ベイズ統計の導入と行動モデル

2025/09/23 行動モデル夏の学校 @ 本郷

東京科学大学 中西航

nakanishi.w.aa [at] m.titech.ac.jp

はじめに

ベイズが難しいいくつかの理由

- 問題1: 「ベイズ」が多義的
  - 人や文脈によって何を指しているのかが異なる
- 問題2: 講義できちんと扱えない
  - 全体像を教える時間がない
- 問題3: 「頻度主義」「ベイズ主義」という空論
  - 50年前に終わったはずの議論がいまだに行われる
- 問題4: 理論がいまだに日進月歩
  - わかっていないことがたくさんある

いろいろなベイズ

- 歴史上の人物: **Bayes**, Thomas (1701?-1761) イギリスの牧師
- **ベイズ**の定理 by Bayes: $P(X,Y)=P(X|Y)P(Y)=P(Y|X)P(X)$
  - 内容は単なる式変形

$$P(X|Y)=\frac{P(X,Y)}{P(Y)},\quad P(Y|X)=\frac{P(X,Y)}{P(X)}$$

- **ベイズ**推定 by Laplace:
  - ベイズの定理を用いた確率的推測[推定・学習]の方法
- **ベイズ**統計学: より一般の枠組みを指すことば(?)
  - 具体的なものには
    - ベイジアンフィルタ
    - ベイジアンネットワーク
    - 状態空間，隠れマルコフ
    - …

改めて，ベイズの定理

- $P(X)$, $P(Y)\neq 0$のときXとYの同時確率$P(X,Y)$を考えると，

$$P(X,Y)=P(X|Y)P(Y)=P(Y|X)P(X)$$

$$P(X|Y)=\frac{P(X,Y)}{P(Y)},\quad P(Y|X)=\frac{P(X,Y)}{P(X)}$$

単なる式変形

- これが数百年にわたる(不毛な)議論を巻き起こした理由
  - 確率とは何か?という哲学(式の意味解釈)における対立
  - 物事が$X\rightarrow Y$の順に物事が起きるとして，

これらはOK

$$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)}$$

これは何?? Yが起きたときにXが起きる確率??

→因果の逆転を認めることになる

  - 伝統的哲学“確率は無限回試行したときの比率”に反する
  - 「データの情報を取り込む」などと表現される

ベイズの定理からベイズ推定へ

ベイズの定理を用いると
統計モデルの推定(モデル特定やパラメータ推定)が行える

使われてきたいくつかの理由:

1. 歴史的な理由
   - 最尤法の正当性提示(Fisher, 1912-)以前の有力選択肢
   - 哲学的対立による主義主張
2. 実務の理由
   - 尤度関数の最大化が困難でも利用可能 [**アルゴリズム**]
   - 事前分布に分析者の知識を反映可能 [**異質性表現**]
3. 数理科学の理由
   - 最尤法より優れるケースがある(Watanabe, 2009-)

ベイズ推定の方法

ベイズ推定(とりあえず)

- (何かはさておき)パラメータの事前分布 $\varphi(\theta)$
- (最尤推定と同じ)尤度関数 $\prod_i p(X_i|\theta)$
- (掛けて正規化すると)パラメータの事後分布 $p(\theta|X_{1,\dots,N})$

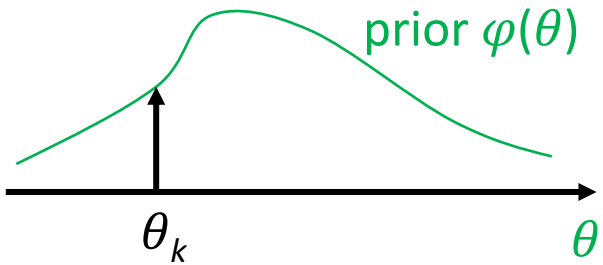
$$p(\theta|X_{1,\dots,N}) = \frac{1}{Z}\prod_i p(X_i|\theta)\varphi(\theta) \longleftrightarrow P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

- 必要ならパラメータの点推定量も求まる
  - 事後確率最大化，事後平均，…
- 当然の疑問 (一旦保留):
  - 事前分布って何??
  - 尤度関数使うなら最尤推定でいいのでは??
- とりあえず“2.実務の理由”に基づいて説明

事後分布の計算方法

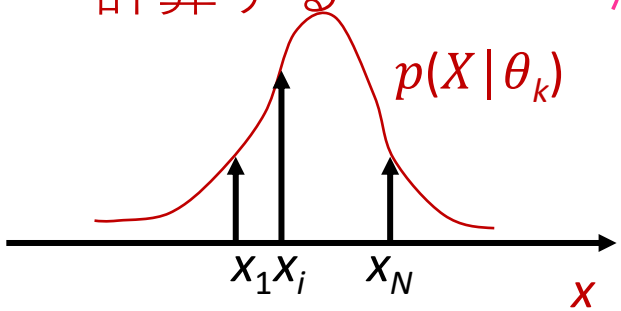
- どうやったら計算できる?? $p(\theta|X_{1,\dots,N}) = \frac{1}{Z}\prod_i p(X_i|\theta)\,\varphi(\theta)$
- 解析的に行えるケースは稀(共役事前分布: 事前分布と事後分布の関数形が同一になる特別な組み合わせ)
- 一般には数値的に近似する

1: 事前分布に従って θをひとつ選ぶ



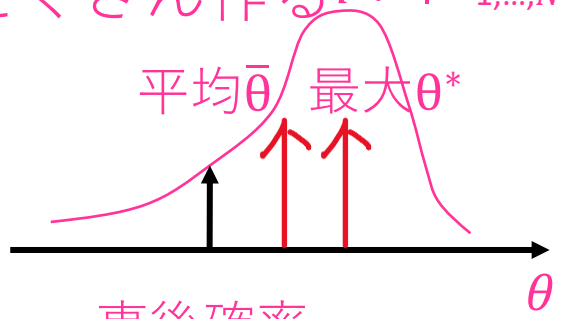
$\theta=\theta_k$が発生する確率(密度)

2. そのθで尤度を計算する



$\theta=\theta_k$のときの$X_{1,\dots,N}$の尤度

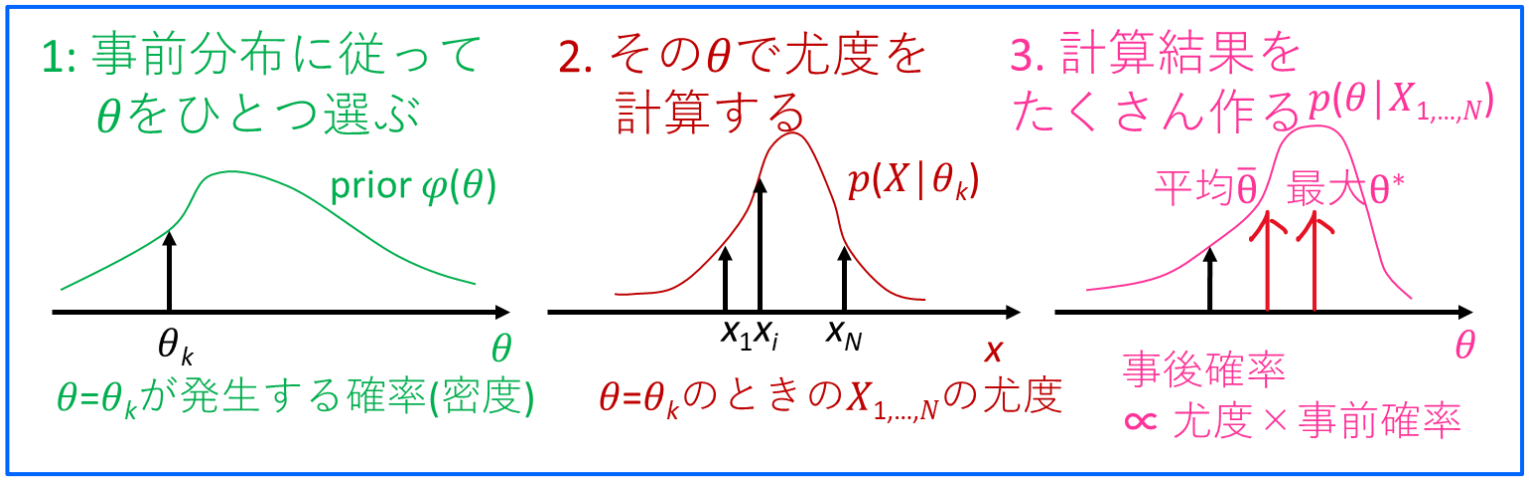
3. 計算結果をたくさん作る$p(\theta|X_{1,\dots,N})$



事後確率
$\propto$ 尤度 × 事前確率

- Zを求めるのは難しいが，Zがなくても事後分布の形はわかる
- 最大や平均もわかる

実務の理由_1: アルゴリズム



- 事後確率を単純に計算するのは難しいが，良い方法がある

$$p(\theta|X_{1,\ldots,N}) = \frac{1}{Z}\prod_i p(X_i|\theta)\,\varphi(\theta)$$

Zが不明

- マルコフ連鎖モンテカルロ(MCMC)の応用
  - ギブスサンプラー
  - メトロポリス・ヘイスティングス
  - (ほか，高速な計算方法の開発は最先端の研究課題)

