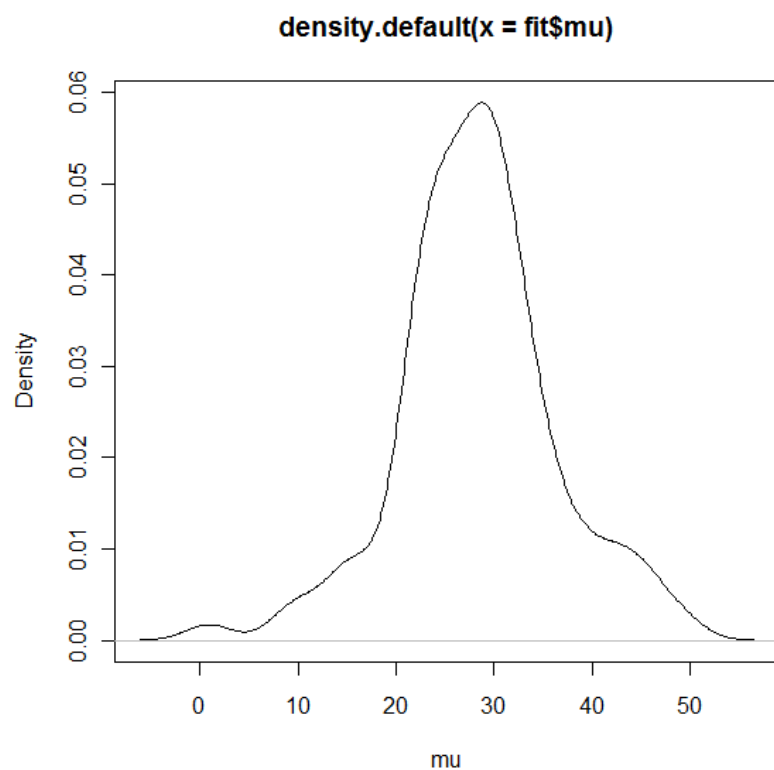
# ギブスサンプリングアルゴリズム

- `robustt()` 関数
  - ギブスサンプリングを行う関数
  - 第1引数: データベクトル
  - 第2引数: 自由度
  - 第3引数: シミュレーション回数 (100回、100万回)

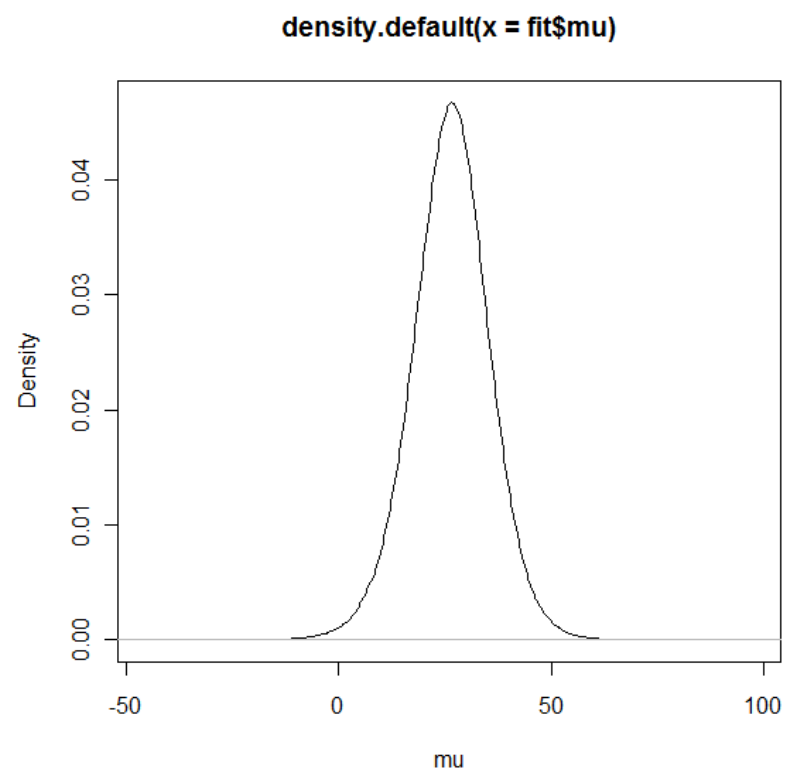
```
> data(darwin)
> attach(darwin)
> fit <- robustt(difference, 4, 100)
> plot(density(fit$mu), xlab="mu")
```

