

① EM アルゴリズム

1 EM アルゴリズム

- EM アルゴリズムの解釈
- 混合ガウス分布の再解釈
- K-Means 法との関連
- 一般の EM アルゴリズム
- MAP 推定に対する EM アルゴリズム
- EM アルゴリズムの拡張
- EM アルゴリズムに関する補足
- 逐次型の EM アルゴリズム

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 ($\gamma(z_{ik}) = p(z_k = 1|x_i)$)

- 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 \mathbf{X} とする (第 i 行が \mathbf{x}_i^T)
 - 全ての潜在変数をまとめた行列を \mathbf{Z} とする (第 i 行が \mathbf{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

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

EM アルゴリズムの解釈

- 潜在変数 z が連続変数の場合は

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

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 $p(\mathbf{X}|\boldsymbol{\theta})$ 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
 - X だけでなく、 Z も観測できるとする
 - $\{X, Z\}$ の組を、**完全データ集合**という
 - 実際には X しか見えないので、実際の観測データ X は**不完全**である
- Z に関する知識は、潜在変数についての事後確率分布 $p(Z|X, \theta)$ のみからしか得られない

重要な仮定と考え方

- 1 完全データ対数尤度関数 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ の最大化は、 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ の最大化よりも、簡単であると仮定
- 2 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ の代わりに、完全データ対数尤度関数 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ を最大化したいが、 \mathbf{Z} に関する情報は $\ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ 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 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ は使えない
- 4 そこで、事後確率分布 $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ による、 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ 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 アルゴリズムへの落とし込み
 - パラメータ θ を適当に初期化する
 - **E ステップ**では、事後確率分布 $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ を、現在のパラメータ θ^{old} を使って求める
 - $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ 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 $\ln p(\mathbf{X}, \mathbf{Z}|\theta)$ の、事後確率分布 $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ に関する期待値 $\mathcal{Q}(\theta, \theta^{\text{old}})$ を計算

$$\mathcal{Q}(\theta, \theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\theta) \quad (3)$$

連続潜在変数の場合は次のようになる

$$\mathcal{Q}(\theta, \theta^{\text{old}}) = \int p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\theta) d\mathbf{Z} \quad (4)$$

EM アルゴリズムの解釈

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

$$\theta^{\text{new}} = \arg \max_{\theta} Q(\theta, \theta^{\text{old}}) \quad (5)$$

- 注意点
 - $Q(\theta, \theta^{\text{old}})$ において、対数 \ln は、同時分布 $p(\mathbf{X}, \mathbf{Z}|\theta)$ 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 $p(\mathbf{Z}|\mathbf{X}, \theta)$ 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

一般の EM アルゴリズム

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

- パラメータを θ^{old} に初期化する
- E ステップ**: 事後確率分布 $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ を計算する

- 3 **M ステップ**: 次式で与えられる θ^{new} を計算する

$$\theta^{\text{new}} = \arg \max_{\theta} Q(\theta, \theta^{\text{old}}) \quad (6)$$

但し

$$Q(\theta, \theta^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\theta) \quad (7)$$

- 4 対数尤度の変化量、あるいはパラメータの変化量を見て、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

$$\theta^{\text{old}} \leftarrow \theta^{\text{new}} \quad (8)$$

1 EM アルゴリズム

- EM アルゴリズムの解釈
- 混合ガウス分布の再解釈
- K-Means 法との関連
- 一般の EM アルゴリズム
- MAP 推定に対する EM アルゴリズム
- EM アルゴリズムの拡張
- EM アルゴリズムに関する補足
- 逐次型の EM アルゴリズム

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

$$\begin{aligned} & p(X, Z|\theta) \\ = & p(Z|\theta)p(X|Z, \theta) \\ = & \prod_i p(z_i|\theta)p(x_i|z_i, \theta) \\ = & \prod_i \left(\prod_k \pi_k^{z_{ik}} \right) \left(\prod_k \mathcal{N}(x_i|\mu_k, \Sigma_k)^{z_{ik}} \right) \\ = & \prod_i \prod_k \pi_k^{z_{ik}} \mathcal{N}(x_i|\mu_k, \Sigma_k)^{z_{ik}} \\ = & \prod_i \prod_k (\pi_k \mathcal{N}(x_i|\mu_k, \Sigma_k))^{z_{ik}} \end{aligned} \tag{9}$$

混合ガウス分布の再解釈

- ここで、データ点 x_i に対応する潜在変数を z_i 、また z_i の k 番目の要素を z_{ik} とする
- データ点 x_i, z_i は、 $p(X, Z|\theta)$ 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

$$\begin{aligned} & \ln p(X, Z|\theta) \\ &= \ln \left(\prod_i \prod_k (\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k))^{z_{ik}} \right) \\ &= \sum_i \sum_k \ln ((\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k))^{z_{ik}}) \\ &= \sum_i \sum_k z_{ik} \ln (\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k)) \\ &= \sum_i \sum_k z_{ik} (\ln \pi_k + \ln \mathcal{N}(x_i | \mu_k, \Sigma_k)) \end{aligned} \tag{10}$$

混合ガウス分布の再解釈

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

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

- $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ と $\ln p(\mathbf{X}|\boldsymbol{\theta})$ 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 $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ に**直接作用している**
- よって、 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ の最大化は、 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ 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、 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ 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

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

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

混合ガウス分布の再解釈

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

これより

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

混合ガウス分布の再解釈

であるから、両辺に左から Σ_k を掛けて

$$\begin{aligned}\sum_i z_{ik} \mu_k &= \sum_i z_{ik} \mathbf{x}_i \\ \mu_k \sum_i z_{ik} &= \sum_i z_{ik} \mathbf{x}_i \\ \mu_k &= \frac{1}{\sum_i z_{ik}} \sum_i z_{ik} \mathbf{x}_i\end{aligned}\tag{14}$$

のようになる

- 上式をみると、完全データ $\{X, Z\}$ について、 μ_k は陽な形で求まっていることが分かる
- 但し実際は 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