実務の理由_1: アルゴリズム

- メトロポリス・ヘイスティングス

yはデータ(前ページまでのX)

10.2 メトロポリスアルゴリズム

モデル $Y \sim p(y|\theta)$ と事前分布 $p(\theta)$ がある一般的な状況を考えよう．ほとんどの問題では，y と θ の任意の値に対して $p(y|\theta)$ と $p(\theta)$ を計算できるが，$p(\theta|y) = p(\theta)p(y|\theta)/\int p(\theta')p(y|\theta')\,d\theta'$ は，分母に積分が含まれているため計算が困難なことが多い．仮に $p(\theta|y)$ からサンプリングできたとすると，$\theta^{(1)},\ldots,\theta^{(S)}$ を独立同一に $p(\theta|y)$ から生成することで

$$\mathrm{E}[g(\theta)|y] \approx \frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)})$$

のモンテカルロ近似が得られる．

$p(\theta|y)$ から直接サンプリングできない場合はどうすればよいだろうか．事後分布を近似するという意味では，重要なことは $p(\theta|y)$ からの独立同一なサンプルがあることではなく，大量の θ の集合 $\{\theta^{(1)},\ldots,\theta^{(S)}\}$ を生成できて，その経験分布が $p(\theta|y)$ を近似していることである．大雑把にいうと，任意の二つの異なる値 θ_a，θ_b に対して以下が成り立つことである．

$$\frac{\#\{\theta_a \text{ の値をとる } \theta^{(s)}\}}{\#\{\theta_b \text{ の値をとる } \theta^{(s)}\}} \approx \frac{p(\theta_a|y)}{p(\theta_b|y)}.$$

このような集合をどのように構成するか，直感的に考えてみよう．いま，$\{\theta^{(1)},\ldots,\theta^{(s)}\}$ という集合が手元にあり，新しい値 $\theta^{(s+1)}$ を追加することを考える．$\theta^{(s)}$ と近い値をとる θ^* があったときに，これを新しい値として追加すべきだろうか．もし $p(\theta^*|y) > p(\theta^{(s)}|y)$ ならば，θ^* の方が $\theta^{(s)}$ よりも事後確率が高く，すでに $\theta^{(s)}$ は集合に入っているので，θ^* も含めるべきだと考えられる．一方で，もし $p(\theta^*|y) < p(\theta^{(s)}|y)$ ならば，必ずしも θ^* を追加する必要はないため，$p(\theta^*|y)$ と $p(\theta^{(s)}|y)$ を具体的に比較することで決めるのが妥当であるだろう．具体的には

$$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y)}\frac{p(y)}{p(y|\theta^{(s)})p(\theta^{(s)})} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(s)})p(\theta^{(s)})} \quad (10.1)$$

を計算すればよい．その際に，事後分布 $p(\theta|y)$ を計算する必要がないことに注意する．次に r を計算した後にどのような決断を下せばよいか考えてみよう．

$r > 1$ の場合

直感：$\theta^{(s)}$ がすでに集合に入っているので，より高い確率をもつ θ^* を集合に加えるべきである．

手順：θ^* を受容する．すなわち $\theta^{(s+1)} = \theta^*$ として集合に加える．

$r < 1$ の場合

直感：集合内において「θ^* と値が等しい要素」の「$\theta^{(s)}$ と値が等しい要素」に対する相対的な頻度は $p(\theta^*|y)/p(\theta^{(s)}|y) = r$ となる．これは，各 $\theta^{(s)}$ に対して θ^* を部分的に採用すべきであることを意味している．

手順：確率 r および $1-r$ で $\theta^{(s+1)}$ をそれぞれ θ^* および $\theta^{(s)}$ に設定する．

これは，有名なメトロポリス (Metropolis) アルゴリズムの基本的な考え方である．メトロポリスアルゴリズムは，現在の値 $\theta^{(s)}$ から対称な提案分布 (symmetric proposal distribution) $J(\theta^*|\theta^{(s)})$ を用いて $\theta^{(s)}$ に近い値 θ^* を生成することを繰り返していく．ここでいう対称性とは，$J(\theta_b|\theta_a) = J(\theta_a|\theta_b)$，すなわち，$\theta^{(s)} = \theta_a$ のもとで $\theta^* = \theta_b$ を提案する確率は $\theta^{(s)} = \theta_b$ のもとで $\theta^* = \theta_a$ を提案する確率と等しいことを意味している．通常 $J(\theta^*|\theta^{(s)})$ は非常に単純で，$J(\theta^*|\theta^{(s)})$ からのサンプルは高確率で $\theta^{(s)}$ の近くになるものが用いられる．具体的には，以下のようなものが例として挙げられる．

- $J(\theta^*|\theta^{(s)}) = \mathrm{uniform}(\theta^{(s)} - \delta, \theta^{(s)} + \delta)$.
- $J(\theta^*|\theta^{(s)}) = \mathrm{normal}(\theta^{(s)}, \delta^2)$.

ホフ; 入江ら訳: 標準ベイズ統計学

実務の理由_1: アルゴリズム

- メトロポリス・ヘイスティングス

10.2 メトロポリスアルゴリズム

モデル $Y \sim p(y|\theta)$ と事前分布 $p(\theta)$ がある一般的な状況を考えよう．ほとんどの問題では，y と θ の任意の値に対して $p(y|\theta)$ と $p(\theta)$ を計算できるが，$p(\theta|y) = p(\theta)p(y|\theta)/\int p(\theta')p(y|\theta')\,d\theta'$ は，分母に積分が含まれているため計算が困難なことが多い．仮に $p(\theta|y)$ からサンプリングできたとすると，$\theta^{(1)}, \ldots, \theta^{(S)}$ を独立同一に $p(\theta|y)$ から生成することで

$$\mathrm{E}[g(\theta)|y] \approx \frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)})$$

のモンテカルロ近似が得られる．

$p(\theta|y)$ から直接サンプリングできない場合はどうすればよいだろうか．事後分布を近似するという意味では，重要なことは $p(\theta|y)$ からの独立同一なサンプルがあることではなく，大量の θ の集合 $\{\theta^{(1)}, \ldots, \theta^{(S)}\}$ を生成できて，その経験分布が $p(\theta|y)$ を近似していることである．大雑把にいうと，任意の二つの異なる値 θ_a，θ_b に対して以下が成り立つことである．

$$\frac{\#\{\theta_a \text{ の値をとる } \theta^{(s)}\}}{\#\{\theta_b \text{ の値をとる } \theta^{(s)}\}} \approx \frac{p(\theta_a|y)}{p(\theta_b|y)}.$$

このような集合をどのように構成するか，直感的に考えてみよう．いま，$\{\theta^{(1)}, \ldots, \theta^{(s)}\}$ という集合が手元にあり，新しい値 $\theta^{(s+1)}$ を追加することを考える．$\theta^{(s)}$ と近い値をとる θ^* があったときに，これを新しい値として追加すべきだろうか．もし $p(\theta^*|y) > p(\theta^{(s)}|y)$ ならば，θ^* の方が $\theta^{(s)}$ よりも事後確率が高く，すでに $\theta^{(s)}$ は集合に入っているので，θ^* も含めるべきだと考えられる．一方で，もし $p(\theta^*|y) < p(\theta^{(s)}|y)$ ならば，必ずしも θ^* を追加する必要はないため，$p(\theta^*|y)$ と $p(\theta^{(s)}|y)$ を具体的に比較することで決めるのが妥当であるだろう．具体的には

$$r = \frac{p(\theta^*|y)}{p(\theta^{(s)}|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(y)}\frac{p(y)}{p(y|\theta^{(s)})p(\theta^{(s)})} = \frac{p(y|\theta^*)p(\theta^*)}{p(y|\theta^{(s)})p(\theta^{(s)})} \quad (10.1)$$

を計算すればよい．その際に，事後分布 $p(\theta|y)$ を計算する必要がないことに注意する．次に r を計算した後にどのような決断を下せばよいか考えてみよう．

$r > 1$ の場合

直感：$\theta^{(s)}$ がすでに集合に入っているので，より高い確率をもつ θ^* を集合に加えるべきである．

手順：θ^* を受容する．すなわち $\theta^{(s+1)} = \theta^*$ として集合に加える．

$r < 1$ の場合

直感：集合内において「θ^* と値が等しい要素」の「$\theta^{(s)}$ と値が等しい要素」に対する $p(\theta^*|y)/p(\theta^{(s)}|y) = r$ となる．これは，各 $\theta^{(s)}$ に対して θ^* を確率的に採用すべきであることを意味している．

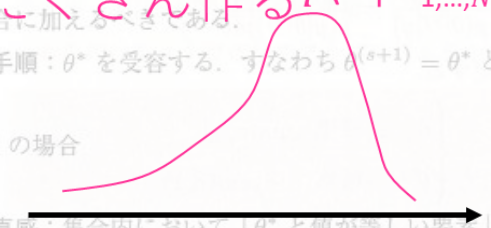
手順：確率 r および $1-r$ でそれぞれ θ^* および $\theta^{(s)}$ に設定する．

