判別方式のコンピュータ側面からの検討

丘 本 正*, 谷 口 るり子**

Computer-aided Comparison of Two Discrimination Procedures

Masashi OKAMOTO and Ruriko TANIGUCHI

Suppose that some data set consisting of several groups with variables of continuous type is given and discrimination of the data is required. Then two discrimination procedures are available. The first is canonical discriminant analysis (CDA) which uses the variables as they are, while the second is Hayashi's second method of quantification (Q2), canonical discriminant analysis for categorized data.

This paper compares the two procedures from two points of view; the value of correlation ratio and the misclassification rate when the results of the analysis are applied to the original data themselves.

As an illustration two data sets were used. The first is based on medical claim units for the elderly compiled by Kōsū Eisei Shinkōkai. The second is the famous iris data treated often in discriminant analysis. Neither of the two procedures can be said better than the other.

1. 序

多くの群からなる判別分析において変数が連続型であるとき、データの数値そのままですべて判別分析（略称 CDA, Rao (1952) 参照）を行うと、データをカテゴリー化したあとで数量化法第2類（略称 Q2, 岩坪 (1987) 参照）を適用するのと2つの解析方式が考えられる。

公衆衛生振興会 (1994) の資料（府川他 (1994) 参照）を材料として、CDA と Q2 とを比較検討する。比較の規準として、まず相関比の値がある。次に、判別結果を自分自身に適用したときの誤判別の割合がある。

上記資料にはいくつかの道府県が含まれ、各道府県はそれぞれいくつかの（2次）医療圏からなっている。

判別分析においては道府県を群とし、医療圏を個体とみなすことにする。したがって問題は医療の地域差の判別ということになるが、本研究の狙いは当面のデータにとどまらず、より一般に統計的な手法としての方法論の比較に貢献することにある。そのため判別分析でしばしば取り上げられるあやめデータをも検討した。

Q2 においてデータをカテゴリー化するとき情報をいくらか失うので、この方法は CDA に比べて一般に判別力が減少するのではないかとされているが、実際はいつもそうとは限らないようである。また Q2 は CDA と違ってデータが離散的な場合にも利用できるという大きい長所をもっているが、一つの問題点は、データが連続型の変数をもつ場合に、解析の結果がカテゴリー化における境界値の選び方に依存することである。本論文においては恣意性を除くため、各カテゴリーが現れる度数ができるだけ等しくなるように境界値を選ぶことにした。

2つの手法の比較に関してはシミュレーションによる評価も考えられるが、多変量正規分布に基づくデータでは CDA が Q2 にまさるであろうと予想され、適切なデータの設計方針が思い当たらないので、他日を期

* 大阪大学 (Osaka University), 追手門学院大学 (Otemon Gakuin University) とともに名誉教授

** 大阪国際女子短期大学 (Osaka International College for Women)

連絡先氏名: 丘本 正

郵便宛先: 〒665-0831 兵庫県宝塚市米谷 2-15-21

電話番号: 0797-86-9353

することにした。

判別のために用いた群と変数を以下に示す。

- (1) 群: 北海道(21), 青森(6), 福島(6), 富山(4), 石川(4), 福井(4), 滋賀(7), 大阪(4), 和歌山(6)

ただし, カッコ内は医療圏の個数。全個数 62。

- (2) 変数: 70 歳以上の老人について

医療費 2 変数: 入院, 入院外別 1 人当たり医療費の医療圏別平均。(単位: 千円)

傷病件数率 7 変数: X_2 (新生物), X_6 (神経系), X_{71} (高血圧), X_{72} (心疾患), X_{73} (脳血管疾患), X_9 (消化系), X_{13} (筋骨格系) (単位: 0.1%)

のそれぞれを入院, 入院外別とする 14 変数。

合計 16 変数である。

データの全体を表 1 に, 府県別平均を表 2 に示す。医療圏の規模は表 2 では考慮されているが, 以後の解析では考慮されていない。群の並べ方は北海道と大阪に関して前述(1)の順序と少し異なっている。これは後に示す散布図(図 1)の都合のためである。

2. 正準判別分析による分析

まず CDA で解析した。9 群, 16 変数であるから得られる軸の個数は $\min(9-1, 16)=8$ である。相関比と変数の標準化スコアを第 4 軸まで表 3 に示すが, 入費と外費は医療費の入院と入院外をさし, 入 X_n と外 X_n は傷病件数率 X_n の入院と入院外をさす。また正準変量の群平均を表 4 に, 軸 1~3 による個体の散布図を図 1 に示す。図 1 において北海道を表す円が小さいのは北海道の順位を下げた効果である。

