ベイズ統計の導入と行動モデル

2024/09/11 行動モデル夏の学校 @ 本郷

金沢大学 中西航

nakanishi [at] se.kanazawa-u.ac.jp

ベイズが難しいいくつかの理由

- 問題1: ベイズ多義的すぎ
  - 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
  - 子供だまし程度しか教える時間がない
- 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

1: 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
  - …

きょうの「ベイズ推定」

- サンプルに基づいて(行動)モデルのパラメータを推定する手法のひとつ
  - 「最尤推定」の仲間(対立相手ではない，単なる代替)
  - もっというと「熟練者の直感」「偉い人の決め打ち」なども仲間
- そのような手法の中で，ベイズ統計学に基づいたもの
  - ベイズ統計学に基づく方法=ベイズ推定にしかない優れた点がある
  - もちろん劣る点もある

2: 教える時間がない

- 結果的にこういう説明になりがち
  - (なんかよくわからんが)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 $p(\theta|X_{1,\ldots,N})$

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

  - 必要ならパラメータの推定量として
    - 事後確率最大化
    - 事後平均
- 当然以下の疑問がある
  - 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??

その場しのぎの説明

- 事前分布って何??
  - 分析者の先見的知識を反映(?)
  - 階層化して個人差を表現(?)
- 事後分布って何??
  - 尤度関数だけでなく事前の情報を反映できるから良い(?)

- 説明すればするほど，そんな主観的でよいの? という批判を際立たせる悪手かもしれない

でも，便利な例で押し切る

- 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

でも，便利な例で押し切る

- 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$\varphi(q)$~$N(0.9,0.01)$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

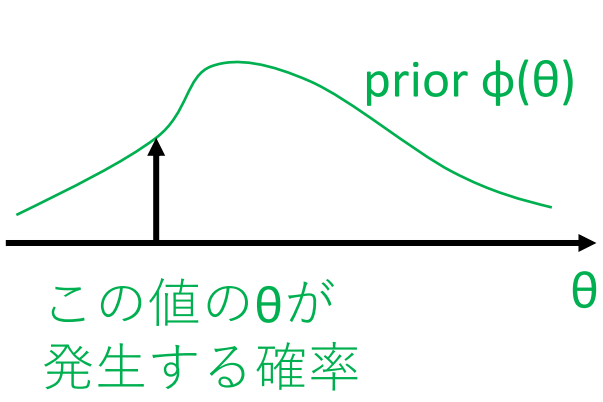
よく見かけるベイズ推定の1つ目の顔:

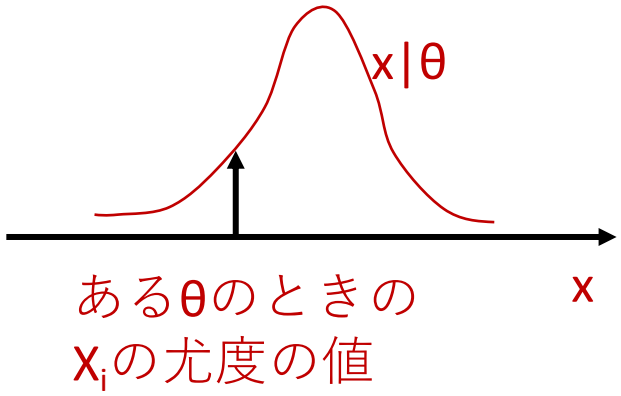
異質性の表現としてのベイズ推定

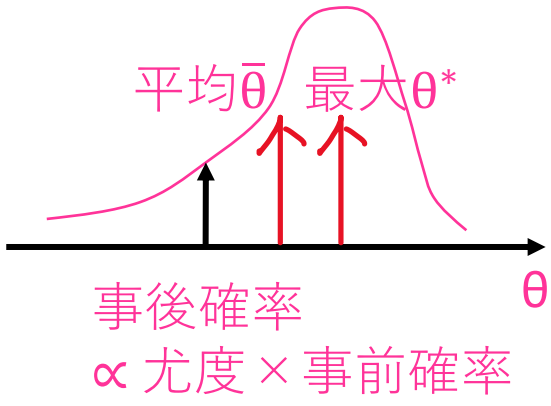
事後分布の計算方法

- どうやったら計算できる??
$$p(\theta \mid X_{1,\ldots,N}) = \frac{1}{Z} \prod_i p(X_i \mid \theta)\, \varphi(\theta)$$
- 解析的に行えるケースは稀(共役事前分布)
- 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
2．そのθで尤度を計算する
3．計算結果をたくさん作る



