データ解析

第６回：クラスタ分析

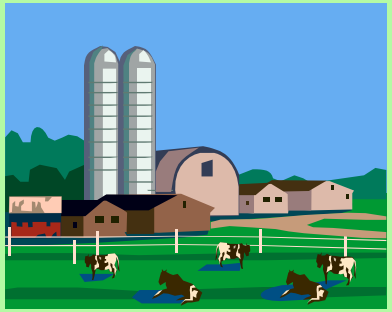
渡辺澄夫

実世界は
複数の
群から
なる？

観測データ

回帰・判別分析

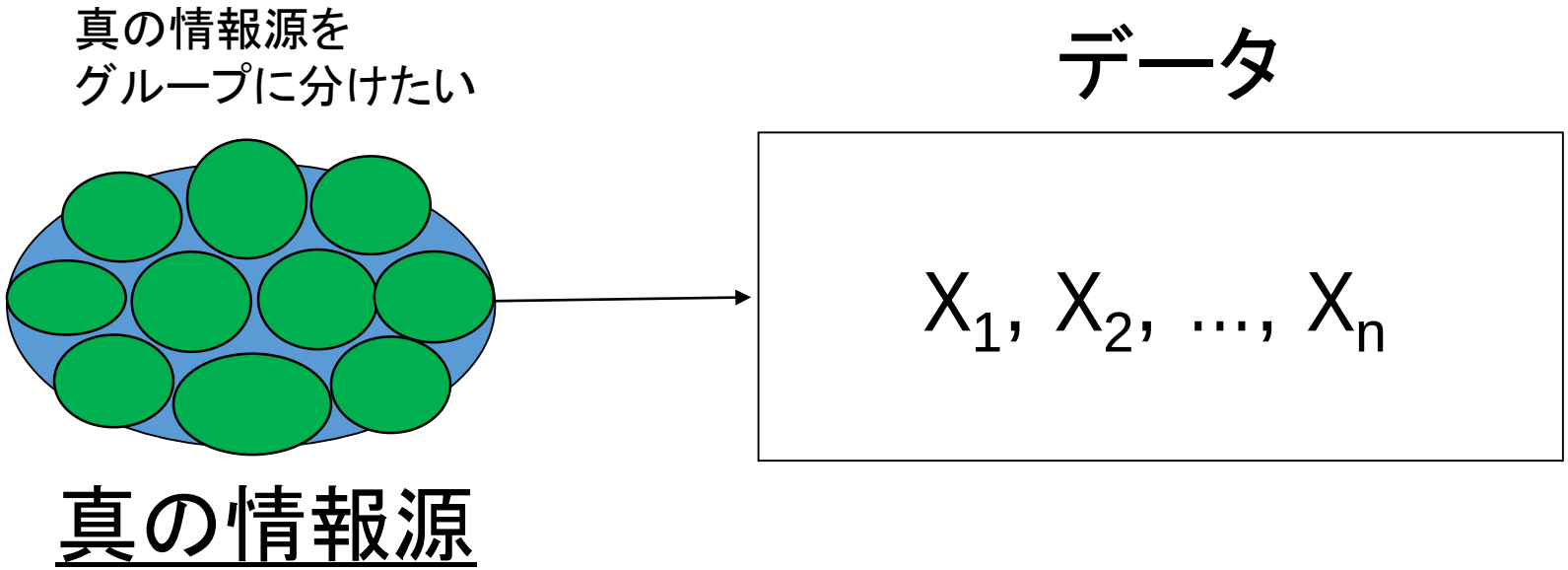
解析方法



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

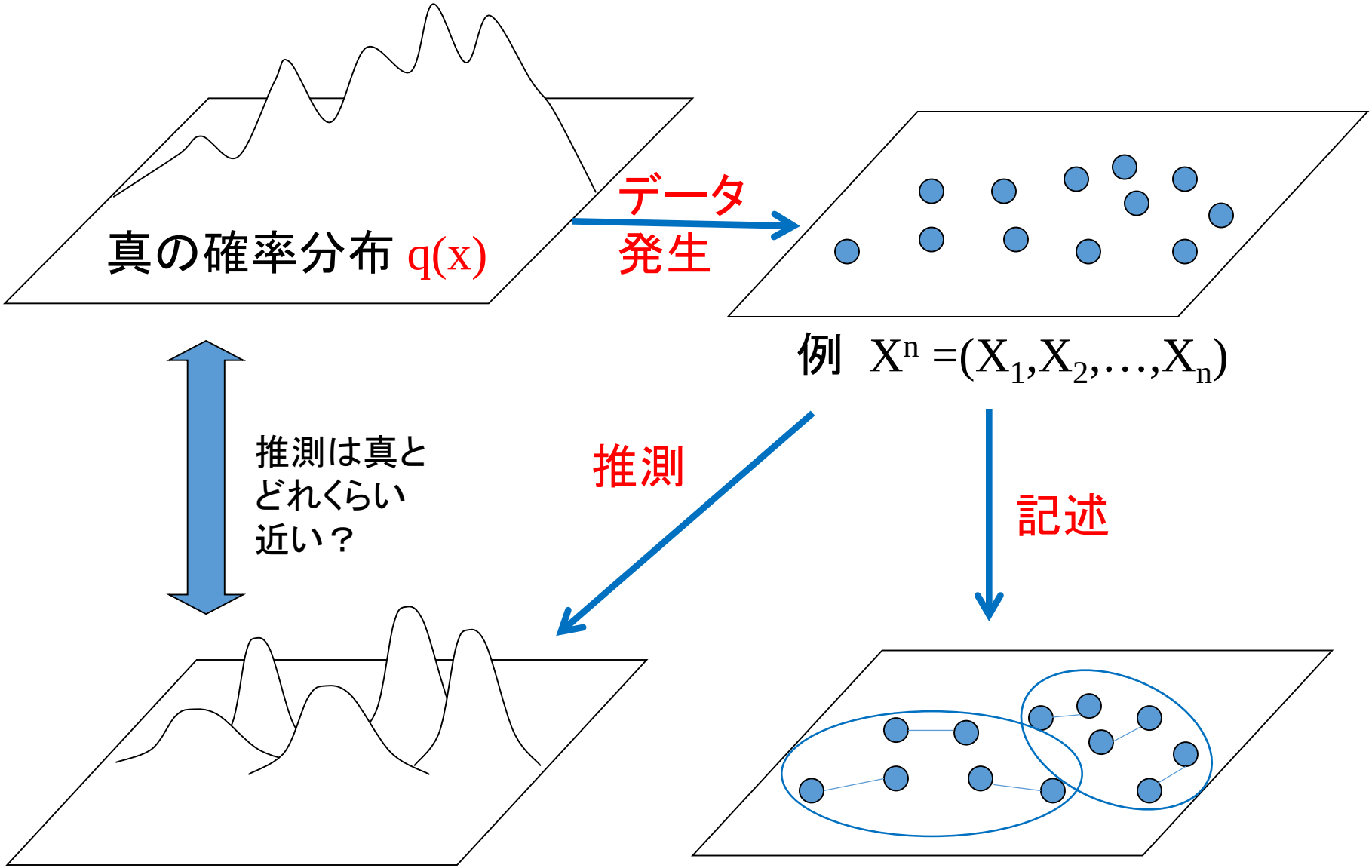


クラスタリング

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

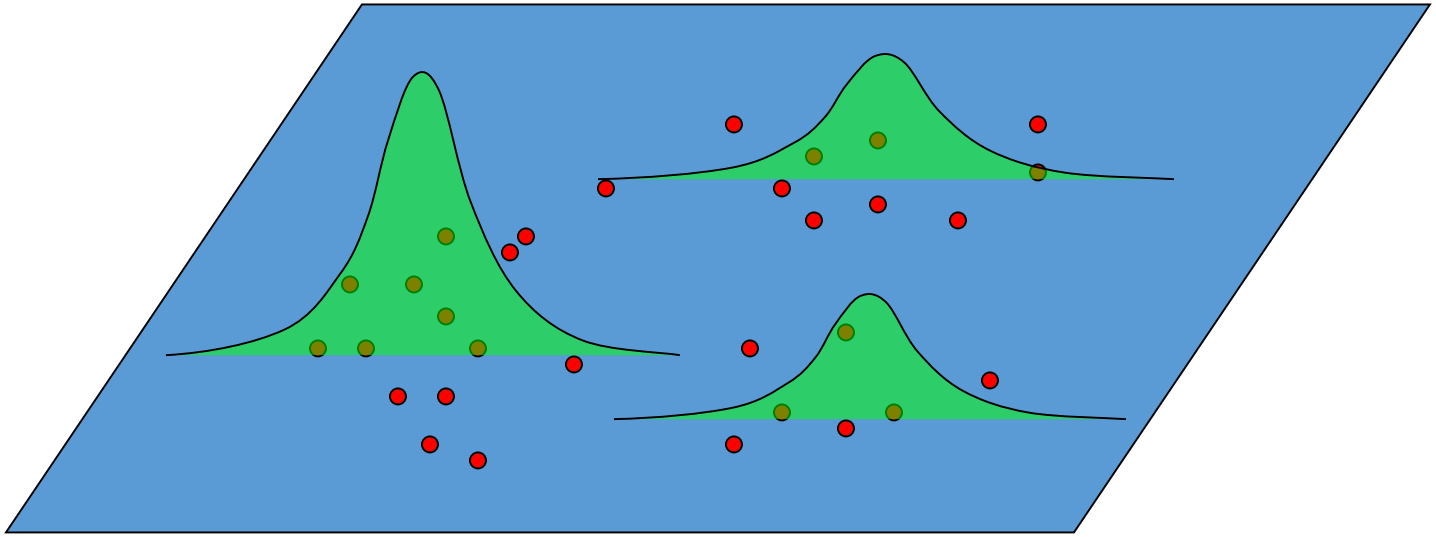
今日は、統計的推測におけるクラスタリングの問題を考える。そのモデルとして混合正規分布を使う。


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

