双対尺度法 (Dual Scaling) について

1　質的データの多変量解析的分析

1.1　対応する分析手法

数量化は、質的なデータに数値を与える方法である。数値の与え方については、目的によって以下の四種類に分かれる (表.1 参照)。

表 1　林の数量化

手法	外的基準	データ	目的	関連する手法
数量化 I 類	量的変数	質的変数	外的基準の予測	重回帰分析
数量化 II 類	質的変数	質的変数	外的基準の判別	判別分析
数量化 III 類	なし	質的変数	変数間の関係の要約と記述	正準相関分析
数量化 IV 類	なし	対象間の類似度	対象間の関係の要約と記述	多次元尺度法

2　双対尺度法

2.1　量的変数の場合はどのように分析するか

質的なデータを分析する前に、演算操作が簡単な量的変数の場合を復習しておこう。量的変数の場合、その内部構造を明らかにする方法として、主成分分析や因子分析があげられよう。主成分分析は、$X_1, X_2, X_3, \ldots$ という量的変数に重み $w_1, w_2, w_3, \ldots$ をつけることで得られる合成変数 Y を作る方法である (式 1)。この合成変数 Y の分散を最大にするように重みを決定するのが主成分分析であった。

$$Y = w_1X_1 + w_2X_2 + w_3X_3 + \ldots + w_nX_n \tag{1}$$

$$\text{ただし、} w_1^2 + w_2^2 + w_3^2 + \ldots + w_n^2 = 1.0$$

合成得点の分散の最大値は固有値に等しく、このとき作られる合成変数を主成分、重みを負荷量と呼ぶ。これが量的変数の場合であり、この背景には得られた量的データが、少なくとも間隔尺度以上の水準である必要がある。しかし、実際に得られたデータがどうして間隔尺度以上の水準であるといえるのか。ここで少し歴史をひもといてみよう。

2.2 尺度に値を与えるには

2.2.1 Likert 法

Likert 法と呼ばれる調査項目のデータ例は、例えば「滅多にない」、「たまにある」、「しばしばある」、「いつもある」という項目に、それぞれ 1 点、2 点、3 点、4 点を与える。では、なぜそんなことができるのか。例えば、以下のような回答例が得られたとする。

表 2　Likert 尺度の数値化

カテゴリ	回答数	相対度数	累積相対度数
滅多にない	5	0.1	0.1
たまにある	10	0.2	0.3
しばしばある	20	0.4	0.7
いつもある	15	0.3	1.0

さて、被験者の散らばりが正規分布すると仮定すると、得られたデータからどのように点数を与えるのがよいだろうか。方法として、相対度数に従って正規分布を分割し、分割されたエリアの代表値を用いてその得点とするというものが考えられよう。

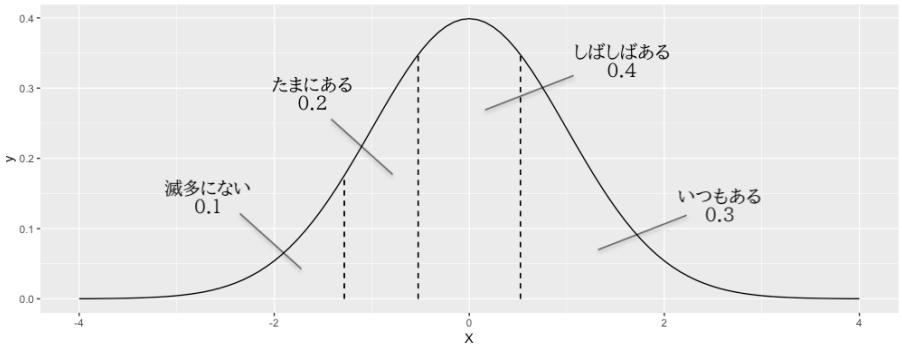

図 1　正規分布の分割

代表値として、平均値もしくは中央値を用いたとすると、得られる尺度値は以下のようになる。

表 3　正規分布を用いた数値化

整数値	1	2	3	4
平均値による数値化	1.0	1.7	2.8	4.0
中央値による数値化	1.0	1.8	2.6	3.7

このように、平均値や中央値を用いて得られた値を尺度値とする方法が、Likert の考えたものであった。しかし、数値化した値を四捨五入すると、整数値と同じ値が得られる。このように、Likert の方法を用いると、尺度値としてだいたい順番につけた数値が得られることが経験的に明らかになっているので、昨今の調査法で

