

# 多重対応分析の因子分析的使用

——正規直交主成分分析とそのプロ野球の観客調査データへの適用<sup>1</sup>——

村 上 隆<sup>2</sup>

## 1.1 社会調査データの分析方法

本研究は社会調査データ、すなわち特定の集団から（可能な限り無作為に近い方法で）選択された多数の成員に対し、決められた複数の質問項目への応答を求めるという形式の調査の結果として得られるデータの、分析方法について論じる。ここで提案する手法は、多数の項目への反応を多次元の軸にもとづいて解釈することを通じて、データ全体の様相を容易に把握することを可能にする。より具体的にいえば、従来から社会調査データの分析において多用されている多重対応分析（multiple correspondence analysis: MCA）に、自然なやり方で主成分負荷行列と斜交回転を導入し、計量心理学で開発された（回転を伴う）主成分分析（principal component analysis: PCA）と同じ形式の出力が得られるようにすることによって実現される（村上, 2018）。

### 1.1.1 社会調査データの特徴：心理測定と比較して

社会調査データの分析は、対象とする集団全体の様相を明らかにすることを通じて、社会問題の理解に資することを目的として行われる（Groves, Fowler, Couper, Lepkowski, Singer, & Tourangeau, 2009）。他方、心理測定データの分析の目的は（究極的には）特性次元上での個人の位置の assessment にある（Levy & Mislevy, 2016）。この2つの目的は必ずしも

見かけほど異なるものではない。実際、心理測定がその目的を達成するためには、対象者の中の個人差を際立たせる必要がある。他方、社会調査データの分析の基本は、単純集計とクロス集計にあるが、これらも対象者の反応に個人差があるからこそ成立する分析である。特にクロス集計による項目間関係の検討は、心理測定における相関係数にもとづく推論と同様、個人差を通じてデータの構造を明らかにするために不可欠である。端的に言えば、社会調査データの分析は心理測定データの分析と同様、項目間の関係性を手がかりにデータの構造を明らかにすることを使命としている。

しかしながら、調査票、あるいは質問紙（どちらも inventory の訳語）の質問項目の内容やその選択基準に注目すると、その間の違いは大きい。これも端的に言えば、心理測定においては測定の信頼性を確保するために、項目の内的整合性（internal consistency）が重視されるから、項目は相互に相関の高い、内容的には比較的似かよったものが集められる傾向がある。また、類似した内容の研究が数多くなされる結果、データの構造についての仮説を立てやすく、その検証が目指されることが多い。近年、探索的因子分析よりも確認的因子分析の使用が増加していることもそれを裏付けている。

他方、社会調査においては、社会問題の多様性、多面性を反映して、項目内容は多様なものになる。また、サンプリングのための時間と経費、項目内容・構成が調査を取り巻く環境の変化に大きく影響されること等のために、共通の項目を多数含む調査はあまり行われなない。したがって、データの構造に関する仮説を設定することも難しく、確認的方法よりは探索的方法が優位となる。本研究は、そうした社会調査データの特質を考慮した分析を、心理測定において開発されてきた（広義の）探索的因子分析（たとえば、Nunnally, 1978；芝, 1979；Pet, Lackey & Sullivan, 2003）の手法を援用して実現しようとするものである。

### 1.1.2 多重対応分析とその限界

多重対応分析（数量化3類、等質性分析、双対尺度法等を含む、Nishisato, 1980；Greenace, 1984；Gifi, 1991；Benzecri, 1992；林, 1993；Gower & Hand, 1996；Blasius & Greenacre, 2006；村上, 2011；足立・村上, 2011；Blasius & Greenacre, 2014；Greenacre, 2017）は、従来、社会調査データの分析に多用され、多くの成果をあげてきた。しかしながら、そこには3つの限界があったと考えられる。

第1に、得られる解が多くの場合2次元、せいぜい3次元までに限られることである。これは、この方法が基本的に結果の視覚化を目指し、空間表現にもとづく解釈を重視しているためである。フーリエ解析等を援用した多次元の表現も工夫されている（Rovan, 1994）が、その読み取りは難しく、広く用いられるには至っていない。社会調査の項目が、対象者の多様性の記述を目指して収集されている以上、結果は心理測定以上に多次元的であると予想されるから、この制約はデータの重要な部分を検出できないままで終わる危険性を内包している。

第2に、さまざまな原因によって、多重対応分析の結果には現実の現象を反映するものとは解されない結果のゆがみ、アーチファクトが出現すると考えられていることである。典型的には、最大の説明力をもつ2次元空間における個体の分布が、通常期待される多変量正規分布とは異なる放物線状を呈することであり、これはしばしば馬蹄現象（horse-shoe phenomenon）と呼ばれている（Gifi, 1990；Murakami, 2012）。この現象が出現する十分条件の1つは、データが圧倒的な1次元性的変動を含むことであり、そのことからそれと曲線的関係にある第2次元は無視するのが適当とする議論もある（Bekker & De Leeuw, 1988）。しかし、馬蹄現象の原因はそれには限られず、そこに経験的に意味のある内容が含まれる可能性も探求されるべきである（足立・村上, 2011；村上, 2016）。

第3に、多重対応分析から出力される係数等の推測統計学的な扱いが難しく、専ら記述的な方法として扱われてきたことである。しかしながら、

近年特に、探索的方法といえども出力される係数等の標準誤差等の情報が求められる傾向が強まり、単なる記述だけの結果は受け入れられなくなりつつある。比較的最近になって、コンピュータ・シミュレーションを用いた推測統計学的扱いが進められてきた（大隅・ルバル・モリノウ・ワーウィック・馬場, 1994; Lebart, 2006 など）が、これも広く利用されるには至っていない。

### 1.1.3 本研究の目的・概要

本研究は、基本的には質的データの探索的分析方法としての多重対応分析の数学的枠組みは崩すことなく、多次元データの扱いと係数の標準誤差の計算を可能にしようとする試みである。基本的な構想は、多重対応分析が依拠する平面、あるいは3次元空間の布置を用いた視覚化によるのではなく、相対的に相互に独立な（軸としての）合成変量への負荷量を通じて解釈を行えるようにする。そのための多重対応分析に対する負荷行列（パターン行列）の自然な導入を考え、さらにその回転による単純構造の達成を通して軸単位での解釈を促進することを目指す。また、bootstrap法（Efron & Tibshirani, 1993; Timmerman, Kiers & Smilde, 2007）を用いた係数の標準誤差の推定と有意性の検定について扱う。ただし、馬蹄現象については、適用例の中で若干の示唆が得られるにとどまる。

新たな方法の導入は以下の順序による。1.2.1では本研究における主成分分析を定義し、1.2.2では多重対応分析を定義する。いずれも、単位長さのベクトルからなる重み行列に回転に関する自由度を認めている点で、伝統的な定式化と若干の隔たりがある。1.2.3では各変数のカテゴリーに中心化、正規直交化された数量化ベクトル（複数）を与えて主成分分析を行う新たな方法を導入する。数量化の値は任意であるにもかかわらず、相関行列の固有値は多重対応分析のそれと一致し、主成分得点は重みの直交回転の結果生じる直交変換（結果的には斜交回転となる）を除き、多重対応分析の数量化得点と一致することが示される。1.2.4では重みの回転を

扱う。古典的な Quartimax 回転により、解は一意に定められる。1.2.5 では、主成分負荷行列等、結果の解釈に役立つ統計測度を導入する。1.2.6 では解釈を促進するために、順序のあるカテゴリカル変数のカテゴリーを直交多項式によって数量化する方法について説明する。1.2.7 ではモデルの説明力について議論する。1.2.8 では分析の流れを説明する。

2 節では、1 節で定式化された方法を実データに適用した結果を示し、解釈について述べる。3 節では bootstrap 法による推測統計学的扱いについて、2 節のデータを用いて検討する。4 節では方法と実データへの適用の結果について討論し、残された問題について言及する。

## 1.2 多重対応分析への因子分析的観点の導入

### 1.2.1 主成分分析

本研究における主成分分析とは次のようなものである。まず、適用の対象となるデータの記述から始めよう。調査対象者の数を  $n$ 、量的な変数の数を  $p$  とし、 $n \times p$  のデータ行列を  $\mathbf{X}$  と書く。 $\mathbf{X}$  の各列の平均値のベクトルを  $\bar{\mathbf{x}}$ 、すべての要素が 1 であるような  $n$  次元ベクトルを  $\mathbf{1}_n$ 、 $\mathbf{X}$  の各列の標準偏差のベクトルを  $\mathbf{s}$  として  $\mathbf{Z} = (\mathbf{X} - \mathbf{1}_n \bar{\mathbf{x}})(\text{diag } \mathbf{s})^{-1}$  とすると、この  $\mathbf{Z}$  は、 $\mathbf{Z}'\mathbf{1}_n = \mathbf{0}$ 、 $\text{dg}\mathbf{Z}'\mathbf{Z} = n\mathbf{I}_p$  という条件を満たすように標準化される。

続いて、主成分を定義する。 $\mathbf{W}$  を  $p \times r$  ( $r < p$ ) の正規直交な重み行列、すなわち、 $\mathbf{W}'\mathbf{W} = \mathbf{I}$  をみたすものとする。主成分得点は、標準化された変数の 1 次合成変量 (linear composites) として次のように定義される。

$$\mathbf{F} = \mathbf{Z}\mathbf{W} \quad (1)$$

最適化基準としては、主成分得点の分散の和の最大化を採用する。すなわち、

$$\psi(\mathbf{W}|\mathbf{Z}) = \text{tr } n^{-1}\mathbf{F}'\mathbf{F} = \text{tr } \mathbf{W}'\mathbf{R}\mathbf{W} \quad (2)$$

ここで、 $\mathbf{R} = n^{-1}\mathbf{Z}'\mathbf{Z}$  であり、これは変数間相関行列である。

この解は、相関行列の固有値分解、 $R = KAK'$ 、すなわち  $A$  を  $R$  の固有値を降順に並べた対角行列、 $K$  を対応する正規直交な固有ベクトルの行列 ( $KK' = I = K'K$ ) としたとき、

$$W = KT \quad (3)$$

である。ただし、 $K_r$  は  $K$  の  $r$  番目までの列からなる正規直交行列、 $T$  は任意の直交行列 ( $T'T = I = TT'$ ) である。なお、 $R$  は正定値で固有値はすべて正で相互に異なる (distinctive) ものとする。このとき、最大化された主成分の分散の和は、

$$\text{tr } n^{-1}F'F = \text{tr } W'RW = \text{tr } T'K_r'A_rK_rT = \text{tr } A_r$$

すなわち、降順に  $r$  番目までの固有値の和となる。

この定式化 (たとえば, Ten Berge, 1993) は、通常の主成分得点の分散の最大化という基準と比べると、変数の標準化を行っていることと、主成分得点の直交性に関する制約 (あるいは、自然に成立する性質) を外した点で異なる。標準化は、後に多重対応分析と対応をつけるために必要な変換である。主成分得点の直交しない (相互に相関をもつ) 点については、 $T$  による直交回転により重み行列の直交性は  $(WT)'WT = T'WWT = T'T$  によって保たれる一方、主成分得点に関しては、 $\text{tr } n^{-1}F'F = W'RW = T'K_r'A_rK_rT$  が、対角行列とはならないことによる。この点については、1.2.4 において再説する。

また、 $\|Z - FW'\|^2 = p - \text{tr } A_r$  だから、この解は標準得点の行列を  $FW'$  という形の因子モデルで近似するという意味ももつ。

### 1.2.2 多重対応分析

続いて、本研究における多重対応分析を定式化する。カテゴリカル変数の行列の第  $k$  列  $x_k$  を 2 値の形に変換した行列  $G_k$  を、変数  $k$  のダミー変数行列と呼ぶ。カテゴリカルな変数の値が連続した正の整数で与えられてい

れば、

$$g_{ijk} = \begin{cases} 1 & \text{if } x_{ik} = j \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

である。変数  $x_k$  の値の種類 (最大値) を  $c_k$  とすると、 $\mathbf{G}_k$  は  $n \times c_k$  の行列で、 $\mathbf{G}_k \mathbf{1}_{c_k} = \mathbf{1}_n$  であり  $c_k$  次元ベクトル、 $\mathbf{d}_k = \mathbf{G}'_k \mathbf{1}_n$  は各カテゴリーへの反応数を要素とする。また、

$$\mathbf{G} = [\mathbf{G}_1 \quad \mathbf{G}_2 \quad \cdots \quad \mathbf{G}_p] \quad (5)$$

は  $n \times c$  の行列である。ここで、 $c = c_1 + c_2 + \cdots + c_p$ 、 $\mathbf{D}_k = \text{diag } \mathbf{d}_k$  とする。また、 $\mathbf{D}$  は  $\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_p$  を要素とする対角行列とする。ここで、 $\mathbf{G} \mathbf{1}_c = p \mathbf{1}_n$  であり、 $\mathbf{d} = \mathbf{G}' \mathbf{1}_n$  は全カテゴリーの反応数の  $c$  次元ベクトルである。また、 $\mathbf{d} = \mathbf{D} \mathbf{1}_c$ 、 $\mathbf{1}'_c \mathbf{D} \mathbf{1}_c = np$  である。

各カテゴリーに  $r$  次元の数量を与え、各個体が選択したカテゴリーの数量の和をその個体の数量化得点 (quantified variables) とする。すなわち、

$$\ddot{\mathbf{F}} = \mathbf{G} \mathbf{V} \quad (6)$$

である。数量化行列 (quantification matrix)  $\mathbf{V}$  には、 $n^{-1} \mathbf{V}' \mathbf{D} \mathbf{V} = \mathbf{I}$  の制約条件を課す。また、数量化得点の各列は中心化されているものとする。すなわち  $\ddot{\mathbf{F}}' \mathbf{1}_n = \mathbf{0}$  である。

最適化基準は、 $\ddot{\mathbf{F}}$  の各列の分散の平方和、

$$\varphi(\mathbf{V} | \mathbf{G}) = \text{tr } n^{-1} \ddot{\mathbf{F}}' \ddot{\mathbf{F}} = \text{tr } \mathbf{V}' (n^{-1} \mathbf{G}' \mathbf{G}) \mathbf{V} \quad (7)$$

の最大化である。この解は次のようにして求められる。まず、 $\mathbf{B} = \mathbf{D}^{-1/2} \mathbf{G}' \mathbf{G} \mathbf{D}^{-1/2}$  と定義する。この行列  $\mathbf{B}$  は基準化 Burt 行列と呼ばれる (Gower & Hand, 1996)。

ここで、 $c$  次元ベクトル  $\mathbf{l}_0 = (np)^{-1/2} \mathbf{D}^{1/2} \mathbf{1}_c$  を定義する。 $\mathbf{1}'_c \mathbf{D} \mathbf{1}_c = np$  により、 $\mathbf{l}_0$  は単位長さのベクトル、すなわち、 $\mathbf{l}'_0 \mathbf{l}_0 = 1$  である。これを  $\mathbf{B}$  の右からかけると、

$$\mathbf{B}\mathbf{l}_0 = (np)^{-1/2}\mathbf{D}^{-1/2}\mathbf{G}'\mathbf{G}\mathbf{1}_c = p^{1/2}n^{-1/2}\mathbf{D}^{-1/2}\mathbf{G}'\mathbf{1}_n = p^{1/2}n^{-1/2}\mathbf{D}^{1/2}\mathbf{1}_c = p\mathbf{l}_0$$

すなわち、 $\mathbf{B}\mathbf{l}_0 = p\mathbf{l}_0$ であるから $\mathbf{l}_0$ は $\mathbf{B}$ の(基準化された)固有ベクトルであり、対応する固有値は $p$ である。この $p$ は、 $\mathbf{B}$ の最大固有値であることが証明できる(Gower & Hand, 1996)。そこで、 $\mathbf{C} = \mathbf{B} - p\mathbf{l}_0\mathbf{l}_0'$ と定義し、その固有値分解を、 $\mathbf{C} = \mathbf{L}\mathbf{A}\mathbf{L}'$ とする。 $\mathbf{C}$ の固有値は $\mathbf{B}$ の最大固有値である $p$ を除き、 $\mathbf{B}$ の固有値と一致する。このとき、数量化行列は、

$$\mathbf{V} = n^{1/2}\mathbf{D}^{-1/2}\mathbf{L}_r\ddot{\mathbf{T}} \quad (8)$$

によって与えられる。ここで、 $\mathbf{L}_r$ は $\mathbf{C}$ の大小順に $r$ 番目までの固有値に対応する正規直交化された固有ベクトルの行列である。また、 $\ddot{\mathbf{T}}$ は任意の直交行列である。

これによって求められた数量化得点 $\ddot{\mathbf{F}}$ が中心化されていることは、各列の和が、

$$\ddot{\mathbf{F}}'\mathbf{1}_n = \mathbf{V}'\mathbf{G}'\mathbf{1}_n = n^{1/2}\ddot{\mathbf{T}}'\mathbf{L}'_r\mathbf{D}^{-1/2}\mathbf{d} = n^{1/2}\ddot{\mathbf{T}}'\mathbf{L}'_r\mathbf{D}^{1/2}\mathbf{1}_c = n\ddot{\mathbf{T}}'\mathbf{L}'_r\mathbf{l}_0$$

となり、 $\mathbf{L}'_r\mathbf{l}_0$ は $\mathbf{B}$ の最大固有値に対応する固有ベクトルと2番目から $r+1$ 番目までの固有値に対応する固有ベクトルとの内積で、これは(固有値が相異なるものである限り) $\mathbf{0}$ となることから確認できる。したがって、最大化された数量化得点の分散の和は、

$$\text{tr } n^{-1}\ddot{\mathbf{F}}'\ddot{\mathbf{F}} = \text{tr } (\mathbf{D}^{-1/2}\mathbf{L}_r\ddot{\mathbf{T}})'\mathbf{G}'\mathbf{G}(\mathbf{D}^{-1/2}\mathbf{L}_r\ddot{\mathbf{T}}) = \text{tr } \ddot{\mathbf{T}}'\mathbf{L}'_r\mathbf{B}\mathbf{L}_r\ddot{\mathbf{T}} = \text{tr } \mathbf{A}_r$$

で与えられる。ただし、 $\mathbf{A}_r$ は $\mathbf{C}$ の最大 $r$ 番目までの( $\mathbf{B}$ の2番目から $r+1$ 番目までの)固有値を要素とする対角行列である。