混合正規分布の定義

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。

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

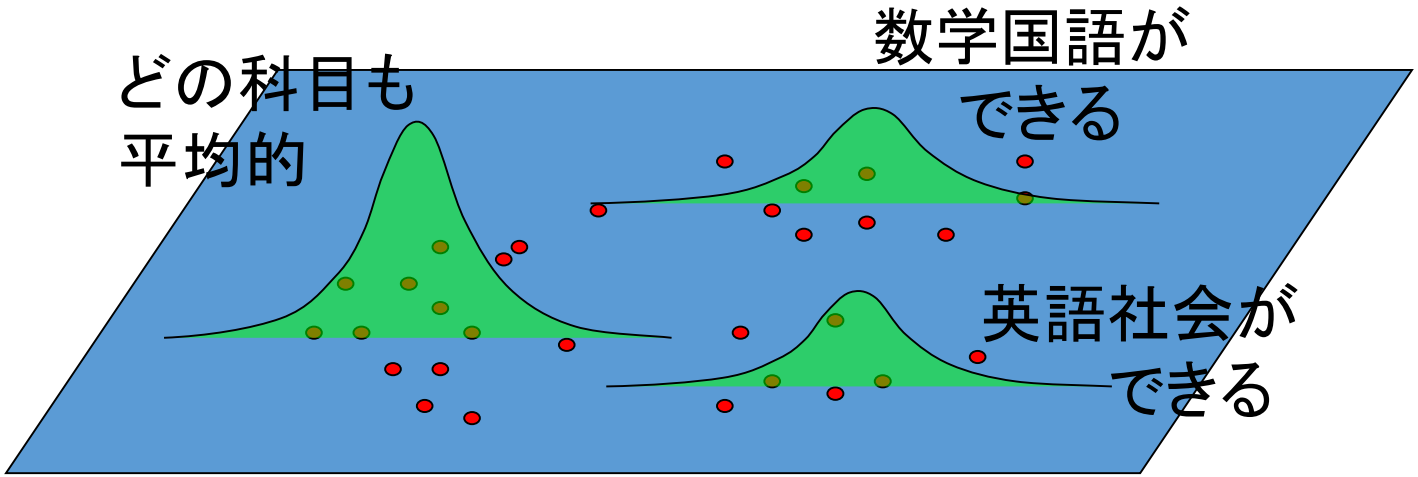
平均 b_k ,分散σ_k^2 の
正規分布

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。



最尤推定法と
カルバックライブラ・情報量

カルバック・ライブラ情報量

定義. 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あるいは相対エントロピーという。

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) $K(q||p)=0 \Leftrightarrow q(x)=p(x)\ (\forall x)$

