

① 混合ガウス分布

混合ガウス分布

- 混合ガウス分布を導入する理由

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

- 混合ガウス分布とは

- 各ガウス分布の線形の重ね合わせ

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

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

混合ガウス分布

- パラメータ π_k を **混合係数** といい、以下の条件を満たす

$$\sum_k \pi_k = 1 \quad (2)$$

これは、 $p(x)$ を x について積分すれば明らかである

$$\int p(x) dx = 1$$

$$\int \sum_k \pi_k \mathcal{N}(x | \mu_k, \Sigma_k) dx = 1$$

$$\sum_k \pi_k \int \mathcal{N}(x | \mu_k, \Sigma_k) dx = 1$$

$$\sum_k \pi_k = 1$$

混合ガウス分布

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

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

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

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

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

$$0 \leq \pi_k \leq 1 \tag{3}$$

混合ガウス分布

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

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

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

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

混合ガウス分布

- 混合ガウス分布を導入する理由 (再確認)
 - データに対して、**各クラスタに属する確率**が分かるようにするため
- 問題設定
 - x の N 個の観測点で構成されるデータ集合 $\mathcal{D} = \{x_1, \dots, x_N\}$ ($x_i \in \mathbb{R}^D$)
 - データ集合 $\mathcal{D} = \{x_1, \dots, x_N\}$ を、 K 個のクラスタに分割
 - K は、**既知の定数**であるとする
 - k 番目のクラスタが、**平均 μ_k 、共分散行列 Σ_k の正規分布 $\mathcal{N}(x|\mu_k, \Sigma_k)$ で表現できる**とする
 - 各クラスタの分布 $\mathcal{N}(x|\mu_k, \Sigma_k)$ を、 π_k で重み付けして足し合わせた混合分布が、データ全体を表す分布である

混合ガウス分布

- 最尤推定を試みる

- 最尤推定によって、混合ガウス分布のパラメータ π, μ, Σ が分かったとする
- このとき次のようにすれば、クラスタリングが可能
- 新たなデータ x が得られたとき、全ての $k(k = 1, \dots, K)$ について $\mathcal{N}(x|\mu_k, \Sigma_k)$ を計算する
- これを最大にするような k が、データ x が属するクラスタである

- 最尤推定を試みる

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

混合ガウス分布

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

$$\begin{aligned} & \ln p(\mathcal{D}|\boldsymbol{\theta}) \\ = & \ln \prod_i p(\mathbf{x}_i|\boldsymbol{\theta}) \\ = & \ln \prod_i \left(\sum_k \pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \\ = & \sum_i \ln \left(\sum_k \pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) \\ = & \sum_i \ln \left(\sum_k \pi_k \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \right. \\ & \left. \exp \left\{ -\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \right) \end{aligned} \quad (4)$$

混合ガウス分布

- 上式の最初の変形では、各データ $\mathcal{D} = \{x_1, \dots, x_N\}$ は、確率分布 $p(x)$ から、**独立に得られている**という仮定を用いた
- このようなデータ \mathcal{D} を、**i.i.d 標本**という (independently and identically distributed)
- 対数 \ln は単調増加関数であるため、対数を適用しても、関数の極値は変化しない
- 尤度関数 $p(\mathcal{D}|\theta)$ の最大化は、対数尤度関数 $\ln p(\mathcal{D}|\theta)$ の最大化と等価
- 重大な問題点
 - **対数関数 \ln の内部に、総和 (\sum) が入っている**
 - **log-sum の形状になっているため、これ以上式を簡単にできない!**
 - クラスタ数 $K = 1$ であれば、対数 \ln と、ガウス分布の指数 \exp が打ち消し合って、式が簡潔になる
- しかしここでは、このまま最尤推定を続けてみる

- パラメータ μ_k の最尤推定

- 対数尤度関数 $\ln p(\mathcal{D}|\theta)$ において、 θ はパラメータ (定数) で、 \mathcal{D} は変数であるが、実際は \mathcal{D} にはデータが入っているので、パラメータ θ を変数とみなす
- $\ln p(\mathcal{D}|\theta)$ をパラメータ μ_k で微分してみる
- これを 0 と等置することで、最適な μ_k が満たすべき式が得られる
- その前に、関数 f の対数の微分について、以下が成立することを確認しておく

$$(\ln f)' = \frac{f'}{f}, \quad f' = f \cdot (\ln f)' \quad (5)$$

- このとき次のようになる

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

$$\begin{aligned}
 & \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \\
 = & \sum_i \frac{1}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \Sigma_k)} \pi_k \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \\
 & \frac{\partial}{\partial \boldsymbol{\mu}_k} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \quad (6)
 \end{aligned}$$

