

Radex モデルを利用した Guttman Scale の視覚化

宮 寺 貴 之
博士（学術）

総合研究大学院大学
複合科学研究科
統計科学専攻

平成 27 年度
(2015)

目次

第1章 研究の目的	1
1.1 研究の背景	1
1.2 研究の目的	2
1.2.1 項目の該当割合への着目	3
1.2.2 行動科学における二値データ	3
1.2.3 二値変数間の連関	4
1.3 本論文の構成	5
第2章 Guttman Scale と連関係数	6
2.1 Guttman Scale	6
2.2 Guttman Scale の評価指標に連関係数を利用できない理由	9
2.3 Guttman Scale の構成度の指標	14
2.3.1 Reproducibility 係数	14
2.3.2 Scalability 係数	16
2.3.3 Loevinger による Homogeneity 係数	20
第3章 多次元尺度構成法—Guttman Scale に関連して	25
3.1 等質性分析の数理モデル	25
3.2 Kruskal の多次元尺度法 (MDS) の数理モデル	27
3.3 Guttman の SSA	29
3.3.1 SSA の数理モデル	29
3.3.2 距離順位数の写像	32
3.4 Facet 理論と Radex モデル	36
3.5 適用するデータ	40
3.5.1 人工データ	40
3.5.2 メディア暴力調査データ	44
3.6 データへの適用	49
3.6.1 等質性分析の適用	49
3.6.2 Kruskal の MDS の適用	53
3.6.3 Guttman の SSA の適用	56
3.7 Facetted SSA	58
第4章 Radex モデル探索支援手法 (RADXEA) の提案	60
4.1 問題と目的	60
4.2 Radex モデルにおける Modular Facet の探索方法	62
4.3 Radex モデルにおける Polar Facet の探索方法	63

4.4	人工データへの適用	69
4.5	メディア暴力調査データへの適用	74
4.6	まとめ	80
第 5 章	Radex モデル多重尺度構成法 (RADXMS) の提案	81
5.1	問題と目的	81
5.2	Radex モデル多重尺度構成法の理論的背景	81
5.2.1	Radex モデルにおける Modular Facet の表現	82
5.2.2	Radex モデルにおける Polar Facet の表現	82
5.2.3	Homogeneity 係数を変換する関数	83
5.2.4	ブートストラップ法による布置の安定性	90
5.2.5	3 次元への拡張	91
5.3	人工データへの適用	92
5.4	メディア暴力調査データへの適用	97
5.4.1	メディア暴力調査データによる布置の作成	97
5.4.2	ブートストラップ法による布置の安定性の検討	99
5.4.3	他の視覚化手法との比較	108
5.5	まとめ	118
第 6 章	総括	119
6.1	Radex モデル探索支援手法 (RADXEA)	119
6.2	Radex モデル多重尺度構成法 (RADXMS)	120
6.3	本研究における提案手法の意義	121
6.4	多値データへの適用	122
6.5	データの視覚化の意義	123
文献	127
謝辞	132
付録 A	2 変数間の Homogeneity 係数の理論上の最小値	133
付録 B	よく見る映画・ドラマの特徴に関する質問文及び質問項目	136
付録 C	Facetted SSA の数理モデル	137
付録 D	Homogeneity 係数を変換する関数 g について	140
付録 E	η^2 の θ_i による偏微分 (2 次元)	141
付録 F	η^2 の θ_i による偏微分 (3 次元)	142
付録 G	R プログラム	143
G.1	人工データ 2 の作成	143
G.2	等質性分析の実施	144
G.3	Kruskal の MDS の実施	144
G.4	Radex モデル探索支援手法	145

G.5 Radex モデル多重尺度構成法.....	149
---------------------------	-----

第1章 研究の目的

本研究は、行動科学的データを対象に、項目のカテゴリの割合を考慮した 2 次元上の布置を作成し、一次元性を満たす尺度を構成する項目群を視覚的に抽出する方法について検討する。特に、二値データに焦点を当て、Guttman Scale に着目し、Facet 理論で仮定される領域モデルの一つである Radex モデルを利用した新たな視覚化手法を提案する。

1.1 研究の背景

人間の行動を対象とした研究の例として、殺人に関する犯行時の行動の類型化に関心がある場合を考える。類型化を行うことによって、新たに殺人事件が発生した際に、捜査によって被疑者の絞り込みを効率的に行い、初動捜査を支援することが必要であるためである。殺人の類型化を行うための一つの指標として、犯行の計画性の程度を測定したいものとしよう。例えば、計画性が高い犯行は、犯行準備から逃走経路の確認まで用意周到に準備し、怨恨から殺人を行ったケースが考えられる。一方、計画性の低い犯行は、言い争いの中でカッとなって衝動的に殺人を行ったケースが考えられる。また、犯人が成人の場合と少年の場合では、計画性の程度が異なることも考えられる。例えば、少年による犯行では、犯行の準備は入念だが、犯行を行った後の逃走や証拠隠滅のことは全く考えていないといった、計画性が中程度のケースも考えられる。

このように、人間の行動を理解しようとする際には、観察される行動の背後に抽象的な概念（上の例では、犯行の計画性）を想定することが多い。その抽象的な概念は、測定の対象となる人間に直接尋ねることで測定する場合もある。しかし、上の例のような犯行の計画性の程度を犯罪者に 10 件法で尋ねたとしても、適切なデータは得られにくい。このような抽象的な概念は、顕在的な行動として直接的に観察できる対象ではないこともその理由の一つである。そこで、こうした抽象的な概念は潜在的な行動特性と仮定して考える方が都合がよい。

潜在的な行動特性を表す構成概念を定義し、その構成概念と関連すると思われる具体的な顕在的な行動を記述した項目を作成する。構成概念を単一の顕在的な行動で表現することは難しいため、通常は顕在的な行動を調べる項目を複数作成することが一般的である。そして、それらの項目への回答を求めることによってデータを収集する。項目に対する「該当あり」または「該当なし」として回答を求めることで、二値データが収集される。

先ほどの「犯行の計画性」の例でいえば、被害者は面識関係のある相手であったか、犯行時間帯は夜間であったか、凶器を準備していたか、犯行後の証拠の隠滅があったか、逃走経路を事前に確認していたか、といった項目が考えられる。各事件について、これらの項目によって表される各変数に対して、該当の有無として回答を求め、「該当あり」または「該当なし」という二値にコードすることによってデータセットを作成し、犯行の計画性を測定する方法を検討することになる。

このように、観察対象となる行動特性を仮定し、その特性を測定するために用意された複数の項目から、特性の度合いを表現するために数値を割り当てる仕組みを尺度と呼び、その尺度の作成に用いる項目を尺度項目と呼ぶ。さらに、これら複数の項目を用いて数値化する尺度が、単一の特性または構造について対象間の関係を適切に数値化しているとき、一次元性が成立しているといえる (McDonald, 1981; Hattie, 1985)。通常、尺度はある一つの特性を測定するために項目を用意して作成されるため、尺度は一次元性をもつことが意図されている。

人間の行動のある行動特性を測定するために尺度を作成する際、直接的に観察できる顕在的な項目を収集して、尺度項目の候補を準備する。そして、それらの項目間の相関を検討することによって、測定したい行動特性と関連を持ち、互いに関連性の強い項目を選ぶことによって、尺度項目を選定する。一方、行動特性に関わる構成概念が広く、それらに関連する項目群の中から、構成概念としてのまとまりを探索的に抽出する、というアプローチも考えられる。例えば、殺人の手口に関する顕在的項目に関するデータを分析する中で、殺人の動機の違いによって互に関連する項目が異なるため、それらの項目をクラスターとして分けていくことによって、いくつかの殺人の動機に関する構成概念が抽出されるような場合である。これは、様々な行動を調べる項目の分析を通じて、一次元性を示す尺度の構成に利用できる項目群を特定し、現象の背後に想定できる行動特性を定式化するというケースであり、尺度作成における探索的手法といえよう。

1.2 研究の目的

本研究は、人間の行動を測定する複数の項目の中から、ある一つの行動特性を測定する項目群を特定することによって、一次元性をもつ尺度の定式化を支援する分析手法を検討する。さらに、項目群の中から複数の尺度が定式化できる場合、それら複数の尺度項目群同士の関係を視覚的に把握することを目指す。視覚的な提示手法は、尺度同士の関係に対する直感的な把握を可能とし、その測定の対象となる構成概念に関わる項目を要素とする全体集合について構造的に理解することにつながる。そのため構成概念のよりよい理解と概念の精緻化をもたらし、探索的手法として望ましいと考えられる。

特に、仮説生成を目的とした研究では、複数の変数によって表されたデータの分析を通じて複数の尺度項目群を抽出するという作業が行われる。例えば、因子分析は複数の変数から尺度を構成する変数群を特定することによって、複数の尺度を検出する方法として利用できる。しかし、因子分析では複数の尺度間の関係を視覚的に表現することは難しい。尺度間の関係をより直感的に把握することを可能とする視覚的な分析手法として、本研究では多次元尺度構成法 (MDS) の考え方を利用する。具体的には、変数を空間内に配置することで、変数間の関係を示す布置を作成する分析である。なお「布置」については岡太・今泉 (1994, p.4) を参照されたい。

1.2.1 項目の該当割合への着目

さらに、従来の MDS の分析とは異なり、本研究が目指す分析手法は、各行動を示す人の割合に着目する。割合とは、項目のカテゴリに該当する人数の割合を意味するが、本研究で焦点を当てる二値データの場合は、項目への該当（または非該当）の割合を指す。二値データにおける項目への該当の割合のことを以下では項目の「該当割合」あるいは単に項目の「割合」と表現することとする。

人間の行動には、当然のことながら多くの人間が示す行動と、少数の人間のみが示す行動がある。ここで、尺度の構成において該当割合の低い項目に着目すると、尺度の特徴がより把握しやすくなる場合がある。

例えば、先ほどの殺人における犯行の計画性について、犯行の予行演習を行う、あるいは、犯行の計画メモを何ヶ月も前から作成する、という項目は、殺人の犯人が行った行動として割合は低いと考えられる。しかし、そうした割合の低い項目に該当する犯人ほど、事前の計画性は極めて高かったと考えられる。このように、該当割合の低い項目が測定の対象とする行動特性をより強く示すことが考えられる。

犯罪行動の分析に MDS を利用した先行研究 (Canter, 2000) では、変数の布置を視覚的に吟味する際、犯罪者の行動の背後に想定される犯行テーマを探索的に検討し、犯行の動機に関わる尺度を抽出するが、該当割合の低い変数ほど犯行テーマの特徴をよく表していることが指摘されている。つまり、行動科学のデータ分析では、項目の該当割合も考慮すべき重要な要素となる。そこで本研究では、各項目の該当割合を項目の布置に反映させる分析手法を提案する。

1.2.2 行動科学における二値データ

先に述べたとおり、本研究では二値データを対象とする。ある行動特徴に該当する程度は、例えば「よくある」、「ときどきある」、「ない」というように 3 件法で尋ねることもできる。回答者が項目に回答するための十分な情報を持っていることが期待できる場合は、回答の評定段階の数を増やすことで、より詳細なデータを測定することが可能となる。

しかし、回答者が十分な情報を持っていない場合、評定段階の数を増やしても、確からしさをもって回答がなされるわけではないため、回答の変動は誤差によって生ずることになる。このようなときは、回答の評定段階の数を少なくすることが必要である。特に、人間の行動特徴を第三者が評価する場合、回答者が評定を行うに足るだけの情報を持っていない状況はよくみられる。

例えば、本章の初めの部分で例示した犯行の計画性の研究において、犯人について記述された捜査資料から特定の特徴を測定する項目について評定を行う、という場合がこれにあたる。このようなときは、回答の評定段階の数を少なくする方が、誤差の変動を拾ってしまうリスクが少ない。評定段階の数がもっとも少ないのは 2 件法である。

行動科学において、二値データが扱われることは非常に多い (Mokken, 1971)。本研究で

は2件法で測定したデータの分析手法について論じる。なお、3件法以上で測定された多値データの分析については第6章4節で言及する。

二値データを構成する各変数は「該当」と「非該当」という2つのカテゴリから構成されることから二値変数と呼ばれる。本研究では、項目に該当する割合のことを「項目の該当割合」と言及する。また、データにおける変数と項目との関係についてみると、項目は3カテゴリ以上の場合があり、それを変数として表すには2つ以上のダミー変数が必要になる。しかし、本論文では項目への該当または非該当を一つの二値変数で表すことから、項目と変数は同じ意味で用いる。そのため、「項目の該当割合」と同じ意味を持つものとして「二値変数の割合」という表現を用いることとする。

二値変数の割合はテスト理論における通過率を意味し、項目間に想定される潜在的な困難度（豊田, 2002）を反映していると考えられる。項目が一次元の尺度を構成しているとき、困難度の高い項目に該当する個人は、困難度の低い項目にも該当しやすいという関係が考えられる。なぜならば、ある能力特性を測る尺度を構成している項目においては、その能力特性が高い個人は、困難度が高い項目に該当しやすく、困難度の低い項目には一層該当しやすいといえるからである。

逆に、その能力特性が低い個人は、困難度の低い項目にしか該当しない、という状況が想定される。そのため、割合の低い項目に当てはまる個人は割合の高い項目にも当てはまる、という関係が想定されるわけである。これは、いわゆる態度尺度の一つとして発展した Guttman Scale (Guttman, 1944) の考え方である。Guttman Scale では、一連の尺度項目について、一つの項目への該当は必ずそれより割合の高い全ての項目への該当を前提とする。これは類積尺度とも呼ばれる（竹内, 1989, p. 798）。本研究では、項目の該当割合を考慮する方法として、この Guttman Scale に着目する。

1.2.3 二値変数間の連関

二値変数間の関係はクロス集計表によって考えることができる。通常、2つの二値変数間の関係は、例えば四分点相関係数（ファイ係数）などの連関係数によって測定されることが多い。二値変数間の連関を考える場合、クロス集計表では2変数ともに該当する者のセル度数及び2変数ともに該当しない者のセル度数が多いほど、2変数の連関係数は高くなる。つまり、どちらか一方の変数のみに該当する者のセル度数が少ないほど、2変数の連関係数は高くなる。これは、それぞれの二値変数の割合の差が小さいほど2変数の連関係数は高くなることを意味する。そのため、割合の差が大きい2変数では、連関係数は高くない傾向がある。

しかし、割合が異なる変数群から Guttman Scale を構成する変数を特定するためには、割合の低い変数に該当するケースは割合の高い変数にも該当するが、その逆は必ずしも当てはまらない、という Guttman Scale のもつ「非対称」な関係を満たすような連関を検討できる指標が必要となる。本研究では、このような Guttman Scale の特徴の程度を表す指標を利

用することによって、変数の 2 次元上の布置から Guttman Scale の性質を満たす変数の特定を視覚的に行う手法を提案する。

1.3 本論文の構成

以下、これ以降の章の概要を説明する。第 2 章では、Guttman Scale の性質について具体的に説明する。また、割合の異なる変数の間にみられる Guttman Scale の性質を示す度合いを表現するために、Guttman Scale の構成度を表す指標について概説する。特に、二値変数の割合を考慮した分析を行う上で、二値変数の連関係数として相関係数を利用する場合の問題点を併せて指摘する。第 3 章では、尺度抽出を行うための視覚化手法について、関連する各種の分析方法について概観する。しかし、視覚化を行う分析手法では、得られた布置に対する解釈が主観的にならざるを得ないという問題点が指摘される(岡太・今泉, 1994)。そこで、布置の解釈を支援するため、データを利用した客観的手法を検討する。その考え方として、空間領域に関するモデルを仮定する Facet 理論を利用する。Facet 理論については第 3 章の最後で説明する。第 4 章では、この Facet 理論における領域の考え方をを用いて、MDS の布置に対する解釈を支援する新たな分析手法の提案を行う。さらに第 5 章では、この MDS の考え方を拡張して、二値変数の割合を反映した複数の尺度間の関係を視覚化する新たな MDS の手法を提案する。最後となる第 6 章では、本研究のまとめ及び今後の課題について述べる。

第 2 章 Guttman Scale と連関係数

この章では、第 1 章で言及した Guttman Scale の性質について説明するとともに、その Guttman Scale の構成度を表す指標を概観する。また、二値変数の割合を考慮した変数間の関係を示す指標を利用する上で、Guttman Scale の構成度を用いなければならない理由について言及する。そして、本章で説明する Guttman Scale の構成度の一つである Homogeneity 係数 (Loevinger, 1947, 1948) を 4 章以降で利用する。

2.1 Guttman Scale

二値データとして収集された項目から尺度を構成する方法として、Guttman (1944, 1950) が提案した Guttman Scale を以下に説明する。ここで、Guttman (1944) による例示を引用し、次のような数学の問題に対する解答を考える。

問題 a 半径 r の円の面積を求めよ。

問題 b 式 $ax^2 + bx + c = 0$ を満たす x の値を求めよ。

問題 c $\frac{dx}{dx}$ を求めよ。

これら 3 つの問題に対する解答パターンは、それぞれの問題に対する正答または誤答しかないとする、全部で $2^3 = 8$ 通りである。しかし Guttman は、これらの問題を例えばアメリカ社会学会の会員に尋ねた場合、解答パターンは概ね次の 4 つになると述べている。

- (1) 3 つの問題 a、b、c の全てに正答するパターン
- (2) 問題 a 及び b に対して正答し、問題 c に誤答するパターン
- (3) 問題 a のみに正答し、問題 b 及び c に誤答するパターン
- (4) いずれの問題にも誤答するパターン

Guttman の主張によれば、問題 a と問題 c には正答するが、問題 b には誤答する、というパターンは生じにくい。

上記 3 つの問題に対する解答パターンがこの 4 種類にほぼ限定されることの背景には、次の 2 つの仮定がある。一つは、これら 3 つの問題が、全て数学に関する知識を測っているという仮定である。つまり、この 3 つの問題は一次元性を満たしていることを仮定している。もう一つは、それぞれの問題に難易度を想定しているということである。問題 a は円の面積の求積に関する知識、問題 b は二次方程式の解に関する知識、問題 c は解析における微分に関する知識を問うものであり、日本の学校教育ではそれぞれ概ね小学校、高等学校 1 年、高等学校 3 年で学習する知識である。そのため、問題 c がもっとも難易度が高いと予想

される。このことから、問題 c に正答した者は、それよりも難易度が低い問題 a と問題 b に対しても正答するだろう、という仮定がある。ここでいう難易度は、いわゆるテスト理論では困難度（豊田, 2002）に該当する。

この数学の知識を問う 3 つの問題が異なる難易度をもつ、という想定は、問題の困難度という一次元の潜在変数を仮定していることと同じ意味を持つ。つまり、連続した値をもつ困難度において、問題 a、b、c はそれぞれ δ_a 、 δ_b 、 δ_c という値を示し、 $\delta_a \leq \delta_b \leq \delta_c$ という関係があることが仮定されているわけである。このとき、ある個人の数学的知識に関する能力が、この問題の困難度と同じ尺度において θ と表されるとき、この θ の値によって、この個人の正答や誤答が決まってくると考える。例えば、 $\delta_a \leq \delta_b < \theta < \delta_c$ である場合、この個人は問題 a、b には正答するが、問題 c には誤答すると考えられる。このときの正答・誤答が、現象として観測できる変数であり、顕在変数といわれる。そして、これらの問題に解答する人のサンプルにおいて、各問題に正答した人の割合は正答率または通過率（豊田, 2002）と呼ばれる。

このとき、得られた解答を正答または誤答によってコードする項目に関して、その正答率が低い項目に正答する者は、それよりも正答率の高い項目にも正答する、という規則が当てはまる場合、これらの項目群から構成される尺度を **Guttman Scale** と呼ぶ（van Schuur, 2011）。これは、正答や誤答以外に、該当の有無によって尋ねる項目についても成り立つ。つまり、該当者の割合が高い項目と低い項目の間で、割合が低い項目に該当する者は、それよりも割合の高い項目に必ず該当する、という規則が当てはまる場合、これも **Guttman Scale** といえる。

ここで、先ほどの数学の知識に関する問題に戻り、どの項目に正答し、どの項目に正答しないかに関する回答パターンを考える。正答及び誤答の状況は、あくまで仮想データではあるが、帯グラフで図 2-1 のように描くことができる。

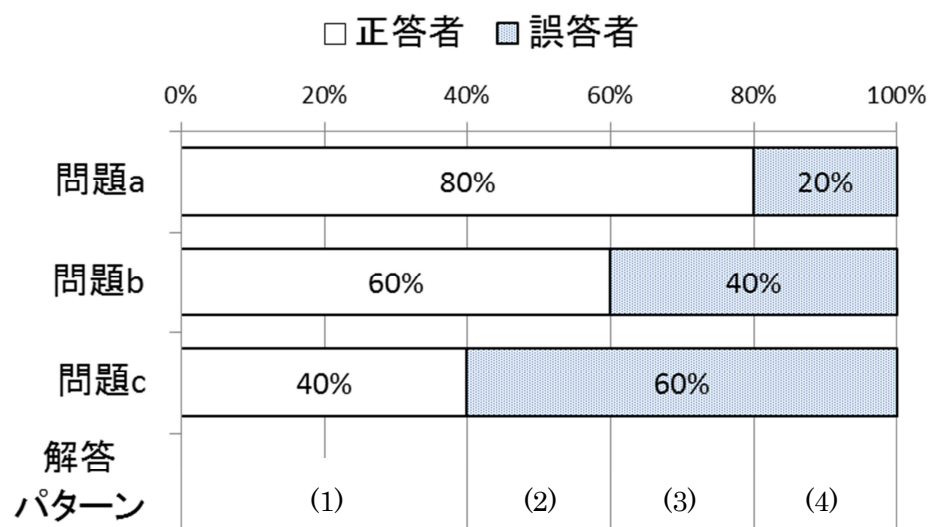


図 2-1 3つの数学的知識の問題に対する解答パターン（帯グラフは正答者及び誤答者の割合を示す）

図 2-1 の帯グラフのうち、白い部分は正答者の割合、ドット柄の部分は各問題に対する誤答者の割合を示す。また、帯グラフの下に示された括弧付きの数字は、先ほど言及した Guttman によるアメリカ社会学会の会員に尋ねた結果として期待される 4 種類の解答パターンに対応する。解答パターンを示すために正答を 1、誤答を 0 とコードし、各問題に正しく解答した場合に各 1 点を与えるとすると、(1) ～ (4) に分類された者の解答パターン {問題 a, 問題 b, 問題 c} と得点は (1) {1,1,1} で 3 点、(2) {1,1,0} で 2 点、(3) {1,0,0} で 1 点、(4) {0,0,0} で 0 点となる。これをみるとわかるように、解答パターンと得点は、それぞれ一対一で対応している。そのため、受験者の得点を知ることによって、どの問題に正答し、どの問題に誤答したのかという解答パターンを知ることが可能となる。また、正答率によって問題の難易度を分けることも可能となる。

