データ解析

第６回：クラスタ分析

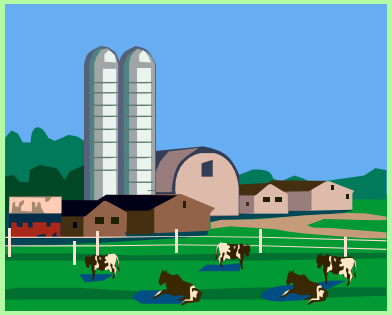
渡辺澄夫

実世界は
複数の
共同体から
なるのか

観測データ

回帰・判別分析

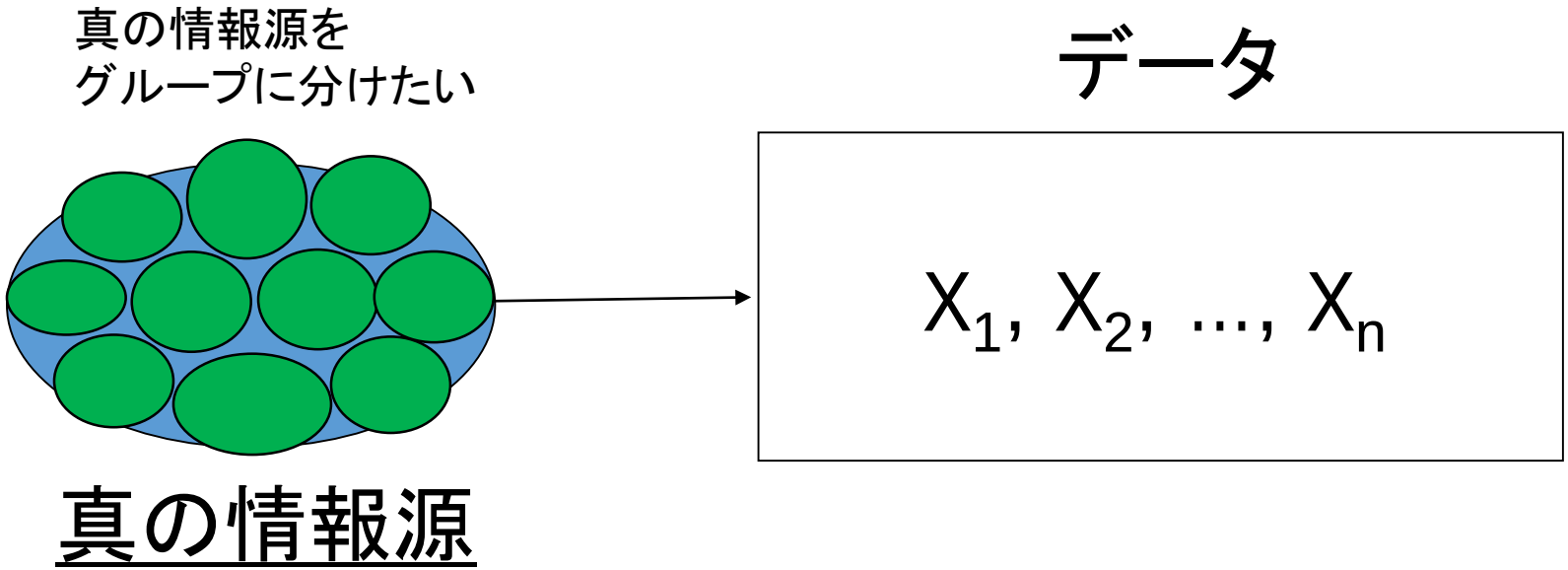
解析方法



主成分・因子
・クラスタ分析

ラベルを持たないデータ

主成分分析と因子分析では、与えられたデータを少数の因子で説明する方法を考察した。クラスタ分析では、データが複数のグループからできているときにそれぞれのグループを推測する。＝クラスタリング、教師なし学習

