目次

EM アルゴリズムの解釈

- ここまででの話の流れ
  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

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln \left(\sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \right) \tag{1}$$

EM アルゴリズムの解釈

- 潜在変数 z が連続変数の場合は

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \ln\left(\int p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})d\boldsymbol{Z}\right) \tag{2}$$

  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

$$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathrm{old}}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\mathrm{old}}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) \tag{3}$$

連続潜在変数の場合は次のようになる

$$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathrm{old}}) = \int p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\mathrm{old}}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) d\boldsymbol{Z} \tag{4}$$

EM アルゴリズムの解釈

- 上式において、$p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\mathrm{old}})$ におけるパラメータ $\boldsymbol{\theta}^{\mathrm{old}}$ は、変数ではなく定数であることに注意

- 更に、$\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathrm{old}})$ 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 $\boldsymbol{\theta}^{\mathrm{new}}$ を得る

$$\boldsymbol{\theta}^{\mathrm{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathrm{old}}) \tag{5}$$

- 注意点
  - $\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\mathrm{old}})$ 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 $\boldsymbol{\theta}^{\text{old}}$ に初期化する
2. E ステップ: 事後確率分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})$ 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 $\boldsymbol{\theta}^{\mathrm{new}}$ を計算する

$$\boldsymbol{\theta}^{\mathrm{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) \tag{6}$$

但し

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \tag{7}$$

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

$$\boldsymbol{\theta}^{\mathrm{old}} \leftarrow \boldsymbol{\theta}^{\mathrm{new}} \tag{8}$$

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

$$
\begin{aligned}
& p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \\
= \; & p(\boldsymbol{Z}|\boldsymbol{\theta}) p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\theta}) \\
= \; & \prod_i p(\boldsymbol{z}_i|\boldsymbol{\theta}) p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta}) \\
= \; & \prod_i \left(\prod_k \pi_k^{z_{ik}} \right) \left(\prod_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{ik}} \right) \\
= \; & \prod_i \prod_k \pi_k^{z_{ik}} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{ik}} \\
= \; & \prod_i \prod_k (\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{z_{ik}}
\end{aligned}
\tag{9}
$$

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

$$\begin{aligned}
& \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \\
= & \ln \left(\prod_i \prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_{ik}} \right) \\
= & \sum_i \sum_k \ln \left(\left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_{ik}} \right) \\
= & \sum_i \sum_k z_{ik} \ln \left(\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \\
= & \sum_i \sum_k z_{ik} \left(\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \qquad (10)
\end{aligned}$$

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

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_i \ln \left(\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \tag{11}$$

- $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ 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

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

混合ガウス分布の再解釈

$$
\begin{aligned}
&= \sum_i z_{ik} \frac{\partial}{\partial \boldsymbol{\mu}_k} \ln \left(\frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \right. \\
&\qquad \left. \exp \left\{ -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right\} \right) \\
&= \sum_i z_{ik} \frac{\partial}{\partial \boldsymbol{\mu}_k} \left(-\frac{D}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_k| \right. \\
&\qquad \left. -\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right) \\
&= \sum_i z_{ik} \frac{\partial}{\partial \boldsymbol{\mu}_k} \left(-\frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right) \\
&= \sum_i z_{ik} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) = 0
\end{aligned} \tag{12}
$$

これより

$$
\sum_i z_{ik} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k = \sum_i z_{ik} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{x}_i \tag{13}
$$

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

$$
\begin{aligned}
\sum_i z_{ik}\boldsymbol{\mu}_k &= \sum_i z_{ik}\boldsymbol{x}_i \\
\boldsymbol{\mu}_k \sum_i z_{ik} &= \sum_i z_{ik}\boldsymbol{x}_i \\
\boldsymbol{\mu}_k &= \frac{1}{\sum_i z_{ik}} \sum_i z_{ik}\boldsymbol{x}_i
\end{aligned}
\tag{14}
$$

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

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \\
= \;& \sum_i z_{ik} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
= \;& \sum_i z_{ik} \frac{\partial}{\partial \boldsymbol{\Sigma}_k} \left(-\frac{1}{2} \ln |\boldsymbol{\Sigma}_k| - \frac{1}{2} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k) \right) \\
= \;& \sum_i z_{ik} \left(-\frac{1}{2} \left(\boldsymbol{\Sigma}_k^{-1} \right)^T + \frac{1}{2} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \right) \\
= \;& \frac{1}{2} \sum_i z_{ik} \left(-\boldsymbol{\Sigma}_k^{-1} + \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \right) \qquad (15) \\
= \;& 0
\end{aligned}
$$

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

$$\frac{\partial}{\partial \boldsymbol{X}} \ln |\boldsymbol{X}| = \left(\boldsymbol{X}^{-1}\right)^T \tag{16}$$

- これより

$$\sum_i z_{ik} \boldsymbol{\Sigma}_k^{-1} = \sum_i z_{ik} \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} \tag{17}$$

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

$$\begin{aligned} \sum_i z_{ik} \boldsymbol{\Sigma}_k &= \sum_i z_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \\ \boldsymbol{\Sigma}_k \sum_i z_{ik} &= \sum_i z_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \\ \boldsymbol{\Sigma}_k &= \frac{1}{\sum_i z_{ik}} \sum_i z_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}_k)(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^T \end{aligned} \tag{18}$$

