社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－MPS－67 (13)
2007－BIO－11 (13)
2007／12／20

擬似的最尤推定を用いたEMアルゴリズムの混合コーシ分布問題への適用

石樽 彩乃
お茶の水女子大学大学院人間文化研究科

不完全データの解析手法の代表的なものとしてEMアルゴリズムがある．本研究では，EMアルゴリズムのMステップにおいて，最尤解が陽に求まらない分布の典型として混合コーシ分布を取りあげる．以前に提案した，中央値と四分位偏差でもってMステップを擬似的最尤推定に置き換えた手法の有効性は，分布数3以下の場合には示されている．ここでは，さらに分布数が多い場合においてこのアルゴリズムの適用性を検証する．

An application of the modified EM algorithm by pseudo ML to mixture Cauchy distribution

Ayano Ishigure

Ochanomizu University
Graduate school of Humanitics and Sciences

The EM algorithm is known as one of tools for the data analysis of incomplete data set. In this study we shall take up mixture Cauchy distribution as a typical model that is quite difficult to estimate the parameters on the maximization step (M-step) of the EM algorithm. We gave the modified EM algorithm for mixture Cauchy distribution, in which the maximum likelihood (ML) estimators of parameters for Cauchy distributions on the M-step are replaced by the median and quartiles. We have also seen that this method can be applicable to the case where the number of mixed distributions is smaller than or equal to three. Here as further study, we shall see the effectivity of our method in case of increasingly more mixed distributions.

1 はじめに

実世界から得られるデータは，常に完全ではなく，ある変量に対しデータが得られていないような欠測値を含んでいる場合も多い．比較的単純な統計モデルにおいても，不完全データの存在により解析が困難になってしまう．本研究で扱う混合分布問題も，どの分布から生成されているかという情報が欠測した不完全データと考えることができる．このようなデータの解析には欠測値を考慮に入れた手法が必要となる．その代表的な手法としてEMアルゴリズムがあげられる．[1]

EMアルゴリズムは一般に，混合正規分布のパラメータの最尤推定で広く用いられる．しかし，密度関数がわかっていればどのような混合分布にも理論的には応用は可能である．本研究では混合コーシ分布への応用を考えた．しかし，コーシ分布の場合，Mステップにおける尤度方程式は，解析的に解けないため，最尤推定値を求めることは困難である．そこで，コーシ分布のパラメータの最尤推定に代わる方法を提案し，この手法をMステップに用いた改良EMアルゴリズムについての考察を行ってきた．[2] なお，実験より，分布数3までの混合分布問題においてはこのアルゴリズムが十分に最尤推定の代用として用いることが出来るものと考えられる．実際に，分子分光スペクトルの分野では多数のコーシ分布の効率的な分解手法が求められている．現在，波形分離に使われている手法は，最小二乗近似を用いたもので，密度関数の重ね合わせをフィッティングして近似するというものである．この手法は多次元の最適化を行うため，計算に時間がかかるという難点がある．[3] それに対し，今回提案する手法は，観測されるスペクトル曲線を十分近似する乱数ヒストグラムを用意し，EMアルゴリズムにより確率的

にパラメータを推定するものである．EMアルゴリズムを基にしているので，比較的計算が単純で時間がかからないという長所を受け継いでいる．そこで本研究ではさらに分布数が多い場合への適用性を検証する．

2 混合分布問題

K 個のクラスからなる混合分布とは，k 番目のクラスの確率密度関数を $f_k(x|\theta_k)(k=1,\cdots,K)$，混合比を $p_k\,(k=1,\cdots,K)$ とするとき，確率密度関数が

$$f(x|\theta,p)=\sum_{k=1}^{K}p_k f_k(x|\theta_k)$$

と与えられるような分布である．ただし，$p_k\geq 0$ かつ $\sum_{k=1}^{K}p_k=1$ である．

混合分布問題とは，ここからの標本 $\{x_1,\cdots,x_N\}$ が与えられたとき，パラメータ $\{\theta_1,\cdots,\theta_K\}$ と混合比 $\{p_1,\cdots,p_K\}$ 及び分布数 K を推定することである．ここでそれぞれのデータが属しているクラスを表す潜在変量として $z_i=(z_{i1},\cdots,z_{iK})$ を導入し，x_i がクラス k からの標本であるとき，z_{ik} を 1，そうでなければ 0 とする．すなわち理想的な完全データを (x_i,z_i) とし，観測データ x_i を z_i が欠測した不完全データと考える．

分布を特徴付けるパラメータ θ_k と混合比 p_k を推定する方法として EM アルゴリズムを用い，最適分布数 K を推定する方法としては，統計モデルの選択基準 BIC を用いることにする．[4]

