混合正規分布モデルの成分数推定に関する数値的検証

中村　永友[1]

要　旨

統計的モデルによる分類手法としての混合分布モデルの成分数の推定を，ブートストラップ法による情報量規準で行う方法の数値的な検証を行う．比較的有名なデータセットに対して我々が提唱した方法を適用し，その詳細を報告する．

キーワード：混合分布モデル，ブートストラップ，情報量規準，EM 法

1　は じ め に

統計的分類モデルとしての混合正規分布モデルは様々な分野で広く用いられている（中村 他，1993; 中村，1995; 中村，2007; 中村，2009; 中村・小西，1998; 中村 他，2005; McLachlan and Basford, 1988; McLachlan and Peel, 2000; McLachlan, 1992; Everitt, 1993; Everitt and Hand, 1981; Tatarinova and Schumitzky, 2015; Mengresen *et al*., 2011; Aggarwal and Reddy, 2014）．データが高次元であったり，成分分布が入り組んでいる，１つの成分あたりのデータ密度が低い，モデルの構造が複雑であるなどのとき，対数尤度関数には複数の極値が存在し，大域的な最適解を得ることは困難な問題として存在している．本報告は，中村・小西（1998）で提案された混合分布モデルの成分数を推定する手続きに基づいて，いくつかのグループからなる多変量のデータセットに対する成分数の推定結果を示すものである．とくに，情報量規準として EIC（Ishigro *et al*., 1997），CEIC（Konishi and Kitagawa, 1997），AIC（Akaike, 1973, 1974），BIC（Schwarz, 1978）を対象として，様々な特徴を持つデータセットに対して，モデルの推定結果を通して，情報量規準の挙動を明らかにしていくことが，本報告の目的である．

[1] 札幌学院大学経済学部; nagatomo＠sgu.ac.jp.

2　混合分布モデル

混合正規分布モデルは，

$$f(x|\theta)=\sum_{k=1}^{m}\pi_k f_k(x|\mu_k, \Sigma_k)$$

で定義される．ここで，$f_k(\cdot)$ は第 k 成分の確率分布を表し，本論文では多変量正規分布を仮定する．π_k は混合比率，μ_k は平均ベクトル，Σ_k は分散共分散行列である．$\theta=\{\pi_1, \ldots, \pi_m, \mu_1, \ldots, \mu_m, \Sigma_1, \ldots, \Sigma_m\}$ はモデルの全パラメータで，m は仮定する成分数である．また，$\sum\pi_k=1$ である．このモデルを推定する，すなわち目的のパラメータを求めるためには，対数尤度関数を構成し，最大化を行うことで達成される．そのために，EM 法（Dempster *et al*., 1977）を用いて推定する．しかし，対数尤度関数に複数の極大値が存在し，現実的な大域的最適解を得なければそのモデルの仮定自体が無意味になることもある．そのためには，最適な初期値の設定が必要不可欠であるが，EM 法はデータに依存する推定法であることから，決定論的なアルゴリズムではなく，データに依存したある種の確率的な初期値設定法が現実的である．その一つの方法が中村（1995）で提案された複数のクラスタリングによる初期値設定法である．

3　情報量規準と推定手続き

混合分布モデルは混合比率をすべて足して１になるという制約条件があるために，モデルとしての正則条件を満足しないという性質がある．このために，成分

表1：情報量規準の各種バイアス

情報量規準	バイアス
AIC	p
EIC	$\frac{n}{B}\sum_{b=1}^{B}\left\{L(\hat{\theta}_b^*\|X_b^*)-L(\hat{\theta}_b^*\|X)\right\}$
CEIC	$\frac{n}{B}\sum_{b=1}^{B}\left\{L(\hat{\theta}_b^*\|X_b^*)-L(\hat{\theta}\|X_b^*)+L(\hat{\theta}\|X)-L(\hat{\theta}_b^*\|X)\right\}$
BIC	$p\log n$

p：自由パラメータ数，n：データ数，B：ブートストラップ反復回数，$L()$：対数尤度関数，X：オリジナルのデータ，X_b^*：b 回目のブートストラップ標本，$\hat{\theta}$：X に対するパラメータ推定値，$\hat{\theta}_b^*$：X_b^* に対するパラメータ推定値．