証明。$t>0$ の関数を $F(t) = \log t + 1/t - 1$ とおき、微分して増減表を書くとで $F(t) \geqq 0$ であり、$F(t)=0 \Leftrightarrow t=1$ がわかる。また$q(x), p(x)$ の積分値が1になることから $K(q||p)=\int q(x)\, F(q(x)/p(x))\, dx$.
これより(1),(2) が得られた。(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わからないときには計算できないが

$$G = \int q(x) \log \; q(x) \, dx \quad - \int q(x) \log p(x) \; dx$$

なので、対数損失（右辺第２項）が小さいことと汎化誤差が小さいことは等価。

できるだけ G を小さくする方法を目指して今日でも多くの研究がなされています。

最尤法と汎化誤差

真の密度関数 $q(x)$ に独立に従う確率変数を $X_1, X_2, \ldots, X_n$ とする。パラメータ w を持つ統計モデルを $p(x|w)$ とする。対数尤度を $\log L(w)$ とかくと、大数の法則から概収束

$$\log L(w) = (1/n) \Sigma_i \log p(X_i|w) \rightarrow \int q(x) \log p(x|w)\, dx$$

が各 w ごとに成立する。もしも上記の収束がパラメータ w について一様性を持っていれば、尤度を最大にすることは汎化誤差を最小にすることに近いと思ってよい（ w についての最大化と $n \rightarrow \infty$ の極限の順番を変えられる）。その場合には最尤推定量はそれほど悪い推定量にはならない。

混合正規分布と EM 法

混合正規分布の対数尤度

データ $\{X_i; i=1,2,\ldots,n\}$ がえられたとき、混合正規分布

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

の対数尤度 $\log L(w) = \Sigma_i \log p(X_i|w)$ を最大にする w を探したい。logL(w) についての最急上昇法でも解けるが、今回は EM 法を用いることにする。

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
$y \in \{(1,0,..,0), (0,1,0,\ldots,0),\ldots,(0,0,\ldots,1)\} \equiv C^K$