3 EMアルゴリズム

EM アルゴリズムは，不完全データに関する様々な処理法を 1977 年に Dempster らが，統一的にまとめたものであり，その基本的発想は，観測された不完全データをいったん扱いやすい擬似的完全データに置き換え (Expectation Step)，この擬似完全データからパラメータの最尤推定値を求め (Maximization Step)，さらに得られたパラメータの推定値から再び完全データを作り直し，それよりまたパラメータの推定値を求め直すという手続きを反復するというものである．

3.1 コーシ分布

コーシ分布は，種々の放射線の線スペクトルの強度分布など共鳴現象を表すのにしばしば用いられており，原子核物理の分野では，ローレンツ分布とも，ブライト・ウィグナー分布とも呼ばれる．

その確率密度関数は

$$f(x)=\frac{1}{\pi}\cdot\frac{c}{(x-m)^2+c^2}$$

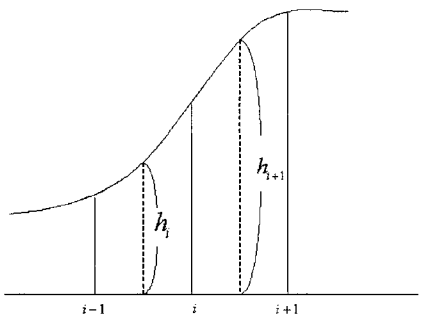
と与えられる．

正規分布が頂点が丸く，裾の減退が速いのに対し，コーシ分布は頂点が鋭く，裾が重たい．また，各モーメントも発散するため，コーシ分布は平均や分散ではなく，中央値と四分位偏差を以って，分布のパラメータ m, c を特徴づけられる．

実際，データから分布のパラメータ m, c の最尤推定値を得ることは非常に困難である．そこで，m を中央値，c を四分位偏差として考え，これらの順序統計量の推定値でもって m, c の最尤推定の代わりとする手法を提案した．

3.2 標本の発生

計測器からの観測データは，曲線として出力されるものとする．そこでまず，概形をみて乱数データを発生させる区間を決定する．次にその区間を適当な単位で n 等分割し，各分割区間の高さを求める．



$h_1,\cdots,h_n$ の和を S とおき，区間 $[i-1,i]$ で発生させる乱数の個数を，

$$N\times\frac{h_i}{S}$$

とする．ただし N は用いる乱数の総数である．また，ここでは適当なスケーリングを行い，各区間幅は 1 とした．

3.3 混合コーシ分布におけるEMアルゴリズム

N 個の標本 $\{x_1,\cdots,x_N\}$ が与えられ，混合分布が K 個のコーシ分布からなるとしたとき，各データの欠測値 z_{ik} とそれぞれの密度関数のパラメータ m_k, c_k と混合比 p_k を次のように逐次推定する．

1. 初期設定
   観測値 $x_i\,(i=1,\cdots,N)$ に対し，どの分布からの標本であるかをランダムに決める．すなわち，各 i に対して z_{i1} から z_{iK} のうちどれか一つのみを 1，それ以外を 0 とする．
2. M ステップ (m_k, c_k, p_k の推定)．

   m_k :分布 k の中央値

   c_k :分布 k の第 1 四分位と第 3 四分位の中央値からの距離の平均 (四分位偏差)

   $$p_k = \frac{1}{N}\sum_{i=1}^{N} z_{ik}$$

3. E ステップ (z_{ik} の推定)．

   $$z_{ik} = \frac{p_k f_k(x_i; m_k, c_k)}{\sum_{k=1}^{K} p_k f_k(x_i; m_k, c_k)}$$

   $f_k(x; m_k, c_k)$:クラス k の確率密度関数

以下，上記の M ステップ，E ステップを収束するまで繰り返す．

4 分布数の推定

分布数を推定する方法として，BIC(ベイズ情報量基準) を用いる．BIC は，複数個のモデルの中から観測データに最適なモデルを選択する評価基準のひとつで，次のように表される．

$$BIC = -2\sum_{i=1}^{N} \log f(x_i) + m \times \log(N),$$

f :推定された混合分布の確率密度関数，

m :自由パラメータ数

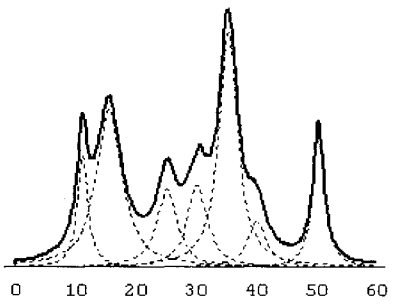
この BIC の値を最小とするモデルが最適なモデルと考えられる．コーシ分布の混合分布では，分布数 K の場合には自由パラメータ数 m は，$3K-1$ となる．

