データ解析

第10回：階層ベイズ法

渡辺澄夫

構造を
推測に
活用する

実世界

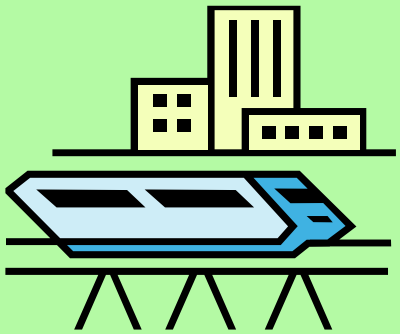
観測データ

回帰・判別分析

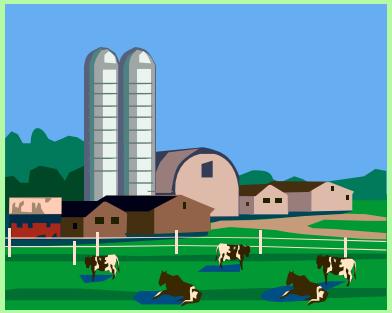
推定と検定

ベイズ法
・階層ベイズ法

解析方法

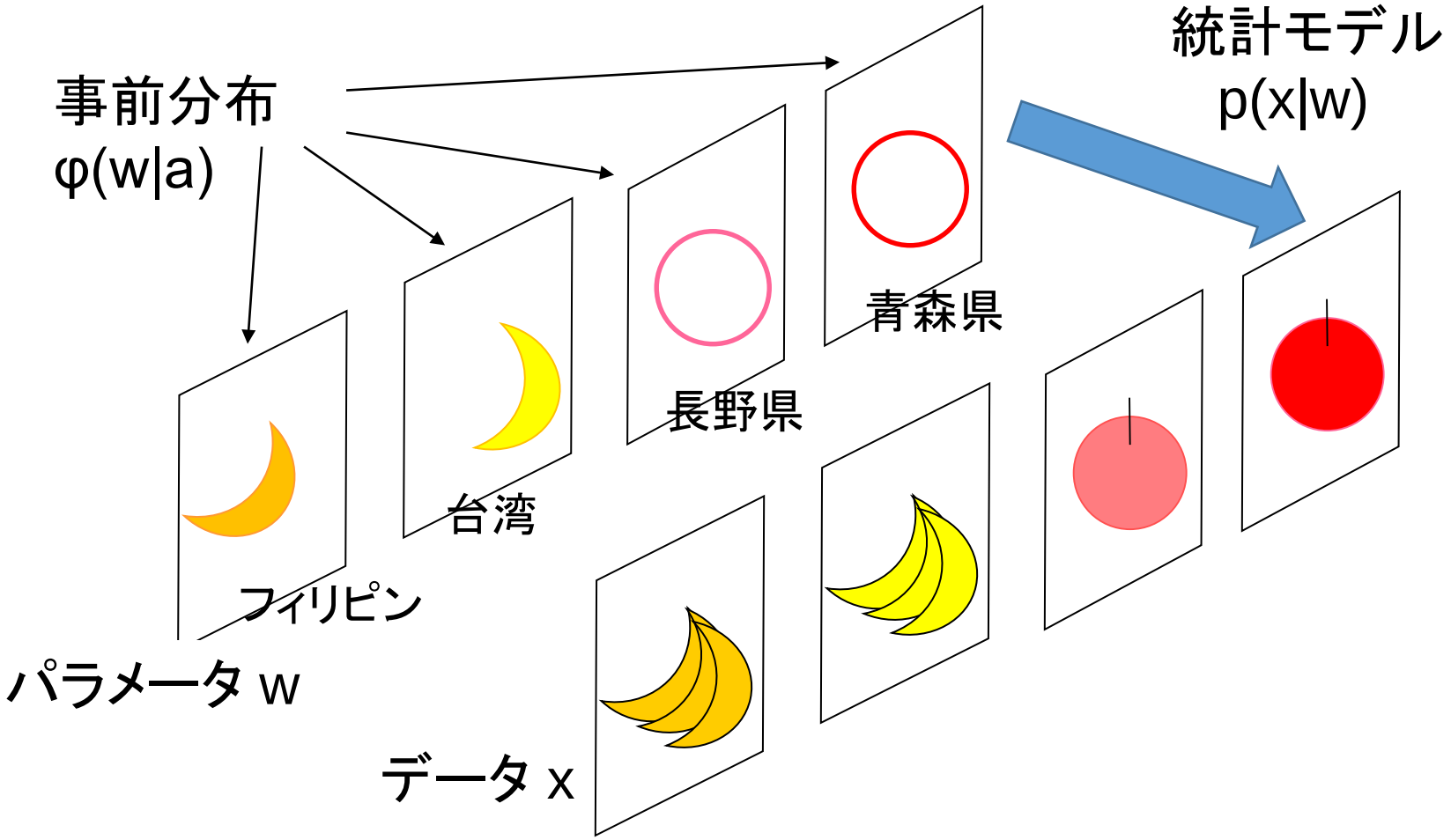


時系列予測



主成分・因子
・クラスタ分析

階層ベイズ法



統計的推測を行なう対象が構造を持つとき、その構造を用いてモデリングを行なうことで精度の良い推測ができることがある。今日の統計モデリングの基本技法のひとつである。

復習：統計的推測について

復習：事前分布と事後分布

データ $X^n=(X_1,X_2,\ldots,X_n)$ が得られる。

パラメータ w が与えられたときの x の確率密度関数 p(x|w) を統計モデルという。

パラメータ集合の上に定義された確率密度関数 φ(w) を事前確率密度関数という。

データ X^n が与えられたときのパラメータの事後確率密度関数を

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

と定義する。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。

復習：推測法

真の密度関数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) 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事後確率密度関数を最大にする 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。