混合分布の最尤推定

混合正規分布の推測は

各 x がどのコンポーネントから発生したかの情報 y (隠れ変数, 潜在変数）

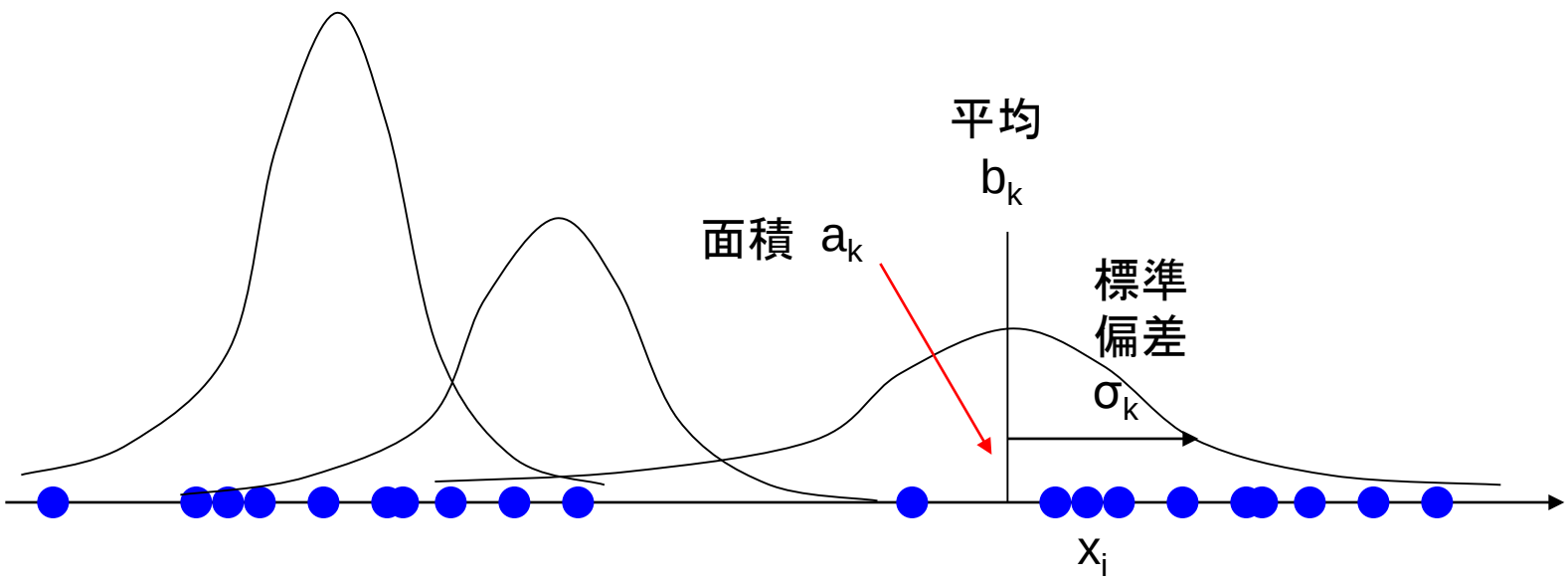
を計測できない場合の推測と等価である。

$$p(x|w) \quad \Leftrightarrow \quad p(x,y|w),\ \ y \text{ は隠れ変数}$$

Expectation -Maximization法

(1) パラメータ初期化
(2) パラメータを固定して隠れ変数推定
(3) 隠れ変数を固定してパラメータを推定

このとき尤度が単調非減少であることを証明できる（付録）。
局所的に尤度の大きな解が得られる。



パラメータ→隠れ変数推定

パラメータ w を固定して、データ X_i に対する隠れ変数 Y_k の平均値を計算する。

$$E[\, y_k | x_i , w]$$

$$= \Sigma_y \; y_k \, p(y | x_i, w)$$

$$= \Sigma_y \; y_k \, p(x_i, y | w) \,/\, p(x_i | w)$$

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

隠れ変数→パラメータ推定

隠れ変数の平均値を固定してパラメータを最尤推定する。

$$
\begin{cases}
a_k = \dfrac{\Sigma\, E[y_k|\, x_i\,,\, w]}{n} \\[2ex]
b_k = \dfrac{\Sigma\, E[y_k|\, x_i\,,\, w]\, x_i}{\Sigma\, E[y_k|\, x_i\,,\, w]} \\[2ex]
\sigma_k^{\,2} = \dfrac{\Sigma\, E[y_k|\, x_i\,,\, w]\, \|\, x_i - b_k \,\|^{\,2}}{\Sigma\, M\, E[y_k|\, x_i\,,\, w]}
\end{cases}
$$

