ML 輪講: 19 章 近似推論

杉浦 圭祐

慶應義塾大学理工学部情報工学科 松谷研究室

April 12, 2019

目次

目次

目次

K-Means 法によるクラスタリング

- 扱う問題
  - 多次元空間上のデータ点集合を考える
  - 各データが属するクラスタを決定する問題を考える
- 問題設定
  - D 次元ユークリッド空間における、確率変数 $\boldsymbol{x}$ を観測
  - $\boldsymbol{x}$ の N 個の観測点で構成されるデータ集合 $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ $(\boldsymbol{x}_i \in \mathbb{R}^D)$
  - データ集合 $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ を、K 個のクラスタに分割
  - K は、既知の定数であるとする
- クラスタとは
  - クラスタとは、簡単に言えば近接するデータの集合である
  - クラスタの内部のデータ点間の距離が、クラスタの外側のデータとの距離と比べて、小さいようなデータの集合

K-Means 法によるクラスタリング

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

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

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

K-Means 法によるクラスタリング

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

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

K-Means 法によるクラスタリング

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

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

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

K-Means 法によるクラスタリング

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

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

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

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

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

K-Means 法によるクラスタリング

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

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

K-Means 法によるクラスタリング

- ここで、次の関係を用いている

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

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

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

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

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

K-Means 法によるクラスタリング

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

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

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

K-Means 法によるクラスタリング

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

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

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

K-Means 法によるクラスタリング

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

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

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

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

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

K-Means 法によるクラスタリング

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

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

K-Means 法によるクラスタリング

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

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

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

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

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

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

K-Means 法によるクラスタリング

- データ点へのクラスタの再割り当てと、クラスタの平均ベクトルの再計算

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

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

K-Means 法によるクラスタリング

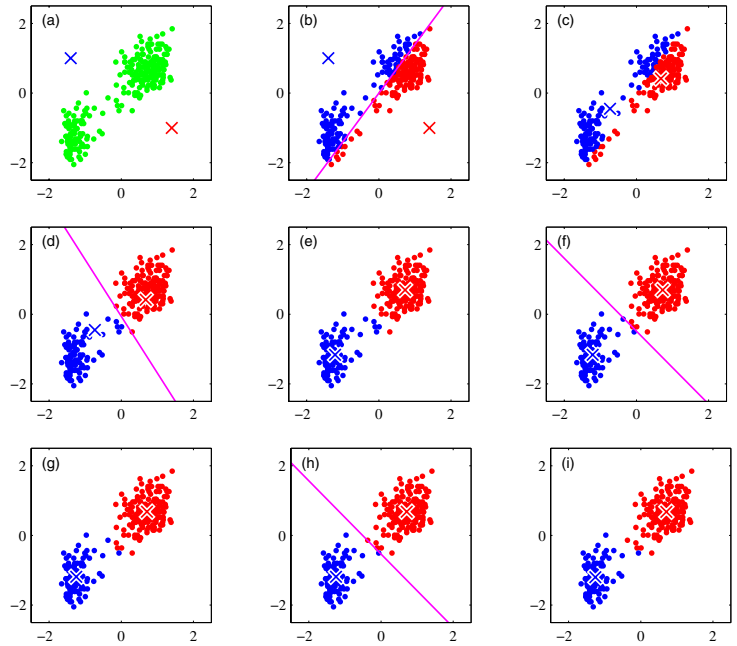

Figure 9.1 Illustration of the K-means algorithm using the re-scaled Old Faithful data set. (a) Green points denote the data set in a two-dimensional Euclidean space. The initial choices for centres μ_1 and μ_2 are shown by the red and blue crosses, respectively. (b) In the initial E step, each data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

図 1: K-Means アルゴリズムの動作

K-Means 法によるクラスタリング

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

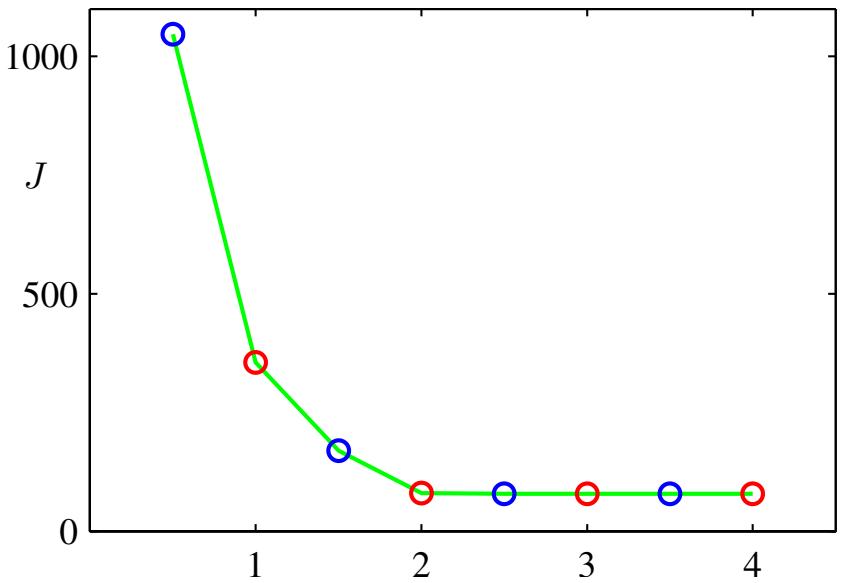

図 2: 目的関数 J の値の遷移

K-Means 法によるクラスタリング

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

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

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

逐次版の K-Means 法

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

Robbins-Monro 法

- Robbins-Monro 法
  - 逐次学習アルゴリズムを導出するための手法
  - 関数 $f(\theta^*) = 0$ を満たす根 θ^* を、逐次的に計算するための式を与える
  - 同時分布 $p(z, \theta)$ に従う確率変数のペア θ, z について、関数 $f(\theta)$ は、θ が与えられたときの z の条件付き期待値 $\mathbb{E}[z|\theta]$ として、定義される

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

  - このとき、根 θ^* の逐次計算式は、次のように記述される

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

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

Robbins-Monro 法

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

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

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

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

Robbins-Monro 法

3 学習率の系列 $\{\eta_N\}$ は次の条件を満たす

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

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

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

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

Robbins-Monro 法

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

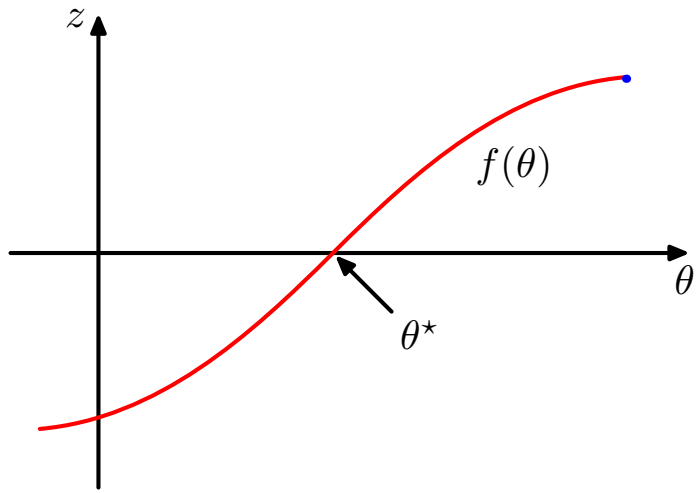

図 3: $f(\theta)$ のグラフ

逐次版の K-Means 法

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

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

これは以下の式と等価である $(N_k = \sum_i r_{ik})$

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

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

逐次版の K-Means 法

- このとき、上式から r_{ik} を省略でき、次のようになる

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

- $N_k \to \infty$ の極限を取って次のように変形する

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

逐次版の K-Means 法

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

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

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

K-Means アルゴリズムの特徴

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

K-Means アルゴリズムのまとめ

- K-Means アルゴリズムの目的
  - 各データが属するクラスタを決定する (ハード割り当て)

- K-Means アルゴリズムで行っていること
  - 各クラスタに対して、中心となるベクトル $\boldsymbol{\mu}_k$ を考えた
  - $\boldsymbol{\mu}_k$ は、対応するクラスタに属する、全てのデータベクトルの平均として得られた
  - 各データ点は、それと最も距離が近い $\boldsymbol{\mu}_k$ に対応するクラスタ k に割り当てた
  - 各データ点が属するクラスタを、r_{ik} という変数 (1-of-K 符号化法) で表した

K-Means アルゴリズムのまとめ

- ここまでの話の流れ
  - クラスタリングの問題を、目的関数 J の最小化として表現できた
  - 目的関数 J を、$\boldsymbol{\mu}_k$ と r_{ik} の両方について一度に最適化することはできなかった
  - そこで J を、$\boldsymbol{\mu}_k$ と r_{ik} について交互に最適化することを考えた

  - 交互に行う最適化は、後ほど説明する EM アルゴリズムの、E ステップと M ステップに対応していた

  - 逐次版の K-Means 法を、Robbins-Monro 法を使って導出した

目次

混合ガウス分布

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

- 混合ガウス分布とは
  - 各ガウス分布の線形の重ね合わせ

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

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

混合ガウス分布

- パラメータ π_k を混合係数といい、以下の条件を満たす

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

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

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

混合ガウス分布

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

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

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

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

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

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

混合ガウス分布

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

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

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

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

混合ガウス分布

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

- 問題設定
  - $\boldsymbol{x}$ の N 個の観測点で構成されるデータ集合 $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ $(\boldsymbol{x}_i \in \mathbb{R}^D)$
  - データ集合 $\mathcal{D} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ を、K 個のクラスタに分割
  - K は、既知の定数であるとする
  - k 番目のクラスタが、平均 $\boldsymbol{\mu}_k$、共分散行列 $\boldsymbol{\Sigma}_k$ の正規分布 $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ で表現できるとする
  - 各クラスタの分布 $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ を、π_k で重み付けして足し合わせた混合分布が、データ全体を表す分布である

混合ガウス分布

- 最尤推定を試みる
  - 最尤推定によって、混合ガウス分布のパラメータ $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ 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 $\boldsymbol{x}$ が得られたとき、全ての $k(k = 1, \ldots, K)$ について $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ を計算する
  - これを最大にするような k が、データ $\boldsymbol{x}$ が属するクラスタである

混合ガウス分布

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

混合ガウス分布

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

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

混合ガウス分布

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

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

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

  - しかしここでは、このまま最尤推定を続けてみる

混合ガウス分布

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

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

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

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

混合ガウス分布

- このとき次のようになる

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

混合ガウス分布

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

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

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

ここで、以下のようにできる

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

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

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

混合ガウス分布

更に、次が成立する

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

ここで、以下を用いた

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

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

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

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

各項の微分は次のようになる

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

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

混合ガウス分布

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

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

混合ガウス分布

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

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

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

- 上と似たような式が、後ほど登場する

混合ガウス分布

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

混合ガウス分布

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

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

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

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

混合ガウス分布

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

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

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

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

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

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

混合ガウス分布

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

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

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

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

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

混合ガウス分布

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

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

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

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

混合ガウス分布

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

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

  - これは、混合ガウス分布と同じ形になっている

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

混合ガウス分布

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

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

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

  - $\sum_k \gamma(z_k) = 1$ であることに注意

  - 潜在変数を導入したので、最尤推定について再度考えてみる

混合ガウス分布

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

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

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

混合ガウス分布

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

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

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

混合ガウス分布

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

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

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

混合ガウス分布

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

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

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

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

混合ガウス分布

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

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

このとき、$\boldsymbol{\mu}_k$ の式は

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

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

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

混合ガウス分布

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

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

混合ガウス分布

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

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

混合ガウス分布

ここで、以下の部分を求める

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

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

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

混合ガウス分布

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

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

これを利用して、先程の微分を求める

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

混合ガウス分布

- 各項の微分を順番に求める

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

混合ガウス分布

また

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

であって

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

混合ガウス分布

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

のように求まる

- 行列のトレースについて、一般に以下が成り立つことを利用している

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

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

混合ガウス分布

- また先程の微分では以下の公式を用いている

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

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

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

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

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

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

混合ガウス分布

$$
\begin{aligned}
&= \sum_j \left(\frac{\partial A_{ij}}{\partial x} B_{jk} + A_{ij} \frac{\partial B_{jk}}{\partial x} \right) \\
&= \sum_j \frac{\partial A_{ij}}{\partial x} B_{jk} + \sum_j A_{ij} \frac{\partial B_{jk}}{\partial x} \qquad (62)
\end{aligned}
$$

であるから、結局

$$
\frac{\partial}{\partial x} \boldsymbol{AB} = \frac{\partial \boldsymbol{A}}{\partial x} \boldsymbol{B} + \boldsymbol{A} \frac{\partial \boldsymbol{B}}{\partial x} \qquad (63)
$$

となる

混合ガウス分布

- 上記の公式から、以下の公式を簡単に導ける

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

これに右から $\boldsymbol{A}^{-1}$ を掛ければ

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

混合ガウス分布

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

より

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

を得る (逆行列の微分公式)

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

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

混合ガウス分布

- この逆行列の微分公式を使えば、以下の微分公式を導出できる

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

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

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

混合ガウス分布

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

であるから

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

を得られる

混合ガウス分布

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

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

混合ガウス分布

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

混合ガウス分布

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

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

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

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

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

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

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

混合ガウス分布

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

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

混合ガウス分布

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

  - 以下を最大化する π_k を求める

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

混合ガウス分布

- π_k で微分して 0 と等置すると、次のようになる

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

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

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

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

両辺に π_k を掛けて

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

混合ガウス分布

k についての和を取ると

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

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

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

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

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

これより

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

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

混合ガウス分布

ここで、以下が成立することに注意

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

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

混合ガウス分布

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

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

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

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

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

混合ガウス分布

$\gamma(z_{ik})$ の更新式

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

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

混合ガウス分布

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

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

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

混合ガウス分布

混合ガウス分布に対する EM アルゴリズム

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

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

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

混合ガウス分布

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

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

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

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

但し

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

混合ガウス分布

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

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

パラメータの変化量、あるいは対数尤度の変化量をみて、収束性を判定

5. 収束基準を満たしていなければ、(2) に戻る

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

混合ガウス分布

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

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

混合ガウス分布

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

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

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

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 $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ 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 $\boldsymbol{\theta}$ は、各ガウス要素の平均 $\boldsymbol{\mu}_k$ と共分散行列 $\boldsymbol{\Sigma}_k$、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 $\ln p(\mathcal{D}|\boldsymbol{\theta})$ の最大化) によって、パラメータ $\boldsymbol{\theta}$ 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 $\boldsymbol{z}$ を導入し、$\boldsymbol{z}$ 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 $p(z_k = 1|\boldsymbol{x}_i)$ 即ち負担率 $\gamma(z_{ik})$ の計算と、パラメータ $\boldsymbol{\theta}$ 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

目次

目次

EM アルゴリズムの解釈

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

EM アルゴリズムの解釈

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

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

  - 対数尤度関数は次のようになる

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

EM アルゴリズムの解釈

- 潜在変数 z が連続変数の場合は

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

  のように、単に総和を積分に置き換えればよい

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

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

EM アルゴリズムの解釈

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

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

EM アルゴリズムの解釈

重要な仮定と考え方

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

EM アルゴリズムの解釈

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

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

連続潜在変数の場合は次のようになる

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

EM アルゴリズムの解釈

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

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

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

EM アルゴリズムの解釈

一般の EM アルゴリズム

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

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

EM アルゴリズムの解釈

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

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

但し

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

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

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

目次

混合ガウス分布の再解釈

- 先程の EM アルゴリズムの解釈で、混合ガウス分布を見直す

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

混合ガウス分布の再解釈

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

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

混合ガウス分布の再解釈

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

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

混合ガウス分布の再解釈

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

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

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

混合ガウス分布の再解釈

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

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

混合ガウス分布の再解釈

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

これより

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

混合ガウス分布の再解釈

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

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

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

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

のようになる

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

混合ガウス分布の再解釈

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

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

となる

混合ガウス分布の再解釈

- ここで、以下の微分公式を用いた

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

- これより

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

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

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

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

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

のようになる

混合ガウス分布の再解釈

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

混合ガウス分布の再解釈

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

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

  - π_k について微分して 0 とおくと、次のようになる

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

混合ガウス分布の再解釈

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

- これより、両辺に π_k を掛けて

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

全ての k について総和を取ると

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

混合ガウス分布の再解釈

- よって

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

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

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

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

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

混合ガウス分布の再解釈

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

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

混合ガウス分布の再解釈

- 事後確率分布は次のように書ける

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

以上より

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

混合ガウス分布の再解釈

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

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

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

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

混合ガウス分布の再解釈

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

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

- ここで

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

  と書けることに注意する

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

混合ガウス分布の再解釈

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

- また

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

  であることにも注意する

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

混合ガウス分布の再解釈

- これより

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

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

混合ガウス分布の再解釈

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

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

である

混合ガウス分布の再解釈

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

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

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

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

混合ガウス分布の再解釈

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

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

目次

K-Means 法との関連

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

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

K-Means 法との関連

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

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

K-Means 法との関連

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \sum_k \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{136}$$

$$= \sum_k \pi_k \frac{1}{(2\pi\epsilon)^{\frac{D}{2}}} \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x} - \boldsymbol{\mu}_k||^2\right\} \tag{137}$$

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

$$\gamma(z_{ik}) = \frac{\pi_k \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

$$= \frac{\pi_k \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x} - \boldsymbol{\mu}_k||^2\right\}}{\sum_j \pi_j \exp\left\{-\frac{1}{2\epsilon}||\boldsymbol{x} - \boldsymbol{\mu}_j||^2\right\}} \tag{138}$$

K-Means 法との関連

- 負担率は、以下のように変形できる

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

K-Means 法との関連

- ここで、k^* を次で定める

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

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

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

を考えると、全ての $j \neq k^*$ について

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

K-Means 法との関連

が成立するので

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

である

- 従って、$k = k^*$ のとき

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

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

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

- $k \neq k^*$ のとき

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

であって、両辺の $\epsilon \to 0$ による極限を取れば

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

K-Means 法との関連

となるから

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

が明らかに成立するほか、以下の不等式が

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

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

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

従って、$k \neq k^*$ の場合は

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

である

K-Means 法との関連

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

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

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

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

K-Means 法との関連

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

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

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

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

K-Means 法との関連

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

  - 次のように計算する

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

K-Means 法との関連

- 両辺に ϵ を掛けると

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

- $\epsilon \to 0$ の極限を取ると

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

であるから

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

K-Means 法との関連

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

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

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

K-Means 法との関連

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

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

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

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

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

K-Means 法との関連

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

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

目次

一般の EM アルゴリズム

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

一般の EM アルゴリズム

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

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

- 連続潜在変数の場合は、次のように、総和を積分に置き換えればよい

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

- ここでは、連続潜在変数の場合を考える

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

一般の EM アルゴリズム

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

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

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

一般の EM アルゴリズム

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

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

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

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

一般の EM アルゴリズム

- x が連続変数であれば、イェンセンの不等式は次のように書ける

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

例えば、$f(x) = -\ln x$ は凸関数であるから

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

よって

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

一般の EM アルゴリズム

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

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

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

一般の EM アルゴリズム

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

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

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

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

一般の EM アルゴリズム

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

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

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

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

一般の EM アルゴリズム

- 分解は次のように行える

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

一般の EM アルゴリズム

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

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

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

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

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

- $\mathcal{L}(q, \boldsymbol{\theta})$ は、$q(\boldsymbol{Z}), \boldsymbol{\theta}$ によらず、常に $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の下界をなす
- 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 $\mathrm{KL}(q\|p) \geqslant 0$, we see that the quantity $\mathcal{L}(q, \boldsymbol{\theta})$ is a lower bound on the log likelihood function $\ln p(\mathbf{X}|\boldsymbol{\theta})$.

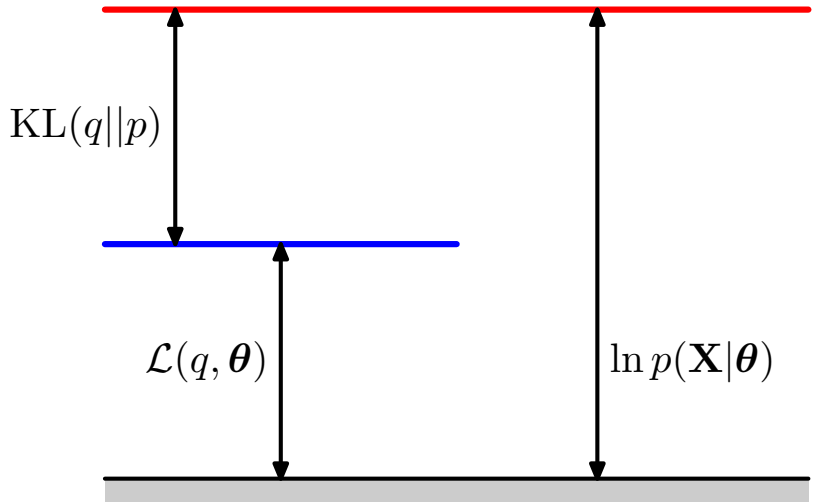

図 4: $\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ の分解

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
$\Leftrightarrow$ 集合 A に属する任意の要素 x は、a 以下である
$\Leftrightarrow \forall x \in A, x \leq a$ (175)

実数 $a \in \mathbb{R}$ が、実数の部分集合 $A \subset \mathbb{R}$ の下界である
$\Leftrightarrow$ 集合 A に属する任意の要素 x は、a 以上である
$\Leftrightarrow \forall x \in A, a \leq x$ (176)

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

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

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

$$
\begin{aligned}
& \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{old}}) \\
= \quad & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) + \mathrm{KL}(q||p) && (177) \\
= \quad & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) + \mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})) && (178)
\end{aligned}
$$

(E ステップ前)

一般の EM アルゴリズム

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

$$
\begin{aligned}
& \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{old}}) && (179)\\
= \; & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) + \mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})) \\
= \; & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) + \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})) && (180)\\
= \; & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) && (181)
\end{aligned}
$$

(E ステップ後)

一般の EM アルゴリズム

- 次の図 5 には E ステップの概要が示されている

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

一般の EM アルゴリズム

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

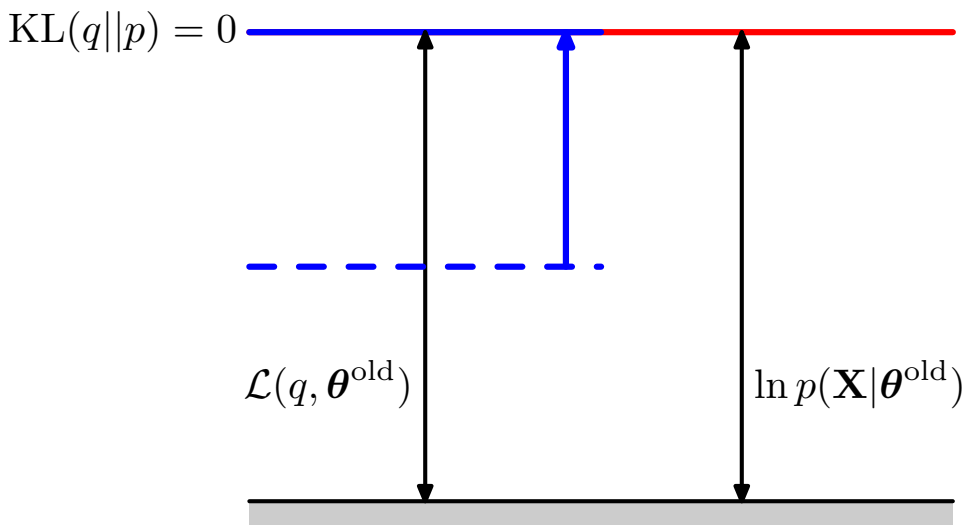

図 5: EM アルゴリズムの E ステップ

一般の EM アルゴリズム

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

$$
\begin{aligned}
& \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \\
= \quad & \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q||p) && (182) \\
= \quad & \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})) && (183) \\
= \quad & \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})) && (184) \\
& (\text{M ステップ前})
\end{aligned}
$$

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

一般の EM アルゴリズム

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

$$\begin{aligned} & \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{new}}) \\ = \quad & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}}) + \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{new}})) \quad (185) \\ & (\text{M ステップ後}) \end{aligned}$$

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

$$\begin{aligned} & \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{new}}) - \ln p(\boldsymbol{X}|\boldsymbol{\theta}^{\mathrm{old}}) \\ = \quad & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}}) + \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{new}})) - \\ & \qquad \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) \quad (186) \\ = \quad & \left(\mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}}) - \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}})\right) + \\ & \qquad \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})||p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{new}})) \quad (187) \\ \geq \quad & \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{new}}) - \mathcal{L}(q, \boldsymbol{\theta}^{\mathrm{old}}) \quad (188) \end{aligned}$$

一般の EM アルゴリズム

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

一般の EM アルゴリズム

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

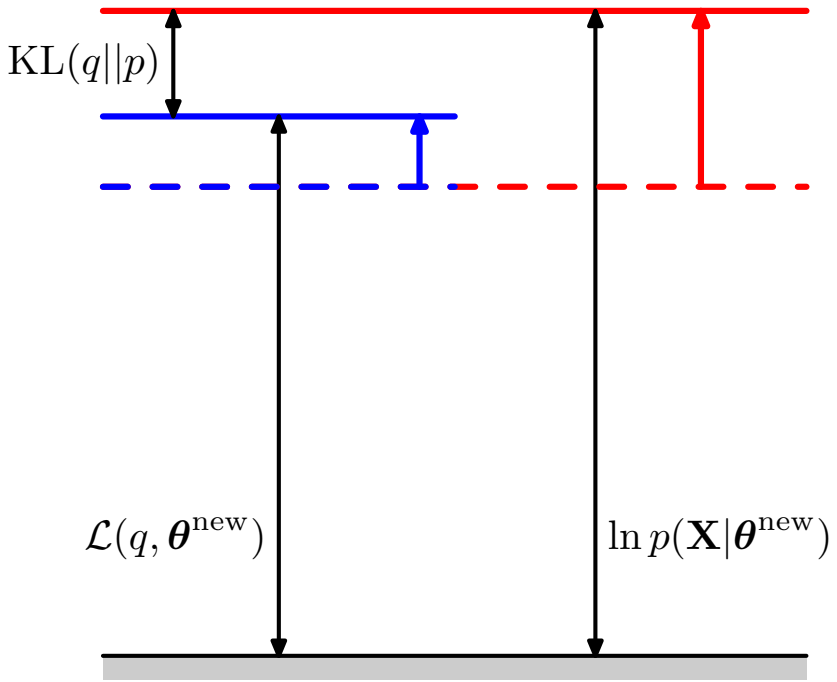

図 6: EM アルゴリズムの M ステップ

一般の EM アルゴリズム

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

$$
\begin{aligned}
& \mathcal{L}(q, \boldsymbol{\theta}) \\
= & \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
= & \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}})} && (189) \\
= & \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) - \\
& \qquad \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}}) && (190) \\
= & \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) + \text{Const.} && (191)
\end{aligned}
$$

一般の EM アルゴリズム

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

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

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

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

一般の EM アルゴリズム

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

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

一般の EM アルゴリズム

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

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

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

一般の EM アルゴリズム

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

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

一般の EM アルゴリズム

$$
\begin{aligned}
&= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}^{\text{new}})}{p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}})} + \\
&\qquad \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{new}})} \\
&= \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}^{\text{new}}) - \\
&\qquad \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}^{\text{old}}) - \\
&\qquad \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}})}{p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{new}})} \\
&= \mathcal{Q}(\boldsymbol{\theta}^{\text{old}},\boldsymbol{\theta}^{\text{new}}) - \mathcal{Q}(\boldsymbol{\theta}^{\text{old}},\boldsymbol{\theta}^{\text{old}}) + \\
&\qquad \text{KL}\left(p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{old}})||p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{\text{new}})\right) \qquad (192) \\
&\geq \mathcal{Q}(\boldsymbol{\theta}^{\text{old}},\boldsymbol{\theta}^{\text{new}}) - \mathcal{Q}(\boldsymbol{\theta}^{\text{old}},\boldsymbol{\theta}^{\text{old}}) \qquad (193) \\
&\geq 0
\end{aligned}
$$

一般の EM アルゴリズム

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

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

一般の EM アルゴリズム

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

$$\begin{aligned}
\ln p(\boldsymbol{X}|\boldsymbol{\theta}) &= \ln \sum_{\boldsymbol{Z}} p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \\
&= \ln \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
&\geq \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
&= \mathcal{L}(q, \boldsymbol{\theta}) \qquad (194)
\end{aligned}$$

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

一般の EM アルゴリズム

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

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

一般の EM アルゴリズム

$$
\begin{aligned}
&= -\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})}{q(\boldsymbol{Z})} \\
&= \mathrm{KL}(q||p) \qquad (195)
\end{aligned}
$$

ゆえ、$\mathrm{KL}(q||p)$ となることが分かったので

$$
\ln p(\boldsymbol{X}|\boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta}) + \mathrm{KL}(q||p) \qquad (196)
$$

のように分解できることが分かる

一般の EM アルゴリズム

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

一般の EM アルゴリズム

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

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

一般の EM アルゴリズム

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

$$
\begin{aligned}
& \frac{\partial}{\partial \boldsymbol{\theta}} \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}^{\mathrm{old}}} \\
= & \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}^{\mathrm{old}}} + \frac{\partial}{\partial \boldsymbol{\theta}} \mathrm{KL}(q||p) \Big|_{\boldsymbol{\theta}^{\mathrm{old}}} \\
= & \frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{L}(q, \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}^{\mathrm{old}}}
\end{aligned} \tag{197}
$$

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

  - 直感的には、次のように考えればよい

一般の EM アルゴリズム

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

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

一般の EM アルゴリズム

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

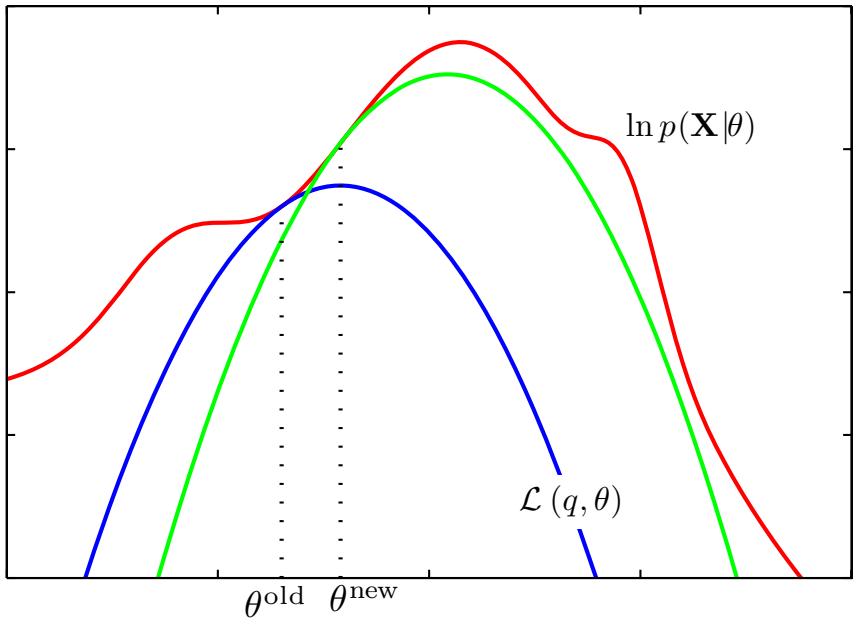

図 7: EM アルゴリズムの手続き

一般の EM アルゴリズム

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

$$p(\boldsymbol{X}, \boldsymbol{Z}) = \prod_i p(\boldsymbol{x}_i, \boldsymbol{z}_i) \tag{198}$$

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

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

一般の EM アルゴリズム

$$
\begin{aligned}
&= \frac{\prod_i p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})}{\prod_i \sum_{\boldsymbol{Z}} p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})} \\
&= \frac{\prod_i p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})}{\prod_i p(\boldsymbol{x}_i|\boldsymbol{\theta})} \\
&= \prod_i \frac{p(\boldsymbol{x}_i, \boldsymbol{z}_i|\boldsymbol{\theta})}{p(\boldsymbol{x}_i|\boldsymbol{\theta})} \\
&= \prod_i p(\boldsymbol{z}_i|\boldsymbol{x}_i, \boldsymbol{\theta}) \qquad (199)
\end{aligned}
$$

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

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

一般の EM アルゴリズム

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

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

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

目次

MAP 推定に対する EM アルゴリズム

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

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

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

$$= \arg\max_{\boldsymbol{\theta}} \frac{p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{X})} \tag{202}$$

$$= \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{203}$$

MAP 推定に対する EM アルゴリズム

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

$$
\begin{aligned}
\ln p(\boldsymbol{\theta}|\boldsymbol{X}) &= \ln \frac{p(\boldsymbol{X},\boldsymbol{\theta})}{p(\boldsymbol{X})} && (204)\\
&= \ln \frac{p(\boldsymbol{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{X})} \\
&= \ln p(\boldsymbol{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - p(\boldsymbol{X}) && (205)\\
&= \mathcal{L}(q,\boldsymbol{\theta}) + \mathrm{KL}(q||p) + \ln p(\boldsymbol{\theta}) - p(\boldsymbol{X}) && (206)\\
&\geq \mathcal{L}(q,\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - p(\boldsymbol{X}) && (207)\\
&= \mathcal{L}(q,\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) + \mathrm{Const.} && (208)
\end{aligned}
$$

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

MAP 推定に対する EM アルゴリズム

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

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

目次

EM アルゴリズムの拡張

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

EM アルゴリズムの拡張

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

目次

パラメータについての補足

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

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

目次

逐次型の EM アルゴリズムの例

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

$$p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}) = \prod_i p(\boldsymbol{z}_i|\boldsymbol{x}_i,\boldsymbol{\theta}) \tag{209}$$

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

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

逐次型の EM アルゴリズムの例

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

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

$$d = \gamma^{\mathrm{new}}(z_{mk}) - \gamma^{\mathrm{old}}(z_{mk}) \tag{210}$$

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

$$N_k^{\mathrm{new}} = N_k^{\mathrm{old}} + \gamma^{\mathrm{new}}(z_{mk}) - \gamma^{\mathrm{old}}(z_{mk}) = N_k^{\mathrm{old}} + d \tag{211}$$

