多次元データの視覚的表現について

妻　鳥　敏　彦

統計学では，いわゆる「データの整理」とか「データの加工」という言葉がよく使われます。これらは母集団（データの源泉）に関する情報を得るために重要な作業です。データの特徴を把握する方法には，おおまかにいって①計算（Calculation）②図示（graph）の2通りが考えられます。①はデータを幾つかの代表値に凝縮（又は変換）して，その代表値でデータの全体的様相を把握することになります。平均値，最頻値，中央値，あるいは範囲（レインジ），偏差（平均偏差，標準偏差，四分位偏差）などが代表値として，よく使われます。②はデータを平面上に図示して，その構造や変化の様子などを視覚的に（あるいはイメージとして）把握しようというのです。グラフ化の効用については，法則性の発見や現象の解明等に関する諸科学の歴史を想起すれば，ここで改めて言及する必要はないでしょう。①，②はそれぞれにちがった表現方法で利点や欠点がありますが，データ自身の内包している実体をあるがままに表現する方法であることが肝要です。さて，ここでは，②について筆者の体験に基づいて，面白い方法を紹介しよう。

（i）　星座グラフ

まず表1のようなデータが与えられているとする。これを適当な変換で0度から180度までの角度に変換する。

例えば　$\max_{1\leq j\leq k}\{x_{1j},x_{2j},\cdots\cdots,x_{nj}\}=z_u$

$\min_{1\leq j\leq k}\{x_{1j},x_{2j},\cdots\cdots x_{nj}\}=z_e$

とすると

$$\alpha_{ij}=\frac{x_{ij}-z_e}{z_u-z_e}\times 180^\circ$$

$$(i=1,2,\cdots\cdots,n)$$

の形で表Ⅰのデータは，表Ⅱに変換される。（このような具体的な意味づけ

＜表Ⅰ＞

変数 / 個数	1	2	3 ……	k
1	x_{11}	x_{12}	x_{13} ……	x_{1k}
2	x_{21}	x_{22}	x_{23} ……	x_{2k}
3	x_{31}	x_{32}	x_{33} ……	x_{3k}
⋮	⋮			
n	x_{n1}	x_{n2}	x_{n3} ……	x_{nk}

が行い易いような変換がよい。)

この<表Ⅱ>のデータに対して

$$\begin{cases} u_i = \sum_{j=1}^{k} w_j \cos \alpha_{ij} \\ v_i = \sum_{j=1}^{k} w_j \sin \alpha_{ij} \end{cases} \qquad (i=1,2,\cdots\cdots,n)$$

なる変換を考える

<表Ⅱ>

個数＼変数	1	2	3 ……	k
1	α_{11}	α_{12}	α_{13} ……	α_{1k}
2	α_{21}	α_{22}	α_{22} ……	α_{2k}
⋮				
n	α_{n1}	α_{n2}	α_{n3} ……	α_{nk}

但し w_j は，各変数の重みと考え $\sum_{j=1}^{k} w_j = 1$ （$w_j \geqq 0$）を仮定する。

この変換により（u_i, v_j）（$i=1,\cdots\cdots,n$）の u 個の点を半円の中にプロット（特に星印をつける）したものを星座グラフと呼ぶ。

さて，実際の作図法を具体例で示そう。データとして，5段階評価された学業成績を用いてみよう。

<表1>

生徒＼教科	英，	数，	国，	理，	社
1（A君）	4	5	4	5	5
2（B君）	3	2	3	1	4
3（C君）	5	4	3	4	3

まず最初の変換式は

$$\alpha_{ij} = \frac{x_{ij}-1}{5-1} \times 180° = (x_{ij}-1) \times 45°$$

これから<表Ⅱ>が作られる。

次に各教科間の重みを，同等とみなして $w_1 = w_2 = \cdots\cdots = w_5$ とおくと $\sum_{j=1}^{5} w_j = 1$ ，

$$w_1 = w_2 = \cdots\cdots = w_5 = \frac{1}{5}$$

となるから，変換

$$u_i = \frac{1}{5}\sum_{j=1}^{5} \cos \alpha_{ij}, \quad v_i = \frac{1}{5}\sum_{j=1}^{5} \sin \alpha_{ij}$$