これは、一般的な項目への回答状況で考えた場合、項目への該当数が回答パターンによって異なる状況として理解できる。このように、該当数と回答パターンに 1 対 1 の関係がみられ、回答者個人を項目への該当数によって順序づけるような回答パターンがみられることを Scalability と呼ぶ (Guttman, 1944; Menzel, 1953)。つまり、Guttman Scale では、回答パターンによって回答者を順序づけることができる。これは、項目への該当数から回答者の回答パターンを再現することができるともいえる。これを Guttman は Reproducibility と呼んだ (Guttman, 1944, 1950)。

こうした Guttman Scale の特徴を完全に満たすのは理想的な状況である。現実にはデータを収集した場合、このような状況が生じることはほとんどない。そこで、実際のデータを用いて Guttman Scale を構成する場合は、Guttman Scale の特徴を満たす度合いを考える必要が

ある。そこで、2つの問題の間で Guttman Scale を構成するかどうかを判断する方法について説明するため、2×2 のクロス表を考える。例えば、図 2-1 における問題 a と問題 b の回答者の割合は、次の表 2-1 のように示すことができる。

表 2-1 問題 a 及び問題 b に対する解答者の割合 (%)

		問題 a	
		正答	誤答
問題 b	正答	60	0
	誤答	20	20

表 2-1 における各セルの数字は、全体を 100 としたときの問題 a 及び b に対する解答者の割合（相対度数）を示す。問題 a と問題 b を比較した場合、問題 a の方が難易度が低い（正答率 80%）、より難易度の高い問題 b（正答率 60%）に正答した者は、問題 a に正答しやすく誤答しにくい、と考えるのは妥当であろう。このことは、表 2-1 におけるグレーのセルの割合が 0% であることに示されている。つまり、グレーのセルは難易度に関する 2 つの問題の順序性を反映している。このグレーのセルに該当する部分は Guttman Error と呼ばれる（van Schuur, 2011）。完全な Guttman Scale の性質をもつ変数の間では、この Guttman Error のセル度数は 0 となる。

このように、それぞれの項目の割合（顕在変数）が各項目の困難度（潜在変数）に応じて低くなるという関係が常に観察される状況が、Guttman Scale の構成度がもっとも高い状況である。特に、困難度の高い項目に該当する者が、より困難度の低い項目に常に該当している場合、これらの項目からなる尺度は完全な Guttman Scale を構成している、といえる。そして、項目の困難度による順序性に従わない該当パターンは Guttman Error となる。この Guttman Error を考慮することによって、Guttman Scale の構成度を評価することができる。

一般的に、二値データの項目間の関係を測定するために各種の連関係数がある。次節の表 2-3 に代表的なものを例示する。しかし、表 2-3 に示すこれらの連関係数は、上記の Guttman Scale の性質を評価するための指標として用いるには問題がある。次節では、その点について説明する。

2.2 Guttman Scale の評価指標に連関係数を利用できない理由

二値変数間の関係を MDS で視覚化する際に、二値変数間の関係の指標として連関係数が利用される。しかし、これらの連関係数は Guttman Scale の構成度を示す目的で作成されたわけではないため、Guttman Scale を示すための性質を持たない。本研究が着目する二値変数の割合の観点から、各種の連関係数の性質の違いについて説明する。

二値変数である変数 i と変数 j について表 2-2 のようなクロス集計表を考える。ここで

クロス表の各セルの度数を a 、 b 、 c 、 d とする。さらに各変数の該当の度数をサンプルサイズ n によって割った割合を、それぞれ $p_i = (a + b)/n$ 、 $p_j = (a + c)/n$ と表記する。さらに変数 i の方が変数 j よりも割合が高いとき、顕在的に観測される該当数については $b > c$ となることが想定される。このとき、2変数の割合の差は $p_i - p_j (> 0)$ となる。以下、 $b > c$ の場合を仮定する。

通常の二値変数における連関係数には、表 2-3 に示されるような指標がある。例えば四分点相関係数 (Yule, 1912) や Jaccard 係数 (Jaccard, 1908) 等では、係数が最大値 1 をとるとき、表 2-2 のクロス集計表で非対角のセル度数 b 、 c がともに 0 となることから、 $p_i - p_j = 0$ である。表 2-2 のセル度数 a 、 b 、 c 、 d を用いて各係数を算出する式を表 2-3 に併せて示す。

表 2-2 変数 i 及び変数 j のクロス集計表

		変数 j		
		該当	非該当	
変数 i	該当	a	b	$a + b$
	非該当	c	d	$c + d$
		$a + c$	$b + d$	n

表 2-3 二値変数に関する各種の連関係数

(a) 四分点相関係数（ファイ係数） (Yule, 1912)	$\frac{ad - bc}{\sqrt{(a + b)(a + c)(b + d)(c + d)}}$
(b) Jaccard 係数 (Jaccard, 1908)	$\frac{a}{a + b + c}$
(c) 単純見合い（単純マッチング）係数 (Sokal & Michener, 1958; Rand, 1971)	$\frac{a + d}{a + b + c + d}$
(d) Yule's Q (Goodman & Kruscal, 1954)	$\frac{ad - bc}{ad + bc}$
(e) Hamann (Hamann, 1961)	$\frac{(a + d) - (b + c)}{(a + d) + (b + c)}$
(f) Sorenson (Dice) (Gleason, 1920; Dice, 1945; Sorensen, 1948)	$\frac{2a}{2a + b + c}$
(g) Rogers & Tanimoto (Rogers & Tanimoto, 1960)	$\frac{a + d}{a + 2(b + c) + d}$
(h) Sokal & Sneath (Sokal & Sneath, 1963)	$\frac{2(a + d)}{2(a + d) + b + c}$
(i) Russel & Rao (Russel & Rao, 1940)	$\frac{a}{a + b + c + d}$
(j) Baroni-Urbani & Buser (Baroni-Urbani & Buser, 1976)	$\frac{a + \sqrt{ad}}{a + b + c + \sqrt{ad}}$
(k) Ochiai (Driver & Kroeber, 1932; Ochiai, 1957)	$\frac{a}{\sqrt{(a + b)(a + c)}}$

各連関係数の計算式から明らかなように、Russel & Rao の係数を除き、非対角のセル度数 b と c がともに 0 であるときに各々の連関係数は最大値 1 をとる。このとき、2 変数の割合の差は、 $p_i - p_j = 0$ である。そこで、サンプルサイズ n における変数 i の該当の度数を $x (=np_i)$ 、変数 j の該当の度数を $y (=np_j)$ として、クロス集計表をあらためて表 2-4 のように設定する。

表 2-4 変数 i 及び変数 j のクロス集計表

		変数 j		
		該当	非該当	
変数 i	該当	a	$x - a$	x
	非該当	$y - a$	$n - x - y + a$	$n - x$
		y	$n - y$	n

ここで、表 2-4 のクロス集計表におけるセル度数について、

$$\begin{aligned} 1 &\leq x \leq n, \\ 1 &\leq y \leq x, \\ x + y - n &\leq a \leq y \text{ かつ } a \geq 0 \end{aligned}$$

として $n = 200$ に設定し、 x 、 y 、 a を動かす。それぞれのセル度数に対して表 2-3 に示す各連関係数の値を縦軸に、2 変数の割合の差 $p_i - p_j$ を横軸とした散布図を図 2-2 に示す。ただし、 $p_i - p_j \geq 0$ の範囲に限定する。

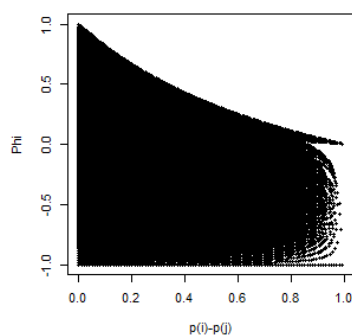
これらをみるとわかるように、2 変数の割合の差が小さいほど連関係数の最大値は 1 に近づく一方、2 変数の割合の差が大きくなるほど連関係数の最大値は小さくなることがわかる。ただし、Yule's Q については、2 変数の割合の差が大きくなっても、連関係数の最大値は 1 の値を取り得ることが示されている。このように、2 変数間の連関としてクロス集計表における対角セルの度数を利用した四分点相関係数をはじめとする連関係数を利用する場合、変数の割合の差が大きくなるほど連関係数は小さくなるという一定の関係がみられる。

2 変数の割合に大きな差が存在する場合でも、これらの間に完全な Guttman Scale の関係は存在し得る。つまり、2 変数の割合に大きな差がある場合でも、Guttman Scale の構成度はその最大値をとらなければならない。そのため、Guttman Scale の構成度を表す指標としては、表 2-3 に示した連関係数のうち、Yule's Q を除いた連関係数は適切ではないことがわかる。なお、Yule's Q についても Guttman Scale の構成度の指標として適切ではない。その理由については、次節において言及する。

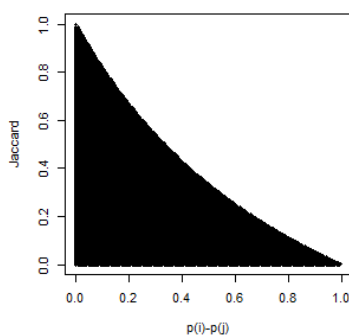
これは、これらの連関係数では 2 変数のクロス集計表における 2 つの非対角のセル度数

を差別化せずに連関を算出しているためである。割合が異なる 2 変数間に Guttman Scale の関係が存在することを指標として表現するためには、2 つの非対角のセル度数の間の差別化もしくは順序関係の考慮が必要となるわけである。次節では、この Guttman Scale の構成度を表す指標について説明する。

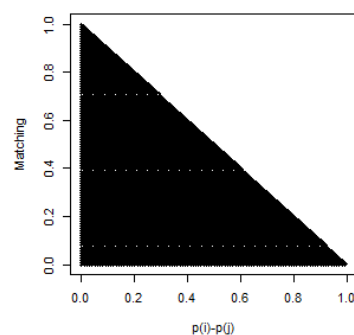
(a) 四分点相関係数



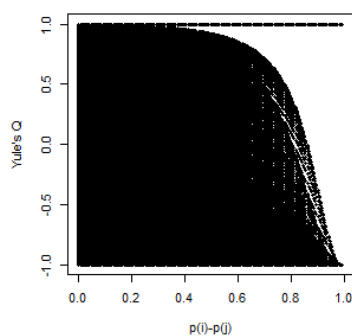
(b) Jaccard 係数



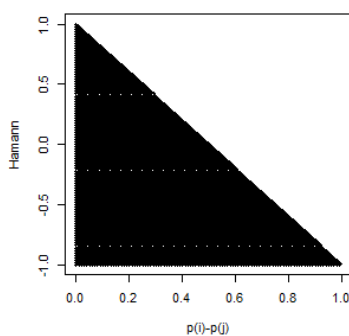
(c) 単純見合い係数



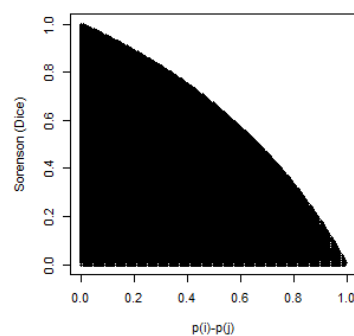
(d) Yule's Q



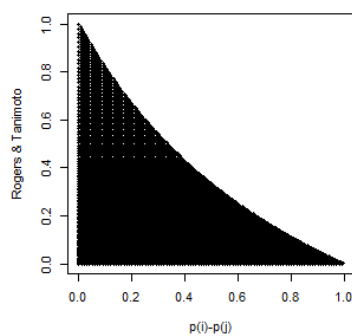
(e) Hamann



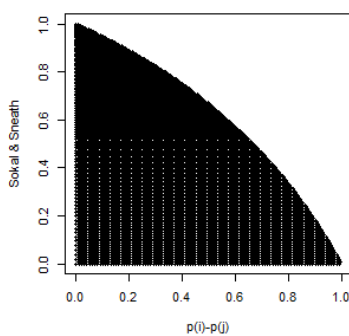
(f) Sorenson (Dice)



(g) Rogers & Tanimoto



(h) Sokal & Sneath



(i) Russel & Rao

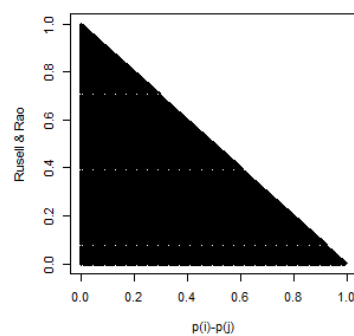


図 2-2 各種連関係数と二値変数の割合の差の関係を表す散布図 (a) ~ (i)

(j) Baroni-Urbani & Buser (k) Ochiai

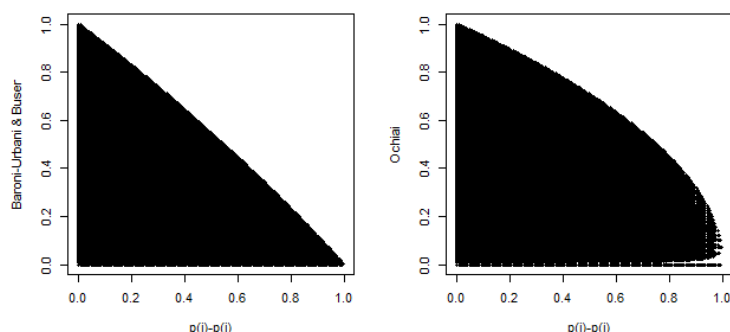


図 2-2（続き） 各種連関係数と二値変数の割合の差の関係を表す散布図 (j) ～ (k)

2.3 Guttman Scale の構成度の指標

複数の項目がある場合に、これらが Guttman Scale を構成しているかどうかを表現する指標として、Guttman Scale の構成度がある。第 2 章 1 節で提示した 2 変数のケースの場合、表 2-1 に示したクロス集計表におけるグレーの部分で示されたセル度数によって Guttman Error を評価することが可能であり、この Guttman Error を用いた各種の指標が提案されている。しかし、Guttman Scale を構成する項目を評価する際、項目が 2 つ以上あることが通常であるため、3 つ以上の変数との関連で Guttman Error を評価する必要がある。そこで、3 つ以上の変数について同時に Guttman Error を考慮し、Guttman Scale を構成する度合いを評価する測度を説明する。当然ながら、3 つ以上の変数における Guttman Scale の構成度の指標の算出方法は、2 変数の場合にも適用できる。

Guttman Scale の構成度は、Scalability 基準 (van Schuur, 2011) とも呼ばれる。従来の先行研究で提案されている Guttman Scale の構成度の指標には、大きく 3 つのタイプがあるとされ、Mokken (1971) はその 3 つを Guttman 自身が提唱した Reproducibility 係数 (Guttman, 1950; Clogg & Sawyer, 1981)、及びその問題点を修正したいくつかの係数をタイプ S1 やタイプ S2 として整理している。そこで、それぞれのタイプに該当する Reproducibility 係数、Scalability 係数、Homogeneity 係数について具体的な計算方法と併せて、以下に説明する。

2.3.1 Reproducibility 係数

この係数は、Guttman 自身によって提唱された基準であり (Guttman, 1950)、次式によって定義される (Menzel, 1953; White & Saltz, 1957; van Schuur, 2011)。

$$[2.3.1] \quad \text{C. of R.} = 1 - \frac{E_{\text{obs}}}{nM}.$$

ただし、 E_{obs} とは Guttman Error の観測度数、 n はサンプルサイズ、 M は項目数を表すものとする。

ここで、実際の計算方法を考えるため、次のような例を考える。5つの項目について6名の調査協力者に回答を求め、各項目への該当を1、非該当を0とコードする二値データによって測定する状況を想定する。このときのデータが表2-5のような形で得られたとする。表頭に項目、表側に協力者のIDを記載し、便宜的に1という回答の多寡によって、項目と協力者の順序の並べ替えを行っている。

これをみると、項目5については6名全員が該当しており、項目1に対する該当者は1名のみ(ID1)である。該当者数が多くなる順に変数が並び、最下行に各項目に対する該当者数が示されている。Guttman Scale では、この該当者の数を項目の困難度（テスト理論における通過率）を示す指標とみなすため、項目5の通過率をもっとも高く、困難度が低いと判断する。一方、項目1の通過率をもっとも低く、困難度が高いと判断する。また、5つの項目に対する該当数に応じて回答者を並べ替えると、ID番号1の者の該当数をもっとも多く、ID番号6の者の該当数をもっとも少ない。そのため、項目と回答者との関係は、該当 (= 1) の部分を実線で囲むと、図のような階段型の区分線として描かれる。その際、この階段型の区分に違反する部分、図中ではグレーで網掛けされた2箇所が Guttman Error を示すことになる。

表 2-5 6名の調査協力者に5項目を実施した測定結果（二値データ）

	項目 1	項目 2	項目 3	項目 4	項目 5	
ID1	1	1	1	1	1	5
ID2	0	1	1	0	1	3
ID3	0	0	1	1	1	3
ID4	0	1	0	1	1	3
ID5	0	0	0	1	1	2
ID6	0	0	0	0	1	1
	1	3	3	4	6	

ここで Guttman Error の数え方には大きく2つの方法がある (Mokken, 1971)。一つは、Goodenough (1944) が提唱した方法であることから、Mokken (1971) にならって Guttman-Goodenough (G-G) 定義と呼ぶことにする。これは、各回答者の該当項目数（表2-5 でいえば表の右側に示された5項目中の該当項目数）を数え、完全な Guttman Scale のパターンからどれだけ逸脱しているかをエラー数とするものである。表2-5の例で該当項目数が3であるID3とID4の回答パターンに着目すると、ID3が示す {0,0,1,1,1} が理想的な回答パターンになる一方、ID4の回答パターンは {0,1,0,1,1} であるため、項目2、または項

目 3 の回答が逸脱を生じていることになることから、逸脱部分を構成する項目 2 のセルまたは項目 3 のセルの数をエラー数と数えることから、エラー数は 1 となる。

もう一つの定義は Suchman (1950) によるもので、Mokken (1971) にならって Guttman-Suchman (G-S) 定義と呼ぶことにする。これは、回答を変更する数をもっとも少なくすることによって理想的な Guttman Scale のパターンに近づける場合の変更数をエラー数とするものである。表 2-5 における ID4 {0,1,0,1,1} の場合、{0,0,0,1,1} または {0,1,1,1,1} と変更することで、理想的な Guttman Scale のパターンになる。この G-S 定義によるエラー数の数え方にもいくつかの方法が提案されているが、そのエラーの数え方のアルゴリズムを提唱した Green (1956) の方法を説明する。Green (1956) は、各回答者の回答パターンについて、通過率の順に並べた項目の順番に沿って回答パターンを検討し、逸脱を生じた箇所をエラーとして数えていく。その際に、隣り合う逸脱パターンの数を数える。例えば、表 2-5 の ID4 {0,1,0,1,1} の場合、通過率の順から逸脱した回答パターン {1,0} を隣り合う回答ごとに数えることから、ここではエラー数は 1 となる。これを一次エラーペアと呼ぶ。次にそのエラーペア全てを除外し、残った回答パターンについて同様に隣り合う回答ごとに逸脱した回答パターン {1,0} の数を数える。これを二次エラーペアと呼ぶ。以下、同様にこのペアを除外した回答配列からさらに高次の逸脱回答ペアの個数を数えていき、最終的に逸脱回答ペアがなくなるまで継続する。ただし、3 次以上のエラーは生じにくいことから、サンプルサイズが大きい場合、2 次のエラーまでを数えることで代替できている (Green, 1956)。このアルゴリズムの利用方法は研究者によってまちまちである。例えば Sagi (1959) は、このうち一次エラーペアの数のみを採用している。また、Loevinger (1947, 1948) は全ての項目間の組み合わせについて逸脱回答ペアの個数を数える方法を採用している。