数推定をAICで推定した場合には，その数を過大推定する傾向がある（中村・小西, 1998）．平均対数尤度の適応的な推定をできるのが，ブートストラップ法（Efron, 1979）による情報量規準EICである．

モデル自体の未知パラメータの推定はEM法で行い，成分数推定には情報量規準EIC（Ishigro *et al*., 1997），CEIC（Konishi and Kitagawa, 1997），AIC（Akaike, 1973, 1974），BIC（Schwarz, 1978）を用いてこれらの比較をする．情報量規準の基本形を

$$-2\times(\text{最大尤度})+2\times(\text{バイアス})$$

とすると，各情報量規準では「バイアス」が特徴となる．本報告で使用する情報量規準とそのバイアスを表1にまとめる．

ここで重要な視点は平均対数尤度の推定をするということである．これにしたがえばBICはこれを行うものではないが，成分数の推定をするアプリケーション・ソフトウェアである，例えばRのmclust関数などや，多くの論文で頻繁に用いられている．しかし，限られた条件下で機能する情報量基準であるため（中村・小西，1998; Nakamura and Konishi, 2016），その比較を行うため，BICも対象とする．AICも真のモデルが想定したモデルの中に含まれているなどの条件下でうまく機能するが（Konishi and Kitagawa, 1996），正則条件を満足しない混合分布モデルでは成分数が多めに推定されるということがあるので，これを同時に確認する．

次に，成分数の推定アルゴリズムは以下の通りである．まず，与えられたデータセット X に対して仮定した成分数で混合分布モデルを推定するために，**推定手続きA**で大域的最適解を求める．

推定手続きA　(1)様々なクラスタリングで初期分類を行いEM法の初期値設定をする．(2)複数の初期値からEM法を実行する．(3)複数得られた解の中から対数尤度の最大値を解とする．

次に，ブートストラップ法を用いた情報量規準EICやCEICを求めるために，次の**推定手続きB**を通して計算する．

推定手続きB　(1)対象データ X に対して推定手続きAを適用し，推定されたパラメータを $\hat{\theta}$ とする．(2)X から b 回目のブートストラップ標本 X_b^* を作成する．(3)ブートストラップ標本 X_b^* に対して，推定手続きAを適用し，推定されたパラメータを $\hat{\theta}^*$ とする．(4)(2)と(3)を B 回繰り返す（$b=1, 2, ..., B$）．(5)$\hat{\theta}$ や $\hat{\theta}^*$ などを使って，各種情報量規準を計算する．

CEICとEICは同じものを推定しているが，もし成分数の推定結果が異なる場合は，CEICのほうが分散が小さいので，こちらの推定結果を採用する．

4　データへの適用

本報告では，以下の5種類のデータセットに対して，既知の成分数を推定できるか否かを検証した．

- Seeds dataset,
- Wine dataset,
- Iris dataset,
- Crabs datase,
- Swiss Bank Notes dataset.

最初の3種類は，UCI Machine Learning Repository（http://mlr.cs.umass.edu/ml/index.html）に登録されていて，あとの2種類はRに登録されているサンプルデータである．これらの各種データに対して，多数の変数がある場合は適宜いくつか組み合わせて，成分

数の推定を行い，推定結果の特徴を見ていく．推定の際には，モデルの成分数を１から４，場合によっては５とし，分散共分散行列は任意と共通として推定を行い，それらの中から各種情報量規準の最小の値に対応する成分数を推定値とする．結果の表中では最小の値を太字で表した．

以下に，各データセットの簡単な説明と推定結果の傾向を示す．

4.1 Seeds データセット

３種類の小麦（Kama, Rosa and Canadian）の3×70＝210個体に対して，穀粒(kernel)の７種類の計測値（area (A), perimeter (P), compactness ($4\pi A/P^2$), length of kernel, width of kernel, asymmetry coefficient, length of kernel groove）を記録したものである．穀粒内部の可視化は，X線により検出したものである．