によって点 $p(u_j, v_j)$ が半円内に与えられることになる。実際A君の星座上の点 P_A は，

<表2>

生徒＼教科	英，	数，	国，	理，	社
A君	135°	180°	135°	180°	180°
B君	90°	45°	90°	0°	135°
C君	180°	135°	90°	135°	90°

$$u_1 = \frac{1}{5}\{\cos 135° + \cos 180° + \cos 135° + \cos 180° + \cos 180°\} = -\frac{3+\sqrt{2}}{5}$$

$$v_1 = \frac{1}{5}\{\sin 90° + \sin 45° + \sin 90° + \sin 0° + \sin 135°\} = \frac{2+\sqrt{2}}{5}$$

によって与えられる。これを半円内に具体的にプロットしやすいように，次の様な整理を行う。

	x 座標	y 座標
点 P_1	$\frac{1}{5}\cos 135°$	$\frac{1}{5}\sin 90°$
点 P_2	$\frac{1}{5}(\cos 135°+\cos 180°)$	$\frac{1}{5}(\sin 90°+\sin 45_0)$
点 P_3	$\frac{1}{5}(\cos 135°+\cos 180°+\cos 135°)$	$\frac{1}{5}(\sin 90+\sin 45_0+\sin 90°)$
点 P_4	$\frac{1}{5}(\cos 135°+\cos 180°+\cos 135°+\cos 180°)$	$\frac{1}{5}(\sin 90°+\sin 45°+\sin 90°+\sin\ 0°)$
点 P_5	$\frac{1}{5}(\cos 135°+\cos 180°+\cos 135°+\cos 180°+\cos 180°)$	$\frac{1}{5}(\sin 90°+\sin 45°+\sin 90°+\sin 0°+\sin 135°)$

プロット手順

手順1.　点P_1は原点から4の方向へ平行に$\frac{1}{5}$だけ進む。(ことに同じ)

手順2.　点P_2は点P_1から5の方向へ平行に$\frac{1}{5}$だけ進む。(ことに同じ)

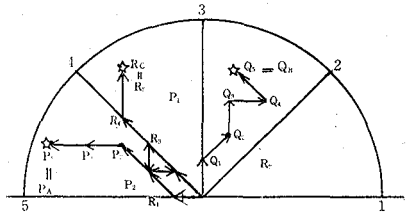
手順3.　点P_3は点P_2から4の方向へ平行に$\frac{1}{5}$だけ進む。(ことに同じ)

手順4.　点P_4は点P_3から5の方向へ平行に$\frac{1}{5}$だけ進む。(ことに同じ)

手順5.　点P_5は点p_4から5の方向へ平行に$\frac{1}{5}$だけ進む。(ことに同じ)

$P_A=P_5$であるから，ここで点P_Aがプロットできる。(下図参照)

もう1つB君の点R_BとC君の点R_Cをプロットしておこう。



星座における1－5の横線をx軸とし，3の縦線をy軸としたとき，点P_5の座標は$\left(-\frac{3+\sqrt{2}}{5},\ \frac{2+\sqrt{2}}{5}\right)$である。

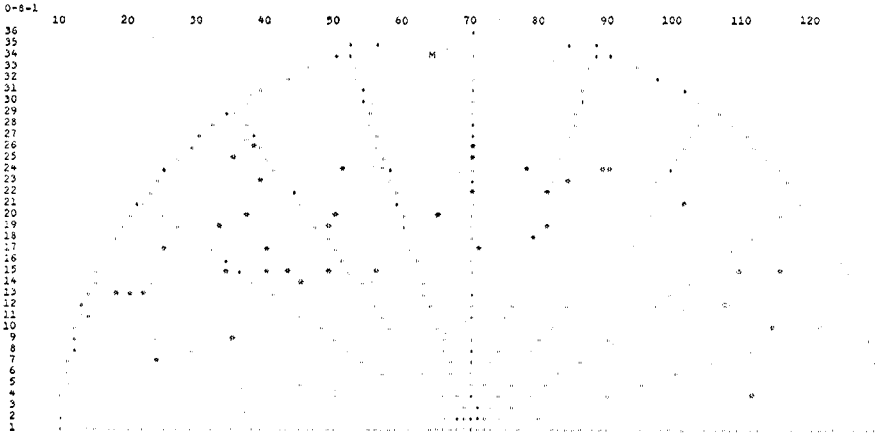
(但し，円の半径は長さ1とする。)

