

昭和43年8月13日第3種郵便物認可 昭和50年12月10日国氏首都特別選定第8503号
第20巻第3号通巻237号 毎月1回1日発行 昭和62年3月1日発行 現代数学社発行
1554 9366 0419

■自然・社会科学を学ぶための数学誌

数学

1987
3月号

特集 行列と変換の問題

行列の計算をしてみよう——石川 洋 文

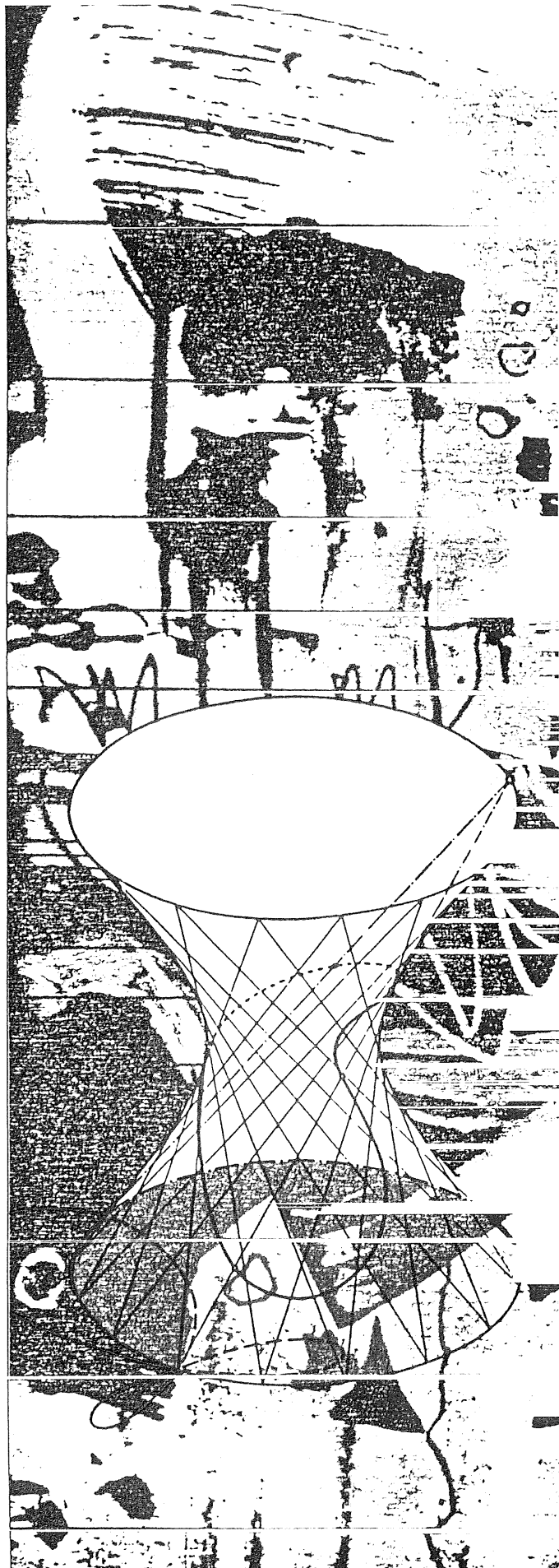
ダメ解答からみたベクトル空間——矢ヶ部 巖

ユニタリ・エルミートのエッセンス——大塚 道明

点の移動から1次変換へ——竹内 康滋

行列の世界での「相似」と「対等」——山下 純一

現代数学社



多変量解析の概論

柳井晴夫 (大学入試センター研究部)

高木廣文 (聖路加看護大学)



多変量解析とは

多変量解析 Multivariate Analysis とは、いくつかの個体が、複数個の変量によって特徴づけられる多変量データを、変量相互の相関関係を考慮しながら分析する手法である。より詳しくいえば、多変量解析とは事象そのもの、または、その事象の背後にあると想定される要因の多元的測定から

- (1) 事象の簡潔な記述と情報の圧縮 (次元の縮小),
- (2) 事象の背後にある潜在因子の探索 (次元の意味づけ),
- (3) 事象に対する複雑に絡み合った多くの要因の影響を総合化 (変数の重みづけ),
- (4) 未知のデータの判別と分類,

