一変量確率分布における複峰性とクラスター分割基準

中村　永友[1]　　土屋　高宏[2]

要　旨

統計的分類手法が提案されるたびに，クラスター数を決める基準が多く提案されてきた．本稿は多次元データが何らかの方法で分類されたことを前提として，それを 1 次元に射影し，それが 2 峰性のとき分割を否定しないという分割基準の考察をする．その判断をするためのいくつかの指標について検証する．

キーワード：クラスタリング，正規混合分布モデル，判別関数，Biaverage

1　はじめに

多次元データが分類されたときのクラスターの分割方法について考察する．ここで想定する状況としては，データが切れ目なく散布していて，散布図等の目視でも明確な切れ目がないときである．これまでにもクラスタリングの際にデータを分割すべきか否かを判断する基準として，種々の方法が提案されてきた．例えば，情報量規準（EIC：中村・小西，1998；BIC：Schwarz, 1978），GAP 統計量（Tibshirani *et al.*, 2001），尤度比検定（Wolfe, 1971）等々である．また，近年のレビューペーパーとして，McLachlan & Rathnayake (2014) を挙げておく．これらの種々の方法の特徴として一般的に言えることは，g 群と $g+1$ 群の何らかの統計量を比較して，+1 すべきか否かを決めている点である．ここで言及する方法は，2 群に分割して 1 次元データの散布状況から分割すべきか元に戻すかを決める方法で，分割前の状況との比較をしない方法である．

本稿で取りあげるクラスター分割基準の基本的な着目点は 1 次元データの複峰性（multimodality, or bimodality）である．多次元データが 2 分割され，それを判別関数により 1 次元に射影したデータに対して，1 次元混合分布モデルの密度関数の単峰–複峰（2 峰）性に関するいくつかの指標による判定方法を紹介する．また，1 次元で 2 峰性をもつ確率分布から得られたと思われるデータに対して，峰の場所を推定する方法として，4 次までのモーメントを利用した biaverage という推定量の紹介をする．

多次元データを 1 次元データに射影する方法を次節で述べる．第 3 節では単峰・複峰性と biaverage を説明し，第 4 節で数値実験の結果を示す．

2　1 次元への射影

通常，判別分析における判別関数の値は，その値の正負の値によりどちらの群に属するか判定をするために用いる．この値を分類対象の全データに対して求めた 1 次元データを対象として，分割の可否を決める方法の検討をする．

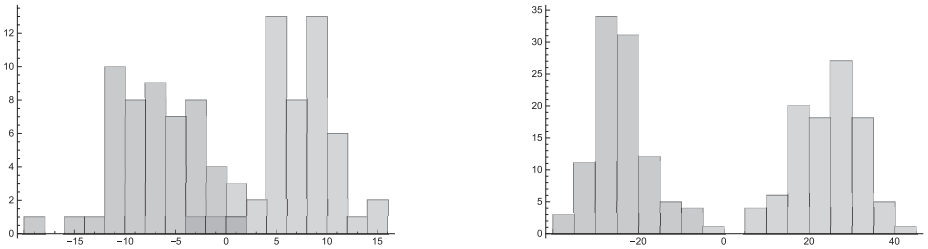
まず，多次元データを判別関数により 1 次元データに射影する方法を説明する．

対象とするデータは，何らかの分類手法によって 2 群に分類された多次元データである．分類されているので，個別のデータには群を識別するラベルが付いていて，この状況はあらかじめラベルが付いているデータに対して判別分析を行う状況と同じである．このとき，2 群を分ける線形判別超平面（あるいは 2 次曲面）へのすべてのデータの距離を求める．これによって多次元データが 1 次元データに射影されるのである．

