EMアルゴリズムと情報量を用いた要因数の自動決定

小田部　修斗† 　三浦　孝夫‡

†法政大学理工学部創生科学科 〒 184-8584 東京都小金井市梶野町 3-7-2

E-mail: †shuto.kotabe.8b@stu.hosei.ac.jp, ‡miurat@hosei.ac.jp

あらまし　確率モデルの最尤推定の逐次繰り返しと情報量基準の値比較による停止基準を用いることにより，確率的クラスタリングの要因数を自動決定するアルゴリズムを提案する．これにより，不確かさやあいまいさが含まれるデータでも最適と考えられる要因数を少ない計算量で導くことができる．最後に，提案手法と既存の発見的手法の比較・評価を，混合多次元正規分布を用いて実施する．

キーワード　クラスタリング，ファジィ，教師なし学習

1. まえがき

近年の情報技術の発展により，多様かつ大量のデータを得ることが可能となっている．それに伴い，膨大なデータから有益な情報を機械的に発見するための情報処理技術の必要性が高まっている．未知のデータから有益な情報を発見するためには，予めクラスタリングを行うことが必須である．クラスタリングとは，関連するオブジェクト同士をグループ化する手法であり，機械学習や人工知能分野において最も重要なタスクのひとつである．

クラスタリングの典型的な生成モデルは正規混合モデルである．正規混合モデルでは，データは複数の正規分布の加重平均によって表現される．このモデルは，柔軟な近似能力を持ち，多くの正規分布を組み合わせそのパラメータの値を調整することで，任意の滑らかな密度関数を精度よく近似できる．このため，統計解析のみならずデータマイニング，パターン認識，機械学習など幅広く応用されている．

xの所属確率$p(x,\theta)$が，

$$p(x,\theta) = \mu p_1(x,\theta_i) + (1-\mu)p_2(x,\theta_2)$$

(ただし，$\theta =< \theta_1, \theta_2, \mu >$は互いに独立)

として定式化される問題を，混合分布問題という．通常解析的に解を得ることが難しく，反復近似などによる数値計算で処理する．

正規混合モデルのパラメータを推定するための数値計算法としては，EM アルゴリズムが知られている[3]．しかし，EM アルゴリズムはあらかじめ要因数 k を与えておく必要があり，未知のデータの要因数を推定することは一般に困難である．さらに，EM アルゴリズムは一次収束であるため，パラメータ推定するデータによっては収束までの反復回数が多大になるという問題もある．事前に要因数を設定せずにクラスタリングを行う手法として，x-means 法が提案されている[4]．x-means 法では，再帰的に k-means 法を行いクラスターの分割前後の状態を，情報量基準を用いて評価することで，要因数を自動決定する．

本論文では，これらの手法を組み合わせて x-EM アルゴリズムを提案する．分割数=2 とした EM アルゴリズムを繰り返し適用し，その改善を情報量基準を用いて評価することで最適な分割数を自動的に決定する．

以下，**2.**では EM アルゴリズム，**3.**では k-means 法とその拡張の原理について述べ，**4.**に本論文が提案する x-EM アルゴリズムを示す．**5.**で多次元正規分布への適用例を示す．**6.** では気象データを用いて混合正規モデルによる分布推定問題への適用実験結果と考察を示し，**7.**をまとめとする．

2. EM アルゴリズム

EM アルゴリズムとは，観測不可能な潜在変数に確率モデルが依存する場合を想定し，確率モデルのパラメータを最尤推定する手法である[3]．

観測可能なデータx_iと非観測データy_i からなるn個のデータXがあるとし，確率密度関数が未知のパラメータθを用いて$p(x_i, y_i; \theta)$ と与えられているとする．未知のパラメータθ の最尤推定値は，観測データの対数尤度関数