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

$$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) + \lambda \left(\sum_k \pi_k - 1 \right) \tag{19}$$

  - 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

$$= \sum_i z_{ik} \frac{1}{\pi_k} + \lambda = 0 \tag{20}$$

- これより、両辺に π_k を掛けて

$$\sum_i z_{ik} + \lambda \pi_k = 0 \tag{21}$$

全ての k について総和を取ると

$$\begin{aligned}
&\sum_k \sum_i z_{ik} + \sum_k \lambda \pi_k = 0 \\
&\sum_i \left(\sum_k z_{ik} \right) + \lambda \sum_k \pi_k = 0 \\
&\sum_i 1 + \lambda = 0 \\
&N + \lambda = 0 \\
&\therefore \lambda = -N
\end{aligned} \tag{22}$$

混合ガウス分布の再解釈

- よって

$$
\begin{aligned}
\sum_i z_{ik}\frac{1}{\pi_k} - N &= 0 \\
\sum_i z_{ik} - N\pi_k &= 0 \\
N\pi_k &= \sum_i z_{ik} \\
\therefore \pi_k &= \frac{1}{N}\sum_i z_{ik}
\end{aligned} \tag{23}
$$

- 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

$$
\begin{aligned}
& p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \\
= & \frac{p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta})p(\boldsymbol{z}_i|\boldsymbol{\theta})}{p(\boldsymbol{x}_i|\boldsymbol{\theta})} && (24) \\
\propto & \; p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{\theta})p(\boldsymbol{z}_i|\boldsymbol{\theta}) && (25) \\
& \qquad (\because p(\boldsymbol{x}_i|\boldsymbol{\theta}) \text{ は、} \boldsymbol{z}_i \text{ には依存しない定数項}) \\
= & \left(\prod_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_{ik}}\right)\left(\prod_k \pi_k^{z_{ik}}\right) \\
= & \prod_k (\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{z_{ik}} && (26)
\end{aligned}
$$

以上より

$$
p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \propto \prod_k (\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{z_{ik}} \tag{27}
$$

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

$$p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) \propto \prod_i \prod_k (\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{z_{ik}} \tag{28}$$

- $p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta})$ 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

$$p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) = \frac{\prod_k (\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{z_{ik}}}{\sum_{\boldsymbol{z}_i} \prod_k (\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{z_{ik}}} \tag{29}$$

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

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}_i\sim p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta})}\left[z_{ik}\right] \\
=&\sum_{\boldsymbol{z}_i} z_{ik}\, p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta}) \\
=&\sum_{\boldsymbol{z}_i} z_{ik}\frac{\prod_k\left(\pi_k\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right)^{z_{ik}}}{\sum_{\boldsymbol{z}_i}\prod_k\left(\pi_k\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right)^{z_{ik}}}
\end{aligned} \tag{30}$$

- ここで

$$\sum_{\boldsymbol{z}_i}\prod_k\left(\pi_k\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right)^{z_{ik}}=\sum_k\left(\pi_k\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right) \tag{31}$$

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
$$\sum_{\boldsymbol{z}_i} z_{ik} \prod_k \left(\pi_k \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right)^{z_{ik}} = \pi_k \mathcal{N}(\boldsymbol{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{32}$$
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

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}_i \sim p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta})}[z_{ik}] \\
= \quad & \frac{\displaystyle\sum_{\boldsymbol{z}_i} z_{ik} \prod_k (\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{z_{ik}}}{\displaystyle\sum_{\boldsymbol{z}_i} \prod_k (\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{z_{ik}}} \\
= \quad & \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \equiv \gamma(z_{ik}) \qquad (33)
\end{aligned}
$$

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

$$
\begin{aligned}
& \mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}\left[\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})\right] \\
= \; & \mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}\left[\sum_i\sum_k z_{ik}\left(\ln\pi_k+\ln\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right)\right] \\
= \; & \sum_i\sum_k \mathbb{E}_{\boldsymbol{z}_i\sim p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta})}\left[z_{ik}\left(\ln\pi_k+\ln\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right)\right] && (34) \\
= \; & \sum_i\sum_k \mathbb{E}_{\boldsymbol{z}_i\sim p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta})}\left[z_{ik}\right]\left(\ln\pi_k+\ln\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right) && (35) \\
= \; & \sum_i\sum_k \gamma(z_{ik})\left(\ln\pi_k+\ln\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\right) && (36)
\end{aligned}
$$

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

$$\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) = \sum_i \sum_k z_{ik} \left(\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \qquad (37)$$

