データ解析

第９回：ベイズ法

渡辺澄夫

ベイズ法と最尤法

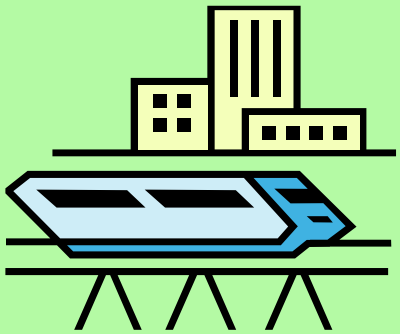
観測データ

回帰・判別分析

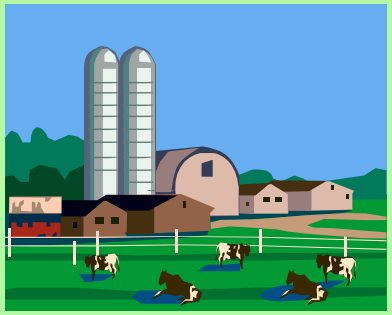
推定と検定

解析方法

ベイズ法
・階層ベイズ法



時系列予測



主成分・因子
・クラスタ分析

統計的推測　今昔物語

復習：統計的推測とは

実ユークリッド空間に値をとる確率変数 $X^n = (X_1, X_2, \ldots, X_n)$ の実現値（データ）が得られたとき、このデータを発生している確率密度関数 $q(x)$ を知りたい。しかし確率分布の集合は一般に無限次元であり、データは有限かつランダムであるから $q(x)$ を完全に知ることはできない。

☆　ここでは、パラメータ w をもつ統計モデル $p(x|w)$ とパラメータ w の事前分布 $\varphi(w)$ を使って推測を行なう方法を考える。

問題：　統計モデルと事前分布として何を使ってもよいが、推測された結果が真の密度関数をできるだけよく推測するようにするにはどうしたらよいだろうか。

「主義による統計学」から「数理科学としての統計学」へ

２０世紀前半、統計学が未熟だった時代には、「主義に基づく推測法の正当化」が争われていました。例えば「最尤主義に基づいて最尤推測の結果だけが正しい」とか「ベイズ主義に基づいてベイズ推測の結果だけが正しい」とか「お前はどちらの味方か」という学問とは思えない不毛な論争が世界を覆っていたのです。統計学をこの荒廃から救ったのは赤池弘次博士です。1970年ころ、赤池博士は「**統計的推測は真の分布に対する適切さで評価するとよい**」という自然な考えかたに立って、「主義による統計学」ではなく「学問としての統計学」を確立しました。

☆　今日でも論争のための論争を続けている人がいますが意味はありません。

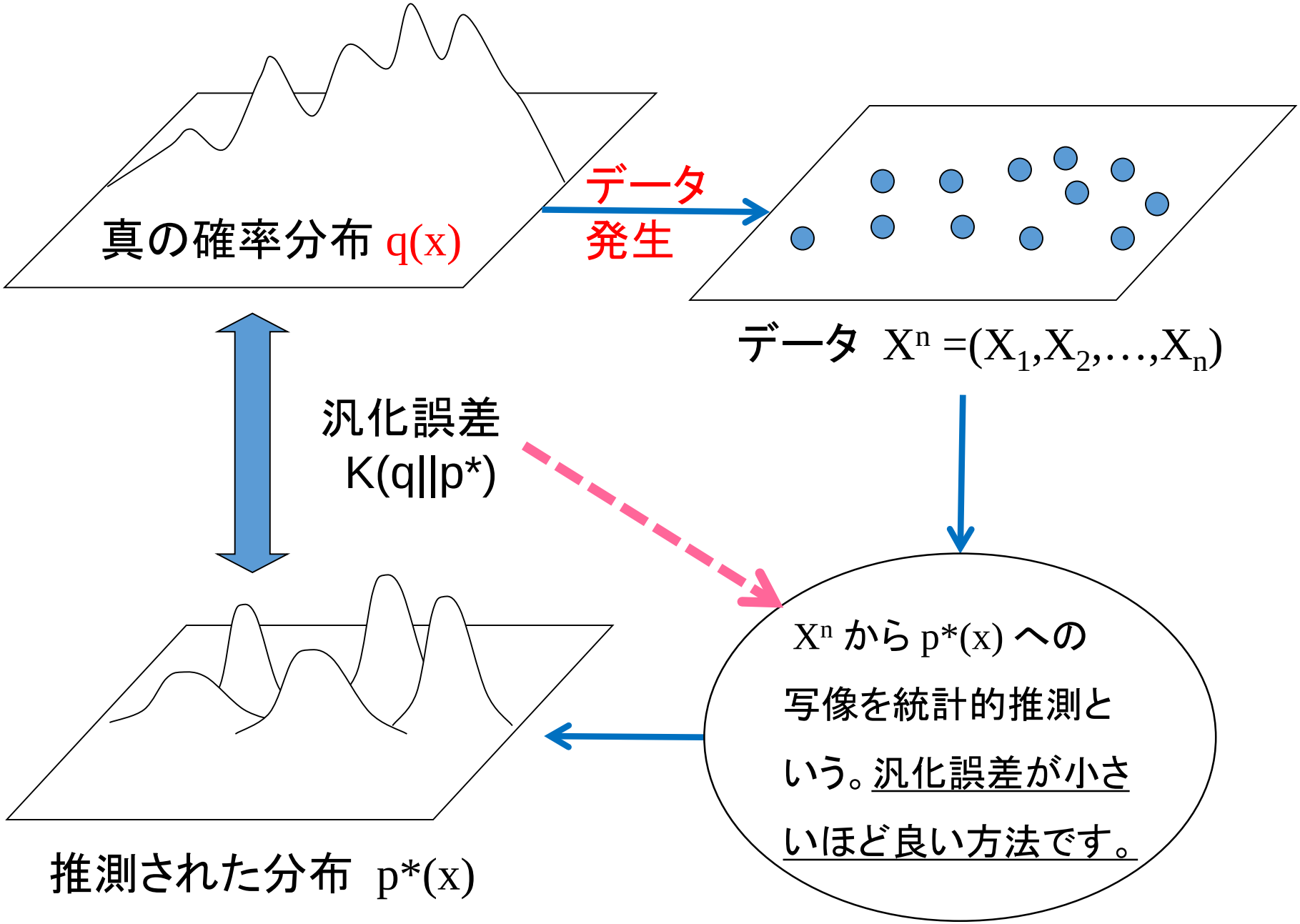
☆　赤池博士の自然な着想の基盤には、そのことを実現可能にした**数理科学の存在**がとても大切です。実際、【未知である真の分布に対する統計的推測の適切さを定量的に測ること】は数学的法則によって初めて可能になることであり、それまでは、「できないことを考えてもしかたない」と思われていたのでした。