具体的には，分析対象のデータに対して次の手順の計算を経る．

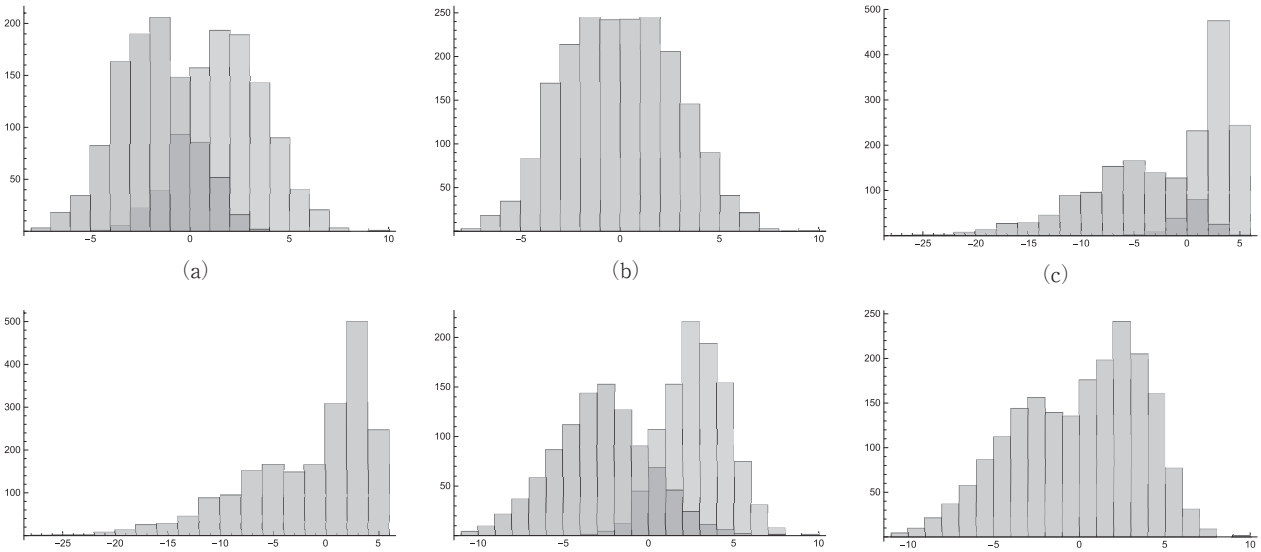
線形判別を行うときと同じ状況で，共通の分散共分

[1] 札幌学院大学　経済学部；nagatomo@sgu.ac.jp
[2] 城西大学　理学部；takahiro@josai.ac.jp



（a）アヤメデータ　　（b）スイス銀行データ

図1：よく知られたデータセットの群間データのヒストグラム



（a）　（b）　（c）

（d）　（e）　（f）

図2：1次元スコアの複峰性

7次元の正規分布を2つ混合させて，判別関数の値を1次元で示した．（a），（b），（e），（f）は線形判別関数によるスコア．（c）と（d）は2次判別関数によるスコア．（a）と（b）のパラメータ：$\mu_1=\{0,0,0,0,0,0,0\}^T$，$\mu_2=\{2,0,0,0,0,0,0\}^T$，$\Sigma_1=\Sigma_2=\boldsymbol{I}$，$\pi_1=\pi_2=\frac{1}{2}$．（c）～（f）のパラメータ：$\mu_1=\{0,0,0,0,0,0,0\}^T$，$\mu_2=\{2,0,0,0,0,0,0\}^T$，$\Sigma_1=\boldsymbol{I}$，$\Sigma_2=4\boldsymbol{I}$，$\pi_1=\pi_2=\frac{1}{2}$．

散行列を

$$\boldsymbol{S}_p=\frac{(n_1-1)\boldsymbol{S}_1+(n_2-1)\boldsymbol{S}_2}{n_1+n_2-2}$$

により求める．ここで，各群の分散共分散行列は

$$\boldsymbol{S}_k=\frac{1}{n_k-1}\sum_{i=1}^{n_k}(\boldsymbol{x}_{ki}-\overline{\boldsymbol{x}}_k)(\boldsymbol{x}_{ki}-\overline{\boldsymbol{x}}_k)^T$$

である（$k=1,2$）．すべてのデータ$\boldsymbol{x}$に対して，次の式で判別平面までの距離を計算する．

$$D(\boldsymbol{x})=(\overline{\boldsymbol{x}}_1-\overline{\boldsymbol{x}}_2)^T\boldsymbol{S}_p^{-1}\left[\boldsymbol{x}-\frac{1}{2}(\overline{\boldsymbol{x}}_1+\overline{\boldsymbol{x}}_2)\right].$$

この段階で多次元のデータは1次元のデータに射影される．

一方，分散共分散行列が等しいという仮定をおかずに判別する方法として2次判別関数がある．この場合は次の通りとなる．

$$Q(x)=\frac{1}{2}\log\left(\frac{|\boldsymbol{S}_2|}{|\boldsymbol{S}_1|}\right)-\frac{1}{2}(\boldsymbol{x}-\overline{\boldsymbol{x}}_1)^T\boldsymbol{S}_1^{-1}(\boldsymbol{x}-\overline{\boldsymbol{x}}_1)$$

$$+\frac{1}{2}(\boldsymbol{x}-\overline{\boldsymbol{x}}_2)^T\boldsymbol{S}_2^{-1}(\boldsymbol{x}-\overline{\boldsymbol{x}}_2).$$