さて，このような星座へのプロットの効用は，5次元のデータを2次元の平面上にプロットして視覚的にも1定の区別がなされる点にあるが，更にプロットした点の位置が，円周上の5つのメモリに従って読むとき，丁度5教科の平均点を示し，円周上からのへだたり具合が，平均点のまわりのバラツキ具合を示唆していることである。又星印への道(path)についても着目することができる。我々の例でみると，A君の5教科の平均点は$\frac{4+5+4+5+5}{5}=4.6$，平均偏差値$(|4-4.6|+|5-4.6|+|4-4.6|+|5-4.6|+|5-4.6|)\frac{1}{5}=\frac{2.4}{5}=0.48$

B君の場合も同様にして平均点2.6，平均偏差値0.88であるから，A君は平

均点が高く5の軸に近くなっており偏差値（この場合，各教科の安定度とみなせよう）も小さいので円周上に近くなっている。B君の場合，平均は3以下で，偏差値も0.88とA君より大きいので，教科間にバラツキありと見受けられる。そのことはQ_Bの方がP_Aより円周上から，やや遠ざかっていることによって確かめられる。

教育現場では，学期の区切りで生徒の成績表が作成されるであろうが，それをこの星座グラフにプロットしてみるならば，各個人の平均点（全体的学力）やバラツキ（学力の不安定さ）が視覚的に把握できるだけでなく，個人のクラス全体における位置づけも一層よく分る筈である。更に毎学期これを行えば，個人及びクラスの学習成果の記録ともなりうるだろう。



もう1つこの考えが判別分析へ応用できることを示しておこう。

今，k次元データの2群が，それぞれ変換されて次のように与えられたとする。

＜1群＞

	1	2	……	k
1	α_{11}	α_{12}		α_{1k}
2	α_{21}	α_{22}	……	α_{2k}
⋮				
n	α_{n1}	α_{n2}	……	α_{nk}

＜2群＞

	1	2	……	k
1	β_{11}	β_{12}	……	β_{1k}
2	β_{21}	β_{22}	……	β_{2k}
⋮				
m	β_{m1}	β_{m2}	……	β_{m}

$P_i(u_i^{(1)}, v_i^{(1)})$に対して　　$Q_i(u_i^{(2)}, v_i^{(2)})$ に対して

$$\begin{cases} u_i^{(1)} = \sum_{j=1}^{k} w_j \cos \alpha_{ij} \\ v_i^{(1)} = \sum_{j} w_j \sin \alpha_{ij} \end{cases} \quad (i=1,\cdots,n) \qquad \begin{cases} u_i^{(2)} = \sum_{j} w_j \cos \beta_{ij} \\ v_i^{(2)} = \sum_{j} w_j \sin \beta_{ij} \end{cases} \quad (i=1,\cdots,m)$$

これら P_i, Q_i を星座の中にプロットしたとき，２つの群が星座の中で最もよく判別できるような点 $(w_1, w_2, \cdots w_k)$ を求める問題である。

換言すれば，k 次元超平面 $w_1+w_2+\cdots+w_k=1$ 上の部分平面 $s_k=\{(w_1, w_2, \cdots, w_k), w_i \geqq 0, i=1,\cdots,k\}$ 上の１つの点に対して１群，２群の k 次のデータ $(n+m)$ 個が星座上にプロットすることができ，そのような s_k の点の中で星座グラフの２群が最もよく判別できるような点 $(w_1,\cdots,w_k)$ を捜すことである。

具体例で説明すれば，ある大学の入学試験は k 教科であり，ある高校からは $(n+m)$ 人受験し n 人の合格者があったとする。この $(n+m)$ 人の在学中(高校) の成績を用いて，上記の点 $(w_1, w_2, \cdots, w_k)$ を求めたとすれば，これらの数値から，どの教科が合格に有効であるか，等が推察されるであろう。

