日本行動計量学会第４３回大会抄録集、2015、298－301（原稿）

２段共通数量化分析法による分析例

ビッグデータ分析のために

○岡本安晴
日本女子大学人間社会学部

1　はじめに

２段共通数量化分析法（岡本、2013、2014）の特色の１つは、複数の分析法に対して共通の数量化が用いられることである。多変量解析などは第２段において行われるが、第１段における数量化は第２段の分析法とは独立して行われる。変数間の関係は第２段において分析されるので、第１段における数量化は第２段において行われる分析法の選択の影響を受けないように変数ごとに行われ、ビッグデータにも適している。数量化が第１段において行われるので、数量化後の第２段において連続変量と数量化変量とを一緒に分析することができる。数量化の対象となる変数は名義尺度水準であるとするので、統計学的情報は各カテゴリの頻度分布で表される。数量化は、この頻度分布を最適に反映するものとして求められる。いま、k番目のカテゴリ変量のindicator行列G_kに対する数量化ベクトルをω_kとおくとき、頻度分布情報を最適に反映するω_kは数量化後の分散$Var(G_k\omega_k)$を制約条件

$$\omega'_k\omega_k = 1$$

のもとで最大にするものであると考える。このω_kを求めるための目的関数は、

$$H_k = G_k - \frac{1}{n}\mathbf{1}_n\boldsymbol{f}_k、 \quad \boldsymbol{f}_k = \mathbf{1}_nG_k$$

とおくとき、

$$Q_k^{(s)} = \left(H_k\omega_k^{(s)}\right)'\left(H_k\omega_k^{(s)}\right)$$

$$-\lambda_k^{(s)}\left(\omega_k^{(s)\prime}\omega_k^{(s)} - 1\right) - \mu_{k,s}^{(1)}\left(\omega_k^{(1)\prime}\omega_k^{(s)} - 0\right) - \cdots - \mu_{k,s}^{(s-1)}\left(\omega_k^{(s-1)\prime}\omega_k^{(s)} - 0\right)$$

と設定することができる。ここで、$\omega_k^{(1)}$、…、$\omega_k^{(s-1)}$は、すでに求められた数量化ベクトルであり、上式の$Q_k^{(s)}$における$\omega_k^{(s)}$はs番目の数量化ベクトルω_kである。これらの数量化ベクトルに対して次式が成り立つ。

$$H'_kH_k\omega_k^{(s)} = \lambda_k^{(s)}\omega_k^{(s)}、 \qquad Var\left(G_k\omega_k^{(s)}\right) = Var\left(H_k\omega_k^{(s)}\right) = \lambda_k^{(s)}/n.$$

ここで、$\lambda_k^{(1)} \geq \cdots \geq \lambda_k^{(s)}$である。$\lambda_k^{(t)} = \cdots = \lambda_k^{(t+u)}$であれば、$\omega_k^{(t)}$、…、$\omega_k^{(t+u)}$は、$(u+1)$次元空間の任意の直交基底ベクトルとして与えられる。標準化された数量化は

$$z^{(s)} = H_k \omega_k^{(s)} / \sqrt{\lambda_k^{(s)} / n}$$

で与えられる。いま、データが

$$X = (X_{ij}) = (X_1, \cdots, X_p, X_{p+1}, \cdots, X_{p+q})$$

の形式で表され、$X_1, \cdots, X_p$が連続変量、$X_{p+1}, \cdots, X_{p+q}$がカテゴリ変量であるとする。X_{p+k}がK_k個の数量化ベクトルにより

$$G_k^z = \left(z_k^{(1)}, \cdots, z_k^{(K_k)}\right)$$

と標準化された値に数量化されたとき、与えられたデータの数量化後の値は

$$X_G = \left(X_1, \cdots, X_p, G_1^z, \cdots, G_q^z\right)$$

と表される。このX_Gに対して、第２段においていろいろな多変量解析による分析を行うことができる。分析例を次に示す。

表１　カテゴリ変数を含むデータ

Group	ID	Score	Attr.	CA	CB
GA	A1	45	1	X	alpha
GA	A2	51	3	X	alpha
GA	A3	39	2	X	gamma
GA	A4	67	5	Z	alpha
GA	A5	50	2	X	alpha
GB	B1	85	8	Y	alpha
GB	B2	91	9	Y	alpha
GB	B3	72	6	Z	alpha
GB	B4	75	7	Y	gamma
GC	C1	52	7	Y	beta
GC	C2	43	9	Y	beta
GC	C3	65	7	Y	gamma
GC	C4	38	5	Z	beta
GD	D1	15	1	X	beta
GD	D2	17	3	X	beta
GD	D3	37	4	Z	beta
GD	D4	35	2	X	gamma

