統計数理研究所彙報
第30巻 第2号 (1983)

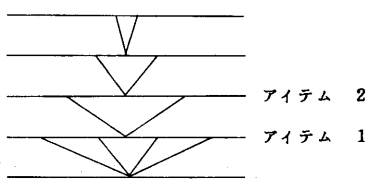
ベクトル変換を用いた数量化法

統計数理研究所 馬 場 康 維
岡山大学 脇 本 和 昌

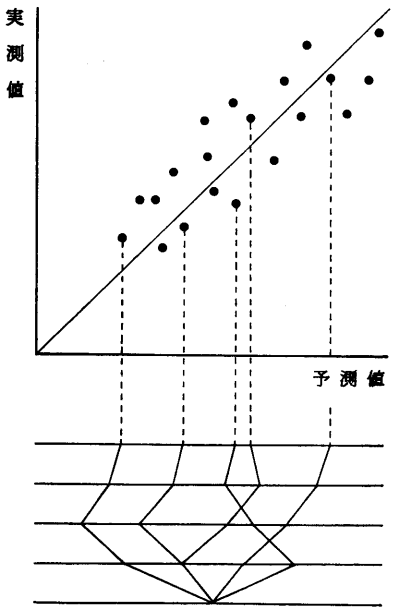
(1982年11月 受付)

1. 序

統計的データ解析において，データの視覚化はデータ構造の把握のための極めて有用な手段である．単なる数値を見るよりもグラフを見る方が，データの構造を直観的に理解しやすい．ヒストグラムや散布図などはその代表的な例であろう．最近多変量データのグラフ表現に関する多くの試みがなされている（たとえば脇本他 [1] を参照）．その一つに脇本，田栗 [2], [3] の星座グラフがある．この手法の本質は各変量の値をベクトルにおきかえそれを連結してグラフ化することにある．この方法をカテゴリカルデータに適用して従来からの数量化法と比較するのがこの論文の目的である．ここでは特にカテゴリー化したデータから目的の変量を予測するといった従来の数量化理論第Ⅰ類に対応する場合について考える．



第1図 アイテム・カテゴリーに与えられる数量に対応する線分



第2図 数量化の結果のグラフ表現

従来の数量化Ⅰ類の結果をグラフ化するとすれば以下のような方法が考えられる．アイテム・カテゴリーに与えられた数量によりレインジの大きい順にアイテムを並べ，アイテムに対応する平行な数直線を引く．カテゴリーに与えられる数量に対応する数直線上の点と一つ下の数直線の原点を結んでできる線分を各アイテム・カテゴリーに対応させる（第1図）．各個体（標本）のアイテム・カテゴリーに対する反応パターンに応じてこの線分を連結すればその特徴を表わす径路が構成できる（第2図）．予測値を求めることは径路の終点の与える数量を求めることにほかならない．ところで上記の方法では，二つのカテゴリー間の数量の差に線分間の角度が比例していない．またカテゴリーによって線分の長さが異る．したがって線分の長さや角度に意味を持たせにくいという欠点がある．こういう欠点を補う方法として

数直線のかわりに円周を用いるベクトル変換法を考える．これが本論文の主題であり次節でその方法を述べる．

2. ベクトル変換による数量化法

性別，学歴，職業などの質的な特性，あるいは20代，30代，…と区分された年齢のようにカテゴリー化された特性からそれらに関連した目的の変量を予測するという問題を考える．数量化理論にならい，特性項目をアイテム，その項目中の区分をカテゴリーと呼び，アイテム・カテゴリーに対する該当，非該当の反応パターンを表わす記号として δ 変数を用いる．数量化理論で求める数量と同じ役割を果すものはここではアイテム・カテゴリーに付与される角度である．

