ML 輪講: 19 章 近似推論

杉浦 圭祐

慶應義塾大学理工学部情報工学科 松谷研究室

April 2, 2019

目次

目次

K-Means 法によるクラスタリング

- 扱う問題
  - 多次元空間上のデータ点集合を考える
  - 各データが属するクラスタを決定する問題を考える

- 問題設定
  - D 次元ユークリッド空間における、確率変数 $\boldsymbol{x}$ を観測
  - $\boldsymbol{x}$ の N 個の観測点で構成されるデータ集合 $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ $(\boldsymbol{x}_i \in \mathbb{R}^D)$
  - データ集合 $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ を、K 個のクラスタに分割
  - K は、既知の定数であるとする

- クラスタとは
  - クラスタとは、簡単に言えば近接するデータの集合である
  - クラスタの内部のデータ点間の距離が、クラスタの外側のデータとの距離と比べて、小さいようなデータの集合

K-Means 法によるクラスタリング

- クラスタを代表するベクトルの表現
  - 各クラスタを代表する K 個の D 次元ベクトル $\mathcal{M} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K\}$ $(\boldsymbol{\mu}_k \in \mathbb{R}^D)$ を導入する
  - これらのベクトル $\mathcal{M} = \{\boldsymbol{\mu}_k\}$ を、プロトタイプという
  - k 番目のクラスタ ($\boldsymbol{\mu}_k$ が支配するクラスタ) を $\mathcal{M}(\boldsymbol{\mu}_k)$ と記す

  - ベクトル $\boldsymbol{\mu}_k$ は、k 番目のクラスタに対応するプロトタイプである
  - $\boldsymbol{\mu}_k$ は、$\mathcal{M}(\boldsymbol{\mu}_k)$ に属するデータ点の平均、即ち k 番目のクラスタの中心である

- 解くべき問題
  - N 個の全データ点を、うまくクラスタに割り振る
  - 各データ点から、対応する (そのデータ点が属するクラスタの) プロトタイプ $\boldsymbol{\mu}_k$ への、二乗距離の総和を最小化する

K-Means 法によるクラスタリング

- データ点のクラスタへの割り当てを表す変数
  - 各データ $\boldsymbol{x}_i$ $(1 \leq i \leq N)$ に対して、二値の指示変数 $r_{ik} \in \{0, 1\}$ $(k = 1, \ldots, K)$ を定める
  - r_{ik} は、データ $\boldsymbol{x}_i$ が、K 個あるクラスタのうちの、どれに割り当てられるのかを表す
  - データ点 $\boldsymbol{x}_i$ がクラスタ k に割り当てられるときに、$r_{ik} = 1$ となり、$j \neq k$ について $r_{ij} = 0$ である (1-of-K 符号化法という)

$$r_{ik} = \begin{cases} 1 & (\boldsymbol{x}_i \in \mathcal{M}(\boldsymbol{\mu}_k) \text{ の場合}) \\ 0 & \text{それ以外} \end{cases} \tag{1}$$

K-Means 法によるクラスタリング

- 目的関数の定義
  - 目的関数を以下のように定義する
  - 各データ点 $\boldsymbol{x}_i$ と、$\boldsymbol{x}_i$ に割り当てたベクトル $\boldsymbol{\mu}_k$($\boldsymbol{x}_i$ が属するクラスタのプロトタイプ) との、二乗距離の総和

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2 \tag{2}$$

  - 上式の J を最小化するような、$\{r_{ik}\}$ と $\{\boldsymbol{\mu}_k\}$ を求めるのが目標

K-Means 法によるクラスタリング

- 目的関数 J の $\{r_{ik}\}$ に関する最小化
  - 目的関数の式は、各 i について独立である

$$J = \sum_i \sum_k r_{ik} ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2 \tag{3}$$

  - $\{r_{ik}\}$ を決定する方法は簡単
  - 各 i について、$||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2$ が最小となるような k に対し、$r_{ik} = 1$ とする
  - それ以外のクラスタ $j \neq k$ については、$r_{ik} = 0$ とする

$$r_{ik} = \begin{cases} 1 & (k = \arg\min_j ||\boldsymbol{x}_i - \boldsymbol{\mu}_j||^2 \text{のとき}) \\ 0 & \text{それ以外} \end{cases} \tag{4}$$

  - 各データ点 $\boldsymbol{x}_i$ を、$\boldsymbol{x}_i$ と最も近い $\boldsymbol{\mu}_k$ に割り当てることに相当
  - 各データ点 $\boldsymbol{x}_i$ を、クラスタを代表するベクトル (クラスタの中心) との二乗距離が最小になるような、クラスタに割り当てる

K-Means 法によるクラスタリング

- 目的関数 J の $\{\boldsymbol{\mu}_k\}$ に関する最小化
  - J を $\boldsymbol{\mu}_k$ について偏微分して 0 とおく

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}_k} J &= \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_i r_{ik} ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2 \\
&= \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_i r_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \\
&= \sum_i r_{ik} \frac{\partial}{\partial \boldsymbol{\mu}_k} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \\
&= \sum_i r_{ik} \frac{\partial}{\partial \boldsymbol{\mu}_k} (\boldsymbol{x}_i^T \boldsymbol{x}_i - \boldsymbol{x}_i^T \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T \boldsymbol{x}_i + \boldsymbol{\mu}_k^T \boldsymbol{\mu}_k) \\
&= \sum_i r_{ik} \frac{\partial}{\partial \boldsymbol{\mu}_k} (\boldsymbol{x}_i^T \boldsymbol{x}_i - \boldsymbol{x}_i^T \boldsymbol{\mu}_k - \boldsymbol{x}_i^T \boldsymbol{\mu}_k + \boldsymbol{\mu}_k^T \boldsymbol{\mu}_k) \\
&= \sum_i r_{ik} (2\boldsymbol{\mu}_k - 2\boldsymbol{x}_i) = 0 \qquad (5)
\end{aligned}
$$

K-Means 法によるクラスタリング

- ここで、次の関係を用いている

$$
\begin{aligned}
& \boldsymbol{\mu}_k^T \boldsymbol{x}_i \\
= \quad & \left(\boldsymbol{\mu}_k^T \boldsymbol{x}_i\right)^T \qquad (\because \text{スカラーであるため転置してもよい}) \\
= \quad & \boldsymbol{x}_i^T \left(\boldsymbol{\mu}_k^T\right)^T \qquad (\because (\boldsymbol{ab})^T = \boldsymbol{b}^T \boldsymbol{a}^T) \\
= \quad & \boldsymbol{x}_i^T \boldsymbol{\mu}_k
\end{aligned}
$$

$$
\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{a} = \frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{a}^T \boldsymbol{x} = \boldsymbol{a} \tag{6}
$$

$$
\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{B} \boldsymbol{x} = \left(\boldsymbol{B} + \boldsymbol{B}^T\right) \boldsymbol{x} \quad (\text{2 次形式}) \tag{7}
$$

$$
\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{x} = 2\boldsymbol{x} \tag{8}
$$

$$
\because \frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{x} = \frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^T \boldsymbol{I} \boldsymbol{x} = \left(\boldsymbol{I} + \boldsymbol{I}^T\right) \boldsymbol{x} = 2\boldsymbol{I}\boldsymbol{x} = 2\boldsymbol{x}
$$

K-Means 法によるクラスタリング

- これより、$\boldsymbol{\mu}_k$ について解くと次のようになる

$$
\begin{aligned}
\sum_i 2r_{ik}(\boldsymbol{\mu}_k - \boldsymbol{x}_i) &= 0 \\
\sum_i r_{ik}(\boldsymbol{\mu}_k - \boldsymbol{x}_i) &= 0 \\
\sum_i r_{ik}\boldsymbol{\mu}_k &= \sum_i r_{ik}\boldsymbol{x}_i \\
\boldsymbol{\mu}_k \left(\sum_i r_{ik} \right) &= \sum_i r_{ik}\boldsymbol{x}_i \\
\boldsymbol{\mu}_k &= \frac{1}{\sum_i r_{ik}} \sum_i r_{ik}\boldsymbol{x}_i
\end{aligned} \tag{9}
$$

- $\sum_i r_{ik}$ は、クラスタ k に属するデータの個数である
- $\sum_i r_{ik} = N_k$ と表すことがある

K-Means 法によるクラスタリング

- r_{ik} は、i 番目のデータ $\boldsymbol{x}_i$ が、クラスタ k に割り当てられているときにのみ、1 となる
- $\sum_i r_{ik}\boldsymbol{x}_i$ は、クラスタ k に属しているデータのベクトル $\boldsymbol{x}_i$ の合計である

- 従って、$\boldsymbol{\mu}_k$ は、k 番目のクラスタに割り当てられた、全てのデータ点 $\boldsymbol{x}_i$ の平均値である

- この意味で、$\boldsymbol{\mu}_k$ のことを、クラスタ k の平均ベクトル、重心、セントロイドということもある

K-Means 法によるクラスタリング

- $\{\boldsymbol{\mu}_k\}$ と $\{r_{ik}\}$ についての最適化
  - 目的関数 J を、$\{\boldsymbol{\mu}_k\}$ と $\{r_{ik}\}$ について最小化する式は次のようになる

$$\boldsymbol{\mu}_k = \frac{1}{\sum_i r_{ik}} \sum_i r_{ik} \boldsymbol{x}_i$$

$$r_{ik} = \begin{cases} 1 & (k = \arg\min_j ||\boldsymbol{x}_i - \boldsymbol{\mu}_j||^2 \text{のとき}) \\ 0 & \text{それ以外} \end{cases}$$

  - $\boldsymbol{\mu}_k$ を計算する式の中に r_{ik} が、r_{ik} を計算する式の中に $\boldsymbol{\mu}_k$ が入っている
  - $\boldsymbol{\mu}_k$ を求めるためには r_{ik} が、r_{ik} を求めるためには $\boldsymbol{\mu}_k$ が既知でなければならない
  - 従って、$\boldsymbol{\mu}_k$ と r_{ik} の両方を同時に最適化することはできない

  - どうすれば目的関数 J を、パラメータ $\{\boldsymbol{\mu}_k\}$ と $\{r_{ik}\}$ の両方について最小化できるか?

K-Means 法によるクラスタリング

- $\{\boldsymbol{\mu}_k\}$ と $\{r_{ik}\}$ についての最適化
  - $\boldsymbol{\mu}_k$ と r_{ik} の両方を同時に最適化することはできない
  - $\boldsymbol{\mu}_k$ と r_{ik} を交互に最適化すればよい
  - $\boldsymbol{\mu}_k$ と r_{ik} のそれぞれを最適化する、2 つのステップを交互に繰り返す
  - $\boldsymbol{\mu}_k$ と r_{ik} の最適化は、次のように行うことができる

1. $\mathcal{M} = \{\boldsymbol{\mu}_k\}$ の初期値を設定
   - N 個のデータ $\mathcal{D} = \{\boldsymbol{x}_i\}$ を、ランダムなクラスタに割り振って、各クラスタの平均ベクトル $\{\boldsymbol{\mu}_k\}$ を求める
   - データ $\mathcal{D}$ からランダムに選択した K 個のデータ点を、各クラスタの中心 $\boldsymbol{\mu}_k$ $(k = 1, \ldots, K)$ とすることもできる

K-Means 法によるクラスタリング

2. 第 1 フェーズでは、$\boldsymbol{\mu}_k$ を固定しつつ、r_{ik} について J を最小化

$$r_{ik} = \begin{cases} 1 & (k = \arg\min_j ||\boldsymbol{x}_i - \boldsymbol{\mu}_j||^2 \text{のとき}) \\ 0 & \text{それ以外} \end{cases}$$

3. 第 2 フェーズでは、r_{ik} を固定しつつ、$\boldsymbol{\mu}_k$ について J を最小化

$$\boldsymbol{\mu}_k = \frac{1}{\sum_i r_{ik}} \sum_i r_{ik} \boldsymbol{x}_i$$

4. (2) と (3) を、$\boldsymbol{\mu}_k$ と r_{ik} が収束するまで繰り返す

- 上記の 2 つのステップが、後述する EM アルゴリズムの E(Expectation) ステップと M(Maximization) ステップに対応する

K-Means 法によるクラスタリング

- データ点へのクラスタの再割り当てと、クラスタの平均ベクトルの再計算

- この 2 段階の処理を、クラスタの再割り当てが起こらなくなるまで (2 段階の処理を行っても、データが属するクラスタが変化しなくなるまで) 繰り返す

- 各フェーズで J の値が減少するので、アルゴリズムの収束が保証される
- 大域的最適解ではなく、局所最適解に収束する可能性はある

K-Means 法によるクラスタリング

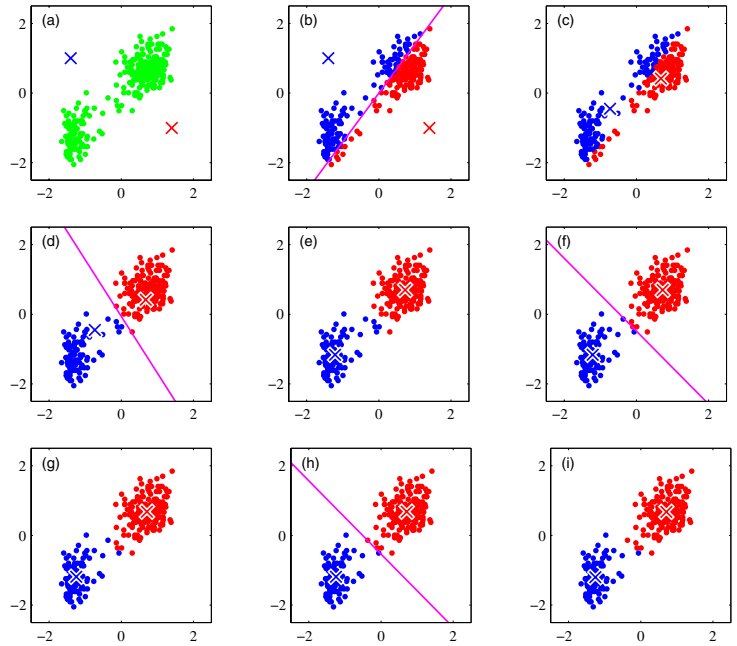

Figure 9.1 Illustration of the K-means algorithm using the re-scaled Old Faithful data set. (a) Green points denote the data set in a two-dimensional Euclidean space. The initial choices for centres μ_1 and μ_2 are shown by the red and blue crosses, respectively. (b) In the initial E step, each data point is assigned either to the red cluster or to the blue cluster, according to which cluster centre is nearer. This is equivalent to classifying the points according to which side of the perpendicular bisector of the two cluster centres, shown by the magenta line, they lie on. (c) In the subsequent M step, each cluster centre is re-computed to be the mean of the points assigned to the corresponding cluster. (d)–(i) show successive E and M steps through to final convergence of the algorithm.

図 1: K-Means アルゴリズムの動作

K-Means 法によるクラスタリング

Figure 9.2 Plot of the cost function J given by (9.1) after each E step (blue points) and M step (red points) of the K-means algorithm for the example shown in Figure 9.1. The algorithm has converged after the third M step, and the final EM cycle produces no changes in either the assignments or the prototype vectors.

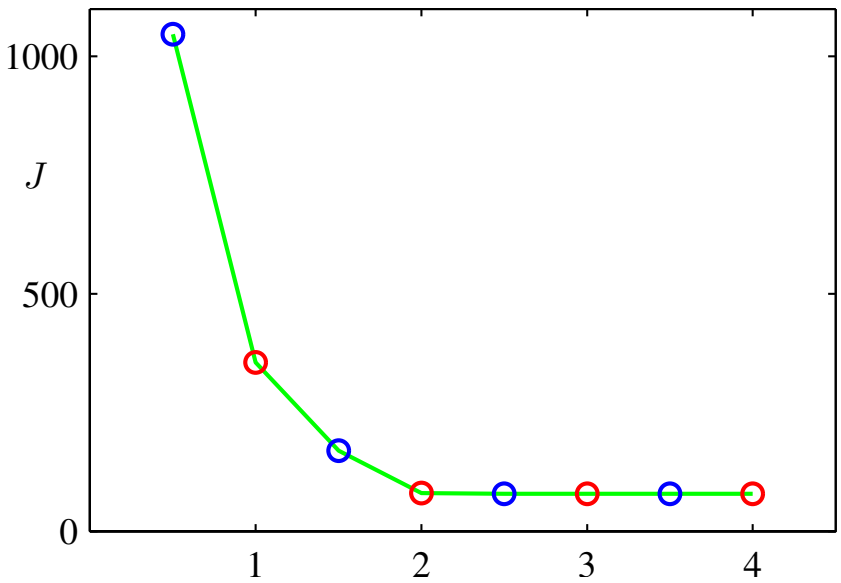

図 2: 目的関数 J の値の遷移

K-Means 法によるクラスタリング

- 素朴な実装では速度が遅いことがある
  - $\{r_{ik}\}$ の更新 (E ステップ) において、各データ点と、各平均ベクトルの全ての組み合わせ間の距離を計算する必要がある

$$r_{ik} = \begin{cases} 1 & (k = \arg\min_j ||\boldsymbol{x}_i - \boldsymbol{\mu}_j||^2 \text{のとき}) \\ 0 & \text{それ以外} \end{cases}$$

  - K-Means の高速化について、これまでに様々な手法が提案されてきた
  - 近接するデータ点同士が、同一の部分木に属するような木構造を採用する方法
  - 距離の三角不等式を利用して、不必要な距離計算を避ける方法

逐次版の K-Means 法

- 逐次版の K-Means 法の導出
  - これまでに紹介した K-Means 法は、利用する全てのデータ $\mathcal{D} = \{\boldsymbol{x}_i\}$ が、最初から用意されていることが前提であった
  - ここでは、Robbins-Monro 法を使って、オンラインのアルゴリズムを導出する

Robbins-Monro 法

- Robbins-Monro 法
  - 逐次学習アルゴリズムを導出するための手法
  - 関数 $f(\theta^*) = 0$ を満たす根 θ^* を、逐次的に計算するための式を与える

  - 同時分布 $p(z, \theta)$ に従う確率変数のペア θ, z について、関数 $f(\theta)$ は、θ が与えられたときの z の条件付き期待値 $\mathbb{E}[z|\theta]$ として、定義される

$$f(\theta) \equiv \mathbb{E}[z|\theta] = \int zp(z|\theta)dz \tag{10}$$

  - このとき、根 θ^* の逐次計算式は、次のように記述される

$$\theta^{(N)} = \theta^{(N-1)} - \eta_{N-1} z(\theta^{(N-1)}) \tag{11}$$

  - η は学習率
  - $z(\theta^{(N)})$ は、確率変数 θ が値 $\theta^{(N)}$ をとるときに観測される z の値

Robbins-Monro 法

- Robbins-Monro 法を適用するための条件
  - Robbins-Monro 法を使用するためには、満たすべき条件が幾つか存在する

1. z の条件付き分散 $\mathbb{E}\left[(z-f)^2|\theta\right]$ が、有限でなければならない

$$\mathbb{E}\left[(z-f)^2|\theta\right] = \int (z-f(\theta))^2 p(z|\theta)dz < \infty \tag{12}$$

2. $\theta > \theta^*$ では $f(\theta) > 0$、$\theta < \theta^*$ では $f(\theta) < 0$ を仮定 (このように仮定しても一般性は失われない)

Robbins-Monro 法

3 学習率の系列 $\{\eta_N\}$ は次の条件を満たす

$$\lim_{N\to\infty} \eta_N = 0 \tag{13}$$

$$\sum_{N=1}^{\infty} \eta_N = \infty \tag{14}$$

$$\sum_{N=1}^{\infty} \eta_N^2 < \infty \tag{15}$$

- 最初の条件は、推定系列 $\theta^{(N)}$ が、目標の根 θ^* に収束していくように、θ の修正量を減らしていくことを保証
- 次の条件は、根 θ^* 以外の値に収束しないことを保証
- 最後の条件は、分散を有限に抑えることで、いつまで経っても収束しないことを防止

Robbins-Monro 法

Figure 2.10 A schematic illustration of two correlated random variables z and θ, together with the regression function $f(\theta)$ given by the conditional expectation $\mathbb{E}[z|\theta]$. The Robbins-Monro algorithm provides a general sequential procedure for finding the root $\theta^\star$ of such functions.

