第9章　混合モデルとEM

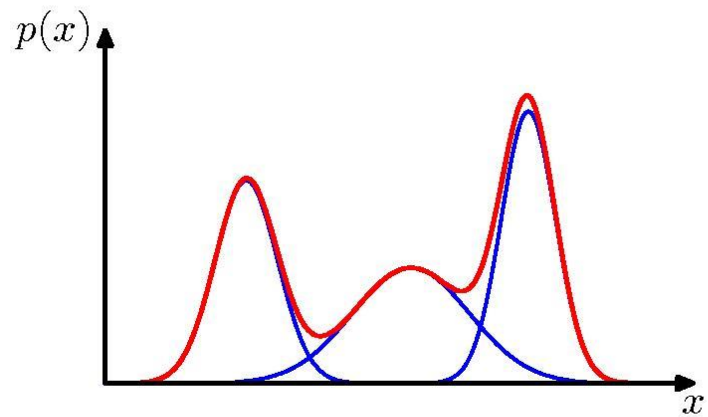
Pattern
Recognition
and
Machine
Learning

修士２年
北川直樹

BehaviorInNetworks
都市生活学・ネットワーク行動学研究室

この章で学ぶこと

- ある赤のデータ分布p(x)がある.
- これは3つの青のガウス分布$N(X|\mu_k,\Sigma_k)$が集まっている.
- では，どんな平均μ_kと分散Σ_kを持つガウス分布がどの割合π_kで集まった分布か？
- これをEMアルゴリズムで推定しよう.



$$p(x)=\sum_{k=1}^{3}\pi_k N(X|\mu_k,\Sigma_k)$$

目次

- 9.1 K-meansクラスタリング
  - 9.1.1 画像分割と画像圧縮
- 9.2 混合ガウス分布
  - 9.2.1 最尤推定
  - 9.2.2 混合ガウス分布のEMアルゴリズム
- 9.3 EMアルゴリズムのもう一つの解釈
  - 9.3.1 混合ガウス分布再訪
  - 9.3.2 K-meansとの関係
  - 9.3.3 混合ベルヌーイ分布
  - 9.3.4 ベイズ線形回帰に関するEMアルゴリズム
- 9.4 一般のEMアルゴリズム

9.1 K-meansクラスタリング

- N個のデータ集合$\{x_1,...x_n\}$をK個のクラスターに分割する.
- Kの値は既知とする.
- クラスターとは、データ点間距離が小さいグループを表す.
- μ_kをk番目クラスターの中心をする。
- 各クラスターに存在するデータからμ_kへの二乗距離の総和を最小にする.

9.1 K-meansクラスタリング

- データ点のクラスターへの割り当てを表現する.
- 各データx_nに対応する二値指示変数$r_{nk}\in\{0,1\}$ (k=1,…K)を定める.
- x_nがクラスターkに割り当てられる場合$r_{nk}=1$, $j\neq k$の場合は$r_{nj}=0$とする.
- これを一対K符号化法という.
- 目的変数Jを定義する.

$$J=\sum_{n=1}^{N}\sum_{k=1}^{K}r_{nk}\|x_n-\mu_k\|^2$$

9.1 K-meansクラスタリング

- これは，歪み尺度とも呼ばれる．
- Jを最小にするr_{nk}とμ_kを求める．
- r_{nk}とμ_kを最適化するステップを繰り返す．
- 最初にμ_kの初期値を選ぶ．
- μ_kを固定して，Jを最小化するr_{nk}を求める．
- r_{nk}を固定して，Jを最小化するμ_kを求める．
- 収束するまで繰り返す．

$$J = \sum_{n=1}^{N}\sum_{k=1}^{K} r_{nk} \left\| x_n - \mu_k \right\|^2$$

9.1 K-meansクラスタリング

- μ_kを固定した上で，r_{nk}の決定を考える．
- $r_{nk}=1$としたときに$\|x_n-\mu_k\|$が最小になるkに対して，r_{nk}を選んで1とする．
- つまり，n番目のデータ点を最も近いクラスター中心に割り当てる．

$$r_{nk}=\begin{cases}1 & if\ \ k=\arg min_j \left\|x_n-\mu_j\right\|^2\\ 0 & otherwise\end{cases}$$