（Σは i=1,..,n の和）

実験例

データ:　4個の正規分布から発生。

混合正規分布（K=4）で学習（EM法）。

初期値

a_k=1/K,

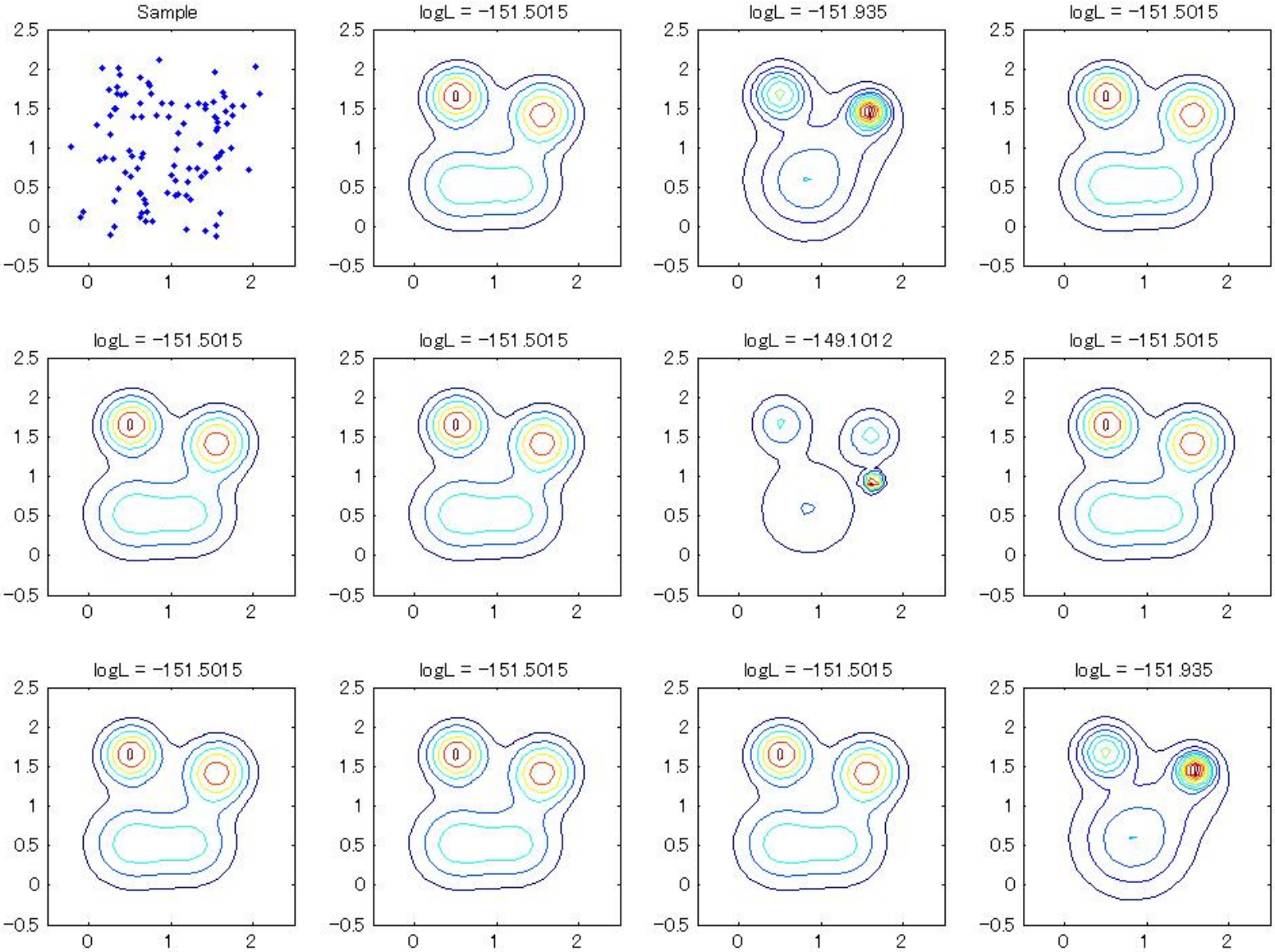
b_k= 全体平均＋乱数

σ_k= 全体標準偏差

から初めて500回繰り返し

初期値を変えて11回学習した。尤度が最も大きい解は、σ_kが小さくなっている。

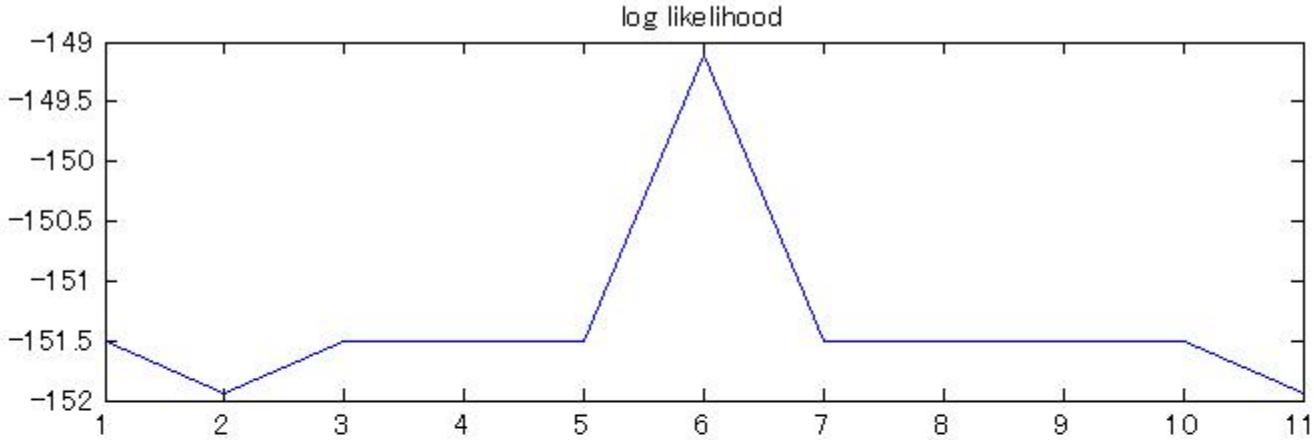
推定された分布と対数尤度



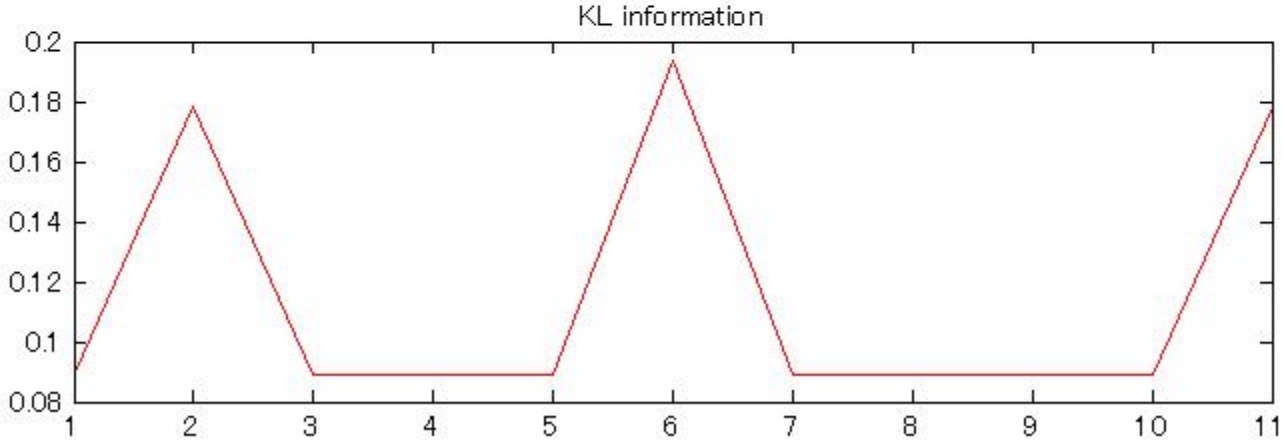
実験例

同じデータに対して１１回EM法を適用して得た局所解の対数尤度と汎化誤差。対数尤度が大きくても汎化誤差が小さいとは限らない。

対数尤度



汎化誤差



注意：混合正規分布の対数尤度関数の形状

このページに書いてあることは基本的なことですが、誰でも知っているとは限らないので、知らない人（統計学の専門家も含む）には説明する必要があります。

(1) 混合正規分布では最尤推定量は発散している。（$x_i=b_k$, $\sigma_k \to 0$）。

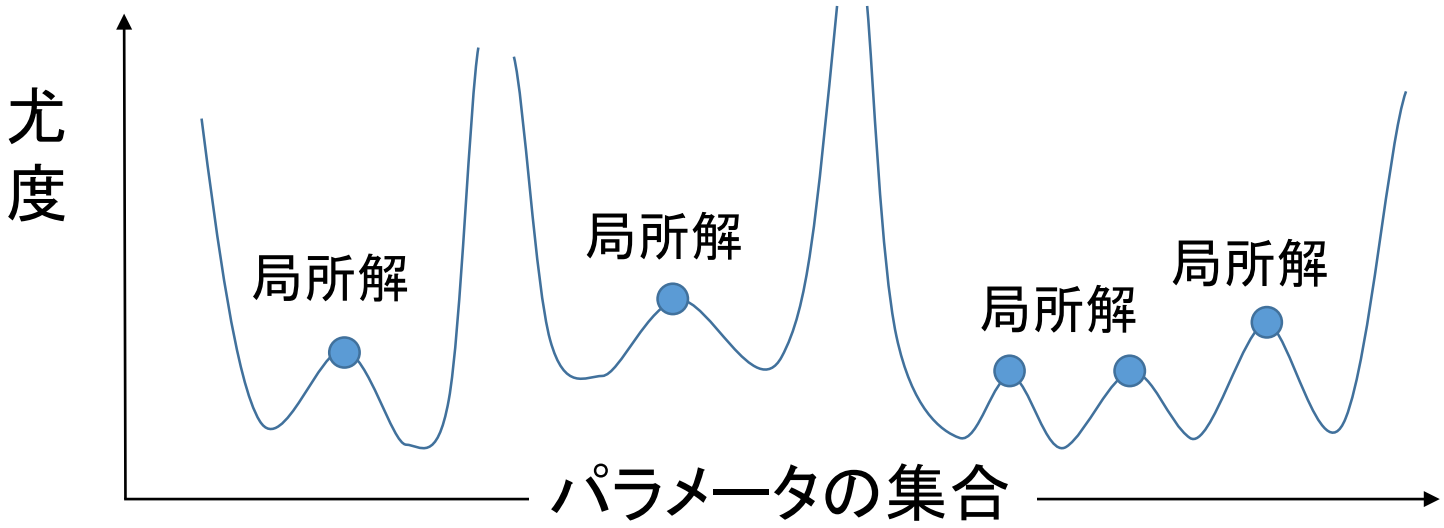
(2) 局所最適解はとてもたくさんあると思われる。

(3) 尤度が大きい局所解が推測の意味で良いのではないので、尤度で局所解の良さの比較はできない。AIC, BIC は使えない。クロスバリデーションも使えない。

→ 最尤法やEM法は「ダマシダマシ使う」必要がある（最尤法の弱点）。

(4) EM法が失敗するのは局所解に落ちたときではなく良い局所解に落ちないとき。

(5) 安心して使いたい場合は VB に変更する。EM を VB に変更して損失になることは何もない（演算量も）。ただし数式と理論は少し難しくなる。



参考:アルゴリズムの発展

(1) EM 法(1977)の問題点

EM法は最尤推定量をみつけようとする方法であるが
混合正規分布では最尤推定量は存在しないので
EM法が何をしているのかは実はよくわからない。
データの揺らぎに対して弱く、汎化誤差も大きい。