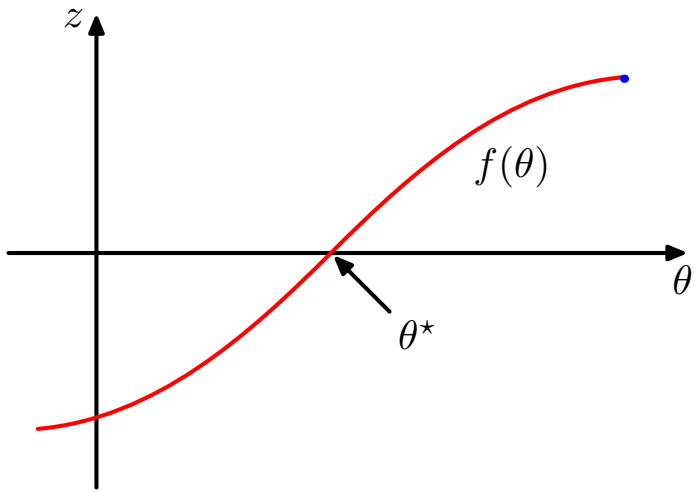

図 3: $f(\theta)$ のグラフ

逐次版の K-Means 法

- 逐次版の K-Means 法の導出
  - 先程のパラメータ θ は、$\{r_{ik}\}$ と $\{\boldsymbol{\mu}_k\}$ である
  - パラメータの最適解 $\{\boldsymbol{\mu}_k^*\}$ は、以下の式を満たす

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} J \Big|_{\boldsymbol{\mu}_k^*} = \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_{i=1}^{N} r_{ik} ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2 \Big|_{\boldsymbol{\mu}_k^*} = 0 \tag{16}$$

これは以下の式と等価である $(N_k = \sum_i r_{ik})$

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \frac{1}{N_k} \sum_{i=1}^{N} r_{ik} ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2 \Big|_{\boldsymbol{\mu}_k^*} = 0 \tag{17}$$

これ以降、クラスタ k に属するデータのみを考えることにして、これを $\boldsymbol{x}_j$ と書く $(j = 1, \ldots, N_k)$

逐次版の K-Means 法

- このとき、上式から r_{ik} を省略でき、次のようになる

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \frac{1}{N_k} \sum_{j=1}^{N_k} ||\boldsymbol{x}_j - \boldsymbol{\mu}_k||^2 \Bigg|_{\boldsymbol{\mu}_k^*} = 0 \tag{18}$$

- $N_k \to \infty$ の極限を取って次のように変形する

$$\begin{aligned}
& \lim_{N_k \to \infty} \frac{\partial}{\partial \boldsymbol{\mu}_k} \frac{1}{N_k} \sum_j ||\boldsymbol{x}_j - \boldsymbol{\mu}_k||^2 \\
= & \lim_{N_k \to \infty} \frac{1}{N_k} \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_j ||\boldsymbol{x}_j - \boldsymbol{\mu}_k||^2 \\
\simeq & \mathbb{E}\left[\frac{\partial}{\partial \boldsymbol{\mu}_k} ||\boldsymbol{x} - \boldsymbol{\mu}_k||^2 \middle| \boldsymbol{\mu}_k \right] \quad (\boldsymbol{x} \text{ はクラスタ } k \text{ に属する}) \\
= & \mathbb{E}\left[2(\boldsymbol{x} - \boldsymbol{\mu}_k) | \boldsymbol{\mu}_k \right] = f(\boldsymbol{\mu}_k)
\end{aligned} \tag{19}$$

逐次版の K-Means 法

- これより、K-Means 法は、関数 $f(\boldsymbol{\mu}_k^*) = 0$ をみたす根 $\boldsymbol{\mu}_k^*$ を求める問題に帰着させられる
- 従って、Robbins-Monro 法を適用すると、$\boldsymbol{\mu}_k$ の逐次更新式は、以下のようになる

$$\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{old}} - \eta_j(\boldsymbol{x}_j - \boldsymbol{\mu}_k^{\text{old}}) \quad (\boldsymbol{x}_j \text{はクラスタ } k \text{ に属する}) \tag{20}$$

- η_j は学習率パラメータであり、一般に、j の増加に伴って単調減少するように設定される
- クラスタ k に属するデータ $\boldsymbol{x}_j$ が 1 つずつ到着するときの、$\boldsymbol{\mu}_k$ のオンライン更新式が得られた

K-Means アルゴリズムの特徴

- K-Means アルゴリズムの特徴
  - 各データを、たった 1 つのクラスタにのみ割り当てる (ハード割り当て)
  - あるクラスタの中心ベクトル $\boldsymbol{\mu}_k$ に非常に近いデータ点もあれば、複数のクラスタの中間領域にあるようなデータ点も存在する
  - データ $\boldsymbol{x}_i$ が、クラスタ k に属するという結果を得たときに、そのクラスタに属することがほぼ確実なのか、それとも他のクラスタに割り振っても大差はないのかが、区別できない
  - 後者のようなデータ点の場合、単一のクラスタへのハード割り当ては最適でない (不正確) かもしれない
  - データ点を単一のクラスタに割り当てるのではなく、各クラスタに属する確率を、計算できれば良さそう
  - 各クラスタへの割り当ての不明瞭さを反映できる
  - このように曖昧さを含んだ割り当てを、ソフト割り当てという

目次

混合ガウス分布

- 混合ガウス分布を導入する理由
  - 曖昧さを含んだクラスタリング (ソフト割り当て) を実現するため
  - 言い換えると、データに対して、各クラスタに属する確率が分かるようにするため

- 混合ガウス分布とは
  - 各ガウス分布の線形の重ね合わせ

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{21}$$

  - 各ガウス分布 $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ は混合要素とよばれる
  - 各ガウス分布は個別に、平均 $\boldsymbol{\mu}_k$ と共分散 $\boldsymbol{\Sigma}_k$ のパラメータをもつ

混合ガウス分布

- パラメータ π_k を混合係数といい、以下の条件を満たす

$$\sum_k \pi_k = 1 \tag{22}$$

これは、$p(\boldsymbol{x})$ を $\boldsymbol{x}$ について積分すれば明らかである

$$\begin{aligned}
\int p(\boldsymbol{x}) d\boldsymbol{x} &= 1 \\
\int \sum_k \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) d\boldsymbol{x} &= 1 \\
\sum_k \pi_k \int \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) d\boldsymbol{x} &= 1 \\
\sum_k \pi_k &= 1
\end{aligned}$$

混合ガウス分布

各ガウス分布 $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ は、正規化されている

$$\forall k \quad \int \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) d\boldsymbol{x} = 1$$

- $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \geq 0$ であるので、$p(\boldsymbol{x}) \geq 0$ となるための十分条件は、全ての k について、$\pi_k \geq 0$ が成立することである

$$\forall k \in \{1, \ldots, K\} \quad \pi_k \geq 0 \Rightarrow p(\boldsymbol{x}) \geq 0$$

- これと $\sum_k \pi_k = 1$ から、結局全ての π_k について以下が成り立つ

$$0 \leq \pi_k \leq 1 \tag{23}$$

混合ガウス分布

$$p(\boldsymbol{x}) = \sum_k \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$\sum_k \pi_k = 1, \quad \forall k \quad 0 \leq \pi_k \leq 1$$

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k)\right\}$$

- 混合ガウス分布を決定づけるパラメータは、$\boldsymbol{\pi} \equiv \{\pi_1, \pi_2, \ldots, \pi_K\}$、$\boldsymbol{\mu} \equiv \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K\}$、$\boldsymbol{\Sigma} \equiv \{\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_K\}$

混合ガウス分布

- 混合ガウス分布を導入する理由 (再確認)
  - データに対して、各クラスタに属する確率が分かるようにするため

- 問題設定
  - $\boldsymbol{x}$ の N 個の観測点で構成されるデータ集合 $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ $(\boldsymbol{x}_i \in \mathbb{R}^D)$
  - データ集合 $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ を、K 個のクラスタに分割
  - K は、既知の定数であるとする

  - k 番目のクラスタが、平均 $\boldsymbol{\mu}_k$、共分散行列 $\boldsymbol{\Sigma}_k$ の正規分布 $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ で表現できるとする

  - 各クラスタの分布 $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ を、π_k で重み付けして足し合わせた混合分布が、データ全体を表す分布である

混合ガウス分布

- 最尤推定を試みる
  - 最尤推定によって、混合ガウス分布のパラメータ π, μ, Σ が分かったとする
  - このとき次のようにすれば、クラスタリングが可能

  - 新たなデータ x が得られたとき、全ての $k(k = 1, \ldots, K)$ について $\mathcal{N}(x|\mu_k, \Sigma_k)$ を計算する
  - これを最大にするような k が、データ x が属するクラスタである

混合ガウス分布

- 最尤推定を試みる
  - 結論から先に言うと、いきなり最尤推定を試すと失敗する
  - 最尤推定によって、混合ガウス分布 $p(\boldsymbol{x})$ のパラメータ $\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}\}$ を求めてみる
  - 尤度関数 $p(\mathcal{D}|\boldsymbol{\theta})$ を、パラメータ $\boldsymbol{\theta}$ の関数とみなして、$\boldsymbol{\theta}$ について最大化することにより、$\boldsymbol{\theta}$ を求めるという考え方
  - 尤度関数 $p(\mathcal{D}|\boldsymbol{\theta})$ は、パラメータ $\boldsymbol{\theta}$ が与えられたときの、データの条件付き確率である
  - パラメータを 1 つに決めたときに、データ $\mathcal{D}$ が得られる確率

混合ガウス分布

- 対数尤度関数 $\ln p(\mathcal{D}|\boldsymbol{\theta})$ は次のようになる

$$
\begin{aligned}
& \ln p(\mathcal{D}|\boldsymbol{\theta}) \\
= \quad & \ln \prod_i p(\boldsymbol{x}_i|\boldsymbol{\theta}) \\
= \quad & \ln \prod_i \left(\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \\
= \quad & \sum_i \ln \left(\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \\
= \quad & \sum_i \ln \left(\sum_k \pi_k \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \right. \\
& \qquad \left. \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} \right) \qquad (24)
\end{aligned}
$$

混合ガウス分布

- 上式の最初の変形では、各データ $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ は、確率分布 $p(\boldsymbol{x})$ から、独立に得られているという仮定を用いた
- このようなデータ $\mathcal{D}$ を、i.i.d 標本という (independently and identically distributed)

- 対数 $\ln$ は単調増加関数であるため、対数を適用しても、関数の極値は変化しない
- 尤度関数 $p(\mathcal{D}|\boldsymbol{\theta})$ の最大化は、対数尤度関数 $\ln p(\mathcal{D}|\boldsymbol{\theta})$ の最大化と等価

- 重大な問題点
  - 対数関数 $\ln$ の内部に、総和 $(\sum)$ が入ってしまっている
  - log-sum の形状になっているため、これ以上式を簡単にできない!
  - クラスタ数 $K = 1$ であれば、対数 $\ln$ と、ガウス分布の指数 $\exp$ が打ち消し合って、式が簡潔になる

  - しかしここでは、このまま最尤推定を続けてみる

混合ガウス分布

- パラメータ $\boldsymbol{\mu}_k$ の最尤推定
  - 対数尤度関数 $\ln p(\mathcal{D}|\boldsymbol{\theta})$ において、$\boldsymbol{\theta}$ はパラメータ (定数) で、$\mathcal{D}$ は変数であるが、実際は $\mathcal{D}$ にはデータが入っているので、パラメータ $\boldsymbol{\theta}$ を変数とみなす

  - $\ln p(\mathcal{D}|\boldsymbol{\theta})$ をパラメータ $\boldsymbol{\mu}_k$ で微分してみる
  - これを 0 と等置することで、最適な $\boldsymbol{\mu}_k$ が満たすべき式が得られる

  - その前に、関数 f の対数の微分について、以下が成立することを確認しておく

$$(\ln f)' = \frac{f'}{f}, \qquad f' = f \cdot (\ln f)' \tag{25}$$

混合ガウス分布

- このとき次のようになる

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\mu}_k} \ln p(\mathcal{D}|\boldsymbol{\theta}) \\
= & \frac{\partial}{\partial \boldsymbol{\mu}_k} \sum_i \ln \left(\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \\
= & \sum_i \frac{\partial}{\partial \boldsymbol{\mu}_k} \ln \left(\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \\
= & \sum_i \frac{1}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \frac{\partial}{\partial \boldsymbol{\mu}_k} \left(\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \\
= & \sum_i \frac{1}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \frac{\partial}{\partial \boldsymbol{\mu}_k} \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
= & \sum_i \frac{1}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \left(\frac{\partial}{\partial \boldsymbol{\mu}_k} \pi_k \right.
\end{aligned}
$$

混合ガウス分布

$$\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right\}\Bigg)$$

$$= \sum_i \frac{1}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}\pi_k\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}}$$

$$\frac{\partial}{\partial\boldsymbol{\mu}_k}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right\} \tag{26}$$

ここで、以下のようにできる

$$\frac{\partial}{\partial\boldsymbol{\mu}_k}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right\}$$

$$= \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right\}$$

$$\frac{\partial}{\partial\boldsymbol{\mu}_k}\left(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right) \tag{27}$$

混合ガウス分布

更に、次が成立する

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\mu}_k}\left(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right) \\
= & -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\mu}_k}\left(\boldsymbol{x}_i^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x}_i-\boldsymbol{x}_i^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k-\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x}_i+\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right) \\
= & -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\mu}_k}\left(-\boldsymbol{x}_i^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k-\boldsymbol{x}_i^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k+\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right) \\
= & -\frac{1}{2}\left(-2\boldsymbol{x}_i^T\boldsymbol{\Sigma}_k^{-1}+2\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k\right) \\
= & \boldsymbol{\Sigma}_k^{-1}\boldsymbol{x}_i-\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k \\
= & \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k) \qquad (28)
\end{aligned}
$$

ここで、以下を用いた

$$
\begin{aligned}
& \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x}_i \\
= & \left(\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{x}_i\right)^T \quad (\because \text{スカラーであるため転置してもよい})
\end{aligned}
$$

$$
\begin{aligned}
&= \boldsymbol{x}_i^T \left(\boldsymbol{\Sigma}_k^{-1}\right)^T \left(\boldsymbol{\mu}_k^T\right)^T \\
&= \boldsymbol{x}_i^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k
\end{aligned} \tag{29}
$$

共分散行列 $\boldsymbol{\Sigma}_k$ は対称行列 $\left(\boldsymbol{\Sigma}_k^T = \boldsymbol{\Sigma}_k\right)$ であるため、以下が成立

$$
\left(\boldsymbol{\Sigma}_k^{-1}\right)^T = \left(\boldsymbol{\Sigma}_k^T\right)^{-1} = \boldsymbol{\Sigma}_k^{-1} \tag{30}
$$

各項の微分は次のようになる

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\mu}_k} \boldsymbol{x}_i^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \\
= \quad & \left(\boldsymbol{x}_i^T \boldsymbol{\Sigma}_k^{-1}\right)^T = \left(\boldsymbol{\Sigma}_k^{-1}\right)^T \left(\boldsymbol{x}_i^T\right)^T = \boldsymbol{\Sigma}_k^{-1} \boldsymbol{x}_i
\end{aligned} \tag{31}
$$

$$
\begin{aligned}
&\frac{\partial}{\partial \boldsymbol{\mu}_k} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \\
= \quad & \left(\boldsymbol{\Sigma}_k^{-1} + \left(\boldsymbol{\Sigma}_k^{-1}\right)^T\right) \boldsymbol{\mu}_k = 2\boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k
\end{aligned} \tag{32}
$$

混合ガウス分布

結局、対数尤度関数 $\ln p(\mathcal{D}|\boldsymbol{\theta})$ の $\boldsymbol{\mu}_k$ による微分は

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\mu}_k} \ln p(\mathcal{D}|\boldsymbol{\theta}) \\
= & \sum_i \frac{1}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \pi_k \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \\
& \qquad \frac{\partial}{\partial \boldsymbol{\mu}_k} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right\} \\
= & \sum_i \frac{1}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \pi_k \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \\
& \qquad \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right\} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k) \\
= & \sum_i \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k) \qquad (33) \\
= & \ 0
\end{aligned}
$$

混合ガウス分布

- 未知のパラメータ $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ が、分母と分子の双方に出現する複雑な式
- 直接この連立方程式を解いて、パラメータの最尤推定量を求めるのは難しそうである

- 勾配 $\nabla_{\boldsymbol{\mu}_k} \ln p(\mathcal{D}|\boldsymbol{\theta})$ を利用した最適化も可能である
- この勾配の方向に、パラメータ $\boldsymbol{\mu}_k$ を少しだけ更新する
- ここでは、EM アルゴリズムという別の手法を導出しようとしている

- $\boldsymbol{x}$ のほかに、潜在変数という仮想的な変数 $\boldsymbol{z}$ を導入することで、簡単に解けるようになる

- 上と似たような式が、後ほど登場する

混合ガウス分布

- ここまでの話の流れ
  1. K-Means では、データを単一のクラスタに割り当てた (ハード割り当て)
  2. データが属するクラスタだけではなく、より多くの情報 (各クラスタに属する確率) を手に入れたい
  3. ソフト割り当てを実現するためには、クラスタリングを統計的機械学習 (確率分布) の観点から見直して、再定式化を行う必要があった
  4. 各クラスタをガウス分布として、データ全体を混合ガウス分布に当てはめることを考えた
  5. 混合ガウス分布のパラメータを、最尤推定により求めようとしたが、困難であることが分かった
  6. そこで、潜在変数を導入して、最尤推定を簡単に解こうと考えている

混合ガウス分布

- 潜在変数 $\boldsymbol{z}$ の導入
  - 各データ $\boldsymbol{x}_i$ につき、1 つのベクトル $\boldsymbol{z}_i \in \mathbb{R}^K$ が対応しているとする
  - $\boldsymbol{z}_i$ は、データ $\boldsymbol{x}_i$ が属するクラスタ を表現する

- 潜在変数 $\boldsymbol{z}$ の表現
  - $\boldsymbol{z}_i$ は、K 次元の二値確率変数 $\boldsymbol{z}$ の観測値である
  - $\boldsymbol{z}$ の k 番目の要素を、z_k と表すことにする

  - 確率変数 $\boldsymbol{z}$ は、1-of-K 符号化法により表現されるとする
  - 即ち、ある 1 つの $k \in \{1, \ldots, K\}$ について $z_k = 1$ で、$j \neq k$ に対し $z_j = 0$ となる
  - $z_k (k = 1, \ldots, K)$ は、$z_k \in \{0, 1\}$ かつ $\sum_k z_k = 1$ をみたす