この Green (1956) の方法を利用すると、表 2-5 の場合、G-G 定義でも G-S 定義でもエラーの数は変わらない。さらに、エラーペアの次数を考えた場合、高次のエラーペアは存在しない。Guttman Scale から逸脱した回答パターン {1,0} が存在する解答者は ID2 {0,1,1,0,1} 及び ID4 {0,1,0,1,1} であり、それぞれエラー数は 1 となり、集計表全体ではエラー数は 2 となる。このデータについて、Reproducibility 係数を計算すると、

$$[2.3.2] \quad C. of R. = 1 - \frac{2}{6 \times 5} = 0.93$$

となる。ちなみに、Guttman (1944) は、この値が 0.90 以上であれば Guttman Scale の構成として望ましいとしている。

2.3.2 Scalability 係数

前節で説明した Reproducibility 係数は Scalability を判断する指標として多用されているが、それ自身がもつ問題点も指摘される。Reproducibility 係数は、その定義式から 0~1 の範囲の値をとると期待されるが、実際に適用した場合の下限値が高くなるという問題がある

(Mokken, 1971)。そこで、高い下限値の問題を修正した係数を Mokken (1971) はタイプ S1 として整理している。

$$[2.3.3] \quad S_1 = 1 - \frac{E}{E_{\max}} = \frac{\text{Rep} - \text{Rep}_{\min}}{1 - \text{Rep}_{\min}}.$$

ここで、 E は Guttman Error として定義されるエラー数、 E_{\max} はエラーの理論上の最大値、 Rep は Reproducibility 係数の値、 Rep_{\min} は Reproducibility 係数の理論上の最小値である。この式の最後の項の関係を図 2-3 に示す。

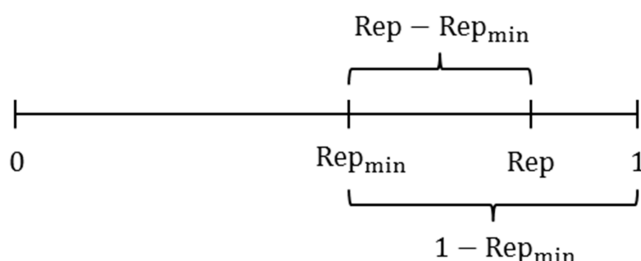


図 2-3 S_1 の算出における Rep 、 Rep_{\min} の関係

このように、 S_1 は Reproducibility 係数の値を修正する。このタイプ S_1 の係数にはいくつかあるが、ここでは Menzel (1953) が定義した Scalability 係数について説明する。

ある項目における該当または非該当のカテゴリに回答者のほとんどが当てはまり（回答者全員が該当であったり、非該当であったりするケース）、あるいは、ある特定の回答者個人の回答がほぼ全ての項目について同じカテゴリに該当したりするような状況（ある回答者の回答パターンが全ての項目について該当であったり、非該当であったりするケース）を Extremeness という。Reproducibility 係数は、Scalability がみられるかどうかによらず、こうした Extremeness がみられる場合に値が高くなる、という問題点がある (Menzel, 1953)。

Reproducibility 係数は最大値 1 をとるが、二値データの場合、最小値は 0.5 を下回ることはない。なぜならば、ある項目において割合の高いカテゴリが Scalability を判断する際の基準カテゴリとなるため、エラーとなりうる回答は、高々もう一方の割合の低いカテゴリ（以下、低割合カテゴリ）に該当する回答にみられることになるからである (Stouffer, 1950)。

このことを G-S 定義によるエラー数で説明する。項目への該当ありを 1、該当なしを 0 としてコードしたデータにおいて、各回答者の該当項目数と各項目への該当者数によって、それぞれ行と列を並べ替えた状態を考える。すなわち、先の 6 名に 5 項目を尋ねた結果では、データの並び順が表 2-5 のような状態にあるとする。このとき、表 2-5 の状態になれば、G-S 定義では $\{1,0\}$ の個数をカウントすることになる。例えば、6 項目のデータの場合、

最大のエラー数をとる回答者の回答パターンは例えば $\{1,1,1,0,0,0\}$ であり、エラー数は 3 である。また、5 項目の場合、最大のエラー数をとる回答者の回答パターンは例えば $\{1,1,1,0,0\}$ であり、エラー数は 2 である。つまり、 $\{1,0\}$ のパターンがもっとも多い回答パターンが最大のエラー数を含むことになる。そして、その数は、各パターンにおいて割合の低い方のカテゴリの数に一致する。6 項目の例ではカテゴリ 1 またはカテゴリ 0 の数 (= 3)、5 項目の例ではカテゴリ 0 の数 (= 2) に一致する。

ここで項目数を M 、回答者数を n とすると、エラー数 ($\{1,0\}$ の個数) は

$$[2.3.4] \quad \begin{cases} \frac{M}{2} \text{ 個} & \text{if } M \text{ が偶数} \\ \frac{M-1}{2} \text{ 個} & \text{if } M \text{ が奇数} \end{cases}$$

となり、データ全体における理論上最大のエラー数を含むデータは、回答者全員がこのエラー数を示す解答パターンを示す場合であり、

$$[2.3.5] \quad \begin{cases} \frac{Mn}{2} \text{ 個} & \text{if } M \text{ が偶数} \\ \frac{(M-1)n}{2} \text{ 個} & \text{if } M \text{ が奇数} \end{cases}$$

となる。このときの Reproducibility 係数は上の値を式[2.3.1]に代入して、

$$[2.3.6] \quad \text{C. of R.} = \begin{cases} 0.5 & \text{if } M \text{ が偶数} \\ 0.5 + \frac{1}{2M} & \text{if } M \text{ が奇数} \end{cases}$$

となり、最小値は 0.5 となる。

以上の説明は、表 2-5 のデータ形式において、列方向における各項目の該当者の割合について説明した場合であるが、行方向における各回答者の該当項目の割合についても同様の議論が当てはまる。各項目や各回答者の回答カテゴリ（該当または非該当）のうち、割合の高いカテゴリを基準にしてエラーが判断されるため、エラーの数の最大値は、割合の低いカテゴリの反応数を超えることはない（Menzel, 1953）。そこで、列方向または行方向において低い割合を示すカテゴリのセルの数をエラーの取り得る理論上の最大数（Maximum Error, ME）として、Reproducibility 係数を次式のように修正したものが Scalability 係数である。このようにエラーの割合を算出する際にエラーが取り得る理論上の最大数で除すことによって取り得る値の幅が広がることから、Scalability の程度の違いを比較する上でより客観的な指標となることが期待できる。

$$[2.3.7] \quad \text{C. of S.} = 1 - \frac{E_{\text{obs.}}}{\text{ME}}$$

このとき、ME の算出方法は 2 通りが考えられる。表 2-5 のデータにおいて、行方向（回答者ごと）にみた場合の ME と列方向（項目ごと）にみた場合の ME である。そのうち、値の小さい方が Scalability 係数の理論上の下限値を下げることから、小さい値の ME を採用する。

実際に Menzel（1953）の算出方法に従って、表 2-5 のデータに式 [2.3.7] を適用する。項目 i におけるカテゴリ 0 と 1 のそれぞれの度数の大きい方を f_i^{col} 、回答者 k が各項目のカテゴリ 0 と 1 に回答した数の大きい方を f_j^{row} とする。 f_i^{col} については、項目 1 についてはカテゴリ 0 の方が多いことから、その数は $6 - 1 = 5$ であり、項目 2 についてはカテゴリ 0 と 1 の回答数が同一であるため、どちらか一方の反応数 3 を利用する、というように、項目 5 までそれぞれ計算して足し合わせると、

$$\sum_{i=1}^5 f_i^{\text{col}} = (6 - 1) + 3 + 3 + 4 + 6 = 21$$

となる。同様に、回答者ごとに行方向で f_j^{row} についてみると、ID1 の数の多いカテゴリは 1 となるため、その数は 5 であり、ID2 の数の多いカテゴリは 1 であることから、その数は 3 である、というように、ID6 までそれぞれ計算して足し合わせると、

$$\sum_{j=1}^6 f_j^{\text{row}} = 5 + 3 + 3 + 3 + (5 - 2) + (5 - 1) = 21$$

となる。ME は少ないカテゴリの回答数の総和になることから、全回答数（回答者数×項目数、ここでは $6 \times 5 = 30$ ）から f_i^{col} の総和、または f_j^{row} の総和を差し引いて ME を計算すると、

$$\text{ME}(\text{項目}) = 6 \times 5 - \sum_{i=1}^5 f_i^{\text{col}} = 9,$$

$$\text{ME}(\text{回答者}) = 6 \times 5 - \sum_{j=1}^6 f_j^{\text{row}} = 9$$

となる。この定義は、カテゴリ数が 3 以上の場合を想定した定義となっているが、二値データの場合は、単純に割合の高いカテゴリ数の総和を計算することと同じである。

このうち、値の小さい方を ME とするが、この場合、項目方向、回答者方向ともに $\text{ME} = 9$ であることから、

$$\text{C. of S.} = 1 - \frac{2}{9} = 0.78$$

となる。

Scalability 係数については、この係数が発表された当時、数値を評価する経験的な基準値が十分に蓄積しておらず (Menzel, 1953)、それ以降も多く利用されているわけではないため、係数としての明確な判断基準がないというデメリットがある。様々な実際のデータに適用することで、経験的な基準値の策定が求められる。Menzel (1953) によれば Guttman Scale としての判断基準として、C. of S. = .60～.65 以上の値が適当ではないかと示唆されている。

Scalability 係数は Reproducibility 係数と同じく最大値 1 をとり、常に Scalability 係数 ≤ Reproducibility 係数という関係がある。Reproducibility 係数の最小値は 0.5 を下回らない (Menzel, 1953; White & Saltz, 1957) 一方、Scalability 係数は 0～1 の値を取るという特性がある。なお、Scalability 係数は Jackson (1949) による Plus Percentage Ratio と本質的には同じである。計算方法の説明は White & Saltz (1957) に示されるとおりである。

2.3.3 Loevinger による Homogeneity 係数

前節のタイプ S1 の係数は、Reproducibility 係数における高い下限値を修正する際に、割合の高いカテゴリがエラーを数える場合の基準になる点を修正したものである。しかし、Reproducibility 係数における高い下限値はもう一つの別の原因でも生じうる。それは、各項目に対する回答がお互いに独立である場合に期待される係数を考慮しないために生じている下限値の高さである。そこで、各項目に対する回答がお互いに独立であるという帰無仮説のもとで計算される係数を下限値とみなして修正するものが、タイプ S2 の係数である (Mokken, 1971)。

$$[2.3.8] \quad S_2 = \frac{S_1 - S_{10}}{1 - S_{10}} = 1 - \frac{E}{E_0}.$$

ここで、 S_1 はあるタイプ S1 の係数の値、 S_{10} は各項目に対する回答が互いに独立であるという帰無仮説の下で計算される同じタイプ S1 の係数の値、 E は Guttman Error として定義されるエラー数、 E_0 は各項目に対する回答が互いに独立であるという帰無仮説の下で予測されるエラーの理論値である。この式に示されるタイプ S2 とタイプ S1 の各係数の値の関係を図 2-4 に示す。

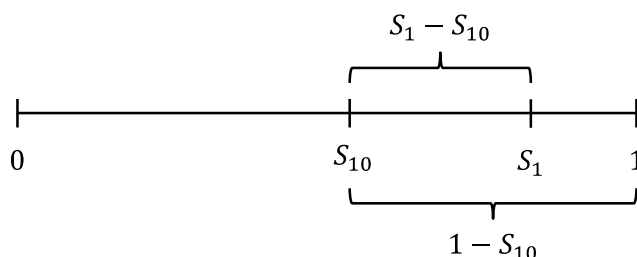


図 2-4 S2 の算出における S_1 、 S_{10} の関係

このように、タイプ S2 の係数の値はタイプ S1 の値を修正する。定義式 [2.3.8] にもみられるように、タイプ S2 の係数は、タイプ S1 によって修正されている Reproducibility 係数の下限値の問題に対処しつつ、さらに回答の独立性を考慮した修正をも行っている点で、タイプ S1 よりも望ましいといえる (Mokken, 1971)。また、この定義では、 S_1 の値が S_{10} の値を下回ることにより、 S_2 の値が負となることもあり得る。その場合には、これらの項目から構成される尺度の仮説が棄却されるという意味で、尺度として許容できないと判断する (Mokken, 1971)。

このタイプの係数にもいくつかあるが、ここでは Loevinger (1947, 1948) による Homogeneity 係数 (Coefficient of Homogeneity) について説明する。Loevinger (1947, 1948) による手法はこれまでに説明した係数とは異なり、項目 i と j について、両項目に該当すると回答した者の割合 p_{ij} を計算し、両項目が互いに独立であると仮定した場合の周辺度数から算出される理論値 $p_i p_j$ を差し引いた $p_{ij} - p_i p_j$ を、全項目の全組み合わせについて足し合わせる (White & Saltz, 1957)。

$$[2.3.9] \quad T = \sum_i \sum_{j \neq i} (p_{ij} - p_i p_j).$$

このとき、全ての項目がペアワイズに独立である場合、理論的に $T = 0$ となる。 T の上限は、式 [2.3.9] における p_{ij} を項目 i と j のうち困難度が高い (通過率が低い) 項目の割合 $\min(p_i, p_j)$ で置きかえた、次式で定義される T_{\max} となる。

$$[2.3.10] \quad T_{\max} = \sum_i \sum_{j \neq i} (\min(p_i, p_j) - p_i p_j).$$

このとき、Homogeneity 係数 (H) は

$$[2.3.11] \quad H = \frac{T}{T_{\max}}$$

となる。表 2-5 のデータについて Homogeneity 係数を計算すると、

$$\begin{aligned} T &= \left(\frac{1}{6} - \frac{1}{12}\right) + \left(\frac{1}{6} - \frac{1}{12}\right) + \left(\frac{1}{6} - \frac{1}{9}\right) + \left(\frac{1}{6} - \frac{1}{6}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \left(\frac{1}{3} - \frac{1}{3}\right) + \left(\frac{1}{2} - \frac{1}{2}\right) + \left(\frac{1}{3} - \frac{1}{3}\right) \\ &\quad + \left(\frac{1}{2} - \frac{1}{2}\right) + \left(\frac{2}{3} - \frac{2}{3}\right) = \frac{11}{36}, \end{aligned}$$

$$T_{\max} = \left(\frac{1}{6} - \frac{1}{12}\right) + \left(\frac{1}{6} - \frac{1}{12}\right) + \left(\frac{1}{6} - \frac{1}{9}\right) + \left(\frac{1}{6} - \frac{1}{6}\right) + \left(\frac{1}{2} - \frac{1}{4}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{2} - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) \\ + \left(\frac{1}{2} - \frac{1}{2}\right) + \left(\frac{2}{3} - \frac{2}{3}\right) = \frac{29}{36}$$

以上から、

$$H = \frac{11}{29} = 0.38$$

となる。 H は項目の組み合わせごとに算出することが可能であり、

$$[2.3.12] \quad H_{ij} = \frac{p_{i|j} - p_i}{1 - p_i} \quad (p_i \geq p_j)$$

として計算される (Loevinger, 1947, 1948)。ただし、 $p_{i|j} = p_{ij}/p_j$ は項目 j への該当の下で項目 i に該当する場合の条件付き確率を示す。なお、式[2.3.12]は変数 i 及び変数 j という 2 つの変数における Homogeneity 係数であることから、式[2.3.9]、[2.3.10]、[2.3.11]を用いると、

$$[2.3.13] \quad H_{ij} = \frac{T}{T_{\max}} = \frac{p_{ij} - p_i p_j}{\min(p_i p_j) - p_i p_j} = \frac{p_{ij}/\min(p_i p_j) - \max(p_i p_j)}{1 - \max(p_i p_j)}$$

$$= \begin{cases} \frac{p_{ij}/p_j - p_i}{1 - p_i} = \frac{p_{i|j} - p_i}{1 - p_i} & \text{if } p_i \geq p_j \\ \frac{p_{ij}/p_i - p_j}{1 - p_j} = \frac{p_{j|i} - p_j}{1 - p_j} & \text{otherwise} \end{cases}$$

として導かれる。

Reproducibility 係数や Scalability 係数に比較すると Homogeneity 係数の値は低いが、経験的な基準値として 0.3 未満の値を示す場合、それら 2 つ以上の項目から構成される尺度は Guttman Scale として意味をなさない (van Schuur, 2011) とされる。なお、Homogeneity 係数の最大値は 1 である。2 項目間を考えた場合、Homogeneity 係数の最小値は $1 - n$ (n : サンプルサイズ) であり負の値を取り得る (付録 A)。

タイプ S2 の係数はタイプ S1 の係数の修正を含めている点でより望ましいと先に述べたが、Reproducibility 係数や Scalability 係数と比較して、この Loevinger の Homogeneity 係数は回答集計表の周辺度数に依存しない点で優れているとされる (Mokken, 1971; Warrens, 2008)。より正確に言うならば、Homogeneity 係数の値は 2 つの二値変数の周辺度数に依存するが、

それらの周辺度数が所与の下で、2変数が互いに独立である場合は0の値を示し、また2変数が関連する場合に最大値1を示す（Warrens, 2008）。最小値そのものは周辺度数に依存する（付録A）。つまり、2変数が独立である場合に期待されるセル度数を差し引いて調整している点で、Homogeneity 係数は Reproducibility 係数や Scalability 係数よりも優れているといえる。そこで本論文の第4章以降では、Guttman Scale の構成度の指標として、この Loevinger の Homogeneity 係数を利用することとする。

ここで、表2-4に示したクロス集計表において、第2章2節において設定した条件のもとで、 $n = 200$ として i, j, a を動かして生じるクロス集計表ごとに、2変数間の割合の差 $p_i - p_j$ を横軸に、その際の Reproducibility 係数、Scalability 係数、Homogeneity 係数の各値を縦軸にした散布図を図2-5に示す。ただし、Homogeneity 係数の縦軸については、下限値を0で切って示している。

各係数の取り得る下限値をみると、実際に、Reproducibility 係数 = 0.75、Scalability 係数 = 0.50、Homogeneity 係数 < 0 となっており、各係数の理論上の下限値が高くなる問題点が、それぞれ修正されていることが示されている。

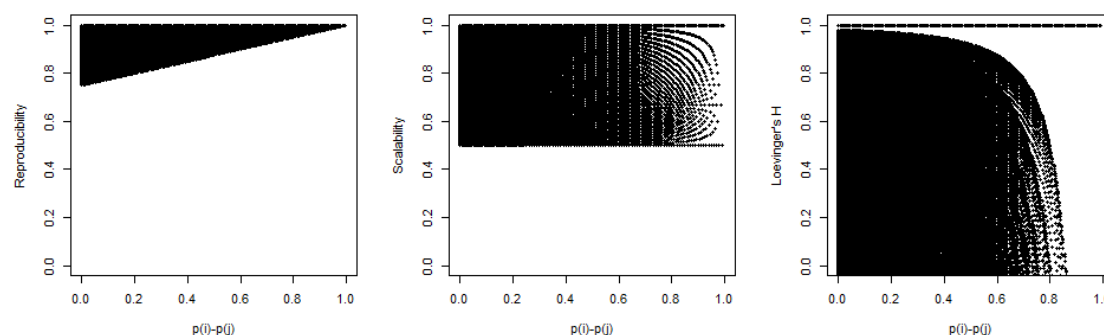


図 2-5 尺度の Scalability を示す各係数と二値変数の割合の差の関係を表す散布図

最後に、前節で取り上げた Yule's Q が、尺度の Scalability として利用するのは適切ではない点について言及する。尺度の Scalability を評価する3つの指標に対して、Yule's Q がどのような値を示すかをみるため、縦軸に Yule's Q の値、横軸に3つの Scalability の指標を配した散布図を図2-6に示す。

これをみるとわかるように Scalability 係数、Loevinger の Homogeneity 係数の値が下がってくる場合でも、Yule's Q の値が1に近い値を示すケースがかなり多く見られている。これ

は、Guttman Error が発生しているにもかかわらず、Yule's Q はそれを適切に評価していない
 ということを示している。Guttman Error が発生していないときに Yule's Q は理論上の上限
 値 1 を示す点では適切であるが、Guttman Error が発生していても高い値を示すことから、
 Guttman Scale の構成度を表す指標としては適切ではない。これは、Yule's Q が本来は尺度の
 Scalability を測定することを意図していない点で、当然であるといえる。

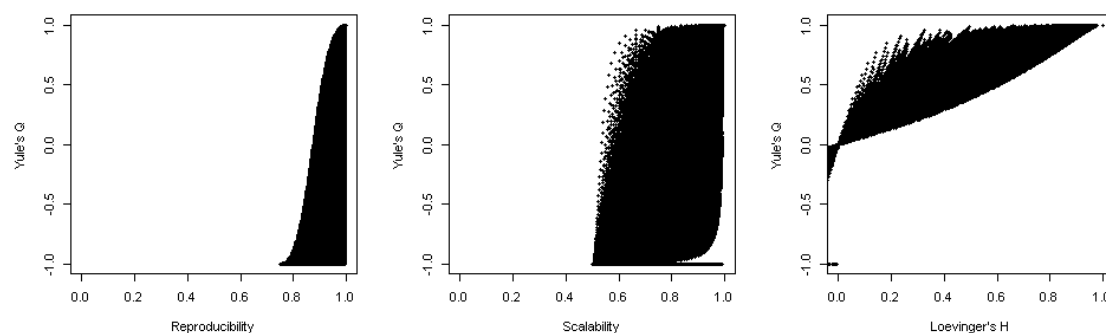


図 2-6 Reproducibility 係数、Scalability 係数、Homogeneity 係数に対する Yule's Q の値を示す散布図

第3章 多次元尺度構成法—Guttman Scale に関連して

本章では、データの視覚化に関わる統計的手法として用いられることの多い多次元尺度法 (MDS; 高根, 1980; Cox & Cox, 1994) について、その数理モデルを説明する。また、MDS の2次元の布置の結果を解釈する枠組みとして利用できる Facet 理論について説明する。さらに、実際に MDS に対してデータを適用した結果を提示し、各種の MDS の視覚化手法の特徴について言及する。

