1　混合ガウス分布

1.1　混合ガウス分布の EM アルゴリズム

EM アルゴリズムについて、観測データのモデルとして混合ガウス分布を設定したときの例で振り返る。まず EM アルゴリズムの目標は観測データの尤度関数に対する最尤推定であった。その中で尤度関数として混合ガウス分布を選んだ（9.14 式）。アルゴリズムの仕組みとしては、混合ガウス分布のパラメタ：各要素[*1]の平均、分散、そして混合係数を動かして、尤度関数[*2]を最大化するというものであった。

ここでパラメタについて初期値が与えられている場合に、EM アルゴリズム内でどのような計算が行われるか、詳細なステップを確認する（See. 教科書 p.154 下段）。そのあとで、EM アルゴリズムによってデータがどのように分類されるのか、具体例として図 9.8 を見ながら、その挙動を再確認しよう。

2　EM アルゴリズムのもう一つの解釈

ここでは、潜在変数に関する事後分布を導入することで、拡張された EM アルゴリズムを構築する。これを理解することで、観測データ[*3]のみが given であるという状況のもとで、EM アルゴリズムを適用し、分布のパラメタを定式化できるようになる。

本日の流れは次の通りである：まず一般的な[*4]尤度関数に対して、潜在変数を導入したときの EM アルゴリズムの挙動について調べる。そのあと、尤度関数として混合ガウス分布を選択したときに、潜在変数の事後分布および、各パラメタがどのように表されるのか確かめる。最後に、EM アルゴリズムと K-means 法との関係性について、クラスタへの割り当て方法の観点などに注目しながら議論を行う。

それでは、はじめに一般的な尤度関数を観測データ集合 $\boldsymbol{X}$ と潜在変数 $\boldsymbol{Z}$ を用いて表すことにしよう。尤度関数は次のように書くことができる。

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

$\boldsymbol{X}$ や $\boldsymbol{Z}$ の詳細な設定については読み合わせ。式 1 を見ると、ln が Σ に直接作用していることがわかる。このようなケースでは、尤度方程式を解こうとしても陽な解が得られないなんとなく解を得るのが大変な気はするんだけど、はっきりわかってないのと文字で説明しきれんのとで明日議論させて。イメージとしては、尤度方程式をパラメタについて解くときに、等式変形しやすいとかしにくいとかそんな話だと思ってる。伝わらなかったらごめん。

一方で、実は観測データ、潜在変数が完全に与えられたもとでの尤度関数については最大化が簡単に実行できることが知られている[*5]。ちなみに観測データ、潜在変数が完全に与えられたときの尤度関数を完全データ

[*1] 混合分布の各要素という意味である。

[*2] ここではデータ点ベクトル $\boldsymbol{x}_n$ が尤度関数に引数となっている。このあとで一般のデータ集合に対して尤度関数を構成するときには、その引数がデータ集合 $\boldsymbol{X}$ となっていることに注意せよ。

[*3] 観測データのみが与えられる場合と、観測データと潜在変数のどちらもが与えられる場合との違いについては、のちほど説明する。

[*4] ここでいう「一般的な」とは分布の形を仮定しない、という意味である。

[*5] 最大化が簡単に実行できる理由について（9.30）式をベースに確かめよう。いま対数は同時分布の部分に直接作用している。この場合、尤度方程式において陽な解が得られる、つまり尤度を最大化するパラメタが簡単な形で表現される（と信じよう）。なお、具体的な分布形を定めたケースについては、混合ガウス分布再訪のサブセクションで確かめることにしよう。

尤度関数と呼ぶ。しかしながら、実際には変数として与えられるのは観測データ集合 $\boldsymbol{X}$ のみであるため、完全データ尤度関数の最大化を実行することは現実的ではない。

そこで、我々が取るアプローチは次のようになる：まず、潜在変数の事後分布に関する期待値を最大化するというステップを挟み込んでやる。実は、この事後期待値最大化のステップを踏むことで、のちに完全データ対数尤度関数の期待値が求められることがわかる。そしてこの期待値を最大化するようなパラメタが EM アルゴリズムで求めるパラメタとなる、と考えてやればよい[*6]。