これは，有名なメトロポリス（Metropolis）アルゴリズムの基本的な考え方である．メトロポリスアルゴリズムは，現在の値 $\theta^{(s)}$ から対称な提案分布（symmetric proposal distribution）$J(\theta^*|\theta^{(s)})$ を用いて $\theta^{(s)}$ に近い値 θ^* を生成することを繰り返していく．ここでいう対称性とは，$J(\theta_b|\theta_a) = J(\theta_a|\theta_b)$，すなわち，$\theta^{(s)} = \theta_a$ のもとで $\theta^* = \theta_b$ を提案する確率は $\theta^{(s)} = \theta_b$ のもとで $\theta^* = \theta_a$ を提案する確率と等しいことを意味している．通常 $J(\theta^*|\theta^{(s)})$ は非常に単純で，$J(\theta^*|\theta^{(s)})$ からのサンプルは高確率で $\theta^{(s)}$ の近くになるものが用いられる．具体的には，以下のようなものが例として挙げられる．

- $J(\theta^*|\theta^{(s)}) = \mathrm{uniform}(\theta^{(s)} - \delta, \theta^{(s)} + \delta)$.
- $J(\theta^*|\theta^{(s)}) = \mathrm{normal}(\theta^{(s)}, \delta^2)$.

yはデータ(前ページまでのX)

3. 計算結果を
たくさん作る $p(\theta|X_{1,\ldots,N})$



事後確率
∝ 尤度×事前確率

ホフ; 入江ら訳: 標準ベイズ統計学

実務の理由_1: アルゴリズム

- 事後確率と同じ分布形状からのサンプリングになるようなルールでθの値を順番につくる
- とてもたくさん繰り返す
- すると，事後確率を(≒Zを)直接計算しなくても，事後確率を近似できる

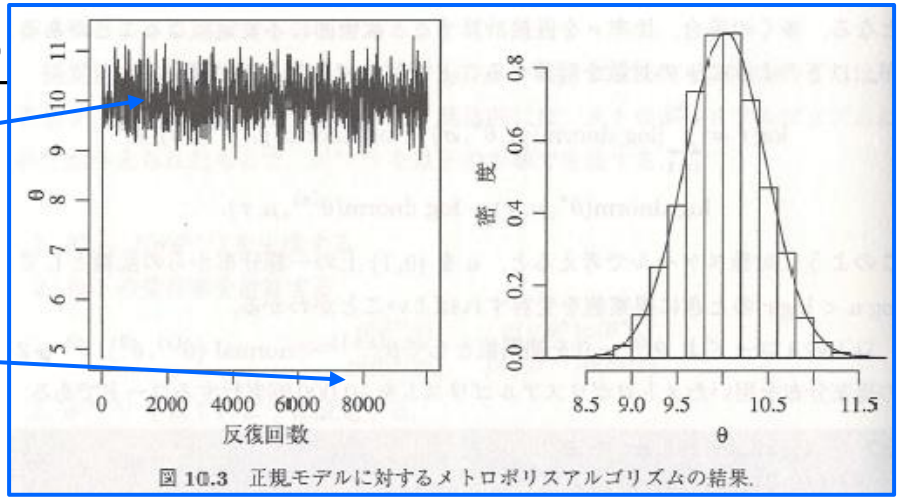


図 10.3 正規モデルに対するメトロポリスアルゴリズムの結果.

ホフ; 入江ら訳: 標準ベイズ統計学

- 仮に尤度関数の形が複雑でも，**尤度関数を直接最大化しなくても**，**Zを求めなくても**，たくさん計算すれば関数の形がだいたい求まる

よく見かけるベイズ推定の顔:

アルゴリズムとしてのベイズ推定

実務の理由_2: 異質性表現

- 階層ベイズ あるいは ハイパーパラメータ
- サッカー選手の(真の)パス成功率を、ある試合で観測したパス本数・パス試行数から推定したい
  - その日の好不調や対戦相手に依存する
  - フォワードの選手は良い選手でもパス成功率は少ない
- 各選手の真のパス成功率q_i，パス試行本数Nのとき，パス成功本数kとなる確率 ${}_N\mathrm{C}_k\{q\}^k\{1-q\}^{N-k}$
  - q=0.9，N=30のときk=27は24%，k=24は5%
  - たとえばk=24ならqの最尤推定量は0.8だが，何人もいたら真のqと遠い観測量はたくさん得られるはず

実務の理由_2: 異質性表現

- 各選手のk_i/N_iだけからq_iをもう少しよく推定したい
  - 事前分布$\varphi(q)$: 試合に出てる選手はある程度上手い**と思う**ので，たとえば$\varphi(q)\sim N(0.9,0.01)$などとする
  - もう少しきちんと$q_i=a_i+b_i$ などを考えることもできる
  - aやbが従う分布を考える: たとえば,
    - a_iは全選手の平均(共通の定数)とする
    - b_iは正規分布$N(0,\sigma^2)$に従うとする
  - σの事前分布を考える
    →階層化/ハイパーパラメータと呼ばれるもの

よく見かけるベイズ推定の顔:

異質性の表現としてのベイズ推定

保留された疑問

- 当然の疑問:
  - 事前分布って何??
  - 尤度関数使うなら最尤推定でいいのでは??
- いろいろな回答方法
  1. 歴史的理由(哲学に依拠)
     - 現代では無意味．“事前確率なるものが存在する”という主張は単なる数理的な仮定
  2. 実務の理由(便利なものは使えば良い)
     - 直前で説明したこと．意味はあるが，諸刃の剣
  3. 数理科学の理由(ベイズ推定にしか無い良い性質)
     - 明確・強力・正当だが，初見では難しい
     - 昨年の中西の講義資料参照

(行動)モデルのベイズ推定

行動モデルにおけるベイズ推定

- アルゴリズムとしてのベイズ推定が発端
  - Probit等のopen formの推定を行う際の手段
    = 尤度関数の最大化が難しい場合
- たとえばTrain本の12章はほとんどこの話

Train: Discrete Choice Methods with Simulation

行動モデルにおけるベイズ推定

- アルゴリズムとしてのベイズ推定が発端
  - Probit等のopen formの推定を行う際の手段
    = 尤度関数の最大化が難しい場合
- Train本, 12.1

12.1 Introduction

A powerful set of procedures for estimating discrete choice models has been developed within the Bayesian tradition. The breakthough concepts were introduced by Albert and Chib (1993) and McCulloch and Rossi (1994) in the context of probit, and by Allenby and Lenk (1994) and Allenby (1997) for mixed logits with normally distributed coefficients. These authors showed how the parameters of the model can be estimated without needing to calculate the choice probabilities. Their procedures provide an alternative to the classical estimation methods described in Cha 7), and

- 尤度の計算には選択確率の計算が必要だが，open formでは難しい
- ベイズ推定でこれを回避

行動モデルにおけるベイズ推定

- 異質性表現としてのベイズ推定の話もある
- Train本, 12.1, つづき

described in Chapter 10. Rossi *et al.* (1996), Allenby (1997), and Allenby and Rossi (1999) showed how the procedures can also be used to obtain information on individual-level parameters within a model with random taste variation. By this means, they provide a Bayesian analog to the classical [illegible] of these procedu[illegible]en numerous. For[illegible]git procedure to ta[illegible]nd choice in each purchase occasion. Bradlow and Fader (2001) showed how similar methods can be used to examine rankings data at an aggregate level rather than choice data at the individual level. Chib and Greenberg (1998) and Wang *et al.* (2002) developed methods for interrelated discrete responses. Chiang *et al.* (1999) examined situations where the choice set that the decision maker considers is unknown to the researcher. Train (2001) extended the Bayesian procedure for mixed logit to nonnormal distributions of coefficients, including lognormal, uniform, and triangular distributions.

- 個人レベルのtaste variationの情報を得ることにも使える

補足: これらは実務の理由なので…

> Two important notes are required before proceeding. First, the Bayesian procedures, and the term "hierarchical Bayes" that is often used in the context of discrete choice models, refer to an estimation method, not a behavioral model. Probit, mixed logit, or any other model that the researcher specifies can, in principle, be estimated by either classical or Bayesian procedures. Second, the Bayesian perspective from which these procedures arise provides a rich and intellectually satisfying

- 「ここでいうBayesとは，行動モデルがBayesということではなく，推定方法がBayesということ」
- 「分析者はあらゆるモデルをClassicalな方法でもBayesの方法でも推定できる」

- 数理科学の立場で(=統計的に)きちんと説明するならば，
  - "(行動)モデルがBayesか否か"は概念であり，統計的な区別ではない
  - ここでいう"Bayesの方法"はあくまでも尤度関数の何らかの意味での最大化

実例

まあでも，役に立つなら使えば良い

- 例: 多項プロビットモデル(MNP)のベイズ推定
  - McCulloch, R. and P. E. Rossi (1994). An exact likelihood analysis of the multinomial probit model. Journal of Econometrics 64(1-2), 207–240.
  - Imai, K. and D. A. van Dyk (2005). A Bayesian analysis of the multinomial probit model using marginal data augmentation. Journal of Econometrics 124(2), 311–334.
- 経済系/IIAが明らかに満たされない場合に対する研究が盛ん
- ベイズ推定はMNPの推定上の課題を回避する方法のひとつ
- (選択肢数-1)回の積分必要
  - 構造化プロビット等で軽減
- 分散共分散行列の
  - 識別性: (1,1)成分を1に，等
  - 正定値性: 何かしら工夫

$$\begin{cases} U_{1n} = V_{1n} + \epsilon_{1n} \\ U_{2n} = V_{2n} + \epsilon_{2n} \\ U_{3n} = V_{3n} + \epsilon_{3n} \end{cases} where \begin{pmatrix} \epsilon_{1n} \\ \epsilon_{2n} \\ \epsilon_{3n} \end{pmatrix} \sim \mathbf{MVN}\left(\mathbf{0}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{bmatrix}\right)$$

Choice probability of alternative 1 is

$$P_n(1) = \int_{-\infty}^{V_{1n}-V_{2n}} \int_{-\infty}^{V_{1n}-V_{3n}} \cdots \int_{-\infty}^{V_{1n}-V_{Jn}} n(\boldsymbol{q}; \mathbf{0}, \boldsymbol{\Sigma_1}) dq$$

$n(\mathbf{q}; \mathbf{0}, \mathbf{\Sigma}_1)$: multivariate normal density with mean vector **0** and variance-covariance matrix $\mathbf{\Sigma}_1$

まあでも，役に立つなら使えば良い

- ベイズ推定はMNPの推定上の課題を回避する方法のひとつ
- McCulloch and Rossi (1994)

yは効用(前ページではU)

A Baycsian analysis of the MNP model requires the specification of a prior over the parameters (β, Σ) and computation of the posterior density.

