データ解析

第9回：ベイズ法

渡辺澄夫

ベイズ法と最尤法

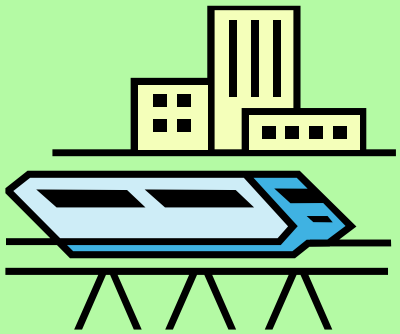
観測データ

回帰・判別分析

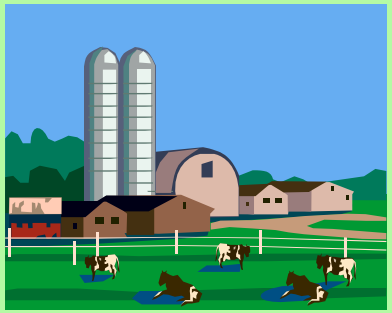
推定と検定

ベイズ法
・階層ベイズ法

解析方法



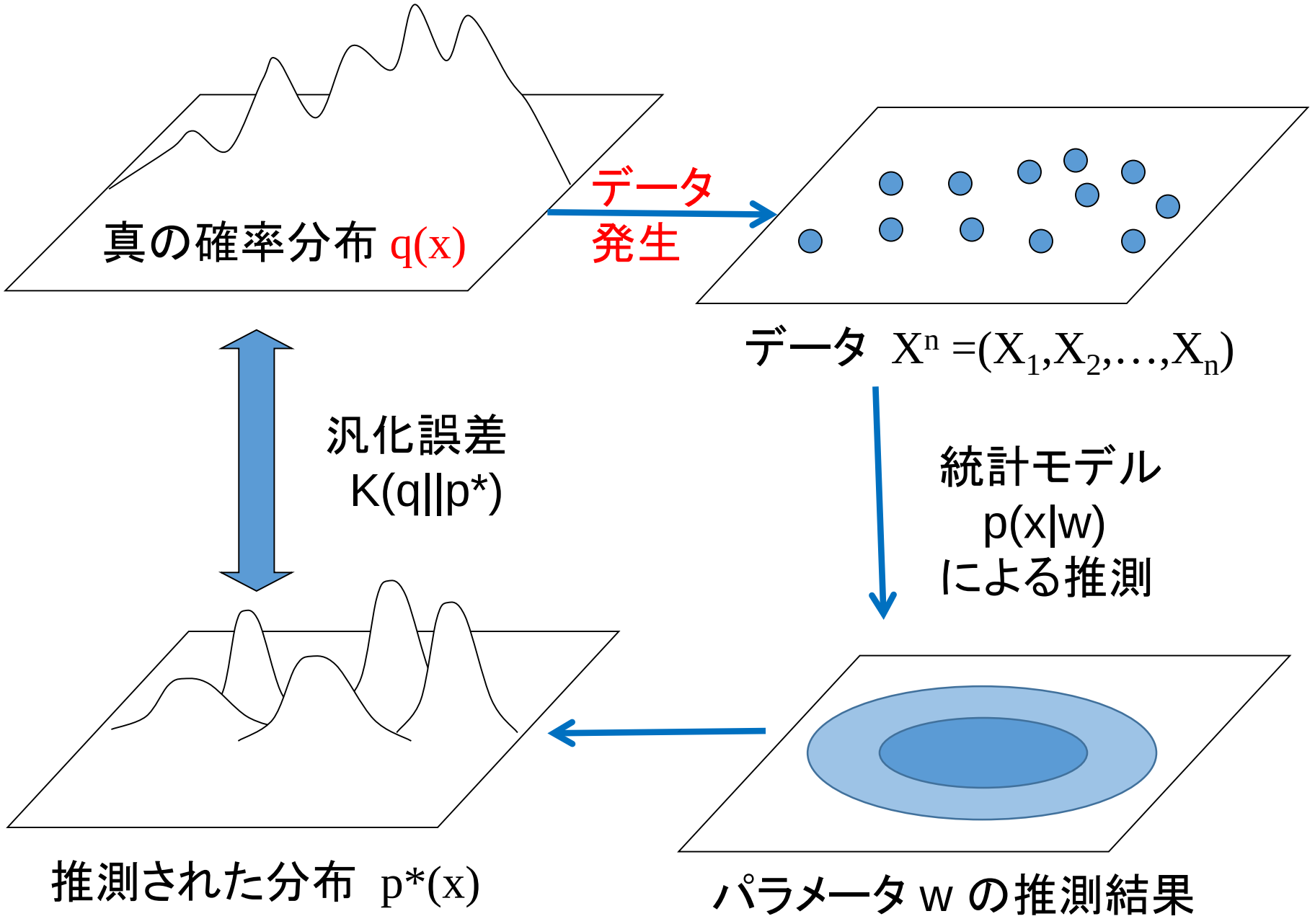
時系列予測



主成分・因子
・クラスタ分析

復習：統計的推測とは

データ　$X^n=(X_1,X_2,\ldots,X_n)$ が得られたとき、このデータを発生した確率密度関数 $q(x)$ を知りたい。しかし確率分布の集合は一般には無限次元であり、データは有限であるから $q(x)$ を完全に知ることはできない。そこでパラメータ w をもつ統計モデル $p(x|w)$ を使って推測を行なう。


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

復習：カルバック・ライブラ情報量

定義. 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という。

定理. 関数 $q(x), p(x)$ が連続であるとする。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) $K(q||p)=0 \Leftrightarrow q(x)=p(x)\ (\forall x)$

統計的推測

事前分布と事後分布

パラメータ集合の上に定義された確率分布を事前分布という。

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

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

