ベイズ統計

古谷知之

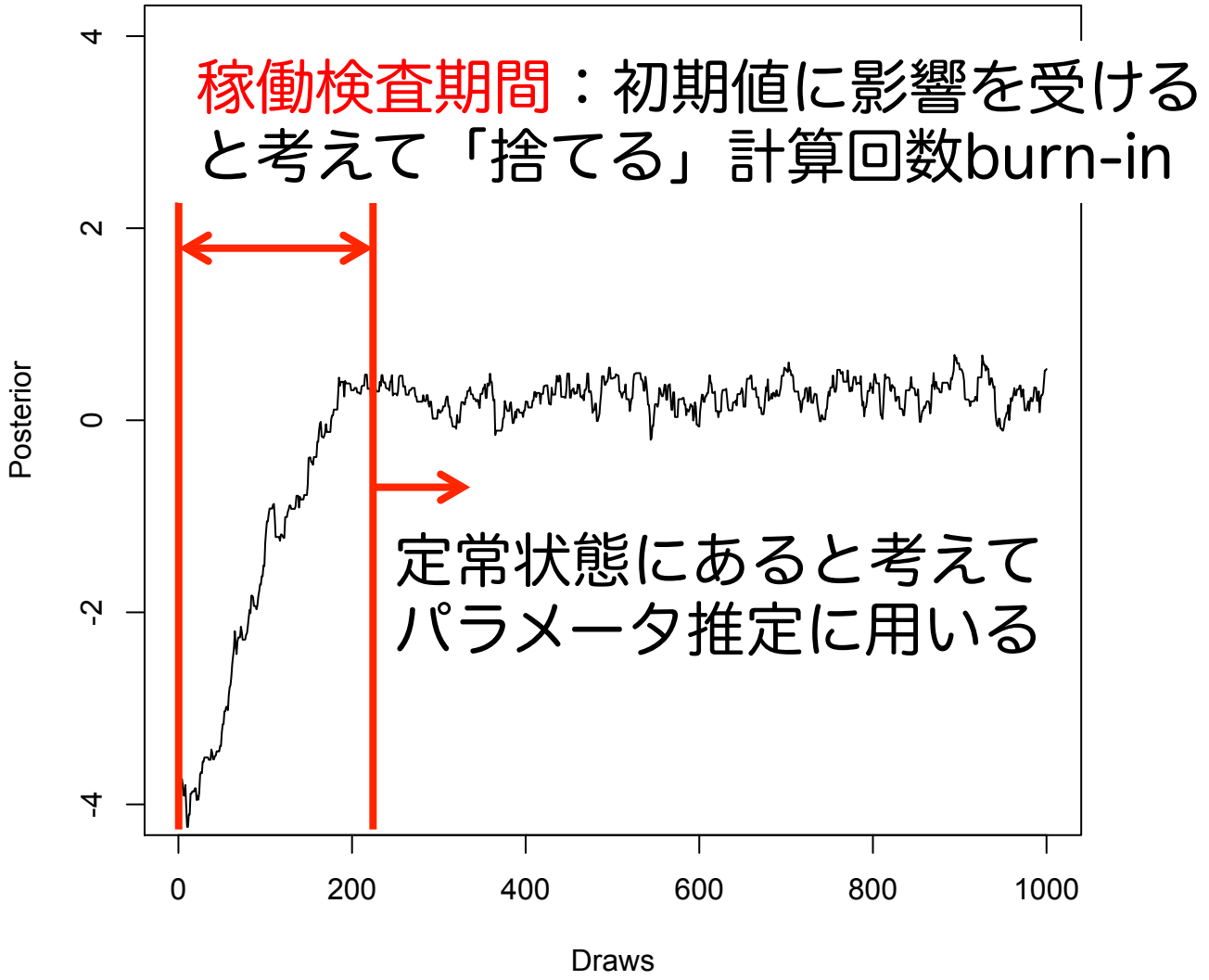
講義概要

- MCMCの結果の要約
  - 事後平均・標準偏差と標準誤差
  - 確信区間と最高事後密度
  - トレース図と自己相関図
- MCMCの収束判定法
  - Gelman & Rubinの診断方法
  - Gewekeの診断方法
  - Raftery & Lewisの診断方法

ベイズ更新の繰り返し計算

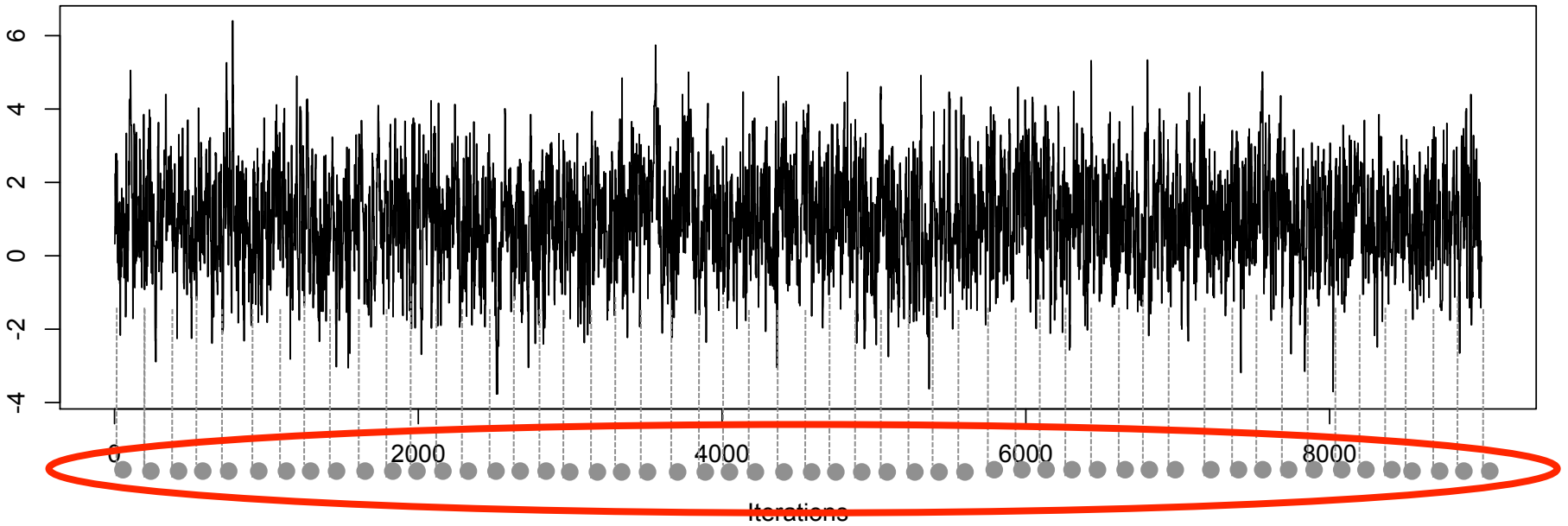
- 何回繰り返せばいいのか？→今日はココと
- 計算結果の信頼性は？→ココ
- 効率的な計算方法は？