カルバック・ライブラ距離

定義. $\mathbf{R}^N$ 上の確率密度関数 $q(x), p(x) > 0$ が与えられたとき
$K(q||p) = \int q(x) \log (q(x)/p(x))\, dx$ を $q(x)$ と $p(x)$ の
カルバック・ライブラ情報量(KL情報量）という。

定理. 連続な確率密度 $q(x), p(x)$ について、次が成り立つ。
(1) 任意の $q(x), p(x)>0$ について $K(q||p) \geqq 0$.
(2) $K(q||p)=0 \Leftrightarrow q(x)=p(x)\ (\forall x)$

未知である真の分布 $q(x)$ から得られたデータに対して何らかの方法で推測を行って得た結果 $p^*(x)$ の適切さを測る量としてよく用いられるのがKL情報量です。$D(q||p^*)$ のことを汎化誤差と呼びます。汎化誤差を小さくするための推測法とその理論を作ることが統計学と機械学習の最大の目標です。


真の確率分布 q(x)
データ
発生
データ $X^n = (X_1, X_2, \ldots, X_n)$
汎化誤差
$K(q||p^*)$
X^n から $p^*(x)$ への
写像を統計的推測と
いう。汎化誤差が小さ
いほど良い方法です。
推測された分布 $p^*(x)$

ベイズ推測の定義

事前分布と事後分布

パラメータ集合の上の確率密度関数 $\varphi(w)$ を事前密度関数という。

仮説 「パラメータ w が事前密度関数 $\varphi(w)$ から発生し、データ X^n がモデル $p(x|w)$ から独立に発生した」のもとでは、(w,X^n) の同時密度関数は

$$\varphi(w) \prod_{i=1}^{n} p(X_i|w)$$

である。データX^nが与えられたときのパラメータの条件つき密度関数は

$$p(w|X^n) = (1/Z)\, \varphi(w) \prod_{i=1}^{n} p(X_i|w)$$

である。これを事後密度関数という。ここで

$$Z = \int \varphi(w) \prod_{i=1}^{n} p(X_i|w)\, dw$$

は X^n の周辺密度関数である。

ベイズ推測

真の密度関数 q(x) を推測する密度関数 p*(x) を次式で定義する。

ベイズ推測. 統計モデルを事後分布で平均して推測結果とする.

$$p^*(x) = \int p(x|w)\, p(w|X^n)\, dw.$$

推測された密度関数 p*(x) を予測分布と呼ぶ。

☆ p*(x) は推測されたものであり真のものではない。
☆ 推測のよさは、K(q||p*) を比べることで比較できる。
☆ K(q||p*) は確率変数であるが、その挙動は数学的に解明されている。
☆ 事前分布 φ(w) がパラメータ a を持ち φ(w|a) と書かれるときがある。このとき a をハイパーパラメータと呼ぶ。p*(x) ができるだけ q(x) と近くなるように (q(x)が不明であるにも関わらず）ハイパーパラメータを決める方法がある。

推測法3種類

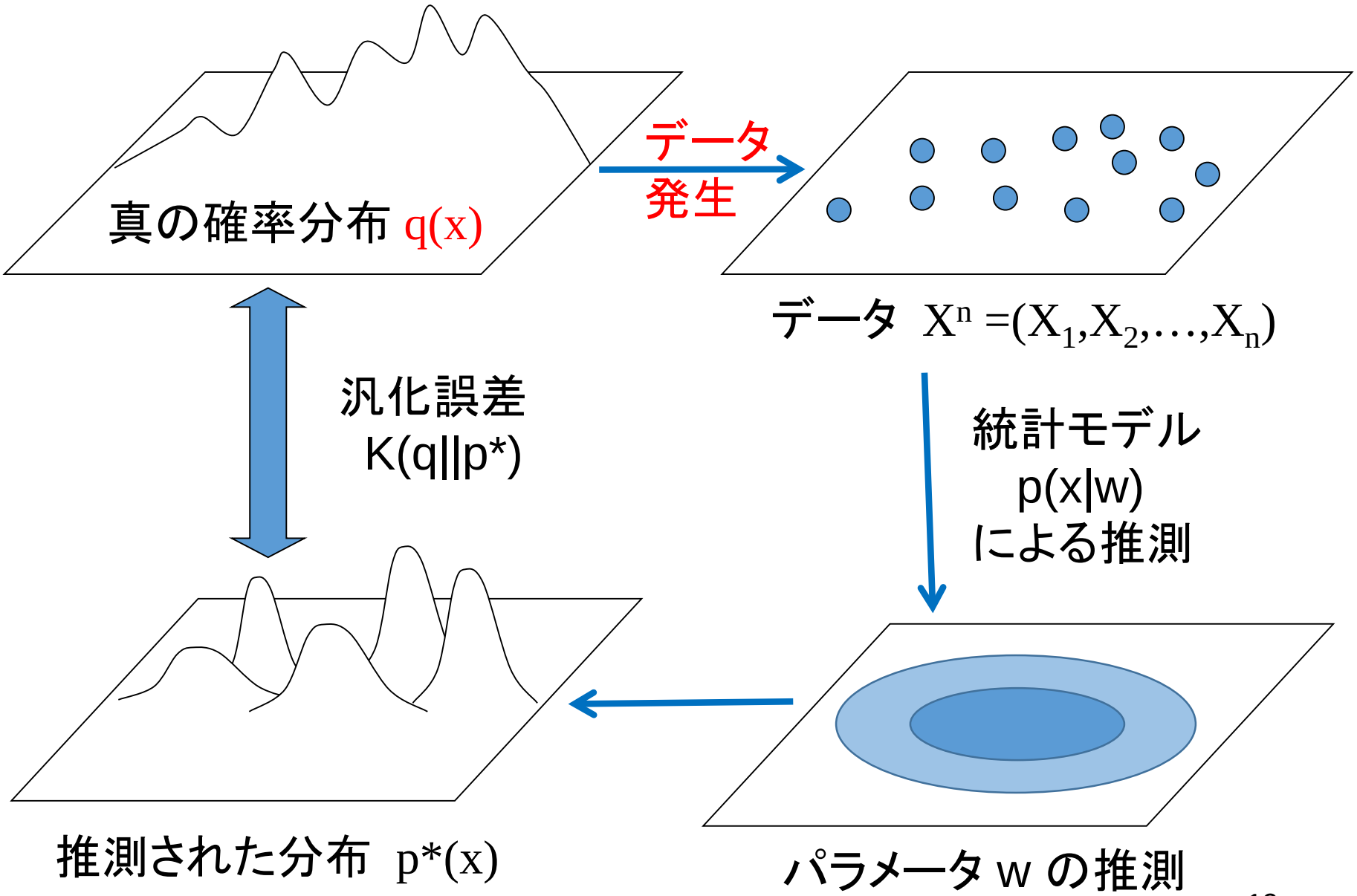
真の密度関数 $q(x)$ を推測する密度関数 $p^*(x)$ を定める方法としてよく用いられるものに3つのものがある。