であるから、データ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。

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標準的に用いられるものに以下のものがある。

(1) 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.

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

(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。

推測された密度関数 p*(x) を予測分布と呼ぶ。

注意：仮説と推測

仮説「パラメータ w が事前密度関数 φ(w) から発生し、データ X^n がモデル p(x|w) から発生した」のもとでは、事後密度関数 $p(w|X^n)$ を用いて推測を行うことが最も自然である。またKL情報量の意味で最も精度の良い推測を与えることも知られている。しかしながら、現実には事前密度関数もモデルも人間が与えたものに過ぎないので、そこから定義される事後密度関数も適切な推測を与えているわけではない。

統計的推測とは、人間が設定した統計モデルと事前分布を用いて真の分布 q(x) を推測することである。推測の結果はどれも q(x) と等しくはない。
与えられた推定法に対してその推定法の精度を解明することが必要になります。

統計モデリングとは

現実の問題ではデータだけが与えられて統計モデルも事前分布も不明である。そこでデータ解析者は、まず回帰・判別・主成分・因子・クラスタ分析を行ってデータのおおよその様子を考察する。次に統計モデルと事前分布をモデリングして真の分布を推測する。推測の結果が適切であれば（注意）予測分布を結論とする。推測の結果が適切でなければ再度モデリングを繰り返す。

