混合分布モデルの推定結果の記述統計的表現の工夫

中村　永友[1]

要　旨

ボックスプロットは提案されてから40年以上たつが，この記述統計的なデータの表現方法は今やデータ解析のための基本ツールである．１次元データの分布状況をヒストグラムを描かずに，その散布状況をある程度把握できるというメリットがある．この論文は，混合分布モデルをあてはめるような複峰性を有するデータにこのモデルをあてはめて，各成分分布をボックスプロットで表現する方法を提案する．

キーワード：混合分布モデル，箱ひげ図・ボックスプロット，パーセンタイル，四分位数，推定値，事後確率

1　はじめに

箱ひげ図＝ボックスプロットは四分位数と共にTukey（1977）が40年以上前に提案した（Wickham and Stryjewski, 2011）．コンピュータによるデータ解析が今や常識となり，この記述統計的なデータの表現方法は今や誰しもが知るツールである．ボックスプロットは１次元データの分布状況をヒストグラムを描かずにある程度把握できる（想像できる）というメリットがある．しかしながら，これらはデータの分布状況が単峰性を前提としたものであり，複峰性を有する場合には，必ずしも四分位数からその様相を想像することは困難であるというデメリットもあるが，とりあえずデータ分析の第１歩では非常に有用なツールである．

では複峰性を有するとわかったデータに対しては，クラスタリングなどの手法で排反にグルーピング（分類）され，各クラスターの分析が進められる．一般的なクラスタリングの手法ではその問題とは関係のないと思われる基準によって分類が行われることもあるため，その分類結果が不自然になることも否めない．これに対して混合分布モデル（中村他，2005）によるクラスタリングは，このモデルがオーバーラップを許容するため，柔軟な分析をすることが可能であることから，今やデータ分析の基本的な分類手法となった．このような分類結果を表現する手段は，ヒストグラムと推定結果の密度関数やその成分分布を同時に重ねて描かれることも多い．この論文は，この混合分布モデルをあてはめた分類結果に対して，各成分分布をボックスプロットで表現する方法を提案する．これは通常のボックスプロットに加えて混合分布モデルで推定した混合比率と平均，標準偏差を同時に描画する．類似の表現方法はQarmalah *et al.*（2016）が提案しているが，本提案はより見栄えがシンプルであり，単色で表現していることが特徴である．

以下，第２節では提案手法の概要，第３節では本提案の根幹となる混合分布モデルに対するパーセンタイルの定義について，第４節では基本的な統計量に対するそれらの描画上の表現方法の特色について，第５節は先行研究との違い，第６節は人工データに対しての適用例を示す．

2　提案方法の概要

ボックスプロットは四分位数を描画したものである．四分位数やパーセンタイルなどはデータを順序統計量に直し，データの個数によってそれを決めることになる．しかし，混合分布をあてはめたときは，データの各点に対して，各成分分布に対する所属確率としての事後確率が推定される．この状況は１つのデータは複数の分布で共有されることになる．この理由により，事後確

[1] 札幌学院大学 経済学部；nagatomo@sgu.ac.jp.