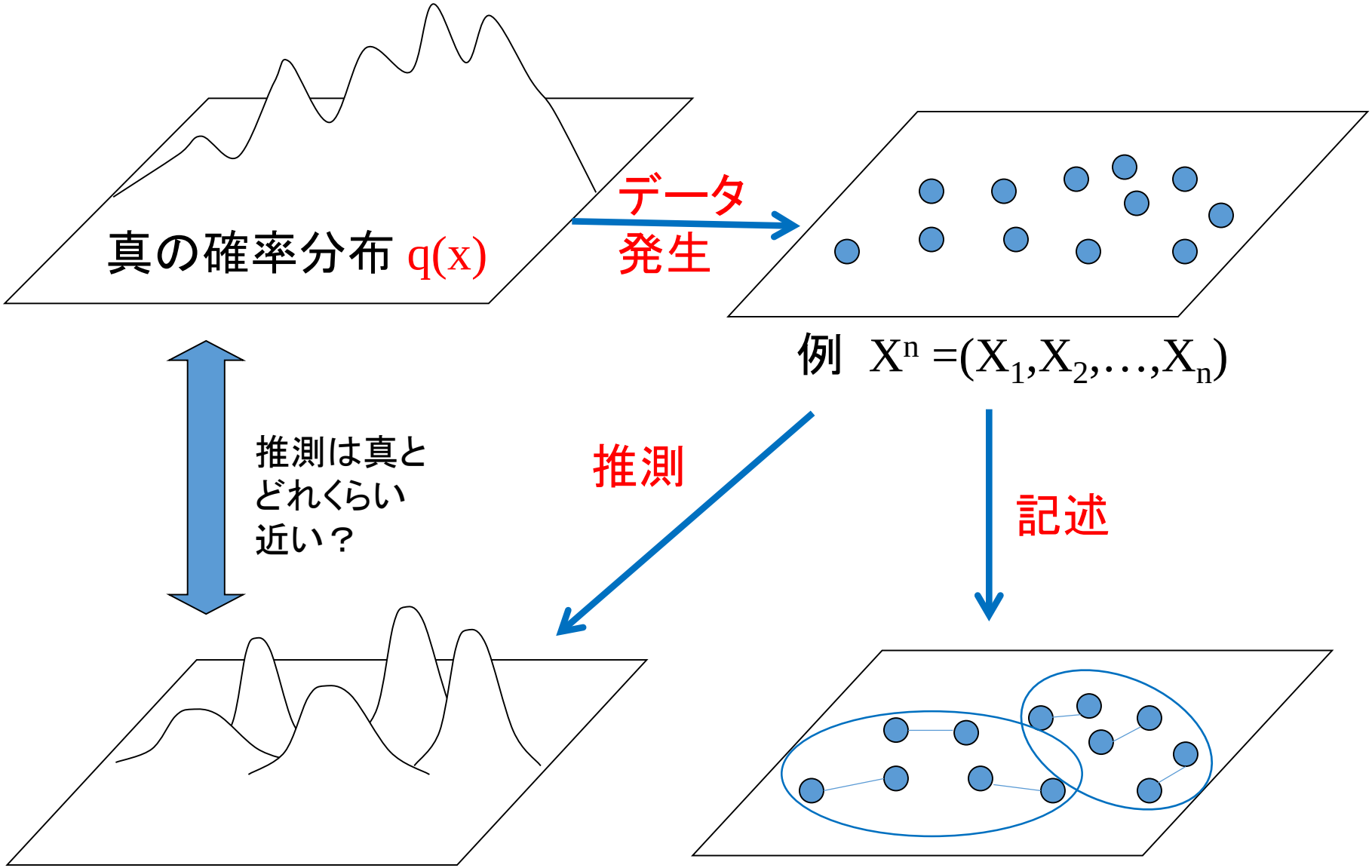


クラスタリング

統計的推測　データを発生した真の分布を推測したい
推定された分布が真の分布に対してどのくらい正確か

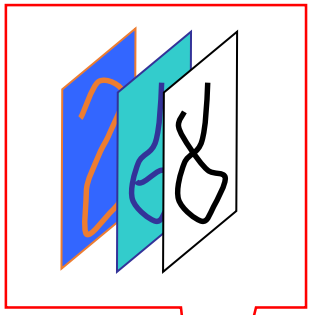
統計的記述　似ているデータをまとめたい
データをまとめることで発見できることはあるか

今日は、統計的推測におけるクラスタリングの問題を考える。そのモデルとして混合正規分布を使う。


真の確率分布 q(x)
データ
発生
例 $X^n = (X_1, X_2, \ldots, X_n)$
推測は真と
どれくらい
近い？
推測
記述
推測された分布 p(x|w)
似ているデータをまとめる

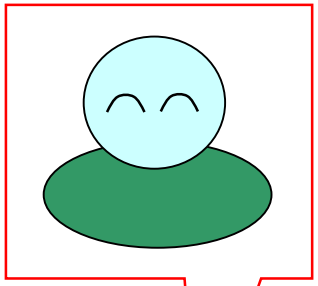
最尤推定法と
カルバックライブラ・情報量

統計的推測 と 機械学習



学習データ

$X_1, X_2, \ldots, X_n$



$q(x)$

$p(x|w)$

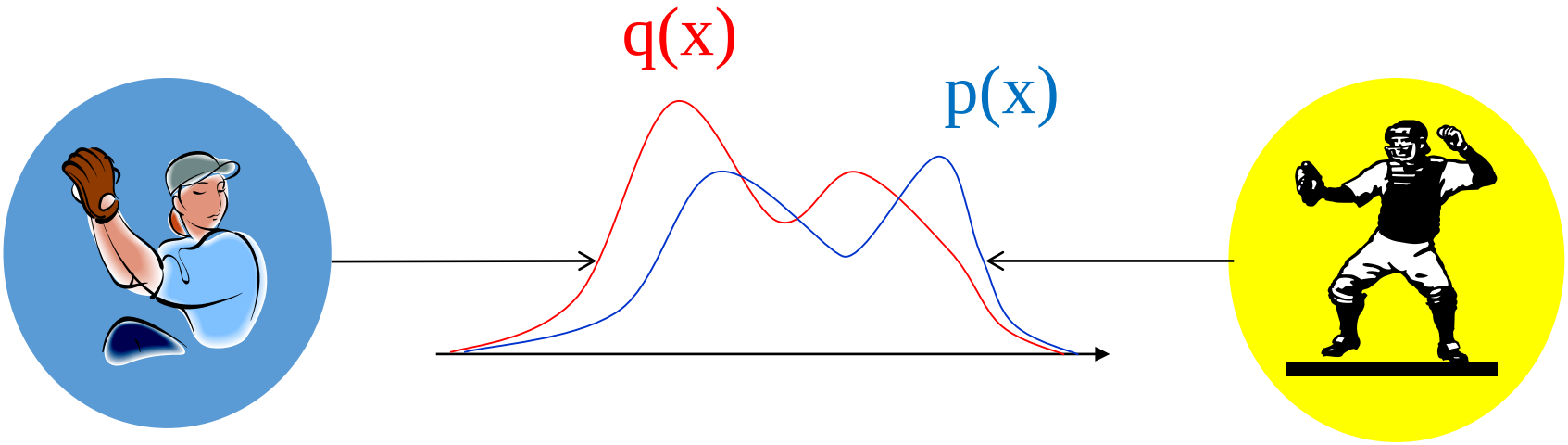
テストデータ

X

情報源

統計モデル
人工知能

真の分布と学習結果の距離



投手 q(x).

保守 p(x).

カルバック・ライブラ距離

定義. 二つの確率密度関数 q(x) と p(x) のカルバック・ライブラ距離を次の式で定義する。

$$D(q\|p) = \int q(x) \log \frac{q(x)}{p(x)} \, dx.$$

注意. 歴史上初めてこの量の大切さに気づいたのは物理学者ボルツマンである. 相対エントロピーという。ボルツマンは時間の不可逆性を示そうとしたがその夢は今でも実現されていない。後、統計学や情報理論で最も重要な量であることが発見された（フィッシャー、シャノン、赤池）。

定理. q(x), p(x)は連続関数, q(x),p(x)>0 を仮定する。

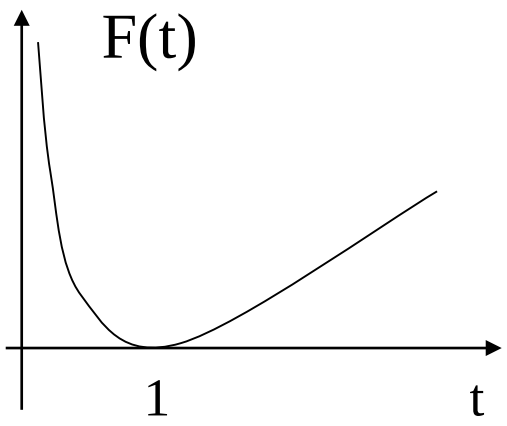
(1) 任意の q(x) について p(x), D(q||p) >=0.