$$L(X;\theta) = log\, p(X;\theta)$$

$$= \log \int p(x_i, y_i; \theta)\, dy_i$$

を最大化するθとして求められる．しかし，これの最大値を解析的に求めるのは困難であることが多い．そのため，EM アルゴリズムでは，完全データの対数尤度関数

$$L_c(\theta;X) = \log p(x_i, y_i; \theta)$$

の条件付き期待値(Q 関数)の逐次最大化により，観測データの対数尤度関数の最大化を間接的に行う．$\theta^{(t)}$を，第 t 回目の反復後のパラメータの推定値を表すものとすると，第 t + 1 回目の反復において，まず，E ステップで Q 関数

$$Q(\theta|\theta^{(t)}) = E[Lc(\theta;X)|x_i;\, \theta^{(t)}]$$

を計算し,M ステップで Q 関数を最大にするθを求め,それを$\theta^{(t+1)}$ とする.

EM アルゴリズム整理すると，以下のようになる.

1. 初期値$\theta^{(0)}$を設定し，t =0 とする.
2. 以下を収束するまで繰り返す.

E-step: $Q(\theta|\theta^{(t)})$ を計算.

M-step: $\theta^{(t+1)} = argmax_{\theta} Q(\theta|\theta^{(t)})$, t ← t + 1 とする.

EM アルゴリズムは，各反復動作において，L(X;θ)は単調増加し，$L(X;\theta^{(t+1)}) \geq L(X;\theta^{(t)})$が成立する[4]. このとき,等号は$Q(\theta^{(t+1)}|\theta^{(t)}) = Q(\theta^{(t)}|\theta^{(t)})$のときに限る. つまり,EM アルゴリズムを実行することにより，L を極大にするようなθが得られる.

3. k-means 法とその拡張

本章では，基本的なクラスタリングアルゴリズムである k-means 法を示し，その問題点とクラスター数推定のための改善方法を示す.

k-means 法は,データの平均をクラスターの代表点とし，代表点との距離が最小となるクラスターに割り当てる方法である[9].

k-means 法は，以下のようなアルゴリズムからなる.

1.データ集合の中の最初の k 個を 1 メンバのクラスターとして取る.

2.残りのデータを最短距離の重心をもったクラスターに割り当てる.

3.新たなクラスターの重心を取り,2.の操作を行う.

4.クラスターの重心が変化しなくなった場合，アルゴリズムを終了する.

ここで，k-means では，クラスター数にあたる k の値はあらかじめ決めておく必要がある.

x-means 法 は，k-means 法を拡張した手法であり，再帰的に k-means 法を実行することでクラスターの分割数を自動決定する方法である[4].

x-means 法のアルゴリズムは,k-means 法による分割と,情報量基準による評価の 2 段階で構成されている.

1. k =2 として，k-means 法で 2 個のクラスターに分割する.

2. 現在存在しているクラスターから 1 個を選択し,それに対し k=2 で k-means を実行する.

3 2.の実行前後の情報量基準を比較し，後のほうが優れているならば，その分割を採択し，2.へ戻る．そうでなければ，分割は棄却され，他のクラスターの分割を試行する．分割すべきクラスターがなくなった場合，アルゴリズムを終了する.

x-means 法はトップダウン型分割の排他的クラスタリングであり，一度分割されたクラスターは他のクラスターに移動することはないため，初期値への依存性が大きく，局所解に陥りやすい．また，同じクラスター数の場合，k-means 法の結果より x-means 法の結果が改善されていることは少ない.

k-means 法では，クラスター同士の境界付近に存在するデータ点であっても，無理矢理片方に分類してしまうため，適切なクラスタリングが行えない場合がある．このため，データ点それぞれに対して各クラスターへの存在確率を考えたファジィ c-means 法が考案されている[10]．その中でも，エントロピー正則化によるファジィ c-means 法と呼ばれるものは，存在確率z_{ij}を以下のように定義している.

$$z_{ij} = \frac{exp\left(\frac{(x_i - \mu_j)^2}{2\sigma^2}\right)}{\sum_j exp\left(\frac{(x_i - \mu_j)^2}{2\sigma^2}\right)}$$

これは，後に示す混合正規分布における EM アルゴリズムにおいて,分散共分散行列$\Sigma_j = \sigma^2 E$,混合比$\pi_j = 1/k$としたものと一致しており，これは EM アルゴリズムにおいて分散と混合比を固定した場合に相当している[11].