を目的とする一連の統計的手法の総称で、1960年以降、法律学、政治学、心理学、教育学、農学、医学、生物学など、いわゆる行動科学 Behavioral Sciences とよばれる幅広い分野における統計的データ解析 Statistical Data Analysis の手法として脚光を浴びようになってきた。特に、1970年代後半における、SPSS、BMDP、SAS といった、汎用統計プログラムパッケージの普及、およびここ数年来のパーソナルコンピュータによる多変量解析プログラムパッケージの普及は、複雑な計算過程が必要とされる多変量解析の各種技法を、もはや、特殊なものではなく、一変量の統計的仮説検定の技法と同様に、きわめて身近なものに変貌させたといって過言ではない。図1に、1971年から1985年の間に、我が国と、アメリカで医学、公衆衛生学、疫学 (Epidemiology) の分野に於ける研究論文が多数掲載されている代表的雑誌である日本公衆衛生雑誌と American Journal of Epidemiology に、少なくとも一つの多変量解析の技法がデータ解析の手法として用いられた割合を示した。

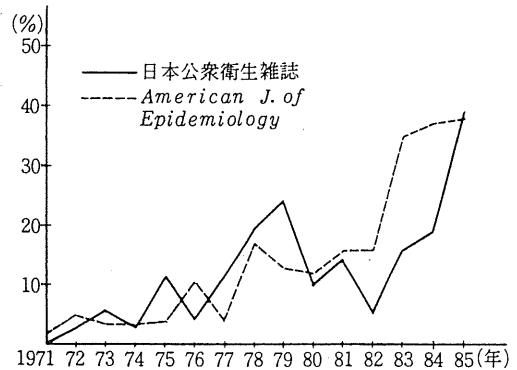


図1 最近15年における「日本公衆衛生雑誌」と「American J. of Epidemiology」における多変量解析の技法を用いた論文の割合

合を示した。多少の振動はあるにせよ、いずれの雑誌においても、多変量解析の手法が用いられた割合は近年加速度的に増加していることが読み取れ、ここ数年のうちに両誌とも二つに一つの論文には多変量解析の手法が用いられるようになることが推測される。一方、詳細は省略するが、文科系の専門分野で多変量解析の手法が比較的多く利用されている学会誌である、心理学研究、教育心理学研究においては、1960年代から多変量解析を用いた論文が散見されたが、1980年以降では、約25%の論文に多変量解析の手法が適用されるようになってきている。このように、多変量解析の手法は、各種領域におけるデータ解析の主要な手法として定着しつつあるといえよう。

多変量データについて

多変量データは、一般に、表1のように、縦行に個体、横列に変量を割りあてた二次元データとして

表1: p 個の変量, n 個のデータからなる多変量データ

変量 個体	1	2	3	...	j	...	p
1	x_{11}	x_{21}	x_{31}	...	x_{j1}	...	x_{p1}
2	x_{12}	x_{22}	x_{32}	...	x_{j2}	...	x_{p2}
3	x_{13}	x_{23}	x_{33}	...	x_{j3}	...	x_{p3}
...	...	...	...	...	...	...	...
i	x_{1i}	x_{2i}	x_{3i}	...	x_{ji}	...	x_{pi}
...	...	...	...	...	...	...	...
n	x_{1n}	x_{2n}	x_{3n}	...	x_{jn}	...	x_{pn}

表わすことができる。すなわち、全部で n 個の個体が p 個の変量で特徴づけられる場合、第 i 番目の個体の p 個のデータを、 $x_{1i}, x_{2i}, \dots, x_{pi}$ これらを i 番目の成分とする n 次元ベクトルを $x_1, x_2, \dots, x_p$ とし、それらを行ベクトルとした (n, p) 型行列

$$(1) X = (x_1, x_2, \dots, x_p)$$

を、多変量データとみなすことができる。この操作により、記述的な立場からみると多変量解析の多くの技法は $n(n > p)$ 次元線型空間にある種の線形結合 linear combination を求める課題に帰着されることになる。ここで、表2、表3、表4の三種の多変量データを概観しよう。表2は、労働省一般職業適性検査の6つの下位尺度に関する12人のデータを示したものである。 $X_1 \sim X_6$ は、テストの結果の得点で、間隔尺度をなす量的データとみなすことができる。なお、 Y は、学生の所属が「文科系」の場合