マルコフ連鎖によるシミュレーション



Thinning interval: 間伐期間

- MCMCの標本は互いに独立していることが前提
- MCMCの標本を一定間隔ごとに抽出することで、標本の独立性が担保される(はず)
- 抽出する間隔のことをthinning intervalという



抽出したこちらを使う

MCMCで予め決めておくこと

- MCMCの回数 number of draws
  - MCMCの全期間
- Burn-inの回数 number of burn-in
  - 初期値に依存していると考えて捨てる部分
  - 使うのは # draws - # burn-in
- チェーン数 number of chains
  - 同時に走らせるMCMCの数
- 間伐期間の回数 number of thinning interval
  - MCMCの標本の独立性を担保するために標本を抽出する間隔

MCMCの結果の要約

- 事後平均（posterior mean）
- 事後標準偏差（posterior SD）
- Naive SE
- Time-series SE
- 95%確信区間
- 密度分布
- トレース図
- 自己相関

MCMCの結果の要約例

```
Iterations = 1:9000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 9000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:
```

事後平均　事後標準偏差　ナイーブSE　Time-series SE

```
          Mean             SD       Naive SE Time-series SE
      0.093799       0.175659       0.001852       0.008551

2. Quantiles for each variable:
```

95%確信区間

```
    2.5%      25%      50%      75%    97.5%
-0.24837 -0.02343  0.09072  0.21010  0.44556
```

事後平均・事後標準偏差

- 事後分布の平均と標準偏差
- MCMCの結果を要約する上で代表的な統計量
- 事後平均⇒推測統計学の点推定値とも解釈できる

Naive SEとTime-Series SE

- Naive SE

$$Naive\ SE = \frac{Posterior\ SD}{\sqrt{n_{iter} \cdot n_{chains}}}$$

- Time-series SE

$$Time\ series\ SE = \frac{Time\ series\ SD}{\sqrt{n_{iter} \cdot n_{chains}}}$$

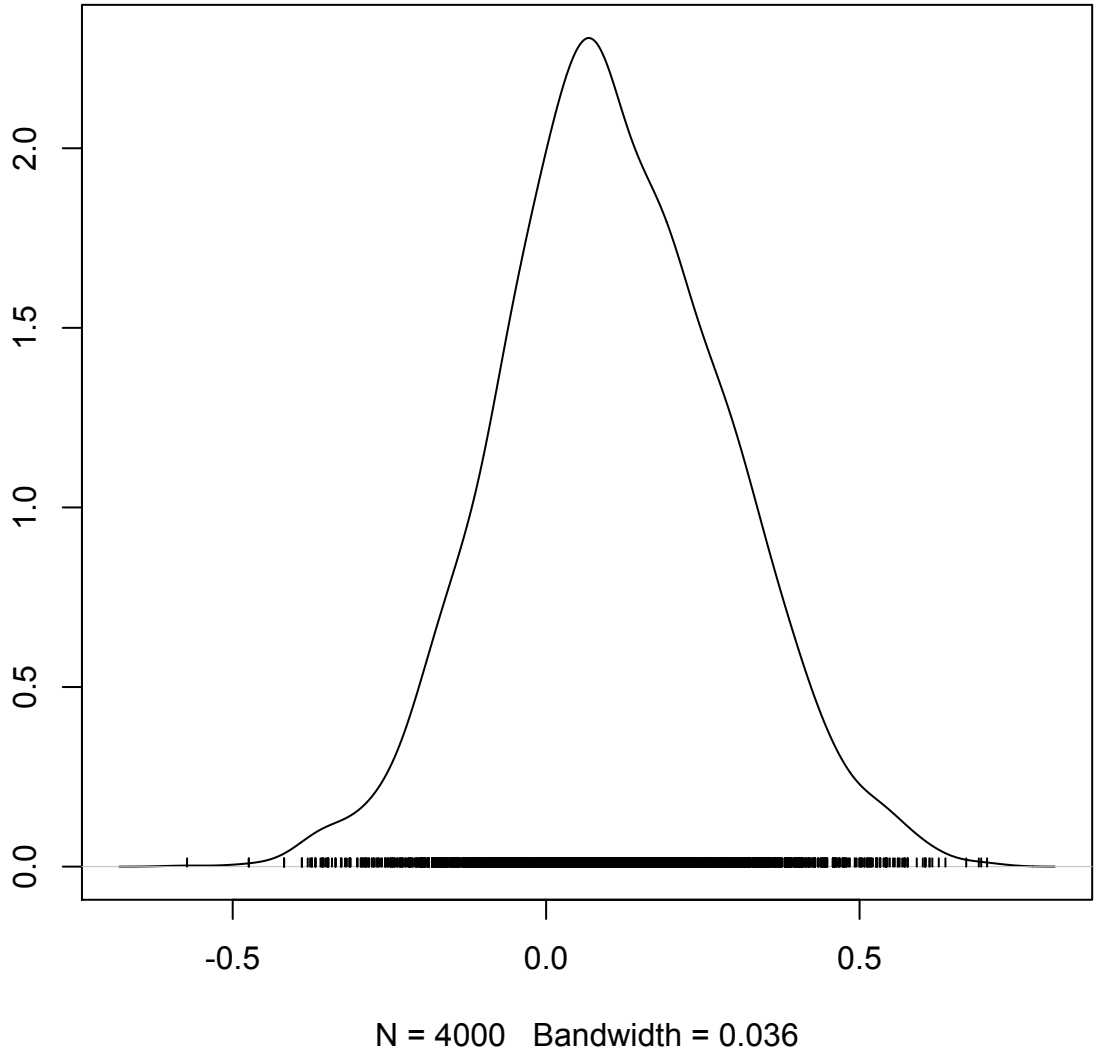
- なぜ標準誤差(SE)を用いるのか？
  - 母集団の平均の区間推定を行えるから