まず MDS を含む各種の視覚化手法の数理モデルを説明する。一つは、二値データの2次元平面への布置による視覚化として多く利用される等質性分析 (Gifi, 1990; Michailidis & de Leeuw, 1998) をとりあげる。この等質性分析は、他の視覚化手法である対応分析 (Greenacre, 2007)、双対尺度法 (西里, 2007, 2010)、数量化Ⅲ類 (林, 1974; 岩坪, 1987) と同種の分析手法である。その後、二値データの類似度 (非類似度) に対して適用されることの多い非計量的な多次元尺度法である Kruskal の MDS (Kruskal, 1964) 及び Guttman の最小空間分析 (Smallest Space Analysis: SSA; Guttman, 1968) を取り上げる。

3.1 等質性分析の数理モデル

等質性分析では、サンプルサイズを n とし、 m 番目の変数のカテゴリ数を l_m とするとき、調査対象者 (ここでは、以下オブジェクトとする) と変数のカテゴリそれぞれに数量を割り当てる数量化を行う。 P 次元の数量化について、オブジェクトスコア行列を次式のように表す。

$$[3.1.1] \quad \mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1P} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nP} \end{pmatrix}.$$

また、 m 番目のカテゴリ数量化行列を次式のように表す。

$$[3.1.2] \quad \mathbf{Y}_m = \begin{pmatrix} y_{11} & \cdots & y_{1P} \\ \vdots & \ddots & \vdots \\ y_{l_m 1} & \cdots & y_{l_m P} \end{pmatrix}, \quad m = 1, \dots, M.$$

ただし、 M は変数の数を表す。さらに、各オブジェクトのカテゴリへの該当の有無を 0 または 1 で示すインディケーター行列を次式のように表す。

$$[3.1.3] \quad \mathbf{G}_m = \begin{pmatrix} \delta_{11} & \cdots & \delta_{1l_m} \\ \vdots & \ddots & \vdots \\ \delta_{n1} & \cdots & \delta_{nl_m} \end{pmatrix}, \quad m = 1, \dots, M.$$

ただし、 $\delta_{ij} = 0$ or 1 である。このとき、オブジェクト及びカテゴリの数量化を行う上で最小化すべき目的関数は、次式で定義される σ で表される。行列要素の平方和を SSQ とすると、行列 $\mathbf{A} = (a_{ij})$ について $\text{SSQ}(\mathbf{A}) = \text{tr}\mathbf{A}'\mathbf{A}$ である。このとき

$$\begin{aligned} [3.1.4] \quad \sigma(\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_M) &= M^{-1} \sum_{m=1}^M \text{SSQ}(\mathbf{X} - \mathbf{G}_m \mathbf{Y}_m) \\ &= M^{-1} \sum_{m=1}^M \text{tr}(\mathbf{X} - \mathbf{G}_m \mathbf{Y}_m)'(\mathbf{X} - \mathbf{G}_m \mathbf{Y}_m) \end{aligned}$$

となる。

ここで $\mathbf{X} = \mathbf{0}$ 、 $\mathbf{Y}_m = \mathbf{0}$ のような無意味解を除くため、次のような制約をかける。まず、オブジェクトスコア行列の各次元を表す列ベクトルを直交基底としノルムを n とする。

$$[3.1.5] \quad \mathbf{X}'\mathbf{X} = n\mathbf{I}_p.$$

ただし、 \mathbf{I}_p は p 次の単位行列である。さらに、オブジェクトスコア行列の各次元を表す列ベクトルを中心化する。各要素を 1 とする n 次元のベクトルを \mathbf{u}_n として、

$$[3.1.6] \quad \mathbf{u}_n' \mathbf{X} = \mathbf{0}.$$

これによって、オブジェクトを空間上に布置する際、各点は原点を中心にプロットされる。

これらの条件式[3.1.5]と[3.1.6]の下で[3.1.4]を最小化すると、 \mathbf{X} を固定して \mathbf{Y}_m に関して σ を最小化する $\hat{\mathbf{Y}}_m$ は次式の通りとなる。

$$[3.1.7] \quad \hat{\mathbf{Y}}_m = (\mathbf{G}_m' \mathbf{G}_m)^{-1} \mathbf{G}_m' \mathbf{X} \quad (m = 1, \dots, M).$$

次に、 \mathbf{Y}_m 固定の下で σ を最小化する $\hat{\mathbf{X}}$ は次式の通りとなる。

$$[3.1.8] \quad \hat{\mathbf{X}} = M^{-1} \sum_{m=1}^M \mathbf{G}_m \mathbf{Y}_m.$$

つまり、変数の各カテゴリ数量化は、そのカテゴリに属するオブジェクトのオブジェクトスコアの重心になり、オブジェクトスコアはそれ自身が属するカテゴリ数量化の平均となる。ここで、回答に欠損値がある場合、[3.1.4]式に欠損があるカテゴリを除いて平方和を計算したものを目的関数とする。

$$[3.1.9] \quad \sigma(\mathbf{X}; \mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_M) = M^{-1} \sum_{m=1}^M \text{tr}(\mathbf{X} - \mathbf{G}_m \mathbf{Y}_m)' \mathbf{Z} (\mathbf{X} - \mathbf{G}_m \mathbf{Y}_m).$$

ただし、行列 \mathbf{Z} は対角行列で、 i 番目のオブジェクトに欠損がない場合に (i, i) 対角要素

に 1、 i 番目のオブジェクトに欠損がある場合に (i, i) 対角要素に 0 を割り当てる。

3.2 Kruskal の多次元尺度法 (MDS) の数理モデル

次に、MDS の数理モデルについて説明する。ここでは二値データについて、類似度に基づいた変数間の近接性を計算し、その近接性データに MDS を適用する場合を想定する。ただし、類似度とは変数同士の関係の類似性を連関係数として測定したものを指す。MDS では、この類似度を間隔尺度とみなすか、順序尺度とみなすかによって用いる分析が異なる。本研究の第 4 章以降では、類似度の指標として Jaccard 係数を使用する（その理由は第 4 章 4 節を参照されたい）。なお、二値データに MDS を適用する際の近接性の指標については、通常、二値変数の割合を考慮した連関を考えることはないため、Guttman Scale の Scalability の指標を使う必要はない。行動科学における二値データの連関係数として Jaccard 係数を用いる場合、これを順序尺度とみなして分析することが多い（例えば、Canter, 2000 など）。そこで、変数間の連関係数を順序尺度とみなし、係数の順序関係に基づいて変数をユークリッド空間に表示する Kruskal の MDS について説明する。

ユークリッド空間 R^P において、2 点 i, j の p 次元 ($p = 1, \dots, P$) における座標を x_{ip} 、 x_{jp} と表す時、2 点間の距離 d_{ij} は次式によって定義できる。

$$[3.2.1] \quad d_{ij} = [\sum_{p=1}^P (x_{ip} - x_{jp})^2]^{\frac{1}{2}}.$$

2 次元平面における距離を表す場合、 $P = 2$ となる。ここで、2 点 i, j 間の類似度を s_{ij} とすると、ある変換 f を用いて点 i, j について次式のように残差を定義する。

$$[3.2.2] \quad e_{ij}^2 = [d_{ij} - f(s_{ij})]^2.$$

これは、類似度を変換したものと MDS 空間内の距離の誤差の二乗である。これを全ての点の組み合わせ (i, j) について足し合わせたものが適合の悪さを示す測度となり、一般に Raw Stress (Kruskal, 1964) と呼ばれる。

$$[3.2.3] \quad \sigma_r = \sum_i \sum_{i < j} [d_{ij} - f(s_{ij})]^2.$$

Raw Stress が大きくなるほど適合が悪くなるが、この測度の値の大小は単位に依存するため、その値の大小から単純に適合の程度を評価することができない。そのため、次のような基準化を行う。

$$[3.2.4] \quad \sigma_1^2 = \frac{\sum_i \sum_{i < j} [d_{ij} - f(s_{ij})]^2}{\sum_i \sum_{i < j} d_{ij}^2}.$$

この平方根は Stress-1 と呼ばれ (Kruskal, 1964)、類似度を変換した値が MDS 空間内の距離と完全に一致する場合、最小値 0 をとる。このような形で定義された目的関数を最小化することで点の空間内の位置を決定する。

対象 i と j の関係を示す指標として非類似度 δ_{ij} の順序を考えると、対象が全て M 個を想定するとき、 (i, j) の組み合わせは $MC = M(M - 1)/2$ 通りになる。このとき、

$$[3.2.5] \quad \delta_{i_1 j_1} < \delta_{i_2 j_2} < \cdots < \delta_{i_{MC} j_{MC}}$$

とする。一方、 M 個の対象を P 次元空間の M 個の点で表し、それらを x_i ($1 \leq i \leq M$) とする。このとき、点 i と点 j の距離 d_{ij} は式 [3.2.1] で表される。Kruskal の方法では、この d_{ij} の大小関係が、それと対応する δ_{ij} の大小関係と一致するように空間内の配置を決定する。

d_{ij} を横軸に、 δ_{ij} を縦軸に布置する散布図を考えた場合、両者の大小関係が一致するには、各点を d_{ij} の順に結んだ時、縦軸の値に関して直線が単調増加を示す必要がある。そこで、横軸 d 方向に値をブロックにまとめて、まとめられた $\{d_{ij}\}$ の平均値を算出することによる調整をおこなうことで、単調増加を示すように修正した \hat{d}_{ij} を用い、それに基づいて空間内に布置する方法が Kruskal の方法となる。

$$[3.2.6] \quad S^* = \sum_i \sum_{i < j} (d_{ij} - \hat{d}_{ij})^2$$

$$[3.2.7] \quad T^* = \sum_i \sum_{i < j} d_{ij}^2$$

とおくとき、

$$[3.2.8] \quad \frac{S^*}{T^*} = \frac{\sum_i \sum_{i < j} (d_{ij} - \hat{d}_{ij})^2}{\sum_i \sum_{i < j} d_{ij}^2}$$

を最小化する空間内の布置が求めるべき MDS 空間である。このとき、式 [3.2.8] は距離 d_{ij} の単位に依存しない。これについて平方根をとった次式の S

$$[3.2.9] \quad S = \sqrt{\frac{S^*}{T^*}}$$

によってデータと MDS 空間内の距離の適合度が定義され、 S の値が小さいほど適合がよいとされる（林・飽戸, 1976; Kruskal, 1964）。

3.3 Guttman の SSA

3.3.1 SSA の数理モデル

本手法も先の Kruskal の方法と同じくデータの値の順序に基づく非計量的多次元尺度法であり、その性質として単調性及び最小次元性を有する。単調性とは MDS 空間内の距離の順序関係がデータの値の順序関係を反映することであり、最小次元性とは最小の次元でデータを空間内に布置することである。

M 個の対象についてお互いの非類似度 δ_{ij} が与えられているとする。このとき、この非類似度の値を順序になおした距離順位数（Distance Ranking Numbers） D_{ij} を考え、次のように定義する。

$$[3.3.1] \quad D_{ij} \geq 0, \quad D_{ii} = 0.$$

また、一般には $D_{ij} \neq D_{ji}$ である。 D_{ij} の異なる値に基づく個数を R 個とする。 $\{D_{ij}\}$ が対称であるとして、 M 個の対象の全ての組み合わせについて $\{D_{ij}\}$ が与えられていれば、 R の最大値は組み合わせの数 $\binom{M}{2} = M(M-1)/2$ となる。

一方、この距離順位数 D_{ij} に対応する SSA 空間内の点間の距離を d_{ij} とし、 d_{ij} の異なる値に基づく個数を Q 個とする。距離 d_{ij} は 2 点間の距離を示すという定義から $\{d_{ij}\}$ は対称であるため、 Q の最大値も $\binom{M}{2}$ となる。そして、 R 、 Q とともに順序関係についてタイが存在する場合、 R 及び Q は $\binom{M}{2}$ 以下の値となる。

次節で説明する順位写像による変換によって、 D_{ij} が D'_{ij} となったときの D'_{ij} の異なる値に基づく個数を R' 個とし、 d_{ij} が d'_{ij} となったときの d'_{ij} の異なる値に基づく個数を Q' 個とする。このとき順位写像によって得られた単調性は、タイの多寡に応じて

$$[3.3.2] \quad \begin{array}{lll} R' \leq Q' & \cdots & \text{準単調性} \\ R' = Q' & \cdots & \text{強単調性} \\ R' \geq Q' & \cdots & \text{弱単調性} \end{array}$$

と分類される。この様子を図示したものが図 3-1 になる。

図 3-1 (a) は強単調性を示したグラフであり、 $\{D'_{ij}\}$ 及び $\{d'_{ij}\}$ にはタイが含まれておらず、 $R' = Q'$ となる。図 3-1 (b) は準単調性を示したグラフであり、 $\{D'_{ij}\}$ にタイ

($D'_{ij} = D'_{i'j'} = 2$) が存在し、それに対して d'_{ij} では異なる値が対応しているため、 $R' \leq Q'$ となっている。図 3-1 (c) は弱単調性を示したグラフであり、 $\{d'_{ij}\}$ に 2 つのタイ ($d'_{ij} = d'_{i'j'} = 3, d'_{ij} = d'_{i'j'} = 5$) が存在し、各タイにおいて D'_{ij} として異なる値が対応しているため、 $R' \geq Q'$ となっている。

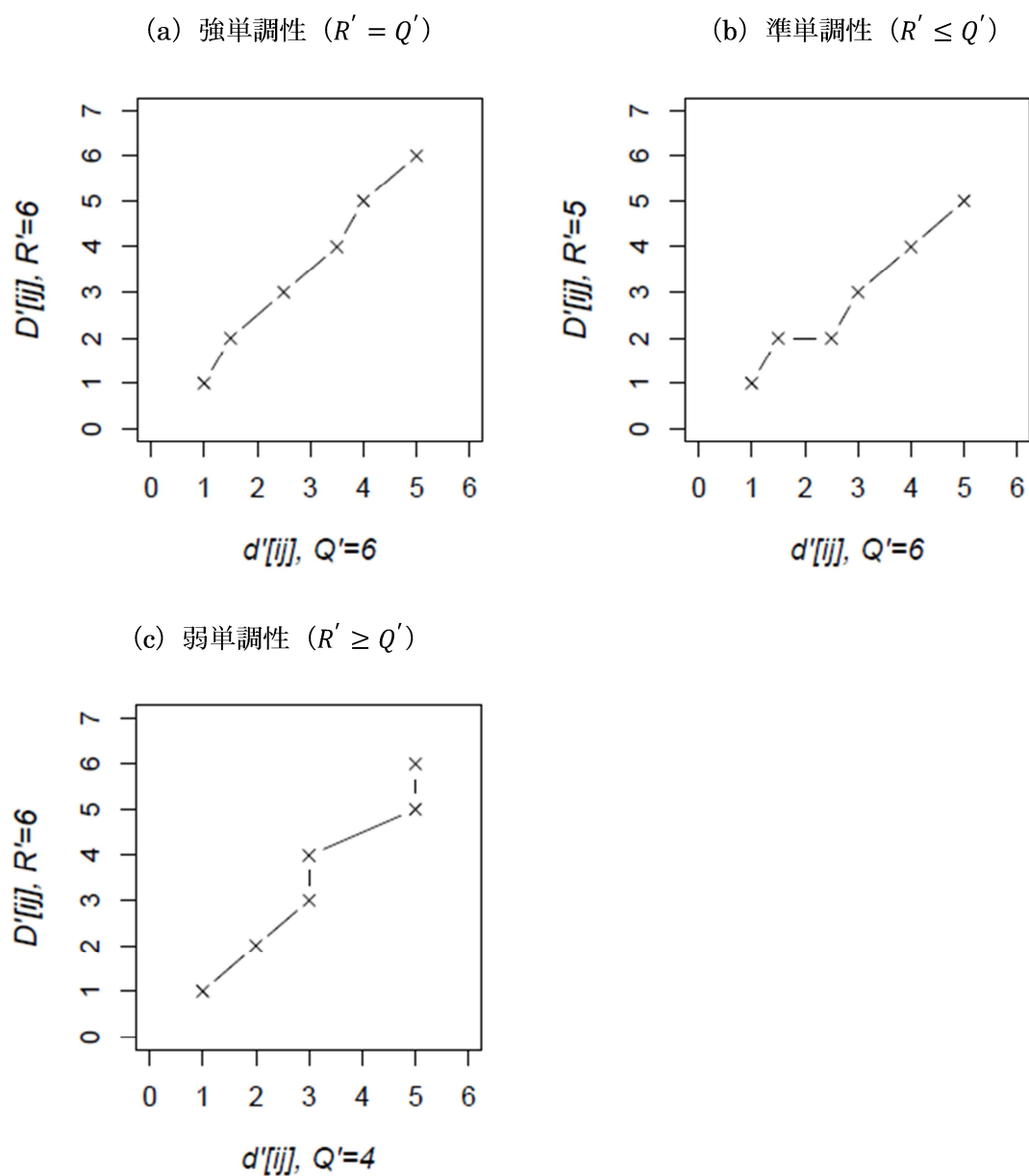


図 3-1 単調性のタイプ別に図示したもの

このように図 3-1 (a) ～ (c) にみられるように $\{D'_{ij}\}$ と $\{d'_{ij}\}$ それぞれにおける順序関係がタイを含めて一致している場合には、 $\{D'_{ij}\}$ と $\{d'_{ij}\}$ の関係は単調関数になる。この両者の関係の単調性の程度を示す指標として Guttman (1968) は単調性係数 μ を次式によって定義した。

$$[3.3.3] \quad \mu = \frac{u}{\sqrt{vw}}.$$

ただし、

$$u = \sum_i \sum_j e_{ij} \hat{D}_{ij} d_{ij}, \quad v = \sum_i \sum_j e_{ij} \hat{D}_{ij}^2, \quad w = \sum_i \sum_j e_{ij} d_{ij}^2,$$

$$e_{ij} = \begin{cases} 1 & \text{if } D_{ij}(i \neq j) \text{ の値が存在する} \\ 0 & \text{if } i = j \text{ または } D_{ij}(i \neq j) \text{ の値が存在しない} \end{cases}$$

とする。また、 \hat{D}_{ij} は D_{ij} を写像によって変換した像である。 D_{ij} の順序情報を利用して d_{ij} を変換したものを \hat{D}_{ij} とし、これと d_{ij} との相関を最大にするところが、SSA の最大の特徴といえる (林・鮑戸, 1976)。 \hat{D}_{ij} はタイの扱いによって異なる写像となる。先に示した準単調性、強単調性、弱単調性のそれぞれに対する \hat{D}_{ij} の求め方は、次節において説明する。

式[3.3.3]で定義される μ を用いて、

$$[3.3.4] \quad \text{CoA} = \sqrt{1 - \mu}$$

を定義し、これを逸脱係数 (Coefficient of Alienation; CoA) と呼ぶ。この CoA の値が小さいほど、データの SSA 空間内の布置に対する適合度がよいことを示す。 μ は $\{D_{ij}\}$ を所与として $\{d_{ij}\}$ の関数となっていると考えられるから、

$$[3.3.5] \quad \mu = \mu(\{d_{ij}\} | \{D_{ij}\})$$

とすると、 p を次元として

$$[3.3.6] \quad \frac{\partial \mu(\{d_{ij}\} | \{D_{ij}\})}{\partial x_{ip}} = 0, \quad i = 1, \dots, M; \quad p = 1, \dots, P$$

を解くことによって、SSA 空間内の布置を得ることができる。ただし、 d_{ij} と x_{ip} との関係は式[3.2.1]に示されるとおりである。

3.3.2 距離順位数の写像

Guttman (1968) は順位写像という写像を説明している。これは、ある順序関係を持つ 2 つの数列に対して、一方の順序情報に基づいて他方の並べ替えを行うという変換である。それぞれの数列に含まれるタイの扱いに応じて、準単調性、強単調性、弱単調性の 3 種類が提案されている。以下に各種の具体的な方法を説明する。なお、この写像は特段 D_{ij} や d_{ij} に固有のものではないことに注意されたい。

(1) 準単調性における写像

d_{ij} の順位写像による像を d_{ij}^* とし、 $\{d_{ij}\}$ から $\{d_{ij}^*\}$ への写像 f^* を考える。この写像は $\{D_{ij}\}$ のタイをなくした順位に基づいて $\{d_{ij}\}$ の値の順序を並べ替えたものである。したがって関数として象徴的に表せばこの $\{d_{ij}\}$ から $\{d_{ij}^*\}$ に変換する順序関数を $\{f_{ijkl}^*\}$ とする。 d_{kl} を f_{ijkl}^* で変換した値を d_{ij}^* とすると、

$$[3.3.7] \quad d_{ij}^* = f^*(\{d_{ij}\}, \{D_{ij}\})$$

となるが、より具体的には

$$[3.3.8] \quad d_{ij}^* = \sum_k \sum_l d_{kl} f_{ijkl}^*$$

となる。以下で、 f_{ijkl}^* の定義について考える。

$\{D_{ij}\}$ の要素のうち $e_{ij} = 1$ である要素の数を Z とすると、 $Z = \sum_i \sum_j e_{ij}$ である。ここで対応する $\{d_{ij}\}$ の値のセットがあり、それらの要素が小さいものから大きいものへと順に並べられているとする。このうちいくつかはタイであるとし、これらに対応した $\{D_{ij}\}$ にはタイがないとする（あるいは、タイがあっても、その同順位間で適当に順位を決めてやる）。このとき d_{ij} の順位はそれに対応する D_{ij} の順位に従って決めるものとする。

$$[3.3.9] \quad e_{ijz}^* = \begin{cases} 1 & \text{if } d_{ij} \text{が } \{d_{ij}\} \text{ で } z \text{ 番目にくるとき} \\ 0 & \text{otherwise} \end{cases}$$