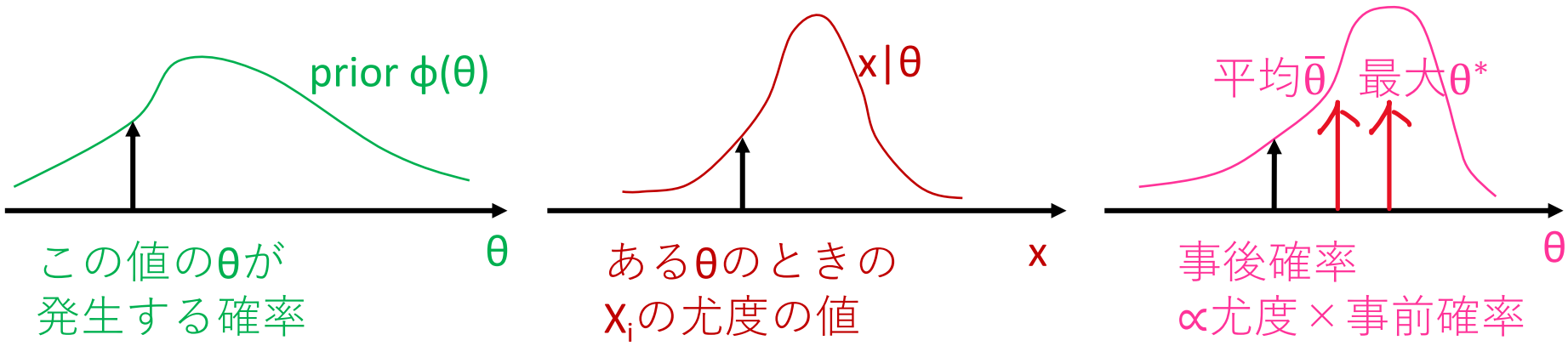




- Zを求めるのは難しいが，Zがなくても事後分布の形はわかる
- 最大や平均もわかる

そして，関係ない文脈で押し切る

- 数値的に近似する方法の別の見方…



- もし，事後分布の平均値を与える$\bar{\theta}$を得たいだけなら，良い計算の方法がある(説明省略)
- 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2つ目の顔:

アルゴリズムとしてのベイズ推定

3: 確率的主義という虚構

説明の前に…

- これから，よく聞くであろう，もしかすると講義で習うであろうことはおよそ間違いだ，という話をします
- このことは統計学・数理科学の人々は当たり前のように認識していることです
- が，それ以外の分野ではいまだに根強い誤解です

