2次合成変量のアルファ係数の和を最大化する 多集合データの階層的主成分分析 I

村 上 隆
(名古屋大学)

Hierarchical component analysis of multiset data
maximizing the sum of alpha coefficients of second order composites I

Takashi MURAKAMI
(Nagoya University)

A component analytical model treating the several sets of variables is proposed in this paper. The model has the hierarchical property; The first order composites are defined as linear combinations of variables for each set, and the second order composites are constructed as linear combinations of the first order ones of all the sets. The optimization criterion is the maximization of the sum of alpha coefficients of the second order composites.

The model may be regarded as a natural extension of the ordinary component analysis in Lord's(1958) sense. Through changing the number of the second order composites, the first order composites could be explored in such a way that the latter would be well be well balanced between the internal consistency coefficients and the correlations between the sets.

For this study, an alternating optimization algorithm of the model was formulated as a first part, but the need for examining relationship with other models and demonstrating numerical examples based on this model was stressed for the next study.

1. 序 — 問題の所在

本研究は、質問紙を用いて複数の(心理学的)構成概念を測定し、概念間の関係を追求しようとするタイプの研究において役立つような、データ分析の方法について検討することを目的とする。このような研究では、一つの質問紙に含まれる項目全体を、明確に区別される内容領域に対応した複数の集合に分割することができる。このように、幾つかの項目の集合からなる質問紙によって得られるデータを多集合データと呼ぶ。例えば、質問紙が全体として2つの部分に分かれ、第1部は、性格特性に関連する項目群からなり、第2部は、職業興味に関わる項目群からなるとすれば、1つの多集合データが得られることになる。

多集合データの分析にあたっては、通常まず各集合の変数を何らかの形で合成して、内容領域の概念に対応する(一般には集合ごとに複数の)尺度を構成する。次に、それらの尺度の間の相関係数を計算することによって、概念間の関係を検討する。

ここで、研究として主要な部分は第2の段階である。

すなわち、2つ、またはそれ以上の概念間の関係を明らかにすることが、一般に「相関研究」と呼ばれるタイプの研究の主要な目標である。例えば、先の例では、外向的性格の個人が、対人・社会的職業に興味をもつ傾向があるか否かが検討される、といったことがあろう。もし、単一の項目への反応の意味が極めて明確であり、かつ信頼性も高ければ、ここで言う集合は、それぞれただ1つの項目だけを含めば足りることになり、特に項目の集合という概念を導入する必要はなくなる。しかし、質問紙を用いた研究においては、多くの場合、個々の項目水準の反応は、必ずしも高い信頼性をもたないから、多少とも意味の似た複数の項目に対する反応の合成・縮約は必須である。また、一般に社会科学で扱われる概念なるものは、必ずしも唯一絶対の測定操作に対応する(あるいは、複数の測定操作が存在するにしても、同一の結果に収束する)わけではなく、多面的な含意をもつ抽象的なものである(例えば、ラザーズフェルド1984)。実際、先にあげた、性格特性としての「外向性」にせよ、職業興味の次元としての「対人的・社会的職業」にせよ、これらを測定する唯一の項目が存在するとは考えられない。

(信頼性の問題は無視したとしても。)換言すれば、多くの相関研究でとりあげられる「概念」は、比較的多数の項目によって間接的に定義され、その定義には常に任意性が残る、ということである。この意味で、社会科学における「量」の「測定」は、少数の基本次元が確定している自然科学におけるそれとは、決定的に異なる。

したがって、質問紙を用いた相関研究には、どうしても第1の段階が必要となる。この段階では、しばしば(主成分分析を含む広い意味での)因子分析が用いられる。多くの場合、バリマックス回転された各因子に高く負荷する項目を選択し、それを単純加算する形で尺度が構成される。一般に、同一の因子に高い負荷を示す項目間の相互相関は高いと考えられるから、このやり方によって、内的整合性の高い尺度が構成されることが期待される。内的整合性は、一定の条件の下で信頼性と同義である(例えば、池田, 1973) から、このやり方は、信頼性の高い尺度を構成する方法であるとも言える。

しかしながら、この方法では、尺度と他の集合に属する別の尺度との相関には関心が払われていない。すなわち、尺度の妥当性はこの方法ではほとんど考慮されていない。

もう少し具体的に考えてみよう。例えば、前述のように、集合1は、被験者の性格特性を測定するための項目

群からなり、集合2は、被験者の職業興味を測定するための項目群からなるとしよう。話を簡単にするために、集合1で測定される性格特性は「向性」ただ1つとし、集合2で測定される職業興味も「対人・社会的職業」に対する興味のみであるとする。集合1の各項目に対する反応は、外向的なほど大きい値に、集合2の各項目に対する反応は、興味があるほど大きい値になるように、それぞれコーディングされるものとする。

集合1、集合2ともそれぞれ8つの項目からなるものとし、これらすべての項目間の相関が、表1のように与えられたとする。それぞれの集合の項目は4つずつ2つのグループに分かれるとみることができる。第1のグループ(項目1~4)は、相互相関は0.18とさほど高くはないものの、他方の集合の項目群ともある程度の相関(0.15~0.18)をもっている。第2のグループ(項目5~8)は、相互相関は0.25と第1のグループより高いが、他方の集合の項目との相関は概して低い(0.00~0.15)。

このデータを集合ごとに主成分分析すると、結果として得られる構造ベクトル(因子負荷ベクトル)は、表2のようになる。(どちらの集合も同じ。)当然、負荷の値は、集合内の項目との相互相関の高い第2のグループ(0.65)の方が第1のグループ(0.58)よりも大きい。この結果にしたがって、第2のグループの項目だけを選

表1 仮想的な多集合データの相関行列

		集 合 1 (向 性)								集 合 2 (職業興味)							
		1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
集 合 1	1	100	18	18	18	15	15	15	15	18	18	18	18	15	15	15	15
	2	18	100	18	18	15	15	15	15	18	18	18	18	15	15	15	15
	3	18	18	100	18	15	15	15	15	18	18	18	18	15	15	15	15
	4	18	18	18	100	15	15	15	15	18	18	18	18	15	15	15	15
	5	15	15	15	15	100	25	25	25	15	15	15	15	00	00	00	00
	6	15	15	15	15	25	100	25	25	15	15	15	15	00	00	00	00
	7	15	15	15	15	25	25	100	25	15	15	15	15	00	00	00	00
	8	15	15	15	15	25	25	25	100	15	15	15	15	00	00	00	00
集 合 2	1	18	18	18	18	15	15	15	15	100	18	18	18	15	15	15	15
	2	18	18	18	18	15	15	15	15	18	100	18	18	15	15	15	15
	3	18	18	18	18	15	15	15	15	18	18	100	18	15	15	15	15
	4	18	18	18	18	15	15	15	15	18	18	18	100	15	15	15	15
	5	15	15	15	15	00	00	00	00	15	15	15	15	100	25	25	25
	6	15	15	15	15	00	00	00	00	15	15	15	15	25	100	25	25
	7	15	15	15	15	00	00	00	00	15	15	15	15	25	25	100	25
	8	15	15	15	15	00	00	00	00	15	15	15	15	25	25	25	100

(小数点省略)

択し、これら4項目の単純和の形で尺度を構成したとすると、その α 係数は、(全項目の分散は同一であるとして、) 0.57となるが、2つの集合間の相関係数は0となる。一方、主成分分析の結果にこだわらず、第1のグループの項目の単純和で尺度を構成したとすれば、 α 係数は、0.47と低下するものの、集合間の相関は、0.47にまで上昇する。すなわち、この場合、主成分分析にしたがって機械的に項目選択を行うと、「外向性」と「対人・社会的職業への興味」とのあいだには相関がない、という結論になり、集合間相関にだけ注目して項目選択をおこなえば、かなりの相関があるという結論を得ることになる。