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

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

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

$$\frac{\partial}{\partial \mathbf{X}} \ln |\mathbf{X}| = (\mathbf{X}^{-1})^T \quad (16)$$

- これより

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

であるから、両辺に左右から Σ_k を掛けて

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

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 $\{X, Z\}$ について、 Σ_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

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

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

- π_k について微分して 0 とおくと、次のようになる

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

混合ガウス分布の再解釈

$$= \sum_i z_{ik} \frac{1}{\pi_k} + \lambda = 0 \quad (20)$$

- これより、両辺に π_k を掛けて

$$\sum_i z_{ik} + \lambda \pi_k = 0 \quad (21)$$

全ての k について総和を取ると

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

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、陽な形で求める
- EM アルゴリズムにおける μ_k, Σ_k, π_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 $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ に関する期待値の計算
 - 完全データ対数尤度関数 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ 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 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ の、事後確率分布 $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ 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

$$\begin{aligned} & p(z_i | x_i, \theta) \\ = & \frac{p(x_i | z_i, \theta) p(z_i | \theta)}{p(x_i | \theta)} \end{aligned} \quad (24)$$

$$\begin{aligned} \propto & p(x_i | z_i, \theta) p(z_i | \theta) \\ & (\because p(x_i | \theta) \text{ は、} z_i \text{ には依存しない定数項}) \end{aligned} \quad (25)$$

$$\begin{aligned} = & \left(\prod_k \mathcal{N}(x_i | \mu_k, \Sigma_k)^{z_{ik}} \right) \left(\prod_k \pi_k^{z_{ik}} \right) \\ = & \prod_k (\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k))^{z_{ik}} \end{aligned} \quad (26)$$

以上より

$$p(z_i | x_i, \theta) \propto \prod_k (\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k))^{z_{ik}} \quad (27)$$

混合ガウス分布の再解釈

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

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

- $p(z_i | \mathbf{x}_i, \boldsymbol{\theta})$ を等式で表すためには、 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**

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

混合ガウス分布の再解釈

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

$$\begin{aligned} & \mathbb{E}_{z_i \sim p(z_i | x_i, \theta)} [z_{ik}] \\ &= \sum_{z_i} z_{ik} p(z_i | x_i, \theta) \\ &= \sum_{z_i} z_{ik} \frac{\prod_k (\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k))^{z_{ik}}}{\sum_{z_i} \prod_k (\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k))^{z_{ik}}} \end{aligned} \quad (30)$$

- ここで

$$\sum_{z_i} \prod_k (\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k))^{z_{ik}} = \sum_k (\pi_k \mathcal{N}(x_i | \mu_k, \Sigma_k)) \quad (31)$$

と書けることに注意する

- 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

- $\prod_k (\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{z_{ik}}$ は、 $z_{ik} = 1$ の場合、 $j \neq k$ に対して $z_{ij} = 0$ であるから、 $\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ という単一の項として書ける
- 全ての z_i についての総和は、 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

$$\sum_{z_i} z_{ik} \prod_k (\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{z_{ik}} = \pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (32)$$

であることにも注意する

- \sum_{z_i} の総和の中身は、 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、 z_i が $z_{ik} = 1$ となるときの項 $\prod_k (\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))^{z_{ik}}$ 、即ち $\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ だけが出現する

混合ガウス分布の再解釈

- これより

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

であるから、データ点 \mathbf{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 $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ に関する、完全データ対数尤度関数 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ の期待値は

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

$$= \sum_i \sum_k \mathbb{E}_{\mathbf{z}_i \sim p(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\theta})} [z_{ik}] (\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)) \quad (35)$$

$$= \sum_i \sum_k \gamma(z_{ik}) (\ln \pi_k + \ln \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)) \quad (36)$$

である

混合ガウス分布の再解釈

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

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

- 先ほどは、 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ 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 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ の最大化よりも、完全データ対数尤度関数 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ 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 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ の代わりに、 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ の最大化を考えた
- 4 しかし \mathbf{Z} に関する情報がないので、代わりに、事後確率分布 $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ による、 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ の期待値を最大化しようとするのが、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

1 EM アルゴリズム

- EM アルゴリズムの解釈
- 混合ガウス分布の再解釈
- K-Means 法との関連
- 一般の EM アルゴリズム
- MAP 推定に対する EM アルゴリズム
- EM アルゴリズムの拡張
- EM アルゴリズムに関する補足
- 逐次型の EM アルゴリズム

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 $\gamma(z_{ik}) \equiv p(z_k = 1 | \mathbf{x}_i)$ 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 法の導出

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

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

$$= \frac{1}{(2\pi\epsilon)^{\frac{D}{2}}} \exp \left\{ -\frac{1}{2\epsilon}(\mathbf{x} - \boldsymbol{\mu}_k)^T (\mathbf{x} - \boldsymbol{\mu}_k) \right\} \quad (39)$$

$$\begin{aligned} & (\because |\epsilon \mathbf{I}|^{\frac{1}{2}} = (\epsilon^D |\mathbf{I}|)^{\frac{1}{2}} = (\epsilon^D)^{\frac{1}{2}} = \epsilon^{\frac{D}{2}}) \\ &= \frac{1}{(2\pi\epsilon)^{\frac{D}{2}}} \exp \left\{ -\frac{1}{2\epsilon} \|\mathbf{x} - \boldsymbol{\mu}_k\|^2 \right\} \quad (40) \end{aligned}$$

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_k \pi_k p(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (41)$$

$$= \sum_k \pi_k \frac{1}{(2\pi\epsilon)^{\frac{D}{2}}} \exp \left\{ -\frac{1}{2\epsilon} \|\mathbf{x} - \boldsymbol{\mu}_k\|^2 \right\} \quad (42)$$

- 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 \mathbf{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、 $\epsilon \rightarrow 0$ についての極限を取ってみる