モデル推定・評価・選択

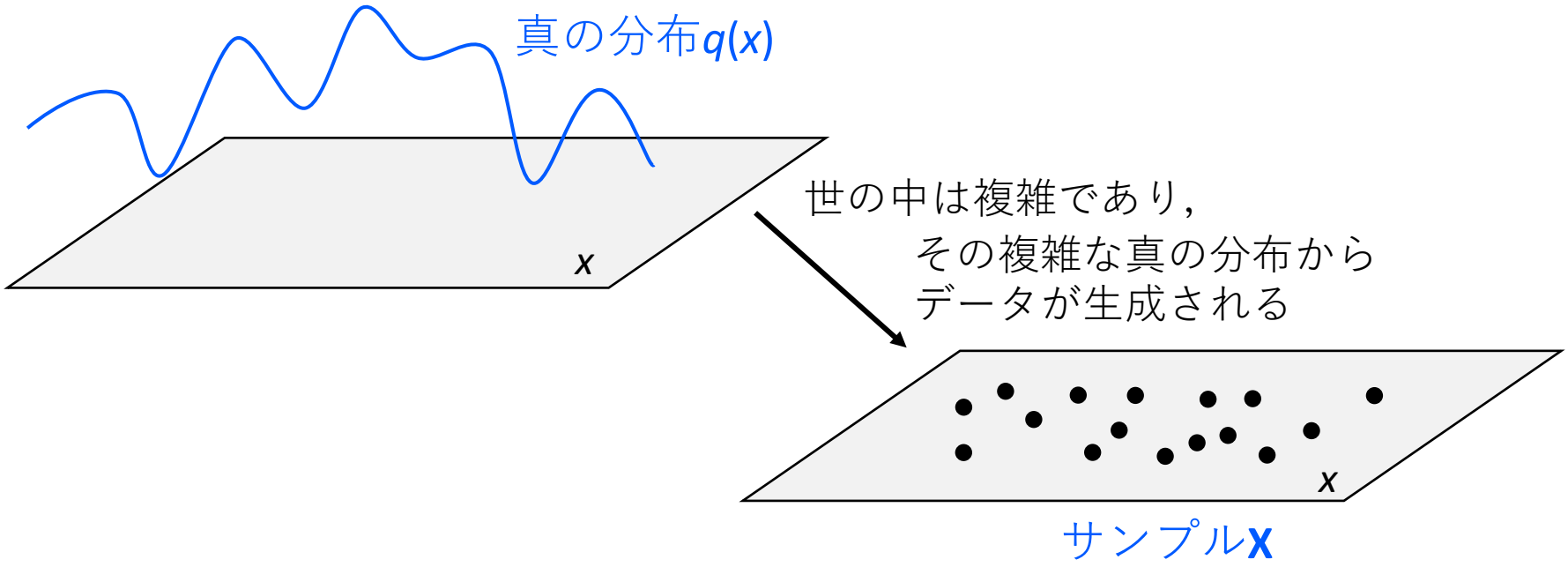
- モデル推定(統計的推測・学習)
  - 得られているサンプル(データ)に基づいて,
    真の確率分布(モデル)に接近しようとする試み

- 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”の意味は後述

モデル推定とモデル評価

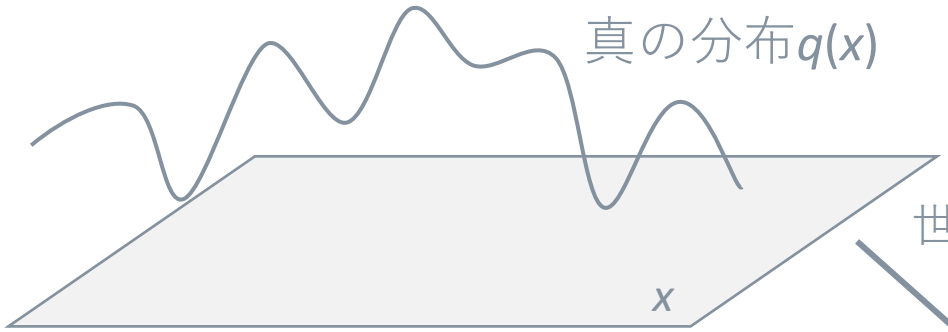
- 真の分布$q(x)$(=モデル)をサンプル$\mathbf{X}$(=データ)から推定したい



