HOKKAIDO UNIVERSITY

Title	三次相関係数とそれによる数量化
Author(s)	新保, 勝; 佐藤, 義治; 山ノ井, 高洋 他
Citation	北海道大學工學部研究報告, 82, 125-132
Issue Date	1976-12-07
Doc URL	https://hdl.handle.net/2115/41393
Type	departmental bulletin paper
File Information	82_125-132.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

北海道大学工学部研究報告
第82号（昭和51年）

Bulletin of the Faculty of Engineering,
Hokkaido University, No. 82 (1976)

三次相関係数とそれによる数量化

新保　勝*　佐藤義治*
山ノ井高洋*　河口至商*

（昭和51年6月30日受理）

Correlation Coefficient for Three Variables and the Quantification of Categorical Data Thereby

Masaru SHIMBO　Yoshiharu SATO
Takahiro YAMANOI　Michiaki KAWAGUCHI

(Received June 30, 1976)

Abstract

A three dimensional correlation coefficient was newly defined in this paper to represent the degree of the linearity in the distribution of three dimensional data. It is different from that of the moment of third order defined elsewhere, which may represent the degree of their inclination. It follows an extension of the quantification theory III, in multivariate analysis proposed by C. Hayashi, which is originally restricted to two dimensional categorical data, to three dimensional ones with an example, where the correlation coefficient plays an important role.

1. 緒　　言

多変量解析には，カテゴリカルなデータを分類・整理する手法として，クラスター分析，数量化理論等がある。特に，後者に含まれる数量化理論第III類[1]は，いわゆる2-WAYのカテゴリカルなデータに対して数量化を行なうもので，数量化の基準としては，相関係数が重要な役割を演ずる。

近年，三つの観点（個体，属性，時間等）から観測されたデータ，いわゆる3-WAYデータの解析の必要性から，数量化理論第III類の3-WAYデータへの適用が問題となってきている。従来提案されていた，三次元の場合の相関係数の定義[2]を基に，3-WAYデータの数量化が提案された[3]。しかし，この相関係数は，二次元の場合の形式的な拡張であり，3-WAYデータの分類に用いるには必ずしも妥当なものとは思われない。

本論文では，相関係数をデータ間の直線性を表わす尺度とみなし，新たに，三次相関係数の定義を与える。また，これを用いて3-WAY離散データの数量化を行ない，数量化理論第III類の三次元への拡張を試みる。

2. 三次相関係数の定義

二次元における相関は，二変量のデータ間の線形性の度合の多寡を意味する。いま，座標系

* 北海道大学工学部　情報数理工学第一講座

を x, y として，一つの直線の方程式を

$$\frac{x-\bar{x}}{n_x}=\frac{y-\bar{y}}{n_y} \tag{2-1}$$

で与える。ただし，$\bar{x}$, $\bar{y}$ はデータの平均値，n_x, n_y は直線の方向余弦である。線形性の度合は，データの直線の式 (2-1) へのあてはめによって得られる。これによって，y の x への回帰直線として，

$$y-\bar{y}=\rho_{xy}\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \tag{2-2}$$

を得る。ただし，ρ_{xy} は相関係数，σ_x, σ_y はそれぞれ，x, y の標準偏差である。したがって，式 (2-1) から，二次元の場合の直線へのあてはめには，比 n_y/n_x の値が意味のある量となり，式 (2-2) から，データが与えられたときにそれを保証するものは，$\rho_{xy}\cdot\sigma_y/\sigma_x$ なる量である。あるいは，$\sigma_x=\sigma_y=1$ に標準化されたデータに対しては，相関係数 ρ_{xy} が唯一の意味のある量となる。