はじめに，ここで用いる記号をまとめておこう．なお以下では解析の対象となるものがなんであってもそれを「個体」ということにする．

N: 個体数（標本数）
R: アイテム数
K_j: j アイテムのカテゴリー数
α: 個体番号を表わす添字 $(\alpha = 1, \cdots, N)$
j: アイテムを表わす添字 $(j = 1, \cdots, R)$
k: カテゴリーを表わす添字 $(k = 1, \cdots, K_j)$
$\delta(jk)$: アイテム・カテゴリーへの反応を表わす変数．j アイテム k カテゴリーに該当するとき1，しないとき0をとる
y: 外的基準（目的変量）
θ: 数量化の結果 j アイテム k カテゴリーに与えられる角度

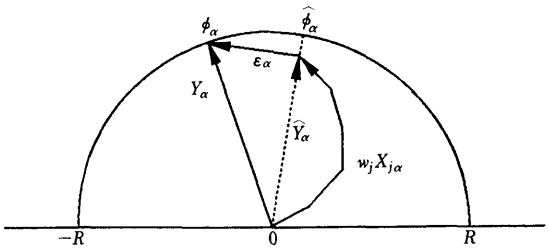
なお以下では添字 α のない変数はその変数を代表し添字 α の付いた変数は標本値を表わすものとする．

まず適当な変換

$$\phi = G(y) \qquad (0 \leqq \phi \leqq \pi) \tag{1}$$

により目的変量 y を角度に変換する（応用例参照）．

$$Y = R \exp(i\phi) = R(\cos\phi + i \sin\phi)$$



第3図 目的変量のベクトル表現およびアイテム・カテゴリーへの反応パターンの連結ベクトル表現

とすると，Y は長さがアイテム数に等しいベクトルである．ここに i は虚数単位である．個体 α には ϕ_α, Y_α が対応するものとする．次に

$$X_j = \exp\left\{ i \sum_{k=1}^{K_j} \theta_{jk} \delta(jk) \right\}$$

とする．X_j は個体が j アイテムの k カテゴリーに該当したとき θ_{jk} の方向を向く単位ベクトルである．この X_j の合成ベクトル

$$\hat{Y} = \sum_{j=1}^{R} w_j X_j \tag{2}$$

によって表わされる $\hat{Y}$ を Y の予測値（ベクトル）と定義する．$w_j X_j$ を連結してできる径路が個体のアイテム・カテゴリーへの反応パターンを表わしている（第3図）．

アイテムベクトルにかかる係数 w_j, アイテム・カテゴリーに与えられる角度 θ_{jk} は残差平方和が最小になるように決める．即ち残差（ベクトル）を

$$\varepsilon_\alpha = Y_\alpha - \hat{Y}_\alpha = R \exp(i\phi) - \sum_{j=1}^{R} w_j \exp\left\{ \sum_{k=1}^{K_j} \theta_{jk} \delta(jk) \right\}$$

としたとき

$$Q = \sum_{\alpha=1}^{N} |\varepsilon_\alpha|^2$$

を最小にするように決めるものとする．

y の予測値 $\hat{y}$ は合成ベクトルの方向から求める．即ち

$$\hat{Y} = |\hat{Y}| \exp(i\hat{\phi}) \tag{3}$$

としたとき

$$\hat{y} = G^{-1}(\hat{\phi}) \tag{4}$$

ここで G^{-1} は G の逆変換である．

Q を最小にする w_j, θ_{jk} を求めるアルゴリズムについて述べよう．

$$\xi_{j\alpha} = \sum_{k=1}^{K_j} \theta_{jk} \delta_\alpha(jk) - \phi_\alpha \tag{5}$$

とおくと $\xi_{j\alpha}$ は個体 α のアイテムベクトル $X_{j\alpha}$ と目的変量（ベクトル）Y_α とのなす角である．Q は

$$Q = \sum_{j=1}^{R} \sum_{j'=1}^{R} \sum_{\alpha=1}^{N} \left\{ (1 - w_j \cos\xi_{j\alpha})(1 - w_{j'} \cos\xi_{j'\alpha}) + w_j w_{j'} \sin\xi_{j\alpha} \sin\xi_{j'\alpha} \right\} \tag{6}$$