復習：MCMC法による事後分布の実現

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。

復習：ギブスサンプラー法

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

(1) 初期値 v を決める。

(2) u を $p(u|v)$ からサンプリング

(3) v を $p(v|u)$ 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 $f(u,v)$ で

$$(1/K)\ \Sigma\ f(u_k,v_k) \rightarrow \int f(u,v)\ p(u,v)\ du\ dv$$

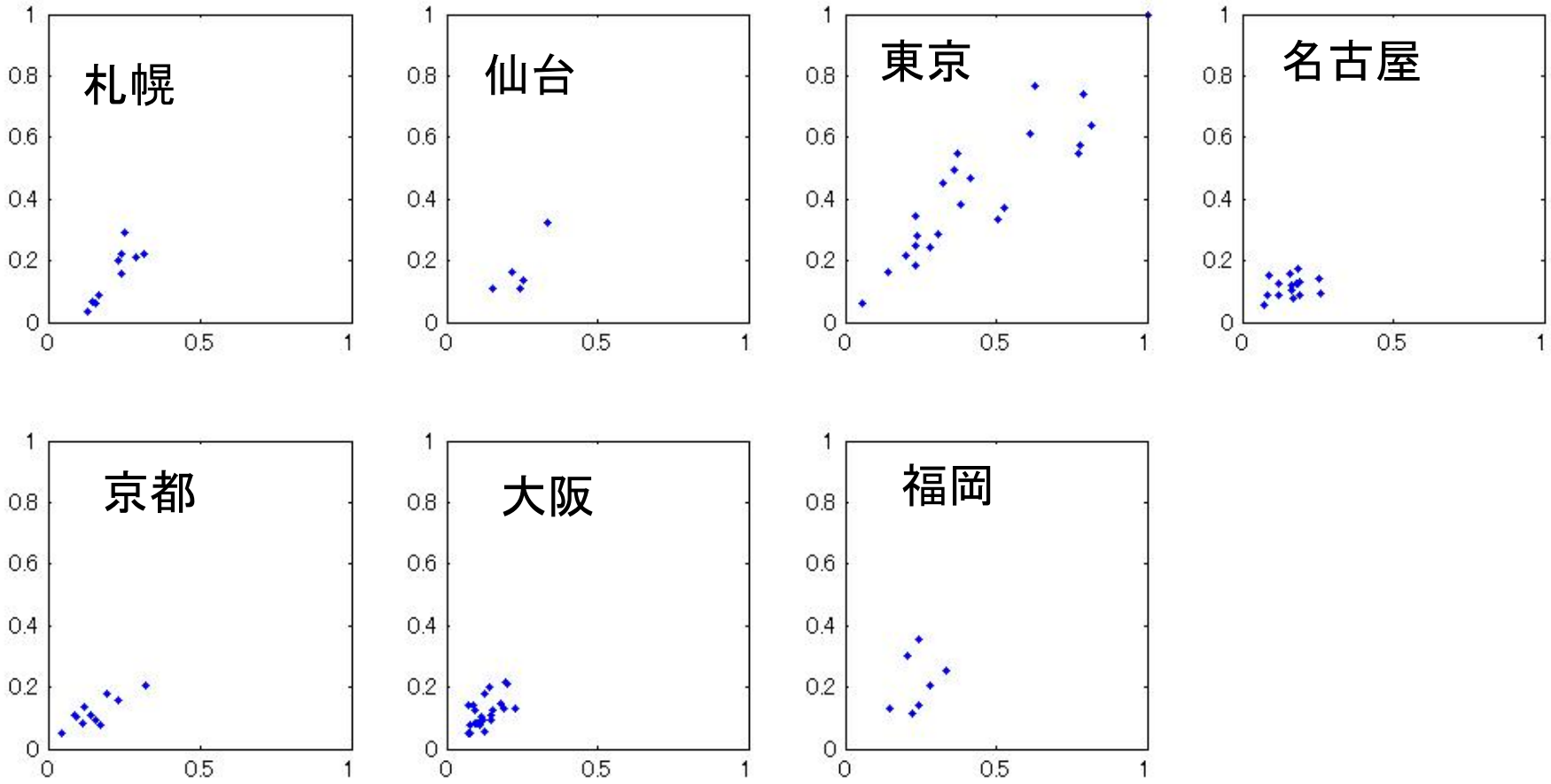
が成り立つ。（ただし右辺が有限なとき）。

階層ベイズ法の例

具体例を使ったほうがわかりやすいと思いますので
具体例で説明します。

構造を持つ問題

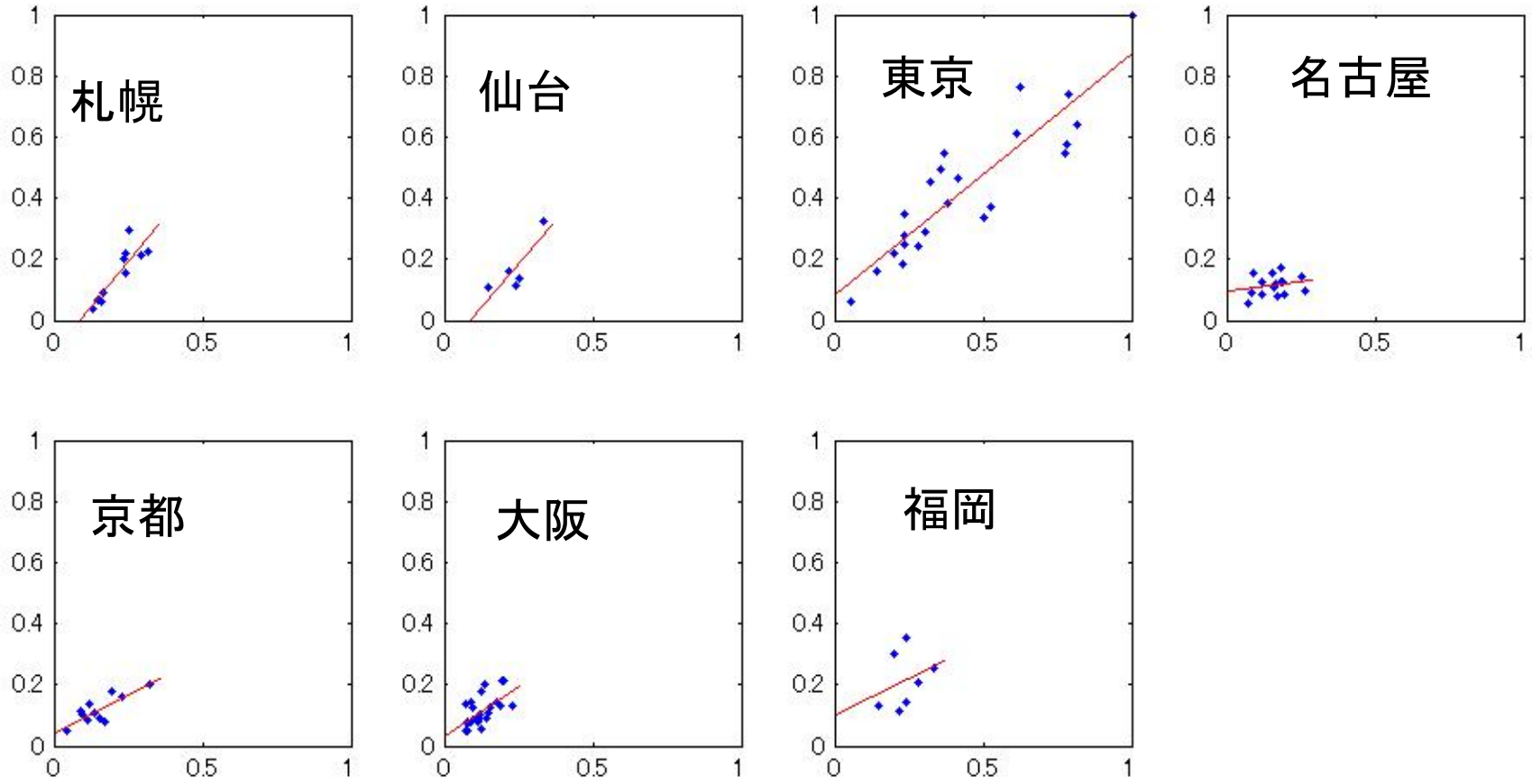
下記は、札幌(1)、仙台(2)、東京(3)、名古屋(4)、京都(5)、大阪(6)、福岡(7)の各区の人口(X軸)と単独世帯の数(Y軸)を正規化して表示したものである。



政府統計の総合窓口より http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do

最尤推定してみたが・・・。

それぞれの都市で、回帰直線 Y=a(k)X+b(k) (ただし k=1,2,…,7) を求めたい。二乗誤差最小推定（最尤推定と同じ）で求めてみた。



データが少なすぎて、うまく推定できない。例:b(k) が0から遠すぎる。

同種の推測が複数の場所で行なわれる問題

同じ構造を持つ問題はとても多い。

例１．１学年８クラスで英語のテストをする。

例２．アメリカ大統領選挙

例３．コンビニエンスストア・チェーン店での商品の売り上げ

共通する情報を統合抽出し、それぞれの推測に役立てたい。

→ 独立した推測ではなく、パラメータを共通のハイパー事前分布から得られたものとしてモデリングする。（モデリングの正しさの評価については、評価の項で述べる）。

階層ベイズのモデリング例

(0) k を都市の番号とする。(k=1,2,…,7)

(1) 第 k 都市のモデル「Y=a(k)X+b(k)+正規分布」を表す確率密度関数は

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

(2) (a(k),b(k))の事前分布として次のものを用いる。

$$\varphi(a(k),b(k)|m_a,m_b,t) = (t / 2\pi) \exp[- t \{(a(k)-m_a)^2+(b(k)-m_b)^2\}/2]$$