正準変量(個体スコア)の各軸はプールされた群内分散行列の意味で無相関かつ長さ 1 となっている。また変数の標準化スコアは正準変量における変数ごとのスコアに, その変数の標準偏差をかけることによって標準化した。(例えば, 田中他(1984), p. 179 参照。)このため標準化スコアは変数の尺度に依存せず, それらの大小を見て変数の寄与を相互に比較することができる。

第 1 軸の相関比 0.893 は十分に大きく, 図 1 の 2 つの散布図は 9 個の群をかなりよく判別している。群内の医療圏の個体差は地域差に比べて小さいことがわかる。

表 3 から最初の 3 つの正準変量について変数のスコアが正負それぞれに著しいものに着目すると表 5 が得

a. 第 1, 2 軸

b. 第 1, 3 軸

図 1. CDA の散布図

られる。

また表 4 から各正準変量に対する群のスコアの中で正負の著しいものは表 6 のようである。この表は表 2 と表 5 からいくらか説明することができる。例えば北海道が表 6 の Z_1 に現れる理由は, 表 5 の Z_1 に現れる入 X_{71} について, 表 2 の北海道の値 20 は入 X_{71} の列の中に現れる 9 個の値の最大値であり, また外 X_{71} の値はその列の中の第 2 最大値であるためである。

3. 数量化法第 2 類による解析

3.1 カテゴリー化

16 個の項目(変数)は次の 2 つの境界値によってそれぞれ 3 個のカテゴリーに分けた。

項目 1: 1150, 1410 項目 2: 255, 300
項目 3: 8, 10 項目 4: 5, 6

表 1. 医療圏別医療費平均と傷病件数率 (1992 年度, 2+14 変数)

	医療費		傷病件数率 (入院)							傷病件数率 (入院外)						
	入院	外	X_2	X_6	X_{71}	X_{72}	X_{73}	X_9	X_{13}	X_2	X_6	X_{71}	X_{72}	X_{73}	X_9	X_{13}
青 森	1149	254	8	5	16	10	26	11	12	35	232	315	74	125	113	211
	1216	264	7	5	11	12	32	9	12	31	247	322	91	126	117	189
	1152	270	8	6	6	16	25	10	13	42	289	282	115	84	138	246
	1137	225	9	6	17	13	27	12	14	39	154	320	77	124	156	175
	1075	265	8	6	9	11	27	11	12	38	229	295	103	119	117	210
	1011	228	5	5	8	11	14	7	9	31	175	299	140	103	99	217
福 島	1020	233	9	4	7	10	16	6	5	34	177	359	106	122	126	153
	1081	248	9	4	6	11	16	9	7	35	158	381	104	117	113	151
	1010	235	10	5	8	13	20	8	6	35	133	354	117	103	116	146
	1177	235	10	5	7	11	21	8	9	36	227	325	116	117	109	214
	938	217	9	5	4	7	12	7	4	40	147	350	106	114	103	208
	1075	248	11	4	6	13	18	8	6	40	211	295	133	116	116	162
大 阪	1495	388	11	5	11	11	23	9	7	48	345	297	95	93	157	192
	1445	335	9	5	9	10	20	7	7	45	337	301	94	80	143	206
	1477	358	13	5	12	10	23	8	6	52	335	285	92	102	140	178
	1512	322	11	5	11	11	26	9	8	50	275	282	97	101	137	182
富 山	1614	259	10	6	7	11	51	5	14	42	285	263	91	106	92	205
	1478	265	10	7	10	12	46	7	21	39	259	253	99	107	98	239
	1436	277	7	5	9	10	28	7	13	36	270	226	112	112	105	247
	1307	263	5	3	9	8	18	8	10	34	200	304	89	109	102	194
石 川	1383	299	12	5	11	15	23	11	20	39	223	210	140	122	110	227
	1571	311	12	5	10	16	30	9	13	44	178	245	144	115	118	161
	1333	273	9	6	14	18	23	11	14	36	169	234	150	103	108	196
	1141	270	10	5	10	12	15	9	13	38	235	303	115	105	130	205
福 井	1147	259	13	5	10	12	25	9	18	33	212	177	106	140	124	240
	931	269	12	4	9	12	21	4	10	31	183	156	105	131	102	248
	1079	218	14	3	11	11	25	8	14	27	185	176	85	89	130	246
	1142	239	9	4	8	10	13	5	7	40	243	258	131	89	99	214
滋 賀	1543	273	10	6	7	10	24	6	9	42	224	281	98	90	101	223
	1253	223	10	4	5	11	16	8	5	39	170	256	111	89	101	182
	1164	224	8	4	9	10	14	7	6	28	154	338	81	103	99	177
	1186	220	6	5	6	10	15	5	4	31	174	316	102	102	100	150
	1163	213	7	4	5	10	14	6	5	33	212	272	103	103	110	205
	1059	225	8	4	4	8	9	5	4	33	205	304	102	126	112	173
	1068	206	8	5	9	8	13	9	9	38	179	298	104	84	91	180
北海道	1654	342	8	7	18	13	28	12	14	31	297	296	94	107	132	197
	1553	334	10	10	27	16	31	17	19	21	145	322	91	96	106	167
	1568	349	6	8	30	13	28	17	18	22	186	314	110	81	106	152
	1941	344	10	8	21	17	50	14	17	35	223	291	98	87	119	172
	1762	356	8	6	18	20	40	12	18	23	211	312	104	70	115	189
	1813	308	7	7	25	15	44	14	17	25	194	309	84	90	129	177
	1746	318	7	8	20	13	41	15	18	27	192	292	93	113	133	200
	1648	312	6	7	21	14	71	14	17	32	163	295	85	129	108	208
	1660	317	9	5	17	14	42	13	16	35	208	308	99	83	112	172
	1623	335	8	6	22	17	33	15	15	33	161	324	100	99	100	177
	1354	283	10	8	23	17	33	9	18	32	117	350	84	81	80	159
	1317	265	7	8	16	11	24	9	14	25	188	347	97	101	100	161
	1392	270	8	6	15	11	23	12	14	38	197	321	90	89	124	187
	1319	331	6	5	21	16	32	15	16	20	109	355	82	101	114	177
	1504	308	13	9	19	18	25	15	14	32	123	307	149	83	119	190
	1314	299	7	7	27	11	17	9	13	23	93	379	91	68	95	131
	1237	286	10	6	13	15	22	9	13	33	176	283	127	102	107	174
	1367	316	9	8	32	21	22	14	19	26	128	297	110	63	98	168
	1215	273	7	5	18	10	23	6	10	27	219	363	86	82	105	156
	1465	306	8	7	20	12	25	11	13	29	205	363	105	86	120	163
	1203	297	6	2	22	9	22	13	12	23	176	375	100	97	110	123
和歌山	1203	296	8	5	7	8	16	9	9	43	274	263	100	94	171	198
	1114	305	7	5	7	8	14	7	8	42	249	290	94	92	153	196
	1114	223	7	5	7	9	13	7	8	43	231	324	75	77	157	171
	1056	235	8	3	4	7	8	5	4	41	195	278	102	110	142	213
	1015	221	8	4	6	8	11	6	5	36	227	311	96	86	117	183
	1037	292	6	4	5	7	17	5	10	34	214	285	96	109	110	197

