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

2006－MPS－62 (22)
2006－BIO－7 (22)
2006／12／22

EMアルゴリズムの混合コーシ分布への応用とその改良

石榑 彩乃, 吉田 裕亮
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混合分布問題として，混合コーシ分布を考え，分布を特徴づけるパラメータである，中央値と四分位偏差によりMステップを擬似的最尤推定に置きかえた手法を提案する．EMアルゴリズムにおいて，分布数は既知であることが前提となっているため，分布の推定にはよく知られたAICを用いる．また，混合分布数を2とし，KL情報量により，真の分布と推定されたモデルとの距離を測り，2つの分布の分解能に関する数値実験を行った．

An application of the improved EM algorithm to mixture Cauchy distributions

Ayano Ishigure, Hiroaki Yoshida

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 give a technical method in the maximization step of the EM algorithm for the problem of mixture Cauchy distributions. It is quite difficult to estimate the parameters for a Cauchy distribution from given sampling data in maximum likelihood (ML), explicitly. Instead of ML estimator, we will use the median and the quartile, and estimate them by using the bootstrap method. We shall also give some numerical experimentation for the mixture of two Cauchy 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．

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，EMアルゴリズムの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た．

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，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である AIC を用いることにする．

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．

本研究では，各分布がコーシ分布である混合分布問題を考える．

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，すその減退が速いのに対し，コーシ分布は頂点が鋭く，すそが重たい．正規分布で値が，平均から標準偏差 σ 以内に存在する確率は約 68%，2σ 以内にある確率は約 96% であるのに対し，コーシ分布では，$[m-c, m+c]$ に存在する確率がちょうど 50% で，また $[m-2c, m-2c]$ に約 70%．さらに，$[m-3c, m+3c]$ でも約 80% であり，正規分布よりもかなり裾が厚いことが分かる．また，各モーメントも発散するため，コーシ分布は平均と分散ではなく，中央値と四分位偏差を以って，分布のパラメータ m, c を特徴づけられると考えられる．

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，本研究では上記のコーシ分布において区間 $[m-c, m+c]$ における出現確率が 50% であるという特性に着目して，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する．

3.2 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)．
   (x_i, z_i) を完全データとして，m_k, c_k, p_k を推定する．

   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

一つ前のサイクルで得られたパラメータの推定値から各データの欠測値 z_{ik} を期待値で補うことで推定する.

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

ただし, $f_k(x; m_k, c_k)$ はクラス k の確率密度関数である.

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.

本研究では, 順位統計量である中央値, 四分位偏差の推定値をより正確に推定するためにブートストラップ法を援用することにした.

3.3 ブートストラップ法

ブートストラップ法とは, 経験分布に基づいて, 大量にリサンプリングした標本から, ある統計量の分布を推定し, 推定値の誤差評価をしたり, 区間推定を行う手法の一つである.

本実験におけるブートストラップ法の手順

1. 各分布ごとに $z_{1k}, \ldots, z_{Nk}$ に基づきランダムに重複も含めてリサンプリングする. ただし, リサンプルの大きさは元の標本の大きさと同じとする.
2. リサンプリング標本から中央値, 四分位偏差を推定し $(T_m^{(j)}, T_c^{(j)})$ とする.
3. このプロセスを J 回繰り返すと, それぞれ J 個の推定値 $(T_m^{(1)}, T_c^{(1)}), \ldots, (T_m^{(J)}, T_c^{(J)})$ が得られる. その平均をとり, m_k, c_k の推定値とする.

4 分布数の推定

分布数を推定する方法として, AIC(赤池情報量基準) を用いる. AIC とは, 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で, 次のように表される.

$$AIC = -2 \times (\text{モデルの最大対数尤度}) + 2 \times (\text{モデルの自由パラメータ数})$$

よって

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

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

が, モデル選択の基準であり, この AIC を最小とするモデルが最適なモデルと考えることができる. コーシ分布の混合分布では, 分布数 K の場合には $m = 3K - 1$ となる.