表2: 適性検査の6つの下位尺度の得点:

	X_1	X_2	X_3	X_4	X_5	X_6	Y
1:	53	16	24	29	46	15	1
2:	50	14	21	23	31	10	1
3:	50	17	18	24	40	15	1
4:	29	12	26	14	20	13	1
5:	29	12	26	14	20	13	2
6:	43	15	25	21	37	12	2
7:	56	19	24	25	43	14	2
8:	40	13	18	18	31	10	1
9:	46	17	22	23	35	11	2
10:	35	17	20	20	36	10	2
11:	44	12	18	23	42	10	1
12:	50	12	19	29	47	11	1

X_1 : 名詞比較 X_2 : 平面図判断 X_3 : 計算 X_4 : 語意
 X_5 : 文章完成 X_6 : 算数応用 Y : 文科系(1); 理科系(2)

表3: 世界12カ国の一日平均食物供給量 (1961~65年の平均年; 単位, cal) とがん死亡率 (1974年, 人口10万対)

	X_1	X_2	X_3	X_4	Y_1	Y_2
1:	1271	21	40	26	18.19	2.98
2:	670	55	575	23	12.87	14.60
3:	1217	68	168	20	50.80	3.92
4:	881	15	120	9	49.46	3.99
5:	976	19	205	18	10.90	6.75
6:	480	13	79	18	2.80	1.28
7:	575	56	98	13	23.93	3.51
8:	998	14	55	5	5.45	0.89
9:	840	49	178	77	8.46	13.76
10:	1313	10	144	6	10.05	1.94
11:	1112	17	140	24	16.02	4.12
12:	1109	14	151	44	17.89	8.53

X_1 : 穀類 X_2 : 野菜 X_3 : 肉類 X_4 : 魚類 Y_1 : 胃がん死亡率 Y_2 : 大腸がん死亡率

表4: 4つの項目に対するカテゴリー反応

	A	B	C	D
1:	3	4	2	5
2:	1	3	2	4
3:	1	2	2	4
4:	2	1	2	4
5:	2	1	2	1
6:	1	3	1	6
7:	1	3	2	4
8:	2	4	2	6
9:	2	1	2	4
10:	3	3	2	4
11:	4	4	3	2
12:	1	3	2	2

A: 支持政党 1: 自民 2: 社会 3: 公明 4: 共産
 B: 不支持政党 1: 自民 2: 公明 3: 共産 4: なし
 C: 生活満足度 1: 満足 2: やや満足 3: 不満
 D: 親の職業 1: 農業 2: 自営 3: サービス 4: 事務
 5: 経営 6: 専門職

1, 「理科系」の場合2としたもので、二値データ binary data と呼ばれる。

表3は、世界12ヶ国における一日一人当たりの平均食物供給量 ($X_1 \sim X_4$) (1961年~1965年の平均, 単位 cal) と、胃癌 (Y_1) および、大腸癌死亡率 (Y_2) を示したものである。これらも、表2の場合と同様に、量的データである。なお、このように地域単位のデータによって得られる相関は、表2の場合の個人相関と区別されて、地域相関または生態学的相関と呼ばれることがある。

ところで、表2のようなデータを各変量毎の平均値がゼロになるように基準化しておけば、変量 i と変量 j のピアソンの積率相関係数は、

$$(2) \quad r_{ij} = \frac{\sum_{k=1}^n (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^n (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^n (x_{jk} - \bar{x}_j)^2}}$$

$$= (\mathbf{x}_i, \mathbf{x}_j) / (\|\mathbf{x}_i\| \|\mathbf{x}_j\|) = \cos \theta_{ij}$$