は点数化の際にややこしいことを教えず、単純に 1〜4 点を与えるようになっている（下手の考え休むに似たり、ということか)。

2.2.2 重みづけられた得点

複数の項目から、一人分の点数を計算する際、Byrd は項目得点の総和を個人の得点とする方法を考えた（式 2)。これは、どの項目にも同じ重みをつけて足し合わせることになっている。

$$Y = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \ldots + \frac{1}{n}X_n \tag{2}$$

もちろんこの方法よりも、情報量に応じて重みを変更する、式 1 の主成分分析の方が優れていると言えるだろう。しかし、主成分分析だけではわからないことがある。例えば血圧と年齢のデータがあって、クロス集計表を作ったとする (表.4)。

表 4　血圧と年齢

血圧	20-34 歳	35-49 歳	50-65 歳
高い	0	0	4
普通	1	4	1
低い	3	1	1

このような線形関係が明らかな場合は、主成分分析をすることに意味があるだろう。しかし、血圧と頭痛（「ない」、「たまにある」、「時々ある」）の表のような場合 (表.5)、これは情報量がないといえるだろうか？

表 5　血圧と頭痛

血圧	ない	たまに	時々
高い	0	0	4
普通	3	3	0
低い	0	0	5

このような場合は、「血圧が高い人か低い人は、頭痛がある」という傾向が見て取れる。が、二つの項目の相関係数としては $r = -0.06$ なので、主成分分析では意味のある主成分を抽出しにくい。

これはなぜか、というと主成分分析は線形性の制限があるからである。すなわち、「1. 血圧高い」、「2. 血圧普通」、「3. 血圧低い」という順番が保存されていなければならない。例えばこれが、表.6 のようであれば、綺麗な線形関係になっているし、主成分分析もその威力を発揮したであろう。

表 6　並び替えられた「血圧と頭痛」

血圧	ない	たまに	時々
高い	0	0	4
低い	0	0	5
普通	3	3	0

双対尺度法は、このように**行・列の双方を並び替えることによって、非線形関係において線形性が最大になるものを見つける方法**であると言える。

3 アルゴリズム

3.1 双対尺度法のイメージ

双対尺度法は、分割表の χ^2 値を、独立した成分に分割するものといえる。

$$\chi^2 = \chi_1^2 + \chi_2^2 + \chi_3^2 + \dots \tag{3}$$

主成分分析が固有値分解したように、双対尺度法は特異値分解を使う。すなわち、

$$f_{ij} = \underbrace{\underbrace{\underbrace{\frac{f_{i.}f_{.j}}{f_t}}_{\text{第 0 次近似}}(1 + \rho_1 y_{1i}x_{1j}}_{\text{第 1 次近似}} + \rho_2 y_{2i}x_{2j} + \dots + \rho_k y_{ki}x_{kj})}_{\text{第 2 次近似}} \tag{4}$$

このとき、y_{1i}は y 列の第 1 ウェイト、x_{1j}は j 行の第 1 ウェイト、ρ^2を特異値という。

特に、どの項までで近似するかを示すのに、第 n 次近似、という表現を用い、表が $m \times n$ の場合、第 $k = min(m, n) - 1$ 次近似でデータを完全に説明する。

3.2 交互平均法

ここでは具体的に、どのようなアルゴリズムで行・列に対する重みをつけるか見ておこう。必要ない方は読み飛ばしていただいて結構である。
例えば、以下のような分割表が得られたとする（表.7）。