標準偏差と標準誤差

- 標準偏差s
  - 標本の標準偏差
- 標準誤差σ_m
  - 標本平均の標準偏差
  - $\sigma_m = s/\sqrt{n}$
- 推測統計での平均の95%信頼区間(ベイズ統計では95%確信区間)
  - 平均±1.96x標本標準偏差/√n

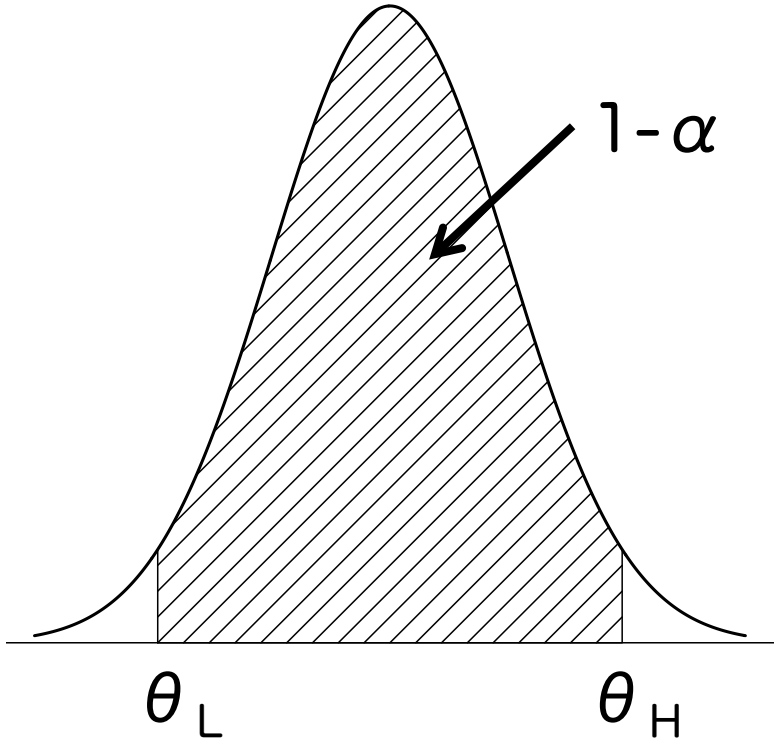
密度分布

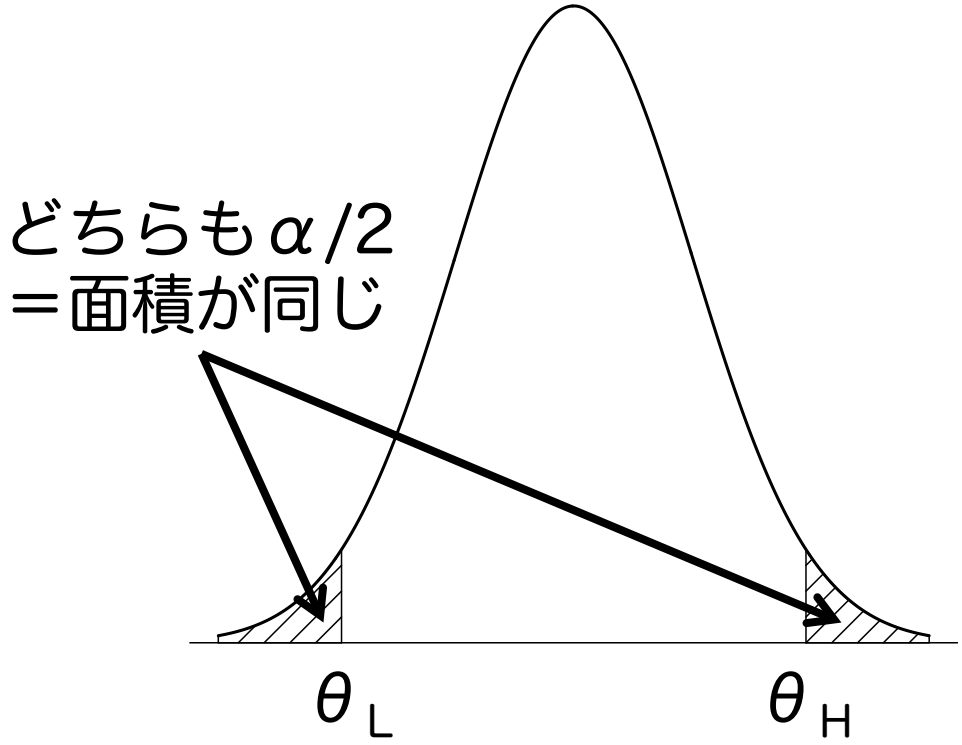
- 事後分布の密度を表したもの



95%確信区間

- 事後分布の(両側)(α=)95%確信区間
- パラメータθが95%の確率で含まれる区間 (Equal-tail interval)
- 最頻値が確信区間に含まれる場合がある