(2) VB 変分ベイズ法(1999)の発見

EM法とよく似た学習アルゴリズムを与える変分ベイズ法は
ベイズ事後分布の平均場近似により導かれるが、
データの揺らぎに対して強く、汎化誤差も小さくできる。
隠れ変数を持つシステムにおいて広く使われている。

参考:変分ベイズ法　あれこれ

(1)「パラメータと隠れ変数が独立な分布」から事後分布までのKL情報量を最小化する方法を平均場近似という。

(2) 事後分布の数値実現にはマルコフ連鎖モンテカルロ法が用いられるが、平均場近似は繰り返し代入法で実行できる。

(3) 平均場近似によるパラメータの分布は真の事後分布よりも縮んでいる。

(4) 平均場近似の自由エネルギーは真の自由エネルギーよりも少し大きい。ハイパーパラメータの変化による相転移が存在する(渡辺一帆)。

(5) 平均場近似の汎化誤差の挙動は未解決。縮小ランク回帰モデルではランダム行列理論を用いて数学的に解ける(中島伸一)。

(6) 自由エネルギー・汎化誤差・相転移は数値実験では解析できないので専門書でも論文でも誤った記載が多い(PRMLなども間違っている)。間違っていることを知らない研究者も多いので注意しましょう。

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。

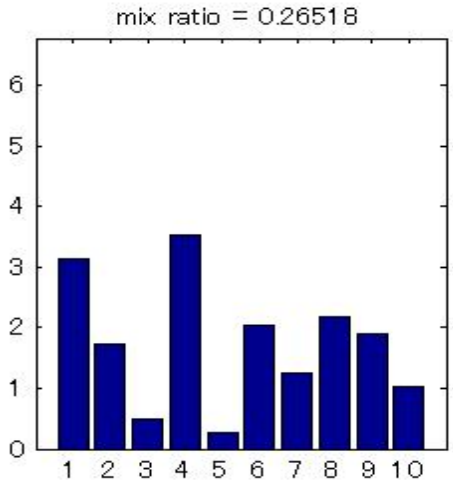
1 建設業
2 製造業
3 運輸郵便業
4 卸売小売業
5 金融保険業
6 不動産物品賃貸業
7 学術研究専門・技術サービス業
8 宿泊飲食サービス業
9 生活関連サービス娯楽業
10 教育学習支援業

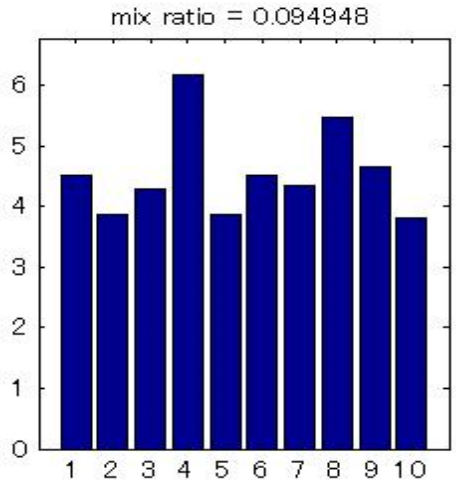
政府統計の総合窓口（e－Stat）を使わせて頂きました。
http://www.e-stat.go.jp/estat/html/spec.html
http://www.e-stat.go.jp/SG1/estat/eStatTopPortal.do