  - ベクトル $\boldsymbol{z}$ は K 種類の状態を取る

混合ガウス分布

- 潜在変数 $\boldsymbol{z}$ の例
  - 例えば、データ点 $\boldsymbol{x}_i$ に対して $\boldsymbol{z}_i$ があるとする
  - $z_{i1} = 1$ $(\boldsymbol{z}_i = [1, 0, 0, \ldots, 0])$ ならば、$\boldsymbol{x}_i$ は 1 番目のクラスタ出身
  - $z_{i2} = 1$ $(\boldsymbol{z}_i = [0, 1, 0, \ldots, 0])$ ならば、$\boldsymbol{x}_i$ は 2 番目のクラスタ出身

- データ $\boldsymbol{x}_i$ が作られるまでの流れ
  - $\boldsymbol{z}$ に関する確率分布 $p(\boldsymbol{z})$ から、$\boldsymbol{z}_i$ がサンプルされる
  - $\boldsymbol{z}$ が与えられた下での条件付き分布 $p(\boldsymbol{x}|\boldsymbol{z})$ から、$\boldsymbol{x}_i$ がサンプルされる

  - 即ち、$\boldsymbol{x}, \boldsymbol{z}$ の同時分布は、周辺分布 $p(\boldsymbol{z})$ と、条件付き分布 $p(\boldsymbol{x}|\boldsymbol{z})$ を用いて次のように書ける

$$p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}) \tag{34}$$

  - 潜在変数 $\boldsymbol{z}_i$ が最初に決められ、その $\boldsymbol{z}_i$ に応じて $\boldsymbol{x}_i$ が決まると考える
  - $\boldsymbol{z}_i$ は実際には存在しない、仮想的なものである

  - $\boldsymbol{z}_i$ は、実際に観測される $\boldsymbol{x}_i$ の裏側に潜んでいる

混合ガウス分布

- $p(\boldsymbol{x})$ の表現 (予想)
  - $p(\boldsymbol{x})$ は混合ガウス分布になってほしい

$$p(\boldsymbol{x}) = \sum_k \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{35}$$

- 周辺分布 $p(\boldsymbol{z})$ の定義
  - $p(\boldsymbol{z})$ は次のように定める

$$p(\boldsymbol{z}) = \prod_k \pi_k^{z_k}, \quad p(z_k = 1) = \pi_k \tag{36}$$

  - 但し π_k は混合係数であり、$\sum_k \pi_k = 1, 0 \leq \pi_k \leq 1$ をみたす
  - $\boldsymbol{z}$ の表現には 1-of-K 符号化法を使うため、左側のようにも書ける

混合ガウス分布

- 条件付き分布 $p(\boldsymbol{x}|\boldsymbol{z})$ の定義
  - $p(\boldsymbol{x}|\boldsymbol{z})$ は次のように定める

$$p(\boldsymbol{x}|\boldsymbol{z}) = \prod_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \tag{37}$$

$$p(\boldsymbol{x}|z_k = 1) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{38}$$

- $p(\boldsymbol{x})$ の導出
  - $\sum_{\boldsymbol{z}}$ は、可能な全ての $\boldsymbol{z}$ についての総和を取るということ
  - $\boldsymbol{z} = [1, 0, \ldots, 0]^T, [0, 1, 0, \ldots, 0]^T, \ldots, [0, \ldots, 0, 1]^T$ についての和
  - これは、ベクトル $\boldsymbol{z}$ の中で、1 である要素のインデックス k についての総和 $\sum_k$ を取ることに相当

混合ガウス分布

- $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$ を、$\boldsymbol{z}$ について周辺化すればよい

$$
\begin{aligned}
p(\boldsymbol{x}) &= \sum_{\boldsymbol{z}} p(\boldsymbol{x}, \boldsymbol{z}) && (39)\\
&= \sum_{\boldsymbol{z}} p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z}) && (40)\\
&= \sum_{k} p(z_k = 1)p(\boldsymbol{x}|z_k = 1) && (41)\\
&= \sum_{k} \pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) && (42)
\end{aligned}
$$

  - これは、混合ガウス分布と同じ形になっている

- 何が嬉しいのか?
  - 潜在変数を陽に含む表現 $p(\boldsymbol{x}, \boldsymbol{z}) = p(\boldsymbol{z})p(\boldsymbol{x}|\boldsymbol{z})$ を得たことで、この同時分布を使った議論が可能になった

混合ガウス分布

- $p(z_k = 1|\boldsymbol{x})$ の表現
  - $\boldsymbol{x}$ が与えられた下での、$\boldsymbol{z}$ の条件付き確率
  - 実は、$p(z_k = 1|\boldsymbol{x})$ は、データ $\boldsymbol{x}$ がクラスタ k に属する確率 を表す
  - 求めようとしているのは、この値である!

  - $\gamma(z_k) = p(z_k = 1|\boldsymbol{x})$ とすると、ベイズの定理から次のように書ける

$$\begin{aligned} \gamma(z_k) &\equiv p(z_k = 1|\boldsymbol{x}) && (43) \\ &= \frac{p(z_k = 1)p(\boldsymbol{x}|z_k = 1)}{\sum_j p(z_j = 1)p(\boldsymbol{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} && (44) \end{aligned}$$

  - $\sum_k \gamma(z_k) = 1$ であることに注意

  - 潜在変数を導入したので、最尤推定について再度考えてみる

混合ガウス分布

- 最尤推定を再挑戦
  - パラメータ $\boldsymbol{\theta}$ は、$\boldsymbol{\pi} \equiv \{\pi_1, \pi_2, \ldots, \pi_K\}$、$\boldsymbol{\mu} \equiv \{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K\}$、$\boldsymbol{\Sigma} \equiv \{\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_K\}$ をまとめたもの
  - 対数尤度関数 $\ln p(\mathcal{D}|\boldsymbol{\theta})$ は以下に示す通りであった

$$\begin{aligned} & \ln p(\mathcal{D}|\boldsymbol{\theta}) \\ = & \sum_i \ln \left(\sum_k \pi_k \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \right. \\ & \left. \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} \right) \end{aligned} \tag{45}$$

  - 尤度関数を、$\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ のそれぞれについて最大化する
  - ここでは、尤度関数を最大化する $\boldsymbol{\mu}_k$ が、満たすべき条件を考える

混合ガウス分布

- $\ln p(\mathcal{D}|\boldsymbol{\theta})$ を $\boldsymbol{\mu}_k$ について偏微分して 0 と等置すると、以下を得る

$$\begin{aligned} & \frac{\partial}{\partial \boldsymbol{\mu}_k} \ln p(\mathcal{D}|\boldsymbol{\theta}) \\ = & \sum_i \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k) \qquad (46) \\ = & \sum_i \gamma(z_{ik}) \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k) = 0 \qquad (47) \end{aligned}$$

- 途中までは、先程導出したものを利用
- 負担率 $\gamma(z_{ik})$ が現れていることに注意

混合ガウス分布

- 共分散行列 $\boldsymbol{\Sigma}_k$ が正則であると仮定して、両辺に左から掛けて整理する

$$
\begin{aligned}
& \sum_i \gamma(z_{ik})(\boldsymbol{x}_i - \boldsymbol{\mu}_k) = 0 \\
\Rightarrow \quad & \sum_i \gamma(z_{ik})\boldsymbol{\mu}_k = \sum_i \gamma(z_{ik})\boldsymbol{x}_i \\
\Rightarrow \quad & \boldsymbol{\mu}_k \sum_i \gamma(z_{ik}) = \sum_i \gamma(z_{ik})\boldsymbol{x}_i \\
\Rightarrow \quad & \boldsymbol{\mu}_k = \frac{1}{\sum_i \gamma(z_{ik})} \sum_i \gamma(z_{ik})\boldsymbol{x}_i
\end{aligned}
\tag{48}
$$

- これより、$\boldsymbol{\mu}_k$ を導出する式が得られた

混合ガウス分布

- K-Means 法との比較
  - K-Means 法における、平均ベクトル $\boldsymbol{\mu}_k$ の更新式と見比べてみる

$$\boldsymbol{\mu}_k = \frac{1}{\sum_i \gamma(z_{ik})} \sum_i \gamma(z_{ik}) \boldsymbol{x}_i$$

$$\boldsymbol{\mu}_k = \frac{1}{\sum_i r_{ik}} \sum_i r_{ik} \boldsymbol{x}_i \tag{49}$$

  - r_{ik} を、$\gamma(z_{ik})$ に置き換えたものとなっている
  - $\gamma(z_{ik})$ は、データ $\boldsymbol{x}_i$ が、クラスタ k に属する確率である
  - $\gamma(z_{ik})$ を、全てのデータ $\boldsymbol{x}_i$ について足し合わせたもの $\sum_i \gamma(z_{ik})$ は、実質的に、k 番目のクラスタに割り当てられるデータの数を表している (整数になるとは限らない)

混合ガウス分布

- そこで、K-Means 法のときと同じように、N_k を次のように定める

$$N_k = \sum_i \gamma(z_{ik}) \tag{50}$$

このとき、$\boldsymbol{\mu}_k$ の式は

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_i \gamma(z_{ik}) \boldsymbol{x}_i \tag{51}$$

- 例えば、$\gamma(z_{ik})$ が $0, 1$ のいずれかであれば、$\sum_i \gamma(z_{ik})$ は、k 番目のクラスタに属するデータの数と完全に一致

- クラスタ k に対応するガウス分布の平均 $\boldsymbol{\mu}_k$ は、各データ $\boldsymbol{x}_i$ の重み付き平均
- 重み因子は、事後確率 $p(z_k = 1|\boldsymbol{x}_i) \equiv \gamma(z_{ik})$ である

混合ガウス分布

- $\gamma(z_{ik})$ は、$\sum_k \gamma(z_{ik}) = 1$ となることからも分かるように、$\boldsymbol{x}_i$ を生成するために、k 番目のガウス分布が、どの程度貢献したかを表す
- 言い換えると、k 番目のガウス分布が、$\boldsymbol{x}_i$ の出現を説明する度合いである

- この意味で、$\gamma(z_{ik})$ のことを負担率 (Responsibility) という

混合ガウス分布

- 尤度関数を最大化する $\boldsymbol{\Sigma}_k$ の導出
  - $\boldsymbol{\mu}_k$ の場合と同様に、対数尤度関数 $\ln p(\mathcal{D}|\boldsymbol{\theta})$ を、$\boldsymbol{\Sigma}_k$ に関して微分して、0 と等置すればよい
  - かなり導出が長くなるので注意

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln p(\mathcal{D}|\boldsymbol{\theta}) \\
= & \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \sum_i \ln \left(\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \\
= & \sum_i \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln \left(\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \\
= & \sum_i \frac{1}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \left(\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \\
= & \sum_i \frac{1}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad (52)
\end{aligned}
$$

混合ガウス分布

ここで、以下の部分を求める

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \pi_k \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
= & \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \pi_k \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} \\
= & \pi_k \frac{1}{(2\pi)^{\frac{D}{2}}} \left(\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \right. \\
& \left. \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} \right) \qquad (53)
\end{aligned}
$$

微分の中身は、$\boldsymbol{\Sigma}_k$ についての合成関数となっている

$$
\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\}
$$

混合ガウス分布

- 一般に、行列 $\boldsymbol{X}$ についてのスカラー関数 $f(\boldsymbol{X}), g(\boldsymbol{X})$ があるとき、以下の連鎖律が成立

$$\frac{\partial}{\partial \boldsymbol{X}} f(\boldsymbol{X}) g(\boldsymbol{X}) = f(\boldsymbol{X}) \frac{\partial g(\boldsymbol{X})}{\partial \boldsymbol{X}} + g(\boldsymbol{X}) \frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}} \tag{54}$$

これを利用して、先程の微分を求める

$$\begin{aligned} & \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right\} \\ = & \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right\} + \\ & \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right\} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \end{aligned} \tag{55}$$

混合ガウス分布

- 各項の微分を順番に求める

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \\
= & \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \exp\left(\ln \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \right) \\
= & \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \exp\left(-\frac{1}{2} \ln |\boldsymbol{\Sigma}_k| \right) \\
= & \exp\left(-\frac{1}{2} \ln |\boldsymbol{\Sigma}_k| \right) \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \left(-\frac{1}{2} \ln |\boldsymbol{\Sigma}_k| \right) \\
= & \exp\left(-\frac{1}{2} \ln |\boldsymbol{\Sigma}_k| \right) \left(-\frac{1}{2} \right) \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln |\boldsymbol{\Sigma}_k| \\
= & -\frac{1}{2} \exp\left(-\frac{1}{2} \ln |\boldsymbol{\Sigma}_k| \right) \left(\boldsymbol{\Sigma}_k^{-1} \right)^T \\
= & -\frac{1}{2} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \boldsymbol{\Sigma}_k^{-1}
\end{aligned}
\tag{56}
$$

混合ガウス分布

また

$$\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right\} \\
= \quad & \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right\} \\
& \frac{\partial}{\partial \boldsymbol{\Sigma}_k}\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right) \qquad (57)
\end{aligned}$$

であって

$$\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\Sigma}_k}\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right) \\
= \quad & -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}_k}\left((\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right) \\
= \quad & -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}_k}\mathrm{Tr}\left((\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)\right)
\end{aligned}$$

混合ガウス分布

$$
\begin{aligned}
&= -\frac{1}{2}\frac{\partial}{\partial \boldsymbol{\Sigma}_k}\mathrm{Tr}\left(\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\right)\\
&= -\frac{1}{2}\left(-\left(\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}\right)^T\right)\\
&= \frac{1}{2}\left(\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}\right)^T\\
&= \frac{1}{2}\left(\boldsymbol{\Sigma}_k^{-1}\right)^T\left((\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\right)^T\left(\boldsymbol{\Sigma}_k^{-1}\right)^T\\
&= \frac{1}{2}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1} \qquad (58)
\end{aligned}
$$

のように求まる

- 行列のトレースについて、一般に以下が成り立つことを利用している

$$\mathrm{Tr}\left(\boldsymbol{X}\boldsymbol{Y}\right) = \mathrm{Tr}\left(\boldsymbol{Y}\boldsymbol{X}\right) \qquad (59)$$

$$\mathrm{Tr}\left(\boldsymbol{X}\boldsymbol{Y}\boldsymbol{Z}\right) = \mathrm{Tr}\left(\boldsymbol{Y}\boldsymbol{Z}\boldsymbol{X}\right) = \mathrm{Tr}\left(\boldsymbol{Z}\boldsymbol{X}\boldsymbol{Y}\right) \qquad (60)$$

混合ガウス分布

- また先程の微分では以下の公式を用いている

$$\frac{\partial}{\partial \boldsymbol{X}} \mathrm{Tr}\left(\boldsymbol{X}^{-1}\boldsymbol{Y}\right) = -\left(\boldsymbol{X}^{-1}\boldsymbol{Y}\boldsymbol{X}^{-1}\right)^{T} \tag{61}$$

- この公式を証明するためには、いくつかの段階を踏む必要がある
- まずは以下の微分公式の導出から始める

$$\frac{\partial}{\partial x}\boldsymbol{A}\boldsymbol{B}$$

上式の微分は、行列の積 $\boldsymbol{A}\boldsymbol{B}$ の i, k 成分について考えれば

$$\frac{\partial}{\partial x}\sum_{j} A_{ij}B_{jk}$$

$$= \sum_{j}\frac{\partial}{\partial x}A_{ij}B_{jk}$$

混合ガウス分布

$$= \sum_j \left(\frac{\partial A_{ij}}{\partial x} B_{jk} + A_{ij} \frac{\partial B_{jk}}{\partial x} \right)$$

$$= \sum_j \frac{\partial A_{ij}}{\partial x} B_{jk} + \sum_j A_{ij} \frac{\partial B_{jk}}{\partial x} \tag{62}$$

であるから、結局

$$\frac{\partial}{\partial x} \boldsymbol{AB} = \frac{\partial \boldsymbol{A}}{\partial x} \boldsymbol{B} + \boldsymbol{A} \frac{\partial \boldsymbol{B}}{\partial x} \tag{63}$$

となる

混合ガウス分布

- 上記の公式から、以下の公式を簡単に導ける

$$\frac{\partial}{\partial x}\boldsymbol{A}^{-1}\boldsymbol{A} = \frac{\partial \boldsymbol{A}^{-1}}{\partial x}\boldsymbol{A} + \boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}$$
$$\frac{\partial}{\partial x}\boldsymbol{I} = \frac{\partial \boldsymbol{A}^{-1}}{\partial x}\boldsymbol{A} + \boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}$$
$$0 = \frac{\partial \boldsymbol{A}^{-1}}{\partial x}\boldsymbol{A} + \boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x} \tag{64}$$

これに右から $\boldsymbol{A}^{-1}$ を掛ければ

$$0 = \frac{\partial \boldsymbol{A}^{-1}}{\partial x}\boldsymbol{A}\boldsymbol{A}^{-1} + \boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{A}^{-1}$$
$$0 = \frac{\partial \boldsymbol{A}^{-1}}{\partial x}\boldsymbol{I} + \boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{A}^{-1}$$
$$0 = \frac{\partial \boldsymbol{A}^{-1}}{\partial x} + \boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{A}^{-1}$$

混合ガウス分布

$$\frac{\partial}{\partial x}\boldsymbol{A}^{-1} = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{A}^{-1} \tag{65}$$

より

$$\frac{\partial}{\partial x}\boldsymbol{A}^{-1} = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{A}^{-1} \tag{66}$$

を得る (逆行列の微分公式)

- 逆行列 $\boldsymbol{A}^{-1}$ の k, l 成分 $\left(\boldsymbol{A}^{-1}\right)_{kl}$ については、以下のように書ける

$$\begin{aligned}
& \frac{\partial}{\partial x}\left(\boldsymbol{A}^{-1}\right)_{kl} \\
= & -\sum_{m,n}\left(\boldsymbol{A}^{-1}\right)_{km}\left(\frac{\partial \boldsymbol{A}}{\partial x}\right)_{mn}\left(\boldsymbol{A}^{-1}\right)_{ml} \\
= & -\sum_{m,n}\left(\boldsymbol{A}^{-1}\right)_{km}\frac{\partial A_{mn}}{\partial x}\left(\boldsymbol{A}^{-1}\right)_{ml}
\end{aligned} \tag{67}$$

混合ガウス分布

- この逆行列の微分公式を使えば、以下の微分公式を導出できる

$$\frac{\partial}{\partial \boldsymbol{X}} \mathrm{Tr}\left(\boldsymbol{X}^{-1}\boldsymbol{Y}\right)$$

行列 $\boldsymbol{X}$ の i, j 要素による微分を考えれば

$$\begin{aligned}
& \frac{\partial}{\partial X_{ij}} \mathrm{Tr}\left(\boldsymbol{X}^{-1}\boldsymbol{Y}\right) \\
= & \frac{\partial}{\partial X_{ij}} \sum_{k,l} \left(\boldsymbol{X}^{-1}\right)_{kl} Y_{lk} \\
= & \sum_{k,l} \frac{\partial}{\partial X_{ij}} \left(\left(\boldsymbol{X}^{-1}\right)_{kl}\right) Y_{lk} \\
= & \sum_{k,l} \left(-\sum_{m,n} \left(\boldsymbol{X}^{-1}\right)_{km} \frac{\partial X_{mn}}{\partial X_{ij}} \left(\boldsymbol{X}^{-1}\right)_{ml} \right) Y_{lk} \\
= & \sum_{k,l} \left(-\left(\boldsymbol{X}^{-1}\right)_{ki} \left(\boldsymbol{X}^{-1}\right)_{jl} \right) Y_{lk}
\end{aligned}$$