ここで、以下のようにできる

$$\begin{aligned}
 & \frac{\partial}{\partial \boldsymbol{\mu}_k} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \\
 = & \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \\
 & \frac{\partial}{\partial \boldsymbol{\mu}_k} \left(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right) \quad (7)
 \end{aligned}$$

更に、次が成立する

$$\begin{aligned}& \frac{\partial}{\partial \boldsymbol{\mu}_k} \left(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right) \\&= -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\mu}_k} (\mathbf{x}_i^T \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_i - \mathbf{x}_i^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k - \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_i + \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k) \\&= -\frac{1}{2} \frac{\partial}{\partial \boldsymbol{\mu}_k} (-\mathbf{x}_i^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k - \mathbf{x}_i^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k + \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k) \\&= -\frac{1}{2} (-2\mathbf{x}_i^T \boldsymbol{\Sigma}_k^{-1} + 2\boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k) \\&= \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_i - \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \\&= \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)\end{aligned}\tag{8}$$

ここで、以下を用いた

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

混合ガウス分布

$$\begin{aligned} &= \mathbf{x}_i^T (\boldsymbol{\Sigma}_k^{-1})^T (\boldsymbol{\mu}_k^T)^T \\ &= \mathbf{x}_i^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \end{aligned} \quad (9)$$

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

$$(\boldsymbol{\Sigma}_k^{-1})^T = (\boldsymbol{\Sigma}_k^T)^{-1} = \boldsymbol{\Sigma}_k^{-1} \quad (10)$$

各項の微分は次のようになる

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{\mu}_k} \mathbf{x}_i^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \\ &= (\mathbf{x}_i^T \boldsymbol{\Sigma}_k^{-1})^T = (\boldsymbol{\Sigma}_k^{-1})^T (\mathbf{x}_i^T)^T = \boldsymbol{\Sigma}_k^{-1} \mathbf{x}_i \end{aligned} \quad (11)$$

$$\begin{aligned} &\frac{\partial}{\partial \boldsymbol{\mu}_k} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \\ &= \left(\boldsymbol{\Sigma}_k^{-1} + (\boldsymbol{\Sigma}_k^{-1})^T \right) \boldsymbol{\mu}_k = 2\boldsymbol{\Sigma}_k^{-1} \boldsymbol{\mu}_k \end{aligned} \quad (12)$$

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

$$\begin{aligned} & \frac{\partial}{\partial \boldsymbol{\mu}_k} \ln p(\mathcal{D}|\boldsymbol{\theta}) \\ = & \sum_i \frac{1}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \pi_k \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \\ & \frac{\partial}{\partial \boldsymbol{\mu}_k} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \\ = & \sum_i \frac{1}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \pi_k \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\Sigma}_k|^{\frac{1}{2}}} \\ & \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \\ = & \sum_i \frac{\pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \quad (13) \\ = & 0 \end{aligned}$$

混合ガウス分布

- 未知のパラメータ π, μ, Σ が、分母と分子の双方に出現する複雑な式
- 直接この連立方程式を解いて、パラメータの最尤推定量を求めるのは難しそうである
- 勾配 $\nabla_{\mu_k} \ln p(\mathcal{D}|\theta)$ を利用した最適化も可能である
- この勾配の方向に、パラメータ μ_k を少しだけ更新する
- ここでは、**EM アルゴリズム**という別の手法を導出しようとしている
- x のほかに、**潜在変数**という仮想的な変数 z を導入することで、**簡単に解けるようになる**
- 上と似たような式が、後ほど登場する

混合ガウス分布

- ここまでの話の流れ

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

- 潜在変数 z の導入

- 各データ x_i につき、1つのベクトル $z_i \in \mathbb{R}^K$ が対応しているとする
- z_i は、データ x_i が属するクラスタ を表現する