のように、ベクトル $\mathbf{x}_i$ とベクトル $\mathbf{x}_j$ のなす角度 θ_{ij} の余弦として表わすことができる。この操作により、 p 個の変量に対応するベクトル $\mathbf{x}_1, \dots, \mathbf{x}_p$ は n 次元ユークリッド空間において最大 p 次元の線形部分空間を生成することになる。つづいて、表 4 のデータについて解説しよう。この表は、アイテムと呼ばれる 4 つの質問項目（それぞれの項目には、カテゴリと呼ばれる 4、4、3、6 個ずつの選択肢がある）に対する各被験者の回答をカテゴリ番号として示したののである。これらは名義尺度と呼ばれるもので、明らかに数値の順序には意味がない。また、加減乗除の演算が不可能である。そこで、表 4 の第 1、第 2 列から、

$$(3) \quad G_A = \begin{matrix} & \text{アイテム A} \\ & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{matrix}$$

$$G_B = \begin{matrix} & \text{アイテム B} \\ & \begin{matrix} 1 & 2 & 3 & 4 \end{matrix} \\ \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} & \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{matrix}$$

というように、成分が 1 と 0 の要素からなる行列を構成する。このような行列は、ダミー変数行列と呼ばれるもので、各列における 1 の数の合計が、それぞれのカテゴリへの反応の合計数、さらに、二つの行列の積は、次のようになる。

$$(4) \quad (G_A)' G_B = \begin{bmatrix} 0 & 1 & 3 & 0 \\ 3 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

上記は二つの項目の各カテゴリに関するクロス集計表を与える。

外的基準について

表 2 における食物供給量と癌死亡率のように、ある事象とその要因に関する変数が二組の変数群に分けられる場合、一方の変数と他方の変数を関連させて分析することが必要となる。ここで、2 つの変数を X および Y と表わそう。このとき、 X の変動が原因となって、その結果 Y の変動が生じた場合、通常、 X は独立変数 independent variable、 Y は従属変数 dependent variable、または、 X は説明変数 explanatory variable、 Y は基準変数 criterion variable と呼ばれる。 X と Y との間に厳密な意味での因果関係が成立しなくても、 X の変動に基づいて、 Y の変動が説明できる場合、やはり、 X を説明変数、 Y を基準変数と呼んでもよい。基準変数は、説明変数 X の測定目的となる変数で、外的基準と呼ばれることがある。最近激増している癌や脳卒中、心筋梗塞などの非感染性の疾患の原因のひとつに、生活習慣、食習慣などが関与しているものと想定して、相互の関連を分析する場合、表 3 の例における、食物供給量 X_1, X_2, X_3, X_4 が説明変数、癌死亡率 Y_1, Y_2 が外的基準となる。また、表 2 における適性検査 ($X_1 \sim X_6$) を高校時代に受け、その結果に基づいて、理科系、文科系の進路を判定するものとすれば、 Y の二値変数は、外的基準である。

また、入学試験の学業成績をより高い精度で予測できる入学試験を作成するために、ある大学の入学者について、入学試験の成績と入学後の成績の関連を分析する場合、入学試験の成績が説明変数で、これに対する外的基準は、入学後の成績である。また、患者の医療検査の成績によってある種の疾病の診断をする場合、その疾病の重症度（または、疾病の有無）が外的基準となる。

上記の場合のように、二組の変数群が与えられている場合の多変量解析の手法は外的基準のある場合の手法に分類されるが、事象またはその説明要因のいずれか、すなわち、基準変数群、または説明変数

群の一方のデータしか与えられていない場合、それぞれの変数群に含まれる変数間の相互関連の強さを分析する必要がある。この種の変数解析の手法は、外的基準のない場合の手法に分類されるもので、分析以前には不明確であった事象の内容構造をあばき出す役割を負わされているものといえよう。次に、上記の二つの場合のそれぞれに含まれる多変数解析の手法について解説しよう。