表 2. 府県別医療費, 傷病件数率の平均 (1992 年度, 2+14 変数)

	医療費		傷病件数率 (入院)							傷病件数率 (入院外)						
	入院	外	X_2	X_6	X_{71}	X_{72}	X_{73}	X_9	X_{13}	X_2	X_6	X_{71}	X_{72}	X_{73}	X_9	X_{13}
青 森	1146	255	8	6	12	12	26	10	12	36	232	307	94	114	124	209
福 島	1077	238	10	4	7	11	18	8	7	36	184	345	113	117	115	170
大 阪	1487	357	11	5	11	11	23	8	7	49	325	292	95	94	146	189
富 山	1462	267	9	6	9	10	37	7	16	38	256	255	100	109	100	230
石 川	1443	296	11	5	11	15	23	10	15	39	201	244	138	112	116	199
福 井	1107	247	12	4	10	11	23	8	14	33	209	190	106	118	119	238
滋 賀	1246	230	8	5	6	10	15	6	6	35	193	293	101	101	103	186
北海道	1655	321	9	7	20	15	37	12	15	30	206	312	98	90	116	175
和歌山	1131	274	8	5	6	8	14	8	8	41	248	281	97	94	151	194

医療費: 単位千円, 傷病件数率: 単位 0.1%

表 3. 原データの正準判別分析 (1992 年度, 2+14 変数)