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

- 負担率は、以下のように変形できる

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

- ここで、 k^* を次で定める

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

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

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

を考えると、全ての $j \neq k^*$ について

$$\frac{\exp \{-\|\mathbf{x} - \boldsymbol{\mu}_j\|^2\}}{\exp \{-\|\mathbf{x} - \boldsymbol{\mu}_k\|^2\}} < 1 \quad (47)$$

が成立するので

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

である

- 従って、 $k = k^*$ のとき

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

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

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

- $k \neq k^*$ のとき

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

であって、両辺の $\epsilon \rightarrow 0$ による極限を取れば

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

となるから

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

が明らかに成立するほか、以下の不等式が

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

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

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

従って、 $k \neq k^*$ の場合は

$$\lim_{\epsilon \rightarrow 0} \gamma(z_{ik}) = 0 \quad (55)$$

である

K-Means 法との関連

- これより、データ点 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

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

- 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
- $k^* = \arg \min_j \|x - \mu_j\|^2$ であるから、結局、各データ点は、**平均ベクトル μ 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 μ_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

$$\text{K-Means :} \quad \mu_k = \frac{1}{\sum_i r_{ik}} \sum_i r_{ik} \mathbf{x}_i \quad (57)$$

$$\text{EM アルゴリズム :} \quad \mu_k = \frac{1}{\sum_i \gamma(z_{ik})} \sum_i \gamma(z_{ik}) \mathbf{x}_i \quad (58)$$

- 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 $\epsilon \mathbf{I}$ としたとき、 $\epsilon \rightarrow 0$ の極限を取ると、K-Means 法が得られる

- 期待完全データ対数尤度の計算

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

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

- 両辺に ϵ を掛けると

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

- $\epsilon \rightarrow 0$ の極限を取ると

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

であるから

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

K-Means 法との関連

$$= -\frac{1}{2} \sum_i \sum_k r_{ik} \|\mathbf{x}_i - \boldsymbol{\mu}_k\|^2 \quad (62)$$

$$= -J \quad (63)$$

- よって、期待完全データ対数尤度 $\mathbb{E}_{\mathbf{Z}} [\ln p(\mathbf{X}, \mathbf{Z} | \boldsymbol{\theta})]$ 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 ϵI 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

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

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

$$\pi_k = \frac{\sum_i r_{ik}}{N} = \frac{N_k}{N} \quad (65)$$

- 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

- $\ln p(\mathbf{X}|\boldsymbol{\theta})$ の代わりに、 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ を最大化してもよい根拠を明らかにする
- $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ の、 $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ 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 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ 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

1 EM アルゴリズム

- EM アルゴリズムの解釈
- 混合ガウス分布の再解釈
- K-Means 法との関連
- 一般の EM アルゴリズム
- MAP 推定に対する EM アルゴリズム
- EM アルゴリズムの拡張
- EM アルゴリズムに関する補足
- 逐次型の EM アルゴリズム

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 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ 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 $p(\mathbf{X}|\boldsymbol{\theta})$ の記述
 - 全ての観測変数と、潜在変数をそれぞれ \mathbf{X}, \mathbf{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 $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ とすると、尤度関数は次のようになる

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

- 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

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

- ここでは、連続潜在変数の場合を考える
- 重要な仮定**
 - $\ln p(\mathbf{X}|\boldsymbol{\theta})$ の最大化よりも、完全データ対数尤度関数 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ の最大化の方が、容易である
 - 以前に見た尤度関数 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ 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

- 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

- $\ln p(\mathbf{X}|\boldsymbol{\theta})$ ではなく $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ を最適化しようとしたが、 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 $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ による、 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ の期待値 $\mathbb{E}_Z [\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})]$ 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

$$f\left(\sum_i \lambda_i x_i\right) \leq \sum_i \lambda_i f(x_i) \quad (68)$$

- 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

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

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

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

例えば、 $f(x) = -\ln x$ は凸関数であるから

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

よって

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

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

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

- 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、入力として関数を取り、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 $p(\mathbf{x})$ と $q(\mathbf{x})$ の間の、カルバック-ライブラーダイバージェンスを、 $\text{KL}(p||q)$ と表す
- 確率分布 $p(\mathbf{x})$ と $q(\mathbf{x})$ の間の、(擬似的な) 距離を表す指標である

$$\text{KL}(p||q) = - \int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} d\mathbf{x} \quad (74)$$

- $\text{KL}(p||q) \geq 0$ であり、等号成立は $p(\mathbf{x}) = q(\mathbf{x})$ 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 $\text{KL}(p||q) \neq \text{KL}(q||p)$ となる

一般の EM アルゴリズム

- $\ln p(\mathbf{X}|\boldsymbol{\theta})$ の分解

- EM アルゴリズムについて考察するために、まずは $\ln p(\mathbf{X}|\boldsymbol{\theta})$ を分解してみよう
- 潜在変数についての分布を $q(\mathbf{Z})$ とおく
- $q(\mathbf{Z})$ の設定の仕方によらず、 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ を次のように分解できる

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

- $\mathcal{L}(q, \boldsymbol{\theta})$ は、分布 $q(\mathbf{Z})$ の汎関数であり、かつパラメータ $\boldsymbol{\theta}$ の関数である
- $\text{KL}(q||p)$ は、確率分布 $q(\mathbf{Z})$ と $p(\mathbf{X}|\boldsymbol{\theta})$ 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

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

一般の EM アルゴリズム

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

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

$$\text{KL}(q||p) = - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z} | \mathbf{X}, \theta)}{q(\mathbf{Z})} \quad (78)$$

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

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

- $\mathcal{L}(q, \theta)$ は、 $q(\mathbf{Z}), \theta$ によらず、常に $\ln p(\mathbf{X} | \theta)$ の下界をなす
- 下界について確認した後に、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 $\text{KL}(q||p) \geq 0$, we see that the quantity $\mathcal{L}(q, \theta)$ is a lower bound on the log likelihood function $\ln p(\mathbf{X}|\theta)$.