5 分布間の距離

本研究では, 正規分布の混合の場合と比較して, コーシ分布の場合にどの程度の分解能があるかを, 数値実験的に調べるために混合条件を変化させ, 推定された分布が真の分解に一致するかを分布間の距離を用いて判断することとした. 分布間の距離としては, Kullback-Leiber (KL) 情報量を用いた. KL 情報量は真の分布を $p(x)$, 観測データから推測されたモデルの分布を $q(x)$ とする. このとき, 真の分布 $p(x)$ から見たモデル $q(x)$ のずれを測るための KL 情報量とは

$$D(p, q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

で与えられる. この KL 情報量には次のような性質がある.

$$D(p, q) > 0,$$
$$D(p, q) = 0 \quad (p(x) \equiv q(x))$$

KL 情報量がある値より小さいとき推定したパラメータが真の分布に近くつまり, 分解ができたと判断した. もちろん, 混合された元の分布間が近いほど, 分解が困難になることは容易に予想される.

6 実験

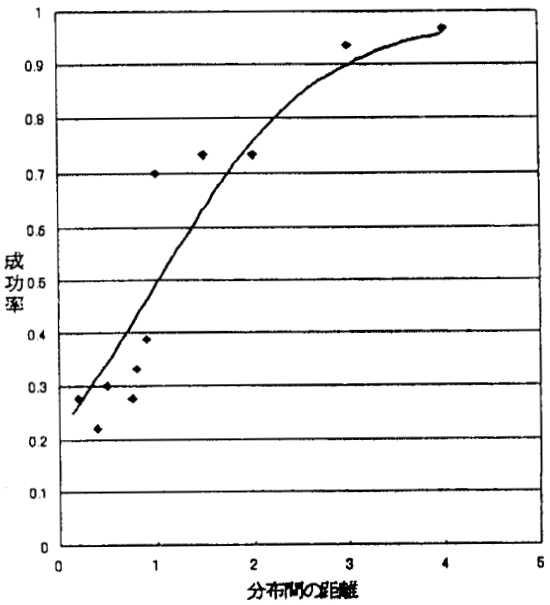
6.1 分解能の実験

本実験では, 2つの分布が 1 : 1 の比率で混合している状況を設定し, 混合される 2 つのコーシ分布のパラメータ m の差と分解能の関係を調べた.

次のような条件で乱数データを用意する.

- データ数は 1000
- パラメータ $c = 1.0$ に固定
- m の値を変化させ，成功率の変化をみた．

以下に，結果のグラフを示す．



6.2 実験結果について

グラフから，2 つの分布のパラメータの差が 2 よりだんだんと分離が困難になることがわかる．グラフに入れた曲線はロジスティック曲線である．一般に 0 か 1 の 2 値の反応を起こす現象における比率の観測データの解析にはロジスティック回帰が多く用いられる．例えば，ある薬物による生物の生死による致死率などが代表的なものである．回帰モデルとしては比率 Y のオッズ比の対数を独立変数 X で線形回帰する．すなわち，

$$\log\left(\frac{Y}{1-Y}\right) = \beta_0 + \beta_1 X$$

である．

コーシ分布の場合の，下がり具合 (係数 β_1 が対応する) は正規分布の同様な実験 [1] と比較して，緩慢であることがわかった．これはコーシ分布は裾の部分に値が多く存在し，データにばらつき大きい現れるためと考えられる．

7 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を提案した．実験の結果からこれらが十分にパラメータの最尤推定に置きかえることの出来るものであると考えられる．

また，分解能を見るため，分布数を 2 に特定して，数値実験を行い正規分布の混合の場合と，ロジスティック回帰による比較を行った．

今後の課題として，様々な条件で実験を行い，詳細な評価を行う必要があると考えられる．また，実データへの応用は以上に重要な課題となると考えられる．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 (ローレンツ分布) の効率的な分解手法が求められている．

参考文献

1. 奥律子，樺沢由希子，EM アルゴリズムと情報量基準の混合分布問題への応用，お茶の水女子大学 理学部・情報科学科 卒業研究 (2002).
2. 下平英寿，伊藤秀 ·，久保川達也，竹内啓，モデル選択，岩波書店，東京 (2004).
3. 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，東京 (2003).