

空気調和・衛生工学における数学の利用(8)

多変量解析の基礎

平手小太郎 東京大学大学院工学系研究科建築学専攻 正会員

キーワード：数学 (Mathematics), 多変量解析 (Multivariate Analysis)

1. 多変量解析とは

多変量解析 (multivariate analysis) とは、複数の対象 (サンプル) についての複数の変数 (カテゴリー) による測定値 (データ) に対し、これらの変数を独立させず特徴を集約し、所与の目的に応じて各変数間の関連性を分析および総合化する統計的手法の総称である。

どのような対象も多様な特性を有しており、その特徴の真の姿を把握するためには、多様な特性データを同時並行的に測定・解析することが望ましいのはいうまでもない。ただ、多変量解析の基本的な理論は 1950 年以前に確立されていたものの計算能力が伴わず、従来はいずれか一つの変数の特性として個別の解析をせざるを得なかった。1960 年代後半以降のコンピュータの急速な発展に伴い、幅広く使用されるようになったのである。

多変量解析の目的は、大きく、

- 1) 情報を圧縮し事象を簡潔に記述すること
- 2) 要因の効果と影響の強さを推定し、事象を説明・予測すること
- 3) 新たなデータを判別し分類・類型化すること

などに要約できる。また、多変量解析は、

- 1) 外的基準の有無
- 2) 変数の尺度の水準 (質的データか量的データか)
- 3) 変数の数

の三つの基準の組合せによって分類可能である。

以下、目的別に各手法を概観していきたい。

2. 説明予測手法

2.1 重回帰分析

(1) 概要

重回帰分析 (multiple regression analysis) とは、一つの基準変数を複数の説明変数の線形結合によって、説明・予測する方法である。なお、結果に相当する予測される変数を基準変数 (目的変数, 従属変数), 原因に相当する特性に関する変数を説明変数 (予測変数, 独立変数) と呼ぶ。変数の関係の線形性を統計的に確認するだけで、分析上両変数の間の因果関係は存在しなくてもよい。また、手法の手続きによる関係とデータそのものが有する関係をとり違えないで解釈する必要がある。

(2) 手順

重回帰分析の基本モデルは、

$$y = Xa + e$$

$$\hat{y} = Xa$$

ここで、

$y = [y_i \ (i=1, 2, \dots, I)]$: 基準変数 (実測値) ベクトル (I 列)

$\hat{y} = [\hat{y}_i \ (i=1, 2, \dots, I)]$: 基準変数 (予測値) ベクトル (I 列)

$X = [x_{ij} \ (i=1, 2, \dots, I; j=0, 1, \dots, J)]$

ただし、 $x_{0i} = 1$: 説明変数行列 ($I \times (J+1)$)

$a = [a_j \ (j=0, 1, \dots, J)]$: 偏回帰係数ベクトル ($J+1$ 列)

$e = [e_i \ (i=1, 2, \dots, I)]$: 残差ベクトル (I 列)

I: 対象数

J: 説明変数の数

a 重回帰式・偏回帰係数の算出

最小二乗法を用い、予測値と実測値との差の平方和が最小、すなわち、実測値と予測値の相関が最大となるように重回帰式は定式化され、予測式の各説明変数に与えられる重み係数である偏回帰係数 (partial regression coefficient) および定数が求められる。

ここで、予測値と実測値との差の平方和 S は、

$$S = (y - \hat{y})(y - \hat{y})$$

$$= y'y - 2a'X'y + a'X'Xa$$

最小二乗法を用い S を最小とする a を求めるには、 a で偏微分して 0 とおき、

$$(X'X)a = X'y$$

が得られる。これを正規方程式という。

$|X'X| \neq 0$ のとき

$$a = (X'X)^{-1}X'y$$

で、偏回帰係数が求められる。なお、2 変数の場合には、重回帰式が示す直線を回帰直線と呼ぶ。

各説明変数を規準化 (平均 0, 分散 1) した場合の重回帰式の各変数の重み係数を標準偏回帰係数 (standardized partial regression coefficient) といい、この場合、定数項に相当する項は 0 となる。標準偏回帰係数は各説明変数の分散の影響を回避しているので、基準変数に対する各説明