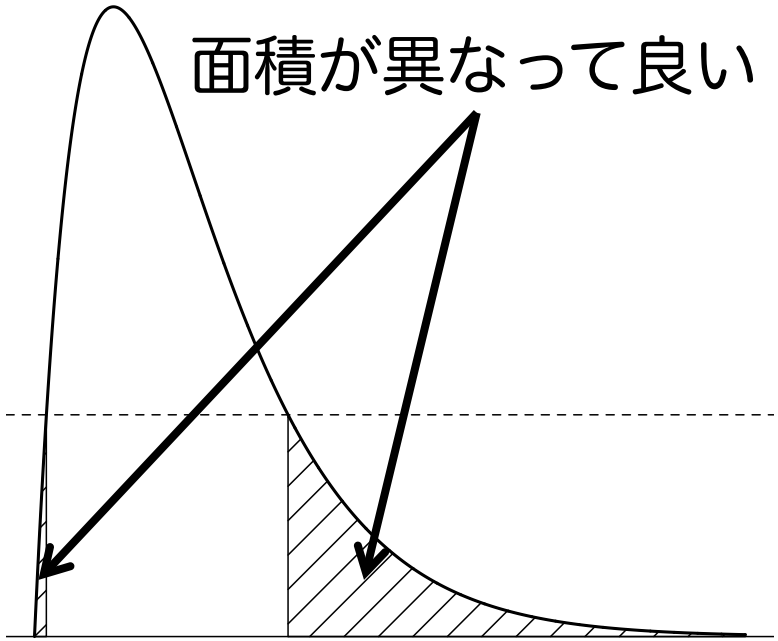


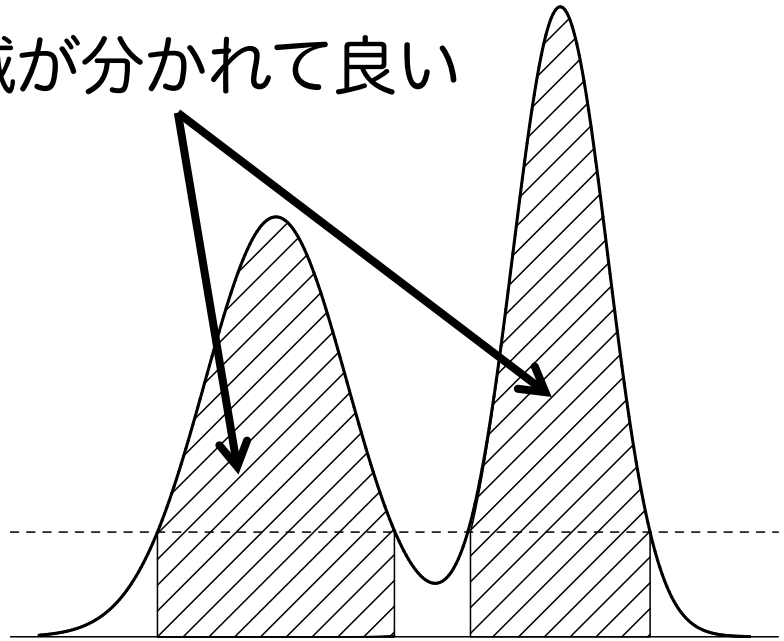
信頼区間と確信区間

- 頻度統計学の95%信頼区間
  - パラメータθを100回推定し、100回の信頼区間を得たらそのうち95回は信頼区間の中にパラメータθが含まれる
- ベイズ統計の95%確信区間
  - パラメータθが確信区間の中に含まれる確率は95%である