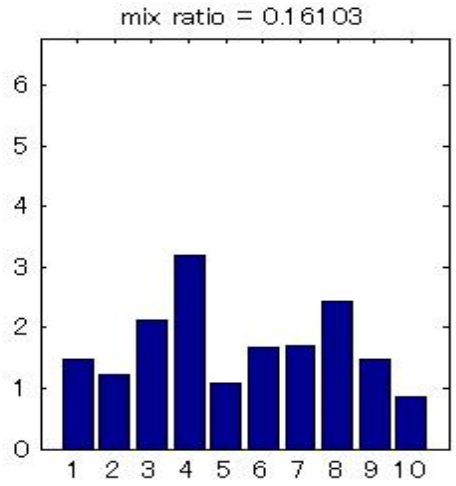
データ　X について、log(1+X)を混合数6個の正規分布で推測を行った。

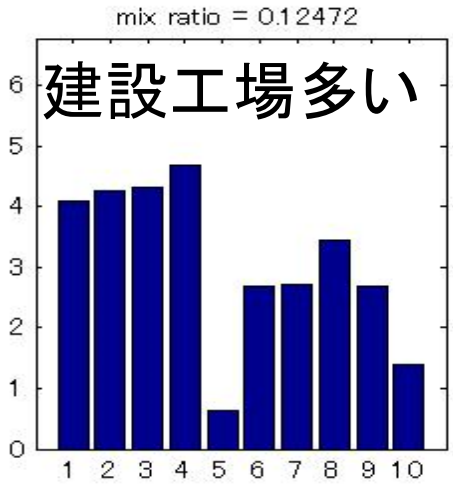
例2

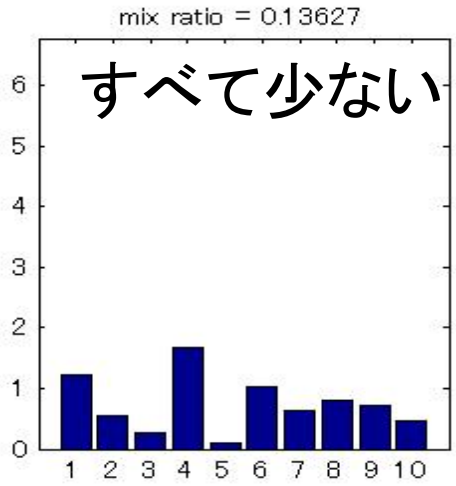
金融・運輸が少ない　　すべて多い　　特徴なし

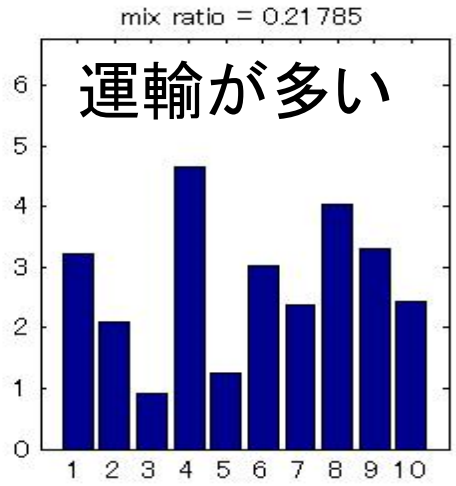






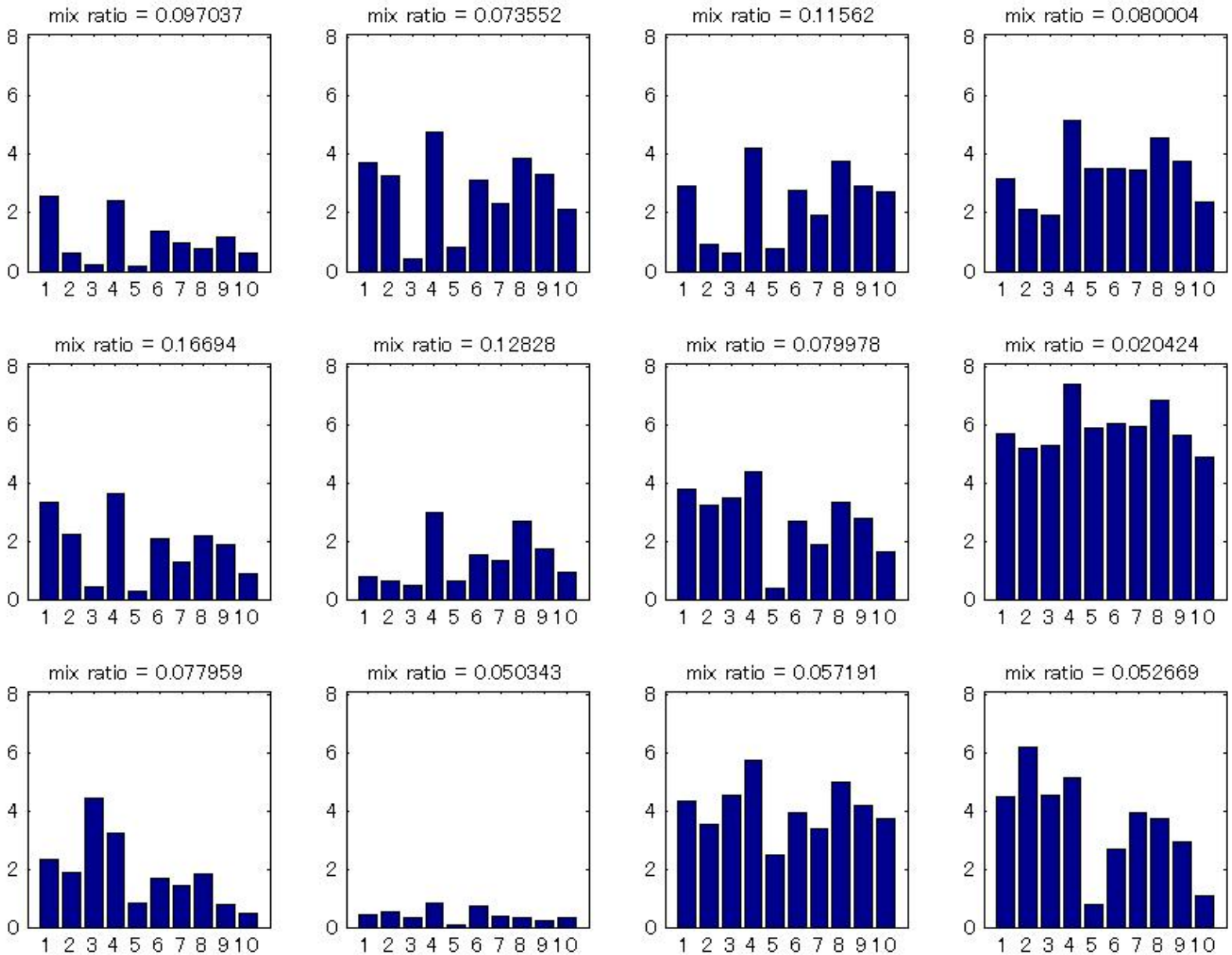






例2

コンポーネント数を12にしてみた。労働者が多いとすべての仕事が多いが少なくなるにつれて街ごとの特徴がでてくるように見える。



現実の問題では

現実のデータが、ぴったりＫ個の正規分布から出ているということはめったに起こらない。

データの個数は有限なので、その個数に応じた解像度で真の情報源がわかる。

データの個数が増えるにつれて少しずつ詳しい情報が分かっていく。適切なクラスターの個数も少しずつ増えていく。

汎化誤差を最小にするようにＫを定めるにはどうしたらよいのだろうか。

付録:　EM法の導出

なぜEM法で尤度が単調非減少かを説明する。

準備

任意の w_1 , w_2 について

$$\int p(y|\, w_1) \log p(y|w_2)\, dy \leqq \int p(y|\, w_1) \log p(y|w_1)\, dy$$

(証明） カルバック情報量の性質からすぐ。

$$\int p(y|\, w_1) \log [p(y|w_1)/\, p(y|w_2)\,]\, dx \geqq 0$$

（応用） $\int p(y|\, w_1) \log p(y|w_2)\, dy$ の w_2 に
w_1 を代入すると大きくなる。

EM法のアルゴリズム

$$G(w_1,w_2) = \sum_{i=1}^{n} \Sigma_y \, p(y| x_i, w_1) \log p(x_i,y \mid w_2)$$