他方、 $\mathbf{B}$ の最大の固有値 $p$ に対応する固有ベクトルである $\mathbf{l}_0$ を重みとした得点は中心化されない。実際、(8)式にならって、 $\mathbf{v}_0 = n^{1/2}\mathbf{D}^{-1/2}\mathbf{l}_0 = p^{-1/2}\mathbf{1}_c$ によって重みを定義すると、 $\mathbf{G}\mathbf{1}_c = p\mathbf{l}_n$ により、 $\mathbf{G}\mathbf{v}_0 = p^{1/2}\mathbf{1}_n$ となり、全個



体の得点が  $p^{1/2}$  と一定になる。 $\mathbf{G}\mathbf{v}_0$  は MCA の自明解と呼ばれ、解釈の対象とはならない。

この定式化は、数量化得点の単位<sup>3</sup>、および直交回転の可能性を導入している点を除けば、MCA のオーソドックスな導出手続きと言える (Gower & Hand, 1996)。

### 1.2.3 正規直交主成分分析

質的変数のカテゴリーに数量を与えて主成分分析的に扱う方法として比較的よく知られているのが非線形主成分分析 (nonlinear principal component analysis; De Leeuw, 2006; 足立・村上, 2011; De Leeuw, 2014) である。この方法は、どのカテゴリーにも 1 次元の数量を与えて主成分分析を行う。もちろん、数量は (本研究の表現では) 分散の和が最大になるように与えられるので、そのための反復的アルゴリズムが必要になる。

しかし、後の適用例で見ると、すべての変数に 1 次元の数量しか与えなければ、データにおける重要な変動 (変数間の関係) が見逃される恐れがある。それ以上に、単なるアーチファクト以上の馬蹄現象が含まれるようなデータでは、カテゴリーに割り当てられる成分がすべて非線形成分となり、内容的側面が全く無視される危険性すらある (村上, 2016)。ここでは、すべての変数に  $(c_k - 1)$  次元の数量を与える方法を提案する。

そのために、各変数について、 $c_k \times (c_k - 1)$  の行列  $\mathbf{U}_k$  を導入する。この要素は、 $\mathbf{U}'_k \mathbf{d}_k = \mathbf{0}$ ,  $\mathbf{U}'_k \mathbf{D}_k \mathbf{U}_k = n\mathbf{I}$  を満たすという意味で、中心化、正規直交化がなされている (単なる  $c_k$  個の数値の平均値や共分散でなく、度数で重みづけた加重平均と共分散で考える) 以外は任意であるものとする。これによって、各変数を次式により  $c_k - 1$  次元の数量化を行う。

$$\mathbf{Z}_k = \mathbf{G}_k \mathbf{U}_k \quad k = 1, \dots, p \quad (9)$$

この行列を要素とする  $n \times (c - p)$  のデータ行列  $\mathbf{Z} = [\mathbf{Z}_1 \quad \mathbf{Z}_2 \quad \dots \quad \mathbf{Z}_p]$

と、正規直交な  $(c-p) \times r$  の重み行列  $\mathbf{W}$  を用いて、2.2.1 の (1) 式によって主成分得点  $\mathbf{F} = \mathbf{Z}\mathbf{W}$  を定義する。 $\mathbf{Z}$  は新たに定義される  $c-p$  個の列からなるわけであるが、これらには、本来の ( $p$  個の) 「変数 (variable)」と区別して「変量 (variate)」の語を充てることにしたい。

最大化基準は (2) 式であるが、この場合、(変量間) 相関行列は  $\mathbf{R}_{kl} = n^{-1}\mathbf{Z}'_k\mathbf{Z}_l$  を部分行列とする  $c-p$  次の正方行列、

$$\mathbf{R} = n^{-1}\mathbf{Z}'\mathbf{Z} = \begin{bmatrix} \mathbf{I} & \mathbf{R}_{12} & \cdots & \mathbf{R}_{1p} \\ \mathbf{R}_{21} & \mathbf{I} & \cdots & \mathbf{R}_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{R}_{p1} & \mathbf{R}_{p2} & \cdots & \mathbf{I} \end{bmatrix} \quad (10)$$

という形をとる。対角ブロックは数量化行列  $\mathbf{U}_k$  の (重みのついた) 正規直交性により、すべて  $c_k - 1$  次の単位行列となる。

固有値分解  $\mathbf{R} = \mathbf{K}\mathbf{\Lambda}\mathbf{K}'$  により、重み行列は (3) 式によって得られる。ダミー変数行列からの算出には、

$$\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{U}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{U}_p \end{bmatrix} \quad (11)$$

として、

$$\mathbf{F} = \mathbf{Z}\mathbf{W} = \mathbf{G}\mathbf{U}\mathbf{K}'\mathbf{T}$$

となる。また、この際の最大化基準の値が、 $\mathbf{R}$  を  $\mathbf{R} = \mathbf{K}\mathbf{\Lambda}\mathbf{K}'$  と固有値分解した場合における  $\text{tr } \mathbf{\Lambda}_r$  によって得られることも同様である。

こうして、この正規直交主成分分析 (orthonormal principal component analysis) と仮に名づける方法は、データ行列  $\mathbf{Z}$  の定義と相関行列  $\mathbf{R}$  の大きさの違いを除けば、通常の主成分分析と同じであるともいえる。ただし、数量化行列  $\mathbf{U}_k$  の値は中心化と正規直交化の制約条件を除いて任意としたから、このままでは実際に有用な解釈を生み出すことはできない。

しかしながらこの方法は、このままで  $\mathbf{R}$  の固有値が多重対応分析のそれと一致し、主成分得点行列  $\mathbf{F}$  は、直交回転を除き多重対応分析の数量化得点行列  $\mathbf{F}$  と一致するという性質をもつ。この意味で、質的データの分

析方法である多重対応分析と量的データの分析方法である主成分分析の間の橋渡しをする役割を果たすのである。以下にそのことを示そう。

まず、数量化行列  $U_k$  にベクトル  $\mathbf{1}_c$  を付け加えた  $c_k \times c_k$  の (正方) 行列、 $U_k^+ = [U_k \ \mathbf{1}_{c_k}]$  を考えよう。 $U_k$  の中心化と標準化に関する条件  $U_k^+ \mathbf{d}_k = \mathbf{0}$ ,  $U_k^+ D_k U_k = nI$  により、行列  $U_k^* = n^{-1/2} D_k^{1/2} U_k^+$  は直交行列、すなわち、 $U_k^{*'} U_k^* = I_c = U_k^* U_k^{*'}$  である。そこで、 $U_k^*$  を対角ブロックにもつ行列

$$U^* = \begin{bmatrix} U_1^* & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & U_2^* & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & U_p^* \end{bmatrix}$$

も直交行列 ( $U^{*'} U^* = I = U^* U^{*'}$ ) である。このとき、行列  $U^{*'} B U^*$  は  $B$  と同じ固有値をもつ。そのことを示そう。まず  $B$  は半正定値の行列でその固有値分解は  $B = L \Delta L'$  である。ただし、 $\Delta$  は  $C$  の固有値を降順に並べた対角行列  $\Delta$  に、自明解に対応する固有値  $p$  を加えた  $c-p+1$  次の対角行列である。また、 $U^*$  は  $B$  と同じ大きさの直交行列で、 $U^{*'} B U^* = U^{*'} L \Delta L' U^*$  である。さらに、このとき  $L' U^*$  もまた直交行列であり、加えて  $B$  のすべての固有値は相異なると仮定したから、この分解は一意的であり、 $U^{*'} L$  は  $U^{*'} B U^*$  のすべての固有ベクトルである。

次に、行列  $U^*$  の部分行列  $U_k^*$  の第  $c_k$  列、すなわち  $n^{-1/2} D_k^{1/2} \mathbf{1}_{c_k}$  の形をとる列を行列の最右端に入れ替える。この入れ替えによって、 $U^*$  は次の形になる。

$$U^+ = n^{-1/2} D^{1/2} \begin{bmatrix} U_1 & \mathbf{0} & \cdots & \mathbf{0} & J_1 \\ \mathbf{0} & U_2 & \cdots & \mathbf{0} & J_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & U_p & J_p \end{bmatrix} = n^{-1/2} D^{1/2} [U \ J]$$

ここで、 $J_k$  は  $c_k \times p$  の行列で、その第  $k$  列は  $\mathbf{1}_{c_k}$ 、他の列は  $\mathbf{0}$  の行列である。 $U^+$  は  $U^*$  の列を入れ替えただけのものだから、やはり直交行列である。

そこで、 $R^+$  を

$$R^{\dagger} = U^{\dagger'} B U^{\dagger} = \begin{bmatrix} R & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_p \mathbf{1}'_p \end{bmatrix}$$

と定義する。行列  $\mathbf{1}_p \mathbf{1}'_p$  は、すべての要素が1であるような  $p$  次の正方行列で、その最大固有値は  $p$ 、対応する固有ベクトルは  $\mathbf{1}_p$  であり、残りの固有値はすべて0、固有ベクトルは  $\mathbf{1}_p$  と直交する限り任意となる。

したがって、固有値分解  $R^{\dagger} = K^{\dagger} \Lambda^{\dagger} K^{\dagger'}$  においては、 $R$  の部分と  $\mathbf{1}_p \mathbf{1}'_p$  の部分が完全に分離され、

$$\Lambda^{\dagger} = \begin{bmatrix} \Lambda & \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & p & \mathbf{0}' \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

および、

$$K^{\dagger} = \begin{bmatrix} K & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & p^{-1/2} \mathbf{1}_p & K_a \end{bmatrix}$$

となる。すなわち、(10)式で定義される相関行列  $R = K \Lambda K'$  の分解において、 $\Lambda = \Delta$ 、すなわち、正規直交主成分分析の固有値と多重対応分析の（自明解に対応する  $p$  を除く）固有値とが一致することがわかった。

次に、数量化得点と主成分得点の関係について考えよう。ダミー変数行列から計算するとして、数量化得点は、

$$\ddot{F} = G V = n^{1/2} G D^{-1/2} L_r \ddot{T}$$

主成分得点は、

$$F = G U K_r T = n^{1/2} G D^{-1/2} (n^{-1/2} D^{1/2} U K_r) T$$

であり、これらの式の  $L_r$  と  $(n^{-1/2} D^{1/2} U K_r)$  は、ともに基準化 Burt 行列  $B = D^{-1/2} (G' G) D^{-1/2}$  の降順に2番目から  $r+1$  番目までの固有値に対応する基準化された固有ベクトルを要素とする。したがって、直交回転  $\ddot{T}$ 、 $T$  に関する不定性を除いて、主成分得点は数量化得点と等しいことがわかる。

最後に行列  $U_k$  の求め方であるが、ここでは最も初等的な Gram-Schmidt の直交化法を用いる。まず任意の乱数による  $c_k \times (c_k - 1)$  の行列を  $Q_k$  とする。 $Q_k$  の階数は  $c_k - 1$  と想定できるから、その各列は 1 次独立である。まず、中心化のためのベクトルを  $u_k^{(0)} = \mathbf{1}_{c_k}$  と定義する。その上で、

$$\dot{u}_k^{(l)} = q_k^{(l)} - n^{-1} \sum_{m=0}^{l-1} (q_k^{(l)'} D_k u_k^{(m)}) u_k^{(m)} \quad (12)$$

ただし、 $q_k^{(l)}$  は  $Q_k$  の第  $l$  列、 $u_k^{(m)}$  ( $m > 0$ ) は行列  $U_k$  の第  $m$  列である。また、

$$u_k^{(l)} = \dot{u}_k^{(l)} \left( n^{-1} \dot{u}_k^{(l)'} D_k \dot{u}_k^{(l)} \right)^{-1/2} \quad (13)$$

とする。すなわち、 $m = 0$  において平均値  $n^{-1} d_k' q_k^{(l)}$  を引いて中心化した上で、先行するすでに中心化、(13 式の意味で) 基準化されている数量化ベクトルと直交するように、順次変換し、最後に (13 式) によって基準化するという手順である。

最終的にこれらを並べた  $U_k = [u_k^{(1)} \quad u_k^{(2)} \quad \dots \quad u_k^{(c_k-1)}]$  を (11 式) のブロック対角行列の形に並べたものが  $U$  となる。

以上の過程を、2 値変数を除くすべての変数について繰り返し、(11 式) によって配列した上で、 $Z = GU$ 、 $R = n^{-1} Z'Z$ 、 $R = K\Lambda K'$  によって主成分分析の手続きを進めることになる。

2 値変数に関しては、 $p_k$  を項目 (変数)  $k$  の反応率とすると、制約条件だけで、

$$u_k' = \left[ -p_k / \sqrt{p_k(1-p_k)} \quad (1-p_k) / \sqrt{p_k(1-p_k)} \right] \quad (14)$$

と定まる。このため、上の過程を適用する必要がない<sup>4</sup>。

### 1.2.4 重みの直交回転による解の一意性の確立

前述のように、正規直交多項式主成分分析の数量化行列  $U_k$  の要素は、 $U_k \mathbf{d}_k = \mathbf{0}$ ,  $U_k' \mathbf{D}_k U_k = nI$  という制約条件を満たす限り任意である。変数が 2 値の場合に限れば、 $u_k$  の値は一意に定まるが、それ以外の場合には値を定める根拠がない。たとえば、分析を実行する都度、乱数を制約条件に合うように変換しても解としては成立するが、そこから何らかの意味のある解釈を導き出すことはできないであろう。

主成分得点を求めるための重み  $W$  には右からの直交変換  $T$  に加えて、

$$U_k W_k T = U_k \tilde{T}_k \tilde{T}_k' W_k T$$

という左からの直交変換  $\tilde{T}_k$  も存在し、これら 2 種（合計  $p+1$  個）の直交回転を定めることができれば、解は一意に定まることになる。

直交回転を定める基準としては、解釈の基礎を空間から軸に移すという目的から考えれば、項目の重みを特定の主成分に集中させ、その集中した（目立つ salient）負荷量を多くの主成分に分散させるという単純構造（simple structure）が適当と思われる。単純構造の最も簡明な定義は芝（1979）による次のものである。

- 1) 各変数は 1 つの因子にだけ高い負荷を示す。
- 2) 各因子において高い負荷を示す変数の数はほぼ等しくなる。

単純構造を実現するための解析的直交回転の方法としては、次の Orthomax 基準を最大化する方法がよく知られている。 $W$  の右からの回転行列  $T$  について考える。また、2 次元の組み合わせに限定する（平面上での回転であり planar rotation と呼ばれる）と、変量  $j$  の主成分  $l$  への重みを  $w_{jl}$ 、回転後の対応する要素を  $w_{jl}^*$  として、

$$w_{j1}^* = w_{j1} \cos\theta - w_{j2} \sin\theta$$

$$w_{j2}^* = w_{j1} \sin\theta + w_{j2} \cos\theta$$

であるが、 $\mathbf{w}_j = [w_{j1} \ w_{j2}]'$ 、 $\mathbf{t}_1 = [\cos\theta \ -\sin\theta]'$ 、 $\mathbf{t}_2 = [\sin\theta \ \cos\theta]'$ と定義すれば、 $w_{j1}^* = \mathbf{w}_j' \mathbf{t}_1$ 、 $w_{j2}^* = \mathbf{w}_j' \mathbf{t}_2$ と書けるが、このとき Orthomax 基準とは、

$$f(\theta|\mathbf{W}) = \sum_{l=1}^2 \sum_{j=1}^p (\mathbf{w}_j' \mathbf{t}_l)^4 - \frac{\gamma}{p} \left( \sum_{l=1}^2 \sum_{j=1}^p (\mathbf{w}_j' \mathbf{t}_l)^2 \right)^2 \quad (15)$$

である。ここで、 $\gamma = 1$  とすると Varimax 回転、 $\gamma = 0$  とした場合が Quartimax 回転である。基準を考えるにあたっては、対象となる行列（ここでは重み行列  $\mathbf{W}$ ）の要素の符号の影響を回避するために、要素はすべて2乗したうえでその分散を考えているのであるが、Quartimax 回転では  $\mathbf{W}$  の各行の分散の和を、Varimax 回転では各列の分散の和の最大化を行う。通常行われる負荷行列の回転では、Varimax 基準の方が上の単純構造の基準2)を満たす上で有利なり、それが探索的分析における回転方法として広く使われる理由になった (村上、1998)。しかしながら、重み行列の場合、 $\mathbf{W}'\mathbf{W} = \mathbf{I}$  であるから、行列の各列の要素の2乗の平均値は、直交回転の下では常に1となる。また、行方向の2乗和も回転によって変化しない。なぜなら、 $\mathbf{W}\mathbf{W}' = \mathbf{K}_r \mathbf{T}\mathbf{T}' \mathbf{K}_r' = \mathbf{K}_r \mathbf{K}_r'$  だからである。すなわち、(15)式の右辺第2項は一定であり、Varimax 基準と Quartimax 基準に違いは生じない。そこで、ここでは Quartimax 回転、すなわち、(15)式の右辺第1項である重み行列の要素の4乗和の最大化を行うものとする。

左からの回転については、必ずしも上の議論は成立しないが、両方向からの回転に同じ基準を適用するのが自然であろう。両方向からの回転は、交互に収束するまで反復される。この過程は単調減少であることが容易にわかり、かつ上限が存在するから収束は保証されている。

なお、この重み行列によって回転された因子得点行列  $\mathbf{F}$  の分散共分散行列は、 $n^{-1}\mathbf{F}'\mathbf{F} = \mathbf{T}'\mathbf{A}\mathbf{T}$  が対角行列とはならないことから、因子得点間にゼロでない相関が発生する。すなわちこれは斜交回転である。

### 1.2.5 パターン行列とその他の解釈の手掛かり

負荷行列でなく重み行列を回転するという手続きは、我々の定式化から自然に出てきたものであるが、この方式は必ずしも新しいものではない (Harris & Kaiser, 1964; Kiers & Ten Berge, 1994)。かつ、主成分分析の回転方法としては、複数の長所もある。回転が行われた結果を解釈するための統計測度について議論しよう。まず、多重対応分析による数量化得点  $\dot{F}$  と回転を除いて同じである主成分得点  $F$  の (列) 間の相関係数を定義する。これは  $n^{-1}F'F$  が共分散行列であることから、

$$\Phi = (\text{dg } W'RW)^{-1/2} W'RW (\text{dg } W'RW)^{-1/2}$$

によって得られる。

主成分、あるいは一般に因子分析によって得られた因子の解釈には、通常、個別変量の主成分への標準偏回帰係数である因子負荷量が用いられる。これは、

$$A = n^{-1/2} Z'F (\text{dg } n^{-1/2} F'F)^{-1/2} \Phi^{-1} = RW(W'RW)^{-1} (\text{dg } W'RW)^{1/2}$$