（注意）真の分布が不明であるにも関わらずモデリングの結果が適切であるかどうかについて調べる方法がある（モデルの評価・検定で述べる）。

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。

ベルヌーイ分布の例

事前分布

ベルヌーイ分布　$p(x|a) = a^x(1-a)^{1-x}$ の用いた推測において、
事前分布としてディリクレ分布

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

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

定理. 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) ベイズ推測の予測分布は

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

(2) 事後確率最大化推測(MAP)の予測分布は

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

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

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

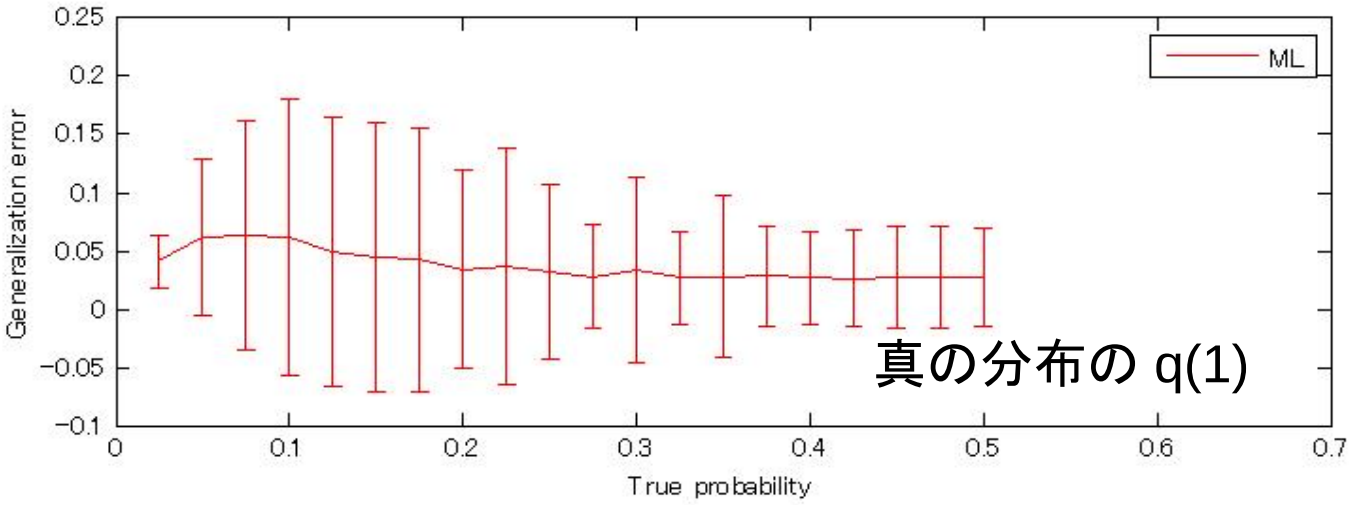
(8)それぞれの推測に対して汎化誤差を計算