- 潜在変数 z の表現

- z_i は、 K 次元の二値確率変数 z の観測値である
- z の k 番目の要素を、 z_k と表すことにする
- 確率変数 z は、1-of- K 符号化法により表現されるとする
- 即ち、ある1つの $k \in \{1, \dots, K\}$ について $z_k = 1$ で、 $j \neq k$ に対し $z_j = 0$ となる
- $z_k (k = 1, \dots, K)$ は、 $z_k \in \{0, 1\}$ かつ $\sum_k z_k = 1$ をみたす
- ベクトル z は K 種類の状態を取る

- 潜在変数 z の例

- 例えば、データ点 x_i に対して z_i があるとする
- $z_{i1} = 1$ ($z_i = [1, 0, 0, \dots, 0]$) ならば、 x_i は 1 番目のクラスタ出身
- $z_{i2} = 1$ ($z_i = [0, 1, 0, \dots, 0]$) ならば、 x_i は 2 番目のクラスタ出身

- データ x_i が作られるまでの流れ

- z に関する確率分布 $p(z)$ から、 z_i がサンプルされる
- z が与えられた下での条件付き分布 $p(x|z)$ から、 x_i がサンプルされる
- 即ち、 x, z の同時分布は、周辺分布 $p(z)$ と、条件付き分布 $p(x|z)$ を用いて次のように書ける

$$p(x, z) = p(z)p(x|z) \quad (14)$$

- 潜在変数 z_i が最初に決められ、その z_i に応じて x_i が決まると考える
- z_i は実際には存在しない、仮想的なものである
- z_i は、実際に観測される x_i の裏側に潜んでいる

混合ガウス分布

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

$$p(\mathbf{x}) = \sum_k \pi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (15)$$

- 周辺分布 $p(z)$ の定義
 - $p(z)$ は次のように定める

$$p(z) = \prod_k \pi_k^{z_k}, \quad p(z_k = 1) = \pi_k \quad (16)$$

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

混合ガウス分布

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

$$p(\mathbf{x}|\mathbf{z}) = \prod_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k} \quad (17)$$

$$p(\mathbf{x}|z_k = 1) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (18)$$

- $p(\mathbf{x})$ の導出

- \sum_z は、可能な全ての \mathbf{z} についての総和を取ること
- $\mathbf{z} = [1, 0, \dots, 0]^T, [0, 1, 0, \dots, 0]^T, \dots, [0, \dots, 0, 1]^T$ についての和
- これは、ベクトル \mathbf{z} の中で、1 である要素のインデックス k についての総和 \sum_k を取ることに相当

混合ガウス分布

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

$$p(\mathbf{x}) = \sum_z p(\mathbf{x}, z) \quad (19)$$

$$= \sum_z p(z)p(\mathbf{x}|z) \quad (20)$$

$$= \sum_k p(z_k = 1)p(\mathbf{x}|z_k = 1) \quad (21)$$

$$= \sum_k \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \quad (22)$$

- これは、混合ガウス分布と同じ形になっている
- 何が嬉しいのか?
 - 潜在変数を陽に含む表現 $p(\mathbf{x}, z) = p(z)p(\mathbf{x}|z)$ を得たことで、この同時分布を使った議論が可能になった

- $p(z_k = 1|\mathbf{x})$ の表現

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

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

$$\gamma(z_k) \equiv p(z_k = 1|\mathbf{x}) \quad (23)$$

$$\begin{aligned} &= \frac{p(z_k = 1)p(\mathbf{x}|z_k = 1)}{\sum_j p(z_j = 1)p(\mathbf{x}|z_j = 1)} \\ &= \frac{\pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \end{aligned} \quad (24)$$

- $\sum_k \gamma(z_k) = 1$ であることに注意
- 潜在変数を導入したので、最尤推定について再度考えてみる

- 最尤推定を再挑戦

- パラメータ θ は、 $\pi \equiv \{\pi_1, \pi_2, \dots, \pi_K\}$ 、 $\mu \equiv \{\mu_1, \mu_2, \dots, \mu_K\}$ 、 $\Sigma \equiv \{\Sigma_1, \Sigma_2, \dots, \Sigma_K\}$ をまとめたもの
- 対数尤度関数 $\ln p(\mathcal{D}|\theta)$ は以下に示す通りであった

$$\begin{aligned} & \ln p(\mathcal{D}|\theta) \\ = & \sum_i \ln \left(\sum_k \pi_k \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \right. \\ & \left. \exp \left\{ -\frac{1}{2} (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right\} \right) \end{aligned} \quad (25)$$