のように算出される。ここで、 $RW(W'RW)^{-1} = K_r ATT' A_r^{-1} T = W$  だから、

$$A = W (\text{dg } W'RW)^{1/2} \quad (16)$$

となり、 $A$  は列ごとに  $W$  に比例することになる。これは他の回転方法を適用した主成分分析では成立しない性質であるが、少なくとも、記述の節約(求めるべきパラメータが少なくなる)という点で優れた性質と言える。さらに (16) から

$$A'A = \text{dg } W'RW = \text{dg } n^{-1} F'F \quad (17)$$

であり、パターン行列の各列の要素の 2 乗和は主成分得点の分散、あるいは主成分寄与を表すことになる<sup>5</sup>。これは、一般の斜交回転のパターン行列では成り立たない性質であり、回転後の主成分の「重要度」を評価でき



るという意味で、極めて貴重な性質である。

各変数の主成分への  $A$  の対応する行を標準回帰係数とする重回帰方程式の決定係数 (重相関係数の 2 乗) は、

$$R_j^2 = \mathbf{w}_j' \Lambda_r \mathbf{w}_j \quad j = 1, \dots, c-p$$

によって得られる。ただしこれは、通常の直交解における負荷行列の行の平方和  $\mathbf{a}_j' \mathbf{a}_j$  とは一致しない。

### 1.2.6 正規直交多項式の利用

以上のようにして、一応、多重対応分析の多次元の解を、心理測定における (広義の) 因子分析的に取り扱う方法を確立することができたと考えられる。しかしながら、実データにこれを適用してみると、あまり使い勝手がよくないことがわかる。その原因は、負荷行列 (パターン行列) を解釈するにあたって、いちいち数量化行列  $\mathbf{U}_k$  を参照しなければならないことにある (2 値変数以外)。通常の因子分析では、扱われているのは量的変数であり、変数の意味はそのままで明らかなのであるが、上で述べた方法の場合、 $c_k - 1$  個の変数の意味、すなわち変数  $k$  の各カテゴリーと  $c_k - 1$  次元の変量として与えられた数値との関係は  $\mathbf{U}_k$  の中身を見ない限りわからないからである。

その点の改善のために、順序のあるカテゴリカル項目については、直交多項式を導入する可能性について考える (村上, 2016)。すなわち、2.2.3 で定義した中心化、直交化する前の数量化ベクトルとして、

$$\mathbf{q}_k^{(l)} = [1^l \quad 2^l \quad \dots \quad c_k^l] \quad l = 1, \dots, c_k - 1 \quad (18)$$

とすることである。順序のあるカテゴリーに 1 から  $c_k - 1$  次式をあてはめるのである。もちろん、中心化、正規直交化するために (12)、(13) 式の手順を踏んで  $\mathbf{q}_k^{(l)}$  は  $\mathbf{u}_k^{(l)}$  に変換されて一種の正規直交関数になる。

多重対応分析は、時にカテゴリカルデータの非計量的多変量解析法に分

類される（足立・村上，2011）が、(18)式は完全に「計量的」であり、本来の目的に反していると考えられるかもしれない。

しかしながら、すでに示されたように、1.2.3の正規直交主成分分析は、数量化行列  $U_k$  をどのように与えても非計量的な方法とされる多重対応分析と実質的に同じ解を与えるのである。実際、上記の手順で導出される重み行列の加重の仕方によって、カテゴリーの（単調増加、あるいは単調減少のものに限らず）すべての可能な関数が表現できる。正規直交多項式の実例については、村上（2016）、あるいは、本論文2.2の図1で、年齢段階に対して定義された1から5次式を見ていただきたい。これらの新たに定義された変量は、次数の低いものから順に、1次変量、2次変量等とよぶことにする。

なお、正規直交多項式によって正規直交主成分分析の解釈は大幅に促進されるが、主成分ごとに、変量の次数ごとの寄与を計算できるという副産物も得られる。なお、順序のあるカテゴリカル変数に直交多項式を当てはめるというアイデアは、通常の（2変数の）対応分析への適用例が Beh & Rombaro（2014）に記載されているものの、多重対応分析への適用は村上（2016）を除いて先行研究には見当たらない。

### 1.2.7 説明力の検討

通常の主成分分析と同様に、数量化されたデータを主成分得点と重みの積によって近似するという考え方をとれば、 $\|GU - FW'\|^2 = c - p - \text{tr} A_r$  であるから、採用した  $r$  個の固有値の和を変量の数で割った比率が当てはまりのよさの指標となりえる。この数値は回転後における  $\text{tr} A'A / (c - p)$  を用いても同じである。

しかしながら、多重対応分析ではより質的データに密着した当てはまりの指標が考えられている。ダミー変数行列にその転置を左から掛けて得られる次の Burt 行列を考えよう。

$$G'G = \begin{bmatrix} G'_1G_1 & G'_1G_2 & \cdots & G'_1G_p \\ G'_2G_1 & G'_2G_2 & \cdots & G'_2G_p \\ \vdots & \vdots & \ddots & \vdots \\ G'_pG_1 & G'_pG_2 & \cdots & G'_pG_p \end{bmatrix}$$

この要素である  $G'_kG_{k'}$  は変数 (項目)  $k$  と  $k'$  の間のクロス集計表である。Burt 行列はこのようにクロス集計表を配列したものである。この左右から  $D^{-1/2}$  を掛けて得られるのが基準化 Burt 行列であり、これから  $p\mathbf{l}_0\mathbf{l}'_0$  を引いた行列  $C$  の同じブロックの形は、 $D_k^{-1/2}G'_kG_{k'}D_{k'}^{-1/2} - n^{-1}D_k^{1/2}\mathbf{1}_{c_k}\mathbf{1}'_{c_{k'}}D_{k'}^{1/2}$  となるが、この要素は、

$$\chi_{ij} = \frac{n_{ij} - n_i \cdot n_j / n}{\sqrt{n_i \cdot n_j}}$$

と書ける。ただし、 $i$  は変数  $k$  の、 $j$  は変数  $k'$  のそれぞれ 1 つのカテゴリーとし、 $n_{ij}$  はクロス集計表の  $i$  行  $j$  列の要素 (項目  $k$  においてカテゴリー  $i$ 、項目  $k'$  においてカテゴリー  $j$  を選択した個体の数)、 $n_i$  と  $n_j$  は対応する行、列の周辺度数である。ここで、

$$In(k, k') = \sum_{i=1}^{c_k} \sum_{j=1}^{c_{k'}} \chi_{ij}^2$$

は、イナーシア (inertia) と呼ばれ、2 つの項目間の関連度の 1 つの指標となる。なお、イナーシアの  $n$  倍が当該のクロス集計表のカイ 2 乗統計量にあたる。

このとき、任意の中心化、正規直交化された重み  $U_k$  によって数量化された得点行列  $Z_k = G_k U_k$  による  $(c_k - 1) \times (c_{k'} - 1)$  の相関行列  $R_{kk'}$  の要素の 2 乗和についても、

$$In(k, k') = \text{tr } R'_{kk'} R_{kk'}$$

が成立する。なぜなら、2.2.3 で導入した  $c_k$  次の直交行列  $\mathbf{U}_k^* = n^{-1/2} \mathbf{D}_k^{1/2} \mathbf{U}_k^+$  により、

$$\mathbf{U}_k^{*'} \mathbf{C}_{kk'} \mathbf{U}_{k'}^* = \begin{bmatrix} \mathbf{R}_{kk'} & \mathbf{0} \\ \mathbf{0}' & \mathbf{0} \end{bmatrix}$$

となるが、 $\ln(k, k') = \text{tr } \mathbf{C}_{kk'}' \mathbf{C}_{kk'}$   $= \text{tr } \mathbf{U}_k^{*'} \mathbf{C}_{kk'}' \mathbf{U}_{k'}^* \mathbf{U}_{k'}^{*'} \mathbf{C}_{kk'} \mathbf{U}_k^* = \text{tr } \mathbf{R}_{kk'}' \mathbf{R}_{kk'}$  だからである。このように変数間の連関に関しても、正規直交主成分分析によって、質的データの指標であるイナーシアを量的データの指標である相関係数の関数に置き換えることができた。

また、

$$\text{tr } \mathbf{R}^2 = \text{tr} \sum_{k=1}^p \sum_{k'=1}^p \mathbf{R}_{kk'}' \mathbf{R}_{kk'}$$

であり、かつ  $\mathbf{R}^2 = \mathbf{K} \mathbf{A} \mathbf{K}' \mathbf{K} \mathbf{A} \mathbf{K}' = \mathbf{K} \mathbf{A}^2 \mathbf{K}'$  であることから、 $\mathbf{R}^2$  の固有値は  $\mathbf{R}$  の固有値の 2 乗であるから、すべての変数の組み合わせに対するイナーシアの総和は、

$$\sum_{k=1}^p \sum_{k'=1}^p \ln(k, k') = \text{tr } \mathbf{A}_r \quad (19)$$

となる。これが分析中のデータの変数間の連関性のすべてを要約しており、説明されるべきものになる。このとき、主成分分析は、 $\|\mathbf{R} - \mathbf{A} \mathbf{\Phi} \mathbf{A}'\|^2 = \text{tr } \mathbf{R}^2 - \text{tr } \mathbf{A}_r^2$  の最小化と見ることもでき、したがって、説明されるイナーシアの大きさとデータに含まれるイナーシアの比率である  $\text{tr } \mathbf{A}_r^2 / \text{tr } \mathbf{A}^2$  をモデルの当てはまりのよさの指標と見ることもできる (Gower & Hand, 1996)。

しかしながら、(19) 式には本来説明の対象ではない  $k = k'$  のイナーシアも含まれていることが問題になる (Greenacre, 2017)。これは、モデルにおいて誤差分散を切り離して定義していない主成分分析において、相関行列の対角要素が 1 になっていることが、因子分析モデルとしては常に批

判の対象となることと同じである (たとえば, Pett, Lackey & Sullivan, 2003)。(10)式で見られるように、正規直交主成分分析では、対角要素は単位行列のブロックとなるので、このことの影響はさらに大きくなる。

本研究では、説明されるべきイナーシア (当てはまりのよさの指標の分母) を

$$2 \sum_{k=1}^{p-1} \sum_{k'=k+1}^p \ln(k, k') = \text{tr } \mathbf{A}^2 - (c - p)$$

と定義し、分子は、個々のブロックごとに、 $\mathbf{R}_{kk'} - \mathbf{A}_k \Phi \mathbf{A}'_{k'}$  の平方和を計算して ( $k \neq k'$  のケースについて) 総和したものをを用いる。ただし、この指標そのものは、多重対応分析によって最大化されたものではないので、注意が必要である<sup>6</sup>。

## 1.2.8 アルゴリズムの流れ

正規直交主成分分析の計算の流れは次のようになる。

- 1) データ行列  $\mathbf{X}$  を (4) 式により、ダミー変数行列  $\mathbf{G}$  に変換する。
- 2) 変数 (項目) を順序のないカテゴリカル項目 (タイプ1)、順序のあるカテゴリカル項目 (タイプ2)、2値項目に分類する (タイプ1とタイプ2の分類には任意性がある)。
- 3) タイプ1の項目のカテゴリには任意の乱数、タイプ2の変数には  $1 \sim (c_k - 1)$  次の多項式の数量を与える。さらに、(12)、(13) 式によって中心化、正規直交化し  $\mathbf{U}_k$  を定める。2値項目については、(14) 式によってベクトル  $\mathbf{u}_k$  を定める。
- 4) (9) 式によって各変数を正規直交化した変数を計算し、それをならべてできる標準得点の行列  $\mathbf{Z}$  を用いて、(10) 式によって正規直交変量間相関行列を求める。
- 5) 固有値分解により、固有値の対角行列  $\mathbf{\Lambda}$  と固有ベクトルを要素とする行列  $\mathbf{K}$  を得る。スクリープロット等を参照しながら、暫定的な主

成分数  $r$  を定める。最大  $r$  番目までの固有値に対応する固有ベクトルからなる行列  $K_r$  を重み行列  $W$  とする。

- 6)  $W$  を Quartimax 基準により回転する。タイプ 1 の項目については、左からの回転も行う。右からの回転と左からの回転を交互に収束するまで反復する。
- 7) (16) 式によってパターン行列  $A$ 、(17) 式によって主成分間相関行列  $\Phi$ 、各変数の決定係数  $R^2$ 、各主成分の寄与等を計算する。あてはまりのよさを 2 種の方法で計算する。
- 8) この段階で、主成分の符号、数量化ベクトルの符号を適切に変換すると、解釈が容易になる。
- 9) 主成分得点  $F$  を算出、外部変数との関連等を検討する。
- 10) 3 節で説明する bootstrap 法による負荷量の有意性を検討する。
- 11) 解の解釈を行い、必要に応じて 5) に戻り、異なる主成分数の数についても検討する。

## 2. 適用例：球場におけるプロ野球ファンの調査

### 2.1 適用例として用いられるデータ

1 節で説明された正規直交主成分分析を、2017 年 5 月 21 日にナゴヤドームにおける公式戦中日ドラゴンズ対ヤクルトスワローズの観客を対象に行われた調査データを用いる。この調査は、学生の「社会調査実習」の一環とし実施された<sup>7</sup>。調査の目的は、メディアの発達した今日、プロ野球ファンが球場で観戦する理由を探ることであり、「多様な人々がともに楽しめるナゴヤドームを目指して」をコンセプトとして、受講生とともに質問内容を構成した。具体的な項目内容に関しては、球場、球団とも十分な調整を行って最終的な調査項目を確定している。

実施にあたっては、中京大学現代社会学部の学生が 2 名 1 組で、調査参加者（回答者）となることを応諾してくれた観客に調査を実施する。手続きとしては、A4 版 3 ページ、26 項目からなる調査表を筆記用具とともに

参加者に手渡し、参加者自身に回答を記入してもらった。調査は開場から試合開始 30 分前までの 1 時間半にわたって行われた。学生の担当箇所は、観客席全体を座席数がほぼ均等になるように 40 箇所に分けて 1 組 1 箇所を指定した。調査担当者として参加した学生は、バックアップ要員も含めて 85 名であった。受講者以外にも、ボランティアとして調査に参加した学生も多いが、このうち 30 名は中京大学硬式野球部員であった。調査実施時間帯は相手チーム（ヤクルト）の打撃練習時間であり、観客席にはしばしばファウルボールが飛び込むため、その危険のある箇所については、調査担当者の 1 名は、回答者ともう 1 名の調査担当者を保護するために、グローブを持参していた。その主要メンバーが硬式野球部員であった。この日の調査に関する限り、特に危険な状況は発生しなかった。

調査終了後、受講者によってデータ入力が行われ、入力ミスや疑問のある回答については、慎重に修正作業がなされた。

分析のための計算には、自作のプログラム（MATLAB 言語による）を用いているが、ここでの計算の基礎となった多重対応分析の 8 次元解に関しては、SPSSv25/CATEGORIES により、固有値ならびに数量化得点が完全に一致していることを確認した<sup>8</sup>。

## 2.2 分析の対象となる変数

### 2.2.1 変数の選択

調査の質問項目のうちから、球場に訪れた観客の多様性を記述するのにふさわしいと思われる変数を選択した。できるだけ多くの側面から複数の次元を抽出するために、行動や意識に関わる項目だけでなく、デモグラフィックな変数も含めて以下の 16 の変数を用いた。

#### 1) デモグラフィックな変数

1. 性別
2. 年齢

### 3. 職業

#### 2) 来場以前のプロ野球関係の特性

4. 応援しているチーム
5. プロ野球関係のインターネットのアクセス頻度
6. ナゴヤドームでの今までの観戦回数

#### 3) 来場時の状況・総合的な満足度

7. 誰と来場しているか
- 8~15. (今日の) 観戦で楽しみにしていること (チェックリスト)
16. ナゴヤドームでの観戦満足度

これ以外にも、個々の観客の特性を表すのに適切と思われる項目（居住地域、プロ野球への関心度、ファンクラブ加入の有無、観覧席の位置）はあるが、これらのうち可能なものは抽出された次元の妥当性検討に用いることとした。

以下において、個々の変数について説明する。なお、この2017年調査の有効回答者数は874名であったが、分析対象とした変数のうち1つでも欠損値があった回答者は分析から除外したので、分析に用いられた回答者数は825であった。以下の反応%の数値の分母は825の方を用いている。以下、個々の変数について説明する。

#### 2.2.2 性別（2値変数）

2値変数であり、男性を1、女性を2とコーディングした。したがって、この変数が正に負荷する主成分は一般に女性の得点が高いことになる。男性の割合が、66.2%で、全体の約3分の2を占めている。2値変数では、平均値が0、標準偏差が1という制約条件により、カテゴリーに与えられる数量は反応率だけにもとづいて(14)式により一意に定まる。したがって、男性には-0.72、女性には1.40という数値が与えられる。



### 2.2.3 年齢段階 (タイプ2)

「10代」から「70代以上」の7段階で回答を求めているが、「60代」と「70代以上」は合併して「60代～」として扱った。6つの段階の回答者の割合は、それぞれ、11.2%、22.2%、13.8%、19.9%、17.9%、15.0%であった(回答者の3分の1は30歳未満である)。これらに、1次から5次までの直交多項式による5次元の数量が与えられる(図1)。

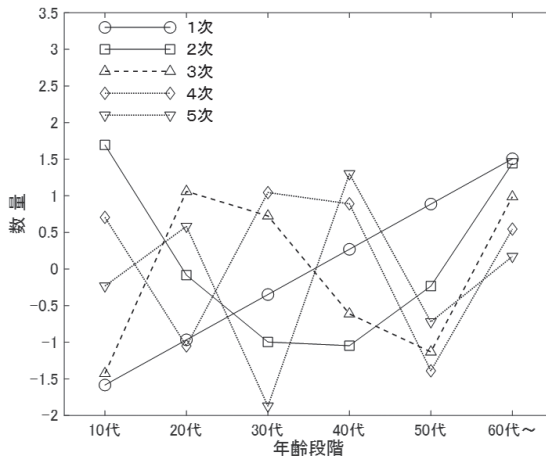


図1 年齢段階に対して定義された1～5次の直交多項式

年齢的变化は、比較的ゆるやかに起こると考えられるから、1次変量(直線)、2次変量(放物線)まではともかく、それ以上高次の数量は現実に役立つかどうか疑問に思われるかもしれない。しかし、少なくとも今回のデータに関するかぎり、文字通りギザギザな5次関数まで解釈可能な主成分を生み出すのに役立つのである。