9.1 K-meansクラスタリング

- r_{nk}を固定した下で，μ_kを最適化する．
- 目的関数Jはμ_kの二次関数なので偏微分=0を解くと最小化できる．

$$2\sum_{n=1}^{N} r_{nk}(x_n - \mu_k) = 0$$

- μ_kについて解くと，

$$\mu_k = \frac{\sum_n r_{nk} x_n}{\sum_n r_{nk}}$$

- k番目クラスターに割り当てられた全データの平均値である．→K-meansアルゴリズム

9.1 K-meansクラスタリング

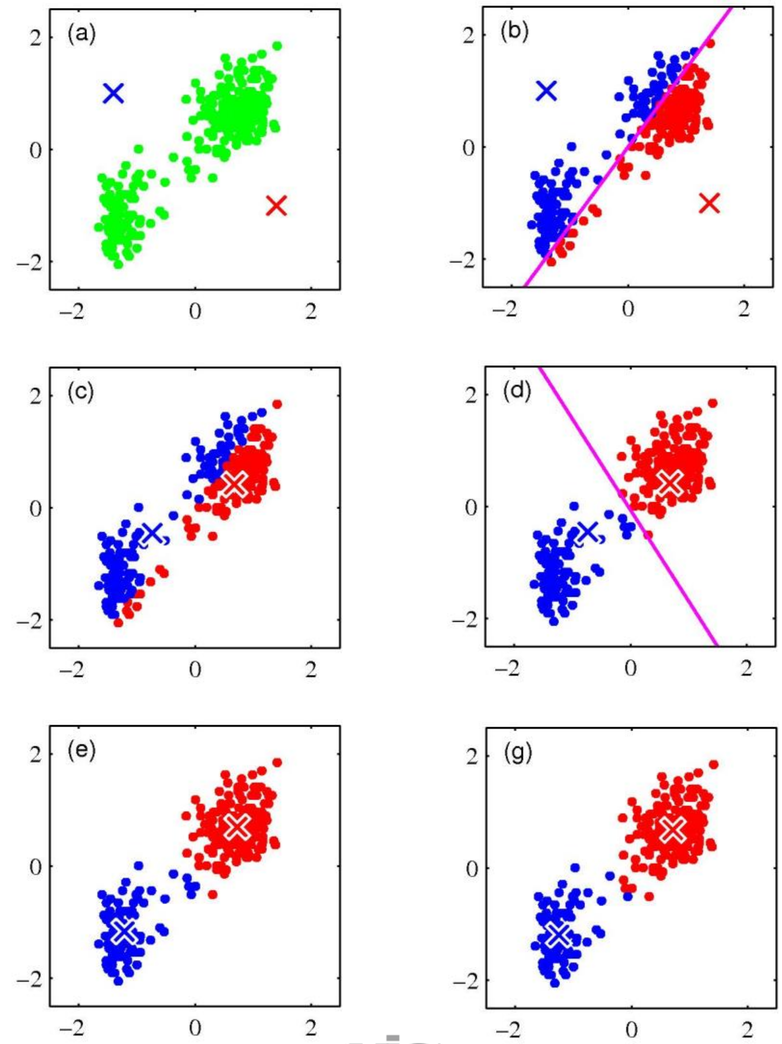
- 2クラスターに分割

(a)×印はμ_1とμ_2の初期選択を表す.

(b)各データを近いクラスターに割り当てる.

(c)割り当てられたデータの平均値をクラスターの中心とする.

(d)収束するまで繰り返す.



9.1.1 画像分割と画像圧縮

- 画像分割の目的は，一つの画像を複数の領域に分割すること．
- 画像の画素は，赤，青，緑の3つ組．
- 各画素ベクトルを割り当てられてクラスター中心{R,G,B}で置き換える．
- つまり，K色のみのパレットを用いる．

9.1.1 画像分割と画像圧縮

- クラスタリングを画像圧縮に使う.
- N個のデータ点について，各々が割り当てられるクラスターkの情報を保存する.
- クラスターkの中心μ_kの値を保存する必要があるが，K≪Nならば少ないデータ数で済む.
- つまり，各データを最も近い中心μ_kで近似する.
- この枠組みをベクトル量子化，μ_kを符号表ベクトルと呼ぶ.