逐次型の EM アルゴリズムの例

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

$$\boldsymbol{\mu}_k^{\text{old}} = \frac{1}{N_k^{\text{old}}} \sum_i \gamma^{\text{old}}(z_{ik}) \boldsymbol{x}_i \tag{212}$$

$$\boldsymbol{\Sigma}_k^{\text{old}} = \frac{1}{N_k^{\text{old}}} \sum_i \gamma^{\text{old}}(z_{ik}) (\boldsymbol{x}_i - \boldsymbol{\mu}_k^{\text{old}})(\boldsymbol{x}_i - \boldsymbol{\mu}_k^{\text{old}})^T \tag{213}$$

$$\pi_k^{\text{old}} = \frac{N_k^{\text{old}}}{N} \tag{214}$$

但し

$$N_k^{\text{old}} = \sum_i \gamma^{\text{old}}(z_{ik}) \tag{215}$$

逐次型の EM アルゴリズムの例

- 平均の更新式は

$$\begin{aligned}
\boldsymbol{\mu}_k^{\text{new}} &= \frac{1}{N_k^{\text{new}}} \left(\sum_{i \neq m} \gamma^{\text{old}}(z_{ik}) \boldsymbol{x}_i + \gamma^{\text{new}}(z_{mk}) \boldsymbol{x}_m \right) && (216) \\
&= \frac{1}{N_k^{\text{new}}} \left(N_k^{\text{old}} \boldsymbol{\mu}_k^{\text{old}} - \gamma^{\text{old}}(z_{mk}) \boldsymbol{x}_m + \gamma^{\text{new}}(z_{mk}) \boldsymbol{x}_m \right) \\
&= \frac{1}{N_k^{\text{new}}} \left(\left(N_k^{\text{new}} - \gamma^{\text{new}}(z_{mk}) + \gamma^{\text{old}}(z_{mk}) \right) \boldsymbol{\mu}_k^{\text{old}} - \right. \\
&\qquad \left. \gamma^{\text{old}}(z_{mk}) \boldsymbol{x}_m + \gamma^{\text{new}}(z_{mk}) \boldsymbol{x}_m \right) \\
&= \boldsymbol{\mu}_k^{\text{old}} + \frac{1}{N_k^{\text{new}}} \left(-(\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})) \boldsymbol{\mu}_k^{\text{old}} + \right. \\
&\qquad \left. (\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})) \boldsymbol{x}_m \right) \\
&= \boldsymbol{\mu}_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right) && (217)
\end{aligned}$$

逐次型の EM アルゴリズムの例

$$= \boldsymbol{\mu}_k^{\text{old}} + \frac{d}{N_k^{\text{new}}} \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right) \tag{218}$$

- 分散の更新式は

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{1}{N_k^{\text{new}}} \sum_i \gamma^{\text{new}}(z_{ik}) \boldsymbol{x}_i \boldsymbol{x}_i^T - \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \tag{219}$$

$$= \frac{1}{N_k^{\text{new}}} \left(\sum_{i \neq m} \gamma^{\text{old}}(z_{ik}) \boldsymbol{x}_i \boldsymbol{x}_i^T + \gamma^{\text{new}}(z_{mk}) \boldsymbol{x}_m \boldsymbol{x}_m^T \right) - \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \tag{220}$$

$$= \frac{1}{N_k^{\text{new}}} \left(\left(\sum_i \gamma^{\text{old}}(z_{ik}) \boldsymbol{x}_i \boldsymbol{x}_i^T - \gamma^{\text{old}}(z_{mk}) \boldsymbol{x}_m \boldsymbol{x}_m^T \right) + \gamma^{\text{new}}(z_{mk}) \boldsymbol{x}_m \boldsymbol{x}_m^T \right) - \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \tag{221}$$

逐次型の EM アルゴリズムの例

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

逐次型の EM アルゴリズムの例

ここで

$$
\begin{aligned}
& N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{new}} (\boldsymbol{\mu}_k^{\text{new}})^T \\
= \quad & \left(N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{old}} + d(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}})\right) (\boldsymbol{\mu}_k^{\text{new}})^T \\
= \quad & \frac{1}{N_k^{\text{new}}} \left(N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{old}} + d(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}})\right) \\
& \qquad \left(N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{old}} + d\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\right)^T \\
= \quad & N_k^{\text{new}} \boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T + 2\boldsymbol{\mu}_k^{\text{old}} d \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T + \\
& \qquad \frac{1}{N_k^{\text{new}}} d^2 \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right) \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \qquad (224) \\
= \quad & (N_k^{\text{old}} + d) \boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T + \\
& \qquad 2d\boldsymbol{\mu}_k^{\text{old}} \boldsymbol{x}_m^T - 2d\boldsymbol{\mu}_k^{\text{old}} (\boldsymbol{\mu}_k^{\text{old}})^T + \\
& \qquad \frac{1}{N_k^{\text{new}}} d^2 \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right) \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \qquad (225)
\end{aligned}
$$

逐次型の EM アルゴリズムの例

であるから

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

逐次型の EM アルゴリズムの例

$$
\begin{aligned}
& \frac{1}{N_k^{\text{new}}} d^2 \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \\
= \quad & \frac{d}{N_k^{\text{new}}} \left(N_k^{\text{new}} - d\right)\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \\
= \quad & \frac{d}{N_k^{\text{new}}} N_k^{\text{old}} \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)\left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \qquad (226)
\end{aligned}
$$

以上より

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

逐次型の EM アルゴリズムの例

$$
\begin{aligned}
&= \frac{N_k^{\text{old}}}{N_k^{\text{new}}} \left(\boldsymbol{\Sigma}_k^{\text{old}} + \frac{d}{N_k^{\text{new}}} \right. \\
&\qquad \left. \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right) \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right)^T \right) \\
&= \frac{N_k^{\text{old}}}{N_k^{\text{new}}} \left(\boldsymbol{\Sigma}_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{ik}) - \gamma^{\text{old}}(z_{ik})}{N_k^{\text{new}}} \right. \\
&\qquad \left. \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right) \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}} \right)^T \right) \qquad (227)
\end{aligned}
$$

- 混合係数の更新式は

$$
\pi_k^{\text{new}} = \frac{N_k^{\text{new}}}{N} = \frac{N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N} \qquad (228)
$$

逐次型の EM アルゴリズムの例

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

$$\boldsymbol{\mu}_k^{\text{new}} = \boldsymbol{\mu}_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N_k^{\text{new}}} \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right) \tag{229}$$

$$\boldsymbol{\Sigma}_k^{\text{new}} = \frac{N_k^{\text{old}}}{N_k^{\text{new}}} \left(\boldsymbol{\Sigma}_k^{\text{old}} + \frac{\gamma^{\text{new}}(z_{ik}) - \gamma^{\text{old}}(z_{ik})}{N_k^{\text{new}}} \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right) \left(\boldsymbol{x}_m - \boldsymbol{\mu}_k^{\text{old}}\right)^T \right) \tag{230}$$

$$\pi_k^{\text{new}} = \frac{N_k^{\text{new}}}{N} = \frac{N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk})}{N} \tag{231}$$

但し

$$N_k^{\text{new}} = N_k^{\text{old}} + \gamma^{\text{new}}(z_{mk}) - \gamma^{\text{old}}(z_{mk}) \tag{232}$$

逐次型の EM アルゴリズムの例

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

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

逐次型の EM アルゴリズムの例

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

EM アルゴリズムのまとめ

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

EM アルゴリズムのまとめ

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

目次

目次

EM アルゴリズムが困難な場合

- EM アルゴリズムで行う計算
  - E ステップでは、潜在変数の事後確率分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}})$ を計算
  - M ステップでは、完全データ対数尤度 $\ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta})$ の期待値を計算

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \tag{233}$$

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \int_{\boldsymbol{Z}} p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) d\boldsymbol{Z} \tag{234}$$

そして、$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$ を最大化するパラメータ $\boldsymbol{\theta}^{\text{new}}$ を求める

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) \tag{235}$$

EM アルゴリズムが困難な場合

- EM アルゴリズムの困難さ
  - 実際に扱うモデルでは、事後分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta}^{\mathrm{old}})$ の計算や、事後分布に従った期待値 $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\mathrm{old}})$ の計算が、不可能であることが多い
  - 隠れ変数の次元が高すぎる
  - 事後分布が複雑な形をしていて、期待値を解析的に計算できない
  - 連続変数であれば、積分が閉形式の解を持たないかもしれない
  - 空間の次元の問題や、被積分項の複雑さから、数値積分すら困難かもしれない
  - 離散変数であれば、期待値を計算するためには、潜在変数の可能な全ての組み合わせについての和を取る必要がある
  - 隠れ変数の次元が高くなると、組み合わせ数が指数的に増大する
  - 計算量が大きすぎて、期待値の厳密な計算がもはや不可能

EM アルゴリズムが困難な場合

- 近似法
  - EM アルゴリズムが困難であるとき、何らかの方法で近似しなければならない
  - 近似法は、確率的な近似と、決定的な近似の 2 つに分けられる
- 確率的な近似
  - マルコフ連鎖モンテカルロ法などの手法がある
  - 無限の計算資源があれば、厳密な結果が得られる
  - 実際には計算量が有限であるため、得られる解は近似解となる
- 決定的な近似
  - 事後分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ を解析的に近似する
  - 事後分布に対して、何らかの仮定をおく
  - 例えば、単純な項の積として分解できる、あるいは、(ガウス分布などの特別な) パラメトリックな分布であるといった仮定

目次

変分推論

- ここで扱う近似法
  - 変分推論法 (Variational inference) あるいは変分ベイズ法 (Variational Bayes) について扱う

- 変分推論 (Variational inference)
  - 18 世紀のオイラー、ラグランジュらによる変分法 (Calculus of variations) に起源をもつ

  - まずは、変分法について説明をしていく

変分法

- 関数と汎関数の違い
  - 通常の関数は、入力として値をとり、出力として関数の値を返す
  - 通常の関数は、値から値への写像である
  - 関数の導関数は、入力値を微小に変えたときに、出力の関数値がどの程度変わるかを表す
  - 汎関数 (Functional) とは、入力として関数をとり、出力として汎関数の値を返す
  - 汎関数は、関数から値への写像である
  - 汎関数微分 (Functional derivative) とは、入力関数が微小に変わったときに、出力の汎関数値がどの程度変わるかを表す
  - 汎関数の微分を、変分という

変分法

- 汎関数の例
  - エントロピー $H[p]$ は、確率分布 $p(x)$ を入力として、以下の量を返す汎関数

$$H[p] = -\int p(x) \ln p(x) dx \tag{236}$$

- 汎関数の最適化
  - 多くの問題は、汎関数の値を最適化する問題として定式化できる
  - 汎関数の最適化とは、可能な全ての入力関数の中から、汎関数の値を最大化、あるいは最小化するような関数を選び出すことである
  - 通常の最適化では、可能な全てのパラメータ (入力値) の中から、関数を最大化、あるいは最小化するような 1 つのパラメータを選び出す
  - 次は、いよいよ変分の計算について説明する

変分法

- 変分法
  - 通常の微分を使えば、ある関数 $y(x)$ を最大化 (最小化) するような x の値が求められる
  - 変分法を使えば、汎関数 $F[y]$ を最大化 (最小化) するような、関数 $y(x)$ が求められる
  - 従って、可能な全ての関数 $y(x)$ の中から、$F[y]$ を最大 (最小) にするような関数が得られる
- 変分法によって解ける問題の例
  - 2 点を結ぶ最短経路は? (答えは直線)
  - 最速降下曲線は? (答えはサイクロイド)
  - エントロピーが最大になるような確率分布は? (答えはガウス分布)

変分法

- 通常の微分の表現
  - 関数 $y(x+\epsilon)$ のテイラー展開は次のように記述できた

$$y(x+\epsilon) = \sum_{n=0}^{\infty} \frac{y^{(n)}(x)}{n!}\epsilon^n \tag{237}$$

$$= y(x) + \frac{dy}{dx}\epsilon + \frac{1}{2!}\frac{d^2y}{dx^2}\epsilon^2 + \frac{1}{3!}\frac{d^3y}{dx^3}\epsilon^3 + \cdots \tag{238}$$

$$= y(x) + \frac{dy}{dx}\epsilon + O(\epsilon^2) \tag{239}$$

  - これより微分 dy/dx は、次のように求められる
  - 変数 x に微小な変化 ϵ を加え、このときの関数値 $y(x+\epsilon)$ を ϵ の累乗形として表現する
  - 最後に $\epsilon \to 0$ の極限をとればよい

$$\frac{dy}{dx} = \lim_{\epsilon \to 0} \frac{y(x+\epsilon) - y(x)}{\epsilon} \tag{240}$$

変分法

- 多変数関数 $y(x_1, \ldots, x_D)$ の偏微分の表現
  - 多変数関数 $y(x_1, \ldots, x_D)$ のテイラー展開は次のように記述できた

$$\boldsymbol{D}^n = \left(\epsilon_1 \frac{\partial y}{\partial x_i} + \cdots + \epsilon_D \frac{\partial y}{\partial x_D} \right)^n \tag{241}$$

上記のような演算子 $\boldsymbol{D}$ を考えれば

$$\begin{aligned} & y(x_1 + \epsilon_1, \ldots, x_D + \epsilon_D) \\ = & \sum_{n=0}^{\infty} \frac{1}{n!} (\boldsymbol{D}^n y)(x_1, \ldots, x_D) \qquad (242) \\ = & y(x_1, \ldots, x_D) + \sum_{i=1}^{D} \frac{\partial y}{\partial x_i} \epsilon_i + \frac{1}{2!} \sum_{i=1}^{D} \sum_{j=1}^{D} \frac{\partial^2 y}{\partial x_i x_j} \epsilon_i \epsilon_j + \\ & \frac{1}{3!} \sum_{i=1}^{D} \sum_{j=1}^{D} \sum_{k=1}^{D} \frac{\partial^3 y}{\partial x_i x_j x_k} \epsilon_i \epsilon_j \epsilon_k + \cdots \qquad (243) \end{aligned}$$

であるから

$$\begin{aligned} & y(x_1 + \epsilon_1, \ldots, x_D + \epsilon_D) \\ = \quad & y(x_1, \ldots, x_D) + \sum_{i=1}^{D} \frac{\partial y}{\partial x_i} \epsilon_i + O(\epsilon^2) \end{aligned} \tag{244}$$

- これより偏微分 $\partial y / \partial x_i$ は、次のように求められる

$$\begin{aligned} \frac{\partial y}{\partial x_i} = \lim_{\epsilon_i \to 0} \frac{1}{\epsilon_i} & (y(x_1, \ldots, x_{i-1}, x_i + \epsilon_i, x_{i+1}, \ldots, x_D) - \\ & y(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_D)) \end{aligned} \tag{245}$$

変分法

- 変分の表現
  - 多少不正確だが、変分をどのように定義すればよいか考えてみる
  - ここで、各 x_i に対する関数の値 $z_i = y(x_i)$ を個別の変数とみなして、次の関数 $F(z_1, \ldots, z_D)$ について考えてみよう

$$\begin{aligned} & F(z_1 + \epsilon\eta(x_1), \ldots, z_D + \epsilon\eta(x_D)) \\ = \quad & F(z_1, \ldots, z_D) + \sum_{i=1}^{D} \frac{\partial F}{\partial z_i} \epsilon\eta(x_i) + O(\epsilon^2) \end{aligned} \tag{246}$$

$z_i = y(x_i)$ を代入してみると

$$\begin{aligned} & F(y(x_1) + \epsilon\eta(x_1), \ldots, y(x_D) + \epsilon\eta(x_D)) \\ = \quad & F(y(x_1), \ldots, y(x_D)) + \sum_{i=1}^{D} \frac{\partial F}{\partial y(x_i)} \epsilon\eta(x_i) + O(\epsilon^2) \end{aligned} \tag{247}$$

変分法

- ここで $D \to \infty$ の極限を取り、$x_1, \ldots, x_D$ が、ある連続した区間 $[a, b]$ に含まれる、全ての実数を表すことにする
- このとき $y(x_1), \ldots, y(x_D)$ は、実数の区間 $[a, b]$ で定義される連続関数 $y(x)$ として書けることが分かる
- 同様に $y(x_1) + \epsilon\eta(x_1), \ldots, y(x_D) + \epsilon\eta(x_D)$ は、実数の区間 $[a, b]$ で定義される連続関数 $y(x) + \epsilon\eta(x)$ として、まとめることができる
- 関数 $\eta(x)$ も、実数の区間 $[a, b]$ で定義される連続関数
- $\epsilon\eta(x)$ は、$y(x)$ に加わる摂動として、考えることができる

変分法

- 関数 F は、関数 $y(x)$ や $y(x) + \epsilon\eta(x)$ を入力として受け取る、汎関数 $F[y]$ として解釈できるから、次のように書ける

$$F(y(x_1) + \epsilon\eta(x_1), \ldots, y(x_D) + \epsilon\eta(x_D)) = F[y(x) + \epsilon\eta(x)] \tag{248}$$

$$F(y(x_1), \ldots, y(x_D)) = F[y(x)] \tag{249}$$

- 以下の項は、入力を $y(x)$ に摂動を加えて $y(x) + \epsilon\eta(x)$ へと微小に変化させたときの、汎関数の ($F[y(x)]$ から $F[y(x) + \epsilon\eta(x)]$ への) 変化量を表している

$$\sum_{i=1}^{D} \frac{\partial F}{\partial y(x_i)} \epsilon\eta(x_i) \tag{250}$$

- 点 x_i における汎関数 F の変化量を、$x_1, \ldots, x_D$ の範囲について、即ち、実数の区間 $[a, b]$ について足し合わせていると解釈する

変分法

- $D \to \infty$ のとき、$x_1, \ldots, x_D$ は区間 $[a, b]$ における全ての実数を表すから、総和を積分に置き換えられそうである
- 汎関数の微分 $\dfrac{\delta F}{\delta y(x)}$ を使えば、次のように書ける

$$\sum_{i=1}^{D} \frac{\partial F}{\partial y(x_i)} \epsilon \eta(x_i)$$

$$\Rightarrow \quad \int_a^b \frac{\delta F}{\delta y(x)} \epsilon \eta(x) dx = \epsilon \int_a^b \frac{\delta F}{\delta y(x)} \eta(x) dx \tag{251}$$

- 結局、変分 $\dfrac{\delta F}{\delta y(x)}$ は次のように定義できる

$$F[y(x) + \epsilon \eta(x)] = F[y(x)] + \epsilon \int_a^b \frac{\delta F}{\delta y(x)} \eta(x) dx + O(\epsilon^2) \tag{252}$$

変分法

- $F[y]$ は、区間 $[a, b]$ で定義される関数 y を受け取るとする

- 変分 $\delta F/\delta y$ は、入力関数 $y(x)$ に、任意の微小な変化 $\epsilon\eta(x)$ を加えたときの、汎関数 $F[y]$ の変化量として定義できる
- $\eta(x)$ は x についての任意の関数

変分法

Figure D.1 A functional derivative can be defined by considering how the value of a functional $F[y]$ changes when the function $y(x)$ is changed to $y(x) + \epsilon\eta(x)$ where $\eta(x)$ is an arbitrary function of x.

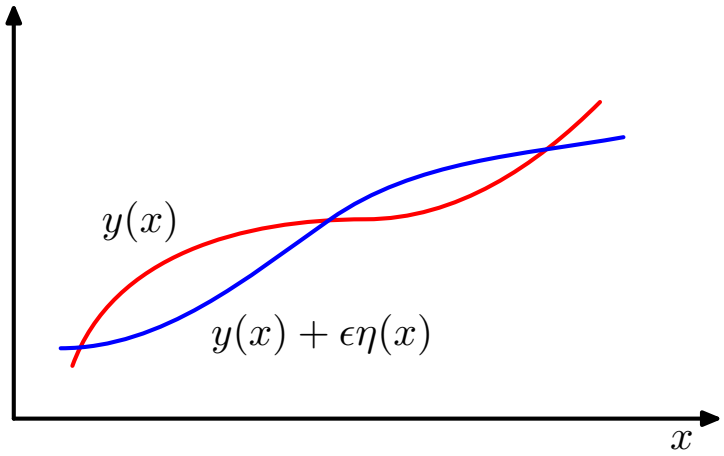

図 8: $y(x)$ と $y(x) + \epsilon\eta(x)$ の表現

変分法

- 変分法の例
  - 次の図 9 を使って、実際に変分を求めてみよう
  - 汎関数 $F[y]$ は、次のように定義されるとする

$$F[y] = \int_a^b y(x)dx \tag{253}$$

  - 汎関数の値 $F[y(x)], F[y(x) + \epsilon\eta(x)]$ は次のようになる

$$F[y(x)] = \int_a^b y(x)dx \tag{254}$$

$$F[y(x) + \epsilon\eta(x)] = \int_a^b (y(x) + \epsilon\eta(x))\, dx \tag{255}$$

変分法

- $F[y(x) + \epsilon\eta(x)]$ は次のように分解できる

$$F[y(x) + \epsilon\eta(x)] = \int_a^b y(x)dx + \epsilon \int_a^b \eta(x)dx \tag{256}$$

$$= F[y(x)] + \epsilon \int_a^b \eta(x)dx \tag{257}$$

- ここで、変分の定義式は

$$F[y(x) + \epsilon\eta(x)] = F[y(x)] + \epsilon \int_a^b \frac{\delta F}{\delta y(x)} \eta(x)dx + O(\epsilon^2) \tag{258}$$

であったので、上の 2 つの式を見比べれば、変分 $\delta F/\delta y$ は結局

$$\frac{\delta F}{\delta y(x)} = 1 \tag{259}$$

となることが分かる

変分法

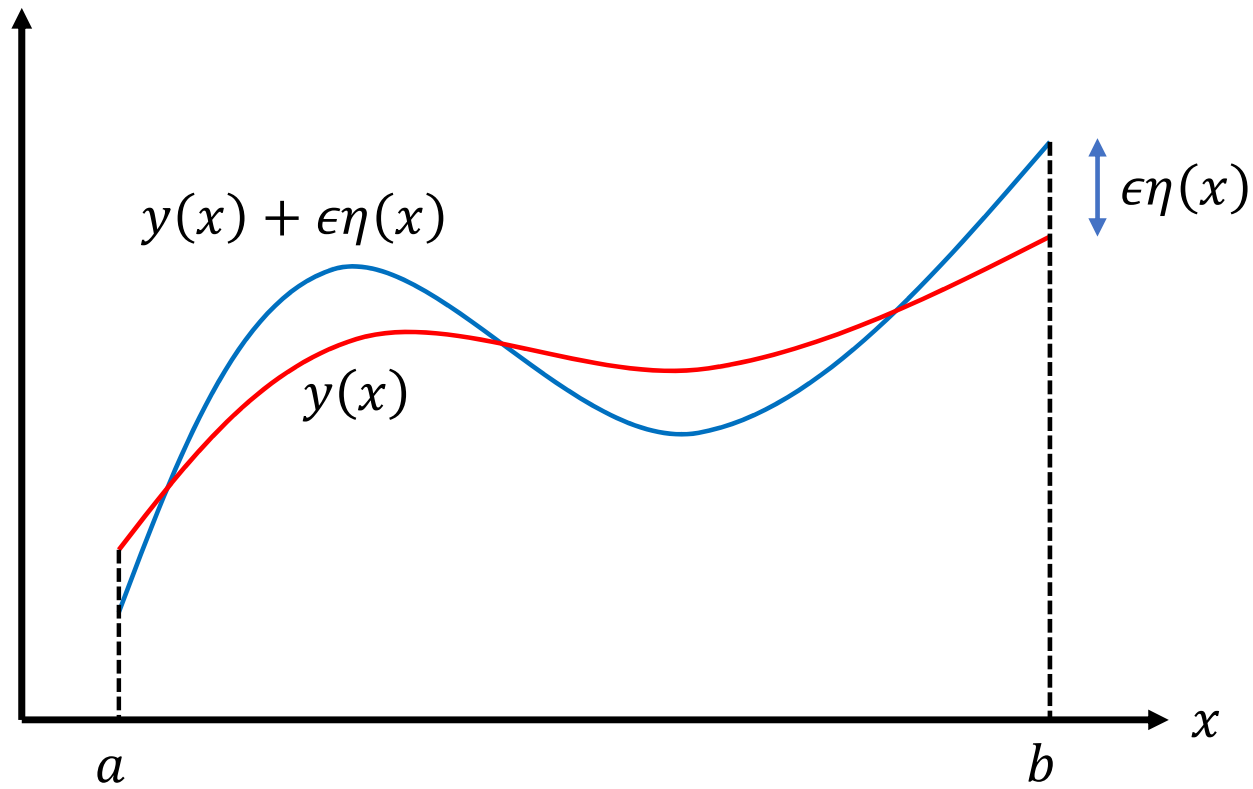

図 9: 区間 $[a, b]$ で定義された関数 $y(x)$ の表現

変分法

- 汎関数の最適化
  - 汎関数 $F[y]$ が最大 (最小) となるとき、関数 $y(x)$ の微小な変化に対して、汎関数は変化しないはず
  - 即ち、汎関数が最大 (最小) となるとき、$F[y(x) + \epsilon\eta(x)] = F[y(x)]$ が成り立つ
  - 従って、変分の定義式から、以下が成り立つ

$$\int_a^b \frac{\delta F}{\delta y(x)} \eta(x) dx = 0 \tag{260}$$

  - 上式は任意の $\eta(x)$ について成立しなければならない
  - 従って、変分 $\delta F/\delta y$ は、任意の x について 0 とならなければならない

  - 汎関数 $F[y]$ が最大 (最小) となるとき、$\delta F/\delta y = 0$ が成立することが分かった (通常の微分と同じ)

変分法

- 変分法の例
  - 様々な汎関数について、変分を導出してみよう
  - また、その汎関数が最大 (最小) となるときに成り立つ条件を、導出してみよう

変分法

- 汎関数の例 (1)
  - $y(x)$ とその微分 $y'(x) = dy/dx$、そして x によって決まる関数 $G(y(x), y'(x), x)$ があるとする
  - 汎関数 $F[y]$ を、$G(y(x), y'(x), x)$ を区間 $[a, b]$ にわたって積分した結果を出力する関数として、次のように定める

$$F[y] = \int_a^b G(y(x), y'(x), x)dx \tag{261}$$

  - 積分区間は無限であってもよいとする ($a = -\infty, b = \infty$ でもよい)

変分法

- $y(x)$ に摂動 $\epsilon\eta(x)$ を加えたときの、汎関数の値 $F[y(x)+\epsilon\eta(x)]$ を使って、変分 $\delta F/\delta y$ を調べてみる

$$F[y(x)+\epsilon\eta(x)] = \int_a^b G(y(x)+\epsilon\eta(x), y'(x)+\epsilon\eta'(x), x)dx \quad (262)$$

ここで、被積分項のテーラー展開を考えれば

$$\begin{aligned} & G(y(x)+\epsilon\eta(x), y'(x)+\epsilon\eta'(x), x) \\ = \; & G(y(x), y'(x), x) + \frac{\partial G}{\partial y}\epsilon\eta(x) + \\ & \frac{\partial G}{\partial y'}\epsilon\eta'(x) + \frac{\partial G}{\partial x}\cdot 0 + O(\epsilon^2) \quad (263) \\ = \; & G(y(x), y'(x), x) + \epsilon\left(\frac{\partial G}{\partial y}\eta(x) + \frac{\partial G}{\partial y'}\eta'(x)\right) + O(\epsilon^2) \quad (264) \end{aligned}$$

であるから

$$
\begin{aligned}
& F[y(x)+\epsilon\eta(x)] \\
= & \int_a^b G(y(x)+\epsilon\eta(x), y'(x)+\epsilon\eta'(x), x)dx \\
= & \int_a^b \left(G(y(x), y'(x), x) + \right. \\
& \qquad \left. \epsilon\left(\frac{\partial G}{\partial y}\eta(x) + \frac{\partial G}{\partial y'}\eta'(x)\right) + O(\epsilon^2) \right) dx \qquad (265) \\
= & \int_a^b G(y(x), y'(x), x)dx + \\
& \qquad \epsilon\int_a^b \left(\frac{\partial G}{\partial y}\eta(x) + \frac{\partial G}{\partial y'}\eta'(x)\right) dx + O(\epsilon^2) \qquad (266) \\
= & F[y(x)] + \epsilon\int_a^b \left(\frac{\partial G}{\partial y}\eta(x) + \frac{\partial G}{\partial y'}\eta'(x)\right) dx + O(\epsilon^2) \qquad (267)
\end{aligned}
$$

変分法

ここで

$$
\begin{aligned}
& \int_a^b \left(\frac{\partial G}{\partial y} \eta(x) + \frac{\partial G}{\partial y'} \eta'(x) \right) dx \\
= & \int_a^b \left(\frac{\partial G}{\partial y} \eta(x) \right) dx + \int_a^b \left(\frac{\partial G}{\partial y'} \eta'(x) \right) dx && (268) \\
= & \int_a^b \left(\frac{\partial G}{\partial y} \eta(x) \right) dx + \\
& \left[\frac{\partial G}{\partial y'} \eta(x) \right]_a^b - \int_a^b \frac{d}{dx} \left(\frac{\partial G}{\partial y'} \right) \eta(x) dx && (269) \\
= & \left[\frac{\partial G}{\partial y'} \eta(x) \right]_a^b + \int_a^b \left(\frac{\partial G}{\partial y} - \frac{d}{dx} \left(\frac{\partial G}{\partial y'} \right) \right) \eta(x) dx && (270)
\end{aligned}
$$

である

変分法

- 途中の式変形では、部分積分を使っていることに注意

- いま、積分区間の両端において、$y(x)$ の値は固定されているとする
- これを固定端条件という (図 10)

- このとき、$\eta(a) = \eta(b) = 0$ であるから、上式の最初の項が消えて

$$\begin{aligned} & \int_a^b \left(\frac{\partial G}{\partial y}\eta(x) + \frac{\partial G}{\partial y'}\eta'(x) \right) dx \\ = & \left[\frac{\partial G}{\partial y'}\eta(x) \right]_a^b + \int_a^b \left(\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'} \right) \right) \eta(x) dx \qquad (271) \\ = & \int_a^b \left(\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'} \right) \right) \eta(x) dx \qquad (272) \end{aligned}$$

のようになる

変分法

- 従って、摂動を加えたときの汎関数の値 $F[y(x)+\epsilon\eta(x)]$ は

$$\begin{aligned} & F[y(x)+\epsilon\eta(x)] \\ = \quad & F[y(x)]+\epsilon\int_a^b\left(\frac{\partial G}{\partial y}-\frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right)\right)\eta(x)dx+O(\epsilon^2) \end{aligned} \tag{273}$$

となる

- 上式を、変分の定義式と比べれば

$$F[y(x)+\epsilon\eta(x)]=F[y(x)]+\epsilon\int_a^b\frac{\delta F}{\delta y(x)}\eta(x)dx+O(\epsilon^2) \tag{274}$$

変分 $\delta F/\delta y$ は次のように書ける

$$\frac{\delta F}{\delta y(x)}=\frac{\partial G}{\partial y}-\frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right) \tag{275}$$

変分法

- 汎関数 $F[y]$ が最大 (最小) になるとき、変分 $\delta F/\delta y$ が 0 になる
- 従って、汎関数が最大 (最小) になるとき、以下の方程式が成り立つ

$$\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right) = 0 \tag{276}$$

- これをオイラー-ラグランジュ方程式という

- オイラー-ラグランジュ方程式は、次のような考え方で導出することもできる

- $F[y]$ が最大 (最小) であれば、摂動 $\epsilon\eta(x)$ によって $y(x)$ が少し変化しても、$F[y]$ の値は変化しないはず
- 従って、$F[y]$ が最大 (最小) であるとき、$F[y]$ の ϵ による微分は 0 になるはず

変分法

- これを数式で表現すると、次のようになる

$$\left.\frac{\partial F[y]}{\partial \epsilon}\right|_{\epsilon=0} = 0 \tag{277}$$

左辺は通常の偏微分であり、これを計算すると

$$\begin{aligned}
& \frac{\partial F[y]}{\partial \epsilon} \\
= & \frac{\partial}{\partial \epsilon}\int_a^b G(y, y', x)dx & (278)\\
= & \int_a^b \frac{\partial}{\partial \epsilon} G(y, y', x)dx & (279)\\
= & \int_a^b \left(\frac{\partial G}{\partial y}\frac{\partial y}{\partial \epsilon} + \frac{\partial G}{\partial y'}\frac{\partial y'}{\partial \epsilon} + \frac{\partial G}{\partial x}\frac{\partial x}{\partial \epsilon}\right) dx & (280)\\
= & \int_a^b \left(\frac{\partial G}{\partial y}\eta(x) + \frac{\partial G}{\partial y'}\eta'(x)\right) dx & (281)
\end{aligned}$$

$$= \left[\frac{\partial G}{\partial y'}\eta(x)\right]_a^b + \int_a^b \left(\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right)\right)\eta(x)dx \quad (282)$$

$$(\because \eta(a) = \eta(b) = 0)$$

$$= \int_a^b \left(\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right)\right)\eta(x)dx \quad (283)$$

$$= 0$$

- 上の式変形では、$y = y(x) + \epsilon\eta(x)$ であるから

$$\frac{\partial y}{\partial \epsilon} = \eta(x) \quad (284)$$

$$\frac{\partial y'}{\partial \epsilon} = \frac{\partial}{\partial \epsilon}\left(\frac{\partial y}{\partial x}\right) = \frac{\partial}{\partial \epsilon}\left(y'(x) + \epsilon\eta'(x)\right) = \eta'(x) \quad (285)$$

$$\frac{\partial x}{\partial \epsilon} = 0 \quad (286)$$

が成立することを利用している

変分法

- 任意の $\eta(x)$ について、上式が恒等的に成り立つためには

$$\frac{\partial G}{\partial y} - \frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right) = 0 \tag{287}$$