同様のことは，三次元の場合についても言えよう。この場合には，座標系を x, y, z とし，$\bar{z}$, n_z を z の平均値，方向余弦とすれば，直線の方程式は

$$\frac{x-\bar{x}}{n_x}=\frac{y-\bar{y}}{n_y}=\frac{z-\bar{z}}{n_z}$$

であり，この直線関係を表わす意味のある量は，方向余弦 n_x, n_y, n_z の任意の二つの比である。それ故，三次元における直線性の推定には

$$\rho_{xy}\frac{\sigma_y}{\sigma_x},\quad \rho_{yz}\frac{\sigma_z}{\sigma_y},\quad \rho_{zx}\frac{\sigma_x}{\sigma_z}$$

あるいは，$\sigma_x=\sigma_y=\sigma_z=1$ と標準化したデータに対しては，任意の二つの変量間の相関係数

$$\rho_{xy},\quad \rho_{yz},\quad \rho_{zx}$$

が意味のある量となる。

そこで，三変量の相関係数を表現する量として，これらの三個の二次相関係数の関数

$$f=f(\rho_{xy},\ \rho_{yz},\ \rho_{zx}) \tag{2-3}$$

を考えてみる。式 (2-3) で，$z=$const. 面を考えるなら

$$f=f(\rho_{xy},\ 0,\ 0)$$

となる。ただし，座標軸に平行にデータが並ぶ場合には，相関は 0 としてある。同様に，$x=$const. 面，$y=$const. 面を考えると，

$$f=f(0,\ \rho_{yz},\ 0)$$
$$f=f(0,\ 0,\ \rho_{zx})$$

となる。それゆえ，ρ_{xy}, ρ_{yz}, ρ_{zx} に関して高次の項を無視すれば，関数 f は

$$f=k_1\rho_{xy}+k_2\rho_{yz}+k_3\rho_{zx}$$

の形になる。ただし，k_1, k_2, k_3 は定数である。座標軸間の対称性を考慮し，また，f の最大値を 1 となるように基準化すれば，

$$f=\frac{1}{3}(\rho_{xy}+\rho_{yz}+\rho_{zx})\equiv\rho_{xyz} \tag{2-4}$$

を一つの相関を表現する量と見なせるであろう。我々はこの値を ρ_{xyz} と表わし，三次相関係数と呼ぶことにする。簡単な計算から，ρ_{xyz} は

$$-\frac{1}{3} \leqq \rho_{xyz} \leqq 1$$

の間にあることがわかる。また，$\rho_{xyz}=0$ としては，三次元的に無相関な場合と，座標平面上で無相関な場合が考えられることになる。

3.　三次相関係数の幾何学的解釈

相関係数は，以上で述べたように，データが直線上に分布している度合を示すものである。したがって，平均0，分散1に標準化したデータを三次元空間の点として考えるならば，相関係数は，原点を通り，方向比 $(1/\sqrt{3},\ 1/\sqrt{3},\ 1/\sqrt{3})$ の直線と，データの各点とのずれを示すものでもある。各点 $(x_i,\ y_i,\ z_i)$ から，この直線上へ垂線を下した距離の二乗の総和を U とすれば，幾何学的関係から，U は

$$U=\sum_i\left\{x_i^2+y_i^2+z_i^2-\frac{(x_i+y_i+z_i)^2}{3}\right\} \tag{3-1}$$

で与えられる。ここで，各変量は標準化してあるから，$\sum x_i^2=\sum y_i^2=\sum z_i^2=1$ の関係が成り立つことを考慮すれば，式 (3-1) は

$$U=2\left\{1-\frac{1}{3}\sum_i(x_iy_i+y_iz_i+z_ix_i)\right\}$$

となる。それゆえ，U は

$$\rho_{xyz}\equiv\frac{1}{3}\sum_i(x_iy_i+y_iz_i+z_ix_i) \tag{3-2}$$

なる量に依存するものであり，この量をもってどの程度データに直線性があるかを計ることができる。このことから，式 (2-4)，あるいは式 (3-2) で定義した ρ_{xyz} は三次の相関を表わしていることがわかる。