### 2.2.4 職業 (タイプ1)

まず、職業については、どのようなカテゴリー化が適切であるかについて、論争がありうる。この調査では、試行錯誤的に、以下のフォーマット

に絞り込んできた。

Q.12 職業：【○は1つだけ】

1. 学生・生徒	2. 公務員	3. 会社員	4. パート・アルバイト
5. 専門職（教員・勤務医など）		6. 会社経営・会社役員	7. 自営業
8. 専業主婦・主夫	9. 無職・家事手伝い	10. その他（ ）	

さらに、反応数の少ない5～7は合併して「自営等」、さらに9と10も合併して「その他」として扱っている。

これらの選択肢は順序さえも定めることのできない質的カテゴリーなものであるから別の基準で数値を与えなければならない。後に実施した正規直交主成分分析によって与えられた数量を、各カテゴリーの反応率とともに表1に示した。

表1 「職業」の各カテゴリーに対して付与された数量と反応

変量	学生	公務員	会社員	パート等	自営等	専業主婦	その他	解釈
1	<b>-2.04</b>	-0.16	-0.04	0.17	0.38	1.00	<b>2.47</b>	学生(-) vs 無職(+)
2	-1.72	<b>0.64</b>	<b>0.59</b>	<b>0.45</b>	<b>0.36</b>	-0.46	-2.46	有職
3	0.10	-0.50	-0.43	<b>2.01</b>	-0.28	<b>2.39</b>	-1.29	パート・主婦
4	-0.01	<b>3.73</b>	-0.43	-0.82	0.56	0.64	-0.21	公務員
5	0.05	-1.16	0.23	-2.69	0.60	1.86	-0.82	
6	0.12	-0.77	-0.34	0.40	3.53	-0.57	-0.23	
反応率	12.2%	5.7%	53.0%	7.5%	6.9%	8.2%	6.4%	

例えば変量1は、「学生」に（絶対値の）大きな負の値、「その他」（無職含む）に大きな正の値が付与されており、学生（若者）と無職（高齢者）を対極にもつ次元である。変量2では、変量1で両端に分かれていた「学生」と「無職」がともに負の目立つ値をとる一方、「公務員」から「自営等」までの「収入を得るための職がある」人達が正の値をとっているので、「有職」と命名した。変量の分散は1に制約しているため、反応数の大きいカテゴリーには絶対値の大きい数値が付きにくい（過半数が「会社員」であ

る) ため、正の値はそれほど大きくないがこの解釈は納得されよう。変量 3 は、「パート等」と「専業主婦」という女性に選択されるカテゴリーに目立った値が付与された。変量 4 は、「公務員」と残りすべてとが対立する次元である。これは後に見るように主成分の寄与としては小さいが、興味ある解釈を生み出す。「公務員」は全体の 6% 弱 (47 人) なので、変量 2 の「会社員」とは逆にかなり極端な数値が与えられている。変量 5 と 6 については、主成分による説明力が極めて小さく、解釈上は無視される。

### 2.2.5 応援しているチーム (タイプ 2)

「ドラゴンズファンですか?」という質問への回答であるが、「いいえ」の場合には、応援しているチーム名を (あれば) 書いてもらった。相手チーム (この日はヤクルト・スワローズ) に 1、それ以外のチーム名、あるいは無記入の場合に 2、「はい」(ドラゴンズファン) の場合に 3 とコード化した。これを順序のあるカテゴリーと考えて、正規直交多項式により表 2 のような数量が付与された。

ナゴヤドームでの試合であり、ドラゴンズファンが全体の 4 分の 3 に達しているため、ドラゴンズファンに付される数量は小さく (0.54)、2 次変量も左右対称な放物線にならないが、1 次変量は 3 つのカテゴリーが完全な等間隔である。2 次変量は「対戦するどちらかのチームのファン」と「どちらでもない」を両極にもち、応援が目的である人とそうでない人を区別しているものとみられる。この点が後の主成分分析では重要になる。

表 2 「応援しているチーム」に付与された数量と反応率

変量	ドラゴンズ	どちらでもない	スワローズ
1 次	0.54	-0.88	-2.30
2 次	0.20	-2.56	1.18
反応率	75.0%	12.0%	13.0%

### 2.2.6 インターネット・アクセス (タイプ2)

「インターネットでプロ野球関連のウェブサイト、ブログ、SNS(ソーシャル・ネットワーキング・サービス、ツイッター、フェイスブック、インスタグラムなど)をどのくらいご覧になっていますか」という質問への回答である。これも順序のあるカテゴリー反応とみなして1～3次の正規直交多項式で数量化している(表3)。3次変量は、主成分による説明率が低く、無視されることになるが、2次変量にはユニークな解釈が見つかることになる。

表3 「インターネット・アクセス」に付与された数量と反応率

変量	毎日見ている	よく見ている	あまり見ない	ほとんど見ない
1次	1.10	0.18	-0.73	-1.65
2次	0.82	-1.07	-0.93	1.22
3次	0.32	-1.03	1.81	-0.59
反応率	33.5%	31.0%	17.6%	17.9%

### 2.2.7 ナゴヤドームでの観戦回数 (タイプ2)

「これまでに、何回くらいナゴヤドームで試合を観戦されましたか? (今シーズンに限らず、ドームの開業以来今日まで)」という質問への回答である。詳細な回数について回答を求めたいところであるが、人の記憶力の限界等も考慮し、2011年調査以降、表4のような刻みとしている。これも順序のあるカテゴリー反応として、4次までの正規直交多項式による数量化を行った(表4)。

このようなかなり主観的な量的評価項目への回答については、調査対象者の反応の構え(response set)が影響することが知られている(たとえば、Hamilton, 1968)。この問題については主成分の解釈をするところで考えることとする。その際に、2次変量に注目することになる。

表 4 ナゴヤドームでの観戦回数に付与された数量と反応率

変量	今日初めて	1～2回	3～5回	6～10回	11回以上
1次	-2.23	-1.47	-0.71	0.05	0.80
2次	2.08	-0.31	-1.37	-1.10	0.50
3次	-1.17	2.15	0.49	-1.65	0.24
4次	0.52	-2.19	1.61	-1.04	0.08
反応率	8.4%	7.9%	16.1%	16.6%	51.0%

## 2.2.8 誰と来たか (タイプ1)

ここからが基本的に調査当日の状況となる。「今日は、お一人でご観戦ですか？それともどなたかと一緒にですか？」という問いについて、次の選択肢を設けた。

1. 一人で	2. 家族と	3. 友人・知人と	4. 恋人と	5. 仕事関係の方と
6. 団体ツアー	7. その他 (具体的に )			

このうち、選択率の低い4以降をまとめて「その他」として扱っている。「職業」と同様、順序をつけることもできないので、主成分分析の結果として付与された数量を反応率とともに表5に示した。

表 5 「誰と来たか」の数量と反応率

変量	一人で	家族と	友人と	その他	解釈
1	-1.73	-0.47	<b>1.28</b>	-0.07	友人
2	-1.89	<b>0.56</b>	<b>-0.61</b>	1.80	友人(-)vs家族(+)
3	-0.92	<b>0.82</b>	-0.07	-2.47	家族
反応率	11.9%	45.2%	33.2%	9.7%	

これは実に解釈が難しい数量である。反応率の低い「一人で」と、さらに反応率が低く、異質な内容のカテゴリーをむりやり併合した「その他」はとりあえず無視して、主成分の解釈時に必要に応じて参照することとし、基本的に家族と友人 (合わせて80%近い反応率になる) のどちらか、

あるいはその対比と考えた。

### 2.2.9 どんな観戦がしたいか（2値変数）

前述のように、調査は試合開始以前に行われており、当日の観戦の感想を問うのは早すぎるので、期待の方を訊ねている。「今日は、どんな観戦がしたいですか。下の中から選んでください」という指示にもとづく、いわゆるチェックリスト型の項目であり、提示順序、位置等の影響を受けやすいきらいがあることを認識したうえで（特に、選択率については）見る必要がある。選択されたかされないかの2値変数であり、性別の場合と同様、平均値が0、標準偏差が1という制約の下では、カテゴリーへの数量は反応率にもとづいて(14)式により一意的に定まる。それらをまとめて表6に示した。

表6 「どんな観戦がしたいか」の8項目の反応率と数量

項目	反応率	選択	非選択
プレーそのものを間近で見たい	50.9%	0.98	-1.02
応援しているチームの勝利を見たい	61.3%	0.79	-1.26
気になっている選手を間近で見たい	28.7%	1.58	-0.63
勝敗に関係なくいい試合を見たい	23.2%	1.82	-0.55
会話を楽しみながら観戦したい	18.9%	2.07	-0.48
飲食を楽しみながら観戦したい	19.9%	2.01	-0.50
球場の雰囲気・ワクワク感を味わいたい	47.0%	1.06	-0.94
ファン同士の一体感を味わいたい	17.9%	2.14	-0.47

なお「プレーそのものを間近で見たい」、「応援しているチームの勝利を見たい」、「球場の雰囲気・ワクワク感を味わいたい」の3項目は、これらの項目を調査表に含めて以後7年間を通じて常に高い反応率を示すとともに、「期待」の異なる側面を代表するものとなってきた。

## 2.2.10 観戦満足度 (タイプ2)

「全体として現在のナゴヤドームでの野球観戦に満足しておられますか」という問いへの反応で、次のフォーマットで反応を求めた。

1. 不満	2. やや不満	3. どちらとも言えない	4. ほぼ満足	5. 満足
-------	---------	--------------	---------	-------

過去の結果と比較するために、「満足」を4、「ほぼ満足」を3、「どちらとも言えない」を2、反応率の低い「やや不満」と「不満」を併合してそれらを1とコード化し、正規直交関数により数量を与えた。それらを反応率とともに表7に示す。

以上、分析の対象となる項目数は16であり、総カテゴリー数は51 (=2+6+7+3+4+5+4+2×8+4) である。主成分分析の対象となるのは、各項目のカテゴリー数から1を引いたものの総和、すなわち、35 (= 51 - 16) 個の変数ということになる。

表7 「観戦満足度」の反応カテゴリーに付与された数量および反応率

変量	満足	ほぼ満足	どちらでもない ( やや) 不満	
1次	1.45	0.33	-0.79	-1.91
2次	1.54	-0.65	-0.65	1.54
3次	0.68	-0.69	1.68	-0.90
反応率	16.8%	49.8%	20.6%	12.7%

## 2.3 主成分分析の結果

### 2.3.1 主成分の数の決定

まず、通常の主成分分析で行われるように、scree-plot (相関行列の固有値を大小順に並べてプロットしたもの、図2) を観察して因子数の見当をつけることから始めた。固有値の値は大小順に、2.29、2.00、1.88、1.74、1.59、1.40、1.30、1.25、1.17、1.11、1.09、1.07である。

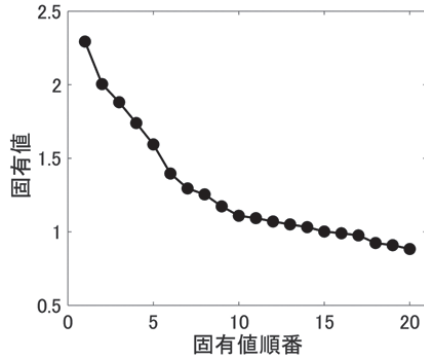


図2 Scree-plot

この図からは、主成分の数として6または9が示唆されるであろう。できるだけ多面的な個人差を表現するために9主成分解が検討されたが、9番目の主成分の解釈が困難であった。主成分数を1つ減らした8主成分解については、納得のいく解釈が可能であったため、これを採用することとした。1節のアルゴリズムの流れの中でも示唆したように、これはあくまでも暫定的な結論であり、現実の応用場面ではもっと多くの主成分数や項目のタイプの変更等を試みる必要がある。

解の寄与（説明力）を、通常の主成分分析と同様に固有値の和で考えると、大小順に8番目までの固有値の和は13.460となる。やはり主成分分析と同様に、データの全分散を全変量の分散の和で考えると35であるから、主成分によって説明される分散の割合（寄与率）は、ほぼ38.6%となる。これは通常の心理測定の項目に対する寄与としては必ずしも高いものとは言えないが、多様性を目的とした社会調査の項目としては、満足すべき水準ではないかと考える。なお、イナーシャの総和に対する説明力（カイ2乗統計量の減少分）は49.0%である。カイ2乗統計量で測られる項目間の関連度のほぼ半分がこの解によって説明されていることになる。

次に、主成分はHarris & Kaiser (1967)の方法によって回転された。



図3は回転前の各主成分の寄与である固有値と回転後の寄与であるパターン行列の各列の平方和を比較したものである。回転によって寄与が平準化していることがわかる。

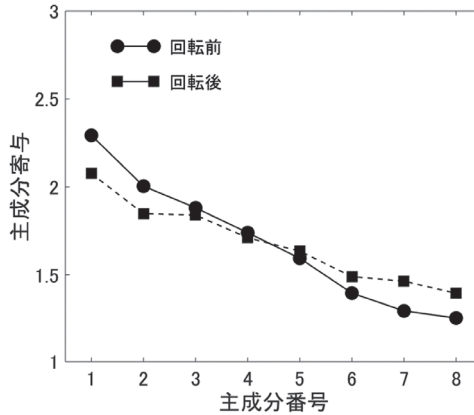


図3 回転の前後の寄与の比較

### 2.3.2 結果の解釈

斜交因子に対する負荷行列であるパターン行列と主成分間相関行列が主たる解釈の対象となる。表8にはそのパターン行列を各変量の決定係数の2乗  $R^2$  とともに示した。パターン行列の要素である負荷量については、3節で報告する bootstrap 法により検定の多重度を考慮して5%水準で有意(0でない)と判断した数値を、暫定的に解釈の対象とするものとしてボード体で示している。この手続きにより、負荷量は必ずしもその絶対値にしたがって評価していない。また、やや恣意的ではあるが、検定統計量が3を超える負荷量はイタリック体で示して解釈の参考にした。これらの点については、3節を参照のこと。また、 $R^2$ の値が0.15に達しない変量は解釈の対象から外した(灰色の網掛け)。

パターン行列の下には、各列の平方和である主成分の分散、あるいは主

表 8 負荷量・主成分寄与・主成分相関等

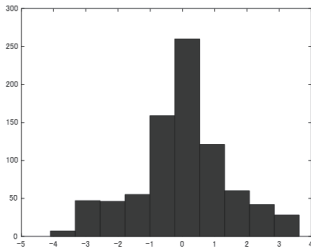
変量	項目		I	II	III	IV	V	VI	VII	VIII	R <sup>2</sup>	
1	1	性別	女性	0.02	0.01	0.02	<b>0.80</b>	0.05	0.01	-0.11	0.05	0.655
2	2	年齢	1次	<b>0.85</b>	0.01	0.00	0.02	-0.07	0.03	0.14	0.05	0.767
3			2次	-0.01	<b>-0.84</b>	0.01	-0.02	0.03	-0.04	-0.01	-0.01	0.704
4			3次	<b>0.34</b>	-0.06	0.06	-0.14	0.19	-0.09	<b>-0.63</b>	-0.15	0.562
5			4次	<b>0.06</b>	<i>-0.24</i>	<i>-0.04</i>	<i>0.04</i>	<i>0.02</i>	<i>0.11</i>	<i>0.02</i>	<i>0.10</i>	<i>0.091</i>
6			5次	0.00	-0.05	0.11	-0.09	0.08	-0.02	0.09	<i>-0.43</i>	0.217
7	3	職業	学生(-)vs無職(+)	<b>0.87</b>	0.01	0.02	-0.03	0.07	-0.05	<b>-0.19</b>	-0.01	0.776
8			有職	0.01	<b>0.82</b>	-0.06	-0.04	0.05	0.01	0.00	-0.02	0.689
9			パート・主婦	-0.03	-0.04	0.01	<b>0.80</b>	-0.01	0.00	0.14	-0.06	0.664
10			公務員	-0.01	0.00	0.06	0.00	0.17	<b>0.36</b>	-0.02	-0.11	0.159
11				<b>0.00</b>	<b>0.00</b>	<i>0.01</i>	<i>0.00</i>	<i>-0.02</i>	<i>-0.02</i>	<i>-0.04</i>	<i>0.22</i>	<i>0.047</i>
12				<b>0.00</b>	<b>0.00</b>	<i>-0.01</i>	<i>0.00</i>	<i>-0.03</i>	<i>0.00</i>	<i>0.07</i>	<i>0.00</i>	<i>0.006</i>
13	4	ドラゴンズファンか	1次	0.20	-0.04	-0.08	-0.12	0.06	-0.12	<b>0.57</b>	-0.17	0.411
14			2次	0.05	-0.08	<b>0.71</b>	0.05	-0.14	0.13	-0.08	0.05	0.526
15	5	インターネット	1次	<b>-0.38</b>	<b>0.29</b>	<b>0.40</b>	-0.10	0.07	-0.17	-0.07	0.12	0.459
16			2次	0.05	<i>-0.20</i>	-0.13	-0.01	0.08	<b>-0.52</b>	-0.04	-0.15	0.340
17			3次	<b>-0.05</b>	<b>-0.06</b>	<b>-0.05</b>	<b>-0.14</b>	0.18	0.24	0.14	<b>-0.16</b>	<b>0.144</b>
18	6	観戦回数	1次	<i>0.14</i>	0.02	<b>0.39</b>	-0.11	0.10	<b>-0.33</b>	<i>0.31</i>	-0.11	0.477
19			2次	0.04	<i>0.23</i>	-0.08	0.18	-0.03	<b>-0.46</b>	-0.13	0.04	0.309
20			3次	<b>-0.09</b>	0.02	<b>-0.05</b>	0.01	0.13	-0.05	0.20	-0.19	0.094
21			4次	<b>-0.04</b>	<b>-0.10</b>	0.11	0.05	-0.05	0.02	-0.13	0.16	0.063
22	7	誰と来たか	友人	<b>-0.37</b>	-0.13	-0.08	-0.08	<b>0.50</b>	0.03	-0.12	-0.13	0.441
23			家族(+) vs友人(-)	0.04	0.12	-0.10	<b>0.31</b>	0.02	<b>0.43</b>	-0.01	-0.14	0.357
24			家族	0.02	<i>-0.18</i>	0.12	<i>0.24</i>	-0.06	-0.04	<i>0.39</i>	0.06	0.300
25	8	プレーを間近で	2値	0.02	0.02	-0.10	<b>-0.22</b>	0.04	0.14	0.19	<b>0.62</b>	0.492
26	9	チームの勝利	2値	0.02	-0.03	<b>0.73</b>	0.05	0.11	-0.05	0.06	-0.04	0.585
27	10	気になっている選手	2値	0.02	-0.05	0.10	0.11	0.06	-0.12	-0.07	<b>0.64</b>	0.470
28	11	いい試合	2値	0.10	0.00	<b>-0.47</b>	0.12	0.22	-0.26	0.14	0.14	0.359
29	12	会話を楽しみながら	2値	0.05	0.01	-0.05	0.01	<b>0.66</b>	-0.06	-0.07	0.10	0.483
30	13	飲食を楽しみながら	2値	0.14	0.09	-0.01	-0.02	<b>0.48</b>	0.27	-0.06	0.08	0.338
31	14	雰囲気・わくわく感	2値	-0.04	0.10	0.02	<i>0.21</i>	<b>0.51</b>	0.00	0.15	-0.06	0.344
32	15	一体感	2値	0.04	-0.07	<b>0.32</b>	0.05	<b>0.41</b>	-0.20	-0.01	0.10	0.361
33	16	満足度	1次	-0.08	<b>-0.31</b>	-0.11	-0.05	0.07	0.26	-0.09	0.13	0.220
34			2次	<i>-0.22</i>	<i>-0.08</i>	<i>-0.12</i>	<i>-0.08</i>	0.03	-0.32	0.07	0.11	0.210
35			3次	-0.16	-0.08	-0.11	0.15	-0.17	-0.22	<i>-0.42</i>	-0.16	0.340
			主成分寄与の分解	2.08	1.85	1.84	1.71	1.63	1.49	1.46	1.40	13.460
			1次変量	0.97	0.21	1.21	0.80	1.19	0.44	0.56	0.92	6.306
			2次変量	0.05	0.81	0.55	0.04	0.03	0.60	0.03	0.04	2.150
			3次変量	0.15	0.01	0.02	0.06	0.11	0.12	0.62	0.11	1.214
			4次変量	0.00	0.07	0.01	0.00	0.00	0.01	0.02	0.04	0.157
			5次変量	0.00	0.00	0.01	0.01	0.01	0.00	0.01	0.18	0.223
			順序なし	0.89	0.74	0.04	0.80	0.29	0.32	0.23	0.10	3.410
			主成分間相関係数									
			I	1.00								
			II	0.05	1.00							
			III	0.05	0.01	1.00						
			IV	<b>0.12</b>	0.01	0.04	1.00					
			V	0.01	0.03	0.02	0.06	1.00				
			VI	0.05	0.00	-0.10	0.07	-0.06	1.00			
			VII	0.06	<i>-0.10</i>	<b>0.16</b>	0.04	-0.02	0.02	1.00		
			VIII	-0.03	-0.07	-0.02	0.01	0.09	0.00	0.02	1.00	
			歪度	-0.11	-1.00	-0.78	0.87	0.74	0.29	-0.49	0.18	