表2 因子負荷

	I	h^2
1	58	34
2	58	34
3	58	34
4	58	34
5	65	42
6	65	42
7	65	42
8	65	42
2乗和	304	304

このデータは人工的なものであり、結果は誇張されたものとなっているが、実際のデータにおいてもこれに近い現象は起こり得る。ここで、第1のグループのような項目群は、内容的に極めてよく似た、結果的には、測定しようとする概念の一側面だけに対応するようなものになっていることが多い。このような項目からなる尺度は内的整合性は高いが、他の変数との相関は低いものとなりがちである。逆に、第2のグループのような項目は、概念の多くの側面に対応する、悪く言えば意味のあいまいなものであることが多い。しかし、この種の項目からなる尺度は、内的整合性は低い、他の特性を予測するためにはむしろ有用なものであることがある。

Cronbach (1970) は、内的整合性にあたる性質を fidelity、もう1つの内容の広がりに対応する性質を bandwidth と呼び、この2つが両立し難いこと (bandwidth-fidelity dilemma) を論じている。単一の集合に適用される通常の項目分析の手続き、あるいは因子分析の方法は、前者のみを専ら数値的に評価するものであり、後者の観点を欠いている。したがって、それについては、研究者の直感に依拠して項目選択が行われている

と思われる。

複数の集合を同時に扱うことにより、この問題はある程度解決される。村上 (1985) は、多集合データのための探索的因子分析の一方法である多集合因子分析 (multiset factor analysis) を提案した。この方法は、村上・斎藤 (1986) において、組織イメージと組織行動の関連の分析に、斎藤・村上・若林 (1986) において、大学生の職業志向性、ライフ・スタイルと就職希望の企業との関連の分析にそれぞれ応用された。

村上 (1986) では、同様のデータに対する主成分分析の方法として多集合データのための階層的成分分析 (hierarchical component analysis) が提案された。主成分分析には、(共通性推定を含む) 因子分析のように、因子得点の不定性 (Guttman, 1955) がなく、因子得点 (正確には主成分得点) が一義的に定まるという利点がある。また実用上は、因子得点が正確に分散1、相互相関0という性質をもつことが、結果の取り扱い上便利であると言える。さらに、この方法では、bandwidth-fidelity dilemma について、ある程度意図的な調整が可能である。すなわち、bandwidth と fidelity は、原理的に両方を最大限に満足することはできず、どこかの点で妥協せざるを得ないが、階層的な主成分分析では、二次主成分の数を変化させることにより、その間のバランスを変化させることができる。しかしながら、階層的な主成分分析は、主成分分析のもつこの特徴をかならずしも十全に生かしたものとはいえなかった。

階層的な主成分分析のモデルは、

$$\mathbf{Z}_k = \mathbf{A}_k \mathbf{C}_k \mathbf{F} + \mathbf{E}_k \quad k = 1, \dots, p$$

$$\left(\text{ただし, } \mathbf{C}_k \mathbf{C}_k' = \mathbf{I}, \quad \frac{1}{N} \mathbf{F} \mathbf{F}' = \mathbf{I} \right)$$

$$(1-1)$$

のように書かれる。ここで、 $\mathbf{Z}_k$ は、項目 $\times$ 被験者という形の集合 k のデータ行列、 $\mathbf{A}_k$ は、集合 k の因子負荷行列、 $\mathbf{C}_k$ は、集合 k の高次因子負荷行列、 $\mathbf{F}$ は、(高次の) 因子得点行列である。また、 $\mathbf{E}_k$ は、誤差行列で、アルゴリズムは、この行列の要素である誤差の2乗和、 $\sum \text{tr } \mathbf{E}_k \mathbf{E}_k'$ を最小化するように構成されている。ただし、 N は被験者の数とし、データ行列 $\mathbf{Z}_k$ の各列の和は0、分散は1となるように前処理済みとする。

このモデルでは、各集合における因子得点を、

$$\tilde{\mathbf{F}}_k = \mathbf{C}_k \mathbf{F} \quad k = 1, \dots, p$$

によって定義した。この $\tilde{\mathbf{F}}_k$ は、

$$\mathbf{A}_k = \frac{1}{N} \sum \mathbf{Z}_k \tilde{\mathbf{F}}_k' \quad k = 1, \dots, p \quad (1-2)$$

を満足する, すなわち, 素データとの相関係数が, (その集合の) 因子負荷と一致する, という意味で, 通常の伝統的な因子分析における因子得点と同じ意味をもつ。しかしながら, この集合ごとの因子得点 $\tilde{F}_k$ は, 集合 k のみでなく, 全データの線型結合として定義されている点が問題になる。実際, 異なる集合間の因子得点間相関行列,

$$\frac{1}{N} \tilde{F}_k \tilde{F}_l' \quad (k \neq l) \text{ の要素は, } \tilde{F}_k \text{ と } \tilde{F}_l \text{ が共通データ}$$

の線型結合であるために, 過大な方向に「偏る」傾向がある。因子得点間相関は, モデルのパラメータとして定義されているわけではないから, 「偏り」というのは, 数理統計学的な意味で正確な言い方ではない。しかしこれが, 異なる領域の概念間の関係を検討しようとする目的には, 必ずしも適切でない性質であることは確かである。

そこで本研究では, 多集合データのための主成分分析の方法として, 集合ごとの主成分得点が, その集合のデータだけの線型結合として算出されるような方法を考える。すなわち, 被験者 s の集合 k における項目 i に対する反応を x_{isk} とするとき, まず, 集合ごとの合成得点を,

$$f_{ask} = \sum_{i=1}^{n_k} v_{iak} x_{isk} \quad (1-3)$$

によって求める。これを, 集合 k の a 番目の1次合成変量 (first order composite) と呼ぶ。ここで, v_{iak} は, 集合 k の a 番目の1次合成変量を求めるために, 項目 i にかかる重み, n_k は集合 k の項目数である。この1次合成変量は, 当該の項目集合のみにもとづいて算出されており, その集合間相関には共通部分の存在による「偏り」がない。

次に, 1次合成変量の線型結合として, 2次合成変量 (second order composite),

$$g_{bs} = \sum_{k=1}^p \sum_{a=1}^{q_k} w_{abk} f_{ask} \quad (1-4)$$

を定義する。ここで w_{abk} は, b 番目の2次合成変量を求めるために, 集合 k の a 番目の1次合成変量にかかる重みであり, q_k は集合 k の1次合成変量の数である。2次合成変量は, 項目集合を越えたデータ全体によって算出される。本研究で提案する方法は, この2次合成変量の α 係数の和を最大化するように, 重み v_{iak} と w_{abk} を求めようとするものである。

ここで, 2次合成変量の数 Q を増減することにより, bandwidth と fidelity の間の調整が可能となる。すなわち, Q を小さくすれば, 少数の2次合成変量の α 係数を大きくするためには, 異なる集合の共通成分を最大限に抽出せねばならず, そのためには, 1次合成変量の

重みを, 集合間で相関が高くなるように決めざるをえない。他方, Q を大きくすれば, 多数の2次合成変量の α 係数を全体として大きくするためには, 各集合の「特殊成分」を抽出してくる必要があり, 結果的に, 集合間の相関は低下するであろう。