事後確率　事前確率　尤度

$$p(\beta, \Sigma \mid y_1, \ldots, y_N, X) \propto p(\beta, \Sigma) l(\beta, \Sigma). \quad (2)$$

The MNP likelihood is formed by the product of the N independent multinomial distributions, 尤度 $l(\beta, \Sigma \mid y_1, \ldots, y_N, X) = \prod_{i=1}^{N} \prod_{j=1}^{p} Pr_{ij}^{y_{ij}}$. As for the prior, it is convenient to employ a normal prior on β, $\beta \sim N(\bar{\beta}, A^{-1})$ and an independent Wishart prior on Σ^{-1} (denoted G below),

$$p(\beta \mid \bar{\beta}, A) \propto |A|^{0.5} \exp\{-\tfrac{1}{2}(\beta - \bar{\beta})' A (\beta - \bar{\beta})\} \quad (3a)$$

and

事前確率を規定する
ハイパーパラメータ

$$p(G \mid \nu, V) \propto |G|^{(\nu - p - 1)/2} \operatorname{etr}\{-\tfrac{1}{2} G V\}. \quad (3b)$$

共役事前分布:
βに正規分布
Σに逆Wishart分布
⇔ $G=\Sigma^{-1}$にWishart分布

まあでも，役に立つなら使えば良い

- ベイズ推定はMNPの推定上の課題を回避する方法のひとつ
- McCulloch and Rossi (1994)

Thus, the conditional posterior distribution of β is normal,

βの事後確率

$$\beta | w, G \sim \mathrm{N}(\hat{\beta}, \Sigma_\beta), \quad \Sigma_\beta = (X^{*\prime} X^* + A)^{-1}, \quad \hat{\beta} = \Sigma_\beta (X^{*\prime} w^* + A\bar{\beta}). \tag{7}$$

d is a function of w so that it is not necessary to add d to the conditioning arguments.

The third conditional distribution of G given β, w, X can be obtained from standard Bayesian analysis of a covariance matrix and Wishart theory. Given β, w, and X, we actually observe $\varepsilon_i = w_i - X_i\beta$. We combine the conjugate Wishart prior with the likelihood to obtain a Wishart posterior,

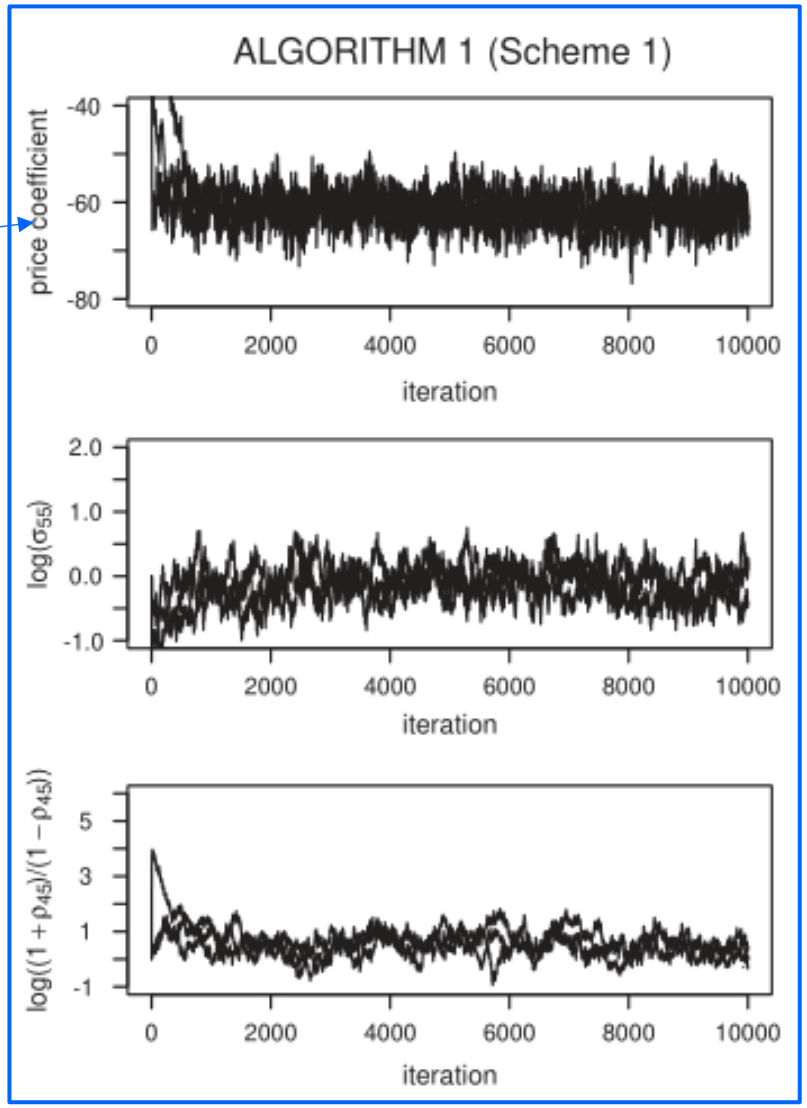
$G=\Sigma^{-1}$の事後確率

$$G | \beta, w \sim W\left(v + N, V + \sum_{i-1}^{N} \varepsilon_i \varepsilon_i'\right). \tag{8}$$

- 詳細は当該文献等を参照

計算の様子 (Imai and van Dyk)

- Sioux Falls洗剤選択モデル(6肢)
- 効用関数の係数
  - 例: 価格の係数
- 分散共分散行列の中身
  - 対角項の例
  - 非対角項(の変数変換)の例



補足: 事前分布

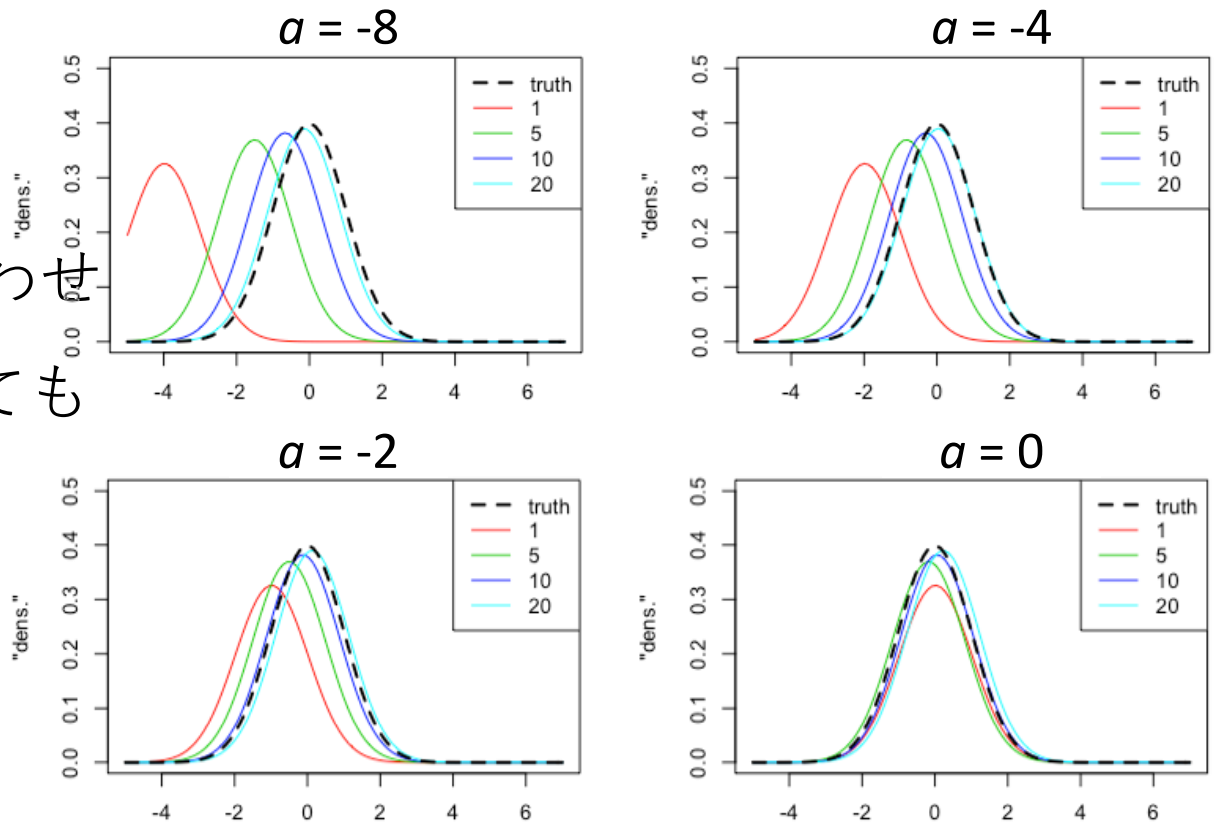
- 実務の理由で使っていると，理想的には事前分布は使いたくないという気持ちがある(?)ので，妥当性が気になる(??)
  - そういうヒトは，数理科学の理由に根拠を求めましょう
- そのうえで冷静に考えると，
  - 有意水準1%でも100回に1回くらいは的外れなのだから，サンプル数>100もあれば，事前確率が的外れになることを気にしても仕方ない
  - サンプル1個の尤度と事前確率が同等の影響力