でなければならないことが分かり、先程と同様に、オイラー-ラグランジュ方程式を得る

変分法

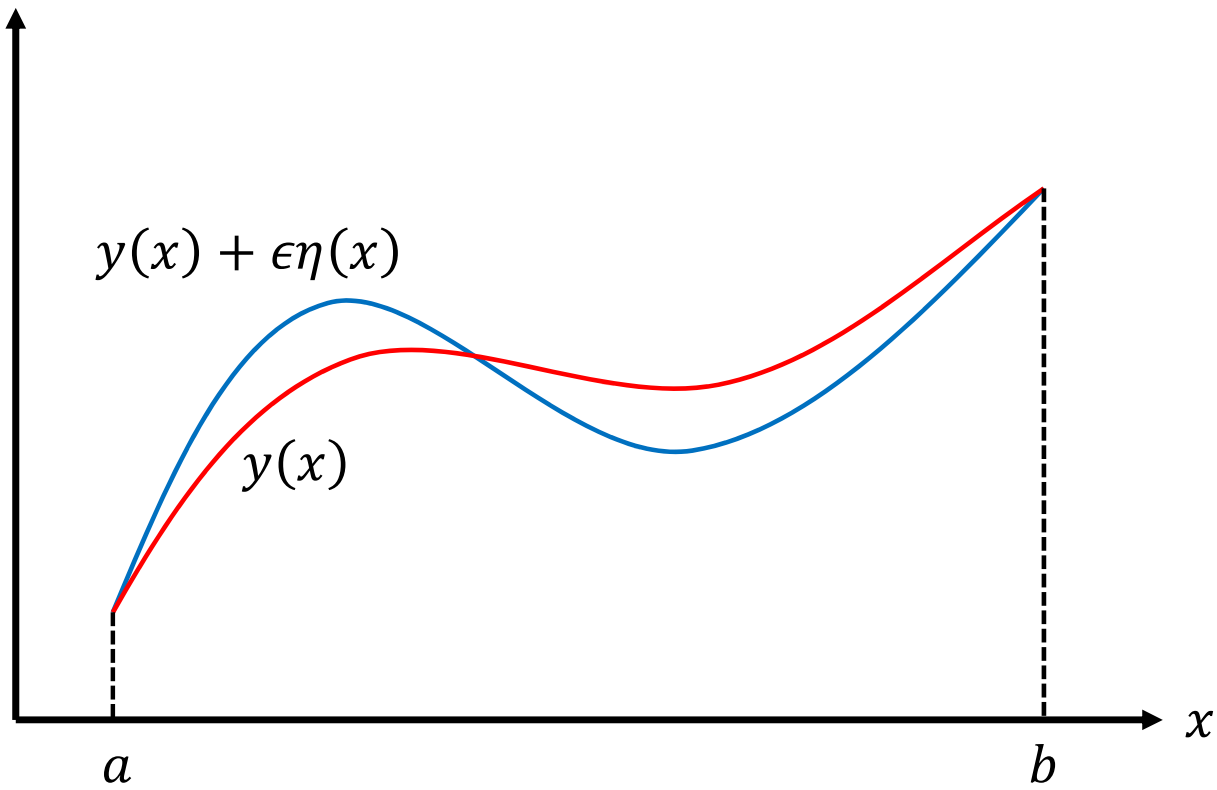

図 10: 制約条件を含んでいる場合の表現

変分法

- 汎関数の例 (2)
  - 上では $G(y(x), y'(x), x)$ について考えて、変分を導出した
  - $y(x)$ と x のみによって決まり、$y'(x)$ には依存しない関数 $G(y(x), x)$ を考えよう

  - 汎関数 $F[y]$ は、先程と同様に以下で表されるとする

$$F[y] = \int_a^b G(y(x), x)dx \tag{288}$$

  - このとき変分 $\delta F/\delta y$ を求めるのは、非常に簡単である
  - 先程の式に、$\partial G/\partial y' = 0$ を代入すれば直ちに得られる

$$\frac{\delta F}{\delta y(x)} = \frac{\partial G}{\partial y} \tag{289}$$

変分法

- あるいは以下のように書ける

$$\frac{\delta}{\delta y(x)} \int_a^b G(y(x), x) dx = \frac{\partial}{\partial y} G(y(x), x) \tag{290}$$

- $F[y]$ が最大 (最小) であるとき、以下のオイラー-ラグランジュ方程式が成り立つ

$$\frac{\partial G}{\partial y} = 0 \tag{291}$$

変分法

- 汎関数の例 (3)
  - 今度は、$y'(x)$ と x のみによって決まり、$y(x)$ には依存しない関数 $G(y'(x), x)$ を考えよう
  - この場合も変分 $\delta F/\delta y$ を求めるのは簡単である
  - $G(y(x), y'(x), x)$ の変分の式に、$\partial G/\partial y = 0$ を代入すればよい

$$\frac{\delta F}{\delta y(x)} = -\frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right) \tag{292}$$

  - オイラー-ラグランジュ方程式は次のようになる

$$-\frac{d}{dx}\left(\frac{\partial G}{\partial y'}\right) = 0 \tag{293}$$

変分法

- 汎関数の例 (4)
  - $y(x)$ と $y'(x)$ によって決まる関数 $G(y(x), y'(x))$ を考えよう
  - このときのオイラー-ラグランジュ方程式を導出してみよう

  - $G(y(x), y'(x))$ を x で微分すれば

$$\begin{aligned} & \frac{d}{dx} G(y, y') \\ = & \frac{\partial}{\partial y} G(y, y') \frac{dy}{dx} + \frac{\partial}{\partial y'} G(y, y') \frac{dy'}{dx} \quad (294) \\ = & y' \frac{\partial}{\partial y} G(y, y') + \frac{\partial}{\partial y'} G(y, y') \frac{dy'}{dx} \quad (295) \end{aligned}$$

  となるから

$$y' \frac{\partial}{\partial y} G(y, y') = \frac{d}{dx} G(y, y') - \frac{\partial}{\partial y'} G(y, y') \frac{dy'}{dx} \quad (296)$$

変分法

また、オイラー-ラグランジュ方程式の両辺に y' を掛けたものは

$$y' \frac{\partial}{\partial y} G(y, y') - y' \frac{d}{dx} \left(\frac{\partial}{\partial y'} G(y, y') \right) = 0 \tag{297}$$

これらを連立させて

$$y' \frac{d}{dx} \left(\frac{\partial}{\partial y'} G(y, y') \right) = \frac{d}{dx} G(y, y') - \frac{\partial}{\partial y'} G(y, y') \frac{dy'}{dx} \tag{298}$$

$$y' \frac{d}{dx} \left(\frac{\partial}{\partial y'} G(y, y') \right) + \frac{\partial}{\partial y'} G(y, y') \frac{dy'}{dx} = \frac{d}{dx} G(y, y') \tag{299}$$

$$\frac{d}{dx} \left(y' \cdot \frac{\partial}{\partial y'} G(y, y') \right) = \frac{d}{dx} G(y, y') \tag{300}$$

$$\int \left(\frac{d}{dx} \left(y' \cdot \frac{\partial}{\partial y'} G(y, y') \right) \right) dx = \int \left(\frac{d}{dx} G(y, y') \right) dx + C \tag{301}$$

変分法

$$G(y, y') = y' \cdot \frac{\partial}{\partial y'} G(y, y') + C \tag{302}$$

となるので、結局オイラー-ラグランジュ方程式は

$$G - y' \frac{\partial G}{\partial y'} = \text{Const.} \tag{303}$$

と書ける

目次

変分法で解ける問題の例

- 変分法で解ける問題の例 (1)
  - 2 点 $P(0,0)$、$Q(a,b)$ を結ぶ最短経路は?
  - 2 点を結ぶ経路 $y = f(x)(0 \leq x \leq a)$ の長さ l は、次のようになる

$$l = \int_0^a \sqrt{1 + \left(\frac{dy}{dx}\right)^2} dx = \int_0^a \sqrt{1 + y'^2} dx \tag{304}$$

  - 被積分項が y' のみの関数となっていることが分かる

変分法で解ける問題の例

- $G(y'(x), x)$ の場合の公式を使えば、l の $y = f(x)$ による変分が求まる

$$\frac{\delta l}{\delta f(x)} = -\frac{d}{dx}\left(\frac{\partial}{\partial y'}\sqrt{1+y'^2}\right) \tag{305}$$

$$= -\frac{d}{dx}\left(\frac{1}{2}\frac{1}{\sqrt{1+y'^2}}\frac{\partial}{\partial y'}\left(1+y'^2\right)\right) \tag{306}$$

$$= -\frac{d}{dx}\frac{y'}{\sqrt{1+y'^2}} \tag{307}$$

- l を最小化するような $y = f(x)$ は、上式の変分を 0 と等置すれば

$$-\frac{d}{dx}\frac{y'}{\sqrt{1+y'^2}} = 0 \tag{308}$$

$$\therefore \quad \frac{y'}{\sqrt{1+y'^2}} = \text{Const.} \tag{309}$$

変分法で解ける問題の例

- これは、y' が定数であることを意味している
- 従って、$y = C_0 x + C_1$ と書ける

- 以上より、2 点間を結ぶ最短経路は直線である

- $y = f(x)$ の形について、具体的な仮定は特に置いていないことに注意
- 変分法では、関数そのものを最適化する
- 従って、関数の具体的な形については、特に仮定する必要がない

目次

変分法のまとめ

- 変分のまとめ
  - これまでの計算で、次の変分が明らかとなった

$$\frac{\delta}{\delta y(x)} \int G(y(x), y'(x), x) dx =$$
$$\frac{\partial}{\partial y} G(y(x), y'(x), x) - \frac{d}{dx} \left(\frac{\partial}{\partial y'} G(y(x), y'(x), x) \right) \quad (310)$$
$$\frac{\delta}{\delta y(x)} \int G(y(x), x) dx = \frac{\partial}{\partial y} G(y(x), x) \quad (311)$$
$$\frac{\delta}{\delta y(x)} \int G(y'(x), x) dx = -\frac{d}{dx} \left(\frac{\partial}{\partial y'} G(y'(x), x) \right) \quad (312)$$

変分法のまとめ

- ここまでの話の流れ
  1. 変分の定義や、変分の計算法について調査した
  2. 変分 (汎関数の微分) とは、入力関数が微小に変化したときの、出力値の変化量として定義される
  3. 汎関数が特定の形で表せるとき、変分がどのようになるか計算した
  4. 汎関数が最大 (最小) になるとき、オイラー-ラグランジュ方程式が成立した
  5. 変分法を用いて、2 点間を結ぶ最短経路が直線になることを確認した
- これからの話の流れ
  - 変分最適化を、どのように推論問題に適用するのかについて調べていく

目次

目次

変分推論

- 変分推論が必要だった理由
  - 潜在変数に関する事後分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ の計算は、困難であることが多い
  - どのような場合に困難になるのか、次の図 11 に示す
  - $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ の厳密な計算は諦める代わりに、別の確率分布で近似したい
  - 別の確率分布で近似するとき、単純な項の積として表現できるといった、何らかの仮定を置く

変分推論

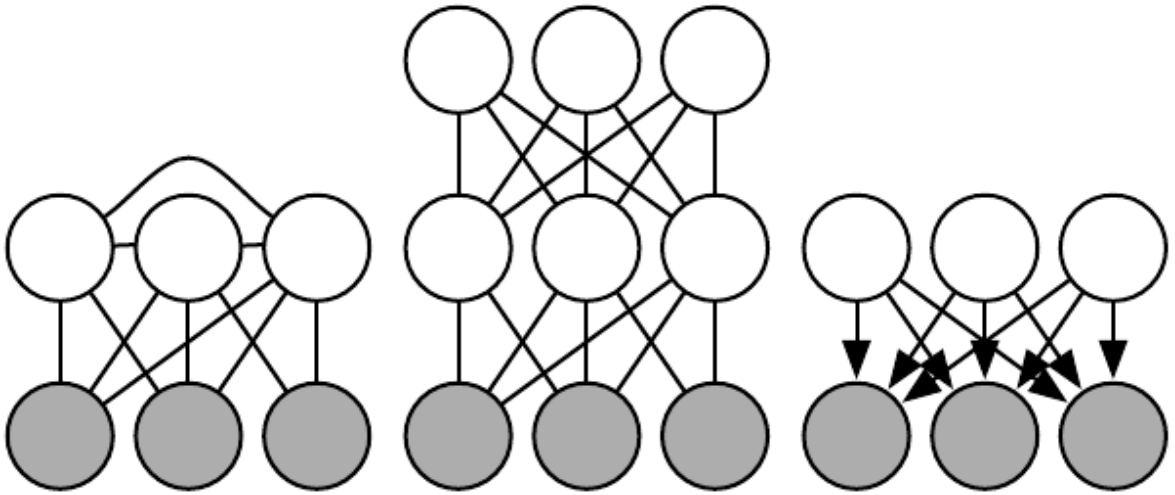

図 11: 事後分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ の計算が困難な場合

変分推論

Figure 19.1: Intractable inference problems in deep learning are usually the result of interactions between latent variables in a structured graphical model. These can be due to edges directly connecting one latent variable to another, or due to longer paths that are activated when the child of a V-structure is observed. *(Left)*A **semi-restricted Boltzmann machine** (Osindero and Hinton, 2008) with connections between hidden units. These direct connections between latent variables make the posterior distribution intractable due to large cliques of latent variables. *(Center)*A deep Boltzmann machine, organized into layers of variables without intra-layer connections, still has an intractable posterior distribution due to the connections between layers. *(Right)*This directed model has interactions between latent variables when the visible variables are observed, because every two latent variables are co-parents. Some probabilistic models are able to provide tractable inference over the latent variables despite having one of the graph structures depicted above. This is possible if the conditional probability distributions are chosen to introduce additional independences beyond those described by the graph. For example, probabilistic PCA has the graph structure shown in the right, yet still has simple inference

図 12: 事後分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ の計算が困難な場合

変分推論

- 事後分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ の計算が困難な場合 1
  - グラフィカルモデルにおいて、潜在変数間の相互作用がある場合、計算が困難になる
  - 左側の半制限付きボルツマンマシンでは、全ての潜在変数の組み合わせ間で、接続がある
  - 従って、潜在変数間に依存関係が存在し、事後分布の計算が手に負えない
- 事後分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{\theta})$ の計算が困難な場合 2
  - 中央は、層間の結合がない潜在変数の層で構成される、深層ボルツマンマシンを表す
  - 潜在変数の層間の結合があるため、事後分布の計算が手に負えない

目次

変分推論

- 変分推論の目的
  - 同時分布 $p(\boldsymbol{X}, \boldsymbol{Z})$ が分かっているときに、事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ の近似を求めたい

- 注意点
  - $\boldsymbol{X}$ は、観測データ $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ をまとめた行列 (i 行ベクトルが $\boldsymbol{x}_i^T$)
  - $\boldsymbol{X}$ の定義は、これまでと全く同様である
  - $\boldsymbol{Z}$ は、データ $\boldsymbol{X}$ とは異なり、観測することができない潜在変数を全てまとめた行列
  - $\boldsymbol{Z}$ の定義は、これまでとは異なる
  - 今まで、各データ $\boldsymbol{x}_i$ に対して存在する変数 $\boldsymbol{z}_i$ のみを潜在変数として扱ってきた
  - これ以降、確率モデルのパラメータも潜在変数に含めることにする

変分推論

- 即ち、潜在変数には、各データ $\boldsymbol{x}_i$ の裏側に潜んでいて表には現れない、データの実体 $\boldsymbol{z}_i$(今までに考えてきた潜在変数) が含まれるかもしれない
- また、各データ $\boldsymbol{x}_i$ について存在する本質的な情報 $\boldsymbol{z}_i$ だけでなく、モデル全体の構造に組み込まれた、何らかの変数 (データに結び付いているのではなく、モデルに対して存在するパラメータ) も含まれるかもしれない
- 要するに、$\boldsymbol{Z}$ は、データ $\boldsymbol{X}$ としては観測できない変数を一般に指す
- EM アルゴリズムにおける潜在変数とは、意味が異なっているので注意

- 事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ の近似
  - 事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ を解析的に導出するのは困難である
  - $p(\boldsymbol{Z}|\boldsymbol{X})$ を直接計算するのは諦めて、別の分布 $q(\boldsymbol{Z})$ で近似する
  - $q(\boldsymbol{Z})$ は、$p(\boldsymbol{Z}|\boldsymbol{X})$ よりは簡単に求められることが前提

変分推論

- 分布 $q(\boldsymbol{Z})$ は、$q(\boldsymbol{Z}|\boldsymbol{X})$ のように、データ $\boldsymbol{X}$ を明示的に書くこともある (両方を適宜使い分けているので注意)

- $q(\boldsymbol{Z})$ を、$p(\boldsymbol{Z}|\boldsymbol{X})$ にできるだけ近づけたい
- そのためには、KL ダイバージェンス $\mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}||\boldsymbol{X}))$ を最小化すればよい

- KL ダイバージェンス $\mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}||\boldsymbol{X}))$ の最小化
  - 当然であるが、KL ダイバージェンス $\mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}||\boldsymbol{X}))$ を最小化するためには、KL ダイバージェンス $\mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}||\boldsymbol{X}))$ を計算できなければならない
  - KL ダイバージェンス $\mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}||\boldsymbol{X}))$ を計算するためには、$p(\boldsymbol{Z}|\boldsymbol{X})$ が分かっている必要がある
  - しかし $p(\boldsymbol{Z}|\boldsymbol{X})$ を計算するのは困難であったため、KL ダイバージェンスは求められない

変分推論

- これでは本末転倒である

- そこで、KL ダイバージェンスの最小化を、別の問題で置き換えることにする
- 後述するように、KL ダイバージェンスの最小化は、変分下界の最大化と等価である

- これを示すために、まずは周辺分布の対数 $\ln p(\boldsymbol{X})$ を 2 つの項に分解してみよう
- $p(\boldsymbol{X})$ は、確率モデルからデータ $\boldsymbol{X}$ が生起する確率である
- データからみたモデルの好みと解釈できるから、$p(\boldsymbol{X})$ をモデルエビデンスという

変分推論

- 周辺分布の対数 $\ln p(\boldsymbol{X})$ の分解
  - EM アルゴリズムのときと同様に計算でき、次のようになる

$$\ln p(\boldsymbol{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p) \tag{313}$$

  - 但し、$\mathcal{L}(q)$ と $\mathrm{KL}(q||p)$ は次のように定義した

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{314}$$

$$\mathrm{KL}(q||p) = -\int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{315}$$

  - ここでは連続潜在変数について考えるが、離散潜在変数であれば、積分を $\boldsymbol{Z}$ に関する総和に置き換えればよい

変分推論

- 下界 $\mathcal{L}(q)$ を最適化する動機
  - $\mathcal{L}(q)$ はエビデンスの対数 $\ln p(\boldsymbol{X})$ の下界であるから、エビデンス下界 (Evidence lower bound, ELBO) ともいう
  - または、変分下界、変分下限、負の変分自由エネルギー (Variational free energy) などという

  - $\ln p(\boldsymbol{X})$ は q には依存しないため、定数項とみなせる
  - 従って、$\mathcal{L}(q)$ を q について最大化することは、$\mathrm{KL}(q||p)$ の最小化に相当
  - このとき、分布 $q(\boldsymbol{Z})$ を真の事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ に近づけられる

  - $q(\boldsymbol{Z}) = q(\boldsymbol{Z}|\boldsymbol{X}) = p(\boldsymbol{Z}|\boldsymbol{X})$ が分かれば、データ $\boldsymbol{X}$ から、潜在変数やパラメータ $\boldsymbol{Z}$ が得られる

  - これにより、KL ダイバージェンス $\mathrm{KL}(q||p)$ の最小化という問題は、変分下界 $\mathcal{L}(q)$ の最大化に読み替えてよいことが分かった

変分推論

- 下界 $\mathcal{L}(q)$ の最適化
  - EM アルゴリズムのときと同じように、下界 $\mathcal{L}(q)$ を、分布 $q(\boldsymbol{Z})$ について最大化する
  - これは、KL ダイバージェンス $\mathrm{KL}(q||p)$ を最小化することと等価である
  - 従って、もし $q(\boldsymbol{Z})$ を任意の分布にしてよければ、$q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$ とおいて、KL ダイバージェンスを 0 にすればよい
  - しかしここでは、真の事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ を求めることは不可能という状況を想定していた
- 分布 $q(\boldsymbol{Z})$ の近似
  - 計算コストを削減するために、$q(\boldsymbol{Z})$ の形をある程度制限する
  - 制限したクラスの $q(\boldsymbol{Z})$ の中で、KL ダイバージェンス $\mathrm{KL}(q||p)$ を最小化するものを探す

変分推論

- 変分推論の目的
  - 分布のクラスを制限することで、$q(\boldsymbol{Z})$ を計算可能にすること
  - 表現力が豊かな分布のクラスを使えば、真の事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ を良く近似できる

  - 計算可能な分布のクラスの中で、可能な限り豊かな表現力を持つものを選びたい
  - 表現力が豊かな分布を使うことは、真の事後分布を、精度良く近似することにつながるのであって、過学習は起こらない

変分推論

- 分布 $q(\boldsymbol{Z})$ のクラスを制限する方法
  - 例えば分布 $q(\boldsymbol{Z})$ を、パラメトリックな分布に限定することができる
  - 即ち、パラメータベクトル $\boldsymbol{\omega}$ によって $q(\boldsymbol{Z}|\boldsymbol{\omega})$ と記述されるような、分布に制限する

  - 分布 $q(\boldsymbol{Z})$ を、ガウス分布などの、何らかの特別なパラメトリックな分布と仮定することに相当

変分推論

- クラスを制限する別の方法 (平均場近似)
  - 分布 $q(\boldsymbol{Z})$ のクラスを制限する別の方法として、平均場近似がある
  - 潜在変数 $\boldsymbol{Z}$ を、M 個の互いに排反なグループ $\{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_M\}$ に分割
  - 分布 $q(\boldsymbol{Z})$ が、これらのグループによって分解されると仮定

$$q(\boldsymbol{Z}) = \prod_{i=1}^{M} q_i(\boldsymbol{Z}_i) \tag{316}$$

  - 分布 q について、これ以上の仮定はしない
  - 従って、各因子 $q_i(\boldsymbol{Z}_i)$ の関数形については、何の制限も課さない
  - 平均場近似とは、元々は物理学における用語である

変分推論

- 下界 $\mathcal{L}(q)$ の最大化
  - $q(\boldsymbol{Z}) = \prod_i q_i(\boldsymbol{Z}_i)$ と分解できるような分布 $q(\boldsymbol{Z})$ の中で、下界 $\mathcal{L}(q)$ を最大にするものを探す
  - $\mathcal{L}(q)$ を $q(\boldsymbol{Z})$ について最大化するために、$\mathcal{L}(q)$ を各因子 $q_i(\boldsymbol{Z}_i)$ について順番に最大化していくことが考えられる
  - $\mathcal{L}(q)$ に $q(\boldsymbol{Z}) = \prod_i q_i(\boldsymbol{Z}_i)$ を代入して、因子の一つ $q_j(\boldsymbol{Z}_j)$ に関する依存項を抜き出してみよう

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{317}$$

$$= \int \left(\prod_i q_i(\boldsymbol{Z}_i) \right) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{318}$$

$$= \int \prod_i q_i(\boldsymbol{Z}_i) \left(\ln p(\boldsymbol{X}, \boldsymbol{Z}) - \sum_i \ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z} \tag{319}$$

$$= \int \prod_i q_i(\boldsymbol{Z}_i) \left(\ln p(\boldsymbol{X}, \boldsymbol{Z})\right) d\boldsymbol{Z} - \int \prod_i q_i(\boldsymbol{Z}_i) \left(\sum_i \ln q_i(\boldsymbol{Z}_i)\right) d\boldsymbol{Z} \tag{320}$$

ここで第 1 項は

$$\int \prod_i q_i(\boldsymbol{Z}_i) \left(\ln p(\boldsymbol{X}, \boldsymbol{Z})\right) d\boldsymbol{Z}$$

$$= \int q_j(\boldsymbol{Z}_j) \left(\prod_{i \neq j} q_i(\boldsymbol{Z}_i)\right) \left(\ln p(\boldsymbol{X}, \boldsymbol{Z})\right) d\boldsymbol{Z} \tag{321}$$

$d\boldsymbol{Z} = d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M$ であるから

$$= \int q_j(\boldsymbol{Z}_j) \left(\prod_{i \neq j} q_i(\boldsymbol{Z}_i)\right) \left(\ln p(\boldsymbol{X}, \boldsymbol{Z})\right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M \tag{322}$$

$$= \int q_j(\boldsymbol{Z}_j) \left(\ln p(\boldsymbol{X}, \boldsymbol{Z})\right) \left(\prod_{i \neq j} q_i(\boldsymbol{Z}_i)\right) \left(\prod_{i \neq j} d\boldsymbol{Z}_i\right) d\boldsymbol{Z}_j \quad (323)$$

$$= \int q_j(\boldsymbol{Z}_j) \left(\int \left(\ln p(\boldsymbol{X}, \boldsymbol{Z})\right) \prod_{i \neq j} q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i\right) d\boldsymbol{Z}_j \quad (324)$$

$$= \int q_j(\boldsymbol{Z}_j) \ln \widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) d\boldsymbol{Z}_j \quad (325)$$

- 但し、新しい分布 $\widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)$ は以下の式で定義した (積分の結果であるため、定数項が出現する)

$$\ln \widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{Const.} \quad (326)$$

変分推論

- 記法 $\mathbb{E}_{i \neq j}$ は、$i \neq j$ をみたす全ての分布 $q_i(\boldsymbol{Z}_i)$ による、期待値を取ることを表す

$$\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] = \int (\ln p(\boldsymbol{X}, \boldsymbol{Z})) \prod_{i \neq j} q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i \tag{327}$$

- 第 2 項は

$$\begin{aligned}
& \int \prod_i q_i(\boldsymbol{Z}_i) \left(\sum_i \ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z} \\
= & \sum_i \int \prod_i q_i(\boldsymbol{Z}_i) \left(\ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z} && (328) \\
= & \sum_i \int \prod_k q_k(\boldsymbol{Z}_k) \left(\ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M && (329) \\
= & \int \prod_k q_k(\boldsymbol{Z}_k) \left(\ln q_j(\boldsymbol{Z}_j) \right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M + &
\end{aligned}$$

$$\sum_{i \neq j} \int \prod_k q_k(\boldsymbol{Z}_k) \left(\ln q_i(\boldsymbol{Z}_i)\right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M \quad (330)$$

但し

$$\int \prod_k q_k(\boldsymbol{Z}_k) \left(\ln q_j(\boldsymbol{Z}_j)\right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M \quad (331)$$

$$= \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) \left(\prod_{k \neq j} q_k(\boldsymbol{Z}_k) \right) \left(\prod_{k \neq j} d\boldsymbol{Z}_k \right) d\boldsymbol{Z}_j \quad (332)$$

$$= \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) \left(\int \prod_{k \neq j} q_k(\boldsymbol{Z}_k) d\boldsymbol{Z}_k \right) d\boldsymbol{Z}_j \quad (333)$$

$$= \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) \prod_{k \neq j} \underbrace{\left(\int q_k(\boldsymbol{Z}_k) d\boldsymbol{Z}_k \right)}_{=1} d\boldsymbol{Z}_j \quad (334)$$

$$= \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j \tag{335}$$

であるほか

$$\sum_{i \neq j} \int \prod_k q_k(\boldsymbol{Z}_k) \left(\ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M$$

$$= \sum_{i \neq j} \int q_i(\boldsymbol{Z}_i) q_j(\boldsymbol{Z}_j) \left(\prod_{k \neq i,j} \ln q_k(\boldsymbol{Z}_k) \right)$$

$$\left(\ln q_i(\boldsymbol{Z}_i) \right) \left(\prod_{k \neq i,j} d\boldsymbol{Z}_k \right) d\boldsymbol{Z}_i d\boldsymbol{Z}_j \tag{336}$$

$$= \sum_{i \neq j} \underbrace{\left(\int q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j \right)}_{=1} \left(\int \prod_{k \neq i,j} \ln q_k(\boldsymbol{Z}_k) d\boldsymbol{Z}_k \right)$$

変分推論

$$\int \ln q_i(\boldsymbol{Z}_i) q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i \tag{337}$$

$$= \sum_{i \neq j} \prod_{k \neq i,j} \underbrace{\left(\int \ln q_k(\boldsymbol{Z}_k) d\boldsymbol{Z}_k \right)}_{=\text{Const.}} \underbrace{\int \ln q_i(\boldsymbol{Z}_i) q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i}_{=\text{Const.}} \tag{338}$$

$$= \sum_{i \neq j} \text{Const.} = \text{Const.} \tag{339}$$

となるから結局

$$\int \prod_i q_i(\boldsymbol{Z}_i) \left(\sum_i \ln q_i(\boldsymbol{Z}_i) \right) d\boldsymbol{Z}$$

$$= \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j + \text{Const.} \tag{340}$$

変分推論

- 従って、下界 $\mathcal{L}(q)$ から $q_j(\boldsymbol{Z}_j)$ に依存する項を取り出すと

$$\mathcal{L}(q) = \int q_j(\boldsymbol{Z}_j) \ln \widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) d\boldsymbol{Z}_j - \int q_j(\boldsymbol{Z}_j) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j + \text{Const.} \tag{341}$$

但し

$$\ln \widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] = \int \ln p(\boldsymbol{X}, \boldsymbol{Z}) \prod_{i \neq j} q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i \tag{342}$$

- $\mathcal{L}(q)$ を、$i \neq j$ である全ての $q_i(\boldsymbol{Z}_i)$ について固定した上で、$q_j(\boldsymbol{Z}_j)$ について最大化することになる
- $q_j(\boldsymbol{Z}_j)$ について可能な全ての分布の中で、$\mathcal{L}(q)$ を最大にするようなものを選ぶ

変分推論

- $\mathcal{L}(q)$ は次のように変形できる

$$\mathcal{L}(q) = \int q_j(\boldsymbol{Z}_j) \ln \frac{\widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)}{q_j(\boldsymbol{Z}_j)} d\boldsymbol{Z}_j + \text{Const.} \quad (343)$$

$$= -\text{KL}\left(q_j(\boldsymbol{Z}_j) || \widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)\right) + \text{Const.} \quad (344)$$

- これより、$\mathcal{L}(q)$ は $q_j(\boldsymbol{Z}_j)$ と $\widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)$ の間の、負の KL ダイバージェンスとなっている
- そして、$\mathcal{L}(q)$ の $q_j(\boldsymbol{Z}_j)$ に関する最大化は、KL ダイバージェンスの最小化と等価

- KL ダイバージェンスを最小にするためには、$q_j(\boldsymbol{Z}_j) = \widetilde{p}(\boldsymbol{X}, \boldsymbol{Z}_j)$ とすればよい

- 従って、$q_j(\boldsymbol{Z}_j)$ の最適解は、次のように書ける

$$\ln q_j^*(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{Const.} \quad (345)$$

変分推論

- 下界 $\mathcal{L}(q)$ を最大化する $\ln q_j(\boldsymbol{Z}_j)$ の解
  - $q_j(\boldsymbol{Z}_j)$ の最適解は次のように書けた

$$\ln q_j^*(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{Const.} \tag{346}$$

  - 上式は次のことを意味している
  - 因子 $q_j(\boldsymbol{Z}_j)$ の最適解の対数 $\ln q_j^*(\boldsymbol{Z}_j)$ は、観測データ $\boldsymbol{X}$ と潜在変数 $\boldsymbol{Z}$ の同時分布の対数 $\ln p(\boldsymbol{X}, \boldsymbol{Z})$ を考え、$i \neq j$ である他の因子 $q_i(\boldsymbol{Z}_i)$ について期待値を取ったものである

  - 定数項は、正規化することで得られるので、結局次のようになる

$$q_j^*(\boldsymbol{Z}_j) = \frac{\exp(\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})])}{\int \exp(\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})]) d\boldsymbol{Z}_j} \tag{347}$$

  - 正規化定数は必要に応じて計算すればよいので、取り敢えず無視できる

変分推論

- 最適解の式は、$q(\boldsymbol{Z})$ の分解の数だけ得られるので、$\{q_i(\boldsymbol{Z}_i)\}$ に関する M 本の連立方程式となる
- この方程式は、分布 $q(\boldsymbol{Z})$ が M 個の因子に分解されるという仮定の下で、下界 $\mathcal{L}(q)$ の最大値が満たすべき条件である

- $\ln q_j^*(\boldsymbol{Z}_j)$ の右辺は、$i \neq j$ である $q_i(\boldsymbol{Z}_i)$ の期待値に依存するため、$q_j^*(\boldsymbol{Z}_j)$ を陽に求めることができない
- そこで、下界 $\mathcal{L}(q)$ は次のように最適化される (重要)

- $i \neq j$ である全ての $q_i(\boldsymbol{Z}_i)$ を固定した状態で、$q_j(\boldsymbol{Z}_j)$ を最適化することを、全ての $j = 1, \ldots, M$ について繰り返す手続きを、座標降下法という

変分推論

- 下界 $\mathcal{L}(q)$ の最適化 (座標降下法)
  1. 全ての因子 $q_j(\boldsymbol{Z}_j)$ を適当に初期化する
  2. 各因子を、以下の式を使って更新する

$$\ln q_j(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{Const.} \tag{348}$$

$$\mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] = \int \ln p(\boldsymbol{X}, \boldsymbol{Z}) \prod_{i \neq j} q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}_i \tag{349}$$

     即ち、因子 $q_j(\boldsymbol{Z}_j)$ を、他の全ての因子の現在の値 $q_i(\boldsymbol{Z}_i)$ を使って改良する

  3. (2) を、下界 $\mathcal{L}(q)$ が収束するまで繰り返す

変分推論

- ここまでの話の流れ
  1. $\ln p(\boldsymbol{X}) = \mathcal{L}(q) + \mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X}))$ であるから、エビデンス下界 $\mathcal{L}(q)$ を q について最大化すれば、$\mathrm{KL}(q||p) = 0$ とでき、従って $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$ を得る
  2. $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$ が分かれば、データ $\boldsymbol{X}$ から、潜在変数やパラメータ $\boldsymbol{Z}$ が得られる (パラメータは潜在変数 $\boldsymbol{Z}$ に含まれている)
  3. しかし、事後分布 $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$ は計算不可能なので、何らかの方法で近似するしかない
  4. 近似するといっても、計算可能でなければならないので、$q(\boldsymbol{Z})$ の形には、通常何らかの制限を課す
  5. $q(\boldsymbol{Z})$ を、パラメトリックな分布 $q(\boldsymbol{Z}|\boldsymbol{\omega})$ と仮定することがある