(2) 事後確率最大化推測. $p(w|X^n)$ を最大にする w^*（事後確率最大化推定量）を用いて $p(x|w^*)$ を推測結果とする。

(3) 最尤推測. 事前分布をパラメータ集合上で一定値としたときの w^*（最尤推定量）を用いて $p(x|w^*)$ を推測結果とする。

最尤推定量は簡単なモデルに対しては計算が容易であるため、期末試験や資格試験において出題されることが多いが、統計的推測の方法としての精度はよくないことが知られている。特に機械学習のように多くのパラメータを持つモデルに最尤法を適用すると予測精度が著しく劣化するので注意しましょう。また簡単なモデルでは最尤推定量は $n \to \infty$ の漸近極限で正規分布に法則収束するが、これも機械学習で現れるモデルにおいては成立しません。今日の統計学では数学的に未完成な課題がたくさんあります。

ベイズ推測の手順


真の確率分布 q(x)
データ
発生
データ $X^n = (X_1, X_2, \ldots, X_n)$
汎化誤差
K(q||p*)
統計モデル
p(x|w)
による推測
推測された分布 p*(x)
パラメータ w の推測

真の分布と推測された分布

事前密度関数 $\varphi(w)$ も統計モデル $p(x|w)$ も人間が仮に定めたものに過ぎないので、そこから定義される予測密度関数 $p^*(w)$ は、真の密度関数 $q(x)$ とは同じではない。汎化損失を次の式で定義する。

$$-\int q(x) \log p^*(x)\, dx$$

汎化損失は $p^*(x)=q(x)$ のときに限り最小値（真の分布のエントロピー）

$$-\int q(x) \log q(x)\, dx$$

をとる。汎化損失はデータ X^n に依存する確率変数であるが、その平均値を平均汎化損失という。統計的推測の適切を調べるときには $K(q\|p^*)$ の代わりに汎化損失を用いることができる（真の分布のエントロピーだけ異なる）。

統計モデリングの基礎実験

現実の問題では真の分布は不明であるが、現実の問題を考えるよりも前に推測方法や統計モデルがどのような性質を持っているのかを数学的に明らかにしておく必要がある。その場合には、真の分布を人間が設定し、データをランダムに発生し、定められた推測法・統計モデル・事前分布を用いて推測を行い、その結果と真のモデルとの汎化誤差を求める。データがランダムにばらつくので汎化誤差もランダムにばらつく。汎化誤差の挙動を比べれば、推測方法・統計モデル・事前分布の選択が推測にどのような影響を及ぼすのかが明らかになる。このことを理論・実験により、十分に調べておいてから現実の問題への適用を考えるのである。

問1

$x, w \in \mathbf{R}$ である。次の三組に対して、真の分布のエントロピー、事後分布、ベイズ予測分布、ベイズ汎化誤差、ベイズ平均汎化誤差を求めよ。

$$\begin{cases} \text{統計モデル } p(x|w) = 1/(2\pi)^{1/2} \exp(-(x-w)^2/2) \\ \text{事前分布} \quad \varphi(w) = 1/(2\pi)^{1/2} \exp(-w^2/2) \\ \text{真の分布} \quad p(x|0) \end{cases}$$

ただしデータを $\{X_1, X_2, \ldots, X_n\}$ とする。次の式を用いてよい。

平均 a 分散 s^2 の正規分布: $G(x|a,s^2) = 1/(2\pi s^2)^{1/2} \exp(-(x-a)^2/2s^2)$

$\Rightarrow$ $G(x|a, s^2)=G(a|x, s^2)$, $\int G(x|a,s^2)\, G(x|b,t^2)\, dx = G(a|b,s^2+t^2)$.

問1の解き方（A,B,Cを求めるとよい）

確率変数 $x^* = 1/(n+1) \sum_{i=1}^{n} X_i$ とおく。$E[(x^*)^2] = n / (n+1)^2$.

$\varphi(w) \prod_{i=1}^{n} p(X_i|w) \propto \exp(- (n+1) w^2/2 + (n+1) x^* \cdot w)$ だから

(1) 事後分布は $p(w|X^n) = G(w \mid x^*, 1/(n+1))$.

(2) 予測分布は $p^*(x) = \int p(x|w)\, p(w|X^n)\, dw$

$$= \int G(w|x,1)\, G(w \mid x^*,1/(n+1))\, dw = G(x|A,B^2)$$

(3) 汎化誤差は $D(q||p^*) = E_X[\log \{ q(X) / p^*(X) \}]$

$= E_X[\log \{ G(X|0,1) / G(X|A,B^2) \}]$

$=E_X[\log B + (1/2)\{ – X^2 + (X-A)^2/B^2 \}] = \log B + (1/2)(-1+(1+A^2)/B^2)$

(4) 平均汎化誤差は

$E[D(q||p^*)] =E [\log B + (1/2)(-1+(1+A^2)/B^2)] = C$.