(2) “D(q||p)=0” と “q(x) = p(x) (for all x)” は同値。

証明。

変数 $t>0$, について関数 $F(t) = \log t + 1/t - 1$ を定義すると

常に $F(t)>=0$ であり $F(t) =0$ は $t=1$のときに限る。



$$D(q\|p) = \int q(x)\{ \log \frac{q(x)}{p(x)} \} dx$$

$$= \int q(x)\{ \log \frac{q(x)}{p(x)} + \frac{p(x)}{q(x)} - 1\} dx$$

$$= \int q(x) F(\frac{q(x)}{p(x)}) dx >= 0,$$ より (1) が示せた。

(2) $D(q\|p)=0$ が成り立つのは 任意の x について $F(q(x)/p(x))=0$ のときであるから $q(x)=p(x)$ が得られる。 証明終

汎化誤差

真の分布を q(x) とし学習の結果を p(x) とするとき汎化誤差は

$$G = \int q(x) \log (q(x) / p(x)) dx$$

汎化誤差は真の分布が不明であるときには計算できないが

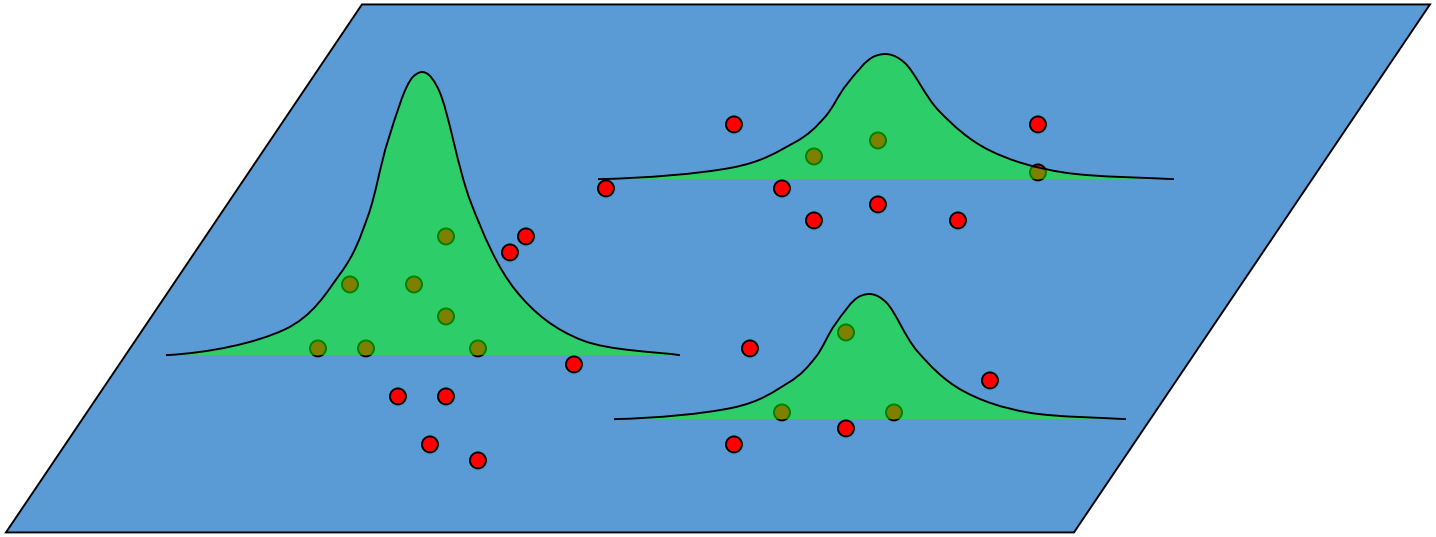
計測された値だけを使って汎化誤差を推定する方法を作ることができる。そのための方法が情報量規準および交差検証である。

☆　真の分布を推定するためには、統計モデルや事前分布を人間が用意することになる。統計モデルや事前分布が何であっても、その統計モデルと事前分布による汎化誤差を推定することができるということである。

混合正規分布と潜在変数

混合正規分布

データを発生した真の分布をいくつかの正規分布の混合を用いて推測する。



平均 b_k ,分散σ_k^2 の正規分布の重み $\{a_k\}$ の和

1個の正規分布

x : N 次元ベクトル
パラメータ　　w = (b ,σ)

平均 b ,分散σ^2 (標準偏差はσ)の正規分布

$$p(x|w) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(- \frac{\| x - b \|^2}{2\sigma^2} \right)$$

（注意）一個の正規分布を推測する場合には分散だけでなく分散共分散行列も容易に推測できる。混合正規分布の場合でも分散共分散は数式的には推測できるが、高次元になると推測が不安定になる。

混合正規分布

x : N 次元ベクトル

パラメータ　$w = \{(a_k, b_k, \sigma_k) ; k=1,2,\ldots,K\}$

→ データからパラメータをうまく決めたい。

$$p(x|w) = \sum_{k=1}^{K} a_k \frac{1}{(2\pi\sigma_k^2)^{N/2}} \exp\left(-\frac{\| x - b_k \|^2}{2\sigma_k^2}\right)$$

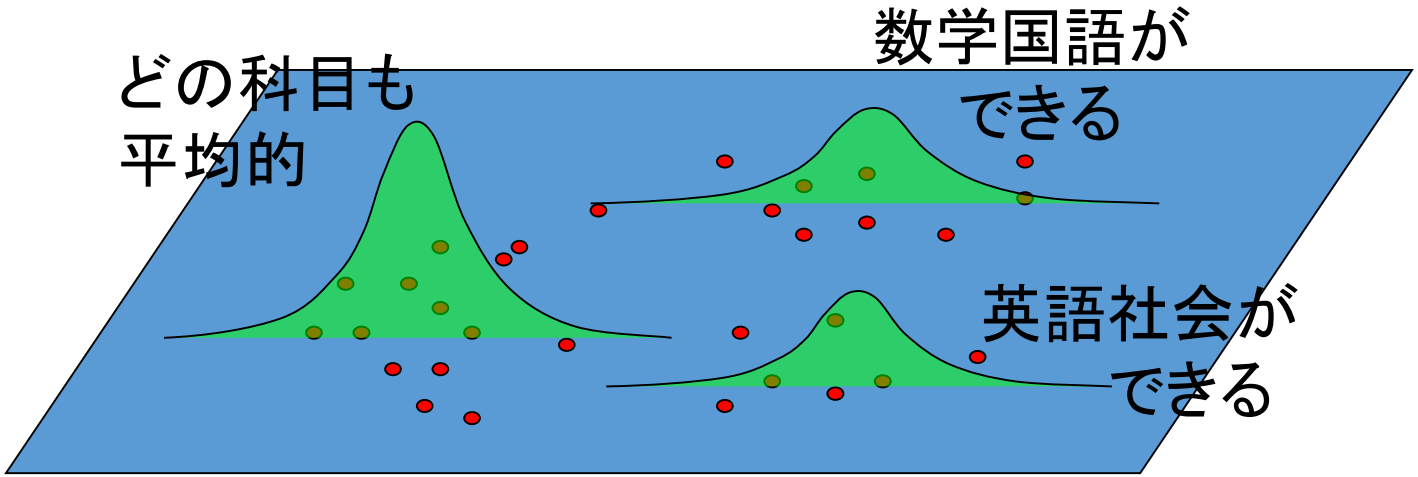
平均 b_k ,分散σ_k^2 の
正規分布

確率密度関数となるための条件

$$ak \geqq 0, \qquad \sum_{k=1}^{K} a_k = 1$$

混合正規分布の使い方

（例）1000人の中学生について5次元のデータ(国語、数学、理科、社会、英語）があったとき、どのような群がどのくらいの割合($\{a_k\}$)であるか知りたい。また、それぞれの群の平均と分散($\{b_k,\sigma_k^2\}$) を知りたい。