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

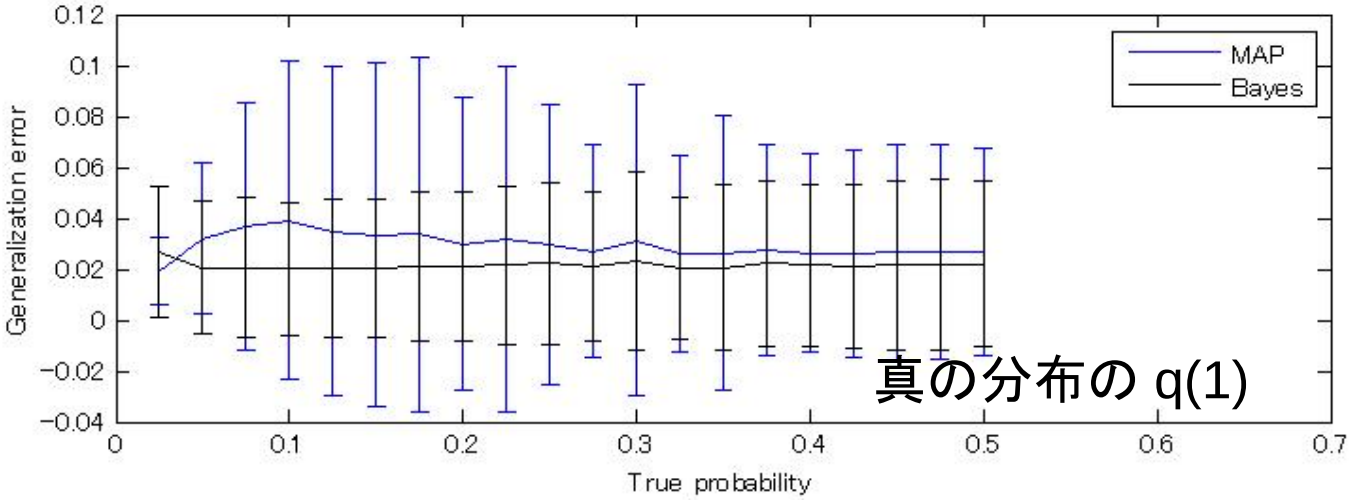
汎化誤差を計算してみた。

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

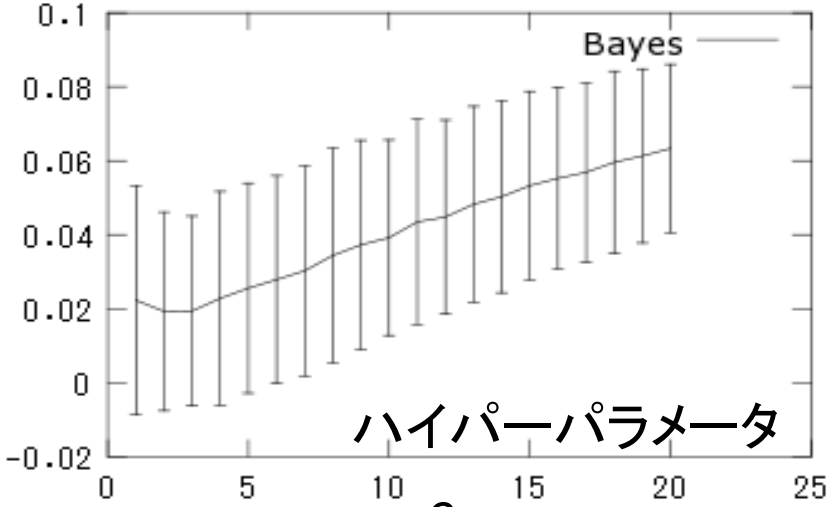


事後確率
最大化推定と
ベイズ推定
の汎化誤差
α＝β=1.1,
n=20

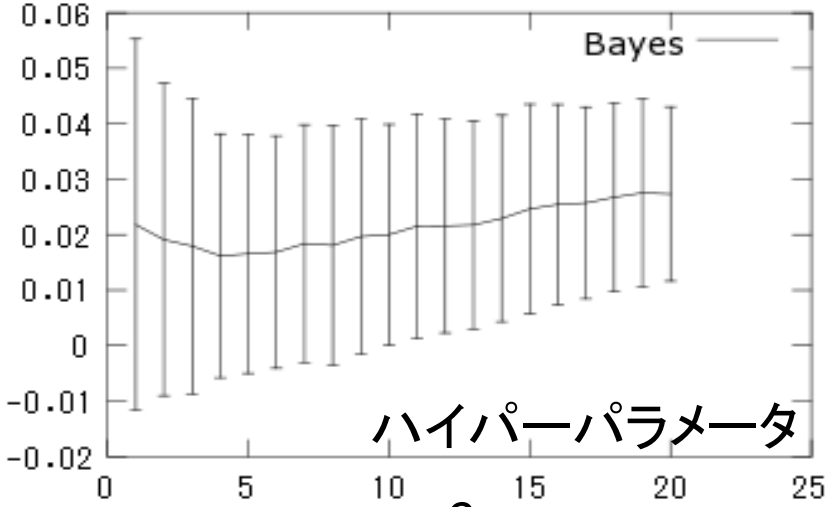


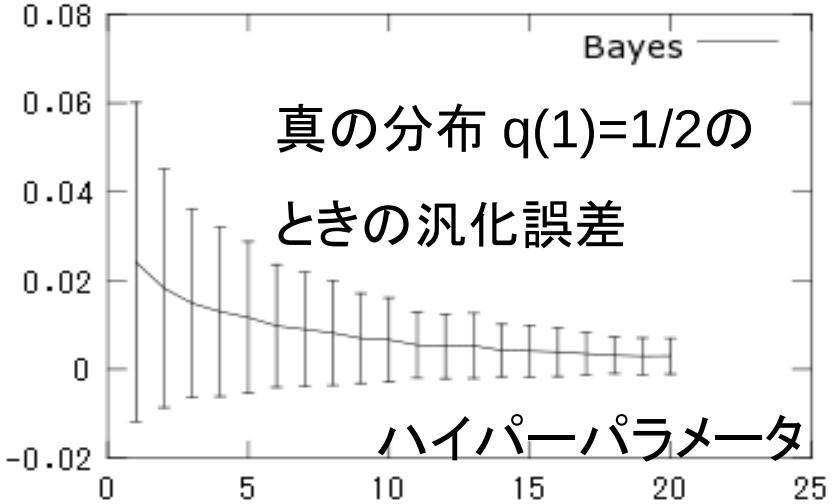
ベイズ汎化誤差とハイパーパラメータ

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。

注意

サイコロは出る目の種類が6個であるが、目の種類が千個もある

サイコロを考えると、すべての目が出る確率が正であったとしても、

すべての目が1回以上出るためには、ものすごく多くの試行が必要になる。

１０万人の人が千種類の商品を買う問題ではどうしたらよいだろうか。

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を作った。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。

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