ベルヌーイ分布の例

$x \in \{0,1\}$ とする。ベルヌーイ分布　$p(x|a) = a^x(1-a)^{1-x}$

コインを振ったら確率 a で表(1)がでる。

事前分布

ベルヌーイ分布　$p(x|a) = a^x(1-a)^{1-x}$ を用いた推測において、
事前分布としてディリクレ分布（この場合はベータ分布ともいう）

$$\varphi(a) = C\, a^{\alpha-1}(1-a)^{\beta-1} \quad (0 \leqq a \leqq 1), \qquad \alpha,\beta>0,$$

を用いる場合を考える。ここで C は定数であり

$$1/C=B(\alpha,\beta) = \int_0^1 a^{\alpha-1}(1-a)^{\beta-1}\, da$$

はベータ関数である。定数(α,β)はハイパーパラメータである。

（注意）ベータ関数は次の恒等式を満たす。

$$B(\alpha,\beta) = \Gamma(\alpha)\,\Gamma(\beta) / \Gamma(\alpha+\beta) = B(\beta,\alpha)$$

$$B(\alpha+1,\beta) = \alpha/(\alpha+\beta)\, B(\alpha,\beta)$$

ベルヌーイ分布の推測

<u>定理</u>. ベルヌーイ分布 $p(x|a) = a^x(1-a)^{1-x}$ の事前分布としてディリクレ分布 $\varphi(a) = C\, a^{\alpha-1}(1-a)^{\beta-1}$ $(0 \leqq a \leqq 1)$ を用いる。データ $X^n = (X_1, X_2, \ldots, X_n)$ が得られたとき $n_1 = \Sigma X_i$, $n_2 = n - n_1$ とおく。

(1) <u>ベイズ推測の予測分布</u>は

$$p^*(1) = (n_1+\alpha) / (n+\alpha+\beta), \quad p^*(0) = (n_2+\beta) / (n+\alpha+\beta).$$

(2) <u>事後確率最大化推測(MAP)の予測分布</u>は

$$p^*(1) = (n_1+\alpha-1) / (n+\alpha+\beta-2), \quad p^*(0) = (n_2+\beta-1) / (n+\alpha+\beta-2).$$

(3) <u>最尤推測の予測分布</u>は $p^*(1) = n_1 / n$, $p^*(0) = n_2/n$.

この定理の導出は、このファイルの最後に載せる。

（注意）真の分布が $q(x)$ で推測結果が $p^*(x)$ であるとき、汎化誤差は $K(q||p^*) = q(1) \log(q(1) / p^*(1)) + q(0) \log (q(0) / p^*(0))$ であるから、定理から汎化誤差が計算できる。

計算の手順

(1) モデル$p(x|a) = a^x(1-a)^{1-x}$ 事前分布$\varphi(a) = C\, a^{\alpha-1}(1-a)^{\beta-1}$

(2) 真の分布 $q(x)$ を設定

(3) $q(x)$ からデータ $X^n = (X_1, X_2, \ldots, X_n)$ を生成

(4) $n_1 = \Sigma X_i$, $n_2 = n - n_1$ とおく

(5) ベイズ: $p^*(1) = (n_1+\alpha) / (n+\alpha+\beta)$, $p^*(0) = (n_2+\beta) / (n+\alpha+\beta)$.

(6) MAP: $p^*(1) = (n_1+\alpha-1) / (n+\alpha+\beta-2)$, $p^*(0) = (n_2+\beta-1) / (n+\alpha+\beta-2)$.

(7) 最尤: $p^*(1) = n_1 / n$, $p^*(0) = n_2/n$.