このようにして得られたデータのことを以後**1次元スコア**，あるいは単に**スコア**と呼ぶ．図1（a）はアヤメデータ（Fisher, 1936）の3つの種のうち，Virginicaと Versicolor の2種について，4つの変数すべてを使って得られた1次元スコアである．図1（b）はスイス銀行データ（Flury and Riedwyl, 1988）の全6変数に対する1次元スコアである．両方のデータセット共にきれいに分かれていて，2群に分ける事が非常に望ましい状況といえる．

一方，図2は7次元の正規分布を2つ混合させた結果の1次元スコアである．（a），（b），（e），（f）は線形

判別関数による1次元スコア，(c) と (d) は2次判別関数によるものである．(a) は2つの群をそれぞれオーバーラップさせたもので，(b) は同データを群分けしないものである．(a) は2峰に見えるが実際のデータは (b) で単峰である．一方，(c) と (d)，さらに (e) と (f) は同じ関係で，(d) と (f) は2峰である．

3　単峰-複峰性

3.1　混合分布モデルでの単峰性

1次元の正規混合分布モデルにおける単峰性と複峰性を見極める条件が研究されている．

$$\Delta=\mu_2-\mu_1$$

として，次の条件を満足するとき，全体の確率分布は単峰である（Eisenberger，1964）：

$$\Delta^2<\frac{27\sigma_1^2\sigma_2^2}{4(\sigma_1^2+\sigma_2^2)}.$$

もし，$\sigma_1=\sigma_2$のときは，$\Delta<1.837$である．

一方 Behboodian（1970）は十分条件として

$$\Delta\leq 2\min\{\sigma_1,\sigma_2\}$$

を示し，もし$\sigma_1=\sigma_2$が仮定できるときは，$\Delta<2$となる．

これらを受けて，Sitek（2016）はより詳細な単峰性の条件を精査している．例えば，

(1) $\mu_1\neq\mu_2$ and $\sigma_1=\sigma_2=\sigma$ and $p=\frac{1}{2}$のとき，

$$|\mu_2-\mu_1|\leq 2\sigma,$$

(2) $\mu_1\neq\mu_2$ and $\sigma_1\neq\sigma_2$のとき，

$$\Delta^2\leq\frac{8\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2},$$

等である．

これらの条件を以下のように指標化する：

$$\mathrm{UI1}=(\mu_2-\mu_1)^2-\frac{27\sigma_1^2\sigma_2^2}{4(\sigma_1^2+\sigma_2^2)}<0,$$

$$\mathrm{UI2}=|\mu_2-\mu_1|-2\min\{\sigma_1,\sigma_2\}\leq 0,$$

$$\mathrm{UI3}=(\mu_2-\mu_1)^2-\frac{8\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\leq 0.$$

いずれも，この条件を満足したとき単峰である．

3.2　Biaverage

2つのモード（峰）のある確率分布に対して，biaverage という統計量がある．これを一般化した k 個のモードに対する k-average がある（Antoniewicz, 2005）．Biaverage は，2つのモードに対応した2つ組みのパラメータ（m_1, m_2）により定義される．それは次の条件を満足するモーメントである：

$$\min_{a,b} E[(X-a)(X-b)]^2=E[(X-m_1)(X-m_2)]^2.$$

確率変数が4次のモーメントを持つとき，2次モーメントである分散は0ではない．このとき上式は解を持ち（Antoniewicz, 2005），以下のようになる：

$$m_1=\frac{1}{2}\left(P-\sqrt{P^2+4Q}\right),$$

$$m_2=\frac{1}{2}\left(P+\sqrt{P^2+4Q}\right),$$

$$P=\frac{E(X^3)-E(X^2)E(X)}{E(X^2)-E(X)^2},$$

$$Q=\frac{E(X^2)^2-E(X^3)E(X)}{E(X^2)-E(X)^2}.$$

また，biaverage の2つの平均周りの分散は，次式で計算される：

$$V_0=E[(X-m_1)(X-m_2)]^2.$$

そしてこの標準偏差は，

$$\sigma_0=\sqrt[4]{V_0}$$

となる．

これらの式は，2峰性の確率分布からの確率変数 $\{X_1,\ldots,X_n\}$ の実現値 $\{x_1,\ldots,x_n\}$ が与えられたとき，次式で計算される．

$$p=\frac{n\sum_{i=1}^n x_i^3-\sum_{i=1}^n x_i^2\sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2-\left(\sum_{i=1}^n x_i\right)^2},$$

$$q=\frac{\left(\sum_{i=1}^n x_i^2\right)^2-\sum_{i=1}^n x_i^3\sum_{i=1}^n x_i}{n\sum_{i=1}^n x_i^2-\left(\sum_{i=1}^n x_i\right)^2}.$$