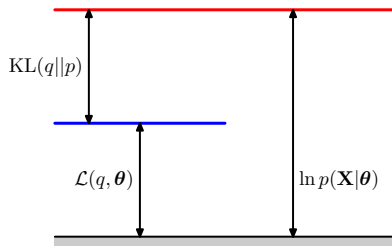


図 1: $\ln p(\mathbf{X}|\theta)$ の分解

下界と上界の定義

- 下界と上界の定義

- 下界と上界は、次のように定義される (実数の集合を \mathbb{R} とする)

実数 $a \in \mathbb{R}$ が、実数の部分集合 $A \subset \mathbb{R}$ の上界である

$$\Leftrightarrow \text{集合 } A \text{ に属する任意の要素 } x \text{ は、} a \text{ 以下である}$$
$$\Leftrightarrow \forall x \in A, x \leq a \quad (80)$$

実数 $a \in \mathbb{R}$ が、実数の部分集合 $A \subset \mathbb{R}$ の下界である

$$\Leftrightarrow \text{集合 } A \text{ に属する任意の要素 } x \text{ は、} a \text{ 以上である}$$
$$\Leftrightarrow \forall x \in A, a \leq x \quad (81)$$

- 上記の定義は、実数の集合 \mathbb{R} だけでなく、一般の半順序集合について成り立つ
- 例えば、集合 A を $A = \{x | 0 \leq x\}$ と定義すると、0 以下である実数、例えば -1 、 -3 などは、いずれも集合 A の下界となる

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、 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ 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 $\boldsymbol{\theta}^{\text{old}}$ とする

- **E ステップ**

- E ステップでは、下界 $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ を、 $\boldsymbol{\theta}^{\text{old}}$ を固定しながら、 $q(\mathbf{Z})$ について最大化する
- この問題は、 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ の分解をみれば簡単に解ける

$$\begin{aligned} & \ln p(\mathbf{X}|\boldsymbol{\theta}^{\text{old}}) \\ = & \mathcal{L}(q, \boldsymbol{\theta}^{\text{old}}) + \text{KL}(q||p) \end{aligned} \tag{82}$$

$$\begin{aligned} = & \mathcal{L}(q, \boldsymbol{\theta}^{\text{old}}) + \text{KL}(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})) \\ & (\text{E ステップ前}) \end{aligned} \tag{83}$$

一般の EM アルゴリズム

- 上式において、左辺の $\ln p(\mathbf{X}|\boldsymbol{\theta}^{\text{old}})$ 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、 $\text{KL}(q||p)$ を最小化するしかない
- $\text{KL}(q||p)$ を最小化するためには、 $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$ とおいて、 $\text{KL}(q||p) = 0$ とすればよい ($\text{KL}(q||p) \geq 0$ 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 $\ln p(\mathbf{X}|\boldsymbol{\theta}^{\text{old}})$ に一致する

$$\ln p(\mathbf{X}|\boldsymbol{\theta}^{\text{old}}) \tag{84}$$

$$\begin{aligned} &= \mathcal{L}(q, \boldsymbol{\theta}^{\text{old}}) + \text{KL}(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})) \\ &= \mathcal{L}(q, \boldsymbol{\theta}^{\text{old}}) + \text{KL}(p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})||p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})) \end{aligned} \tag{85}$$

$$\begin{aligned} &= \mathcal{L}(q, \boldsymbol{\theta}^{\text{old}}) \\ &\quad (\text{E ステップ後}) \end{aligned} \tag{86}$$

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
- $KL(q||p) = 0$ 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 $\mathcal{L}(q, \theta^{\text{old}})$ が、赤線で示されている対数尤度 $\ln p(\mathbf{X}|\theta^{\text{old}})$ 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 θ^{old} , causing the lower bound to move up to the same value as the log likelihood function, with the KL divergence vanishing.

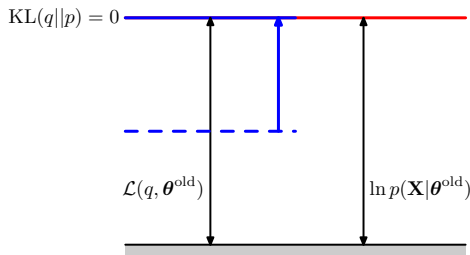


図 2: EM アルゴリズムの E ステップ

一般の EM アルゴリズム

● M ステップ

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

$$\begin{aligned} & \ln p(\mathbf{X}|\theta) \\ = & \mathcal{L}(q, \theta) + \text{KL}(q||p) \end{aligned} \tag{87}$$

$$= \mathcal{L}(q, \theta) + \text{KL}(q(\mathbf{Z})||p(\mathbf{Z}|\mathbf{X}, \theta)) \tag{88}$$

$$= \mathcal{L}(q, \theta) + \text{KL}(p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})||p(\mathbf{Z}|\mathbf{X}, \theta)) \tag{89}$$

(M ステップ前)

- 分布 $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ は、古いパラメータ θ^{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

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

$$\begin{aligned} & \ln p(\mathbf{X}|\boldsymbol{\theta}^{\text{new}}) \\ = & \mathcal{L}(q, \boldsymbol{\theta}^{\text{new}}) + KL(p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})||p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{new}})) \quad (90) \\ & (\text{M ステップ後}) \end{aligned}$$

- 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)

$$\begin{aligned} & \ln p(\mathbf{X}|\boldsymbol{\theta}^{\text{new}}) - \ln p(\mathbf{X}|\boldsymbol{\theta}^{\text{old}}) \\ = & \mathcal{L}(q, \boldsymbol{\theta}^{\text{new}}) + KL(p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})||p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{new}})) - \\ & \mathcal{L}(q, \boldsymbol{\theta}^{\text{old}}) \quad (91) \end{aligned}$$

$$\begin{aligned} = & (\mathcal{L}(q, \boldsymbol{\theta}^{\text{new}}) - \mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})) + \\ & KL(p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})||p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{new}})) \quad (92) \end{aligned}$$

$$\geq \mathcal{L}(q, \boldsymbol{\theta}^{\text{new}}) - \mathcal{L}(q, \boldsymbol{\theta}^{\text{old}}) \quad (93)$$

