OR メモランダム

このコラムは，ORにかかわる概念，知識(手法，原理)，それらの図解，よい教材や問題，実学ORの実施経験，そこから得られた智恵やアドバイス，失敗談と教訓，新しい視点，視座，フレームワーク，未だ解けていない問題，面白い研究テーマなどを，“新鮮に”，しかも“コンパクトに”表現し，提示していただくものです．ユニークなアイディア，フレッシュな見方，発想，だれかと意見をたたかわせたい問題提起など，ふるってご投稿ください．(原稿は，刷り上がり，半ページから3ページに納まるようにお書きください．簡単に！ 加筆訂正をお願いする場合があります)

多変量解析における簡便法

－Sに学ぶ－

上田　太一郎

1．はじめに

判別分析，数量化理論II類（以下II類）は回帰分析，数量化理論I類と同様に多変量解析のなかで広く活用されている有効な手法です．また数量化理論III類（以下III類）も有効な手法としていろいろな分野で使われています．筆者も一企業のORマンとしていろいろな局面でパソコンソフトSを用いて適用しています．このSで簡単にできる判別分析（II類）の変数（要因）選択法を考えました．実際に適用し，有効な結果が得られているのでご紹介します．また，III類の簡便法についてもご紹介します．

2．判別分析，II類の変数（要因）選択法

判別分析はサンプル(データ)が，たとえば，合格・不合格のようにグループに分類されていて，同時にサンプルの数値データが所与のとき，数値データを説明変数として，できるだけ明確にグループ化するように未知係数（判別関数の係数）を求める方法です．さらに所属不明なサンプルがどのグループに属すのか判別（予測）にも使用されます．II類と判別分析との違いはII類は説明変数がアイテム・カテゴリと呼ばれる不連続な定性的なデータとなっている点です．判別分析，II類はともに要因分析と予測に役立つ手法です．

実際に適用する上で重要な点は，判別分析では説明変数，II類では要因（アイテム）をいかにして選択するか，つまり最適な説明変数，要因を求めることだと感じています．

ところで，パソコンで使用可能なソフトウェアSAS，SPSS等では判別分析の変数選択がサポートされています．しかしながら，II類の要因選択はサポートされていないようです．Sでもサポートされていません．そこでSを用いて簡単にできる判別分析(II類)の変数（要因）選択法を考えました．実際に適用し有効な結果が得られているのでご紹介します．

変数（要因）選択規準は，回帰分析の変数選択規準のアナロジーから思いつきました．回帰分析の変数選択規準の1つに佐和の予測用修正重相関係数R_sがあります [1]．R_sは次式で表わされます．

$$Rs=1-(1-R^2)(n-1)(n-2)/\{(n-p-2)(n-p-1)\}$$

ここで，n：データ数（サンプル数），p：変数の個数，R：重相関係数です．この規準値が最大となる変数の組合せを最適な変数としています．

提案する判別分析（II類）の変数（要因）選択規準D_u（仮に予測用修正正準相関係数と呼びます）は

$$D_u=1-(1-\eta_1)(n-1)(n-2)/\{(n-p-2)(n-p-1)\}$$

となります．ここでη_1は第1正準相関係数です．グループ数が2のときは判別関数は回帰式と同等（係数が比例関係にある）であることがわかっていて [2]，しかも$\eta_1=R$です．したがって，重相関係数のかわりに第1正準相関係数をもってきたわけです．このD_uが最大となる変数の組み合せを最適な変数とします [3]．II類ではD_uが最大となるアイテムの組合せです．またII類ではpは{カテゴリ総数－アイテム数}です．Sによるソースは以下のようになります．

```
function(x, y){n<-nrow(x)
p<-ncol(x) ; syusei<-((n-2)*(n-1))/
(n-p-2)/(n-p-1) ; 1-syusei *(1-discr
(x, y) $ co [1])  # co [1] は第1正準相関係数}
```

うえだ　たいちろう
三菱電機東部コンピュータシステム㈱　生産管理部
〒244　横浜市戸塚区川上町87-1

$discr(x, y)$はSの正準判別分析関数です．(相関比を最大にする判別分析では正準相関係数のかわりに相関比を用いることになります．)簡便法ですのでパソコンで簡単に使用できます．**2グループはもちろん，3グループ以上でも適用可能であることを強調しておきます．**