変数の影響の強さが比較できる。

b 決定係数・重相関係数

重回帰式の適合性の良否、精度(信頼性)を表すため、最小二乗法によって求められた実測値と予測値の間の相関係数を重相関係数(multiple correlation coefficient)、その平方値を決定係数(coefficient of determination)という。決定係数は、基準変数の分散のうち説明変数による回帰予測変数が説明する分散の割合を示し、1に近いほど適合性が高い。

$$\hat{y} = X(X'X)^{-1}X'y$$

より、

$$H = X(X'X)^{-1}X'$$

とすると、決定係数 R^2 は、

$$R^2 = (\hat{y}'\hat{y})/(y'y) = (y'Hy)/(y'y)$$

ここで、

R : 重相関係数

重相関係数の大きさは、説明変数と基準変数との相関だけでなく、説明変数間の内部相関にも左右される。また、説明変数の個数が対象の個数よりも多くなると、無条件に重相関係数は1になる。よって、対象数に比べて説明変数の数が多い場合には、重相関係数が過大評価されることになる。このような場合には、自由度調整済み重相関係数(multiple correlation coefficient adjusted for the degrees of freedom)を用いたほうがよい。

c 検定

重回帰式の予測有用性については、予測値の変動を、回帰による変動と誤差による変動(回帰からの変動)に分け、分散分析を用い前者が後者より有意か否かという検定を行う。具体的には、“重回帰式が予測に役立たない(全偏回帰係数=0)”という帰無仮説をF検定⁶⁾で棄却するという手順による。

また、偏回帰係数の有用性(変数選択の際の新たに追加する変数の有用性など)については、誤差項が独立に正規分布に従うという仮定の下で、“(新たに追加する説明変数が予測に役立たない(偏回帰係数=0))”という帰無仮説をt検定⁶⁾によって棄却するという手順をとる。

d 重要な説明変数の探索

重回帰式において、説明変数の選定は最も重要な問題である。経験などにに基づき、研究目的に対し重要な説明変数から順次採択していくことになるが、一応の目安としては、

1) 基準変数と相関が高い

2) 説明変数同士が独立

などが挙げられ、最終的には、標準偏回帰係数の大小などで総合的判断することになる。

e 多重共線性

モデルに含まれるすべての説明変数ベクトルが一次独立でなく線形関係($X'X$ の逆行列が存在しない)が生じ、偏回帰係数が不安定で一定値に定められなくなる状態を多重共線性(multicollinearity)という。説明変数の幾つかの間に強い相関がみられる場合に起こり、得られた推定値の信頼性が低くなる。他の変数と明らかに高い相関を持ついずれか一方の変数を除去する、リッジ推定量⁶⁾⁹⁾を導入して、各変数間の相関を理論的に低下させるなどの工夫が必要となる。

f 適切な予測式

適切な説明変数の探索を行う場合、変数を増減させて分析を行うことがある。この場合、変数追加・削減の打ち切り基準として、自由度調整済重相関係数⁶⁾、自由度二重調整済重相関係数⁶⁾、予測平方和(PSS)⁷⁾C_p統計量、赤池情報量基準(AIC)⁷⁾などが用いられる。変数選択において、これらの指標は、重回帰式に加える説明変数の増減によって極値になったところで最適、すなわち、説明変数の追加・削減を打ち切るようにする。

2.2 数量化I類

(1) 概要

数量化I類(quantification method of the first type)とは、量的変数(間隔尺度以上)の外的基準(基準変数)を、質的変数(名義尺度)の説明変数によって説明・予測するための方法で、形式的には重回帰分析で説明変数が名義尺度からなるダミー変数を用いた場合に相当する。

(2) 手順

数量化I類の基本モデルは、

$$y = Da + e$$

$$\hat{y} = Da$$

ここで、

$y = [y_i] (i=1, 2, \dots, I)$: 外的基準(実測値)ベクトル (I列)

$\hat{y} = [\hat{y}_i] (i=1, 2, \dots, I)$: 外的基準(予測値)ベクトル (I列)

$D = [d_{ij}] (i=1, 2, \dots, I; j=0, 1, \dots, J)$

ただし、 $d_{i0} = 1$: 説明変数ダミー行列($I \times (J+1)$)

$a = [a_j] (j=0, 1, \dots, J)$: 偏回帰係数ベクトル($J+1$ 列)

$e = [e_i] (i=1, 2, \dots, I)$: 残差ベクトル(I列)

I : 対象数

J : 独立したダミー変数の数

外的基準に対する合成変数(各対象ごとに反応したカテゴリーの数値の和)の予測誤差が最小(外的基準と合成変数の相関が最大)になるように、各カテゴリー変数に数値(カテゴリーウェイト、偏回帰係数)を与える。

反応をダミー変数として(0, 1)データとするが、各対象

ごとでみた場合、各アイテム(項目)の全カテゴリーに与えられた数値の和は常に1となるので、各カテゴリーに対応する列ベクトルがそのアイテム内で線型従属となる。よって、一意的には求めることができないため、実際の計算では、各アイテムから一つずつのカテゴリー変数を取り除く。取り除いたカテゴリー変数には重み0を与える。

ここで、説明変数ダミー行列とは、各アイテムごとのカテゴリーへの反応を(カテゴリー数-1)個の独立したダミー変数(0,1表現)に変換し、定数項に相当する列ベクトル(全成分が1)を加え並べて行列にしたもの。 J は独立したダミー変数の数(全カテゴリー数-全アイテム数)となる。