- 尤度関数を、 π, μ, Σ のそれぞれについて最大化する
- ここでは、尤度関数を最大化する μ_k が、満たすべき条件を考える

混合ガウス分布

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

$$\begin{aligned} & \frac{\partial}{\partial \boldsymbol{\mu}_k} \ln p(\mathcal{D}|\boldsymbol{\theta}) \\ &= \sum_i \frac{\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \end{aligned} \quad (26)$$

$$= \sum_i \gamma(z_{ik}) \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) = 0 \quad (27)$$

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

混合ガウス分布

- 共分散行列 Σ_k が正則であると仮定して、両辺に左から掛けて整理する

$$\begin{aligned}\sum_i \gamma(z_{ik})(\mathbf{x}_i - \boldsymbol{\mu}_k) &= 0 \\ \Rightarrow \sum_i \gamma(z_{ik})\boldsymbol{\mu}_k &= \sum_i \gamma(z_{ik})\mathbf{x}_i \\ \Rightarrow \boldsymbol{\mu}_k \sum_i \gamma(z_{ik}) &= \sum_i \gamma(z_{ik})\mathbf{x}_i \\ \Rightarrow \boldsymbol{\mu}_k &= \frac{1}{\sum_i \gamma(z_{ik})} \sum_i \gamma(z_{ik})\mathbf{x}_i\end{aligned}\tag{28}$$

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

- K-Means 法との比較

- K-Means 法における、平均ベクトル μ_k の更新式と見比べてみる

$$\begin{aligned}\mu_k &= \frac{1}{\sum_i \gamma(z_{ik})} \sum_i \gamma(z_{ik}) \mathbf{x}_i \\ \mu_k &= \frac{1}{\sum_i r_{ik}} \sum_i r_{ik} \mathbf{x}_i\end{aligned}\tag{29}$$

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

混合ガウス分布

- そこで、K-Means 法のとおり、 N_k を次のように定める

$$N_k = \sum_i \gamma(z_{ik}) \quad (30)$$

このとき、 μ_k の式は

$$\mu_k = \frac{1}{N_k} \sum_i \gamma(z_{ik}) \mathbf{x}_i \quad (31)$$

- 例えば、 $\gamma(z_{ik})$ が 0, 1 のいずれかであれば、 $\sum_i \gamma(z_{ik})$ は、 k 番目のクラスタに属するデータの数と完全に一致
- クラスタ k に対応するガウス分布の平均 μ_k は、各データ \mathbf{x}_i の重み付き平均
- 重み因子は、事後確率 $p(z_k = 1 | \mathbf{x}_i) \equiv \gamma(z_{ik})$ である

混合ガウス分布

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

- 尤度関数を最大化する Σ_k の導出

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

$$\begin{aligned} & \frac{\partial}{\partial \Sigma_k} \ln p(\mathcal{D}|\theta) \\ = & \frac{\partial}{\partial \Sigma_k} \sum_i \ln \left(\sum_k \pi_k \mathcal{N}(\mathbf{x}_i | \mu_k, \Sigma_k) \right) \\ = & \sum_i \frac{\partial}{\partial \Sigma_k} \ln \left(\sum_k \pi_k \mathcal{N}(\mathbf{x}_i | \mu_k, \Sigma_k) \right) \\ = & \sum_i \frac{1}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_i | \mu_k, \Sigma_k)} \frac{\partial}{\partial \Sigma_k} \left(\sum_k \pi_k \mathcal{N}(\mathbf{x}_i | \mu_k, \Sigma_k) \right) \\ = & \sum_i \frac{1}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_i | \mu_k, \Sigma_k)} \frac{\partial}{\partial \Sigma_k} \pi_k \mathcal{N}(\mathbf{x}_i | \mu_k, \Sigma_k) \end{aligned} \quad (32)$$