(3) (s(k), t) については正値で十分大きな広がりを持つ事前分布を固定。

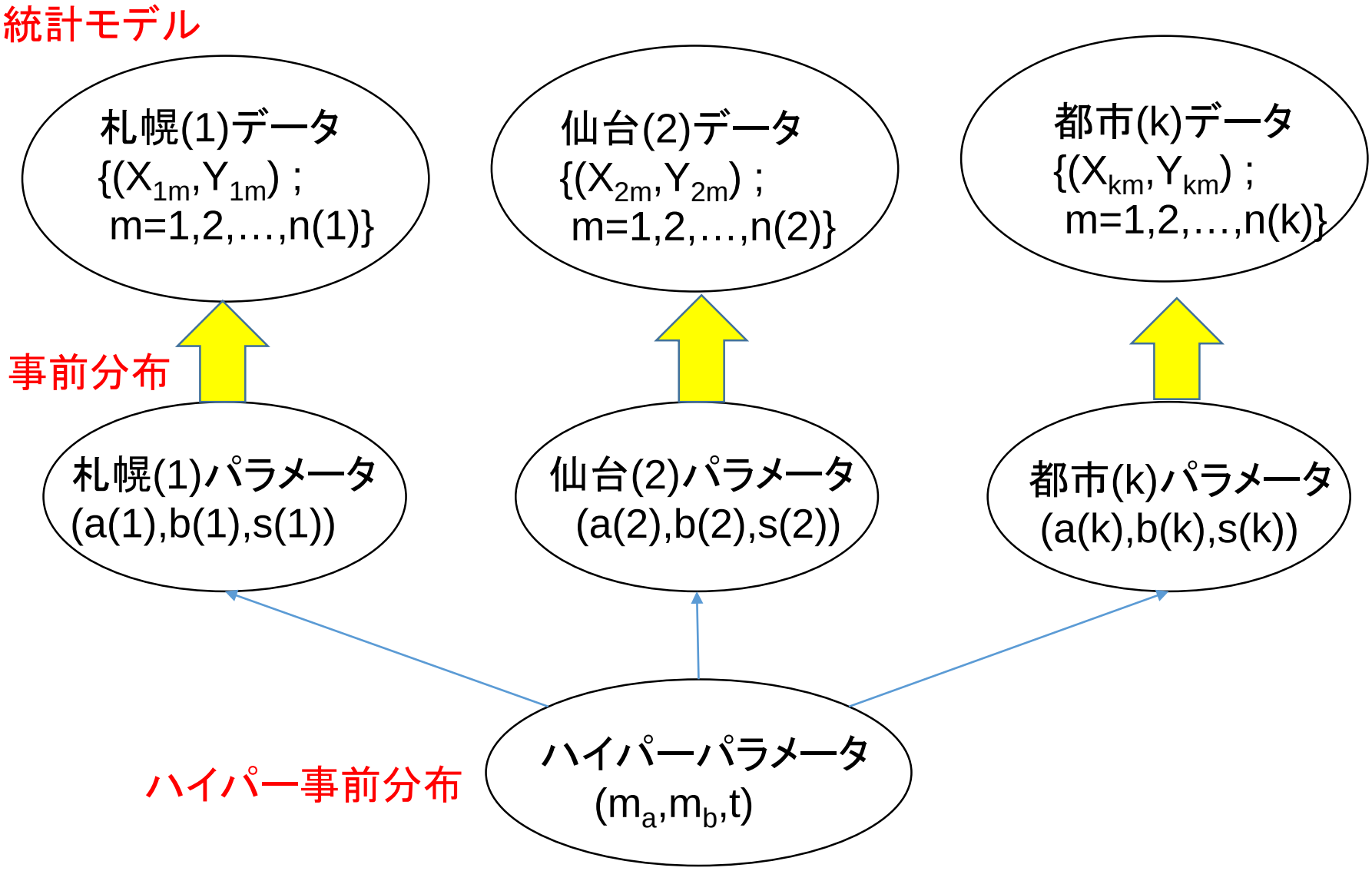
ε=0.001とする。

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

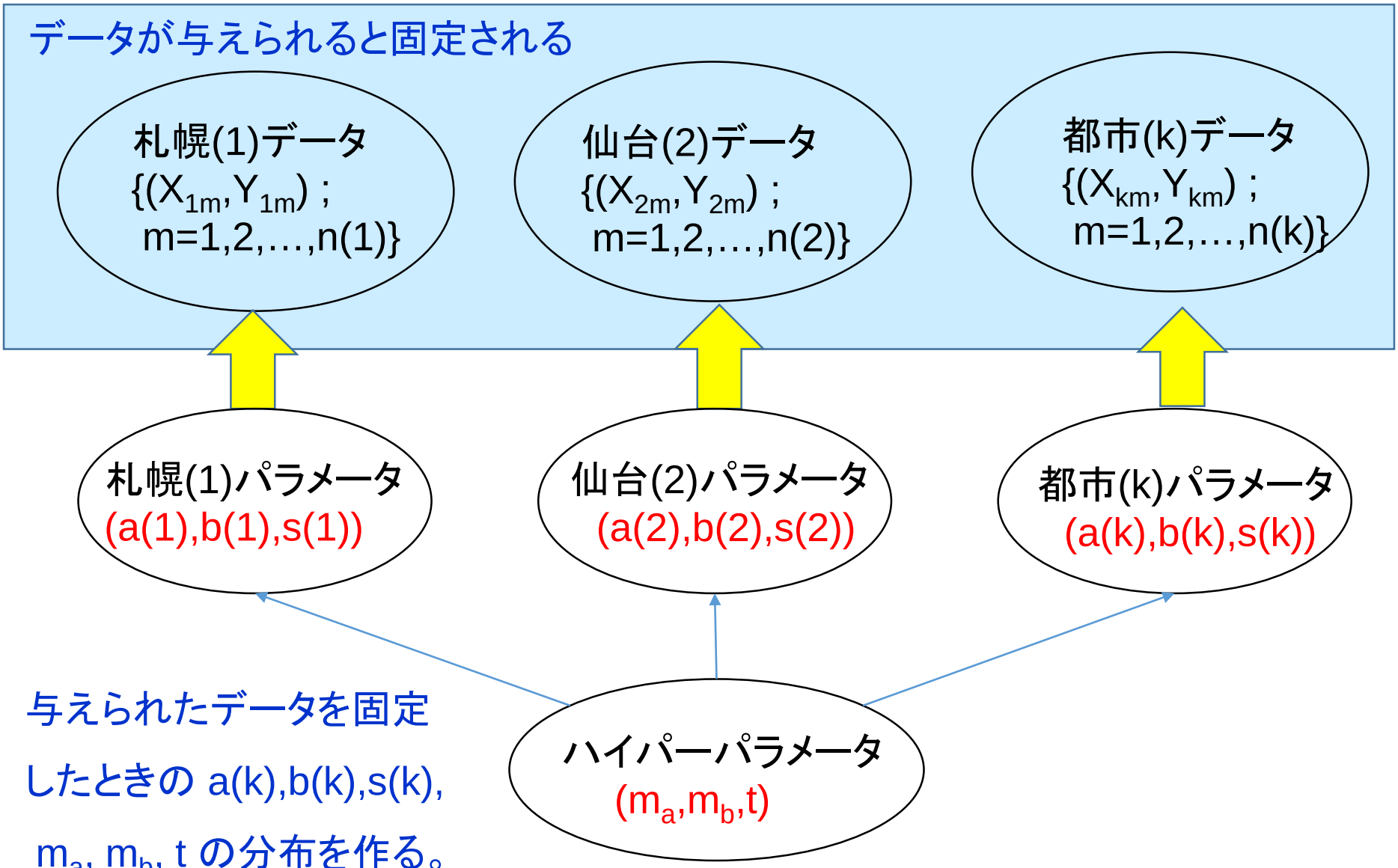
$$\psi(t) = \varepsilon \exp(- \varepsilon t)$$