3．適 用 例

3.1 2グループの判別分析の例［4］

ある球団の採用テストの結果と合否の判定の結果は**表1**のようになっています．変数の組合せごとのD_u値，η_1を求めると**表2**のようになります．
D_uが最大となるX_1，X_3，X_4，X_5の組合せを最適な説明変数とします．［4］でも変数増
減法の結果は同様にX_1，X_3，X_4，X_5を選択しています．

3.2 3グループの判別分析の例

フィッシャーによるアイリス・データです．3種のアヤメの4部位の測定値（$X_1 \sim X_4$）が掲載されています（データは［5］を参照してください）．変数の組合せごとのD_u値，η_1を求めると**表3**のようになります．D_uが最大なのは$X_1 \sim X_4$です（0.984）．$X_2 \sim X_4$でも0.983とあまり違いがありませんので$X_2 \sim X_4$を選択してもよいでしょう．

3.3 2グループのII類の例［4］

電子レンジ保有世帯と非保有世帯2グループを外的規準とし，アイテムを世帯年収（I_1），主婦の就業状況（I_2），家族全員での食事回数（I_3）で判別する例です（データは**表4**）．アイテムの組合せごとのD_u値，η_1を求めると，**表5**のようになります．

I_1，I_2を選択したとき0.619と最大になりまた．最適なアイテムは世帯年収，主婦の就業状況ということです．

前述したように2グループのときの判別分析，II類は回帰分析で可能であることがわかっていますので，念のため回帰分析でアイテム選択をやりました．AICと佐和の予測用修正重相関係数Rsによると，ともにI_1，I_2を選択しました．

II類は判別分析の特殊なケースといわれています．判別分析プログラムでII類を実行するときは各アイテムのたとえば先頭カテゴリデータを削除して実行し，カテゴリスコアを求め，先頭カテゴリのカテゴリスコアは0とすればよいことが知られています．したがって，判別分析プログラムさえあればここで紹介した簡便法で判別分析の変数選択，II類のアイテム選択が容易に可能となります．判別分析はパソコンでサポートされているので簡便法を活用されてみてはいかがでしょうか．

4．III類と特異値分解

次のようなデータを考えます．年齢・性別の異なる人々に好きなスポーツを回答してもらった仮想のデータです（**表6**）．

表6のデータを特性値反応データとみなし，「もの」と「特性」との反応パターンからみて似たスポーツ同士は近くに，そうでないスポーツは遠くに配置するようにしたのがIII類です（同時に年齢・性別も配置し直してくれます）．

さて，III類と特異値分解とは数学的に同一のものであることがわかっています（鷲尾・大橋［6］，西里［7］，統計学辞典等）．そこで，このデータをIII類と特異値分解によりプロットしてみました（**図1，図2**）．ここで，III類ではλ_2（第2固有値）とλ_3に，特異値分解ではd_2（第2特異値）とd_3とに対応させています（d_2とd_3とに対応させたのがミソです）．図1を見ると右にいくほど高齢，左にいくほどヤングになっています．上半分は女性，下半分は男性がプロットされています．このようにヨコ軸，タテ軸で解釈することもできます．また，たとえばスキー，サッカー，女20未満，男20未満，男21～30は比較的近くにプロットされていることがわかるのでパターン分類をする手法ともいえます．図2でも同様な解釈が可能なことがわかります．簡単な例だと目視でも可能でしょうが，**表6**のデータでさえコンピュータの力が必要になります．「もの」や「特性」が多くなると，コンピュータプログラムにたよらざるをえないことになります．

このように数量化理論III類は大変有用であるにもかかわらず，コンピュータプログラムがない等の理由で利用したくても必ずしもうまく利用されているとは言えないようです．特異値分解はFORTRAN，C言語等でサブルーチンとしてサポートされている（たとえば［8］）ので活用されてみてはいかがでしょうか．