さらに，$c=\frac{1}{2}\sqrt{p^2+4q}$ として，biaverage の推定値は以下の通り求められる：

$$\widehat{m}_1=\frac{1}{2}p-c,$$

$$\widehat{m}_2=\frac{1}{2}p+c.$$

表1：高次元データの1次元スコアの平均差

次元	$\bar{x}_1$	$\bar{x}_2$	s_1^2	s_2^2	平均差	分散比
2	2.02	−2.02	4.03	4.04	4.03	1.00
5	2.03	−2.03	4.05	4.06	4.05	1.00
7	2.01	−2.01	4.03	4.02	4.02	1.00
10	2.04	−2.04	4.07	4.07	4.07	1.00
20	2.08	−2.08	4.16	4.16	4.16	1.00
50	2.21	−2.21	4.42	4.41	4.42	1.00
100	2.45	−2.45	4.90	4.90	4.90	1.00

4　数値実験

4.1　1次元スコア

まず，1次元スコアのふるまいを調べる．データの次元を2，5，7，10，20，50，100として，2つの多次元正規分布から乱数を発生させて1次元スコアに射影し，2平均の距離などを求めた．実験の設定は次の通り：$n=1000$，$\mu_1=0$，$\mu_2=\{2,0,\ldots,0\}$，$\boldsymbol{\Sigma}_1=\boldsymbol{I}$，$\boldsymbol{\Sigma}_2=\boldsymbol{I}$，$\pi_1=\pi_2=\frac{1}{2}$．結果を表1に示す．

データ数を一定にしていて，高次元になるほど平均間距離がより広がっている．次元に関する影響があることが観察される．この事実とデータ解析の整合性については別の機会に議論したい．

4.2　Biaverage

次に，biaverageのふるまいを探るため，2つの正規分布を平均間距離，混合比率，データ数を変えながら混合させて，2つの推定値（biaverage）間の距離を見た．実験結果は表2に示す．実験の設定は，$\mu_1=0$，$\mu_2=1, 3, 5, 10$，　$\sigma_1=\sigma_2=1$，　$n=50, 100, 500, 1000$，$\pi_1=0.5, 0.75, 0.95, 0.99$とした．

Biaverageは全く分類に関する事前情報のないままで2つのモードの値を推定する方法である．平均間距離が十分大きいときには，とくに$\mu_2=10$，$\pi_1=0.5$や0.75のときは，かなり真値に近く推定されている．しかし，距離が小さくなり，第2の確率分布のデータ数が多くなると，推定精度が悪くなることがわかる．さらに表3には1成分のみのとき，すなわち単峰のデータに対するbiaverageの推定結果を示す．$N(0,1)$から$n=50, 100, 500, 1000$を抽出したこととなる．データ数によらず$\widehat{m}_1=-1.0$，$\widehat{m}_2=1.0$と推定されている．複峰性のないデータに対しても同様に推定されるということも確認できる．この結果はprincipal points（Flury, 1990）と似た性質を持つと考えられるので，これらとの関係は今後の研究課題としたい．