成分寄与の合計値をまず示し、その下に変量の次数別の寄与を示した。右端の合計値で見ると、2 値変数も含む 1 次成分の寄与がやはり大きい、順序のないカテゴリカル変数（主に職業）がそれに次ぎ、2 次変量、3 次変量の寄与も無視できない水準にある。主成分ごとに見ると、それぞれの特徴が読み取れるが、それは主成分の解釈を行った上で見る方がわかりやすいであろう。

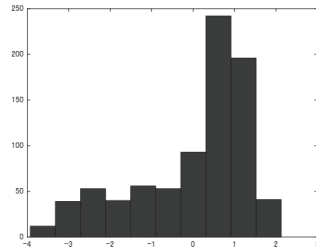
その下には主成分間相関行列、さらに主成分の歪度を示した。図 4 には各主成分得点の分布（ヒストグラム）を示した。大部分の主成分得点は多かれ少なかれ分布の偏りを示している。それぞれの解釈には多数派の側（分布のピークがある側）から解釈した方が理解しやすいものと、少数派の側（分布の裾野の側）から解釈する方がわかりやすいものがあることがわかった。ここでは、そうした解釈の「基準の極」が正の値になるように符号を整えた。以下、表 8 と図 4 にもとづき、順番に主成分の解釈について述べる。

### 2.3.3 主成分 I : 年齢段階

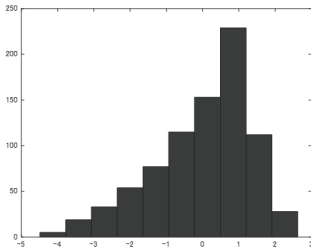
表 8 の I の列から有意な値を拾ってみると、「年齢」の 1 次変量への負荷と「職業」の第 1 変量への負荷がともに 0.8 に達している。すなわちこの主成分の得点は高齢で「無職」の人たちが高く、若い「学生・生徒」が低くなっていることから、「年齢段階」と命名できるであろう。年齢段階の 3 次変量にも低い負荷があるが、これは得点が年齢段階に対して直線的に増加するのではなく、後に図 9 で見るように、30～50 歳代での変動が小さい、いわば踊り場のような状況になるからである。「インターネット・アクセス」に対する負の負荷は高齢者ほどインターネットを閲覧しなくなる傾向を示している。また、「誰と来たか」の第 1 変量への低い負の負荷は、高齢になるほど、一緒に観戦する相手が「友人・知人」から「家族」に変わっていくことを反映している（図 5）。この主成分に大きく寄与しているのは 1 次変量（主に年連段階）と順序のない変量（主に職業）である。



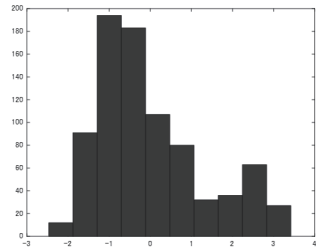
主成分I 年齢



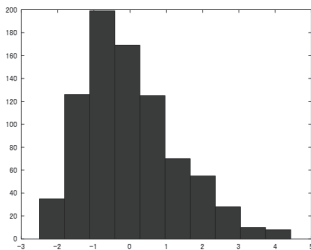
主成分II 有職



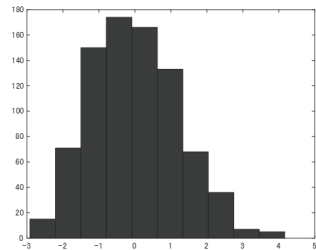
主成分III サポーター



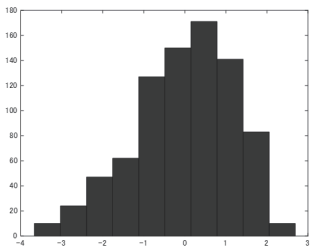
主成分IV 女性客



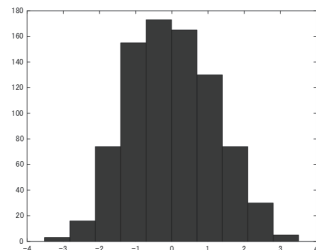
主成分V 居場所希求



主成分VI 中庸反応



主成分VII ドラゴンズファン



主成分VIII ふれあい期待

図4 8つの主成分の度数分布(ヒストグラム)

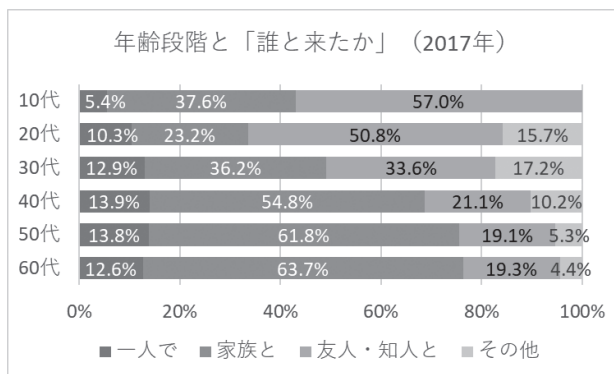


図5 年齢段階ごとに見た一緒に観戦する相手の割合

なお参考程度にとどまるが、「観戦満足度」の2次変量も低い負の負荷を示している。これは、放物線を上下逆にした形に対応し、若者と高齢者が比較的満足度が高いことの反映である（図6参照）。

いずれにせよ、この主成分は年齢というデモグラフィックな要因を主に反映しており、年齢に随伴する行動の変化をいくらか反映するものであ

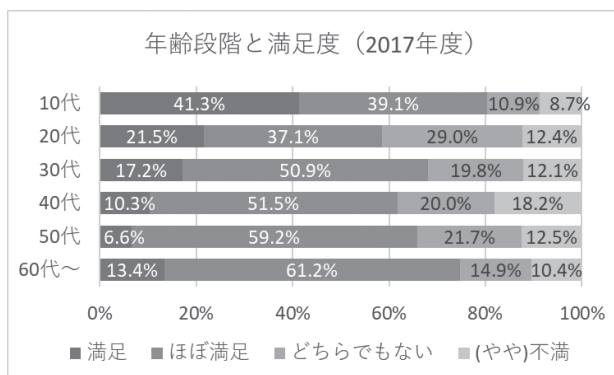


図6 年齢段階と観戦満足度の関係

る。また、図4のヒストグラムから、この主成分得点の分布が10代の若者と60代以上の高齢者を両端にするほぼ左右対称な形状をとっていること（歪度は-0.11）、回答者の大半は平均値（0である）付近に集中していることが読み取れる。

### 2.3.4 主成分Ⅱ：有職

第Ⅱの主成分で目立つのは、年齢段階の2次変量への負の負荷と職業の第2変量への正の負荷である。図1で見たように年齢段階の2次変量は下に凸の形になっているから、中間の年齢段階の主成分得点が高く、かつ「職業」の第2変量は「公務員」、「会社員」、「パート・アルバイト」、「自営等」といったいわば自ら稼いでいるカテゴリーに属する人たちの数量が高いことからそれが主成分得点を押し上げていることがわかる。図3からこの主成分得点の分布は著しく負に偏っている（歪度は-1.00）ことが見てとれる。つまり球場に来る人の多くが中年のサラリーマンやOLであることを意味している。この主成分もデモグラフィックな変数によって定義されている。大きな寄与は2次変量（年齢段階）と順序のない変量（職業）による。

この主成分でデモグラフィックな要因以外で有意な負荷をもつ変量は「インターネット・アクセス」と「観戦満足度」の1次変量である。観客の多数派である有職者はプロ野球関係のサイトを頻繁に閲覧する傾向があると同時に、観戦には必ずしも満足していないことを示している。ただし、この年齢と不満の放物線状の関係は、現代日本の大部分の社会調査の結果に見られるものではある（たとえば、NHK放送文化研究所、2015）。また、図6からもわかるように、不満という反応は全体として見てもそれほど多いわけではなく、「ほぼ満足」も加えれば、来場者の過半数は満足していると見てよいであろう。

ところで、第Ⅰ主成分とこの第Ⅱ主成分の主成分得点の間の散布図は図7のようになる。これはまさしく典型的な2次曲線、すなわち馬蹄現象

を示している。これは、方法の性質に起因するアーチファクトなのだろうか。特にこの場合、馬蹄の内側にはほとんど点が存在しないことから、この2つの次元は非線形の変換によって1つの次元にまとめられるのではないか、すなわち、次元の縮約が可能になるのではないかと考えられよう。さらに、この馬蹄はほとんど傾いていないことから、その1つにまとめられた次元はほぼ年齢段階（の1次変量）を反映するものになるであろう。しかしながら、そうすると圧倒的多数を占める有職者は、変換の結果生まれた年齢次元の中で平凡な位置を占めることになり、彼らの特徴を特に外部変数との関係で検討する場合には、2次関数が必要となってしまうことになる。これもそれなりに煩雑な作業となる。少なくとも2つの主成分の寄与の大きさを考えれば、第II主成分は単なるアーチファクトとは言えない。有職者と（学生と高齢者という）無職者を対極にもつという明確な意味をもつ第II主成分は維持するに値するであろう。

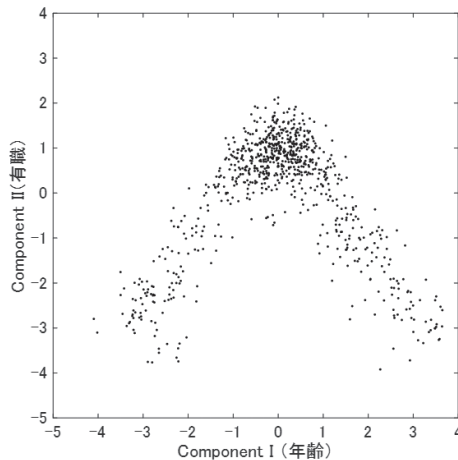


図7 主成分Iと主成分IIの間の散布図

### 2.3.5 主成分Ⅲ：サポーター

この主成分で目立つ負荷量は、「ドラゴンズファンか」の2次変量、および「どんな観戦がしたいか」の「応援しているチームの勝利が見たい」で、ともに0.7を上回っている。表2で見たように、前者の2次変量は対戦している2チームのどちらかのファンであれば正の値をとる。また後者はひいきのチームがこれから対戦するとすれば、当然の期待であろう。それ以外にも、「インターネット・アクセス」の1次変量、「観戦回数」の1次変量に正の負荷をもつことは、この主成分の高得点者のプロ野球への関りの深さを示している。また、「どんな観戦がしたいか」の「ファン同士の一体感が味わいたい」というのも、彼らが球場で観戦する大きな理由を示しているであろう。この次元の高得点者には「サポーター」の名を冠するのが適当と考えるゆえんである。他方で、「勝敗に関係なくいい試合が見たい」への負の負荷は、この選択肢にチェックするのがむしろ低得点者であることを意味している。寄与は主に1次変量と2次変量（「どこのファンか」）から来る。この主成分はデモグラフィックな変量がほとんど負荷しておらず、様々な年齢、性別にわたってまんべんなく存在する特徴であると思われる。

図4のヒストグラムからは、主成分Ⅱほどではないにしても、この主成分では高得点者が多数派であることを示している。このようにいわば勝負にこだわる人々が球場への来訪者の多くを占めているのである。対極にあるのは、必ずしも初心者ではなく、試合の結果よりも内容を重視する、どちらのチームにも与しない少数者である。

図8は、「どんな観戦がしたいか」というチェックリストの中の4つの項目についてここ7年間の反応率の推移を示したものである（「勝敗に関係なくいい試合が見たい」という選択肢は2014年版にはなかった）。どうやら最近では、この主成分の高得点者が勢いを増してきているようである。



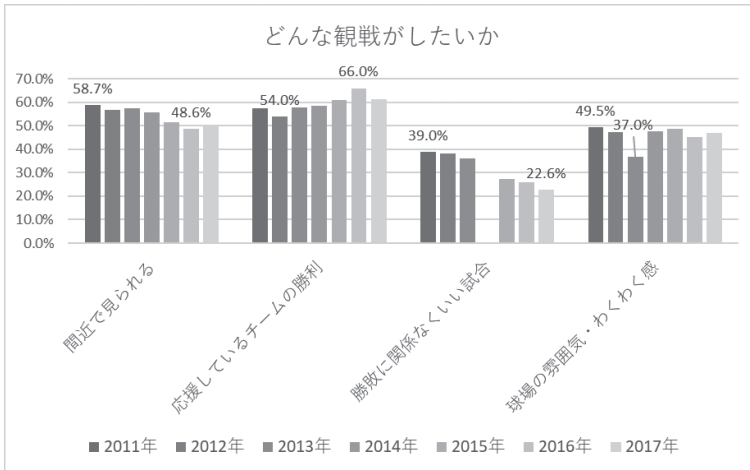


図 8 「どんな観戦がしたいか」における代表的な 4 項目の反応率の推移

### 2.3.6 主成分Ⅳ：女性客

表 8 に見られるように、「性別」と「職業」の第 3 変量に高い正の負荷をもつ。「パート」あるいは「専業主婦」の女性が高得点となる。「年齢段階」への負荷からは読み取れないが、比較的年齢は高い。若い女性は、フルタイムで働いているか学生である。図 4 からは（当然ながら）全観客の中では少数派になる。この主成分上では、このグループだけで他と離れた分布をなしているようである（二峰性の分布）。対極は以上の属性をもたない人全部である。「誰と来たか」の第 2 変量にも低い負荷があり、イタリック体の第 3 変量とも併せて、この主成分の高得点者は家族と一緒に来ることが多いことを示している。また、「間近でプレーを見ること」への関心は低い一方、「球場の雰囲気」は楽しんでいられるらしい。

大変わかりやすい主成分であるが、これもほぼデモグラフィックな変数によって強く定義される主成分である。1 次変量（2 値の「性別」と順序のない変量（「職業」）の寄与が大きい。

### 2.3.7 主成分Ⅴ：居場所希求

表8から「誰と来たか」の第1変量にやや目立つ正の負荷があって「友人・知人」と来場しており、さらに「どんな観戦がしたいか」の「会話を楽しみながら」、「飲食を楽しみながら」、「球場の雰囲気・ワクワク感を楽しむ」人が高得点をとる。主成分Ⅲと同様、デモグラフィックな要因は関わりが薄い。高得点側に位置するのは、いわゆるサード・スペースとして球場を一種の居場所的に利用している人々であるが、主成分Ⅱの場合ほどではないとしても全体の中では少数派である。ほとんど1次変量の負荷だけからなっている。

### 2.3.8 主成分Ⅵ：中庸反応傾向

これは、非常に特異な主成分である。おそらく、正規直交主成分分析なしにこのような次元を発見することはできなかったであろう。まず「職業」の第4変量「公務員」に低い負荷がある。ただ、第4変量において「公務員」に付された数値は非常に大きいので、公務員のこの主成分の平均値は他の職業グループに比して相当に高い。また、「インターネット・アクセス」、「観戦回数」のそれぞれ2次主成分に負の負荷をもち、順序のついたカテゴリーのうち中央付近のものを選択する傾向を示している。これを中庸反応傾向と呼ぶことにしよう。参考レベルにも達しないが、「観戦満足度」の2次変量への負荷は、この主成分の高得点者がほとんど「ほぼ満足」を選択することを反映している。