隠れ変数（潜在変数）の導入

$$p(x|w) = \sum_{k=1}^{K} a_k \frac{1}{(2\pi\sigma_k^2)^{N/2}} \exp\left(-\frac{\|x - b_k\|^2}{2\sigma_k^2}\right)$$

y について和をとる

$$p(x,y|w) = \prod_{k=1}^{K} \left[a_k \frac{1}{(2\pi\sigma_k^2)^{N/2}} \exp\left(-\frac{\|x - b_k\|^2}{2\sigma_k^2}\right) \right]^{y_k}$$

$y = (y_1, y_2, .., y_k)$ はひとつだけ1で残りは0。つまり
$y \in \{(1,0,..,0), (0,1,0,\ldots,0),\ldots,(0,0,\ldots,1)\} \equiv C^K$

混合正規分布の推定 = 隠れ変数とパラメータの推定

問1

$y = (y_1, y_2, .., y_K)$ はひとつだけ1で残りは0。つまり
$y \in \{(1,0,..,0), (0,1,0,\ldots,0),\ldots,(0,0,\ldots,1)\} \equiv C^K$
とする。次の和を求めよ。

$$\sum_{y \in C^K} \prod_{j=1}^{K} (j)^{y_j}$$

混合正規分布の推定

ギブスサンプラ法

(X,Y) が確率密度関数 p(x,y) に従う確率変数であるとき

p(x|y) に従う X を発生
p(y|x) に従う Yを発生

を繰り返して { (X_j,Y_j), j=1,2,…,J} を得ると

任意の関数 f(x,y) について平均が有限ならば

$$E_{XY}[\, f(X,Y)\,] = \lim_{J \to \infty} (1/J) \sum_{j=1}^{J} f(X_j,Y_j),$$

ベイズ推定の方法

(1) (a,b,c) の事前分布 φ(a) φ(b) φ(c) を用意する。

(2) $(x_1,\ldots,x_n, y_1,\ldots,y_n, a,b,c)$ の同時確率密度は

$$\prod_{i=1}^{n} p(x_i,y_i|a,b,c)\, \varphi(a)\, \varphi(b)\, \varphi(c\ ).$$

(3) $(x_1,\ldots,x_n)$ が与えられたときの$(y_1,\ldots,y_n,a,b,c)$ の条件つき確率密度はこの式に比例する。

$$p(y_1,\ldots,y_n,a,b,c|x_1,\ldots,x_n) \propto \prod_{i=1}^{n} p(x_i,y_i|a,b,c)\, \varphi(a)\, \varphi(b)\, \varphi(c\ )$$

ベイズ推定の方法（つづき）

(4) $(y_1,\ldots,y_n)$ と (a,b,c) をギブスサンプラ[（あ）（い）の繰り返し] で推定。
（あ） $p(y_1,\ldots,y_n|a,b,c,x_1,\ldots,x_n)$ から $(y_1,\ldots,y_n)$ をサンプリングし
（い） $p(a,b,c|x_1,\ldots,x_n,y_1,\ldots,y_n)$ から (a,b,c) をサンプリングする。

(5) 得られた $\{(a_j,b_j,c_j)\ ;\ j=1,2,\ldots,J\}$ を用いて

$$p(x) = (1/J) \sum_{j=1}^{J} p(\,x \mid a_j,b_j,c_j\,)$$

をデータを発生した分布を推定したもの（学習したもの）とする。

(6) 真の分布と推定した結果の誤差（汎化誤差）を推定する量を計算する。
この計算の方法については、講義の後の回で述べる。

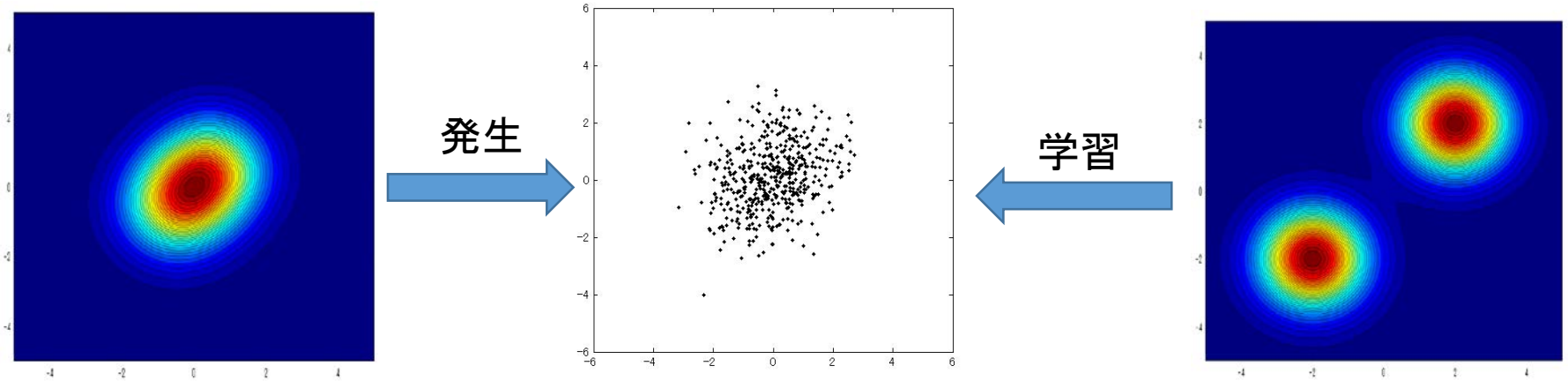
簡単な例で説明する

混合正規分布:　$x \in \mathbf{R}^2$ の確率密度関数で $w=(a,b,c)$ をパラメータに持つもの

$$p(x|w) = (1-a)\,N(x-b) + a\,N(x-c)$$

パラメータ w は $(0 \leqq a \leqq 1,\ b, c \in \mathbf{R}^2)$.　$N(x)$ は $x \in \mathbf{R}^2$ の正規分布とします.