(8)それぞれの推測に対して汎化誤差を計算

$$K(q||p^*) = q(1) \log(q(1) / p^*(1)) + q(0) \log (q(0) / p^*(0))$$

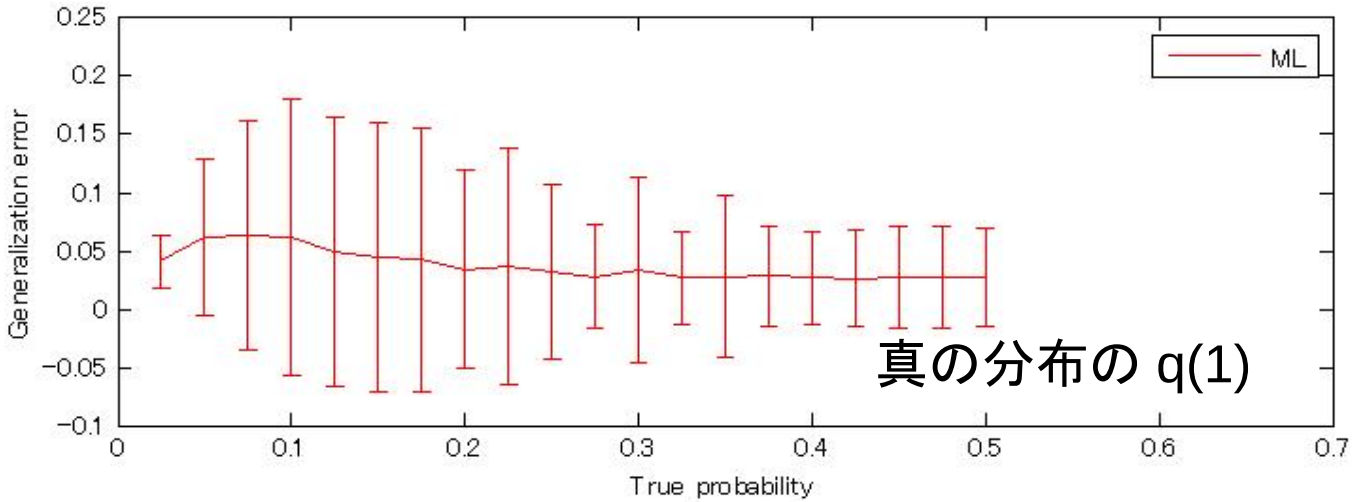
汎化誤差を計算してみた。

同じ真の分布に対してもデータがばらつくから汎化誤差もばらつく。その平均と標準偏差を示した。

最尤推定の

汎化誤差

n=20



事後確率

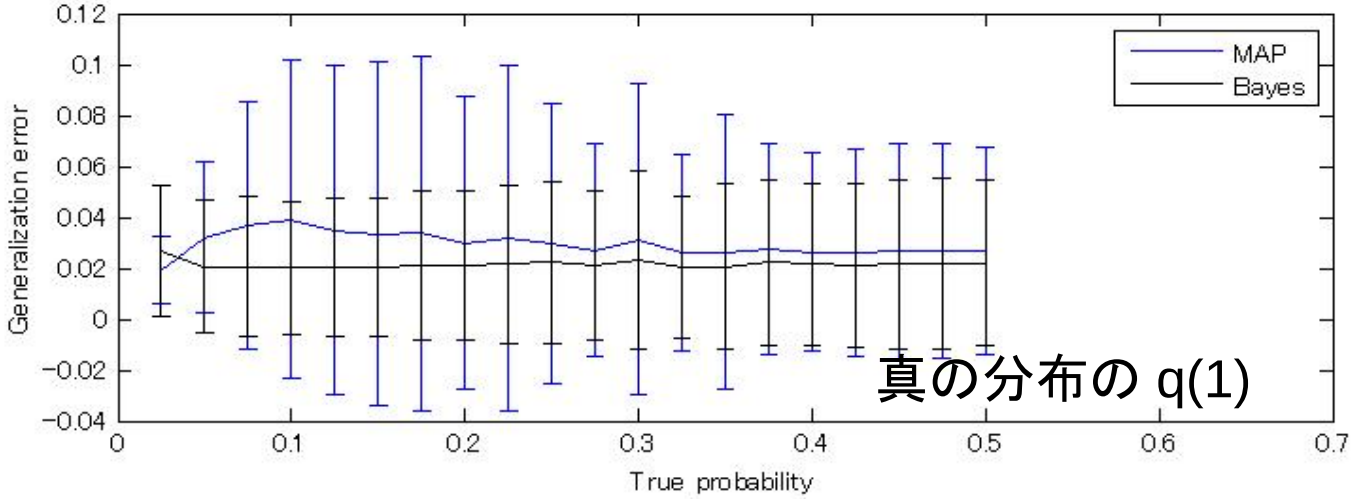
最大化推定と

ベイズ推定

の汎化誤差

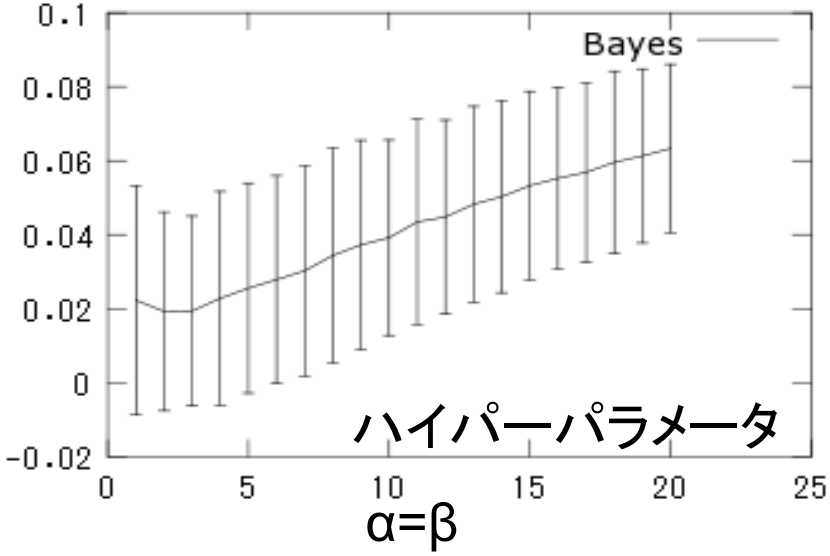
α＝β=1.1,

n=20

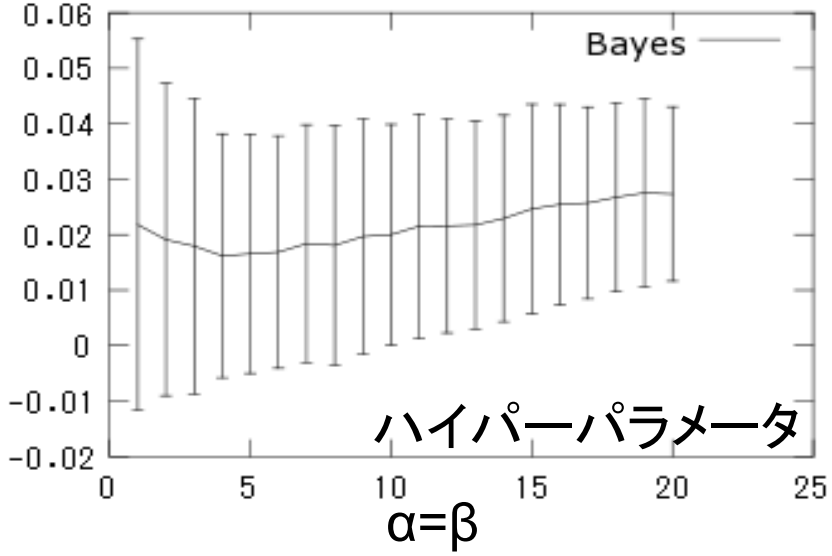


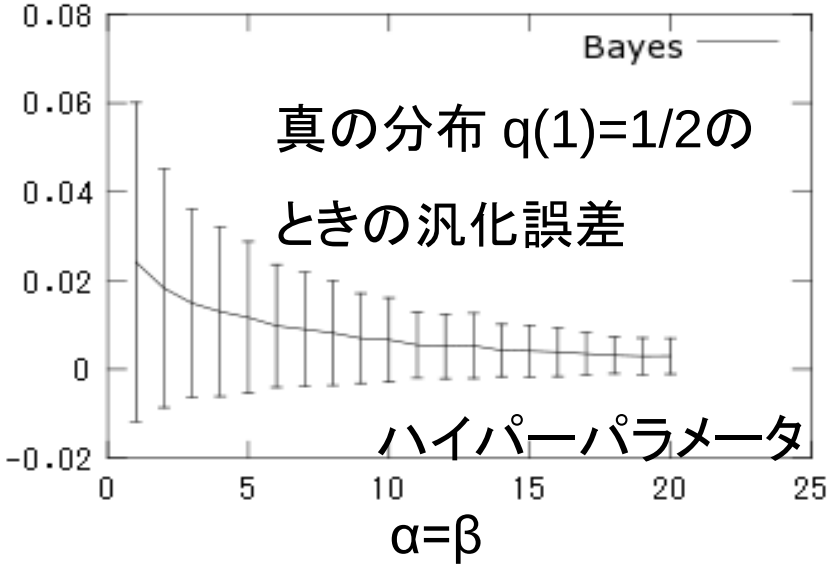
参考:ベイズ汎化誤差とハイパーパラメータ

真の分布 q(1)=1/4のときの汎化誤差



真の分布 q(1)=1/3のときの汎化誤差





ベイズ法の精度はハイパーパラメータに依存する。最適なハイパーパラメータは真の分布に依存する。真の分布が不明のときハイパーパラメータを定める方法については「モデルの評価」の回に説明する。

問２

(1,0)がでる確率が(q,1-q)のコインを n 回ふって1がでた回数を n_1 とする。

定数 $\alpha \geqq 0$ について $p^*=(n_1+\alpha)/(n+2\alpha)$ と定義する。

このコインを1回振ったときの平均と分散は (q,q(1-q)) である。

(1) 確率変数 n_1 の平均と分散を求めよ。

(2) 確率変数 p^* の平均と分散を求めよ。

(3) 確率変数 $(q-p^*)^2$ の平均を求めよ。[ヒント : $q-p^* = q-E(p^*)+E(p^*)-p^*$]

事後分布と事後平均

MCMC法による事後分布の実現

ベルヌーイ分布では事後分布と予測分布を解析的に求めることができたが、一般的に事後分布や予測分布の計算では、解析的な計算はできないことのほうが多い。その場合に用いられるのがマルコフ連鎖モンテカルロ法(MCMC法）である。積分計算

$$p^*(x) = \int p(x|w)\, p(w|X^n)\, dw$$

の実行が困難であるとき、 $p(w|X^n)$ に従う $\{w_k\}$ をたくさん(K個）採取し、

$$p^*(x) = (1/K)\, \Sigma_k\ p(x|w_k)$$

によって数値的に近似する。

いろいろなMCMC法

MCMC法は、原理的には任意の形の事後分布を実現することができるので今日の統計モデリングの発展の基礎になっている。最近では極めて高度な手法も実装されデータサイエンスの現場で使われている。代表的な方法は次の通り。

1. メトロポリス法
2. ギブスサンプラー法
3. ハミルトニアン法
4. ランジュバン法

この他にも、局所的な分布に限定されないようにする方法なども考案されている。

ギブスサンプラー法

確率密度関数 $p(x,y)$ に従う $\{(x_k,y_k);k=1,2,\ldots,K\}$ を生成するために次の方法を繰り返して使う。