ここで、以下の部分を求める

$$\begin{aligned} & \frac{\partial}{\partial \Sigma_k} \pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \Sigma_k) \\ = & \frac{\partial}{\partial \Sigma_k} \pi_k \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \\ = & \pi_k \frac{1}{(2\pi)^{\frac{D}{2}}} \left(\frac{\partial}{\partial \Sigma_k} \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \right. \\ & \left. \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \right) \end{aligned} \quad (33)$$

微分の中身は、 Σ_k についての合成関数となっている

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

混合ガウス分布

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

$$\frac{\partial}{\partial \mathbf{X}} f(\mathbf{X}) g(\mathbf{X}) = f(\mathbf{X}) \frac{\partial g(\mathbf{X})}{\partial \mathbf{X}} + g(\mathbf{X}) \frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \quad (34)$$

これを利用して、先程の微分を求める

$$\begin{aligned} & \frac{\partial}{\partial \Sigma_k} \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \\ = & \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \frac{\partial}{\partial \Sigma_k} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} + \\ & \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \frac{\partial}{\partial \Sigma_k} \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \end{aligned} \quad (35)$$

- 各項の微分を順番に求める

$$\begin{aligned} & \frac{\partial}{\partial \Sigma_k} \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \\ = & \frac{\partial}{\partial \Sigma_k} \exp \left(\ln \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \right) \\ = & \frac{\partial}{\partial \Sigma_k} \exp \left(-\frac{1}{2} \ln |\Sigma_k| \right) \\ = & \exp \left(-\frac{1}{2} \ln |\Sigma_k| \right) \frac{\partial}{\partial \Sigma_k} \left(-\frac{1}{2} \ln |\Sigma_k| \right) \\ = & \exp \left(-\frac{1}{2} \ln |\Sigma_k| \right) \left(-\frac{1}{2} \right) \frac{\partial}{\partial \Sigma_k} \ln |\Sigma_k| \\ = & -\frac{1}{2} \exp \left(-\frac{1}{2} \ln |\Sigma_k| \right) (\Sigma_k^{-1})^T \\ = & -\frac{1}{2} \frac{1}{|\Sigma_k|^{\frac{1}{2}}} \Sigma_k^{-1} \end{aligned} \tag{36}$$

また

$$\begin{aligned} & \frac{\partial}{\partial \Sigma_k} \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \\ = & \exp \left\{ -\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right\} \\ & \frac{\partial}{\partial \Sigma_k} \left(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right) \end{aligned} \quad (37)$$

であって

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

$$\begin{aligned} &= -\frac{1}{2} \frac{\partial}{\partial \Sigma_k} \text{Tr} \left(\Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \right) \\ &= -\frac{1}{2} \left(- \left(\Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} \right)^T \right) \\ &= \frac{1}{2} \left(\Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} \right)^T \\ &= \frac{1}{2} \left(\Sigma_k^{-1} \right)^T \left((\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \right)^T \left(\Sigma_k^{-1} \right)^T \\ &= \frac{1}{2} \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} \end{aligned} \quad (38)$$

のように求まる

- 行列のトレースについて、一般に以下が成り立つことを利用している

$$\text{Tr}(\mathbf{X}\mathbf{Y}) = \text{Tr}(\mathbf{Y}\mathbf{X}) \quad (39)$$

$$\text{Tr}(\mathbf{X}\mathbf{Y}\mathbf{Z}) = \text{Tr}(\mathbf{Y}\mathbf{Z}\mathbf{X}) = \text{Tr}(\mathbf{Z}\mathbf{X}\mathbf{Y}) \quad (40)$$

混合ガウス分布

- また先程の微分では以下の公式を用いている

$$\frac{\partial}{\partial \mathbf{X}} \text{Tr}(\mathbf{X}^{-1} \mathbf{Y}) = -(\mathbf{X}^{-1} \mathbf{Y} \mathbf{X}^{-1})^T \quad (41)$$

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

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

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

$$\begin{aligned} & \frac{\partial}{\partial x} \sum_j A_{ij} B_{jk} \\ &= \sum_j \frac{\partial}{\partial x} A_{ij} B_{jk} \end{aligned}$$

$$\begin{aligned} &= \sum_j \left(\frac{\partial A_{ij}}{\partial x} B_{jk} + A_{ij} \frac{\partial B_{jk}}{\partial x} \right) \\ &= \sum_j \frac{\partial A_{ij}}{\partial x} B_{jk} + \sum_j A_{ij} \frac{\partial B_{jk}}{\partial x} \end{aligned} \quad (42)$$