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

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

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

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

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

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

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

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

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

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 $(a^*,b^*)^{\mathrm{T}}$ 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

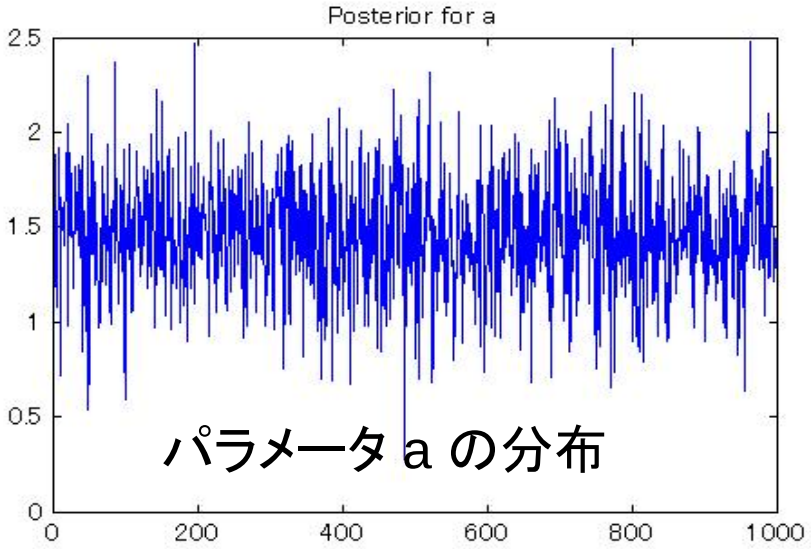
$$\varphi(a) = (t / 2\pi) \exp(- t (a^2+b^2)/2)$$