(4) (m_a,m_b) については原点を含む十分に大きな集合上の一様分布を固定

統計モデリングを図示


統計モデル
札幌(1)データ
{(X1m,Y1m) ;
m=1,2,…,n(1)}
仙台(2)データ
{(X2m,Y2m) ;
m=1,2,…,n(2)}
都市(k)データ
{(Xkm,Ykm) ;
m=1,2,…,n(k)}
事前分布
札幌(1)パラメータ
(a(1),b(1),s(1))
仙台(2)パラメータ
(a(2),b(2),s(2))
都市(k)パラメータ
(a(k),b(k),s(k))
ハイパーパラメータ
(ma,mb,t)
ハイパー事前分布

推測されるものの確率分布



事後分布

データが $\{(X_{ki},Y_{ki});\ k=1,2,\ldots,K,\ i=1,2,\ldots,n(k)\}$ であるとき

事後分布は次の式になる。

$$p(a(k),b(k),s(k),m_a,m_b,t \mid \text{データ})$$

$$\propto \psi(t)\ [\ \Pi_k\ \psi(s(k))\ \varphi(a(k),b(k)|m_a,m_b,t)\ ]$$

$$\times\ [\ \Pi_k\ \Pi_i\ p(Y_i|X_i,a(k),b(k),s(k))\ ]$$

この分布に従う $\{\ a(k),b(k),s(k),m_a,m_b,t\ \}$ をサンプリングすればよい。

個々の分布は次の形をしている。

$$p(Y_i|X_i,a(k),b(k),s(k)) = (\ s(k)\ /\ 2\pi\ )^{1/2}\ \exp[\ -s(k)(Y_i-a(k)X_i-b(k))^2/2\ ].$$

$$\varphi(a(k),b(k)|m_a,m_b,t) = (\ t\ /\ 2\pi\ )\ \exp[\ -\ t\ \{(a(k)-m_a)^2+(b(k)-m_b)^2\}/2\ ]$$

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

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

ギブスサンプラー法

準備:ガンマ分布と正規分布

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

ガンマ分布に従う x をサンプリングする関数を準備する。

例　　x = gamrnd(α,β);

(2) 標準正規分布（平均0、標準偏差1)に従う x をサンプリングする関数を準備する。例 x= randn;

平均ベクトル が2次元ベクトル $v =(v_1,v_2)^T$ で

分散共分散行列が S の2次元正規分布に従う2次元確率変数

$x=(x_1,x_2)^T$　は次式で生成できる。（注）式では $\exp(- (1/2)||S^{-1/2}(x-v)||^2)$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + S^{1/2} \begin{pmatrix} randn \\ randn \end{pmatrix}$$

事後分布

ギブスサンプラーでは、ある変数に着目しているときは他の変数は固定して考えることができる。$p(a(k),b(k),s(k),m_a,m_b,t \mid データ)$ からのサンプリングを次の順番で生成する。

(1) m_a,m_b

(2) t

(3) $s(k)$

(4) $a(k), b(k)$

それぞれの変数の生成法を以下で述べる。

変数 (m_a,m_b) の生成

まず (m_a,m_b) の生成を考える。

事後分布は $p(a(k),b(k),s(k),m_a,m_b,t \mid データ)$

$$\propto \psi(t)\ [\Pi_k \psi(s(k))\ \varphi(a(k),b(k)|m_a,m_b,t)\ ]\ [\Pi_k \Pi_i\ p(Y_i|X_i,a(k),b(k),s(k))\ ]$$

(m_a,m_b)がないところは1と考えてよい。必要なところだけ取り出すと

$$p(m_a,m_b|\ 他の変数、データ) \propto [\Pi_k \varphi(a(k),b(k)|m_a,m_b,t)\ ]$$

ここで $\varphi(a(k),b(k)|m_a,m_b,t) = (t / 2\pi) \exp[- t \{(a(k)-m_a)^2+(b(k)-m_b)^2\}/2]$

だから、(m_a,m_b) は正規分布を使って生成できる。

m_a は平均が $(1/K)\ \Sigma_k\ a(k)$ で標準偏差が $1/(tK)^{1/2}$ の正規分布

m_b は平均が $(1/K)\ \Sigma_k\ b(k)$ で標準偏差が $1/(tK)^{1/2}$の正規分布

変数 t の生成

次に t の生成を考える。

事後分布は $p(a(k),b(k),s(k),m_a,m_b,t \mid データ)$

$\propto \quad \psi(t) \ [\Pi_k \psi(s(k)) \varphi(a(k),b(k)|m_a,m_b,t)] \ [\Pi_k \Pi_i p(Y_i|X_i,a(k),b(k),s(k))]$

t がないところは1と考えてよいので、必要なところだけ取り出すと

$p(t \mid 他の変数、データ) \quad \propto \quad \psi(t) \ [\Pi_k \varphi(a(k),b(k)|m_a,m_b,t)]$