混合ガウス分布

$$
\begin{aligned}
&= \sum_{k,l} \left(- \left(\boldsymbol{X}^{-1} \right)_{jl} Y_{lk} \left(\boldsymbol{X}^{-1} \right)_{ki} \right) \\
&= - \left(\boldsymbol{X}^{-1} \boldsymbol{Y} \boldsymbol{X}^{-1} \right)_{ji} \\
&= - \left(\left(\boldsymbol{X}^{-1} \boldsymbol{Y} \boldsymbol{X}^{-1} \right)^T \right)_{ij} \qquad (68)
\end{aligned}
$$

であるから

$$
\frac{\partial}{\partial \boldsymbol{X}} \operatorname{Tr} \left(\boldsymbol{X}^{-1} \boldsymbol{Y} \right) = - \left(\boldsymbol{X}^{-1} \boldsymbol{Y} \boldsymbol{X}^{-1} \right)^T \qquad (69)
$$

を得られる

混合ガウス分布

- 尤度関数を最大化する $\boldsymbol{\Sigma}_k$ の導出
  - これより、結局次のようになる

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln p(\mathcal{D}|\boldsymbol{\theta}) \\
= & \sum_i \frac{1}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
= & \sum_i \frac{1}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \pi_k \frac{1}{(2\pi)^{\frac{D}{2}}} \\
& \left(\frac{\partial}{\partial \boldsymbol{\Sigma}_k} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} \right) \\
= & \sum_i \frac{1}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \pi_k \frac{1}{(2\pi)^{\frac{D}{2}}} \\
& \left(\frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} + \right.
\end{aligned}
$$

混合ガウス分布

$$
\begin{aligned}
& \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right\}\frac{\partial}{\partial\boldsymbol{\Sigma}_k}\frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}}\Bigg) \\
= & \sum_i \frac{1}{\sum_k \pi_k\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}\pi_k\frac{1}{(2\pi)^{\frac{D}{2}}} \\
& \left(\frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right\}\right. \\
& \left(\frac{1}{2}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}\right)- \\
& \left.\frac{1}{2}\frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}}\boldsymbol{\Sigma}_k^{-1}\exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right\}\right) \\
= & \sum_i \frac{1}{\sum_k \pi_k\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}\pi_k\frac{1}{(2\pi)^{\frac{D}{2}}}\frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \\
& \exp\left\{-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)\right\}
\end{aligned}
$$

混合ガウス分布

$$\left(\frac{1}{2}\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}-\frac{1}{2}\boldsymbol{\Sigma}_k^{-1}\right)$$

$$= \sum_i \frac{\pi_k\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_k \pi_k\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)} \frac{1}{2}\boldsymbol{\Sigma}_k^{-1}\left((\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}-\boldsymbol{I}\right)$$

$$= \frac{1}{2}\sum_i \gamma(z_{ik})\boldsymbol{\Sigma}_k^{-1}\left((\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}-\boldsymbol{I}\right)=0 \quad (70)$$

両辺に左右から $\boldsymbol{\Sigma}_k$ を掛けて、整理すれば

$$\frac{1}{2}\sum_i \gamma(z_{ik})\left((\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T-\boldsymbol{\Sigma}_k\right)=0$$

$$\Rightarrow \sum_i \gamma(z_{ik})\boldsymbol{\Sigma}_k = \sum_i \gamma(z_{ik})(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T$$

$$\Rightarrow \boldsymbol{\Sigma}_k = \frac{1}{\sum_i \gamma(z_{ik})}\sum_i \gamma(z_{ik})(\boldsymbol{x}_i-\boldsymbol{\mu}_k)(\boldsymbol{x}_i-\boldsymbol{\mu}_k)^T \quad (71)$$

混合ガウス分布

$$\Rightarrow \quad \boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_i \gamma(z_{ik})(\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \tag{72}$$

- これより、$\boldsymbol{\Sigma}_k$ を導出する式が得られた

混合ガウス分布

- 尤度関数を最大化する π_k の導出
  - $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$ の場合と同様に、対数尤度関数 $\ln p(\mathcal{D}|\boldsymbol{\theta})$ を、π_k に関して微分して、0 と等置すればよい
  - 但し、$\sum_k \pi_k = 1$ という制約条件を考慮しなければならない
  - そのため、ラグランジュの未定係数法を用いる

  - 以下を最大化する π_k を求める

$$\ln p(\mathcal{D}|\boldsymbol{\theta}) + \lambda \left(\sum_k \pi_k - 1 \right) \tag{73}$$

混合ガウス分布

- π_k で微分して 0 と等置すると、次のようになる

$$\frac{\partial}{\partial \pi_k}\left(\ln p(\mathcal{D}|\boldsymbol{\theta}) + \lambda\left(\sum_k \pi_k - 1\right)\right)$$

$$= \frac{\partial}{\partial \pi_k}\left(\sum_i \ln\left(\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) + \lambda\left(\sum_k \pi_k - 1\right)\right)$$

$$= \sum_i \frac{\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} + \lambda \tag{74}$$

$$= \sum_i \frac{\gamma(z_{ik})}{\pi_k} + \lambda = 0 \tag{75}$$

両辺に π_k を掛けて

$$\sum_i \gamma(z_{ik}) + \lambda \pi_k = 0 \tag{76}$$

混合ガウス分布

k についての和を取ると

$$\sum_k \left(\sum_i \gamma(z_{ik}) + \lambda \pi_k \right) = 0 \tag{77}$$

$$\sum_i \sum_k \gamma(z_{ik}) + \lambda \sum_k \pi_k = 0 \tag{78}$$

$$\sum_i 1 + \lambda = 0 \tag{79}$$

$$N + \lambda = 0 \tag{80}$$

$$\therefore \lambda = -N \tag{81}$$

これより

$$\sum_i \gamma(z_{ik}) + (-N)\pi_k = 0 \tag{82}$$

$$\pi_k = \frac{1}{N} \sum_i \gamma(z_{ik}) = \frac{N_k}{N} \tag{83}$$

混合ガウス分布

ここで、以下が成立することに注意

$$N_k = \sum_i \gamma(z_{ik}), \quad 1 = \sum_k \gamma(z_{ik})$$

- これより、混合係数 π_k は、全ての要素における、クラスタ k の負担率 $\gamma(z_{ik})$ の平均である
- ここまでで、対数尤度関数 $\ln p(\mathcal{D}|\boldsymbol{\theta})$ を最大化するような、$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ の式が得られた

混合ガウス分布

$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ の更新式

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_i \gamma(z_{ik}) \boldsymbol{x}_i \tag{84}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k} \sum_i \gamma(z_{ik}) (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \tag{85}$$

$$\pi_k = \frac{N_k}{N} \tag{86}$$

$$N_k = \sum_i \gamma(z_{ik}) \tag{87}$$

混合ガウス分布

$\gamma(z_{ik})$ の更新式

$$\gamma(z_{ik}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \tag{88}$$

- 注意点
  - $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ の更新式は、これらのパラメータについての、陽な解は与えていない
  - なぜなら、これらの更新式は全て、負担率 $\gamma(z_{ik})$ に依存しているため
  - そしてその負担率 $\gamma(z_{ik})$ は、$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ の全てに依存する

混合ガウス分布

- これらの更新式の意味
  - 最尤推定の解を求めるための、繰り返し手続きの存在を示唆
  - 即ち、$\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ の初期化後に、(1) $\gamma(z_{ik})$ の更新と、(2) それを用いた $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ の更新という、2 段階の処理を繰り返す手続き

  - これは、混合ガウス分布を確率モデルとして使ったときの、EM アルゴリズムとなっている

  - 混合ガウス分布に対する EM アルゴリズムは重要なので、次にまとめる

混合ガウス分布

混合ガウス分布に対する EM アルゴリズム

- 目的は、混合ガウスモデルが与えられているとき、そのパラメータ (各ガウス分布の平均、分散、そして混合係数) について、尤度関数を最大化することである

1. 平均 $\boldsymbol{\mu}_k^{\text{old}}$、分散 $\boldsymbol{\Sigma}_k^{\text{old}}$、そして混合係数 π_k^{old} を初期化し、対数尤度 $\ln p(\mathcal{D}|\boldsymbol{\theta})$ の初期値を計算
2. E ステップ: 現在のパラメータを用いて、負担率 $\gamma(z_{ik})$ を計算

$$\gamma(z_{ik}) \leftarrow \frac{\pi_k^{\text{old}} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k^{\text{old}}, \boldsymbol{\Sigma}_k^{\text{old}})}{\sum_k \pi_k^{\text{old}} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k^{\text{old}}, \boldsymbol{\Sigma}_k^{\text{old}})} \tag{89}$$

混合ガウス分布

3 M ステップ: 現在の負担率 $\gamma(z_{ik})$ を用いて、パラメータを更新

$$\boldsymbol{\mu}_k^{\text{new}} \leftarrow \frac{1}{N_k}\sum_i \gamma(z_{ik})\boldsymbol{x}_i \tag{90}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} \leftarrow \frac{1}{N_k}\sum_i \gamma(z_{ik})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{\text{new}})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{\text{new}})^T \tag{91}$$

$$\pi_k^{\text{new}} \leftarrow \frac{N_k}{N} \tag{92}$$

但し

$$N_k = \sum_i \gamma(z_{ik}) \tag{93}$$

混合ガウス分布

4. 対数尤度 $\ln p(\mathcal{D}|\boldsymbol{\theta})$ を計算

$$\ln p(\mathcal{D}|\boldsymbol{\theta}) = \sum_i \ln \left(\sum_k \pi_k^{\text{new}} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k^{\text{new}}, \boldsymbol{\Sigma}_k^{\text{new}}) \right) \tag{94}$$

パラメータの変化量、あるいは対数尤度の変化量をみて、収束性を判定

5. 収束基準を満たしていなければ、(2) に戻る

$$\boldsymbol{\mu}_k^{\text{old}} \leftarrow \boldsymbol{\mu}_k^{\text{new}}, \quad \boldsymbol{\Sigma}_k^{\text{old}} \leftarrow \boldsymbol{\Sigma}_k^{\text{new}}, \quad \pi_k^{\text{old}} \leftarrow \pi_k^{\text{new}} \tag{95}$$

混合ガウス分布

- EM アルゴリズムの概要
  - E ステップ (Expectation step) では、事後確率 $p(z_k = 1|\boldsymbol{x}_i)$、即ち負担率 $\gamma(z_{ik})$ を計算
  - M ステップ (Maximization step) では、事後確率を使って、各パラメータ $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ を再計算

- EM アルゴリズムでの注意点
  - 上記 (3) の M ステップにおける、各パラメータの計算順序に注意
  - 最初に新しい平均値 $\boldsymbol{\mu}_k^{\text{new}}$ を計算し、その新しい平均値を使って、新しい共分散行列 $\boldsymbol{\Sigma}_k^{\text{new}}$ を計算する
  - E ステップと M ステップは、対数尤度関数 $\ln p(\mathcal{D}|\boldsymbol{\theta})$ を増加させることが保証されている
  - EM アルゴリズムは、K-Means 法と比べて、収束までに必要な繰り返し回数と、各ステップでの計算量が非常に多くなる

混合ガウス分布

- 混合ガウス分布の良い初期値を見つけるために、最初に K-Means 法を実行し、その後に EM アルゴリズムを利用する、という方法がある

- K-Means 法により得られた平均ベクトル $\boldsymbol{\mu}_k$ を、各ガウス分布の平均 $\boldsymbol{\mu}_k$ の初期値とする
- 各クラスタに属するデータ点の標本分散を、共分散行列 $\boldsymbol{\Sigma}_k$ の初期値とする
- 各クラスタに属するデータ点の割合を、混合係数 π_k の初期値とする

- 一般に対数尤度には、多数の局所解が存在するため、その中で最大のもの (大域的最適解) に収束するとは限らない

目次

EM アルゴリズムの解釈

- ここまでの話の流れ
  1. ソフト割り当てを実現するために、確率モデル (混合ガウスモデル) を導入した
  2. 混合ガウス分布のパラメータを、最尤推定により直接求めるのは困難であった
  3. 潜在変数を導入して再度定式化を行い、混合ガウス分布に対する EM アルゴリズムを自然に導出した
  4. EM アルゴリズムの中で、潜在変数は、負担率 (事後分布) の形で登場しただけであった $(\gamma(z_{ik}) = p(z_k = 1|\boldsymbol{x}))$
- これからの話の流れ
  - 潜在変数が果たす重要な役割を明確にする
  - そのうえで、混合ガウス分布の場合をもう一度見直す

EM アルゴリズムの解釈

- EM アルゴリズムの目的
  - 潜在変数をもつ確率モデルについて、パラメータの最尤解を求める

- 対数尤度関数の記述 (一般的な場合)
  - 全ての観測データをまとめた、データ行列を $\boldsymbol{X}$ とする (第 i 行が $\boldsymbol{x}_i^T$)
  - 全ての潜在変数をまとめた行列を $\boldsymbol{Z}$ とする (第 i 行が $\boldsymbol{z}_i^T$)
  - 確率モデルの全てのパラメータを、$\boldsymbol{\theta}$ と表す

  - 対数尤度関数は次のようになる

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \left(\sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \right) \tag{96}$$

EM アルゴリズムの解釈

- 潜在変数 z が連続変数の場合は

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \left(\int p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) d\boldsymbol{Z} \right) \tag{97}$$

  のように、単に総和を積分に置き換えればよい

- これ以降、離散潜在変数のみを扱うが、総和を積分に置き換えれば、ここでの議論は、連続潜在変数についても同様に成立

- 何が問題だったか
  - 対数の中に、潜在変数に関する総和が含まれる (log-sum の形)
  - 総和が存在するので、対数 $\ln$ が、周辺分布 $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ に直接作用することが妨げられる
  - その結果として、対数尤度関数が複雑な形となる

EM アルゴリズムの解釈

- 完全データと不完全データ
  - $\boldsymbol{X}$ だけでなく、$\boldsymbol{Z}$ も観測できるとする
  - $\{\boldsymbol{X}, \boldsymbol{Z}\}$ の組を、完全データ集合という
  - 実際には $\boldsymbol{X}$ しか見えないので、実際の観測データ $\boldsymbol{X}$ は不完全である

  - $\boldsymbol{Z}$ に関する知識は、潜在変数についての事後確率分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ のみからしか得られない

EM アルゴリズムの解釈

重要な仮定と考え方

1. 完全データ対数尤度関数 $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の最大化は、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の最大化よりも、簡単であると仮定
2. $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の代わりに、完全データ対数尤度関数 $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ を最大化したいが、$\boldsymbol{Z}$ に関する情報は $\ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ からしか得られない
3. そのため、完全データ対数尤度関数 $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ は使えない
4. そこで、事後確率分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ による、$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の期待値を最大化することを考える
5. これが、EM アルゴリズムの考え方である

EM アルゴリズムの解釈

- EM アルゴリズムへの落とし込み
  - パラメータ $\boldsymbol{\theta}$ を適当に初期化する
  - E ステップでは、事後確率分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})$ を、現在のパラメータ $\boldsymbol{\theta}^{\mathrm{old}}$ を使って求める
  - $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})$ を、M ステップでの期待値の計算に使う
  - M ステップでは、完全データ対数尤度関数 $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の、事後確率分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})$ に関する期待値 $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}})$ を計算

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \tag{98}$$

連続潜在変数の場合は次のようになる

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) = \int p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) d\boldsymbol{Z} \tag{99}$$

EM アルゴリズムの解釈

- 上式において、$p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})$ におけるパラメータ $\boldsymbol{\theta}^{\text{old}}$ は、変数ではなく定数であることに注意
- 更に、$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ を $\boldsymbol{\theta}$ について最大化することで、新たなパラメータの推定値 $\boldsymbol{\theta}^{\text{new}}$ を得る

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \tag{100}$$

- 注意点
  - $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ において、対数 $\ln$ は、同時分布 $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ に直接作用していることに注意
  - これにより、期待値の計算が簡単になることが期待される
- なぜ事後確率分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ についての期待値なのか
  - 幾分恣意的にみえるが、後ほど、期待値を取ることの正当性が明らかになる

EM アルゴリズムの解釈

一般の EM アルゴリズム

- 観測変数 $\boldsymbol{X}$ と、潜在変数 $\boldsymbol{Z}$ の同時分布 $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ が与えられているとする
- 目的は、尤度関数 $p(\boldsymbol{X}|\boldsymbol{\theta})$ を、パラメータ $\boldsymbol{\theta}$ について最大化することである

1. パラメータを $\boldsymbol{\theta}^{\mathrm{old}}$ に初期化する
2. E ステップ: 事後確率分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})$ を計算する

EM アルゴリズムの解釈

3. M ステップ: 次式で与えられる $\boldsymbol{\theta}^{\text{new}}$ を計算する

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \tag{101}$$

但し

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \tag{102}$$

4. 対数尤度の変化量、あるいはパラメータの変化量をみて、収束性を判定
5. 収束条件を満たしていなければ、(2) に戻る

$$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}} \tag{103}$$

混合ガウス分布の再解釈

- 先程の EM アルゴリズムの解釈で、混合ガウス分布を見直す

- これまでの話の流れ
  - 目的は、対数尤度関数 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の最大化であった
  - しかし、対数の中に総和が出現するため、最尤推定が困難であった
  - そこで、離散潜在変数 $\boldsymbol{Z}$ を導入し、完全データ集合 $\{\boldsymbol{X}, \boldsymbol{Z}\}$ に関する尤度の最大化を考える

混合ガウス分布の再解釈

- 完全データ集合 $\{\boldsymbol{X}, \boldsymbol{Z}\}$ に関する尤度の最大化
  - 完全データ尤度関数 $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ は次のようになる

$$
\begin{aligned}
& p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \\
= \; & p(\boldsymbol{Z}|\boldsymbol{\theta}) p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\theta}) \\
= \; & \prod_i p(\boldsymbol{z}_i|\boldsymbol{\theta}) p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) \\
= \; & \prod_i \left(\prod_k \pi_k^{z_{ik}} \right) \left(\prod_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{ik}} \right) \\
= \; & \prod_i \prod_k \pi_k^{z_{ik}} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{ik}} \\
= \; & \prod_i \prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)^{z_{ik}}
\end{aligned}
\tag{104}
$$

混合ガウス分布の再解釈

- ここで、データ点 $\boldsymbol{x}_i$ に対応する潜在変数を $\boldsymbol{z}_i$、また $\boldsymbol{z}_i$ の k 番目の要素を z_{ik} とする
- データ点 $\boldsymbol{x}_i, \boldsymbol{z}_i$ は、$p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ から独立にサンプルされているとする (このとき、要素ごとの積として書ける)
- 対数を取ると次のようになる

$$
\begin{aligned}
& \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \\
= & \ln \left(\prod_i \prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)^{z_{ik}} \right) \\
= & \sum_i \sum_k \ln \left(\left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)^{z_{ik}} \right) \\
= & \sum_i \sum_k z_{ik} \ln \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \\
= & \sum_i \sum_k z_{ik} \left(\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \qquad (105)
\end{aligned}
$$

混合ガウス分布の再解釈

- $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ を、元々最大化しようとしていた $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ と比較する

