学習理論の練習　５

復習

復習： 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
$D(q||p) = \int q(x) \log (q(x)/p(x))\, dx$ 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 $D(q||p) \geqq 0$.
(2) $D(q||p)=0 \Leftrightarrow q(x)=p(x)\ (\forall x)$

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。

教師あり学習と教師なし学習は等価

教師あり学習は
「 $q(x)q(y|x)$ からのデータを用いて $q(y|x)$ を $p(y|x,w)$で推定」

教師なし学習は
「 $q(x)$ からのデータを用いて $q(x)$ を $p(x|w)$ で推定」

数理的には、教師あり学習は教師なし学習の特殊な場合でありまた、教師なし学習は教師あり学習の特殊な場合です。

理論を作るとき記述は教師なし学習のほうがシンプルになるので、教師なし学習の数式を考えますが、教師あり学習にそのまま適用できます。

ベイズ推測とベイズ学習

統計的推測と統計的学習

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。

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。

事後分布の定義

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) から独立に発生した」のもとで

(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

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) から独立に発生した」が正しくても正しくなくても

解析者が準備した φ(w) と p(x|w) を用いて真の密度関数 q(x) を推測する密度関数 p*(x) を次式で定義する。

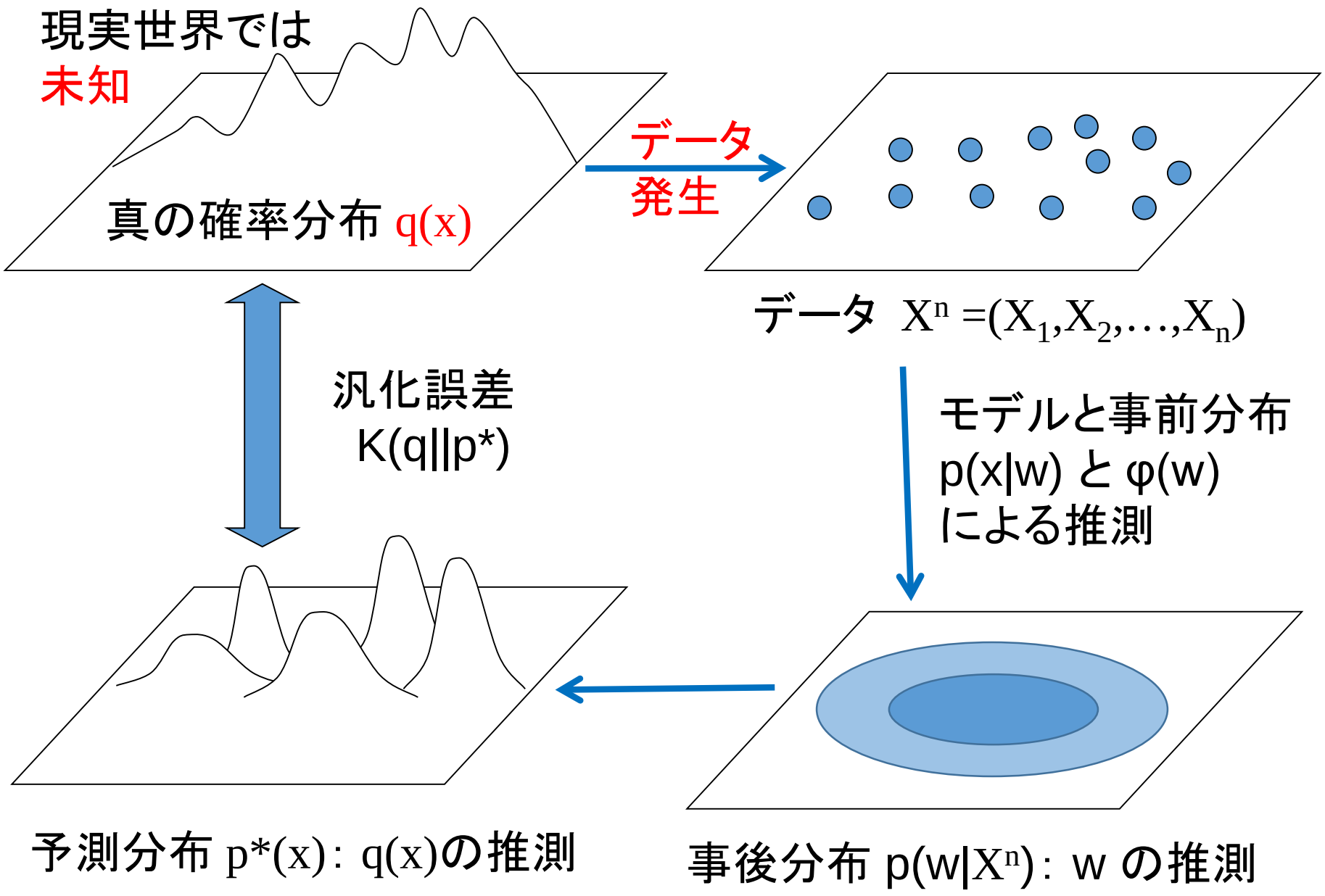
ベイズ推測. 統計モデルを事後分布で平均して推測結果とする.