$$p(\theta|X_{1,\ldots,N}) = \frac{1}{Z}\prod_i p(X_i|\theta)\,\varphi(\theta)$$

  - 次のページで図解します

補足: 事前分布よりもモデルが主役

- 真の分布 $q(x) \sim N(0, 1^2)$: 標準正規分布
- 事前分布 $\varphi(\mu) \sim N(a, 1^2)$
- モデル $p(x|\mu) \sim N(\mu, 1^2)$
- サンプル数
  1, 5, 10, 20のときの
  推定された分布と
  真の分布の重ね合わせ
- a=-8まで離れていても
  サンプル数20で
  ほとんど真の分布



少しだけ，数理科学の理由

モデル推定・評価・選択

- モデル推定(統計的推測・学習)
  - 得られているサンプル(データ)に基づいて,
    真の確率分布(モデル)に接近しようとする試み
- モデル評価
  - 推定したモデルの良さを客観的・相対的に比較すること
- モデル選択
  - 考えうるモデル候補から良いモデルを選択すること
  - “良い”=真のモデルに近い**or**未知データの予測精度が高い

推定法: モデル推定を行う具体的な方法

ベイズ推定も最尤推定も，前例踏襲も決めつけも，全部仲間

それぞれの方法に利点や欠点がある

モデル選択の歴史

-1930　1950　1970　1990　2010-

主義によるモデル選択

最尤法 (Fisher 1912-22)

ベイズ法 (もっと昔から)

歴史的な理由

規準によるモデル選択

AIC (Akaike 1974)　WAIC (Watanabe 2009)

BIC (Schwarz 1978)　WBIC (Watanabe 2013)

正則な場合　一般の場合

- 真の分布が未知だから
  正確さは永遠にわからない
- 確率とはなにかという決め事
  確率についての哲学の違い
- 最尤法は客観的
  ベイズ法は主観的
- 主義主張の対立と停滞

- 真の分布が未知でも
  正確さを見積もれる
- 統計的モデル推定には無関係
  哲学ではなく数理科学の問題
- どちらにせよモデルは主観
  ゆえに客観的な評価規準が重要
- パラダイムシフトによる進展

- モデルの統計的な側面は，工学ではなく理学の興味の範疇と捉えられることがあるが，むしろ，やっと工学で議論できる基盤ができてきた

要点

	最尤推定	ベイズ推定
目的	サンプル$\mathbf{X}$から真の分布qに接近したい	
用意するもの	サンプル$\mathbf{X}$とモデル$p(\mathbf{x}\|\boldsymbol{\theta})$ (→尤度関数)	
仮定	**尤度最大が最良**	**事前分布$\varphi(\boldsymbol{\theta})$の存在**
得られる推定量・分布	最尤推定量$\hat{\boldsymbol{\theta}}$ 必要なら，分散共分散行列 (=尤度関数の形状)	事後分布$p(\boldsymbol{\theta}\|\mathbf{X})$ 必要なら，事後確率最大化・事後平均等の代表値
得られるモデル	最尤推定量を モデルに代入したもの	事後分布で モデルを平均したもの
使えるとき	**正則な場合のみ**	**(比較的)一般の場合**
計算量	問題の性質が良ければ小	一般に大
モデル比較	多数のモデル$p(\mathbf{x}\|\boldsymbol{\theta})$で推定して情報量規準を比較	多数のモデル$p(\mathbf{x}\|\boldsymbol{\theta})$と事前分布$\varphi(\boldsymbol{\theta})$の組み合わせで推定して情報量規準を比較

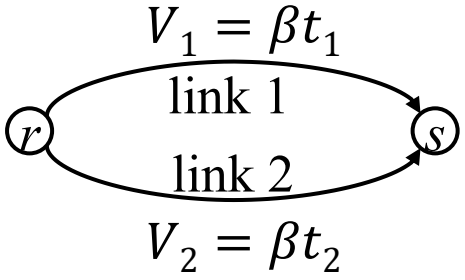
誤解ポイントと正しい説明

	最尤推定	ベイズ推定
目的	共通	
用意するもの	共通	
仮定	**尤度最大が最良**	**事前分布$\varphi(\theta)$の存在**
得られる推定量・分布	最尤推定量$\hat{\boldsymbol{\theta}}$ 必要なら，分散共分散行列 (=尤度関数の形状)	事後分布$p(\boldsymbol{\theta}\|\mathbf{X})$ 必要なら，事後確率最大化・事後平均等の代表値
得られるモデル	最尤推定量を モデルに代入したもの	事後分布で モデルを平均したもの
使えるとき	**正則な場合のみ**	**(比較的)一般の場合**
計算量	問題の性質が良ければ小	一般に大
モデル比較	多数のモデル$p(x\|\theta)$で推定し て情報量規準を比較	多数のモデル$p(x\|\theta)$と事前分布$\varphi(\theta)$の組み合わせで推定して情報量規準を比較

- 「最尤推定は点推定，ベイズ推定は区間推定」
  → 誤りではないが，
  それはこの部分だけ
- 事前分布の存在を認めるor尤度最大が最良と信じるという選択をしている
- 最尤推定量の最良性は正則な場合のみ成立
- 最尤推定も真のモデルに近いモデルを推定する作業
- 最尤推定量はその過程で出てくるもの
- パラメータの事後分布と最終的に求まるモデルは別物

別の興味深い例_設定

- とてもシンプルな2項ロジットリンク選択
  - 所要時間パラメータのみ

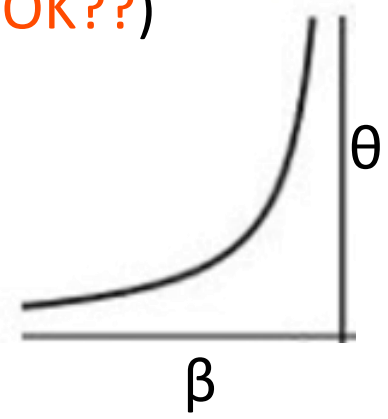


$$U = V + \epsilon$$

$$\epsilon \sim \text{iid } Gumbel(location = 0, scale = \theta)$$

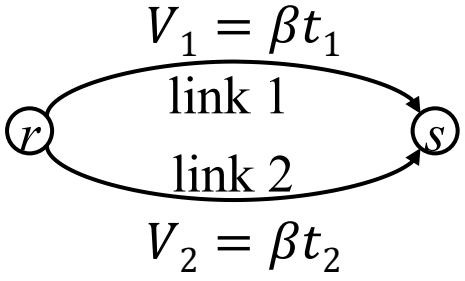
$$P(1) = \frac{\exp(\theta V_1)}{\exp(\theta V_1) + \exp(\theta V_2)} = \frac{\exp(\theta\beta t_1)}{\exp(\theta\beta t_1) + \exp(\theta\beta t_2)}$$

- (θβ)でひとつのパラメータになっている → θとβが不定 (事実)
  - 一般性を失わずθ=1とする (←これはいつでもOK??)
  - 理屈上はもちろんOK
- だが，推定は尤度に基づいて行うし，
  その尤度はサンプルデータから計算される
  - どんなサンプルデータでも常に成立するか?