最高事後密度区間
Hyper Posterior Density (HPD) Interval

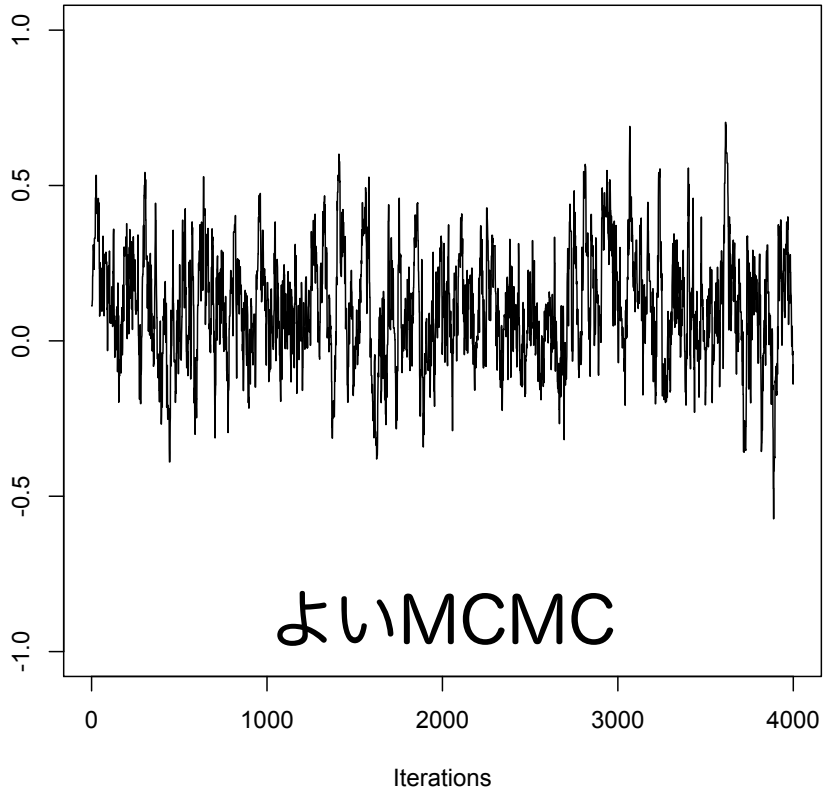
- 確信区間の面積=1-α
- 確信区間内の密度は必ず区間外の密度より高い

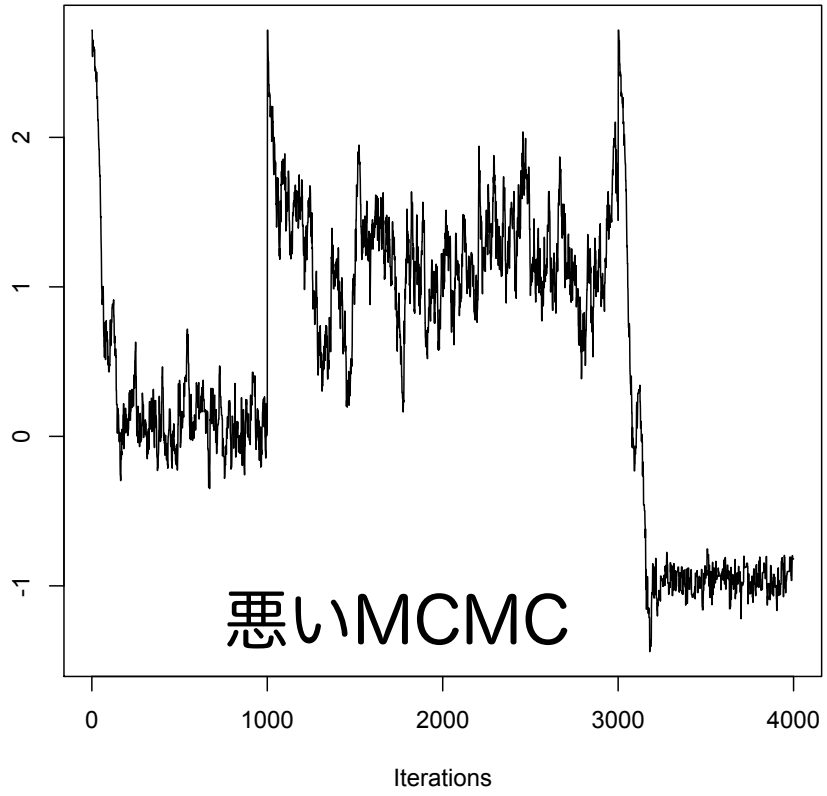




トレース図

- MCMCの乱数を順番にプロットしたもの
- トレース図をみて、burn-inの期間適切さやMCMCの質の良さを確認する





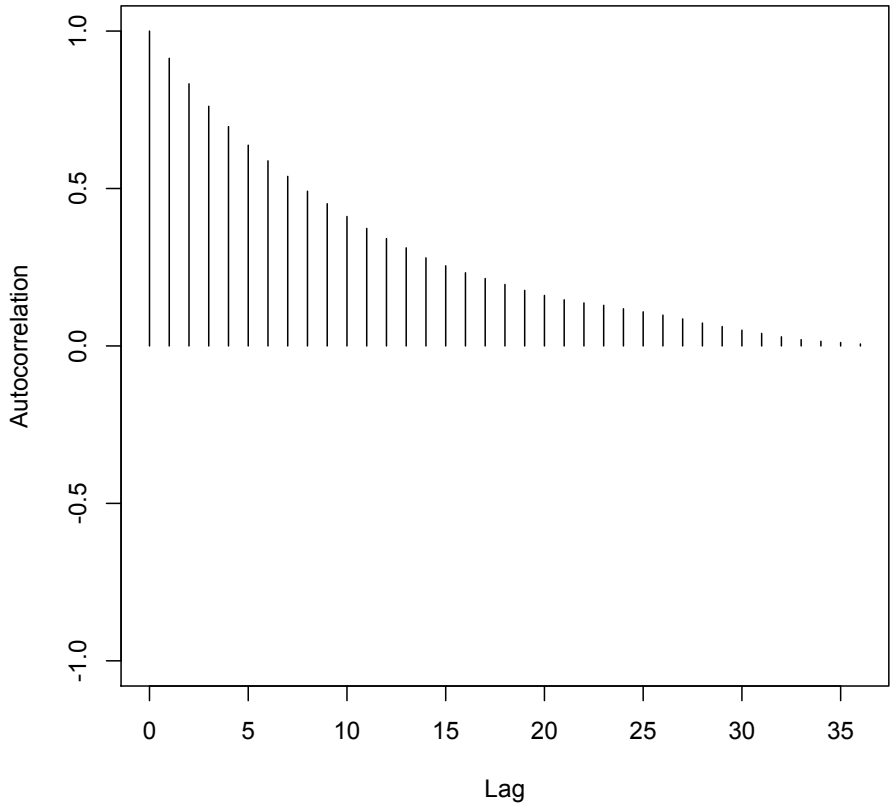
自己相関図

- MCMCを時系列データと見た場合の時系列自己相関を確認する
- 横軸にラグ、縦軸に自己相関をとる
- MCMCのサンプルは互いに独立している（と考える）ので、ラグの増加とともに自己相関がなくなるほうが良い
- ラグkでの自己相関ρ_kは、MCMCの計算回数をi(=1, …, n)とした時、次式で表される

