K14-20

Let's Try 多変量

～プロ野球('94)を分析する～

工学部技術部　松本 英敏

1．はじめに

今年のプロ野球は、黒田や松坂がメジャーからの復帰で話題になっている。直接の関係は無いが今回の技術発表では、古いデータであるイチローのプロ入団 3 年目（'94）に照準をあてて、多変量解析を行ってみた。皆様にとって身近なプロ野球を､表 1 に分類してある分析を通して見ていただきたい。

表 1　多変量解析の分類

多変量解析	目的変数		説明変数	
	質	量	質	量
重回帰分析		○		○
数量化Ⅰ類		○	○	
判別分析	○			○
数量化Ⅱ類	○		○	
ｸﾗｽﾀｰ分析				○
数量化Ⅲ類			○	
数量化Ⅳ類			○	
主成分分析				○
因子分析				○

2．多変量解析

多変量解析（Multivariate analysis）とは何ぞや。変量の種類がたくさんあって変量同志の種々の関係を明らかにして将来の予測とか変量の分類、合成とかを決めていこうとするものである。

表 1 より目的変数（外的基準）が有るか無いか、それらの目的変数が質で与えられているか、量で与えられているか。また、説明変数（要因）も同様に質か量かによって分析法が 9 種類に分かれる。

3．OERA モデル

この OERA(Offensive Earned Run Average)モデルは野球における打者の評価として、コウバーとケイラーによる「野球のための OERA 計算法」という論文で紹介されており、打者の評価を客観的に行うもので、チーム力の強弱による誤差を無くそうとする計算方法である。

表 2　'94 セ・パ両リーグ打撃ベストテン

順位	氏名	打率	試合数	打席数	打数	得点	安打	二塁打	三塁打	本塁打	塁打数	打点	長打率	三振	四球	死球	犠打	犠飛	盗塁	出塁率	OERA
8	ﾊﾟｳｴﾙ （中）	0.324	110	475	423	61	137	23	0	20	220	76	0.520	73	42	5	0	5	3	0.387	7.883
9	前田智 （広）	0.321	123	554	492	66	158	26	0	20	224	66	0.496	56	57	0	2	3	4	0.389	7.601
2	江藤　（広）	0.320	105	466	390	83	125	21	0	28	230	81	0.590	81	69	1	0	6	7	0.418	9.887
16	和田　（神）	0.318	130	601	519	76	165	13	3	2	190	43	0.366	40	65	3	11	3	8	0.395	6.016
3	ﾌﾞﾗｯｸﾞｽ（横）	0.315	122	534	448	84	141	25	1	35	273	91	0.609	83	68	12	0	6	1	0.414	9.866
6	ｵﾏﾘｰ　（神）	0.314	124	524	430	61	135	18	2	15	202	74	0.470	74	89	1	0	4	2	0.429	8.805
5	大豊　（中）	0.310	130	560	477	83	148	24	2	38	290	107	0.608	97	71	3	0	9	1	0.396	9.218
17	野村　（広）	0.303	130	614	558	77	169	20	4	10	227	61	0.407	75	45	4	4	3	37	0.359	5.425
19	川相　（巨）	0.302	130	567	473	69	143	18	4	0	169	33	0.357	51	54	3	35	2	3	0.376	5.350
13	ローズ （横）	0.296	130	574	510	71	151	28	4	15	232	86	0.455	72	55	2	0	7	1	0.362	6.304
1	ｲﾁﾛｰ　（オ）	0.385	130	616	546	111	210	41	5	13	300	54	0.549	53	51	10	7	2	29	0.445	10.605
7	山本　（ダ）	0.317	115	509	420	64	133	21	2	11	191	62	0.455	60	77	6	1	5	3	0.425	8.521
4	石井　（近）	0.316	130	574	487	88	154	31	2	33	288	111	0.591	87	74	7	0	6	1	0.409	9.486
12	松永　（ダ）	0.314	116	541	477	74	150	20	4	8	202	55	0.423	73	60	0	2	2	8	0.390	6.611
18	小川　（オ）	0.303	126	527	459	48	139	17	5	4	178	53	0.388	70	44	3	13	8	2	0.362	5.410
14	福良　（オ）	0.301	114	477	386	49	116	23	1	3	150	50	0.389	39	49	7	33	2	4	0.387	6.268
11	ﾗｲﾏｰ　（ダ）	0.298	127	525	470	57	140	33	2	26	255	97	0.543	91	40	13	0	2	4	0.368	7.428
20	辻　（西）	0.294	105	458	412	63	121	21	1	4	156	45	0.379	51	35	4	2	5	9	0.351	5.090
10	ﾌﾞﾗｲｱﾝﾄ（近）	0.293	105	481	437	80	128	23	1	35	258	106	0.590	153	39	2	0	3	0	0.351	7.490
15	初芝　（ロ）	0.290	129	521	476	66	138	31	5	17	230	75	0.483	90	36	4	1	4	1	0.342	6.044