７つのすべての変数と，３つ変数の組み合わせによる４つの成分数推定結果を表２に示す．

すべての７変数を使って推定した場合は，CEIC & EIC, BIC も，真の３成分ではなく４成分を推定している．Nakamura and Konishi（2016）では第２変数から第７変数を使って推定した結果が示されており，そこではCEIC & EICは３成分（任意の分散共分散行列）を推定している．BICも同様に３成分であるが，共通の分散共分散行列が推定された．これらの変数より少ない変数の組み合わせを，表２に示しているが，このデータでは，CEIC & EICは正しい成分数を推定

表２：Seeds dataset

変数	m	共通分散	p	LL	CEIC	EIC	AIC	BIC
1,2,3,4,5,6,7	1	—	35	−2673.66	5430.69	5441.02	5417.33	5534.47
1,2,3,4,5,6,7	2	no	71	−2298.50	4781.70	4782.21	4739.01	4976.64
1,2,3,4,5,6,7	3	no	107	−2102.68	4498.62	4487.64	4419.36	4777.50
1,2,3,4,5,6,7	4	no	143	−2000.70	**4421.34**	**4426.17**	**4286.14**	**4766.04**
1,2,3,4,5,6,7	5	no	179	−1942.35	4445.09	4445.09	4438.33	4841.83
1,2,3,4,5,6,7	2	yes	43	−2589.53	5285.32	5286.62	5265.05	5408.99
1,2,3,4,5,6,7	3	yes	51	−2523.66	5173.96	5170.90	5149.32	5320.02
1,2,3,4,5,6,7	4	yes	59	−2461.86	5073.03	5071.39	5041.72	5239.20
1,2,3,4,5,6,7	5	yes	67	−2412.77	4974.51	4954.37	4959.54	5183.80
5,6,7	1	—	9	−1972.47	3961.68	3960.53	3962.94	3993.06
5,6,7	2	no	19	−1892.63	3824.89	3820.19	3823.27	3886.86
5,6,7	3	no	29	−1872.26	3808.15	3804.99	3802.52	3899.59
5,6,7	4	no	39	−1856.20	3814.19	3813.65	**3790.40**	3920.94
5,6,7	2	yes	13	−1910.83	3847.61	3848.03	3847.66	3891.17
5,6,7	3	yes	17	−1884.10	3803.06	3804.45	3802.19	**3859.10**
5,6,7	4	yes	21	−1881.55	**3799.09**	**3799.71**	3797.50	3875.39
3,6,7	1	—	9	−1474.75	2966.17	2963.28	2967.49	2997.62
3,6,7	2	no	19	−1398.58	2837.69	2835.05	2835.16	2898.76
3,6,7	3	no	29	−1380.02	2829.88	2835.75	2818.05	2915.11
3,6,7	4	no	39	−1366.93	2830.60	2828.76	**2811.86**	2942.40
3,6,7	2	yes	13	−1414.77	2857.81	2863.59	2855.54	2899.05
3,6,7	3	yes	17	−1397.26	2831.84	2829.96	2828.52	**2885.42**
3,6,7	4	yes	21	−1388.66	**2823.28**	**2821.91**	2819.32	2889.61
3,4,5	1	—	9	−799.703	1618.45	1616.24	1617.41	1647.53
3,4,5	2	no	19	−727.503	1505.84	1506.46	1493.01	**1556.60**
3,4,5	3	no	29	−704.292	**1486.93**	**1484.42**	**1466.58**	1563.65
3,4,5	4	no	39	−685.512	1489.65	1488.33	1489.02	1686.50
3,4,5	2	yes	13	−771.084	1573.64	1575.88	1568.17	1611.68
3,4,5	3	yes	17	−749.117	1542.41	1543.70	1532.23	1589.13
3,4,5	4	yes	21	−733.424	1519.05	1519.50	1508.85	1621.91
3,5,7	1	—	9	−917.683	1855.16	1857.83	1853.37	1883.49
3,5,7	2	no	19	−842.592	1726.04	1725.85	1723.18	1786.78
3,5,7	3	no	29	−841.137	1702.88	1703.57	1686.27	1837.34
3,5,7	4	no	39	−792.327	**1680.58**	**1680.07**	**1662.65**	1793.19
3,5,7	2	yes	13	−856.558	1740.50	1742.31	1739.12	1782.63
3,5,7	3	yes	17	−841.570	1723.95	1721.51	1717.14	**1774.04**
3,5,7	4	yes	21	−832.907	1718.63	1719.34	1707.81	1778.10