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 $\mathcal{L}(q, \theta)$ を、 $q(\mathbf{Z})$ を固定しつつ、 θ について最大化している
- 青の点線で示されている下界 $\mathcal{L}(q, \theta^{\text{old}})$ が、青の実線で示されている下界 $\mathcal{L}(q, \theta^{\text{new}})$ へと、持ち上げられている
- 赤の点線で示される対数尤度 $\ln p(\mathbf{X}|\theta^{\text{old}})$ は、赤の実線で示される対数尤度 $\ln p(\mathbf{X}|\theta^{\text{new}})$ へと、持ち上げられている
- 新たに生じた $\text{KL}(q||p)$ 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 $\mathcal{L}(q, \theta)$ is maximized with respect to the parameter vector θ to give a revised value θ^{new} . Because the KL divergence is nonnegative, this causes the log likelihood $\ln p(\mathbf{X}|\theta)$ to increase by at least as much as the lower bound does.

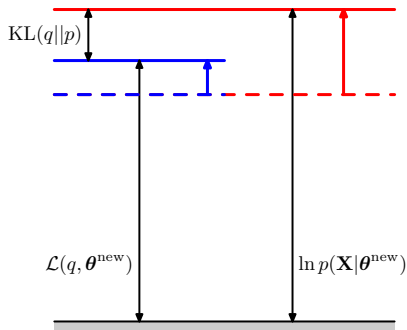


図 3: EM アルゴリズムの M ステップ

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 $\mathcal{L}(q, \theta)$ を、 q を固定しつつ θ 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** $\mathcal{L}(q, \theta)$ であり、これは次のように表せる ($q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ である)

$$\begin{aligned}\mathcal{L}(q, \theta) &= \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{q(\mathbf{Z})} \\ &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})} \quad (94)\end{aligned}$$

$$\begin{aligned}&= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\theta) - \\ &\quad \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \ln p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}}) \quad (95)\end{aligned}$$

$$= \mathcal{Q}(\theta, \theta^{\text{old}}) + \text{Const.} \quad (96)$$

一般の EM アルゴリズム

- 定数項は、単に分布 $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ のエントロピーであって、 θ 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 $\ln p(\mathbf{X}, \mathbf{Z}|\theta)$ の、事後確率分布 $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ による期待値 $Q(\theta, \theta^{\text{old}})$ である
- 最適化しようとしているパラメータ θ は、**対数の中にしか現れない**
- 同時分布 $p(\mathbf{X}, \mathbf{Z}|\theta)$ 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 $\ln p(\mathbf{X}|\theta)$ 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 $\mathcal{L}(q, \theta^{\text{old}})$ を、 θ^{old} を固定しつつ、 q について最大化する
- これは、単に $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ とすればよい
- 即ち、 $p(\mathbf{Z}|\mathbf{X}, \theta^{\text{old}})$ を計算するだけである

- M ステップのまとめ

- 下界 $\mathcal{L}(q, \theta)$ を、 q を固定しつつ、 θ について最大化する
- これは、期待値 $\mathcal{Q}(\theta, \theta^{\text{old}})$ を最大化するような、パラメータ θ 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

- $\ln p(\mathbf{X}|\boldsymbol{\theta})$ の代わりに、 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ を最大化してよい根拠
- そして、 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ の、 $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ 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 $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ 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 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ の下界である
- 下界を最大化することは、 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ 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、 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ の期待値を最適化させることは、 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ 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 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ 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

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

一般の EM アルゴリズム

$$\begin{aligned} &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{\text{new}})}{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{\text{old}})} + \\ &\quad \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{new}})} \\ &= \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{\text{new}}) - \\ &\quad \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^{\text{old}}) - \\ &\quad \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})}{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{new}})} \\ &= \mathcal{Q}(\boldsymbol{\theta}^{\text{old}}, \boldsymbol{\theta}^{\text{new}}) - \mathcal{Q}(\boldsymbol{\theta}^{\text{old}}, \boldsymbol{\theta}^{\text{old}}) + \\ &\quad \text{KL} (p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) || p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{new}})) \end{aligned} \tag{97}$$

$$\geq \mathcal{Q}(\boldsymbol{\theta}^{\text{old}}, \boldsymbol{\theta}^{\text{new}}) - \mathcal{Q}(\boldsymbol{\theta}^{\text{old}}, \boldsymbol{\theta}^{\text{old}}) \tag{98}$$

$$\geq 0$$

一般の EM アルゴリズム

- 最後の変形は、M ステップでは $Q(\theta, \theta^{\text{old}})$ を、 θ について最大化しているから、 $Q(\theta^{\text{old}}, \theta^{\text{new}}) \geq Q(\theta^{\text{old}}, \theta^{\text{old}})$ であることを利用
- 更新によって $\ln p(\mathbf{X}|\theta)$ 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

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

$$\begin{aligned}\ln p(\mathbf{X}|\boldsymbol{\theta}) &= \ln \sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) \\ &= \ln \sum_{\mathbf{Z}} q(\mathbf{Z}) \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \\ &\geq \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})}{q(\mathbf{Z})} \\ &= \mathcal{L}(q, \boldsymbol{\theta})\end{aligned}\tag{99}$$

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

一般の EM アルゴリズム

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

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

一般の EM アルゴリズム

$$\begin{aligned} &= - \sum_{\mathbf{Z}} q(\mathbf{Z}) \ln \frac{p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})}{q(\mathbf{Z})} \\ &= \text{KL}(q||p) \end{aligned} \tag{100}$$

ゆえ、 $\text{KL}(q||p)$ となることが分かったので

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \text{KL}(q||p) \tag{101}$$

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 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ を表す