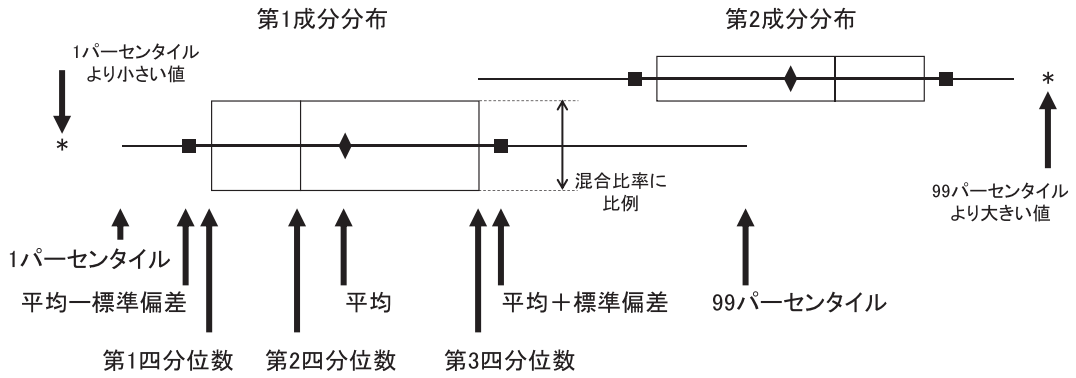

図１：提案方法のボックスプロットの要素の説明

率を負の方向から累積することで，分位点を定義する．

本提案は，推測統計によって得られた情報を記述統計の世界での表現で試みることである．本提案の概要は次の通りである．

(1) 成分分布に対応する複数のボックスプロットを，１次元的ではなく，軸をずらす．

(2) ボックスプロットの重なりがないため，単色の表現で十分である．

(3) 混合分布モデルで推定された平均と標準偏差を打点する．

(4) ひげの端点は１パーセンタイルと99パーセンタイルを表し，それより外側のデータは事後確率最大の成分のひげの外側に打点する．

3　パーセンタイルの定義

r 個の成分分布からなる混合分布モデルは

$$f(x|\Theta)=\sum_{k=1}^{r}\pi_k f(x|\theta_k)$$

として書くことができ，$\Theta=\{\pi_1, \ldots, \pi_r, \theta_1, \ldots, \theta_r\}$ である．任意のデータセットが与えられて，このモデルをあてはめて，$\hat{\Theta}$ が EM 法（Depmster *et al.*, 1977）で推定されたとする．$f(\cdot)$ に正規分布を想定するとき，パラメータの推定量は

$$\hat{\pi}_k=\frac{1}{n}\sum_{i=1}^{n}\hat{P}_{ki},$$

$$\hat{\mu}_k=\frac{1}{n\hat{\pi}_k}\sum_{i=1}^{n}\hat{P}_{ki}x_i,$$

$$\hat{\sigma}_k^2=\frac{1}{n\hat{\pi}_k}\sum_{i=1}^{n}\hat{P}_{ki}(x_i-\hat{\mu}_k)^2,$$

$$\hat{P}_{ki}=\frac{\hat{\pi}_k f(x_i|\hat{\theta}_k)}{\sum_{k=1}^{r}\hat{\pi}_k f(x_i|\hat{\theta}_k)}$$

である．

このとき，第 k 成分のみに注目して，これのみの四分位数などの記述統計であるパーセンタイルは次のように定義できる．

第 k 成分分布の c パーセンタイルを次式で定義する．$x_{(i)}$ は x_i の順序統計量，$\hat{P}_{k(i)}$ は $x_{(i)}$ に対応する第 k 成分分布の事後確率とする．このとき，第 k 成分分布の c パーセンタイルを次式で定義する：

$$Q_{kc}=\frac{x_{(h)}+x_{(h+1)}}{2}.$$

ここで，

$$h:\sum_{i=1}^{j}\hat{P}_{k(i)}<c \text{ を満足する } j \text{ の最大値.}$$

c は一般的に，$0<c\leq 100$ となる整数値を想定する．また，記述統計的には，例えば四分位数であれば，その分位数はデータの個数が奇数個であればデータの値，偶数個であればそれを挟む２つのデータの平均として定義される．ここでは，事後確率の累積によりパーセンタイルを構成するので，定義で示したように求めるパーセンタイルを挟む２つのデータの平均として定義する．第１，２，３四分位数は，それぞれ25，50，75パーセンタイルとなる．

4　表現方法

パーセンタイルが前節で定義されたので，ボックスプロットでの表現が可能になった．図１に２つの成分分布の場合の２つのボックスプロットを示す．基本的には従来のボックスプロットと大きな変更はない．混合分布モデルにより事後確率を潜在変数として混合分布を推定したので，パラメータ推定値の平均(◆)と平均±標準偏差(■)を太線でボックスを縦断して描画する．ほとんどの確率分布では平均と分散は推定できる

表 1：Qarmalah *et al.* との違い

	本提案	Qarmalah *et al.*
配色	単色	異なる成分分布は別の配色
Boxplot の重なり	なし	あり，線種と色により区別
平均・標準偏差	表示あり	表示なし
混合比率	ボックスの幅で対応	同左
事後確率	表示なし	オプションであり
軸の方向	横	縦
ヒストグラム軸との対応	あり	なし
全体的表現	2 次元的	1 軸上にすべて配置