本研究は, 紙面の都合で2つの部分に分けて公表される。第I部である本論文は以下, 2節において, α 係数の和の最大化の原理について説明し, 3節において, 本研究のモデルについて詳述する。4節では, アルゴリズムを扱う。第II部は, ここで提案するモデルと, 関連する他のモデルとの関係を扱う5節, 人工データによる, モデルの性質の検討を行う6節, 実際のデータへの適用例を示す7節, 残された問題を扱う8節からなる。

以下第2節では, 単一のデータ集合における主成分分析を,

$$f_{as} = \sum_{i=1}^n v_{ia} x_{is} \quad (1-5)$$

という合成変量を求める方法として論じ, それを通じて, 主成分分析の定式化の意味について再検討する。

2. α 係数の和の最大化としての主成分分析

ここでは, 主成分分析を, データの線型結合として, 何らかの意味で最適化された合成変量を導く方法として考えるわけであるが, その際, 合成変量を標準得点行列 Z からではなく, 偏差得点行列から定義するものとする。すなわち, $n \times N$ の偏差得点行列 X は,

$$X1 = 0 \quad (2-1)$$

を満足する, すなわち, X の各行は平均値が0となるように基準化される。(分散は1とは限らない。) したがって,

$$S = \frac{1}{N} XX' \quad (2-2)$$

は項目間共分散行列である。

そこで, $n \times q$ の重み行列を V として, $q \times N$ の合成変量行列 F は,

$$F = V' X \quad (2-3)$$

によって定義される。この F の要素が, 式 (1-5) の f_{as} である。 F の各列は平均0, 分散1で相互に直交することを要請する。すなわち, (2-3), (2-2) によって,

$$\frac{1}{N} FF' = V' SV$$

であるから,

$$V' SV = I \quad (2-4)$$

が制約条件となる。(ここでは、通常の変量解析のテキストと少し違った条件での定式化を目指している。通常の定式化との関係は5節で述べる。)

ここで、(2-4)を満足するような $\mathbf{V}$ は無数に存在するから、何らかの最適化基準を設けて解を一義的にする必要がある。しばしばとられる基準は、素データと合成変量との相関係数、すなわち負荷、の2乗和の最大化というものである(芝 1975がほぼこの基準によっている)。負荷行列は、行例 $\mathbf{D}$ を個別変数の分散を要素とする対角行列、すなわち、

$$\mathbf{D} = \text{diag } \mathbf{S} \quad (2-5)$$

と定義するとき、負荷行列は、

$$\mathbf{A} = \mathbf{D}^{-\frac{1}{2}} \mathbf{S} \mathbf{V} \quad (2-6)$$

のように定義されるから、 $\text{tr } \mathbf{A}' \mathbf{A}$ が最大化すべき基準となる。

しかしながら、この定式化は、後に定義するような多集合の場合に一般化することが難しい。そこで、多集合への一般化が容易な別の定式化を探究する必要がある。

Lord (1958) は、一般化された α 係数を最大化する重み $\mathbf{V}$ が、相関行列の最大固有値に対応する固有ベクトルに $\mathbf{D}^{\frac{1}{2}}$ を掛けたものになることを示した。ここでは、これを手掛かりに主成分分析の新しい定式化を考えてみよう。

まず、 $\mathbf{f}$ を n 個の変数の単純和、

$$\mathbf{f}' = \mathbf{1}' \mathbf{X} \quad (2-7)$$

で定義される合成変量ベクトルとすると、その α 係数は、

$$\alpha = \frac{n}{n-1} \left(1 - \frac{\mathbf{1}' \mathbf{D} \mathbf{1}}{\mathbf{1}' \mathbf{S} \mathbf{1}} \right) \quad (2-8)$$

によって定義される。ここで、 $\mathbf{1}' \mathbf{D} \mathbf{1}$ は、全項目の分散の和であり、 $\mathbf{1}' \mathbf{S} \mathbf{1}$ は合成得点の分散であることに注意すれば、これが Cronbach (1951) による α 係数のオリジナルな定義、

$$\alpha = \frac{n}{n-1} \left(1 - \frac{\sum s_i^2}{s^2} \right)$$

と等価であることは理解されよう。つぎに、 $\mathbf{v}$ を一般的な重みベクトルとして、

$$\mathbf{f}' = \mathbf{v}' \mathbf{X} \quad (2-9)$$

という形で、合成変量ベクトル $\mathbf{f}$ を定義すれば、この α 係数は、(2-8)の $\mathbf{1}$ を $\mathbf{v}$ におきかえた、

$$\alpha = \frac{n}{n-1} \left(1 - \frac{\mathbf{v}' \mathbf{D} \mathbf{v}}{\mathbf{v}' \mathbf{S} \mathbf{v}} \right) \quad (2-10)$$

となる。Lord (1958) が示したのは、この α を最大化する $\mathbf{v}$ を求める方法であった。

次に、(2-3)にしたがって、複数の合成変量を考える。 $\mathbf{F}$ の任意の行を $\mathbf{f}_m'$ とする。まず、(2-4)にしたがって、この合成変量の分散が1、すなわち、 $\mathbf{v}_m' \mathbf{S} \mathbf{v}_m = 1$ となることを要請すれば、その α 係数 α_m は、

$$\alpha_m = \frac{n}{n-1} (1 - \mathbf{v}_m' \mathbf{D} \mathbf{v}_m) \quad (2-11)$$

という、簡単な形であらわせることを確認しておこう。

ここで、合成変数の数を q とし、その α 係数の和を A とする。すなわち、

$$A = \sum \alpha_m \quad (2-12)$$

である。式(2-11)を代入して、

$$A = \frac{n}{n-1} (q - \text{tr } \mathbf{V}' \mathbf{D} \mathbf{V}) \quad (2-13)$$

である。この A を制約条件(2-4)の下で最大化することを考える。

A の最大化は、 $\text{tr } \mathbf{V}' \mathbf{D} \mathbf{V}$ の最小化と等価であることは容易にわかるから、 $\mathbf{M}$ をラグランジュの乗数行列として、

$$S = \text{tr } \mathbf{V}' \mathbf{D} \mathbf{V} - \text{tr } \mathbf{M} (\mathbf{V}' \mathbf{S} \mathbf{V} - \mathbf{I}) \quad (2-14)$$

を最小化するのが基準である。途中経過は省略するが、

$$\mathbf{U} = \mathbf{D}^{\frac{1}{2}} \mathbf{V} \quad (2-15)$$

とし、 $\mathbf{T}$ を任意の直交行列とすると、

$$\mathbf{R} \mathbf{U} \mathbf{T} = \mathbf{U} \mathbf{T} \mathbf{A} \quad (2-16)$$

が導かれる。ここで、 $\mathbf{R}$ は $\mathbf{D}^{-\frac{1}{2}} \mathbf{S} \mathbf{D}^{-\frac{1}{2}}$ によって定義される項目間相関行列である。また、 $\mathbf{A}$ は $\mathbf{R}$ の最大 q 番目までの固有値を要素とする対角行列であり、式(2-16)の形から見て、 $\mathbf{U} \mathbf{T}$ は、それに対応する固有ベクトルを要素とする正規直交行列である。なお、

$$\mathbf{T}' \mathbf{U}' \mathbf{U} \mathbf{T} = \mathbf{A}^{-1} \quad (2-17)$$

となるように基準化する。このとき(2-6)の負荷行列は、

$$\mathbf{A} = \mathbf{R} \mathbf{U} = \mathbf{U} \mathbf{T} \mathbf{A}' \mathbf{T}' \quad (2-18)$$