この計算方法では、両リーグを通じて OERA が最も高いのは順位に示してある通りイチロー、江藤、ブラッグス、石井、大豊、…の順となっている。打撃ベストテンに入っていても、長打の少ない選手（16～20 位）は OERA の低い値を示すが、ベストテンに入ってなくても清原（西）（8.664）、松井（巨）（6.617）などは長打力があるので比較的高い OERA 値を示している。以下、この OERA を目的変数として、打率、打点などの説明変数を用いて多変量解析を行っていく。

４．重回帰分析

重回帰分析は表 1 から判るように、目的変数（OERA)、説明変数（打率、本塁打、四死球）とも量で表される。これを次の線形の予測式

$$y(\text{OERA}) = a_0 + a_1 \cdot x_1(\text{打率}) + a_2 \cdot x_2(\text{本塁打}) + a_3 \cdot x_3(\text{四死球})$$

に当てはめた結果

$$y = -9.968 + 46.013x_1 + 0.073x_2 + 0.032x_3 \qquad (1)$$

という回帰式になった。この場合、重相関係数が 0.983 と非常に高い値を示している。また、標準偏回帰係数は打率（0.684)、本塁打（0.463)、四死球（0.304）となり、数値が影響の大きさを表している。そこで、他の影響を排除した偏相関係数は

$$\begin{bmatrix} 1.0 & 0.958 & 0.931 & 0.828 \\ & 1.0 & -0.896 & -0.717 \\ & & 1.0 & -0.761 \\ & & & 1.0 \end{bmatrix}$$

OERA は打率との相関が最も高く、ついで本塁打、四死球の順となっている。以上の結果より、個人のデータ（打率、本塁打、四死球）を式(1)に代入すれば、その OERA 値を求めることができる。

表 3　パリーグベスト 15 の成績表

氏　名	OERA	打率	本塁打	四死球
ｲﾁﾛｰ	10.605	0.385	13	61
山本	8.521	0.317	11	83
石井	9.486	0.316	33	81
松永	6.611	0.314	8	60
小川	5.410	0.303	4	47
福良	6.268	0.301	3	56
ﾗｲﾏｰ	7.428	0.298	26	53
辻	5.090	0.294	4	39
ﾌﾞﾗｲｱﾝﾄ	7.490	0.293	35	41
初芝	6.044	0.290	17	40
田中（日）	6.223	0.286	27	43
佐々木（西）	5.256	0.285	20	35
吉永（ダ）	6.303	0.284	19	44
平井（ロ）	4.446	0.280	5	24
広瀬（日）	4.880	0.280	2	68

５．数量化理論Ｉ類

数量化理論Ｉ類は、目的変数は重回帰分析と同じであるが、説明変数（アイテム）が質的に表される。表 3 のデータについて打率、本塁打、四死球を優、良、可の 3 段階のカテゴリーに分類したのが表 4 である。この表を用いて数量化理論Ｉ類で分析した結果、表 5 のようになった。重相関係数は 0.923 であり、結果の精度はきわめて良好といえる。図 1 はアイテムの範囲で、これは目的変数（外的基準）への影響度を示しており、本塁打、打率、四死球の順に影響している。