ので，この表現は可能となる．箱の高さは混合比率と比例させ，さらに箱と箱の間はオーバーラップさせず，可能な限り隣接させる．さらに，ボックスプロットのひげの定義は様々存在するが，ここではひげの端を 1 パーセンタイルと99パーセンタイルとして，さらにその外側にデータがある場合は，アステリスク＊で打点する．

本提案の特徴は以下の通りである．

(1) 各点に対する事後確率を累積してパーセンタイルを構成する．データ点がパーセンタイルと一致する確率は非常に低いので，それを挟むデータ点の平均で求める．
(2) ボックスの幅（高さ）を π_k の大きさに比例させる．
(3) 推定された成分分布はほとんどの場合重なるので，それを考慮してボックスは重ね合わさない．
(4) ひげの端は 1 パーセンタイルと99パーセンタイル点とする．
(5) ひげの外側のデータ点は事後確率が一番大きなボックスプロットのひげの外にアステリスク＊で打点する．
(6) 成分分布の平均と分散の推定値は，ボックスの中に平均±標準偏差を線分と打点で表現する．
(7) 単色とする．
(8) ヒストグラムの横軸とボックスプロットの軸を平行にする．すなわちヒストグラムの横軸をボックスプロットの軸を共有する．

5　Qarmalah *et al.* との違い

類似の表現方法を Qarmalah *et al.*（2016）が提案している．その特徴は以下のとおりである．(1) 事後確率の累積で分位点を決めている．これは本質的に本提案と同じ考え方である．(2) 複数のボックスプロットを 1 本の軸上に乗せている．(3) 複数のボックスプロットを実線，破線，点線などで見分けられるようにしていて，さらに別の色で描くことで，区別している．(4) 混合比率はボックスの幅（高さ）を変えることで表現している．(5) R 関数を同時に提供していて，4 つの描画オプションがある．その中にデータの事後確率を描くオプションなどがある．(6) 一般的にヒストグラムの横軸はデータの値，縦軸は頻度あるいは相対度数としていることが多い，彼らの方法はボックスプロットの軸を縦にしているため，ヒストグラムと並べた場合にデータの散布状況とボックスプロットによるデータの広がり具合がわかりづらい．

Qarmalah らの提案との違いを表 1 にまとめておく．

6　数値実験

数値実験では混合分布からデータを発生させ，混合分布モデルをあてはめて，その結果に対する提案手法を提示する．

混合分布は

$$\frac{2}{3}N(0, 1)+\frac{1}{3}N(4.5, 1.5)$$

として，これからデータを150個生成し（図 2 上），混合分布モデルをあてはめて，推定されたパラメータは

$$\hat{\boldsymbol{\pi}} = \{0.68, 0.32\},\ \hat{\boldsymbol{\mu}} = \{0.03, 4.8\},\ \hat{\boldsymbol{\sigma}} = \{1.25, 2.95\}$$

であった．また，各成分分布の ｛1 パーセンタイル，四分位数，99パーセンタイル｝ はそれぞれ以下の通りである．

第 1 成分分布：$\{-2.52, -0.82, 0.06, 0.90, 2.51\}$,
第 2 成分分布：$\{0.90, 3.75, 4.46, 5.94, 8.24\}$.