	第 1 軸	第 2 軸	第 3 軸	第 4 軸		
定数項	-26.44	-9.62	-14.83	24.28		
固有値	8.343	3.960	3.279	2.395		
相関比	0.893	0.798	0.766	0.706		
	標準化スコア				平均	標準偏差
入費	0.27	-0.51	-0.50	0.75	1309	158
外費	0.00	0.93	-0.37	-0.07	278	25
入 X_2	0.02	0.06	0.06	-0.20	8.7	1.6
X_6	-0.09	0.25	0.23	0.59	5.5	1.2
X_{71}	0.59	0.18	-0.35	-0.78	12.8	3.5
X_{72}	0.20	-0.15	0.63	-0.20	12.0	2.4
X_{73}	0.31	0.22	0.25	-0.28	24.6	9.2
X_9	0.02	0.00	0.62	0.53	9.4	2.1
X_{13}	0.34	-0.61	-0.47	-0.67	11.5	2.7
外 X_2	-0.16	0.45	0.11	-0.41	34.6	4.6
X_6	0.13	0.08	0.27	-0.28	205	40
X_{71}	1.10	0.18	0.97	-0.42	298	29
X_{72}	0.72	-0.28	0.24	-0.84	103	15
X_{73}	0.24	-0.31	0.51	-0.49	101	15
X_9	0.01	0.35	0.01	-0.54	116	14
X_{13}	0.18	-0.04	0.13	-0.40	189	22

固有値: λ , 相関比: $\lambda/(1+\lambda)$

項目 5: 9, 16 項目 6: 11, 13
 項目 7: 20, 26 項目 8: 8, 10
 項目 9: 9, 14 項目 10: 33, 39
 項目 11: 180, 225 項目 12: 290, 320

項目 13: 95, 105 項目 14: 90, 107
 項目 15: 107, 120 項目 16: 175, 200

例えば項目 1 の変数の値を x で表すとき, $x < 1150$ ならばカテゴリー 1, $1150 \leq x < 1410$ ならばカテゴリー

表4. 正準変量4個の群平均 (1992年度, 2+14変数)

	Z_1	Z_2	Z_3	Z_4
青森	0.01	-0.49	2.18	-1.51
福島	-0.33	-0.23	3.83	-0.31
大阪	-1.60	5.71	-0.83	-0.80
富山	-1.05	-1.88	-1.84	0.06
石川	0.36	-1.60	-0.33	-2.86
福井	-4.10	-2.31	-2.66	-1.78
滋賀	-2.85	-1.26	0.47	2.90
北海道	3.28	0.11	-0.69	0.49
和歌山	-3.57	1.86	-0.37	0.31

表5. 正準変量における変数の特徴

軸	正	負
Z_1	外 X_{71} , 外 X_{72} , 入 X_{71}	
Z_2	外費	入 X_{13} , 入費
Z_3	外 X_{71} , 入 X_{72}	入費

表6. 正準変量における群の特徴

軸	正	負
Z_1	北海道	福井, 和歌山
Z_2	大阪	福井
Z_3	福島, 青森	福井

2, $x \geq 1410$ ならばカテゴリー3が実現したとみなすのである。このとき3個のカテゴリーは項目ごとに個体総数62をほぼ3等分している。

3.2 解析結果

このデータをQ2によって解析すると、カテゴリースコアについて表7, 個体スコアの群平均と相関比について表8, 第1~3軸の個体スコアの散布図として図2が得られる。このとき田中他(1984)による個体スコアの各軸を群内分散が1になるように標準化した。相関比は3つの軸についてCDAの場合よりもかなり増加したことが注目される。地域差は個体差に比べてさらに大きい。

表7において軸ごとに範囲の大きさを見ると、どの項目がその軸の中で大きく貢献しているかがわかる。最初の3つの軸について大きく貢献している2~3個

a. 第1, 2軸

b. 第1, 3軸

図2. Q2の散布図

の変数を選び、第3カテゴリーのスコアの符号によって正と負に分類すると次の表9が得られる。

次に、表8から最初の3つの軸について、個体スコアの群平均の中で正負の著しいものは表10のようになる。

3.3 考察

相関比が大きいためQ2はCDAにまさと簡単に結論してよいかどうかは疑問である。自由度の問題がある。表1のデータは9群, 62個体であって、データの自由度 = $62 - (9 - 1) = 54$ となる。一方、解析方法の自由度は、CDAについて変数の個数16であり、Q2については、3カテゴリーに分けたため、自由度 = $16 * (3 - 1) = 32$ となる。Q2における相関比の増加はかなりの程度まで自由度の増加によるものと考えられる。

表10を表6と比較すると、軸1においてはかなり似ているが、軸番号が増すにつれて違いが大きくなる。また、表9と表5との似かたも残念ながら不十分であると言わざるを得ない。方法が変われば切り口も変わる

表7. 数量化法第2類の結果 (カテゴリー別スコアと範囲: 92年度, 2+14変数)