変分推論

6. または、$q(\boldsymbol{Z})$ を、$\prod_i q_i(\boldsymbol{Z}_i)$ のように分解できるとする (平均場近似)

7. 平均場近似を行うとき、各因子 $q_j(\boldsymbol{Z}_j)$ の最適解は、
$\ln q_j^*(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{Const.}$ であった

8. 全ての因子 $\{q_j(\boldsymbol{Z}_j)\}$ を同時に最適化することはできない

9. 下界 $\mathcal{L}(q)$ を、各因子 $q_j(\boldsymbol{Z}_j)$ について順番に最適化することはできる

- これからの話の流れ
  - MAP 推定と最尤推定は、変分推論の特殊な場合であることを確認する

目次

変分推論

- MAP 推定および最尤推定の変分推論からの導出
  - 変分推論で得たいのは、潜在変数に関する事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ である
  - この変分推論の特殊ケースが、MAP 推定や最尤推定であることを導く

  - 変分推論では、エビデンス下界 $\mathcal{L}(q)$ を q について最大化し、従って $\mathrm{KL}(q||p)$ を最小化する
  - いま、分布 $q(\boldsymbol{Z})$ が、次のデルタ関数であるとする
  - 特定の値 $\boldsymbol{Z} = \widehat{\boldsymbol{Z}}$ についてのみ確率が非零になる、無限に鋭い分布

$$q(\boldsymbol{Z}) = \delta(\boldsymbol{Z} - \widehat{\boldsymbol{Z}}) \tag{350}$$

変分推論

- このとき、KL ダイバージェンス $\mathrm{KL}(q||p)$ は次のようになる

$$
\begin{aligned}
& \mathrm{KL}(q||p) \\
\equiv\ & \mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X})) \\
=\ & -\int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \\
=\ & -\int q(\boldsymbol{Z}) \ln p(\boldsymbol{Z}|\boldsymbol{X}) d\boldsymbol{Z} + \int q(\boldsymbol{Z}) \ln q(\boldsymbol{Z}) d\boldsymbol{Z} \qquad (351) \\
=\ & -\int \delta(\boldsymbol{Z} - \widehat{\boldsymbol{Z}}) \ln p(\boldsymbol{Z}|\boldsymbol{X}) d\boldsymbol{Z} + \\
& \qquad \int \delta(\boldsymbol{Z} - \widehat{\boldsymbol{Z}}) \ln \delta(\boldsymbol{Z} - \widehat{\boldsymbol{Z}}) d\boldsymbol{Z} \qquad (352) \\
=\ & -\ln p(\widehat{\boldsymbol{Z}}|\boldsymbol{X}) + \ln \delta(\widehat{\boldsymbol{Z}} - \widehat{\boldsymbol{Z}}) \qquad (353) \\
=\ & -\ln p(\widehat{\boldsymbol{Z}}|\boldsymbol{X}) + \mathrm{Const.} \qquad (354)
\end{aligned}
$$

変分推論

- これを $q(\boldsymbol{Z})$ について最小化することは、$\widehat{\boldsymbol{Z}}$ について最小化することに相当する
- よって $\widehat{\boldsymbol{Z}}$ の最適解 $\widehat{\boldsymbol{Z}}^*$ は

$$\begin{aligned}
\widehat{\boldsymbol{Z}}^* &= \arg\min_{\widehat{\boldsymbol{Z}}} \Big(-\ln p(\widehat{\boldsymbol{Z}}|\boldsymbol{X}) \Big) && (355)\\
&= \arg\max_{\widehat{\boldsymbol{Z}}} \ln p(\widehat{\boldsymbol{Z}}|\boldsymbol{X}) && (356)\\
&= \arg\max_{\widehat{\boldsymbol{Z}}} p(\widehat{\boldsymbol{Z}}|\boldsymbol{X}) && (357)\\
&= \arg\max_{\widehat{\boldsymbol{Z}}} p(\boldsymbol{X}|\widehat{\boldsymbol{Z}}) p(\widehat{\boldsymbol{Z}}) && (358)
\end{aligned}$$

  となるから、MAP 推定 (最大事後確率推定) の式と一致する

- これより、MAP 推定は、KL ダイバージェンス $\mathrm{KL}(q||p)$ の最小化、従って、エビデンス下界 $\mathcal{L}(q)$ の最大化と等価である

変分推論

- 更に、$p(\widehat{\boldsymbol{Z}}) = \text{Const.}$、即ち $\widehat{\boldsymbol{Z}}$ に関する事前の情報がないとすると

$$\begin{aligned} \widehat{\boldsymbol{Z}}^* &= \arg\max_{\widehat{\boldsymbol{z}}} p(\boldsymbol{X}|\widehat{\boldsymbol{Z}})p(\widehat{\boldsymbol{Z}}) \\ &= \arg\max_{\widehat{\boldsymbol{z}}} p(\boldsymbol{X}|\widehat{\boldsymbol{Z}}) \end{aligned} \tag{359}$$

となるから、これは最尤推定の式に一致する

変分推論

- 下界 $\mathcal{L}(q)$ の式から導くことも可能である

$$
\begin{aligned}
\mathcal{L}(q) &= \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \\
&= \int q(\boldsymbol{Z}) \ln p(\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{Z} - \int q(\boldsymbol{Z}) \ln q(\boldsymbol{Z}) d\boldsymbol{Z} \quad (360) \\
&= \int \delta(\boldsymbol{Z} - \widehat{\boldsymbol{Z}}) \ln p(\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{Z} - \\
&\qquad \int \delta(\boldsymbol{Z} - \widehat{\boldsymbol{Z}}) \ln \delta(\boldsymbol{Z} - \widehat{\boldsymbol{Z}}) d\boldsymbol{Z} \quad (361) \\
&= \ln p(\boldsymbol{X}, \widehat{\boldsymbol{Z}}) - \ln \delta(\widehat{\boldsymbol{Z}} - \widehat{\boldsymbol{Z}}) \quad (362) \\
&= \ln p(\boldsymbol{X}, \widehat{\boldsymbol{Z}}) + \text{Const.} \quad (363) \\
&= \ln p(\boldsymbol{X}|\widehat{\boldsymbol{Z}}) p(\widehat{\boldsymbol{Z}}) + \text{Const.} \quad (364)
\end{aligned}
$$

- これより、下界 $\mathcal{L}(q)$ の最大化は、$\widehat{\boldsymbol{Z}}$ に関する $p(\widehat{\boldsymbol{Z}}|\boldsymbol{X}) \propto p(\boldsymbol{X}|\widehat{\boldsymbol{Z}}) p(\widehat{\boldsymbol{Z}})$ の最大化、即ち MAP 推定と等価である

MAP 推定の例

- スパース符号化
  - ここでは MAP 推定の例として、スパース符号化を扱う (変分推論の例ではない)
  - データ $\boldsymbol{x}$ に対する潜在変数 $\boldsymbol{z}$ に、スパース性を持たせる
  - そのために、スパース性を導く事前分布 (ラプラス事前分布) を、潜在変数に用いる
  - データ $\boldsymbol{x}$ を D 次元、また潜在変数 $\boldsymbol{z}$ を K 次元とする
  - データ $\boldsymbol{x}$ に対応する潜在変数を $\boldsymbol{z}_i$ と表し、$\boldsymbol{z}_i$ の第 k 成分を z_{ik} とする

$$p(z_{ik}|\lambda) = \frac{\lambda}{2}\exp\left(-\lambda|z_{ik}|\right) \tag{365}$$

$$p(\boldsymbol{z}_i|\lambda) = \prod_k p(z_{ik}) = \prod_k \frac{\lambda}{2}\exp\left(-\lambda|z_{ik}|\right) \tag{366}$$

MAP 推定の例

- データに関する事後分布 $p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{W}, \boldsymbol{b}, \beta)$ を、次で定義する

$$
\begin{aligned}
& p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{W}, \boldsymbol{b}, \beta) \\
= \quad & \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{b}, \beta^{-1}\boldsymbol{I}) && (367) \\
= \quad & \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\beta^{-1}\boldsymbol{I}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - (\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{b}))^T \right. \\
& \left. \left(\beta^{-1}\boldsymbol{I}\right)^{-1}(\boldsymbol{x}_i - (\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{b}))\right) && (368) \\
= \quad & \frac{1}{(2\pi\beta^{-1})^{\frac{D}{2}}} \exp\left(-\frac{\beta}{2}(\boldsymbol{x}_i - (\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{b}))^T \right. \\
& \left. (\boldsymbol{x}_i - (\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{b}))\right) && (369)
\end{aligned}
$$

- データ $\boldsymbol{x}_i$ は、対応する潜在変数 $\boldsymbol{z}_i$ に、線形変換 $\boldsymbol{W}\boldsymbol{z}_i + \boldsymbol{b}$ を施し、更に分散 $\beta^{-1}\boldsymbol{I}$ のガウスノイズを足すことで、生成されると考える

MAP 推定の例

- パラメータについての補足
  - 今回は、$\boldsymbol{b} = 0$ とおいてバイアスを無視したものを考える
  - λ と、精度 β はハイパーパラメータであり、予め決められているとする
  - そこでこれ以降、次のように確率分布を表現する

$$p(\boldsymbol{z}_i|\lambda) = p(\boldsymbol{z}_i) \tag{370}$$

$$p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{W}, \boldsymbol{b}, \beta) = p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{W}) \tag{371}$$

  - $\boldsymbol{X}$ を、全てのデータ $\{\boldsymbol{x}_i\}$ を集めた行列とする (第 i 行ベクトルが $\boldsymbol{x}_i^T$)
  - $\boldsymbol{Z}$ についても同様に、全ての潜在変数 $\{\boldsymbol{z}_i\}$ を集めた行列として定める (第 i 行ベクトルが $\boldsymbol{z}_i^T$)

- スパース符号化の学習
  - 事後分布 $p(\boldsymbol{z}_i|\boldsymbol{x}_i)$ は、表現することすら困難であるため、最尤推定による手法 (EM アルゴリズムなど) は利用できない

MAP 推定の例

- そこで、最尤推定の代わりに MAP 推定を利用することで、最適なパラメータが得られる
- 最大化するのは、以下の事後分布 $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{W})$ である

$$
\begin{aligned}
p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{W}) &= \frac{p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{W})}{p(\boldsymbol{X})} && (372)\\
&\propto p(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{W}) && (373)\\
&= p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{W})p(\boldsymbol{Z}) && (374)\\
&= \left(\prod_i p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{W})\right)\left(\prod_i p(\boldsymbol{z}_i)\right) && (375)\\
&= \prod_i p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{W})p(\boldsymbol{z}_i) && (376)
\end{aligned}
$$

MAP 推定の例

- 対数を取って最大化してもよいので、最大化する量は結局

$$\ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{W}) = \ln\left(\prod_i p(\boldsymbol{x}_i|\boldsymbol{z}_i,\boldsymbol{W})p(\boldsymbol{z}_i)\right) \quad (377)$$

$$= \sum_i \left(\ln p(\boldsymbol{x}_i|\boldsymbol{z}_i,\boldsymbol{W}) + \ln p(\boldsymbol{z}_i)\right) \quad (378)$$

$$= \sum_i \ln p(\boldsymbol{x}_i|\boldsymbol{z}_i,\boldsymbol{W}) + \sum_i \ln p(\boldsymbol{z}_i) \quad (379)$$

各項を求めると

$$\sum_i \ln p(\boldsymbol{x}_i|\boldsymbol{z}_i,\boldsymbol{W})$$

$$= \sum_i \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i,\beta^{-1}\boldsymbol{I})$$

$$= \sum_i \ln\left(\frac{1}{(2\pi\beta^{-1})^{\frac{D}{2}}}\exp\left(-\frac{\beta}{2}(\boldsymbol{x}_i-\boldsymbol{W}\boldsymbol{z}_i)^T(\boldsymbol{x}_i-\boldsymbol{W}\boldsymbol{z}_i)\right)\right)$$

MAP 推定の例

$$
\begin{aligned}
&= \sum_i \left(-\frac{D}{2} \ln 2\pi - \frac{D}{2} \ln \beta^{-1} - \right. \\
&\qquad \left. \frac{\beta}{2} \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i \right)^T \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i \right) \right) \\
&= \sum_i \left(-\frac{\beta}{2} \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i \right)^T \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i \right) \right) + \text{Const.} \\
&= -\frac{\beta}{2} \sum_i \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i \right)^T \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i \right) + \text{Const.} \\
&= -\frac{\beta}{2} \sum_i \left(\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T \right)_{i:} \left(\left(\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T \right)_{i:} \right)^T + \text{Const.} \qquad (380) \\
&= -\frac{\beta}{2} \sum_{i,j} \left(\left(\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T \right) \odot \left(\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T \right) \right)_{i,j} + \text{Const.} \qquad (381) \\
&= -\frac{\beta}{2} \sum_{i,j} \left(\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T \right)^2_{i,j} + \text{Const.} \qquad (382)
\end{aligned}
$$

MAP 推定の例

$$= -\frac{\beta}{2}||\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T||_2^2 + \text{Const.} \tag{383}$$

また

$$\begin{aligned}
& \sum_i \ln p(\boldsymbol{z}_i) \\
= & \sum_i \ln \prod_j \left(\frac{\lambda}{2} \exp\left(-\lambda |z_{ij}|\right) \right) \\
= & \sum_i \sum_j \left(\ln \frac{\lambda}{2} - \lambda |z_{ij}| \right) \\
= & -\lambda \sum_{i,j} |z_{ij}| + \text{Const.} \qquad (384) \\
= & -\lambda \sum_{i,j} |\boldsymbol{Z}_{i,j}| + \text{Const.} \qquad (385) \\
= & -\lambda ||\boldsymbol{Z}||_1 + \text{Const.} \qquad (386)
\end{aligned}$$

MAP 推定の例

のようになる

- $\odot$ はアダマール積、$\boldsymbol{A}_{i:}$ は行列 $\boldsymbol{A}$ の第 i 行目のベクトルを表す
- また、行列のノルム $||\boldsymbol{A}||_p$ は次で定義される

$$||\boldsymbol{A}||_p = \left(\sum_{i,j} |A_{i,j}|^p \right)^{\frac{1}{p}} \tag{387}$$

$p = 1, p = 2$ のときは次のようになる

$$||\boldsymbol{A}||_1 = \sum_{i,j} |A_{i,j}|, \quad ||\boldsymbol{A}||_2 = \sqrt{\sum_{i,j} A_{i,j}^2} \tag{388}$$

- これより、$||\boldsymbol{A}||_1$ は、行列 $\boldsymbol{A}$ の各要素の絶対値の和を表す
- また $||\boldsymbol{A}||_2$ は、行列 $\boldsymbol{A}$ の各要素の二乗和の平方根を表し、フロベニウスノルム $||\boldsymbol{A}||_F$ ともよばれる

MAP 推定の例

- 従って、$||\boldsymbol{A}||_2^2$ は、行列 $\boldsymbol{A}$ の各要素の二乗和である
- 上記から

$$\begin{aligned} & \ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{W}) \\ = & \sum_i \ln p(\boldsymbol{x}_i|\boldsymbol{z}_i, \boldsymbol{W}) + \sum_i \ln p(\boldsymbol{z}_i) \\ = & -\frac{\beta}{2}||\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T||_2^2 - \lambda||\boldsymbol{Z}||_1 + \text{Const.} \end{aligned} \tag{389}$$

となるので、$\ln p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{W})$ の最大化は、以下の関数 $J(\boldsymbol{Z}, \boldsymbol{W})$ の最小化である

$$J(\boldsymbol{Z}, \boldsymbol{W}) = ||\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T||_2^2 + ||\boldsymbol{Z}||_1 \tag{390}$$

- スパース符号化の学習方法のまとめ
  - 関数 $J(\boldsymbol{Z}, \boldsymbol{W})$ を、$\boldsymbol{Z}$ と $\boldsymbol{W}$ について交互に最小化すればよい

MAP 推定の例

- $p(\boldsymbol{Z}|\boldsymbol{X}, \boldsymbol{W})$ の最大化は、$p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{W})p(\boldsymbol{Z})$ の最大化、即ち $J(\boldsymbol{Z}, \boldsymbol{W})$ の最小化と等価であった
- 従って、$J(\boldsymbol{Z}, \boldsymbol{W})$ の各パラメータ $\boldsymbol{Z}, \boldsymbol{W}$ についての最小化は、事後確率を大きくする方向に働く

- 関数 $J(\boldsymbol{Z}, \boldsymbol{W})$ をもう一度みてみよう

$$J(\boldsymbol{Z}, \boldsymbol{W}) = ||\boldsymbol{X} - \boldsymbol{Z}\boldsymbol{W}^T||_2^2 + ||\boldsymbol{Z}||_1 \tag{391}$$

- 第 1 項は、明らかに再構成誤差を表している
- 第 2 項は、データ $\boldsymbol{X}$ の表現 $\boldsymbol{Z}$ がスパースになるように付加した、正則化項である

- $\boldsymbol{Z}\boldsymbol{W}^T$ は、データ $\boldsymbol{X}$ の内部表現 $\boldsymbol{Z}$ に、重み $\boldsymbol{W}$ を掛けることで、データ $\boldsymbol{X}$ を復元したもの

目次

変分推論

- これまでの話の流れ
  1. 変分推論から、MAP 推定と最尤推定が導出できることを確認した
  2. MAP 推定は、$q(\boldsymbol{Z})$ を無限に鋭い確率分布 (デルタ関数) として、$\boldsymbol{Z}$ が特定の値しか取らないと仮定した場合であった
  3. 最尤推定は、MAP 推定において、$\boldsymbol{Z}$ の事前確率分布を設けない場合であった
  4. MAP 推定の例として、スパース符号化を扱った
- これからの話の流れ
  - 話題を変えて、分解 $q(\boldsymbol{Z}) = \prod_i q_i(\boldsymbol{Z}_i)$ によって $p(\boldsymbol{Z}|\boldsymbol{X})$ を近似するときの、弊害を調べる

変分推論

- $q(\boldsymbol{Z})$ の分解による近似の性質
  - 変分推論では、真の事後分布 $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$ を、分解により近似する
  - 分解で近似することによって、どのような不正確さが生じるのか?

- ガウス分布の分解による近似
  - ガウス分布を、分解されたガウス分布で近似することを考えてみよう
  - 分解による近似で、どのような問題が起こるのか考えてみよう

  - 2 つの変数 $\boldsymbol{z} = (z_1, z_2)$ 間には、相関があるとする
  - $\boldsymbol{z}$ はガウス分布 $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ に従っているとする

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \qquad \boldsymbol{\Lambda} = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix} \tag{392}$$

  - 精度行列 $\boldsymbol{\Lambda}$ は対称行列であるから、$\Lambda_{12} = \Lambda_{21}$

変分推論

- この分布 $p(\boldsymbol{z})$ を、分解されたガウス分布 $q(\boldsymbol{z}) = q_1(z_1)q_2(z_2)$ で近似する
- 各因子 $q_1(z_1), q_2(z_2)$ の関数形については何の仮定も置いていないことに注意

- 最適な因子 $q_1(z_1), q_2(z_2)$ の計算
  - 最適な因子 $q_1^*(z_1)$ を、先程の結果を使って求める

$$\ln q_j(\boldsymbol{Z}_j) = \mathbb{E}_{i \neq j}[\ln p(\boldsymbol{X}, \boldsymbol{Z})] + \text{Const.} \tag{393}$$

  - 従って、$q_1^*(z_1)$ を計算する式は次のようになる

$$\ln q_1^*(z_1) = \mathbb{E}_{z_2}[\ln p(\boldsymbol{z})] + \text{Const.} \tag{394}$$

$$\mathbb{E}_{z_2}[\ln p(\boldsymbol{z})] = \int \ln p(\boldsymbol{z}) q_2(z_2) dz_2 \tag{395}$$

  - 上式の右辺では、z_1 に依存する項だけを考えればよい

変分推論

- z_1 の関数を求めようとしているため
- z_1 に依存しない項は、全て定数項 (正規化定数) に含まれてしまうため

- 従って $q_1^*(z_1)$ は

$$\begin{aligned} \ln q_1^*(z_1) &= \mathbb{E}_{z_2}[\ln p(\boldsymbol{z})] + \text{Const.} && (396) \\ &= \mathbb{E}_{z_2}\left[\ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})\right] + \text{Const.} && (397) \end{aligned}$$

但し

$$\begin{aligned} & \ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \\ = & \ln \left(\frac{1}{(2\pi)^{\frac{2}{2}}} \frac{1}{|\boldsymbol{\Lambda}^{-1}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \left(\boldsymbol{\Lambda}^{-1}\right)^{-1} (\boldsymbol{z}-\boldsymbol{\mu}) \right) \right) \\ = & \ln \left(\frac{1}{2\pi} \frac{1}{|\boldsymbol{\Lambda}|^{-\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Lambda} (\boldsymbol{z}-\boldsymbol{\mu}) \right) \right) \\ = & -\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Lambda} (\boldsymbol{z}-\boldsymbol{\mu}) + \text{Const.} \qquad (398) \end{aligned}$$

z_1 に依存する項だけを取り出せば

$$
\begin{aligned}
& -\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu}) \\
= \quad & -\frac{1}{2}\left[z_1-\mu_1, z_2-\mu_2\right]\begin{bmatrix}\Lambda_{11} & \Lambda_{12}\\ \Lambda_{21} & \Lambda_{22}\end{bmatrix}\begin{bmatrix} z_1-\mu_1 \\ z_2-\mu_2\end{bmatrix} \\
= \quad & -\frac{1}{2}\Big[\Big\{(\Lambda_{11}(z_1-\mu_1)+\Lambda_{21}(z_2-\mu_2)\Big\}(z_1-\mu_1)+ \\
& \qquad \Big\{\Lambda_{12}(z_1-\mu_1)+\Lambda_{22}(z_2-\mu_2)\Big\}(z_2-\mu_2)\Big] \\
= \quad & -\frac{1}{2}\left(\Lambda_{11}(z_1-\mu_1)^2+2\Lambda_{12}(z_1-\mu_1)(z_2-\mu_2)+\right. \\
& \qquad \left.\Lambda_{22}(z_2-\mu_2)^2\right) \quad (\because \Lambda_{21}=\Lambda_{12}) \\
= \quad & -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2-\Lambda_{12}(z_1-\mu_1)(z_2-\mu_2)+\text{Const.} \quad (399)
\end{aligned}
$$

これを代入して

$$\begin{aligned} & \ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \\ = \quad & -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2 - \Lambda_{12}(z_1-\mu_1)(z_2-\mu_2) + \text{Const.} \end{aligned} \tag{400}$$

従って

$$\begin{aligned} & \ln q_1^*(z_1) \\ = \quad & \mathbb{E}_{z_2}\left[\ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})\right] + \text{Const.} \\ = \quad & \mathbb{E}_{z_2}\left[-\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2 - \Lambda_{12}(z_1-\mu_1)(z_2-\mu_2)\right] + \text{Const.} \\ = \quad & -\frac{1}{2}\Lambda_{11}z_1^2 + \Lambda_{11}\mu_1 z_1 - \Lambda_{12}z_1\left(\mathbb{E}[z_2]-\mu_2\right) + \text{Const.} \end{aligned} \tag{401}$$

変分推論

- これより、$q_1^*(z_1)$ は次のように書ける

$$
\begin{aligned}
& q_1^*(z_1) \\
\propto \quad & \exp\left(-\frac{1}{2}\Lambda_{11}z_1^2 + \Lambda_{11}\mu_1 z_1 - \Lambda_{12}z_1\left(\mathbb{E}[z_2] - \mu_2\right)\right) \\
= \quad & \exp\left(-\frac{1}{2}\Lambda_{11}\left(z_1 - (\mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2] - \mu_2))\right)^2 + \right. \\
& \left. \frac{1}{2}\Lambda_{11}\left(\mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2] - \mu_2)\right)^2\right) \\
\propto \quad & \exp\left(-\frac{1}{2\Lambda_{11}^{-1}}\left(z_1 - (\mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2] - \mu_2))\right)^2\right) \\
= \quad & \mathcal{N}(z_1|m_1, \Lambda_{11}^{-1}) \qquad (402)
\end{aligned}
$$

但し m_1 は次のようにおいた

$$
m_1 = \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2] - \mu_2) \qquad (403)
$$

変分推論

- 対称性から、$q_2^*(z_2)$ も次のように求められる

$$\begin{aligned} \ln q_2^*(z_2) &= \mathbb{E}_{z_2}[\ln p(\boldsymbol{z})] + \text{Const.} && (404) \\ &= \mathbb{E}_{z_2}[\ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})] + \text{Const.} && (405) \end{aligned}$$

$$q_2^*(z_2) = \mathcal{N}(z_2|m_2, \Lambda_{22}^{-1}) \quad (406)$$

但し m_2 は次のようにおいた

$$m_2 = \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathbb{E}[z_1] - \mu_1) \quad (407)$$

- これより $q_1^*(z_1), q_2^*(z_2)$ は

$$\begin{aligned} q_1^*(z_1) &= \mathcal{N}(z_1|m_1, \Lambda_{11}^{-1}) && (408) \\ m_1 &= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(\mathbb{E}[z_2] - \mu_2) && (409) \\ q_2^*(z_2) &= \mathcal{N}(z_2|m_2, \Lambda_{22}^{-1}) && (410) \\ m_2 &= \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(\mathbb{E}[z_1] - \mu_1) && (411) \end{aligned}$$

変分推論

- $\mathbb{E}[z_2]$ は、z_2 の $q_2(z_2) = \mathcal{N}(z_2|m_2, \Lambda_{22}^{-1})$ による平均であるから、$\mathbb{E}[z_2] = m_2$ である (同様に、$\mathbb{E}[z_1] = z_1$)

- これより m_1, m_2 を連立させれば

$$
\begin{aligned}
m_1 &= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(m_2 - \mu_2) && (412)\\
m_2 &= \mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(m_1 - \mu_1) && (413)
\end{aligned}
$$

であるから、m_1 について解けば

$$
\begin{aligned}
m_1 &= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(m_2 - \mu_2) && (414)\\
&= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}\left(\mu_2 - \Lambda_{22}^{-1}\Lambda_{21}(m_1 - \mu_1) - \mu_2\right) && (415)\\
&= \mu_1 + \Lambda_{11}^{-1}\Lambda_{12}\Lambda_{22}^{-1}\Lambda_{21}(m_1 - \mu_1) && (416)\\
&= \mu_1 + \Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^2(m_1 - \mu_1) && (417)\\
&= \left(1 - \Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^2\right)\mu_1 + \Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^2 m_1 && (418)
\end{aligned}
$$

変分推論

従って

$$\left(1-\Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)\mu_1=\left(1-\Lambda_{11}^{-1}\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)m_1 \tag{419}$$

精度行列 $\mathbf{\Lambda}$ が正則であれば

$$\begin{aligned}
& \Lambda_{11}\Lambda_{12}-\Lambda_{12}\Lambda_{21}\neq 0 && (420)\\
\Rightarrow\quad & \Lambda_{11}\Lambda_{12}-\Lambda_{12}^{2}\neq 0 && (421)\\
\Rightarrow\quad & \left(\Lambda_{11}\Lambda_{12}\right)\left(1-\Lambda_{11}^{-1}\Lambda_{12}^{-1}\Lambda_{12}^{2}\right)\neq 0 && (422)\\
\Rightarrow\quad & \left(1-\Lambda_{11}^{-1}\Lambda_{12}^{-1}\Lambda_{12}^{2}\right)\neq 0 && (423)
\end{aligned}$$

が成立するから、分布 $p(\boldsymbol{z})$ が非特異 (精度行列が正則) ならば

$$m_1=\mu_1 \tag{424}$$

が唯一の解であるほか、対称性から、m_2 についても以下を得る

$$m_2=\mu_2 \tag{425}$$

変分推論

- ゆえに、$q_1^*(z_1), q_2^*(z_2)$ は

$$q_1^*(z_1) = \mathcal{N}(z_1|\mu_1, \Lambda_{11}^{-1}) \quad (426)$$
$$q_2^*(z_2) = \mathcal{N}(z_2|\mu_2, \Lambda_{22}^{-1}) \quad (427)$$

- 注意点 1
  - $q_1^*(z_1)$ は、$q_2^*(z_2)$ を使って計算される $p(\boldsymbol{z})$ の期待値 $\mathbb{E}[z_2]$ に依存 (逆も成り立つ)
  - $q_1^*(z_1)$ と、$q_2^*(z_2)$ は相互に依存しているため、2 つを同時に求めることはできない
  - その代わりに、次のように最適化すればよい
  - $q_1(z_1), q_2(z_2)$ を適当に初期化したあと、$q_1^*(z_1), q_2^*(z_2)$ の式を使って、交互に $q_1(z_1)$ と $q_2(z_2)$ を更新していく (収束するまでこれを繰り返す)
- 注意点 2

変分推論

- $q_1(z_1), q_2(z_2)$ の具体的な関数形については、何の仮定も置かなかった
- $q_i^*(z_i)$ がガウス分布だという仮定は置いていないが、$\mathrm{KL}(q||p)$ を最適化する変分推論によって、結果的にガウス分布が得られた

変分推論

- $\mathrm{KL}(q||p)$ の最適化と $\mathrm{KL}(p||q)$ の最適化の比較
  - 上記の結果は、$\mathrm{KL}(q||p)$ の最適化 (エビデンス下界 $\mathcal{L}(q)$ の最適化) によって得た
  - $\mathrm{KL}(q||p)$ ではなく、$\mathrm{KL}(p||q)$ を最適化したらどうなるか?
  - 変分推論ではない、もう一つの近似推論の方法である、EP 法で使われる考え方

  - $q(\boldsymbol{Z})$ を $p(\boldsymbol{Z}|\boldsymbol{X})$ に近づけたいのであれば、$\mathrm{KL}(q||p)$ と $\mathrm{KL}(p||q)$ のどちらを最小化してもよいはず
  - なぜなら、KL ダイバージェンスは、確率分布間の (擬似的な) 距離を表すため

- $\mathrm{KL}(p||q)$ の最適化
  - $q(\boldsymbol{Z})$ が平均場近似によって分解できるとき、$\mathrm{KL}(p||q)$ を最適化したい

変分推論

- KL ダイバージェンス $\mathrm{KL}(p||q)$ は、次のように書ける

$$
\begin{aligned}
\mathrm{KL}(p||q) &\equiv \mathrm{KL}(p(\boldsymbol{Z}|\boldsymbol{X})||q(\boldsymbol{Z})) && (428)\\
&= -\int p(\boldsymbol{Z}|\boldsymbol{X}) \ln \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})} d\boldsymbol{Z} && (429)\\
&= -\int p(\boldsymbol{Z}|\boldsymbol{X}) \left(\ln q(\boldsymbol{Z}) - \ln p(\boldsymbol{Z}|\boldsymbol{X})\right) d\boldsymbol{Z} && (430)\\
&= -\int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q(\boldsymbol{Z}) d\boldsymbol{Z} - \\
&\qquad \underbrace{-\int p(\boldsymbol{Z}|\boldsymbol{X}) \ln p(\boldsymbol{Z}|\boldsymbol{X}) d\boldsymbol{Z}}_{q \text{ には依存しない定数項}} && (431)\\
&= -\int p(\boldsymbol{Z}|\boldsymbol{X}) \ln \prod_i q_i(\boldsymbol{Z}_i) d\boldsymbol{Z} + \mathrm{Const.} && (432)\\
&= -\int p(\boldsymbol{Z}|\boldsymbol{X}) \sum_i \ln q_i(\boldsymbol{Z}_i) d\boldsymbol{Z} + \mathrm{Const.} && (433)
\end{aligned}
$$

$$= -\sum_i \int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q_i(\boldsymbol{Z}_i) d\boldsymbol{Z} + \text{Const.} \quad (434)$$

定数項は、$p(\boldsymbol{Z}|\boldsymbol{X})$ のエントロピーであり、q には依存しない

- 各因子 $q_j(\boldsymbol{Z}_j)$ について $\mathrm{KL}(p||q)$ を最適化したい
- このとき、$i \neq j$ となる、全ての $q_i(\boldsymbol{Z}_i)$ は固定する

- $q_j(\boldsymbol{Z}_j)$ に依存する項を取り出せば、次のようになる

$$\sum_i \int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q_i(\boldsymbol{Z}_i) d\boldsymbol{Z}$$

$$= \int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z} \quad (435)$$

$$= \int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_1 d\boldsymbol{Z}_2 \cdots d\boldsymbol{Z}_M \quad (436)$$

変分推論

$$= \int p(\boldsymbol{Z}|\boldsymbol{X}) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j \left(\prod_{i \neq j} d\boldsymbol{Z}_i \right) \tag{437}$$

$$= \int \left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i \right) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j \tag{438}$$

- $q_j(\boldsymbol{Z}_j)$ は確率分布であるから、以下の条件を満たさなければならない

$$\int q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j = 1 \quad (\text{規格化条件}) \tag{439}$$

$$q_j(\boldsymbol{Z}_j) \geq 0 \tag{440}$$

- 従って $\mathrm{KL}(p||q)$ を $q_j(\boldsymbol{Z}_j)$ について最適化するとき、ラグランジュの未定乗数法を使って、規格化条件を組み込む必要がある

変分推論

- $q_j(\boldsymbol{Z}_j) \geq 0$ という条件は、$\ln q_i(\boldsymbol{Z}_i)$ という項が既にあるから、何もしなくても常に満たされる (ラグランジュ関数に、制約条件を改めて取り入れる必要がない)

- 結局、ラグランジュ汎関数 $\mathcal{L}[q_j]$ は、次のようになる

$$\begin{aligned}\mathcal{L}[q_j] &= -\int \left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i \right) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j + \\ &\quad \lambda \left(\int q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j - 1 \right) \end{aligned} \tag{441}$$