9.2 混合ガウス分布

- 離散的な潜在変数を用いた混合ガウス分布を定式化する.

$$p(x)=\sum_{k=1}^{K}\pi_k N(x|\mu_k,\Sigma_k)$$

- K次元の2値確率変数zを導入する.
- 1つのz_kだけ1，他は0の1-of-K表現
- z_kは，$z_k \in \{0,1\}$かつ$\Sigma_k z_k=1$を満たす.
- Zの周辺分布は，混合係数π_kで定まる.

x \ k	1	2	3	
1	0	0	1	
2	1	0	0	
3	1	0	0	z
4	0	0	1	
5	0	1	0	
π	0.4	0.2	0.4	

$$p(z_k=1)=\pi_k$$

9.2 混合ガウス分布

- ただし，パラメータπ_kは以下を満たす．

$$0 \le \pi_k \le 1, \quad \sum_{k=1}^{K} \pi_k = 1$$

- Zは，1-of-K表現なので，

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k}$$

- Zの値が与えられた下でのxの条件付き確率は，

$$p(x|z_k = 1) = N(x|\mu_k, \Sigma_k)$$

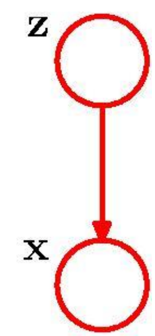
- これは、以下の形にも書ける．

$$p(x|z) = \prod_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{z_k}$$

9.2 混合ガウス分布

- Xの周辺分布は，zの取り得る状態全ての総和を取り，以下となる.

$$p(x) = \sum_{z} p(z) p(x|z) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$$



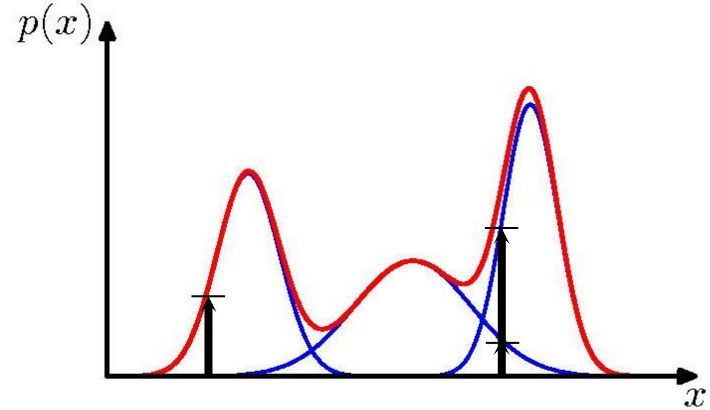
- これは，混合ガウス分布と同じ形である.
- こうして，潜在変数を含む別な混合ガウス分布の表現をした.
- これにより、EMアルゴリズムの単純化ができる.

9.2 混合ガウス分布

- Xが与えられた下でのzの条件付き確率は$\gamma(z_k)$はベイズの定理を用いて得られる.

$$\gamma(z_k) \equiv p(z_k = 1|x) = \frac{p(x|z_k = 1)p(z_k = 1)}{p(x)} = \frac{p(z_k = 1)p(x|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(x|z_j = 1)} = \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma_j)}$$

- π_kはz_k=1なる事象の事前確率, $\gamma(z_k)$はxを観測したときの事後確率
- $\gamma(z_k)$は, 混合要素kがxの観測を説明する程度を表す負荷率



9.2 混合ガウス分布

(a) 同時分布p(z)p(x|z)からのサンプル.