4.　3-WAY 離散データの数量化

3-WAY 離散データ（カテゴリカルデータ）の三方向の要因をそれぞれ，$X=\{x_1,\ x_2,\cdots\cdots,\ x_l\}$，$Y=\{y_1,\ y_2,\cdots\cdots,\ y_m\}$，$Z=\{z_1,\ z_2,\cdots\cdots,\ z_n\}$ と表わす。このとき，X について，たとえば，$x_i,\ x_j$ を固定したときの $(y,\ z)$ 面のパターンが互いに類似しているものには互いに近い実数値を与え，他と比べて類似していないもの同志には互いに離れた実数値を割り当てる。$Y,\ Z$ についても同様である。このように，それぞれの要因に実数値を割り当てる手法が数量化といわれているものである。以下では，$X,\ Y,\ Z$ を同時に数量化する際の基準として，式 (2-4) で与えた三次相関係数を用いる。

いま，要因間の同時確率 $p_{ijk}=P_r\{X=x_i,\ Y=y_j,\ Z=z_k\}$ は既知とすると，式 (2-4) は

$$\rho_{xyz}=\frac{1}{3}\sum_i^l\sum_j^m\sum_k^n p_{ijk}(u_iv_j+v_jw_k+w_ku_i) \tag{4-1}$$

と表わせる。ここに

$$u_i=\frac{x_i-\sum_i^l p_{i\cdot\cdot}x_i}{\left\{\sum_i^l p_{i\cdot\cdot}x_i^2-\left(\sum_i^l p_{i\cdot\cdot}x_i\right)^2\right\}^{\frac{1}{2}}},\qquad v_j=\frac{y_j-\sum_j^m p_{\cdot j\cdot}y_j}{\left\{\sum_j^m p_{\cdot j\cdot}y_j^2-\left(\sum_j^m p_{\cdot j\cdot}y_j\right)^2\right\}^{\frac{1}{2}}},$$

$$w_k=\frac{z_k-\sum_k^n p_{\cdot\cdot k}z_k}{\left\{\sum_k^n p_{\cdot\cdot k}z_k^2-\left(\sum_k^n p_{\cdot\cdot k}z_k\right)^2\right\}^{\frac{1}{2}}}$$

$$p_{i\cdot\cdot}=\sum_j^m\sum_k^n p_{ijk}\,,\qquad p_{\cdot j\cdot}=\sum_i^l\sum_k^n p_{ijk}\,,\qquad p_{\cdot\cdot k}=\sum_i^l\sum_j^m p_{ijk}$$

である。

したがって，X, Y, Z の数量化は，式 (4-1) を最大にするように，u_i, v_j, w_k を決定する問題に帰着するので，2-WAY 離散データに対する林の数量化理論 III 類と同様である。以下，各変量が標準化されているとして

$$\sum_i^l p_{i\cdot\cdot}u_i^2=1\,,\qquad \sum_j^m p_{\cdot j\cdot}v_j^2=1\,,\qquad \sum_k^n p_{\cdot\cdot k}w_k^2=1$$

の条件の下で，ρ_{xyz} を最大にするような u, v, w を定める。そのためには，Lagrange の未定乗数 λ_1, λ_2, λ_3 を用いて，

$$\begin{aligned}f(u,v,w)=&\sum_i^l\sum_j^m\sum_k^n p_{ijk}(u_iv_j+v_jw_k+w_ku_i)\\&+\lambda_1(1-\sum_i^l p_{i\cdot\cdot}u_i^2)+\lambda_2(1-\sum_j^m p_{\cdot j\cdot}v_j^2)+\lambda_3(1-\sum_k^n p_{\cdot\cdot k}w_k^2)\end{aligned}\tag{4-2}$$

とし，これを u_r で偏微分し，0 とおくと，

$$\frac{\partial f}{\partial u_r}=\sum_j^m p_{rj\cdot}v_j+\sum_k^n p_{r\cdot k}w_k-2\lambda_1p_{r\cdot\cdot}u_r=0$$