外的基準のある場合の手法

ある基準変数 Y が p 個の説明変数 $X_1, \dots, X_p$ によってほぼ説明される場合、通常

$$(5) \quad Y = \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon$$

という関係式を想定する。これは Y の変動が X_j ($j=1, \dots, p$) の一次式によってほぼ説明されるもので、線形モデル linear model と呼ばれ、このとき、 Y 、 X_j の第 k 番目の、個体に関するデータを、 y_k, x_{jk} とするとき、 $\beta_1, \beta_2, \dots, \beta_p$ の推定値 $b_1, b_2, \dots, b_p$ は、

$$(6) \quad \sum_{k=1}^n (y_k - \sum_{j=1}^p b_j x_{jk})^2$$

を最小にする解として求められる。これは、次号で解説される重回帰分析 multiple regression analysis の基本モデルであるが、説明変数 $X_1, \dots, X_p$ 間にある種の構造モデルが仮定される場合には、仮定的に設定された因果関係の分析を定量的に行うことを意図した手法であるパス解析 path analysis が有用である。

基準変数 Y は必ずしも $X_1, \dots, X_p$ の線型モデルによってのみ説明される必要はない。事象 Y が、適当な非線形関数 $f(X_1, \dots, X_p)$ を用いて、よりよく説明できるならば、

$$(7) \quad Y = f(X_1, \dots, X_p) + \varepsilon$$

といった非線形モデルを構築することができる。その一つの例として

$$(8) \quad f(X_1, \dots, X_p) = 1/[1 + \exp(-\sum_j \beta_j X_j)]$$

によって表される関数を想定する多重ロジスティックモデルがある。しかし、非線形モデルは対数変換、ロジット変換、またはテラー展開などにより、局部的に、線形モデルに変換することができる。この他、一組の説明変数 $(X_1, X_2, \dots, X_p)$ に対して、基準変数 $(Y_1, Y_2, \dots, Y_q)$ が二つ以上存在する場合の多変数解析の手法に正準相関分析 canonical correlation analysis がある。

基準変数 Y が、連続量ではなく、職業、病名などのように、名義尺度の分類基準で表わされる場合、 p 個の説明変数 $X_1, \dots, X_p$ に関する線形結合

$$(9) \quad f = a_1 X_1 + a_2 X_2 + \dots + a_p X_p$$

のグループ別平均値を $\bar{f}_j$ としたとき、グループ間の平均値の差の平方和

$$(10) \quad \sum_i \sum_j (\bar{f}_i - \bar{f}_j)^2$$

をある制約条件のもとで、最大にするような、重み $(a_1, a_2, \dots, a_p)$ を求める方法は、判別分析法 discriminant analysis と呼ばれる。判別分析法は、これまで、人類学における、人骨の多元的計測にもとづく種族の判定、医学における計量診断、心理学における適性診断の方法等に適用されてきた。

説明変数が潜在変数で与えられている場合

多変数解析の多くの技法においては、上記の線形モデルが採用されているが、その特殊な場合として、説明変数 X (または基準変数 Y) を

$$(11) \quad X_j = a_{j1} f_1 + a_{j2} f_2 + \dots + a_{jr} f_r + \varepsilon_j$$

と分解する双線形モデル bilinear model がある。上記のモデルにおいて f_i ($i=1, \dots, r$) は直接に観測されない潜在変数 latent variable を表わす。潜在変数があるらかの方法で推定されれば(11)式も推定された変数 f_i に関して線形モデル (5)式)となる。なお、(11)式のモデルは知能の潜在因子、性格の多面的特性を多元的に分析する手法として開発された因子分析の基本モデルとみなすことができる。因子分析モデルを、さらに一般化した形のモデルとして LISREL (Linear Structural Relationship) がある。因子分析法は、応用的な観点からいえば、次節で解説する主成分分析法と類似した手法である。

外的基準のない場合の手法

多変数解析の役割のひとつとして、事象の簡潔な記述と情報の圧縮ということを述べた。その代表的な手法が、主成分分析と呼ばれるもので、与えられた変数のもつ情報をなるべく少数の合成変数に縮約する手法である。線形数学の術語を用いれば、主成分分析とは、 p 個の変量ベクトル $x_1, x_2, \dots, x_p$ をなるべく低次元の空間に収める手法で、次式を満たす相関係数行列 R_{xx} の固有値 λ 、固有ベクトル t_j を求

	X2・平面図判断	X3・計算	X4・語意	X5・文章完成	X6・算数応用
X1・名詞比較	0.519	-0.307	0.904	0.825	0.234
	X2・平面図判断	X3・計算	X4・語意	X5・文章完成	X6・算数応用
	0.034	-0.360	0.934	0.114	