表 4　質的に表した成績表

OERA	打率			本塁打			四死球		
	優	良	可	優	良	可	優	良	可
10.605	1				1		1		
8.521	1				1		1		
9.486	1			1			1		
6.611	1					1	1		
5.410	1					1		1	
6.268		1				1		1	
7.428		1		1				1	
5.090		1				1			1
7.490		1		1				1	
6.044		1			1			1	
6.223			1	1				1	
5.256			1		1				1
6.303			1		1			1	
4.446			1			1			1
4.880			1			1	1		

表 5　数量化理論Ｉ類による分析結果

アイテム	カテゴリー	ｶﾃｺﾞﾘｰ数量	範囲	偏相関係数
打率	優	0.841	1.866	0.729
	良	0.184		
	可	−1.025		
本塁打	優	1.032	2.240	0.838
	良	0.623		
	可	−1.207		
四死球	優	0.909	1.430	0.664
	良	−0.426		
	可	−0.521		

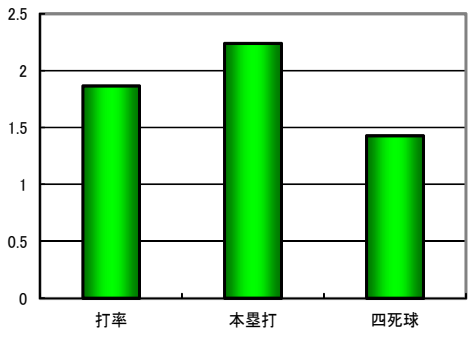

図 1　目的変数への影響度

6．判別分析

ここでは、パリーグ打撃 20 傑を質的な目的変数として 2 つのグループ、説明変数は量的なもので打率、本塁打、四死球を取り上げる。表 6 に示す 2 つのグループは OERA の上位 10 人と下位 10 人に分けてある。

判別分析を適用した結果、表 7 のようになり 2 つのグループの分散共分散行列が等しいときは、線形判別関数によって判別を行う。その等しいかどうかの検定は χ^2 検定を用いておこない

$$\chi_0^{\,2} < \chi^2\{q(q+1)/2, a\}$$

を満足すれば 2 つのグループは等しいといえる。q は説明変数の数で、a は有意水準 5%を検定するので、$a = 0.05$ である。χ^2 分布表より求めた結果

$$\chi_0^{\,2} = 8.702$$

$$\chi^2(6, 0.05) = 12.59$$

となり、検定式を満足するので線形判別関数により判別を行った。その結果

$$y = 71.305x_1 + 0.165x_2 + 0.073x_3 - 27.438 \qquad (2)$$

そこで、各選手のデータを式(2)に代入して適中率を求めてみた。表 8 がその結果であり、グループ 1 では正が真であり、グループ 2 は負が真となる。グループ 1 で 8 割、グループ 2 では 10 割適中している。

表 6　グループ 1,グループ 2 の打撃 10 傑

氏　名	打率	本塁打	四死球
ｲﾁﾛｰ	0.385	13	61
石井	0.316	33	81
清原	0.279	26	105
山本	0.317	11	83
ﾌﾞﾗｲｱﾝﾄ	0.293	35	41
ﾗｲﾏｰ	0.298	26	53
松永	0.314	8	60
吉永	0.284	19	44
福良	0.301	3	56
田中	0.286	27	43

氏　名	打率	本塁打	四死球
平井	0.281	5	24
ﾄﾗｯｸｽﾀｰ	0.263	15	41
石毛	0.266	11	41
広瀬	0.281	2	68
辻	0.294	4	39
佐々木	0.285	20	35
小川	0.303	4	47
鈴木	0.263	19	51
初芝	0.290	17	40
ﾎｰﾙ	0.277	22	44

表 7　グループ 1,グループ 2 の平均値、分散、共分散行列

ｸﾞﾙｰﾌﾟ1	平均	分散	標準偏差
打率	0.307	0.001	0.031
本塁打	20.1	119.9	10.95
四死球	62.7	432.7	20.80
ｸﾞﾙｰﾌﾟ2	平均	分散	標準偏差
打率	0.28	0.000	0.014
本塁打	11.9	58.3	7.637
四死球	43.0	129.3	11.37