- 上記は $q_j(\boldsymbol{Z}_j)$ についての汎関数となっていることに注意

- 次の公式を使って、$\mathcal{L}[q_j]$ を変分最適化する

$$\frac{\delta}{\delta y(x)} \int G(y(x), x) dx = \frac{\partial}{\partial y} G(y(x), x) \tag{442}$$

変分推論

- 従って

$$
\begin{aligned}
& \frac{\delta}{\delta q_j(\boldsymbol{Z}_j)} \mathcal{L}[q_j] \\
= & -\frac{\delta}{\delta q_j(\boldsymbol{Z}_j)} \int \left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i \right) \ln q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j + \\
& \frac{\delta}{\delta q_j(\boldsymbol{Z}_j)} \lambda \left(\int q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j - 1 \right) \qquad (443) \\
= & -\frac{\partial}{\partial q_j} \left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i \right) \ln q_j(\boldsymbol{Z}_j) + \\
& \lambda \frac{\partial}{\partial q_j} q_j(\boldsymbol{Z}_j) \qquad (444) \\
= & -\left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i \right) \frac{1}{q_j(\boldsymbol{Z}_j)} + \lambda = 0 \qquad (445)
\end{aligned}
$$

変分推論

- これより、未定乗数 λ は

$$\lambda q_j(\boldsymbol{Z}_j) = \left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i \right) \frac{1}{q_j(\boldsymbol{Z}_j)} \tag{446}$$

$$\Rightarrow \quad \int \lambda q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j = \int \underbrace{\left(\int p(\boldsymbol{Z}|\boldsymbol{X}) \prod_{i \neq j} d\boldsymbol{Z}_i \right)}_{=p(\boldsymbol{Z}_j|\boldsymbol{X})} d\boldsymbol{Z}_j \tag{447}$$

$$\Rightarrow \quad \lambda \underbrace{\int q_j(\boldsymbol{Z}_j) d\boldsymbol{Z}_j}_{=1} = \underbrace{\int p(\boldsymbol{Z}_j|\boldsymbol{X}) d\boldsymbol{Z}_j}_{=1} \tag{448}$$

$$\Rightarrow \quad \lambda = 1 \tag{449}$$

変分推論

- 結局、最適解 $q_j^*(\boldsymbol{Z}_j)$ は次のようになる

$$-\left(\int p(\boldsymbol{Z}|\boldsymbol{X})\prod_{i\neq j} d\boldsymbol{Z}_i\right)\frac{1}{q_j(\boldsymbol{Z}_j)}+\lambda=0 \tag{450}$$

$$\Rightarrow \quad q_j^*(\boldsymbol{Z}_j)=\underbrace{\int p(\boldsymbol{Z}|\boldsymbol{X})\prod_{i\neq j} d\boldsymbol{Z}_i}_{=p(\boldsymbol{Z}_j|\boldsymbol{X})} \tag{451}$$

$$\Rightarrow \quad q_j^*(\boldsymbol{Z}_j)=p(\boldsymbol{Z}_j|\boldsymbol{X}) \tag{452}$$

- $q_j^(\boldsymbol{Z}_j)$ の最適解は、$p(\boldsymbol{Z}|\boldsymbol{X})$ を、$i\neq j$ である全ての $\boldsymbol{Z}_i$ について周辺化した分布
- これは閉じた解であり、繰り返しを必要としない

- 最適な因子 $q_1(z_1), q_2(z_2)$ の計算

変分推論

- 今回は $p(\boldsymbol{z}|\boldsymbol{x}) = p(\boldsymbol{z})$ の場合を考えており、かつ $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1})$ であった

- 従って、$q_1^*(z_1)$ は、$p(\boldsymbol{z})$ を z_2 について周辺化すればよいから

$$
\begin{aligned}
& q_1^*(z_1) \\
= & \int p(\boldsymbol{z}) dz_2 \\
= & \int \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) dz_2 && (453) \\
= & \frac{1}{2\pi} \frac{1}{|\boldsymbol{\Lambda}|^{-\frac{1}{2}}} \int \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu})\right) dz_2 && (454)
\end{aligned}
$$

変分推論

- ここで、指数の内側を、積分変数 z_2 に依存する項と、そうでない項に分ける

$$
\begin{aligned}
& -\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu}) \\
= \quad & -\frac{1}{2}\left(\Lambda_{11}(z_1-\mu_1)^2+\right. \\
& \qquad \left. 2\Lambda_{12}(z_1-\mu_1)(z_2-\mu_2)+\Lambda_{22}(z_2-\mu_2)^2\right) \qquad (455) \\
= \quad & -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2+\Lambda_{12}(z_1-\mu_1)\mu_2+ \\
& \qquad -\frac{1}{2}\left(2\Lambda_{12}(z_1-\mu_1)z_2+\Lambda_{22}(z_2-\mu_2)^2\right) \qquad (456)
\end{aligned}
$$

そして

$$
\begin{aligned}
& -\frac{1}{2}\left(2\Lambda_{12}(z_1-\mu_1)z_2+\Lambda_{22}(z_2-\mu_2)^2\right) \\
= \quad & -\frac{1}{2}\left(\Lambda_{22}z_2^2-2\Lambda_{22}\mu_2 z_2+2\Lambda_{12}(z_1-\mu_1)z_2+\Lambda_{22}\mu_2^2\right) \qquad (457)
\end{aligned}
$$

$$= -\frac{1}{2}\left(\Lambda_{22}z_2^2 - 2\left(\Lambda_{22}\mu_2 - \Lambda_{12}(z_1-\mu_1)\right)z_2 + \Lambda_{22}\mu_2^2\right) \quad (458)$$

$$= -\frac{1}{2}\Big(\Lambda_{22}\left(z_2 - \Lambda_{22}^{-1}\left(\Lambda_{22}\mu_2 - \Lambda_{12}(z_1-\mu_1)\right)\right)^2 - \Lambda_{22}\left(\Lambda_{22}^{-1}\left(\Lambda_{22}\mu_2 - \Lambda_{12}(z_1-\mu_1)\right)\right)^2 + \Lambda_{22}\mu_2^2\Big) \quad (459)$$

$$= -\frac{1}{2}\left(\Lambda_{22}\left(z_2 - \Lambda_{22}^{-1}m\right)^2 - \Lambda_{22}^{-1}m^2 + \Lambda_{22}\mu_2^2\right) \quad (460)$$

ゆえ ($m = \Lambda_{22}\mu_2 - \Lambda_{12}(z_1-\mu_1)$ とおいた)

$$-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu})$$

$$= -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2 + \Lambda_{12}(z_1-\mu_1)\mu_2 - \frac{1}{2}\Lambda_{22}\mu_2^2 - \frac{1}{2}\Lambda_{22}\left(z_2 - \Lambda_{22}^{-1}m\right)^2 + \frac{1}{2}\Lambda_{22}^{-1}m^2 \quad (461)$$

変分推論

- これより、積分変数 z_2 の依存項だけを取り出せたので

$$\begin{aligned}
&\int \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu})\right)dz_2 \\
= \quad &\exp\left(-\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2+\Lambda_{12}(z_1-\mu_1)\mu_2-\right. \\
&\left.\frac{1}{2}\Lambda_{22}\mu_2^2+\frac{1}{2}\Lambda_{22}^{-1}m^2\right) \\
&\int \exp\left(-\frac{1}{2}\Lambda_{22}\left(z_2-\Lambda_{22}^{-1}m\right)^2\right)dz_2 \qquad (462)
\end{aligned}$$

であって、右側の積分は、中身が (正規化されていない) ガウス分布であるから

$$\int \exp\left(-\frac{1}{2}\Lambda_{22}\left(z_2-\Lambda_{22}^{-1}m\right)^2\right)dz_2$$

$$
\begin{aligned}
&= (2\pi\Lambda_{22}^{-1})^{\frac{1}{2}} \cdot \int \frac{1}{(2\pi\Lambda_{22}^{-1})^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\Lambda_{22}\left(z_2 - \Lambda_{22}^{-1}m\right)^2\right) dz_2 \\
&= \left(2\pi\Lambda_{22}^{-1}\right)^{\frac{1}{2}} \qquad (463)
\end{aligned}
$$

となって、z_2 を積分により消去できる

- また指数の残りの部分から、z_1 に依存する項だけを取り出して

$$
\begin{aligned}
& -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2 + \Lambda_{12}(z_1-\mu_1)\mu_2 - \frac{1}{2}\Lambda_{22}\mu_2^2 + \frac{1}{2}\Lambda_{22}^{-1}m^2 \\
= & -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2 + \Lambda_{12}(z_1-\mu_1)\mu_2 - \frac{1}{2}\Lambda_{22}\mu_2^2 + \\
& \quad \frac{1}{2}\Lambda_{22}^{-1}\left(\Lambda_{22}\mu_2 - \Lambda_{12}(z_1-\mu_1)\right)^2 \\
= & -\frac{1}{2}\Lambda_{11}(z_1-\mu_1)^2 + \Lambda_{12}(z_1-\mu_1)\mu_2 - \frac{1}{2}\Lambda_{22}\mu_2^2 + \\
& \quad \frac{1}{2}\Lambda_{22}\mu_2^2 - \mu_2\Lambda_{12}(z_1-\mu_1) +
\end{aligned}
$$

$$\frac{1}{2}\Lambda_{22}^{-1}\Lambda_{12}^2(z_1-\mu_1)^2 \tag{464}$$

$$= -\frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^2\right)(z_1-\mu_1)^2 \tag{465}$$

$$= -\frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^2\right)z_1^2 + \frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^2\right)\mu_1 z_1 + \text{Const.} \tag{466}$$

- これより結局、z_2 による積分は次のようになる

$$\int \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu})\right)dz_2$$

$$= \left(2\pi\Lambda_{22}^{-1}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^2\right)(z_1-\mu_1)^2\right) \tag{467}$$

変分推論

- 従って、$q_1^*(z_1)$ は次のようになる

$$
\begin{aligned}
& q_1^*(z_1) \\
= & \frac{1}{2\pi} \frac{1}{|\mathbf{\Lambda}|^{-\frac{1}{2}}} \int \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T \mathbf{\Lambda}(\boldsymbol{z}-\boldsymbol{\mu})\right) dz_2 && (468) \\
= & \frac{1}{2\pi} \frac{1}{\left(\Lambda_{11}\Lambda_{22}-\Lambda_{12}^2\right)^{-\frac{1}{2}}} \left(2\pi\Lambda_{22}^{-1}\right)^{\frac{1}{2}} \\
& \exp\left(-\frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^2\right)(z_1-\mu_1)^2\right) && (469) \\
= & \frac{1}{(2\pi)^{\frac{1}{2}}} \frac{1}{\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^2\right)^{-\frac{1}{2}}} \\
& \exp\left(-\frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^2\right)(z_1-\mu_1)^2\right) && (470) \\
= & \mathcal{N}(z_1|\mu_1, (\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^2)^{-1}) && (471)
\end{aligned}
$$

変分推論

- 共分散行列 $\boldsymbol{\Sigma}$ を、精度行列 $\boldsymbol{\Lambda}$ を使って次のように定めれば

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} \quad (\Sigma_{12} = \Sigma_{21}) \tag{472}$$

次が成り立つから

$$\begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \begin{pmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{pmatrix} = \boldsymbol{I} \tag{473}$$

各成分に注目すれば

$$\begin{cases} \Sigma_{11}\Lambda_{11} + \Sigma_{12}\Lambda_{21} = 1 \\ \Sigma_{11}\Lambda_{12} + \Sigma_{12}\Lambda_{22} = 0 \end{cases} \tag{474}$$

これを Σ_{11} について解けば

$$\Sigma_{11}\Lambda_{11} + \left(-\Lambda_{22}^{-1}\Lambda_{12}\Sigma_{11}\right)\Lambda_{21} = 1 \tag{475}$$

$$\Rightarrow \quad \Sigma_{11}\left(\Lambda_{11} - \Lambda_{22}^{-1}\Lambda_{12}\Lambda_{21}\right) = 1 \tag{476}$$

$$\Rightarrow \quad \Sigma_{11}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)=1 \tag{477}$$

$$\Rightarrow \quad \Sigma_{11}=\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)^{-1} \tag{478}$$

- これから、$q_1^*(z_1)$ は次のようにも書ける

$$\begin{aligned} & q_1^*(z_1) \\ = & \frac{1}{(2\pi)^{\frac{1}{2}}}\frac{1}{(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^{2})^{-\frac{1}{2}}} \\ & \exp\left(-\frac{1}{2}\left(\Lambda_{11}-\Lambda_{22}^{-1}\Lambda_{12}^{2}\right)(z_1-\mu_1)^2\right) \end{aligned} \tag{479}$$

$$= \frac{1}{(2\pi)^{\frac{1}{2}}}\frac{1}{\Sigma_{11}^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\Sigma_{11}^{-1}(z_1-\mu_1)^2\right) \tag{480}$$

$$= \mathcal{N}(z_1|\mu_1,\Sigma_{11}) \tag{481}$$

変分推論

- $q_2^*(z_2)$ は、対称性から次のようになる

$$q_2^*(z_2) = \mathcal{N}(z_2|\mu_2, (\Lambda_{22} - \Lambda_{11}^{-1}\Lambda_{12}^2)^{-1}) = \mathcal{N}(z_2|\mu_2, \Sigma_{22}) \tag{482}$$

- 2 つの解の比較
  - $\mathrm{KL}(q||p)$ の最小化によって次の解を得た

$$q_1^*(z_1) = \mathcal{N}(z_1|\mu_1, \Lambda_{11}^{-1}) \tag{483}$$
$$q_2^*(z_2) = \mathcal{N}(z_2|\mu_2, \Lambda_{22}^{-1}) \tag{484}$$

変分推論

- $\mathrm{KL}(p||q)$ の最小化では、次の解を得た

$$
\begin{aligned}
q_1^*(z_1) &= \mathcal{N}(z_1|\mu_1, \Sigma_{11}) && (485)\\
\Sigma_{11} &= \left(\Lambda_{11} - \Lambda_{22}^{-1}\Lambda_{12}^2\right)^{-1} && (486)\\
q_2^*(z_2) &= \mathcal{N}(z_2|\mu_2, \Sigma_{22}) && (487)\\
\Sigma_{22} &= (\Lambda_{22} - \Lambda_{11}^{-1}\Lambda_{12}^2)^{-1} && (488)
\end{aligned}
$$

- $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ の平均は $\boldsymbol{\mu} = [\mu_1, \mu_2]^T$ であったので、いずれの場合も、平均は正しく捉えている
- しかし、両者の間では、分散が異なっている
- また、変数 z_1 と z_2 の間の相関は消えてなくなっている

- 両者の違いを、次の図 13 に示す

- 緑色の線が真の分布 $p(\boldsymbol{z})$ を表す
- 左側の赤線は、$\mathrm{KL}(q||p)$ の最小化によって得られた分布 $q(\boldsymbol{z})$
- 右側の赤線は、$\mathrm{KL}(p||q)$ の最小化によって得られた分布 $q(\boldsymbol{z})$

Figure 10.2 Comparison of the two alternative forms for the Kullback-Leibler divergence. The green contours corresponding to 1, 2, and 3 standard deviations for a correlated Gaussian distribution $p(\mathbf{z})$ over two variables z_1 and z_2, and the red contours represent the corresponding levels for an approximating distribution $q(\mathbf{z})$ over the same variables given by the product of two independent univariate Gaussian distributions whose parameters are obtained by minimization of (a) the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$, and (b) the reverse Kullback-Leibler divergence $\mathrm{KL}(p\|q)$.

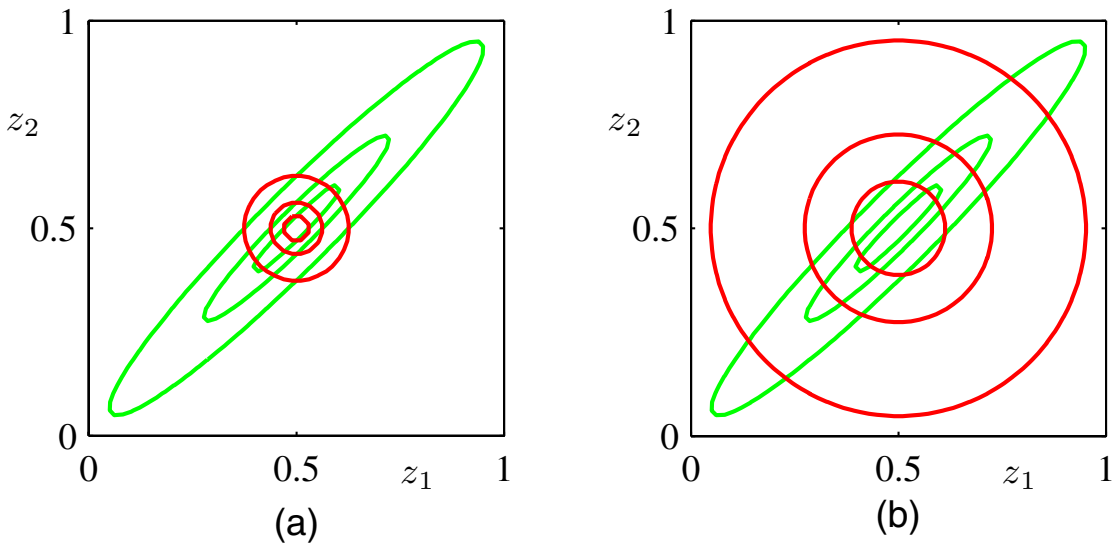

図 13: KL ダイバージェンスの 2 つの形の比較

変分推論

- 2 つの解の比較
  - $\mathrm{KL}(q||p)$ の最小化で得られる $q(\boldsymbol{z})$ は、分散が小さくなる方向に制御されていて、それと直交する方向の分散は、大きく過小評価されている
  - 分解による近似では、一般に事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ をコンパクトに近似しすぎる

  - $\mathrm{KL}(p||q)$ の最小化で得られる $q(\boldsymbol{z})$ は、分散が大きくなる方向に制御されていて、それと直交する方向の分散は、過大評価されている
  - 非常に低い確率しか持たないはずの領域にも、多くの確率質量が割り当てられている

  - 変分推論では、計算コストの観点から $\mathrm{KL}(q||p)$ の方を用いる
  - $\mathrm{KL}(q||p)$ の計算には、q に関する期待値の評価が含まれるので、q を単純な形に制限することで、必要な期待値を単純化できる
  - $\mathrm{KL}(p||q)$ の計算には、真の事後分布 p に関する期待値の計算が必要である

変分推論

- 違いが生じる理由
  - KL ダイバージェンス $\mathrm{KL}(q||p)$ は次のようであった

$$\mathrm{KL}(q||p) = -\int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z}|\boldsymbol{X})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{489}$$

  - $\mathrm{KL}(q||p)$ が大きくなる主要因は、$p(\boldsymbol{Z})$ がほとんど 0 で、$q(\boldsymbol{Z})$ はそうでない領域
  - 従って、$\mathrm{KL}(q||p)$ を最小化すると、$q(\boldsymbol{Z})$ は、$p(\boldsymbol{Z})$ が小さい領域を避けるようになる

  - また、$\mathrm{KL}(p||q)$ は次のようであった

$$\mathrm{KL}(p||q) = -\int p(\boldsymbol{Z}|\boldsymbol{X}) \ln \frac{q(\boldsymbol{Z})}{p(\boldsymbol{Z}|\boldsymbol{X})} d\boldsymbol{Z} \tag{490}$$

変分推論

- $\mathrm{KL}(p||q)$ が大きくなる主要因は、$q(\boldsymbol{Z})$ がほとんど 0 で、$p(\boldsymbol{Z})$ はそうではない領域
- 従って、$\mathrm{KL}(p||q)$ を最小化すると、$q(\boldsymbol{Z})$ は、$p(\boldsymbol{Z})$ が 0 でない領域にも、必ず確率を持たせるようになる

- 別の分布について、この両者の振舞いの違いを観察してみよう

変分推論

- 多峰性のある分布を、単峰の分布で近似する場合
  - $\mathrm{KL}(q||p)$ を最小化する変分近似では、多数ある峰のうちの 1 つを再現
  - $\mathrm{KL}(p||q)$ を最小化する変分近似では、全ての峰を平均したような分布が得られる
  - 多峰性のある分布を平均してしまうと、予測性能の悪化をもたらす
  - これらの比較を次の図 14 に示す

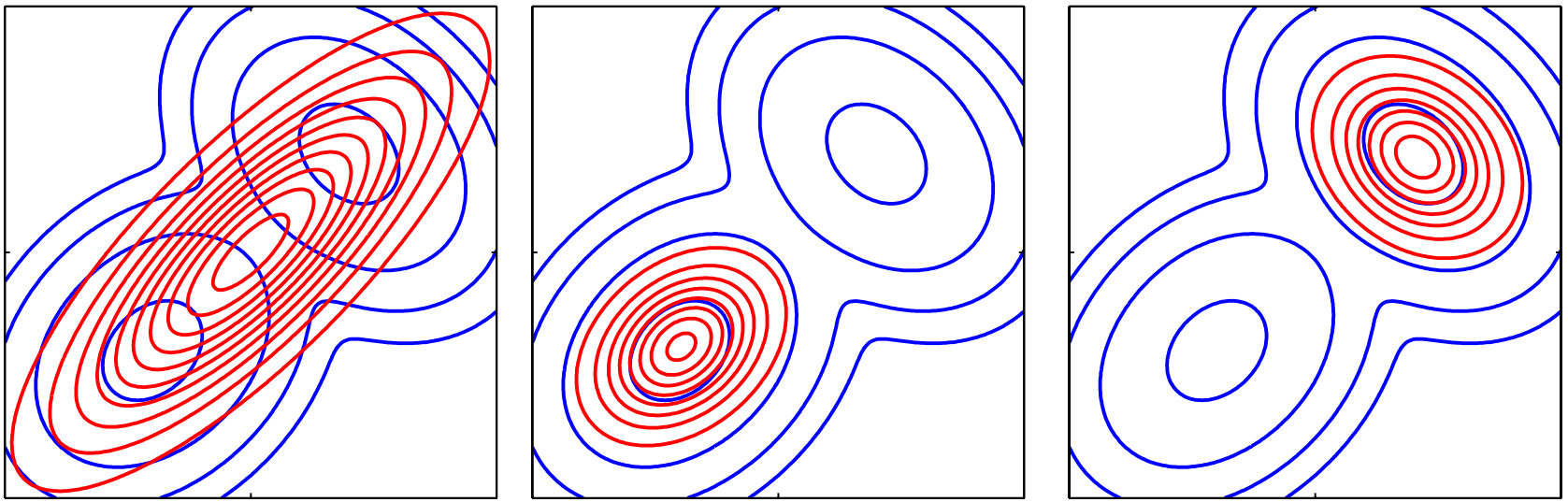

Figure 10.3 Another comparison of the two alternative forms for the Kullback-Leibler divergence. (a) The blue contours show a bimodal distribution $p(\mathbf{Z})$ given by a mixture of two Gaussians, and the red contours correspond to the single Gaussian distribution $q(\mathbf{Z})$ that best approximates $p(\mathbf{Z})$ in the sense of minimizing the Kullback-Leibler divergence $\mathrm{KL}(p\|q)$. (b) As in (a) but now the red contours correspond to a Gaussian distribution $q(\mathbf{Z})$ found by numerical minimization of the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$. (c) As in (b) but showing a different local minimum of the Kullback-Leibler divergence.

図 14: KL ダイバージェンスの 2 つの形の別の比較

変分推論

- ここまでの話の流れ
  1. 分解 $q(\boldsymbol{Z}) = \prod_i q_i(\boldsymbol{Z}_i)$ を使ったエビデンス下界 $\mathcal{L}(q)$ の最適化は、$\mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z}|\boldsymbol{X}))$ の最小化と等価である
  2. $\mathrm{KL}(q||p)$ と、$\mathrm{KL}(p||q)$ を最小化する変分近似を、2 変数のガウス分布を例として試した
  3. $\mathrm{KL}(q||p)$ の最小化を使って求めた $q(\boldsymbol{Z})$ は、事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ をコンパクトに近似する傾向にあった
  4. $\mathrm{KL}(p||q)$ の最小化によって求めた $q(\boldsymbol{Z})$ は、事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ を大きく捉えて近似する傾向にあった
- これからの話の流れ
  - 変分推論の具体的な例について更に見ていく

変分推論

- 離散潜在変数のモデル (二値スパース符号化モデル) に変分推論を適用する
- 連続潜在変数の場合は、簡単な確率モデルを使って、変分推論を試してみる

目次

離散潜在変数の変分推論

- 離散潜在変数をもつ変分推論の概要
  - ここでは単純な場合を扱う
  - データ $\boldsymbol{x}_i$ に対する潜在変数 $\boldsymbol{z}_i$ は、各要素が二値であるとする
  - データ $\boldsymbol{x}$ は D 次元、潜在変数 $\boldsymbol{z}$ は K 次元とする

  - 分布 $q(\boldsymbol{z}_i|\boldsymbol{x}_i)$ は、平均場近似によって、次のように分解できるとする

$$q(\boldsymbol{z}_i|\boldsymbol{x}_i) = \prod_k q(z_{ik}|\boldsymbol{x}_i) \tag{491}$$

  - 潜在変数は二値であるから、$q(z_{ik} = 1|\boldsymbol{x}_i) = \widehat{z_{ik}}$ と書くことにする
  - このとき、$q(z_{ik}|\boldsymbol{x}_i)$ は次のようになる

$$q(z_{ik}|\boldsymbol{x}_i) = \widehat{z_{ik}}^{z_{ik}} \left(1 - \widehat{z_{ik}}\right)^{(1-z_{ik})} \tag{492}$$

  - 各パラメータ $\widehat{z_{ik}}$ について、下界 $\mathcal{L}(q)$ を順番に最適化することを、$\mathcal{L}(q)$ が収束するまで繰り返し行う

離散潜在変数の変分推論

- 即ち、以下の不動点方程式を、各 $\widehat{z_{ik}}$ について繰り返し解く

$$\frac{\partial}{\partial \widehat{z_{ik}}} \mathcal{L}(q) = 0 \tag{493}$$

- 離散潜在変数の場合は、単なる標準的な最適化問題を解くことになる

- 二値スパース符号化モデル
  - 二値スパース符号化モデルでのデータ生成過程は、次のようになる
  - 各データ $\boldsymbol{x}_i$ には、潜在変数 $\boldsymbol{z}_i \in \{0,1\}^K$ が対応する
  - $\boldsymbol{z}_i$ に対し、重み $\boldsymbol{W}$ を用いて線形変換を施し、更にガウスノイズを足し合わせることで、データ $\boldsymbol{x}_i$ が生成される
  - 潜在変数 $\boldsymbol{z}_i$ は、K 次元ベクトルであり、その各要素は 0 または 1 である

離散潜在変数の変分推論

- 従って、確率分布は次のようになる

$$p(z_{ik}=1) = \sigma(b_{ik}) \tag{494}$$
$$p(\boldsymbol{x}_i|\boldsymbol{z}_i) = \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i, \boldsymbol{\beta}^{-1}) \tag{495}$$

- $\sigma(\cdot)$ はシグモイド関数、$\boldsymbol{b}_i = [b_{i1}, \ldots, b_{iK}]^T$ は学習可能なバイアス、$\boldsymbol{W}$ は学習可能な重み行列、$\boldsymbol{\beta}$ は学習可能な対角精度行列である
- $\boldsymbol{\beta} = \mathrm{diag}\left(\beta_1, \beta_2, \ldots, \beta_D\right)$ と書ける

- $p(\boldsymbol{z}_i)$ は次のように計算できる

$$p(\boldsymbol{z}_i) = \prod_k p(z_{ik}) \tag{496}$$
$$= \prod_k \left(\sigma(b_{ik})\right)^{z_{ik}} \left(1-\sigma(b_{ik})\right)^{1-z_{ik}} \tag{497}$$
$$= \prod_k \left(\sigma(b_{ik})\right)^{z_{ik}} \left(\sigma(-b_{ik})\right)^{1-z_{ik}} \tag{498}$$

離散潜在変数の変分推論

- 事後分布 $p(\boldsymbol{z}_i|\boldsymbol{x}_i)$ は複雑であり、表現することも、計算することもできない
- 従って、最尤推定の手法 (EM アルゴリズム) により学習することはできない

- 例えばバイアス b_{ik} に関する微分を考えると

$$\begin{aligned} & \frac{\partial}{\partial b_{ik}} \ln p(\boldsymbol{z}_i|\boldsymbol{x}_i) \\ = & \frac{\partial}{\partial b_{ik}} \ln \frac{p(\boldsymbol{x}_i, \boldsymbol{z}_i)}{p(\boldsymbol{x}_i)} \quad (499) \\ = & \frac{\partial}{\partial b_{ik}} \left(\ln p(\boldsymbol{x}_i, \boldsymbol{z}_i) - \ln p(\boldsymbol{x}_i)\right) \quad (500) \end{aligned}$$

であって、第 1 項の $p(\boldsymbol{x}_i, \boldsymbol{z}_i)$ は容易に計算できるが、第 2 項について考えると

$$\frac{\partial}{\partial b_{ik}} \ln p(\boldsymbol{x}_i)$$

$$= \frac{1}{p(\boldsymbol{x}_i)} \frac{\partial}{\partial b_{ik}} p(\boldsymbol{x}_i) \tag{501}$$

$$= \frac{1}{p(\boldsymbol{x}_i)} \frac{\partial}{\partial b_{ik}} \sum_{\boldsymbol{z}_i} p(\boldsymbol{x}_i, \boldsymbol{z}_i) \tag{502}$$

$$= \frac{1}{p(\boldsymbol{x}_i)} \frac{\partial}{\partial b_{ik}} \sum_{\boldsymbol{z}_i} p(\boldsymbol{z}_i) p(\boldsymbol{x}_i|\boldsymbol{z}_i) \tag{503}$$

$$= \frac{1}{p(\boldsymbol{x}_i)} \sum_{\boldsymbol{z}_i} p(\boldsymbol{x}_i|\boldsymbol{z}_i) \frac{\partial}{\partial b_{ik}} p(\boldsymbol{z}_i) \tag{504}$$

$$= \sum_{\boldsymbol{z}_i} \frac{1}{p(\boldsymbol{x}_i)} \frac{p(\boldsymbol{x}_i, \boldsymbol{z}_i)}{p(\boldsymbol{z}_i)} \frac{\partial}{\partial b_{ik}} p(\boldsymbol{z}_i) \tag{505}$$

$$= \sum_{\boldsymbol{z}_i} p(\boldsymbol{z}_i|\boldsymbol{x}_i) \frac{1}{p(\boldsymbol{z}_i)} \frac{\partial}{\partial b_{ik}} p(\boldsymbol{z}_i) \tag{506}$$

$$= \sum_{\boldsymbol{z}_i} p(\boldsymbol{z}_i|\boldsymbol{x}_i) \frac{\partial}{\partial b_{ik}} \ln p(\boldsymbol{z}_i) \tag{507}$$

離散潜在変数の変分推論

$$= \quad \mathbb{E}_{\boldsymbol{z}_i \sim p(\boldsymbol{z}_i|\boldsymbol{x}_i)} \left[\frac{\partial}{\partial b_{ik}} \ln p(\boldsymbol{z}_i) \right] \tag{508}$$

何とか $p(\boldsymbol{z}_i)$ と $p(\boldsymbol{x}_i|\boldsymbol{z}_i)$ を使って計算できないかと試行錯誤したが、結局、$p(\boldsymbol{z}_i|\boldsymbol{x}_i)$ に関する期待値の計算が必要になった

- $p(\boldsymbol{x}_i, \boldsymbol{z}_i)$ と、$p(\boldsymbol{z}_i|\boldsymbol{x}_i)$ のグラフ構造を次の図 15 に示す
- 図では、$\boldsymbol{h} = \boldsymbol{z}$、$\boldsymbol{v} = \boldsymbol{x}$ のように読み替える必要がある

離散潜在変数の変分推論

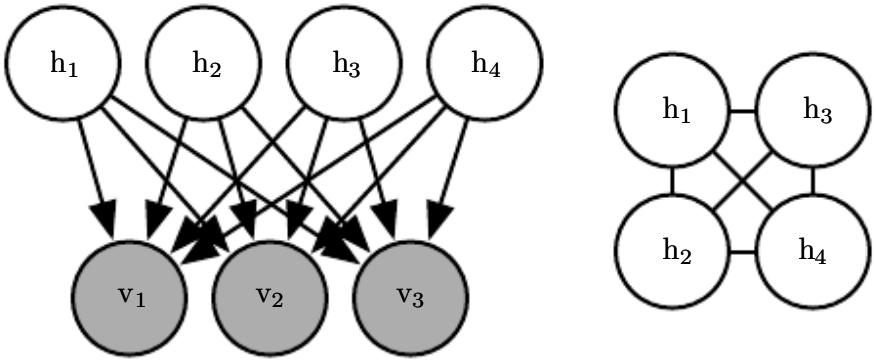

Figure 19.2: The graph structure of a binary sparse coding model with four hidden units. *(Left)*The graph structure of $p(\boldsymbol{h}, \boldsymbol{v})$. Note that the edges are directed, and that every two hidden units are co-parents of every visible unit. *(Right)*The graph structure of $p(\boldsymbol{h} \mid \boldsymbol{v})$. In order to account for the active paths between co-parents, the posterior distribution needs an edge between all of the hidden units.