(1) 初期値 y を決める。

(2) x を $p(x|y)$ からサンプリング

(3) y を $p(y|x)$ からサンプリングして (2) に戻る。

初期値の影響が消えるまでの区間(Burn-in)は捨てて、それ以後のものを使う。

このとき、サンプリングする個数 K が無限大になる極限で任意の関数 $f(x,y)$ で

$$(1/K)\ \Sigma\ f(x_k,y_k) \rightarrow \int f(x,y)\ p(x,y)\ dx\ dy$$

が成り立つ。（ただし右辺が有限なとき）。

回帰分析の例

統計モデルと事前分布

$N(0,1/s)$ を平均 0, 分散 1/s の正規分布を表すものとする。

回帰モデル $Y=aX+b+N(0,1/s)$ を表す確率密度関数は

$$p(y|x,a,b,s) = (s / 2\pi)^{1/2} \exp(-s(y-ax-b)^2/2).$$

事前分布として次のものを用いる。

$$\varphi(a,b) = (t / 2\pi) \exp(- t (a^2+b^2)/2)$$

$$\psi(s)= \varepsilon \exp(- \varepsilon s)$$

ただし ε は十分に小さく設定する。ハイパーパラメータ t の変化が汎化誤差に及ぼす影響を調べてみよう。

事後分布

パラメータ (a,b,s) についての事後分布 p(a,b,s|データ) は定義から

$$p(a,b,s|\text{データ}) = \varphi(a)\,\psi(s)\,\Pi_i\, p(Y_i|X_i,a,b,s)$$

$$= \varepsilon\, s \exp(-\varepsilon\, s) \times (t / 2\pi) \exp(-t\,(a^2+b^2)/2)$$

$$\times\ \Pi_i\, (s / 2\pi)^{1/2} \exp(-s(Y_i-aX_i-b)^2/2)$$

となる。ギブスサンプラーを実行するためには次の分布を求めればよい。

$$p(s|a,b,\text{データ}) \propto \exp(-\varepsilon\, s) \times \Pi_i\, s^{1/2} \exp(-s(Y_i-aX_i-b)^2/2)$$

$$p(a,b|s,\text{データ}) \propto \exp(-t\,(a^2+b^2)/2)\, \Pi_i \exp(-s(Y_i-aX_i-b)^2/2)$$

事後分布(s)

パラメータ s の確率密度関数は

$$p(s|a,b,\text{データ}) \propto \exp(-\varepsilon s) \times \Pi_i s^{1/2} \exp(-s(Y_i-aX_i-b)^2/2)$$

$$= s^{n/2} \exp\{ -s (\varepsilon + \Sigma_i (Y_i-aX_i-b)^2/2) \}$$

従って　$p(s|a,b,\text{データ}) = G(s|1+n/2, \varepsilon + \Sigma_i (Y_i-aX_i-b)^2/2)$

ここでガンマ分布を用いた。　$G(x|\alpha,\beta) = 1/(\beta^\alpha \Gamma(\alpha)) x^{\alpha-1} \exp(-x/\beta)$