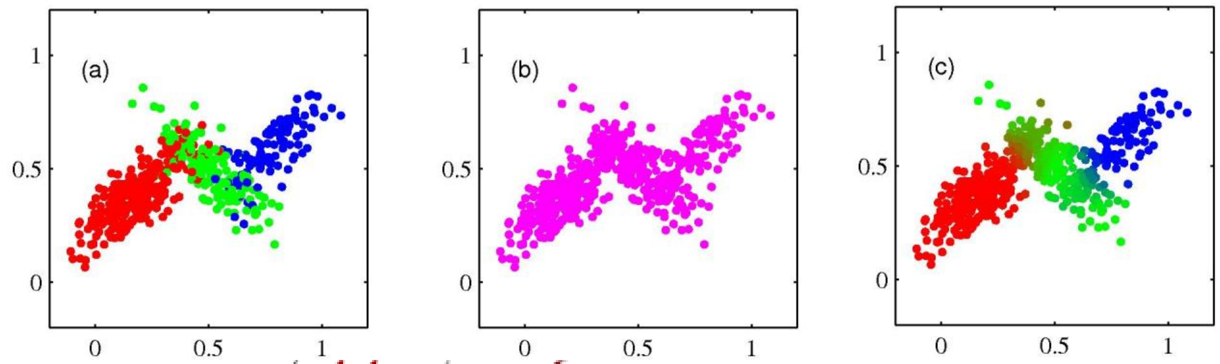
- 混合要素に対応するZの状態を赤,緑,青で描写.

(b) 同サンプルを周辺分布(x)から生成.

- Zの値を無視し，xの値のみ描写.

(c) 同サンプルの負担率$\gamma(z_{nk})$を表現

- $\gamma(z_{nk})(k=1,2,3)$に比例する量の赤,青,緑のインク



9.2.1 最尤推定

- 観測したデータ集合$\{x_1,\dots x_N\}$に混合ガウス分布を当てはめる.
- 混合ガウス分布は以下の通りである.

$$p(x)=\sum_{k=1}^{K}\pi_k N(x|\mu_k,\Sigma_k)$$

- このとき，対数尤度関数は以下のように表せる.

$$\ln p(X|\pi,\mu,\Sigma)=\sum_{n=1}^{N}\ln\left\{\sum_{k=1}^{K}\pi_k N(x_n|\mu_k,\Sigma_k)\right\}$$

$$\begin{aligned}&\mathcal{N}(x_i|\mu_k,\Sigma_k)\\&=\frac{1}{(2\pi\Sigma_k)^{1/2}}\exp\left\{-\frac{1}{2\Sigma_k}(x-\mu_k)^2\right\}\end{aligned}$$

9.2.2 混合ガウス分布のEMアルゴリズム

- 尤度関数の最大点が満たす条件
- 対数尤度lnp(X|π,μ,Σ)をガウス要素の平均μ_kに関して微分し，0とおくと，

$$0=\sum_{n=1}^{N}\underbrace{\frac{\pi_k N(x_n|\mu_k,\Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j,\Sigma_j)}}_{\gamma(z_{nk})}\Sigma_k^{-1}(x_n-\mu_k)$$

- 負担率が自然と右辺に現れる.
- 両辺にΣ_kを掛けて整理すると,

$$\mu_k=\frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})x_n,\quad N_k=\sum_{n=1}^{N}\gamma(z_{nk})$$

9.2.2 混合ガウス分布のEMアルゴリズム

- N_kは，kk番目クラスターに割り当てられたデータの実効的な数である．
- つまり，k番目のガウス要素の平均μ_kはデータ集合各点の重み付きへ平均である．
- データ点x_nの重み係数は，k番目ガウス要素がx_nを生成を負担した事後確率$\gamma(z_{nk})$である．

9.2.2 混合ガウス分布のEMアルゴリズム

- 対数尤度lnp(X|π,μ,Σ)をΣ_kに関して微分して0とおき，整理すると，

$$\Sigma_k = \frac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n - \mu_k)(x_n - \mu_k)^T$$

- 共分散も，各データは負担した事後確率$\gamma(z_{nk})$で重み付けられており，分母はk番目要素に割り当てられたデータの実効的な数である.

9.2.2 混合ガウス分布のEMアルゴリズム

- 最後に対数尤度lnp(X|π,μ,Σ)を混合係数について最大化する.
- このとき，各パラメータの総和が1であるという制約条件が必要なため，ラグランジュ未定係数法を用いる.

$$\ln p(X|\pi,\mu,\Sigma)+\lambda\left(\sum_{k=1}^{K}\pi_k-1\right)$$

- 上記の式をπ_k(k=1,…K)で微分し0とおくと,

$$0=\sum_{n=1}^{N}\frac{N(x_n|\mu_k,\Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j,\Sigma_j)}+\lambda$$

9.2.2 混合ガウス分布のEMアルゴリズム

- 両辺にπ_kを掛けてkについて和を取り，$\sum_{k=1}^{K}\pi_k = 1$を用いると，$\lambda=-N$が得られる.
- これを用いてλを消去し，変形すると，

$$\pi_k = \frac{N_k}{N}$$

- つまり，k番目要素の混合係数は，全データ数に対する，k番目要素に含まれるデータの負担率の総和である.