図 15: 二値スパース符号化モデルのグラフ構造 (4 つの潜在変数をもつ場合)

離散潜在変数の変分推論

- 変分推論による二値スパース符号化モデルの学習
  - 最尤推定の手法を諦める代わりに、変分推論を使うことで、困難を解決できる
  - 分布 $p(\boldsymbol{z}_i|\boldsymbol{x}_i)$ を $q(\boldsymbol{z}_i)$ で表現し、更に平均場近似を行う
  - 但し $\boldsymbol{z}_i = [z_{i1}, z_{i2}, \ldots, z_{iK}]^T$ とする

$$q(\boldsymbol{z}_i|\boldsymbol{x}_i) = \prod_k q(z_{ik}|\boldsymbol{x}_i) \tag{509}$$

  - 潜在変数の各要素は二値であるから、各 $q(z_{ik}|\boldsymbol{x}_i)$ はベルヌーイ分布とすればよい
  - 即ち、$q(z_{ik} = 1|\boldsymbol{x}_i) = \widehat{z_{ik}}$ とする
  - $\widehat{z_{ik}} \neq 0, 1$ という制約を課すことで、$\ln \widehat{z_{ik}}$ を計算できる

  - このようにすれば、潜在変数 $z_{ik}, z_{il}(k \neq l)$ 間の相関を断ち切ることができる
  - 先程の図 15 の右側から、潜在変数間の辺を、全て消し去ることになる

離散潜在変数の変分推論

- これで、平均場近似により、因数分解可能な q を表現することができる

$$q(\boldsymbol{z}_i|\boldsymbol{x}_i) = \prod_k q(z_{ik}|\boldsymbol{x}_i) \tag{510}$$

$$= \prod_k \widehat{z_{ik}}^{z_{ik}} \left(1 - \widehat{z_{ik}}\right)^{1-z_{ik}} \tag{511}$$

$$q(\boldsymbol{Z}|\boldsymbol{X}) = \prod_i q(\boldsymbol{z}_i|\boldsymbol{x}_i) \tag{512}$$

$$= \prod_i \prod_k \widehat{z_{ik}}^{z_{ik}} \left(1 - \widehat{z_{ik}}\right)^{1-z_{ik}} \tag{513}$$

- ソフトウェア上では、丸め誤差などによって $\widehat{z_{ik}}$ が 0 や 1 になり、計算を続行できなくなるかもしれない
- これを回避するためには、パラメータ $\widetilde{z}_i$ を使って二値スパース符号化モデルを学習させる

離散潜在変数の変分推論

- そして、$\widehat{\boldsymbol{z}}_i = \sigma(\widetilde{\boldsymbol{z}}_i)$ の関係によって、$\widehat{\boldsymbol{z}}_i = [\widehat{z_{i1}}, \ldots, \widehat{z_{iK}}]^T$ を得るようにする
- $\ln \widehat{z_{ik}} = \ln \sigma(\widetilde{z_{ik}}) = -\zeta(-\widetilde{z_{ik}})$ によって、コンピュータ上で安全に $\ln \widehat{z_{ik}}$ を計算できる ($\zeta(\cdot)$ はソフトプラス関数)

- 変分推論のために、まずはエビデンス下界 $\mathcal{L}(q)$ を計算する

$$
\begin{aligned}
& \mathcal{L}(q) \\
= & \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z}|\boldsymbol{X})} && (514) \\
= & \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \ln \frac{p(\boldsymbol{Z})p(\boldsymbol{X}|\boldsymbol{Z})}{q(\boldsymbol{Z}|\boldsymbol{X})} && (515) \\
= & \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \left(\ln p(\boldsymbol{Z}) + \ln p(\boldsymbol{X}|\boldsymbol{Z}) - \ln q(\boldsymbol{Z}|\boldsymbol{X})\right) && (516) \\
= & \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \left(\ln p(\boldsymbol{Z}) - \ln q(\boldsymbol{Z}|\boldsymbol{X})\right) + \\
\end{aligned}
$$

離散潜在変数の変分推論

$$\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \ln p(\boldsymbol{X}|\boldsymbol{Z}) \tag{517}$$

- $q(\boldsymbol{Z}|\boldsymbol{X})$ についての期待値を取っていることに注意する
- 全ての $\boldsymbol{Z}$ についての和を取ればよいので、第 1 項は次のようになる

$$\begin{aligned}
& \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \left(\ln p(\boldsymbol{Z}) - \ln q(\boldsymbol{Z}|\boldsymbol{X})\right) \\
= & \sum_{i} \sum_{k} \sum_{z_{ik}} q(z_{ik}|\boldsymbol{x}_i) \left(\ln p(z_{ik}) - \ln q(z_{ik}|\boldsymbol{x}_i)\right) && (518) \\
= & \sum_{i} \sum_{k} q(z_{ik}=1|\boldsymbol{x}_i) \left(\ln p(z_{ik}=1) - \ln q(z_{ik}=1|\boldsymbol{x}_i)\right) + \\
& \qquad q(z_{ik}=0|\boldsymbol{x}_i) \left(\ln p(z_{ik}=0) - \ln q(z_{ik}=0|\boldsymbol{x}_i)\right) && (519) \\
= & \sum_{i} \sum_{k} \{\widehat{z_{ik}} \left(\ln \sigma(b_{ik}) - \ln \widehat{z_{ik}}\right) + \\
& \qquad (1-\widehat{z_{ik}}) \left(\ln\left(1-\sigma(b_{ik})\right) - \ln\left(1-\widehat{z_{ik}}\right)\right)\} && (520)
\end{aligned}$$

離散潜在変数の変分推論

$$= \sum_i \sum_k \{\widehat{z_{ik}} \left(\ln \sigma(b_{ik}) - \ln \widehat{z_{ik}}\right) + \left(1 - \widehat{z_{ik}}\right) \left(\ln \sigma(-b_{ik}) - \ln \left(1 - \widehat{z_{ik}}\right)\right)\} \tag{521}$$

- また第 2 項は、次のようになる

$$\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \ln p(\boldsymbol{X}|\boldsymbol{Z})$$

$$= \sum_i \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i|\boldsymbol{x}_i) \ln p(\boldsymbol{x}_i|\boldsymbol{z}_i) \tag{522}$$

$$= \sum_i \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i|\boldsymbol{x}_i) \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i, \boldsymbol{\beta}^{-1}) \tag{523}$$

離散潜在変数の変分推論

- 但し、$\ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i, \boldsymbol{\beta}^{-1})$ は次のように分解できる

$$
\begin{aligned}
& \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i, \boldsymbol{\beta}^{-1}) \\
= \quad & \ln \left(\frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\boldsymbol{\beta}^{-1}|^{\frac{1}{2}}} \right. \\
& \qquad \left. \exp\left(-\frac{1}{2} \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i\right)^T \boldsymbol{\beta} \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i\right) \right) \right) \qquad (524) \\
= \quad & -\frac{D}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\beta}^{-1}| - \\
& \qquad \frac{1}{2} \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i\right)^T \boldsymbol{\beta} \left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i\right) \qquad (525)
\end{aligned}
$$

離散潜在変数の変分推論

- ここで、$\boldsymbol{\beta} = \mathrm{diag}\left(\beta_1, \beta_2, \ldots, \beta_D\right)$ であるので

$$\begin{aligned} \ln\left|\boldsymbol{\beta}^{-1}\right| &= \ln\left|\boldsymbol{\beta}\right|^{-1} = -\ln\left|\boldsymbol{\beta}\right| \\ &= -\ln\prod_{j=1}^{D}\beta_j = -\sum_{j=1}^{D}\ln\beta_j \end{aligned} \tag{526}$$

となるほか

$$\left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i\right)^T \boldsymbol{\beta}\left(\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{z}_i\right) = \sum_{j=1}^{D}\beta_j\left(x_{ij} - \boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2 \tag{527}$$

であるから

$$\begin{aligned} &\ln\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i, \boldsymbol{\beta}^{-1}) \\ = &-\frac{1}{2}\sum_j \ln 2\pi + \frac{1}{2}\sum_j \ln\beta_j - \frac{1}{2}\sum_j \beta_j\left(x_{ij} - \boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2 \end{aligned} \tag{528}$$

離散潜在変数の変分推論

$$= \frac{1}{2}\sum_j \left(-\ln 2\pi + \ln \beta_j - \beta_j \left(x_{ij} - \boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2\right) \tag{529}$$

$$= \frac{1}{2}\sum_j \left(\ln \frac{\beta_j}{2\pi} - \beta_j \left(x_{ij} - \boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2\right) \tag{530}$$

$$= \frac{1}{2}\sum_j \left(\ln \frac{\beta_j}{2\pi} - \beta_j \left(x_{ij} - \sum_k W_{jk} z_{ik}\right)^2\right) \tag{531}$$

- 上式の対数を外すと、確率分布の積になっていることが分かる

$$\mathcal{N}(\boldsymbol{x}_i | \boldsymbol{W}\boldsymbol{z}_i, \boldsymbol{\beta}^{-1})$$

$$= \exp\left(\sum_j \left(\frac{1}{2}\ln \frac{\beta_j}{2\pi} - \frac{1}{2}\beta_j \left(x_{ij} - \boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2\right)\right) \tag{532}$$

$$= \prod_j \exp\left(\frac{1}{2}\ln \frac{\beta_j}{2\pi} - \frac{1}{2}\beta_j \left(x_{ij} - \boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2\right) \tag{533}$$

離散潜在変数の変分推論

$$= \prod_j \exp\left(\frac{1}{2}\ln\frac{\beta_j}{2\pi}\right)\exp\left(-\frac{1}{2}\beta_j\left(x_{ij}-\boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2\right) \tag{534}$$

$$= \prod_j \left(\frac{\beta_j}{2\pi}\right)^{\frac{1}{2}}\exp\left(-\frac{1}{2}\beta_j\left(x_{ij}-\boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2\right) \tag{535}$$

$$= \prod_j \left(\frac{1}{(2\pi\beta_j^{-1})^{\frac{1}{2}}}\exp\left(-\frac{1}{2}\beta_j\left(x_{ij}-\boldsymbol{W}_{j:}\boldsymbol{z}_i\right)^2\right)\right) \tag{536}$$

$$= \prod_j \mathcal{N}(x_{ij}|\boldsymbol{W}_{j:}\boldsymbol{z}_i,\beta_j^{-1}) \tag{537}$$

- 精度行列 $\boldsymbol{\beta}$ は対角行列であるから、$\boldsymbol{x}_i$ の各成分は無相関である
- 従って、$\mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i,\boldsymbol{\beta}^{-1})$ は、各成分 x_{ij} についての確率 $\mathcal{N}(x_{ij}|\boldsymbol{W}_{j:}\boldsymbol{z}_i,\beta_j^{-1})$ の積として、記述できる

離散潜在変数の変分推論

- さて、第 2 項は

$$
\begin{aligned}
& \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \ln p(\boldsymbol{X}|\boldsymbol{Z}) \\
= & \sum_{i} \sum_{\boldsymbol{z}_i} q(\boldsymbol{z}_i|\boldsymbol{x}_i) \ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i, \boldsymbol{\beta}^{-1}) && (538) \\
= & \sum_{i} \mathbb{E}_{\boldsymbol{z}_i \sim q(\boldsymbol{z}_i|\boldsymbol{x}_i)} \left[\ln \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{W}\boldsymbol{z}_i, \boldsymbol{\beta}^{-1}) \right] && (539) \\
= & \sum_{i} \mathbb{E}_{\boldsymbol{z}_i} \left[\frac{1}{2} \sum_{j} \left(\ln \frac{\beta_j}{2\pi} - \beta_j \left(x_{ij} - \sum_{k} W_{jk} z_{ik} \right)^2 \right) \right] && (540) \\
= & \frac{1}{2} \sum_{i} \sum_{j} \mathbb{E}_{\boldsymbol{z}_i} \left[\ln \frac{\beta_j}{2\pi} - \beta_j \left(x_{ij} - \sum_{k} W_{jk} z_{ik} \right)^2 \right] && (541) \\
= & \frac{1}{2} \sum_{i} \sum_{j} \left(\ln \frac{\beta_j}{2\pi} - \beta_j \mathbb{E}_{\boldsymbol{z}_i} \left[\left(x_{ij} - \sum_{k} W_{jk} z_{ik} \right)^2 \right] \right) && (542)
\end{aligned}
$$

離散潜在変数の変分推論

ここで

$$\mathbb{E}_{\boldsymbol{z}_i}\left[\left(x_{ij}-\sum_k W_{jk}z_{ik}\right)^2\right]$$

$$= \mathbb{E}_{\boldsymbol{z}_i}\left[x_{ij}^2 - 2x_{ij}\sum_k W_{jk}z_{ik} + \left(\sum_k W_{jk}z_{ik}\right)^2\right] \quad (543)$$

$$= x_{ij}^2 - 2x_{ij}\mathbb{E}_{\boldsymbol{z}_i}\left[\sum_k W_{jk}z_{ik}\right] + \mathbb{E}_{\boldsymbol{z}_i}\left[\left(\sum_k W_{jk}z_{ik}\right)^2\right] \quad (544)$$

各項を順番に計算すると

$$\mathbb{E}_{\boldsymbol{z}_i}\left[\sum_k W_{jk}z_{ik}\right]$$

$$= \sum_k W_{jk} \mathbb{E}_{\boldsymbol{z}_i} [z_{ik}] \tag{545}$$

$$= \sum_k W_{jk} \mathbb{E}_{z_{ik}} [z_{ik}] \tag{546}$$

$$= \sum_k W_{jk} \left(q(z_{ik} = 1 | \boldsymbol{x}_i) \cdot 1 + q(z_{ik} = 0 | \boldsymbol{x}_i) \cdot 0 \right) \tag{547}$$

$$= \sum_k W_{jk} \widehat{z_{ik}} \tag{548}$$

また

$$\mathbb{E}_{\boldsymbol{z}_i} \left[\left(\sum_k W_{jk} z_{ik} \right)^2 \right]$$

$$= \mathbb{E}_{\boldsymbol{z}_i} \left[\sum_k \sum_l W_{jk} W_{jl} z_{ik} z_{il} \right] \tag{549}$$

離散潜在変数の変分推論

$$= \mathbb{E}_{\boldsymbol{z}_i}\left[\sum_k \left(W_{jk}^2 z_{ik}^2 + \sum_{l\neq k}(W_{jk}W_{jl}z_{ik}z_{il})\right)\right] \tag{550}$$

$$= \sum_k \left(W_{jk}^2 \mathbb{E}_{\boldsymbol{z}_i}\left[z_{ik}^2\right] + \sum_{l\neq k} W_{jk}W_{jl}\mathbb{E}_{\boldsymbol{z}_i}\left[z_{ik}z_{il}\right]\right) \tag{551}$$

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}_i}\left[z_{ik}^2\right] \\
=\ &\mathbb{E}_{z_{ik}}\left[z_{ik}^2\right] && (552)\\
=\ &q(z_{ik}=1|\boldsymbol{x}_i)\cdot 1^2 + q(z_{ik}=0|\boldsymbol{x}_i)\cdot 0^2 && (553)\\
=\ &\widehat{z_{ik}} && (554)
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z}_i}\left[z_{ik}z_{il}\right] \\
=\ &\mathbb{E}_{\boldsymbol{z}_i}\left[z_{ik}\right]\mathbb{E}_{\boldsymbol{z}_i}\left[z_{il}\right] \quad (\because z_{ik} \text{ と } z_{il} \text{ は独立であるため}) && (555)
\end{aligned}$$

$$= \mathbb{E}_{z_{ik}}[z_{ik}]\mathbb{E}_{z_{il}}[z_{il}] \tag{556}$$
$$= \widehat{z_{ik}}\widehat{z_{il}} \tag{557}$$

従って

$$\mathbb{E}_{\boldsymbol{z}_i}\left[\left(x_{ij} - \sum_k W_{jk}z_{ik}\right)^2\right]$$
$$= x_{ij}^2 - 2x_{ij}\mathbb{E}_{\boldsymbol{z}_i}\left[\sum_k W_{jk}z_{ik}\right] + \mathbb{E}_{\boldsymbol{z}_i}\left[\left(\sum_k W_{jk}z_{ik}\right)^2\right] \tag{558}$$
$$= x_{ij}^2 - 2x_{ij}\sum_k W_{jk}\widehat{z_{ik}} +$$
$$\sum_k\left(W_{jk}^2\widehat{z_{ik}} + \sum_{l\neq k} W_{jk}W_{jl}\widehat{z_{ik}}\widehat{z_{il}}\right) \tag{559}$$

離散潜在変数の変分推論

これより

$$\sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \ln p(\boldsymbol{X}|\boldsymbol{Z})$$

$$= \frac{1}{2} \sum_i \sum_j \left(\ln \frac{\beta_j}{2\pi} - \beta_j \mathbb{E}_{\boldsymbol{z}_i} \left[\left(x_{ij} - \sum_k W_{jk} z_{ik} \right)^2 \right] \right) \quad (560)$$

$$= \frac{1}{2} \sum_i \sum_j \left(\ln \frac{\beta_j}{2\pi} - \beta_j \left(x_{ij}^2 - 2x_{ij} \sum_k W_{jk} \widehat{z_{ik}} + \right. \right.$$

$$\left. \left. \sum_k \left(W_{jk}^2 \widehat{z_{ik}} + \sum_{l \neq k} W_{jk} W_{jl} \widehat{z_{ik}} \widehat{z_{il}} \right) \right) \right) \quad (561)$$

離散潜在変数の変分推論

- よって、エビデンス下界 $\mathcal{L}(q)$ は

$$
\begin{aligned}
& \mathcal{L}(q) \\
= & \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \left(\ln p(\boldsymbol{Z}) - \ln q(\boldsymbol{Z}|\boldsymbol{X}) \right) + \\
& \quad \sum_{\boldsymbol{Z}} q(\boldsymbol{Z}|\boldsymbol{X}) \ln p(\boldsymbol{X}|\boldsymbol{Z}) \qquad (562) \\
= & \sum_{i} \sum_{k} \{ \widehat{z_{ik}} \left(\ln \sigma(b_{ik}) - \ln \widehat{z_{ik}} \right) + \\
& \quad (1 - \widehat{z_{ik}}) \left(\ln \sigma(-b_{ik}) - \ln (1 - \widehat{z_{ik}}) \right) \} + \\
& \quad \frac{1}{2} \sum_{i} \sum_{j} \left(\ln \frac{\beta_j}{2\pi} - \beta_j \left(x_{ij}^2 - 2 x_{ij} \sum_{k} W_{jk} \widehat{z_{ik}} + \right. \right. \\
& \quad \left. \left. \sum_{k} \left(W_{jk}^2 \widehat{z_{ik}} + \sum_{l \neq k} W_{jk} W_{jl} \widehat{z_{ik}} \widehat{z_{il}} \right) \right) \right) \qquad (563)
\end{aligned}
$$

離散潜在変数の変分推論

- この式にはあまり美的魅力がない (Somewhat unappealing aesthetically) が、事後分布よりは計算しやすい
- 事後分布 $\ln p(\boldsymbol{Z}|\boldsymbol{X})$ を最大化する代わりに、上記の下界 $\mathcal{L}(q)$ を q について最大化することができる

- データ $\boldsymbol{x}_i$ に対するパラメータ $\{\widehat{z_{i1}}, \ldots, \widehat{z_{iK}}\}$ をまとめて、ベクトル $\widehat{\boldsymbol{z}}_i$ として記述する
- パラメータ $\widehat{\boldsymbol{z}}_i = [\widehat{z_{i1}}, \ldots, \widehat{z_{iK}}]^T$ は、データ $\boldsymbol{x}_i$ に対応する潜在変数 $\boldsymbol{z}$ の各要素が、1 となる確率を集めたベクトルである

- $\widehat{z_{ik}} = q(z_{ik} = 1|\boldsymbol{x}_i)$ と定義されていることに注意

- よってベクトル $\widehat{\boldsymbol{z}}_i$ は、データ $\boldsymbol{x}_i$ に対応する二値のスパース符号である

- 勾配上昇法を利用しない理由
  - データ $\boldsymbol{X}$ と、潜在変数 $\boldsymbol{Z}$ についての勾配上昇法を用いれば、学習することが可能

離散潜在変数の変分推論

- 但し、その方法では、各 $\boldsymbol{x}_i$ について平均場パラメータ $\widehat{\boldsymbol{z}_i}$ を保管する必要がある
- 各事例について、動的に更新されるベクトルが必要であるため、そのアルゴリズムを、数十億もの事例に対して適用することは困難である

- また、収束するまで繰り返し計算を行うため、データ $\boldsymbol{x}$ から、パラメータ $\widehat{\boldsymbol{z}_i}$ を素早く抽出することができない
- 現実にデプロイされるときは、$\widehat{\boldsymbol{z}_i}$ をリアルタイムで計算できなければならない

- 不動点方程式による平均場パラメータ $\widehat{\boldsymbol{z}_i}$ の推定
  - 勾配上昇法の代わりに、不動点方程式を使ってパラメータ $\widehat{z_{ik}}$ を素早く推定できる

  - $\nabla_{\widehat{\boldsymbol{z}_i}} \mathcal{L}(q) = 0$ をみたす、$\widehat{\boldsymbol{z}_i}$ の極大値を見つけ出す

離散潜在変数の変分推論

- $\widehat{z}_i$ の全ての成分について同時に解くことはできないので、各成分 $\widehat{z_{ik}}$ について繰り返し解く
- 即ち、各パラメータ $\widehat{z_{ik}}$ について、下界 $\mathcal{L}(q)$ を順番に最適化する手続きを、$\mathcal{L}(q)$ の収束基準を満たすまで繰り返す

$$\frac{\partial}{\partial \widehat{z_{ik}}} \mathcal{L}(q) = 0 \tag{564}$$

- 平均場不動点方程式を導くためには、$\mathcal{L}(q)$ を $\widehat{z_{ik}}$ で微分する必要がある

$$\begin{aligned} & \frac{\partial}{\partial \widehat{z_{ik}}} \mathcal{L}(q) \\ = & \frac{\partial}{\partial \widehat{z_{ik}}} \sum_i \sum_k \{\widehat{z_{ik}} \left(\ln \sigma(b_{ik}) - \ln \widehat{z_{ik}}\right) + \\ & \qquad (1 - \widehat{z_{ik}}) \left(\ln \sigma(-b_{ik}) - \ln \left(1 - \widehat{z_{ik}}\right)\right)\} + \end{aligned}$$

離散潜在変数の変分推論

$$\frac{\partial}{\partial \widehat{z_{ik}}} \frac{1}{2} \sum_i \sum_j \left(\ln \frac{\beta_j}{2\pi} - \beta_j \left(x_{ij}^2 - 2x_{ij} \sum_k W_{jk} \widehat{z_{ik}} + \right. \right.$$
$$\left. \left. \sum_k \left(W_{jk}^2 \widehat{z_{ik}} + \sum_{l \neq k} W_{jk} W_{jl} \widehat{z_{ik}} \widehat{z_{il}} \right) \right) \right) \tag{565}$$

- 前半部分は

$$\frac{\partial}{\partial \widehat{z_{ik}}} \sum_i \sum_k \{ \widehat{z_{ik}} \left(\ln \sigma(b_{ik}) - \ln \widehat{z_{ik}} \right) +$$
$$(1 - \widehat{z_{ik}}) \left(\ln \sigma(-b_{ik}) - \ln (1 - \widehat{z_{ik}}) \right) \} \tag{566}$$
$$= \frac{\partial}{\partial \widehat{z_{ik}}} \{ \widehat{z_{ik}} \left(\ln \sigma(b_{ik}) - \ln \widehat{z_{ik}} \right) +$$
$$(1 - \widehat{z_{ik}}) \left(\ln \sigma(-b_{ik}) - \ln (1 - \widehat{z_{ik}}) \right) \} \tag{567}$$
$$= \frac{\partial}{\partial \widehat{z_{ik}}} \widehat{z_{ik}} \left(\ln \sigma(b_{ik}) - \ln \widehat{z_{ik}} \right) +$$

離散潜在変数の変分推論

$$\frac{\partial}{\partial \widehat{z_{ik}}} \left(1 - \widehat{z_{ik}}\right) \left(\ln \sigma(-b_{ik}) - \ln\left(1 - \widehat{z_{ik}}\right)\right) \tag{568}$$

$$\begin{aligned} = \quad & \left(\ln \sigma(b_{ik}) - \ln \widehat{z_{ik}}\right) + \widehat{z_{ik}} \left(-\frac{1}{\widehat{z_{ik}}}\right) + \\ & \left(-\left(\ln \sigma(-b_{ik}) - \ln\left(1 - \widehat{z_{ik}}\right)\right)\right) + \\ & \left(1 - \widehat{z_{ik}}\right) \frac{1}{1 - \widehat{z_{ik}}} \end{aligned} \tag{569}$$

$$= \quad \ln \sigma(b_{ik}) - \ln \widehat{z_{ik}} - 1 - \ln \sigma(-b_{ik}) + \ln(1 - \widehat{z_{ik}}) + 1 \tag{570}$$

$$= \quad \ln \sigma(b_{ik}) - \ln \widehat{z_{ik}} - \ln \sigma(-b_{ik}) + \ln(1 - \widehat{z_{ik}}) \tag{571}$$

$$= \quad -\zeta(-b_{ik}) - \ln \widehat{z_{ik}} + \zeta(b_{ik}) + \ln(1 - \widehat{z_{ik}}) \tag{572}$$

$$= \quad (\zeta(b_{ik}) - \zeta(-b_{ik})) - \ln \widehat{z_{ik}} + \ln(1 - \widehat{z_{ik}}) \tag{573}$$

$$= \quad b_{ik} - \ln \widehat{z_{ik}} + \ln(1 - \widehat{z_{ik}}) \tag{574}$$

離散潜在変数の変分推論

- ここで、シグモイド関数 $\sigma(\cdot)$ と、ソフトプラス関数 $\zeta(\cdot)$ に関する、以下の公式を用いた

$$\ln \sigma(x) = -\zeta(-x) \tag{575}$$

$$\zeta(x) - \zeta(-x) = x \tag{576}$$

- 後半部分は

$$\frac{\partial}{\partial \widehat{z_{ik}}} \frac{1}{2} \sum_i \sum_j \left(\ln \frac{\beta_j}{2\pi} - \beta_j \left(x_{ij}^2 - 2x_{ij} \sum_k W_{jk} \widehat{z_{ik}} + \sum_k \left(W_{jk}^2 \widehat{z_{ik}} + \sum_{l \neq k} W_{jk} W_{jl} \widehat{z_{ik}} \widehat{z_{il}} \right) \right) \right) \tag{577}$$

$$= \frac{1}{2} \sum_j \beta_j \frac{\partial}{\partial \widehat{z_{ik}}} \left(2x_{ij} \sum_k W_{jk} \widehat{z_{ik}} - \right.$$

離散潜在変数の変分推論

$$\sum_k \left(W_{jk}^2 \widehat{z_{ik}} + \sum_{l \neq k} W_{jk} W_{jl} \widehat{z_{ik}} \widehat{z_{il}} \right) \Bigg) \tag{578}$$

$$= \frac{1}{2} \sum_j \beta_j \frac{\partial}{\partial \widehat{z_{ik}}} \Bigg(2 x_{ij} \sum_k W_{jk} \widehat{z_{ik}} -$$

$$\left(W_{jk}^2 \widehat{z_{ik}} + \sum_{l \neq k} W_{jk} W_{jl} \widehat{z_{ik}} \widehat{z_{il}} \right) -$$

$$\sum_{m \neq k} \left(W_{jm}^2 \widehat{z_{im}} + \sum_{l \neq m} W_{jm} W_{jl} \widehat{z_{im}} \widehat{z_{il}} \right) \Bigg) \tag{579}$$

$$= \frac{1}{2} \sum_j \beta_j \left(2 x_{ij} W_{jk} - W_{jk}^2 - \right.$$

$$\left. \sum_{l \neq k} W_{jk} W_{jl} \widehat{z_{il}} - \sum_{m \neq k} W_{jm} W_{jk} \widehat{z_{im}} \right) \tag{580}$$

離散潜在変数の変分推論

$$= \frac{1}{2}\sum_j \beta_j \left(2x_{ij}W_{jk} - W_{jk}^2 - 2\sum_{l\neq k} W_{jk}W_{jl}\widehat{z_{il}} \right) \tag{581}$$

$$= \sum_j \beta_j \left(x_{ij}W_{jk} - \frac{1}{2}W_{jk}^2 - \sum_{l\neq k} W_{jk}W_{jl}\widehat{z_{il}} \right) \tag{582}$$

$$= \sum_j x_{ij}\beta_j W_{jk} - \frac{1}{2}\sum_j W_{jk}\beta_j W_{jk} - \sum_j \beta_j \sum_{l\neq k} W_{jk}W_{jl}\widehat{z_{il}} \tag{583}$$

$$= \sum_j x_{ij}\beta_j W_{jk} - \frac{1}{2}\sum_j W_{jk}\beta_j W_{jk} - \sum_{l\neq k} \left(\sum_j W_{jl}\beta_j W_{jk} \right) \widehat{z_{il}} \tag{584}$$

離散潜在変数の変分推論

$$= \quad \boldsymbol{x}_i^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \frac{1}{2} \boldsymbol{W}_{:k}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \sum_{l \neq k} \boldsymbol{W}_{:l}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} \widehat{z_{il}} \tag{585}$$

- これより、$\mathcal{L}(q)$ の $\widehat{z_{ik}}$ による微分は次のようになる

$$\begin{aligned} & \frac{\partial}{\partial \widehat{z_{ik}}} \mathcal{L}(q) \\ = \quad & b_{ik} - \ln \widehat{z_{ik}} + \ln(1 - \widehat{z_{ik}}) + \\ & \boldsymbol{x}_i^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \frac{1}{2} \boldsymbol{W}_{:k}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \sum_{l \neq k} \boldsymbol{W}_{:l}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} \widehat{z_{il}} \end{aligned} \tag{586}$$

離散潜在変数の変分推論

- これを 0 と等置して、$\widehat{z_{ik}}$ について解くと次のようになる

$$\ln \widehat{z_{ik}} - \ln(1 - \widehat{z_{ik}}) = b_{ik} + \boldsymbol{x}_i^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \frac{1}{2} \boldsymbol{W}_{:k}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \sum_{l \neq k} \boldsymbol{W}_{:l}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} \widehat{z_{il}} \tag{587}$$

右辺を A とおけば

$$\ln \widehat{z_{ik}} - \ln(1 - \widehat{z_{ik}}) = A \tag{588}$$

$$\Rightarrow \quad \ln \frac{\widehat{z_{ik}}}{1 - \widehat{z_{ik}}} = A$$

$$\Rightarrow \quad \frac{\widehat{z_{ik}}}{1 - \widehat{z_{ik}}} = \exp A$$

$$\Rightarrow \quad \frac{1}{1 - \widehat{z_{ik}}} - 1 = \exp A$$

$$\Rightarrow \quad 1 - \widehat{z_{ik}} = \frac{1}{1 + \exp A}$$

離散潜在変数の変分推論

$$\Rightarrow \quad \widehat{z_{ik}} = 1 - \frac{1}{1 + \exp A}$$

$$\Rightarrow \quad \widehat{z_{ik}} = \frac{\exp A}{1 + \exp A}$$

$$\Rightarrow \quad \widehat{z_{ik}} = \frac{1}{1 + \exp(-A)} \tag{589}$$

$$\Rightarrow \quad \widehat{z_{ik}} = \sigma(A) \tag{590}$$

従って、不動点方程式は

$$\widehat{z_{ik}} = \sigma\left(b_{ik} + \boldsymbol{x}_i^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \frac{1}{2} \boldsymbol{W}_{:k}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \sum_{l \neq k} \boldsymbol{W}_{:l}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} \widehat{z_{il}}\right) \tag{591}$$

- 不動点方程式の観察

離散潜在変数の変分推論

- 不動点方程式は次で表された

$$\widehat{z_{ik}} = \sigma\left(b_{ik} + \boldsymbol{x}_i^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \frac{1}{2}\boldsymbol{W}_{:k}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} - \sum_{l \neq k} \boldsymbol{W}_{:l}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} \widehat{z_{il}}\right) \quad (592)$$

- 第 2 項 $\boldsymbol{x}_i^T \boldsymbol{\beta} \boldsymbol{W}_{:k}$ は、潜在変数のユニット k に対する入力
- 第 3 項 $-\frac{1}{2}\boldsymbol{W}_{:k}^T \boldsymbol{\beta} \boldsymbol{W}_{:k}$ は、隠れユニット k から自身への入力
- 第 4 項 $-\sum_{l \neq k} \boldsymbol{W}_{:l}^T \boldsymbol{\beta} \boldsymbol{W}_{:k} \widehat{z_{il}}$ は、他の隠れユニット $l \neq k$ から隠れユニット k への入力

- これより、平均場不動点方程式は、回帰結合型ニューラルネットワーク (RNN) との関係があることが分かる

- 隠れユニット k と l は、それらの重みベクトル $\boldsymbol{W}_{:l}$ と $\boldsymbol{W}_{:k}$ が互いに同調するとき (似たような重みを持っているとき) に、互いに抑制し合う

離散潜在変数の変分推論

- 即ち、2 つの隠れユニット k, l が、共に入力を説明するとき (入力から同じような表現を抽出するとき)、入力を最もよく説明するユニットのみがアクティブになる (強く活性化される)
- これはユニット間の競合の一形態である

- 従って、実際には多峰性の事後分布かもしれないが、そのうちの 1 つのみが選択される (図 14 の (b) と (c) 参照)

- 不動点方程式を更に以下のように変形する

$$\widehat{z_{ik}} = \sigma\left(b_{ik} + \left(\boldsymbol{x}_i - \sum_{l \neq k} \boldsymbol{W}_{:l}\widehat{z_{il}}\right)^T \boldsymbol{\beta}\boldsymbol{W}_{:k} - \frac{1}{2}\boldsymbol{W}_{:k}^T\boldsymbol{\beta}\boldsymbol{W}_{:k}\right) \tag{593}$$

- これより、ユニット k への入力は、$\boldsymbol{x}_i$ ではなく $\boldsymbol{x}_i - \sum_{l \neq k} \boldsymbol{W}_{:l}\widehat{z_{il}}$ であるとみなせる

離散潜在変数の変分推論

- ユニット k への入力は、他の全てのユニットによる $\boldsymbol{x}_i$ の再構成と、実際の入力 $\boldsymbol{x}_i$ との誤差である
- ユニット k は、この残差誤差を符号化していると分かるので、スパース符号化は、反復自己符号化器とみなせる

- スパース符号化では、入力 $\boldsymbol{x}_i$ の符号化 ($\widehat{z_{ik}}$ の計算) と、復号 ($\sum_{l \neq k} \boldsymbol{W}_{:l} \widehat{z_{il}}$ の計算) を繰り返す
- この反復のたびに、再構成の誤差を修正していく

- ダンピング
  - 1 つのユニットの更新則を (不動点方程式として) 導出した
  - 複数のユニットを同時に更新することは、二値スパース符号化モデルでは通常できない
  - 但し、ダンピングという発見的手法を使えば可能になる