項目	カテゴリー	軸1		軸2		軸3	
		スコア	範囲	スコア	範囲	スコア	範囲
入 費	1	-0.93	1.59	-1.24	2.52	1.19	2.12
	2	0.65		-0.05		-0.93	
	3	0.31		1.28		-0.30	
外 費	1	-0.45	1.13	-1.09	2.03	-0.03	1.22
	2	0.68		0.13		0.60	
	3	-0.30		0.94		-0.62	
入 X_2	1	0.17	0.75	0.37	0.79	0.55	1.38
	2	0.29		-0.42		0.32	
	3	-0.46		0.14		-0.83	
X_6	1	-0.39	1.06	0.97	1.70	1.69	2.81
	2	-0.42		0.14		0.07	
	3	0.64		-0.73		-1.12	
X_{71}	1	-1.88	4.17	-2.61	4.35	-0.90	1.94
	2	-0.42		0.90		1.04	
	3	2.28		1.75		-0.09	
X_{72}	1	-0.82	1.28	0.18	1.04	-0.72	1.26
	2	0.43		0.47		0.53	
	3	0.46		-0.56		0.29	
X_{73}	1	0.42	1.20	0.17	1.68	-0.14	0.34
	2	-0.78		0.78		-0.07	
	3	0.32		-0.91		0.20	
X_9	1	0.09	0.98	0.86	1.58	0.11	0.30
	2	-0.56		-0.07		-0.19	
	3	0.42		-0.72		0.08	
X_{13}	1	-0.78	1.29	1.06	2.11	-2.16	3.88
	2	0.51		-1.05		0.08	
	3	0.20		0.03		1.72	
外 X_2	1	-0.06	0.18	-0.66	1.10	-0.39	1.15
	2	0.12		0.44		-0.29	
	3	-0.06		0.28		0.76	
X_6	1	0.22	0.84	0.14	0.85	-0.50	1.72
	2	0.28		-0.48		-0.54	
	3	-0.57		0.37		1.18	
X_{71}	1	-0.65	0.98	1.45	3.16	0.81	1.38
	2	0.33		0.14		-0.24	
	3	0.29		-1.70		-0.57	
X_{72}	1	0.01	1.44	0.54	1.55	0.18	1.89
	2	-0.75		0.49		-1.06	
	3	0.70		-1.01		0.83	
X_{73}	1	0.26	1.18	0.55	1.22	-0.13	0.54
	2	0.49		0.20		0.33	
	3	-0.69		-0.67		-0.21	
X_9	1	-0.47	1.02	0.53	1.26	0.30	0.70
	2	0.55		-0.73		0.06	
	3	-0.13		0.26		-0.40	
X_{13}	1	0.51	1.13	0.11	0.30	0.36	0.88
	2	-0.63		0.08		-0.53	
	3	0.15		-0.19		0.19	

表 8. Q2 の結果(個体スコアの群平均と相関比:
92 年度, 2+14 変数)

群	軸 1	軸 2	軸 3
青 森	0.79	-3.87	1.84
福 島	-3.27	-5.53	-1.40
大 阪	-4.25	7.70	-1.47
富 山	-0.94	2.82	3.19
石 川	0.99	-0.09	3.07
福 井	-3.27	1.08	4.47
滋 賀	-2.97	0.30	-2.51
北海道	4.46	0.58	-0.83
和歌山	-4.67	-0.64	-0.80
相関比	0.938	0.913	0.831

表 9. Q2 の軸における項目の特徴

軸	正	負
軸 1	入 X_{71}	
軸 2	入 X_{71} , 入費	外 X_{71}
軸 3	入 X_{13}	入 X_6 , 入費

表 10. Q2 の軸における群の特徴

軸	正	負
軸 1	北海道	大阪, 和歌山
軸 2	大阪	福島, 青森
軸 3	福井, 富山, 石川	滋賀

ということであるが, 地域差は個体差に比べて大きい
という結論は変わらない。

4. カテゴリー数の削減

表 7 は Q2 に関する 1 つの重要な問題点を示してい
る。例えば軸 1 について項目 1 (入費) の 3 つのカテ
グリーのスコア

-0.93, 0.65, 0.31

を見ると, 増加のあと減少に転じ, 項目 4 (X_6) につ
いては

-0.39, -0.42, 0.64

のように減少のあと増加に転じている。固有ベクトル

a. 第 1, 2 軸

b. 第 1, 3 軸

図 3. Q2R の散布図

には符号の自由性があるから, この両者は同じ現象で
あり, 「折り返し」と呼びたい。このような現象は Q2
では許されるが, CDA では許されていない。この点に
Q2 の CDA に対する優位性があるようである。

筆者の考えでは, 第 2 軸以降はともかくとして, 特
別な事情がない限り, 第 1 軸においては折り返しを許
さない方がデータ解析として自然であると思われる。
そこで折り返しをおこなった 14 個の項目については, 境
界値 1 個を用いて 2 カテゴリーに再編成した。その際,
新しい境界値は 2 つのカテゴリーの度数がほぼ等しく
なるように選んだ。それらの項目番号と境界値は次の
ようである。