- E ステップ

- パラメータの初期値 $\boldsymbol{\theta}^{\text{old}}$ 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 $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ を計算
- このとき、**青の実線**で示す下界 $\mathcal{L}(q, \boldsymbol{\theta}^{\text{old}})$ が q について更新され、下界 \mathcal{L} は、 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ と $\boldsymbol{\theta}^{\text{old}}$ において一致する
- 下界 \mathcal{L} の曲線は、 $\boldsymbol{\theta}^{\text{old}}$ において $\ln p(\mathbf{X}|\boldsymbol{\theta})$ と**接する**ことに注意する
- 下界 \mathcal{L} と対数尤度 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ は、 $\boldsymbol{\theta}^{\text{old}}$ 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 $\mathcal{L}(q, \theta)$ が θ について最大化されて、パラメータ θ^{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 $\mathcal{L}(q, \theta^{\text{new}})$ が計算される
- 下界 $\mathcal{L}(q, \theta^{\text{new}})$ は、 $\ln p(\mathbf{X}|\theta)$ と θ^{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、 θ による微分を考えれば明らか

$$\begin{aligned} & \left. \frac{\partial}{\partial \theta} \ln p(\mathbf{X}|\theta) \right|_{\theta^{\text{old}}} \\ &= \left. \frac{\partial}{\partial \theta} \mathcal{L}(q, \theta) \right|_{\theta^{\text{old}}} + \left. \frac{\partial}{\partial \theta} \text{KL}(q||p) \right|_{\theta^{\text{old}}} \\ &= \left. \frac{\partial}{\partial \theta} \mathcal{L}(q, \theta) \right|_{\theta^{\text{old}}} \end{aligned} \tag{102}$$

- E ステップによって $\text{KL}(q||p)$ が最小化されるので、 θ による勾配も当然 0 になるはずである
- このとき、 $\ln p(\mathbf{X}|\theta)$ と $\mathcal{L}(q, \theta)$ の、 θ^{old} における微分値が等しくなる
- 従って、 θ^{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 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ が、下界 \mathcal{L} を上回る ($\mathcal{L}(q, \boldsymbol{\theta}) > \ln p(\mathbf{X}|\boldsymbol{\theta})$) ような $\boldsymbol{\theta}$ が存在する
- これは、 $\text{KL}(q||p) < 0$ 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.

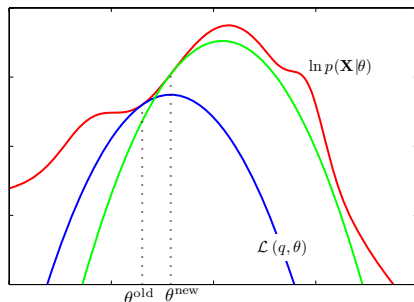


図 4: EM アルゴリズムの手続き

一般の EM アルゴリズム

- i.i.d 標本である場合

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

$$p(\mathbf{X}, \mathbf{Z}) = \prod_i p(x_i, z_i) \quad (103)$$

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

$$\begin{aligned} p(\mathbf{Z}|\mathbf{X}, \theta) &= \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{p(\mathbf{X}|\theta)} \\ &= \frac{p(\mathbf{X}, \mathbf{Z}|\theta)}{\sum_{\mathbf{Z}} p(\mathbf{X}, \mathbf{Z}|\theta)} \\ &= \frac{\prod_i p(x_i, z_i|\theta)}{\sum_{\mathbf{Z}} \prod_i p(x_i, z_i|\theta)} \end{aligned}$$

$$\begin{aligned} &= \frac{\prod_i p(\mathbf{x}_i, \mathbf{z}_i | \boldsymbol{\theta})}{\prod_i \sum_{\mathbf{z}} p(\mathbf{x}_i, \mathbf{z}_i | \boldsymbol{\theta})} \\ &= \frac{\prod_i p(\mathbf{x}_i, \mathbf{z}_i | \boldsymbol{\theta})}{\prod_i p(\mathbf{x}_i | \boldsymbol{\theta})} \\ &= \prod_i \frac{p(\mathbf{x}_i, \mathbf{z}_i | \boldsymbol{\theta})}{p(\mathbf{x}_i | \boldsymbol{\theta})} \\ &= \prod_i p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\theta}) \end{aligned} \tag{104}$$

各データ点に対する事後確率 $p(\mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\theta})$ の積として、 $p(\mathbf{Z} | \mathbf{X}, \boldsymbol{\theta})$ 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 \mathbf{x}_i に対する各ガウス分布の負担率は、データ \mathbf{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
 - $\ln p(\mathbf{X}|\boldsymbol{\theta})$ の代わりに、 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ を最大化してもよい根拠
 - $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ の、 $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ 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 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ 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

1 EM アルゴリズム

- EM アルゴリズムの解釈
- 混合ガウス分布の再解釈
- K-Means 法との関連
- 一般の EM アルゴリズム
- MAP 推定に対する EM アルゴリズム
- EM アルゴリズムの拡張
- EM アルゴリズムに関する補足
- 逐次型の EM アルゴリズム

MAP 推定に対する EM アルゴリズム

- 事後分布の対数 $\ln p(\boldsymbol{\theta}|\mathbf{X})$ の最大化
 - 今までは、尤度関数 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ 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 $p(\boldsymbol{\theta})$ 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 $p(\boldsymbol{\theta}|\mathbf{X})$ を最大化するパラメータ $\boldsymbol{\theta}^*$ を求める問題である

$$\boldsymbol{\theta}^* = \arg \max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{X}) \quad (105)$$

$$= \arg \max_{\boldsymbol{\theta}} \frac{p(\mathbf{X}, \boldsymbol{\theta})}{p(\mathbf{X})} \quad (106)$$

$$= \arg \max_{\boldsymbol{\theta}} \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X})} \quad (107)$$

$$= \arg \max_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \quad (108)$$

MAP 推定に対する EM アルゴリズム

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

$$\ln p(\boldsymbol{\theta}|\mathbf{X}) = \ln \frac{p(\mathbf{X}, \boldsymbol{\theta})}{p(\mathbf{X})} \quad (109)$$

$$\begin{aligned} &= \ln \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X})} \\ &= \ln p(\mathbf{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - p(\mathbf{X}) \end{aligned} \quad (110)$$

$$= \mathcal{L}(q, \boldsymbol{\theta}) + \text{KL}(q||p) + \ln p(\boldsymbol{\theta}) - p(\mathbf{X}) \quad (111)$$

$$\geq \mathcal{L}(q, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - p(\mathbf{X}) \quad (112)$$

$$= \mathcal{L}(q, \boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) + \text{Const.} \quad (113)$$