であるから、結局

$$\frac{\partial}{\partial x} \mathbf{A} \mathbf{B} = \frac{\partial \mathbf{A}}{\partial x} \mathbf{B} + \mathbf{A} \frac{\partial \mathbf{B}}{\partial x} \quad (43)$$

となる

- 上記の公式から、以下の公式を簡単に導ける

$$\begin{aligned}\frac{\partial}{\partial x} \mathbf{A}^{-1} \mathbf{A} &= \frac{\partial \mathbf{A}^{-1}}{\partial x} \mathbf{A} + \mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x} \\ \frac{\partial}{\partial x} \mathbf{I} &= \frac{\partial \mathbf{A}^{-1}}{\partial x} \mathbf{A} + \mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x} \\ 0 &= \frac{\partial \mathbf{A}^{-1}}{\partial x} \mathbf{A} + \mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x}\end{aligned}\tag{44}$$

これに右から \mathbf{A}^{-1} を掛ければ

$$\begin{aligned}0 &= \frac{\partial \mathbf{A}^{-1}}{\partial x} \mathbf{A} \mathbf{A}^{-1} + \mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x} \mathbf{A}^{-1} \\ 0 &= \frac{\partial \mathbf{A}^{-1}}{\partial x} \mathbf{I} + \mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x} \mathbf{A}^{-1} \\ 0 &= \frac{\partial \mathbf{A}^{-1}}{\partial x} + \mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x} \mathbf{A}^{-1}\end{aligned}$$

$$\frac{\partial}{\partial x} \mathbf{A}^{-1} = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x} \mathbf{A}^{-1} \quad (45)$$

より

$$\frac{\partial}{\partial x} \mathbf{A}^{-1} = -\mathbf{A}^{-1} \frac{\partial \mathbf{A}}{\partial x} \mathbf{A}^{-1} \quad (46)$$

を得る (逆行列の微分公式)

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

$$\begin{aligned} & \frac{\partial}{\partial x} (\mathbf{A}^{-1})_{kl} \\ &= -\sum_{m,n} (\mathbf{A}^{-1})_{km} \left(\frac{\partial \mathbf{A}}{\partial x} \right)_{mn} (\mathbf{A}^{-1})_{ml} \\ &= -\sum_{m,n} (\mathbf{A}^{-1})_{km} \frac{\partial A_{mn}}{\partial x} (\mathbf{A}^{-1})_{ml} \end{aligned} \quad (47)$$

混合ガウス分布

- この逆行列の微分公式を使えば、以下の微分公式を導出できる

$$\frac{\partial}{\partial \mathbf{X}} \text{Tr}(\mathbf{X}^{-1} \mathbf{Y})$$

行列 \mathbf{X} の i, j 要素による微分を考えれば

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

$$\begin{aligned} &= \sum_{k,l} \left(-(\mathbf{X}^{-1})_{jl} Y_{lk} (\mathbf{X}^{-1})_{ki} \right) \\ &= -(\mathbf{X}^{-1} \mathbf{Y} \mathbf{X}^{-1})_{ji} \\ &= -\left((\mathbf{X}^{-1} \mathbf{Y} \mathbf{X}^{-1})^T \right)_{ij} \end{aligned} \tag{48}$$

であるから

$$\frac{\partial}{\partial \mathbf{X}} \text{Tr}(\mathbf{X}^{-1} \mathbf{Y}) = -(\mathbf{X}^{-1} \mathbf{Y} \mathbf{X}^{-1})^T \tag{49}$$

を得られる

- 尤度関数を最大化する Σ_k の導出
 - これより、結局次のようになる

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

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

$$\begin{aligned}
 & \left(\frac{1}{2} \Sigma_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} - \frac{1}{2} \Sigma_k^{-1} \right) \\
 = & \sum_i \frac{\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \Sigma_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \Sigma_k)} \\
 & \frac{1}{2} \Sigma_k^{-1} ((\mathbf{x}_i - \boldsymbol{\mu}_k) (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} - \mathbf{I}) \\
 = & \frac{1}{2} \sum_i \gamma(z_{ik}) \Sigma_k^{-1} ((\mathbf{x}_i - \boldsymbol{\mu}_k) (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \Sigma_k^{-1} - \mathbf{I}) = 0 \quad (50)
 \end{aligned}$$