ブートストラップ反復回数 $B=200$，m：混合分布の成分数，p：モデルのパラメータ数．

表3：Wine dataset

変数	r	共通分散	p	LL	CEIC	EIC	AIC	BIC
2,6,7,11,12,13	1	—	27	−1075.62	2213.66	2212.82	2205.25	2291.15
2,6,7,11,12,13	2	no	55	−950.037	2035.86	2034.05	2010.07	2185.07
2,6,7,11,12,13	3	no	83	−893.741	2022.38	2023.29	1953.48	2217.57
2,6,7,11,12,13	4	no	111	−839.517	**1982.88**	**1983.85**	**1901.03**	2254.21
2,6,7,11,12,13	2	yes	34	−1020.03	2118.22	2117.32	2108.07	2216.24
2,6,7,11,12,13	3	yes	41	−989.652	2076.84	2077.20	2061.30	2191.76
2,6,7,11,12,13	4	yes	48	−957.948	2030.32	2033.31	2011.90	**2164.62**

ブートストラップ反復回数 $B=200$，m：混合分布の成分数，p：モデルのパラメータ数．

することは困難なようであり，BIC は比較的成績が良いことが観察できる．

4.2　Wine データセット

このデータは，イタリアの同じ地方で成長した3つの異なる品種由来のワインの化学分析の結果である．分析では3種類のワインにおける13の成分を計測したものである．その内訳は，Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols, Proanthocyanins, Color intensity, Hue, OD280/OD315 of diluted wines, Proline である．

13変数からの6変数の成分数推定結果を表3に示す．真の群数は3であるが，すべての情報量規準は4と推定し，BIC の推定結果は共通の分散共分散行列であった．その推定結果に基づくオリジナルの分類と推定の分類結果のクロス表を以下に示す．

任意の分散共分散行列

		オリジナル分類		
		1	2	3
推定結果	1	58	8	0
	2	0	47	0
	3	1	13	1
	4	0	3	47

共通の分散共分散行列

		オリジナル分類		
		1	2	3
推定結果	1	56	6	0
	2	3	59	0
	3	0	0	26
	4	0	6	22

3群と推定してほしいところだが，任意の分散共分散行列ではオリジナルの第2群が2つに分かれ，共通の分散共分散行列では第3群が2つに分かれる結果となった．第2群にせよ第3群にせよ，2つに分かれた方がよりあてはまりがよいということである．

4.3　Iris データセット

フィッシャーによって計測され，アンダーソンの引用によって有名になったこのデータセットは，判別分析やクラスタリング，パターン認識等の論文で最もよく知られたデータである．データセットは4つの変数（がく片長：Sepal Length，がく片幅：Sepal Width，花びら長：Petal Length，花びら幅：Petal Width），50個体×3つのアヤメの種（setosa, versicolor, virginica）からなり，そのうち1つのクラスは他の2種から明らかに分かれていて，2種は完全に分離していない．

4変数の推定結果は中村・小西（1998）にあるので，3変数（表4）と2変数（表5）の結果をここでは示す．傾向として，AIC はこのデータセットに対してほぼ全体で CIEC & EIC に比べて多めに成分数を推定している．BIC はその逆で，同じか少なめに推定する傾向がある．

変数1と2の結果で，CEIC と EIC の推定結果が異なるので，CIEC の3群を採用する．しかし，2群と3群の情報量規準の値の差が非常に小さいことを記しておく．

4.4　Crabs データセット

このデータは，Campbell and Mahon（1974）に掲載されたデータセットで，2種類の種とオスとメスで4分類されている．測定されているのは5つの部位からなる（FL: frontal lobe size，RW: rear width，CL: carapace length，CW: carapace width，BD: body depth，いずれも単位は mm）．

データセットを観察すると，オスとメスがそれぞれ直線的な傾向を示し，全体がV字型をしている．

Crabs データの成分数推定結果を表6に示す．CEIC & EIC，AIC は任意の分散共分散行列で4を推定し，BIC は共通の分散共分散行列で5を推定している．