謝辞　原稿を読んでいただき，貴重なコメントをいただきました編集委員会に感謝いたします．

表1　採用テスト結果と合否

1：合格　－1：不合格

	100m走 x1	ボール投速度 x2	ボール速度 x3	ヒット数 x4	エラー数 x5	懸垂回数 x6	視力 x7	握力 x8	合否
1	13.8	62	120	3	2	71	1.0	80	-1
2	12.2	90	130	2	0	32	1.2	79	1
3	13.4	53	95	1	1	20	1.5	32	-1
4	12.7	88	141	3	1	28	0.9	68	1
5	12.9	79	128	2	0	45	1.2	50	1
6	11.9	88	118	1	1	30	1.3	70	1
7	10.9	83	108	5	2	22	1.2	56	1
8	15.0	53	87	0	0	40	1.4	38	-1
9	12.8	92	120	4	0	15	1.3	62	1
10	14.2	70	110	0	3	10	1.1	43	-1
11	11.7	70	100	3	1	10	1.2	78	1
12	11.3	82	127	2	1	31	1.5	76	1
13	13.5	87	112	1	2	13	1.3	47	-1
14	14.5	63	130	2	1	18	1.0	65	-1
15	15.0	79	99	4	3	15	1.2	72	-1
16	14.7	69	102	0	2	22	0.9	50	-1
17	12.5	77	130	3	2	28	0.9	48	1
18	13.2	78	110	1	0	17	1.4	39	-1
19	14.0	68	120	0	3	33	1.6	43	-1
20	12.9	81	128	3	0	42	1.5	52	1

出所：菅民郎『初心者がらくらく読める多変量解析の実践（上）』p.94

表2　変数の組合せと D_u 値，η_1

変数の組合せ	D_u	η_1
$X_1 \sim X_8$	0.729	0.913
$X_1 \sim X_7$	0.774	0.913
$X_1 \sim X_6$	0.798	0.908
$X_1 \sim X_5$	0.826	0.907
$X_1 \sim X_4$	0.813	0.885
X_1, X_3, X_4, X_5	0.838	0.901
X_1, X_3, X_4	0.828	0.880

表3　変数の組合せと D_u 値，η_1

変数の組合せ	D_u	η_1
$X_1 \sim X_4$	0.984	0.985
$X_1 \sim X_3$	0.980	0.981
$X_2 \sim X_4$	0.983	0.984
X_1, X_3, X_4	0.981	0.982

表4　電子レンジ保有・非保有世帯

電子レンジ保有世帯			電子レンジ非保有世帯		
世帯年収（2468万円以下／1000万円以上）	就業状況（主婦が有職／主婦が無職）	食事回数（週に5回以上／週に3回以下／週に2回以下）	世帯年収（2468万円以下／1000万円以上）	就業状況（主婦が有職／主婦が無職）	食事回数（週に5回以上／週に3回以下／週に2回以下）
0 1 0 0 0	0 1	1 0 0	1 0 0 0 0	0 1	1 0 0
0 1 0 0 0	1 0	0 1 0	1 0 0 0 0	0 1	1 0 0
0 0 1 0 0	0 1	0 1 0	1 0 0 0 0	1 0	1 0 0
0 0 1 0 0	1 0	0 0 1	1 0 0 0 0	0 1	1 0 0
0 0 1 0 0	1 0	0 1 0	1 0 0 0 0	0 1	0 1 0
0 0 1 0 0	1 0	0 0 1	1 0 0 0 0	1 0	0 1 0
0 0 1 0 0	1 0	0 1 0	1 0 0 0 0	1 0	0 0 1
0 0 0 1 0	0 1	0 0 1	0 1 0 0 0	0 1	0 0 1
0 0 0 1 0	0 1	1 0 0	0 1 0 0 0	0 1	0 1 0
0 0 0 1 0	1 0	0 0 1	0 1 0 0 0	0 1	0 1 0
0 0 0 1 0	1 0	0 0 1	0 1 0 0 0	0 1	1 0 0
0 0 0 1 0	0 1	0 0 1	0 1 0 0 0	0 1	1 0 0
0 0 0 1 0	1 0	0 1 0	0 1 0 0 0	1 0	1 0 0
0 0 0 0 1	1 0	0 1 0	0 0 1 0 0	0 1	1 0 0
0 0 0 0 1	0 1	0 0 1	0 0 1 0 0	0 1	1 0 0
0 0 0 0 1	0 1	0 0 1	0 0 1 0 0	0 1	0 0 1
0 0 0 0 1	1 0	0 0 1	0 0 1 0 0	1 0	1 0 0
0 0 0 0 1	1 0	0 0 1	0 0 1 0 0	1 0	0 1 0
0 0 0 0 1	1 0	0 1 0	0 0 0 1 0	0 1	1 0 0
0 0 0 0 1	1 0	1 0 0	0 0 0 1 0	0 1	0 1 0