そして、重回帰分析と同様に正規方程式

$$(D'D)a = D'y$$

が得られ、

$$a = (D'D)^{-1}D'y$$

で、偏回帰係数が求められる。

各アイテム内のカテゴリーウェイト、偏回帰係数の最大値と最小値の差をレンジ(範囲)と呼び、各アイテムが外的基準にどの程度寄与しているかの判断基準となる。通常、各カテゴリーの重みを比較するために、予測式としては標準化したカテゴリーウェイトを用いる。予測式の信頼性は、重回帰分析同様、重相関係数により評価できる。

3. 判別手法

3.1 判別分析

(1) 概要

判別分析(discriminant analysis)とは、複数の群に関する既知の多変数データを利用し、対象があらかじめ設定された群のどれに属するのかを判別するための多変数の一次結合による合成変数をつくる手法で、名義尺度で表された基準変数を幾つかの説明変数の一次結合で表現する。所属が不明の対象を想定された群のうちどの群に判別すればよいかを分析することが可能となる。

判別関数(discriminant function)とは、群を判別するため、各群間の平均値の差を最もよく識別できるように構成される合成変数で、2群の判別で分散が等しい場合は判別関数は線形となる。線形判別関数を幾何学的に説明すると、ある直線を引き、この直線上の座標上で完全に2群に分離された場合、この直線と直交する直線上の値を示す関数である。群数を2以上にして分析することも可能である。なお、判別関数に用いる変数の確率分布は正規分布であると仮定されている。

(2) 手順

2群の分散が等しい場合について説明する。

判別対象 x から2群 G_1, G_2 (分散は等しい)へのマハラ

ノビス距離(変数間の相関を考慮した距離)をそれぞれ D_1^2, D_2^2 とすると、

$$D_1^2 = (x - m_1)' V^{-1} (x - m_1)$$

$$D_2^2 = (x - m_2)' V^{-1} (x - m_2)$$

$$m_1 = [m_{1j}] [j=1, 2, \dots, J]: G_1 \text{ 平均ベクトル}(J \text{ 列})$$

$$m_2 = [m_{2j}] [j=1, 2, \dots, J]: G_2 \text{ 平均ベクトル}(J \text{ 列})$$

$$V = [\sigma_{jj}] [j=1, 2, \dots, J]: \text{共分散行列}(J \times J)$$

$$x = [x_j] [j=1, 2, \dots, J]: \text{判別対象ベクトル}(J \text{ 列})$$

$$J: \text{説明変数の数}$$

よって、

$$D_1^2 - D_2^2 = -\alpha(m_1 - m_2)' V^{-1} x \\ - 1/2(m_1 - m_2)' V^{-1} (m_1 + m_2)}$$

であるから

$$z = (m_1 - m_2)' V^{-1} x \\ - 1/2(m_1 - m_2)' V^{-1} (m_1 + m_2)}$$

が線形判別関数となり、

$$z \geq 0 \text{ のとき } D_1^2 < D_2^2 \text{ で } x \text{ は } G_1 \text{ に属する}$$

$$z < 0 \text{ のとき } D_1^2 > D_2^2 \text{ で } x \text{ は } G_2 \text{ に属する}$$

と判別される。

また、相関比(correlation ratio)に基づいた固有値問題を解き、最大固有値に対する固有ベクトルを求める方法もある。

群への所属の情報がない場合は、クラスター分析などを用いて、暫定的に対象を幾つかの群に分類し、その結果に基づき判別分析を適用することも可能である。また、効率的な判別関数を作成するためには、できるかぎり少数で、かつ判別精度を最大にするような変数を選ばなくてはならない。

3.2 数量化Ⅱ類

(1) 概要

数量化Ⅱ類(quantification method of the second type)とは、質的変数(名義尺度)の外的基準(基準変数)を、質的変数(名義尺度)の説明変数によって説明・予測するための方法で、形式的には判別分析で説明変数が名義尺度からなるダミー変数のデータを用いた場合に相当する。

(2) 手順

数量化Ⅰ類の場合と同様に、質的変数の各カテゴリーに数値を与え、各対象ごとに反応したカテゴリーの数値の和を合成変数とする。外的基準に対する合成変数の相関比が最大(群間の差を最大、同一群では合成変数値が類似)になるように各カテゴリー変数に数値を与える。

数量化理論Ⅰ類と同じく、ダミー変数を用い、説明変数ダミー行列 D を作成する。その共分散行列を V とし、判別分析と同じ手順で判別を行う。判別分析と同様、相関比に基づいた固有値問題を解く方法もある。

判別に対する各変数の寄与の大きさは、各変数の各カテ

ゴリーに与えられる数値の範囲(レンジ)や、外的基準と変数との偏相関係数の大小で判断する。ただし、変数間で相関が高いものが存在すれば、寄与を過小評価する変数が存在することになるので注意が必要である。なお、正確な判別、判別の精度を示す基準としては、判別の中率、相関比などが用いられる。