ｸﾞﾙｰﾌﾟ1	打率	本塁打	四死球
打率	0.001	−0.117	0.034
本塁打	−0.117	119.9	−3.937
四死球	0.034	−3.967	432.7
ｸﾞﾙｰﾌﾟ2	打率	本塁打	四死球
打率	0.000	−0.045	−0.013
本塁打	−0.045	58.3	−14.89
四死球	−0.013	14.89	129.3

表 8　判別結果の適中率

氏　名	y	判定
ｲﾁﾛｰ	6.608	○
石井	6.444	○
清原	4.403	○
山本	3.035	○
ﾌﾞﾗｲｱﾝﾄ	2.215	○
ﾗｲﾏｰ	1.964	○
松永	0.648	○
吉永	−0.845	×
福良	−1.395	×
田中	0.543	○

氏　名	y	判定
平井	−4.833	○
ﾄﾗｯｸｽﾀｰ	−3.256	○
石毛	−3.666	○
広瀬	−2.146	○
辻	−2.969	○
佐々木	−1.265	○
小川	−1.744	○
鈴木	−1.853	○
初芝	−1.039	○
ﾎｰﾙ	−0.850	○

7．数量化理論Ⅱ類

数量化理論Ⅱ類は、目的変数、説明変数ともに質的に与えるデータである。これは表 4 で OERA が 7 以上を優、6 代を良、6 以下を可として質的に表したものである。

分析したところ、表 10 のような結果となり重相関係数は 0.930 と良好であった。また、グループ間の判別のよさを意味する指標の相関比は $\eta^2 = 0.865$ であり、よい判別をしたといえる。

この結果、OERA が約-0.14 以下のときは優、0.15 以上のときは可、その中間を良と予測することができる。

表 9　質的な目的変数

OERA	打率			本塁打			四死球		
	優	良	可	優	良	可	優	良	可
1	1				1		1		
1	1				1		1		
1	1			1			1		
2	1					1	1		
3	1					1		1	
2		1				1		1	
1		1		1				1	
3		1				1			1
1		1		1				1	
2		1			1			1	
2			1	1				1	
3			1		1				1
2			1		1			1	
3			1			1			1
3			1			1	1		

8．クラスター分析

これからは目的変数が無い場合について分析を行っていく。クラスター分析は説明変数を量的に扱うのでここでは、5 変量（打率、安打、本塁打、打点、OERA）にパリーグ 12 選手について行った。

分析に先立って非類似度の定義が必要である。非類似度は 5 つ存在するが、今回はユークリッド平方距離を用いることにする。また、クラスター分析も 7 種類存在するが、最短距離法によるデンドログラムを自動作成した。

表 11　クラスター分析の量的な説明変数

氏　名	打率	安打	本塁打	打点	OERA
ｲﾁﾛｰ	0.385	210	13	54	10.605
山本	0.317	133	11	62	8.521
石井	0.316	154	33	111	9.486
松永	0.314	150	8	55	6.611
小川	0.303	139	4	53	5.410
福良	0.301	116	3	50	6.268
ﾗｲﾏｰ	0.298	140	26	97	7.428
辻	0.294	121	4	45	5.090
ﾌﾞﾗｲｱﾝﾄ	0.293	128	35	106	7.490
初芝	0.29	138	17	75	6.044
田中	0.286	148	27	87	6.223
佐々木	0.285	150	20	84	5.256

表 10　数量化理論Ⅱ類による分析結果

アイテム	カテゴリー	ｶﾃｺﾞﾘｰ基準	範囲	偏相関係数
打率	優	−0.042	0.259	0.730
	良	−0.109		
	可	0.151		
本塁打	優	−0.227	0.440	0.873
	良	−0.074		
	可	0.213		
四死球	優	−0.148	0.292	0.733
	良	0.044		
	可	0.144		
外的基準	優		相関比	
OERA	良		0.865	
	可			