- 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

混合ガウス分布の再解釈

- これからの話の流れ
  - 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)

$$
\begin{aligned}
& p(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \\
= \; & p(\boldsymbol{x}|\boldsymbol{\mu}_k, \epsilon\boldsymbol{I}) \\
= \; & \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \epsilon\boldsymbol{I}) \\
= \; & \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\epsilon\boldsymbol{I}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T(\epsilon\boldsymbol{I})^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\} && (38) \\
= \; & \frac{1}{(2\pi\epsilon)^{\frac{D}{2}}} \exp\left\{-\frac{1}{2\epsilon}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\} && (39) \\
& \qquad \left(\because |\epsilon\boldsymbol{I}|^{\frac{1}{2}} = \left(\epsilon^D|\boldsymbol{I}|\right)^{\frac{1}{2}} = \left(\epsilon^D\right)^{\frac{1}{2}} = \epsilon^{\frac{D}{2}}\right) \\
= \; & \frac{1}{(2\pi\epsilon)^{\frac{D}{2}}} \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x}-\boldsymbol{\mu}_k||^2\right\} && (40)
\end{aligned}
$$

K-Means 法との関連

$$
\begin{aligned}
p(\boldsymbol{x}|\boldsymbol{\theta}) &= \sum_k \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) && (41) \\
&= \sum_k \pi_k \frac{1}{(2\pi\epsilon)^{\frac{D}{2}}} \exp\left\{ -\frac{1}{2\epsilon} ||\boldsymbol{x} - \boldsymbol{\mu}_k||^2 \right\} && (42)
\end{aligned}
$$

- 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

$$
\begin{aligned}
\gamma(z_{ik}) &= \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \\
&= \frac{\pi_k \exp\left\{ -\frac{1}{2\epsilon} ||\boldsymbol{x} - \boldsymbol{\mu}_k||^2 \right\}}{\sum_j \pi_j \exp\left\{ -\frac{1}{2\epsilon} ||\boldsymbol{x} - \boldsymbol{\mu}_j||^2 \right\}} && (43)
\end{aligned}
$$

K-Means 法との関連

- 負担率は、以下のように変形できる

$$
\begin{aligned}
&\frac{\pi_k \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x}-\boldsymbol{\mu}_k||^2\right\}}{\sum_j \pi_j \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x}-\boldsymbol{\mu}_j||^2\right\}} \\
=&\left(\frac{\sum_j \pi_j \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x}-\boldsymbol{\mu}_j||^2\right\}}{\pi_k \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x}-\boldsymbol{\mu}_k||^2\right\}}\right)^{-1} \\
=&\left(\sum_j \frac{\pi_j}{\pi_k} \frac{\left(\exp\left\{-||\boldsymbol{x}-\boldsymbol{\mu}_j||^2\right\}\right)^{\frac{1}{2\epsilon}}}{\left(\exp\left\{-||\boldsymbol{x}-\boldsymbol{\mu}_k||^2\right\}\right)^{\frac{1}{2\epsilon}}}\right)^{-1} \\
=&\left(\sum_j \frac{\pi_j}{\pi_k} \left(\frac{\exp\left\{-||\boldsymbol{x}-\boldsymbol{\mu}_j||^2\right\}}{\exp\left\{-||\boldsymbol{x}-\boldsymbol{\mu}_k||^2\right\}}\right)^{\frac{1}{2\epsilon}}\right)^{-1} \\
=&\left(1+\sum_{j\neq k} \frac{\pi_j}{\pi_k} \left(\frac{\exp\left\{-||\boldsymbol{x}-\boldsymbol{\mu}_j||^2\right\}}{\exp\left\{-||\boldsymbol{x}-\boldsymbol{\mu}_k||^2\right\}}\right)^{\frac{1}{2\epsilon}}\right)^{-1} \qquad (44)
\end{aligned}
$$

K-Means 法との関連

- ここで、k^* を次で定める

$$k^* = \arg\min_j ||\boldsymbol{x} - \boldsymbol{\mu}_j||^2 = \arg\max_j \left(-||\boldsymbol{x} - \boldsymbol{\mu}_j||^2\right) \tag{45}$$

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

$$\lim_{\epsilon \to 0} \left(\sum_{j \neq k} \frac{\pi_j}{\pi_k} \left(\frac{\exp\left\{-||\boldsymbol{x} - \boldsymbol{\mu}_j||^2\right\}}{\exp\left\{-||\boldsymbol{x} - \boldsymbol{\mu}_k||^2\right\}} \right)^{\frac{1}{2\epsilon}} \right) \tag{46}$$

を考えると、全ての $j \neq k^*$ について

$$\frac{\exp\left\{-||\boldsymbol{x} - \boldsymbol{\mu}_j||^2\right\}}{\exp\left\{-||\boldsymbol{x} - \boldsymbol{\mu}_k||^2\right\}} < 1 \tag{47}$$

K-Means 法との関連

が成立するので

$$\lim_{\epsilon \to 0} \left(\sum_{j \neq k} \frac{\pi_j}{\pi_k} \left(\frac{\exp\left\{-||\boldsymbol{x} - \boldsymbol{\mu}_j||^2\right\}}{\exp\left\{-||\boldsymbol{x} - \boldsymbol{\mu}_k||^2\right\}} \right)^{\frac{1}{2\epsilon}} \right) = 0 \tag{48}$$

である

- 従って、$k = k^*$ のとき

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

$$= \quad (1+0)^{-1} = 1 \tag{49}$$

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

- $k \neq k^*$ のとき

$$1 = \sum_k \gamma(z_{ik}) = \gamma(z_{ik^*}) + \sum_{k \neq k^*} \gamma(z_{ik}) \tag{50}$$