9.2.2 混合ガウス分布のEMアルゴリズム

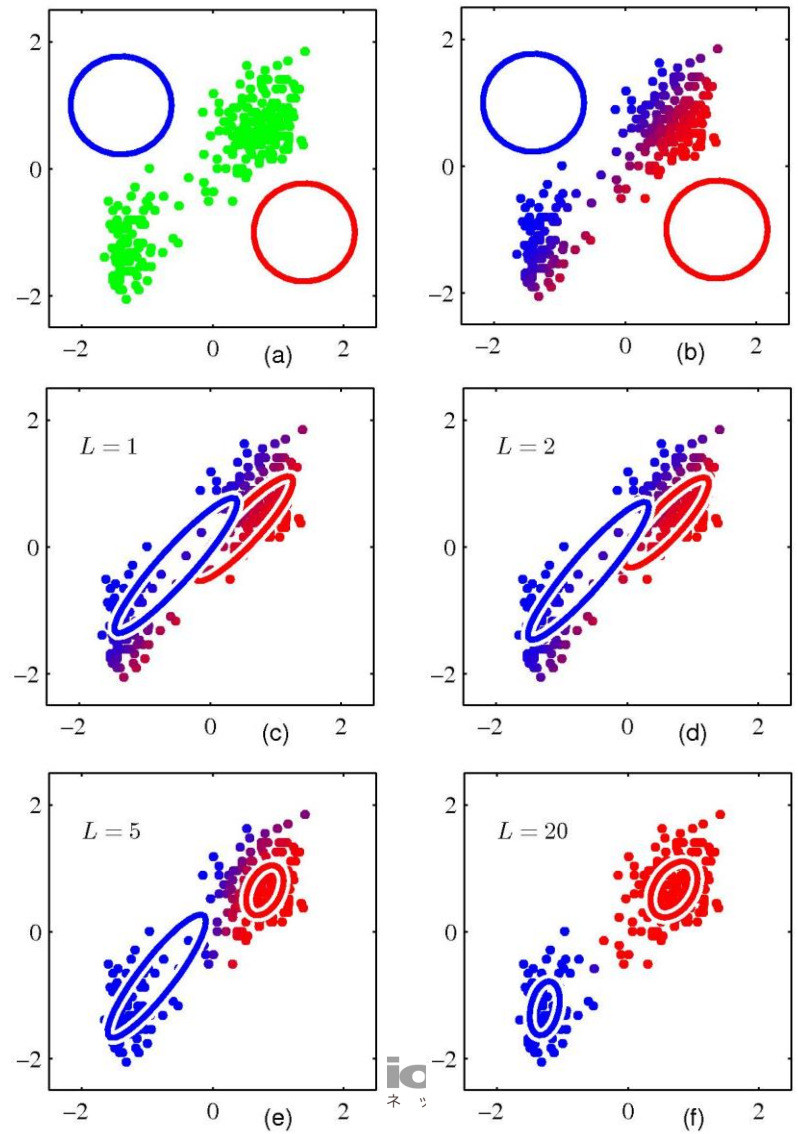
- μ_k, Σ_k, π_kをEMアルゴリズムを用いた最尤推定法で解を見付ける.
- 最初に，平均，分散，混合係数の初期値を選ぶ.
- Eステップ(expectation)では，初期パラメータを用いて負担率$\gamma(z_{nk})$を計算する.
- Mステップ(maximization)では，負担率に基づき平均，分散，混合係数のパラメータを再計算する.
- 対数尤度，またはパラメータの変化量が閾値より小さくなったとき，収束したとする.