4. 情報集約手法

4.1 主成分分析

(1) 概要

主成分分析(principal component analysis)とは、多数の量的変数(多変数データ)の相互関係を分析し、多数の変数の持つ変動をなるべく保存しながら、また相関のあるデータから冗長な情報を除くことにより簡略化し、互いに無相関の少数の合成変数(主成分という)に要約する手法で、一種の重みつき合計点で、GNPなどの経済指標なども一種の主成分といえる。幾何学的には、分散を最大限保持する形式で、多次元空間のデータをより少ない次元の超平面に射影することにあたる。

(2) 手順

まず、

$$V = (1/I) X X$$

$$R = (1/I) Z Z$$

ここで、

$X = [x_{ij} \mid i=1, 2, \dots, I; j=1, 2, \dots, J]$: データ行列 ($I \times J$)

$Z = [z_{ij} \mid i=1, 2, \dots, I; j=1, 2, \dots, J]$: 規準化データ行列 ($I \times J$)

V : 分散共分散行列 ($J \times J$)

R : 相関行列 ($J \times J$)

で、主成分分析の基本モデルは、

$$p_1 = \mathbf{a}_1 \mathbf{x}$$

ここで、

p_1 : 第1主成分(スカラー)

$\mathbf{a}_1 = [a_{1j} \mid j=1, 2, \dots, J]$: 第1主成分重み係数ベクトル (J 列)

$\mathbf{x} = [x_j \mid j=1, 2, \dots, J]$: 変数ベクトル (J 列)

と置ける。

もとの変数の変動を最もよく説明することが可能な第1の合成変数として第1主成分を定めていく。すなわち、 p_1 の分散 $V[p_1]$ が

$$V[p_1] = V[\mathbf{a}_1 \mathbf{x}] = \mathbf{a}_1 V[\mathbf{x}] \mathbf{a}_1 = \mathbf{a}_1 V \mathbf{a}_1$$

であり、条件 $\mathbf{a}_1 \mathbf{a}_1 = 1$ のもとで p_1 の分散を最大にする \mathbf{a}_1 を求めることになる。

ラグランジュの未定乗数法を用いて、

$$u = \mathbf{a}_1 V \mathbf{a}_1 - \lambda (\mathbf{a}_1 \mathbf{a}_1 - 1)$$

で u を最大、したがって両辺で微分して、

$$\partial u / \partial \mathbf{a}_1 = 2 V \mathbf{a}_1 - 2 \lambda \mathbf{a}_1 = 0$$

$$V \mathbf{a}_1 = \lambda \mathbf{a}_1$$

よって、 \mathbf{a}_1 は V の固有ベクトルとなる。

以下、互いに直交するという条件の下で、残差分散を最も説明することが可能な第2以下の主成分を求めていく。これも同様に、行列 V の固有値問題に帰着する。

なお、主成分 p_k と変数 x_j の相関係数(因子負荷量)は、 $\sqrt{\lambda_k} a_{jk}$ となる。また、

$$\sum_{k=1}^J \lambda_k = J$$

の関係があり、

$$\alpha_k = \lambda_k / J$$

を主成分 p_k の因子寄与率という。

各変数を、平均0、分散1と規準化した分散共分散行列、すなわち相関行列 R の固有値問題とする場合もある。

すなわち、主成分は、分散共分散行列 V あるいは相関行列 R から、固有値の大きいほうから順に互いに直交する固有ベクトル(主成分重み係数ベクトル)を求めることになる。

ただし、分散共分散行列と相関行列に基づく結果は異なる。前者の結果は、測定単位に依存するので、抽出される主成分は各変数の標準偏差の大きさの影響を受ける。後者の場合には、分散の幅が均一化される。よって、変数間の測定単位が異なる場合には、相関行列を用いるのが一般的である。

次に、第1, 第2, ... 第 K 主成分の寄与率の和(累積寄与率) $\sum_{k=1}^K \lambda_k$ が十分に大きくなったところで、主成分の抽出を止める。この判断基準として、

- 1) 主成分の累積寄与率がある数値を超えること
- 2) 各主成分の寄与率がもとの変数の1個分以上あること
- 3) 検定を行い固有値が有意であること

などが目安となる。

4.2 因子分析

(1) 概要

因子分析(factor analysis)とは、多数の量的変数を少数の因子と呼ばれる仮想的・仮説的な潜在変数によって解釈することにより、データの背後に潜む共通性を客観的に明らかにし、多変数の変動を一次結合のモデルによって表すことで情報の集約を図り、データの持つ構造を明らかにしようとする方法である。因子分析によって抽出される共通因子(common factor)と呼ばれる仮説的変数は、説明変数を分解・合成したものである。主として知能、性格の分析などの心理学研究の分野での手法として理論的に発展してきた。