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_i \ln \left(\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \tag{106}$$

- $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ と $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ を比較すると、対数 $\ln$ と、総和 $\sum_k$ の、順番が入れ替わっている
- そして、対数 $\ln$ が、ガウス分布 $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ に直接作用している
- よって、$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の最大化は、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の最大化よりも、遥かに容易である (そして、パラメータは陽な形で解ける)
- そこで、$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ を最大化するようなパラメータを求めてみる

混合ガウス分布の再解釈

- $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の $\boldsymbol{\mu}_k$ に関する最大化
  - 以下のように、$\boldsymbol{\mu}_k$ で微分して 0 とおけば、簡単に解ける
  - ガウス分布の微分については、先程の EM アルゴリズムの導出時に求めたものを利用している

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\mu}_k} \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \\
= & \frac{\partial}{\partial \boldsymbol{\mu}_k} \left(\sum_i \sum_k z_{ik} \left(\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \right) \\
= & \sum_i \frac{\partial}{\partial \boldsymbol{\mu}_k} \left(\sum_k z_{ik} \left(\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \right) \\
= & \sum_i \frac{\partial}{\partial \boldsymbol{\mu}_k} z_{ik} \left(\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \\
= & \sum_i z_{ik} \frac{\partial}{\partial \boldsymbol{\mu}_k} \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)
\end{aligned}
$$

混合ガウス分布の再解釈

$$
\begin{aligned}
&= \sum_i z_{ik} \frac{\partial}{\partial \boldsymbol{\mu}_k} \ln \left(\frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \right. \\
&\qquad \left. \exp\left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} \right) \\
&= \sum_i z_{ik} \frac{\partial}{\partial \boldsymbol{\mu}_k} \left(-\frac{D}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_k| \right. \\
&\qquad \left. -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right) \\
&= \sum_i z_{ik} \frac{\partial}{\partial \boldsymbol{\mu}_k} \left(-\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right) \\
&= \sum_i z_{ik} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) = 0 \qquad (107)
\end{aligned}
$$

これより

$$
\sum_i z_{ik} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k = \sum_i z_{ik} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{x}_i \qquad (108)
$$

混合ガウス分布の再解釈

であるから、両辺に左から $\boldsymbol{\Sigma}_k$ を掛けて

$$\sum_i z_{ik}\boldsymbol{\mu}_k = \sum_i z_{ik}\boldsymbol{x}_i$$

$$\boldsymbol{\mu}_k \sum_i z_{ik} = \sum_i z_{ik}\boldsymbol{x}_i$$

$$\boldsymbol{\mu}_k = \frac{1}{\sum_i z_{ik}} \sum_i z_{ik}\boldsymbol{x}_i \tag{109}$$

のようになる

- 上式をみると、完全データ $\{\boldsymbol{X}, \boldsymbol{Z}\}$ について、$\boldsymbol{\mu}_k$ は陽な形で求まっていることが分かる
- 但し実際は $\boldsymbol{Z}$ が分からないので、z_{ik} をどうにかして得る必要がある

混合ガウス分布の再解釈

- $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の $\boldsymbol{\Sigma}_k$ に関する最大化
  - $\boldsymbol{\Sigma}_k$ について微分して 0 とおくと、次のようになる

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \\
= & \sum_i z_{ik} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
= & \sum_i z_{ik} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \left(-\frac{1}{2} \ln |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right) \\
= & \sum_i z_{ik} \left(-\frac{1}{2} \left(\boldsymbol{\Sigma}_k^{-1} \right)^T + \frac{1}{2} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \right) \\
= & \frac{1}{2} \sum_i z_{ik} \left(-\boldsymbol{\Sigma}_k^{-1} + \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \right) \qquad (110) \\
= & 0
\end{aligned}
$$

となる

混合ガウス分布の再解釈

- ここで、以下の微分公式を用いた

$$\frac{\partial}{\partial \boldsymbol{X}} \ln |\boldsymbol{X}| = \left(\boldsymbol{X}^{-1}\right)^T \tag{111}$$

- これより

$$\sum_i z_{ik} \boldsymbol{\Sigma}_k^{-1} = \sum_i z_{ik} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \tag{112}$$

であるから、両辺に左右から $\boldsymbol{\Sigma}_k$ を掛けて

$$\sum_i z_{ik} \boldsymbol{\Sigma}_k = \sum_i z_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T$$

$$\boldsymbol{\Sigma}_k \sum_i z_{ik} = \sum_i z_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T$$

$$\boldsymbol{\Sigma}_k = \frac{1}{\sum_i z_{ik}} \sum_i z_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \tag{113}$$

のようになる

混合ガウス分布の再解釈

- 上式をみても、やはり、完全データ $\{\boldsymbol{X}, \boldsymbol{Z}\}$ について、$\boldsymbol{\Sigma}_k$ は陽な形で求まっていることが分かる

混合ガウス分布の再解釈

- $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の π_k に関する最大化
  - $\sum_k pi_k = 1$ という制約条件を考慮し、ラグランジュの未定乗数法で解く
  - 従って、以下の量を最大化する

$$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) + \lambda \left(\sum_k \pi_k - 1 \right) \tag{114}$$

  - π_k について微分して 0 とおくと、次のようになる

$$\begin{aligned} & \frac{\partial}{\partial \pi_k} \left(\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) + \lambda \left(\sum_k \pi_k - 1 \right) \right) \\ = & \frac{\partial}{\partial \pi_k} \left(\sum_i \sum_k z_{ik} \left(\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) + \lambda \left(\sum_k \pi_k - 1 \right) \right) \\ = & \sum_i z_{ik} \frac{\partial}{\partial \pi_k} \ln \pi_k + \lambda \end{aligned}$$

混合ガウス分布の再解釈

$$= \sum_i z_{ik} \frac{1}{\pi_k} + \lambda = 0 \tag{115}$$

- これより、両辺に π_k を掛けて

$$\sum_i z_{ik} + \lambda \pi_k = 0 \tag{116}$$

全ての k について総和を取ると

$$\begin{aligned}
&\sum_k \sum_i z_{ik} + \sum_k \lambda \pi_k = 0 \\
&\sum_i \left(\sum_k z_{ik} \right) + \lambda \sum_k \pi_k = 0 \\
&\sum_i 1 + \lambda = 0 \\
&N + \lambda = 0 \\
&\therefore \lambda = -N
\end{aligned} \tag{117}$$

混合ガウス分布の再解釈

- よって

$$\sum_i z_{ik} \frac{1}{\pi_k} - N = 0$$

$$\sum_i z_{ik} - N\pi_k = 0$$

$$N\pi_k = \sum_i z_{ik}$$

$$\therefore \pi_k = \frac{1}{N} \sum_i z_{ik} \tag{118}$$

- π_k も、完全データ (特に潜在変数) が与えられていれば、陽な形で求まる
- EM アルゴリズムにおける $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ の更新式は、ここで求めた式の z_{ik} を、負担率 $\gamma(z_{ik})$ にそのまま置き換えたものである

混合ガウス分布の再解釈

- 事後確率分布 $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$ に関する期待値の計算
  - 完全データ対数尤度関数 $\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$ の最大化は、陽な形で解けた
  - これらの全ての式には z_{ik} が登場したが、実際には潜在変数は分からないので、z_{ik} を何かで代用しなければならない

  - 結局、完全データ対数尤度関数 $\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$ の、事後確率分布 $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$ に関する期待値を考えるしかない

混合ガウス分布の再解釈

- 事後確率分布は次のように書ける

$$
\begin{aligned}
& p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \\
= & \frac{p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) p(\boldsymbol{z}_i|\boldsymbol{\theta})}{p(\boldsymbol{x}_i|\boldsymbol{\theta})} && (119) \\
\propto & p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) p(\boldsymbol{z}_i|\boldsymbol{\theta}) && (120) \\
& \qquad (\because p(\boldsymbol{x}_i|\boldsymbol{\theta}) \text{ は、} \boldsymbol{z}_i \text{ には依存しない定数項}) \\
= & \left(\prod_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{ik}} \right) \left(\prod_k \pi_k^{z_{ik}} \right) \\
= & \prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)^{z_{ik}} && (121)
\end{aligned}
$$

以上より

$$
p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \propto \prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right)^{z_{ik}} \tag{122}
$$

混合ガウス分布の再解釈

であるので、$p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$ は

$$p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}) \propto \prod_i \prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right)^{z_{ik}} \tag{123}$$

- $p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta})$ を等式で表すためには、$\boldsymbol{z}_i$ で総和を取って 1 になる (確率としての条件を満たす) ように、正規化すればよい

$$p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta}) = \frac{\prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right)^{z_{ik}}}{\sum_{\boldsymbol{z}_i} \prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right)^{z_{ik}}} \tag{124}$$

混合ガウス分布の再解釈

- まずは、事後確率 $p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta})$ に関する、z_{ik} の期待値を求めてみる

$$\begin{aligned} & \mathbb{E}_{\boldsymbol{z}_i \sim p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta})}\left[z_{ik}\right] \\ = & \sum_{\boldsymbol{z}_i} z_{ik} p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \\ = & \sum_{\boldsymbol{z}_i} z_{ik} \frac{\prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_{ik}}}{\sum_{\boldsymbol{z}_i} \prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_{ik}}} \end{aligned} \tag{125}$$

- ここで

$$\sum_{\boldsymbol{z}_i} \prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_{ik}} = \sum_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \tag{126}$$

  と書けることに注意する

- $\boldsymbol{z}_i$ は、1-of-K 符号化法で表現されている

混合ガウス分布の再解釈

- $\prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_{ik}}$ は、$z_{ik} = 1$ の場合、$j \neq k$ に対して $z_{ij} = 0$ であるから、$\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ という単一の項として書ける
- 全ての $\boldsymbol{z}_i$ についての総和は、$\boldsymbol{z}_i$ の中で、要素が 1 になるインデックス k についての総和を意味する

- また

$$\sum_{\boldsymbol{z}_i} z_{ik} \prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_{ik}} = \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{127}$$

  であることにも注意する

- $\sum_{\boldsymbol{z}_i}$ の総和の中身は、$\boldsymbol{z}_i$ が $z_{ik} = 1$ となるとき以外は、0 である (総和の中に z_{ik} があるため)
- 従って、$\boldsymbol{z}_i$ が $z_{ik} = 1$ となるときの項 $\prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_{ik}}$、即ち $\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ だけが出現する

混合ガウス分布の再解釈

- これより

$$
\begin{aligned}
& \mathbb{E}_{\boldsymbol{z}_i \sim p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta})}(z_{ik}) \\
= & \frac{\sum_{\boldsymbol{z}_i} z_{ik} \prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_{ik}}}{\sum_{\boldsymbol{z}_i} \prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_{ik}}} \\
= & \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \equiv \gamma(z_{ik}) \qquad (128)
\end{aligned}
$$

であるから、データ点 $\boldsymbol{x}_i$ に対する、k 番目のガウス要素の負担率に一致

混合ガウス分布の再解釈

- これより、事後確率 $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$ に関する、完全データ対数尤度関数 $\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$ の期待値は

$$
\begin{aligned}
& \mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}\left[\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})\right] \\
= \quad & \mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}\left[\sum_i\sum_k z_{ik}\left(\ln\pi_k + \ln\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right)\right] \\
= \quad & \sum_i\sum_k \mathbb{E}_{\boldsymbol{z}_i\sim p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta})}\left[z_{ik}\left(\ln\pi_k + \ln\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right)\right] \quad (129) \\
= \quad & \sum_i\sum_k \mathbb{E}_{\boldsymbol{z}_i\sim p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta})}\left[z_{ik}\right]\left(\ln\pi_k + \ln\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right) \quad (130) \\
= \quad & \sum_i\sum_k \gamma(z_{ik})\left(\ln\pi_k + \ln\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right) \quad (131)
\end{aligned}
$$

である

混合ガウス分布の再解釈

- これは $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ において、z_{ik} を $\gamma(z_{ik})$ に置き換えたものと等しい

$$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) = \sum_i \sum_k z_{ik} \left(\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \tag{132}$$

- 先ほどは、$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ を最大化するような、パラメータ $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ の式を導出した
- これらの式について、z_{ik} を $\gamma(z_{ik})$ に置き換えれば、そのまま期待値を最大化する式として使える

- $\gamma(z_{ik})$ に置き換えた式は、EM アルゴリズムにおける更新式と一致

混合ガウス分布の再解釈

- ここまでの話の流れ
  1. 対数尤度関数 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の最大化よりも、完全データ対数尤度関数 $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の最大化の方が簡単であると仮定した
  2. この仮定は、混合ガウス分布の場合について成り立っていた
  3. $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の代わりに、$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の最大化を考えた
  4. しかし $\boldsymbol{Z}$ に関する情報がないので、代わりに、事後確率分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ による、$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の期待値を最大化しようと考えるのが、EM アルゴリズムであった
  5. 混合ガウス分布の場合について実際に試すと、期待値の最大化によって、パラメータの更新式が再び導出できた
- これからの話の流れ

混合ガウス分布の再解釈

- K-Means 法と、混合ガウス分布に対する EM アルゴリズムを比較する

K-Means 法との関連

- K-Means 法と、混合ガウス分布に対する EM アルゴリズムの関係
  - K-Means 法では、各データ点は、ただ一つのクラスタに割り当てられる (ハード割り当て)
  - EM アルゴリズムでは、事後確率 $\gamma(z_{ik}) \equiv p(z_k = 1|\boldsymbol{x}_i)$ に基づいて、各データをソフトに割り当てる (ソフト割り当て)

  - K-Means 法は、混合ガウス分布に対する EM アルゴリズムの、ある極限として得られる

K-Means 法との関連

- K-Means 法の導出
  - 次のように、各ガウス分布の共分散行列が $\epsilon \boldsymbol{I}$ で与えられる、混合ガウスモデル $p(\boldsymbol{x}|\boldsymbol{\theta})$ を考える (ϵ は定数とする)

$$
\begin{aligned}
& p(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
= \; & p(\boldsymbol{x}|\boldsymbol{\mu}_k, \epsilon \boldsymbol{I}) \\
= \; & \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \epsilon \boldsymbol{I}) \\
= \; & \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\epsilon \boldsymbol{I}|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_k)^T (\epsilon \boldsymbol{I})^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_k) \right\} \quad (133) \\
= \; & \frac{1}{(2\pi\epsilon)^{\frac{D}{2}}} \exp\left\{ -\frac{1}{2\epsilon} (\boldsymbol{x} - \boldsymbol{\mu}_k)^T (\boldsymbol{x} - \boldsymbol{\mu}_k) \right\} \quad (134) \\
& \qquad \left(\because |\epsilon \boldsymbol{I}|^{\frac{1}{2}} = \left(\epsilon^D |\boldsymbol{I}| \right)^{\frac{1}{2}} = \left(\epsilon^D \right)^{\frac{1}{2}} = \epsilon^{\frac{D}{2}} \right) \\
= \; & \frac{1}{(2\pi\epsilon)^{\frac{D}{2}}} \exp\left\{ -\frac{1}{2\epsilon} ||\boldsymbol{x} - \boldsymbol{\mu}_k||^2 \right\} \quad (135)
\end{aligned}
$$

K-Means 法との関連

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\theta}) &= \sum_k \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) && (136)\\
&= \sum_k \pi_k \frac{1}{(2\pi\epsilon)^{\frac{D}{2}}} \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x}-\boldsymbol{\mu}_k||^2\right\} && (137)
\end{aligned}
$$

- この混合ガウスモデルについて、EM アルゴリズムを実行する
- 最初に、データ点 $\boldsymbol{x}_i$ に対する、k 番目のガウス要素の負担率 $\gamma(z_{ik})$ を求めて、$\epsilon \to 0$ についての極限を取ってみる

$$
\begin{aligned}
\gamma(z_{ik}) &= \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}\\
&= \frac{\pi_k \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x}-\boldsymbol{\mu}_k||^2\right\}}{\sum_j \pi_j \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x}-\boldsymbol{\mu}_j||^2\right\}} && (138)
\end{aligned}
$$

K-Means 法との関連

- 負担率は、以下のように変形できる

$$
\begin{aligned}
& \frac{\pi_k \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x}-\boldsymbol{\mu}_k||^2\right\}}{\sum_j \pi_j \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x}-\boldsymbol{\mu}_j||^2\right\}} \\
= & \left(\frac{\sum_j \pi_j \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x}-\boldsymbol{\mu}_j||^2\right\}}{\pi_k \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x}-\boldsymbol{\mu}_k||^2\right\}}\right)^{-1} \\
= & \left(\sum_j \frac{\pi_j}{\pi_k} \frac{\left(\exp\left\{-||\boldsymbol{x}-\boldsymbol{\mu}_j||^2\right\}\right)^{\frac{1}{2\epsilon}}}{\left(\exp\left\{-||\boldsymbol{x}-\boldsymbol{\mu}_k||^2\right\}\right)^{\frac{1}{2\epsilon}}}\right)^{-1} \\
= & \left(\sum_j \frac{\pi_j}{\pi_k} \left(\frac{\exp\left\{-||\boldsymbol{x}-\boldsymbol{\mu}_j||^2\right\}}{\exp\left\{-||\boldsymbol{x}-\boldsymbol{\mu}_k||^2\right\}}\right)^{\frac{1}{2\epsilon}}\right)^{-1} \\
= & \left(1+\sum_{j\neq k} \frac{\pi_j}{\pi_k} \left(\frac{\exp\left\{-||\boldsymbol{x}-\boldsymbol{\mu}_j||^2\right\}}{\exp\left\{-||\boldsymbol{x}-\boldsymbol{\mu}_k||^2\right\}}\right)^{\frac{1}{2\epsilon}}\right)^{-1} \qquad (139)
\end{aligned}
$$

K-Means 法との関連

- ここで、k^* を次で定める

$$k^* = \arg\min_j ||\boldsymbol{x} - \boldsymbol{\mu}_j||^2 = \arg\max_j \left(-||\boldsymbol{x} - \boldsymbol{\mu}_j||^2\right) \tag{140}$$

$k = k^*$ であるとき、以下の、$\epsilon \to 0$ による極限

$$\lim_{\epsilon \to 0} \left(\sum_{j \neq k} \frac{\pi_j}{\pi_k} \left(\frac{\exp\left\{-||\boldsymbol{x} - \boldsymbol{\mu}_j||^2\right\}}{\exp\left\{-||\boldsymbol{x} - \boldsymbol{\mu}_k||^2\right\}} \right)^{\frac{1}{2\epsilon}} \right) \tag{141}$$

を考えると、全ての $j \neq k^*$ について

$$\frac{\exp\left\{-||\boldsymbol{x} - \boldsymbol{\mu}_j||^2\right\}}{\exp\left\{-||\boldsymbol{x} - \boldsymbol{\mu}_k||^2\right\}} < 1 \tag{142}$$

K-Means 法との関連

が成立するので

$$\lim_{\epsilon \to 0} \left(\sum_{j \neq k} \frac{\pi_j}{\pi_k} \left(\frac{\exp\left\{-||\boldsymbol{x} - \boldsymbol{\mu}_j||^2\right\}}{\exp\left\{-||\boldsymbol{x} - \boldsymbol{\mu}_k||^2\right\}} \right)^{\frac{1}{2\epsilon}} \right) = 0 \tag{143}$$