(ii)　Faces method (H. Chernoff)

この方法は簡単に云えば，多次元データを人間の顔に似せて配列することによって作られる似顔の表情を通じて，２次元空間内に表現しようというのである。

従ってデータの特徴は似顔の表情に現れ，我々がその表情をみて，データの特徴をつかむことになる。表情のデリケートな変化や特徴把握に関しては，我々個々の人生を通じて，さまざまな表情に接してきており，その無意識に培われた経験が大いに役立つのである。早速その概略を述べてみよう。

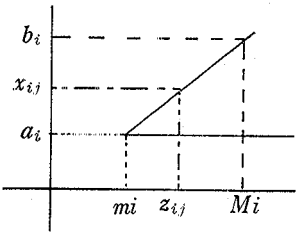
まず実際のデータを顔を構成する変数 $(x_1, x_2, \cdots x_{18})$ に当てはめる場合には各変数項目が，異なったレインジをとるので，次の様な基準化が必要である。

すなわちデータの最大，最小値を，基準化の最大，最小値へ変換する１次変換を求めると，それは，

$$x_{ij}=\frac{b_i-a_i}{M_i-n_i}(z_{ij}-m_i)+a_i$$

である。ここに

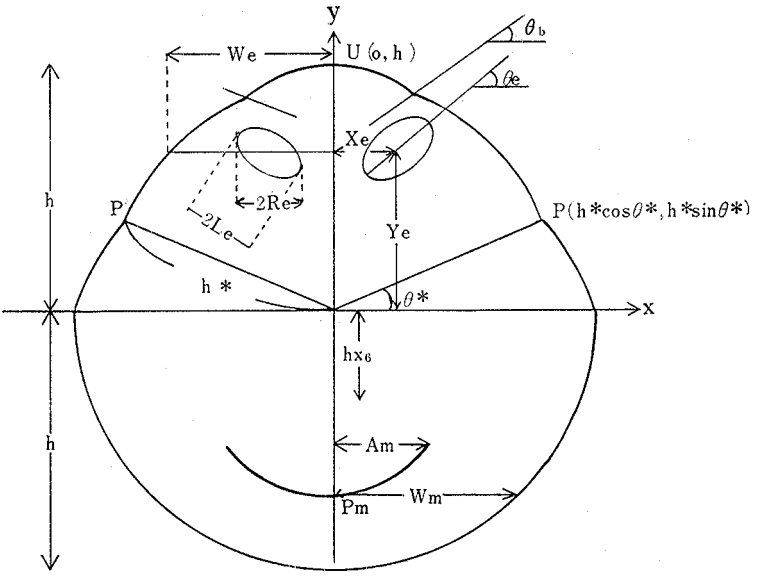
- $i=1,\cdots,18$（変数の個数）
- $j=1,\cdots n$（データ数）
- M_i, m_i：i番目の変数項目に対応する生のデータの最大値，最小値
- b_i, a_i：基準化されたi番目の変数のレインジの最大値，最小値



但し，実際にデータを用いる場合，規準化されたレインジが閉区間の場合，少し不都合が起こりうるので少し広めの開区間が望ましい。

さて，具体的な顔の構成は ①顔の輪郭 ②鼻 ③口 ④目 ⑤ひとみ ⑥眉の6つの部分からなる。

右図がおよその骨格であるので，以下の説明はこの図に基づくことにしよう。



① 顔の輪郭

顔の中心（原点0）と距離h^*とx軸との角度θ^*（ラジアン）で決められる点Pとy軸対称の点P^1によって顔の上部，下部に分ける。又，顔の

中心から上端，下端の距離をいずれも h とし，それぞれの点を U，L とする。上顔部 PUP' 下顔部 PLP' は長軸 a_i，短軸 b_i（$i=U,L$）の比が異なる楕円をあてはめる。

$$\begin{cases} h^*=\dfrac{1}{2}(1+x_1)\cdot H \quad (H:\text{顔の大きさの倍率}) \begin{cases} a_u/b_u=x_4 \\ \quad (\text{上顔部の楕円の比}) \end{cases} \\ \theta^*=(2,x_2-1)\dfrac{\pi}{4} \qquad\qquad \begin{cases} a_l/b_2=x_5 \\ \quad (\text{下顔部の楕円の比}) \end{cases} \\ h=\dfrac{1}{2}(1+x_3)\cdot H \qquad \begin{pmatrix} \text{特に } a/b=1 \text{ のとき円であり } a/b<1 \text{ なら} \\ \text{縦長の楕円になる } a/b>1 \text{ なら横長の楕} \\ \text{円である。} \end{pmatrix} \end{cases}$$