両辺に左右から Σ_k を掛けて、整理すれば

$$\begin{aligned}
 & \frac{1}{2} \sum_i \gamma(z_{ik}) ((\mathbf{x}_i - \boldsymbol{\mu}_k) (\mathbf{x}_i - \boldsymbol{\mu}_k)^T - \Sigma_k) = 0 \\
 \Rightarrow & \sum_i \gamma(z_{ik}) \Sigma_k = \sum_i \gamma(z_{ik}) (\mathbf{x}_i - \boldsymbol{\mu}_k) (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \\
 \Rightarrow & \Sigma_k = \frac{1}{\sum_i \gamma(z_{ik})} \sum_i \gamma(z_{ik}) (\mathbf{x}_i - \boldsymbol{\mu}_k) (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \quad (51)
 \end{aligned}$$

$$\Rightarrow \Sigma_k = \frac{1}{N_k} \sum_i \gamma(z_{ik})(\mathbf{x}_i - \boldsymbol{\mu}_k)(\mathbf{x}_i - \boldsymbol{\mu}_k)^T \quad (52)$$

- これより、 Σ_k を導出する式が得られた

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

$$\ln p(\mathcal{D}|\theta) + \lambda \left(\sum_k \pi_k - 1 \right) \quad (53)$$

混合ガウス分布

- π_k で微分して 0 と等置すると、次のようになる

$$\begin{aligned} & \frac{\partial}{\partial \pi_k} \left(\ln p(\mathcal{D}|\boldsymbol{\theta}) + \lambda \left(\sum_k \pi_k - 1 \right) \right) \\ &= \frac{\partial}{\partial \pi_k} \left(\sum_i \ln \left(\sum_k \pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) + \lambda \left(\sum_k \pi_k - 1 \right) \right) \\ &= \sum_i \frac{\mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} + \lambda \end{aligned} \quad (54)$$

$$= \sum_i \frac{\gamma(z_{ik})}{\pi_k} + \lambda = 0 \quad (55)$$

両辺に π_k を掛けて

$$\sum_i \gamma(z_{ik}) + \lambda \pi_k = 0 \quad (56)$$

混合ガウス分布

k についての和を取ると

$$\sum_k \left(\sum_i \gamma(z_{ik}) + \lambda \pi_k \right) = 0 \quad (57)$$

$$\sum_i \sum_k \gamma(z_{ik}) + \lambda \sum_k \pi_k = 0 \quad (58)$$

$$\sum_i 1 + \lambda = 0 \quad (59)$$

$$N + \lambda = 0 \quad (60)$$

$$\therefore \lambda = -N \quad (61)$$

これより

$$\sum_i \gamma(z_{ik}) + (-N) \pi_k = 0 \quad (62)$$

$$\pi_k = \frac{1}{N} \sum_i \gamma(z_{ik}) = \frac{N_k}{N} \quad (63)$$

混合ガウス分布

ここで、以下が成立することに注意

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

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

μ_k, Σ_k, π_k の更新式

$$\mu_k = \frac{1}{N_k} \sum_i \gamma(z_{ik}) \mathbf{x}_i \quad (64)$$

$$\Sigma_k = \frac{1}{N_k} \sum_i \gamma(z_{ik}) (\mathbf{x}_i - \mu_k)(\mathbf{x}_i - \mu_k)^T \quad (65)$$

$$\pi_k = \frac{N_k}{N} \quad (66)$$

$$N_k = \sum_i \gamma(z_{ik}) \quad (67)$$

$\gamma(z_{ik})$ の更新式

$$\gamma(z_{ik}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_k \pi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \quad (68)$$

● 注意点

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

- これらの更新式の意味

- 最尤推定の解を求めるための、**繰り返し手続きの存在**を示唆
- 即ち、 μ_k, Σ_k, π_k の初期化後に、(1) $\gamma(z_{ik})$ の更新と、(2) それを用いた μ_k, Σ_k, π_k の更新という、**2段階の処理を繰り返す**手続き
- これは、混合ガウス分布を確率モデルとして使ったときの、**EM アルゴリズム**となっている
- 混合ガウス分布に対する EM アルゴリズムは重要なので、次にまとめる