こうした傾向は、通常はアーチファクトとされる馬蹄状の分布（馬蹄現象）の底辺に位置する反応で、内容的な解釈が可能かどうか疑わしいとみなされる可能性もある。しかしながら、2つの理由でこの主成分がアーチファクトであるという可能性は否定される。まず、こうした次元は対応する変数の1次変量が同じまとまりをもち、それに付随して出現するものであるが、先の3つの変数の1次変量は同一の主成分に負荷していない。第2に、通常は「中庸」とは逆の極端反応傾向が、項目内容とは独立な反応

傾向として注目され、そうした傾向をもつ回答者は少数派である。しかし、図4からわかるように、現在分析中のデータでは、中庸反応傾向の方が少数派なのである。

これ以外には、「観戦回数」の1次変量に負、「誰と来たか」の第2変量の正の負荷を示し、来場回数が少ないこと、家族と来場する傾向があることを示している。これらをまとめて「公務員的反応」と見ることにはさほど無理がないと言えるかもしれない。内容の性質上、2次変量の寄与が大きい。

この主成分は、今回は分析の対象としていない項目である外部変数との関係でも、いくつか注目すべき傾向を示しているが、これについては後述する。

### 2.3.9 主成分Ⅶ：ホームチーム（ドラゴンズ）ファン

この主成分で一番目立つのは年齢段階の3次変量への負荷であるが、これについては後述することとし、「ドラゴンズファンか」の1次成分への正の負荷に注目しよう。この主成分の高得点者は、ナゴヤドームのホームチームである中日ドラゴンズのファンである。図4のヒストグラムから見ても、当然高得点の側が多数派である（歪度は-0.49）。

ドラゴンズファンの対極は当然相手チームのファンであり、そのことを考えると「観戦満足度」の3次変量へのかなり大きな負の負荷量（参考程度）も理解できる。すなわち、表7で見たように、この変量で目立つのは「どちらでもない」への（絶対値の）大きい負の値である。相手チームのファンは、当然、ナゴヤドームでの観戦経験が少なく、「どちらでもない」は「わからない」という意味に解した方がよさそうである。

さて、この主成分の「年齢段階」の3次変量の負荷であるが、これを理解するために、図9をご覧ください。この図には、ここまでの主成分のうち年齢段階の変量に目立った負荷量をもつ3つの主成分得点の平均値を年齢段階の関数として示したものである。「年齢段階」と命名された第

I 主成分には、1 次変量が正に負荷していたが、平均値も年齢段階に関してほぼ直線的に増加している。ただし、先にも指摘したように 30～50 代のところが平坦な踊り場状を呈しており、そのために 3 次変量も低く負荷している。3 次変量によって年齢による変化がより忠実にあらわされているわけである。第 II 主成分は、図 1 の 2 次変量をちょうど上下に反転した形であるが、20～50 代の間のほぼ平坦な状態を作るために 4 次変量がいくらか寄与している。

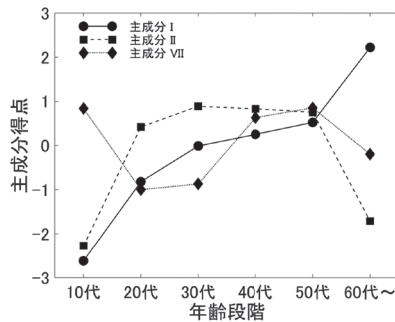


図 9 年齢段階と主成分得点の平均値

問題の第 VII 主成分であるが、これも平均値の曲線が図 1 の 3 次変量を上下反転した形とほぼ一致することが見てとれるであろう。具体的には 20～30 代で「ドラゴンズファン度」が低下しているところが 3 次変量でうまくあらわされるのである。実際、図 10 に見られるように、この年齢段階は他の年齢段階と比べてドラゴンズファンの割合が際立って少ない。ただ、これが若い世代のドラゴンズ離れを示しているかどうかは、この 1 回の調査から即断すべきではない。

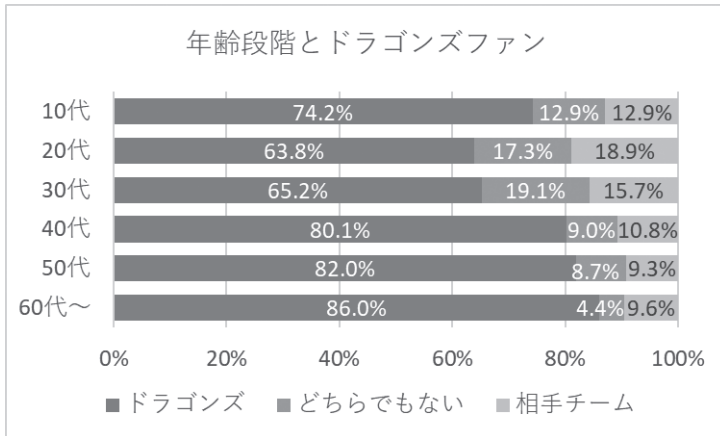


図 10 年齢段階によるドラゴンズファンの割合の差異

なお、この主成分の高得点者には、学生が多いこと（有意）、家族と来ているケースが多らしいこと（参考）にも注意しておこう。

### 2.3.10 主成分Ⅷ：ふれあい期待

この主成分において目立つ負荷量は、「年齢」の5次変量（ただし参考）と、「どんな観戦がしたいか」の「プレーそのものを間近で見たい」と「気になっている選手を間近で見たい」の2つである。後者については、選手たちを同じ場所で「生（なま）で」見たいという希望であり、村上（2014）が「生の体験」と呼んだものに近い。これ以前の調査においても、「選手と触れ合えるイベント」を望む声は常に強く、この次元も同様の期待を反映するものと思われる。

問題は、5次変量のようなギザギザな変量とこれらの項目への反応が連関するという事実である。実際、上記の2つの項目への反応率を年齢段階別に集計すると、5次変量を上下反転したような形状を示す（図 11）。したがって、2017年のデータに関しては、このような結果になることは納得せざるを得ないとしても、他の年の調査でも同様な傾向があるかどうか

を検討する必要がある。

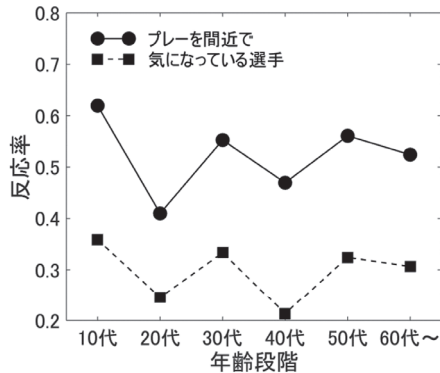


図 11 主成分Ⅳに負荷する項目の集計結果

### 2.3.11 「観戦満足度」の変動因

分析の対象となった16番目の変数である「観戦満足度」は、研究目的からすれば最も重要な変数ともいえるが、負荷量の点では特にどの主成分にも際立って高い負荷を示していない。これについては、複数の主成分が少しずつ負荷しているためとも考えられる。そこで、図12に多少とも「観戦満足度」のどれかの変量に負荷していた4つの主成分について、4つの選択肢ごとの平均値を示した。まず主成分Ⅰでは低得点（若い）が「満足」を選択、高得点者（高齢）が「ほぼ満足」を選択、中間の年齢層には特に特徴がないことが読み取れる。第Ⅰ主成分に対して放物線状を呈していた第Ⅱ主成分は、まず低得点（若者と高齢者）が満足を、それ以外では得点が高い相（有職者）は「不満」を選んでいる。主成分Ⅵは「中庸反応傾向」であるが高得点者が「ほぼ満足」に集中しているようである（無難な反応と言うべきか）。主成分Ⅶのドラゴンズファンは高得点者が「不満」、低得点者（相手チームのファン）は、「どちらでもない」を選択して

いる。結局、低い負荷量が示していた傾向は、こうした平均値のプロットからも読み取ることができるわけである。

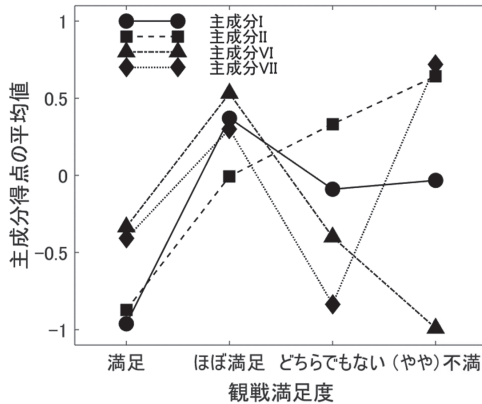


図 12 観戦満足度と 4 つの主成分得点の平均値

### 2.3.12 主成分間相関

表 8 の下方に示された主成分間相関であるが、大部分の数値は 0 に近く、解全体が直交解に近いことを示している。かすかながら、一応解釈に値すると思われるのは主成分 I (年齢) と主成分 IV の 0.12、並びに主成分 III (サポーター) と主成分 VII (ドラゴンズファン) の 0.16 の 2 つのゴシック体で示した数字 (有意である) に加えて、主成分 II (有職) と主成分 VII の逆相関だけである。第 1 の相関は、パートや専業主婦の女性が比較的年齢が高い (若い女性は学生・生徒か有職者である場合が多い) ことを示しており、第 2 はサポーターの過半数はドラゴンズファンでもあることを。第 3 は有職者には相対的にドラゴンズファンが少ないことを示している。いずれにしても数値としては小さく、特に問題にする必要はないであろう。

## 2.4 外部変数との関係

ここまで分析の対象としてきた調査の調査票には、正規直交主成分分析には含めなかった変数がいくつか存在している。ここでは、上で「抽出」された8つの主成分と、それらの残された変数のいくつかとの関係を検討する。

多重対応分析の結果に対し、外部の変数を投影してその関係（構造）を検討することはよく行われているが、それを形式化して示したのが Le Roux & Rouanet (2010) の幾何学的データ分析 (Geometric data analysis) である。基本的な手順としては、意識・行動レベルの変数で多重対応分析を行い、その結果に対して主としてデモグラフィックなカテゴリーを投影してその関係を見ることのようにである。たとえば、個人の文化に関連した行動が、その個人の属する社会階層（学歴、職業等）と関連するかどうかを、イギリスにおける調査結果を用いてその手続きで検証したのが、Le Roux, Rouanet, Savage, & Warde (2018) であり、その結果にもとづく著書も刊行されている（ベネット・サヴィッジ・シルヴァ・ワード・ガヨ＝カル・ライト, 2017）。この種の手続きは、近年心理学において多用されている構造方程式モデルのように明確な因果関係に関する仮説検証というフレームワークはもっておらず、あくまでもデータの構造の解明に目的を限定している。

他方、Lebart (2006) は同じ手続きを、多重対応分析の妥当性検討 (validation) であるとしている。この場合の妥当性は、個人のアセスメントに限定された心理学における用法とは異なり、多重対応分析の軸の解釈の適切性の検討といった意味合いのようである。本研究も同様の意味でこの手続きに従うことにしたい。

ここで検討の対象とする変数と選択肢は、以下のようなものである（( ) 内は選択パーセント）。

・プロ野球への関心度

1. すべてのスポーツの中で一番 (66.2)



2. プロ野球と同じくらい関心のあるスポーツがある (23.1)
  3. プロ野球よりも関心のあるスポーツがある (5.5)
  4. スポーツ観戦にあまり関心がない (1.8)
- ・ファンクラブ加入の有無
    1. 入会している (36.2)
    2. 入会したことがないが入会したい (15.4)
    3. 入会する気はない (34.2)
    4. 以前は入会していたが、今は入会していない (11.3)
  - ・初めてプロ野球を観戦した時期
    1. 小学校入学前 (13.5)
    2. 小・中学生 (52.3)
    3. 高・大学生, 専門学校生 (9.3)
    4. 社会人 (結婚前) (14.3)
    5. 社会人 (結婚後) (7.4)
  - ・初めての観戦のきっかけ (複数選択可)
    1. 好きな選手がいる (いた) (21.1)
    2. 好きなチームの試合である (あった) (25.5)
    3. 家族と一緒に (45.8)
    4. 友人・知人に誘われた (15.7)
    5. チケットをもらった (7.9)
    6. 野球そのものに興味がある (あった) (14.3)
    7. 球場に入りたかった (4.6)
  - ・ナゴヤドーム以外での観戦経験
    1. 一度もない (26.4)
    2. 1～2回 (23.3)
    3. 3～5回 (16.5)
    4. 6～10回 (10.0)
    5. 11回以上 (22.8)

- ・どんな観戦がしたいか（正規直交主成分分析に含めなかったもの）
  1. 野球以外のファンサービスを楽しみたい (4.6)
  2. 応援でストレスを解消したい (12.5)
- ・球場までの所要時間（時間・分を記入してもらったが、以下の6カテゴリーに整理）
  1. 30分未満 (4.5)
  2. 30～59分 (23.3)
  3. 60分 (29.6)
  4. 61～119分 (19.4)
  5. 120～179分 (12.6)
  6. 180分～ (10.6)
- ・観覧席（調査員の担当箇所から判断、次の7箇所に整理）
  1. ネット裏 (6.2)
  2. 内野1塁側 (24.7)
  3. 内野3塁側 (25.7)
  4. 外野ライト側 (10.8)
  5. 外野レフト側 (12.1)
  6. パノラマ1塁側 (8.7)
  7. パノラマ3塁側 (11.8)

なお、上記項目の選択肢等の反応率は、それぞれの質問への欠損値のない回答者全員について計算されており、正規直交主成分分析で分析の対象となった回答者の値とは一致しない。

これらの外部変数との関連の分析は、主成分ごとに、主成分得点の平均値を上記外部変数のカテゴリーごとに計算し、1元配置分散分析によって有意性を検定するとともに、効果量として相関比の2乗 ( $\eta^2$ ) を求めた。

ここで問題になるのが検定の多重性である (Lebart, 2006)。すなわち、上の手順で実施される検定数は膨大であるから、通常の5%水準の検定

では第1種の誤りが不当に増加してしまうのである。この問題は、次のBootstrap法ではより深刻な問題となるが、ここでは基本的に単独では0.1%水準で有意となる $p < 0.001$ のものを取り上げ、特に興味深いものについては、これを上まわる $p$ 値であっても記載するという方針をとることにする。これらを一覧表にしたのが、表9である。それぞれの平均値を表やグラフにすれば膨大なものになるので、値の様相は言葉で略述するという方針にした。

それぞれの主成分についてコメントしておこう。主成分Iの低得点者は、年齢段階の低い層であり、最初の観戦時期が早いのは、球場に来ている以上それ以前に観戦経験があるのに対し、年齢が高い層には比較的遅く

表9 各主成分と外部変数との関係

平均値比較で目立った変数	$p$ 値	$\eta^2$	差の内容・方向
主成分 I (年齢)			
初めての観戦時期	$p < 0.001$	0.145	低得点者(若い層)ほど早い
観覧席	$p = 0.002$	0.025	高得点者はネット裏、低得点者は外野レフト側にやや多い
主成分 II (有職)			
他球場での観戦	$p < 0.001$	0.047	高得点者ほど他球場での観戦回数が多い
主成分 III (サポーター)			
野球以外のファンサービス	$p < 0.001$	0.013	高得点者が楽しみにしている
野球への関心度	$p < 0.001$	0.067	低得点者は関心が低い
ファンクラブ	$p < 0.001$	0.171	高得点者が加入、低得点者は加入する気がない
観覧席	$p < 0.001$	0.046	高得点者は外野応援席にいる
主成分 IV (女性)			
他球場での観戦	$p < 0.001$	0.035	高得点者は他球場の観戦経験が少ない
初観戦の時期	$p < 0.001$	0.178	高得点者は時期が遅い
主成分 V (居場所希求)			
野球以外のファンサービス	$p < 0.001$	0.044	高得点者が楽しみにしている
主成分 VI (中席反応)			
他球場での観戦	$p < 0.001$	0.046	高得点者は1~2回、低得点者は11回以上
ストレス解消	$p = 0.002$	0.011	低得点者が肯定
野球への関心度	$p < 0.001$	0.036	高得点者は4段階の2と3に多く反応
ファンクラブ	$p < 0.001$	0.035	上と同じ傾向
初観戦の時期	$p < 0.001$	0.029	高得点者は遅い傾向
観覧席	$p < 0.001$	0.035	高得点者は外野ライト側にはいない
主成分 VII (ドラゴンズファン)			
他球場での観戦	$p < 0.001$	0.028	低得点者ほど多い(相手チームのファンだから)
ファンクラブ	$p = 0.013$	0.013	高得点者に「やめた」が多い
球場までの所要時間	$p < 0.001$	0.045	高得点者ほど短い 地元ファンが多い
観覧席	$p < 0.001$	0.104	高得点者はレフト側にはいない
主成分 VIII (ふれあい期待)			
好きな選手がいた	$p < 0.001$	0.016	高得点者が肯定
球場までの所要時間	$p < 0.001$	0.035	高得点者ほど長い 他球場での経験から期待が高い?
観覧席	$p = 0.001$	0.027	高得点者は内野3塁側にやや多い ライト側、ネット裏には少ない

初観戦した人たちが混入するせいであって、論理的な帰結である。観覧席に関しては、年齢が高い層がネット裏におり、低い層は外野のそれもレフト側にいることが多いのは、経済的な「格差（年齢差）」の反映であろう。主成分Ⅱでは高得点者（有職者）が他球場での観戦経験が多いのも、他地域への「遠征」を可能にするだけの経済的基盤があることの反映と考えられる。圧倒的多数を占める有職者は、必ずしもスポーツ観戦に関心が高いわけではないので、それ以外の外部変数と目立った関係を示さないのは理解可能である。

それに対して、主成分Ⅲはまさに「野球ファン度」の指標であり、野球への関心度、ファンクラブへの加入率、外野応援席を選ぶ比率など、納得のいく関係が見られる。野球以外のファンサービスへの期待は、村上（2014）が示したように、比較的高齢なファンだけが目を向ける方向だからであると解される。主成分Ⅳについては、比較的高齢の女性が初観戦の時期が遅く、他球場での観戦経験が少ないことで、これも納得できる方向である。