ここで $\psi(t) = \varepsilon \exp(- \varepsilon t)$

$\varphi(a(k),b(k)|m_a,m_b,t) = (t / 2 \pi) \exp[- t \{(a(k)-m_a)^2+(b(k)-m_b)^2\}/2]$

従って t はガンマ分布 $G(K+1, 1 / [\varepsilon + (1/2) \Sigma_k \{ (a(k)-m_a)^2+(b(k)-m_b)^2\}])$

を使って生成できる。

変数 s(k) の生成

さらに s(k) の生成を考える。

事後分布は $p(a(k),b(k),s(k),m_a,m_b,t \mid データ)$

$$\propto \psi(t) \left[\Pi_k \psi(s(k)) \varphi(a(k),b(k)|m_a,m_b,t) \right] \left[\Pi_k \Pi_i p(Y_i|X_i,a(k),b(k),s(k)) \right]$$

s(k) がないところは1と考えてよい。Kごとの掛け算で欠けているから

各 k ごとに独立に考えてよい。

$$p(s(k) \mid 他の変数、データ) \propto \psi(s(k)) \left[\Pi_i p(Y_i|X_i,a(k),b(k),s(k)) \right]$$

ここで $\psi(s(k)) = \varepsilon \exp(- \varepsilon s(k))$

$$p(Y_i|X_i,a(k),b(k),s(k)) = (s(k) / 2\pi)^{1/2} \exp[-s(k)(Y_i-a(k)X_i-b(k))^2/2]$$

従って s(k) はガンマ分布を使って生成できる。

$$G(n(k)/2+1, 1 / [\varepsilon + (1/2) \Sigma_i (Y_i-a(k)X_i-b(k))^2])$$

変数 (a(k),b(k)) を生成

最後に (a(k),b(k)) の生成を考える。

事後分布は $p(a(k),b(k),s(k),m_a,m_b,t \mid データ)$

$\propto \quad \psi(t) \ [\ \Pi_k \ \psi(s(k))\ \varphi(a(k),b(k)|m_a,m_b,t)\ ] \ \ [\ \Pi_k \ \Pi_i \ p(Y_i|X_i,a(k),b(k),s(k))\ ]$

(a(k),b(k)) がないところは1と考えてよく、各 k ごとに独立に考えてよい。

$p(a(k),b(k)|他の変数、データ) \propto \varphi(a(k),b(k)|m_a,m_b,t) \ \Pi_i \ p(Y_i|X_i,a(k),b(k),s(k))$

ここで $\varphi(a(k),b(k)|m_a,m_b,t) = (t / 2\pi) \exp[- t \{(a(k)-m_a)^2+(b(k)-m_b)^2\}/2]$

$p(Y_i|X_i,a(k),b(k),s(k)) = (s(k) / 2\pi)^{1/2} \exp[-s(k)(Y_i-a(k)X_i-b(k))^2/2]$

p(a(k),b(k)) は exp(-2次式) の形をしているので正規分布で生成できる。

変数 (a(k),b(k)) の生成つづき

パラメータ (a(k),b(k)) の確率密度関数は次のように変形できる。

$$\propto \exp[- t \{(a(k)-m_a)^2+(b(k)-m_b)^2\}/2] \Pi_i \exp[-s(k)(Y_i-a(k)X_i-b(k))^2/2]$$

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

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

ここで行列 H とベクトル v は次のように定義した。

$$H = \begin{pmatrix} t + s(k)\ \Sigma_i\ X_i^2 & s(k)\ \Sigma_i\ X_i \\ s(k)\ \Sigma_i\ X_i & t + s(k)n(k) \end{pmatrix} \qquad v = \begin{pmatrix} t\ m_a + s(k)\ \Sigma_i\ X_i Y_i \\ t\ m_b + s(k)\ \Sigma\ Y_i \end{pmatrix}$$

従って (a(k),b(k)) は次の式で生成できる。

$$\begin{pmatrix} a(k) \\ b(k) \end{pmatrix} = H^{-1}\ v\ +\ H^{-1/2} \begin{pmatrix} randn \\ randn \end{pmatrix}$$

データの正規化・初期値の設定など

データ $\{(X_{ki},Y_{ki})\}$ は何でもよいが、数値の発散などが心配なので[0, 1] 区間の程度の値に変換しておくと安心である。例 log (exp(-10000)) = NaN.

理論上はパラメータの初期値は何でもよい。実験上は次のものを使った。

最尤推定で (a(k),b(k)) を推定し、その {a*(k),b*(k):k=1,2,…,K} を用いてさらに t を最尤推定して得た t* を (a(k),b(k)) と t の初期値として用いる。

最尤推定量 (a*(k),b*(k))　は t=0, s(k)=1 のときの H^{-1} v と等しいので計算できる。
最尤推定量　t* は　t*=(V(a*)+V(b*))/2 である。ここで V(a*) と V(b*) は {a*(k);k=1,2,…,K} と {b*(k);k=1,2,…,K} のそれぞれの分散である。

計算の手順

(1) データ $\{(X_{ki},Y_{ki})\ ;\ k=1,2,\ldots,K,\ i=1,2,\ldots,n(k)\}$ を得る。適切な大きさに変換。

(2) 最尤推定量 $(a^*(k),b^*(k))$ と t^* を計算し、$(a(k),b(k),t)$ の初期値とする。

(3) m_a を平均 $(1/K)\ \Sigma_k\ a(k)$ で標準偏差 $1/(tK)^{1/2}$ の正規分布から生成。

(4) m_b を平均 $(1/K)\ \Sigma_k\ b(k)$ で標準偏差 $1/(tK)^{1/2}$の正規分布から生成。

(5) $s(k)$ をガンマ分布 $G(\ n(k)/2+1,\ 1\ /\ [\ \varepsilon + (1/2)\ \Sigma_i\ (Y_i-a(k)X_i-b(k))^2\ ]\ )$で生成