であって、両辺の $\epsilon \to 0$ による極限を取れば

$$\begin{aligned} & 1 = \lim_{\epsilon \to 0} \left(\gamma(z_{ik^*}) + \sum_{k \neq k^*} \gamma(z_{ik}) \right) \\ \Rightarrow \quad & 1 = \lim_{\epsilon \to 0} \gamma(z_{ik^*}) + \lim_{\epsilon \to 0} \sum_{k \neq k^*} \gamma(z_{ik}) \\ \Rightarrow \quad & 1 = 1 + \lim_{\epsilon \to 0} \sum_{k \neq k^*} \gamma(z_{ik}) \end{aligned} \tag{51}$$

となるから

$$\lim_{\epsilon \to 0} \sum_{k \neq k^*} \gamma(z_{ik}) = 0 \tag{52}$$

が明らかに成立するほか、以下の不等式が

$$0 \leq \gamma(z_{ik}) \leq \sum_{k \neq k^*} \gamma(z_{ik}) \tag{53}$$

$\gamma(z_{ik}) \geq 0$ 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

$$0 \leq \lim_{\epsilon \to 0} \gamma(z_{ik}) \leq \lim_{\epsilon \to 0} \sum_{k \neq k^*} \gamma(z_{ik}) = 0 \tag{54}$$

従って、$k \neq k^*$ の場合は

$$\lim_{\epsilon \to 0} \gamma(z_{ik}) = 0 \tag{55}$$

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

$$\gamma(z_{ik}) \equiv p(z_{ik}=1|\boldsymbol{x}_i) = \begin{cases} 1 & (k=k^*\text{の場合}) \\ 0 & (\text{それ以外の場合}) \end{cases} \tag{56}$$

- 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

$$\text{K-Means :} \qquad \boldsymbol{\mu}_k = \frac{1}{\sum_i r_{ik}} \sum_i r_{ik} \boldsymbol{x}_i \tag{57}$$

$$\text{EM アルゴリズム :} \qquad \boldsymbol{\mu}_k = \frac{1}{\sum_i \gamma(z_{ik})} \sum_i \gamma(z_{ik}) \boldsymbol{x}_i \tag{58}$$

- 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

$$
\begin{aligned}
& \mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}\left[\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})\right] \\
= & \sum_i \sum_k \gamma(z_{ik}) \left(\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) \\
= & \sum_i \sum_k \gamma(z_{ik}) \left(\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \epsilon \boldsymbol{I})\right) \\
= & \sum_i \sum_k \gamma(z_{ik}) \left(\ln \pi_k + \ln \left(\frac{1}{(2\pi\epsilon)^{\frac{D}{2}}} \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2\right\}\right)\right) \\
= & \sum_i \sum_k \gamma(z_{ik}) \left(\ln \pi_k - \frac{D}{2}\ln(2\pi\epsilon) - \frac{1}{2\epsilon}||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2\right)
\end{aligned} \tag{59}
$$

K-Means 法との関連

- 両辺に ϵ を掛けると

$$\begin{aligned} & \epsilon \cdot \mathbb{E}_{\boldsymbol{Z}\sim p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}\left[\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})\right] \\ = & \sum_i \sum_k \gamma(z_{ik}) \left(\epsilon \ln \pi_k - \right. \\ & \left. \quad \frac{D}{2}\epsilon \ln(2\pi\epsilon) - \frac{1}{2}||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2 \right) \end{aligned} \tag{60}$$

- $\epsilon \to 0$ の極限を取ると

$$\gamma(z_{ik}) \to r_{ik}, \quad \epsilon \ln \pi_k \to 0, \quad \epsilon \ln(2\pi\epsilon) \to 0 \tag{61}$$

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

$$
\begin{aligned}
&= -\frac{1}{2}\sum_i \sum_k r_{ik}||\boldsymbol{x}_i - \boldsymbol{\mu}_k||^2 && (62)\\
&= -J && (63)
\end{aligned}
$$

- 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

$$\pi_k = \frac{\sum_i \gamma(z_{ik})}{N} \tag{64}$$

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

$$\pi_k = \frac{\sum_i r_{ik}}{N} = \frac{N_k}{N} \tag{65}$$

  - 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
  - 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

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \tag{66}$$

- 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

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = \int_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) d\boldsymbol{Z} \tag{67}$$

- 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

$$f\left(\sum_i \lambda_i x_i\right) \leq \sum_i \lambda_i f(x_i) \tag{68}$$

  - 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

$$\begin{aligned} f\left(\sum_i p(x_i) x_i\right) &\leq \sum_i p(x_i) f(x_i) \\ f\left(\mathbb{E}[x]\right) &\leq \mathbb{E}[f(x)] \end{aligned} \tag{69}$$

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

$$f\left(\int \boldsymbol{x}p(\boldsymbol{x})d\boldsymbol{x}\right) \leq \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \tag{70}$$

例えば、$f(x) = -\ln x$ は凸関数であるから

$$-\ln\left(\int \boldsymbol{x}p(\boldsymbol{x})d\boldsymbol{x}\right) \leq \int (-\ln \boldsymbol{x})\, p(\boldsymbol{x})d\boldsymbol{x} \tag{71}$$