9.2.2 混合ガウス分布のEMアルゴリズム

(a) 緑はデータ点の中心．青と赤の円は，ガウス分布の標準偏差の等高線．

(b) 青と赤の両クラスターの負担率に比例したインクで描写．

(c) 青のガウス分布の平均は，各データ点が持つ青インクの重み付き平均(重心)．共分散は，インクの共分散である．



Pattern Recognition and Machine Learning

9.2.2 混合ガウス分布のEMアルゴリズム

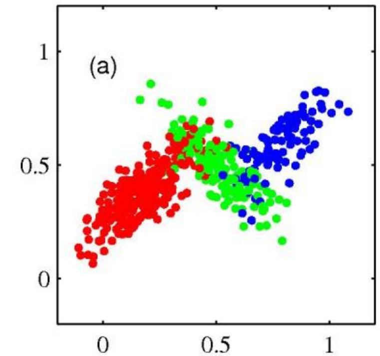
- EMアルゴリズムは，K-meansより収束するまでの繰り返し回数と計算量が多い.
- そのため，混合ガウスモデルの初期値を発見するために，K-meansを実行した後，EMアルゴリズムを行う.
- 共分散は各クラスターのサンプル分散，混合係数は各クラスターに属する点の割合.
- ただし，一般に対数尤度は多数の極大値を持ち，EM解がその中で最大とは限らない.

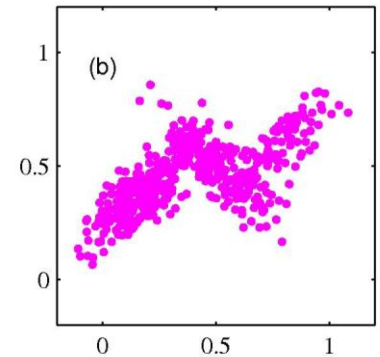
9.3 EMアルゴリズムのもう一つの解釈

- 潜在変数を持つモデルの最尤解を見付けることがEMアルゴリズムの目的.
- データ集合をX，潜在変数の集合をZ，パラメータをθとする，

$$\ln p(X|\theta) = \ln\left\{\sum_{Z} p(X,Z|\theta)\right\}$$

- 完全データ集合{X，Z}が与えられれば対数尤度関数の最大化ができる.
- しかし実際は，不完全データXのみ.





9.3 EMアルゴリズムのもう一つの解釈

- 完全データ尤度関数が使えないため，潜在変数の事後確率に関する期待値を考える．
- Eステップでは，現在のパラメータθ_{old}を用いて潜在変数の事後分布$p(Z|X,\theta_{old})$を計算する．
- これを完全データ対数尤度$\ln p(X,Z|\theta)$の期待値$Q(\theta,\theta_{old})$を計算するのに用いる．

$$Q(\theta,\theta^{old}) = \sum_{z} p(Z|X,\theta^{old}) \ln p(X,Z|\theta)$$

- Mステップでは，この関数をθについて最大化し新しいθ_{new}を決定する．

$$\theta^{new} = \arg\max_{\theta} Q(\theta,\theta^{old})$$

9.3.2 K-meansとの関係

- K-meansとEMは，強い類似性がある.
- K-meansはデータ点を1つのクラスターに割り当てるが，EMは事後確率に基づいて割り当てる.
- 混合ガウス分布に関するEMの極限としてK-meansを導出できる.
- 各ガウス要素の共分散がεの混合ガウス分布を考える.

$$p(x|\mu_k,\Sigma_k) = \frac{1}{(2\pi\varepsilon)^{D/2}} \exp\left\{-\frac{1}{2\varepsilon}\|x-\mu_k\|^2\right\}$$

- この形のK個混合ガウス分布のEMを考える.
- ただし，εは推定しない固定定数とする.

9.3.2 K-meansとの関係

- データ点x_nに関するk番目混合要素の負担率は,

$$\gamma(z_{nk}) = \frac{\pi_k \exp\{-\|x_n - \mu_k\|^2 / 2\varepsilon\}}{\sum_j \exp\{-\|x_n - \mu_j\|^2 / 2\varepsilon\}}$$

- $\varepsilon \to 0$の極限を考えると，データ点x_nに関する負担率$\gamma(z_{nk})$は，$\|x_n - \mu_j\|$が最小となるj番目の要素が1に，その他は0に収束する.
- これにより，K-meansと同様に$\gamma(z_{nk}) \to r_{nk}$という{1,0}の割り当てが実現する.
- K-meansではクラスターの平均のみ推定し，分散は推定しないが，楕円K-meansアルゴリズムは{1,0}割り当てで分散も推定する.