あるいは

$$Q = NR^2 - 2R \sum_{j=1}^{R} \sum_{\alpha=1}^{N} w_j \cos\xi_{j\alpha} + \sum_{j=1}^{R} \sum_{j'=1}^{R} \sum_{\alpha=1}^{N} w_j w_{j'} \cos(\xi_{j\alpha} - \xi_{j'\alpha})$$

と表わされる。したがって Q を最小にする問題は非線型の最適化の問題になる．

上記の Q を最小にする w_j, θ_{jk} を解析的に求めることは難かしい．数値的解法については現

在検討中である．非線型の最適化の問題については稿を改めることにし，ここでは特殊な場合を考え線型化した方程式によって近似解を求める方法について述べる．まず

$$w_j = 1 \quad (j = 1, \cdots, R)$$

とおくことにする．これは径路を構成するベクトルの長さを変えないことに対応する．このとき

(7)
$$Q = Q_c + Q_s$$
$$Q_c = \sum_{j=1}^{R}\sum_{j'=1}^{R}\sum_{\alpha=1}^{N}(1-\cos\xi_{j\alpha})(1-\cos\xi_{j'\alpha})$$
$$Q_s = \sum_{j=1}^{R}\sum_{j'=1}^{R}\sum_{\alpha=1}^{N}\sin\xi_{j\alpha}\sin\xi_{j'\alpha}$$

である．ここで予測が十分に良くアイテム・カテゴリーに適正な角度が与えられている場合を考えよう．この場合には $\xi_{j\alpha}$ が十分小さいと考えられる．Q を $\xi_{j\alpha}$ で展開すると Q_c は $\xi_{j\alpha}$ の4次のオーダー，Q_s は2次のオーダーである．したがって

(8)
$$Q \simeq Q_s \simeq \sum_{j=1}^{R}\sum_{j'=1}^{R}\sum_{\alpha=1}^{N}\xi_{j\alpha}\xi_{j'\alpha}$$

を得る．(5) を (8) の右辺に代入して θ_{jk} で微分することにより θ_{jk} に対する連立方程式

(9)
$$\sum_{j'=1}^{R}\sum_{k'=1}^{K_{j'}} n_{jk,\ j'k'}\theta_{j'k'} = R\sum_{\alpha=1}^{N}\phi_\alpha\delta_\alpha(jk)$$
$$(j = 1, \cdots, R;\ k = 1, \cdots, K_j)$$

を得る．ここで

$$n_{jk,\ j'k'} = \sum_{\alpha=1}^{N}\delta_\alpha(jk)\,\delta_\alpha(j'k')$$

である．(9) 式は右辺に R がかかっていることを除いて数量化 I 類と同じ方程式（駒澤 [4] の15頁参照）になっている．

3. 応　用　例

表1は血圧の測定の際に5つの項目について調べたものである．年齢はカテゴリー化してある．調査の際のカテゴリーはもう少し細分化されているが標本数が少いことからカテゴリーを再編したものが表に示したものである．例えば，塩からいものを食べるかどうかは調査の際は4段階である．この例をもとにグラフ化の方法を説明しよう．調査の項目に対する答のパターンから最大血圧値を推測するものとする．