よって

$$\ln\left(\int \boldsymbol{x}p(\boldsymbol{x})d\boldsymbol{x}\right) \geq \int (\ln \boldsymbol{x})\, p(\boldsymbol{x})d\boldsymbol{x} \tag{72}$$

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

$$\mathrm{H}[p] = -\int p(\boldsymbol{x}) \ln p(\boldsymbol{x}) d\boldsymbol{x} \tag{73}$$

  - 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

$$\mathrm{KL}(p||q) = -\int p(\boldsymbol{x}) \ln \left\{ \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} \right\} d\boldsymbol{x} \tag{74}$$

  - $\mathrm{KL}(p||q) \geq 0$ 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

$$\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q||p) \tag{75}$$

  - $\mathcal{L}(q, \boldsymbol{\theta})$ 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

$$
\begin{aligned}
\ln p(\boldsymbol{X}|\boldsymbol{\theta}) &= \underbrace{\left(\sum_{\boldsymbol{Z}} q(\boldsymbol{Z})\right)}_{=1} \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \\
&= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \\
&= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})} \\
&= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \left(\frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})} \right) \\
&= \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} - \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
&= \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q||p) \qquad (76)
\end{aligned}
$$

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

$$\mathcal{L}(q,\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \tag{77}$$

$$\mathrm{KL}(q||p) = -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})} \tag{78}$$

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

$$\mathcal{L}(q,\boldsymbol{\theta}) \leq \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \tag{79}$$

- $\mathcal{L}(q,\boldsymbol{\theta})$ 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 $\ln p(\mathbf{X}|\boldsymbol{\theta})$.

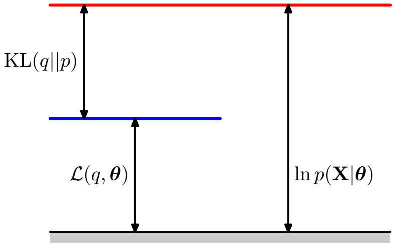

図 1: $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の分解

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

$$
\begin{aligned}
& \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{old}}) \\
= \; & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) + \mathrm{KL}(q||p) && (80) \\
= \; & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) + \mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})) && (81) \\
& \text{(E ステップ前)}
\end{aligned}
$$

一般の EM アルゴリズム

- 上式において、左辺の $\ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\text{old}})$ は、q には依存しない定数である
- 従って、q について $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ 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、$q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})$ 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 $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ は、対数尤度 $\ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\text{old}})$ に一致する

$$\begin{aligned}
& \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\text{old}}) && (82) \\
= \; & \mathcal{L}(q, \boldsymbol{\theta}^{\text{old}}) + \mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})) \\
= \; & \mathcal{L}(q, \boldsymbol{\theta}^{\text{old}}) + \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})) && (83) \\
= \; & \mathcal{L}(q, \boldsymbol{\theta}^{\text{old}}) && (84) \\
& (\text{E ステップ後})
\end{aligned}$$

一般の EM アルゴリズム

- 次の図 2 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.

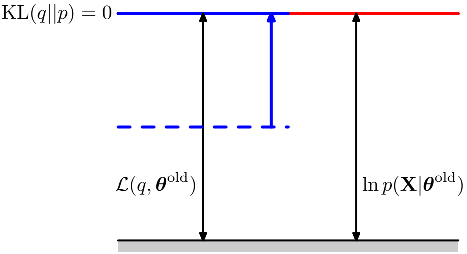

図 2: EM アルゴリズムの E ステップ

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

$$
\begin{aligned}
& \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \\
= \; & \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q||p) && (85) \\
= \; & \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})) && (86) \\
= \; & \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})) && (87) \\
& \text{(M ステップ前)}
\end{aligned}
$$

  - 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

$$
\begin{aligned}
& \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{new}}) \\
= \; & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}}) + \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{new}})) \qquad (88) \\
& (\text{M ステップ後})
\end{aligned}
$$

- 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 3)

$$
\begin{aligned}
& \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{new}}) - \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{old}}) \\
= \; & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}}) + \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{new}})) - \\
& \quad \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) \qquad (89) \\
= \; & \left(\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}}) - \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})\right) + \\
& \quad \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{new}})) \qquad (90) \\
\geq \; & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}}) - \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) \qquad (91)
\end{aligned}
$$

一般の EM アルゴリズム

- 次の図 3 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.

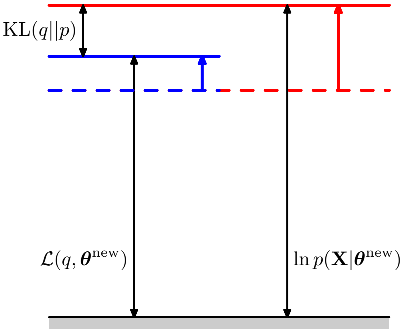

図 3: EM アルゴリズムの M ステップ

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)

$$
\begin{aligned}
& \mathcal{L}(q, \boldsymbol{\theta}) \\
= & \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
= & \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}})} \qquad (92) \\
= & \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) - \\
& \qquad \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}}) \qquad (93) \\
= & \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) + \text{Const.} \qquad (94)
\end{aligned}
$$

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 2 と図 3 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