2　分析例

表１のデータを２段共通数量化分析法によって分析してみた。

2.1　第１段：共通数量化

カテゴリ変数 CA と CB、それぞれの数量化におけるλとωの値を表２に示す。それぞれの変数における３つのカテゴリに対する頻度分布が同じであるので、λとωはお互いに同じ値が得られている。数量化後の標準得点を、それぞれ CA-0、CA-1、CB-0、CB-1 で表す。H_kのランクがそれぞれ２であるので、標準化数量化得点は２個ずつ与えられている。連続変量 Attr.も第２段では標準化したものを用いる（標準化数量化得点と一緒に独立変数として用いる）が、これを Attr.-0 で表す。このとき、数量化後のデータは表３のようになる。

表２　カテゴリ変数のλとカテゴリ値

λ	6.492	4.566
X/alpha	0.747	-0.331
Y/beta	-0.660	-0.481
Z/gamma	-0.087	0.812

2.2　第２段：多変量解析

表３　数量化後のデータ

Group	ID	Score	Attr.-0	CA-0	CA-1	CB-0	CB-1
GA	A1	45	-1.411	1.120	-0.416	1.120	-0.416
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
GD	D4	35	-1.036	1.120	-0.416	-0.228	1.788

表３のデータに対して、主成分分析、回帰分析、判別分析を行う。

表４　構造行列

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Attr.-0	-0.987	-0.005	0.079	-0.026	0.138
CA-0	0.987	0.017	0.013	-0.086	0.137
CA-1	0.063	-0.810	-0.014	0.583	0.014
CB-0	0.067	0.004	0.997	0.022	-0.012
CB-1	0.035	0.811	-0.018	0.584	0.012

2.2.1　主成分分析

5 個の変数、Attr.-0、CA-0、CA-1、CB-0、CB-1 に対して主成分分析を行い、５つの主成分を
得た。各変量との相関係数を表４に示す。Comp.s で s 番目の主成分を表す。各ケースの第１主成分と第３主成分の値を横座標、縦座標の値として布置を描くと図１のようになる。布置の描画では、ケースの重なりを防ぐため、ランダムな揺らぎを加えてある。グループ、A、B、C、D ごとにまとまっていることがわかる。すなわち、第１主成分と第 3 主成分は、グループ分けの特徴を反映するものである。

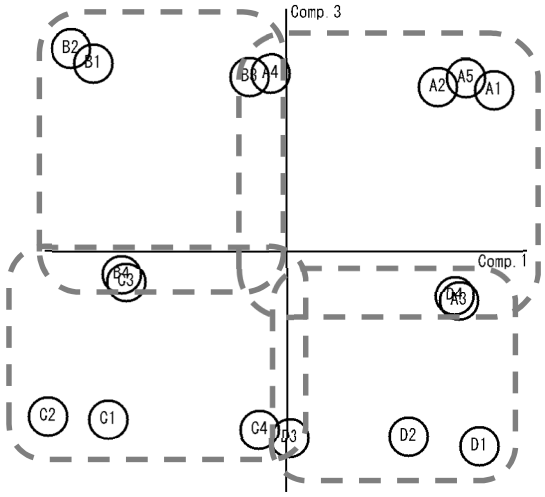

図１　各ケースの主成分値による布置

2.2.2　回帰分析

従属変数 Score に対して、標準化された連続変量 Attr.-0 と共通数量化変量 CA-0、CA-1、CB-0、CB-1 を独立変量として回帰分析を行った。Score とその予測値 $\widehat{Score}$の関係を図示すると図２のようになる。回帰係数の値を表５に示す。独立変数として主成分値 Comp.1、・・・、Comp.5 をとると表６の結果を得る。Comp.1 の係数が負（-14.084）であり、Comp.3 の係数が正（14.667）で