となる。この式の両辺に u_r を乗じて r で和をとり，整理すると，

$$2\lambda_1=\sum_i^l\sum_j^m p_{ij\cdot}u_iv_j+\sum_i^l\sum_k^n p_{i\cdot k}w_ku_i$$

を得る。v, w に関しても同様に

$$2\lambda_2=\sum_i^l\sum_j^m p_{ij\cdot}u_iv_j+\sum_j^m\sum_k^n p_{\cdot jk}v_jw_k$$

$$2\lambda_3=\sum_j^m\sum_k^n p_{\cdot jk}v_jw_k+\sum_k^n\sum_i^l p_{i\cdot k}w_ku_i$$

が得られる。これらの関係より，

$$3\rho_{xyz}=\sum_i^l\sum_j^m p_{ij\cdot}u_iv_j+2\lambda_3$$

となり，ρ_{xyz} を最大にする u_i, v_j, w_k を求めるには，上式の

$$\rho_{uv}\equiv\sum_i^l\sum_j^m p_{ij\cdot}u_iv_j$$

を最大にする u, v を求め，さらに式 (4-1) を最大にするように w を決定すればよい。故に，極値をもつ条件，

$$\frac{\partial\rho_{uv}}{\partial x_\alpha}=0\,,\qquad \frac{\partial\rho_{uv}}{\partial y_\beta}=0$$

より，

$$\begin{cases}\sum_j^m p_{\alpha j\cdot}v_j=p_{\alpha\cdot\cdot}u_\alpha\rho_{uv}\\ \sum_i^l p_{i\beta\cdot}u_i=p_{\cdot\beta\cdot}v_\beta\rho_{uv}\end{cases}\tag{4-3}$$

なる関係式を得る。上式からつぎの固有方程式が導かれる。

$$\sum_j^m a_{j\beta}v_j=\rho^2{}_{uv}v_\beta\,,\tag{4-4}$$

$$a_{j\beta}\equiv\sum_i^l\left\{p_{i\beta\cdot}\frac{p_{ij\cdot}}{p_{i\cdot\cdot}p_{\cdot\beta\cdot}}\right\}$$

したがって，ρ_{uv} を最大にする u, v を求めるためには，式 (4-4) の最大固有値（1を除く）に対応する固有ベクトル v_j を求め，式 (4-3) の関係より u_i を決定する。ここで求まった ρ_{uv}, u_i, v_j を用いて，w_k を決定する手順は次の通りである。式 (4-1) を書きかえると，

$$3\rho_{xyz} = \rho_{uv} + \sum_{k}^{n}\sum_{i}^{l}\sum_{j}^{m}(p_{ijk}v_j + p_{ijk}u_i)w_k \tag{4-5}$$

となり，これを z_τ で偏微分して，0とおくことにより，

$$\sum_{i}^{l}\sum_{j}^{m}(p_{ij\tau} - p_{i\cdot\tau}p_{ij\cdot})(u_i + v_j) = p_{\cdot\cdot\tau}w_\tau\sum_{i}^{l}\sum_{j}^{m}\sum_{k}^{n}(p_{ijk}v_j + p_{ijk}u_i)w_k$$

表—1 3-WAY 離散データ ($i=1\sim8$, $j=1\sim9$, $k=1\sim10$)

$k=1$

i \ j	1	2	3	4	5	6	7	8	9
1							1		
2				1			1		
3						1	1		
4	1								
5									
6	1			1		1	1		
7				1		1	1		
8	1		1	1		1	1		

$k=2$

i \ j	1	2	3	4	5	6	7	8	9
1		1	1					1	1
2									
3		1	1		1	1		1	1
4									
5		1	1		1	1		1	1
6									
7					1				
8									

$k=3$

i \ j	1	2	3	4	5	6	7	8	9
1	1			1					
2	1			1			1		
3									
4	1			1		1	1		
5									
6	1			1			1		
7	1								
8	1						1		