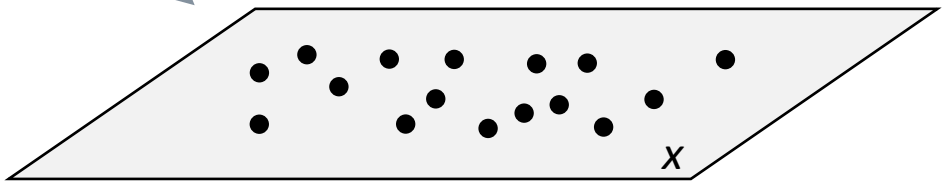
$q()$: 確率分布

モデル推定とモデル評価

- 真の分布$q(x)$(=モデル)をサンプル**X**(=データ)から推定したい

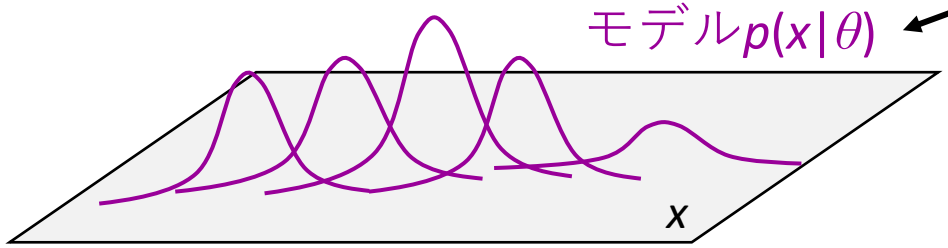


世の中は複雑であり，
その複雑な真の分布から
データが生成される



サンプル**X**

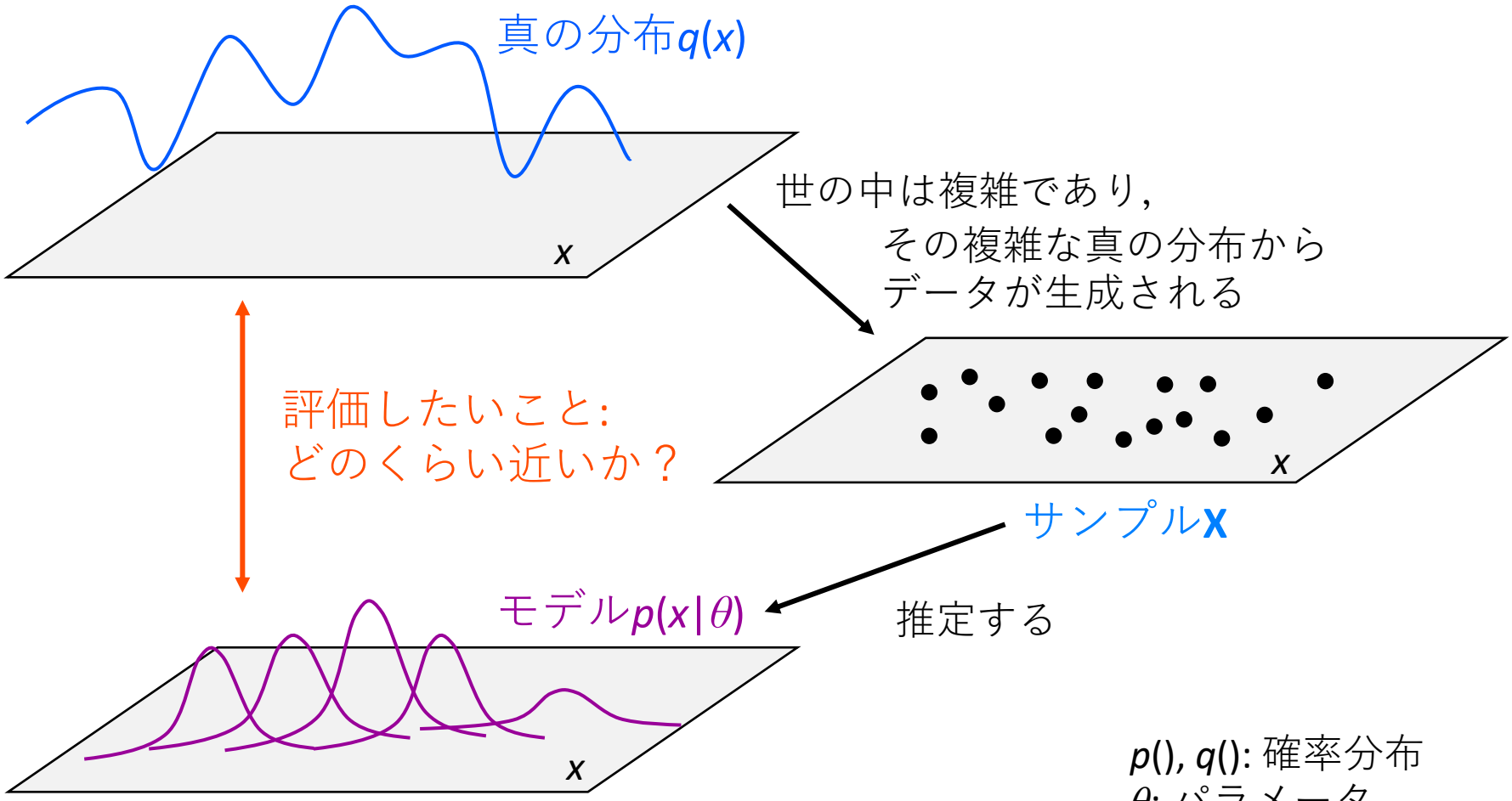
推定する



$p()$, $q()$: 確率分布
θ: パラメータ

モデル推定とモデル評価

- 真の分布$q(x)$(=モデル)をサンプル$\mathbf{X}$(=データ)から推定したい



$p()$, $q()$: 確率分布
θ: パラメータ

モデル推定とモデル評価の本質

- モデル推定: $q(x)$になるべく近い$p(x|\theta)$を求めること
- モデル評価: $q(x)$に$p(x|\theta)$がどの程度近いかを見積もること
  - 近いほど良いモデルと考える
- 統計的モデル評価の本質的な意味:
  真の分布$q(x)$が未知でも，サンプル$\mathbf{X}$のみから，
  (1) モデル$p(x|\theta)$と真の分布$q(x)$との近さがわかる
  (2) モデル$p(x|\theta)$による未知データの予測精度がわかる
  - より正確には，それらの期待値が分かる
  - そうでなければ意味がない(手元のデータで評価不可能)
- これはビッグデータでもスモールデータでも一緒

2種類の良いモデル

- モデル候補(モデルA, モデルB, …)のなかで
  - (1) 真の分布$q(x)$に最も近いモデル
  - (2) 未知データの予測精度が最も良いモデル
- 真の分布がモデルに対して**正則な場合**(条件は省略),
  - (1) は Bayesian Information Criterion (BIC) 最小のモデル
  - (2) は Akaike Information Criterion (AIC) 最小のモデル
- この2つは一般に一致しない
  - 善し悪しではなく，分析者が何を求めるか
- これらが唯一(唯二?)の方法ではない
  - 尤度も，統計的検定(尤度比)も，信念もダメではない