とすると、 $z = 1, \dots, Z$ として、

$$[3.3.10] \quad \sum_z e_{ijz}^* = e_{ij}, \quad \sum_i \sum_j e_{ijz}^* = 1$$

となる。次に $\{d_{ij}\}$ における z 番目の値 d_z^* は、

$$[3.3.11] \quad d_z^* = \sum_i \sum_j d_{ij} e_{ijz}^*$$

として得られる。このとき、

$$[3.3.12] \quad d_1^* \leq d_2^* \leq \cdots \leq d_z^*$$

である。

同様に、 $\{D_{ij}\}$ についても次のように定義する。

$$[3.3.13] \quad e_{ijz} = \begin{cases} 1 & \text{if } D_{ij} \text{が}\{D_{ij}\}\text{で}z\text{番目にくるとき} \\ 0 & \text{otherwise.} \end{cases}$$

このとき

$$[3.3.14] \quad \sum_z e_{ijz} = e_{ij}, \quad \sum_i \sum_j e_{ijz} = 1$$

が成り立つ。以上のように設定された d_{ij} と D_{ij} との関連を d_{ij} の写像である d_{ij}^* を用いて次式のように定義する。

$$[3.3.15] \quad d_{ij}^* = (1 - e_{ij})d_{ij} + \sum_z d_z^* e_{ijz}.$$

[3.3.8]と[3.3.15]を結び付けて考えるため、次のような δ_{ijkl} を定義する。

$$[3.3.16] \quad \delta_{ijkl} = \begin{cases} 1 & \text{if } i = k \text{ かつ } j = l \text{ であるとき} \\ 0 & \text{otherwise.} \end{cases}$$

これにより、[3.3.14]を用いて

$$[3.3.17] \quad f_{ijkl}^* = (1 - e_{ij})\delta_{ijkl} + \sum_z e_{klz}^* e_{ijz}$$

と定義された写像変換が、準単調性における写像となる。以上のように定義された d_{ij}^* を用いて

$$[3.3.18] \quad \hat{D}_{ij} = d_{ij}^*$$

とすることで、準単調性における写像が得られる。なお、 $d_{ij}^* \equiv d_{ij}$ あるいは $\{d_{ij}\}$ の順位づけと $\{D_{ij}\}$ の順位づけがタイを除いて一致するとき、単調性係数 $\mu = 1$ となる。

(2) 強単調性における写像

$\{D_{ij}\}$ の r 番目に同じ値の要素が f_r 個あるとする。ここで

$$[3.3.19] \quad \bar{e}_{ijr} = \begin{cases} 1 & \text{if } D_{ij} \text{が} \{D_{ij}\} \text{で} r \text{番目にくるとき} \\ 0 & \text{otherwise} \end{cases}$$

とおき、 $r = 1, \dots, R$ とする。このとき、

$$[3.3.20] \quad \sum_r \bar{e}_{ijr} = e_{ij}, \quad \sum_i \sum_j \bar{e}_{ijr} = f_r, \quad \sum_r f_r = Z$$

が成り立つ。 $\{d_{ij}\}$ についても値の小さいものから大きいものへと並べると、 R 個の区間に区分でき、 r 番目の区間には f_r 個の値が含まれるようにする。ここで、

$$[3.3.21] \quad \bar{e}_{ijr}^* = \begin{cases} 1 & \text{if } d_{ij} \text{が} \{d_{ij}\} \text{で} r \text{番目にくるとき} \\ 0 & \text{otherwise} \end{cases}$$

とすると、

$$[3.3.22] \quad \sum_r \bar{e}_{ijr}^* = e_{ij}, \quad \sum_i \sum_j \bar{e}_{ijr}^* = f_r$$

となる。 $\{d_{ij}\}$ の r 番目の区間に含まれる $\{d_{ij}\}$ の平均を \bar{d}_r とすると、

$$[3.3.23] \quad \bar{d}_r = f_r^{-1} \sum_i \sum_j d_{ij} \bar{e}_{ijr}^*$$

と表される。以上の表記を利用して、強単調性を示す d_{ij} の写像である \bar{d}_{ij} は次式のように定義される。

$$[3.3.24] \quad \bar{d}_{ij} = (1 - e_{ij})d_{ij} + \sum_r \bar{d}_r \bar{e}_{ijr}.$$

また、この写像に対応する順序関数 \bar{f} は、

$$[3.3.25] \quad \bar{f}_{ijkl} = (1 - e_{ij})\delta_{ijkl} + \sum_r \bar{e}_{klr}^* \bar{e}_{ijr}$$

となり、

$$[3.3.26] \quad \bar{d}_{ij} = \sum_k \sum_l d_{kl} \bar{f}_{ijkl}$$

が写像となる。よって $\hat{D}_{ij} = \bar{d}_{ij}$ とおくことで、強単調性の場合の単調性係数 μ が計算できる。

(3) 弱単調性における写像

$\{d_{ij}\}$ の要素のうち $e_{ij} = 1$ を満たすものが複数のブロックにまとめられ、一つのブロック内の値が全て同じであるようにしたい場合を考える。ブロックの個数を B 個、 b 番目のブロック内の値を d'_b 、 d_{ij} を変換した値を d'_{ij} 、 b 番目のブロック内に含まれる $\{d_{ij}\}$ の要素の数を f'_b とする。ここで

$$[3.3.27] \quad e'_{ijb} = \begin{cases} 1 & \text{if } d_{ij} \text{が } b \text{番目のブロックにあるとき} \\ 0 & \text{otherwise} \end{cases}$$

とし、 $b = 1, \dots, B$ とすると

$$[3.3.28] \quad \sum_b e'_{ijb} = e_{ij}, \quad \sum_i \sum_j e'_{ijb} = f'_b, \quad \frac{1}{f'_b} \sum_i \sum_j d_{ij} e'_{ijb} = d'_b$$

である。そこで

$$[3.3.29] \quad d'_{ij} = (1 - e_{ij})d_{ij} + \sum_b d'_b e'_{ijb}$$

とし、弱単調性の写像に対応する順序関数 f' を次式のように定義する。

$$[3.3.30] \quad f'_{ijkl} = (1 - e_{ij})\delta_{ijkl} + \sum_b f_b'^{-1} e'_{klb} e'_{ijb}.$$

すると d'_{ij} は次式のように書ける。

$$[3.3.31] \quad d'_{ij} = \sum_k \sum_l d_{kl} f'_{ijkl}.$$

ただし、

$$[3.3.32] \quad D_{ij} \leq D_{kl} \quad \text{ならば} \quad d'_{ij} \leq d'_{kl}$$

とする。

3.4 Facet 理論と Radex モデル

第 2 章に記述した Guttman Scale をどのような形で視覚化することが可能であろうか。Guttman Scale はこれまでの説明で見てきたように、該当か非該当の二値で回答を求める項目の該当割合に応じて項目の順序づけを行う。そのため、各項目間の関係には、この項目の該当割合の順序を反映した視覚化を行う手法が望ましい。

項目の該当割合を反映させるデータの視覚化手法として、前節までに説明した等質性分析と多次元尺度法が考えられる。ただし、これらの分析では順序尺度の強度が一つの次元として抽出されることから、いわゆる馬蹄形 (Horseshoe Effect, Arch Effect, Guttman Effect) を呈するという問題がある。これは、Simplex 構造 (三角形や四面体等) を 2 次元などの低次元上に反映する場合に現れるとされる (Greenacre, 2007)。

とはいえ、等質性分析では 2 つの変数カテゴリに該当する回答者が重なるほど、それらの変数カテゴリは近接して布置されることから、近接して布置される変数カテゴリの該当割合も類似するという側面がある。また、多次元尺度法では、2 次元もしくは 3 次元の空間に変数間の類似性に対応した距離を反映させる形で変数が布置される。そのため、割合が相対的に類似した変数同士が近接して布置され、割合が相対的に離れている変数同士が離れて布置されれば、項目の割合を反映した布置が可能となる。このとき、異なる Guttman Scale が異なる領域として分離されれば、視覚的に了解可能な布置が得られる。この領域性に着目した枠組みとして Facet 理論 (Guttman, 1982; Canter, 1985; Shye et al., 1994; Borg & Shye, 1995; 木村ら, 2002;) が有用である。

Facet 理論では、平面もしくは空間を構成する次元 (Dimension) ではなく、平面や空間の連続的な領域 (Region) に着目する (Borg & Groenen, 1998)。そのため、異なる変数 (もしくは変数のカテゴリ) が平面や空間の異なる領域に規則的に割り当てられて布置される、というとらえ方で結果を解釈する。この法則を利用することで、割合の異なる変数を異なる領域に規則的に布置することができ、それによって Guttman Scale の視覚化が可能となる。

Facet 理論とは 1954 年に Guttman によって提唱されたものである。これは、測定の対象となる現象の定義、測定そしてデータ分析に関する一体的な方法論である (Canter, 1985)。観測対象となる現象には科学的法則に基づく構造が観察され、その現象を構成する要素全体に対して、その各要素を空間的に位置づけ、要素の区分の仕方に相当する意味合いに対応して、空間を構成する枠組みが描かれるとした (Guttman & Greenbaum, 1998)。具体的には、2 次元または 3 次元の空間内に変数を位置づける際に、変数 (またはそのカテゴリ) の属性に応じた領域によって空間が区分されるというものである。そのため、変数間の距離が近い場合でも、属する領域が異なることがあり得る。Facet 理論は、第 2 章に説明した尺度分析から発展したものであるため、Guttman Scale の視覚化の手法に馴染むと考えられる。

本来、Facet 理論は仮説検証型のモデルである。そのため、測定対象の属性とその構成要素をあらかじめ定義することによって属性を明示し、その属性に沿ってデータを測定する (Guttman & Greenbaum, 1998)。そして、各属性を構成する要素が空間内の異なる領域とし

て現れるとする。領域の図式的なモデルを図 3-2 に示す。

図 3-2 (a) は属性の構成要素 a、b、c が属する領域を直線によって区分するパターンであり、このような領域のパターンを **Axial Facet** と呼ぶ。図 3-2 (b) は、同心円によって領域を区分するパターンであり、これを **Modular Facet** と呼ぶ。図 3-2 (c) は、中央の点から周辺方向に伸びる線によって領域を区分するパターンを示しており、これを **Polar Facet** と呼ぶ。これらは基本的な領域区分のパターンとして提示されている。そして、空間内に布置される変数の区分の仕方（属性）が複数ある場合、その属性に対応する領域区分の様式が空間内に **Facet** として現れる。例えば、犯罪行動を測定する変数を視覚化手法によって布置した場合に、犯罪行動の割合という属性を示す一連の変数は、その割合の程度に応じて異なる領域に区分できる。また、犯行の動機という属性で変数を分類すると、犯行動機の種類に応じて各変数は異なる領域に区分できる。

さらに、これらの領域区分の複数のパターンが同時に観測される。例えば、**Modular Facet** と **Polar Facet** が同時に現れる場合、その複合的なパターンは図 3-2 (d) に示される **Radex** モデルと呼ばれる (Canter, 1985; Coxon, 1982; Guttman, 1954)。また、その **Radex** モデルが存在する x-y 平面に対して、垂直に伸びる z 軸方向に **Axial Facet** が観測された場合、図 3-2 (e) に示される **Cylindrex** と呼ばれる円柱様の構造として空間が把握される例も示されている。

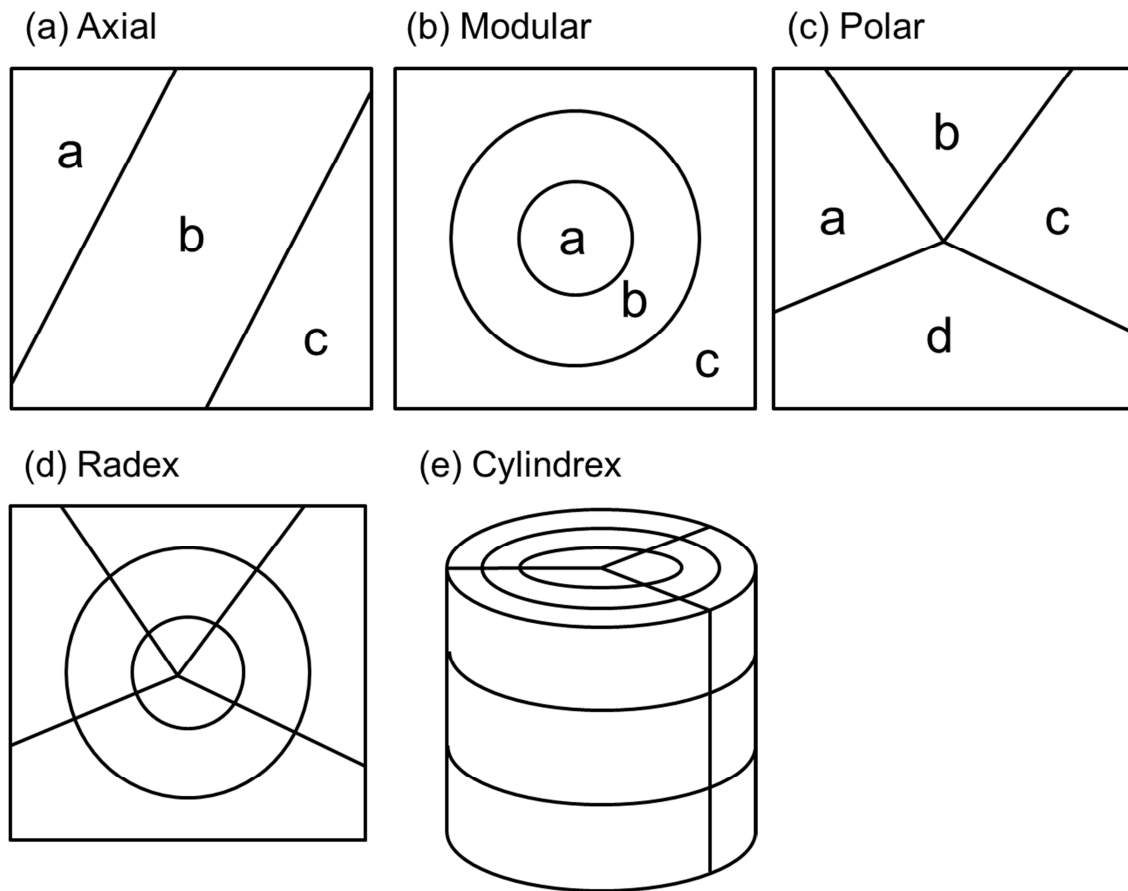


図 3-2 Facet 理論における領域区分の模式図

Facet 理論におけるこうした領域区分の考え方は、MDS で布置された変数の関係を吟味する際に、一つの参照枠を与えるものとして有用であり、この枠組みが変数間の関係を理解する上でより多くの洞察を与える場合がある。ここで、行動科学的データにおける Facet 理論の利用例として、犯罪行動に適用した事例を挙げる。

Canter (2000) は、殺人や強姦といった暴力的な犯罪において、犯罪者がとる行動に関する諸特徴を変数とし、各犯行ケースについて変数への該当の有無を評定したデータに対して MDS を適用した場合、2 次元の布置に Radex モデルがみられると主張した。この Radex モデルは、先述のとおり、Modular Facet と Polar Facet が組み合わさった形で領域が区分されるパターンである。また、二値変数として測定されたデータの場合、割合の高い変数ほど図の中心部分に布置されやすい (Borg & Groenen, 1998; Borg & Shye, 1995)。そのため、犯罪行動として出現頻度の高い変数が図の中央に布置される。一方、出現頻度が低い変数が図の周縁部に布置される。周縁部に布置された変数同士は、類似性も相対的に低く、互いに離れて布置される。この様子を Canter (2000) は模式図として示した (図 3-3)。

図の中央には割合の高い変数が観測され、これらは各犯行ケースに共通してみられる特徴であるといえる。そのため、犯行特徴を分類に利用する場合、その重要度は相対的に低くなる。逆に、割合の低い変数群が図の中央から周縁部に布置される。この周縁部の変数群が犯行ケースの異なる特徴を示すことになる。そのため、これら周縁部の変数の組み合わせを見ることによって、異なった犯行特徴の組み合わせで表現される犯行テーマが得られる。例えば、6つの犯行テーマが出現する場合、図3-3のように図の中心から周縁部に向かう6つのテーマが観測できる。このとき、関連する犯行テーマ同士は隣接し、それらが円環状に配置される。これが Polar Facet である。このような配列は Circumplex (Guttman, 1954) と呼ばれる。

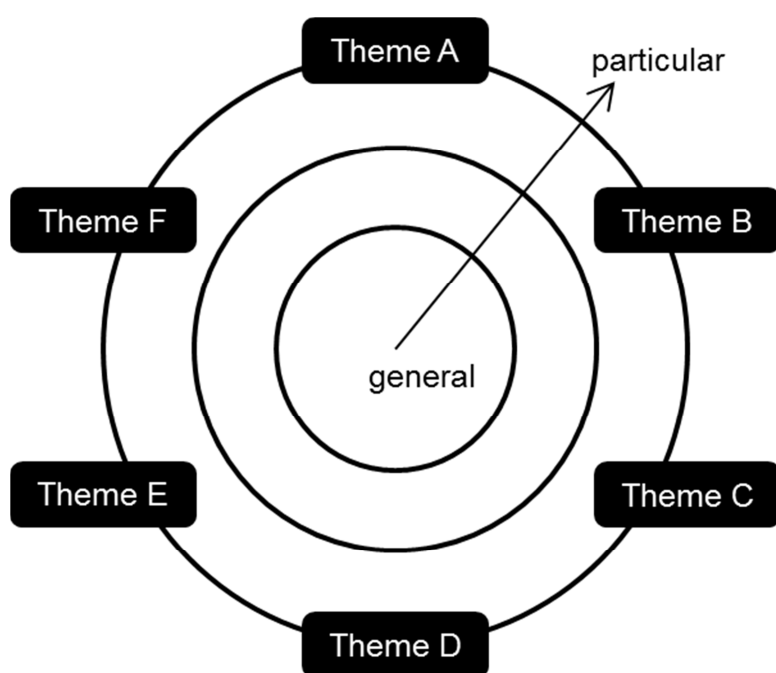


図 3-3 Canter (2000) の示した犯罪行動の Radex モデル

一方、図の中心部から周辺部に向かう同心円状の Modular Facet は、例えば犯行の特殊化 (Specialisation, Yokota & Canter, 2004) として解釈できる。これは、熟達度の高い変数ほど出現頻度が下がっていくことから、犯行を繰り返すうちに犯行スキルが熟達したと考えることができ、熟達度の高い変数ほど周縁部の方向に布置されてゆく。これは、各項目の該当割合の Facet として理解することができる。こうした一方向に向かう配列は、Simplex (Guttman, 1954) と呼ばれる。空間内に布置された変数の関係を、Polar Facet と Modular Facet または Circumplex と Simplex のような2つの区分の仕方としてみることによって、より多義的なデータ構造を把握することが可能となる。

このとき、Facet 理論によって区分された領域において、Guttman Scale をどのように把握することができるだろうか。二値変数で測定されたデータでは、Guttman Scale は各二値変数の割合によって順序づけられる。この順序構造は二値変数の割合を反映する Facet として現れると期待される。上記の Radex モデルでは、図の中心から周縁部に向かって布置される Modular Facet に二値変数の割合が反映される。そのため、Guttman Scale を構成する一連の変数は、同心円の中心から周辺方向に向かって観察されると考えられる。さらに、複数の Guttman Scale がデータ内に存在する場合、異なる Guttman Scale に含まれる割合の高い変数同士は中心部に比較的近接したかたちで布置される一方、異なる Guttman Scale に含まれる割合の低い変数同士は互いに離れたかたちで布置されると考えられる。すなわち、複数の Guttman Scale は、割合の高い二値変数を布置の中心に持ち、各 Guttman Scale の周縁部に向かって布置される（詳細は 4 章を参照されたい）。このことから、行動科学的データを扱う場合、Radex モデルの枠組みの中に Guttman Scale を視覚化することが期待できる。本研究では、行動科学的データに MDS を適用して得られた結果に、Radex モデルを仮定して Guttman Scale を測定する手法について検討する。

MDS によって布置された平面に Radex モデルを当てはめて解釈する場合、次の点に注意する必要がある。二値変数を表す点の位置関係は、2 つの変数間の類似度が高いほど近接して布置され、逆に 2 つの変数間の類似度が低いほど離れて布置される。図 3-3 でいえば、犯罪行動のテーマ A と類似性の高いテーマは B とテーマ F であるため、テーマ A の変数群とテーマ B の変数群は互いに近くに布置され、またテーマ A の変数群とテーマ F の変数群も互いに近くに布置される。一方、テーマ A との類似度が最も低いテーマ D の変数群は、テーマ A からもっとも離れた位置、図では中心から見てテーマ A の変数群の正反対の方向に布置されている。

ただし、この平面では、図の中心からの各変数の位置をベクトルとして、変数相互の関係がベクトルの内積として表現されとは限らないことに注意する必要がある。例えば、調査サンプルにおいて、特定の 2 つの変数への該当の有無に関連がない場合、この 2 つの変数の位置を示すベクトル同士の余弦が 0 になる、すなわち 2 つのベクトルが互いに直角をなす、とは限らない。また、特定の 2 つの変数への該当の有無に負の関連がある場合でも、2 つの変数の位置を示すベクトル同士の余弦が -1 になる、すなわち 2 つのベクトルが図の中心を挟んで 180 度の関係に布置される、とは限らない。MDS 空間における位置関係は、あくまで 2 つの変数間の類似度に対応した 2 点間の距離が表現されているに過ぎないことに留意したい。