真の分布: データの数は n で　$p(x|0.5,(0.5,0.5),(-0.5,-0.5))$　から独立に発生。



真の分布
(2山だが区別しにくい)

データ

モデル

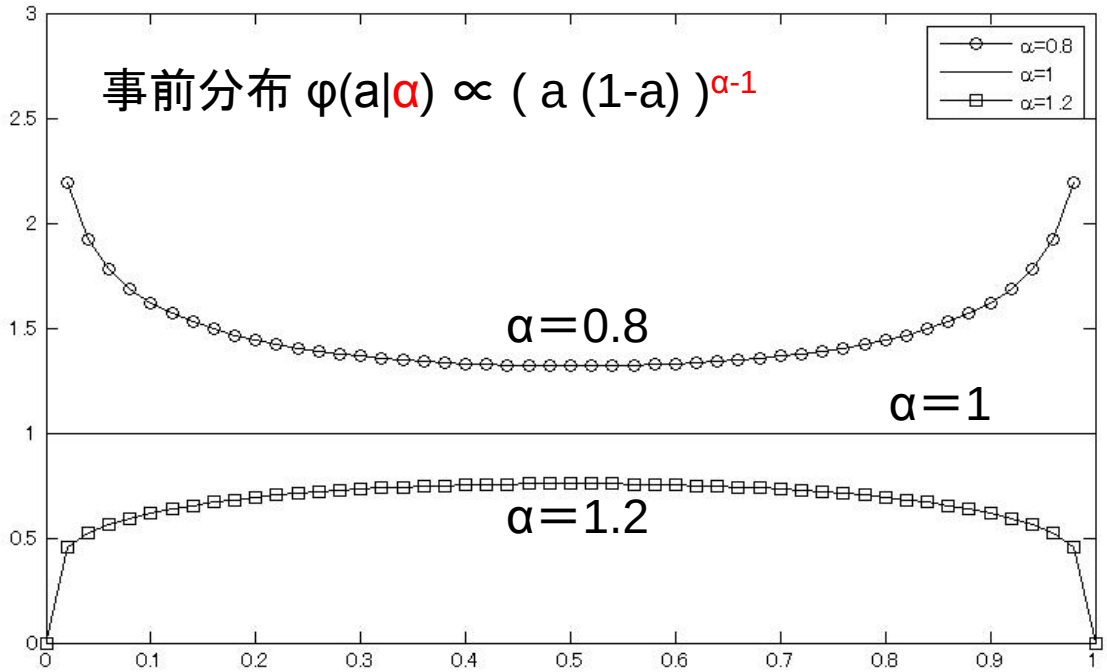
事前分布

事前分布にはディリクレ分布を使います。

事前分布 (α,hをハイパーパラメータに持つ)　(αの影響を調べる. h=0.0001は固定)

事前分布　$\varphi(a|\alpha) \propto (a(1-a))^{\alpha-1}$　　ディリクレ分布（α>0)

事前分布　$\varphi(b,c|h) \propto \exp\{-(h/2)(\|b\|^2 + \|c\|^2)\}$



（注意）α が大きいほど
事前分布の a の値は
0.5 の近くに集まります。

ギブスサンプラーの作り方

データを $\{x_i \in \mathbf{R}^2 ; i=1,2,\ldots,n\}$ とします。

隠れ変数 $\{y_{ik} ; i=1,2,\ldots,n, k=0,1\}$ を用いてギブスサンプラーを作りましょう。

(1) 各 $\{y_{ik}\}$ の初期値を、各 i ごとに $y_{ik}=0$ または $y_{ik}=1$ と決める（任意でよい）.

(2) $n_k= \Sigma\, y_{ik}$ (k=0,1) とする. ただし（ Σ は i=1,2,…,n の和）。

(3) $B_0=\{ \Sigma(1-y_i)x_i \} /(h+n_0)$, $B_1= \{ \Sigma y_i x_i \} /(h+n_1)$ を算出する。

(4) $s_k=1/(h+n_k)$ (k=0,1) とする。

(5) ディリクレ分布 $\propto (a_0)^{\alpha-1+n_0} (a_1)^{\alpha-1+n_1}$ から(a_0,a_1)を発生。

(6)「平均 B_k 共分散 $s_k I$ の正規分布」から b_k を発生 (k=0,1. I は2×2の単位行列)

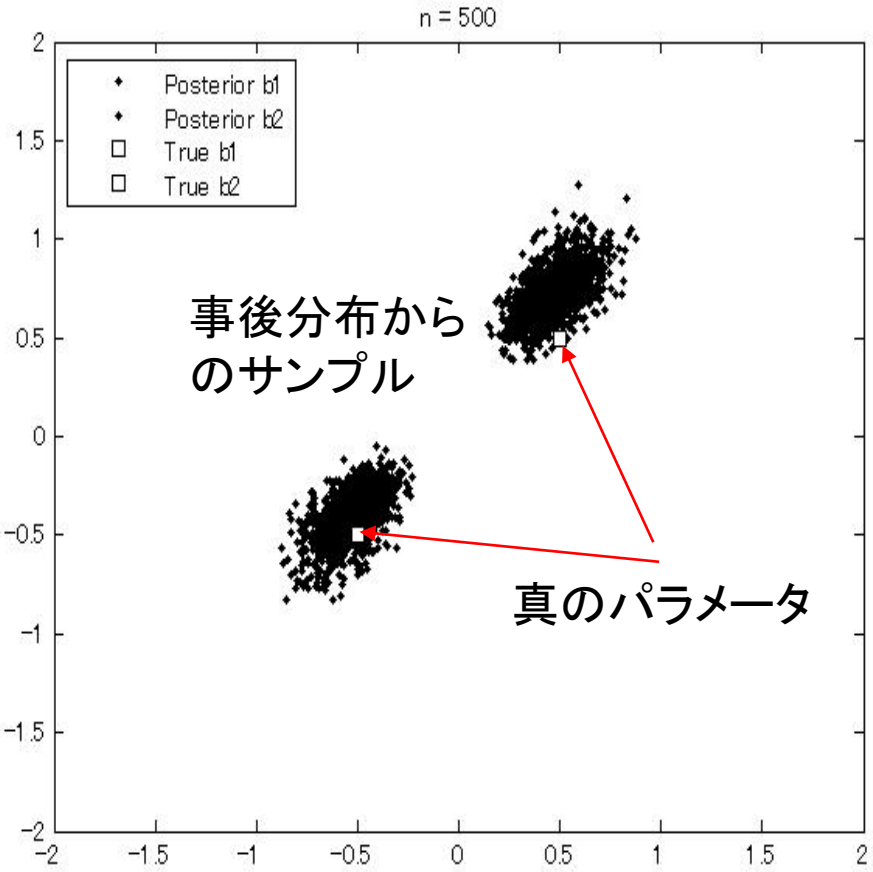
(7) $L_k= a_k \exp(-(1/2)\|x_i-b_k \|^2)$ (k=0,1) を用いて確率 $L_k/(L_0+L_1)$ で $y_{ik}=1$ とする。