項目番号	1	2	3	4	7	8	9
境界値	1,300	272	9	6	24	9	13
項目番号	10	11	12	13	14	15	16
境界値	35	200	300	100	102	113	185

このデータに Q2 を適用し, 表 7 に対応するスコア
の表を作ったとき, 3 カテゴリーをもつ 2 個の項目の

表 11. Q2R の結果 (カテゴリースコア, 92年度, 2+14変数)

項目	カテゴリー	軸1		軸2		軸3	
		スコア	範囲	スコア	範囲	スコア	範囲
入 費	1	-0.39	0.80	-1.77	3.65	0.38	0.78
	2	0.41		1.88		-0.41	
外 費	1	-0.33	0.67	-0.03	0.07	1.35	2.70
	2	0.33		0.03		-1.35	
入 X_2	1	0.34	0.70	-0.84	1.73	-0.19	0.40
	2	-0.36		0.89		0.21	
X_6	1	-0.29	0.74	0.60	1.56	0.14	0.35
	2	0.45		-0.95		-0.21	
X_{71}	1	-1.68	3.59	-0.80	2.15	-1.37	2.15
	2	-0.25		1.36		0.61	
	3	1.91		-0.50		0.79	
X_{72}	1	-0.04	0.09	0.46	1.01	0.04	0.09
	2	0.05		-0.55		-0.05	
X_{73}	1	0.57	1.27	0.37	0.83	-0.83	1.83
	2	-0.70		-0.45		1.00	
X_9	1	-0.21	0.38	0.37	0.65	0.60	1.06
	2	0.17		-0.28		-0.46	
X_{13}	1	-0.87	1.81	-0.54	1.11	-0.47	0.97
	2	0.93		0.57		0.50	
外 X_2	1	0.47	0.92	-0.01	0.01	0.52	1.00
	2	-0.44		0.01		-0.48	
X_6	1	-0.15	0.28	-0.40	0.78	-0.18	0.35
	2	0.14		0.38		0.17	
X_{71}	1	0.08	0.16	-0.05	0.09	-0.50	1.01
	2	-0.08		0.05		0.50	
X_{72}	1	0.03	0.07	-0.42	0.82	-0.35	0.68
	2	-0.03		0.40		0.33	
X_{73}	1	0.49	0.94	-0.37	0.72	0.37	0.72
	2	-0.46		0.35		-0.35	
X_9	1	0.19	0.38	-0.07	0.14	0.69	1.38
	2	-0.19		0.07		-0.69	
X_{13}	1	0.07	0.14	0.12	0.24	-0.40	0.77
	2	-0.07		-0.12		0.38	

うち項目番号6でまたもや折り返しがおこったので、境界値12を用いて再び2カテゴリーに編成し直した。今度は折り返しがおこらず、最終的に3カテゴリーが生き延びたのは項目5(入 X_{71}) だけであった。このデータにQ2を適用した結果をQ2R (R: reduced) とよぶことにする。

Q2のスコアを示す表7, 群平均と相関比の表8に対

応して, Q2Rについてそれぞれ表11と表12ができる。図3は散布図を与える。相関比は表13に見られるようにQ2に比較して減少したものの, CDAに比べると最初の2つの軸についてはかなり大きい。第3軸以後は逆転している。

Q2について項目の特徴を示す表9, 群の特徴を示す表10に対応して, Q2Rについてそれぞれ表14と表15

表 12. Q2R の結果(群平均, 92 年度, 2+14 変数)

群	軸 1	軸 2	軸 3
青 森	-1.64	-2.90	1.28
福 島	-4.43	-0.18	-1.27
大 阪	-0.99	4.71	-2.45
富 山	-0.76	3.46	2.49
石 川	0.15	4.01	-0.47
福 井	-1.86	0.65	3.93
滋 賀	-2.28	-0.85	0.63
北海道	3.91	-0.63	-0.24
和歌山	-2.64	-2.28	-2.25
相関比	0.913	0.849	0.757

表 13. 3 つの方法の相関比の比較

方法\軸	1	2	3	4
CDA	0.893	0.798	0.766	0.706
Q2	0.938	0.913	0.831	0.755
Q2R	0.913	0.849	0.757	0.636