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

推測された密度関数 p*(x) を予測分布と呼ぶ。

真の分布 q(x) は不明であり、解析者の推論 p*(x) が、不明な q(x) に対して、どのくらい正しいかを数学で調べていく。

ベイズ推測の手順


現実世界では
未知
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*)
モデルと事前分布
p(x|w) と φ(w)
による推測
予測分布 p*(x): q(x)の推測
事後分布 p(w|X^n): w の推測

真の分布と推測された分布

事前密度関数 φ(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*(x) 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。

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

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）

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

をとる。次式が成立する。D(q||p*) を汎化誤差という。

$$D(q||p^*) = G(p^*) - S.$$

汎化誤差はある数学的法則に従う（→今後の課題）

回帰分析の例

回帰分析では何をしたいのか

1. 真の分布 $q(x)q(y|x)$ から得られた n 個のデータから $q(y|x)$ を推測する方法を作りたい。

2. データ、学習モデル $p(y|x,w)$ 、事前分布 $\varphi(w)$ があると事後分布を作ることができる。

3. 事後分布があると予測分布 $p^*(y|x)$ を作ることができる。

4. 予測分布 $p^*(y|x)$ が真 $q(y|x)$ からどのくらいずれているかを知りたい。

5. 現実の問題では真 $q(x)\,q(y|x)$ は不明で、データだけが得られる。
モデルと事前分布を人間が用意して予測分布を作る。
不明な $q(y|x)$ に対して予測分布がどのくらい正しいかを知りたい。

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。t はハイパーパラメータ。

事後分布の計算

パラメータ (a,b,s) の事後分布から $\{(a_k,b_k,s_k)\}$をサンプリング

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

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

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

ベイズ法の予測分布 = $p^*(Y|X) = (1/K)\,\Sigma_k\, p(Y|X,a_k,b_k,s_k)$

最尤法の予測分布 = $p(Y|X,a^*,b^*)$

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 = $E_{(X,Y)}[\log(q(Y|X)/p(Y|X,a^*,b^*))]$

計算の手順（なぜこれでいいかは後述）

(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 をガンマ分布 $G(\ s\ |1+n/2,\ \varepsilon+\Sigma_i\ (Y_i-aX_i-b)^2/2\ )$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