- $\ln p(\mathbf{X})$ は定数とみなせるから、 $\ln p(\boldsymbol{\theta}|\mathbf{X})$ 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 θ を固定しつつ、 q について $\mathcal{L}(q, \theta)$ を最大化する
- q は下界 $\mathcal{L}(q, \theta)$ 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 θ について $\mathcal{L}(q, \theta) + \ln p(\theta)$ を最大化する
- 事前分布の項 $\ln p(\theta)$ 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、少ししか変わらない

1 EM アルゴリズム

- EM アルゴリズムの解釈
- 混合ガウス分布の再解釈
- K-Means 法との関連
- 一般の EM アルゴリズム
- MAP 推定に対する EM アルゴリズム
- EM アルゴリズムの拡張
- EM アルゴリズムに関する補足
- 逐次型の EM アルゴリズム

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 $\ln p(\mathbf{X}|\theta)$ 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 $\mathcal{L}(q, \theta)$ を θ について**最大化するのは諦める**代わりに、下界 $\mathcal{L}(q, \theta)$ を**少しでも増加させるように**、 θ を更新する
- $\mathcal{L}(q, \theta)$ は、**常に**尤度関数 $\ln p(\mathbf{X}|\theta)$ 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

1 EM アルゴリズム

- EM アルゴリズムの解釈
- 混合ガウス分布の再解釈
- K-Means 法との関連
- 一般の EM アルゴリズム
- MAP 推定に対する EM アルゴリズム
- EM アルゴリズムの拡張
- EM アルゴリズムに関する補足
- 逐次型の EM アルゴリズム

パラメータについての補足

- パラメータ θ についての補足

- 任意の θ について、下界 $\mathcal{L}(q, \theta)$ は q について**唯一の最大点**をもつ
- それは事後分布 $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{X}, \theta)$ である
- またこのとき、下界 \mathcal{L} は対数尤度関数 $\ln p(\mathbf{X}|\theta)$ に一致する
- θ が下界 $\mathcal{L}(q, \theta)$ の大域的最適解に収束するなら、そのような θ は、対数尤度関数 $\ln p(\mathbf{X}|\theta)$ の大域的最適解でもある
- 任意の下界 $\mathcal{L}(q, \theta)$ の任意の極大点は、 $\ln p(\mathbf{X}|\theta)$ の極大点でもある

1 EM アルゴリズム

- EM アルゴリズムの解釈
- 混合ガウス分布の再解釈
- K-Means 法との関連
- 一般の EM アルゴリズム
- MAP 推定に対する EM アルゴリズム
- EM アルゴリズムの拡張
- EM アルゴリズムに関する補足
- 逐次型の EM アルゴリズム

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 $p(Z|X, \theta)$ 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 $p(z_i|x_i, \theta)$ の積として分解できる

$$p(Z|X, \theta) = \prod_i p(z_i|x_i, \theta) \quad (114)$$

- このとき、**全てのデータ点** 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 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 $\gamma^{\text{new}}(z_{mk})$ 、以前の負担率を $\gamma^{\text{old}}(z_{mk})$ とする
 - d を次のようにおく (前後の負担率の差)

$$d = \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \quad (115)$$

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

$$N_k^{\text{new}} = N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) = N_k^{\text{old}} + d \quad (116)$$

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

$$\boldsymbol{\mu}_k^{\text{old}} = \frac{1}{N_k^{\text{old}}} \sum_i \gamma^{\text{old}}(z_{ik}) \mathbf{x}_i \quad (117)$$

$$\boldsymbol{\Sigma}_k^{\text{old}} = \frac{1}{N_k^{\text{old}}} \sum_i \gamma^{\text{old}}(z_{ik}) (\mathbf{x}_i - \boldsymbol{\mu}_k^{\text{old}})(\mathbf{x}_i - \boldsymbol{\mu}_k^{\text{old}})^T \quad (118)$$

$$\pi_k^{\text{old}} = \frac{N_k^{\text{old}}}{N} \quad (119)$$

但し

$$N_k^{\text{old}} = \sum_i \gamma^{\text{old}}(z_{ik}) \quad (120)$$

逐次型の EM アルゴリズムの例

- 平均の更新式は

$$\boldsymbol{\mu}_k^{\text{new}} = \frac{1}{N_k^{\text{new}}} \left(\sum_{i \neq m} \gamma^{\text{old}}(z_{ik}) \mathbf{x}_i + \gamma^{\text{new}}(z_{mk}) \mathbf{x}_m \right) \quad (121)$$

$$= \frac{1}{N_k^{\text{new}}} \left(N_k^{\text{old}} \boldsymbol{\mu}_k^{\text{old}} - \gamma^{\text{old}}(z_{mk}) \mathbf{x}_m + \gamma^{\text{new}}(z_{mk}) \mathbf{x}_m \right)$$

$$= \frac{1}{N_k^{\text{new}}} \left((N_k^{\text{new}} - \gamma^{\text{new}}(z_{mk}) + \gamma^{\text{old}}(z_{mk})) \boldsymbol{\mu}_k^{\text{old}} - \gamma^{\text{old}}(z_{mk}) \mathbf{x}_m + \gamma^{\text{new}}(z_{mk}) \mathbf{x}_m \right)$$

$$= \boldsymbol{\mu}_k^{\text{old}} + \frac{1}{N_k^{\text{new}}} \left(-(\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})) \boldsymbol{\mu}_k^{\text{old}} + (\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})) \mathbf{x}_m \right)$$

$$= \boldsymbol{\mu}_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} (\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}}) \quad (122)$$