別の興味深い例_実験

- とてもシンプルな2項ロジットリンク選択
  - 所要時間パラメータのみ



$$U = V + \epsilon$$

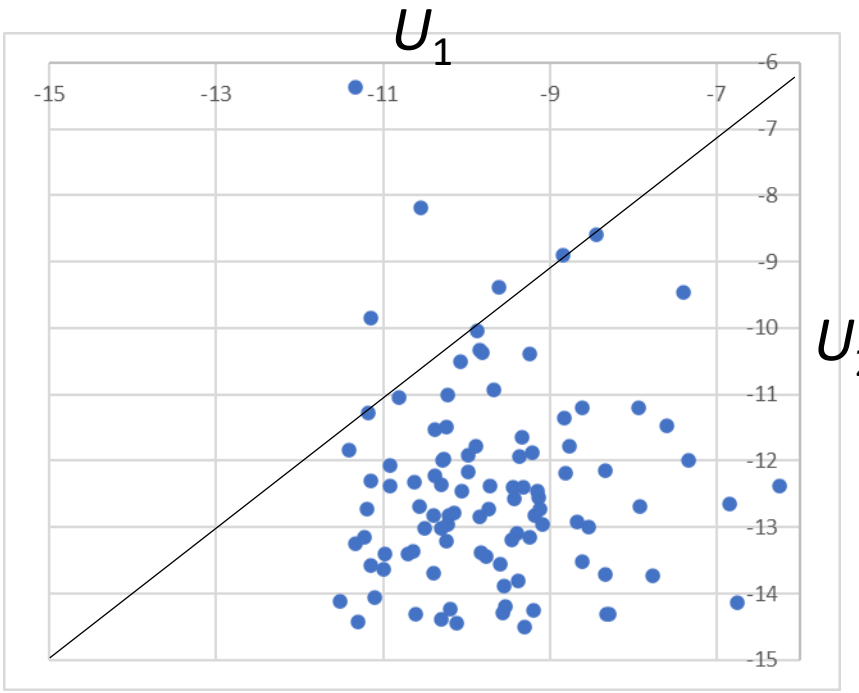
$$\epsilon \sim \text{iid } Gumbel(location = 0, scale = \theta)$$

$$P(1) = \frac{\exp(\theta V_1)}{\exp(\theta V_1) + \exp(\theta V_2)} = \frac{\exp(\theta\beta t_1)}{\exp(\theta\beta t_1) + \exp(\theta\beta t_2)}$$

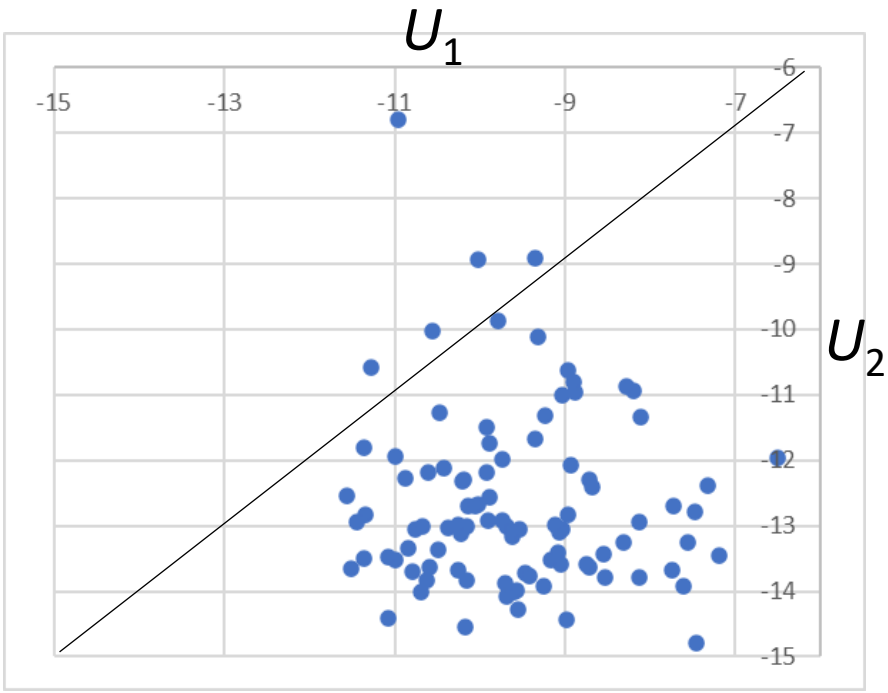
- β=-1, θ=1, t_1=10を固定，$\epsilon \sim \text{iid } Gumbel(0, \theta)$で$\epsilon$を多数発生させ，$t_2$(>$t_1$)の値に応じた$U_1$,$U_2$のサンプルを多数作成
- そのサンプルを元に
  - θ=1を前提にβを推定
  - θを固定せずにβとムリヤリ同時推定

確定効用の差がみえるとき

- β=-1, θ=1, t_1=10を固定，$\epsilon \sim \text{iid } Gumbel(0,\theta)$で$\epsilon$を多数発生させ，$t_2(>t_1)$の値に応じた$U_1$,$U_2$のサンプルを100個作成
- t_2=13のときの例



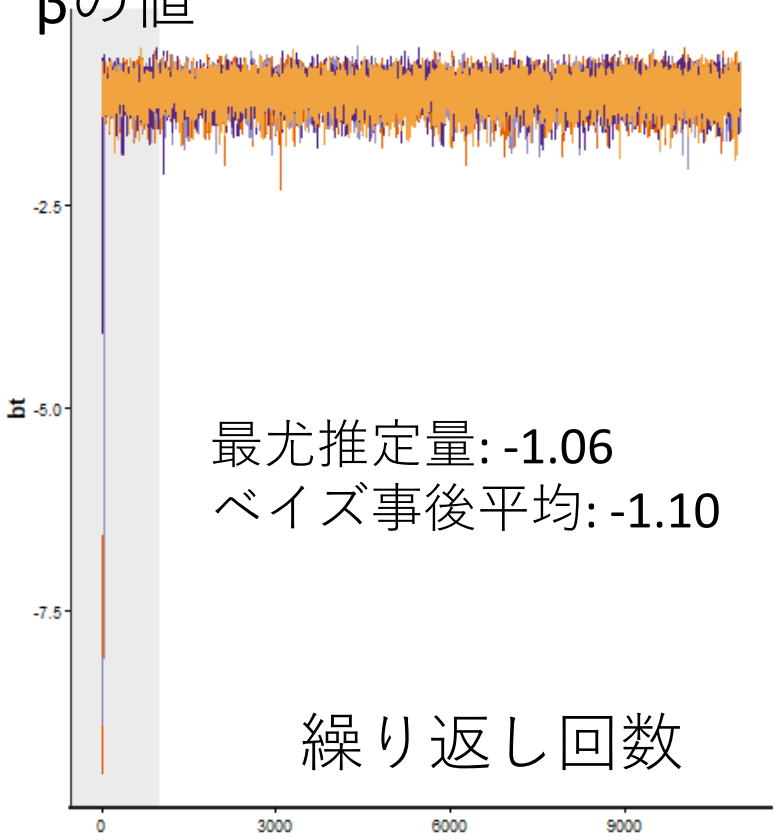
最尤推定量: -1.06
ベイズ事後平均: -1.10



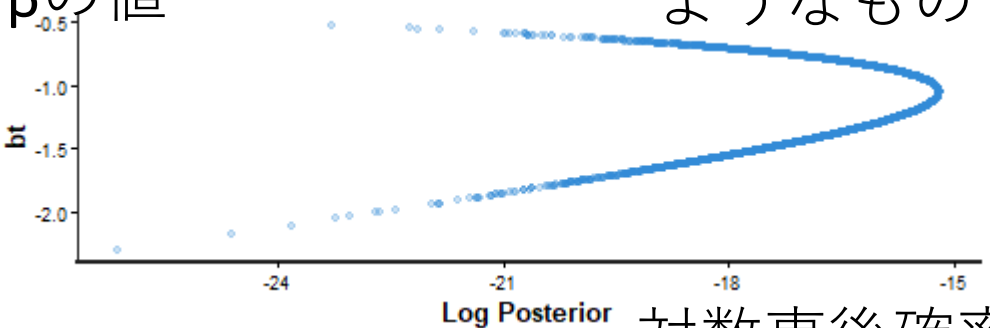
最尤推定量: -0.98
ベイズ事後平均: -1.01

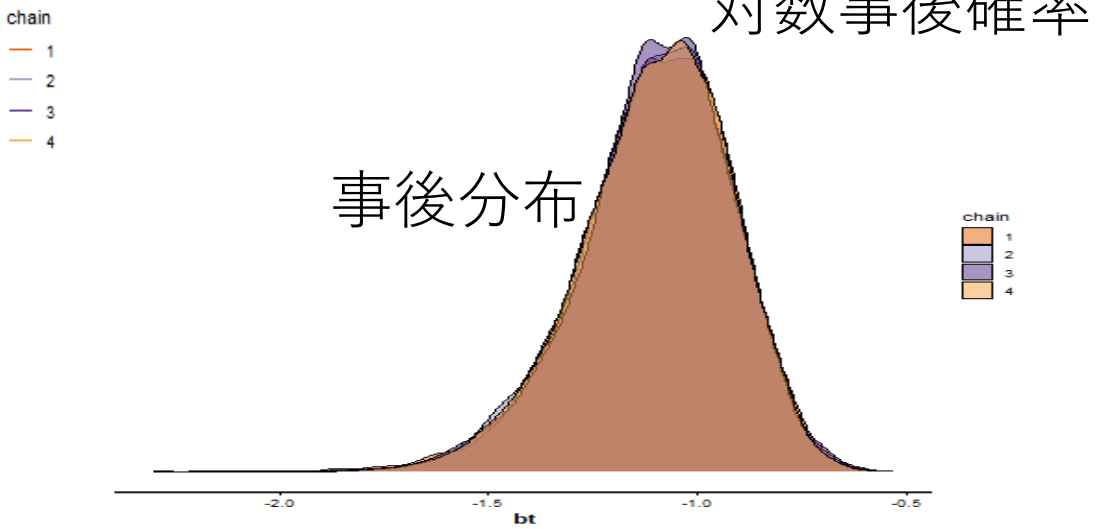
確定効用の差がみえるとき

- β=-1, θ=1, t_1=10を固定，$\epsilon \sim \text{iid } Gumbel(0,\theta)$で$\epsilon$を多数発生させ，$t_2(>t_1)$の値に応じた$U_1$,$U_2$のサンプルを100個作成
- t_2=13のときのベイズ推定の様子