電算機のラインプリンターによる打ち出しの場合，横と縦の文字間隔が同じでないので注意を用する。例えば，上部楕円と下部楕円が P，P' でうまく交わらないようなことが起こる。この辺りは電算機マニアの腕のみせどころと思し召せ。

② 鼻

鼻は原点0を中心として y 軸に上下 h，x_6 の長さとする。

x_6 のレインジは（0，1）であるから，鼻の長さは（0，$2h$）内におさまることになっている。

③ 口

口は原点0から下 P_m の位置に半径 R の円弧を描く

$$P_m=-h\{x_7+(1-x_7)x_6\}$$

$$R=h/|x_8|$$

ここで $x_8=0$ のとき適当に指定した長さの線分を描き $x_8>0$ なら上向きの円弧（スマイル型）$x_8<0$ なら，下向きの円弧（アンゲリ型）を描くようにする。（注 $x_8\to 0$ ならば $R\to\infty$ となり半径の大きな円が想定され，ゆるやかなカーブの円弧になろう。〝直線は円の1部〟が理解される。）

特に R が顔中におさまるときは $A_m=R\times x_9$ とし，はみでるときは P_m から楕円までの距離 W_m を求めて $A_m=W_m\times x_9$ にする。

はみでるかどうかの判断は，口の終点の y 座標と下顔部楕円のそれを比較すればよい。しかし，現実には上の処置だけでは，はみ出ない保障はない。

④ 目

目は適当な大きさの楕円をある角度傾けて描く，楕円の中心の位置は次式で与えられる。

$$Y_e = h\{x_{10} + (1 - x_{10})x_6\}$$

$$X_e = W_2(1 + 2x_{11})/4 \qquad (\text{但し } W_e \text{ は } Y_e \text{ の高さでの顔までの距離})$$

(x_{10} や x_6 のレインジの関係から目の中心は鼻より下方にくることはない!!)

又，目の傾き θ_e，目の幅 L_e は $\theta_e = (2x_{12} - 1)\pi/5$，$L_e = x_{14}, x\min(X_e, W_e - X_e)$ 左目は y 座標はそのままで θ_e，X_e の符号を変える。

⑤ ひとみ

ひとみは目の楕円の中心から $H_e = R_e(2x_{15} - 1)$ だけ x 軸に平行に移動した位置にとる。(この位置に関しては左右対称にしないことが肝要，その理由は読者におまかせします。) $R_e = L_e\sqrt{\cos^2\theta_e + x_{13}^2\sin^2\theta_e}$ (計算略)

⑥ 眉

眉は目の中心から Y_e だけあがったところで傾き θ_b とする。

ここに $Y_b = 2(x_{16} + 0.3)L_e \cdot x_{13}$

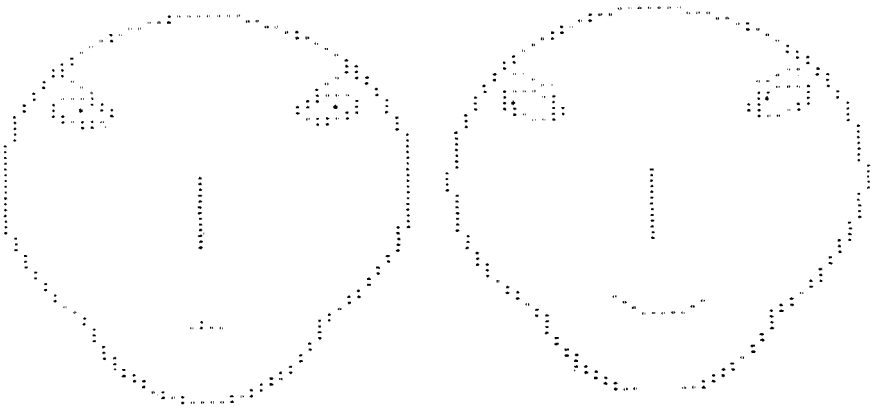
$$\theta_b = \vartheta_e + (2x_{17} - 1)\frac{\pi}{5}$$