出所：菅民郎『初心者がらくらく読める多変量解析の実践（下）』p.44

表5　アイテムの組合せと D_u 値，η_1

アイテムの組合せ	カテゴリ総数	D_u	η_1
I_1, I_2, I_3	7	0.605	0.735
I_1, I_2	5	0.619	0.712
I_1, I_3	6	0.590	0.708
I_2, I_3	3	0.442	0.525

表6　スポーツの好み　1：はい　0：いいえ

もの＼特性	テニス	スキー	サッカー	野球	ゴルフ
男20未満	1	1	1	1	0
男21～30	0	1	1	1	0
男31～40	0	1	0	1	1
男41以上	1	0	0	1	1
女20未満	1	1	1	0	0
女21～30	1	1	0	0	0
女31～40	1	1	0	0	0
女41以上	1	0	0	0	1

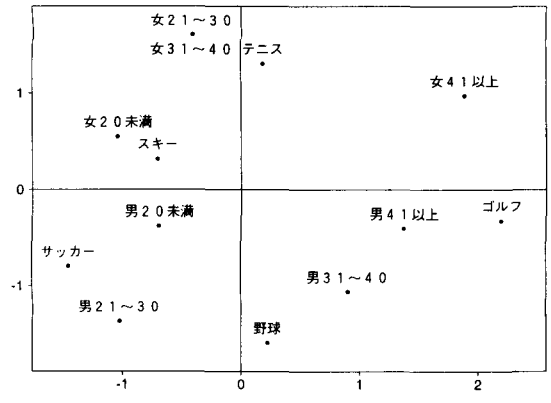

図1　数量化3類によるプロット

図1

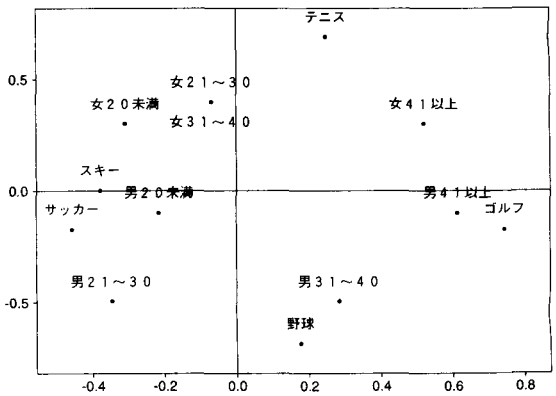

図2　特異値分解によるプロット

図2

参考文献

[1]　佐和隆光：「計量経済学の基礎」(1970)，東洋経済新報社.

[2]　奥野・久米・芳賀・吉澤「多変量解析法」(1971)，日科技連出版社.

[3]　上田太一郎，小林整功(1994)：“ソフトウェア業界の特質について－判別分析変数選択規準とその適用例”，情報処理学会第49回(平成6年後期)全国大会予稿集.

[4]　菅　民郎「初心者がらくらく読める多変量解析の実践上下」(1993)，現代数学社.

[5]　東大教養学部統計学教室編「自然科学の統計学」(1992)，東大出版会.

[6]　鷲尾・大橋「多次元データの解析」(1989)，岩波書店.

[7]　西里静彦「質的データの数量化」(1982)，朝倉書店.

[8]　渡部他監修「FORTRAN 77による数値計算ソフトウェア」，丸善.

会合記録

6月1日（木）	名簿刊行委員会	5名
6月3日（土）	機関誌編集委員会	13名
6月9日（金）	企業サロン企画委員会	4名
6月28日（水）	表彰委員会	6名
6月29日（木）	ＯＡ化委員会	3名