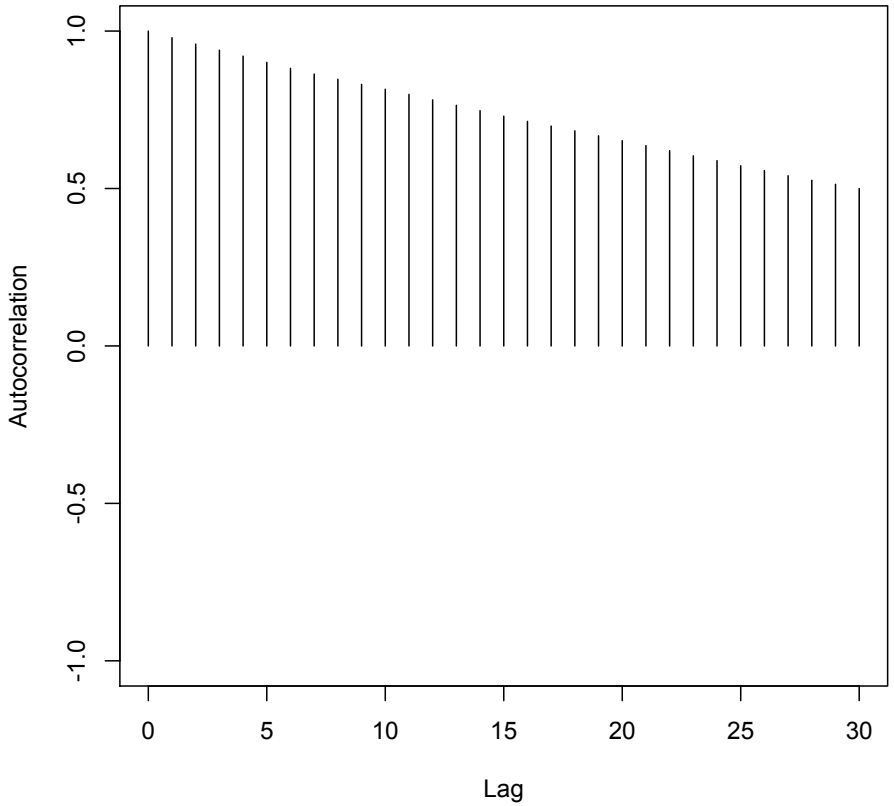
$$\rho_k = \frac{\sum_{i=1}^{n-k}\left(x_i - \overline{x}\right)\left(x_{i+k} - \overline{x}\right)}{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

自己相関図

よいMCMC



悪いMCMC



Gelman-Rubinの診断方法

- MCMCのチェーン数m(≧2)、MCMCの計算回数Sからburn-in期間bを除いたMCMC計算回数をn(=S-b)とする

1) Within-chain variance(W)とBetween-chain variance(B)を計算
2) Potential scale reduction factor (PSRF)値を計算
3) PSRF≦1.05(或いは1に近づく)であれば、MCMCは定常状態に収束したとみなすほど十分長い計算期間であったと判断する。

Gelman-Rubinの診断方法

- Within-chain variance

$$W = \frac{1}{n(m-1)} \sum_{j=1}^{m} \sum_{i=1}^{n} \left(\theta_{ij} - \bar{\theta}_j \right)^2$$

- Between-chain variance

$$B = \frac{n}{(m-1)} \sum_{j=1}^{m} \left(\bar{\theta}_j - \bar{\theta} \right)^2$$

Gelman-Rubinの診断方法

- 定常分布の分散

$$\hat{V} = \left((n-1)/n\right)W + \left(1/n\right)B$$

- Scale reduction factor (SRF)

$$R = \sqrt{\hat{V}/V^2}$$

- Potential scale reduction factor (PSRF)

$$\hat{R} = \sqrt{\hat{V}/W}$$

Gewekeの診断方法

- MCMCのサンプリング(標本)期間が定常状態にあるかどうかを検定
- MCMCの(burn-inを除く)２つのサンプリング期間について、平均値の差に関するZ統計量を計算
- ２つのサンプリング期間として、大抵は最初の10%と最後の50%を選ぶ

Gewekeの診断方法

- 最初の10% : $g_1 = \sum_{i=1}^{m} \theta_i \Big/ m,\ \left(\theta_1, \ldots, \theta_m\right)$
- 最後の50% : $g_2 = \sum_{i=n-h+1}^{n} \theta_i \Big/ h,\ \left(\theta_{n-h+1}, \ldots, \theta_n\right)$
- Z統計量

  $$Z = \left(g_1 - g_2\right) \Big/ \sqrt{V\left(g_1\right) - V\left(g_2\right)}$$