また眉の幅 L_b は，$L_b = R_e\left(\frac{2x_{18}+1}{2}\right)$ で与えられ，特に $Y_b = x_{13} \cdot L_e$ の場合や x_{16} がある範囲内にきたときなど眉が目を "cut" するようだ。

以上で各部分の説明を終えますが，基準化された変数の（指定した）レインジと顔の構成に関する特性の対応を表1に整理しておきましょう。

範 囲	変 数	顔の構成に関する特性	範 囲	変 数	顔の構成に関する特性
(0, 1)	$x_1 \to h^*$	OPの長さ	(0, 1)	$x_{10} \to Y_e$	目の位置
(0. 1)	$x_2 \to \theta^*$	X軸とOPの角度	(0, 1)	$x_{11} \to X_e$	目の中心の離れ具合
(0, 1)	$x_3 \to h$	OU(=OL) の長さ	(0, 1)	$x_{12} \to \theta_e$	目の傾き
(0.5,2)	x^4	顔の上半分のa/b	(0.4,0.8)	x_{13}	目の楕円のa/b
(0.5,2)	x_5	顔の下半分のa/b	(0, 1)	$x_{14} \to L_e$	目の幅の半分
(0, 1)	x_6	鼻の長さ	(0, 1)	x_{15}	ひとみの位置
(0, 1)	$x_7 \to P_m$	口の位置	(0, 1)	$x_{16} \to Y_e$	目から眉の位置
(-5, 5)	x_8	口の曲率	(0, 1)	$x_{17} \to \theta_b - \theta$	眉の傾き
(0, 1)	$x_9 \to a_m$	口の幅	(0, 1)	x_{18}	眉の長さ

以上はC hernoff の Faces method に沿って説明したが，実際は変数の多さによっては耳や髪なども付加することが可能であろう。又，場合によっては変数を折半して2組の顔を描いて，それを1組とみなすことも考えられる。変数が18個以下の場合は，どれかの変数を固定して（つまり定数を与えておく）おけばよい。しかし，どの変数を固定するのがよいか，あるいは，どの変数を，どの部分に移すとよいか等々は，多くは体験に基づく問題であろう。この手法の有用性は，利用目的によって差異があるのはもちろんであるが，データの特徴が表情によって捉えられるばかりでなく，データをいくつかの群に分類する場合や多変量データの系列の変化を指摘する場合などに有効であろう。データの微妙な変化も案外捉えることができるかも知れない。筆者も県下（香川県）の土地分類基本調査のデータを使って，この表現方法を利用した体験がある。（変数の個数12，データ数約600）そして，視覚的分類は，統計的判別分析に十分耐えることを実感している。ただ多量のデータ群を処理するには，この方法は能率的でないので，この辺は今後の研究課題であろう。電算機で描いた表情。（資料2）



以上，データの表現についてグラフ化の観点から2つの方法を紹介致しました。云うまでもなくグラフ化の効用はデータの特徴を視覚的に把握したり，伝達したり，印象づけることにあります。又，データ解析に先立って解析の方法や方向にネライをつけるためにも利用できます。教育現場で発生するデータに

ついても色んな表現を工夫することによって「気付かなかった」ことや「忘れていた」ことどもに遭遇するにちがいありません。

参 考 文 献

1. Wakimoto, K and Taguri, M (1978)
Constellatwn graphical methods for representing multidimensional dota, Anu lvst Stct. Moth Vol. 30
2. Chernoff. H (1973)
The use of faces to represent points in k-dimevsionol space graphically. Jour Amer Statist.Assoc Vol. 63
3. 妻鳥敏彦
メッシュデータの一つの表現とその活用への提案（香川県土地分類基本調査総括報告書 1977）