である

- 従って、$k = k^*$ のとき

$$\begin{aligned} & \lim_{\epsilon \to 0} \gamma(z_{ik}) \\ = & \lim_{\epsilon \to 0} \frac{\pi_k \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x} - \boldsymbol{\mu}_k||^2\right\}}{\sum_j \pi_j \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x} - \boldsymbol{\mu}_j||^2\right\}} \\ = & \lim_{\epsilon \to 0} \left(1 + \sum_{j \neq k} \frac{\pi_j}{\pi_k} \left(\frac{\exp\left\{-||\boldsymbol{x} - \boldsymbol{\mu}_j||^2\right\}}{\exp\left\{-||\boldsymbol{x} - \boldsymbol{\mu}_k||^2\right\}} \right)^{\frac{1}{2\epsilon}} \right)^{-1} \end{aligned}$$

$$= \quad (1+0)^{-1} = 1 \tag{144}$$

から、$\gamma(z_{ik}) \to 1 \ (\epsilon \to 0)$ がいえる

- $k \neq k^*$ のとき

$$1 = \sum_k \gamma(z_{ik}) = \gamma(z_{ik^*}) + \sum_{k \neq k^*} \gamma(z_{ik}) \tag{145}$$

であって、両辺の $\epsilon \to 0$ による極限を取れば

$$1 = \lim_{\epsilon \to 0} \left(\gamma(z_{ik^*}) + \sum_{k \neq k^*} \gamma(z_{ik}) \right)$$

$$\Rightarrow \quad 1 = \lim_{\epsilon \to 0} \gamma(z_{ik^*}) + \lim_{\epsilon \to 0} \sum_{k \neq k^*} \gamma(z_{ik})$$

$$\Rightarrow \quad 1 = 1 + \lim_{\epsilon \to 0} \sum_{k \neq k^*} \gamma(z_{ik}) \tag{146}$$

K-Means 法との関連

となるから

$$\lim_{\epsilon \to 0} \sum_{k \neq k^*} \gamma(z_{ik}) = 0 \tag{147}$$

が明らかに成立するほか、以下の不等式が

$$0 \leq \gamma(z_{ik}) \leq \sum_{k \neq k^*} \gamma(z_{ik}) \tag{148}$$

$\gamma(z_{ik}) \geq 0$ ゆえ成立するので ($\gamma(z_{ik})$ は確率値)、両辺の $\epsilon \to 0$ による極限を再び取れば

$$0 \leq \lim_{\epsilon \to 0} \gamma(z_{ik}) \leq \lim_{\epsilon \to 0} \sum_{k \neq k^*} \gamma(z_{ik}) = 0 \tag{149}$$

従って、$k \neq k^*$ の場合は

$$\lim_{\epsilon \to 0} \gamma(z_{ik}) = 0 \tag{150}$$

である

K-Means 法との関連

- これより、データ点 $\boldsymbol{x}_i$ に関する負担率 $\gamma(z_{ik})$ は、1 に収束する k^* 番目の負担率 $\gamma(z_{ik^*})$ を除き、全て 0 に収束する

$$\gamma(z_{ik}) \equiv p(z_{ik} = 1|\boldsymbol{x}_i) = \begin{cases} 1 & (k = k^* \text{の場合}) \\ 0 & (\text{それ以外の場合}) \end{cases} \tag{151}$$

- これは、k^* 番目のクラスタに確率 1 で属するということ、即ち、クラスタ k^* へのハード割り当てを意味する

- $k^* = \arg\min_j ||\boldsymbol{x} - \boldsymbol{\mu}_j||^2$ であるから、結局、各データ点は、平均ベクトル $\boldsymbol{\mu}$ への二乗ユークリッド距離が最小となるクラスタに割り当てることになる

K-Means 法との関連

- $\gamma(z_{ik})$ を r_{ik} に置き換えれば、EM アルゴリズムにおける $\boldsymbol{\mu}_k$ の更新式は、K-Means における平均ベクトルの更新式に帰着

$$\text{K-Means}: \qquad \boldsymbol{\mu}_k = \frac{1}{\sum_i r_{ik}} \sum_i r_{ik} \boldsymbol{x}_i \tag{152}$$

$$\text{EM アルゴリズム}: \qquad \boldsymbol{\mu}_k = \frac{1}{\sum_i \gamma(z_{ik})} \sum_i \gamma(z_{ik}) \boldsymbol{x}_i \tag{153}$$

- 従って、混合ガウスモデルの EM アルゴリズムにおいて、各ガウス分布の共分散行列を $\epsilon \boldsymbol{I}$ としたとき、$\epsilon \to 0$ の極限を取ると、K-Means 法が得られる

K-Means 法との関連

- 期待完全データ対数尤度の計算
  - $\mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}\left[\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})\right]$ を計算する
  - 完全データ対数尤度関数 $\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})$ の、事後確率 $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$ による期待値

  - 次のように計算する

$$
\begin{aligned}
& \mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}\left[\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})\right] \\
= & \sum_i \sum_k \gamma(z_{ik}) \left(\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right) \\
= & \sum_i \sum_k \gamma(z_{ik}) \left(\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\epsilon\boldsymbol{I})\right) \\
= & \sum_i \sum_k \gamma(z_{ik}) \left(\ln \pi_k + \ln \left(\frac{1}{(2\pi\epsilon)^{\frac{D}{2}}} \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x}_i-\boldsymbol{\mu}_k||^2\right\}\right)\right) \\
= & \sum_i \sum_k \gamma(z_{ik}) \left(\ln \pi_k - \frac{D}{2}\ln(2\pi\epsilon) - \frac{1}{2\epsilon}||\boldsymbol{x}_i-\boldsymbol{\mu}_k||^2\right) \qquad (154)
\end{aligned}
$$

K-Means 法との関連

- 両辺に ϵ を掛けると

$$\begin{aligned} & \epsilon \cdot \mathbb{E}_{\boldsymbol{Z} \sim p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})} \left[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \right] \\ = & \sum_i \sum_k \gamma(z_{ik}) \left(\epsilon \ln \pi_k - \right. \\ & \left. \frac{D}{2} \epsilon \ln(2\pi\epsilon) - \frac{1}{2} ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2 \right) \end{aligned} \tag{155}$$

- $\epsilon \to 0$ の極限を取ると

$$\gamma(z_{ik}) \to r_{ik}, \quad \epsilon \ln \pi_k \to 0, \quad \epsilon \ln(2\pi\epsilon) \to 0 \tag{156}$$

であるから

$$\begin{aligned} & \lim_{\epsilon \to 0} \epsilon \cdot \mathbb{E}_{\boldsymbol{Z} \sim p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})} \left[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \right] \\ = & \sum_i \sum_k r_{ik} \left(-\frac{1}{2} ||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2 \right) \end{aligned}$$

K-Means 法との関連

$$= -\frac{1}{2}\sum_i\sum_k r_{ik}||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2 \tag{157}$$

$$= -J \tag{158}$$

- よって、期待完全データ対数尤度 $\mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})\right]$ の最大化は、K-Means における目的関数 J の最小化と同等である

K-Means 法との関連

- その他のパラメータ
  - K-Means 法では、各クラスタの分散は推定しない
  - 実際に、混合ガウスモデルにおいて、各クラスタの共分散行列は $\epsilon \boldsymbol{I}$ で固定した
  - 混合ガウスモデルの混合係数 π_k の更新式は、次のようであった

$$\pi_k = \frac{\sum_i \gamma(z_{ik})}{N} \tag{159}$$

$\epsilon \to 0$ の極限においては、$\gamma(z_{ik}) \to r_{ik}$ であるから

$$\pi_k = \frac{\sum_i r_{ik}}{N} = \frac{N_k}{N} \tag{160}$$

  - これは、π_k の値を、k 番目のクラスタに割り当てられる、データ数の割合に設定することを意味している
  - π_k の値は K-Means 法においては、もはや何の意味も持たない

K-Means 法との関連

- ここまでの話の流れ
  1. K-Means 法は、混合ガウス分布に対する EM アルゴリズムの、ある極限として得られることが分かった

- これからの話の流れ
  - $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の代わりに、$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ を最大化してもよい根拠を明らかにする
  - $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の、$p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ による期待値を取る理由を明らかにする
  - E ステップと M ステップが、確かに対数尤度関数 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ を増加させることを証明する
  - これらの解明のために、一般的な EM アルゴリズムの取り扱いについて調べる

一般の EM アルゴリズム

- EM アルゴリズムの目的 (再掲)
  - 潜在変数をもつ確率モデルについて、パラメータの最尤解を求める
- 一般的な EM アルゴリズムの取り扱い
  - これまでは、混合ガウスモデルに対して、EM アルゴリズムを発見的に導いた
  - ここでは、EM アルゴリズムが、確かに尤度関数 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ を極大化することを証明する
  - 後述する変分推論の基礎をなす部分
- 尤度関数 $p(\boldsymbol{X}|\boldsymbol{\theta})$ の記述
  - 全ての観測変数と、潜在変数をそれぞれ $\boldsymbol{X}, \boldsymbol{Z}$ と表す
  - 確率モデルの全てのパラメータの組を、$\boldsymbol{\theta}$ と表す

一般の EM アルゴリズム

- 同時確率分布を $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ とすると、尤度関数は次のようになる

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \tag{161}$$

- 連続潜在変数の場合は、次のように、総和を積分に置き換えればよい

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \int_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) d\boldsymbol{Z} \tag{162}$$

- ここでは、連続潜在変数の場合を考える

- 重要な仮定
  - $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の最大化よりも、完全データ対数尤度関数 $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の最大化の方が、容易である
  - 以前に見た尤度関数 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ では、対数の中に総和が含まれており (log-sum)、複雑な形をしていた

一般の EM アルゴリズム

- $\boldsymbol{Z}$ についての情報を加えることで、尤度関数から log-sum の構造を消すことができた
- 対数 $\ln$ がガウス分布に直接作用するようになったため、尤度関数の形が簡単になった

- EM アルゴリズムで行うこと
  - $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ ではなく $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ を最適化しようとしたが、$\boldsymbol{Z}$ に関する情報がないので、それはできない
  - そこで、事後確率 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ による、$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の期待値 $\mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})\right]$ を最大化する

  - これ以降の議論のために、イェンセンの不等式、エントロピー、KL ダイバージェンスについて確認しておく

一般の EM アルゴリズム

- イェンセンの不等式
  - 凸関数 $f(x)$ は、任意の点集合 $\{x_i\}$ について以下を満たす

$$f\left(\sum_i \lambda_i x_i\right) \leq \sum_i \lambda_i f(x_i) \tag{163}$$

  - ここで、$\lambda_i \geq 0,\ \sum_i \lambda_i = 1$ であるとする
  - λ_i を、値 $\{x_i\}$ を取る離散確率変数 x 上の確率分布 $p(x)$ と解釈すると

$$\begin{aligned} f\left(\sum_i p(x_i) x_i\right) &\leq \sum_i p(x_i) f(x_i) \\ f\left(\mathbb{E}[x]\right) &\leq \mathbb{E}[f(x)] \end{aligned} \tag{164}$$

一般の EM アルゴリズム

- x が連続変数であれば、イェンセンの不等式は次のように書ける

$$f\left(\int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x}\right) \leq \int f(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \tag{165}$$

例えば、$f(x) = -\ln x$ は凸関数であるから

$$-\ln\left(\int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x}\right) \leq \int (-\ln \boldsymbol{x})\, p(\boldsymbol{x}) d\boldsymbol{x} \tag{166}$$

よって

$$\ln\left(\int \boldsymbol{x} p(\boldsymbol{x}) d\boldsymbol{x}\right) \geq \int (\ln \boldsymbol{x})\, p(\boldsymbol{x}) d\boldsymbol{x} \tag{167}$$

一般の EM アルゴリズム

- エントロピー
  - 確率分布 $p(\boldsymbol{x})$ について、エントロピーは以下で定義される

$$\mathrm{H}[p] = -\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x} \tag{168}$$

  - エントロピーは、確率分布 $p(\boldsymbol{x})$ を入力として、上記の量を返す、汎関数 (Functional) である
  - 汎関数とは、入力として関数をとり、出力として汎関数の値を返すものである

一般の EM アルゴリズム

- KL ダイバージェンス
  - 確率分布 $p(\boldsymbol{x})$ と $q(\boldsymbol{x})$ の間の、カルバック-ライブラーダイバージェンスを、$\mathrm{KL}(p||q)$ と表す
  - 確率分布 $p(\boldsymbol{x})$ と $q(\boldsymbol{x})$ の間の、(擬似的な) 距離を表す指標である

$$\mathrm{KL}(p||q) = -\int p(\boldsymbol{x}) \ln \left\{ \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} \right\} d\boldsymbol{x} \tag{169}$$

  - $\mathrm{KL}(p||q) \geq 0$ であり、等号成立は $p(\boldsymbol{x}) = q(\boldsymbol{x})$ のときに限る
  - 2 つの分布が完全に同一であれば、KL ダイバージェンスは 0 で最小値を取る

  - また厳密には距離ではないため、対称性は成立しない
  - 従って、一般に $\mathrm{KL}(p||q) \neq \mathrm{KL}(q||p)$ となる

一般の EM アルゴリズム

- $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の分解
  - EM アルゴリズムについて考察するために、まずは $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ を分解してみよう

  - 潜在変数についての分布を $q(\boldsymbol{Z})$ とおく
  - $q(\boldsymbol{Z})$ の設定の仕方によらず、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ を次のように分解できる

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q||p) \tag{170}$$

  - $\mathcal{L}(q, \boldsymbol{\theta})$ は、分布 $q(\boldsymbol{Z})$ の汎関数であり、かつパラメータ $\boldsymbol{\theta}$ の関数である
  - $\mathrm{KL}(q||p)$ は、確率分布 $q(\boldsymbol{Z})$ と $p(\boldsymbol{X}|\boldsymbol{\theta})$ の間の、KL ダイバージェンスである

一般の EM アルゴリズム

- 分解は次のように行える

$$
\begin{aligned}
\ln p(\boldsymbol{X}|\boldsymbol{\theta}) &= \underbrace{\left(\sum_{\boldsymbol{Z}} q(\boldsymbol{Z})\right)}_{=1} \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \\
&= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \\
&= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})} \\
&= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left(\frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}\right) \\
&= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
&= \mathcal{L}(q,\boldsymbol{\theta}) + \mathrm{KL}(q||p) \qquad (171)
\end{aligned}
$$

一般の EM アルゴリズム

- ここで $\mathcal{L}(q, \theta)$ と $\mathrm{KL}(q||p)$ は以下のように定義した

$$\mathcal{L}(q, \boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \tag{172}$$

$$\mathrm{KL}(q||p) = -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \tag{173}$$

- $\mathrm{KL}(q||p) \geq 0$ ゆえ、以下の不等式を得る

$$\mathcal{L}(q, \boldsymbol{\theta}) \leq \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \tag{174}$$

- $\mathcal{L}(q, \boldsymbol{\theta})$ は、$q(\boldsymbol{Z}), \boldsymbol{\theta}$ によらず、常に $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の下界をなす
- EM アルゴリズムの各ステップについて見ていく

一般の EM アルゴリズム

Figure 9.11 Illustration of the decomposition given by (9.70), which holds for any choice of distribution $q(\mathbf{Z})$. Because the Kullback-Leibler divergence satisfies $\mathrm{KL}(q\|p) \geqslant 0$, we see that the quantity $\mathcal{L}(q, \boldsymbol{\theta})$ is a lower bound on the log likelihood function $\ln p(\mathbf{X}|\boldsymbol{\theta})$.

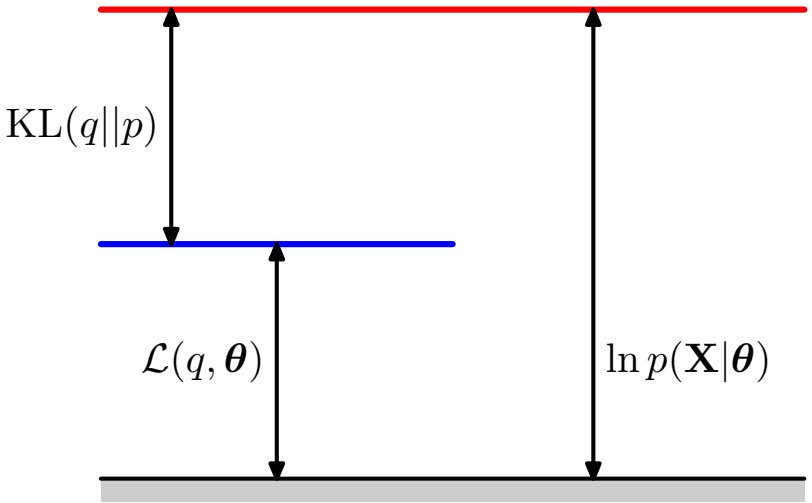

図 4: $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の分解

一般の EM アルゴリズム

- EM アルゴリズムの概要
  - EM アルゴリズムでは、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の最尤解を求めるために、E ステップと M ステップの二段階の処理を、交互に繰り返す
  - パラメータの現在値を $\boldsymbol{\theta}^{\mathrm{old}}$ とする

- E ステップ
  - E ステップでは、下界 $\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})$ を、$\boldsymbol{\theta}^{\mathrm{old}}$ を固定しながら、$q(\boldsymbol{Z})$ について最大化する

  - この問題は、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の分解をみれば簡単に解ける

$$
\begin{aligned}
& \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{old}}) \\
= \quad & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) + \mathrm{KL}(q||p) && (175) \\
= \quad & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) + \mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})) && (176)
\end{aligned}
$$

(E ステップ前)

一般の EM アルゴリズム

- 上式において、左辺の $\ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{old}})$ は、q には依存しない定数である
- 従って、q について $\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})$ を最大化するためには、$\mathrm{KL}(q||p)$ を最小化するしかない
- $\mathrm{KL}(q||p)$ を最小化するためには、$q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})$ とおいて、$\mathrm{KL}(q||p) = 0$ とすればよい ($\mathrm{KL}(q||p) \geq 0$ であるから、最小値は 0)
- このとき、下界 $\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})$ は、対数尤度 $\ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{old}})$ に一致する

$$
\begin{aligned}
& \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{old}}) && (177)\\
= \; & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) + \mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})) \\
= \; & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) + \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})) && (178)\\
= \; & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) && (179)\\
& (\text{E ステップ後})
\end{aligned}
$$

一般の EM アルゴリズム

- 次の図 5 には E ステップの概要が示されている

- $\mathrm{KL}(q||p) = 0$ となるように q を調節している
- 青線で示されている下界 $\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})$ が、赤線で示されている対数尤度 $\ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{old}})$ のところまで、持ち上げられている

一般の EM アルゴリズム

Figure 9.12 Illustration of the E step of the EM algorithm. The q distribution is set equal to the posterior distribution for the current parameter values $\boldsymbol{\theta}^{\text{old}}$, causing the lower bound to move up to the same value as the log likelihood function, with the KL divergence vanishing.