4. x-EM アルゴリズム

EM アルゴリズムを繰り返して適用し、その結果を情報量基準を用いて比較することにより，最適と思われる要因数を推定する.

まず入力データを一つのクラスターと見なし，そこから分割を繰り返していく.

まず，EM ステップを実行し収束した時のパラメータ推定値をθ*とすると，そのときの Q *値*$Q(\theta|\theta^*)$は,

$$Q(\theta|\theta^*) = \sum_{i=1}^{m} Q_i(\theta|\theta^*) = Q_p(\theta|\theta^*) + \sum_{i \neq p}^{m} Q_i(\theta|\theta^*)$$

と表すことができる．この時，クラスターp を新たな 2 つのクラスターs1 と s2 に分割することを考える.

分割後のクラスターs1，s2 の初期パラメータは,

$$\theta_{s1} = \theta_{s2} = \frac{1}{2}\theta_p$$

$$X_{s1} = X_p + \epsilon,\ X_{s2} = X_p - \epsilon,$$

とする．ただし，ϵは十分小さなランダムベクトルである．この後に，EM アルゴリズムを再び実行することで s1,s2 に関してパラメータの再推定を行う．ここで，再推定が s1,s2 以外のモデルに影響を与えないようにするために,

$$\sum_{m'=s1',s2'} p(m'|X;\theta^t) = \sum_{m=s1,,s2} p(m|X;\theta^*)$$

を満たすように値を修正する．このようにすることで,s1,s2 以外のモデルに影響を及ぼさないままに再学習を行うことができる.

最後に，得られたパラメータを初期値として，もう一度通常の EM アルゴリズムを行う．

再学習が終了した時点の情報量基準の値が前よりも大きくなっていれば，新しいクラスターを採用し，そうでなければ再学習前に戻り，別のクラスターの分割を試みる．以上を繰り返し，分割すべきクラスターがなくなった場合，アルゴリズムを終了する．

5. 多次元正規分布における x-EM アルゴリズム

多次元正規分布に対し，x-EM アルゴリズムを適用する場合を考える．

要因数 k の多次元正規分布は，以下のように表現できる．

$$p(x|\theta) = \sum_{i=1}^{k} \pi_i N(x|\mu_i, \Sigma_i)$$

$$\sum_{i=1}^{k} \pi_i = 1,\ \pi_i > 0$$

$$N(x|\mu_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma_j)}} exp(-\frac{1}{2}(x-\mu_j)^T {\Sigma_j}^{-1}(x-\mu_j))$$

ここで，$\mu_i, \sum_i, \pi_i$はそれぞれ正規分布の平均,分散共分散行列,混合比を表す．また，iはデータを生成した分布のインデックスに当たる変数であるが，これは直接観測することができないため，隠れ変数である．

これらのパラメータをまとめて，

$$\theta = \{\mu_i, \Sigma_i, \pi_i | i = 1, \dots k\}$$

と表す． このθを x-EM アルゴリズムで推定する．

このとき，混合正規分布$p(x|\theta)$は,

$$p(x|\theta) = \sum_{i=1}^{k} p\ (x, i|\theta)$$

と表せるので，EM アルゴリズムの E ステップと M ステップは以下のようになる．

E ステップ

・データx_iが分布f_jから得られる確率z_{ij}は，以下のように求められる．

$$z_{ij} = \frac{p_j f_j(x_i)}{\sum_{k=1}^{K} p_k f_k(x_i)}$$

$$f_j = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma_j)}} exp(-\frac{1}{2}(x-\mu_j)^T {\Sigma_j}^{-1}(x-\mu_j))$$