(8) (2)に戻って繰り返す。

初期値の影響が無くなってから $\{(a_1,b_0,b_1)\}$ を取り出して (a,b,c) の事後分布からのサンプリングとする。

事後分布の表示

事後分布はギブスサンプラーで生成します。事後分布が n や α の変化とともに
どのように変わるかを目でみて確かめてみましょう。



事後分布は (a,b_1,b_2,c_1,c_2) で表される5次元の空間の上の確率分布です。5次元空間を描くことはできないので、以下では、次のように表示しています。

(1) a については描画しない。

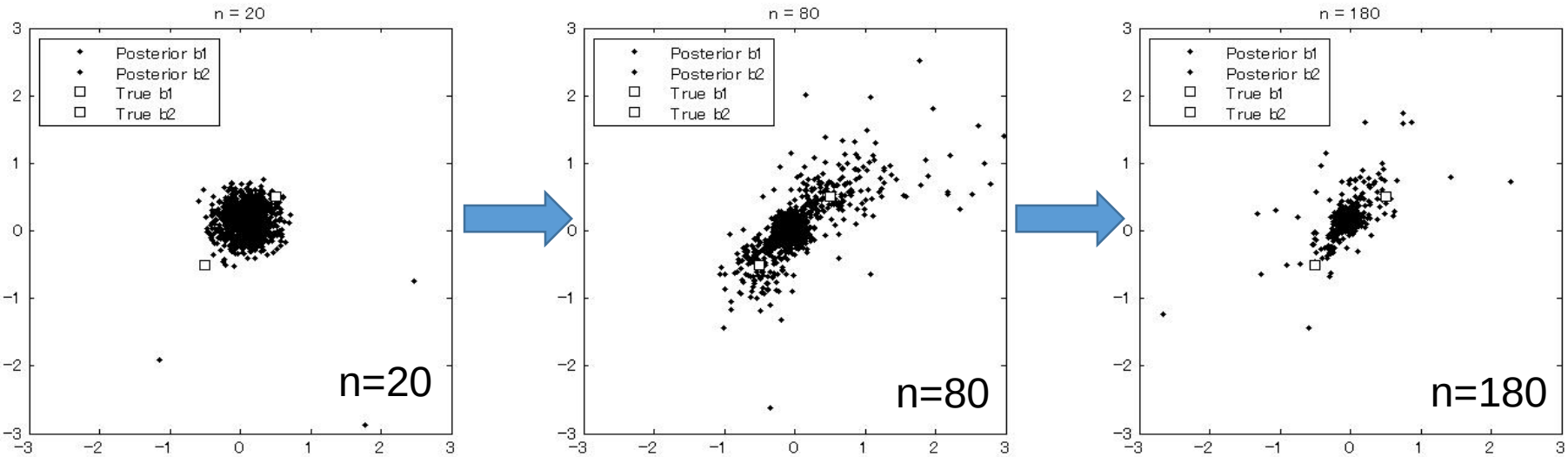
(2) (b_1,b_2) と (c_1,c_2) を同じ図の上に重ねてプロットする。