(2) 基本モデル

因子分析モデルとは、多変数データを少数の共通な潜在変数の線形結合として表すモデルであり、基本モデルの式は、

$$\mathbf{Z} = \mathbf{F}\mathbf{A} + \mathbf{U}\mathbf{D}$$

ここで、

$\mathbf{Z} = [z_{ij}]$ ($i=1, 2, \dots, I; j=1, 2, \dots, J$): 規準化データ行列 ($I \times J$)

$\mathbf{F} = [f_{ik}]$ ($i=1, 2, \dots, I; k=1, 2, \dots, K$): 共通因子得点(潜在変数)行列 ($I \times K$)

$\mathbf{A} = [a_{jk}]$ ($j=1, 2, \dots, J; k=1, 2, \dots, K$): 因子負荷行列 ($J \times K$)

\mathbf{U} : 独自因子得点行列 ($I \times J$)

\mathbf{D} : 独自性行列 ($J \times J$ 対角行列)

I : 対象数

J : 変数数

K : 因子数

このモデルでは、観測可能な J 個の変数が、その数も未知の潜在的な共通因子の線形結合によって説明される共通部と、各変数に固有な変動を表す独自因子 (uniqueness) から構成されると仮定している。共通因子にかかる重み係数は、共通因子に対する影響を示し、因子負荷量 (factor loading) と呼ばれる。通常は、この共通因子によってもたらされる変動が主たる解析対象となる。一般的には、各変数は規準化 (平均 0, 分散 1) し、各共通因子、各独自因子はそれぞれ独立であると仮定する。独自因子を仮定せず誤差に含め、変数の変動を最大限説明する成分を抽出すれば、主成分分析と全く同じになる。

モデル式の右辺には、観測可能な変数も既知の定数もなくすべて未知で、因子分析の課題は、幾つの共通因子を設けたらよいか、どのような変数を選ぶのがよいかなど、これらの未知のものを推定することにある。因子負荷が因子の特徴を表しているので、因子分析の解を求めるということは、通常は因子負荷を求めることを意味する。

また、共通因子が互いに直交するように定める直交因子モデルと、この制約を課さない斜交因子モデルとがある。一般的に、前者は簡明で理解しやすい。直交因子モデルの場合には、因子負荷量は重み係数を表すと同時に、当該因子と各変数との相関を表すことになる。なお、それぞれの段階で個別の解法が提案されており、因子分析とはこれらの方法の総称となっている。

(3) 手順

a 相関行列

まず、変数間相関行列を求め、その対角成分に、変数の分散のうち共通因子によって説明される割合である共通性 (communality) と呼ばれる数値を入れる。

$$\mathbf{R} = (1/I) \mathbf{Z}' \mathbf{Z}$$

より、基本式を代入し

$$\mathbf{R} = (1/I) (\mathbf{A}' \mathbf{F}' \mathbf{F} \mathbf{A} + \mathbf{A}' \mathbf{U}' \mathbf{D} \mathbf{U} \mathbf{A} + \mathbf{D}' \mathbf{U}' \mathbf{U} \mathbf{D})$$

は、 \mathbf{I} を単位行列として、

第 1 項: $(1/I) \mathbf{F}' \mathbf{F} = \mathbf{I}$ (直交解の場合、因子間の相関行列)

第 2 項: 0

第 3 項: 0

第 4 項: $(1/I) \mathbf{U}' \mathbf{U} = \mathbf{I}$

となり、

$$\mathbf{R} = \mathbf{A}\mathbf{A}' + \mathbf{D}^2$$

$$\mathbf{R}_c = \mathbf{R} - \mathbf{D}^2 = \mathbf{A}\mathbf{A}'$$

\mathbf{R}_c : 相関行列 \mathbf{R} の対角成分に共通性を入れた行列 ($J \times J$)

b 共通性の推定

因子解を求めるためには、未知である共通性を知る必要がある。よって、

1) ある変数と他の変数との重相関係数の平方値 (SMC)

2) ある変数 j と他の変数との相関係数の絶対値のなかで最大の値

3) 何らかの初期値から始め、逐次的方法での推定値などの推定値を用い共通性を推定する。

c 因子解

次に、因子解を求める。すなわち因子数を決め、因子負荷量を推定することになる。共通因子の定め方は、互いに独立であればどのようにとってもかまわないので、理論的には無数に存在する。因子解を求める方法には、主因子法、セントロイド法、直接バリマックス法など各種あるが、基本的な相違は共通因子軸の設定法にある。代表的なものを以下に示す。

i) 主因子解

主因子解 (principal factor solution) は、因子解として最も基本的なもので、回転解を求めるときの初期解としてもよく用いられる。多変数の間で共通にみられる変動のうち、いずれの変数に対しても近い変動を表すものを因子として取り出す方法で、共通因子を求める手順は、主成分分析法と同じである。