表 14. Q2R の軸における項目の特徴

軸	正	負
軸 1	入 X_{71}	
軸 2	入費, 入 X_{71}	
軸 3	入 X_{71} , 入 X_{73}	外費

表 15. Q2R の軸における群の特徴

軸	正	負
軸 1	北海道	福島, 和歌山
軸 2	大阪, 石川, 富山	青森
軸 3	福井, 富山	大阪

ができる。

これら 2 つの表は Q2 に関する表 9 と表 10 のそれぞれとはいくらか似ているが, CDA の場合との違いはそれほど小さいとは言えないようである。

5. 自 己 判 別

判別方式 CDA と Q2 のどちらについても判別の結果を自分自身に適用することができる。まず, 各方式

表 16. CDA の自己判別

真	判 別 結 果										誤判別
	A	Fs	O	T	I	Fi	S	H	W		
青 森(6)	6									0	
福 島(6)		6								0	
大 阪(4)			4							0	
富 山(4)				4						0	
石 川(4)	1				3					1	
福 井(4)				1		3				1	
滋 賀(7)							7			0	
北海道(21)					1			20		1	
和歌山(6)							1		5	1	

表 17. Q2R の自己判別

真	判 別 結 果										誤判別
	A	Fs	O	T	I	Fi	S	H	W		
青 森(6)	5						1			1	
福 島(6)		6								0	
大 阪(4)			4							0	
富 山(4)				4						0	
石 川(4)					4					0	
福 井(4)						4				0	
滋 賀(7)							7			0	
北海道(21)								21		0	
和歌山(6)									6	0	

は個体スコアとその群平均とを 8 次元ベクトルとして与える。前述したように個体スコアの各軸はプールされた群内分散行列の意味で無相関かつ長さ 1 であるから, 個体と群との間の距離を個体スコアのベクトルと群平均のベクトルとの間のユークリッド 2 乗距離で測るのは自然であろう。

1 つの個体が与えられたとき, 9 個の群平均との間のユークリッド 2 乗距離の最小を与える群にその個体を判別することができる。これを自己判別と呼ぼう。このときに定まる群は本来その個体が属していた群であるとは限らない。たまたま一致しておればよいが, 一致しなければ誤判断がおこったことになる。

判別方式は誤判別の割合が小さいほどよいというべきであろう。この観点から CDA と Q2 とを比較してみることにした。表 16 は CDA による自己判別の結果であって, 群の名称, 個体数, 自己判別の正誤の度数を

示している。誤判別は

石川→青森, 福井→富山, 北海道→石川,
和歌山→滋賀

の 4 回であった。一方, 表 17 は Q2R の結果であって誤判別は

青森→滋賀

の 1 回だけとなった。Q2R が CDA にまさることになる。ちなみに Q2 の場合の誤判別は「北海道→石川」の 1 回である。

6. あやめデータの場合

これまで考察してきたことを Fisher (1936) によって適用されたあやめデータについて検討してみたい。群はイリス・セトーサ, イリス・ヴェルシコロール, イリス・ヴィルジニカの 3 種類であって, 各群とも個体数は 50 である。データは例えば田中他 (1984) に与えられている。変数はがくの長さ (X_1), がくの幅 (X_2), 花卉の長さ (X_3), 花卉の幅 (X_4) の 4 個である。

まず CDA の場合, 相関比などを表 18 に示す。第 1 軸の相関比 0.970 は非常に大きい。第 1 軸は X_3 に, 第 2 軸は X_2 に支配されていることがわかる。手法の比較に興味があるので, 群平均は省略する。自己判別の結果を下に示す。

真	判別結果			誤判別
	1	2	3	
1	50			0
2		48	2	2
3		1	49	1

表 18. CDA の結果 (あやめデータ)

	第 1 軸	第 2 軸
固有化	32.19	0.285
相関比	0.970	0.222
	標準化スコア	
X_1	-0.43	0.01
X_2	-0.52	0.74
X_3	0.95	-0.40
X_4	0.58	0.58

Q2 の場合, 4 項目に対して次の境界値を用いた。

$$X_1: 5.5, 6.3, \quad X_2: 2.9, 3.2$$

$$X_3: 3.3, 4.9, \quad X_4: 1.0, 1.6$$

これらは個体総数 150 を 3 個のカテゴリにほぼ等分している。しかし残念ながらカテゴリ化したデータをプログラムにかけたとき, エラーがおこってしまった。どれか 1 つの項目を 2 カテゴリに落とせばエラーが解消するのではないかと想像して, 項目 X_4 について成功した。 X_4 の境界値は 1.4 とした。

カテゴリ数 (3, 3, 3, 2) のときの相関比は

$$\eta_1^2 = 0.973, \quad \eta_2^2 = 0.744$$

となり, η_2^2 は CDA のときに比べて大幅に増加した。カテゴリのスコアを表 19 に示す。 X_1 で折り返しがおこっている。自己判別の結果は次のようである。