M ステップ

各パラメータは，以下の式により推定することができる．

・平均μ_j(j=1,2,⋯k)

$$\mu_j = \frac{\sum_{i=1}^{N} z_{ij} x_i}{\sum_{i=1}^{N} z_{ij}}$$

・分散共分散行列Σ_j

$$\Sigma_j = \frac{\sum_{i=1}^{N} z_{ij}(x_i - \mu_j)\ (x_i - \mu_j)^T}{\sum_{i=1}^{N} z_{ij}}$$

・混合比π_j

$$\pi_j = \frac{1}{N}\sum_{i=1}^{N} z_{ij}$$

また，クラスター分割の際の初期パラメータと制約条件は，以下のようになる．

s1,s2 の初期パラメータ

$$\mu_{s1} = \mu_p + \epsilon,\ \mu_{s2} = \mu_p - \epsilon,$$

$$\Sigma_{s1} = \Sigma_{s2} = \frac{1}{2}\Sigma_p$$

$$\pi_{s1} = \pi_{s2} = \frac{1}{2}\pi_p$$

制約条件

$$p^*(i|x_i, \theta) = \frac{p(x_i, i|\theta)}{\sum_k^{s1,s2} p(x_i, k|\theta))} p(p|x_i, \theta)$$

6. 実験

6.1 準備

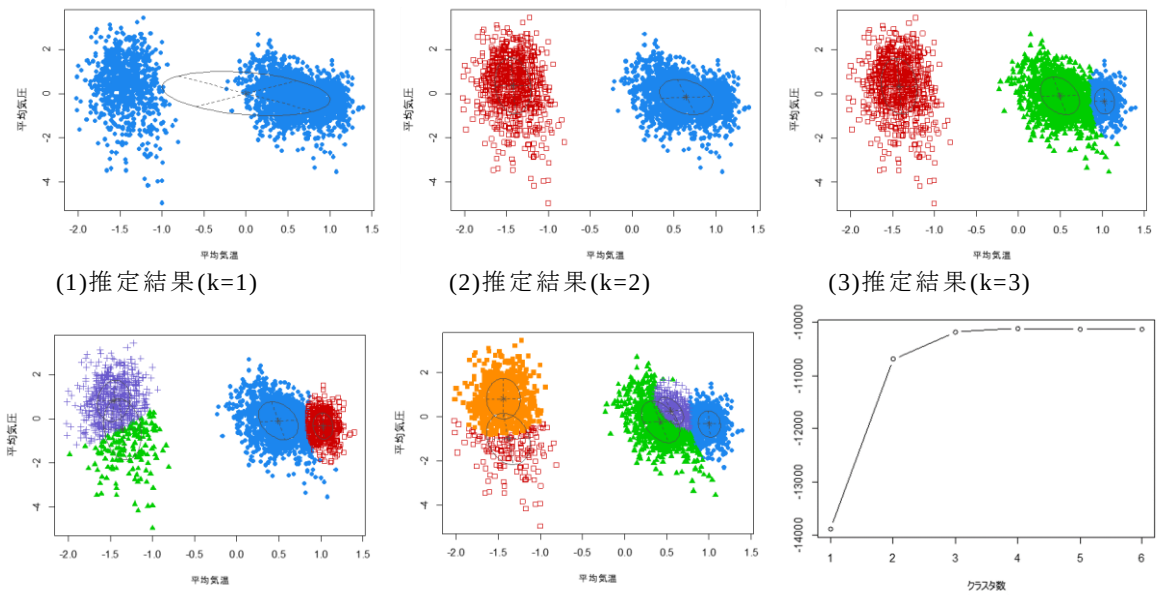
気象庁データから提供されている天気データから，1987 年～2016 年の 30 年分のデータを使用する．