まず始めに目的変量を角度に変換する．ここでは次式を用いた．

(10)
$$\phi = a + b(y - \bar{y})$$
$$a = \pi/2$$
$$b = \pi/180$$

ここで $\bar{y}$ は y の平均である．即ち血圧の単位1が角度1° に対応し平均が 90° になるように変

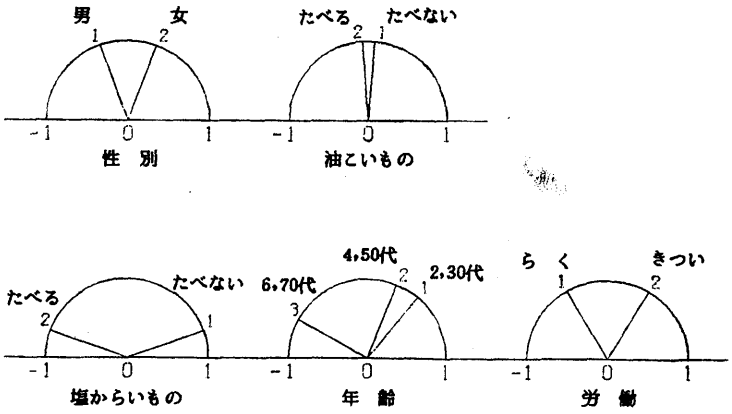
表1　アイテム・カテゴリーへの反応パターンおよび目的変量の予測値と実測値

アイテム	性		年齢			油こいもの		塩からいもの		労働		最大血圧値 (mmHg)		
カテゴリー / 個体	男	女	20～39	40～59	60～79	食べない	食べる	食べない	食べる	らく	きつい	実測値	予[a]測値	予[b]測値
1	V		V				V	V			V	124	122.8	122.7
2		V		V			V		V		V	154	142.9	146.1
3		V		V		V		V		V		130	129.0	128.8
4	V			V			V		V		V	168	152.8	154.2
5		V	V			V		V		V		134	124.9	125.0
6	V				V	V			V	V		180	181.0	180.9
7	V		V				V	V			V	114	122.8	122.7
8		V			V		V		V	V		172	175.1	174.7
9		V		V			V	V		V		120	131.0	130.4
10	V				V	V			V	V		176	181.0	180.9
11		V		V		V			V		V	108	140.7	144.3
12	V				V	V		V		V		162	156.6	153.0
13		V	V			V		V			V	123	113.1	112.9
14	V				V		V		V	V		185	182.7	182.7
15	V		V			V		V			V	110	120.9	120.9
16		V		V		V			V		V	168	140.7	144.3
17		V			V		V		V	V		172	175.1	174.7
18	V		V				V	V			V	122	122.8	122.7

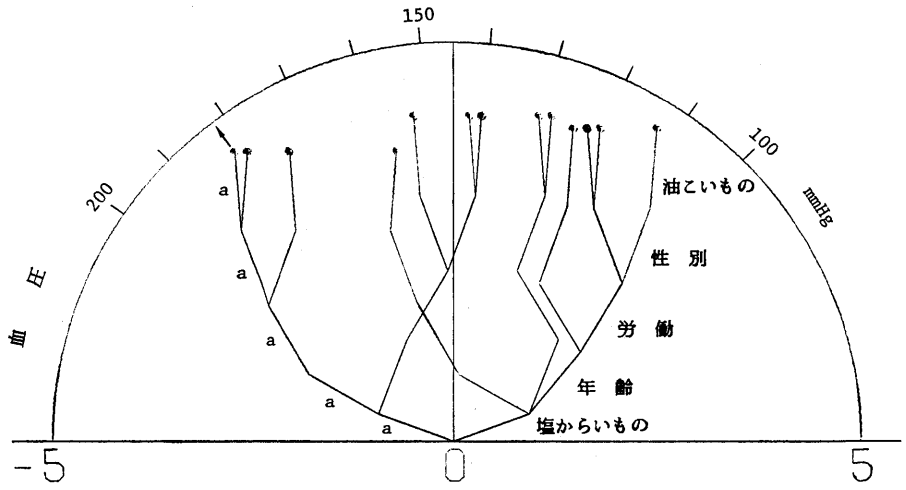
a　ベクトル変換による予測値．　b　数量化Ⅰ類による予測値．

換した．第5図の半円周上の目盛は (10) 式に基くものである．

ここでは (9) 式によってアイテム・カテゴリーに与える角度 θ_{jk} を求めた．この角度を第4図に示す．アイテムは角度のレインジが大きい順に並べかえてある．図上の線分が径路の各部分を構成する線分である．例えば（塩からいものを食べる，6, 70代，労働はらく，男，油こいものを食べる）という答のパターンは第5図のａという記号を付けた径路であらわされる．径



第4図　アイテム・カテゴリーに与えられる角度．アイテムはレインジの大きい順に左下から右上に並べてある．



第5図 数量化の結果のグラフ表現．円周の目盛は最大血圧値を表わす．矢印は径路aで表わされる反応パターンから予測される最大血圧の値．

路の終点の角度からこのパターンの血圧の推測値が求められる．このようにして18人のパターンを描いたものが第5図である．ただし同じパターンを持つ人がいるため異る径路は13本しかない．