逐次型の EM アルゴリズムの例

$$= \boldsymbol{\mu}_k^{\text{old}} + \frac{d}{N_k^{\text{new}}} (\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}}) \quad (123)$$

- 分散の更新式は

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k^{\text{new}}} \sum_i \gamma^{\text{new}}(z_{ik}) \mathbf{x}_i \mathbf{x}_i^T - \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \quad (124)$$

$$= \frac{1}{N_k^{\text{new}}} \left(\sum_{i \neq m} \gamma^{\text{old}}(z_{ik}) \mathbf{x}_i \mathbf{x}_i^T + \gamma^{\text{new}}(z_{mk}) \mathbf{x}_m \mathbf{x}_m^T \right) - \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \quad (125)$$

$$= \frac{1}{N_k^{\text{new}}} \left(\left(\sum_i \gamma^{\text{old}}(z_{ik}) \mathbf{x}_i \mathbf{x}_i^T - \gamma^{\text{old}}(z_{mk}) \mathbf{x}_m \mathbf{x}_m^T \right) + \gamma^{\text{new}}(z_{mk}) \mathbf{x}_m \mathbf{x}_m^T \right) - \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \quad (126)$$

逐次型の EM アルゴリズムの例

$$\begin{aligned} &= \frac{1}{N_k^{\text{new}}} \left(N_k^{\text{old}} \left(\boldsymbol{\Sigma}_k^{\text{old}} + \boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T \right) - \right. \\ &\quad \left. \gamma^{\text{old}}(z_{mk}) \mathbf{x}_m \mathbf{x}_m^T + \gamma^{\text{new}}(z_{mk}) \mathbf{x}_m \mathbf{x}_m^T \right) - \\ &\quad \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \end{aligned} \quad (127)$$

$$\begin{aligned} &= \frac{1}{N_k^{\text{new}}} \left(N_k^{\text{old}} \boldsymbol{\Sigma}_k^{\text{old}} + N_k^{\text{old}} \boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T + \right. \\ &\quad \left(\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \right) \mathbf{x}_m \mathbf{x}_m^T - \\ &\quad \left. N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \right) \\ &= \frac{1}{N_k^{\text{new}}} \left(N_k^{\text{old}} \boldsymbol{\Sigma}_k^{\text{old}} + N_k^{\text{old}} \boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T + \right. \\ &\quad \left. d\mathbf{x}_m \mathbf{x}_m^T - N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \right) \end{aligned} \quad (128)$$

逐次型の EM アルゴリズムの例

ここで

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

$$\begin{aligned} = & (N_k^{\text{old}} + d) \boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T + \\ & 2d \boldsymbol{\mu}_k^{\text{old}} \mathbf{x}_m^T - 2d \boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T + \\ & \frac{1}{N_k^{\text{new}}} d^2 (\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}}) (\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}})^T \end{aligned} \quad (130)$$

逐次型の EM アルゴリズムの例

であるから

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

逐次型の EM アルゴリズムの例

$$\begin{aligned} & \frac{1}{N_k^{\text{new}}} d^2 (\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}}) (\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}})^T \\ = & \frac{d}{N_k^{\text{new}}} (N_k^{\text{new}} - d) (\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}}) (\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}})^T \\ = & \frac{d}{N_k^{\text{new}}} N_k^{\text{old}} (\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}}) (\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}})^T \end{aligned} \quad (131)$$

以上より

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

逐次型の EM アルゴリズムの例

$$\begin{aligned} &= \frac{N_k^{\text{old}}}{N_k^{\text{new}}} \left(\boldsymbol{\Sigma}_k^{\text{old}} + \frac{d}{N_k^{\text{new}}} \right. \\ &\quad \left. (\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}}) (\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}})^T \right) \\ &= \frac{N_k^{\text{old}}}{N_k^{\text{new}}} \left(\boldsymbol{\Sigma}_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{ik}) - \gamma^{\text{old}}(z_{ik})}{N_k^{\text{new}}} \right. \\ &\quad \left. (\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}}) (\mathbf{x}_m - \boldsymbol{\mu}_k^{\text{old}})^T \right) \end{aligned} \quad (132)$$

- 混合係数の更新式は

$$\pi_k^{\text{new}} = \frac{N_k^{\text{new}}}{N} = \frac{N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N} \quad (133)$$

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 μ_k, Σ_k, π_k の逐次更新式が得られた

$$\mu_k^{\text{new}} = \mu_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} (\mathbf{x}_m - \mu_k^{\text{old}}) \quad (134)$$

$$\Sigma_k^{\text{new}} = \frac{N_k^{\text{old}}}{N_k^{\text{new}}} \left(\Sigma_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{ik}) - \gamma^{\text{old}}(z_{ik})}{N_k^{\text{new}}} (\mathbf{x}_m - \mu_k^{\text{old}}) (\mathbf{x}_m - \mu_k^{\text{old}})^T \right) \quad (135)$$

$$\pi_k^{\text{new}} = \frac{N_k^{\text{new}}}{N} = \frac{N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N} \quad (136)$$

但し

$$N_k^{\text{new}} = N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \quad (137)$$

逐次型の EM アルゴリズムの例

- 到着したデータ 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
 - x_m が新しく到着したデータであれば、 $\gamma^{\text{old}}(z_{mk}) = 0$ 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 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ 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 $\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ の事後確率 $\ln p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$ による期待値 $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ 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 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ の下界であるから、 \mathcal{L} の最大化は、対数尤度 $\ln p(\mathbf{X}|\boldsymbol{\theta})$ 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