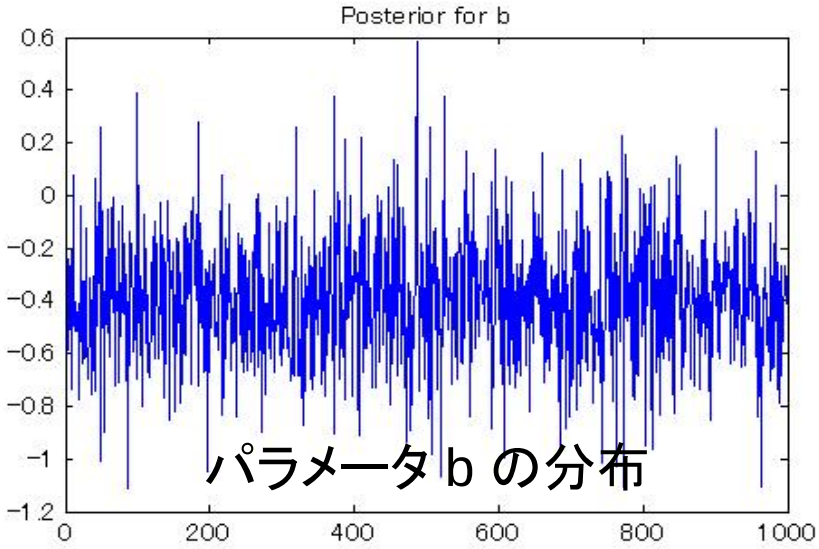
の中のハイパーパラメータ t が推測に与える影響を考察する。

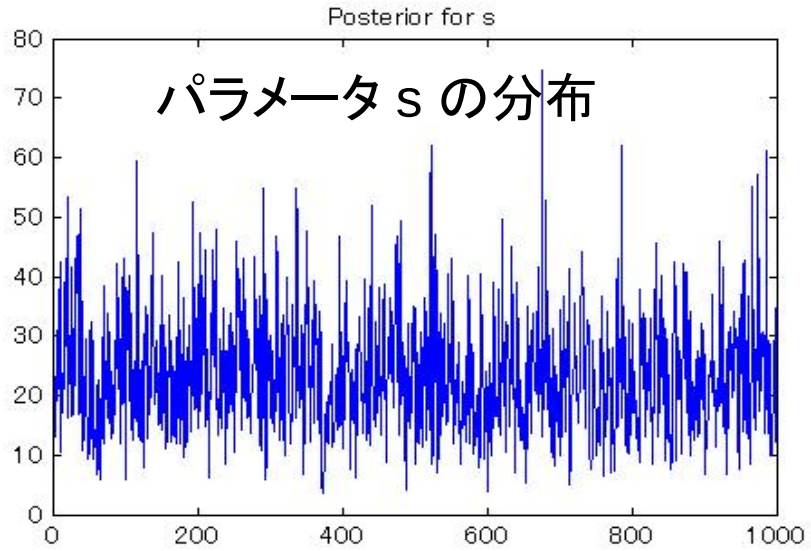
汎化誤差の計算は次のようにした。(A,B) を(a,b)の事後分布での
平均とし、(a*,b*)を最尤推定量とする。