第4図からは，塩からいものを食べるかどうかは血圧に影響を与えるが油こいものはあまり影響を及ぼさないということや，年齢の区分1，2はあまり差がないということなどが読みとれる．労働がらくなものの方が血圧が高いのはむしろ年齢との相関が高いためと考えられる．

この結果をもとに (2)，(3)，(4) 式を用いて求めた予測値と数量化Ⅰ類による予測値とを合わせて表1に示した．我々の方法による予測値と実測値との重相関係数は0.8843であった．それに対して数量化Ⅰ類による場合は0.8847であった．

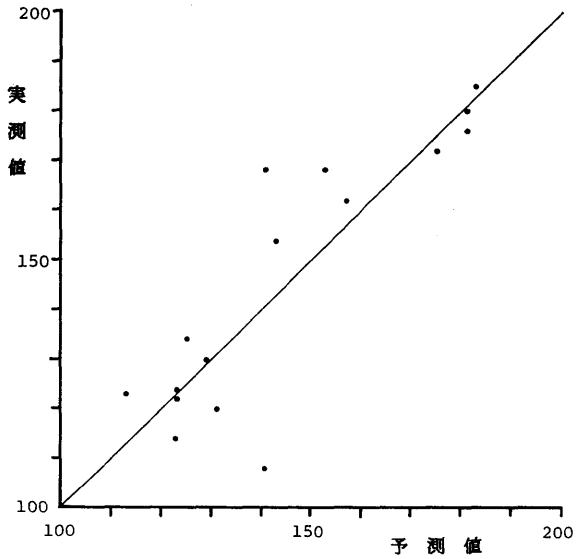
我々の方法による予測値と実測値との関係を第6図に示した．また第7図には我々の方法による予測値と数量化Ⅰ類によるものとの比較を示した．

4. 補　　遺

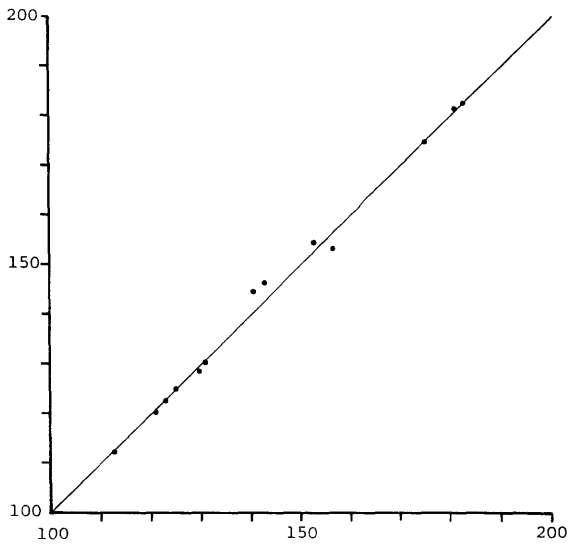
上記の応用例は近似方程式によるものである．(6) 式の最小化の際近似式を用いたのは従来の数量化Ⅰ類との関係をつけるためと，最小化が難かしいことの二つの理由からである．(6) 式の最小化については $w_j=1$ として求めた近似解を初期値として逐次的に解く方法が最適と考え現在試みている．

近似方程式 (9) は形式的には数量化Ⅰ類における方程式と同じものである．それにもかかわらず表1に示されるように我々の方法による予測値と数量化Ⅰ類による予測値とが異るのは次の理由による．(10) 式によって目的変量は角度に変換されているものとしよう．アイテム・カテゴリーに付与される数量も角度として与えられているものとすると我々の方法による予測値はアイテムに関する平均方向 (mean direction) である．一方数量化Ⅰ類による予測値はアイテムに関する算術平均である．一般に平均方向と算術平均とは等しくない（たとえば馬場 [5] 参照）．したがって我々の方法による予測値と数量化Ⅰ類による予測値とが異るのである．