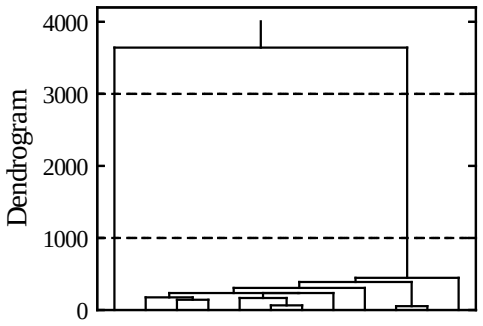

図 2　最短距離法によるデンドログラム

表 12　選手間の 5 変量による非類似度行列（ユークリッド平方距離）

氏名	山本	石井	松永	小川	福良	ﾗｲﾏｰ	辻	ﾌﾞﾗｲｱﾝﾄ	初芝	田中	佐々木
ｲﾁﾛｰ	6001.3	6786.3	3641.9	5149.9	8970.8	6928.1	8223.4	9921.7	5661.8	5148.2	4577.6
山本		3326.9	350.6	175.6	502.1	1500.2	493.7	2538.1	230.1	1111.2	864.6
石井			3785.2	4446.6	6075.3	445.2	6305.3	708.9	1819.8	658.6	931.8
松永				142.4	1206.1	2188.6	959.3	3814.7	625.3	1389.1	986.8
小川					539.7	2425.1	388.1	3895.3	654.4	1766.6	1338
福良						3315.3	52.3	4305.5	1305.1	2969	2602
ﾗｲﾏｰ							3354.4	306.0	570.9	166.4	309.7
辻								4736.7	1358.9	3023.2	2618
ﾌﾞﾗｲｱﾝﾄ									1387.1	826.6	1197.9
初芝										344.0	234.6
田中											62.9

9．数量化理論Ⅲ類

予測すべき外的基準が無い場合に数量化する一つの方法であり、個体（サンプル）とカテゴリーの両方を数量化する方法である。プロ野球に対して適当なデータが無いので省略する。発表時に時間があれば、年代別の趣味や食事の好みについて例題を取り上げる。

10. 数量化理論Ⅳ類

前述の数量化理論Ⅲ類と同様に外的基準が存在せず、多くの対象を似たものどうしに分類する方法である。これもプロ野球に関しては似たものどうしの適用が難しいので、発表時に時間があればアイドルやタレントなど親近性について例題を取り上げる。

11. 主成分分析

主成分分析とは多くの変量に異なる重みを付加して、互いに独立な合成変量を求める方法である。ここでは表13 に示すプロ野球投手の'94 セ・パ両リーグ 18 人の投手成績に適用してみる。

表 14 は勝率、投球回数、試合数、奪三振数の平均、分散、標準偏差を表している。また、相関行列の固有値解析を行った結果、第一主成分は最大固有値 2.506 となり 4 つの変量を最もよく代表している。寄与率についてみれば、このデータを約 63%説明している。そこで、第二主成分まで加えると約 86%説明していることになる。図 3 は第一、第二主成分得点を図化した。

表 14　投手の分析結果

	平均	分散	標準偏差
勝　数	10.56	7.20	2.68
投球回数	164.98	882.2	29.7
試合数	27.0	24.82	4.96
奪三振数	126.0	1744.2	41.76

表 15　主成分分析結果

変量＼主成分	1	2	3	4
勝　数	0.574	−0.049	−0.538	−0.615
投球回数	0.590	−0.130	−0.231	0.763
試合数	0.195	0.977	0.075	0.039
奪三振数	0.534	−0.161	0.807	−0.196
固有値	2.506	0.945	0.392	0.157
寄与率	0.626	0.236	0.098	0.039
累積寄与率	0.626	0.863	0.961	1.000