となる。この $\mathbf{A}$ は $\mathbf{R}$ の、ノルムを固有値に等しくなるように基準化した固有ベクトルを、直交回転したものである、この $\mathbf{A}$ 、すなわち、全合成変量の α 係数の和の最大化という基準によって導き出される合成変量は、後に示すように通常の定式化によるものと実質的に同一である。

この定式化に関しては、2つの点に注目すべきである。

第1に、我々が、標準得点行列 $\mathbf{Z}$ でなく、偏差得点行列 $\mathbf{X}$ から出発したにもかかわらず、結果は相関行列 $\mathbf{R}$ の固有値問題に帰着したことである。すなわち、素データから α 係数最大の合成変量を導こうとする限り、もとのデータの単位にかかわりなく、データを標準化する前処理を行ったのと同じ結果になる。逆に言えば、データを標準化することは、この定式化をとれば、常に正当化される。素データの単位にかかわらず同一の結果が得られるような解法は scale free な解法と呼ばれる。(例えば、Kaiser & Caffrey 1965。この Kaiser らの論文で、提案されているアルファ因子分析は、共通因子空間で、 α 係数の最大化をはかるものである。)

第2は、回転に関連した問題である。式(2-11)によって与えられる個々の主成分得点の α 係数は、直交回転によって変化するが、その和である A の値は、回転に関して不変である。このことは、

$$\text{tr} \mathbf{T}' \mathbf{V}' \mathbf{D} \mathbf{V} \mathbf{T} = \text{tr} \mathbf{V}' \mathbf{D} \mathbf{V} \mathbf{T} \mathbf{T}' = \text{tr} \mathbf{V}' \mathbf{D} \mathbf{V}$$

のようにして示すことができるが、その結果として、解には任意の直交行列 $\mathbf{T}$ に関する不定性が生ずることになる。

なお最後に、この定式化と、因子数の決定のための Guttman-Kaiser の基準との関係に触れておこう。ここまで、因子数(主成分または合成変量の数) q は、あらかじめ決められているものと見なして、最大化された A の値はどのように変化するかを考えてみる。まず、

$$\text{tr} \mathbf{V}' \mathbf{D} \mathbf{V} = \text{tr} \mathbf{U}' \mathbf{U} = \text{tr} A$$

である。したがって、主成分の数を1だけ増やしたとき、式(2-13)は、 q が $q+1$ にかわり、 $\text{tr} \mathbf{V}' \mathbf{D} \mathbf{V}$ は、 $q+1$ 番目の固有値の逆数 $\frac{1}{\lambda_{q+1}}$ だけ増加する。したがって、この値が1を下回る限り、 A は増加することになる。これは、 $\lambda_{q+1} > 1$ を意味している。すなわち、1を越える固有値の数を因子数とする、という Guttman-Kaiser の基準は、 A を最大化するように因子数を決めるということと同等であったわけである。

次の節で、この基準を多集合データに一般化した主成分分析の定式化をはかる。

3. 階層的モデル

1節で述べた多集合モデルを行列を用いて書きかえることから始めよう。まず、集合ごとに合成変量の行列 $\mathbf{F}_k$ を求めることになる。すなわち、集合 k において求めるべき合成変量の数を q_k とし、 $\mathbf{V}_k$ を $n_k \times q_k$ の重み行列、 $\mathbf{X}_k$ を $n_k \times N$ のデータ行列とすると、

$$\mathbf{F}_k = \mathbf{V}_k' \mathbf{X}_k \quad k = 1, \dots, p \quad (3-1)$$

とする。なお、 q_k は $\mathbf{X}_k$ のランクより小さいものとする。この $\mathbf{F}_k$ の要素が(1-3)の1次合成変量である。ここで、データ行列は、行ごとに平均0、すなわち、

$$\mathbf{X}_k \mathbf{1} = \mathbf{0} \quad k = 1, \dots, p \quad (3-2)$$

と基準化してあるものとする。必然的に、

$$\mathbf{F}_k \mathbf{1} = \mathbf{0} \quad k = 1, \dots, p \quad (3-3)$$

が成立する。また、合成変量はすべて分散1で、相互に直交するものとする。これは、

$$\frac{1}{N} \mathbf{F}_k \mathbf{F}_k' = \mathbf{I} \quad k = 1, \dots, p \quad (3-4)$$

と書けるが、(3-1)により、これは、

$$\mathbf{V}_k' \mathbf{S}_{kk} \mathbf{V}_k = \mathbf{I} \quad k = 1, \dots, p \quad (3-5)$$

ということである。なお、

$$\mathbf{S}_{kl} = \frac{1}{N} \mathbf{X}_k \mathbf{X}_l' \quad k = 1, \dots, p \\ l = 1, \dots, p \quad (3-6)$$

は、集合 k の全変数と集合 l の全変数の間の共分散行列であり、(3-5)の $\mathbf{S}_{kk}$ は、 $l = k$ の特別の場合、すなわち、集合 k に属する項目間の相関行列である。

次に、この合成変量を更に合成した2次の合成変量 $\mathbf{G}$ は、

$$\mathbf{G} = \sum \mathbf{W}_k' \mathbf{F}_k \quad (3-7)$$

と定義される。この $\mathbf{G}$ の要素が、(1-4)で定義されたものであり、 $\mathbf{W}_k$ は、 $q_k \times Q$ の重み行列である。なお、 Q は、2次合成変量の数であり、

$$\max q_k \leq Q \leq \sum q_k \quad (3-8)$$

という条件を付しておく。ここで、(3-3)から必然的に、

$$\mathbf{G} \mathbf{1} = \mathbf{0} \quad (3-9)$$

である、すなわち、すべての2次合成変量の平均は0となる。さらに、2次合成変量についても、分散1で相互に直交する、という条件を課すこととしよう。すなわち、

$$\frac{1}{N} \mathbf{G} \mathbf{G}' = \mathbf{I} \quad (3-10)$$

とする。(3-1)と(3-7)をあわせて、

$$\mathbf{G} = \sum \mathbf{W}_k' \mathbf{V}_k \mathbf{X}_k \quad (3-11)$$

となるから、(3-10)は結局、

$$\sum_{k=1}^p \sum_{l=1}^p \mathbf{W}_k' \mathbf{V}_k' \mathbf{S}_{kl} \mathbf{V}_l \mathbf{W}_l = \mathbf{I} \quad (3-12)$$

という制約条件となる。

この段階で、 α 係数の和の最大化、という最適化基準の設定の仕方は、幾つか考えられる。まず、集合ごとの合成変量 $\mathbf{F}_k$ について α 係数の和 A_k を考え、その和、

$$\sum A_k = \sum \frac{n_k}{n_k - 1} (q_k - \text{tr} \mathbf{V}_k \mathbf{D}_k \mathbf{V}_k) \quad (3-13)$$

を最大化するというやり方がありうる。ここで、

$$\mathbf{D}_k = \text{diag} \mathbf{S}_{kk} \quad k = 1, \dots, p \quad (3-14)$$

である。しかしながら、これは集合ごとに $\text{tr} \mathbf{V}_k' \mathbf{D}_k \mathbf{V}_k$ を最小化することに帰着し、結局集合ごとに個別に主成分分析した結果を、任意に重みづけた2次合成変量を求める、ということになるから、多集合データを同時的に分析する積極的意義があまりない。

また、高次合成変量を全集合の1次合成変量の線型結合とみなして、その α 係数の和 A 、

$$A = \frac{\sum q_k}{\sum q_k - 1} (Q - \text{tr} \sum \mathbf{W}_k' \mathbf{W}_k) \quad (3-15)$$