それでは、この場合の EM アルゴリズムの流れをまとめておこう。細かい流れに入る前に、いま一度アルゴリズムの目的が、潜在変数の尤度関数 $\ln p(\boldsymbol{X} \mid \boldsymbol{\theta})$ をパラメタ $\boldsymbol{\theta}$ について最大化することであるということを思い出してほしい。なお、変数およびパラメタの細かい設定については p.156 中段を読み合わせるものとする。

EM アルゴリズム流れ（潜在変数の尤度関数最大化）

・パラメタの初期値 $\boldsymbol{\theta}^{old}$ を適当に決める[*7]。

・潜在変数の事後分布 $p(\boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{\theta}^{old})$ を計算する。そして、潜在変数の事後分布について、完全データ対数尤度で期待値をとったものを次のように定義する。

$$\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{\boldsymbol{Z}} p(\boldsymbol{Z} \mid \boldsymbol{X}, \theta^{old}) \ln p(\boldsymbol{X}, \boldsymbol{Z} \mid \boldsymbol{\theta}) \tag{2}$$

この期待値計算が E ステップに相当している。[*8]

・先に定義した期待値 $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{old})$ を最大化し、そのときのパラメタ $\boldsymbol{\theta}^{new}$ を求める。これは M ステップに相当する。[*9]

・収束条件が満たされていれば計算終了、満たされていなければ $\boldsymbol{\theta}^{old} \leftarrow \boldsymbol{\theta}^{new}$ として E ステップに戻る。

最後にパラメタの事前分布を導入するモデルについて紹介する。事前分布の導入によって、EM アルゴリズムにおいて、MAP 解を求めることができる。その場合、M ステップにおいて最大化する量が変化すること、具体的にいうとパラメタ $\boldsymbol{\theta}$ の事前分布に対応する対数尤度項 $\ln p(\boldsymbol{\theta})$ を加えてやる必要があることに注意せよ。

2.1 混合ガウス分布再訪

ここでは、先に議論した一般の EM アルゴリズム—潜在変数 $\boldsymbol{Z}$ の事後分布を導入したもの—の具体例を取り扱う。その例として、観測データ集合 $\boldsymbol{X}$ の尤度関数に混合ガウス分布を設定してみよう。一般的な議論のときと同様に、我々は観測データ集合で作られた尤度関数を最大化したい。しかしながら既に学んだように、この尤度関数は対数の中に混合分布の和が入っており、微分して 0 とおいてもパラメタについて陽な解が得られないという問題がある。したがいこの最大化を実行するのは難しい。

[*6] 現時点では、このようなフレームワークをとること、特に完全データ対数尤度関数の期待値に関する最大化を行うこと、についての正当性を保証することはできない。だが、ひとまずはこのやり方で上手くいくと信じることにしよう。

[*7] 適当に決める術については、私にはわからない。教科書に説明もない。

[*8] もしかすると一般的な潜在変数で議論をしていると、これが E ステップであるというイメージがつかないかもしれない。実は、のちに見るように、潜在変数を指示変数と読み替えてしまえば、先で計算した負担率そのものになることが確かめられる。負担率はデータ点の割り当てを行うものであり、これは E ステップに他ならない。

[*9] 実際、このステップは新しい期待値を求めるという操作であり、これは今まで習った M ステップ—つまり新しいプロトタイプ（平均）を計算していたもの—と位置づけが同等なものになっている。

ここでいったん完全データ集合に関する尤度の最大化を考える[*10]。この完全データ集合の尤度について尤度方程式を解いて、平均と分散を求めることは難しくない[*11]。その理由は 2 つある：まずひとつは、いまの場合、尤度関数において対数が直接作用するのは単一のガウス分布に対するものとなっているから。もうひとつは、$\boldsymbol{z}_n$ が符号化されているようなベクトル—どこかの成分が 1 でありそれ以外の成分はすべて 0 であるという K 次元ベクトル—となっており、それゆえ尤度関数が各混合要素ごとの独立な和として捉えられるからである。

しかしながら、やはり実際は完全データ集合に関する尤度は与えられない。したがって、一般的なアルゴリズムにおいても議論したように、潜在変数の事後分布に関して、完全データ尤度関数の期待値を考えてやればいい。[*12] $\boldsymbol{Z}$ に関する事後分布を求めるために、式（9.10-11）にベイズの定理を適用する。定理の適用にあたって、$\boldsymbol{Z}$ に関する周辺分布と $\boldsymbol{X}$ に関する尤度関数が必要となる。

$$p(\boldsymbol{Z} \mid \boldsymbol{X} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) \propto \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_{nk}} \tag{3}$$