表 16　主成分スコア

氏名＼主成分	1	2	3	4
郭　（中）	−1.815	−0.860	−0.169	0.072
桑田 （巨）	2.37	−0.279	0.134	0.029
斉藤 （巨）	1.904	0.275	−0.620	0.210
槙原 （巨）	1.130	0.174	0.106	0.072
今中 （中）	1.581	−0.104	−0.145	0.129
岡林 （ヤ）	−0.364	−0.899	−0.815	0.177
湯舟 （神）	−2.256	−0.465	0.999	0.424
斉藤隆(横)	0.575	−0.011	1.033	0.574
佐藤 （中）	−1.484	0.452	0.493	0.302
新谷（ 西）	−0.611	3.013	0.072	−0.536
伊良部(ロ)	3.235	−0.701	0.961	−0.461
長谷川(オ)	−0.667	−0.209	−0.824	−0.153
山崎 （近）	0.077	0.067	−1.197	0.239
工藤 （西）	−0.729	−0.439	0.094	−0.996
佐藤 （オ）	−1.931	−1.048	0.040	−0.204
星野 （オ）	−0.834	−0.849	0.070	−0.435
吉田豊(ダ)	0.975	0.438	−0.387	0.339
河野 （日）	−1.157	1.443	0.156	0.218

表 13　両リーグ投手成績表

氏　名	勝数	投球数	試合数	奪三振
郭　（中）	8	139.3	21	85
桑田 （巨）	14	207.3	28	185
斉藤 （巨）	14	206.3	30	144
槙原 （巨）	12	185	29	153
今中 （中）	13	197	28	156
岡林 （ヤ）	11	171.6	22	95
湯舟 （神）	5	130	23	109
斉藤隆(横)	9	181	28	169
佐藤 （中）	7	140.6	28	104
新谷（ 西）	10	130	41	99
伊良部(ロ)	15	207.3	27	239
長谷川(オ)	11	156.3	25	86
山崎 （近）	12	179.6	27	85
工藤 （西）	11	130.6	24	124
佐藤 （オ）	8	130.6	20	93
星野 （オ）	10	143.3	22	119
吉田豊(ダ)	12	190.6	30	129
河野 （日）	8	143	33	94

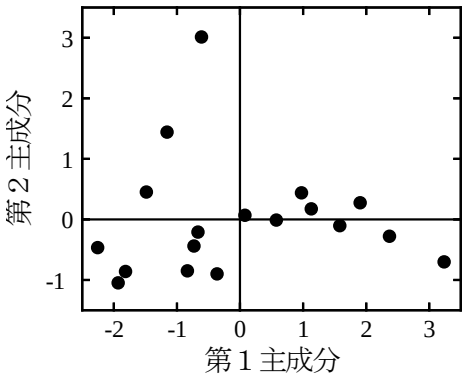

図 3　主成分得点による相関図

表 17　アンケートデータ一覧表

サンプル	長嶋	吉田	王	江夏	落合	バース	野村	川上	江川	掛布
1	2	3	1	2	3	2	3	2	1	2
2	3	1	1	3	3	1	1	1	3	1
3	3	2	2	3	3	2	3	2	3	3
4	2	1	2	3	3	2	3	2	3	1
5	3	3	2	3	3	3	2	1	3	3
6	2	2	2	2	2	2	2	2	2	2
7	3	3	3	1	1	3	1	3	2	3
8	3	2	1	1	2	2	2	2	1	3
9	3	3	2	3	3	3	3	3	2	2
10	3	3	1	3	2	3	2	1	1	3
11	2	3	1	3	3	3	1	1	1	2
12	3	2	3	3	3	2	1	1	3	3
13	1	2	1	2	3	3	2	1	3	3
14	3	3	1	3	3	2	3	1	1	3
15	3	3	2	3	3	2	3	2	3	3

12. 因子分析

最後にデータは古いが往年のプロ野球人を含めて因子分析を行った。因子分析は主成分分析と同様、目的関数は存在せずに説明変数を量的に用いる。

そこで、表 17 は好き（3)、どちらでもない（2)、嫌い（1）として数値で表してみた。人気度を表す平均値は、落合、長嶋、江夏がベスト 3 で下位には王、川上がいる。次のこのデータを基に相関行列を求めたのが表 19 である。比較的相関の高い組み合わせは、

江夏<->落合、吉田<->バース、吉田<->掛布、王<->江川

である。アウトロー同志と監督<->選手の組み合わせになっている。最後に相関行列に対する固有値を求め、共通因子数を 5 として主因子分析法により因子負荷量を求めてみた。図 4 はその一部を図示したものである。