これらを元にボックスプロットを描いた結果を図 2 に示す．上からヒストグラム，次の 2 つが提案したボックスプロット，一番下に全体のボックスプロットである．

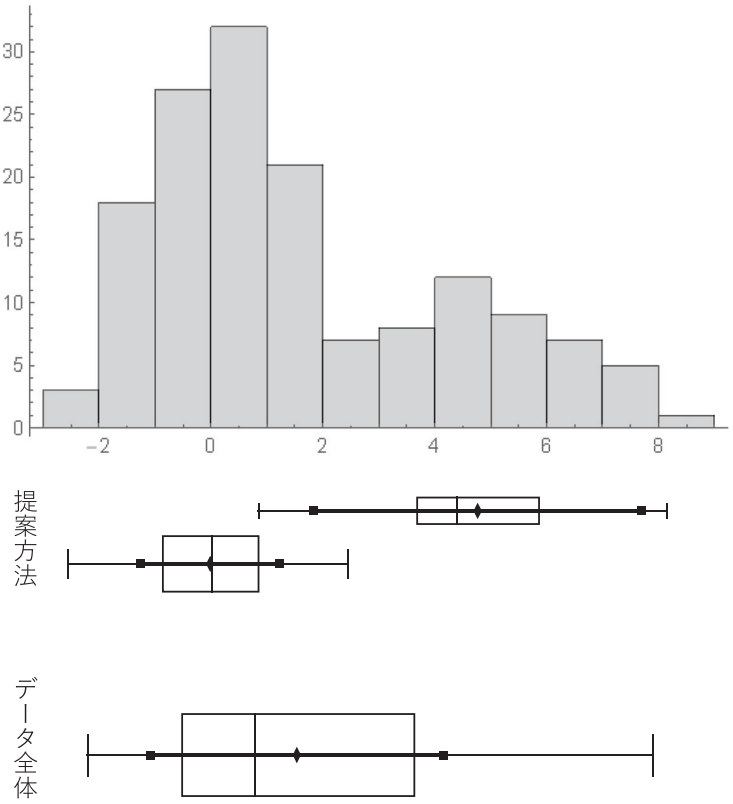

図 2：$\frac{2}{3}N(0,1)+\frac{1}{3}N(4.5,1.5)$ からデータを生成

各ボックスプロットのスケールはヒストグラムと共通

7　おわりに

Qarmalah *et al.*（2018）との違いから，本提案手法はよりシンプルな表現方法であることがわかる．とくに成分分布の重なりについては，本提案手法はより見やすくなっているが，Qarmalah らの方法は配色と点線・破線で表現していて，モノクロで表示したときには非常に見づらい．カラープリンタやカラーディスプレイを前提としているが，環境依存しない方がベターと考える．また彼らの R 関数のオプションで事後確率を出せるようになっている．これは考え方の問題かも知れないが，記述統計の表現方法の中に推測統計による推定結果をどこまで表示するかは，好みの問題かも知れないが，事後確率の表示はやり過ぎのような気もする．もし事後確率まで表示するのであれば，混合分布の曲線，成分分布の曲線，事後確率の曲線，データのヒストグラム，これらを同時に示した方がより情報が豊かになると考えられる．あくまでも記述統計の延長線での表現とするならば，本提案のように平均と ± 標準偏差程度までで良いのではないだろうか．

今後の検討課題としては，R などの関数として提供することで普及を図ることである．

参考文献

[1] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, Journal of Royal Statistical Society Ser. B, **39**, 1–38.

[2] 中村永友・上野玄太・樋口知之・小西貞則（2005）．欠損混合分布モデルとその応用, 応用統計学, 34 (2), 57–75.

[3] Qarmalah, N.M., Einbeck, J., and Coolen, F.P.A. (2018). *k*-Boxplots for mixture data, Statistical Papers, **59**, 513–528, DOI: 10.1007/s00362–016–0774–7.

[4] Tukey J. W. (1977). Exploratory data analysis, Addison-Wesley, Boston.

[5] Wickham, H. and Stryjewski, L. (2011). 40 years of boxplots, https://vita.had.co.nz/papers/boxplots.pdf.

The Proceedings of the Research Institute of Sapporo Gakuin University Vol.7, 1–5 (2020)

Descriptive Presentation for Estimated Mixture Model

Nagatomo NAKAMURA[1]

Abstract

Boxplot was proposed more than 40 years ago. This descriptive statistical data representation is now a basic tool for data analysis. There is an advantage that the distribution state of one-dimensional data can be grasped to some extent without drawing a histogram. This paper proposes a method to represent the component distribution of the mixture model using a boxplot.

Keywords: Mixture Model, Boxplot, Percentile, Quartile, Estimator, Postterior Probability.

[1] Department of Economics, Sapporo Gakuiun University; nagatomo@sgu.ac.jp.