  - どのような根拠に基づくかの認識と使い分けが重要

※ 厳密には，AICは実現可能なときに限る(説明省略)

確率論と主義: モデル評価には無関係[19]

- このような整理を見ることがあるが…

	頻度主義的	ベイズ的
真のモデル	神のみぞ知る 絶対的に正しい単一のもの	分布$q(x)$
サンプル$\mathbf{X}$	頻度主義的な真のモデル+ なんらかの分布に従う誤差	ベイズ的な真のモデルの サンプル
推定モデル	揺らがない$f(x)$とサンプルに よって発生する誤差分布	サンプルによって揺らぐ $q(x)$

こんなものはない!!!

- $f(x)$+(誤差分布)を$q(x)$とみれば両者はまったく一緒
  - $f(x)$が揺らいでいないというのは信念であり識別不可能
- 主義・信念とモデルの推定方法(最尤/ベイズ)は無関係
- 主義・信念はモデルの評価・選択においては無意味

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)

- 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

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

改めて整理すると

	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$\varphi(\theta)$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**	**一般の場合**

いくつかの誤解ポイント

- 「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	ベイズ推定
目的	共通：サンプルXから真の分布に接近したい	
用意するもの	共通：サンプルとモデル$p(x\|\theta)$ (→尤度関数)	
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$\hat{\theta}$ 必要なら，分散共分散行列 (=尤度関数の形状)	事後分布$p(\theta\|X)$ 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**	**一般の場合**