表 7　データ例

血圧	頭痛なし	たまに頭痛	頭痛あり	計
低い	9	2	3	14
普通	3	1	4	8
高い	6	2	5	13
計	18	5	12	35

1. 行方向の「なし (X_1)」を 1 点、「たまに (X_2)」を 0 点、「あり (X_3)」を-1 点と仮におく。
2. 列方向の「低い (Y_1)」、「普通 (Y_2)」、「高い (Y_3)」を算出する。すなわち、

$$Y_1 = \frac{1 \times 9 + 0 \times 2 + (-1) \times 3}{14} = 0.4285$$

$$Y_2 = \frac{1 \times 3 + 0 \times 1 + (-1) \times 4}{8} = -0.125$$

$$Y_3 = \frac{1 \times 6 + 0 \times 2 + (-1) \times 5}{13} = 0.0769$$

3. Y の平均値を 0 にする。すなわち、

$$\bar{Y} = \frac{14 \times Y_1 + 8 \times Y_2 + 13 \times Y_3}{35} = 0.1714$$
$$Y_1 = Y_1 - \bar{Y} = 0.4285 - 0.1714 = 0.2571$$
$$Y_2 = Y_2 - \bar{Y} = -0.125 - 0.1714 = -0.2964$$
$$Y_3 = Y_3 - \bar{Y} = 0.0769 - 0.1714 = -0.0945$$

4. Y のノルムを整える。すなわち、絶対値最大の要素で割ることになるので、

$$Y_1 = Y_1/maxY = 0.2571/|-0.2964| = 0.8674$$
$$Y_2 = Y_2/maxY = -0.2964/|-0.2964| = -1.000$$
$$Y_3 = Y_3/maxY = -0.0945/|-0.2964| = -0.3188$$

5. これらの重みを使って、今度は行の重みを算出する。

$$X_1 = \frac{Y_1 \times 9 + Y_2 \times 3 + Y_3 \times 6}{18} = 0.1608$$
$$X_2 = \frac{Y_1 \times 2 + Y_2 \times 1 + Y_3 \times 2}{5} = 0.0194$$
$$X_3 = \frac{Y_1 \times 3 + Y_2 \times 4 + Y_3 \times 5}{12} = -0.2493$$

6. X の平均値を 0 にする。

$$\bar{X} = \frac{18 \times X_1 + 5 \times X_2 + 12 \times X_3}{35} = -3.806e-17$$
$$X_1 = X_1 - \bar{X} = 0.1608$$
$$X_2 = X_2 - \bar{X} = -0.0194$$
$$X_3 = X_3 - \bar{X} = -0.2493$$

7. X のノルムを整える。

$$X_1 = X_1/maxX = 0.6450$$
$$X_2 = X_2/maxX = 0.0780$$
$$X_3 = X_3/maxX = -1.000$$

8. 2〜7 を収束するまで繰り返す。収束判定は、

$$\delta_Y = \sum Y_i^{t+1} - Y_i^t$$
$$\delta_X = \sum X_i^{t+1} - X_i^t$$
$$\text{として、} \delta_X < \epsilon \quad \text{且つ} \quad \delta_Y < \epsilon$$

こうして行および列に対する重みが得られる。この例では、最終的に (5) の重みが得られるであろう。

$$X = \begin{pmatrix} 0.644 \\ 0.081 \\ -1.00 \end{pmatrix}, Y = \begin{pmatrix} 0.851 \\ -1.00 \\ -0.301 \end{pmatrix} \tag{5}$$

このような、各系列の重みを代入して、平均値を更新し続けていく方法を、交互平均法 (Method of Reciprocal Averages) という。

4　他の手法との違い

さて、双対尺度法の意味、および計算方法が理解できたと思うので、ここで数量化 Ⅲ 類とコレスポンデンス分析、および双対尺度法とよばれる一群の手法が、それぞれどのように異なっているのかについて論じよう。

4.1　数量化 Ⅲ 類とコレスポンデンス分析

数量化 Ⅲ 類とコレスポンデンス分析は、分析のもとになる分割表が異なっている。例えば、「はい」と「いいえ」で回答が得られる項目が三つあったとしよう。このとき、数量化 Ⅲ 類が分析対象にする行列は、「はい」を 1、「いいえ」を 0 とした分割表.8 のようになっている。