第 1 因子の因子寄与 (因子負荷量の平方和) を最大にする解を求め、第 2 因子以下を第 1 因子と直交するという条件で因子寄与が最大となるように定めていくものである。

すなわち、

$$\mathbf{R}_c = \mathbf{A}\mathbf{A}'$$

この条件の下で、因子寄与を最大にする解は、ラグランジュの未定乗数法を通して、

$$R \mathbf{a}_1 = \lambda_1 \mathbf{a}_1$$

ここで,

λ_1 : 第1因子負荷量の平方和 ($\lambda_1 = \mathbf{a}_1' \mathbf{a}_1$)

\mathbf{a}_1 : 第1因子負荷量を成分に持つベクトル (J 列)

となる。すなわち, R の固有値が λ_1 , 固有ベクトルが \mathbf{a}_1 となる。第2因子以下は同様に R の固有値問題となり, 大きい固有値に対する \mathbf{a} から順次, 第1因子, 第2因子と定める。

なお, 第 k 因子と変数 x_j の相関係数 (因子負荷量) は, a_{jk} となる。

$$\sum_{k=1}^J \lambda_k = J$$

の関係があり,

$$c_k = \lambda_k / J$$

を第 k 因子の因子寄与率という。

ii) 直接バリマックス解

直接バリマックス解 (direct varimax solution) は, 単純構造を満たす解を直接求める方法で, 潜在因子をみつけるのは優れている。他の解法による因子よりも, 意味づけしやすい。

iii) 重心解

重心解 (centroid factor solution) は, 計算が比較的容易で, かつては初期解として最もよく用いられた。

d) 因子数の決定

一般的に, 多数の共通因子に科学的な意味づけを与えることには無理があるため, 適当なところで, 新しい因子の抽出をうち切る必要がある。その基準としては,

- 1) 累積因子寄与率が一定値 (95% など) を超えた場合
- 2) 因子負荷量の平方和が1以下になった場合

などが挙げられる。

ここで, 因子寄与率とは, ある因子についてすべての変数の因子負荷量の平方和をすべての変数が有する分散で除した値である。

e) 回 転

因子負荷量を下に, 因子を解釈することになるが, 解釈をしやすくするために, 直接解によって定められた因子を線形変換 (座標軸の回転) することがある。そのために, 通常はバリマックス回転を介し, 単純構造を目指すことになる。単純構造とは, 因子がそれぞれ一部の変数とのみ高い相関を持ち, その他の変数との因子負荷量が0に近くなるように因子軸を設け, 変数と因子との関係を単純化することである。

i) バリマックス回転

バリマックス回転 (varimax rotation) とは, 因子ごとに因子負荷量の2乗値の分散を求め, これの和を最大とするという基準で直交回転を行う方法である。因子負荷量の正負にかかわらず, 因子負荷の絶対値の大きいものが多くま

とまっているような位置に因子軸を回転することに相当する。単純構造への因子軸の回転をする方法のなかで最も代表的なものである。

f) 因子得点

因子負荷量を求めることにより, 目的を達することもあるが, 場合によっては, 各対象ごと共通因子の値, 因子得点を推定し, 個体の分類などに利用することもある,

因子負荷が定められたとしても, 独自性に関する未知数があるため, 因子得点は確定できず, 因子負荷量とものデータとから推定することになる。

幾つかの推定式が提案されているが, 最も代表的な推定式は, モデル上の因子得点と推定された因子得点の差の平方和を最小にするもので,

$$F = ZR^{-1}A$$

である。

4.3 数量化Ⅲ類

(1) 概 要

数量化Ⅲ類 (quantification method of the third type) とは, サンプル (対象) のカテゴリー (特性項目) への反応パターン類似性に基づいて, サンプルとカテゴリーの両方について同時に数量化を行い, 適切な数値を与える方法である。因子分析は定量的データを扱うのに対して, 数量化Ⅲ類は定性的データを数量化して扱うことになる。

(2) 手 順

数量化Ⅲ類では, 反応パターンにおいて, 各サンプルがあるカテゴリーに反応したとき "1", 反応しないときには "0" というダミー変数を導入する。反応パターンが類似したサンプル, カテゴリーが近くなるような並び替えを想定し, この並びに対応し相関係数が最大となるように, サンプル, カテゴリーに数値を与える。そして順次, 別の並び替えに対応する第2, 第3の相関係数 (極値) に対する数値を求めていく。

カテゴリーに対し,

$$\mathbf{x} = [x_j \mid j=1, 2, \dots, J]$$

サンプルに対し,

$$\mathbf{y} = [y_i \mid i=1, 2, \dots, I]$$

を与えるとして, まず, 各サンプル, 各カテゴリーおよび全体の正反応数を,

$$m_i = \sum_{j=1}^J \delta_{ij} \quad (i=1, 2, \dots, I)$$

$$n_j = \sum_{i=1}^I \delta_{ij} \quad (j=1, 2, \dots, J)$$

$$t = \sum_{i=1}^I \sum_{j=1}^J \delta_{ij}$$

δ_{ij} : サンプル i のカテゴリー j に対する反応 (ダミー変数)