- 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

補足: 得られるモデル

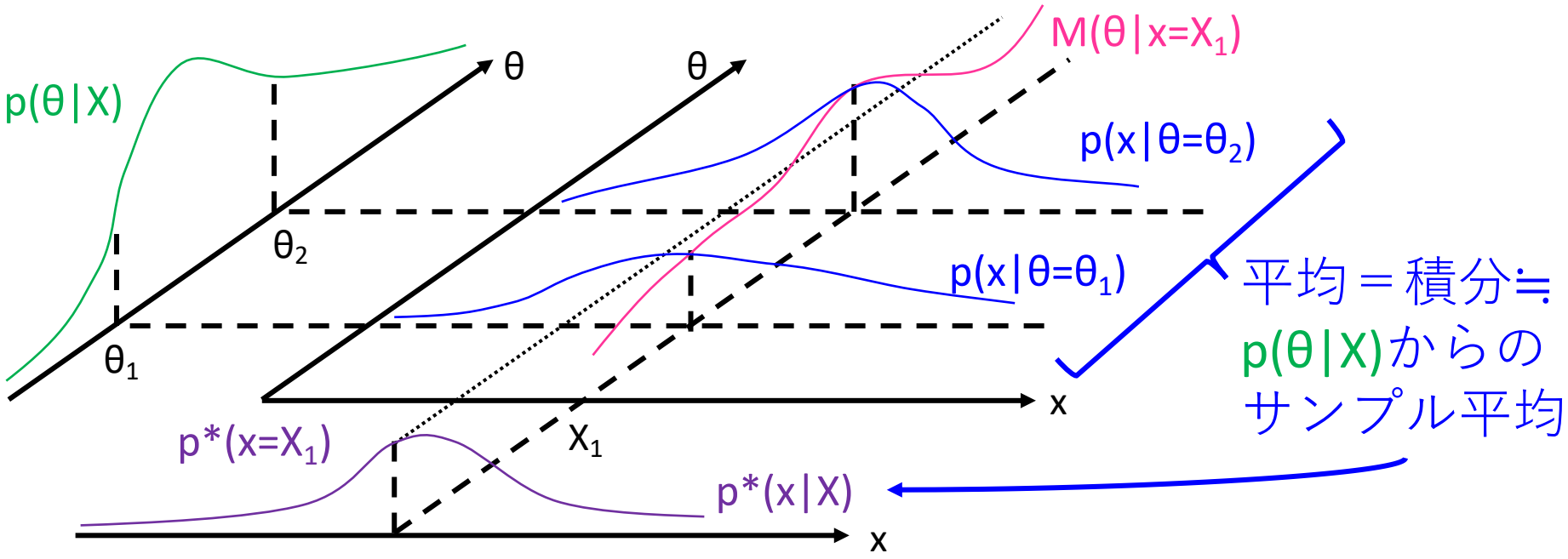
モデル $V = \beta_0 + \beta_1 x + \epsilon, \epsilon \sim$(何らかの確率分布)

パラメータ β_0, β_1 を推定し，モデルを推定する

- 最尤推定
  - 最尤推定量 $\widehat{\beta_0}, \widehat{\beta_1}$
  - それを用いたモデルは $V^* = \widehat{\beta_0} + \widehat{\beta_1} x + \epsilon$
- ベイズ推定
  - 事前分布 $\varphi(\beta_0, \beta_1)$
  - 事後分布 $p(\beta_0, \beta_1 | \mathbf{X})$
  - それを用いたモデルは $V^{**} = \iint p(V) p(\beta_0, \beta_1 | \mathbf{X}) d\beta_0 d\beta_1$

補足の補足: 関数を確率分布で平均

- モデルを事後分布で平均



推定された(真の分布に最接近した)モデル

(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

- 顔2: 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

行動モデルにおけるベイズ推定

- 顔2: 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

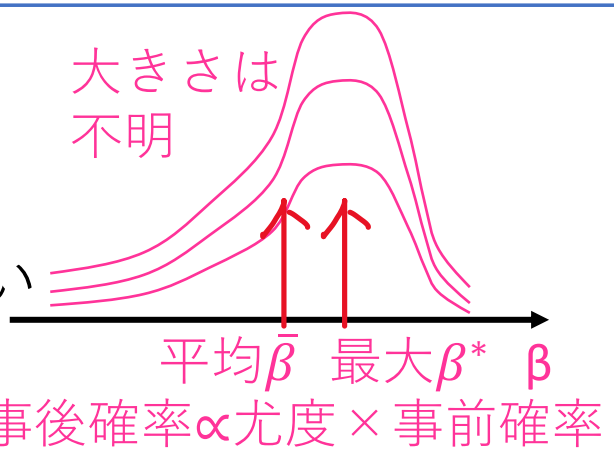
行動モデルにおけるベイズ推定

- 顔1: 異質性の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 procedures that we describe in Chapter 11. Variations of these procedures to accommodate other aspects of behavior have been numerous. For example, Arora *et al.* (1998) generalized the mixed logit procedure to take account of the quantity of purchases as well as brand 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.