  - 各要素 $\widehat{z_{ik}}$ についての最適値を計算し、その値の変化の方向に、他の要素 $\widehat{z_{il}}$ を小さいステップで動かす

離散潜在変数の変分推論

- 下界 $\mathcal{L}(q)$ が増加することはもはや保証されないが、多くの場合はうまくいく

離散潜在変数の変分推論のまとめ

- ここまでの話の流れ
  1. 離散潜在変数における変分推論の具体例として、二値スパース符号化モデルをみた
  2. 事後分布 $p(\boldsymbol{Z}|\boldsymbol{X})$ が複雑になるので、最尤推定 (EM アルゴリズム) が使えない
  3. 代わりに、別の分布 $q(\boldsymbol{Z})$ を使って事後分布を近似することにした
  4. エビデンス下界 $\mathcal{L}(q)$ を苦労して求めた
  5. 更に、下界 $\mathcal{L}(q)$ を (各パラメータについて) 最大化するための、不動点方程式を導出した
  6. 不動点方程式を観察し、回帰結合型ニューラルネットワークや、自己符号化器との関係を考えた

- これからの話の流れ

離散潜在変数の変分推論のまとめ

- 連続潜在変数に対する変分推論を、簡単な確率モデルを使って、試してみよう

目次

連続潜在変数の変分推論

- 連続潜在変数をもつ変分推論の概要
  - 平均場近似を行う場合は、以下の式によって最適な因子 $q_j^*(\boldsymbol{Z}_j|\boldsymbol{X})$ が得られる

$$\ln q_j^*(\boldsymbol{Z}_j|\boldsymbol{X}) = \mathbb{E}_{i\neq j}\left[\ln p(\boldsymbol{X},\boldsymbol{Z})\right] + \text{Const.} \tag{594}$$

$$\mathbb{E}_{i\neq j}\left[\ln p(\boldsymbol{X},\boldsymbol{Z})\right] = \int \ln p(\boldsymbol{X},\boldsymbol{Z}) \prod_{i\neq j} q_i(\boldsymbol{Z}_i|\boldsymbol{X}) d\boldsymbol{Z}_i \tag{595}$$

$$q_j^*(\boldsymbol{Z}_j|\boldsymbol{X}) = \frac{\exp\left(\mathbb{E}_{i\neq j}\left[\ln p(\boldsymbol{X},\boldsymbol{Z})\right]\right)}{\int \exp\left(\mathbb{E}_{i\neq j}\left[\ln p(\boldsymbol{X},\boldsymbol{Z})\right]\right) d\boldsymbol{Z}_j} \tag{596}$$

  - 上式は、下界 $\mathcal{L}(q)$ を最大化する q であり、従って連続潜在変数の場合の不動点方程式とみなせる
  - 各因子 $q_j(\boldsymbol{Z}_j|\boldsymbol{X})$ を、上式を用いて順番に更新していく ($i \neq j$ である全ての q_i を固定した状態で、各 q_j について下界を最適化する)

連続潜在変数の変分推論

- このステップを、下界 $\mathcal{L}(q)$ が収束するまで繰り返し行う (座標降下法)
- 不動点方程式は、下界が最適な値に収束するかどうかには関係なく、q_j の最適解が取る関数形を提供してくれる

- 扱う確率モデルの表現
  - ここでは次のような確率モデルを対象として、変分推論を扱う

$$p(\boldsymbol{z}_i) = \mathcal{N}(\boldsymbol{z}_i|0, \boldsymbol{I}) \tag{597}$$
$$p(x_i|\boldsymbol{z}_i) = \mathcal{N}(x_i|\boldsymbol{w}^T\boldsymbol{z}_i, 1) \tag{598}$$

  - 1 次元のデータ $x \in \mathbb{R}$ に対して、2 次元の潜在変数 $\boldsymbol{z} \in \mathbb{R}^2$ が存在する
  - 同時分布 $p(x_i, \boldsymbol{z}_i) = p(x_i|\boldsymbol{z}_i)p(\boldsymbol{z}_i)$ を $\boldsymbol{z}_i$ で積分消去すれば、x_i についての単なるガウス分布となる

- 真の事後分布 $p(\boldsymbol{z}_i|x_i)$ の計算

連続潜在変数の変分推論

- 正規化定数を無視して次のように計算できる

$$
\begin{aligned}
& p(\boldsymbol{z}_i|x_i) \\
\propto\ & p(\boldsymbol{z}_i|x_i)p(x_i) \\
=\ & p(x_i, \boldsymbol{z}_i) \\
=\ & p(\boldsymbol{z}_i)p(x_i|\boldsymbol{z}_i) \\
=\ & \mathcal{N}(\boldsymbol{z}_i|0, \boldsymbol{I})\mathcal{N}(x_i|\boldsymbol{w}^T\boldsymbol{z}_i, 1) \\
\propto\ & \exp\left(-\frac{1}{2}\boldsymbol{z}_i^T\boldsymbol{z}_i\right)\exp\left(-\frac{1}{2}\left(x_i - \boldsymbol{w}^T\boldsymbol{z}_i\right)^T\left(x_i - \boldsymbol{w}^T\boldsymbol{z}_i\right)\right) \\
=\ & \exp\left(-\frac{1}{2}\left(z_{i1}^2 + z_{i2}^2\right)\right) \\
& \qquad \exp\left(-\frac{1}{2}\left(x_i - w_1 z_{i1} - w_2 z_{i2}\right)^T\left(x_i - w_1 z_{i1} - w_2 z_{i2}\right)\right) \\
=\ & \exp\left(-\frac{1}{2}\left(z_{i1}^2 + z_{i2}^2 + x_i^2 + w_1^2 z_{i1}^2 + w_2^2 z_{i2}^2 -\right.\right.
\end{aligned}
$$

連続潜在変数の変分推論

$$2x_i w_1 z_{i1} - 2x_i w_2 z_{i2} + 2w_1 w_2 z_{i1} z_{i2}) \Bigg) \tag{599}$$

- 正規化項を C とすれば次のように書ける

$$\begin{aligned} & p(\boldsymbol{z}_i | x_i) \\ = \quad & C \exp \Bigg(-\frac{1}{2} \left(z_{i1}^2 + z_{i2}^2 + x_i^2 + w_1^2 z_{i1}^2 + w_2^2 z_{i2}^2 - \right. \\ & \left. 2x_i w_1 z_{i1} - 2x_i w_2 z_{i2} + 2w_1 w_2 z_{i1} z_{i2} \right) \Bigg) \end{aligned} \tag{600}$$

- z_{i1} と z_{i2} を乗算する項が存在する
- 従って、真の事後分布は、z_{i1} と z_{i2} のみの因子には分解できないことが分かる

- 平均場近似の計算

連続潜在変数の変分推論

- 平均場近似を次のように表現する

$$q(\boldsymbol{z}_i|x_i) = q_1(z_{i1}|x_i)q_2(z_{i2}|x_i) \tag{601}$$

- q_2 を固定した状態で、最適な $q_1^*(z_{i1}|x_i)$ を求める

$$\begin{aligned}
& \ln q_1^*(z_{i1}|x_i) \\
= \; & \mathbb{E}_{z_{i2}\sim q_2(z_{i2}|x_i)}\left[\ln p(\boldsymbol{z}_i, x_i)\right] + \text{Const.} \qquad (602) \\
= \; & \mathbb{E}_{z_{i2}}\left[\ln\left(C\exp\left(-\frac{1}{2}\left(z_{i1}^2 + z_{i2}^2 + x_i^2 + w_1^2 z_{i1}^2 + w_2^2 z_{i2}^2 - \right.\right.\right.\right. \\
& \left.\left.\left.\left. 2x_i w_1 z_{i1} - 2x_i w_2 z_{i2} + 2w_1 w_2 z_{i1} z_{i2}\right)\right)\right)\right] + \text{Const.} \\
= \; & \mathbb{E}_{z_{i2}}\left[-\frac{1}{2}\left(z_{i1}^2 + z_{i2}^2 + x_i^2 + w_1^2 z_{i1}^2 + w_2^2 z_{i2}^2 - \right.\right. \\
& \left.\left. 2x_i w_1 z_{i1} - 2x_i w_2 z_{i2} + 2w_1 w_2 z_{i1} z_{i2}\right)\right] + \text{Const.} \qquad (603)
\end{aligned}$$

連続潜在変数の変分推論

- $z_{i2} \sim q_2(z_{i2}|x_i)$ による期待値を取っている
- $q_2(z_{i2}|x_i)$ から得る必要があるのは、結局 $\mathbb{E}_{z_{i2}}[z_{i2}]$ と、$\mathbb{E}_{z_{i2}}[z_{i2}^2]$ の 2 つだけである
- $\langle z_{i2} \rangle = \mathbb{E}_{z_{i2}}[z_{i2}]$、$\langle z_{i2}^2 \rangle = \mathbb{E}_{z_{i2}}[z_{i2}^2]$ と書くことにする

- このとき次式が得られる

$$
\begin{aligned}
& \ln q_1^*(z_{i1}|x_i) \\
= \quad & -\frac{1}{2}\left(z_{i1}^2 + \langle z_{i2}^2 \rangle + x_i^2 + w_1^2 z_{i1}^2 + w_2^2 \langle z_{i2}^2 \rangle - \right. \\
& \quad 2x_i w_1 z_{i1} - 2x_i w_2 \langle z_{i2} \rangle + 2w_1 w_2 z_{i1} \langle z_{i2} \rangle) + \\
& \quad \text{Const.}
\end{aligned} \tag{604}
$$

連続潜在変数の変分推論

- これより、最適な $q_1^*(z_{i1}|x_i)$ はガウス分布の形であると分かる

$$
\begin{aligned}
& q_1^*(z_{i1}|x_i) \\
= \quad & C \exp\left(-\frac{1}{2} \left(z_{i1}^2 + \left\langle z_{i2}^2 \right\rangle + x_i^2 + w_1^2 z_{i1}^2 + w_2^2 \left\langle z_{i2}^2 \right\rangle - \right. \right. \\
& \left. \left. 2x_i w_1 z_{i1} - 2x_i w_2 \left\langle z_{i2} \right\rangle + 2w_1 w_2 z_{i1} \left\langle z_{i2} \right\rangle \right) \right) \\
\propto \quad & \exp\left(-\frac{1}{2} \left(z_{i1}^2 + \left\langle z_{i2}^2 \right\rangle + x_i^2 + w_1^2 z_{i1}^2 + w_2^2 \left\langle z_{i2}^2 \right\rangle - \right. \right. \\
& \left. \left. 2x_i w_1 z_{i1} - 2x_i w_2 \left\langle z_{i2} \right\rangle + 2w_1 w_2 z_{i1} \left\langle z_{i2} \right\rangle \right) \right) \qquad (605)
\end{aligned}
$$

- 対称性から、最適な $q_2^*(z_{i2}|x_i)$ もガウス分布であることが分かる
- ガウス分布同士の積もガウス分布になるので、結局 $q(\boldsymbol{z}_i|x_i) = q_1(z_{i1}|x_i)q_2(z_{i2}|x_i)$ はガウス分布である

連続潜在変数の変分推論

- $q(\boldsymbol{z}_i|x_i)$ が 2 つの因子に分解できるとは仮定したが、各因子の関数形については全く仮定していないことに注意
- ガウス分布は、下界 $\mathcal{L}(q)$ を q について変分最適化する過程で、自然に出現した

連続潜在変数の変分推論のまとめ

- ここまでの話の流れ
  1. 連続潜在変数における変分推論の例として、簡単な確率モデルを扱った
  2. 分布 $q(\boldsymbol{Z}|\boldsymbol{X})$ を、$\prod_i q_i(\boldsymbol{Z}_i|\boldsymbol{X})$ のように因数分解できるという仮定 (平均場近似) のみを置いた
  3. 各因子 $q_i(\boldsymbol{Z}_i|\boldsymbol{X})$ の関数形については全く仮定を置かなかった
  4. 変分推論によって、最適解となる q_i の関数形が自然に導出できた
- これからの話の流れ
  - 変分推論と、正則化との関係をみてみよう

目次

正則化との関わり

- 正則化との関係
  - 変分下界 $\mathcal{L}(q)$ を次のように分解してみよう

$$
\begin{aligned}
& \mathcal{L}(q) \\
= & \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \\
= & \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}|\boldsymbol{Z})p(\boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \\
= & \int q(\boldsymbol{Z}) \left(\ln p(\boldsymbol{X}|\boldsymbol{Z}) + \ln \frac{p(\boldsymbol{Z})}{q(\boldsymbol{Z})} \right) d\boldsymbol{Z} \\
= & \int q(\boldsymbol{Z}) \ln p(\boldsymbol{X}|\boldsymbol{Z}) d\boldsymbol{Z} + \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \\
= & \int q(\boldsymbol{Z}) \ln p(\boldsymbol{X}|\boldsymbol{Z}) d\boldsymbol{Z} - \mathrm{KL}(q||p) \qquad (606)
\end{aligned}
$$

正則化との関わり

- 下界 $\mathcal{L}(q)$ を最大化するとき、第 1 項を最大化し、第 2 項 $\mathrm{KL}(q||p)$ を最小化する必要がある

- 第 1 項は、観測データの対数尤度 $\ln p(\boldsymbol{X}|\boldsymbol{Z})$ の、近似事後分布 $q(\boldsymbol{Z})$ についての期待値である
- この項の最大化は、最尤推定の考え方と似ている
- $q(\boldsymbol{Z})$ を、$\ln(\boldsymbol{X}|\boldsymbol{Z})$ が最大となる $\boldsymbol{Z}^*$ でのみ非零となる、デルタ関数 $\delta(\boldsymbol{Z}-\boldsymbol{Z}^*)$ とすれば、最大化できる

- 第 2 項は、近似事後分布 $q(\boldsymbol{Z})$ と、潜在変数に関する事前分布 $p(\boldsymbol{Z})$ との KL ダイバージェンス
- これを小さくしようとすることは、近似分布 $q(\boldsymbol{Z})$ が事前分布 $p(\boldsymbol{Z})$ から、大きく離れることを防ぐことにあたる

- 第 1 項で、近似分布が際限なく尖ろうとすることに対して、第 2 項では、予め定められた事前分布 $p(\boldsymbol{Z})$ を通じて、ペナルティを与える

正則化との関わり

- 第 2 項は、MAP 推定において、パラメータの事前分布 $p(\boldsymbol{Z})$ を導入することに相当する
- これより、第 2 項は正則化項として作用していることが分かる

正則化との関わりのまとめ

- ここまでの話の流れ
  1. 変分下界 $\mathcal{L}(q)$ の式を変形し、2 つの項に分解した
  2. 最尤推定に相当する項と、パラメータに関する事前分布の導入による正則化に相当する項とに、分かれていることを確認した

- これからの話の流れ
  - 教科書で触れられている雑多な話題を扱う
  - 近似推論が、推論アルゴリズムの精度に影響することについて考える
  - 学習による近似推論の手法を考える

目次

学習と推論の相互作用

- 近似推論が精度に与える影響
  - 近似推論では、以下の下界 $\mathcal{L}(q)$ を q について最適化する

$$\ln p(\boldsymbol{X}) = \mathcal{L}(q) + \mathrm{KL}(q||p) \tag{607}$$

$$\mathcal{L}(q) = \int q(\boldsymbol{Z}) \ln \frac{p(\boldsymbol{X}, \boldsymbol{Z})}{q(\boldsymbol{Z})} d\boldsymbol{Z} \tag{608}$$

  - 下界 $\mathcal{L}(q)$ を次のように分解する

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\boldsymbol{Z}) \ln p(\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{Z} - \int q(\boldsymbol{Z}) \ln q(\boldsymbol{Z}) d\boldsymbol{Z} && (609)\\
&= \int q(\boldsymbol{Z}) \ln p(\boldsymbol{X}, \boldsymbol{Z}) d\boldsymbol{Z} + H[q] && (610)\\
&= \mathbb{E}_{\boldsymbol{Z} \sim q(\boldsymbol{Z})} \left[\ln p(\boldsymbol{X}, \boldsymbol{Z}) \right] + H[q] && (611)\\
&= \mathbb{E}_{\boldsymbol{Z} \sim q(\boldsymbol{Z})} \left[\ln p(\boldsymbol{X}|\boldsymbol{Z}) p(\boldsymbol{Z}) \right] + H[q] && (612)
\end{aligned}$$

  - 分布に含まれるパラメータを明示的に記述する

学習と推論の相互作用

- このとき、モデリングされる関数は $p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\theta})$、$p(\boldsymbol{Z}|\boldsymbol{\theta})$ と表せるので、$p(\boldsymbol{X}|\boldsymbol{Z}) = p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta}) = p(\boldsymbol{X}|\boldsymbol{Z},\boldsymbol{\theta})p(\boldsymbol{Z}|\boldsymbol{\theta})$ と書ける
- 従って、下界 $\mathcal{L}(q)$ は次のようになる

$$\mathcal{L}(q,\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})\right] + H[q] \tag{613}$$

- 下界 $\mathcal{L}(q,\boldsymbol{\theta})$ をパラメータ $\boldsymbol{\theta}$ について増加させるとする
- これは、$\mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})\right] = \mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})\right] + \text{Const.}$ を、q を固定しつつ、$\boldsymbol{\theta}$ について増加させることに繋がる

- このとき、$q(\boldsymbol{Z})$ が高い確率となる $\boldsymbol{Z}$ について $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$ が増大し、$q(\boldsymbol{Z})$ が低い確率となる $\boldsymbol{Z}$ について $p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})$ が減少する

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{X},\boldsymbol{Z}|\boldsymbol{\theta})\right] &= \mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})p(\boldsymbol{X})\right] \\
&= \mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})\right] + \mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{X})\right] \\
&= \mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})\right] + \ln p(\boldsymbol{X}) \\
&= \mathbb{E}_{\boldsymbol{Z}}\left[\ln p(\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})\right] + \text{Const.}
\end{aligned} \tag{614}$$

学習と推論の相互作用

- これより、近似仮定が自己充足的予言となることが分かる
- 自己充足的予言とは、ある事象や状況に関する判断や思い込みが原因となり、その結果として、その判断や思い込みが現実化することである

- 即ち、予め近似事後分布 $q(\boldsymbol{Z})$ に何らかの仮定を置いて訓練することで、結果としてその仮定に沿うような分布が得られてしまう

- 単峰性の近似事後分布 $q(\boldsymbol{Z})$ を使って訓練することを考える
- 得られる事後分布 $q(\boldsymbol{Z}) = p(\boldsymbol{Z}|\boldsymbol{X})$ は、厳密な推論によって得られた $p(\boldsymbol{Z}|\boldsymbol{X})$ と比べて、単峰性がはるかに強くなっている
- 事後分布が多峰性であるという仮定が、推定結果にも現れてしまう

- 変分推論によってモデルに課される損害
  - モデルを訓練した後に、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ を推定し、下界 $\mathcal{L}(q, \boldsymbol{\theta})$ との差が十分に小さい $(\mathcal{L}(q, \boldsymbol{\theta}) \simeq \ln p(\boldsymbol{X}|\boldsymbol{\theta}))$ ことを確認する

学習と推論の相互作用

- このとき、ある特定のパラメータ $\boldsymbol{\theta}$ について、変分近似が正確であるといえる

- 但し、一般的に変分近似が正確であるとはいえない
- また、変分推論が、学習過程に殆ど害を及ぼさないとも結論付けてはいけない

- 変分近似による真の損害は、$\ln p(\boldsymbol{X}|\boldsymbol{\theta})$ を最大にする $\boldsymbol{\theta}^*$ が分からなければ、測定することができない

- ある $\boldsymbol{\theta}$ について $\mathcal{L}(q, \boldsymbol{\theta}) \simeq \ln p(\boldsymbol{X}|\boldsymbol{\theta})$ が成立しても、$\mathcal{L}(q, \boldsymbol{\theta}) \simeq \ln p(\boldsymbol{X}|\boldsymbol{\theta}) \ll \ln p(\boldsymbol{X}|\boldsymbol{\theta}^*)$ であるかもしれない
- あるパラメータ $\boldsymbol{\theta}$ については変分近似は正確だが、実際には近似が全く上手く行っていないことになる

学習と推論の相互作用

- また $\max_q \mathcal{L}(q, \boldsymbol{\theta}^*) \ll \ln p(\boldsymbol{X}|\boldsymbol{\theta}^*)$ であれば、$\boldsymbol{\theta}^*$ のときの事後分布は、制限された q にとって複雑すぎるため、訓練では $\boldsymbol{\theta}^*$ を見つけることができない
- この問題は、変分推論とは別の方法で $\boldsymbol{\theta}^*$ が求まる場合にしか、検知することができない

学習による近似推論

- 近似推論の実行の学習
  - 不動点方程式や勾配に基づく最適化を、繰り返し行う近似推論は、計算コストが非常に高い
  - 近似推論による最適化処理は、入力 $\boldsymbol{X}$ から、近似事後分布 $q^*(\boldsymbol{Z}) = \arg\max_q \mathcal{L}(\boldsymbol{X}, q)$ へと写像する関数 f とみなせる
  - これにより、反復的な最適化処理を単なる関数とみなして、関数 $\widehat{f}(\boldsymbol{X}, \boldsymbol{\theta})$ を、ニューラルネットワークで近似することができる
- Wake-Sleep アルゴリズム
  - データ $\boldsymbol{X}$ から潜在変数 $\boldsymbol{Z}$ を推論するときの困難は、正しい潜在変数 $\boldsymbol{Z}$ が分からず、従って教師あり訓練集合が得られないことである
  - データ $\boldsymbol{X}$ から潜在変数 $\boldsymbol{Z}$ への写像は、モデル $(p(\boldsymbol{X}|\boldsymbol{Z}, \boldsymbol{\theta}), p(\boldsymbol{Z}|\boldsymbol{\theta})$ の選択) に依存し、更に $\boldsymbol{\theta}$ の変化に伴って変わり続ける

学習による近似推論

- Wake-Sleep アルゴリズムでは、$\boldsymbol{X}$ と $\boldsymbol{Z}$ の両方のサンプルを、モデル分布から抽出する
- モデルの分布が高い確率を示すような $\boldsymbol{X}$ に対してしか、推論ネットワークを学習できない

- モデルの分布が、データの分布に似ていないとき、推論ネットワークはデータに似たサンプルを学習できない

- 推論学習の他の形式
  - 変分自己符号化器は、生成モデリングで主要なアプローチとなった

  - 推論ネットワークは、単に下界 $\mathcal{L}(q)$ を定義するために使用される
  - ネットワークのパラメータは、下界 $\mathcal{L}(q)$ が増加するように適用される

目次

目次

生成モデル

- 生成モデルの目的
  - データ $\boldsymbol{x}$ に関する分布 $p(\boldsymbol{x})$ を推定する

- データ生成過程
  - データ $\boldsymbol{x}$ は一般に高次元である
  - 但し、実際にデータが分布しているのは、ごく限られた一部の低次元の領域であると考えられる (多様体仮説)
  - データ $\boldsymbol{x}$ 自体は高次元だが、本質的には低次元の情報しか持たないと考えられる

  - データ $\boldsymbol{x}$ を、より低次元なベクトル $\boldsymbol{z}$ を使って、表現することを考える
  - データに関する分布 $p(\boldsymbol{x})$ を、潜在変数 $\boldsymbol{z}$ に関する分布と、うまく組み合わせて記述する
  - 潜在変数からデータが生成されるまでの過程を組み込んで、$p(\boldsymbol{x})$ を記述する

目次

変分自己符号化器 (VAE) の概要

- 深層学習における生成モデル
  - 主に以下の 2 つの手法が存在する
  - 敵対的生成ネットワーク (Generative Adversarial Networks, GAN)
  - 変分自己符号化器 (Variational Auto Encoders, VAE)
  - ここでは変分自己符号化器 (VAE) について扱う
  - VAE を、異常検知 (不良品の検出など) に使った例がある

変分自己符号化器 (VAE) の概要

- VAE におけるグラフィカルモデル
  - 図 16 のような、潜在変数を含んだグラフィカルモデルを考える

  - データ $\boldsymbol{x}$ について、ある一つの潜在変数 $\boldsymbol{z}$ が対応しているとする
  - 各データ $\boldsymbol{x}$ は、分布 $p(\boldsymbol{x})$ から独立にサンプルされるとする
  - 従って、データ $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ は独立同分布標本とする

  - θ は、潜在変数 $\boldsymbol{z}$ からデータ $\boldsymbol{x}$ を取得する際に使用されるパラメータ
  - ϕ は、データ $\boldsymbol{x}$ から潜在変数 $\boldsymbol{z}$ を生成する際に使用されるパラメータ
  - N は、データ数である

変分自己符号化器 (VAE) の概要

- データ $\boldsymbol{x}$ の生成過程
  - データ $\boldsymbol{x}$ の生成過程は、次のように考える
  - 分布 $p(\boldsymbol{z}|\theta)$ から、潜在変数 $\boldsymbol{z}_i$ がサンプルされる
  - 分布 $p(\boldsymbol{x}|\boldsymbol{z}_i, \theta)$ から、データ $\boldsymbol{x}_i$ がサンプルされる
  - これより、データ $\boldsymbol{x}$ の分布を次のように表現できる

$$p(\boldsymbol{x}|\theta) = \int p(\boldsymbol{x}|\boldsymbol{z}, \theta) p(\boldsymbol{z}|\theta) d\boldsymbol{z} \tag{615}$$

- 潜在変数 $\boldsymbol{z}$ をデータ $\boldsymbol{x}$ から取得する過程
  - 潜在変数 $\boldsymbol{z}_i$ をデータ $\boldsymbol{x}_i$ から得る過程は、次のように考える
  - 分布 $q(\boldsymbol{z}|\boldsymbol{x}_i, \phi)$ から、潜在変数 $\boldsymbol{z}_i$ がサンプルされる

変分自己符号化器 (VAE) の概要

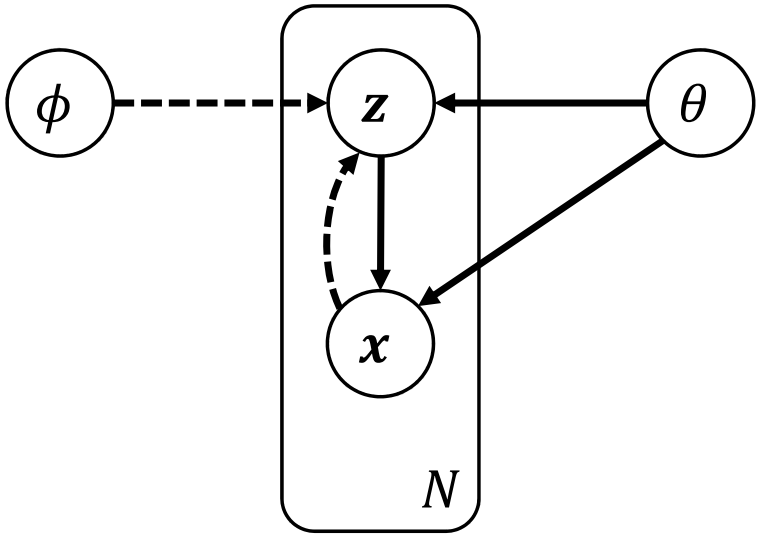

図 16: 変分自己符号化器 (VAE) におけるグラフィカルモデル

変分自己符号化器 (VAE) の概要

- 確率分布のニューラルネットワークによる表現
  - 潜在変数を含む確率モデルについて、パラメータの最尤解を求めるために、EM アルゴリズムを導出した
  - EM アルゴリズムでは、潜在変数に関する事後分布 $p(\boldsymbol{z}|\boldsymbol{x}, \theta)$ を計算する必要があった
  - この事後分布 $p(\boldsymbol{z}|\boldsymbol{x}, \theta)$ の計算が困難であるとき、$p(\boldsymbol{z}|\boldsymbol{x}, \theta)$ を別の分布 $q(\boldsymbol{z}|\boldsymbol{x}, \phi)$ で近似し、変分推論によって $q(\boldsymbol{z}|\phi)$ の最適解を求めた

  - VAE は変分推論の変種であり、近似事後分布 $q(\boldsymbol{z}|\boldsymbol{x}, \phi)$ と、$p(\boldsymbol{x}|\boldsymbol{z}, \theta)$ の 2 つをニューラルネットワークで表現する

  - データ $\boldsymbol{x}$ を潜在変数 $\boldsymbol{z}$ に対応付けるニューラルネットワークを、Encoder という
  - 潜在変数 $\boldsymbol{z}$ からデータ $\boldsymbol{x}$ を復元するニューラルネットワークを、Decoder という
  - 分布 $q(\boldsymbol{z}|\boldsymbol{x}, \phi)$ は Encoder、分布 $p(\boldsymbol{x}|\boldsymbol{z}, \theta)$ は Encoder に相当する

目次

変分自己符号化器 (VAE) の理論

- 変分自己符号化器 (VAE) の理論
  - 変分下界 $\mathcal{L}(q)$ は次のようであった

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\boldsymbol{z}|\boldsymbol{x}) \ln \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z} && (616)\\
&= \int q(\boldsymbol{z}|\boldsymbol{x}) \ln \frac{p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z} && (617)\\
&= \int q(\boldsymbol{z}|\boldsymbol{x}) \ln p(\boldsymbol{x}|\boldsymbol{z}) d\boldsymbol{z} + \int q(\boldsymbol{z}|\boldsymbol{x}) \ln \frac{p(\boldsymbol{z})}{q(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z} && (618)\\
&= \int q(\boldsymbol{z}|\boldsymbol{x}) \ln p(\boldsymbol{x}|\boldsymbol{z}) d\boldsymbol{z} - \mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) && (619)\\
&= \mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})} [\ln p(\boldsymbol{x}|\boldsymbol{z})] - \mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) && (620)
\end{aligned}$$

  - ここでは、単一のデータ $\boldsymbol{x}$ と、それに対応する潜在変数 $\boldsymbol{z}$ を考えている
  - また、パラメータ θ, ϕ は省略している

変分自己符号化器 (VAE) の理論

- 第 1 項 $\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\ln p(\boldsymbol{x}|\boldsymbol{z})\right]$ を大きく、また第 2 項 $\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z}))$ を小さくすることで、変分下界 $\mathcal{L}(q)$ を大きくできる

- VAE では、変分下界 $\mathcal{L}(q)$ を最大化するパラメータ θ, ϕ を求めるために、ニューラルネットを使用する (変分推論にニューラルネットをねじ込んだもの)

- KL ダイバージェンスの項は、後ほど求めることにする (解析的に求められる)
- 第 1 項は、分布 $q(\boldsymbol{z}|\boldsymbol{x})$ に関する期待値であり、VAE ではサンプリングで近似する

$$\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z}|\boldsymbol{x})}\left[\ln p(\boldsymbol{x}|\boldsymbol{z})\right] \simeq \frac{1}{L}\sum_{l=1}^{L}\ln p(\boldsymbol{x}_i|\boldsymbol{z}_{i,l}) \tag{621}$$

変分自己符号化器 (VAE) の理論

- これより、変分下界 $\mathcal{L}(q)$ は以下のように書ける

$$\mathcal{L}(q) \simeq -\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) + \frac{1}{L}\sum_{i=1}^{L} \ln p(\boldsymbol{x}_i|\boldsymbol{z}_{i,l}) \tag{622}$$

- パラメータ θ, ϕ を含めれば、次のように書ける

$$\mathcal{L}(q) \simeq -\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x},\phi)||p(\boldsymbol{z}|\theta)) + \frac{1}{L}\sum_{i=1}^{L} \ln p(\boldsymbol{x}_i|\boldsymbol{z}_{i,l},\theta) \tag{623}$$

変分自己符号化器 (VAE) の理論

- Encoder のニューラルネットの入出力
  - Encoder は、分布 $q(\boldsymbol{z}|\boldsymbol{x}, \phi)$ を表現するニューラルネット
  - 入力 $\boldsymbol{x}$ に対応する潜在変数 $\boldsymbol{z}$ を得る

  - Encoder の入力は、明らかにデータ $\boldsymbol{x}$ である

  - 変分下界 $\mathcal{L}(q)$ において、項 $\mathbb{E}_{\boldsymbol{z}}\left[\ln p(\boldsymbol{x}|\boldsymbol{z})\right]$ は近似する必要があった
  - $\boldsymbol{z}$ を、分布 $q(\boldsymbol{z}|\boldsymbol{x}, \phi)$ から L 回サンプリングした

  - ニューラルネットで、$\boldsymbol{x}$ から $\boldsymbol{z}$ を直接サンプリングするのは困難
  - そこで、Encoder では、サンプルされたデータは出力しないことにする
  - その代わりに、サンプルする分布のパラメータを出力する

  - 例えば、サンプルする分布がガウス分布であれば、平均と分散の 2 つのパラメータを出力する

変分自己符号化器 (VAE) の理論

- 後述のように、分布 $q(\boldsymbol{z}|\boldsymbol{x}, \phi)$ はガウス分布になるので、Encoder は平均ベクトル $\boldsymbol{\mu}$ と共分散行列 $\boldsymbol{\Sigma}$ を出力する
- 共分散行列は、実際には対角行列であるため、実際には行列ではなく、行列の対角成分を要素にもつベクトルを出力する

変分自己符号化器 (VAE) の理論

- Encoder の損失関数
  - VAE では、事前分布として、平均ベクトル $\mathbf{0}$、共分散行列 $\boldsymbol{I}$ のガウス分布 $\mathcal{N}(\boldsymbol{z}|\mathbf{0}, \boldsymbol{I})$ を仮定する
  - データ $\boldsymbol{x}$ は高次元だが、実際にはそのうちの低次元な領域にまとまって存在する (多様体仮説)
  - 従って、データ $\boldsymbol{x}$ の構造を、より低次元な潜在変数 $\boldsymbol{z}$ の空間 $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}|\mathbf{0}, \boldsymbol{I})$ に押し込めることができる
  - Encoder の損失関数は、KL ダイバージェンス $\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x}, \phi)||p(\boldsymbol{z}|\theta))$ で定義できる
  - この KL ダイバージェンスを最小化することは、Encoder の分布 $q(\boldsymbol{z}|\boldsymbol{x}, \phi)$ を、$p(\boldsymbol{z}|\theta) \sim \mathcal{N}(\boldsymbol{z}|\mathbf{0}, \boldsymbol{I})$ に近づける制約に相当する
  - $p(\boldsymbol{z}|\theta)$ がガウス分布であれば、事後分布 $p(\boldsymbol{z}|\boldsymbol{x}, \theta)$ もガウス分布であり、従って $q(\boldsymbol{z}|\boldsymbol{x}, \phi)$ もガウス分布となる

