ISS-A-058

一般 EM 法による混合分布の動作検証

新浪　有茄　　三浦　孝夫
法政大学理工学部創生科学科

1.前書き

本研究では、指数分布や正規分布を含んだ混合分布を、EM アルゴリズムを用い精度よく推定を行う。そしてその精度と推定速度から、 一般 EM 法の有用性を検証する。

2.混合分布推定

混合分布を推定するためには、内部にある分布の各パラメータ、平均や分散などの推定はもちろん、各分布の混合率を推定する必要がある。そのため混合率を求めることのできる EM 法を用いることで、各パラメータ推定に加えて、混合率を得ることも可能にする。

EM アルゴリズムはすべてのクラスタの場合について確率を重みにして尤度と足し合わせたもの、対数尤度を最大化するパラメータθを推定していくものである。

3.実験

誤差 30%以下である場合は有意性があるとし、精度が高いとする。EM 法による分布推定までの速度は、収束までの EM 回数とし、その値同士を比較する。本研究では、実験データとしてデータ数 50000 件ある指数分布、正規分布を各 2 つずつ用意する。指数分布 1 はλ1＝2、平均値μ =0.862 、分散値σ 2 =0.233、指数分布 2 はλ2＝1.5である。正規分布 1 は、$\mu = 0.862$ 、$\sigma^2 = 0.233$、正規分布 2 は、$\mu 2 = 0.862$ 、$\sigma 2^2 = 1$ である。上記の実験データより、各指数・正規分布データから 20000 件抽出した分析データ群を 5 つ作成する。

まず初期値を与えて、期待値を最尤推定する。次に求まった値を最大化し、パラメータ混合率π'、平均μ'、標準偏差σ'、発生率λ'を推定する。その後推定した値を初期値にいれ、上記の行程を収束するまで繰り返す。本研究では EM 法の収束条件を 3 ステップ以上所属確率に変動がない場合とした。

5.実験結果と考察

内分比の異なる指数-正規分布データの、全パラメータ数に対する精度が高いパラメータの割合は、1 対 1、1 対 3、3 対 1 の順に、75%、50%、25%となっている。よって、1 対 1 の時、最も精度が高くなっている。収束平均に対するパーセント誤差の最大は 7%で、10%より少ないという結果から収束回数はほぼ等しいとみなす。　　正規-正規分布と指数-指数分布データの収束後の値は正規-正規分布の混合率以外パーセント誤差が 30%以上であり、精度が高いといえない。

表.1　収束後のパラメータ

	EM(40)		EM(67)
平均1	0.469895	λ1	1.220317
標準偏差1	0.193496	λ2	1.220317
平均2	0.468859		
標準偏差2	0.193126		

表.1 は正規-正規分布と指数-指数分布での EM 法収束後のパラメータ値である。内部分布の各パラメータをパーセント誤差で求めたところ、最大 0.2%であり、同じ分布とみなせる。

表.2　収束までのステップ数

	正規-指数	正規-正規	指数-指数
収束(回)	29	40	67

表.2 は正規-指数分布、正規-正規分布、指数-指数分布の収束までのステップ数を表した表である。正規-指数分布の収束回数に対して、正規-正規分布は＋11 回の 1.3 倍あり、指数-指数分布は＋38 回の 2.3 倍あった。

正規-正規分布、指数-指数分布のパラメータほとんどが、実際の値と 30%以上も誤差があることについては、表.5 より EM によって求まった内部の 2 つの分布は等しいことが原因ではないかと推察する。では次になぜ得られた分布が等しくなってしまったのか。それは分析データを抽出した実験データの分布同士が酷似していることが原因ではないかと考える。今回実験データを用意する際、同じ分布同士は変更するパラメータの誤差が±1 以内になるように設定し作成した。それにより、分類するのが困難になっていたのではないかと考えられる。

7.結論

実験結果の中で内分比 1 対 1 の指数-正規分布の場合が、一番パラメータ数に対して、精度が高い、または非常に高いパラメータが 75%と多く、最も精度が高い結果が得られた。また、分析データを抽出する実験データ同士が酷似すると、EM 法によって求まった正規-正規分布、指数-指数分布内のパラメータ同士がパーセント誤差、最大 0.2%で内部にある分布同士が等しいという結果になる。

次に EM 法推定までの速度については、指数-正規分布は内分比に関係なく、パーセント誤差の最大は 7%で収束速度はほぼ等しかった。一方、混合分布が異なるもの同士での速度については正規-指数分布に対して正規-正規分布が 1.3 倍、指数-指数分布が 2.3 倍収束まで時間がかかり、混合分布の中身が同じである分布の方が、推定速度が遅いと推察できた。