の最大化も考えられる。これは、正準相関分析の一種の一般化につながり、一つの可能性としてはありうる。しかし専ら2次合成変量の内的整合性のみに焦点が合わせられる結果、1次合成変量の内的整合性が犠牲になるという、正準相関分析を質問紙の項目水準のデータに適用したのと同じ問題が生ずると思われる。

ここでは、第3の可能性、すなわち、二次合成変量を、(3-12)にしたがって、素データの1次結合とみなした場合における α 係数の和、

$$A = \frac{\sum n_k}{\sum n_k - 1} (Q - \text{tr} \sum \mathbf{W}_k' \mathbf{V}_k' \mathbf{D}_k \mathbf{V}_k \mathbf{W}_k) \quad (3-16)$$

の最大化として、定式化することにしよう。このやり方は結局、制約条件(3-5)、(3-11)の下における $\text{tr} \sum \mathbf{W}_k' \mathbf{V}_k' \mathbf{D}_k \mathbf{V}_k \mathbf{W}_k$ の最小化ということになる。

この基準では、 $\mathbf{W}_k$ と $\mathbf{V}_k$ の要素をともに小さくすることが求められるから、与えられた合成変量の数の下で、内的整合性と集合間相関とはある程度バランスすることになるであろう。また、1節で述べたように、各集合の1次合成変量の数を固定した上で、2次合成変量の数 Q を増減することによって、意図的に bandwidth-fidelity dilemma についての調整を行うことができるはずである。

4. アルゴリズム

最適解の導出

制約条件(3-5)と(3-12)を含めて、最小化基準は、

$$S = \sum \text{tr} \mathbf{W}_k' \mathbf{V}_k' \mathbf{D}_k \mathbf{V}_k \mathbf{W}_k - \text{tr} \sum \mathbf{M}_k (\mathbf{V}_k' \mathbf{S}_{kk} \mathbf{V}_k - \mathbf{I}) - \text{tr} \mathbf{M} \left(\sum_{k=1}^p \sum_{l=1}^p \mathbf{W}_k' \mathbf{V}_k' \mathbf{S}_{kl} \mathbf{V}_l \mathbf{W}_l - \mathbf{I} \right) \quad (4-1)$$

となる。 $\mathbf{M}$, $\mathbf{M}_k$ はラグランジュの乗数行列である。この S を $\mathbf{V}_k$, $\mathbf{W}_k$, $\mathbf{M}_k$, $\mathbf{M}$ で偏微分して全要素を0とおくことにより、

$$\mathbf{D}_k \mathbf{V}_k \mathbf{W}_k \mathbf{W}_k' - \sum_{l=1}^p \mathbf{S}_{kl} \mathbf{V}_l \mathbf{W}_l \mathbf{L} \mathbf{W}_k' - \mathbf{S}_{kk} \mathbf{V}_k \mathbf{L}_k = \mathbf{0} \quad k = 1, \dots, p \quad (4-2)$$

$$\mathbf{V}_k' \mathbf{D}_k \mathbf{V}_k \mathbf{W}_k - \sum_{l=1}^p \mathbf{V}_k' \mathbf{S}_{kl} \mathbf{V}_l \mathbf{W}_l \mathbf{L} = \mathbf{0} \quad k = 1, \dots, p \quad (4-3)$$

$$\mathbf{V}_k' \mathbf{S}_{kk} \mathbf{V}_k = \mathbf{I} \quad k = 1, \dots, p \quad (4-4)$$

$$\sum_{k=1}^p \sum_{l=1}^p \mathbf{W}_k' \mathbf{V}_k' \mathbf{S}_{kl} \mathbf{V}_l \mathbf{W}_l \mathbf{L} = \mathbf{I} \quad (4-5)$$

ここで、

$$\mathbf{L}_k = \frac{\mathbf{M}_k + \mathbf{M}_k'}{2} \quad k = 1, \dots, p \quad (4-6)$$

$$\mathbf{L} = \frac{\mathbf{M} + \mathbf{M}'}{2} \quad (4-7)$$

である。

まず、(4-2)の左から、 $\mathbf{V}_k'$ をかけ、(4-3)と(4-4)を用いて、

$$\mathbf{L}_k = (\mathbf{V}_k' \mathbf{D}_k \mathbf{V}_k \mathbf{W}_k - \sum_{l=1}^p \mathbf{V}_k' \mathbf{S}_{kl} \mathbf{V}_l \mathbf{W}_l \mathbf{L}) \mathbf{W}_k' = \mathbf{0} \quad k = 1, \dots, p \quad (4-8)$$

となる。つぎに、

$$\mathbf{U}_k = \mathbf{D}_k^{-\frac{1}{2}} \mathbf{V}_k \quad k = 1, \dots, p \quad (4-9)$$

$$\mathbf{R}_{kl} = \mathbf{D}_k^{-\frac{1}{2}} \mathbf{S}_{kl} \mathbf{D}_l^{-\frac{1}{2}} \quad k = 1, \dots, p \quad l = 1, \dots, p \quad (4-10)$$

とおく。 $\mathbf{R}_{kl}$ は、2つの集合間の項目間相関行列である。これを用いて(4-3)は、

$$\sum_{l=1}^p \mathbf{U}_k' \mathbf{R}_{kl} \mathbf{U}_l \mathbf{W}_l \mathbf{L} = \mathbf{U}_k' \mathbf{U}_k \mathbf{W}_k \quad k = 1, \dots, p \quad (4-11)$$

となる。また、(4-2)は、左から $\mathbf{D}_k^{-\frac{1}{2}}$ をかけ、(4-8)を用いて、

$$\sum_{l=1}^p \mathbf{R}_{kl} \mathbf{U}_l \mathbf{W}_l \mathbf{L} \mathbf{W}_k' = \mathbf{U}_k \mathbf{W}_k \mathbf{W}_k' \quad k = 1, \dots, p \quad (4-12)$$

となる。

つぎに、

$$\mathbf{U}_k^* = \mathbf{U}_k (\mathbf{U}_k' \mathbf{U}_k)^{-\frac{1}{2}} \quad k = 1, \dots, p \quad (4-13)$$

$$\mathbf{W}_k^* = (\mathbf{U}_k' \mathbf{U}_k)^{\frac{1}{2}} \mathbf{W}_k \quad k = 1, \dots, p \quad (4-14)$$

とおき、(4-11)の左から $(\mathbf{U}_k' \mathbf{U}_k)^{\frac{1}{2}}$ をかけて、

$$\sum_{l=1}^p \mathbf{U}_k^* \mathbf{R}_{kl} \mathbf{U}_l^* \mathbf{W}_l^* \mathbf{L} = \mathbf{W}_k^* \quad k = 1, \dots, p \quad (4-15)$$

が得られ、また、(4-12)の右から $(\mathbf{U}_k' \mathbf{U}_k)^{\frac{1}{2}}$ をかけて、

$$\sum_{l=1}^p \mathbf{R}_{kl} \mathbf{U}_l^* \mathbf{W}_l^* \mathbf{L} \mathbf{W}_k^{*'} = \mathbf{U}_k^* (\mathbf{W}_k^* \mathbf{W}_k^{*'}) \quad k = 1, \dots, p \quad (4-16)$$

を得る。さらに、(4-15)の左から $\mathbf{W}_k^{*'}$ をかけて、 k について総和し、(4-5)を用いて、

$$\mathbf{L} = \sum \mathbf{W}_k^{*'} \mathbf{W}_k^* = \sum \mathbf{W}_k' \mathbf{V}_k \mathbf{D}_k \mathbf{V}_k \mathbf{W}_k \quad (4-17)$$