$k=4$

i \ j	1	2	3	4	5	6	7	8	9
1		1	1		1			1	1
2									
3					1	1		1	1
4		1							
5									
6									
7		1						1	
8									

$k=5$

i \ j	1	2	3	4	5	6	7	8	9
1							1		
2	1			1		1			
3									
4						1	1		
5									
6	1			1		1	1		
7	1			1		1	1		
8				1		1	1		

$k=6$

i \ j	1	2	3	4	5	6	7	8	9
1		1	1		1			1	1
2									
3		1	1		1			1	1
4									
5		1	1		1			1	1
6									
7		1						1	
8									

$k=7$

i \ j	1	2	3	4	5	6	7	8	9
1									
2	1					1	1		
3									
4	1			1		1	1		
5									
6				1		1	1		
7	1			1		1			
8	1			1		1	1		1

$k=8$

i \ j	1	2	3	4	5	6	7	8	9
1		1	1		1			1	
2									
3		1						1	1
4									
5		1	1		1	1		1	
6					1			1	
7								1	
8									

$k=9$

i \ j	1	2	3	4	5	6	7	8	9
1	1						1		
2	1			1		1	1		
3									
4	1			1		1	1		
5									
6	1			1		1	1		
7							1		
8	1			1		1	1		

$k=10$

i \ j	1	2	3	4	5	6	7	8	9
1	1						1		
2	1			1		1	1		
3									
4	1			1		1	1		
5									
6	1			1		1	1		
7	1			1		1	1		
8	1		1	1		1	1		

を得る。ここで，

$$P_\gamma \equiv \sum_i^l \sum_j^m (p_{ij\gamma} - p_{i\cdot\gamma} p_{ij\cdot})(u_i + v_j)$$

とおき，式 (4-5) を用いることにより

$$P_\gamma = (3\rho_{xyz} - \rho_{uv}) p_{\cdot\cdot\gamma} w_\gamma \tag{4-6}$$

なる関係を得る，一方

$$\bar{\rho} \equiv 3\rho_{xyz} - \rho_{uv}$$

とおくと，

$$\bar{\rho} = \left\{ \sum_i^l \sum_j^m \sum_k^n \left(\frac{p_{ijk}}{p_{\cdot\cdot k}} v_j + \frac{p_{ijk}}{p_{\cdot\cdot k}} u_i \right) P_k \right\}^{\frac{1}{3}}$$

なる式は，簡単な計算により，式 (4-5) から得ることができる。したがって，この値は，u, v が決まれば求まる値であるから，これを式 (4-6) に代入することにより w を決定することができる。以上の手続きにより，u_i, v_j, w_k をそれぞれ，x_i, y_j, z_k に割り当てる実数値とすることができる。

5.　従来の数量化との比較

表-1 のようなデータが与えられたとして岩坪は表-2, 図-1 の結果を得ている[3]。本論文の三次相関係数を用いた数量化の結果として表-3, 図-2 が得られた。

今ここで，二つの手法を比較するために次のように考える。k の分類 ($k=1\sim10$) を考えた場合，i, j に関する二次元平面の 1—0 パターンをもとに，似たもの同志を近づけ，似ていないもの同志は分離するように並べかえを行なう。このとき，k を構成する任意の二つの 1—0 パターンの親縁性を測るものとして，通常 2×2 分割表にもとづく，一致係数とか，類似比が用いられている。これらは，

$$\text{一致係数} = \frac{a+d}{a+b+c+d},$$

$$\text{類 似 比} = \frac{a}{a+b+c}$$

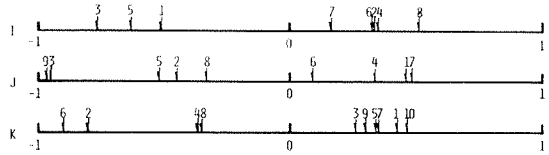
であり，ここで，a: 個体 A と個体 B とがともに 1 である場合の数，b: 個体 A が 1，個体 B が 0 である場合の数，c: 個体 A が 0，個体 B が 1 である場合の数，d: 個体 A と個体 B がとも