3.5 適用するデータ

3.5.1 人工データ

(1) 10 変数からなる人工データ

初めに、人工的に作成したデータを提示する。10 変数について二値データ（該当あり=1、

該当なし=0としてコード)として測定されたものを想定した人工データ 1 (表 3-1) を作成した。これは、Guttman による完全尺度構造データ (Guttman, 1950) に該当する。各二値変数の割合をみると、変数 1 (v1) の割合がもっとも高く、変数の番号が大きくなるほど割合が低くなる。また、この人工データ 1 について、変数間の類似度として Jaccard 係数を算出した結果を表 3-2 に示す。

表 3-1 人工データ 1 (10 変数)

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10
s1	1	1	1	1	1	1	1	1	1	1
s2	1	1	1	1	1	1	1	1	1	0
s3	1	1	1	1	1	1	1	1	0	0
s4	1	1	1	1	1	1	1	0	0	0
s5	1	1	1	1	1	1	0	0	0	0
s6	1	1	1	1	1	0	0	0	0	0
s7	1	1	1	1	0	0	0	0	0	0
s8	1	1	1	0	0	0	0	0	0	0
s9	1	1	0	0	0	0	0	0	0	0
s10	1	0	0	0	0	0	0	0	0	0
割合	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1

表 3-2 人工データ 1 の変数間の連関 (Jaccard 係数による)

	v1	v2	v3	v4	v5	v6	v7	v8	v9
v2	0.90								
v3	0.80	0.89							
v4	0.70	0.78	0.88						
v5	0.60	0.67	0.75	0.86					
v6	0.50	0.56	0.63	0.71	0.83				
v7	0.40	0.44	0.50	0.57	0.67	0.80			
v8	0.30	0.33	0.38	0.43	0.50	0.60	0.75		
v9	0.20	0.22	0.25	0.29	0.33	0.40	0.50	0.67	
v10	0.10	0.11	0.13	0.14	0.17	0.20	0.25	0.33	0.50

(2) 14 変数からなる人工データ

さらに、二値データ（該当あり=1、該当なし=0 としてコード）として測定された 7 変数からなる完全尺度構造データを 2 つ含む、14 変数からなる人工データ 2（表 3-3）を作成し、このデータについても検討する。このデータは、以下のような手順で作成したものである。

まず、変数 1～7 について一つの完全尺度構造データとなるようにする。ここでは、変数 1～7 の該当割合を 0.8、0.7、0.6、0.5、0.4、0.3、0.2、0.1 とした。次に、そのデータを調査対象者に関してランダムに並び替えて、対象者の番号を振り直したデータを作成し、変数を 8～14 として先程のデータに追加する。ただし、14 個の変数全てについて該当がない調査対象者が生じないようにするため、変数 1～14 について該当がない対象者については、変数 1～7 に対して少なくとも 1 変数以上に該当があり、かつ変数 8～14 に対して少なくとも 1 変数以上に該当がある調査対象者のデータと、変数 8～14 のデータ部分を交換することでデータを作成した。データ作成のための R によるプログラムを付録 G.1 に示す。

各二値変数の割合は、データ作成で指定したとおり、変数 1～7 の中では変数 1 がもっとも割合が高く、変数の番号が大きくなるほど割合が低くなる。同様に、変数 8～14 では変数 8 がもっとも割合が高く、変数の番号が大きくなるほど割合が低くなる。

また、この人工データ 2 について、変数間の類似度として Jaccard 係数を算出した結果を表 3-4 に示す。Jaccard 係数をみるとわかるように、変数 1～7、変数 8～14 のそれぞれにおいて、隣り合う変数同士の連関が高く、離れるほどに連関が下がる。一方、変数 1～7 の変数群と変数 8～14 の変数群についてみると、変数 1 と変数 8 の連関が高く、変数 1 と変数 9、変数 1 と変数 10 と順に連関は下がる。つまり、一方の変数群の変数に対して、他方の変数群の変数番号が小さいほど連関が大きく、変数番号が大きくなるほど連関は小さくなる。

表 3-3 人工データ 2 (14 変数)

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	列和
s1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	12
s2	1	1	1	1	1	1	1	1	0	0	0	0	0	0	8
s3	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7
s4	1	1	1	1	1	1	1	1	1	1	1	1	0	0	12
s5	1	1	1	1	1	1	1	1	1	1	1	1	1	0	13
s6	1	1	1	1	1	1	1	1	0	0	0	0	0	0	8
s7	1	1	1	1	1	1	1	1	1	1	0	0	0	0	10
s8	1	1	1	1	1	1	1	1	0	0	0	0	0	0	8
s9	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7
s10	1	1	1	1	1	1	1	0	0	0	0	0	0	0	7
s11	1	1	1	1	1	1	0	1	1	1	1	1	0	0	11
s12	1	1	1	1	1	1	0	1	1	1	0	0	0	0	9
s13	1	1	1	1	1	1	0	1	1	1	1	1	0	0	11
s14	1	1	1	1	1	1	0	1	1	1	0	0	0	0	9
s15	1	1	1	1	1	1	0	1	1	1	1	1	0	0	11
s16	1	1	1	1	1	1	0	1	1	1	1	1	1	0	12
s17	1	1	1	1	1	1	0	1	1	1	1	0	0	0	10
s18	1	1	1	1	1	1	0	1	0	0	0	0	0	0	7
s19	1	1	1	1	1	1	0	1	1	1	0	0	0	0	9
s20	1	1	1	1	1	1	0	1	0	0	0	0	0	0	7
s21	1	1	1	1	1	0	0	1	0	0	0	0	0	0	6
s22	1	1	1	1	1	0	0	1	1	0	0	0	0	0	7
s23	1	1	1	1	1	0	0	1	1	1	1	1	1	0	11
s24	1	1	1	1	1	0	0	0	0	0	0	0	0	0	5
s25	1	1	1	1	1	0	0	0	0	0	0	0	0	0	5
s26	1	1	1	1	1	0	0	1	1	1	1	0	0	0	9
s27	1	1	1	1	1	0	0	0	0	0	0	0	0	0	5
s28	1	1	1	1	1	0	0	1	0	0	0	0	0	0	6
s29	1	1	1	1	1	0	0	0	0	0	0	0	0	0	5
s30	1	1	1	1	1	0	0	0	0	0	0	0	0	0	5
s31	1	1	1	1	0	0	0	1	1	1	1	0	0	0	8
s32	1	1	1	1	0	0	0	1	1	1	1	1	1	1	11
s33	1	1	1	1	0	0	0	1	1	1	1	1	1	0	10
s34	1	1	1	1	0	0	0	1	1	1	1	1	1	1	11
s35	1	1	1	1	0	0	0	1	1	1	1	1	1	1	11
s36	1	1	1	1	0	0	0	1	0	0	0	0	0	0	5
s37	1	1	1	1	0	0	0	1	1	1	1	0	0	0	8
s38	1	1	1	1	0	0	0	1	1	1	1	0	0	0	8
s39	1	1	1	1	0	0	0	0	0	0	0	0	0	0	4
s40	1	1	1	1	0	0	0	1	0	0	0	0	0	0	5
s41	1	1	1	0	0	0	0	0	0	0	0	0	0	0	3
s42	1	1	1	0	0	0	0	1	1	1	1	1	1	0	9
s43	1	1	1	0	0	0	0	1	1	1	0	0	0	0	6
s44	1	1	1	0	0	0	0	1	1	1	1	0	0	0	7
s45	1	1	1	0	0	0	0	1	1	0	0	0	0	0	5
s46	1	1	1	0	0	0	0	0	0	0	0	0	0	0	3
s47	1	1	1	0	0	0	0	1	1	1	0	0	0	0	6
s48	1	1	1	0	0	0	0	1	1	1	1	1	1	1	10
s49	1	1	1	0	0	0	0	1	1	0	0	0	0	0	5
s50	1	1	1	0	0	0	0	0	0	0	0	0	0	0	3
s51	1	1	0	0	0	0	0	1	1	1	1	1	1	0	8
s52	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2
s53	1	1	0	0	0	0	0	1	1	1	0	0	0	0	5
s54	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2
s55	1	1	0	0	0	0	0	1	1	1	0	0	0	0	5
s56	1	1	0	0	0	0	0	1	0	0	0	0	0	0	3
s57	1	1	0	0	0	0	0	1	1	1	1	1	1	0	8
s58	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2
s59	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2
s60	1	1	0	0	0	0	0	1	1	1	1	1	1	1	9
s61	1	0	0	0	0	0	0	1	0	0	0	0	0	0	2
s62	1	0	0	0	0	0	0	1	1	1	1	1	1	1	8
s63	1	0	0	0	0	0	0	1	1	0	0	0	0	0	3
s64	1	0	0	0	0	0	0	1	1	1	1	0	0	0	5
s65	1	0	0	0	0	0	0	1	0	0	0	0	0	0	2
s66	1	0	0	0	0	0	0	1	0	0	0	0	0	0	2
s67	1	0	0	0	0	0	0	1	1	0	0	0	0	0	3
s68	1	0	0	0	0	0	0	1	1	1	0	0	0	0	4
s69	1	0	0	0	0	0	0	1	1	1	1	1	0	0	6
s70	1	0	0	0	0	0	0	1	1	1	1	1	0	0	6
s71	1	0	0	0	0	0	0	1	1	1	1	0	0	0	5
s72	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
s73	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
s74	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
s75	1	0	0	0	0	0	0	1	0	0	0	0	0	0	2
s76	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
s77	1	0	0	0	0	0	0	1	1	1	1	0	0	0	5
s78	1	0	0	0	0	0	0	1	0	0	0	0	0	0	2
s79	1	0	0	0	0	0	0	1	1	1	1	1	1	1	8
s80	1	0	0	0	0	0	0	1	0	0	0	0	0	0	2
s81	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
s82	0	0	0	0	0	0	0	1	1	1	1	1	1	0	6
s83	0	0	0	0	0	0	0	1	1	0	0	0	0	0	2
s84	0	0	0	0	0	0	0	1	1	1	1	1	1	1	7
s85	0	0	0	0	0	0	0	1	1	1	1	1	1	1	7
s86	0	0	0	0	0	0	0	1	1	1	1	1	0	0	5
s87	0	0	0	0	0	0	0	1	1	0	0	0	0	0	2
s88	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
s89	0	0	0	0	0	0	0	1	1	0	0	0	0	0	2
s90	0	0	0	0	0	0	0	1	1	1	1	1	0	0	5
s91	0	0	0	0	0	0	0	1	1	1	1	1	1	1	7
s92	0	0	0	0	0	0	0	1	1	1	0	0	0	0	3
s93	0	0	0	0	0	0	0	1	1	1	1	0	0	0	4
s94	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
s95	0	0	0	0	0	0	0	1	1	1	1	1	0	0	5
s96	0	0	0	0	0	0	0	1	1	0	0	0	0	0	2
s97	0	0	0	0	0	0	0	1	1	1	1	1	1	0	6
s98	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
s99	0	0	0	0	0	0	0	1	1	0	0	0	0	0	2
s100	0	0	0	0	0	0	0	1	1	1	1	1	1	0	6
割合	0.8	0.6	0.5	0.4	0.3	0.2	0.1	0.8	0.6	0.5	0.4	0.3	0.2	0.1	

表 3-4 人工データ 2 の変数間の連関 (Jaccard 係数による)

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13
v2	0.75												
v3	0.63	0.83											
v4	0.50	0.67	0.80										
v5	0.38	0.50	0.60	0.75									
v6	0.25	0.33	0.40	0.50	0.67								
v7	0.13	0.17	0.20	0.25	0.33	0.50							
v8	0.60	0.46	0.41	0.35	0.25	0.20	0.08						
v9	0.46	0.40	0.36	0.28	0.20	0.18	0.06	0.75					
v10	0.43	0.39	0.35	0.30	0.21	0.21	0.07	0.63	0.83				
v11	0.33	0.30	0.29	0.27	0.17	0.15	0.06	0.50	0.67	0.80			
v12	0.24	0.23	0.21	0.21	0.15	0.16	0.08	0.38	0.50	0.60	0.75		
v13	0.16	0.18	0.15	0.13	0.06	0.05	0.03	0.25	0.33	0.40	0.50	0.67	
v14	0.08	0.08	0.07	0.06	0.00	0.00	0.00	0.13	0.17	0.20	0.25	0.33	0.50

3.5.2 メディア暴力調査データ

次に、実施した調査データについて説明する。インターネットや漫画、雑誌、テレビなどのメディアの内容について、青少年の興味、普段の行動や考え方を理解する目的で「青少年の興味に関するアンケート調査」(代表者：宮寺貴之、科学研究費補助金若手研究 (B)、研究課題番号 20730466 による)を実施した。なお、これ以降では本調査データを「メディア暴力調査データ」と呼ぶこととする。

(1) 調査の背景

この調査は、青年期を対象として、メディアに描かれる暴力（以下、メディア暴力）への嗜好性と、日常生活上の不適応との関連を明らかにすることを目的としたものである。日常生活に対する深刻な不適応の一つに犯罪行動がある。犯罪行動は極端ではあるが、未成年者（以下、少年）の背景と問題行動との関連を示唆する重要な事例といえる。ここでは、少年による犯罪行動の一つとして殺人に着目する。

少年による殺人事件は多様であるといわれ (Miyadera, 2005)、殺人を犯した少年の中には決して非行性が高くない者も含まれる。そうした非行性の高くない殺人を犯した少年の背景を調べていくと、そうした者の中に、メディア暴力への強い嗜好性を示す者が散見される。我が国で発生した特異・凶悪な事件の被疑少年の背景を調べた調査 (警察庁・科学警察研究所, 2000) では、ホラー映画や暴力的な内容のビデオ・ゲームに強い嗜好性を示す者、拳銃やナイフ等の武器に強い興味を示す者、過去の重大犯罪の犯罪者に対する興味や憧れを述べる者が見られた。こうした特徴は、米国の校内乱射事件の犯人、いわゆるスクール・シューターにも共通して見られている (McGee & DeBernarde, 1999)。さらに、こうした少年たちの背景に特徴的なものの一つとして、日常生活への不適応感、特に孤立感が指摘さ

れる。そのため、孤立感を中心とした日常生活上の不適応と、メディア暴力への嗜好性との関連が問題となる。

日常生活における不適応とメディア暴力への嗜好性との関連性を示唆する研究はいくつかみられる。東（1999）は日本の中学 2 年生を対象とした調査から、家庭や学校に不満を持つ者ほど、メディア暴力に肯定的な態度を有していることを明らかにした。また、暴力的なゲームへの強い嗜好性は、深刻なレベルの臨床的な問題と関連することも指摘されている（Funk et al., 2002）。

日常生活における不適応とメディア暴力への嗜好性が関連する理由はいくつか考えられる。メディア利用に関する研究として、例えば、Strizhakova & Krcmar (2007) は、否定的な気分を感じている人ほどレンタルビデオで一般のドラマや犯罪もののドラマを借りる者が多く、深刻な内容の映画を一人で鑑賞することと関連することを見出した。社会的比較理論（Social Comparison Theory; Festinger, 1954）を引用し、自分の置かれた状況よりもひどい状況にいる主人公の話を見ることで、自分の置かれた状況のほうがましであると感じることにより、気分を管理しているのではないかと考察している。また、メディア暴力の利用は、日常や現実からの逃避として利用されることもあり（Kirsh, 2006）、寂しさや怒りを紛らす意味がある（Johnston, 1995）ことも指摘される。このほか、不適応や孤立から生じる周囲に対する怒りを代理的に満足するために、攻撃的な内容のメディア暴力を利用することも考えられる。

メディア暴力の利用とその原因については因果関係を特定しにくいものの、生活上の不適応が強い場合、メディアの影響を受けやすい可能性があり、悪循環になることが考えられる（Comstock, 2008）。メディア暴力の視聴によって攻撃性が高まることは実証的に示されていることから（Anderson et al., 2003）、不適応状態でメディア暴力を利用することは、自身の攻撃性を高め、さらなる周囲との問題を生じさせ、不適応を高めることが予想される。

こうしたことから、メディア暴力に対する強い嗜好性は、少年の不適応のサインになり得ると考えられる。もちろん、メディア暴力への嗜好性は、刺激欲求（Zuckerman, 1996）とも関連し、メディア暴力への嗜好性そのものが必ずしも不健康を意味するという訳ではない。しかし、メディア暴力への嗜好性が非常に強く、その利用頻度が著しく高いことは、少年の不適応に対する懸念を生じさせる。

そこで、メディア暴力への嗜好性を測定する尺度を検討するため、本調査は実施された。しかし、メディア暴力と一口にいても、暴力として描かれる内容は多様である。暴力的犯罪者のメディア暴力への関心を調べた先行研究では、暴力的内容として、軍隊や武器等への関心を示すミリタリズム（Militarism）、オカルトや超常現象に対する関心を示すオカルティズム（Occultism）、反社会的行動への関心や犯罪者への憧れ等を示す犯罪同一性（Crime Identity）の 3 つが指摘されている（Egan et al., 1999; Weiss et al., 2004; Charles & Egan, 2009）。そこで、「青少年の興味に関するアンケート調査」では、こうしたメディア暴力の内容を考

慮し、かつ多くの人に該当する項目からメディア暴力への嗜好性が強いごく一部の人のみに該当する項目までを含めて、メディア暴力への嗜好性の程度を測定する尺度の検討を行うことを目的とした。

(2) 調査の方法

調査会社のインターネット調査パネルに登録された全国のモニターのうち、13～18 歳の少年を子どもに持ち、その少年と同居する保護者に対して、メディアへの接触状況やメディアの嗜好性に関するオンライン調査に匿名性を担保した上で少年を参加させてもよい、と保護者の同意が得られた場合に、その少年に対してオンライン調査画面を提示した。結果は集計されて利用されるため個人が特定されることはないこと、どのように回答しても回答者本人や保護者が不利益を被ることはないこと、調査の途中で回答を棄権することも可能であること、調査時間は概ね 25 分程度であり、できるだけ本人一人で回答するように教示した。

調査対象者は 13～18 歳の少年 1400 人（男子 700 人、女子 700 人）の割当数に達するまで調査への協力を依頼した。調査票は、メディア（テレビゲーム、映画やドラマ、漫画、漫画以外の書籍、インターネット）の視聴内容や視聴時間、性格等に関する質問項目から構成される。調査に対する回答を終了した場合に、保護者に対して謝金に当たるポイントを付与した。調査の実施時期は平成 22 年 3 月である。

調査に参加した少年の特徴を表 3-5 に示す。

表 3-5 メディア暴力調査に参加した少年の特徴

出生順位	第 1 子 58.7%、第 2 子 32.5%、第 3～4 子 8.8%
平均年齢	15.5 歳（SD = 1.7）
男女別	男子 15.5 歳（1.7）、女子 15.5 歳（1.7）
学職	中学 49.0%、高校*・高等専門学校 47.8%、専門学校・短大・大学 2.2%、その他 1.1%
男子	中学 48.9%、高校*・高等専門学校 48.1%、専門学校・短大・大学 1.7%、その他 0.9%
女子	中学 49.1%、高校*・高等専門学校 47.4%、専門学校・短大・大学 2.6%、その他 0.6%

* 高校は全日制、定時制、通信制を含む。

本調査は、メディア暴力の視聴行動を測定することを目的としたため、暴力的なシーンを含むメディアのテーマやシーンに関する特徴について候補となる項目を収集した後、KJ 法（川喜田, 1967）によって整理して調査項目を作成した。これらの調査項目のうち、本研究に使用したデータは、「普段よく見る映画・ドラマ（アニメを含む）としてあてはまるものをいくつかも選んでください」という教示により、映画・ドラマの特徴を示す 18 項目について回答（複数回答可）を求めた二値データである。質問の教示文及び項目を付録 B に

示す。

これらの回答のうち、分析に使用するデータは「映画・ドラマはあまり見ない」と回答した者 ($n = 192$) を除く 1208 人の二値データ (該当あり=1、該当なし=0 としてコード) で、項目 18「その他の映画やドラマ (アニメをふくむ)」を除く 17 項目までのデータとする。尋ねた 18 項目とそれぞれに対する回答割合を表 3-6 に示す。暴力的メディアとは関連しない項目として含めた「コメディーもの」や「恋愛もの」の選択割合が高い一方、暴力的なテーマ・シーンに関する特徴を表す項目は概して割合が低かった。

以下、前節までに説明したデータの視覚化に関わる各分析手法を利用して、本節で示した人工データ 2 種類、調査データ 1 種類の計 3 種類のデータを適用した結果について言及する。

表 3-6 よく見る映画・ドラマの特徴に関する項目

項目 番号 a	映画・ドラマの特徴	本文図中 の略号	全体 n=1208	男子 n=596	女子 n=612	b
1	コメディーもの	comedy	64.9	62.4	67.3	+
3	恋愛もの	love	38.0	13.8	61.6	***
14	ファンタジー（空想世界）もの	fantasy	37.3	37.8	36.9	
4	アクションもの	action	30.5	45.0	16.5	***
13	S F もの	sf	27.6	38.8	16.7	***
2	ヒューマンドラマ系	human	24.1	19.1	28.9	***
10	サスペンス、推理もの	suspense	23.5	21.1	25.8	+
6	警察や刑事もの	police	18.5	23.0	14.1	***
11	ホラー（幽霊などオカルト系）	occult	10.0	6.9	13.1	***
7	スパイもの	spy	8.9	13.3	4.6	***
5	戦争や軍隊もの	war	5.5	9.2	1.8	***
16	武器（銃やナイフなど）を使った攻撃場面がでてくるもの	weapon	4.8	7.6	2.1	***
17	血が流れるなどのリアルな場面がでてくるもの	blood	2.9	3.4	2.5	
12	ホラー（残虐な殺害シーンのあるスプラッター系）	splatter	2.9	2.5	3.3	
15	暴力的な場面が出てくるもの	vio	2.6	2.7	2.5	
9	犯罪者が主人公の話	crime	1.9	2.2	1.6	
8	ギャング、暴力団もの	gang	1.7	1.8	1.5	
18	その他の映画やドラマ（アニメをふくむ）	—	22.4	24.8	20.1	*