4.5　Swiss Bank Notes データセット

このデータは Flury and Riedwyl（1988）に掲載さ

表 4 ：Iris dataset（ 3 変数）

変数	m	共通分散	p	LL	CEIC	EIC	AIC	BIC
1,2,3	1	—	9	−416.666	849.853	850.106	851.331	878.428
1,2,3	2	no	19	−273.404	**589.126**	**589.408**	584.808	**642.010**
1,2,3	3	no	29	−259.182	592.324	593.063	576.364	663.672
1,2,3	4	no	39	−248.401	599.829	603.891	**574.802**	692.217
1,2,3	2	yes	13	−333.632	693.677	692.078	693.260	732.402
1,2,3	3	yes	17	−308.070	657.110	659.423	650.140	701.321
1,2,3	4	yes	21	−317.696	646.566	644.975	631.315	740.615
1,2,4	1	—	9	−340.507	697.902	697.587	699.013	726.110
1,2,4	2	no	19	−211.850	467.974	470.010	461.701	**518.902**
1,2,4	3	no	29	−190.528	456.229	456.156	439.057	526.364
1,2,4	4	no	39	−175.596	**441.604**	**439.680**	**429.192**	546.607
1,2,4	2	yes	13	−272.764	573.538	571.987	571.528	610.666
1,2,4	3	yes	17	−237.194	516.830	516.487	508.388	559.569
1,2,4	4	yes	21	−223.155	500.511	499.917	488.310	551.533
1,3,4	1	—	9	−347.815	712.655	710.811	713.630	740.726
1,3,4	2	no	19	−208.358	458.074	459.531	454.716	511.918
1,3,4	3	no	29	−181.283	**433.259**	**430.809**	420.566	**507.874**
1,3,4	4	no	39	−166.182	434.219	436.975	**410.365**	527.779
1,3,4	2	yes	13	−281.487	593.116	595.033	588.974	628.112
1,3,4	3	yes	17	−242.475	524.413	523.482	518.949	570.131
1,3,4	4	yes	21	−208.996	467.333	472.139	459.992	523.215
2,3,4	1	—	9	−342.593	704.184	706.176	703.187	730.282
2,3,4	2	no	19	−186.369	418.023	416.557	410.739	467.940
2,3,4	3	no	29	−155.252	382.124	387.615	368.503	**455.812**
2,3,4	4	no	39	−137.778	**381.403**	**378.710**	**353.556**	470.971
2,3,4	2	yes	13	−260.650	550.018	550.707	547.300	586.438
2,3,4	3	yes	17	−222.272	487.865	485.606	478.545	529.725
2,3,4	4	yes	21	−190.658	434.410	433.894	423.316	486.539

ブートストラップ反復回数 $B=200$，m：混合分布の成分数，p：モデルのパラメータ数．

れたデータセットである．スイスで昔使われた1000フラン紙幣の真贋を検証するためのデータセットで，6つの測定値（Length: Length of bill，Left: Width of left edge，Right: Width of right edge，Bottom: Bottom margin width，Top: Top margin width，Diagonal: Length of diagonal，いずれも単位は mm）と，本物の紙幣と偽物の紙幣の測定値の各100個体からなる．

Swiss Bank Notes データの成分数推定結果を表7に示す．真贋を見極めるものなので，本来は2群で推定してほしいが，いくつかの例外を除いて4群で推定されている．では4群で具体的にどのように推定されているのか，オリジナルの分類とのクロス集計を次に示す．

	(123456)		(23456)		(1456)		(123)		(456)	
	1	2	1	2	1	2	1	2	1	2
1	75	0	79	0	96	0	28	87	96	0
2	24	0	20	0	3	8	70	13	3	2
3	1	15	1	15	1	15	1	0	1	15
4	0	85	0	85	0	77			0	85

括弧内の数字は変数を表し，その下（表頭）の1と2はオリジナルの分類番号，表側の1から4は推定された分類番号である．変数番号が1，2，3を除いて，いずれもオリジナルの群をそれぞれ2つに分ける形で推定されている．すなわち，本来一つの群を2つの多変量正規分布であてはめた方が，情報量規準の観点では効率的＝あてはまりが良いということを示す．真贋の2種類を4群が最適と推定したことは，多変量正規分布の1つの成分で表現できないことを示している．データの出所が詳しく記されていないので想像の域を出ないが，それぞれが2種類の母集団からのデータと考えられないだろうか．