混合ガウス分布に対する EM アルゴリズム

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

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

$$\gamma(z_{ik}) \leftarrow \frac{\pi_k^{\text{old}} \mathcal{N}(\mathbf{x}_i | \mu_k^{\text{old}}, \Sigma_k^{\text{old}})}{\sum_k \pi_k^{\text{old}} \mathcal{N}(\mathbf{x}_i | \mu_k^{\text{old}}, \Sigma_k^{\text{old}})} \quad (69)$$

混合ガウス分布

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

$$\boldsymbol{\mu}_k^{\text{new}} \leftarrow \frac{1}{N_k} \sum_i \gamma(z_{ik}) \boldsymbol{x}_i \quad (70)$$

$$\boldsymbol{\Sigma}_k^{\text{new}} \leftarrow \frac{1}{N_k} \sum_i \gamma(z_{ik}) (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{\text{new}})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{\text{new}})^T \quad (71)$$

$$\pi_k^{\text{new}} \leftarrow \frac{N_k}{N} \quad (72)$$

但し

$$N_k = \sum_i \gamma(z_{ik}) \quad (73)$$

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

$$\ln p(\mathcal{D}|\boldsymbol{\theta}) = \sum_i \ln \left(\sum_k \pi_k^{\text{new}} \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k^{\text{new}}, \boldsymbol{\Sigma}_k^{\text{new}}) \right) \quad (74)$$

パラメータの変化量、あるいは対数尤度の変化量を見て、収束性を判定

5 収束基準を満たしていなければ、(2) に戻る

$$\boldsymbol{\mu}_k^{\text{old}} \leftarrow \boldsymbol{\mu}_k^{\text{new}}, \quad \boldsymbol{\Sigma}_k^{\text{old}} \leftarrow \boldsymbol{\Sigma}_k^{\text{new}}, \quad \pi_k^{\text{old}} \leftarrow \pi_k^{\text{new}} \quad (75)$$

- EM アルゴリズムの概要

- **E ステップ** (Expectation step) では、事後確率 $p(z_k = 1|x_i)$ 、即ち負担率 $\gamma(z_{ik})$ を計算
- **M ステップ** (Maximization step) では、事後確率を使って、各パラメータ μ_k, Σ_k, π_k を再計算

- EM アルゴリズムでの注意点

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

混合ガウス分布

- 混合ガウス分布の良い初期値を見つけるために、最初に K-Means 法を実行し、その後に EM アルゴリズムを利用する、という方法がある
- K-Means 法により得られた平均ベクトル μ_k を、各ガウス分布の平均 μ_k の初期値とする
- 各クラスタに属するデータ点の**標本分散**を、共分散行列 Σ_k の初期値とする
- 各クラスタに属するデータ点の**割合**を、混合係数 π_k の初期値とする
- 一般に対数尤度には、多数の局所解が存在するため、**その中で最大のものの (大域的最適解) に収束するとは限らない**

混合ガウス分布のまとめ

- 混合ガウス分布を導入する理由
 - 曖昧さを含んだ、各データ点のクラスタへの割り当て (ソフト割り当て) を実現するため
 - 各データ点について、各クラスタに属する確率が分かるようにするため
- 混合ガウス分布による表現
 - 各クラスタがガウス分布 $\mathcal{N}(x|\mu_k, \Sigma_k)$ に従うと仮定
 - 各ガウス分布を、混合係数 π_k で重み付けして足し合わせることで、混合ガウス分布を作り、データ全体を表現する
 - パラメータ θ は、各ガウス要素の平均 μ_k と共分散行列 Σ_k 、そして混合係数 π_k である

混合ガウス分布のまとめ

- ここまでの話の流れ

- 最尤推定 (対数尤度 $\ln p(\mathcal{D}|\theta)$ の最大化) によって、パラメータ θ を求める試みは失敗した
- 対数の中に総和が入っているせいで、対数尤度の式が複雑になっていた
- 潜在変数 z を導入し、 z に関する分布を考えたことで、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
- 事後確率 $p(z_k = 1|x_i)$ 即ち負担率 $\gamma(z_{ik})$ の計算と、パラメータ θ の更新という 2 段階の処理を、交互に繰り返していくアルゴリズムであった