(6) t をガンマ分布 $G(\ K+1,\ 1\ /\ [\ \varepsilon + (1/2)\ \Sigma_k\ \{\ (a(k)-m_a)^2+(b(k)-m_b)^2\}\ ]\ )$で生成

(7) $(a(k),b(k))$ を下記の方法で生成

$$H = \begin{pmatrix} t + s(k)\ \Sigma_i\ X_i^2 & s(k)\ \Sigma_i\ X_i \\ s(k)\ \Sigma_i\ X_i & t + s(k)n(k) \end{pmatrix} \qquad v = \begin{pmatrix} t\ m_a + s(k)\ \Sigma_i\ X_iY_i \\ t\ m_b + s(k)\ \Sigma\ Y_i \end{pmatrix}$$

$$\begin{pmatrix} a(k) \\ b(k) \end{pmatrix} = H^{-1}\ v\ +\ H^{-1/2} \begin{pmatrix} randn \\ randn \end{pmatrix}$$

(8) (3)に戻る。

実験の例

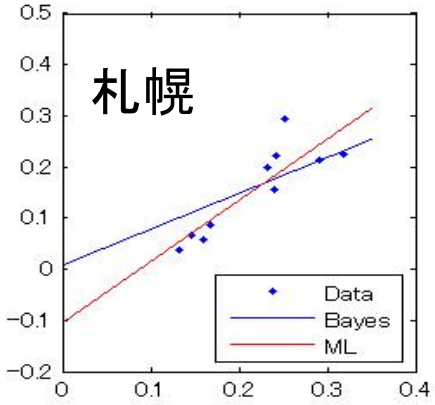
初期値の影響を消すために最初の100回分を捨てる（バーンイン）。
その後1000回の繰り返しを行なって {a(k),b(k),s(k);k=1,2,…,K} のそれぞれについて1000個のサンプリング結果を得る。これを
{A(k,j),B(k,j),S(k,j); k=1,2,…,K, j=1,2,…,1000} と書く。

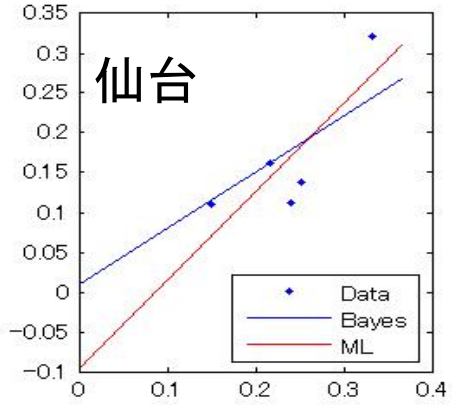
{A(k,j)} と {B(k,j)} の j についての平均を {EA(k)} {EB(k)} と書いてこれを階層ベイズ法によるパラメータの推定結果として用いることにした。