めることに帰着される。

$$(12) R_{xx}t_j = \lambda_j t_j \quad (j=1, \dots, p)$$

(例) 表2のX1~X6の6つの変数についての相関係数行列を求めると上記ようになる。

また、その正規化された固有ベクトル $\|t_j\|=1$ に固有値の平方根を掛けて得られる主成分負荷量ベクトル $a_j = \sqrt{\lambda_j} t_j$ を、第1成分と第2成分に関して図2に示したが、この図から明らかに、平面図判断、計算力、算数応用力と、名詞比較、語彙、文章完成力は異なった能力を測定していることが読みとれる。

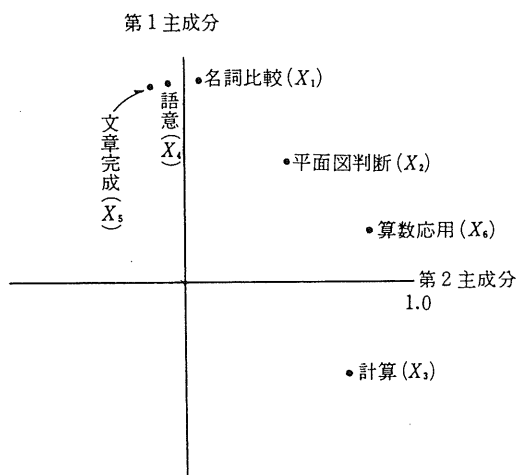


図2 表2のX1~X6の適性検査の下位尺度の主成分分析

上記の例から明らかなように、主成分分析は、全体の変数のもつ変動をなるべく少数の次元の空間で説明しようとするもので上記の例の場合、二次元によって、約81%の情報が説明されたことになる。

上記の主成分分析の手法は、多変量データから、相関係数行列を求め、その固有値、固有ベクトルを求め、それぞれの変量を多次元空間に位置づけるものであった。この他の多変量解析の手法として、多

変量データでなく、変数間（または個体間）の類似性（または非類似性）の指標を分析の対象とする方法がある。そのひとつは、クラスター分析と呼ばれるもので、個体間（変数間）の類似性（近さ）が何らかの方法で測られる場合、これらの個体（変数）をいくつかの個体（変数）のまとまり（クラスター）に分割する手法、より広義の意味では多くの個体間（変数間）の距離を定量的に測定して、それらをいくつかのクラスターに分割する数値分類の手法といってよい。クラスター分析はもともと生物学の分野で発展し、植物の系統分類の客観的手法として脚光をあびるようになったものである。なお、表2の6尺度のデータのクラスター分析の結果を図3に示した。

〈クラスター分析——基準：ウォード法〉

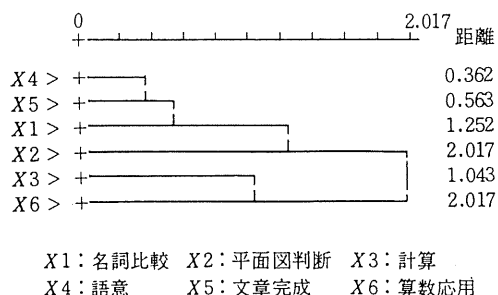


図3 表2の6尺度のクラスター分析

もうひとつの手法は、直接的に変数間（または個体間）の類似性（または非類似性）の指標が与えられている場合、それらの変数ベクトル（または個体）に対応するベクトルが含まれている最小次元の空間を見出す方法で、多次元尺度構成法 (MDS, Multi-Dimensional Scaling) とよばれるものである。MDSの手法は、個体間の距離を間隔尺度（または比尺度）の変数としてそのまま取り扱う計量的多次元尺度構成法 (Metric MDS)、さらに個体間の距離を順序尺度または名義尺度の質的なデータの情報とみなして取る非計量的多次元尺度構成法 (Non-metric MDS) がある。