ガンマ分布に従う s は直接サンプリングできる（ソフトウエアのライブラリ）。

事後分布((a,b))

パラメータ (a,b) の確率密度関数は

$$p(a,b|s,\text{データ}) \propto \exp(-t(a^2+b^2)/2)\,\Pi_i \exp(-s(Y_i-aX_i-b)^2/2)$$

$$\propto \exp(-(1/2)\{H_{11}a^2 + H_{12}ab + H_{21}ba + H_{22}a^2 - 2v_1a - 2v_2b\})$$

$$\propto \exp(-(1/2)\| H^{1/2}\{(a,b) - H^{-1}v\}^T \|^2)$$

ここで行列 H とベクトル $v = (v_1,v_2)^T$ は次のように定義した。

$$H = \begin{pmatrix} t + s\,\Sigma_i X_i^2 & s\,\Sigma_i X_i \\ s\,\Sigma_i X_i & t + n \end{pmatrix} \qquad v = (v_1,v_2)^T = \begin{pmatrix} s\,\Sigma_i X_i Y_i \\ s\,\Sigma Y_i \end{pmatrix}$$

従って $N_2(u,S)$ を平均 u 分散共分散行列が S の2次元正規分布とすると

$$p(a,b|s,\text{データ}) = N_2(H^{-1}v, H^{-1})$$

正規分布に従う (a,b) は直接サンプリングできる（ソフトウエアのライブラリ）。

計算の手順

(1) データ $\{(X_i,Y_i)\ ;\ i=1,2,\ldots,n\}$ を得る。

(2) 最尤推定量 (a^*,b^*) を計算（注意）し、(a,b) の初期値とする。

(3) S を $G(\ s\ |1+n/2,\ \ \varepsilon+\Sigma_i\ (Y_i-aX_i-b)^2/2\ \ )$からサンプリング。

(4) 行列 H とベクトル v を計算。

$$H = \begin{pmatrix} t + s\ \Sigma_i\ X_i^2 & s\ \Sigma_i\ X_i \\ s\ \Sigma_i\ X_i & t + sn \end{pmatrix} \qquad v = \begin{pmatrix} s\ \Sigma_i\ X_iY_i \\ s\ \Sigma\ Y_i \end{pmatrix}$$

(5) (a,b) を $N_2(H^{-1}v,\ H^{-1})$ からサンプリングし(3)に戻る。

(6) 十分に繰り返した後、初期値の影響の分を削除する。

(7) 事後分布に従うパラメータの集合 $\{(a_k,b_k,s_k);\ k=1,2\ldots,K\}$ を得る。

(注意）最尤推定量 $(a^*,b^*)^T$ は、$t=0,s=1$ のときの $H^{-1}v$ と等しい。

実験の例

現実のデータはモデルに従っているわけではないが、ここでは

Xは区間[0,1]の一様分布で、Yは p(Y|X,1,0,25) の場合を考える。

独立なデータ$\{(X_i,Y_i)$; i=1,2,…,10} を得た。(a,b) の事前分布

$$\varphi(a) = (t / 2\pi) \exp(- t (a^2+b^2)/2)$$

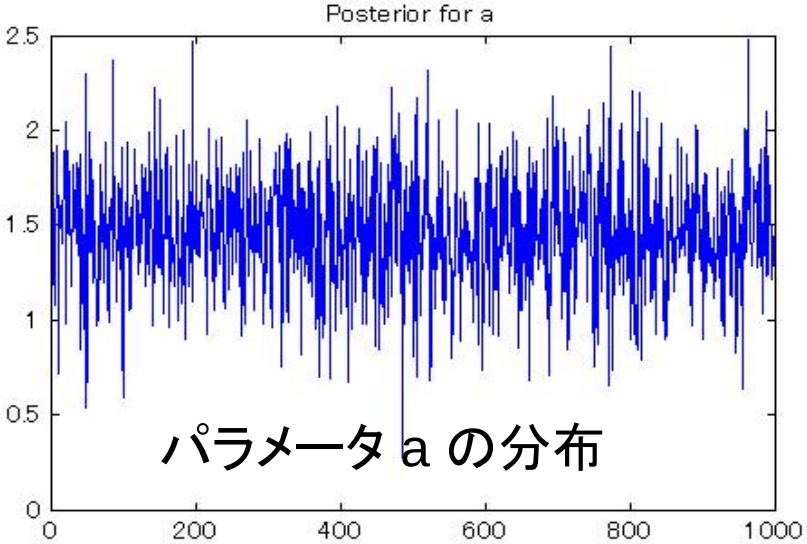
の中のハイパーパラメータ t が推測に与える影響を考察する。

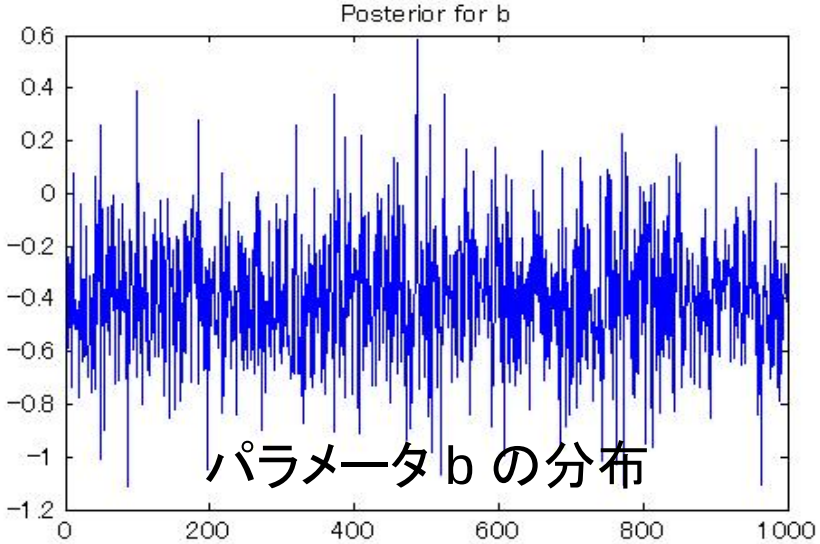
汎化誤差の計算は次のようにした。(A,B) を(a,b)の事後分布での平均とし、(a*,b*)を最尤推定量とする。