・ 事後分布のサンプル点

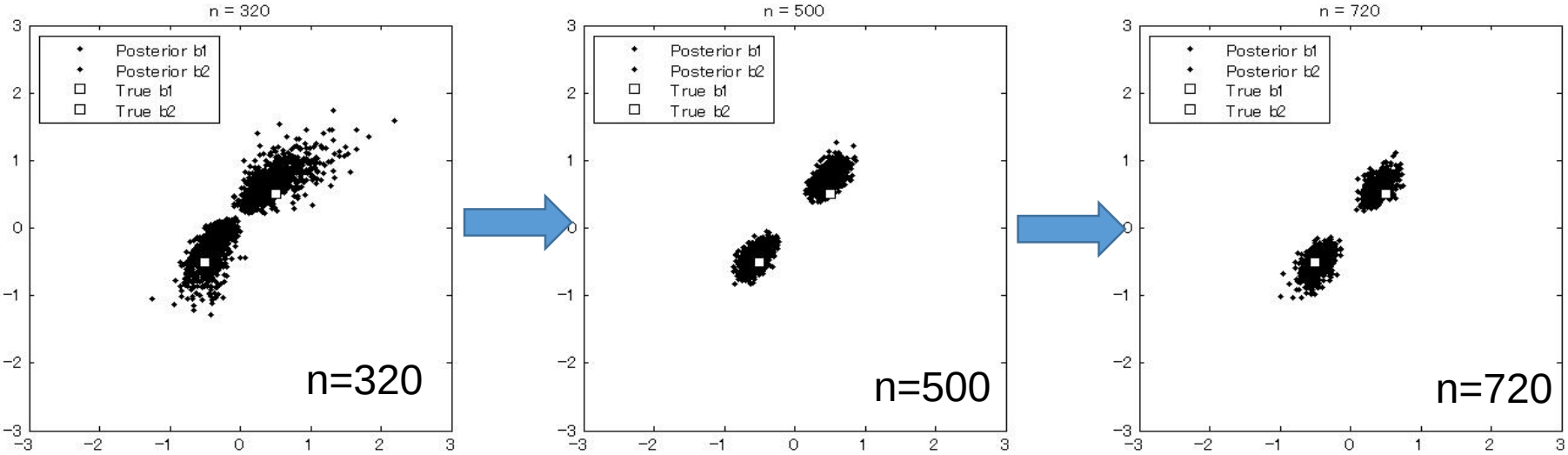
□ 真のパラメータ（2個）

（注) a についても描画するとさらに面白いのですが、わかりにくくなるのでやめました。

事後分布のデータの数 n による変化

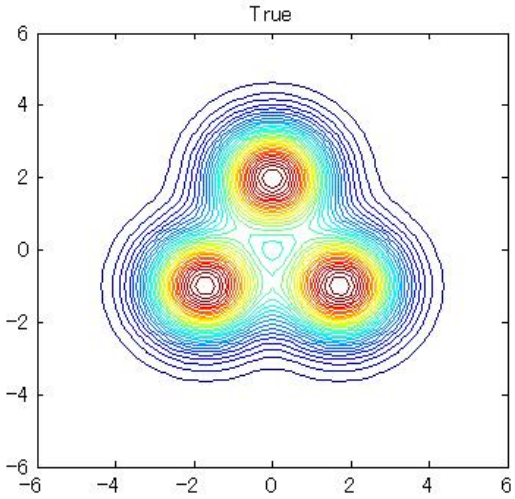


データの数 n が増えると事後分布が1山から2山に分かれます(構造の発見)。

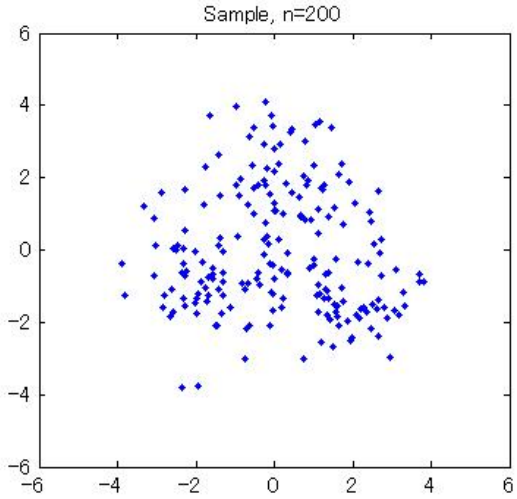


3個の場合の例

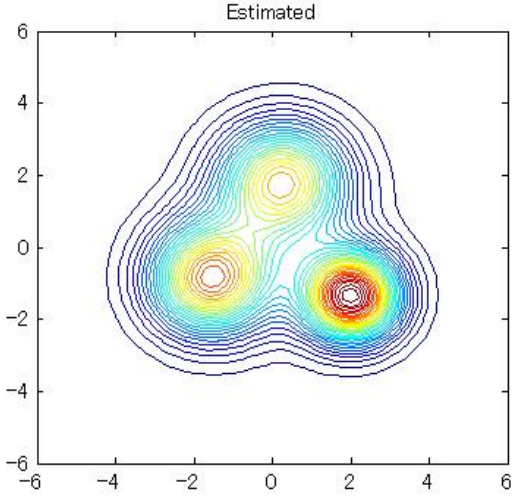
真の分布



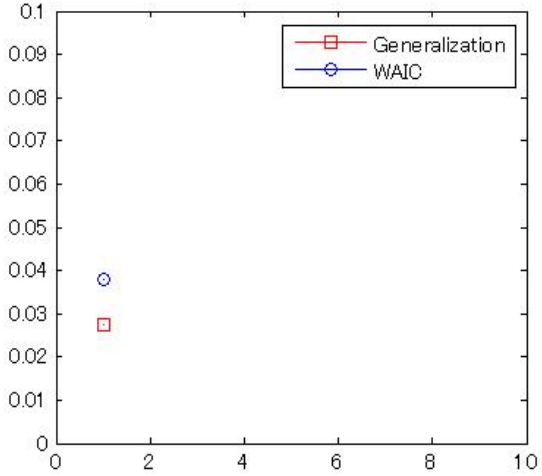
データ



推定された分布

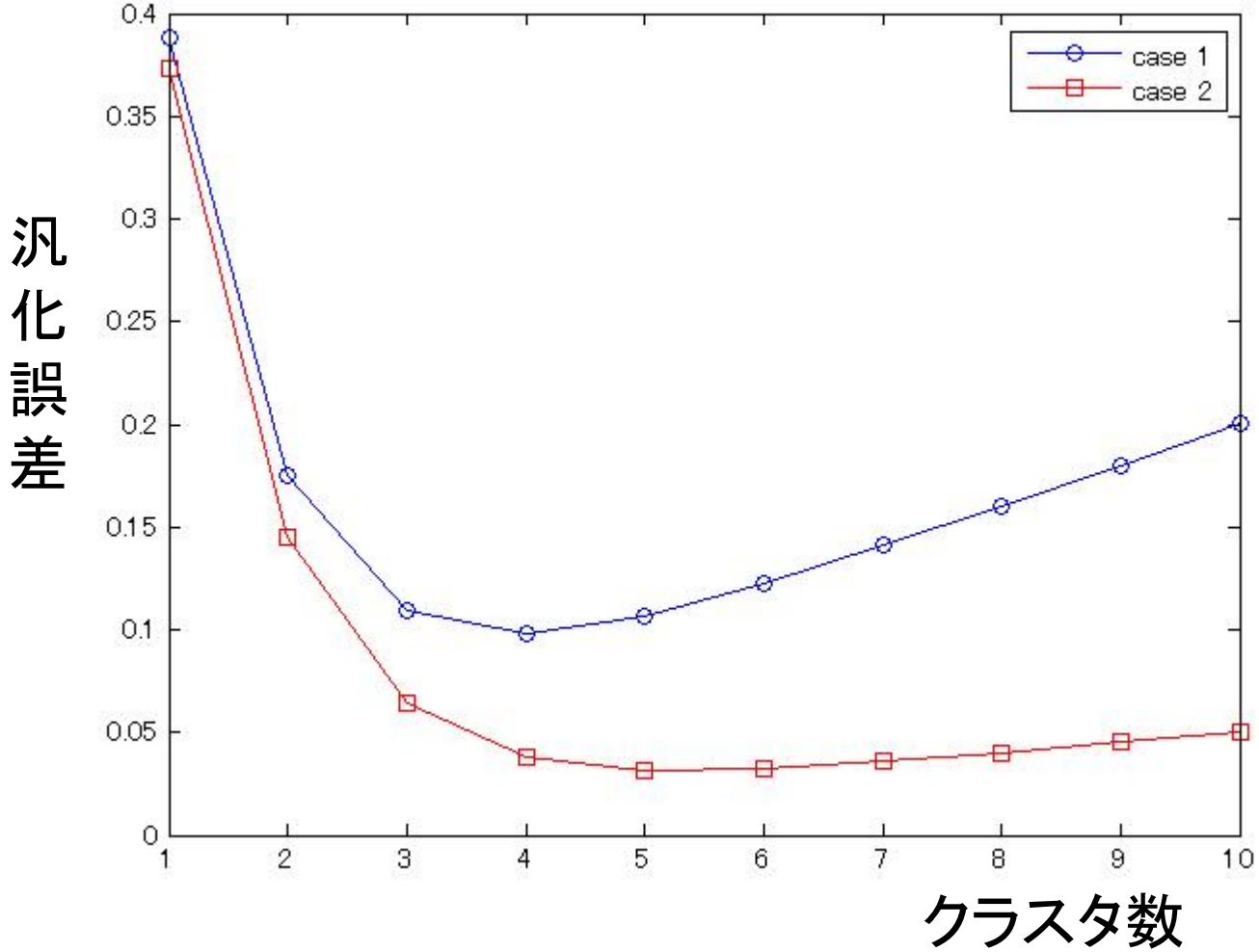


汎化誤差と推定量



問2

与えられたクラスタ数に対して、そのモデルを用いたときの汎化誤差を求めることができると、汎化誤差を最小にするようにモデル（クラスタ数）を定めることができる。下記の例で最適だと思うクラスタ数を求めよ。