表 8　数量化が分析する行列

被験者	項目 1	項目 2	項目 3	合計
S1	1	0	1	2
S2	0	0	1	1
S3	1	1	0	2
S4	1	0	1	2
S5	1	1	1	3
	4	2	4	

これに対して、コレスポンデンス分析（および双対尺度法）が分析対象とする行列は、表.9 のようになっている。

表 9　コレスポンデンス分析、および双対尺度法が分析する行列

被験者	項目 1(Yes)	項目 1(No)	項目 2(Yes)	項目 2(No)	項目 3(Yes)	項目 3(Yes)	合計
S1	1	0	0	1	1	0	3
S2	0	1	0	1	1	0	3
S3	1	0	1	0	0	1	3
S4	1	0	0	1	1	0	3
S5	1	0	1	0	1	0	3
	4	1	2	3	4	1	

要するに、この分析するもととなる行列の違いがあるだけで、それを特異値分解して行・列の両方に重みを与える点では同じ分析手法である。

4.2　コレスポンデンス分析と双対尺度法

コレスポンデンス分析と双対尺度法の違いは、もう少し説明がややこしい。これらの違いについて考える前に、双対の関係（双対性）について知る必要がある。双対の関係とは、式 (6) のようなものである。

$$\begin{cases} Y_i &= \dfrac{1}{\rho}\dfrac{\sum f_{ij}X_j}{f_{i.}} \\ X_j &= \dfrac{1}{\rho}\dfrac{\sum f_{ij}Y_i}{f_{.j}} \end{cases} \tag{6}$$

さて、この双対の関係が意味するのは、Y のベクトルが張る空間と X のベクトルが張る空間は等しくなく、一方が他方に射影されるためには特異値をもちいて変換しなければならない[*1]。

このように列および行の項目を、それぞれ行ベクトルの作る空間か、列ベクトルの作る空間か、どちらかに射影することを選択させるのが双対尺度法であるが、コレスポンデンス分析はこれをひとつの空間に射影する。ρY_i と ρX_j は、ノルムを整えてあるので分散が同じである。そこで、この ρY_i と ρ_j をひとつの空間に射影しようじゃないか、というのがコレスポンデンス分析のやり方なのである。これは、数学的に正しくないが、アウトプットが見やすいというメリットがある。あくまでも、同じ空間にプロットできているわけではない点に注意が必要である。

5　双対尺度法の発展

5.1　順序データの場合

例えば、好きな政党を順に番号をつけてもらうといった順序データを考える（表.10）。

表 10　順序データの場合

被験者	自民党	民主党	公明党	自由党	社民党	合計
S1	1	2	4	5	3	15
S2	2	1	3	4	5	15
⋮	⋮	⋮	⋮	⋮	⋮	⋮

双対尺度法ではドミナンス表に変換してから計算を行う。ドミナンス表とは、R_{ij} を被験者 i の項目 j に対する順位とすると、

$$d_{ij} = n + 1 - 2R_{ij} \tag{7}$$

で表される数値が入った表である。これは、ある項目が他の残りの項目それぞれと比べて、順位が上であれば +1、下であれば-1 である数値で、いわば他の項目と何勝何敗であったかを表す数値であり、直接演算できない順序尺度データを算術的演算の俎上にのせる方法であるといえるだろう。上の例で言うと、ドミナンス表は以下のようになる (表.11)。

[*1] 被験者 Y1 が、X1,X2,X3 という 3 つの項目を選択した場合、射影される重みは X が作る三角形の重心に来る。

表 11　ドミナンス表

被験者	自民党	民主党	公明党	自由党	社民党	合計
S1	4	2	-2	-4	0	0
S2	2	4	0	-2	-4	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮

これを特異値分解して行・列の両方に重みを与えることになる。このようなドミナンス数の考え方は、例えば一対比較法のデータにも対応できる。一対比較法とは、例えば A,B,C の 3 つのカテゴリがあった場合、A と B を比較してどちらが優位であったか、A と C は、B と C は、というように一対一での優劣（似ている、似ていないとか、好ましい、好ましくないなど次元は何でもよい）を競いあったデータである。このようなデータであっても、ドミナンス表に変換すれば双対尺度法を適用できる。

5.2　強制分類法（Forced Classification）

例えば 2 つの項目、F1,F2 があって、それがそれぞれ「はい」・「いいえ」で答えるデータであったとする。このとき、以下の二つの行列は同じ情報量を持っていることは明らかだろう。

$$[F1, F2, F2, F2, F2, F2] = [F1, 5F2] \tag{8}$$

これは、項目 2 を 5 回繰り返したもので、データ表は以下のようになっている (表.12)。

表 12　繰り返された表

x1	x2	x3	x4
1	0	5	0
1	0	0	5
0	1	0	5
0	1	5	0

一般に、ある項目 j を k 回繰り返したとすると、それは項目 j に重み k を掛け合わせたことと同じになり、繰り返しが ∞ に近づくにつれ、項目 j が双対尺度法で得られる第一次元の解に与える寄与率が 1.00 に収束していく。言い換えれば、項目 j が張る空間を基準にして、他の項目を強制的に布置していることになる (図.2)。

5.3　全情報分析

式 (6) に挙げた双対の関係にあるように，列の最適な重みづけによる平均値は行の重みに特異値をかけたもの，行の最適な重みづけによる平均値は列の重みに特異値をかけたもの，である。つまり双対尺度法によるプロットは，「列空間」に行・列の各カテゴリをプロットするか，「列空間」にそれらをプロットするかのどちらかを選ばなければならない。行の空間と列の空間は異なる縮尺を持つマップなのである。

行変数と列変数を同一の空間にプロットすることを考えるなら，行変数と列変数の空間的隔たりを加味した

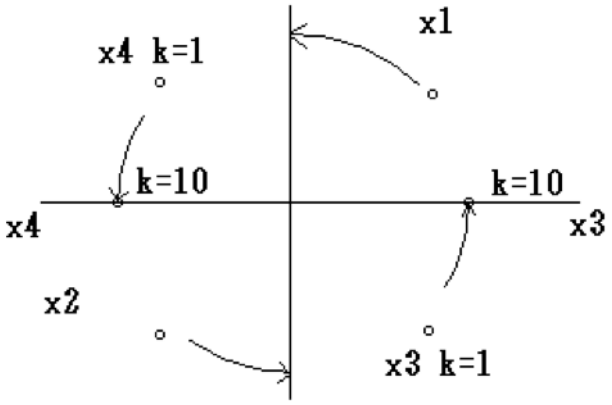

図 2　強制分類法による布置

距離を算出しなければならない。この距離は次式で計算される。

$$d_{ij} = \sqrt{\sum_{k=1}^{K} \rho_k^2 (y_{ik}^2 + x_{jk}^2 - 2\rho_k y_{ik} x_{jk})}$$

また，行変数同士 (y_i, y_{i*}) の距離や，列変数同士の距離 (x_j, x_{j*}) は，

$$d_{ii*} = \sqrt{\sum_{k=1}^{K} \rho_k^2 (y_{ik} - y_{i*k})^2}$$

$$d_{jj*} = \sqrt{\sum_{k=1}^{K} \rho_k^2 (x_{jk} - x_{j*k})^2}$$

で求められる。これらを距離行列 $D_{yy} = \{d_{ii*}\}, D_{xx} = \{d_{jj*}\}, D_{xy} = d_{ij}, D_{yx} = d_{ji}$ からなる拡大行列に含め，

$$D = \begin{bmatrix} D_{yy} & D_{yx} \\ D_{xy} & D_{xx} \end{bmatrix} \tag{9}$$

とすれば，行と列のすべての情報を分析することができる (全情報分析，Total Information Analysis;TIA)。

参考文献

[1] 西里静彦 (2010) 行動科学のためのデータ解析, 培風館.