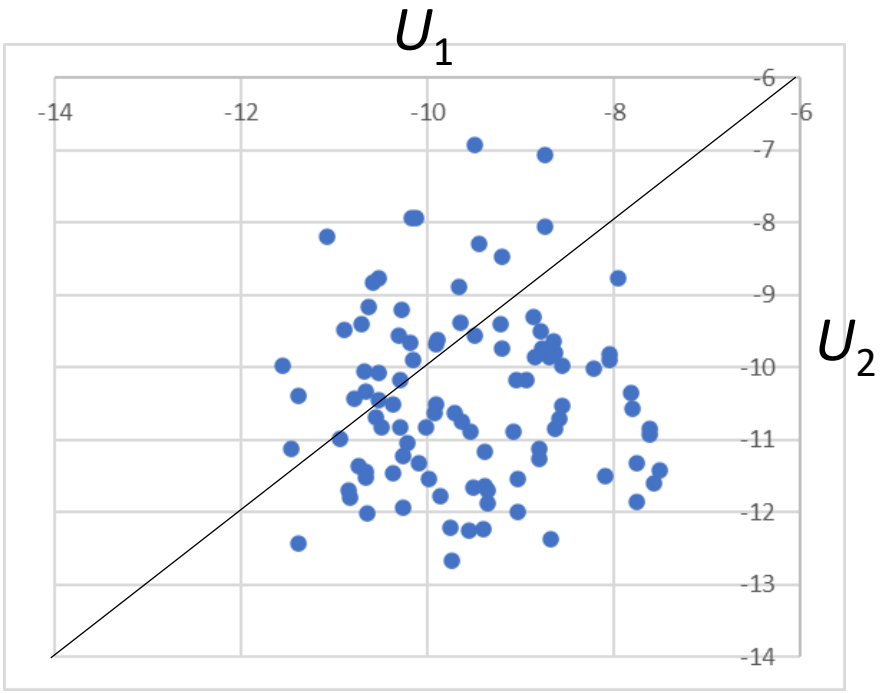
最尤推定量: -1.06
ベイズ事後平均: -1.10



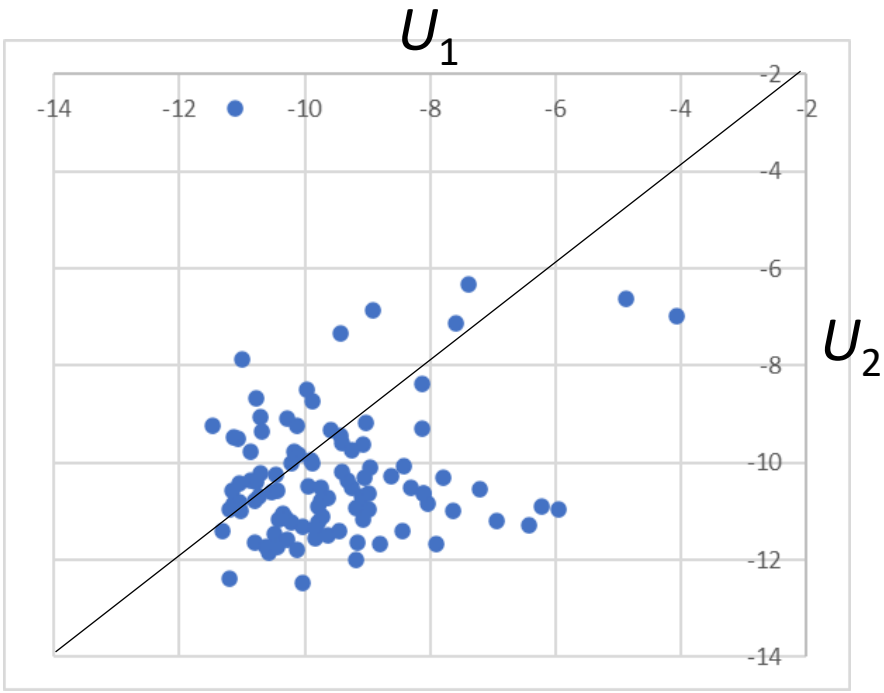


確定効用の差がみえないとき

- β=-1, θ=1, t_1=10を固定，$\epsilon \sim \text{iid } Gumbel(0, \theta)$で$\epsilon$を多数発生させ，$t_2(>t_1)$の値に応じた$U_1$,$U_2$のサンプルを100個作成
- t_2=11のときの例



最尤推定量: -0.85
ベイズ事後平均: -0.86

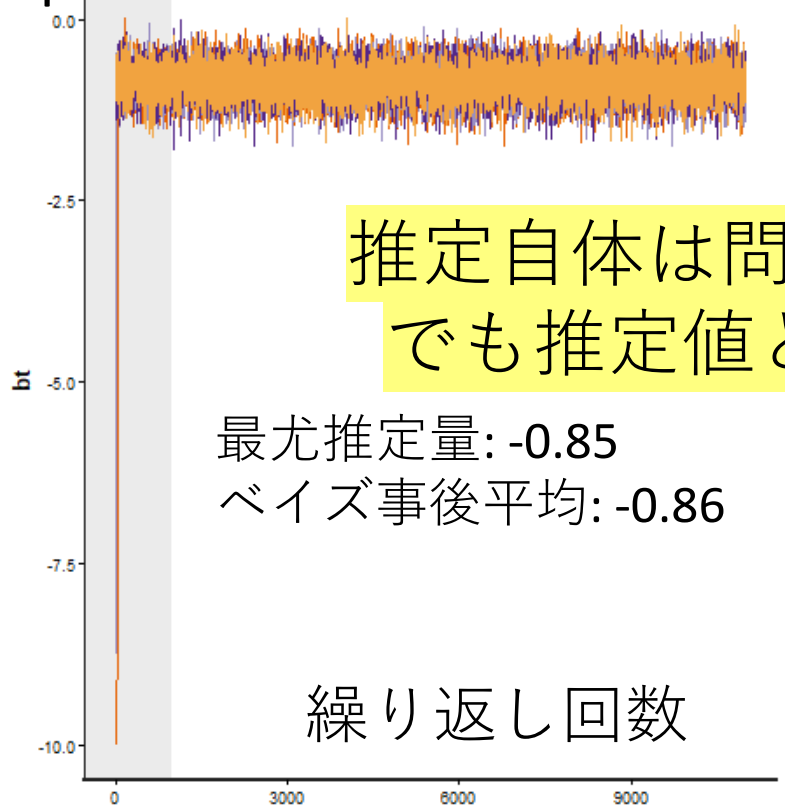


最尤推定量: -0.71
ベイズ事後平均: -0.72

確定効用の差がみえないとき

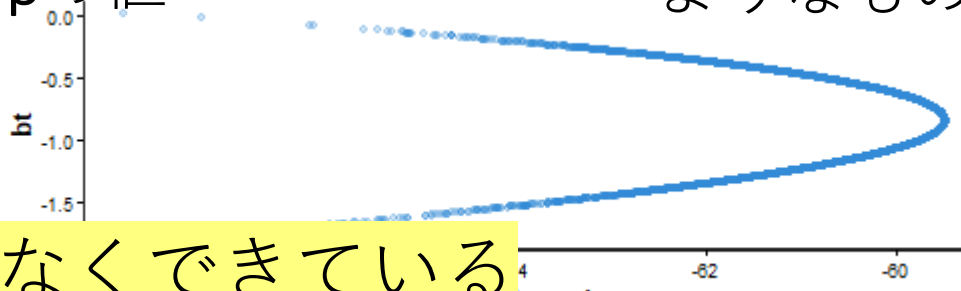
- β=-1, θ=1, t_1=10を固定，$\epsilon \sim \text{iid } Gumbel(0,\theta)$で$\epsilon$を多数発生させ，$t_2(>t_1)$の値に応じた$U_1$,$U_2$のサンプルを100個作成
- t_2=11のときのベイズ推定の様子