# パラメータ $\mu$ の周辺事後密度

- 回数が多いほど、正規分布に近くなる



サイクル100回



サイクル100万回

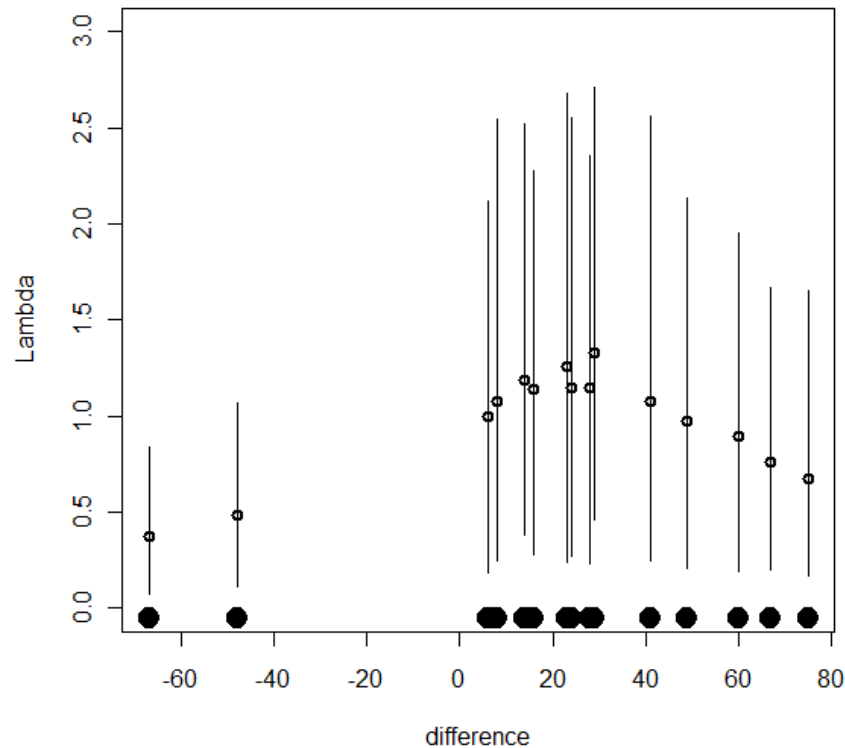
# パラメータ $\lambda$ の分布

- 各 $\lambda_i$ の事後平均を計算
- 各 $y_i$ について、 $\lambda_i$ の90%信頼区間を計算
  - 外れ値は $\lambda_i$ の値が小さくなる傾向

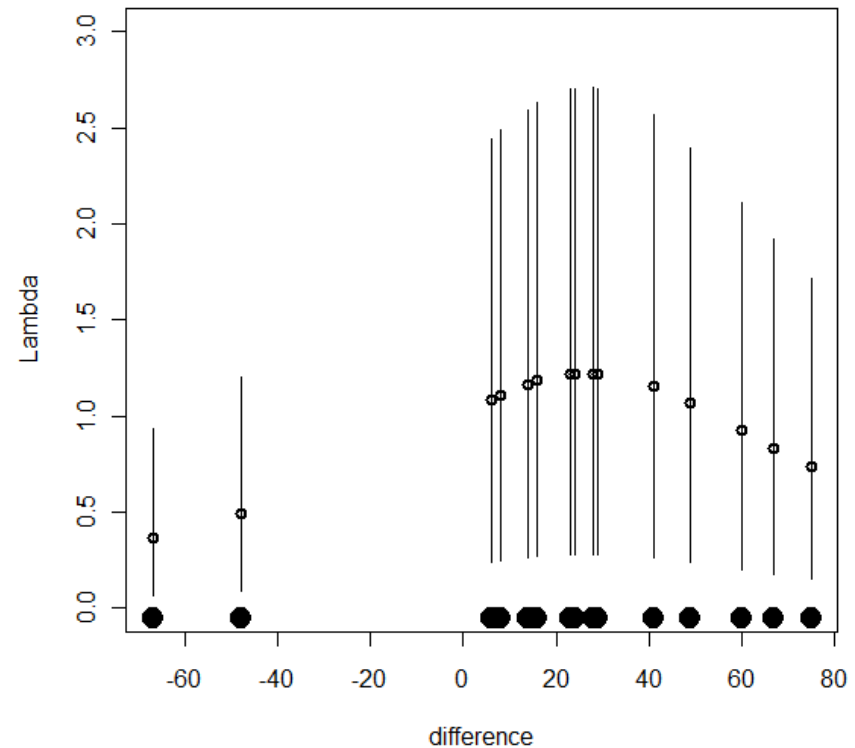
```
> mean.lambda <- apply(fit$lam,2,mean)
> lam5 <- apply(fit$lam,2,quantile,.05)
> lam95 <- apply(fit$lam,2,quantile,.95)
> plot(difference,mean.lambda,lwd=2,ylim=c(0,3),ylab="Lambda")
> for (i in 1:length(difference))
+   lines(c(1,1)*difference[i],c(lam5[i],lam95[i]))
> points(difference,0*difference-.05,pch=19,cex=2)
```

# パラメータ $\lambda_i$ の事後分布

- 100回でも外れ値の特定が可能？



サイクル100回



サイクル100万回