表—2　数量化例[3]

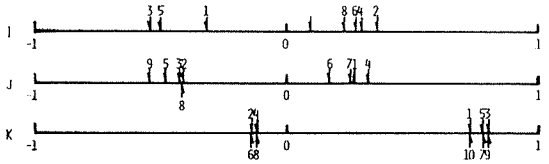
	i	j	k
1	−0.5452	0.4593	0.4226
2	0.3336	−0.4486	−0.8010
3	−0.7669	−0.9505	0.2623
4	0.3518	0.3365	−0.3651
5	−0.6332	−0.5212	0.3391
6	0.3291	0.0899	−0.8990
7	0.1672	0.4833	0.3496
8	0.6248	−0.3288	−0.3528
9		−0.9634	0.3020
10			0.4651

表—3　本手法による数量化

	i	j	k
1	−0.3040	0.3818	0.2455
2	0.3713	−0.2531	−0.4136
3	−0.5295	−0.2640	0.2684
4	0.3137	0.4329	−0.3532
5	−0.4882	−0.3178	0.2592
6	0.2880	0.2895	−0.4223
7	0.1080	0.3698	0.2611
8	0.2407	−0.2589	−0.3566
9		−0.3802	0.2674
10			0.2440



図—1　数量化例[3)]



図—2　本手法による数量化

表—4　比較のための類似比

i	類似比	j	類似比	k	類似比
1と5	0.4839	3と9	0.5714	2と6	0.7000
2と4	0.6000	3と8	0.5000	2と4	0.5715
3と5	0.7619	3と2	0.6000	2と8	0.5500
4と6	0.7037			4と8	0.5000
6と8	0.7142				

に0である場合の数である。この類似比を用いて，i, j, k の分類の比較を行なう。

i については，図-1によれば，1と5，また5と3は同程度の距離であるが，本論文の手法では，1と5より5と3の方がより似ているといえる。類似比は，1と5が0.4839，5と3が0.7619である。また，2と4と6の分類において，図-1によれば，2の方が6よりも4と似ているとしている。本手法では，6の方が2よりも4と似ていると言える。類似比は，2と4が0.6000，4と6が0.7037である。また，2, 4, 6, 8の分類においては，図-1によれば，2と4の方が6と8よりはるかに似ているとなっている。本手法では，逆に2と4より，6と8の方がわずかに似ている。類似比は，2と4が0.6000，6と8が0.7142である。j, k についても同様である。

6. 結　　言

本論文は，三次元空間の直線性を表わすという概念から，三次相関係数を定義して，これを基準として，3-WAY 離散データを数量化する手法を提案した。この手法と，竹内の三次相関係数を用いた岩坪の数量化の手法と比較した。二つの手法の差は，基準となる三次相関係数の式の持っている根本的概念の差によるものと思われる。実際のデータへの応用としては，三次相関係数のそれほど高くない場合には，第二番目に大きい固有値に対する固有ベクトルを求めて，それぞれのカテゴリ i, j, k に対して二次元の平面上の点を対応させることにより分類が可能となる。

数値計算については，3-WAY データの解析には，三次元配列を必要とするので，おのずから計算機の記憶容量に制限されることになる。

本論文では，3-WAY 離散データの数量化に三次相関係数を応用したが，連続量のデータの相関はもちろん，他の多変量解析の手法への応用も考えられよう。

尚，計算実験には，本学大型計算機センター FACOM 230-75 を使用した。

文　　献

1) 林知己夫：統計数理研究所彙報，2 (1954) 1, p. 11.
2) Takeuchi, Kei: Behaviormetrika, No. 1 (1974), p. 59.
3) 岩坪秀一：行動計量学，2 (1974) 1, p. 54.