5 実験

5.1 多分布での数値実験

分布数 7 の場合

下図は分布数 7 の場合の実験例である．実線にしたがうような 10000 個の標本を用意し，パラメータ推定を行った結果が，破線である．



このとき，BIC の結果も，やはり分布数 7 のときに最小値をとった．

分布数	BIC
5	77620
6	77541
7	77328
8	77354
9	77365

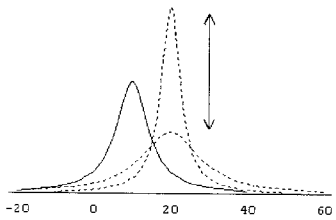
分布数が多くても，かなり分離が可能であることがいえる．ただ，様々な数値実験から分離が困難な場合として，c の値が大きい，つまり裾が広い場合があげられる．逆に，含まれる割合が少なくても，c の値が小さい場合は正常な分離が行われた．

5.2 定量実験

前項で述べた，c の値が大きい場合の分離の困難性について分布数を 2 とし，以下のような条件下で数値実験による定量的な検証を行った．

分布 1(実線) : $m = 10.0, c = 5.0$

分布 2(破線) : $m = 20.0, c = 1.0 \sim 11.0$

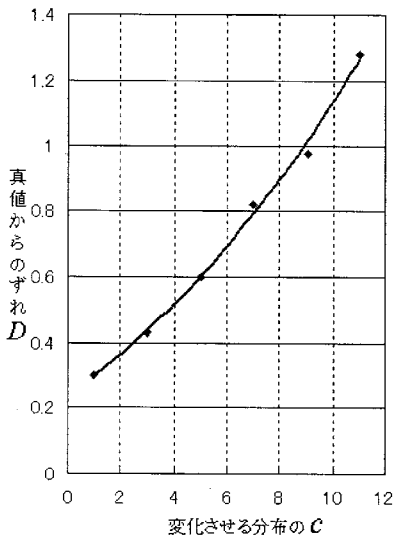


分布1は固定し，分布2はcの値を1.0～11.0まで変化させ，これらを同比率で混合させ，cの値と推定されたパラメータの真値からのずれを調べた．なお，真値からのずれを測る指標として，以下のDを導入した．これはコーシ分布の密度関数の形状から面積の関係に着目したものである．

$$D = \widetilde{p_1}|c_1 - \widetilde{c_1}| + \widetilde{p_2}|c_2 - \widetilde{c_2}|$$

ここで，c_1, c_2は真値，$\widetilde{p_1}, \widetilde{p_2}, \widetilde{c_1}, \widetilde{c_2}$は，推定されたパラメータである．

そして次にその結果を示す．



実験結果について

cの値が大きくなるにつれて比例よりも大きい速度で真値からのずれが大きくなることが読み取れる．また，数値実験よりDが大きい場合は，cの値が小さく推定され，その分，比率を大きくすることでもって真の混合分布に近づくような推定が行われる傾向があることもわかった．今回は分布数を2としているため，結果として合成される分布は真値による混合分布に近い形であらわされるが，分布数が多くなると裾の重たさが正確な分離を妨げる非常に大きな要因になってくるものと考えられる．

6 まとめ

本研究では，EMアルゴリズムのMステップを中央値と四分位偏差で代用するアルゴリズムの多分布への適用性を検証した．実験の結果から，分布数が大きい場合においてもこれらが十分にパラメータの最尤推定に置き換えることのできるものであると考えられる．

また，以下の二点において本アルゴリズムの実データへの有効性がいえる．第1に，標本の発生方法が波形として得られる実データへの解析にも応用可能だということである．そして，第2に実際の計測器からでてくるARモデルのスペクトル分布関数においては，ピークが比較的cが小さい分布の重ね合わせで現れてくる．これらの点においてもこの手法は有効ではないかと思われる．

参考文献

1. 麻生英樹，津田宏治，村田昇, パターン認識と学習の統計学，岩波書店 (2003).
2. 石樽彩乃，吉田裕亮，EMアルゴリズムの混合コーシ分布への応用とその改良，情報処理学会研究報告，2006-MPS-62，pp.85-88 (2006)
3. 南茂夫，科学計測のための波形データ処理，CQ出版社 (1986).
4. 下平英寿，伊藤秀一，久保川達也，竹内啓, モデル選択, 岩波書店 (2004).