このデータの次元数は 2(平均気温・平均湿度)であり，要因数=3 となるよう，3 都市(札幌・東京・大阪)のデータのみを抽出する．

それぞれの都市ごとのパラメータは，以下の表 1 の通りである．また，図 1 は都市ごとに分割した状態の実験データのプロットである．

表 1. 各都市のパラメータ

平均	平均気温	平均湿度
大阪	0.377	0.121
北海道	−1.425	0.348
東京	0.851	−0.375
分散共分散行列		
大阪	0.033	−0.014
	−0.014	0.687
北海道	0.045	−0.04
	−0.04	1.631
東京	0.055	−0.016
	−0.016	0.489
混合比		
大阪		0.302
北海道		0.31
東京		0.385



(1)推定結果(k=1)　(2)推定結果(k=2)　(3)推定結果(k=3)

(4)推定結果(k=4)　(5)推定結果(k=5)　(6)BIC の変化

図 2　x-EM アルゴリズムによる推定の様子

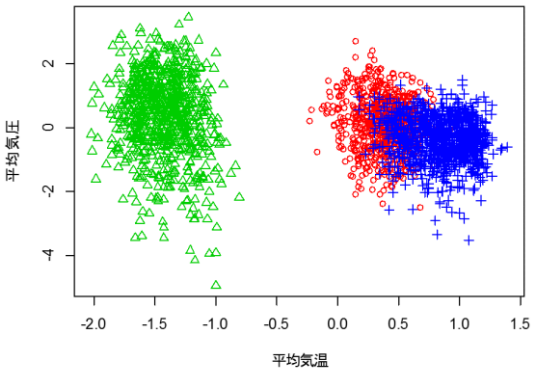


図 1　都市ごとに分割した実験データ

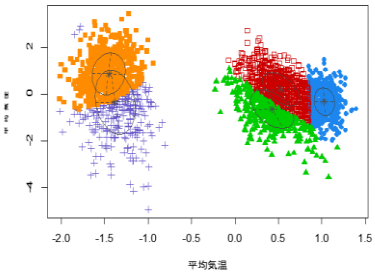


図 3　EM アルゴリズムと発見的手法による推定の結果

表 2　クラスターと都市のクロス表

	大阪	北海道	東京	合計
クラスタ1	0	493	0	493
クラスタ2	0	277	0	277
クラスタ3	1	0	513	514
クラスタ4	749	0	444	1193
合計	750	770	957	2247

表 3　アルゴリズムごとの実験結果

	x-EMアルゴリズム	EMアルゴリズム+発見的手法	x-means
混合数	4	5	2
BIC	-10114	-10143	-10355
実行時間(s)	5.24	6.8	0.21

6.2　評価方法

EM アルゴリズムにおいては，対数尤度$logL$

$$logL = \sum_{j=1}^{k}\sum_{i=1}^{n} logp_j(x_i)$$

を用い，値が大きいほどデータの当てはまりが良いとしている．これは，モデルに含まれるパラメータの数が一定の場合は信頼できるが，パラメータの数が変動する場合には，複雑なモデルの方が大きい値になりやすいということが知られている．そのため，今回の場合には評価法として使うには不適である．

一方，情報量基準の 1 つである BIC には，パラメータ数に応じた罰則項が付加されており，BIC がその導出過程において,指数型分布族の選択を考えているため，これを用いることにした．

$$BIC = 2logL - klogn$$

(k:モデルのパラメータ数,n:データ数)