5 おわりに

情報量規準に基づく多変量混合正規分布モデルの成分数の推定に関する研究成果として，Nakamura and Konishi (2016) で主張するものを補完するものとして本報告をまとめた．

世界中で取得される様々なデータは〝ビッグデータ〟として日々大量に蓄積され，様々な統計的分類モデル

表 5 ：Iris dataset（ 2 変数）

変数	m	共通分散	p	LL	CEIC	EIC	AIC	BIC
1,2	1	—	5	−270.772	551.393	554.684	551.540	566.597
1,2	2	no	11	−225.916	476.002	**474.339**	473.831	**506.949**
1,2	3	no	17	−217.127	**475.554**	475.126	468.255	519.435
1,2	4	no	23	−210.448	484.553	486.077	**466.897**	536.141
1,2	2	yes	8	−245.432	509.774	510.572	506.864	530.949
1,2	3	yes	11	−235.934	501.385	505.282	493.867	526.985
1,2	4	yes	14	−223.547	481.548	485.615	475.095	517.243
1,3	1	—	5	−374.607	758.105	758.187	759.215	774.267
1,3	2	no	11	−259.344	545.552	547.016	540.688	**573.805**
1,3	3	no	17	−250.313	**542.139**	**538.993**	**534.627**	585.807
1,3	4	no	23	−244.999	554.211	555.953	535.998	605.243
1,3	2	yes	8	−309.049	634.176	636.062	634.098	658.183
1,3	3	yes	11	−284.307	595.813	594.860	590.615	623.731
1,3	4	yes	14	−274.228	585.474	585.747	576.456	618.605
1,4	1	—	5	−272.725	555.177	551.303	555.451	570.503
1,4	2	no	11	−200.370	425.551	424.489	422.740	**455.857**
1,4	3	no	17	−190.639	**418.641**	**419.418**	415.278	466.459
1,4	4	no	23	−183.096	422.932	422.805	**412.193**	481.437
1,4	2	yes	8	−248.757	513.893	511.780	513.514	537.599
1,4	3	yes	11	−218.305	463.466	460.250	458.610	491.727
1,4	4	yes	14	−204.997	445.417	444.610	437.995	480.143
2,3	1	—	5	−370.153	749.279	746.912	750.306	765.359
2,3	2	no	11	−237.352	**501.004**	**500.562**	496.705	**529.821**
2,3	3	no	17	−232.478	509.204	508.924	498.956	550.137
2,3	4	no	23	−224.516	509.846	509.589	**495.032**	564.277
2,3	2	yes	8	−289.492	597.558	597.325	594.985	619.069
2,3	3	yes	11	−269.977	568.574	571.286	561.954	595.071
2,3	4	yes	14	−258.027	553.722	551.852	544.054	586.203
2,4	1	—	5	−248.596	506.526	503.748	507.193	522.245
2,4	2	no	11	−140.981	307.394	306.956	303.961	337.079
2,4	3	no	17	−124.187	**293.061**	**293.040**	282.375	**333.555**
2,4	4	no	23	−112.341	298.301	295.879	**270.682**	339.927
2,4	2	yes	8	−185.831	387.776	388.200	387.663	411.747
2,4	3	yes	11	−150.816	326.437	329.147	323.631	356.749
2,4	4	yes	14	−139.201	310.752	308.895	306.402	348.551
3,4	1	—	5	−272.792	554.610	553.814	555.583	570.637
3,4	2	no	11	−154.731	335.708	330.889	331.463	364.579
3,4	3	no	17	−134.136	312.821	313.440	302.271	**353.453**
3,4	4	no	23	−122.924	**304.800**	**304.949**	**291.848**	361.093
3,4	2	yes	8	−223.336	463.458	465.279	462.671	486.757
3,4	3	yes	11	−189.814	407.541	406.094	401.629	434.745
3,4	4	yes	14	−159.053	349.618	348.190	346.107	388.255