βの値



繰り返し回数

βの値



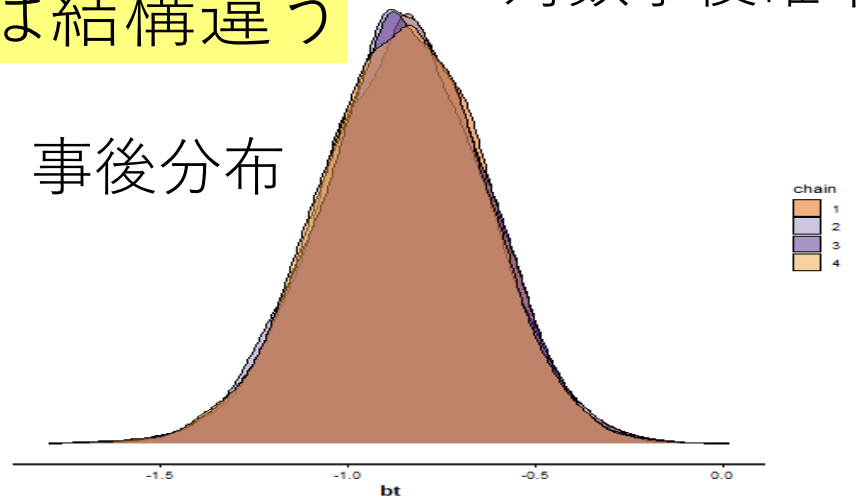
尤度関数のようなもの

対数事後確率

推定自体は問題なくできている
でも推定値と真値は結構違う

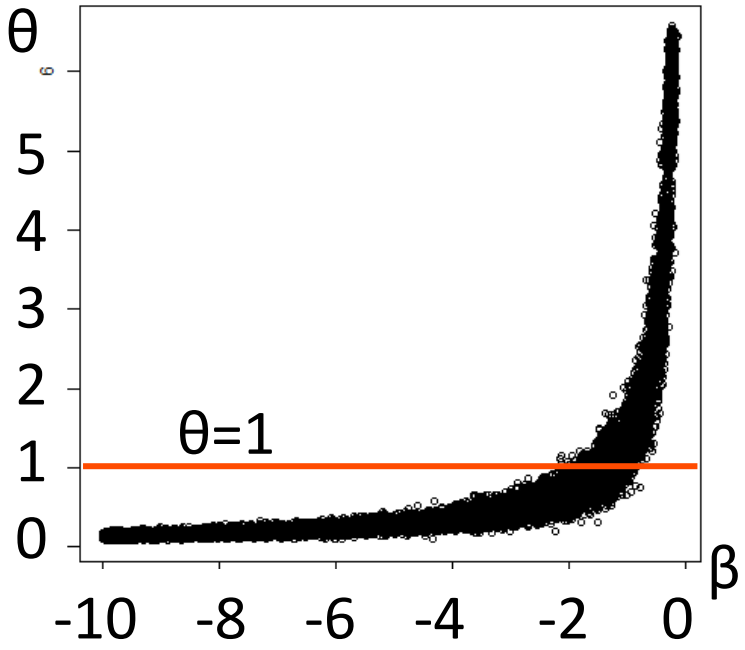
最尤推定量: -0.85
ベイズ事後平均: -0.86

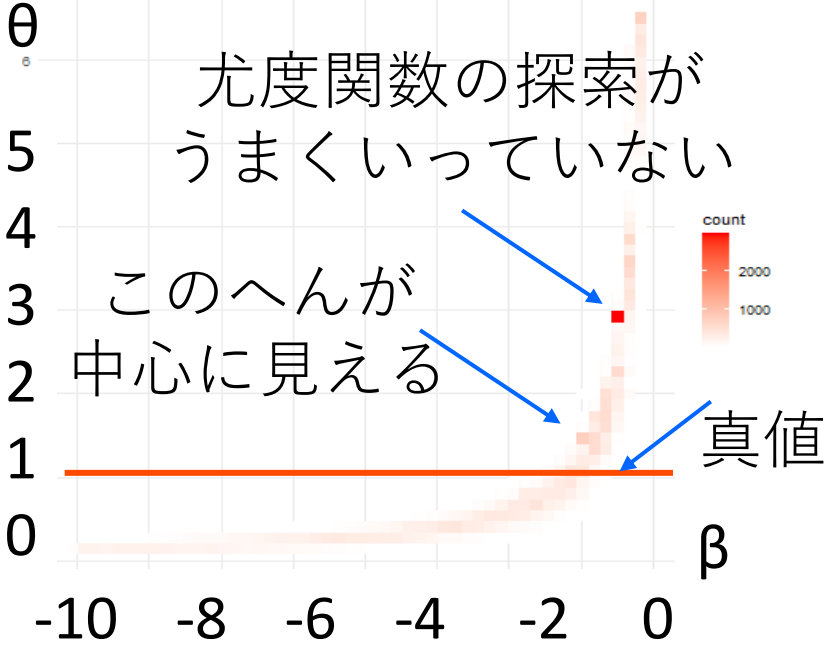
事後分布



事後分布から考える

- t_2=11のときの別の例，θ=1で固定すると最尤推定量は-1.39
  - θを固定せずにムリヤリ同時推定したときの事後分布



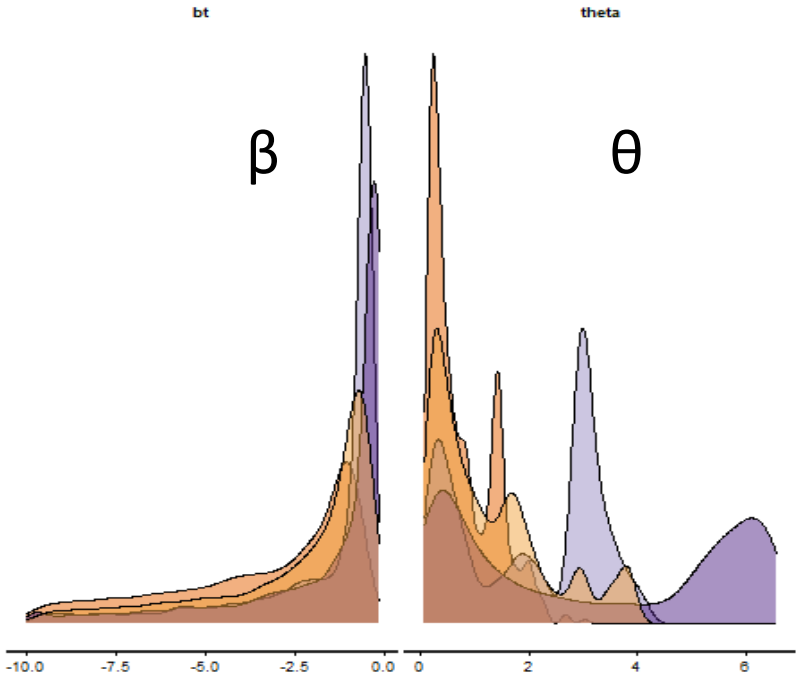


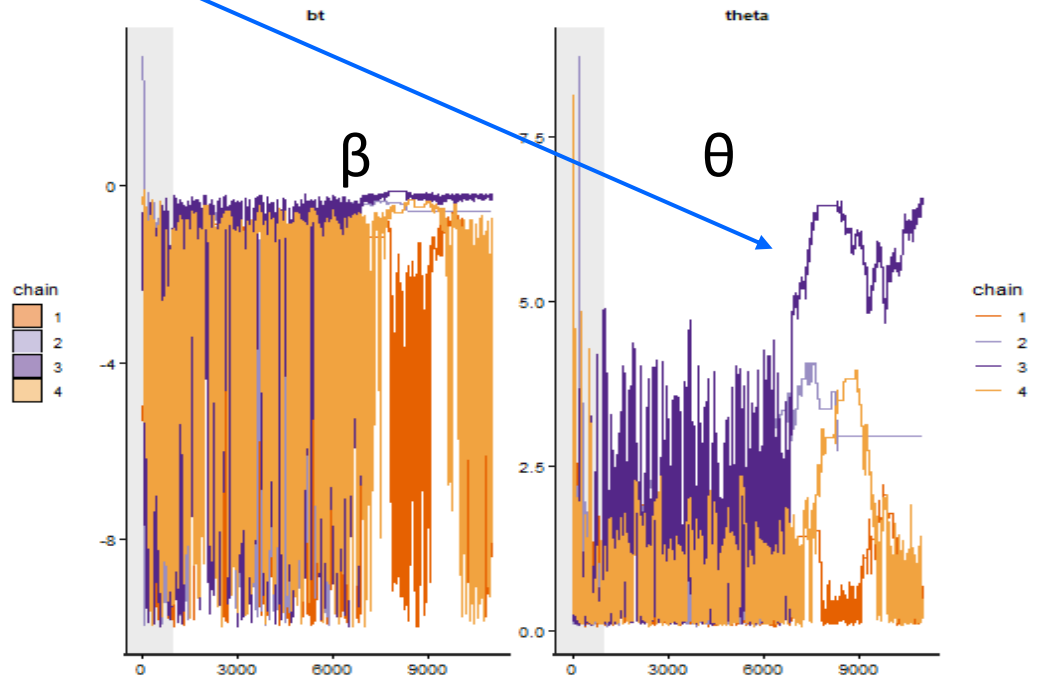
  - (注) そもそも推定が収束していない

- **確定効用の差が小さいとそもそも復元は困難**
- **復元がどのくらい困難か，同時推定すると見通しが良い**

(補足) 収束していない様子

- どんな値でも割と良い尤度を示すので，なかなか収束せず，流浪の旅に出始める





おわりに

おわりに

- **必要なときはベイズ推定を使いましょう**
  - 夏の学校に限らずいつでも中西に連絡ください
  - 実務の理由で用いるベイズ推定，大歓迎
    - 階層化したい，尤度関数の最大化が難しい，...
    - 事前分布批判は気にしない．モデルが支配的に重要.
  - 数理科学の理由が頭の片隅にあると，より説得力あり
    - モデルが正則でないならば，ベイズ推定が望ましい
  - どちらの理由で用いるにせよ，計算量をいかに抑えるかがポイント．これ自体が最先端の研究課題
- どんなときモデルが正則でないかは数学の問題，難しい
  - 端的にはほとんどの場合だが，工学的にはそれだと困る
  - 妥協のしどころはよく分かっていない(分野固有の問題)