^a 項目番号は調査票における提示順序を示す。

^b 男女差に関する χ^2 二乗検定の結果を示す（*** $p < .001$, * $p < .05$, + $p < .10$ ）。

3.6 データへの適用

前節で準備した二値データに本章 1~3 節で説明した各種の統計手法を適用し、その結果に各二値変数の割合がどのように反映されるかについて検討する。特に、Guttman Scale の視覚化として、前節で準備した人工データである完全尺度がどのように 2 次元平面に反映されるかに着目したい。

3.6.1 等質性分析の適用

(1) 人工データ 1

人工データ 1 (10 変数) に等質性分析を適用した結果は、図 3-4 の通りである。なお、R によるプログラムを付録 G.2 に示す。

変数の番号に続く括弧内には、各二値変数の割合 (%) を示した。これをみると、横軸に沿って左から右方向に向かって変数の割合が高くなるように各変数が布置されている様子がわかる。このデータは Guttman の完全尺度であり、1 次元の尺度であるため、この横軸に沿って布置されれば充分であるが、2 次元の平面では縦軸方向に逆 U 字型に布置されている。これは、いわゆる馬蹄形 (Horseshoe または Arch Effect; Greenacre, 2007) といわれ、Guttman の完全尺度に等質性分析を適用する場合に観測されることが知られている (足立・村上, 2011)。このため、変数の割合に沿って、横軸方向に大まかな区分を描くことはできるが、この縦軸方向の歪みを変数間の関係を解釈する際の障害になることが考えられる。

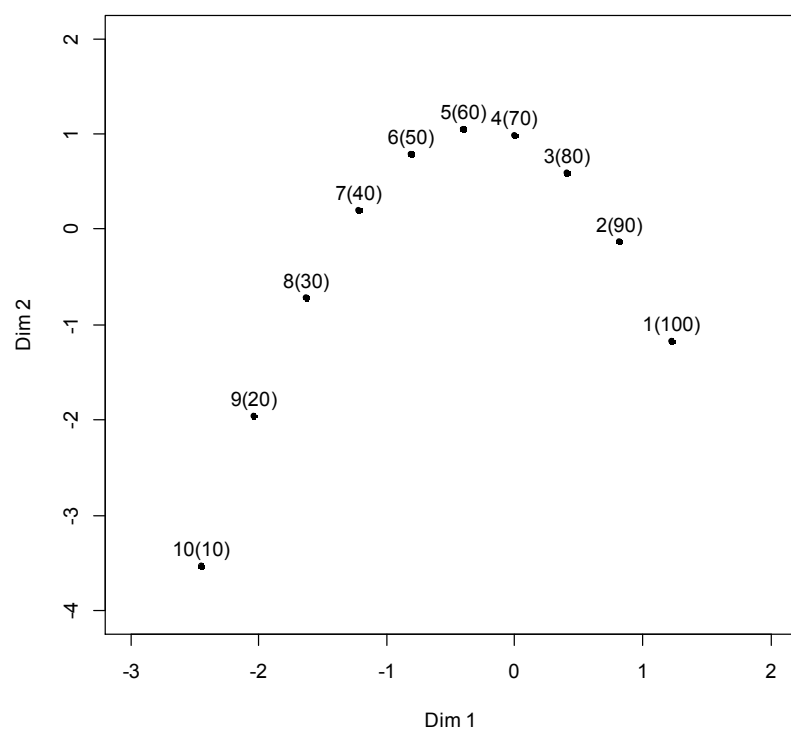


図 3-4 人工データ 1 (10 変数) に等質性分析を適用した結果 (括弧内は割合 (%))

(2) 人工データ 2

人工データ 2 に等質性分析を適用した結果は、図 3-5 の通りである。変数の番号に続く括弧内には各二値変数の割合（％）を示した。プロットをみると、変数 1～7 は図の右側に縦に並び、変数 8～14 は図の左側に同じく縦に並んでいる。そして、布置の下側から上側に向かつて、割合が小さくなる順に変数が布置されている様子がみられる。このことから、各二値変数の割合を反映し、かつ Guttman Scale としてのまとまりをもつ 2 つの変数群（変数 1～7 と変数 8～14）の布置を隔てた領域の区分を策定することが可能である。

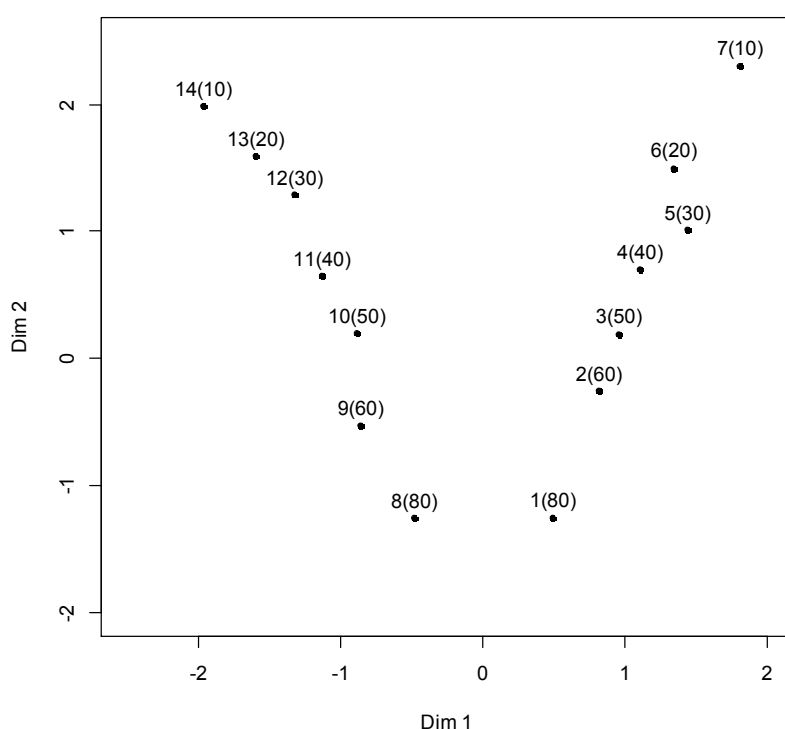


図 3-5 人工データ 2（14 変数）に等質性分析を適用した結果（次元 1×次元 2、変数括弧内は割合（％））

(3) メディア暴力調査データ

調査データ（17 変数）に等質性分析を適用した結果を検討する。結果は、図 3-6 の通りである。まず変数の内容をみると、図の右下にヒューマンドラマや恋愛ものなどの非暴力系（comedy、human、love）、図の左下にアクション系（police、action、war、spy、ただし suspense、fantasy、sf の各変数も含まれる）、図の左上にメディア暴力系（weapon、gang、crime、blood、splatter、vio）と、大きく 3 つの変数群に布置が分類できる様子が見受けられる。し

かし、オカルトの変数（occult）は図の中央に布置されている。

また、二値変数の割合についてみると、割合が最大の変数であるコメディ（65%）を中心に割合の大きい変数が図の右下に布置されており、そこから図の左上方向に離れて布置される変数ほど割合が低くなっていくという傾向がみてとれる。

これらを総合すると、大まかに変数の実質的意味、そして各二値変数の割合の傾向を捉えて視覚化できており、Radex モデルの想定も可能であると考えられる。しかし、細かい部分で意味内容の異なる変数の布置に解釈しづらい部分が生じている。

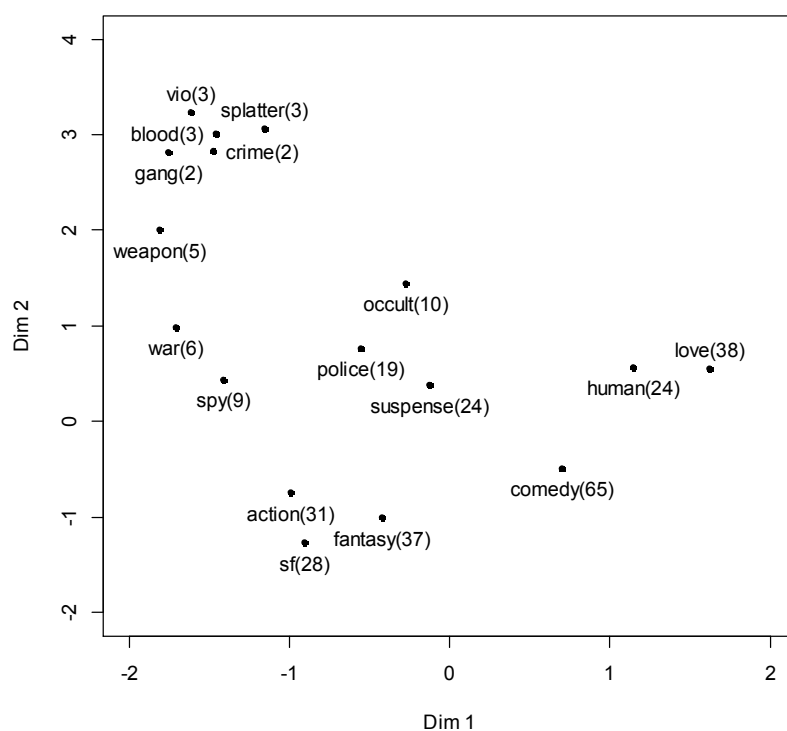


図 3-6 メディア暴力調査データ（17 変数）の二値データに等質性分析を適用した結果（各変数の括弧内は割合（%））

3.6.2 Kruskal の MDS の適用

(1) 人工データ 1

人工データ 1 に適用した結果を検討する。Jaccard 係数の値を 1 から引いた値を非類似度として、Kruskal の MDS を適用した結果は、図 3-7 の通りである。なお、R によるプログラムを付録 G.3 に示す。

10 個の変数が、割合の高いものから低いものへと左から右へと順に並び、その値の順序が横軸に反映されて布置されている。このことから、二値変数の割合を反映した軸が想定できる。しかし、変数の並びが直線ではなく U 字形をしており、等質性分析と同様に縦軸方向に布置の歪みの問題がみられる。

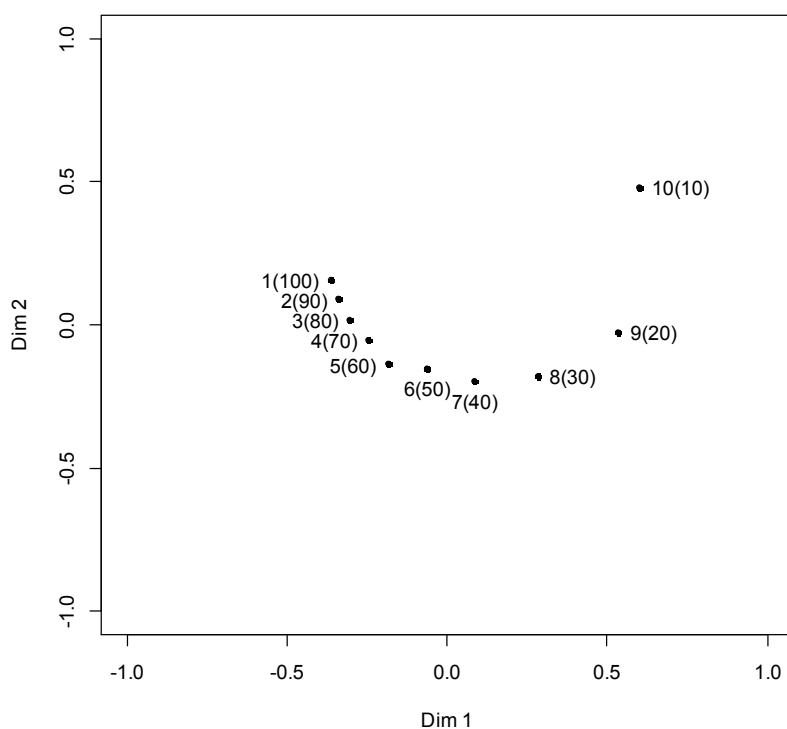


図 3-7 人工データ 1 (10 変数) に Kruskal の MDS を適用した結果 (2 次元、Stress = 1.004799×10^{-13} 、括弧内は割合 (%))

(2) 人工データ 2

人工データ 2 に適用した結果を検討する。Jaccard 係数の値を 1 から引いた値を非類似度として、Kruskal の MDS を適用した結果は、図 3-8 の通りである。

変数 1～7 が図の左側に、変数 8～14 が図の右側に並んでおり、割合の高い変数ほど図の下側に布置され、割合が低くなるほど上側に布置されている。このことから、2 つの Guttman

Scale の並びが区別され、馬蹄形の歪みが若干見られるものの、それぞれが概ね直線的に並んでおり、視覚的な把握は困難ではない。また、割合を反映する領域を考える際、変数 1～7 の布置と、変数 8～14 の布置で割合の等しい変数同士を見較べることで、変数の割合を反映する領域の区分を検討することも可能である。

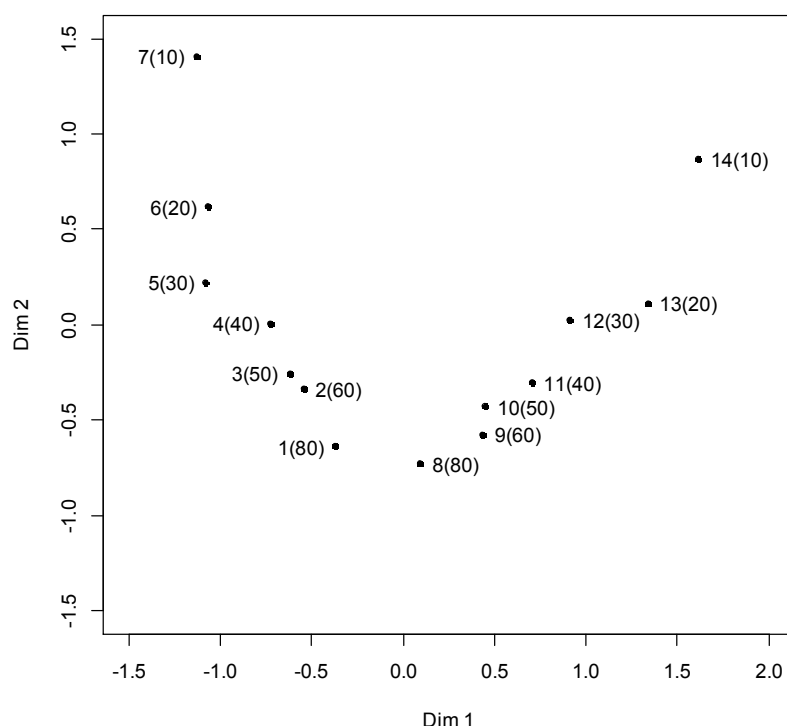


図 3-8 人工データ 2 (14 変数) に Kruskal の MDS を適用した結果 (2 次元、Stress = 1.526256、括弧内は割合 (%))

(3) メディア暴力調査データ

調査データ (17 変数) に適用した結果を検討する。Jaccard 係数の値を 1 から引いた値を非類似度として、Kruskal の MDS を適用した結果を図 3-9 に示す。

図の左側に割合の高い変数群、右側に割合の低い変数群が布置されていることがわかる。また、図の下側には S F (sf)、アクション (action)、警察 (police)、スパイ (spy)、戦争もの (war) が布置されている。図の右側にはメディア暴力と関連すると考えられる変数群 (vio、blood、weapon、gang、crime) が布置され、オカルトもの (occult、splatter) が図の上から右上の位置にみられる。また、図の左上には恋愛もの (love)、ヒューマンドラマ (human) がみられる。

このように、映画・ドラマの特徴によって変数を分類することができ、大まかな割合の

相違も併せて検討できることから、コメディーの変数（comedy）の近傍を中心とした同心円を考えると、Radex モデルの存在を想定できる。しかし、この図から Radex モデルを想定し、Guttman Scale を考慮した変数の組み合わせを検討するためには、Guttman Scale の構成度の指標を利用した補助的な吟味が必要となろう。

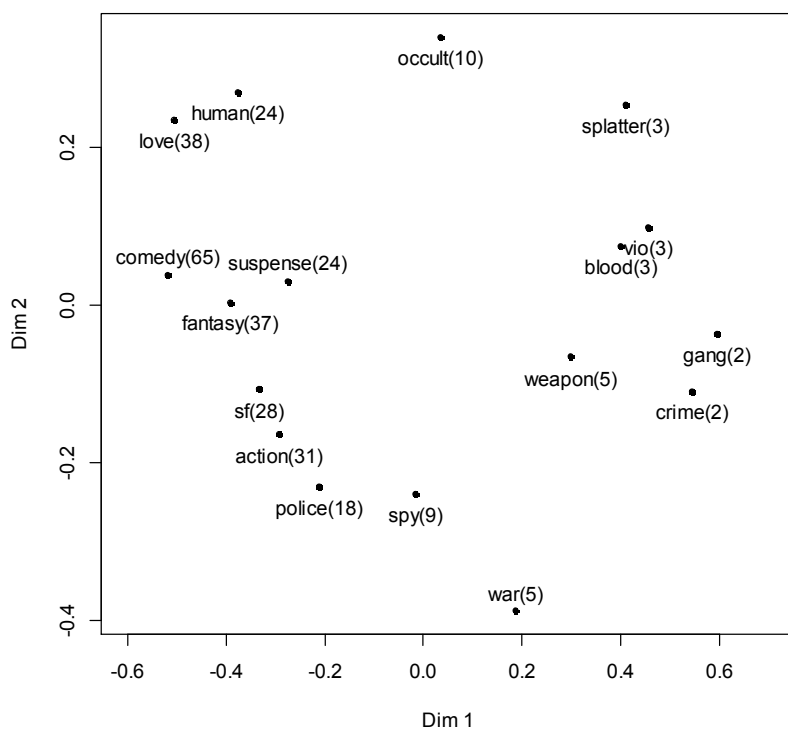


図 3-9 メディア暴力調査データ（17 変数）の二値データに Kruskal の MDS を適用した結果（2 次元、Stress = 8.085452、括弧内は割合（%））

3.6.3 Guttman の SSA の適用

(1) 人工データ 10 変数

人工データ 1 に適用した結果を検討する。Jaccard 係数の値を 1 から引いた値を非類似度として、SSA を適用した結果は図 3-10 の通りである。

10 個の変数が右から左へと順に並び、各二値変数の割合の順序が横軸の値に反映されており、変数の割合が高いほど図の右側に布置されている。このことから、二値変数の割合を反映した軸を想定することができ、これに沿って **Guttman Scale** を構成する並びを把握することが可能である。また、縦軸方向に馬蹄形様の歪みが若干見られている。

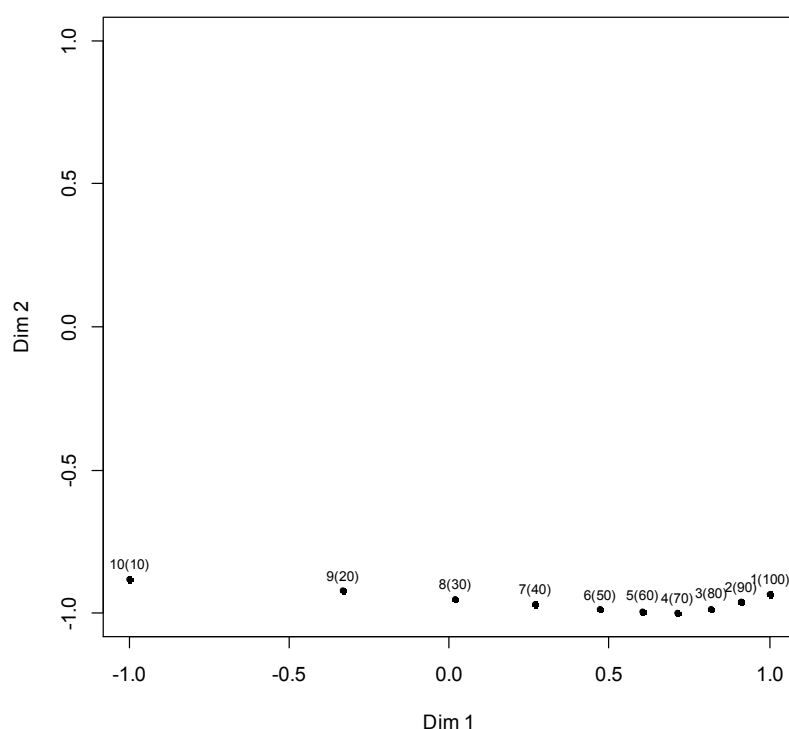


図 3-10 人工データ 1 (10 変数) に SSA を適用した結果 (2 次元、CoA = 0.013890、括弧内は割合 (%))

(2) 人工データ 14 変数

人工データ 2 に適用した結果を検討する。Jaccard 係数の値を 1 から引いた値を非類似度として、SSA を適用した結果は、図 3-11 の通りである。

変数 1～7 が図の右側に図の中心から右方向に向かって並び、その順に二値変数の割合が低くなっている。また、変数 8～14 が図の左側に図の中心から左方向に向かって並び、その順に変数の割合が低くなっている。このことから、割合の高い変数が図の中心に、そこ