ブートストラップ反復回数 $B=200$，m：混合分布の成分数，p：モデルのパラメータ数．

表 6 ：Crabs dataset

変数	m	共通分散	p	LL	CEIC	EIC	AIC	BIC
1,2,3,4,5	1	—	20	−1481.88	3004.41	3004.42	3003.76	3069.73
1,2,3,4,5	2	no	41	−1354.16	2799.15	2798.38	2790.31	2925.55
1,2,3,4,5	3	no	62	−1281.28	2722.71	2728.26	2686.56	2891.06
1,2,3,4,5	4	no	83	−1223.69	**2667.82**	**2670.77**	**2613.39**	2887.14
1,2,3,4,5	5	no	104	−1203.49	2713.77	2716.81	2614.98	2958.01
1,2,3,4,5	2	yes	26	−1413.51	2880.83	2883.75	2879.02	2964.78
1,2,3,4,5	3	yes	32	−1380.58	2836.24	2832.88	2825.16	2930.71
1,2,3,4,5	4	yes	38	−1349.05	2790.52	2789.97	2774.11	2899.44
1,2,3,4,5	5	yes	44	−1317.61	2747.80	2743.88	2723.22	**2868.35**

ブートストラップ反復回数 $B=200$，m：混合分布の成分数，p：モデルのパラメータ数．

表 7 ：Swiss Bank Notes dataset

変数	m	共通分散	p	LL	CEIC	EIC	AIC	BIC
1,2,3,4,5,6	2	no	55	−718.396	1564.79	1563.68	1546.79	1728.20
1,2,3,4,5,6	3	no	83	−628.161	1495.22	1490.54	1422.32	1696.08
1,2,3,4,5,6	4	no	111	−588.708	1509.27	1508.50	**1399.42**	1765.53
1,2,3,4,5,6	2	yes	34	−793.642	1669.92	1670.96	1655.28	1767.43
1,2,3,4,5,6	3	yes	41	−698.121	1489.24	1482.99	1478.24	1613.48
1,2,3,4,5,6	4	yes	48	−676.486	**1468.37**	**1462.20**	1448.97	**1607.29**
2,3,4,5,6	2	no	41	−671.425	1432.67	1427.47	1424.85	1560.08
2,3,4,5,6	3	no	62	−590.616	1352.15	1351.52	1305.23	1509.73
2,3,4,5,6	4	no	83	−557.339	1346.14	1341.88	1280.68	1554.44
2,3,4,5,6	2	yes	26	−729.822	1522.10	1525.18	1511.64	1597.40
2,3,4,5,6	3	yes	32	−640.677	1350.35	1344.04	1345.35	1450.90
2,3,4,5,6	4	yes	38	−621.020	**1317.62**	**1318.49**	**1318.04**	**1443.38**
1,4,5,6	2	no	29	−718.427	1503.08	1504.96	1494.85	1590.51
1,4,5,6	3	no	44	−642.448	1403.32	1407.28	1372.90	1518.02
1,4,5,6	4	no	59	−621.737	**1396.78**	**1398.47**	**1361.47**	1556.07
1,4,5,6	2	yes	19	−766.402	1578.57	1585.48	1570.80	1633.47
1,4,5,6	3	yes	24	−677.048	1408.74	1408.70	1402.10	**1481.26**
1,4,5,6	4	yes	29	−667.362	1405.79	1408.71	1392.72	1488.38
1,2,3	2	yes	19	−174.715	384.598	387.832	375.431	**418.318**
1,2,3	3	yes	24	−166.006	**375.002**	**368.177**	**366.012**	422.083
1,2,3	4	yes	29	−163.062	378.049	371.495	368.124	437.398
4,5,6	2	no	29	−649.140	1339.51	1340.85	1336.28	1398.95
4,5,6	3	no	44	−593.373	1257.78	1257.15	1244.75	1340.40
4,5,6	4	no	59	−578.830	**1243.99**	**1244.06**	**1235.66**	1364.29
4,5,6	2	yes	19	−683.274	1399.28	1405.74	1392.55	1435.43
4,5,6	3	yes	24	−607.876	1249.82	1252.21	1249.75	**1305.82**
4,5,6	4	yes	29	−603.016	1250.96	1251.04	1248.03	1317.30