$\mathbf{L}$ は非負定符号行列だから、適当な非負定符号の対角行列 $\mathbf{A}$ と直交行列 $\mathbf{T}$ を用いて、

$$\mathbf{L} = \tilde{\mathbf{T}} \mathbf{A} \tilde{\mathbf{T}}' \quad (4-18)$$

と分解できる。また、行列 $\mathbf{W}_k^* \mathbf{W}_k^{*'}$ も非負定符号だから、やはり、適当な非負定符号対角行列 $\mathbf{A}_k$ と直交行列 $\mathbf{T}_k$ を用いて、

$$\mathbf{W}_k^* \mathbf{W}_k^{*'} = \mathbf{T}_k \mathbf{A}_k \mathbf{T}_k' \quad k = 1, \dots, p \quad (4-19)$$

と分解できる。(以下、証明の簡略化のために、 $\mathbf{A}$ と $\mathbf{A}_k$ は正定符号と仮定する。これらの行列の対角要素に0のものがあるときは q_k と Q をその数だけ減らせばよい。)そこで、

$$\mathbf{r}_k = \mathbf{U}_k^* \mathbf{T}_k \quad k = 1, \dots, p \quad (4-20)$$

$$\mathbf{\Gamma}_k = \mathbf{T}_k' \mathbf{W}_k^* \tilde{\mathbf{T}} \mathbf{A}^{-\frac{1}{2}} \quad k = 1, \dots, p \quad (4-21)$$

とおき、(4-15)の左から $\mathbf{T}_k'$ 、右から $\mathbf{L}^{-1}$ をあけて整理すると、

$$\sum_{l=1}^p \mathbf{r}_k' \mathbf{R}_{kl} \mathbf{\Gamma}_l \mathbf{\Gamma}_l = \mathbf{\Gamma}_k \mathbf{A}^{-1} \quad k = 1, \dots, p \quad (4-22)$$

となり、また、(4-16)の右から $\mathbf{T}_k$ をかけて、

$$\sum_{l=1}^p \mathbf{R}_{kl} \mathbf{\Gamma}_l \mathbf{\Gamma}_l \mathbf{A}^2 \mathbf{\Gamma}_k' = \mathbf{r}_k \mathbf{A}_k \quad k = 1, \dots, p \quad (4-23)$$

が得られる。ここで、 $\mathbf{r}_k$ と $\mathbf{\Gamma}_k$ は、それぞれ、

$$\mathbf{r}_k' \mathbf{r}_k = \mathbf{I} \quad k = 1, \dots, p \quad (4-24)$$

$$\sum \mathbf{\Gamma}_k' \mathbf{\Gamma}_k = \mathbf{I} \quad (4-25)$$

を満たす。

かくして、(4-22)は、行列

$$\mathbf{\Gamma} = [\mathbf{\Gamma}_1', \mathbf{\Gamma}_2', \dots, \mathbf{\Gamma}_p']' \quad (4-26)$$

が、行列、

$$\mathbf{\Psi} = \begin{bmatrix} \mathbf{r}_1' \mathbf{R}_{11} \mathbf{\Gamma}_1, \dots, \mathbf{r}_1' \mathbf{R}_{1p} \mathbf{\Gamma}_p \\ \dots\dots\dots \\ \mathbf{r}_p' \mathbf{R}_{p1} \mathbf{\Gamma}_1, \dots, \mathbf{r}_p' \mathbf{R}_{pp} \mathbf{\Gamma}_p \end{bmatrix} \quad (4-27)$$

の正規直交な固有ベクトル、すなわち、

$$\mathbf{\Psi} \mathbf{\Gamma} = \mathbf{\Gamma} \mathbf{A}^{-1} \quad (\mathbf{\Gamma}' \mathbf{\Gamma} = \mathbf{I}) \quad (4-28)$$

であることを意味していることがわかる。ここで、(4-17)、および(4-18)から、

$$\text{tr } \mathbf{A} = \sum \text{tr } \mathbf{W}_k' \mathbf{V}_k \mathbf{D}_k \mathbf{V}_k \mathbf{W}_k \quad (4-29)$$

であり、これが最小化すべきものであるから、(4-22)の形から見て、 $\mathbf{\Gamma}$ は行列 $\mathbf{\Psi}$ の最大 Q 番目までの固有値に対応する固有ベクトルを要素とするようにとればよいことになる。

これと、(4-23)を合わせて、次のようなアルゴリズムが示唆される。

- (1) 適当な $\mathbf{r}_k$ の初期値を設定する。 $(\mathbf{R}_{kk})$ の最大 q_k 番目までの固有値に対応する固有ベクトルをとる。
- (2) 行列 $\mathbf{\Psi}$ をもとめ、その固有値と固有ベクトルとして、 $\mathbf{A}$ と $\mathbf{\Gamma}$ の近似値を求める。
- (3) 次の行列、

$$\mathbf{r}_k^* = \sum_{l=1}^p \mathbf{R}_{kl} \mathbf{\Gamma}_l \mathbf{\Gamma}_l \mathbf{A}^2 \mathbf{\Gamma}_k' \quad k = 1, \dots, p \quad (4-30)$$

を算出する。

- (4) $\mathbf{r}_k^*$ をグラム-シュミットの直交化法(例えば、竹内1974 p.42)により、正規直交化したものを新しい $\mathbf{r}_k$ とする。すなわち、

$$\mathbf{r}_k = \mathbf{r}_k^* \mathbf{A}_k \quad k = 1, \dots, p \quad (4-31)$$

ただし、 $\mathbf{A}_k$ は、 $\mathbf{r}_k$ が、(4-24)を満たすような下三角行列である。

- (5) 収束すれば終了。そうでなければ、(2)へもどる。

この交互的アルゴリズムが収束すれば、解は得られる。なお、行列 Ψ に 0 の固有値が出現した場合には Q を、行列 Δ_k の対角要素が 0 となった場合には、 q_k をそれぞれ 1 ずつ減らす。

なお、収束時には、

$$\Gamma_k \Delta_k^{-1} = \sum_{l=1}^p \mathbf{R}_{kl} \Gamma_l \Gamma_l' \Gamma_k' \quad k = 1, \dots, p \quad (4-32)$$

となり、これに Γ_k' を左からかけて、(4-22), (4-24) を利用すると、

$$\Delta_k^{-1} = \Gamma_k A \Gamma_k' \quad k = 1, \dots, p \quad (4-33)$$

となる。 Δ_k は、下三角行列であり、かつ (4-33) はこれが対称行列でもあることを意味しているから、 Δ_k は対角行列となる。また、(4-23) と (4-32) を比較することにより、

$$A_k = \Delta_k^{-1} \quad k = 1, \dots, p \quad (4-34)$$

ということになる。更に、(4-33), (4-34) から、

$$\text{tr } A = \sum \text{tr } A_k \quad k = 1, \dots, p \quad (4-35)$$

であることもわかる。

重み行列の算出

さて、目標となる 2 つの重み行列、 $\mathbf{V}_k$ と $\mathbf{W}_k$ は、

$$\begin{aligned} \mathbf{V}_k &= \mathbf{D}_k^{-\frac{1}{2}} \mathbf{U}_k \\ &= \mathbf{D}_k^{-\frac{1}{2}} \mathbf{U}_k^* (\mathbf{U}_k' \mathbf{U}_k)^{-\frac{1}{2}} \\ &= \mathbf{D}_k^{-\frac{1}{2}} \Gamma_k \mathbf{T}_k' (\mathbf{U}_k' \mathbf{U}_k)^{-\frac{1}{2}} \\ & \quad k = 1, \dots, p \quad (4-36) \end{aligned}$$

$$\begin{aligned} \mathbf{W}_k &= (\mathbf{U}_k' \mathbf{U}_k)^{-\frac{1}{2}} \mathbf{W}_k^* \\ &= (\mathbf{U}_k' \mathbf{U}_k)^{-\frac{1}{2}} \mathbf{T}_k \Gamma_k A^{-\frac{1}{2}} \tilde{\mathbf{T}}' \\ & \quad k = 1, \dots, p \quad (4-37) \end{aligned}$$