実データの例

例２

「街の単位」を「◎町〇丁目」とする。横浜市にある1070個の街で下記の業種で働く人の人数を調査して10次元のベクトルを1070個得た。

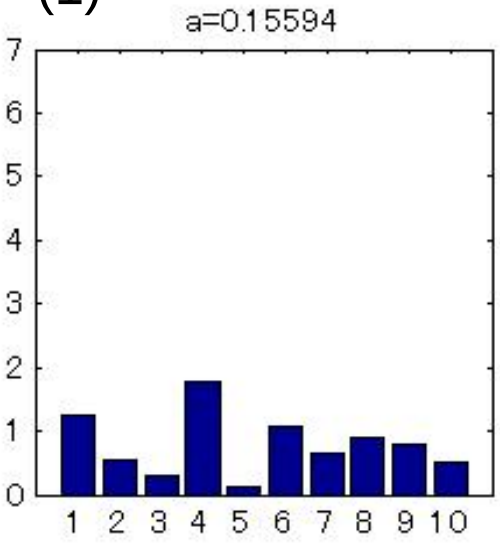
政府統計の総合窓口（ｅ－Stat）を使わせて頂きました。
http://www.e-stat.go.jp/estat/html/spec.html
http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do

1 建設業
2 製造業
3 運輸郵便業
4 卸売小売業
5 金融保険業
6 不動産物品賃貸業
7 学術研究専門・技術サービス業
8 宿泊飲食サービス業
9 生活関連サービス娯楽業
10 教育学習支援業

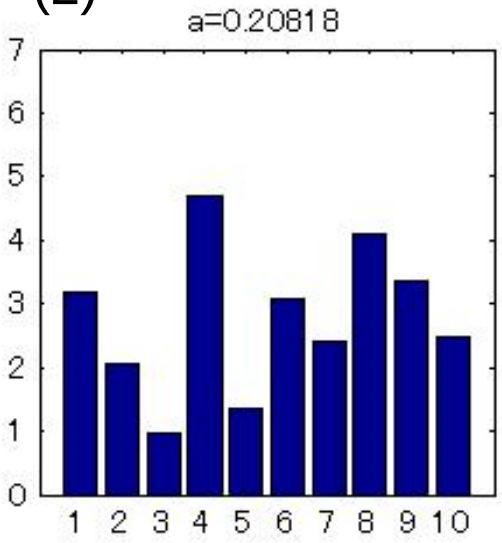
データについて、log(1+X)を混合数 K=6 個の混合正規分布で推測を行った。

例2

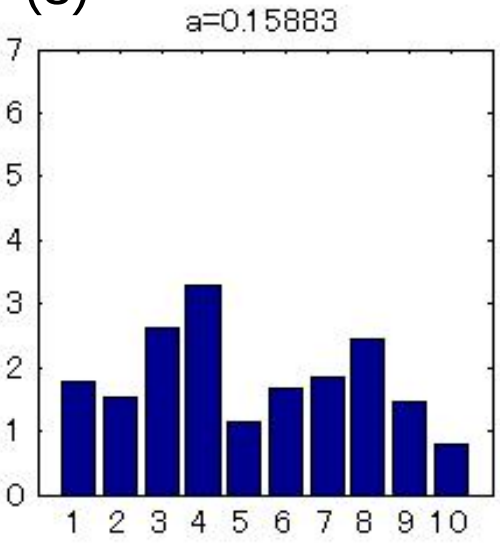
(1)


a=0.15594
0 1 2 3 4 5 6 7
1 2 3 4 5 6 7 8 9 10

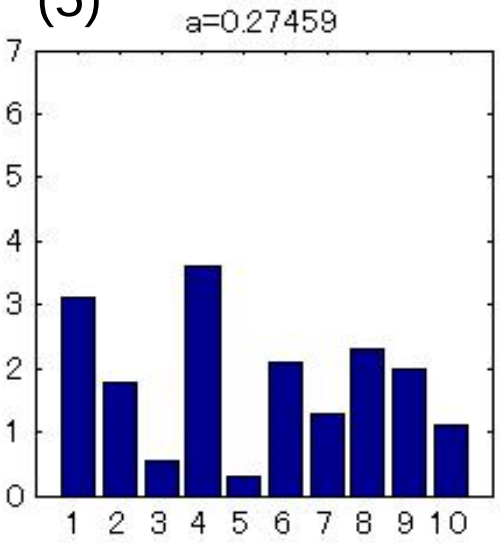
(2)


a=0.20818
0 1 2 3 4 5 6 7
1 2 3 4 5 6 7 8 9 10

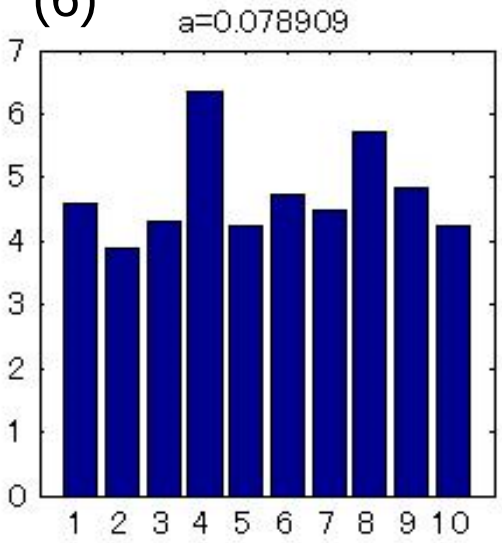
(3)


a=0.15883
0 1 2 3 4 5 6 7
1 2 3 4 5 6 7 8 9 10

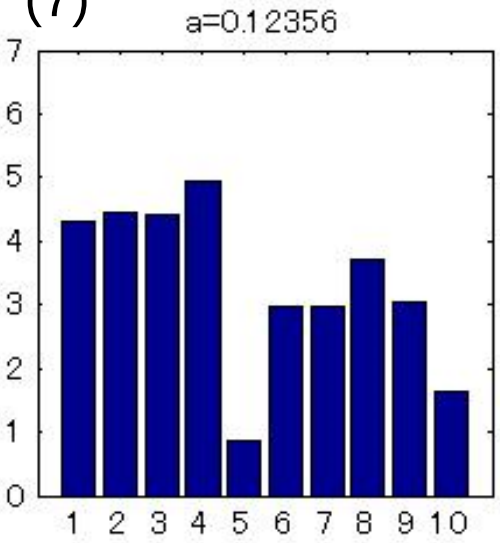
(5)


a=0.27459
0 1 2 3 4 5 6 7
1 2 3 4 5 6 7 8 9 10

(6)


a=0.078909
0 1 2 3 4 5 6 7
1 2 3 4 5 6 7 8 9 10

(7)


a=0.12356
0 1 2 3 4 5 6 7
1 2 3 4 5 6 7 8 9 10

問3

(1)から(6)までの街のクラスタリングの結果から
わかったことを述べよ。