気になる点

- zを求めないと形しかわからないので，
  モデル評価のための統計量は求まらない
  →ベイズ推定の**統計的**利点は発揮されない



- 本来ベイズ推定で得られるモデルは
  $V^{**} = \iint p(V)p(\beta_0, \beta_1|\mathbf{X})d\beta_0 d\beta_1$ だが，
  この使い方は見たことがない．

- むしろ，事後確率の平均を与える$\bar{\boldsymbol{\beta}}$・事後確率の最大を与える$\boldsymbol{\beta}^*$を代入したモデルをよく見かける: $V^{***} = \overline{\beta_0} + \overline{\beta_1}x + \epsilon$
  - これは，厳密にはベイズ推定の結果のモデルではない
  - 正確には「最尤推定の推定プロセスをベイズ推定に，尤度最大を事後確率の平均で置き換えたもの」
  - **世の中ではこれらを「ベイズ推定」と称することも多い**

補足: Train本にもそう書いてある

> 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」

- ただし，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

概念といいつつも…

- モデルがベイズ的であるとはどういうことか?
  - 観測するもの(変数)とシステムに従うもの(変数)が異なり,
  - システムに従うもの(変数)は直接観測ができないけれど,
  - 観測するもの(変数)からシステムに従うもの(変数)とシステムの特性(パラメータ/モデル)を知りたい

  という構造のモデル

- 例: 状態空間モデル

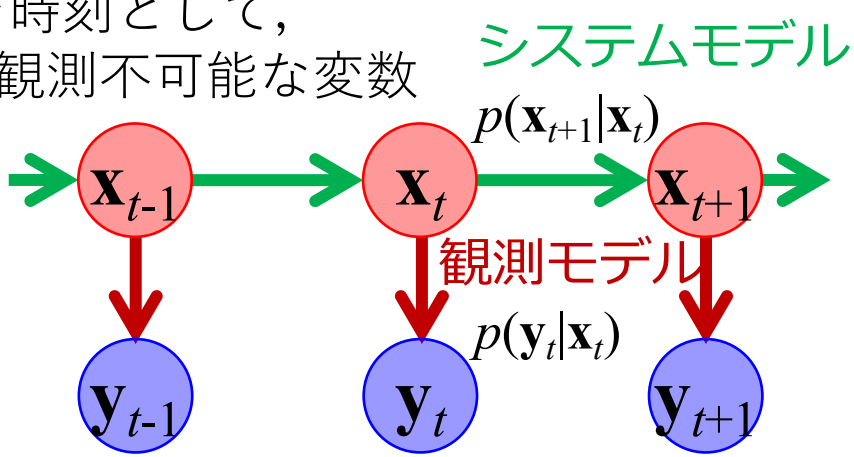
$$p(\mathbf{x}_t \mid \mathbf{y}_{1:t}) \propto p(\mathbf{y}_t \mid \mathbf{x}_t)\, p(\mathbf{x}_t \mid \mathbf{y}_{1:t-1})$$

事後分布　観測モデル　予測分布

$$= p(\mathbf{y}_t \mid \mathbf{x}_t) \int p(\mathbf{x}_t \mid \mathbf{x}_{t-1}) p(\mathbf{x}_{t-1} \mid \mathbf{y}_{1:t-1})\, d\mathbf{x}_{t-1}$$