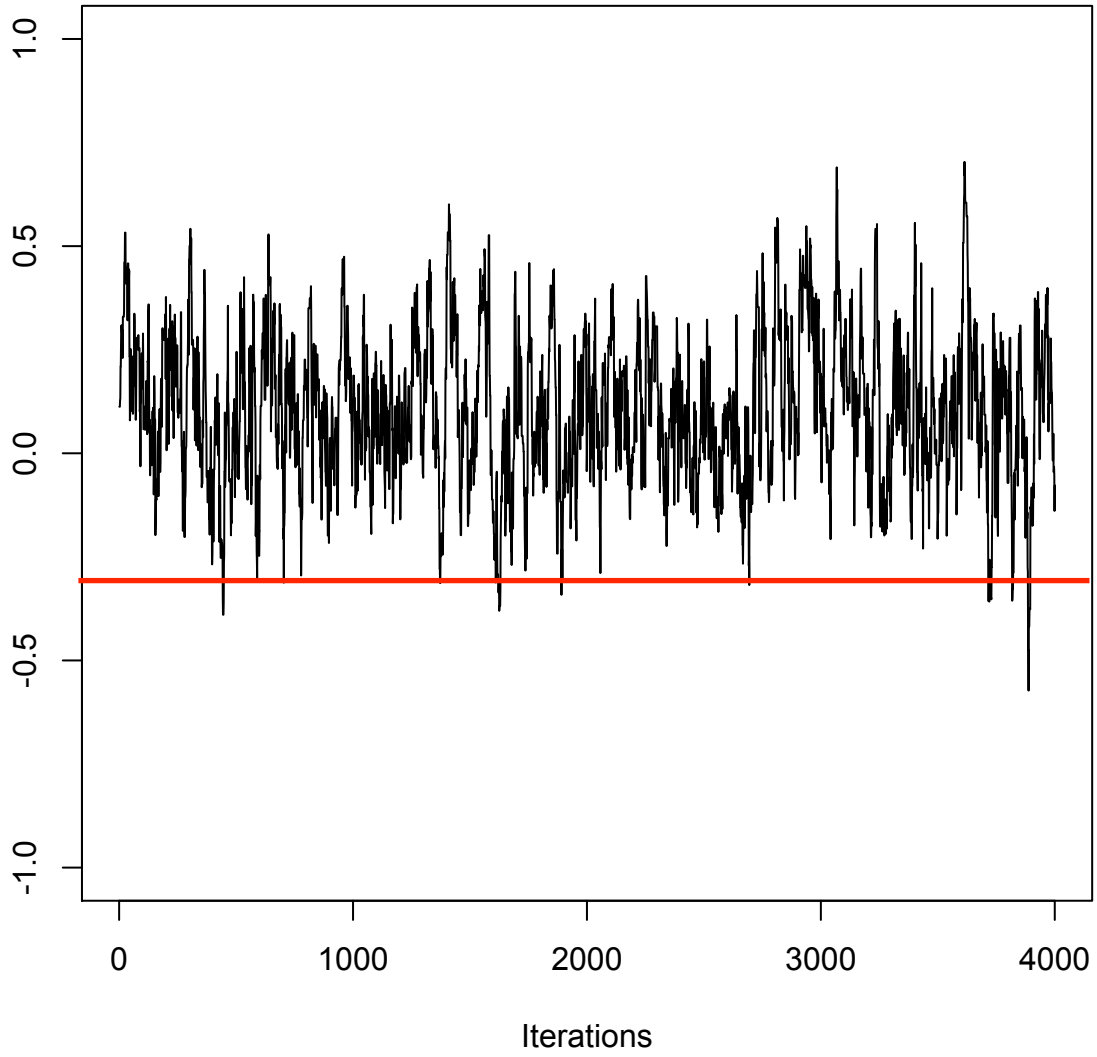
- 平均値の差の検定（ベイズなのに…）

Raftery-Lewisの診断方法

- MCMCの標本期間が適切かどうかを判断
- 事後分布のq分位(例えばq=0.025)を設定
- q分位に対する精度rを設定する(例えばr=0.005ならqの精度は0.025±0.005)
- 区間[q-r, q+r]に含まれる確率sを設定
- 次式で得られるN_{min}回のMCMCをパイロット的に計算する。Φ^{-1}は正規累積密度関数の逆関数

$$N_{\min} = \left[\Phi^{-1}\left(\frac{s+1}{2}\right) \cdot \frac{\sqrt{q(1-q)}}{r} \right]^2$$

Raftery-Lewisの診断方法



>q (=0)

q分位

≦q (=1)