また，ベースラインとして，x-means 法と EM アルゴリズムの発見的手法のそれぞれで実験データのモデル推定を行う．

6.3 実験結果・考察

x-EM アルゴリズムを用いた，2 次元正規分布データのモデル推定の結果，要因数は k=4 と推定された．推定の様子を，図 2 に示す．

図 3 に示した，EM アルゴリズムと発見的手法による推定に近い結果が出ていることが分かる．

表 2 は，クラスターごとに含まれるデータを都市ごとにまとめたものである．

クラスター4 において，大阪のデータが約 63%，東京のデータが約 37%と区別できずに混ざっているが，個々のデータの事後確率を見ると，大阪と東京が 45%～55%とほぼ同確率となっているものが多く，これはクラスター分けする際に単純に最も事後確率が高いクラスターに所属させるようにしたことで生まれたものだと考えられる．ファジイ分割の特性を活かし，個々のデータにおける混合比の大きさによるクラスター分割を行うことで，さらに特徴のつかみやすいクラスタリングが可能になると考える．

表 3 には，アルゴリズムごとの結果の要因数，BIC，実行時間を示す．

ここから，x-EM アルゴリズムは，EM アルゴリズムと発見的手法による手法と同程度のクラスタリングを，より短時間で行うことができることがわかる．実行時間の面で言えば，x-means 法の方がはるかに早いが，x-means では確率モデルに基づくクラスタリングはできないため，今回のような混合分布を仮定した場合では，提案手法の方が有利であるといえる．

今回の手法では，一度分割したクラスターは併合されることなく最後まで残り続けるため，初期値に強く依存してしまっている．この点を改善するため，今後はさらに併合の操作を加えるなどし，初期値に依存しないアルゴリズムを作ることが課題である．

7. 結論

本論文では，混合モデルにおいて，要因数が未知の場合でも EM アルゴリズムの逐次繰り返しにより分割し最適な要因数を推定するアルゴリズムを提案した．

混合正規分布を用いた実験では，発見的手法と比較し，BIC で 0.2%の向上,実行時間で 22.9%の短縮と，良好な結果を得た．このアルゴリズムは，観測不可能な潜在変数に確率モデルが依存する場合ならば混合分布に限らず適用可能である．

EM アルゴリズムに存在した初期値依存性はそのままであるため，分割だけでなく併合の操作を導入するなど，局所最適性の解決が今後の課題である．

参 考 文 献

[1] 宮元定明“クラスター分析入門―ファジィクラスタリングの理論と応用”森北出版 1999

[2] 上田修功，中野良平.“併合・分割操作付き EM アルゴリズムとその混合分布推定への応用”，TECHNICAL REPORT OF IEICE，NC97-141,pp.17-24，1998.

[3] Dempster,A.,Laird,N.&Rubin,D. “Maximum likelyhood from incomplete data via the EM algorithm”, Journal of the Royal Statistical Society B 39, pp.1-38,1977

[4] Pelleg D., Moore A., “X-means: Extending K-means with Efficient Estimation of the Number of Clusters”, Proc. of the Seventeenth International Conference on Machine Learning (ICML2000), pp.727– 734, 2013.

[5] 濵砂 幸裕,遠藤 靖典“ファジィな分割に対する妥当性基準を用いた x-means について”，日本知能情報ファジィ学会 ファジィシステムシンポジウム講演論文集 31(0), 99-100, 2015

[6] 石岡恒憲”クラスター数を自動決定する k-means アルゴリズムの拡張について”,応用統計学 29(3), 141-149, 2001

[7] 中野良平“ニューラル情報処理の基礎数理” 数理工学社, 2005

[8] 安福友浩，吉岡琢，石井信，伊藤実.”高次元データにおける階層的クラスタリング”，電子情報通信学会総合大会講演論文集 2000 年.情報・システム(2), 192, 2000

[9] MacQueen, J.B. “Some methods for classification and analysis of multivariate observations”, Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability 1, University of California Press. ,1967

[10] J.C.Bezdek,“pattern recognition with fuzzy objective function algorithms”,Plenum Pless,1981

[11] 赤穂昭太郎,”EM アルゴリズム:クラスタリングへの適用と最近の発展”日本ファジィ学会誌,Vol12,No5,pp594-602,2000