主成分Ⅴについては、野球以外のファンサービスだけが関わっているが、この主成分が「どんな観戦がしたいか」の項目で、多くのチェックリストに○をつける傾向のある参加者が高得点になる傾向があることを示唆している可能性もある。いずれにせよ、野球ファンとしての熱意の高低とはあまり関係のない次元である。主成分Ⅵについては、多くの外部変数が関係しているが、高得点者が極端な反応を避ける傾向がこちらにも表れている。

主成分Ⅶは高得点者がホームチームのファン、低得点者が相手チームのファンであるというこの次元の意味に明確に対応している。主成分Ⅷについては、遠方からの来訪者が高得点であるといったやや意外な傾向もあるが、初観戦のきっかけが選手にあることなど、やはりこの次元の特徴は現れている。

以上の外部変数との関係は、だいたい各次元の解釈の裏づけとなっている

ると考えられる。8次元という通常の多重対応分析では想定されない多くの次元には、相応の意味を認めてよいのではないと思われる。

### 3. Bootstrap 法による推測統計学的取扱い

#### 3.1 Bootstrap 法の概要

正規直交主成分分析が対象としているようなカテゴリカル・データは、通常の多変量解析の手法が前提としているような多変量正規分布にもとづく標本分布の理論の条件を満たさず、そうした理論にもとづく推測統計学的な取扱いはできない。他方、近年では計算能力の圧倒的向上により、コンピュータ・シミュレーションの技法による標本誤差の推定等の可能性が開けてきた（たとえば、大隅他、1994；Lebart, 2006；Greenacre, 2017）。Bootstrap 法 (Efron & Tibshirani, 1993) はそうした方法の中でも、使用頻度の高いものである。

Bootstrap 法の基本的なアイデアは、データからそれと同じ大きさの標本を無作為に復元抽出する (resampling と呼ばれる) というものである。抽出された標本の中に同一のケースが複数回含まれる一方で、全く抽出されないケースもあり得るため、同じ大きさの標本の間にランダムな変動が生じてくる。そうして抽出された標本にもとづいて同一の分析を行い、それぞれの標本で推定された係数の値の変動を標本変動とみなす。本研究では、係数の標準偏差を標準誤差とみなしている。

Efron & Hastie (2016) は、Bootstrap 法がデータと計算方法さえ確定していれば「自動的」に標準誤差が出力されること、推定バイアスのような他の統計測度にも適用可能であることといった長所を列挙している。しかしながら、主成分分析や多重対応分析では、固有ベクトルの符号が任意である点で、完全に自動的というわけにはいかない。符号のマッチングは最低限必要である (Lebart, 2006)。さらに、回転を伴う場合には、主成分の出力順序が必ずしも一定しないという問題も加わる。したがって、負荷量や重みについては列間の入れ替えや、Procrustes 回転 (Gower &

Dijksterhuis, 2004) によるマッチングといった技法が必要になる (Timmerman, Kiers & Smilde, 2007)。ここで提案した正規直交主成分分析では、さらに変量の符号や順序の問題が加わるため、単なる符号のマッチングと入れ替えのプロセスは非常に複雑なものとなる。したがって、Procrustes 回転の導入は不可避と考えられた。

すなわち、全標本によるターゲットとなる重み行列を  $\bar{\mathbf{W}}'_k$  として

$$g(\mathbf{T}, \tilde{\mathbf{T}}_1, \dots, \tilde{\mathbf{T}}_p | \mathbf{W}_1, \dots, \mathbf{W}_p, \bar{\mathbf{W}}_1, \dots, \bar{\mathbf{W}}_p) = \sum_{k=1}^p \|\bar{\mathbf{W}}_k - \tilde{\mathbf{T}}'_k \mathbf{W}_k \mathbf{T}\|^2$$

を最小化する。ただし、通常の特異値分解によるアルゴリズムを用いた交互最小 2 乗法 (たとえば, Ten Berge, 1993) は、そのままではかなりの頻度で local minima に収束することがわかった。この問題は、下記の適用例においては、順序のあるカテゴリカル変数 (タイプ 2) のみによる予備回転を行うことで、ほぼ 100% 解消できた。

Bootstrap 標本の大きさをどうするかについて、Lebart (2006) のように確率楕円を図示するのに数回でよいとするものもあるが、Efron & Hastie (2016) の標準誤差を求めるなら 200 回で十分、信頼区間を求めるなら 1,000 ~ 2,000 回という示唆にしたがい、安全のために 2,000 回とした。正規直交主成分分析の後述するようなやや特殊な性質からすれば、この判断は適切だったように思われる。筆者の自作プログラム (MATLAB) を通常のラップトップパソコンで実行して、2,000 回の計算に要した実行時間は 30 秒以内であった。

## 3.2 結果の概要

### 3.2.1 固有値の分布

図 13 は、大小順に 8 番目までの固有値の box plot である。相関行列の最大固有値の分布は、この方法以外では見ることができないとして、Efron & Tibshirani (1993)、Efron & Hastie (2016) でも bootstrap 法の

例示に用いられているが、このように2番目以降の固有値についての図示はあまり行われていないようである。この図のボックスの下端と上端はそれぞれ第1四分位数と第3四分位数であり、ボックス内の水平線分は中央値である。さらに第3四分位数と第1四分位数の差(つまり box の高さ)の1.5倍の長さの線分が box の上下に引かれ、その範囲を超える値は外れ値として+で表示されている。固有値の分布はほぼ左右対称であるが外れ値とみなされる値が上下に数個ずつ見られる。線分の上限と下限の間に値が落ちる確率は、正規分布ではほぼ0.993であり、大きさ2,000の標本では上下に7個程度の外れ値が現れることになる。第1固有値では上方に9個、下方1個の外れ値であり、固有値の分布が著しく正規分布と異なっているとは言えない(ヒストグラムからも確認できる)。

それよりも問題は、分布の中央値がもとのデータから算出される固有値よりも、四分偏差1つ分、あるいはそれ以上に以上に過大になっていることであろう。Bootstrap法による固有値は上方にバイアスしている。この原因ははっきりしないが、通常的相关係数にもとづく主成分分析にはほとんど見られない現象である。個々の変数間のクロス集計表のbootstrap標本から算出されるイナーシアが、自由度に比例した偏りをもつことを確

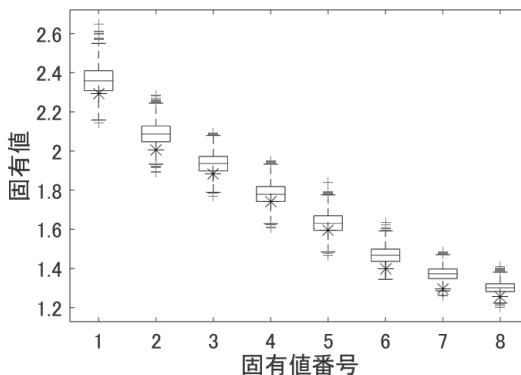


図 13 固有値の box plot \*は全データによる値

認しているが、このあたりに解明のカギがありそうである。このバイアスがこの後の結果にどのように影響するかは不明であるが、結果を見るにあたって常に念頭に置く必要があるだろう。

### 3.2.2 基準化 Burt 行列の固有ベクトルの一致性

次は、対応する固有ベクトルの安定性について見る。こちらは、基準化 Burt 行列  $B$  の固有ベクトルで考える。 $R$  の固有ベクトルは、順序のないカテゴリカル変数に付与される乱数によって変動するからである。全データによる固有ベクトルと bootstrap 標本の固有ベクトルの間の内積をとる。これは一致性係数 (coefficient of congruence) と呼ばれるものになる。これを各固有ベクトルについて図示したのが図 14 である。こちらは、7 番目と 8 番目の固有値に対応するものを除き、かなり負に偏っている。すなわち、過半数の標本で一致性は 0.8 を上回るが、著しく低下する標本もあるということである。これは、同一内容の固有ベクトルに対応する固有値が必ずしも同一の順序で出てこないことによると思われる。第 1 固有ベクトルと第 5 固有ベクトルの一致性が相対的に高いのは、離接する固有値との差が比較的大きいからである。第 7 と第 8 固有ベクトルは、どちらが

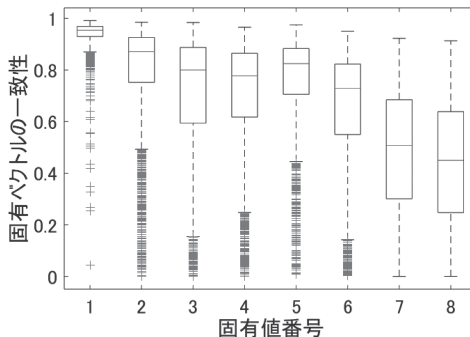


図 14 基準化 Burt 行列の固有ベクトル 全データとの内積



どちらに対応するか (あるいは9番目以降の固有ベクトルとも入れ替わるか) がほとんど予測できないのであろう。ここでは分布はほぼ対称になり、一貫性係数の中央値がだいたい0.5のあたりにとどまっている。

### 3.2.3 正規直交主成分分析の重み行列の一致性

ここまでは、多重対応分析の段階の結果である。次に、正規直交主成分分析の結果に移ろう。図15は、各bootstrap標本の回転後の重み行列  $\tilde{T}'_k W_k T$  を第  $k$  ブロックとする  $(c-p) \times r$  の重み行列の全データによる同じ行列との一貫性係数の box plot である。図14の固有ベクトルと比べれば、全体に一貫性が高く、中央値はほとんど0.9を超えている。これが回転 (と bootstrap 法適用にあたっての Procrustes 回転) の効果である。それでも、デモグラフィックな変数が高く負荷する第I、第II、第IV主成分の一貫性は高く、第VII、第VIII主成分の一貫性は低い。

### 3.2.4 主成分負荷量の標準誤差とバイアス

解釈上は最も重要なパターン (主成分負荷) 行列と主成分間相関行列の要素の標準誤差とバイアスについて検討しよう。表10に2,000回の

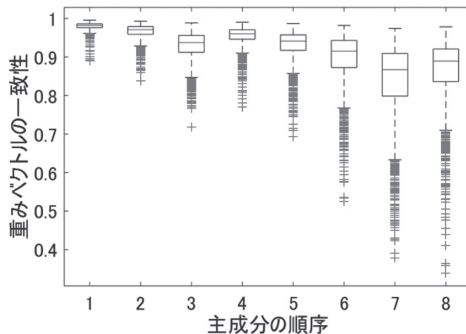


図15 正規直交主成分分析の重みベクトルの一致性

表 10 Bootstrap 法による平均値と標準偏差 (標準誤差)

変量	質問項目	変量内容	I		II		III		IV		V		VI		VII		VIII	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1	1	性別	0.03	0.038	-0.02	0.045	0.02	0.052	-0.77	0.042	0.07	0.053	-0.03	0.062	0.00	0.068	0.04	0.067
2	2	年齢	0.83	0.028	0.81	0.037	0.01	0.038	-0.02	0.035	-0.07	0.038	-0.04	0.045	-0.13	0.047	0.04	0.047
3	3	1次	-0.01	0.028	0.81	0.037	-0.01	0.034	0.02	0.046	0.02	0.053	-0.05	0.064	-0.03	0.071	-0.02	0.067
4	4	2次	0.33	0.037	0.02	0.054	-0.01	0.084	0.73	0.070	0.18	0.078	0.13	0.100	0.52	0.114	-0.14	0.104
5	5	3次	0.06	0.047	0.24	0.073	-0.04	0.124	-0.05	0.119	-0.02	0.119	-0.10	0.155	-0.03	0.171	-0.09	0.159
6	6	4次	0.01	0.039	0.04	0.055	0.12	0.088	0.08	0.085	0.05	0.095	0.03	0.112	-0.07	0.120	-0.38	0.118
7	3	職業	-0.85	0.025	0.02	0.019	-0.01	0.031	-0.02	0.019	-0.07	0.036	-0.05	0.030	-0.18	0.040	-0.02	0.037
8	8	学生(-)や無職(+)	0.02	0.017	-0.81	0.030	-0.05	0.037	0.04	0.021	0.04	0.036	-0.01	0.037	0.00	0.046	-0.02	0.042
9	9	パート・主婦	0.02	0.020	-0.05	0.023	-0.01	0.040	0.78	0.033	0.01	0.038	0.00	0.041	0.13	0.051	0.06	0.049
10	10	公務員	-0.02	0.037	-0.01	0.027	0.10	0.088	-0.01	0.035	0.18	0.088	-0.39	0.101	0.05	0.118	-0.11	0.102
11	11		-0.01	0.038	-0.01	0.020	0.05	0.098	-0.02	0.031	-0.07	0.100	0.00	0.093	0.08	0.129	0.29	0.115
12	12		0.04	0.033	0.01	0.012	-0.08	0.096	0.02	0.020	0.00	0.093	0.00	0.042	-0.19	0.110	0.01	0.045
13	4	ドラゴンズファンか	-0.19	0.048	-0.06	0.058	0.01	0.073	-0.11	0.069	-0.04	0.078	-0.08	0.086	0.48	0.084	0.14	0.100
14	14	1次	0.05	0.041	0.07	0.048	0.65	0.058	-0.05	0.051	-0.13	0.061	-0.10	0.071	0.01	0.087	0.03	0.081
15	5	インターネット	0.38	0.049	0.29	0.056	-0.38	0.063	-0.10	0.056	-0.08	0.065	-0.78	0.071	-0.05	0.085	-0.10	0.083
16	16	1次	0.05	0.051	0.20	0.063	-0.10	0.078	0.03	0.075	-0.08	0.078	0.08	0.071	0.09	0.109	-0.13	0.105
17	17	2次	0.05	0.067	-0.05	0.087	0.06	0.100	-0.11	0.097	-0.14	0.106	0.21	0.126	0.14	0.153	-0.12	0.148
18	6	観戦回数	0.74	0.039	-0.02	0.046	0.41	0.063	0.11	0.051	0.10	0.086	0.31	0.066	0.04	0.103	0.16	0.117
19	19	2次	-0.08	0.075	-0.01	0.101	-0.02	0.138	-0.01	0.109	0.10	0.122	0.04	0.160	-0.15	0.196	-0.16	0.175
20	20	3次	-0.04	0.077	0.09	0.096	0.07	0.119	-0.04	0.106	-0.04	0.116	0.00	0.143	0.08	0.173	0.13	0.156
21	21	4次	-0.36	0.056	0.12	0.073	-0.09	0.088	0.07	0.065	0.46	0.074	-0.01	0.078	0.12	0.095	-0.09	0.103
22	7	誰と来たか	0.05	0.050	-0.12	0.059	-0.10	0.075	-0.23	0.070	-0.06	0.053	0.02	0.073	-0.36	0.103	0.06	0.097
23	23	緊要(+)/vs友人(-)	0.02	0.056	0.78	0.057	0.14	0.085	-0.23	0.070	-0.06	0.053	0.02	0.073	-0.36	0.103	0.06	0.097
24	24	緊要	0.02	0.045	0.00	0.055	-0.10	0.062	0.21	0.057	0.07	0.063	-0.14	0.075	-0.18	0.085	0.56	0.064
25	8	プレーを身近で	0.02	0.038	0.02	0.044	0.69	0.047	-0.05	0.046	0.11	0.051	0.08	0.062	-0.10	0.074	-0.03	0.075
26	9	チームの勝利	0.01	0.048	0.06	0.060	0.08	0.074	-0.10	0.064	0.70	0.079	0.11	0.093	0.04	0.105	0.57	0.095
27	10	気になっている選手	0.10	0.061	0.01	0.080	-0.41	0.087	-0.12	0.081	0.23	0.089	0.21	0.109	-0.07	0.134	0.14	0.123
28	11	いい試合	0.05	0.042	-0.02	0.057	-0.06	0.070	-0.02	0.060	0.63	0.063	0.07	0.090	0.05	0.112	0.12	0.092
29	12	試合を楽しみながら	0.14	0.051	-0.10	0.066	-0.03	0.081	0.00	0.075	0.65	0.075	-0.24	0.099	0.07	0.105	0.10	0.108
30	13	試合を楽しみながら	-0.04	0.054	-0.10	0.066	0.04	0.078	-0.21	0.067	0.48	0.067	0.21	0.075	-0.12	0.106	-0.02	0.102
31	14	雰囲気・わくわく感	0.04	0.051	0.06	0.058	0.31	0.062	-0.05	0.064	0.39	0.067	0.21	0.075	0.00	0.090	0.10	0.092
32	15	一体感	0.08	0.057	-0.30	0.070	0.13	0.090	-0.04	0.077	-0.06	0.092	0.23	0.102	-0.05	0.124	-0.13	0.123
33	16	満足度	-0.22	0.066	0.10	0.079	-0.08	0.105	0.09	0.085	0.14	0.098	0.28	0.117	-0.02	0.149	-0.10	0.141
34	34		0.16	0.054	-0.07	0.062	0.13	0.083	0.13	0.082	0.16	0.081	-0.21	0.096	-0.32	0.110	0.16	0.105
35	35	平均値	0.046	0.055	0.055	0.077	0.077	0.077	0.062	0.062	0.062	0.074	0.074	0.086	0.104	0.104	0.097	0.097
主成分分析			-0.044	0.020														
			0.048	0.030	-0.004	0.031												
			-0.118	0.026	0.007	0.026	-0.035	0.031										
			0.010	0.031	-0.030	0.032	0.022	0.033	-0.050	0.032								
			-0.043	0.031	0.003	0.030	0.080	0.029	0.061	0.031	0.055	0.029						
			-0.056	0.024	-0.087	0.025	-0.132	0.030	0.039	0.029	0.017	0.030	0.011	0.028				
			-0.026	0.032	0.058	0.030	-0.015	0.032	-0.009	0.032	0.071	0.032	0.005	0.030	-0.013	0.029		

resampling にもとづく平均値と標準偏差を示した。2節で表8にもとづいて行った主成分の解釈にあたっては、社会調査における質問項目の多様性(内的整合性の低さ)により、通常の因子分析的方法における基準よりもかなり小さい負荷量に依拠せざるを得なかった。あくまでも記述目的に限定しても、標準誤差にもとづく有意性の判断ができることは重要である。2,000個の推定値から算出されるパラメータの推定値を $\hat{\theta}^*$ 、標準偏差(標準誤差)を $\widehat{se}_{boot}$ と書くとすると、分布の正規性を前提に、帰無仮説 $H_0: \theta = 0$ に対する検定統計量は、

$$z = \frac{\hat{\theta}^*}{\widehat{se}_{boot}}$$

によって算出できる。通常の両側.5%の臨界値は1.96であるから、 $z$ が1.96(おおよそ2)を超えていれば有意とみなしてよいと考えられるかもしれない。

しかしながらここで問題になるのは、2.4でも述べた検定の多重度である。因子負荷量だけでも、変数の数 $\times$ 主成分数 $=35 \times 8=280$ に及ぶから、すべての負荷量の真の値が0であったとしても、その5%である14個程度は帰無仮説が棄却されることになるであろう。一つの考え方としては、Bonferroniの基準、すなわち両側5%水準の確率0.025を検定の数280で割った0.00008926に対応する標準正規分布の値の絶対値、3.7475を臨界値とすることがある。そうした極端な分布の端の値を用いることについての懸念もあるが、これに基づいて有意と判断される数値を太字で示した。また、いささか恣意的であるが、この基準には達しないものの、 $z$ が3を超える数値についてはイタリックとした。これらは、すでに表8において採用し、解釈の参考としている。