ブートストラップ反復回数 $B=200$，m：混合分布の成分数，p：モデルのパラメータ数．

の適用対象とされる．本報告で示した多変量混合正規分布モデルに基づく分類法は古典的なモデルとして位置づけされるであろう．ノンパラメトリック分布に基づく分類方法は，効果的な，ある意味予測的に効率的な分類境界を求めることが目的の方法である．本研究のパラメトリックな分類モデルは，その成分分布に重要な意味がある．つまり，推定された確率分布に何らかの各専門分野（物理学，工学，生物学等）での意味や知見を見いだすことが目的と言える．自然科学系のデータは，データの出所が複雑な母集団分布ではなく，比較的単純な分布であると考えるのが自然であろう．この意味で混合分布モデルは，確率分布の平均対数尤度の推定という観点のモデルであり，データの持つ情報量を表現するための最適なものと言える．

謝辞　本研究は札幌学院大学研究促進奨励金（SGU-S11-198007-01）の支援を受けた．

参考文献

[1] Aggarwal, C.C. and Reddy, C.K. (2014). *Data Clustering, Algorithms and Applications*, Chapman & Halls.

[2] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle, Proc. of 2nd International Symposium on Information Theory (eds. B. Petrov and F. Csaki), 267-281, Akademiai Kiado, Budapest.

[3] Akaike, H. (1974). A new look at statistical model identification, IEEE Transactions on Automatic Control, 6, 716-723.

[4] Campbell, N.A. and Mahon, R.J. (1974). A multivariate study of variation in two species of rock crab of genus Leptograpsus, Australian Journal of Zoology, 22, 417-425.

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

[6] Efron, B. (1979). Bootstrap methods: another look at the jackknife, Annals of Statistics, 7, 1-26.

[7] Everitt, B.S. (1993). *Cluster Analysis*, Third edition, Wiley-Halsted, London.

[8] Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture Distributions*, Chapman and Hall, New York.

[9] Flury, B. and Riedwyl, H. (1988). *Multivariate Statistics: A practical approach*. London: Chapman & Hall, Tables 1.1 and 1.2, pp. 5-8.

［10］Ishiguro, M., Sakamoto, Y., and Kitagawa, G. (1997). Bootstrapping log-likelihood and EIC, an extension of AIC, Annals of the Institute of Statistical Mathematics, 49, 411–434.
［11］Konishi, S. and Kitagawa, G. (1996). Generalized information criterion and bootstrap method, Biometrika, 83, 875–890.
［12］McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley, New York.
［13］McLachlan, G.J. and Basford, K.E. (1988). *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York.
［14］McLachlan, G.J. and Peel, D. (2000). *Finite Mixture Models*, Wiley.
［15］Mengresen, K.L., Robert, C.P., and Titterington, D.M. (2011). *Mixtures: Estimation and Applications*, Wiley.
［16］中村永友（1995）. 多変量正規混合分布モデルに基づく分類法, 計算機統計学, 8, 117–133.
［17］中村永友（2009）. 多次元データ解析法, 共立出版, 東京.
［18］中村永友（2007）. 混合分布モデル,「統計・データ科学事典」, 杉山高一・藤越康祝・杉浦成昭・国友直人編, 朝倉書店.
［19］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.
［20］Nakamura, N. and Konishi, S. (2016). Estimating the number of components for multivariate normal mixture models via bootstrap information criteria, preparing to submit.
［21］中村永友・小西貞則・大隅昇（1993）. 混合分布モデルを用いた画像分類と色彩変換 ── LANDSAT 画像の解析 ──, 統計数理, 41, 149–167.
［22］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, 57–75.
［23］Schwarz, G. (1978). Estimating the dimension of a model, Annals of Statistics, 6, 461–464.
［24］Titterington, D.M., Smith, A.F.M., and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York.

The Proceedings of the Research Institute of Sapporo Gakuin University Vol.3, 7-15 (2016)

Numerical Validations in Estimation of the Number of Components in a Normal Mixture Model for Famous Datasets

Nagatomo NAKAMURA[1]

Abstract

The effectiveness of the bootstrapping information criterion for estimating the number of components in a multivariate normal mixture model used as a statistical classification model is verified by several numerical experiments. We report that the proposed estimation procedure was applied to several famous datasets of problem statistical classification.

Keywords: Normal Mixture Model, Bootstrapping, Information Criteria, EM Algorithm.

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