とし,

$$A = [a_{ik}], \quad B = [b_{jk}]$$

ここで,

$$a_{ik} = \sum_{j=1}^J \delta_{ij} \delta_{kj} / (n_j \sqrt{m_k} \sqrt{m_i})$$

$$b_{jk} = \sum_{i=1}^I \delta_{ij} \delta_{ki} / (m_i \sqrt{n_k} \sqrt{n_j})$$

を算出し、固有方程式

$$A\mathbf{u} = r^2 \mathbf{u}$$

$$B\mathbf{v} = r^2 \mathbf{v}$$

ここで、

$$\mathbf{u} = [\sqrt{m_i} x_i] \quad [i = 1, 2, \dots, I]$$

$$\mathbf{v} = [\sqrt{n_j} y_j] \quad [j = 1, 2, \dots, J]$$

を解くことで、 \mathbf{x} 、 \mathbf{y} が数量化される。原理的に最大固有値が必ず1となりこの解には意味はない。第2以下の固有値 r^2 は相関係数の2乗となり、第2以下の固有値に対する固有ベクトルが相関を最大(極大)にする数値を与えることになる。なお、カテゴリー数がサンプル数より大きい場合には、データ行列の行と列を入れ換えて計算する。

(3) 対応分析

対応分析(correspondence analysis)とは、クロス表の反応度数のデータ形式にも対応できる方法で、数量化Ⅲ類を含んだ一般的なモデルになっている。双対尺度法(dual scaling)も同様の方法である。

5. 分類・類型化手法

クラスター分析

(1) 概要

クラスター分析(cluster analysis)とは、対象間の距離(distance) (相違性)あるいは類似度(similarity)の指標が定量的に測定される場合、それに基づいて、似ている対象をまとめ、多数の対象を幾つかのまとまり(クラスター)に分割し、全対象を客観的に分類し、類型化を行う手法の総称である。この手法の出発点は、植物の系統分類などで、生物の特徴を数値データ化して取り扱う数値分類法である。

クラスター分析は、変数の正規性や線形関係などの仮定が必ずしも必要ではなく、対象間の関係が数値で与えられるだけで分類が可能なので、応用範囲は極めて広い。なお、対象間の距離や類似性の与え方、クラスター作成手順には数多くの手法があり、クラスター分析とはその総称である。

(2) 階層的手法と非階層的手法

a 階層的手法(hierarchical method)

最も広く用いられており、クラスター化の過程が樹形図(dendrogram)で示される。与えられた対象・クラスター間のすべての組合せについて距離を計算し、最も近接した二つの個体(またはクラスター)を一つの新しいクラスターに集約していく過程を順次繰り返していく方法で、計算アルゴリズムが容易である。ただし、得られたクラスターの解が不安定になることもあり、必ずしも分離のよいクラスターが得られるという保証はない。

b 非階層的な方法

あらかじめ同一のクラスター間の個体間分散を小さく、異なったクラスター間の個体間分散を大きくするような最適化基準を設定して、対象を並列的に幾つかのクラスターに分類する方法である。なお、個体・クラスター間の距離、相互関係を明確に理解するためクラスター分析に多変量解析の技法を併用することも多い。

(3) 手順

a 対象の視覚的な把握

対象が何らかの手続きによって二～三次元空間内に布置される場合には、視覚的な分布特性の把握は重要である。類似度・距離の定義の検討にも有効といえる。

b 対象間の距離あるいは類似度の定義・数量化

類似度として求められた数値は、逆数にしたり変換式により距離に変換することになる。分類すべき意図を明確にし、目的に適した特性(変数・測定項目)の選定を行う。多変数で説明される場合、対象間の相関係数を類似度として分析することも多い。

分類の成否は、この定義・数量化に大きく左右される。よって、距離、類似度は対象間の特質を損なわない適切なものを用いる必要がある。また、因子得点、主成分得点など直交する合成変数を利用し、変数間の相関による影響を取り除き次元を縮小する方法もある。

特性が名義尺度で与えられている場合は、類似度として特性の一致数と相違数などにより算出する。順位尺度の場合には、Spearmanの順位相関係数やKendallの順位相関係数などを用いる。

c クラスター化(以下、階層的な方法の手順)

- 1) 最も近接している対象同士をクラスター化し、作成したクラスターと他の対象および他のクラスターとの距離を与える。
- 2) 新たな距離を用いて最も近接している対象もしくはクラスター同士をクラスター化する。
- 3) 最終的に全体が一つのクラスターになるまで、操作を繰り返す。