表 18　平均値と分散

	平均値	分散	標準偏差
長嶋（OB）	2.600	0.400	0.632
吉田（OB）	2.400	0.543	0.737
王 （OB）	1.667	0.524	0.724
江夏（OB）	2.533	0.552	0.743
落合（ロ）	2.667	0.381	0.617
バース（神）	2.333	0.381	0.617
野村（OB）	2.133	0.695	0.834
川上（OB）	1.667	0.524	0.724
江川（巨）	2.133	0.838	0.915
掛布（神）	2.467	0.552	0.743

表 19　因子分析による相関行列

氏名	長嶋	吉田	王	江夏	落合	バース	野村	川上	江川	掛布
長嶋	1.000	0.215	0.312	0.182	−0.183	−0.183	−0.027	0.156	−0.025	0.274
吉田	0.215	1.000	0.000	−0.026	−0.157	0.628	0.140	0.134	−0.508	0.548
王	0.312	0.000	1.000	−0.044	−0.267	0.107	−0.158	0.455	0.503	0.177
江夏	0.182	−0.026	−0.044	1.000	0.727	−0.104	0.223	−0.443	0.308	−0.224
落合	−0.183	−0.157	−0.267	0.727	1.000	−0.250	0.370	−0.426	0.337	−0.260
バース	−0.183	0.628	0.107	−0.104	−0.250	1.000	−0.093	0.107	−0.211	0.415
野村	−0.027	0.140	−0.158	0.223	0.370	−0.093	1.000	0.316	−0.025	0.008
川上	0.156	0.134	0.455	−0.443	−0.426	0.107	0.316	1.000	−0.036	−0.089
江川	−0.025	−0.508	0.503	0.308	0.337	−0.211	−0.025	−0.036	1.000	−0.098
掛布	0.274	0.548	0.177	−0.224	−0.260	0.415	0.008	−0.089	−0.098	1.000

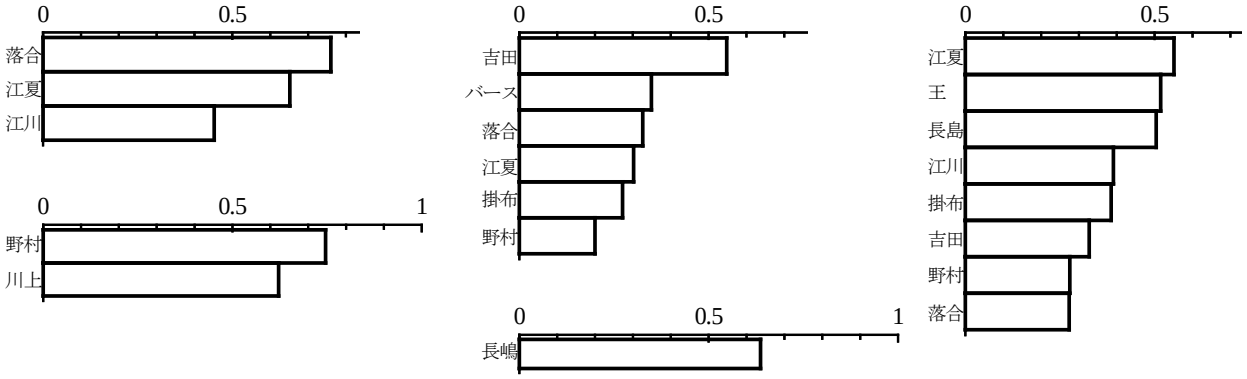

図 4　主因子分析法による因子負荷量

13. おわりに

今回は、多変量解析を用いて～プロ野球 ('94)パリーグを分析する～と銘打ったが、紙面の都合により尻切れトンボの感は否めない。詳細については、発表時に時間があれば言及する予定である。

このような身近な例題を通して、どのような時にどんな分析手法を使ってグループ分するのか？あるいはこの目的に対してその説明変数は果たして有意義なのか？等をご理解いただければと幸いに思います。尚、以上 9 個のプログラムは自作ですので、ご相談いただければ提供できます。

【参考文献】

木下栄蔵：わかりやすい数学モデルによる多変量解析，近代科学社
田中　豊：多変量統計解析，現代数学社