観測モデル　システムモデル　1期前の結果

tを時刻として,
x: 観測不可能な変数

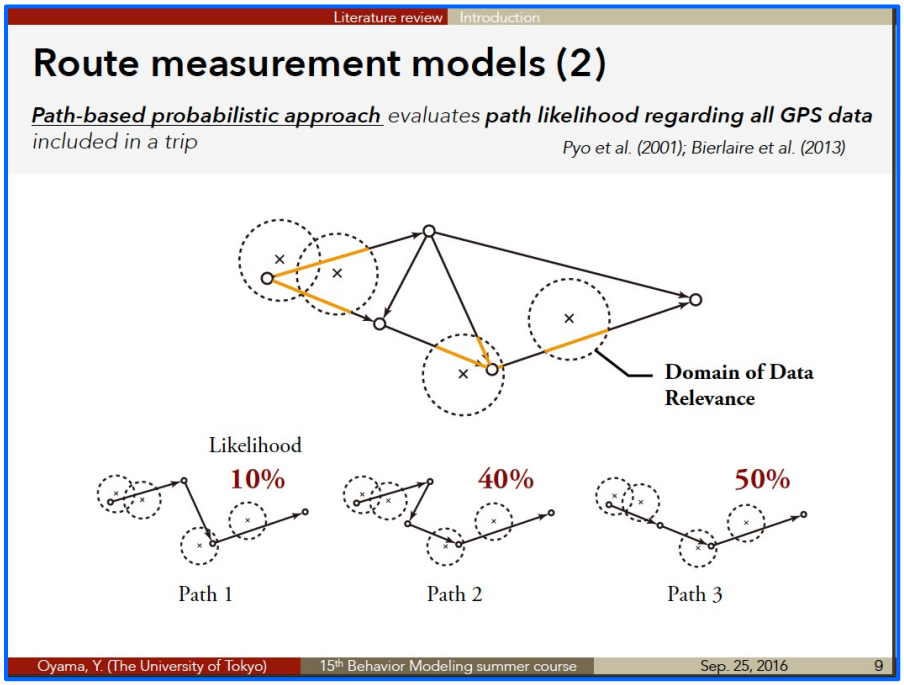


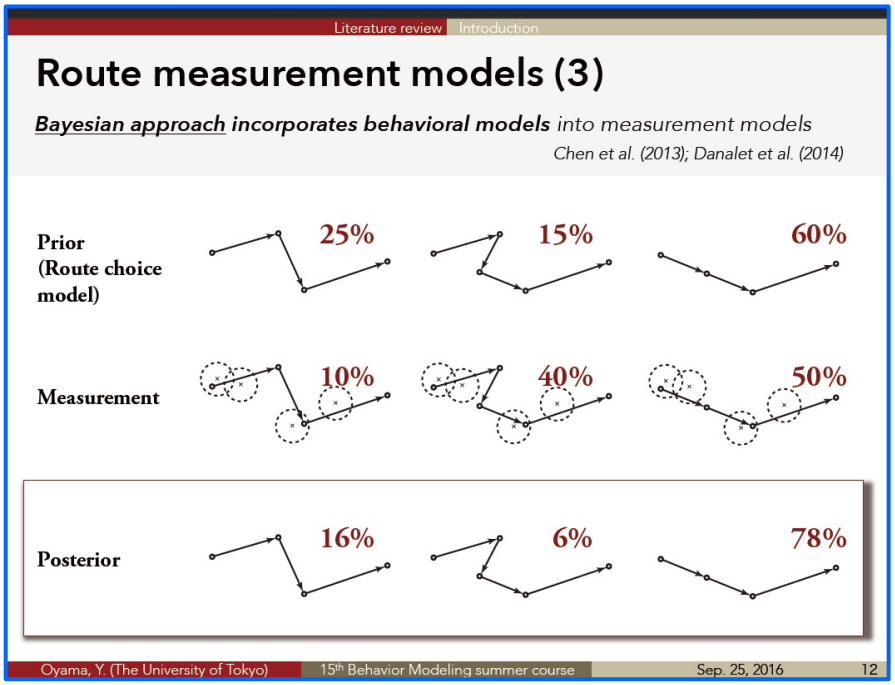
y: 直接観測可能な変数

- ほかにも隠れマルコフモデル, トピックモデル, …

行動モデル in ベイズ的なモデルの例

- 観測された一連のGNSS座標(y)から，直接観測のできない歩行者の存在リンクおよび経路(x)を推定する問題





http://bin.t.u-tokyo.ac.jp/model16/lecture/Oyama.pdf

- (注) 大山さんの上記スライドはあくまで研究背景で，Oyama and Hato (2018)は上記手法に内包される問題を解決した研究

ベイズへの批判

- [正当な批判]
  - 計算時間がかかりすぎる，いくら計算機が進化したといっても無理なものは無理
  - コーディングが面倒で複雑，現場担当者が使えない
  - 収束判断がよく分からない
  - (?)私のモデルは正則だと信じている．最尤推定で十分
- [誤った批判]
  - 事前確率は主観が入っており客観的ではない! 最尤推定こそが客観的である!!!
  - 推定量が不偏性を持たないなんてけしからん!!!

おわりに

- **必要なときはベイズ推定を使いましょう**
  - 階層化したい，尤度関数の最大化が難しい→もちろんOK
    - ただし，ベイズ統計の本質とはあまり関係ない
  - 事前分布への批判は気にしなくてOK
    - モデルの想定が支配的に重要
  - 本当は，想定したモデルにおいてベイズ推定が望ましい(=正則でない)ならば，ベイズ推定すべき
- 一方で，どのようなときにモデルが正則でないかは数学の問題で，簡単ではない
  - 端的には「ほとんどの場合」
  - とはいえ，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)