という手順をとる。

距離の定義やクラスター間の距離の決め方には、以下に示すいくつかの考え方がある。

i) 最短距離法

最短距離法(nearest neighbour method)では、二つのクラスター間の距離をそれぞれのクラスターに含まれている最も近い対象の間の距離とする。一つのクラスターが極端に大きくなったり、逆にある対象がクラスターを形成せずに単独で残ってしまうなど、全体のバランスが悪くなることがある。

ii) 最長距離法

最長距離法(furthest neighbour method)では、二つのクラスター間の距離をそれぞれのクラスターに含まれている最も遠い対象の間の距離とする。対象が単独で残ってしまうことは少なく、クラスターの広がり大きさが極端に違わないように分類される。

iii) メジアン法

メジアン法(median method)では、二つのクラスター間の距離を二つのクラスター間の最短・最長2点の平均とする。

iv) 重心法

重心法(centroid method)では、二つのクラスター間の距離を二つのクラスター間の重心間の距離とする。各クラスターに含まれる対象数を考慮して距離を定義している。個々のクラスターの広がり大きさはまちまちであるが、各々のクラスターの重心は空間的に均等に分布する。

v) 群平均法

群平均法(group average method)では、二つのクラスター間の距離を各クラスターを構成する対象間の距離の2乗の平均とする。球形にまとまった分布となる。

vi) ウォード法

ウォード法(minimum variance method)では、クラスターの重心まわりの偏差平方和が最小になるように、他のクラスターと融合する。対象が密集しているところから、クラスターが形成され、広がり大きさは必ずしも均等にはならない。

d) クラスター数の判定

クラスター数の決定に関しては、一般的な判断基準はないが、分析結果と実際面との照合をしながら、情報が最も多く得られる分類数、全体の対象の把握が容易な数(10以下程度)などを目安とし、クラスター化を終了する。また、樹形図はクラスターリングの過程を示しているため、分類過程でも分類規準などの情報が得られ、各クラスター同士の関連性を検討することが可能となる。

6. 親疎表示手法

多次元尺度構成法

(1) 概要

多次元尺度構成法(multidimensional scaling, MDS)とは、標本間の距離が非類似度に最もよく一致するように、かつできるだけ次元の少ない多次元空間(multidimensional space)で表現できるように、各標本の親疎を位置づけ(座標決定)するための方法で、非類似度の大きい対象同士は遠くに、非類似性の小さい対象同士は近くに配置され

るように、全体の対象の布置を定めるものである。

対象を(非)類似度に応じて空間的に表現することにより、対象間の背後に潜んでいる構造、すなわち対象間の(非)類似度を規定している要因を視覚的に明快な形で提示する。なお、因子分析、主成分分析、数量化Ⅲ類なども広い意味での多次元尺度法である。

(2) 計量的MDSと非計量的MDS

対象間の距離が間隔尺度以上か順序尺度以下によって、計量的と非計量的MDSとに二分される。

a) 計量的多次元尺度構成法

計量的多次元尺度構成法(metric MDS)は、対象間の距離が間隔尺度以上であることが明確なデータに対して適用される。対象間の距離を求め、原点からの対象点までのベクトルの内積を算出し、これを要素とする行列の固有分解から潜在空間内の布置を求める、というTorgersonの方法が代表的である。このほか、INDSCAL(Carroll & Chang)、数量化Ⅳ類、K-L型数量化(林)などがある。

b) 非計量的多次元尺度構成法

非計量的多次元尺度構成法(nonmetric MDS)は、対象間の心理的距離が順序尺度以下であるようなデータに対して適用される。Shepardの方法、MDSCAL(Kruskal)、SSA(Guttman)、ALSCAL(Takane, Young & de Leeuw)などがある。

参考文献

- 1) 奥野忠一ほか著：多変量解析法(1971)、日科技連出版社
- 2) 河口至商：多変量解析入門Ⅰ(1973)、森北出版
- 3) 林 知己夫ほか編：多次元尺度解析法(1976)、サイエンス社
- 4) 芝 祐順：因子分析法第2版(1979)、東京大学出版会
- 5) 竹村伸一編著：システム技法ハンドブック(1981)、日本理工出版会
- 6) 芝 祐順ほか編：統計用語辞典(1984)、新曜社
- 7) 日本建築学会編：建築・都市計画のための調査・分析方法(1987)、井上書院
- 8) 池田 央編：統計ガイドブック(1989)、新曜社
- 9) 杉原敏夫ほか著：多変量解析(1998)、牧野書店
- 10) 朝野熙彦：入門多変量解析の実際(2000)、講談社サイエンスフィク

(2003/5/22 原稿受理)



平手小太郎 ひらてこたろう

生年月日 昭和29年12月24日/最終学歴 東京大学建築学科卒(昭和53年卒)/専門 建築光・視環境、建築環境心理/学位 工学博士