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

一般の EM アルゴリズム

$$
\begin{aligned}
&= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}^{\text{new}})}{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}})} + \\
&\qquad \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{new}})} \\
&= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}^{\text{new}}) - \\
&\qquad \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}}) - \\
&\qquad \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{new}})} \\
&= \mathcal{Q}(\boldsymbol{\theta}^{\text{old}},\boldsymbol{\theta}^{\text{new}}) - \mathcal{Q}(\boldsymbol{\theta}^{\text{old}},\boldsymbol{\theta}^{\text{old}}) + \\
&\qquad \mathrm{KL}\left(p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) || p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{new}})\right) \qquad (95) \\
&\geq \mathcal{Q}(\boldsymbol{\theta}^{\text{old}},\boldsymbol{\theta}^{\text{new}}) - \mathcal{Q}(\boldsymbol{\theta}^{\text{old}},\boldsymbol{\theta}^{\text{old}}) \qquad (96) \\
&\geq 0
\end{aligned}
$$

一般の EM アルゴリズム

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

$$
\begin{aligned}
\ln p(\boldsymbol{X}|\boldsymbol{\theta}) &= \ln \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \\
&= \ln \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
&\geq \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
&= \mathcal{L}(q, \boldsymbol{\theta}) \qquad (97)
\end{aligned}
$$

  - 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

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

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

一般の EM アルゴリズム

$$
\begin{aligned}
&= -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
&= \mathrm{KL}(q||p) \qquad (98)
\end{aligned}
$$

ゆえ、$\mathrm{KL}(q||p)$ となることが分かったので

$$
\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q,\boldsymbol{\theta}) + \mathrm{KL}(q||p) \qquad (99)
$$

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 4)
  - 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 $\boldsymbol{\theta}^{\mathrm{new}}$ 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 $\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}})$ が計算される
  - 下界 $\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}})$ は、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ と $\boldsymbol{\theta}^{\mathrm{new}}$ で接する

- 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

$$
\begin{aligned}
& \left.\frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\boldsymbol{X}|\boldsymbol{\theta})\right|_{\boldsymbol{\theta}^{\text{old}}} \\
= \quad & \left.\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta})\right|_{\boldsymbol{\theta}^{\text{old}}} + \left.\frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{KL}(q||p)\right|_{\boldsymbol{\theta}^{\text{old}}} \\
= \quad & \left.\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta})\right|_{\boldsymbol{\theta}^{\text{old}}}
\end{aligned}
\tag{100}
$$

  - 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、$\boldsymbol{\theta}^{\text{old}}$ における微分値が等しくなる
  - 従って、$\boldsymbol{\theta}^{\text{old}}$ 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.

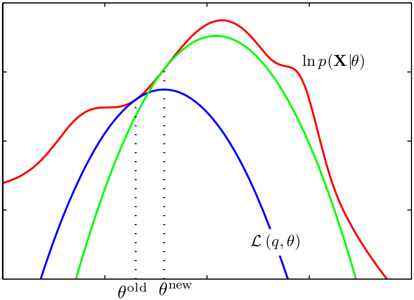

図 4: EM アルゴリズムの手続き

- 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

$$p(\boldsymbol{X}, \boldsymbol{Z}) = \prod_i p(\boldsymbol{x}_i, \boldsymbol{z}_i) \tag{101}$$

  - 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

$$
\begin{aligned}
&= \frac{\prod_i p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})}{\prod_i \sum_{\boldsymbol{Z}} p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})} \\
&= \frac{\prod_i p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})}{\prod_i p(\boldsymbol{x}_i|\boldsymbol{\theta})} \\
&= \prod_i \frac{p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})}{p(\boldsymbol{x}_i|\boldsymbol{\theta})} \\
&= \prod_i p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \qquad (102)
\end{aligned}
$$

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

$$\begin{aligned}
\boldsymbol{\theta}^* &= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\boldsymbol{X}) && (103) \\
&= \arg\max_{\boldsymbol{\theta}} \frac{p(\boldsymbol{X}, \boldsymbol{\theta})}{p(\boldsymbol{X})} && (104) \\
&= \arg\max_{\boldsymbol{\theta}} \frac{p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{X})} && (105) \\
&= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}) && (106)
\end{aligned}$$

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

$$
\begin{aligned}
\ln p(\boldsymbol{\theta}|\boldsymbol{X}) &= \ln \frac{p(\boldsymbol{X}, \boldsymbol{\theta})}{p(\boldsymbol{X})} && (107) \\
&= \ln \frac{p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{X})} \\
&= \ln p(\boldsymbol{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - p(\boldsymbol{X}) && (108) \\
&= \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q||p) + \ln p(\boldsymbol{\theta}) - p(\boldsymbol{X}) && (109) \\
&\geq \mathcal{L}(q, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - p(\boldsymbol{X}) && (110) \\
&= \mathcal{L}(q, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) + \mathrm{Const.} && (111)
\end{aligned}
$$

- $\ln p(\boldsymbol{X})$ 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

$$p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}) = \prod_i p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \tag{112}$$

  - 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)
$$d = \gamma^{\mathrm{new}}(z_{mk}) - \gamma^{\mathrm{old}}(z_{mk}) \tag{113}$$
  - 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)
$$N_k^{\mathrm{new}} = N_k^{\mathrm{old}} + \gamma^{\mathrm{new}}(z_{mk}) - \gamma^{\mathrm{old}}(z_{mk}) = N_k^{\mathrm{old}} + d \tag{114}$$