Raftery-Lewisの診断方法

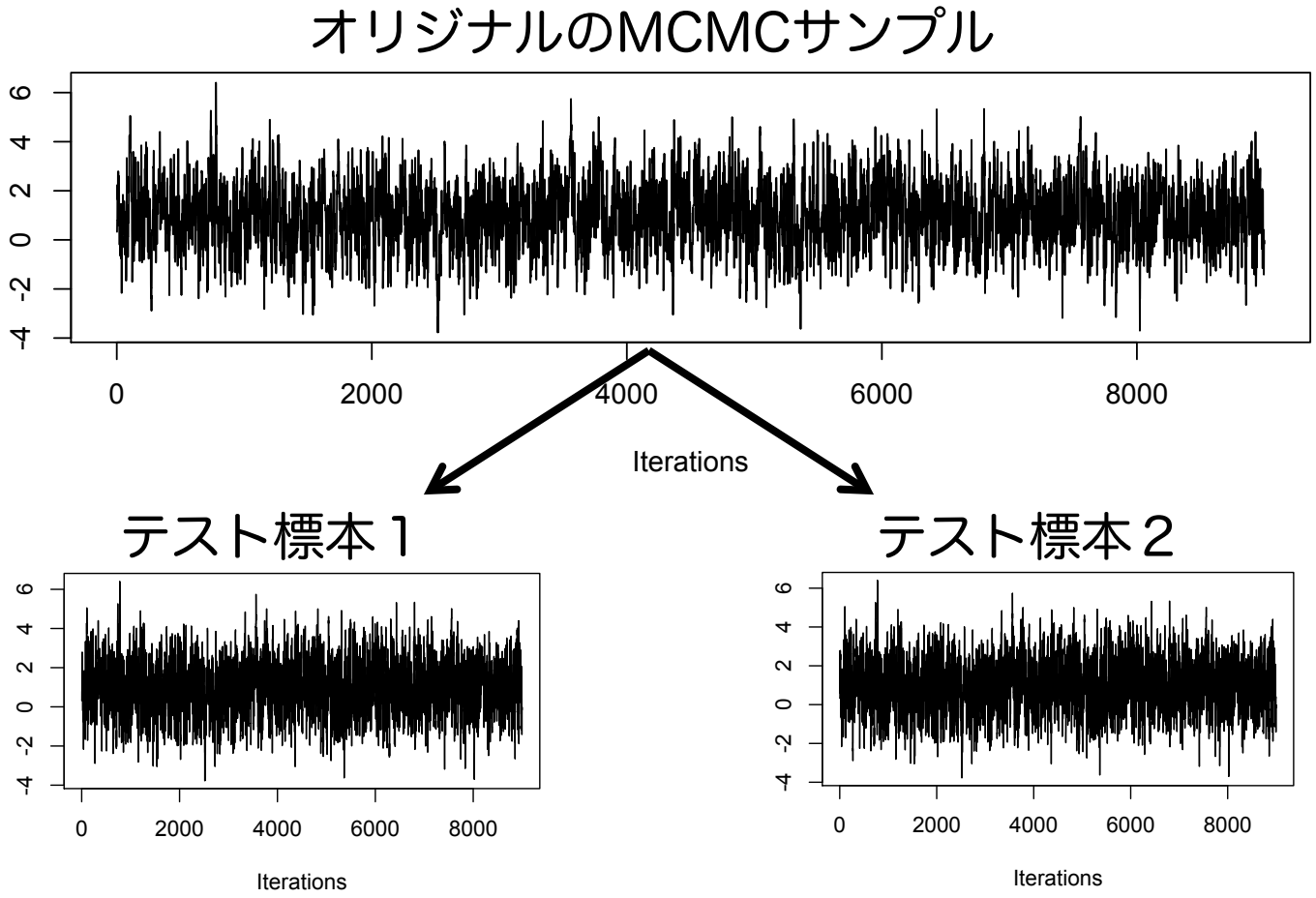
- 次式からDependence factor (I)を計算

$$I = \frac{N - M}{N_{\min}}$$

  - M: burn-in回数
  - N: burn-inを除くMCMC回数
  - M,Nともにthinning intervalを除く
- 係数Iが大きい(例えば５より大きい)とき、MCMCは初期値に依存しているか、悪いMCMCであるとかんがえられる

Raftery-Lewisの診断方法

- MCMCのサンプルは互いに独立
- MCMCを複数の標本に分けてもq分位以下が出現する標本(つまり0/1)は同じ配列で並ぶはず



Raftery-Lewisの診断方法

- 得られたサンプリングからq以上となるものを0、q以下となるものを1とする
- 0/1で置き換えられたthinning-interval=kのサンプリングをZ^kとする。Z^kの要素をz_iとする
  - $z_i=\{0,1\}$
  - $i=1, \cdots, t$

Raftery-Lewisの診断方法

- ここで本来のサンプリングから次の３つのマルコフモデルを抽出したときz_t=0or1によって以下の２つのクロス集計表を作成することができる
  - サンプル１：$z_3, \cdots, z_t$
  - サンプル２：$z_2, \cdots, z_{t-1}$
  - サンプル３：$z_1, \cdots, z_{t-2}$

	z_t=0		z_t=1	
	z_{t-1}=0	z_{t-1}=1	z_{t-1}=0	z_{t-1}=1
z_{t-2}=0	w_{000}	w_{010}	w_{001}	w_{011}
z_{t-2}=1	w_{100}	w_{110}	w_{101}	w_{111}

- ここでw_{ijl}はサンプルの個数である (i,j,lは0or1)

Raftery-Lewisの診断方法

- Thinning interval=kのマルコフ連鎖に対して以下の尤度比検定統計量を考える

$$G_k^2 = 2\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{l=0}^{1} w_{ijl} \log \frac{w_{ijl}}{\overline{w}_{ijl}}$$

- ここで、

$$\overline{w}_{ijl} = \frac{\sum_{i=0}^{1} w_{ijl} \cdot \sum_{j=0}^{1} w_{ijl}}{\sum_{i=0}^{1}\sum_{j=0}^{1} w_{ijl}}$$

Raftery-Lewisの診断方法

- すると、次式のベイズ情報量規準 (Bayesian Information Criteria: BIC) が得られる
  $$BIC = G_k^2 - 2\log\left(n_k - 2\right)$$
- ここでn_kはthinning intervalを間引いた回数

- k=1から開始してkを順に増加させ、BICが負となる最小のkが、望ましいthinning interval

Raftery-Lewisの診断方法

- kが得られたら、２つのマルコフモデルについてやはり0/1を集計し、遷移確率行列を求める
  - サンプル１：$z_2, \cdots, z_t$
  - サンプル２：$z_1, \cdots, z_{t-1}$

	$z_t=0$	$z_t=1$
$z_{t-1}=0$	$1-\alpha$	α
$z_{t-1}=1$	β	$1-\beta$

$$\alpha = \frac{\left\{z_{t-1}=0 \cap z_t=1\right\}}{\left\{z_{t-1}=0 \cap z_t=0\right\} \cup \left\{z_{t-1}=0 \cap z_t=1\right\}}$$

Raftery-Lewisの診断方法

- Burn-in期間M=M'k
- ここで、十分に小さいεに対して

$$M' = \log\left|\frac{(\alpha+\beta)\varepsilon}{\max(\alpha,\beta)}\right| \cdot \frac{1}{\log|1-\alpha-\beta|}$$

- MCMCの期間N=N'k
- ここで、

$$N' = \frac{(2-\alpha-\beta)\alpha\beta}{(\alpha+\beta)^3 r^2} \cdot \Phi^{-1}\left(\frac{s+1}{2}\right)^2$$