表2：複峰データに対するbiaverageの推定値

π_1	μ_2	n	$\widehat{m}_1$	$\widehat{m}_2$	$\widehat{m}_2-\widehat{m}_1$	SD$(\widehat{m}_1)$	SD$(\widehat{m}_2)$
0.5	1	50	−0.62	1.62	2.23	0.25	0.25
		100	−0.62	1.62	2.24	0.18	0.18
		500	−0.62	1.62	2.24	0.08	0.08
		1000	−0.62	1.62	2.24	0.06	0.06
	3	50	−0.30	3.30	3.60	0.24	0.24
		100	−0.30	3.30	3.60	0.17	0.17
		500	−0.30	3.30	3.60	0.08	0.08
		1000	−0.30	3.30	3.61	0.05	0.05
	5	50	−0.19	5.19	5.37	0.22	0.22
		100	−0.19	5.19	5.38	0.16	0.16
		500	−0.19	5.19	5.38	0.07	0.07
		1000	−0.19	5.19	5.38	0.05	0.05
	10	50	−0.10	10.10	10.19	0.21	0.21
		100	−0.10	10.10	10.19	0.15	0.15
		500	−0.10	10.10	10.20	0.07	0.07
		1000	−0.10	10.10	10.20	0.05	0.05
0.75	1	50	−0.79	1.39	2.18	0.24	0.26
		100	−0.80	1.38	2.18	0.17	0.19
		500	−0.80	1.38	2.18	0.08	0.08
		1000	−0.80	1.38	2.18	0.06	0.06
	3	50	−0.48	2.94	3.42	0.20	0.32
		100	−0.49	2.92	3.41	0.14	0.23
		500	−0.49	2.93	3.41	0.06	0.10
		1000	−0.48	2.93	3.41	0.05	0.07
	5	50	−0.31	4.89	5.20	0.18	0.32
		100	−0.32	4.87	5.19	0.13	0.23
		500	−0.32	4.88	5.19	0.06	0.10
		1000	−0.32	4.88	5.20	0.04	0.07
	10	50	−0.16	9.92	10.08	0.17	0.29
		100	−0.16	9.91	10.07	0.12	0.21
		500	−0.16	9.91	10.08	0.05	0.10
		1000	−0.16	9.91	10.08	0.04	0.07
0.95	1	50	−0.95	1.11	2.05	0.24	0.25
		100	−0.95	1.09	2.05	0.17	0.18
		500	−0.95	1.09	2.05	0.08	0.08
		1000	−0.95	1.09	2.05	0.05	0.06
	3	50	−0.70	1.92	2.62	0.22	0.44
		100	−0.72	1.81	2.53	0.16	0.31
		500	−0.71	1.82	2.52	0.07	0.14
		1000	−0.71	1.82	2.52	0.05	0.10
	5	50	−0.44	3.57	4.01	0.18	0.62
		100	−0.45	3.38	3.83	0.13	0.47
		500	−0.45	3.39	3.84	0.06	0.21
		1000	−0.45	3.39	3.84	0.04	0.15
	10	50	−0.20	8.85	9.05	0.15	0.67
		100	−0.21	8.62	8.83	0.11	0.53
		500	−0.21	8.64	8.84	0.05	0.24
		1000	−0.21	8.64	8.85	0.03	0.17
0.99	1	50	−0.98	1.04	2.02	0.24	0.24
		100	−0.99	1.02	2.01	0.17	0.17
		500	−0.99	1.02	2.01	0.08	0.08
		1000	−0.99	1.02	2.01	0.05	0.06
	3	50	−0.85	1.38	2.23	0.24	0.38
		100	−0.91	1.20	2.11	0.18	0.24
		500	−0.90	1.20	2.10	0.08	0.11
		1000	−0.90	1.20	2.10	0.06	0.08
	5	50	−0.60	2.33	2.93	0.22	0.69
		100	−0.70	1.76	2.46	0.18	0.44
		500	−0.69	1.76	2.44	0.09	0.20
		1000	−0.68	1.76	2.44	0.06	0.14
	10	50	−0.24	6.92	7.17	0.15	1.15
		100	−0.29	5.31	5.60	0.12	1.01
		500	−0.28	5.34	5.63	0.05	0.46
		1000	−0.28	5.35	5.63	0.04	0.32

シミュレーションの繰り返す数＝100,000回．$\mu_1=0$，$\sigma_1=\sigma_2=1$として，$n\pi_1$個を$N(0,1)$から，$n(1-\pi_1)$個を$N(\mu_2,1)$からデータを発生させて，当該推定値を計算した．

表 3：単峰データに対する biaverage の推定値

μ_2	n	$\widehat{m}_1$	$\widehat{m}_1$	差分	SD$(\widehat{m}_1)$	SD$(\widehat{m}_2)$
0	50	−1.00	1.00	2.00	0.24	0.24
	100	−1.00	1.00	2.00	0.17	0.17
	500	−1.00	1.00	2.00	0.08	0.08
	1000	−1.00	1.00	2.00	0.05	0.05