ところで (10) の変換には任意性がある．上記のことから，特に b の選び方が結果に及ぼす影響は大きいと考えられるが，s を y の標準偏差として $b=(\pi/4)/s$，あるいは $b=(\pi/8)/s$ と



第6図 予測値と実測値



第7図 ベクトル変換による予測値と数量化Ⅰ類による予測値との比較．横軸はベクトル変換による予測値．縦軸は数量化Ⅰ類による予測値．

しても重相関係数はほとんど変わらない．

$w_j=1$ の場合に限って (7) 式の持つ意味を考えてみよう．Q_e を小さくするということはどんな意味を持つか．Q_e に含まれる項

$$\sum_{j=1}^{R}(1-\cos\xi_{j\alpha})$$

をアイテム数 R で割った式は directional data analysis における円分散 (circular variance) の形をしている ([5] 参照). したがって Q_c を小さくすることは合成ベクトルを構成する径路の折れ曲がりを小さくすることに対応する. 言いかえれば合成ベクトルを円周に近づけることに対応する. これに対して Q_s を小さくすることは合成ベクトルの方向を実測値の方向に近づけることに対応している.

我々が提案した方法のうちグラフ化については標本数が多いときにはそのままでは使えない. その場合径路の最終点のみを描くとかサンプリングした少数の標本に対する径路だけを描くというような配慮が必要であろう.

謝　　辞

数量化の計算プログラムは統計数理研究所駒澤勉室長のもの ([4]) をベースに作成したものであることを付記し謝意を表する. またプログラムの作成に協力していただいた統計数理研究所北村秀子さんに感謝する.

参 考 文 献

[1] 脇本和昌, 後藤昌司, 松原義弘 (1979). 多変量グラフ解析法, 朝倉書店.
[2] 脇本和昌, 田栗正章 (1974). 2次元図式パターンを用いる判別分析, 応用統計学, **3**, 119-135.
[3] Wakimoto, K. and Taguri, M. (1978). Constellation graphical method for representing multidimemsional data, *Ann. Inst. Statist. Math.*, **30**, A, 77-84.
[4] 駒澤　勉 (1982). 数量化理論とデータ解析, 朝倉書店.
[5] 馬場康維 (1981). 角度データの統計, 統計数理研究所彙報, **28**, 41-54.

Some Quantification Method of Qualitative Data Using Vector Transformation

Yasumasa Baba
(The Institute of Statistical Mathematics)

and

Kazumasa Wakimoto
(Okayama University)

A graphical representation method is proposed to represent individual sample scores visually on the constellation graph which can be used for the prediction of the objective variate from the R categorical variables (items). For the procedure of this representation we need an arbitrarily fixed "angle transform function G" with which each objective variate y_α is transformed to

$$Y_\alpha = R \exp \{i G(y_\alpha)\}.$$

Put

$$\delta_\alpha(jk) = \begin{cases} 1, & \text{if the } \alpha\text{-th sample responds to the k-th category in the j-th item} \\ 0, & \text{otherwise.} \end{cases}$$

Then the predictor for Y_α is given by

$$\hat{Y}_\alpha = \sum_{j=1}^{R} w_j \exp\left\{ i \sum_{k=1}^{K_j} \theta_{jk} \delta_\alpha(jk) \right\}$$

where w_j is the weight for j-th item and θ_{jk} is the angle from the horizontal line assigned to the k-th category in the j-th item and they are determined by minimizing the sum of square error

$$\sum_{\alpha=1}^{N} |Y_\alpha - \hat{Y}_\alpha|^2.$$

The α-th sample score is represented by the path corresponding to $\hat{Y}_\alpha$ on the constellation graph and the predictor of y_α for the α-th sample is obtained by

$$\hat{y}_\alpha = G^{-1}\{\arg(\hat{Y}_\alpha)\}$$

where G^{-1} is the inverse transformation of G.

This method is illustrated with an example in the case of $w_j = 1$, where our method is similar to that proposed by Hayashi (Quantification theory of Type I).