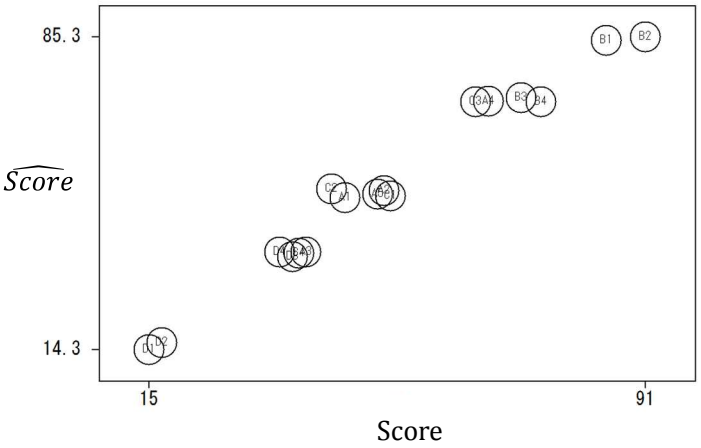

図２　データ値 Score と予測値$\widehat{Score}$

表５　回帰分析（共通数量化変量）

Const.	Attr.-0	CA-0	CA-1	CB-0	CB-1
51.588	2.101	-13.310	0.370	14.773	3.058

あることから、図１の布置において右下から左上の方向に Score の値が増加していることがわか

る。

2.2.3　判別分析

データの各ケースは、表１に示されているように４群、GA、GB、GC、GD、に分かれている。

表６　回帰分析（主成分）

Const.	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
51.590	-14.084	1.998	14.667	3.407	-1.679

このグループ分けに対して、変量 Attr.-0、CA-0、CA-1、CB-0、CB-1 によって Fisher の判別分析を行った。λの値は、$\lambda_1 = 3.880$、$\lambda_2 = 3.135$、$\lambda_3 = 0.012$となり、第１判別関数 Discri.1 と第２判別関数 Discri.2 によって主たる判別が行われたことがわかる。判別関数の値を座標値とし

表７　構造行列（判別関数と変数との相関係数）

	Attri.-0	CA-0	CA-1	CB-0	CB-1
Discri.1	-0.902	0.922	0.070	0.410	0.004
Discri.2	0.411	-0.331	-0.062	0.908	0.096
Discri.3	0.056	0.141	0.277	-0.021	0.257

て、各ケースと各群の平均値の布置を図示すると図３のようになる。ケースの重なりを避けるため、ランダムな揺らぎを加えてある。図３に示されている布置は、図１のものとよく対応している。判別関数と各変数との相関係数を求めると、表７のようになる。第１判別関数は、Attr.-0 および CA-0 と強い関係があり、第２判別関数は CB-0 と強い相関がある。これは、表４に示されている第１主成分および第３主成分と各変量との関係によく対応している。判別関数と主成分との相関を求めると、表８のようになる。第１判別関数は第１主成分と、第２判別関数は第３主成分と強い相関がある。これは図１と図３に描かれている布置がよく似ていることと対応するものであり、第

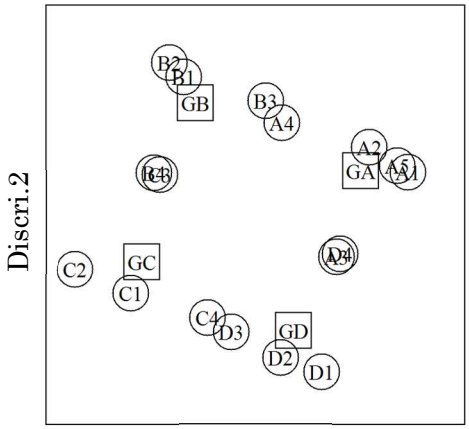

図３　判別分析

表８　判別関数と主成分との相関係数

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Discri.1	0.936	-0.024	0.348	-0.005	-0.047
Discri.2	-0.343	0.094	0.931	0.082	0.007
Discri.3	0.056	-0.011	-0.023	0.432	0.900

１判別関数と第２判別関数は、それぞれ第１主成分と第３主成分によく対応している。

3　引用文献

岡本安晴（2013）、日本行動計量学会大会発表論文抄録集 41、192－195.

岡本安晴（2014）、日本行動計量学会大会発表論文抄録集 42、72－75.