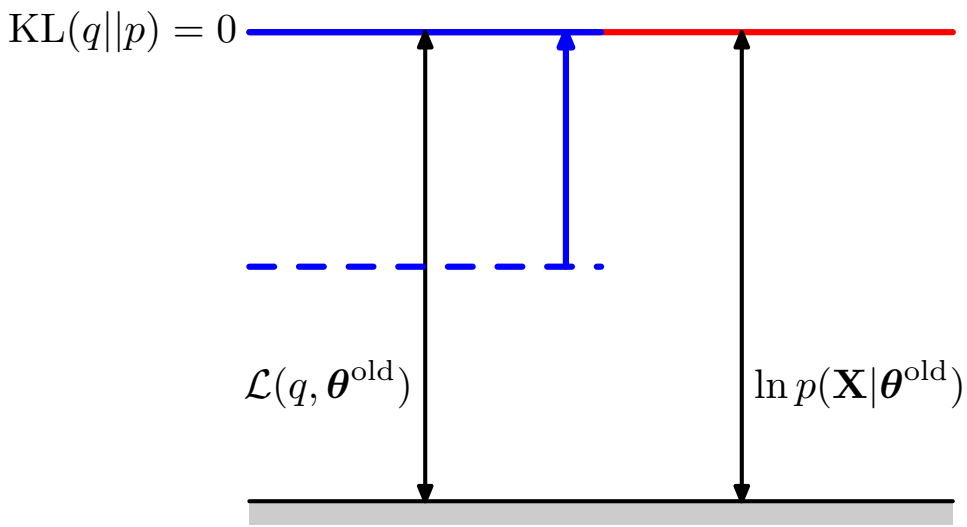

図 5: EM アルゴリズムの E ステップ

一般の EM アルゴリズム

- M ステップ
  - M ステップでは、下界 $\mathcal{L}(q, \boldsymbol{\theta})$ を、分布 $q(\boldsymbol{Z})$ を固定しながら、$\boldsymbol{\theta}$ について最大化し、新たなパラメータ $\boldsymbol{\theta}^{\text{new}}$ を得る
  - M ステップは下界 $\mathcal{L}$ を増加させるが、$\mathrm{KL}(q||p) \geq 0$ であるから、対数尤度 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ も必然的に増加する

$$
\begin{aligned}
& \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \\
= \quad & \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q||p) && (180) \\
= \quad & \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})) && (181) \\
= \quad & \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})) && (182) \\
& (\text{M ステップ前})
\end{aligned}
$$

  - 分布 $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})$ は、古いパラメータ $\boldsymbol{\theta}^{\text{old}}$ によって決められており、M ステップの間は固定されている

一般の EM アルゴリズム

- $\mathrm{KL}(q||p)$ は、$q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})$ と $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ との KL ダイバージェンスである
- $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})$ と、M ステップ後の新しい事後分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{new}})$ とは一致しないため、$\mathrm{KL}(q||p) > 0$ となる

$$\begin{aligned} & \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{new}}) \\ = \quad & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}}) + \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{new}})) \quad (183) \\ & (\text{M ステップ後}) \end{aligned}$$

- 対数尤度の増加量は、下界 $\mathcal{L}$ の増加量よりも大きくなる (図 6)

$$\begin{aligned} & \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{new}}) - \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{old}}) \\ = \quad & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}}) + \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{new}})) - \\ & \qquad \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) \quad (184) \\ = \quad & \left(\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}}) - \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})\right) + \\ & \qquad \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{new}})) \quad (185) \\ \geq \quad & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}}) - \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) \quad (186) \end{aligned}$$

一般の EM アルゴリズム

- 次の図 6 には M ステップの概要が示されている
- 下界 $\mathcal{L}(q, \boldsymbol{\theta})$ を、$q(\boldsymbol{Z})$ を固定しつつ、$\boldsymbol{\theta}$ について最大化している
- 青の点線で示されている下界 $\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})$ が、青の実線で示されている下界 $\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}})$ へと、持ち上げられている
- 赤の点線で示される対数尤度 $\ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{old}})$ は、赤の実線で示される対数尤度 $\ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{new}})$ へと、持ち上げられている
- 新たに生じた $\mathrm{KL}(q||p)$ によって、対数尤度の増加量は、下界 $\mathcal{L}$ の増加量よりも大きくなっている

一般の EM アルゴリズム

Figure 9.13 Illustration of the M step of the EM algorithm. The distribution $q(\mathbf{Z})$ is held fixed and the lower bound $\mathcal{L}(q, \boldsymbol{\theta})$ is maximized with respect to the parameter vector $\boldsymbol{\theta}$ to give a revised value $\boldsymbol{\theta}^{\text{new}}$. Because the KL divergence is nonnegative, this causes the log likelihood $\ln p(\mathbf{X}|\boldsymbol{\theta})$ to increase by at least as much as the lower bound does.

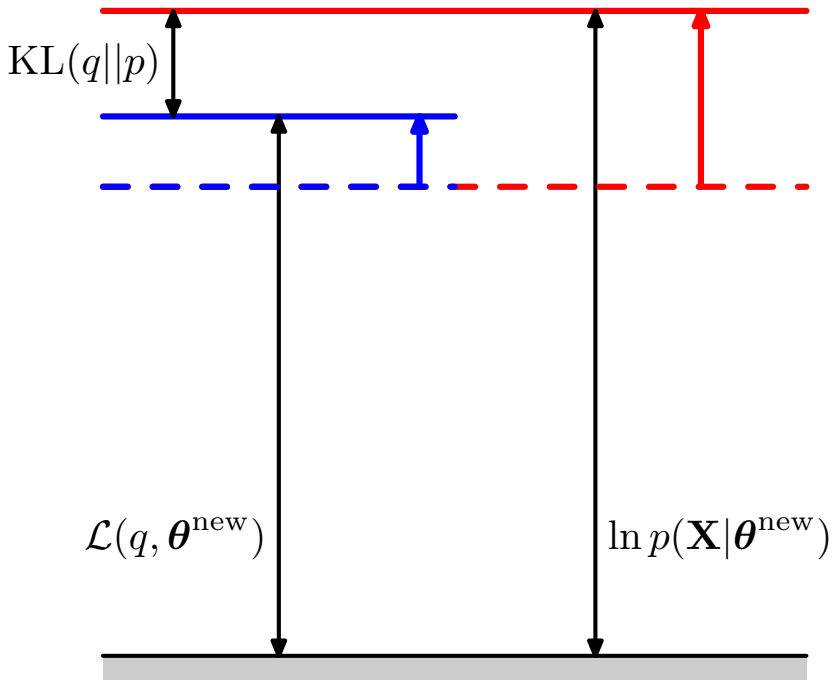

図 6: EM アルゴリズムの M ステップ

一般の EM アルゴリズム

- M ステップで最大化される量
  - M ステップでは下界 $\mathcal{L}(q, \boldsymbol{\theta})$ を、q を固定しつつ $\boldsymbol{\theta}$ について最大化する
  - M ステップで最大化するのは、E ステップ後の下界 $\mathcal{L}(q, \boldsymbol{\theta})$ であり、これは次のように表せる ($q(\boldsymbol{Z}) = p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}})$ である)

$$
\begin{aligned}
& \mathcal{L}(q, \boldsymbol{\theta}) \\
= & \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
= & \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}})} && (187) \\
= & \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) - \\
& \qquad \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}}) && (188) \\
= & \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) + \text{Const.} && (189)
\end{aligned}
$$

一般の EM アルゴリズム

- 定数項は、単に分布 $q(\boldsymbol{Z}) = p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\mathrm{old}})$ のエントロピーであって、$\boldsymbol{\theta}$ には依存しないため無視できる

- M ステップで最大化される量は、結局、完全データ対数尤度関数 $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の、事後確率分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})$ による期待値 $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}})$ である

- 最適化しようとしているパラメータ $\boldsymbol{\theta}$ は、対数の中にしか現れない
- 同時分布 $p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ に対して対数が直接作用するので、同時分布が例えばガウス分布であれば、対数と指数が打ち消されて、簡単な形になる

- その結果として、不完全データ対数尤度関数 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の最適化よりも、非常に単純な手続きとなる

一般の EM アルゴリズム

- E ステップのまとめ
  - 下界 $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ を、$\boldsymbol{\theta}^{\text{old}}$ を固定しつつ、q について最大化する
  - これは、単に $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})$ とすればよい
  - 即ち、$p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})$ を計算するだけである

- M ステップのまとめ
  - 下界 $\mathcal{L}(q, \boldsymbol{\theta})$ を、q を固定しつつ、$\boldsymbol{\theta}$ について最大化する
  - これは、期待値 $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ を最大化するような、パラメータ $\boldsymbol{\theta}$ を求めることに相当

一般の EM アルゴリズム

- 疑問に対する答え
  - $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の代わりに、$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ を最大化してよい根拠
  - そして、$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の、$p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ による期待値を取る理由

  - 期待値を取る操作は、式の導出の中で、極めて自然に現れた
  - 期待値 $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}})$ の最大化は、$\mathcal{L}(q, \boldsymbol{\theta})$ の最大化と等価である
  - $\mathcal{L}(q, \boldsymbol{\theta})$ は、q や $\boldsymbol{\theta}$ によらず、常に $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の下界である
  - 下界を最大化することは、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ を徐々に大きくしていくことにつながる (図 5 と図 6 を参照)

  - これらより、$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の期待値を最適化させることは、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ を最適化させることと等価

一般の EM アルゴリズム

- パラメータの更新によって $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ が常に大きくなることの補足
  - 以下のように式変形を行う

$$
\begin{aligned}
& (\text{M ステップ後の } \ln p(\boldsymbol{X}|\boldsymbol{\theta})) - (\text{E ステップ後の } \ln p(\boldsymbol{X}|\boldsymbol{\theta})) \\
= \quad & \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\text{new}}) - \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\text{old}}) \\
= \quad & \ln \frac{p(\boldsymbol{X}|\boldsymbol{\theta}^{\text{new}})}{p(\boldsymbol{X}|\boldsymbol{\theta}^{\text{old}})} \\
= \quad & \underbrace{\sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})}_{=1} \ln \frac{p(\boldsymbol{X}|\boldsymbol{\theta}^{\text{new}})}{p(\boldsymbol{X}|\boldsymbol{\theta}^{\text{old}})} \\
= \quad & \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{new}})}{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{new}})} \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})}{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}})} \\
= \quad & \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{new}})}{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}})} \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})}{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{new}})}
\end{aligned}
$$

一般の EM アルゴリズム

$$
\begin{aligned}
&= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}^{\text{new}})}{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}})} + \\
&\qquad \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{new}})} \\
&= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}^{\text{new}}) - \\
&\qquad \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}}) - \\
&\qquad \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{new}})} \\
&= \mathcal{Q}(\boldsymbol{\theta}^{\text{old}},\boldsymbol{\theta}^{\text{new}}) - \mathcal{Q}(\boldsymbol{\theta}^{\text{old}},\boldsymbol{\theta}^{\text{old}}) + \\
&\qquad \text{KL}\left(p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}})||p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{new}})\right) \qquad (190) \\
&\geq \mathcal{Q}(\boldsymbol{\theta}^{\text{old}},\boldsymbol{\theta}^{\text{new}}) - \mathcal{Q}(\boldsymbol{\theta}^{\text{old}},\boldsymbol{\theta}^{\text{old}}) \qquad (191) \\
&\geq 0
\end{aligned}
$$

一般の EM アルゴリズム

- 最後の変形は、M ステップでは $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ を、$\boldsymbol{\theta}$ について最大化しているから、$\mathcal{Q}(\boldsymbol{\theta}^{\text{old}}, \boldsymbol{\theta}^{\text{new}}) \geq \mathcal{Q}(\boldsymbol{\theta}^{\text{old}}, \boldsymbol{\theta}^{\text{old}})$ であることを利用

- 更新によって $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ は、収束していない限り常に大きくなる

一般の EM アルゴリズム

- $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の分解の導出の補足
  - イェンセンの不等式を用いて導出してみよう

$$\begin{aligned}
\ln p(\boldsymbol{X}|\boldsymbol{\theta}) &= \ln \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \\
&= \ln \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
&\geq \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
&= \mathcal{L}(q, \boldsymbol{\theta}) \qquad (192)
\end{aligned}$$

  - 不等式の部分でイェンセンの不等式 $\log(\mathbb{E}[x]) \leq \mathbb{E}[\log x]$ を用いた

一般の EM アルゴリズム

- これより、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ と $\mathcal{L}(q, \boldsymbol{\theta})$ の差を調べると

$$
\begin{aligned}
& \ln p(\boldsymbol{X}|\boldsymbol{\theta}) - \mathcal{L}(q, \boldsymbol{\theta}) \\
= \quad & \ln p(\boldsymbol{X}|\boldsymbol{\theta}) - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
= \quad & \underbrace{\sum_{\boldsymbol{Z}} q(\boldsymbol{Z})}_{=1} \ln p(\boldsymbol{X}|\boldsymbol{\theta}) - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
= \quad & \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln p(\boldsymbol{X}|\boldsymbol{\theta}) - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) p(\boldsymbol{X}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
= \quad & \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln p(\boldsymbol{X}|\boldsymbol{\theta}) - \\
& \qquad \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \left(\ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) + \ln p(\boldsymbol{X}|\boldsymbol{\theta}) - \ln q(\boldsymbol{Z}) \right) \\
= \quad & - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \left(\ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) - \ln q(\boldsymbol{Z}) \right)
\end{aligned}
$$

一般の EM アルゴリズム

$$
\begin{aligned}
&= -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
&= \mathrm{KL}(q||p) \qquad (193)
\end{aligned}
$$

ゆえ、$\mathrm{KL}(q||p)$ となることが分かったので

$$
\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q||p) \qquad (194)
$$

のように分解できることが分かる

一般の EM アルゴリズム

- パラメータ空間での図示
  - EM アルゴリズムは、パラメータ空間でも視覚化できる (図 7)
  - 赤の実線は、最大化したい対象である、不完全データ対数尤度関数 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ を表す
- E ステップ
  - パラメータの初期値 $\boldsymbol{\theta}^{\mathrm{old}}$ から始めて、最初の E ステップでは、潜在変数の事後確率分布 $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$ を計算
  - このとき、青の実線で示す下界 $\mathcal{L}(q,\boldsymbol{\theta}^{\mathrm{old}})$ が q について更新され、下界 $\mathcal{L}$ は、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ と $\boldsymbol{\theta}^{\mathrm{old}}$ において一致する
  - 下界 $\mathcal{L}$ の曲線は、$\boldsymbol{\theta}^{\mathrm{old}}$ において $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ と接することに注意する
  - 下界 $\mathcal{L}$ と対数尤度 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ は、$\boldsymbol{\theta}^{\mathrm{old}}$ において同じ勾配を持つ

一般の EM アルゴリズム

- M ステップ
  - 下界 $\mathcal{L}$ が凹関数で、唯一の最大値をもつとする (例えば混合ガウスモデル)
  - M ステップでは、下界 $\mathcal{L}(q, \boldsymbol{\theta})$ が $\boldsymbol{\theta}$ について最大化されて、パラメータ $\boldsymbol{\theta}^{\text{new}}$ が得られる

- 続く E ステップ
  - 続く E ステップでは、緑の実線で示した下界 $\mathcal{L}(q, \boldsymbol{\theta}^{\text{new}})$ が計算される
  - 下界 $\mathcal{L}(q, \boldsymbol{\theta}^{\text{new}})$ は、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ と $\boldsymbol{\theta}^{\text{new}}$ で接する

一般の EM アルゴリズム

- 勾配が等しくなることについての証明
  - 以下の式の、$\boldsymbol{\theta}$ による微分を考えれば明らか

$$\begin{aligned} & \frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}^{\mathrm{old}}} \\ = & \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}^{\mathrm{old}}} + \frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{KL}(q||p) \Big|_{\boldsymbol{\theta}^{\mathrm{old}}} \\ = & \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}^{\mathrm{old}}} \end{aligned} \tag{195}$$

  - E ステップによって $\mathrm{KL}(q||p)$ が最小化されるので、$\boldsymbol{\theta}$ による勾配も当然 0 になるはずである
  - このとき、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ と $\mathcal{L}(q, \boldsymbol{\theta})$ の、$\boldsymbol{\theta}^{\mathrm{old}}$ における微分値が等しくなる
  - 従って、$\boldsymbol{\theta}^{\mathrm{old}}$ において両者は接することが分かる

  - 直感的には、次のように考えればよい

一般の EM アルゴリズム

- 両者が接していなければ、交差しているはずである
- このとき、対数尤度 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ が、下界 $\mathcal{L}$ を上回る $(\mathcal{L}(q, \boldsymbol{\theta}) > \ln p(\boldsymbol{X}|\boldsymbol{\theta}))$ ような $\boldsymbol{\theta}$ が存在する

- これは、$\mathrm{KL}(q||p) < 0$ となる可能性があることを示し、従って有り得ないので、両者は接しているはず

一般の EM アルゴリズム

Figure 9.14 The EM algorithm involves alternately computing a lower bound on the log likelihood for the current parameter values and then maximizing this bound to obtain the new parameter values. See the text for a full discussion.

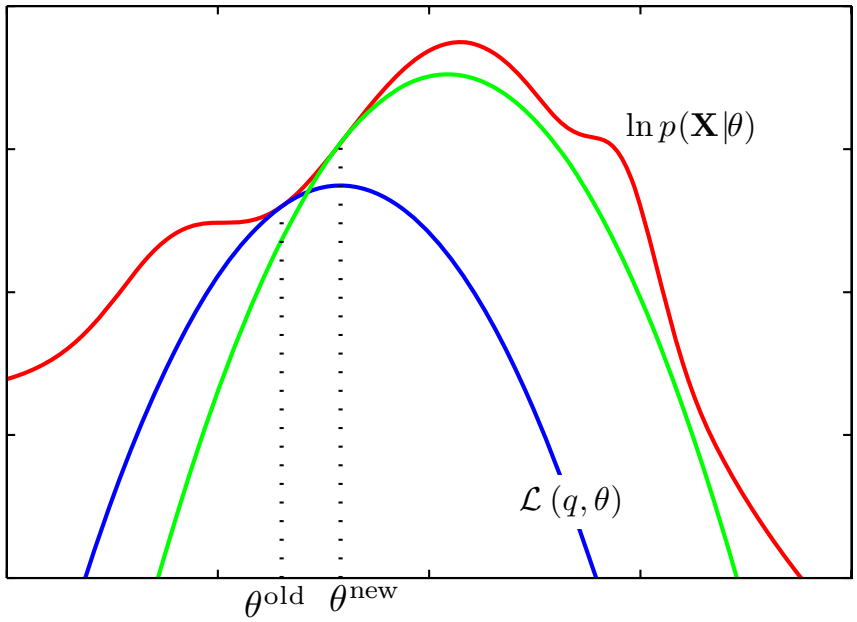

図 7: EM アルゴリズムの手続き

一般の EM アルゴリズム

- i.i.d 標本である場合
  - データ点 $\boldsymbol{x}_i$ と、対応する潜在変数 $\boldsymbol{z}_i$ が、同一の確率分布 $p(\boldsymbol{x}, \boldsymbol{z})$ から独立に得られている場合
  - 以下のように同時分布 $p(\boldsymbol{X}, \boldsymbol{Z})$ を分解できる

$$p(\boldsymbol{X}, \boldsymbol{Z}) = \prod_i p(\boldsymbol{x}_i, \boldsymbol{z}_i) \tag{196}$$

  - 従って、E ステップで計算される事後確率 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ は次のようになる

$$\begin{aligned} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) &= \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{X}|\boldsymbol{\theta})} \\ &= \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{\sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})} \\ &= \frac{\prod_i p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})}{\sum_{\boldsymbol{Z}} \prod_i p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})} \end{aligned}$$

一般の EM アルゴリズム

$$
\begin{aligned}
&= \frac{\prod_i p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})}{\prod_i \sum_{\boldsymbol{Z}} p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})} \\
&= \frac{\prod_i p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})}{\prod_i p(\boldsymbol{x}_i|\boldsymbol{\theta})} \\
&= \prod_i \frac{p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})}{p(\boldsymbol{x}_i|\boldsymbol{\theta})} \\
&= \prod_i p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \qquad (197)
\end{aligned}
$$