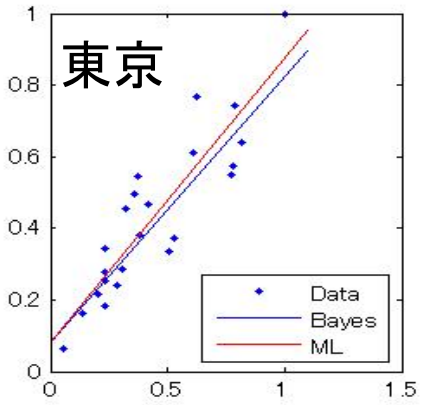
K本の直線 Y= EA(k)X+EB(k) をそれぞれの都市のグラフに描いた。

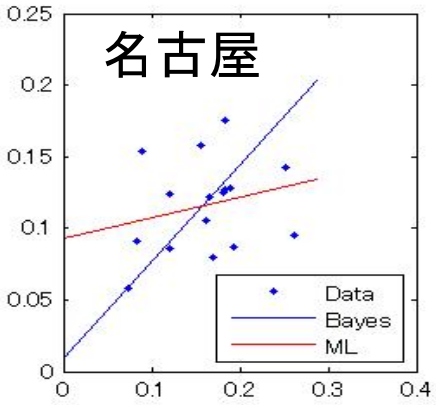
階層ベイズ法による推測結果

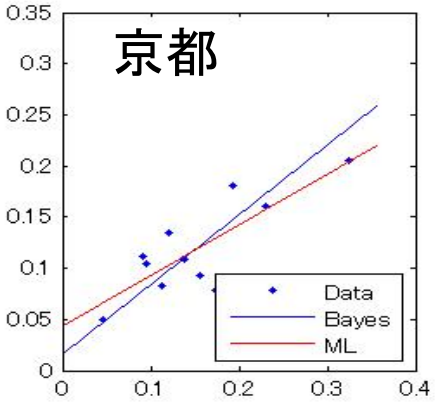
区の数が少ない都市は最尤推定と階層ベイズの結果が異なることが多い。

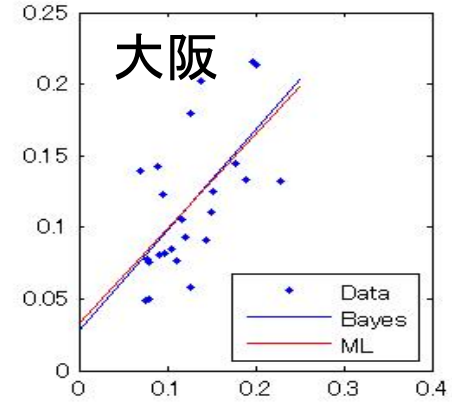


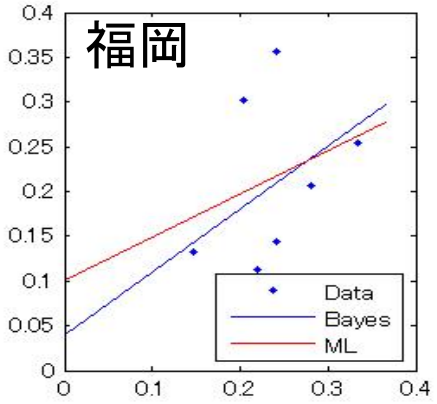






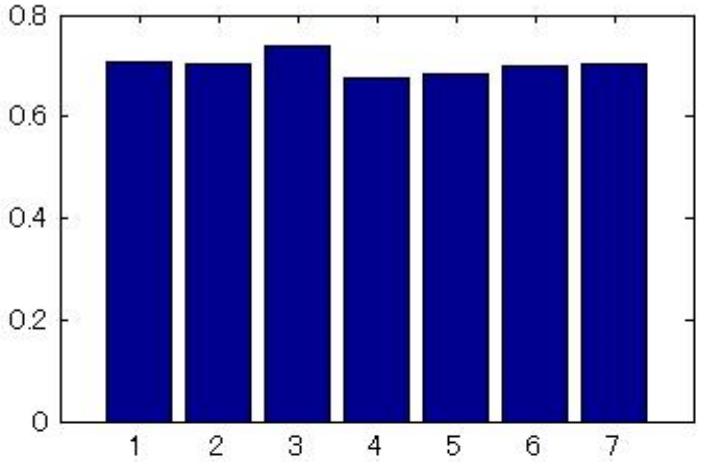




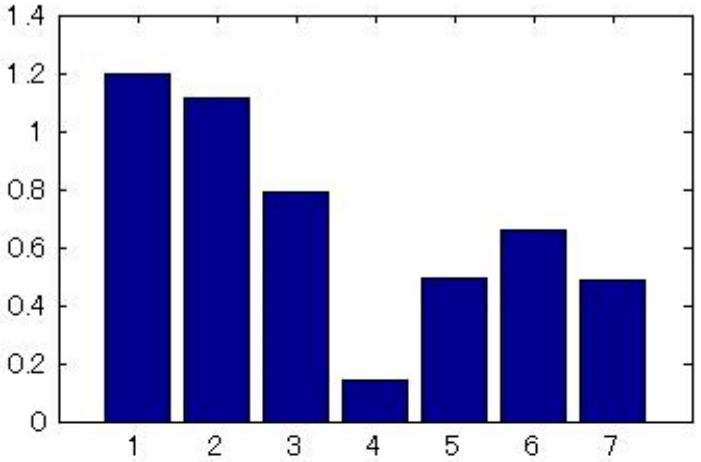


推測されたパラメータの比較

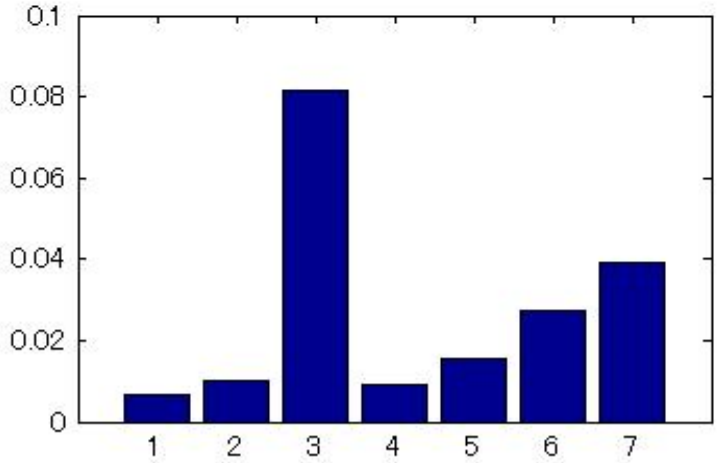
階層ベイズによる a(k)



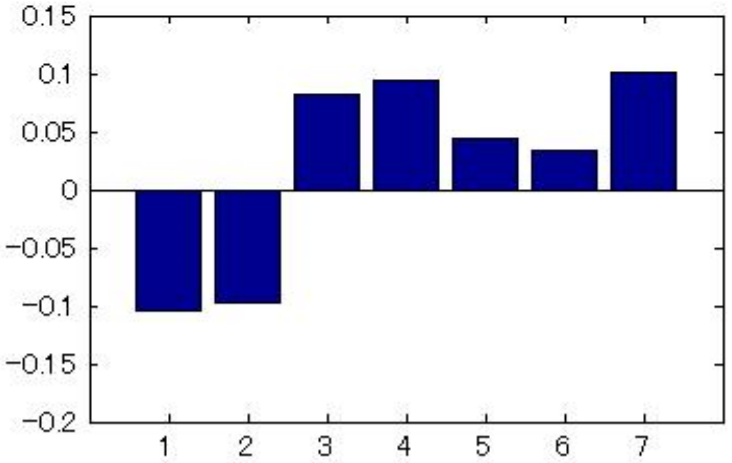
最尤による a(k)



階層ベイズによる b(k)



最尤による b(k)



データについての考察

東京と大阪は、最尤と階層ベイズの違いが小さいが、他の都市は
最尤と階層ベイズの結果は異なっている。名古屋は大きく異なっている。

「人口→単独世帯数」の傾きを各都市で比較する場合、最尤法では
札幌が最も大きいが、階層ベイズ法では東京が最も大きい。

東京と大阪以外では、最尤推定の結果はランダムネスによる影響が
大きいと考えられるので、単独世帯数の人口に対する比が最も大きい
のは東京であると考察される。