q(x) は区間[0,1]の一様分布で　q(y|x) ＝ p(y|x,1,0,25) とする。

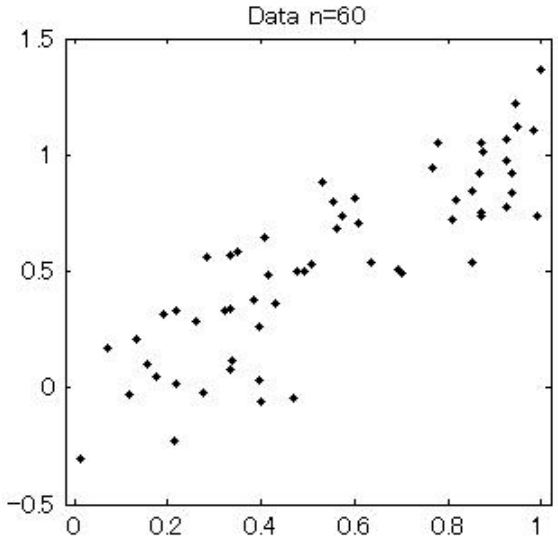
n=12, 12, …, 60　の場合を実験

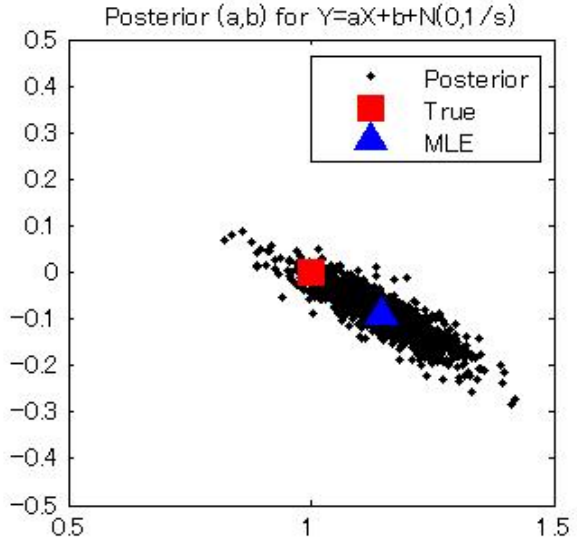
$\{(a_k,b_k)\}$ と $\{(a_k,s_k)\}$ を描画

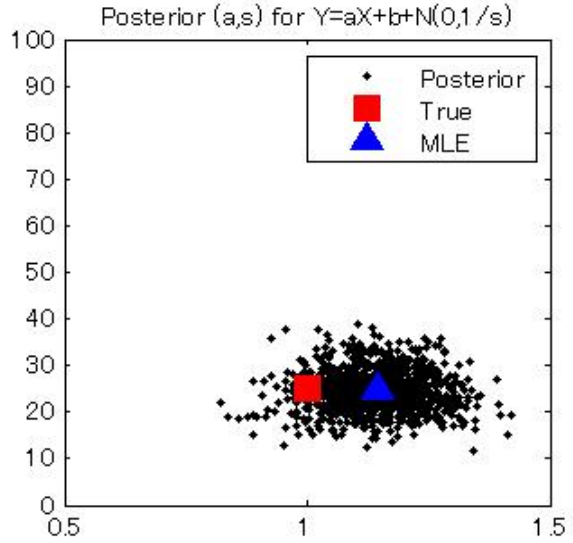
真 q(y|x) 、および最尤とベイズで p(y|x) 描画して比較

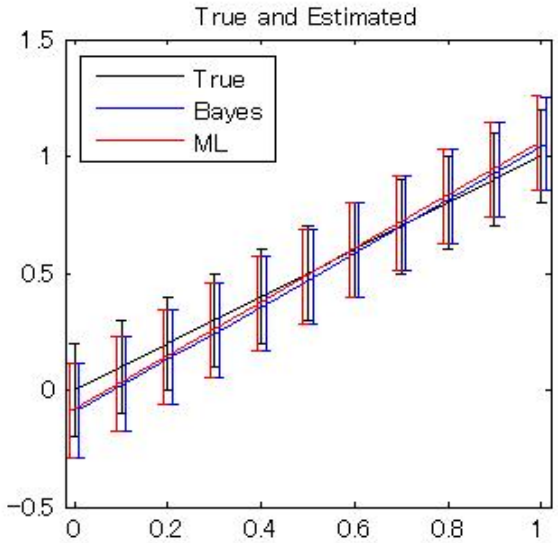
汎化誤差、WAIC、LOOCV を描画して比較

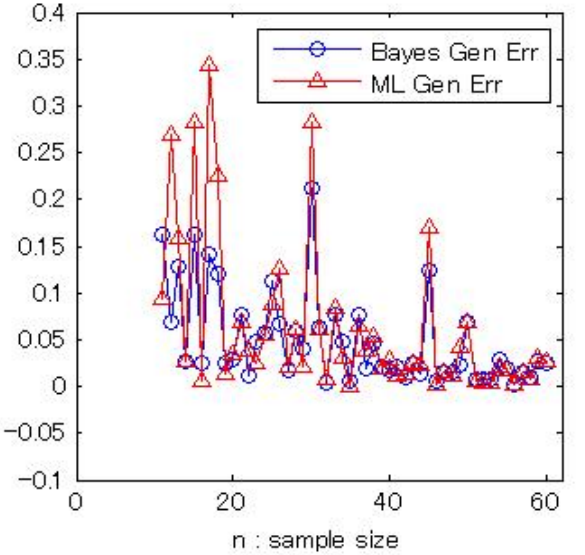
実験結果

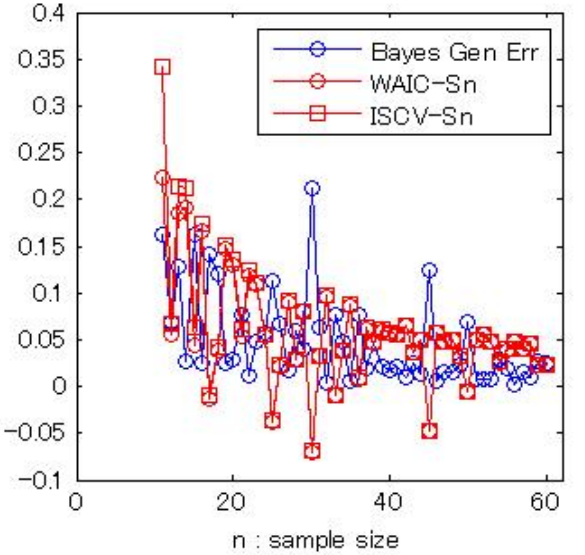

Data n=60


Posterior (a,b) for Y=aX+b+N(0,1/s)
Posterior
True
MLE


Posterior (a,s) for Y=aX+b+N(0,1/s)
Posterior
True
MLE


True and Estimated
True
Bayes
ML


Bayes Gen Err
ML Gen Err
n : sample size


Bayes Gen Err
WAIC-Sn
ISCV-Sn
n : sample size

事後分布とモンテカルロ法

ベイズ法を実現するときの計算法の問題

ベイズ推測の目標は予測分布 $p^*(x)$ を作ること

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

この積分 $\int dw$ は数値的にしか実行できないことが多い。そこで

事後分布 $p(w|X^n)$ に従う $\{w_k;\ k=1,2,\ldots,K\}$ をたくさん用意して

$$p^*(x) \fallingdotseq (1/K) \sum_{k=1}^{K} p(x|w_k)$$

と近似する（モンテカルロ法）。

（※）モデルと事前分布を定義すると、事後分布を自動生成する計算機言語に STAN があり、精度が良く評価が高い。研究や実務に使われている。

使えるようになっておくと、統計解析力が +255 アップ。

ギブスサンプラー法

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

(1) 初期値 w を決める。

(2) u を p(u|w) からサンプリング

(3) w を p(w|u) 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(w,u) で

$$(1/K)\ \Sigma\ f(w_k,u_k) \rightarrow \int f(w,u)\ p(w,u)\ dw\ du$$

が成り立つ。

計算の確認

事後分布の計算

パラメータ (a,b,s) の事後分布 p(a,b,s|データ) は

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

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

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

となる。ギブスサンプラーの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。

$$G(x|\alpha,\beta) = 1/(\beta^{\alpha}\Gamma(\alpha))\, x^{\alpha-1}\exp(-x/\beta)$$

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。