逐次型の EM アルゴリズムの例

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

$$\boldsymbol{\mu}_k^{\mathrm{old}} = \frac{1}{N_k^{\mathrm{old}}} \sum_i \gamma^{\mathrm{old}}(z_{ik}) \boldsymbol{x}_i \tag{115}$$

$$\boldsymbol{\Sigma}_k^{\mathrm{old}} = \frac{1}{N_k^{\mathrm{old}}} \sum_i \gamma^{\mathrm{old}}(z_{ik}) (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{\mathrm{old}})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{\mathrm{old}})^T \tag{116}$$

$$\pi_k^{\mathrm{old}} = \frac{N_k^{\mathrm{old}}}{N} \tag{117}$$

但し

$$N_k^{\mathrm{old}} = \sum_i \gamma^{\mathrm{old}}(z_{ik}) \tag{118}$$

逐次型の EM アルゴリズムの例

- 平均の更新式は

$$
\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k^{\text{new}}}\left(\sum_{i\neq m}\gamma^{\text{old}}(z_{ik})\boldsymbol{x}_i + \gamma^{\text{new}}(z_{mk})\boldsymbol{x}_m\right) && (119)\\
&= \frac{1}{N_k^{\text{new}}}\left(N_k^{\text{old}}\boldsymbol{\mu}_k^{\text{old}} - \gamma^{\text{old}}(z_{mk})\boldsymbol{x}_m + \gamma^{\text{new}}(z_{mk})\boldsymbol{x}_m\right)\\
&= \frac{1}{N_k^{\text{new}}}\left(\left(N_k^{\text{new}} - \gamma^{\text{new}}(z_{mk}) + \gamma^{\text{old}}(z_{mk})\right)\boldsymbol{\mu}_k^{\text{old}} - \right.\\
&\qquad \left.\gamma^{\text{old}}(z_{mk})\boldsymbol{x}_m + \gamma^{\text{new}}(z_{mk})\boldsymbol{x}_m\right)\\
&= \boldsymbol{\mu}_k^{\text{old}} + \frac{1}{N_k^{\text{new}}}\left(-(\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}))\boldsymbol{\mu}_k^{\text{old}} + \right.\\
&\qquad \left.(\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}))\boldsymbol{x}_m\right)\\
&= \boldsymbol{\mu}_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}}\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right) && (120)
\end{aligned}
$$

逐次型の EM アルゴリズムの例

$$= \boldsymbol{\mu}_k^{\text{old}} + \frac{d}{N_k^{\text{new}}} \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right) \tag{121}$$

- 分散の更新式は

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k^{\text{new}}} \sum_i \gamma^{\text{new}}(z_{ik}) \boldsymbol{x}_i \boldsymbol{x}_i^T - \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \tag{122}$$

$$= \frac{1}{N_k^{\text{new}}} \left(\sum_{i \neq m} \gamma^{\text{old}}(z_{ik}) \boldsymbol{x}_i \boldsymbol{x}_i^T + \gamma^{\text{new}}(z_{mk}) \boldsymbol{x}_m \boldsymbol{x}_m^T \right) - \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \tag{123}$$

$$= \frac{1}{N_k^{\text{new}}} \left(\left(\sum_i \gamma^{\text{old}}(z_{ik}) \boldsymbol{x}_i \boldsymbol{x}_i^T - \gamma^{\text{old}}(z_{mk}) \boldsymbol{x}_m \boldsymbol{x}_m^T \right) + \gamma^{\text{new}}(z_{mk}) \boldsymbol{x}_m \boldsymbol{x}_m^T \right) - \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \tag{124}$$

逐次型の EM アルゴリズムの例

$$
\begin{aligned}
&= \frac{1}{N_k^{\mathrm{new}}}\Big(N_k^{\mathrm{old}}\left(\boldsymbol{\Sigma}_k^{\mathrm{old}} + \boldsymbol{\mu}_k^{\mathrm{old}}(\boldsymbol{\mu}_k^{\mathrm{old}})^T\right) - \\
&\qquad \gamma^{\mathrm{old}}(z_{mk})\boldsymbol{x}_m\boldsymbol{x}_m^T + \gamma^{\mathrm{new}}(z_{mk})\boldsymbol{x}_m\boldsymbol{x}_m^T\Big) - \\
&\qquad \boldsymbol{\mu}_k^{\mathrm{new}}(\boldsymbol{\mu}_k^{\mathrm{new}})^T \qquad (125)\\
&= \frac{1}{N_k^{\mathrm{new}}}\Big(N_k^{\mathrm{old}}\boldsymbol{\Sigma}_k^{\mathrm{old}} + N_k^{\mathrm{old}}\boldsymbol{\mu}_k^{\mathrm{old}}(\boldsymbol{\mu}_k^{\mathrm{old}})^T + \\
&\qquad \left(\gamma^{\mathrm{new}}(z_{mk}) - \gamma^{\mathrm{old}}(z_{mk})\right)\boldsymbol{x}_m\boldsymbol{x}_m^T - \\
&\qquad N_k^{\mathrm{new}}\boldsymbol{\mu}_k^{\mathrm{new}}(\boldsymbol{\mu}_k^{\mathrm{new}})^T\Big) \\
&= \frac{1}{N_k^{\mathrm{new}}}\Big(N_k^{\mathrm{old}}\boldsymbol{\Sigma}_k^{\mathrm{old}} + N_k^{\mathrm{old}}\boldsymbol{\mu}_k^{\mathrm{old}}(\boldsymbol{\mu}_k^{\mathrm{old}})^T + \\
&\qquad d\boldsymbol{x}_m\boldsymbol{x}_m^T - N_k^{\mathrm{new}}\boldsymbol{\mu}_k^{\mathrm{new}}(\boldsymbol{\mu}_k^{\mathrm{new}})^T\Big) \qquad (126)
\end{aligned}
$$