によって求められることになるわけであるが、ここで、 $(\mathbf{U}_k' \mathbf{U}_k)^{-\frac{1}{2}}$, $\mathbf{T}_k$, $\tilde{\mathbf{T}}$ が不定のまま残された。このうち、 $(\mathbf{U}_k' \mathbf{U}_k)^{-\frac{1}{2}}$ は、(4-4) から定められる。すなわち、(4-4) は、

$$\begin{aligned} \mathbf{V}_k' \mathbf{S}_{kk} \mathbf{V}_k &= (\mathbf{U}_k' \mathbf{U}_k)^{-\frac{1}{2}} \mathbf{T}_k \Gamma_k' \mathbf{R}_{kk} \Gamma_k \mathbf{T}_k' (\mathbf{U}_k' \mathbf{U}_k)^{-\frac{1}{2}} \\ &= \mathbf{I} \quad k = 1, \dots, p \quad (4-38) \end{aligned}$$

を意味する。したがって、

$$\begin{aligned} \Gamma_k' \mathbf{R}_{kk} \Gamma_k &= \mathbf{T}_k' (\mathbf{U}_k' \mathbf{U}_k)^{-1} \mathbf{T}_k \\ & \quad k = 1, \dots, p \quad (4-39) \end{aligned}$$

この左辺は、 $q_k \times q_k$ の正定符号の行列だから、適当な正定符号の対角行列 $\mathbf{M}_k$ と正規直交行列 $\mathbf{X}_k$ によって、

$$\begin{aligned} \Gamma_k' \mathbf{R}_{kk} \Gamma_k &= \mathbf{X}_k \mathbf{M}_k \mathbf{X}_k' \\ & \quad k = 1, \dots, p \quad (4-40) \end{aligned}$$

と分解できる。これにより、

$$\begin{aligned} \mathbf{U}_k' \mathbf{U}_k &= \mathbf{T}_k \mathbf{X}_k \mathbf{M}_k^{-1} \mathbf{X}_k' \mathbf{T}_k' \\ & \quad k = 1, \dots, p \quad (4-41) \end{aligned}$$

と書ける。したがって、

$$\begin{aligned} (\mathbf{U}_k' \mathbf{U}_k)^{-\frac{1}{2}} &= \mathbf{T}_k \mathbf{X}_k \mathbf{M}_k^{-\frac{1}{2}} \mathbf{X}_k' \mathbf{T}_k' \\ &= \mathbf{T}_k (\Gamma_k' \mathbf{R}_{kk} \Gamma_k)^{-\frac{1}{2}} \mathbf{T}_k' \\ & \quad k = 1, \dots, p \quad (4-42) \end{aligned}$$

が得られる。かくして、最終的な $\mathbf{V}_k$ と $\mathbf{W}_k$ は、

$$\begin{aligned} \mathbf{V}_k &= \mathbf{D}_k^{-\frac{1}{2}} \Gamma_k \mathbf{X}_k \mathbf{M}_k^{-\frac{1}{2}} \mathbf{X}_k' \mathbf{T}_k' \\ &= \mathbf{D}_k^{-\frac{1}{2}} \Gamma_k (\Gamma_k' \mathbf{R}_{kk} \Gamma_k)^{-\frac{1}{2}} \mathbf{T}_k' \\ & \quad k = 1, \dots, p \quad (4-43) \end{aligned}$$

$$\begin{aligned} \mathbf{W}_k &= \mathbf{T}_k \mathbf{X}_k \mathbf{M}_k^{\frac{1}{2}} \mathbf{X}_k' \Gamma_k A^{-\frac{1}{2}} \tilde{\mathbf{T}}' \\ &= \mathbf{T}_k (\Gamma_k' \mathbf{R}_{kk} \Gamma_k)^{-\frac{1}{2}} \Gamma_k A^{-\frac{1}{2}} \tilde{\mathbf{T}}' \\ & \quad k = 1, \dots, p \quad (4-44) \end{aligned}$$

によって求められる。こうして、最終的に直交回転、 $\mathbf{T}_k$ と $\tilde{\mathbf{T}}$ に関する不定性だけが残ることになる。

合成変量の回転

回転行列 $\tilde{\mathbf{T}}$ と $\mathbf{T}_k$ を決定する方法には、さまざまなものが考えられる。

1 つのやり方は、重み行列 $\mathbf{V}_k$ を単純構造にするように行列 $\mathbf{T}_k$ を定め、次に、これによって回転された $\mathbf{W}_k$ を単純構造化するように $\tilde{\mathbf{T}}$ を決める。というものである。これは、合成変量の意味を、個々の変数から解釈していくためには、適切なやり方のようにも思われるが、筆者自身まだ試みていない。また、通常の主成分分析において、このやり方を採用した例もないようである。(ただし、基準化した重みを直交回転して因子そのものの斜交回転とする手法は Harris & Kaiser 1964 が提案している。)

より伝統的なやり方は、素データと合成変量との相関(因子分析の用語で言えば、因子負荷)、あるいは、1 次

合成変量と2次合成変量の間の相関（本研究の立場で言えば2次因子負荷）からこれらを決めていくことである。

まず、素データと1次合成変量との相関行列は、

$$\begin{aligned} \mathbf{A}_k &= \frac{1}{N} \mathbf{D}_k^{-\frac{1}{2}} \mathbf{X}_k \mathbf{F}'_k \\ &= \mathbf{D}_k^{-\frac{1}{2}} \mathbf{S}_{kk} \mathbf{V}_k \\ &= \mathbf{R}_{kk} \mathbf{U}_k \quad k = 1, \dots, p \quad (4-45) \end{aligned}$$

となるが、これを1次負荷行列と呼ぶ。

1次合成変量と2次合成変量との相関行列は、

$$\begin{aligned} \mathbf{C}_k &= \frac{1}{N} \mathbf{F}_k \mathbf{G}' \\ &= \sum_{l=1}^p \mathbf{V}_k \mathbf{S}_{kl} \mathbf{V}_l \mathbf{W}_l \\ &= \sum_{l=1}^p \mathbf{U}'_k \mathbf{R}_{kl} \mathbf{U}_l \mathbf{W}_l \quad k = 1, \dots, p \quad (4-46) \end{aligned}$$

であり、これを2次負荷行列と呼ぶ。

もう1つ、素データと2次合成変量との相関行列を考えることもできる。これは、

$$\begin{aligned} \mathbf{B}_k &= \mathbf{D}_k^{-\frac{1}{2}} \frac{1}{N} \mathbf{X}_k \mathbf{G}' \\ &= \sum_{l=1}^p \mathbf{D}_k^{-\frac{1}{2}} \mathbf{S}_{kl} \mathbf{V}_l \mathbf{W}_l \\ &= \sum_{l=1}^p \mathbf{R}_{kl} \mathbf{U}_l \mathbf{W}_l \quad k = 1, \dots, p \quad (4-47) \end{aligned}$$

この行列の要素を大局的負荷（global loading）と呼ぶ。

ここから、次のような2種の回転方法が示唆される。

(a) まず1次負荷行列を単純構造化するように $\mathbf{T}_k$ を（例えば、バリマックス回転によって）定め、そこからえられる $\mathbf{V}_k$ を用いて算出される2次負荷行列を単純構造化するように $\tilde{\mathbf{T}}$ を定める。この $\mathbf{T}$ によって大局的負荷行列を回転する。