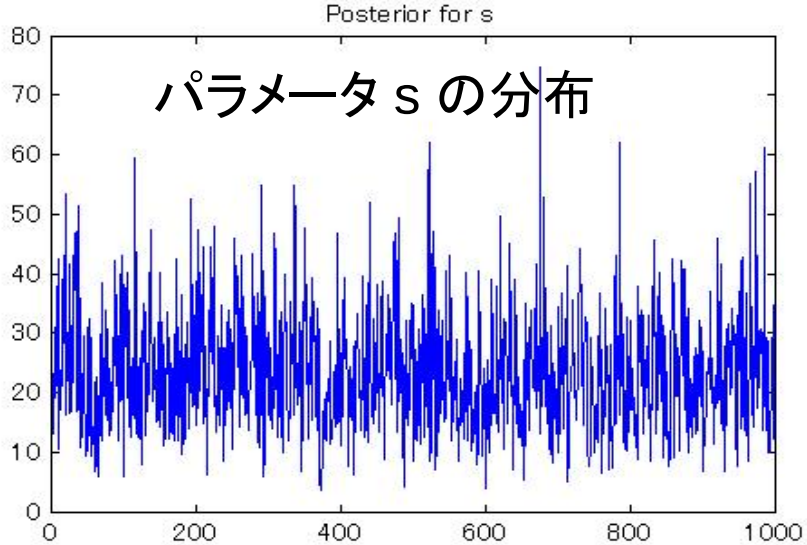
ベイズ法の汎化誤差 = $E_{(X,Y)} \log(q(Y|X)/p^*(Y|X))$

最尤法の汎化誤差 = $E_{(X,Y)} \log(q(Y|X)/p(Y|X,a^*,b^*))$

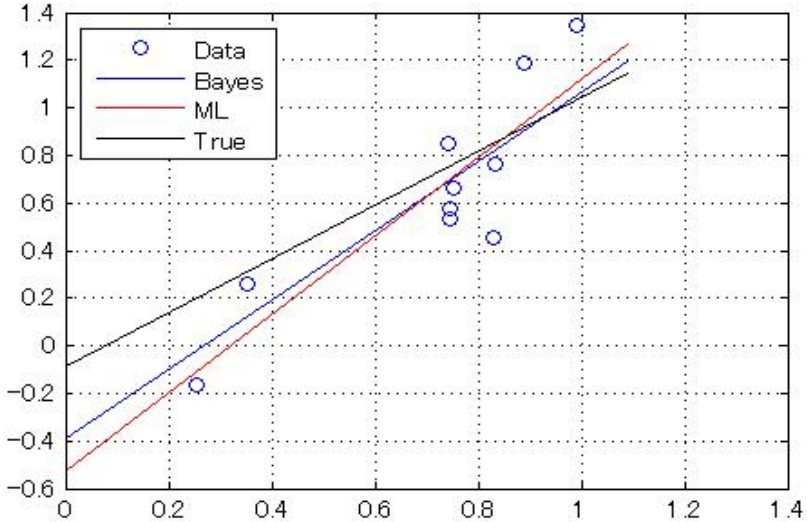
実験の例







真と推測(n=10, t=1)



実験例

問3

次の密度関数についてのギブスサンプラーを作れ。

$$p(x,y) \propto x^3 \exp(-xy^2/2+y) \qquad x>0,\ y\in\mathbf{R}.$$

ただし ガンマ分布　$G(x|\alpha,\beta) =1/(\beta^{\alpha}\Gamma(\alpha))\, x^{\alpha-1}\exp(-x/\beta)$

正規分布　$N(y|a,s^2) = 1/(2\pi s^2) \exp(- (y-a)^2/(2s^2))$

のサンプリング関数が利用できるものとする。

付録:ベルヌーイ分布の予測分布

ベルヌーイ分布の予測分布の導出

証明. まず事後確率最大化と最尤推定を求める。事後分布は

$$p(a|X^n) = (C/Z)\, a^{\alpha-1}(1-a)^{\beta-1} \prod_{i=1}^{n} a^{x_i}(1-a)^{1-x_i}$$

$\Sigma_i x_i = n_1$ とおき $n_2=n-n_1$ とおくと

$$p(a|X^n) = (C/Z) \exp \{ (n_1+\alpha-1) \log a + (n_2+\beta-1) \log (1-a) \}$$

これを最大にする a は $a^*=(n_1+\alpha-1) / (n+\alpha+\beta-2)$である。従って

事後確率最大化推定量 は $a^*=(n_1+\alpha-1) / (n+\alpha+\beta-2)$.

また最尤推定量は $a^*= n_1/ n$.

導出つづき

次にベイズ推定を求める。

モデル $p(x|a) = a^x(1-a)^{1-x}$ と事前分布 $\varphi(a) = C\, a^{\alpha-1}(1-a)^{\beta-1}$ による予測分布は

$$p^*(x) = \int p(x|a)\, p(a|X^n)\, da$$

$$= (C/Z) \int p(x|a) \exp\{ (n_1+\alpha-1) \log a + (n_2+\beta-1) \log (1-a) \}\, da$$

$$= (C/Z) \int \exp\{ (x+n_1+\alpha-1) \log a + (1-x+ n_2+\beta-1) \log (1-a) \}\, da$$

$$= (C/Z)\, B(x+n_1+\alpha, 1-x + n_2+\beta).$$

従って

$$p^*(1)= (C/Z)\, B(1+n_1+\alpha, n_2+\beta) =(C/Z)\, (n_1+\alpha) / (n+\alpha+\beta)\, B(n_1+\alpha, n_2+\beta)$$

$$p^*(0)= (C/Z)\, B(n_1+\alpha, n_2+\beta+1) =(C/Z)\, (n_2+\beta) / (n+\alpha+\beta)\, B(n_1+\alpha, n_2+\beta)$$

確率の和が1であるから

$$p^*(1)= (n_1+\alpha) / (n+\alpha+\beta), \quad p^*(0)= (n_1+\beta) / (n+\alpha+\beta).$$

(証明終)