逐次型の EM アルゴリズムの例

ここで

$$
\begin{aligned}
& N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \\
= & \left(N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{old}} + d(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}})\right) (\boldsymbol{\mu}_k^{\text{new}})^T \\
= & \frac{1}{N_k^{\text{new}}} \left(N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{old}} + d(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}})\right) \\
& \quad \left(N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{old}} + d\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\right)^T \\
= & N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T + 2\boldsymbol{\mu}_k^{\text{old}} d\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T + \\
& \quad \frac{1}{N_k^{\text{new}}} d^2 \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right) \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \qquad (127) \\
= & (N_k^{\text{old}} + d) \boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T + \\
& \quad 2d\boldsymbol{\mu}_k^{\text{old}} \boldsymbol{x}_m^T - 2d\boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T + \\
& \quad \frac{1}{N_k^{\text{new}}} d^2 \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right) \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \qquad (128)
\end{aligned}
$$

逐次型の EM アルゴリズムの例

であるから

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

$$\begin{aligned}
& \frac{1}{N_k^{\text{new}}} d^2 \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \\
= & \frac{d}{N_k^{\text{new}}} \left(N_k^{\text{new}} - d\right)\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \\
= & \frac{d}{N_k^{\text{new}}} N_k^{\text{old}} \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \qquad (129)
\end{aligned}$$

以上より

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

逐次型の EM アルゴリズムの例

$$
\begin{aligned}
&= \frac{N_k^{\mathrm{old}}}{N_k^{\mathrm{new}}} \left(\boldsymbol{\Sigma}_k^{\mathrm{old}} + \frac{d}{N_k^{\mathrm{new}}} \right. \\
&\qquad \left. \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\mathrm{old}}\right) \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\mathrm{old}}\right)^T \right) \\
&= \frac{N_k^{\mathrm{old}}}{N_k^{\mathrm{new}}} \left(\boldsymbol{\Sigma}_k^{\mathrm{old}} + \frac{\gamma^{\mathrm{new}}(z_{ik}) - \gamma^{\mathrm{old}}(z_{ik})}{N_k^{\mathrm{new}}} \right. \\
&\qquad \left. \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\mathrm{old}}\right) \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\mathrm{old}}\right)^T \right) \qquad (130)
\end{aligned}
$$

- 混合係数の更新式は

$$
\pi_k^{\mathrm{new}} = \frac{N_k^{\mathrm{new}}}{N} = \frac{N_k^{\mathrm{old}} + \gamma^{\mathrm{new}}(z_{mk}) - \gamma^{\mathrm{old}}(z_{mk})}{N} \qquad (131)
$$

- 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

$$\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right) \tag{132}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{N_k^{\text{old}}}{N_k^{\text{new}}} \left(\boldsymbol{\Sigma}_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{ik}) - \gamma^{\text{old}}(z_{ik})}{N_k^{\text{new}}} \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right) \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \right) \tag{133}$$

$$\pi_k^{\text{new}} = \frac{N_k^{\text{new}}}{N} = \frac{N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N} \tag{134}$$

但し

$$N_k^{\text{new}} = N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \tag{135}$$

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

EM アルゴリズムのまとめ

- 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

- EM アルゴリズムで行っていること
  - 対数尤度 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の直接の最適化が困難であっても、E ステップと M ステップという 2 段階の簡単な手続きに分割し、交互に繰り返すことで最適化できるようにする
  - 完全データ対数尤度 $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の事後確率 $\ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ による期待値 $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}})$ の最大化を行う
  - 期待値の最大化は、$\mathcal{L}(q, \boldsymbol{\theta})$ の最大化と等価である
  - $\mathcal{L}(q, \boldsymbol{\theta})$ は $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の下界であるから、$\mathcal{L}$ の最大化は、対数尤度 $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の最大化に相当

EM アルゴリズムのまとめ

- ここまでの話の流れ
  - 発見的に導出した、混合ガウスモデルに対する EM アルゴリズムも、期待値の最大化という考え方で解釈可能であった
  - K-Means 法は、混合ガウスモデルに対する EM アルゴリズムの一種の極限として得られた
  - 一般的な EM アルゴリズムの取り扱いについて調べた
  - 最尤推定だけでなく、MAP 推定に対しても EM アルゴリズムを適用できた
  - 混合ガウスモデルに対する、逐次版の EM アルゴリズムを導出した