質的データの解析法

表4のような、質的データ、つまり名義尺度からなるアイテム・カテゴリ型データ、の解析法としては、林知己夫による数量化理論が著名で、質的データが頻繁に現れる法律学、社会学、医学等の分野で、広く適用されてきた。なお、数量化理論は、第一類、第二類、第三類、第四類、と4つの手法に分類されるもので、第一類、第二類の、数学的原理は、説明変数としてアイテム・カテゴリ型データを用いた重回帰分析、判別分析に対応する。第三類は同様に、(3)式のようなアイテムカテゴリカルデータ、(4)式のようなクロス集計表を用いた主成分分析とみなすこともできるが、その端緒に、フィッシャーやガットマンによる尺度解析の方法があり、その後、西里静彦の双対尺度法(dual scaling)；フランスのデータ解析学派による、Correspondence analysis等、さまざまな形式に拡張されている。

以上、概説した多変量解析の手法の相互関連を示したものが、図4である。

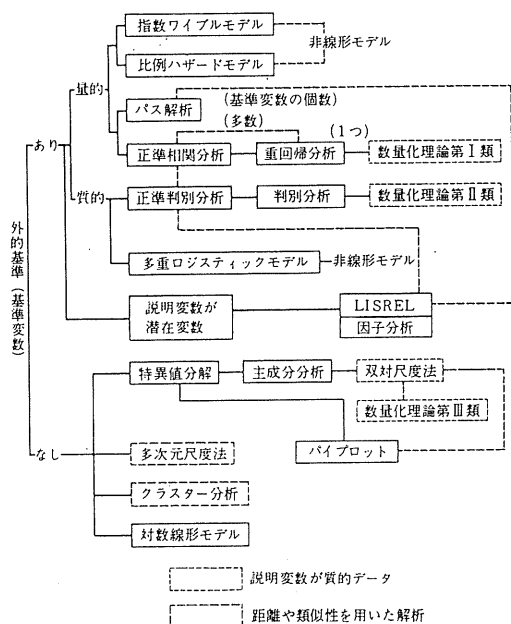


図4 多変量解析の分類〔柳井・高木(1986)の図1.1を1部改変したもの〕

次号から、6回にわたって、10人の執筆者によって以下のような多変量解析の各種技法を解説する。

4月号 重回帰分析法

5月号 判別分析法
6月号 主成分分析法・数量化三類
7月号 因子分析法
8月号 非線形モデル
9月号 多次元尺度法・クラスター分析

なお、上記の各種技法については、筆者らの著書(文献参照)に基づいて構成された多変量解析プログラムパッケージ“HALBAU”によって、解析した計算例を提示しながら、できる限り平易に解説するつもりである。(なお、本稿の図2、3の結果も、“HALBAU”によるものである)

文献

柳井晴夫・高木廣文編 多変量解析ハンドブック
現代数学社、1986

(やない はるお/たかぎ ひろふみ)

大学新入生のための数学

●線型代数・微積分のネックをさぐる●

御園生 善尚著

A5判・定価2,300円

大学へ進学した学生諸君は、数学を学んでいるうちに、その内容が高等学校で学んだ数学とは異質な面を多分にもっていることに気がつくであろう。そして、多くの学生は学習に関して戸惑い、学業からの脱落という運命をたどる者も皆無とはいえない。そのような戸惑いの緩和を意図して著わしたのが本書である。大学新入生にとって、与えられた数学のテキストの内容を理解しようとするとき、単なる計算技術だけではまず、その思想の把握が容易でない。高等学校で自明としてきた命題等を、自明でないとして掘り下げられたりすると、そのようなことを学ぶ必要性に疑問をもったりする。多くのテキストはテキストという性格から、そのような疑問に対する解答を用意していない。本書はこのような高校と大学の断層に苦しむ新入生が乗り越えねばならないネックポイントを選び、わかりやすく、ていねいに解説する。

〈内容〉 1. 高校数学から大学数学へ 2. 文学的記述と数学的記述 3. 1変数から2変数へ 4. 一様連続性 5. 一様収束性 6. 整数数 7. 陰関数とその存在定理 8. 逆関数と関数行列式 9. 重積分の変数変換 10. 行列式展開 11. 線型空間 12. 1次独立、1次従属 13. 内積空間 14. 行列とその固有値 15. 行列の対角化 16. 広義積分 17. 曲線とその特異点 18. 順序交換について

現代数学社