各データ点に対する事後確率 $p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta})$ の積として、$p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ を表現できた

- 例えば混合ガウスモデルであれば、データ点 $\boldsymbol{x}_i$ に対する各ガウス分布の負担率は、データ $\boldsymbol{x}_i$ とガウス分布のパラメータ $\boldsymbol{\theta}$ にのみ依存し、他のデータ点には依存しないということを示している

一般の EM アルゴリズム

- ここまでの話の流れ
  - 一般的な EM アルゴリズムの取り扱いを調べた
  - EM アルゴリズムに対する次の疑問を解決した

  - $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の代わりに、$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ を最大化してもよい根拠
  - $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の、$p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ による期待値を取る理由
  - E ステップと M ステップが、対数尤度関数 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ を増加させる理由

- これからの話の流れ
  - MAP 推定に対する EM アルゴリズムの適用を考える
  - EM アルゴリズムの拡張 (一般化 EM アルゴリズム) について簡単に触れる
  - 混合ガウスモデルについて、逐次型の EM アルゴリズムを導出する

MAP 推定に対する EM アルゴリズム

- 事後分布の対数 $\ln p(\boldsymbol{\theta}|\boldsymbol{X})$ の最大化
  - 今までは、尤度関数 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の最適化を考えてきた
  - 即ち、最尤推定に対する EM アルゴリズムを考えてきた
  - パラメータの事前分布 $p(\theta)$ を導入したモデルであれば、最尤推定だけでなく MAP 推定に対しても、EM アルゴリズムを使える
  - MAP 推定とは、次式のように、事後分布 $p(\boldsymbol{\theta}|\boldsymbol{X})$ を最大化するパラメータ $\boldsymbol{\theta}^*$ を求める問題である

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{X}) \tag{198}$$

$$= \arg\max_{\boldsymbol{\theta}} \frac{p(\boldsymbol{X},\boldsymbol{\theta})}{p(\boldsymbol{X})} \tag{199}$$

$$= \arg\max_{\boldsymbol{\theta}} \frac{p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{X})} \tag{200}$$

$$= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{201}$$

MAP 推定に対する EM アルゴリズム

- 事後分布の対数 $\ln p(\boldsymbol{\theta}|\boldsymbol{X})$ は

$$
\begin{aligned}
\ln p(\boldsymbol{\theta}|\boldsymbol{X}) &= \ln \frac{p(\boldsymbol{X},\boldsymbol{\theta})}{p(\boldsymbol{X})} && (202)\\
&= \ln \frac{p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{X})} \\
&= \ln p(\boldsymbol{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - p(\boldsymbol{X}) && (203)\\
&= \mathcal{L}(q,\boldsymbol{\theta}) + \mathrm{KL}(q||p) + \ln p(\boldsymbol{\theta}) - p(\boldsymbol{X}) && (204)\\
&\geq \mathcal{L}(q,\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - p(\boldsymbol{X}) && (205)\\
&= \mathcal{L}(q,\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) + \mathrm{Const.} && (206)
\end{aligned}
$$

- $\ln p(\boldsymbol{X})$ は定数とみなせるから、$\ln p(\boldsymbol{\theta}|\boldsymbol{X})$ の最大化は、結局 $\mathcal{L}(q,\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})$ の最大化に相当する

MAP 推定に対する EM アルゴリズム

- MAP 推定に対する EM アルゴリズム
  - E ステップでは、パラメータ $\boldsymbol{\theta}$ を固定しつつ、q について $\mathcal{L}(q, \boldsymbol{\theta})$ を最大化する
  - q は下界 $\mathcal{L}(q, \boldsymbol{\theta})$ にしか現れないので、通常の EM アルゴリズムと全く同様である

  - M ステップでは、分布 q を固定しつつ、パラメータ $\boldsymbol{\theta}$ について $\mathcal{L}(q, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})$ を最大化する
  - 事前分布の項 $\ln p(\boldsymbol{\theta})$ が現れているが、大抵は、通常の最尤推定に関する EM アルゴリズムと、少ししか違わない

EM アルゴリズムの拡張

- EM アルゴリズムに対する懸念
  - EM アルゴリズムは、潜在的に困難である尤度関数 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の最大化を、E ステップと M ステップの 2 つに分解してくれる
  - この 2 つのステップは多くの場合、実装が単純になる
  - 但し、複雑なモデルに対しては、2 つのどちらかのステップが、依然として手に負えないかもしれない
- 一般化 EM アルゴリズム
  - 手に負えない M ステップに対処するためのアルゴリズム
  - M ステップで、下界 $\mathcal{L}(q, \boldsymbol{\theta})$ を $\boldsymbol{\theta}$ について最大化するのは諦める代わりに、下界 $\mathcal{L}(q, \boldsymbol{\theta})$ を少しでも増加させるように、$\boldsymbol{\theta}$ を更新する
  - $\mathcal{L}(q, \boldsymbol{\theta})$ は、常に尤度関数 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の下界であるから、$\mathcal{L}$ を押し上げることは、尤度関数の増加につながる

EM アルゴリズムの拡張

- M ステップで制限付きの最適化を行うことができる
- パラメータ θ を幾つかのグループに分割
- 各グループに属するパラメータを、他のグループに属するパラメータを固定しながら、順番に最適化していく

パラメータについての補足

- パラメータ $\boldsymbol{\theta}$ についての補足
  - 任意の $\boldsymbol{\theta}$ について、下界 $\mathcal{L}(q, \boldsymbol{\theta})$ は q について唯一の最大点をもつ
  - それは事後分布 $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ である
  - またこのとき、下界 $\mathcal{L}$ は対数尤度関数 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ に一致する

  - $\boldsymbol{\theta}$ が下界 $\mathcal{L}(q, \boldsymbol{\theta})$ の大域的最適解に収束するなら、そのような $\boldsymbol{\theta}$ は、対数尤度関数 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の大域的最適解でもある
  - 任意の下界 $\mathcal{L}(q, \boldsymbol{\theta})$ の任意の極大点は、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の極大点でもある

逐次型の EM アルゴリズムの例

- 混合ガウスモデルに対する逐次型の EM アルゴリズム
  - E ステップでは、事後確率分布 $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$ を計算する
  - データが i.i.d 集合であれば、次のように、各データ点ごとの事後確率 $p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta})$ の積として分解できる

$$p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}) = \prod_i p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta}) \tag{207}$$

  - このとき、全てのデータ点 $\boldsymbol{X}$ に対して事後確率を求める必要がある
  - これを、1 つのデータ点についてだけ事後確率を求めるように変更する

  - M ステップでも、1 つのデータ点に対して求めた事後確率だけを使って、パラメータを逐次的に更新するように変更を加える
  - 混合ガウスモデルであれば、逐次的な更新式を導出することが可能

逐次型の EM アルゴリズムの例

- 従って、全てのデータ点に対する事後確率を使って、パラメータを再計算する必要がない
- これらの変更によって、逐次版の EM アルゴリズムを導出できる

- 混合ガウスモデルに対する逐次型の EM アルゴリズムの導出
  - データ点 $\boldsymbol{x}_m$ について、事後確率 (負担率)$\gamma(z_{mk})$ を更新したとする
  - 新しい負担率を $\gamma^{\mathrm{new}}(z_{mk})$、以前の負担率を $\gamma^{\mathrm{old}}(z_{mk})$ とする
  - d を次のようにおく (前後の負担率の差)

$$d = \gamma^{\mathrm{new}}(z_{mk}) - \gamma^{\mathrm{old}}(z_{mk}) \tag{208}$$

  - N_k^{new} を次のようにおく (クラスタ k に属するデータの、実質的な個数)

$$N_k^{\mathrm{new}} = N_k^{\mathrm{old}} + \gamma^{\mathrm{new}}(z_{mk}) - \gamma^{\mathrm{old}}(z_{mk}) = N_k^{\mathrm{old}} + d \tag{209}$$

逐次型の EM アルゴリズムの例

- 以前の平均 $\boldsymbol{\mu}_k^{\text{old}}$、共分散行列 $\boldsymbol{\Sigma}_k^{\text{old}}$、混合係数 π_k^{old} を、以下のように書くことにする

$$\boldsymbol{\mu}_k^{\text{old}} = \frac{1}{N_k^{\text{old}}} \sum_i \gamma^{\text{old}}(z_{ik}) \boldsymbol{x}_i \tag{210}$$

$$\boldsymbol{\Sigma}_k^{\text{old}} = \frac{1}{N_k^{\text{old}}} \sum_i \gamma^{\text{old}}(z_{ik}) (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{\text{old}})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{\text{old}})^T \tag{211}$$

$$\pi_k^{\text{old}} = \frac{N_k^{\text{old}}}{N} \tag{212}$$

但し

$$N_k^{\text{old}} = \sum_i \gamma^{\text{old}}(z_{ik}) \tag{213}$$

逐次型の EM アルゴリズムの例

- 平均の更新式は

$$
\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k^{\text{new}}} \left(\sum_{i \neq m} \gamma^{\text{old}}(z_{ik}) \boldsymbol{x}_i + \gamma^{\text{new}}(z_{mk}) \boldsymbol{x}_m \right) && (214) \\
&= \frac{1}{N_k^{\text{new}}} \left(N_k^{\text{old}} \boldsymbol{\mu}_k^{\text{old}} - \gamma^{\text{old}}(z_{mk}) \boldsymbol{x}_m + \gamma^{\text{new}}(z_{mk}) \boldsymbol{x}_m \right) \\
&= \frac{1}{N_k^{\text{new}}} \left(\left(N_k^{\text{new}} - \gamma^{\text{new}}(z_{mk}) + \gamma^{\text{old}}(z_{mk}) \right) \boldsymbol{\mu}_k^{\text{old}} - \right. \\
&\qquad \left. \gamma^{\text{old}}(z_{mk}) \boldsymbol{x}_m + \gamma^{\text{new}}(z_{mk}) \boldsymbol{x}_m \right) \\
&= \boldsymbol{\mu}_k^{\text{old}} + \frac{1}{N_k^{\text{new}}} \left(-(\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})) \boldsymbol{\mu}_k^{\text{old}} + \right. \\
&\qquad \left. (\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})) \boldsymbol{x}_m \right) \\
&= \boldsymbol{\mu}_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right) && (215)
\end{aligned}
$$

逐次型の EM アルゴリズムの例

$$= \boldsymbol{\mu}_k^{\text{old}} + \frac{d}{N_k^{\text{new}}} \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right) \tag{216}$$

- 分散の更新式は

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k^{\text{new}}} \sum_i \gamma^{\text{new}}(z_{ik}) \boldsymbol{x}_i \boldsymbol{x}_i^T - \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \tag{217}$$

$$= \frac{1}{N_k^{\text{new}}} \left(\sum_{i \neq m} \gamma^{\text{old}}(z_{ik}) \boldsymbol{x}_i \boldsymbol{x}_i^T + \gamma^{\text{new}}(z_{mk}) \boldsymbol{x}_m \boldsymbol{x}_m^T \right) - \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \tag{218}$$

$$= \frac{1}{N_k^{\text{new}}} \left(\left(\sum_i \gamma^{\text{old}}(z_{ik}) \boldsymbol{x}_i \boldsymbol{x}_i^T - \gamma^{\text{old}}(z_{mk}) \boldsymbol{x}_m \boldsymbol{x}_m^T \right) + \gamma^{\text{new}}(z_{mk}) \boldsymbol{x}_m \boldsymbol{x}_m^T \right) - \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \tag{219}$$

逐次型の EM アルゴリズムの例

$$
\begin{aligned}
&= \frac{1}{N_k^{\text{new}}} \Big(N_k^{\text{old}} \left(\boldsymbol{\Sigma}_k^{\text{old}} + \boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T \right) - \\
&\qquad \gamma^{\text{old}}(z_{mk}) \boldsymbol{x}_m \boldsymbol{x}_m^T + \gamma^{\text{new}}(z_{mk}) \boldsymbol{x}_m \boldsymbol{x}_m^T \Big) - \\
&\qquad \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \qquad (220) \\
&= \frac{1}{N_k^{\text{new}}} \Big(N_k^{\text{old}} \boldsymbol{\Sigma}_k^{\text{old}} + N_k^{\text{old}} \boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T + \\
&\qquad \left(\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \right) \boldsymbol{x}_m \boldsymbol{x}_m^T - \\
&\qquad N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \Big) \\
&= \frac{1}{N_k^{\text{new}}} \Big(N_k^{\text{old}} \boldsymbol{\Sigma}_k^{\text{old}} + N_k^{\text{old}} \boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T + \\
&\qquad d \boldsymbol{x}_m \boldsymbol{x}_m^T - N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \Big) \qquad (221)
\end{aligned}
$$

逐次型の EM アルゴリズムの例

ここで

$$
\begin{aligned}
& N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \\
= \quad & \left(N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{old}} + d(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}})\right) (\boldsymbol{\mu}_k^{\text{new}})^T \\
= \quad & \frac{1}{N_k^{\text{new}}} \left(N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{old}} + d(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}})\right) \\
& \qquad \left(N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{old}} + d\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\right)^T \\
= \quad & N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T + 2\boldsymbol{\mu}_k^{\text{old}} d \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T + \\
& \qquad \frac{1}{N_k^{\text{new}}} d^2 \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right) \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \qquad (222) \\
= \quad & (N_k^{\text{old}} + d) \boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T + \\
& \qquad 2d\boldsymbol{\mu}_k^{\text{old}} \boldsymbol{x}_m^T - 2d\boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T + \\
& \qquad \frac{1}{N_k^{\text{new}}} d^2 \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right) \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \qquad (223)
\end{aligned}
$$

逐次型の EM アルゴリズムの例

であるから

$$
\begin{aligned}
& N_k^{\text{old}}\boldsymbol{\mu}_k^{\text{old}}(\boldsymbol{\mu}_k^{\text{old}})^T + d\boldsymbol{x}_m\boldsymbol{x}_m^T - N_k^{\text{new}}\boldsymbol{\mu}_k^{\text{new}}(\boldsymbol{\mu}_k^{\text{new}})^T \\
= \quad & N_k^{\text{old}}\boldsymbol{\mu}_k^{\text{old}}(\boldsymbol{\mu}_k^{\text{old}})^T + d\boldsymbol{x}_m\boldsymbol{x}_m^T - \\
& \quad (N_k^{\text{old}} + d)\boldsymbol{\mu}_k^{\text{old}}(\boldsymbol{\mu}_k^{\text{old}})^T - \\
& \quad 2d\boldsymbol{\mu}_k^{\text{old}}\boldsymbol{x}_m^T + 2d\boldsymbol{\mu}_k^{\text{old}}(\boldsymbol{\mu}_k^{\text{old}})^T - \\
& \quad \frac{1}{N_k^{\text{new}}}d^2\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \\
= \quad & d\boldsymbol{x}_m\boldsymbol{x}_m^T + d\boldsymbol{\mu}_k^{\text{old}}(\boldsymbol{\mu}_k^{\text{old}})^T - 2d\boldsymbol{\mu}_k^{\text{old}}\boldsymbol{x}_m^T - \\
& \quad \frac{1}{N_k^{\text{new}}}d^2\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \\
= \quad & d\left(\boldsymbol{x}_m\boldsymbol{x}_m^T + \boldsymbol{\mu}_k^{\text{old}}(\boldsymbol{\mu}_k^{\text{old}})^T - 2\boldsymbol{\mu}_k^{\text{old}}\boldsymbol{x}_m^T\right) - \\
& \quad \frac{1}{N_k^{\text{new}}}d^2\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \\
= \quad & d(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}})(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}})^T -
\end{aligned}
$$

逐次型の EM アルゴリズムの例

$$
\begin{aligned}
& \frac{1}{N_k^{\text{new}}} d^2 \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \\
= \quad & \frac{d}{N_k^{\text{new}}} \left(N_k^{\text{new}} - d\right)\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \\
= \quad & \frac{d}{N_k^{\text{new}}} N_k^{\text{old}} \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \qquad (224)
\end{aligned}
$$

以上より

$$
\begin{aligned}
\boldsymbol{\Sigma}_k^{\text{new}} = \quad & \frac{1}{N_k^{\text{new}}} \Big(N_k^{\text{old}} \boldsymbol{\Sigma}_k^{\text{old}} + N_k^{\text{old}} \boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T + \\
& \qquad d\boldsymbol{x}_m \boldsymbol{x}_m^T - N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \Big) \\
= \quad & \frac{1}{N_k^{\text{new}}} \Big(N_k^{\text{old}} \boldsymbol{\Sigma}_k^{\text{old}} + \\
& \qquad \frac{d}{N_k^{\text{new}}} N_k^{\text{old}} \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \Big)
\end{aligned}
$$

逐次型の EM アルゴリズムの例

$$
\begin{aligned}
&= \frac{N_k^{\text{old}}}{N_k^{\text{new}}} \left(\boldsymbol{\Sigma}_k^{\text{old}} + \frac{d}{N_k^{\text{new}}} \right. \\
&\qquad \left. \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right) \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right)^T \right) \\
&= \frac{N_k^{\text{old}}}{N_k^{\text{new}}} \left(\boldsymbol{\Sigma}_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{ik}) - \gamma^{\text{old}}(z_{ik})}{N_k^{\text{new}}} \right. \\
&\qquad \left. \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right) \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right)^T \right) \qquad (225)
\end{aligned}
$$

- 混合係数の更新式は

$$
\pi_k^{\text{new}} = \frac{N_k^{\text{new}}}{N} = \frac{N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N} \qquad (226)
$$

逐次型の EM アルゴリズムの例

- 混合ガウスモデルにおける逐次型の EM アルゴリズム
  - 上記より、パラメータ $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ の逐次更新式が得られた

$$\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right) \tag{227}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{N_k^{\text{old}}}{N_k^{\text{new}}} \left(\boldsymbol{\Sigma}_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{ik}) - \gamma^{\text{old}}(z_{ik})}{N_k^{\text{new}}} \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right) \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \right) \tag{228}$$

$$\pi_k^{\text{new}} = \frac{N_k^{\text{new}}}{N} = \frac{N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N} \tag{229}$$

但し

$$N_k^{\text{new}} = N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \tag{230}$$

逐次型の EM アルゴリズムの例

- 到着したデータ $\boldsymbol{x}_m$ について、E ステップで負担率 $\gamma(z_{mk})$ を求めた後に、M ステップで (上記の更新式を用いて) パラメータを更新することを、交互に繰り返せばよい

- 逐次型の EM アルゴリズムの特徴
  - $\boldsymbol{x}_m$ が新しく到着したデータであれば、$\gamma^{\mathrm{old}}(z_{mk}) = 0$ とする
  - E ステップと M ステップの計算に必要な時間は、データ点の総数とは無関係に決まる
  - パラメータの更新は、全データについての処理を待たずに、各データ点についての処理の後に行われる
  - そのため、逐次型の EM アルゴリズムは、従来のバッチ型に比べて、速く収束する

逐次型の EM アルゴリズムの例

- ここまでの話の流れ
  - MAP 推定に対する EM アルゴリズムについて考えた
  - EM アルゴリズムの拡張 (一般化 EM アルゴリズム) について簡単に触れた
  - 混合ガウスモデルについて、逐次型の EM アルゴリズムを導出した