(1) w_1 初期化
(2) $G(w_1,w_2)$を w_2 について最大化(w_1 固定）
(3) w_1 ：＝w_2 として(2)に戻る。

以下で、なぜこれで尤度が単調非減少かを説明する。

なぜEM法でよいのかの説明：$G^*(w_1,w_2)$が必要

定義

$$G^*(w_1,w_2) \equiv G(w_1,w_2) - \sum_{i=1}^{n}\sum_y p(y|x_i, w_1) \log p(y|x_i, w_1)$$

$$= \sum_{i=1}^{n}\sum_y p(y|x_i, w_1) \{ \log p(x_i,y \mid w_2) - \log p(y|x_i, w_1)\}$$

$$= \sum_{i=1}^{n}\sum_y p(y|x_i, w_1) \{\log p(x_i \mid w_2) + \log p(y|x_i, w_2) - \log p(y|x_i, w_1) \}$$

$$= \sum_{i=1}^{n} \log p(x_i \mid w_2) - \sum_{i=1}^{n}\sum_y p(y|x_i, w_1) \log \frac{p(y|x_i, w_1)}{p(y|x_i, w_2)}$$

関数 $G^*(w_1,w_2)$ は 対数尤度 ー KL情報量

$G^*(w_1,w_2)$

$$= \underbrace{\underbrace{\sum_{i=1}^{n} \log p(x_i \mid w_2)}_{L(w_2)} - \underbrace{\sum_{i=1}^{n} \sum_y p(y \mid x_i, w_1) \log \frac{p(y \mid x_i, w_1)}{p(y \mid x_i, w_2)}}_{w_2=w_1 \text{ のとき最大値0}}}$$

$L(w)$ がw^*で最大$\Leftrightarrow$ $G^*(w_1,w_2)$が$w_1=w_2=w^*$で最大

$L(w)$ を最大化したい$\rightarrow$ $G^*(w_1,w_2)$を最大化すればよい

EM法でなぜ尤度は大きくなるかの説明

$$G^*(w_1,w_2)= G(w_1,w_2) - \sum_{i=1}^{n}\sum_y p(y|x_i, w_1) \log p(y|x_i, w_1)$$

$$= \sum_{i=1}^{n} \log p(x_i \mid w_2) - \sum_{i=1}^{n}\sum_y p(y|x_i, w_1) \log \frac{p(y|x_i, w_1)}{p(y|x_i, w_2)}$$

(1) w_1 初期化

(2) $G(w_1,w_2)$を w_2 について最大化　$G^*(w_1,w_2)$増加

(3) $w_1 := w_2$ として(2)に戻る。　$G^*(w_1,w_2)$増加

$G^*(w_1,w_2)$は増加➔$L(w_1)$ が大きくなっていく。

ステップ(2)の最大化の計算を次ページ以下で説明。

混合正規分布への応用:
ステップ2で最大化する量

$$G(w_1,w_2)=\sum_{i=1}^{n}\sum_{y} p(y|\, x_i, w_1) \log p(x_i,y \mid w_2)$$

をw_2について最大化する。ここで

$$\log p(x_i,y|w) = -\sum_{k=1}^{K} y_k \{\|x_i-b_k\|^2/2\sigma_k^2 - N\log\sigma_k + \log a_k\}$$

これの $p(y|\, x_i, w_1)$ による平均が必要。

$p(y|\ x_i, w_1)$ による平均は計算できる。

$$\Sigma_y\ \ y_k\, p(y|\ x_i, w)$$

$$= \Sigma_y\ \ y_k\, p(x_i, y|\ w)/p(x_i|w)$$

$$= \frac{a_k\ \ \dfrac{1}{(2\pi\sigma_k{}^2)^{N/2}}\ \exp\left(-\dfrac{\|\ x_i - b_k\|^2}{2\sigma_k{}^2}\right)}{p(x_i|w_1)}$$

$$= E[y_k|\ x_i\ , w_1]$$ とかく

ステップ２で最大化する関数を書くと

$$G(w_1,w_2) = \sum_{i=1}^{n} \Sigma_y \, p(y| x_i, w_1) \log p(x_i,y \mid w_2)$$

$$= -\sum_{i=1}^{n} \sum_{k=1}^{K} E[y_k| x_i , w_1]\{\| x_i - b_k\|^2/2\sigma_k^2 - N\log\sigma_k + \log a_k\}$$

w_1 が与えられたもとで

これを$w_2=(a_k\, b_k \sigma_k)$について最大化。解ける。

結局、EM法のステップ2は

Σはすべてi=1,..,n の和を表す

$$a_k = \frac{\Sigma\, E[y_k|\, x_i\,,\, w_1]}{\Sigma\, 1}$$

$$b_k = \frac{\Sigma\, E[y_k|\, x_i\,,\, w_1]\, x_i}{\Sigma\, E[y_k|\, x_i\,,\, w_1]}$$

$$\sigma_k^{\,2} = \frac{\Sigma\, E[y_k|\, x_i\,,\, w_1]\, \| x_i - b_k \|^2}{N\ \ \Sigma\, E[y_k|\, x_i\,,\, w_1]}$$

$w_1 \rightarrow w_2$ が
計算できる

混合正規分布でのEM法

(1) w_1 初期化

(2) $w_2=(a_k\, b_k \sigma_k)$を次式で計算

$$
\begin{cases}
a_k = \dfrac{\Sigma\, E[y_k|\, x_i\,,\, w_1]}{n} \\[2ex]
b_k = \dfrac{\Sigma\, E[y_k|\, x_i\,,\, w_1]\, x_i}{\Sigma\, E[y_k|\, x_i\,,\, w_1]} \\[2ex]
\sigma_k^{\,2} = \dfrac{\Sigma\, E[y_k|\, x_i\,,\, w_1]\, \| x_i - b_k \|^2}{N \;\; \Sigma\, E[y_k|\, x_i\,,\, w_1]}
\end{cases}
$$

(3) $w_1 := w_2$ として(2)に戻る。