ここで式（9.10-11）における z_k, x_k をそれぞれ z_{nk}, x_{nk} と置き換えた。ここで指示変数 z_{nk} の期待値、つまりデータ点がどのクラスタに割り当てられるかという期待値を計算しよう。いま式 3 を見ると n についての積になっており、各 z_{nk} は n について独立となっていることがわかる。したがって z_{nk} の期待値は n を固定して考えればよろしい。したがって期待値 $E[z_{nk}]$ は次のようになる。分数になる理由がわからん。期待値って分子のところだけ考えればよくない？分母どっから出てきたんだっけ？ちなみに分子の計算もわからん。

$$\begin{aligned} E[z_{nk}] &= \frac{\sum_{\boldsymbol{z}_n} z_{nk} \prod_{k'} [\pi_{k'} \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}'_k, \boldsymbol{\Sigma}_{k'})]^{z_{nk'}}}{\sum_{\boldsymbol{z}_n} \prod_j [\pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)]^{z_{nj}}} \\ &= \frac{\pi_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \\ &= \gamma(z_{nk}) \end{aligned} \tag{4}$$

これは k 番目の混合要素のデータ点 $\boldsymbol{x}_n$ に対する負担率そのものである（E ステップに相当）。

そして、ここで求めた負担率から完全データ対数尤度関数の期待値は次のようになる。潜在変数で期待値をとることに注意すれば、（9.36）より

$$E_{\boldsymbol{Z}}[\ln p(\boldsymbol{x}, \boldsymbol{Z} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi})] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma(z_{nk}) \{\ln \pi_k + \ln \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\} \tag{5}$$

のように計算することができる。この期待値の最大化計算を通じてパラメタ $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$ を更新することができるのである（M ステップに相当）。

あとは、これらのステップを計算が収束するまで繰り返せばよい。

2.2 K-means との関連

ここでは K-means と EM アルゴリズム両者の違いを整理し、そのあとで実はとある極限で両者が一致することを見ていく。

[*10] 結局ここで、この最大化を考えた意義がよくわからなかった。完全データ集合の尤度を導入する必要があることは理解できるのだが。

[*11] 難しくないが、実際のところ、完全データは与えられないことは承知しておいてほしい。

[*12] 繰り返すが、なぜこのフレームワークで良いのかは、正直よくわかっていない。とにかく事後分布を導入し期待値をとれ、ということらしい。詳しい説明は 9.4 節で行われるとの記述あり。

まずは K-means と混合ガウス分布に対する EM アルゴリズムとで違いを以下の表 1 に整理した。これまで学んだ内容をおさらいしよう。

表 1　K-means アルゴリズムと EM アルゴリズムの比較

	K-means	EM アルゴリズム
データ点の割り当て	ハード割り当て	ソフト割り当て
アルゴリズムの目的	歪み尺度の最小化：式（9.1）	尤度関数の最大化：式（9.40）
更新するパラメタ	プロトタイプ $\boldsymbol{\mu}_k$	平均 $\boldsymbol{\mu}_k$、分散 $\boldsymbol{\Sigma}_k$、混合係数 π_k

続いて、両者の極限における一致性を確かめる。準備として、混合ガウス分布モデルにおいて、分散パラメタ ϵ（定数）を式（9.41）のように定義する。実は、この ϵ について $\epsilon \to 0$ と極限をとることで、K-means アルゴリズムとの対応関係—パラメタの一致性と目的関数の一致性—がわかる。

一致性について詳細は追いきれず、結論のみを以下表 2 に整理させてもらう。申し訳ない。 ただし、歪み

表 2　分散 0 極限における EM アルゴリズムと K-means アルゴリズムとの一致性

	EM アルゴリズム	K-means（EM アルゴリズムの分散 0 極限）
パラメタ	負担率 $\gamma(z_{nk})$	指示変数 r_{nk}
	平均 $\boldsymbol{\mu}_k$	プロトタイプ $\boldsymbol{\mu}_k$
目的関数	完全データ対数尤度関数：式（9.40）	歪み尺度：式（9.43）

尺度については定数倍の違いについては許容することに注意されたい。