真	判別結果			誤判別
	1	2	3	
1	50			0
2	1	45	4	5
3		3	47	3

η_2^2 は増加しているのに誤判別の度数がふえたことは筆者には説明できない。

次に X_1 の境界値 5.8 を用い, カテゴリ数 (2, 3,

表 19. Q2 の結果 (あやめデータ)

項目	カテゴリー	軸 1		軸 2	
		スコア	範囲	スコア	範囲
X_1	1	-0.02	0.14	-0.33	0.68
	2	-0.06		-0.01	
	3	0.08		0.35	
X_2	1	0.43	0.73	0.10	0.25
	2	-0.07		0.09	
	3	-0.30		-0.15	
X_3	1	-7.98	12.45	-0.16	4.50
	2	3.73		2.40	
	3	4.47		-2.10	
X_4	1	-0.19	0.39	0.39	0.82
	2	0.20		-0.42	
相関比		0.973		0.744	

表 20. Q2R の結果 (あやめデータ)

項目	カテゴリー	軸 1		軸 2	
		スコア	範囲	スコア	範囲
X_1	1	-0.03	0.06	-0.14	0.27
	2	0.03		0.13	
X_2	1	0.41	0.70	0.04	0.21
	2	-0.06		0.10	
	3	-0.29		-0.11	
X_3	1	-7.97	12.46	-0.40	4.33
	2	3.70		2.44	
	3	4.49		-1.90	
X_4	1	-0.19	0.40	0.43	0.89
	2	0.21		-0.46	
相関比		0.973		0.741	

3, 2) に落として Q2 を行った。これを Q2R とよぶことにする。相関比は

$$\eta_1^2 = 0.973, \quad \eta_2^2 = 0.741$$

となった。カテゴリーのスコアを表 20 に示すが、折り返しはない。自己判別の結果は直前の場合と全く同じであった。折り返しを避ける方針の効果が出ることを期待していたが、実現しなかったのは残念である。

7. 終わりに

2 つの判別方式 CDA と Q2 の比較はデータに依存する。どの方式を採用すべきかという問題に対して、査読者方は必ずしも支持されたのではないが、筆者の意見は次のようである。

まず Q2 については、すべての項目についてカテゴリー数を 3 以下として複雑な折り返しを避けるのがよい。次に、もし相関比について CDA の方が大きいならば、それを採用するのがよいであろう。Q2 の相関比の方が大きいときは、カテゴリースコアに折り返しがおこっているかどうかを見ることを勧めたい。折り返しがおこっていれば、それを消すように 2 カテゴリーに削減したい。本論文のように削減をくり返すのを避け

るためには、各項目のスコアにカテゴリーに関し単調な解を生成する解析方式を利用することができれば、それも有効であろう。このときはカテゴリー数 4 以上も許される。最後に自己判別を行って、誤判別の割合と相関比の値の 2 つを勘案して、どちらの方式がよいかを総合的に判断するのがよいと思われる。

本研究ノートの審査中に著者たちの共同研究が進んで、論文：丘本・谷口 (1998) 「地域差の判別分析」が発表された。そこでは第 5 節で考察した自己判別方式の代替方法として、交差検証法 (cross-validation method) に基づく Lachenbruch-Mickey (1968) の交差判別を検証した。医療データの場合、個体総数 62 に対し誤判別の度数は次の表のようである。

判別方式	CDA	Q2R	Q2
自己判別	4	1	1
交差判別	11	15	25

交差判別の場合、Q2R は Q2 に比べて誤判別の度数が減り、折り返しを除いた効果が表れている。また自己判別の成績が悪かった CDA が名誉を回復している。

参 考 文 献

- Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics.*, 7, 179-188.
- 府川哲夫・星 旦二・郡司篤晃 (1994). 老人医療の構造分析. 季刊・社会保障研究, 30 巻 1 号, 90-98.
- 岩坪秀一 (1987). 数量化法の基礎. 朝倉書店.
- 公衆衛生振興会 (1994). 老人医療年齢階級別分析事業 1993 年度報告書.
- Lachenbruch, P.A. and Mickey, M.R. (1968). Estimation of error rates in discriminant analysis. *Technometrics*, 10, 1-11.
- 丘本 正・谷口るり子 (1998). 地域差の判別分析. 郡司篤晃編『老人医療費の研究』, 丸善プラネット, 第 5 章.
- Rao, C.R. (1952). *Advanced Statistical Methods in Biometric Research*. Wiley, New York.
- 田中 豊・垂水共之・脇本和昌 (1984). パソコン統計解析ハンドブック. 第 2 巻. 多変量解析編. 共立出版.