表 4：実データに対する推定値

	アヤメデータ	スイス銀行データ
UI1	154.5	2168.
UI2	6.677	34.36
UI3	145.7	2139.
1 次元スコア	{−7.109, 7.109}	{−24.12, 24.12}
Biaverage	{−8.355, 7.717}	{−24.72, 25.47}
Mixture1	{−6.929, 7.251}	{−23.93, 24.40}
Mixture2	{−6.367, 7.820}	{−23.93, 24.40}

実データの 2 つ峰に対する推定値．1 次元スコア：オリジナルの分類による 1 次元スコアから求めた値．
Biaverage：この手法による推定値．
Mixture1：1 次元混合分布モデル（等分散の仮定）による 2 つの平均の推定値．
Mixture2：1 次元混合分布モデル（不等分散の仮定）による 2 つの平均の推定値．UI1，UI2，UI3：1 次元スコアによる各種指標．

5　実データ

アヤメデータとスイス銀行データの 1 次元スコアに対して，2 群のオリジナルの平均，混合分布モデル，biaverage の結果を表 4 に示す．

これらの実データは判別分析などの例題としてよく使われていることもあり，明確に分離している（図 1 ）．このこともあり，UI1，UI2，UI3の値はかなり大きな正の値となっている．

2 群の 1 次元スコアで各群の平均がこの場合は基準値となり，2 つの混合分布モデルでの推定値はかなりこれらに近く推定されている．一方，スイス銀行データの分離度が良いことから，biaverage もそれほど悪くはない．

6　今後の課題

多次元データを 1 次元スコアに射影し，その後分割するか否かを判断するための指標等について，いくつかの方法の検討を行った．単峰性のみの基準では分類が保守的になるので，ここで紹介した基準に加えてデータ数や混合比率を考慮する分割基準を考えたい．より，実用的な指標やアルゴリズムを構築していくことが今後の検討課題である．

謝辞

本研究は札幌学院大学　奨励金（SGU-BS2018-02）の補助を受けた．

参考文献

[1] Antoniewicz, R. and Misztal, A. (2001). Biaverage, Statistical Review, 47, 269–274, (in Polish).
[2] Behboodian, J. (1970). On a mixture of normal distributions, Biometrika, 57, 215–217.
[3] Eisenberger, I. (1964). Genesis of bimodal distributions, Technometrics, 6, 357–363.
[4] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems, Annals of Eugenics. 7(2), 179–188. doi:10.1111/j.1469–1809.1936.tb02137.x.
[5] Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics: A Practical Approarch*, Chapman & Hall, London.
[6] Flury, B. (1990). Principal points. Biometrika 77, 1, 33–41.
[7] McLachlan, G. J. and Rathnayake, S. (2014). On the number of components in a Gaussian mixture model, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4, 341–355.
[8] 中村永友・小西貞則（1998）. 情報量規準に基づく多変量混合正規分布モデルのコンポーネント数の推定, 応用統計学, 27, 165–180.
[9] Schwarz, G. E. (1978). Estimating the dimension of a model, Annals of Statistics, 6(2), 461–464, doi:10.1214/aos/1176344136.
[10] Sitek, G. (2016). The modes of a mixture of two normal distoributions, Silesian Journal of Pure and Applied Mathematics, 6(1), 59–67.
[11] Tibshirani R., Walther G., Hastie T. (2001). Estimating the number of clusters in a data set via the gap statistic. Journal of the Royal Statistical Society Seres B Methodology, 63, 411–423.
[12] Wolfe, J. H. (1971). A Monte Carlo study of the sampling distribution of the likelihood ratio for mixture of multinormal distributions, Technical Bulltetin, STB 72-2, Naveal Personnel and Training Research Laboratory, San Diego, CA.

The Proceedings of the Research Institute of Sapporo Gakuin University Vol.6, 1–6 (2019)

Multimodality of the Univariate Probablity Distribution and Clustering Criterion

Nagatomo NAKAMURA[1]
and
Takahiro TSUCHIYA[2]

Abstract

When a new statistical classification method is proposed, many criteria for partitioning clusters have been proposed. In this report, we consider a method that does not deny division when multidimensional data is classified into two groups in some way, when projected one-dimensional data is bimodal. We examined several indicators to make that judgment.

Keywords: Biaverage, Clustering, Discriminant Function, Normal Mixture Model.

[1] Department of Economics, Sapporo Gakuiun University; nagatomo@sgu.ac.jp.
[2] Department of Mathematics, Josai University; takahiro@josai.ac.jp.