変分自己符号化器 (VAE) の理論

- よって $\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x},\phi)||p(\boldsymbol{z}|\theta))$ は、2 つのガウス分布間の KL ダイバージェンスである

- 一般に、2 つのガウス分布 $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$、$q(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ 間の KL ダイバージェンスは、解析的に計算できる
- KL ダイバージェンス $\mathrm{KL}(p(\boldsymbol{z})||q(\boldsymbol{z}))$ を順番に求めてみよう

$$
\begin{aligned}
& \mathrm{KL}(p(\boldsymbol{z})||q(\boldsymbol{z})) \\
= & \int p(\boldsymbol{z}) \ln \frac{p(\boldsymbol{z})}{q(\boldsymbol{z})} d\boldsymbol{z} && (624) \\
= & \int p(\boldsymbol{z}) \left(\ln p(\boldsymbol{z}) - \ln q(\boldsymbol{z})\right) d\boldsymbol{z} && (625) \\
= & \int p(\boldsymbol{z}) \left(\ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) - \ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\right) d\boldsymbol{z} && (626) \\
= & \mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})} \left[\ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) - \ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\right] && (627)
\end{aligned}
$$

- データを D 次元、潜在変数を K 次元とする

変分自己符号化器 (VAE) の理論

- ここで

$$\begin{aligned} & \ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \\ = & \ln \left(\frac{1}{(2\pi)^{\frac{K}{2}}} \frac{1}{|\boldsymbol{\Sigma}_0|^{\frac{1}{2}}} \exp \left(-\frac{1}{2} (\boldsymbol{z} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_0) \right) \right) \\ = & -\frac{K}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_0| - \frac{1}{2} (\boldsymbol{z} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_0) \end{aligned} \tag{628}$$

$$\begin{aligned} & \ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \\ = & -\frac{K}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_1| - \frac{1}{2} (\boldsymbol{z} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_1) \end{aligned} \tag{629}$$

であるので

$$\begin{aligned} & \ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) - \ln \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \\ = & -\frac{K}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_0| - \frac{1}{2} (\boldsymbol{z} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1} (\boldsymbol{z} - \boldsymbol{\mu}_0) - \end{aligned}$$

$$\left(-\frac{K}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}_1| - \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1)\right)$$

$$= \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} - \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_0) + \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1) \tag{630}$$

- よって

$$\begin{aligned}
&\mathrm{KL}(p(\boldsymbol{z})||q(\boldsymbol{z})) \\
=\ &\mathbb{E}_{\boldsymbol{z}\sim p(\boldsymbol{z})}\left[\ln\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) - \ln\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)\right] \\
=\ &\mathbb{E}_{p(\boldsymbol{z})}\left[\frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} - \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_0) + \frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1)\right] \qquad (631) \\
=\ &\frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} - \frac{1}{2}\mathbb{E}_{p(\boldsymbol{z})}\left[(\boldsymbol{z}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_0)\right] +
\end{aligned}$$

変分自己符号化器 (VAE) の理論

$$\frac{1}{2}\mathbb{E}_{p(\boldsymbol{z})}\left[\left(\boldsymbol{z}-\boldsymbol{\mu}_1\right)^T\boldsymbol{\Sigma}_1^{-1}\left(\boldsymbol{z}-\boldsymbol{\mu}_1\right)\right] \tag{632}$$

- ここで、期待値についての式を導出しておく
- $\mathbb{E}\left[\boldsymbol{z}\right]=\boldsymbol{\mu}$、$\mathbb{E}\left[\left(\boldsymbol{z}-\boldsymbol{\mu}\right)\left(\boldsymbol{z}-\boldsymbol{\mu}\right)^T\right]=\boldsymbol{\Sigma}$ とする
- $\boldsymbol{z}$ の i 成分を z_i、$\boldsymbol{\mu}$ の i 成分を μ_i、$\boldsymbol{\Sigma}$ の i,j 成分を Σ_{ij} とする

- このとき

$$\begin{aligned}
\Sigma_{ij} &= \mathbb{E}\left[\left(z_i-\mu_i\right)\left(z_j-\mu_j\right)\right] \\
&= \mathbb{E}\left[z_iz_j-z_i\mu_j-z_j\mu_i+\mu_i\mu_j\right] \\
&= \mathbb{E}\left[z_iz_j\right]-\mu_j\mathbb{E}\left[z_i\right]-\mu_i\mathbb{E}\left[z_j\right]+\mu_i\mu_j \\
&= \mathbb{E}\left[z_iz_j\right]-\mu_j\mu_i-\mu_i\mu_j+\mu_i\mu_j \\
&= \mathbb{E}\left[z_iz_j\right]+\mu_i\mu_j
\end{aligned} \tag{633}$$

であるから

$$\mathbb{E}\left[z_iz_j\right]=\Sigma_{ij}-\mu_i\mu_j \tag{634}$$

変分自己符号化器 (VAE) の理論

- そして、行列 $\boldsymbol{A}$ の i, j 成分を A_{ij} とすれば、以下を得る

$$
\begin{aligned}
\mathbb{E}\left[\boldsymbol{z}^T \boldsymbol{A} \boldsymbol{z}\right] &= \mathbb{E}\left[\sum_i \sum_j z_i A_{ij} z_j\right] \\
&= \sum_i \sum_j A_{ij} \mathbb{E}\left[z_i z_j\right] \\
&= \sum_i \sum_j A_{ij} \left(\Sigma_{ij} + \mu_i \mu_j\right) \\
&= \sum_i \sum_j A_{ij} \Sigma_{ij} + \sum_i \sum_j A_{ij} \mu_i \mu_j \\
&= \sum_i \sum_j A_{ij} \Sigma_{ji} + \sum_i \sum_j \mu_i A_{ij} \mu_j \\
&= \sum_i \left(\boldsymbol{A}\boldsymbol{\Sigma}\right)_{ii} + \boldsymbol{\mu}^T \boldsymbol{A} \boldsymbol{\mu} \\
&= \mathrm{Tr}\left(\boldsymbol{A}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}^T \boldsymbol{A} \boldsymbol{\mu} \qquad (635)
\end{aligned}
$$

変分自己符号化器 (VAE) の理論

- 上式の変形では、共分散行列 $\boldsymbol{\Sigma}$ が対称行列ゆえ、$\Sigma_{ij} = \Sigma_{ji}$ が成立することを用いた
- また、ベクトル $\boldsymbol{a}$ の i 成分を a_i とすれば、以下を得る

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{a}^T\boldsymbol{z}\right] &= \mathbb{E}\left[\boldsymbol{z}^T\boldsymbol{a}\right] \\
&= \mathbb{E}\left[\sum_i z_i a_i\right] \\
&= \sum_i a_i \mathbb{E}\left[z_i\right] \\
&= \sum_i a_i \mu_i \\
&= \boldsymbol{a}^T\boldsymbol{\mu} = \boldsymbol{\mu}^T\boldsymbol{a} \qquad (636)
\end{aligned}$$

変分自己符号化器 (VAE) の理論

- これより、$\boldsymbol{a}, \boldsymbol{B}$ をそれぞれ適当なベクトル、行列とすれば、以下を得る

$$
\begin{aligned}
& \mathbb{E}\left[(\boldsymbol{z}-\boldsymbol{a})^T \boldsymbol{B}(\boldsymbol{z}-\boldsymbol{a})\right] \\
= \; & \mathbb{E}\left[\boldsymbol{z}^T\boldsymbol{B}\boldsymbol{z} - \boldsymbol{z}^T\boldsymbol{B}\boldsymbol{a} - \boldsymbol{a}^T\boldsymbol{B}\boldsymbol{z} + \boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a}\right] \\
= \; & \mathbb{E}\left[\boldsymbol{z}^T\boldsymbol{B}\boldsymbol{z}\right] - \mathbb{E}\left[\boldsymbol{z}^T\boldsymbol{B}\boldsymbol{a}\right] - \mathbb{E}\left[\boldsymbol{a}^T\boldsymbol{B}\boldsymbol{z}\right] + \boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a} \\
= \; & \left(\mathrm{Tr}\left(\boldsymbol{B}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}^T\boldsymbol{B}\boldsymbol{\mu}\right) - \boldsymbol{\mu}^T\boldsymbol{B}\boldsymbol{a} - \boldsymbol{a}^T\boldsymbol{B}\boldsymbol{\mu} + \boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a} \\
= \; & \mathrm{Tr}\left(\boldsymbol{B}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}^T\boldsymbol{B}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\boldsymbol{B}\boldsymbol{a} + \boldsymbol{a}^T\boldsymbol{B}\boldsymbol{a} \qquad (637)
\end{aligned}
$$

特に、$\boldsymbol{a} = \boldsymbol{\mu}, \boldsymbol{B} = \boldsymbol{\Sigma}^{-1}$ とすれば

$$
\begin{aligned}
& \mathbb{E}\left[(\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right] \\
= \; & \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right) + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\
= \; & \mathrm{Tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\right) \\
= \; & \mathrm{Tr}\left(\boldsymbol{I}\right) = K \qquad (638)
\end{aligned}
$$

変分自己符号化器 (VAE) の理論

- これを用いれば、KL ダイバージェンスは次のようになる

$$
\begin{aligned}
& \mathrm{KL}(p(\boldsymbol{z})||q(\boldsymbol{z})) \\
= \; & \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} - \frac{1}{2}\mathbb{E}_{p(\boldsymbol{z})}\left[(\boldsymbol{z}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_0)\right] + \\
& \frac{1}{2}\mathbb{E}_{p(\boldsymbol{z})}\left[(\boldsymbol{z}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{z}-\boldsymbol{\mu}_1)\right] \\
= \; & \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} - \frac{1}{2}K + \frac{1}{2}\left(\mathrm{Tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0\right) + \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_0 - \right. \\
& \left. 2\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1\right) \\
= \; & \frac{1}{2}\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} - \frac{1}{2}K + \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0\right) + \\
& \frac{1}{2}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\mu}_0-\boldsymbol{\mu}_1) \\
= \; & \frac{1}{2}\left(\ln\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} - K + \mathrm{Tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0\right) + \right.
\end{aligned}
$$

変分自己符号化器 (VAE) の理論

$$\left. (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \right) \tag{639}$$

- これで、2 つのガウス分布 $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$、$q(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ 間の KL ダイバージェンスが、次のようになることが分かった

$$\begin{aligned} & \mathrm{KL}(p(\boldsymbol{z})||q(\boldsymbol{z})) \\ = & \frac{1}{2} \left(\ln \frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|} - K + \mathrm{Tr}\left(\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_0\right) + \right. \\ & \left. (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1} (\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1) \right) \end{aligned} \tag{640}$$

- ここでは、$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ は、Encoder のニューラルネットが表現する分布 $q(\boldsymbol{z}|\boldsymbol{x}, \phi)$ である
- そして、分布 $q(\boldsymbol{z}|\boldsymbol{x}, \phi)$ はガウス分布であったので、ここでは平均 $\boldsymbol{\mu}_0$ と共分散行列 $\boldsymbol{\Sigma}_0$ を使って、$q(\boldsymbol{z}|\boldsymbol{x}, \phi) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ と表すことにする

変分自己符号化器 (VAE) の理論

- また $q(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ は、潜在変数に関する事後分布 $p(\boldsymbol{z}|\theta) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})$ である

- 結局、KL ダイバージェンス $\mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x}, \phi)||p(\boldsymbol{z}|\theta))$ は、先程の式に $\boldsymbol{\mu}_1 = \boldsymbol{0}$ と $\boldsymbol{\Sigma}_1 = \boldsymbol{I}$ を代入すれば得られる

$$\begin{aligned} & \mathrm{KL}(q(\boldsymbol{z}|\boldsymbol{x}, \phi)||p(\boldsymbol{z}|\theta)) \\ = \quad & \mathrm{KL}(\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)||\mathcal{N}(\boldsymbol{z}|\boldsymbol{0}, \boldsymbol{I})) \\ = \quad & \frac{1}{2}\left(-\ln|\boldsymbol{\Sigma}_0| - K + \mathrm{Tr}\left(\boldsymbol{\Sigma}_0\right) + \boldsymbol{\mu}_0^T \boldsymbol{\mu}_0\right) \end{aligned} \tag{641}$$

- この KL ダイバージェンスが、Encoder の損失関数として定義される

- Encoder のニューラルネットは、入力としてデータ $\boldsymbol{x}$ を取り、平均 $\boldsymbol{\mu}_0$ と共分散行列 $\boldsymbol{\Sigma}_0$ を出力する
- 従って、Encoder の出力と、潜在変数の次元 K を上式に代入すれば、損失関数を容易に計算できる

変分自己符号化器 (VAE) の理論

- $\boldsymbol{\Sigma}_0$ は対称行列であるため、実際に出力されるのは、$\boldsymbol{\Sigma}_0$ の対角成分を並べたベクトルである

- 一般的な VAE の Encoder は次の図 17 のように表せる

変分自己符号化器 (VAE) の理論

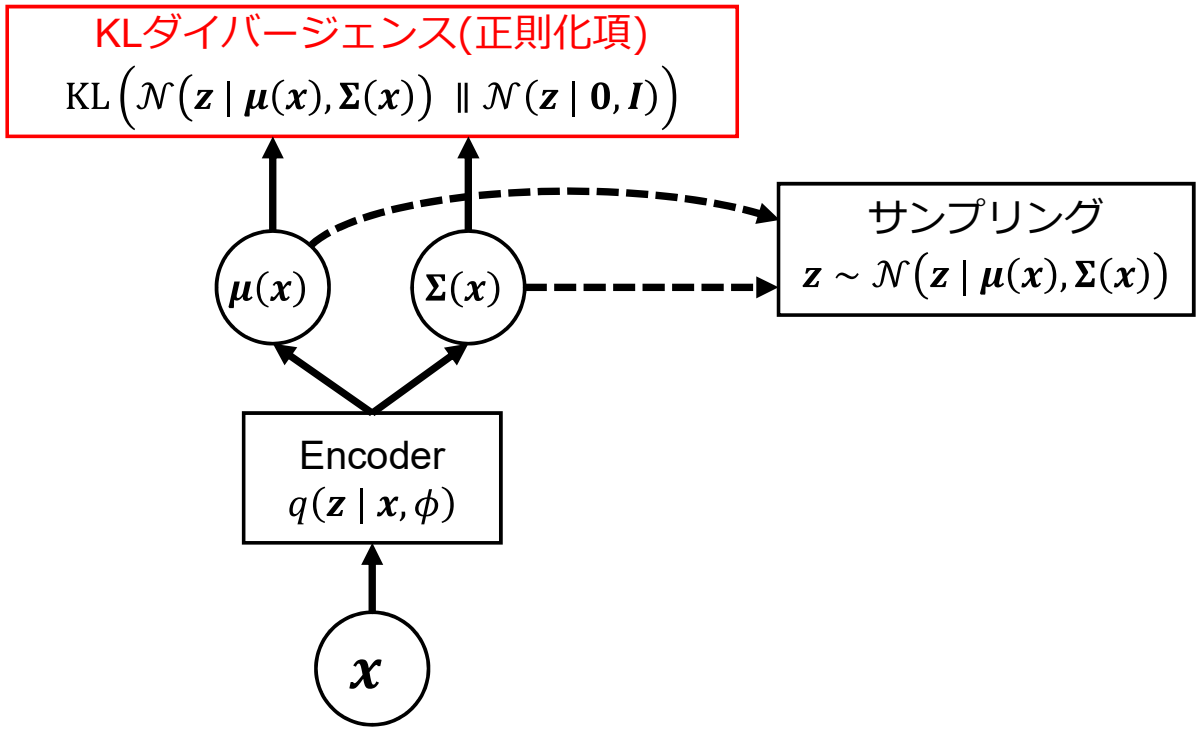

図 17: VAE の Encoder の概要

変分自己符号化器 (VAE) の理論

- Encoder のニューラルネットの構造
  - VAE が最初に提案された論文では、隠れ層は 1 層となっている
  - 最終層は、$\boldsymbol{\mu}_0$ と $\sqrt{\boldsymbol{\Sigma}_0}$ を出力する 2 つのユニットから成る
  - $\sqrt{\boldsymbol{\Sigma}_0}$ とは、行列 $\boldsymbol{\Sigma}_0$ の各要素の平方根を取った行列である

  - 隠れ層の重みを $\boldsymbol{W}_h$、バイアスを $\boldsymbol{b}_h$、活性化関数を $f(\cdot)$、層の出力を $\boldsymbol{h}$ とする
  - 隠れ層で行う処理は、次の式で表される

$$\boldsymbol{h} = f(\boldsymbol{W}_h \boldsymbol{x} + \boldsymbol{b}_h) \tag{642}$$

  - $\boldsymbol{\mu}_0$ は、重み $\boldsymbol{W}_m$ とバイアス $\boldsymbol{b}_m$ を使って、以下のように計算される

$$\boldsymbol{\mu}_0 = \boldsymbol{W}_m \boldsymbol{h} + \boldsymbol{b}_m \tag{643}$$

  - $\sqrt{\boldsymbol{\Sigma}_0}$ は、重み $\boldsymbol{W}_s$ とバイアス $\boldsymbol{b}_s$ から、以下のように計算される

$$\sqrt{\boldsymbol{\Sigma}_0} = \boldsymbol{W}_s \boldsymbol{h} + \boldsymbol{b}_s \tag{644}$$

変分自己符号化器 (VAE) の理論

- Decoder のニューラルネットの入出力
  - Decoder は、分布 $p(\boldsymbol{x}|\boldsymbol{z}, \theta)$ を表現するニューラルネット
  - 潜在変数 $\boldsymbol{z}$ から元のデータ $\boldsymbol{x}$ を復元する
  - Decoder の入力は、Encoder によってサンプリングされた $\boldsymbol{z}$ となる
  - もう少し正確に表現すると、Encoder から出力されるのは、分布のパラメータ $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ である
  - そして、そのパラメータを使って分布 $\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ を構成し、分布から $\boldsymbol{z}$ をサンプリングする
  - Decoder の出力は、復元されたデータ $\boldsymbol{y}$ である

変分自己符号化器 (VAE) の理論

- Decoder の損失関数
  - 画像データでは通常、各ピクセルの値が 0 から 1 までになるようにスケーリングされている
  - このとき、$p(\boldsymbol{x}|\boldsymbol{z})$ はベルヌーイ分布と仮定していることになる

  - 出力層のユニット j は、$y_j = p(x_j = 1|\boldsymbol{z})$ を出力しているとみなせる
  - y_j は、再構成 $\boldsymbol{y}$ の j 番目の要素であり、元のデータ $\boldsymbol{x}$ の j 番目の要素 x_j と対応する

  - VAE が最初に提案された論文では、隠れ層は 1 層のみである
  - 再構成 $\boldsymbol{y}$ は、次のように計算される

$$\boldsymbol{y} = f_\sigma\left(\boldsymbol{W}_o \tanh\left(\boldsymbol{W}_h \boldsymbol{z} + \boldsymbol{b}_h\right) + \boldsymbol{b}_o\right) \tag{645}$$

  - $f_\sigma(\cdot)$ は、行列の各要素にシグモイド関数 $\sigma(\cdot)$ を適用する活性化関数
  - $\boldsymbol{W}_h, \boldsymbol{b}_h$ は隠れ層の重みとバイアス、$\boldsymbol{W}_o, \boldsymbol{b}_o$ は出力層の重みとバイアス

変分自己符号化器 (VAE) の理論

- このとき $\ln p(\boldsymbol{x}|\boldsymbol{z})$ は次のように記述できる

$$\begin{aligned}
\ln p(\boldsymbol{x}|\boldsymbol{z}) &= \ln \prod_{j=1}^{D} \left(p(x_j = 1|\boldsymbol{z})\right)^{x_j} \left(p(x_j = 0|\boldsymbol{z})\right)^{1-x_j} \\
&= \ln \prod_{j} \left(p(x_j = 1|\boldsymbol{z})\right)^{x_j} \left(1 - p(x_j = 1|\boldsymbol{z})\right)^{1-x_j} \\
&= \ln \prod_{j} y_j^{x_j} \left(1 - y_j\right)^{1-x_j} \\
&= \sum_{j} \left(x_j \ln y_j + (1 - x_j) \ln \left(1 - y_j\right)\right) \qquad (646)
\end{aligned}$$

- $\boldsymbol{z}$ は、分布 $q(\boldsymbol{z}|\boldsymbol{x}, \phi)$ からサンプリングされている
- 従って、上記を最大化することは、$\mathbb{E}_{q(\boldsymbol{z})}\left[\ln p(\boldsymbol{x}|\boldsymbol{z})\right]$ を最大化することに等しい

変分自己符号化器 (VAE) の理論

- 上記は、x_j と y_j のいずれもベルヌーイ分布に従う (二値変数) ときの、負の交差エントロピーとなっていることが分かる
- 従って、$\mathbb{E}_{q(\boldsymbol{z})}\left[\ln p(\boldsymbol{x}|\boldsymbol{z})\right]$ を最大化することは、交差エントロピーを最小化することに相当

- Decoder の損失関数は、以下の交差エントロピー誤差として定義できる

$$E = -\sum_j \left(x_j \ln y_j + (1 - x_j) \ln (1 - y_j)\right) \tag{647}$$

- 元データ x_j と、その再構成 y_j との差が大きければ大きいほど、上記の誤差は増大する
- これより、上記の誤差は再構成誤差 (Reconstruction Error) とよばれる

- これより、VAE の Encoder と Decoder は次の図 18 のように表せる

変分自己符号化器 (VAE) の理論

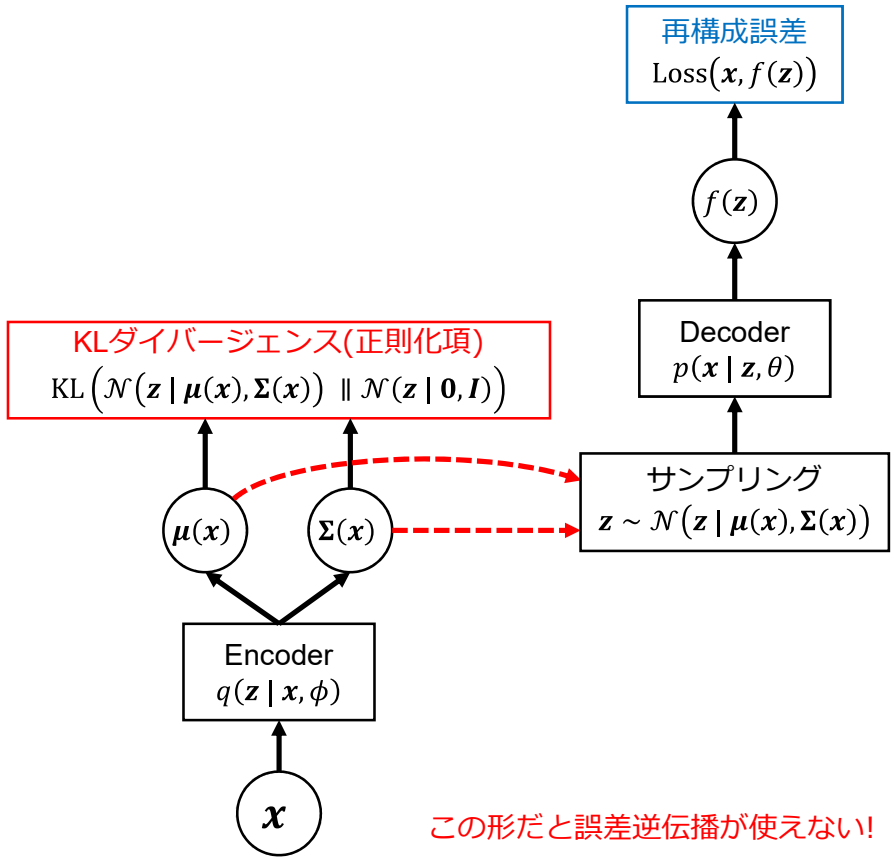

図 18: VAE の Encoder と Decoder の概要

変分自己符号化器 (VAE) の理論

- Reparameterization Trick
  - VAE が先程の図 18 のようであるとき、大きな問題が生じる
  - サンプリングを行うと、計算グラフが途中で途切れるため、誤差逆伝播法を実行できない
  - そこで、次の図 19 のように構成する
  - $\boldsymbol{z} \sim (\boldsymbol{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ として、$\boldsymbol{z}$ を分布から直接サンプリングするのではない
  - $\boldsymbol{z}$ を、決定論的な関数 $g(\boldsymbol{\epsilon}, \boldsymbol{x}|\phi)$ から決定する
  - 但し、$\boldsymbol{\epsilon}$ は、分布 $p(\boldsymbol{\epsilon})$ からサンプリングされる
  - ニューラルネットの最適化には無関係な項 $\boldsymbol{\epsilon}$ と、Encoder のパラメータ ϕ で $\boldsymbol{z}$ を表現することで、誤差逆伝播法を実行可能にする
  - このテクニックを、Reparameterization Trick という
  - $\boldsymbol{\epsilon} \sim (\boldsymbol{\epsilon}|\boldsymbol{0}, \boldsymbol{I})$ とすれば、$\boldsymbol{z}$ は次のように計算できる

$$\boldsymbol{z} = g(\boldsymbol{\epsilon}, \boldsymbol{x}|\phi) = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\epsilon} \tag{648}$$

変分自己符号化器 (VAE) の理論

- 共分散行列 $\boldsymbol{\Sigma}$ が対角行列であれば、上記の $\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\epsilon}$ は、単なる要素ごとの積 (対角行列の各要素を並べたベクトルと、$\boldsymbol{\epsilon}$ の要素ごとの積) として書ける
- $\boldsymbol{z}$ の式は以下のように導出できる
- 確率変数 $\boldsymbol{z}, \boldsymbol{\epsilon}$ 間の関係が、次のようになっているとする

$$\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\epsilon} \tag{649}$$

- 但し、正定値対称行列 $\boldsymbol{\Sigma}$ が、固有値分解によって $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T = \boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\left(\boldsymbol{U}\boldsymbol{\Lambda}\right)^T$ と表せるとする
- $\boldsymbol{U}$ は固有ベクトルを並べた行列、$\boldsymbol{\Lambda}$ は対角成分に固有値をもつ対角行列とする

変分自己符号化器 (VAE) の理論

- 確率分布 $p(\boldsymbol{z})$ と $p(\boldsymbol{\epsilon})$ との関係は、ヤコビ行列 $\boldsymbol{J} = \partial \boldsymbol{\epsilon} / \partial \boldsymbol{z}$ により次のように記述できる

$$p(\boldsymbol{z}) = p(\boldsymbol{\epsilon}) |\det(\boldsymbol{J})| = p(\boldsymbol{\epsilon}) \left| \det \left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\epsilon}} \right) \right| \tag{650}$$

- ヤコビ行列 $\boldsymbol{J}$ を計算すると次のようになる

$$\boldsymbol{J} = \frac{\partial \boldsymbol{\epsilon}}{\partial \boldsymbol{y}} = \frac{\partial}{\partial \boldsymbol{z}} \left(\left(\boldsymbol{U} \boldsymbol{\Lambda}^{\frac{1}{2}} \right)^{-1} (\boldsymbol{z} - \boldsymbol{\mu}) \right) = \left(\left(\boldsymbol{U} \boldsymbol{\Lambda}^{\frac{1}{2}} \right)^{-1} \right)^{T} \tag{651}$$

変分自己符号化器 (VAE) の理論

- ヤコビ行列 $\boldsymbol{J}$ の行列式は次のようになる

$$\begin{aligned}
\det(\boldsymbol{J}) &= \left|\left(\left(\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{-1}\right)^{T}\right| \\
&= \left|\left(\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{-1}\right| \qquad (\because |\boldsymbol{A}^{T}| = |\boldsymbol{A}|) \\
&= \frac{1}{\left|\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\right|} \qquad \left(\because |\boldsymbol{A}^{-1}| = \frac{1}{|\boldsymbol{A}|}\right)
\end{aligned} \tag{652}$$

- $\boldsymbol{\Sigma}$ の行列式は次のように表せる

$$|\boldsymbol{\Sigma}| = \left|\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\left(\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{T}\right| = \left|\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\right|\left|\left(\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{T}\right| = \left|\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\right|^{2} \tag{653}$$

- 従って $\left|\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\right| = |\boldsymbol{\Sigma}|^{\frac{1}{2}}$ である

変分自己符号化器 (VAE) の理論

- $\boldsymbol{\Sigma}$ は正定値である ($|\boldsymbol{\Sigma}| > 0$) から、$|\boldsymbol{\Sigma}|^{\frac{1}{2}} = \left|\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\right| > 0$ が成立し、$\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}$ も正定値行列となる
- これより、逆行列 $\left(\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^{-1}$ が存在するので、ヤコビ行列 $\boldsymbol{J}$ は計算できることが確認される

- ヤコビ行列 $\boldsymbol{J}$ の行列式の絶対値は、次のようになる

$$|\det(\boldsymbol{J})| = \left|\frac{1}{\left|\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\right|}\right| = \left|\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}}\right| = \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \tag{654}$$

- $p(\boldsymbol{\epsilon})$ がガウス分布 $\mathcal{N}(\boldsymbol{\epsilon}|\boldsymbol{0}, \boldsymbol{I})$ であるとすると、$p(\boldsymbol{z})$ は次のようになる

$$\begin{aligned} p(\boldsymbol{z}) &= p(\boldsymbol{\epsilon})\left|\det\left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{\epsilon}}\right)\right| \\ &= \frac{1}{(2\pi)^{\frac{K}{2}}}\exp\left(-\frac{1}{2}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}\right)\frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \end{aligned} \tag{655}$$

変分自己符号化器 (VAE) の理論

- ここで、指数関数の中身は次のように書ける

$$
\begin{aligned}
& (\boldsymbol{z}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{z}-\boldsymbol{\mu}) \\
= & \left(\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\epsilon}\right)^T \left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\right)^{-1} \left(\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\epsilon}\right) \\
= & \boldsymbol{\epsilon}^T \left(\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^T \boldsymbol{U}^T \left(\boldsymbol{U}^T\right)^{-1} \boldsymbol{\Lambda}^{-1}\boldsymbol{U}^{-1}\boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\epsilon} \\
= & \boldsymbol{\epsilon}^T \left(\boldsymbol{\Lambda}^{\frac{1}{2}}\right)^T \boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\epsilon} \\
= & \boldsymbol{\epsilon}^T \boldsymbol{\Lambda}^{\frac{1}{2}} \boldsymbol{\Lambda}^{-1}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\epsilon} \\
= & \boldsymbol{\epsilon}^T \boldsymbol{\epsilon}
\end{aligned}
\tag{656}
$$

変分自己符号化器 (VAE) の理論

- これより、$p(\boldsymbol{z})$ は平均 $\boldsymbol{\mu}$、共分散行列 $\boldsymbol{\Sigma}$ のガウス分布である

$$
\begin{aligned}
p(\boldsymbol{z}) &= \frac{1}{(2\pi)^{\frac{K}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{\epsilon}^T\boldsymbol{\epsilon}\right) \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \\
&= \frac{1}{(2\pi)^{\frac{K}{2}}} \frac{1}{|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right) \quad (657) \\
&= \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \quad (658)
\end{aligned}
$$

- $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ について、共分散行列 $\boldsymbol{\Sigma}$ が既に対角行列であれば、$\boldsymbol{U} = \boldsymbol{I}$、$\boldsymbol{\Lambda} = \boldsymbol{\Sigma}$ であるので、結局以下が言える

$$
\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{U}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{\epsilon} = \boldsymbol{\mu} + \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{\epsilon} \quad (659)
$$

変分自己符号化器 (VAE) の理論

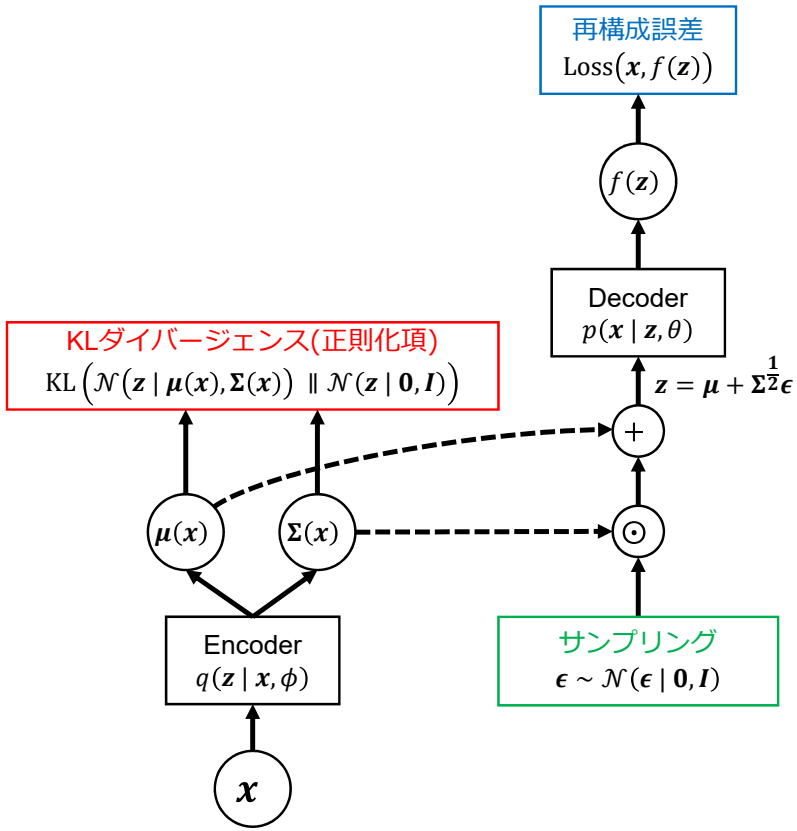

図 19: VAE の構造