(b) 1次負荷行列を単純構造化するように $\mathbf{T}_k$ を定め、それとは独立に、大局的負荷行列を単純構造化するように $\mathbf{T}$ を定める。得られた $\mathbf{T}_k$ と $\tilde{\mathbf{T}}$ を用いて、2次因子負荷を回転する。

一般には、(a)のやり方が好まれよう。ただし、本研究の方法は、2次合成変量よりも、むしろ1次合成変量の集合間相関、

$$\mathbf{P}_{kl} = \frac{1}{N} \mathbf{F}_k \mathbf{F}'_l$$

$$\begin{aligned} &= \mathbf{V}'_k \mathbf{S}_{kl} \mathbf{V}_l \quad k = 1, \dots, p \\ & \quad l = 1, \dots, p \quad (4-48) \end{aligned}$$

の方に関心が集まる場合が多いと思われる。この観点に立てば、2次合成変量は、bandwidth と fidelityの間のバランスを調整するための、一種の方便とも見なされる。もしそうなら、(a)と(b)には大差がないことになる。

$\mathbf{P}_{kl}$ は、 $\tilde{\mathbf{T}}$ とは無関係だからである。

一応、(a)の回転法をとるとして、 $\mathbf{r}_k$ と $\mathbf{\Gamma}$ が得られた後のアルゴリズムをまとめると以下ようになる。

(6) 式(4-40)の分解を行い、それに基づいて行列、

$$\Phi_k = \mathbf{X}_k \mathbf{M}_k^{\frac{1}{2}} \mathbf{X}'_k \quad k = 1, \dots, p \quad (4-49)$$

および、

$$\Phi_k^{-1} = \mathbf{X}_k \mathbf{M}_k^{-\frac{1}{2}} \mathbf{X}'_k \quad k = 1, \dots, p \quad (4-50)$$

を算出する。 Φ_k は $(\mathbf{\Gamma}'_k \mathbf{R}_{kk} \mathbf{\Gamma}_k)^{\frac{1}{2}}$ に相当する。

(7) 回転前の重み行列、

$$\mathbf{U}_k^* = \mathbf{\Gamma}_k \Phi_k^{-1} \quad k = 1, \dots, p \quad (4-51)$$

$$\mathbf{W}_k^* = \Phi_k \mathbf{\Gamma} \quad k = 1, \dots, p \quad (4-52)$$

を算出する。 $\mathbf{U}_k^*$ に基づいて、回転前の1次因子負荷行列、

$$\mathbf{A}_k^* = \mathbf{R}_{kk} \mathbf{U}_k^* \quad k = 1, \dots, p \quad (4-53)$$

を算出する。

(8) $\mathbf{A}_k^* (k = 1, \dots, p)$ をバリマックス回転する。回転された1次因子負荷行列を $\mathbf{A}_k$ とする($\mathbf{A}_k = \mathbf{A}_k^* \mathbf{T}'_k$)。回転行列 $\mathbf{T}'_k$ は保存し、回転後の重み行列、

$$\mathbf{U}_k = \mathbf{U}_k^* \mathbf{T}'_k \quad k = 1, \dots, p \quad (4-54)$$

$$\mathbf{W}_k^\dagger = \mathbf{T}_k \mathbf{W}_k^* \quad k = 1, \dots, p \quad (4-55)$$

を求める。 $\mathbf{W}_k^\dagger$ を用いて、(4-46)にしたがって、回転前の2次因子負荷行列、

$$\mathbf{C}_k^* = \sum_{l=1}^p \mathbf{U}'_k \mathbf{R}_{kl} \mathbf{V}_l \mathbf{W}_l^\dagger \quad k = 1, \dots, p \quad (4-56)$$

を算出する。

(9) $\mathbf{C}_k^*$ を並べた $\sum n_k \times Q$ の行列、

$$\mathbf{C}^* = [\mathbf{C}_1^*, \mathbf{C}_2^*, \dots, \mathbf{C}_p^*]' \quad (4-57)$$

をバリマックス回転する。回転後の行列を $\mathbf{C} (= \mathbf{C}^* \tilde{\mathbf{T}}')$ とする。この要素行列 $\mathbf{C}_k$ が、最終的な、2次因子負荷行列である。回転行列は保存し、これによって、

$$W_k = W_k^\dagger \tilde{T}' \quad k = 1, \dots, p \quad (4-58)$$

を求める。これが、最終的な2次の重み行列となる。

(10) 式(4-47)によって、大局的因子負荷行列

$B_k (k = 1, \dots, p)$ を求める。

$B_k (k = 1, \dots, p)$ を求める。(ただし、これは、(8)で(4-56)により C_k を求める前段階で、

$$B_k^* = \sum R_{kl} U_l W_l^\dagger \quad k = 1, \dots, p \quad (4-59)$$

として求めておく方が効率的であろう。これを用いれば C_k^* は、

$$C_k^* = U_k' B_k^* \quad k = 1, \dots, p \quad (4-60)$$

によって求められるからである。そうすれば、この段階(10)では、

$$B_k = B_k^* \tilde{T}' \quad k = 1, \dots, p \quad (4-61)$$

の回転によって、 B_k が求まることになる。

また、式(4-47)によって、合成変量間相関行列を、

$$P = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1p} \\ \dots & \dots & \dots & \dots \\ P_{p1} & P_{p2} & \dots & P_{pp} \end{bmatrix} \quad (4-62)$$

のように、全集合間の相関行列の形で求める。

文 献

- Cronbach, L. J. 1951 Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Cronbach, L. J. 1970 *Essentials of Psychological Testing*, 3rd ed. New York: Harper & Row.
- Guttman, L. 1955 The determinancy of factor score matrices with implications for five other basic problems of common-factor theory. *British Journal of Statistical Psychology*, 8, 65-81.
- Harris, C. W. & Kaiser, H. F. 1964 Oblique factor analytic solutions by orthogonal transformations. *Psychometrika*, 29, 347-362.
- 池田 央 1973 心理学研究法 8 テストII 東京大学出版社。
- Kaiser, H. F. & Caffrey, C. 1965 Alpha factor analysis. *Psychometrika*, 30, 1-14.
- ラザーズフェルド(西田・高坂・奥川訳) 1984 概念形成史小論 ラザーズフェルド 質的分析法—社会学論集—岩波書店 (Lazarsfeld, P. 1972 *Qualitative Analysis: Historical and Critical Essays*. Boston: Allyn and Bacon.)
- Lord, F. M. 1958 Some relations between Guttman's principal components of scale analysis and other psychometric theory. *Psychometrika*, 23, 291-296.
- 村上 隆 1985 多集合データのための探索的因子分析 名古屋大学教育学部紀要—教育心理学科—, 32, 75-93.
- 村上 隆 1986 多集合データのための階層的主成分分析 名古屋大学教育学部紀要—教育心理学科—, 33, 35-48.
- 村上 隆・斎藤和志 1986 多集合因子分析による組織イメージと組織活動の関連の検討 経営行動科学, 1, 41-53.
- 斎藤和志・村上 隆・若林 満 1986 組織の就職先としての魅力と職業志向性及びライフ・スタイルとの関連 経営行動科学, 1, 101-113.
- 芝 祐順 1975 行動科学のための相関分析法 第2版 東京大学出版会。
- 竹内 啓 1977 線形数学 補訂版 培風館

(1987年3月3日受稿, 1987年3月10日受理)