標準誤差の値を主成分ごとに比較するために平均値を求めて表10の下端に示した。図15の重み行列の変動の大きさからも推測されるように、デモグラフィックな要因が負荷する主成分I、II、IVの標準誤差は概して小さく、寄与の小さい主成分VIIとVIIIでは大きい。それでも解釈のキー

となるような変量については、おおよそ有意と判断できたのは幸いであった。

バイアスについては、全データによる結果（表8）と bootstrap 平均値との差を求めた。有意になるような目立つ負荷については、より0に近い方向に偏りが出ているものの、主成分 I～VI に関してはおおよそ 0.05 以内に収まっている。しかし、主成分 VII と VIII では、一部 0.1 を超える数値も見られる。

いくつかの問題は指摘されるものの、記述的多変量解析の適用場面での推測統計学的取扱いについて、今回の試みは、ある程度の成功例と考えてよいのではなかろうか。

#### 4. 討論

重回帰分析、主成分分析、因子分析、正準相関分析といった多変量解析の手法は、基本的に量的データを対象とし、算出される統計測度にもとづいて結果を解釈し、推論に用いることを目指して開発されてきた。他方、

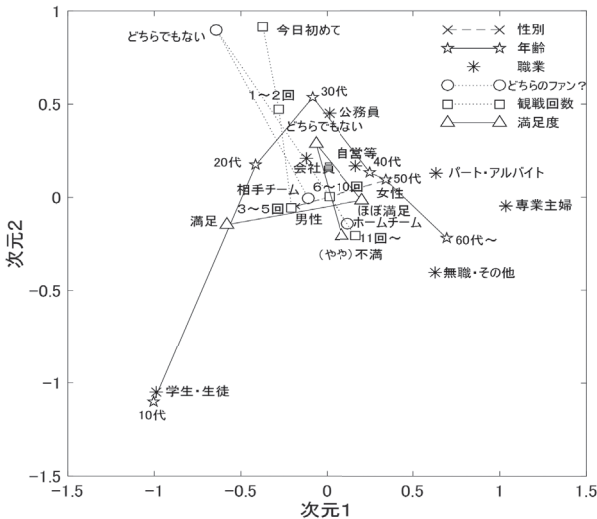


図 16 多重対応分析の重みの図示

対応分析や多次元尺度法は、結果の空間的表現にもとづき、視覚的な現象理解を目指す方法である。それに対して、biplot (Gower & Hand, 1994) のように、広範囲な多変量解析について、視覚的な表現を提供する試みがある。本研究の正規直交主成分分析はそれとは逆に、視覚的な方法に量的データのための多変量解析の特徴を導入する試みであったといえよう。この試みのコンセプトをあえて1点に絞って述べれば、質的データの観点からみた変数間関係の評価尺度であるイナーシアを量的データの変数間関係の尺度である相関係数に読み替えたということになるうか。

#### 4.1 多重対応分析の限界をどれだけ超えられたか

この試みがどこまで成功したか、本研究の冒頭であげた多重対応分析の3つの限界という観点から考えてみよう。まず、多重対応分析がせいぜい3次元の解にとどまっていたのに対して、もっと多くの次元(適用例では8次元)が解釈可能になったという点では、はっきり進歩が認められるであろう。次元ごとにかなり明確な解釈が付き、各次元と外部変数との関係の検討も一元配置分散分析と相関比の2乗を用いて容易に行うことができた。次元の意味を言語化するという観点からは有効に機能したと見てよいのではなかろうか。

同じデータの一部の変数を多重対応分析によって分析し、2次元目までの重み行列を図示したのが図16である。より多くの変数を投入した分析も試みたが、特に2値変数の重みが小さく、それらを図に描きこむと原点付近に点が集中しすぎて到底、読み取ることができなかった。結局、分析の対象は図の右上の凡例に示した6変数にとどめたが、それでも性別、とくに女性の点が年齢段階の50代と重なって隠れていたりしている。この平面は、本研究の主成分IとIIをそれぞれ縦軸と横軸にもち、主成分IIIとIVの成分がいくらか重なっていると読むことはできるが、やはり2次元の限界は現れている。たとえば、この横軸は大体年齢段階に対応しているが、性別もほぼその軸と平行しており、女性の方が全般に高齢者が多い

ように見える。しかし、現実には男性が10代でやや多く、女性の数が30代でやや上回ることを除けば、男女の年齢差はほとんどない。また、縦軸は上方向が観戦ビギナー、下方向がリピーターを表していると読み取れるが、下方にとびぬけた位置にある10代の学生・生徒が必ずしもリピーターであるわけではない。

ただ、この問題は「年齢段階」、「性別」、「職業」という相互に堅固に結びついたデモグラフィックな要因を分析に加えたことに起因するとも言える。これらのカテゴリーが平面上で極端な位置を占めるために、行動や意識レベルの変数の位置が原点付近に集中してしまい、それらを分析に含められなかったからである。デモグラフィックな要因を外した分析を行うことで、よりフェアな比較が可能になるのは事実であろう。

第2の分析上のアーチファクトとしての非線形関係の問題については、今回の主成分VIのような、参加者の反応スタイルを示す次元が現れ、それが参加者の属性（公務員、家族と観戦）と結びつくことが明らかになった点で、現象の一部の説明が可能になった。ただ、従来から特に問題にされている2値データにおける馬蹄現象（たとえば、Guttman, 1954; 村上, 2011; Murakami, 2012）をはじめ、なお多くの解明されるべき問題が残されている。

第3の推測統計学的扱いの不十分性については、bootstrap法の使用によって、一定の成果をあげられたと見てよいであろう。本研究では、Procrustes回転を双方向から用いるという工夫によって、それぞれの次元の統計的な特徴を明らかにすることができた。また、多重度を考慮した上で、主成分負荷量の有意性をともかくも判断することができた。この点は、特に内的整合性の低い多様な質問項目への反応からなる社会調査データの分析では重要である。ただし、固有値に関する無視しえないバイアス、これがパラメータ推定にもたらしている影響等、（筆者の知る限り）未解決の問題が残されている。

## 4.2 正規直交主成分分析の強みと弱み

正規直交主成分分析という方法に限定して、その長所と短所について論じておきたい。この方法は、20世紀半ばにはすでに定式化され、社会調査データの分析について多くの蓄積をもつ多重対応分析の解を、その許容される変換 (admissible transformation) の範囲、具体的には、多重対応分析の重み行列を、数量化ベクトルの行列と重みベクトルの行列に分解した上で、解の説明力 (寄与) を変えない直交回転によって、空間表示にもとづく視覚的な解釈を、パターン行列の数値にもとづく解釈に移行させたものである。その点において、多重対応分析の線形代数にもとづく頑健なアルゴリズムを引き継いでおり、bootstrap 法による 2,000 回の反復も 30 秒以内で実施できる計算のスピードや、イナーシアにもとづく解の評価など多重対応分析の理論的な蓄積も利用できることは強みと言えるであろう。もちろん、本来の目的であった質的データの 3 次元を超える分析については、著者の知識の範囲で同じことのできる方法は存在しないと信ぜられる。

しかし、パターン行列の解釈が、数量化ベクトルを参照しないと行えないことは、正規直交多項式を援用したとしても、やはり視覚的表現による直観的解釈にくらべて煩雑であることは否めない。また、回転により、高次元の解が低次元の解をそのまま含んでいるという解が「ネスト」しているという性質が失われていることも、複数の次元の解を比較しながら最上の解を探す過程を煩瑣なものにする。さらに、数量化ベクトルと重み行列の符号の反転操作も面倒であるが、これを行わないと解釈において錯誤が起りやすい。ともかく、非常に使い勝手のよい方法とは言えないであろう。このあたりは、プログラム上の工夫によって多少とも改善できる余地はあると思われるので検討を続けたい。

今回分析したプロ野球の観客のデータで、何か新たな発見があったか、というのは実用性の評価としては必要な観点であろう。村上 (2014) は 2011 年度のデータを用いて、今回の「どんな観戦がしたいか」という質

問項目に含まれていた13のチェックリストへの反応を単純に主成分分析にかけ、今回の主成分 III、V、VIII に相当する主成分を導き出した上で、外部変数となっていたデモグラフィックな変数等との関連を検討した。手順としてははるかに単純な方法で、「ファンは人間関係を求めて球場に来る」という結論を得ている。今回、それよりはいささか複雑な方法によって得られた8つの主成分のうち、I、II、IVの3つはデモグラフィックな要因に関わるもの、主成分 VI は社会調査の方法論上は興味深いが、ファンの特徴の記述としての意義は小さく、そうしてみると新しいと言えるのは主成分 VII のホームチームのファンである程度の次元であるが、これは常識の域を出るものではない。この意味では、少なくとも今回の分析でかけたコストに見合うだけの成果があったとは言い難いかもしれない。

本研究はいわば試作車の試運転のようなものであり、今後に向けていくつかの問題提起ができたことに、とりあえず満足すべきであろう。幸いこの方法は、データの形について強い制約を課すことがない。標準的な社会調査データへの適用を積み重ねることにより、この方法ならではの結果が得られることを期待ながら、分析を積み重ねていきたい。

## 注

- 
- <sup>1</sup> 本研究は文部科学省科研費 基盤研究(C)15K04197の支援により行ったものである。
  - <sup>2</sup> 中京大学名誉教授、現代社会学部非常勤講師
  - <sup>3</sup> 式(6)の右辺が $1/\sqrt{p}$ となっているのが普通である (Gower & Hand, 1996)。
  - <sup>4</sup> 乱数を与えるところで変数が反転してしまうと解釈が混乱するので、むしろ適用すべきでない。
  - <sup>5</sup>  $A$  の列も相互に直交するから  $A'A$  も対角行列となる。
  - <sup>6</sup> この対策として、因子分析の主因子法と同様の反復計算を用いる Joint correspondence analysis (JCA) が提案されている (Greenacre, 2017)。



- <sup>7</sup> この調査は、2006年以來、(株)中日ドラゴンズと(株)ナゴヤドームの協力のもと「ドラゴンズ・プロジェクト」としてほぼ毎年1回実施されてきた。こうした調査の実施を可能にくださった球団・球場関係者の皆さん、プロジェクトの立ち上げにあたってご尽力いただいた西山哲郎現関西大学教授、調査開始以來、球団との調整にあたっていただいた中京大学の鰻淵真司職員をはじめとする中京大学教職員各位、毎年ゼミ単位で調査に協力して下さってきた齊藤尚文教授とゼミ生の皆さん、中京大学硬式野球部の半田卓也監督と部員の皆さん、そして調査の準備段階から球団事務所でのプレゼンまで全力で取り組んでくれた筆者が担当した「社会調査実習」、「調査研究法」履修者諸君に心からお礼申し上げます。
- <sup>8</sup> こうしたことが可能である点も、MCA とのつながりがはっきりした方法であることの強みであるともいえよう。

## 文 献

- 足立浩平・村上 隆 (2011) 非計量多変量解析法主成分分析から多重対応分析へ  
朝倉書店
- Beh, E. J., & Lombardo, R. (2014). *Correspondence analysis: Theory, practice, and new strategies*. Chichester: Wiley.
- Bekker, P., & De Leeuw, J. (1988). Relations between variants of non-linear principal component analysis. In Van J. L. A. Rijkevorsel, & J. De Leeuw, (Eds.) (1988). *Component and correspondence analysis: Dimension reduction by functional approximation*. Chichester: Wiley, pp.1-31.
- ベネット, T.・サヴィジ, M.・シルヴァ, E.・ワード, A.・ガル=カヨ, M.・ライト,  
D. (磯 直樹・香川めい・森田次郎・知念 渉・相澤真一 訳) (2017) 文化・階級・卓越化 青弓社
- Benzecri, J.P. (1992). *Correspondence analysis handbook*. New York: Marcel Decker.
- Blasius, J. & Greenacre, M. (Eds.) (2014). *Visualization and verbalization of data*. Broken Sound Parkway NW: CRC Press.
- De Leeuw, J. (2006). Nonlinear principal component analysis. In M. Greenacre. & J.

- Blasius (Eds.) *Multiple correspondence analysis and related methods*. Boca Raton FL: CRC Press, pp. 107-133.
- De Leeuw, J. (2014). History of nonlinear principal component analysis. In J. Blasius & M. Greenacre, (Eds.) *Visualization and verbalization of data*. Broken Sound Parkway NW: CRC Press, pp. 45-60.
- Efron, B. & Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidences, and data science*. Cambridge: Cambridge University Press.
- Efron, B. & Tibshirani, R.J. (1993). *An introduction to the Bootstrap*. New York: Chapman & Hall.
- Gifi, A. (1990). *Nonlinear multivariate analysis*. Chichester: Wiley.
- Gower, J.C. & Dijksterhuis, G.B. (2004). *Procrustes problems*. Oxford: Oxford University Press.
- Gower, J.C. & Hand, D.J. (1996). *Biplots*. London: Chapman & Hall.
- Greenacre, M. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.
- Greenacre, M. (2017). *Correspondence analysis in practice. 3rd ed.* Broken Sound Parkway NW: CRC Press.
- Greenacre, M. & Blasius, J. (Eds.) (2006). *Multiple correspondence analysis and related methods*. Boca Raton FL: CRC Press.
- Groves, R.M., Fowler, F.J., Couper, M.P., Lepkowski, J.M., Singer, E., & Tourangeau, R. (2009). *Survey methodology, 2nd ed.* Hoboken, NJ: Wiley.
- Guttman, L. (1954). The principal components of scalable attitudes. In P. F. Lazarsfeld (ed.) *Mathematical thinking in the social sciences*. New York: Russell & Russell, pp.216-257.
- Hamilton, D.L. (1968). Personality attributes associated with extreme response style. *Psychological Bulletin*, 69, 192-203.
- Harris, C.H., & Kaiser, H.F. (1964). Oblique factor analytic solutions by orthogonal transformations. *Psychometrika*, 29, pp.347-362.

- 林 知己夫 (1993) 数量化 理論と方法 朝倉書店
- 伊藤大幸・谷 伊織・平島太郎 (2018) 心理学・社会科学研究のための構造方程式モデリング *Mplus* による実践 ナカニシヤ出版
- Kiers, H.A.L. & Ten Berge, J.M.F. (1994). The Harris-Kaiser independent cluster rotation to simple component weights. *Psychometrika*, 59, 81-90.
- Le Roux, B. & Rouanet, H. (2010). *Geometric data analysis: From correspondence analysis to structural data analysis*. Dordrecht: Kluwer Academic Publishers.
- Le Roux, B., Rouanet, M., Savage, A., & Warde, A. (2008). Class and cultural division in the UK. *Sociology*, 42(6), 1049-1071.
- Lebart, L. (2006). Validation techniques in multiple correspondence analysis. In M. Greenacre & J. Blasius (Eds.) *Multiple correspondence analysis*, Boca Raton FL: CRC Press, pp.179-195.
- Levy, R. & Mislevy, R.J. (2016). *Bayesian psychometric modeling*. Boca Raton, NW: CRC Press.
- 村上 隆 (1998) こだわり派の主成分分析(4) 名古屋大学大型計算機センターニュース, 29(4), 325-353
- 村上 隆 (2011) 特別な構造をもつ2値データの相関行列の性質について 中京大学現代社会学部紀要, 5, No.1, 107-124.
- Murakami, T. (2012). A geometrical interpretation of the horseshoe effect in multiple correspondence analysis of binary data. In W. Gaul, A. Geyer-Schulz, L. Schmidt=Thieme, and J. Kunze (eds.) *Challenges at the interface of data analysis, computer science, and optimization*. Springer, 101-108.
- 村上 隆 (2012) 数量化3類と多重対応分析 社会と調査, No.9, 46-62.
- 村上 隆 (2014) なぜプロ野球ファンは球場で観戦するのか? 「3つの人間関係」の提案 中京大学現代社会学部紀要, 7(2), 175-207
- 村上 隆 (2016) 多重対応分析の変換としての正規直交多項式主成分分析 - Likert型項目の探索的分析のための新たな手続き - データ分析の理論と応用, 5, No.1, 27-47.

- 村上 隆 (2018) 多重対応分析の「因子分析」としての使用 日本分類学会第 37 回大会予稿集, 39-42
- NHK 放送文化研究所 (編) (2015) 現代日本人の意識構造 (第八版) NHK 出版
- Nishisato, S. (1980). *Analysis of categorical data: Dual scaling and its applications*. Toronto: University of Toronto Press
- Nunnally, J.C. (1978). *Psychometric theory*. New York: McGraw Hill.
- 大隅 昇・ルパール, L・モリノウ, A・ワーウィック, K.M・馬場康維 (1994) 記述的多変量解析法 日科技連
- Pett, M.A., Lackey, N.R., & Sullivan, J.J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research*. Thousand Oaks, CA: Sage Publications.
- Rovan, J. (1994). Visualizing solutions in more than two dimensions. In M, Greenacre & J. Blasius (Eds.) *Correspondence analysis in the social sciences*. London: Academic Press, pp.210-229
- 芝 祐順 (1979) 因子分析法 (第 2 版). 東京大学出版会
- Ten Berge, J.M.F. (1993). *Least squares optimization in multivariate analysis*. Leiden: DSWO Press.
- Timmerman, M., Kiers, H.A.L. & Smilde, A.K. (2007). Estimating confidence intervals for principal component loadings: A comparison between the bootstrap and asymptotic results. *British Journal of Mathematical and Statistical Psychology*, **60**, 295-314