ベイズ法の汎化誤差 = $E_{(X,Y)}$ (Y-AX-B)^2
最尤法の汎化誤差 = $E_{(X,Y)}$ (Y-a*X-b*)^2

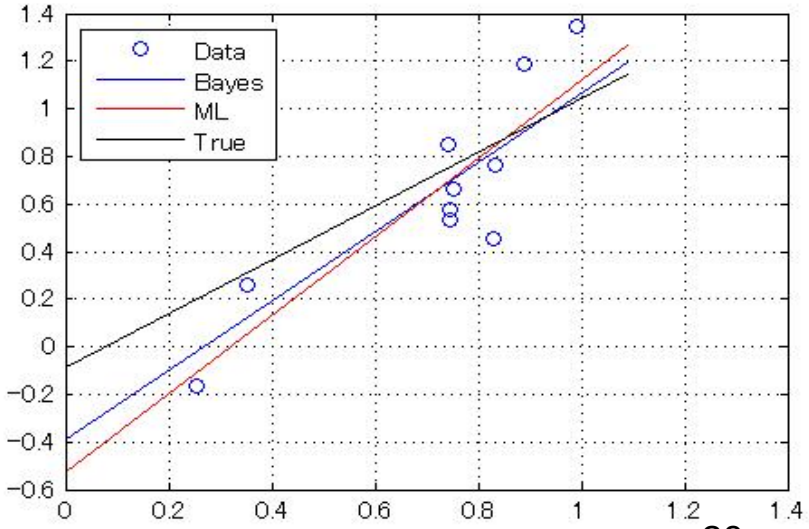
実験の例



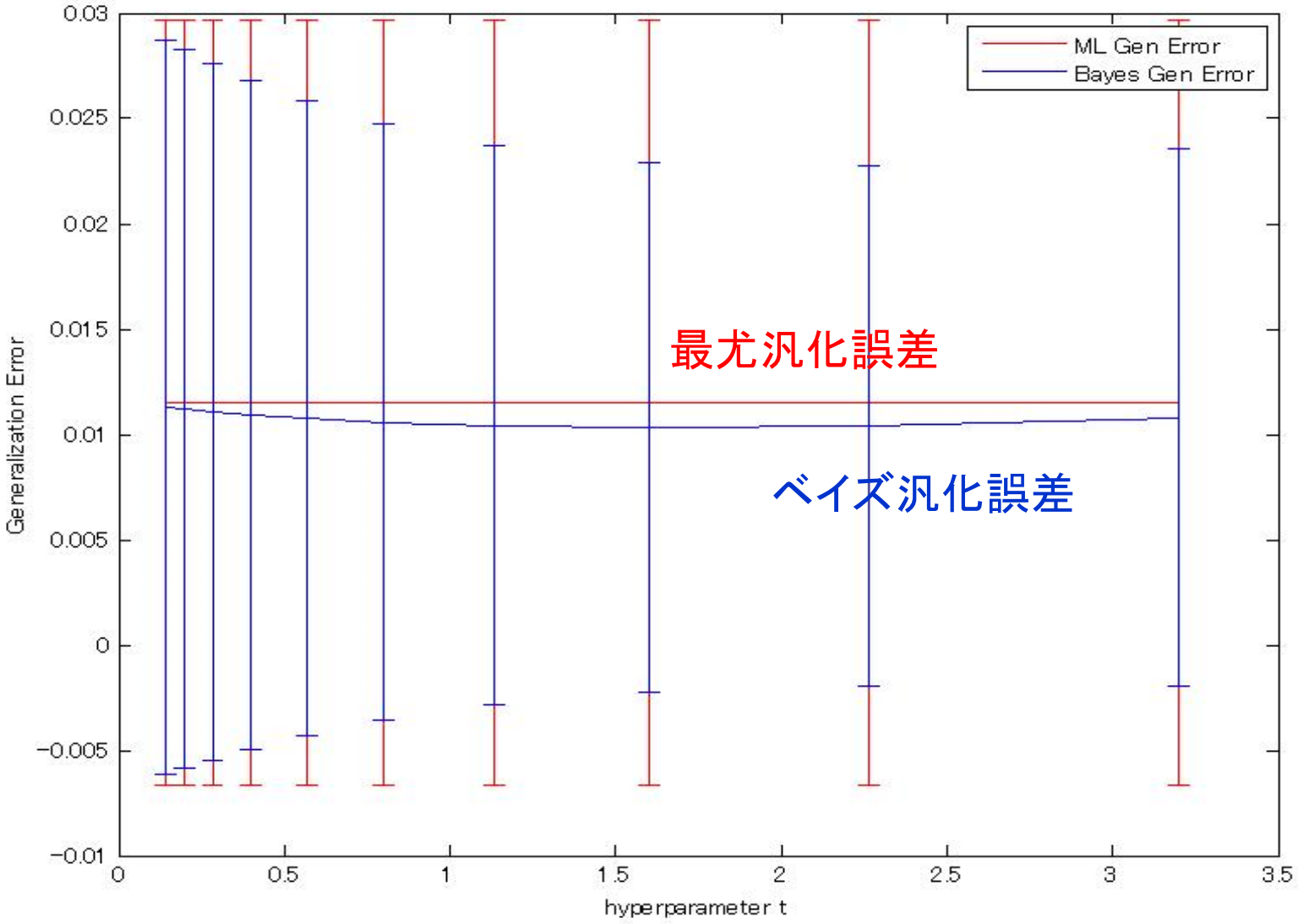




真と推測(n=10, t=1)



実験の例



ハイパーパラメータ t

今後の展開

○　現実には真の分布は不明であるから、人間が定めたモデルや事前分布が適切であるかどうかを真の分布を用いて汎化誤差を求めることはできない。→評価と検定で述べる。方法があります。

○ 構造を持つ問題の統計モデリングにおいては、最尤法よりもベイズ法のほうが優れた推測を与える。→　階層ベイズ法については次回に説明します。（モデリングこそがデータ分析の中心です）。

○MCMC法をどのように作ったらよいか。→今日の統計学において最も重要な問題のひとつ。大学院で学びましょう。最新のプログラム言語STANは、精度のよい事後分布を生成するということで評判になっているようです。使ってみましょう。

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)