から左右に離れるに従って割合の低い変数が布置される様子が見られる。Guttman Scale を Polar Facet、二値変数の割合を Modular Facet として領域を区分した Radex モデルを検討することが可能である。

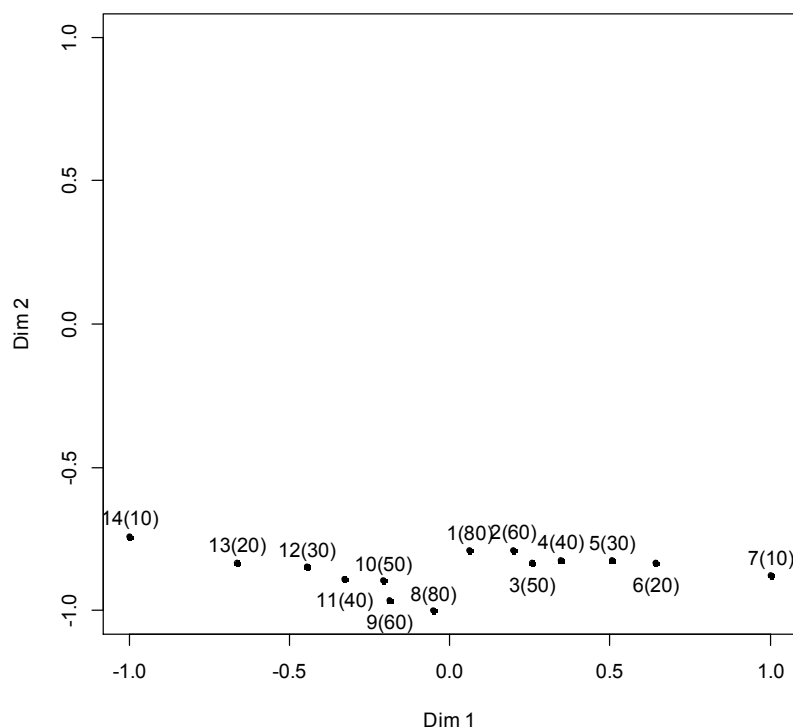


図 3-11 人工データ 2 (14 変数) に SSA を適用した結果 (2 次元、CoA = 0.005580、括弧内は割合 (%))

(3) メディア暴力調査データ

調査データ (17 変数) に適用した結果を検討する。Jaccard 係数の値を 1 から引いた値を非類似度として、SSA を適用した結果は、図 3-12 の通りである。

図を見ると、Kruskal の MDS の適用結果に酷似している。左側に割合の高い変数群、右側に割合の低い変数群が布置されている。図の下側には S F (sf)、アクション (action)、警察 (police)、スパイ (spy)、戦争もの (war) が布置され、図の右側にはメディア暴力と関連する変数群 (weapon、blood、vio、crime、gang)、図の上側にホラー関連の変数群 (occult、splatter) が布置されている。また、恋愛もの (love)、ヒューマンドラマ (human) はコメディ (comedy) の近傍に布置されている。このように、映画・ドラマの特徴によって分類でき、大まかな割合の相違も併せて検討できる。コメディの変数 (comedy) の近くを中心として、メディアの特徴を Polar Facet、二値変数の割合を Modular Facet として見ると、

大まかな Radex モデルが想定できる。しかし、この図だけから Guttman Scale を考慮した変数の組み合わせを検討するためには、Guttman Scale の構成度の指標を利用した補助的な吟味が必要となろう。

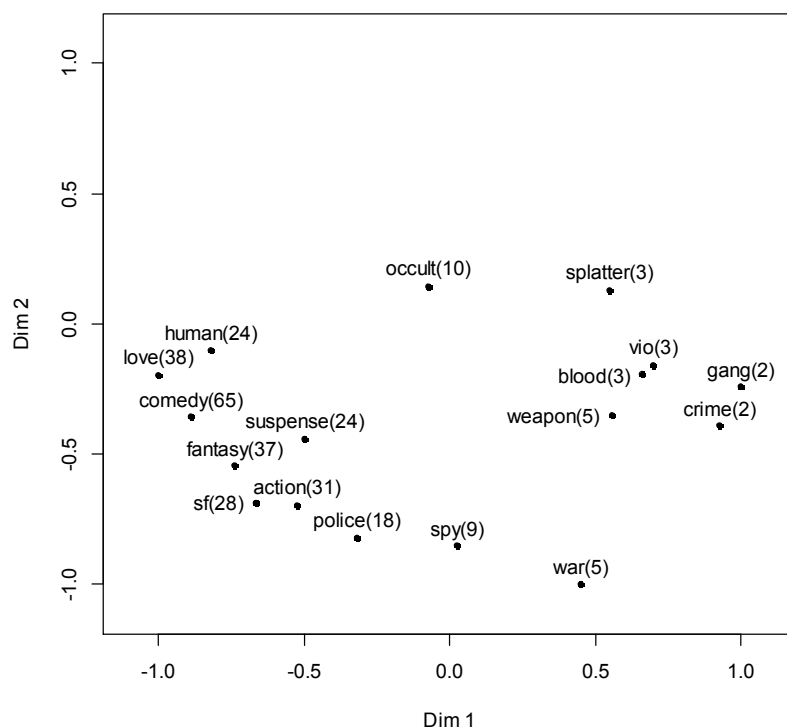


図 3-12 メディア暴力調査データ (17 変数) の二値データに SSA を適用した結果 (2 次元、CoA = 0.115130、括弧内は割合 (%))

3.7 Facetted SSA

本章の最後に、4 節で説明した Facet 理論を 2 次元空間に当てはめる数理モデルについて言及する。ただし、このモデルは平面に布置される対象の属するカテゴリがデータ収集の時点で既に決まっている場合に適用する手法であり、本研究のように布置された結果から対象のカテゴリを探索的に調べる手法ではないことから、あくまで参考として言及するにとどめる。

Faceted SSA (Borg & Shye, 1995) は、4 節で言及した領域に関する 3 つのモデル、すなわち平行する直線によって区分される Axial Facet (Axial モデル)、同心円によって区分される Modular Facet (Radial モデル)、中心から放射状に伸びる半直線によって区分される Polar Facet (Angular モデル) について、どのモデルがもっとも当てはまりがよいかをデータに基づいて評価する。SSA の結果である 2 次元上の点の布置に各モデルを当てはめ、その領域

モデルから逸脱した点を利用して目的関数を定義する。具体的には、点のカテゴリが平面上の領域のカテゴリと一致しない点を逸脱点 (Deviant Point) とし、各逸脱点からその点が属すべきカテゴリの領域までの距離の和を目的関数とする。そして、この目的関数を最小化する領域モデルが、SSA の結果を解釈するのにもっともふさわしい領域モデルである、と考える。各領域モデルの目的関数については、付録 C を参照されたい。

第 4 章 Radex モデル探索支援手法（RADXEA）の提案

本章では、2 次元平面上に示される SSA（最小空間分析）の結果を Facet 理論の領域モデル（Modular、Polar、Radex 等）を利用して解釈する方法を提案する。具体的には Guttman Scale を利用し、所与の SSA の結果に対して Radex モデルを探索的に当てはめる新たな手法を提案する。以下、この Radex モデルを探索的に当てはめる本手法を Radex モデル探索支援手法（Radex Exploration Assistance, RADXEA）と呼ぶことにする。

4.1 問題と目的

本章の狙いは、二値変数である項目間の関係から項目の割合を考慮して複数の一次元尺度を抽出することである。特に行動科学的データにおいては、二値変数の割合に着目することが有用な場合がある。割合の高い変数は調査対象のサンプルにおいて共通性の高い行動特徴を意味し、割合の低い変数は特定の対象者にしか該当しない。例えば、第 3 章で言及した犯罪行動における特殊化（Yokota & Canter, 2004）の例をみると、特殊化が進んだ者ほど犯行スキルの高い特徴を示し、その特徴の該当者の割合は少ない。ある一連の特殊化を想定すると、割合の高い変数と、特殊化の指標となる割合の低い変数は、その特殊化の程度を測定する一次元尺度を構成するわけである。このように、行動科学的データでは二値変数の割合を考慮することが有用となりうる。

しかし、第 3 章に示した各種の分析手法は、二値変数の割合を利用した結果の解釈が積極的に考慮されているわけではない。また、分析に各変数間の連関指標を利用するが、第 2 章で示したように、二値変数の割合が大きく異なる 2 変数は離れて布置されるため、それらの 2 変数間に存在する一次元性を検討する方法も明確には示されていない。そこで、本章は二値変数の割合を用いて一次元尺度を構成する Guttman Scale を利用する。第 3 章で言及したように、この Guttman Scale を視覚化する手法として、Facet 理論における Radex モデルに着目する。

二値変数の割合の視覚的表現としてシンプルなのは、割合の順に変数を点列として並べることである。さらに、複数の尺度が存在する場合、これらの尺度同士を 2 次元平面に配置する。各点列に含まれる変数同士の連関が強ければ、点列の間の類似度は高いことから互いに近くに配置されることが望ましい。Facet 理論の領域モデルでは、点列を分ける横方向の区分線と二値変数の割合を分ける縦方向の区分線が並ぶ Axial Facet として表現できる（図 4-1）。

一方、点列同士の類似度が連続的に循環する場合も考えられる。類似度の離れた点列が、最初の点列に近づくような場合である。この状況を 2 次元で表現する場合、点列は一方の端点を中心として円形に配列されたかたちを考えることができる。これは点列同士を区分する Polar Facet と、二値変数の割合を区分する Modular Facet からなる Radex モデルである（図 4-2）。第 3 章で言及した Canter（2000）が提唱する犯罪行動の Radex モデルがこれに

当たる。割合の高い変数同士ほど連関が強く、割合の低い変数同士ほど連関が弱くなり、通常は図 4-1 よりも図 4-2 の Radex モデルの方が経験的に観察されやすい (Borg & Groenen, 1998; Borg & Shye, 1995)。

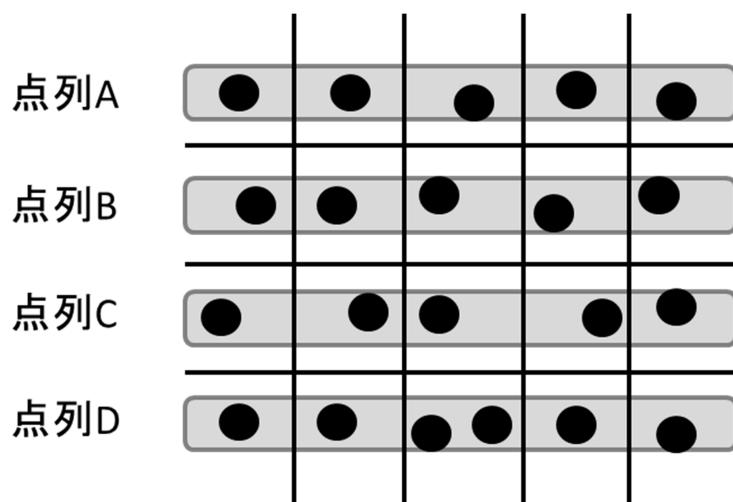


図 4-1 縦方向と横方向に Axial Facet が並んだ領域モデル

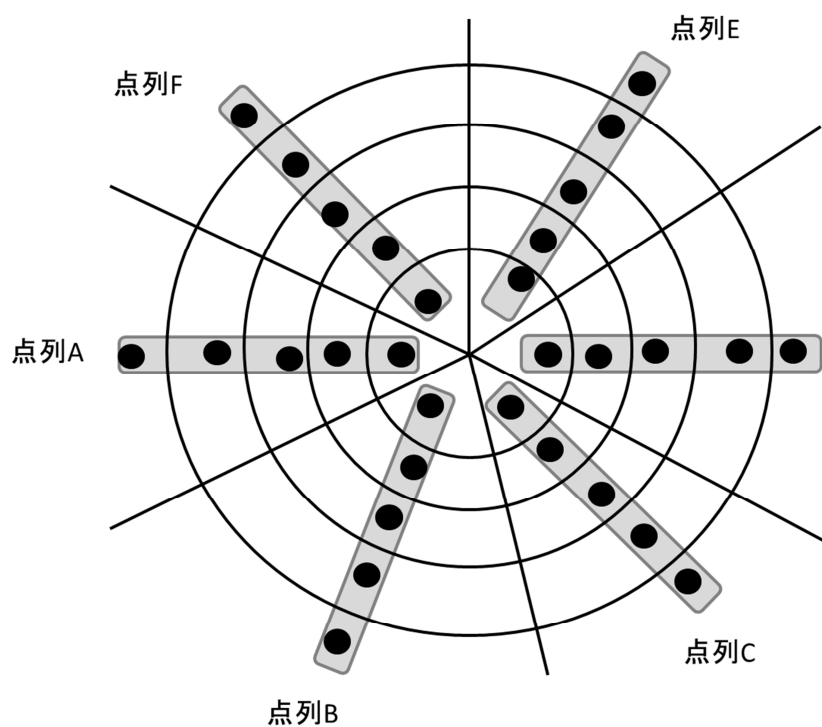


図 4-2 Polar Facet と Modular Facet が組み合わさった Radex モデル

このように、Facet 理論の Radex モデルを当てはめることで、MDS 等の結果の解釈をよりシステマティックに行うことが可能となる。しかし、従来、布置された変数を見ながら行う変数の位置関係の解釈は主観的な作業とならざるを得なかった。例えば、変数 A、B、C が人間の「行動」に関する項目、変数 D、E、F が人間の「認知」に関する項目であり、これらの SSA の結果において、それぞれの変数群が布置される領域が異なることから、両者の間に区分線を想定できるとする。これは「行動」や「認知」という質問の対象が Facet に当たり、それぞれに対応する 2 つの領域が分割できる、と解釈される。その際、領域の区分線をどの変数の位置からどのくらい離れた箇所に引くのか、という作業は主観的な判断となる。

そこで本章では、SSA の 2 次元の布置を所与として、この布置に Radex モデルを当てはめる際、データに基づく客観的な方法によって支援する方法を提案する。まず、二値変数の割合を反映する Modular Facet の当てはめを行う。さらに Guttman Scale の構成度を利用した Polar Facet の領域判断を支援する。それによって Radex モデルを探索的に当てはめ、複数の一次元尺度を構成する変数群を視覚的に抽出する方法 (Miyadera, 2014) を提案する。次節以降では、Radex モデルを構成する Modular Facet、Polar Facet の順に整理し、結果の解釈を支援する分析手法について数式による定式化を行う。

4.2 Radex モデルにおける Modular Facet の探索方法

Modular Facet は同心円の区分線によって分割される領域モデルである。その構造を SSA の結果から把握するには、まず同心円の中心を決定する必要がある。犯罪行動に対する Canter (2000) のモデルを利用した先行研究 (Hakkanen et al., 2003; Santtila, et al., 2003) にみられるように、割合が高い変数ほど同心円の中心近くに布置され、割合の低い変数ほど周縁部に布置される。この性質を利用すると、各二値変数の割合と同心円の中心から各変数までの距離は単調減少の関係を示す。そこで、そのような条件を満たす同心円の中心の位置を探し、Modular Facet の領域モデルを当てはめる。

二値変数 i の割合を p_i ($i = 1, \dots, M$) とし、取り得る値の範囲を $0 \leq p_i \leq 1$ とする。ここで、サンプルサイズを n 、変数 i の頻度を n_i とすれば、 $p_i = n_i/n$ である。このとき、Modular Facet の同心円の中心には $p_i = 1$ の点が布置される。この中心から変数 i までの距離を r_i とする。半径 r_i に対する割合 p_i の関数 h について、x 軸に半径、y 軸に割合をとって表現する。関数の形状を制御するパラメータを ψ とするとき、関数 $h = h(r; \psi)$ が r に対して単調減少で値域が $[0, 1]$ であれば、どのような関数でもよい。このとき、変数 i を表す点の座標を (x_i, y_i) 、同心円の中心の座標を (x_0, y_0) とすると、 r_i は (x_0, y_0) の関数として、

$$[4.2.1] \quad r_i = r_i(x_0, y_0) = \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2}$$

と表される。このとき、次のような目的関数 L を定義する。

$$[4.2.2] \quad L(x_0, y_0, \psi) = \sum_i (p_i - h(r_i; \psi))^2.$$

この目的関数を最小化するパラメータ x_0, y_0, ψ を求めることによって、同心円の中心を決定することができる。

4.3 Radex モデルにおける Polar Facet の探索方法

Polar Facet はある点から放射状に広がる半直線によって分割する領域モデルである。その構造を SSA の結果から把握するためには、中心から伸びる半直線によって区切られる楔形の領域によって、布置される点の属性が異なることを確認する必要がある。ここでは、前節で求めた同心円の中心から周縁方向に向かう点列における Guttman Scale の構成度を利用して、点の属性を弁別する。それによって一次元尺度を構成する変数群の分類が可能となるとともに、それら変数群の領域を分割することで Polar Facet を当てはめることができる。同心円の中心から周縁方向に布置される変数間で Guttman Scale の構成度を計算し、周囲に対して相対的に構成度の低い部分を領域分割の位置と見なせば、区分線を策定できる。この状況を図 4-3 に模式的に示す。

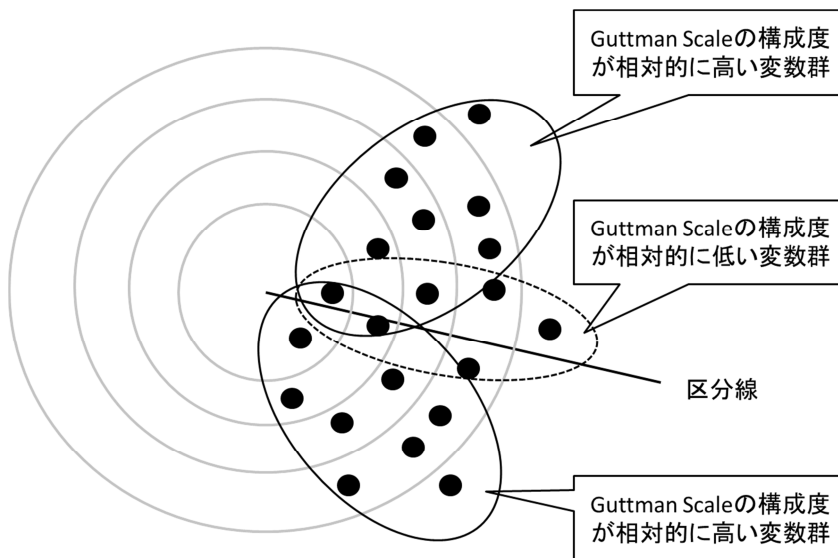


図 4-3 Guttman Scale の構成度から Polar Facet の領域を形成する区分線の策定方法

ここでは、第2章で説明した Loewinger の Homogeneity 係数 H_{ij} を利用する。SSA の結果に布置される変数群について、変数 i 及び変数 j の Homogeneity 係数 H_{ij} を求めることができる。式[2.3.9]によって2変数の組み合わせ全てについて H_{ij} を計算できる。これらの値が全体的に高いほど Guttman Scale の構成度の高い変数群であるといえる。そこで、複数の変数群に対して Guttman Scale の構成度を求める具体的な方法を以下に説明する。

SSA の2次元の布置に対して、Modular Facet の同心円の中心を原点とし、その原点から伸びる半直線を想定する。この半直線を基準として Homogeneity 係数に重みづけを行う方法を考える。半直線の近傍に位置する変数には高い重みづけを与え、半直線から離れるほど低い重みづけを与えることとする。重みの値は0から1の範囲の値とする。このとき、変数 i に対する重みを w_i として、2変数の全ての組み合わせについて次式のような重みづけされた Homogeneity 係数 H_w を定義する。

$$[4.3.1] \quad H_w = \frac{\sum \sum_{i,j \neq i} w_i w_j H_{ij}}{\sum \sum_{i,j \neq i} w_i w_j} \quad (i = 1, \dots, M; j = 1, \dots, M).$$

これにより H_w が各半直線について計算できることから、この H_w の値の高低を調べることによって、Polar Facet の領域を分ける区分線の位置を示唆することができる。 H_w の値が高い半直線は、その位置を中心に Guttman Scale の構成度の高い変数群を含む領域が楔形に広がっていると考えられる。逆に、 H_w の値が低い半直線は、Polar Facet における領域間の区分線の候補になると考えられる。つまり、 H_w の値の高低が、同心円の中心から放射線状に広がる領域をどのように分割すればよいかを表す指標となる。

次に、重みづけの仕方を考える。各点に対する重みを、各点から半直線までの距離に応じて決めるのが合理的であろう。第3章で説明した Faceted SSA の Angular モデルにおける距離の計算 (Borg & Shye, 1995) にならい、円弧を利用して距離を計算することにする。これは極座標を利用した座標特定の際に、計算が容易となる利点がある。

重みづけの仕方には様々な方法が考えられるが、ここではいくつかの方法を提示する。

(1) 角度を利用した重みづけ

この方法は、極座標表示による点と半直線の偏角の差に基準値 θ_0 ($0 < \theta_0 < \frac{\pi}{2}$) を設定し、偏角の差が基準値以下の場合に重み1を与え、基準値を超える場合は重み0を与えるものである。すなわち、SSA の布置における変数 i の極座標を (r_i, θ_i) で表し、半直線の偏角を $\theta(l)$ とする。 θ_i と $\theta(l)$ の差と基準値 θ_0 の関係に応じて、重み w_i に与える値を次式のように定義する。

$$[4.3.2] \quad w_i = \begin{cases} 1 & \text{if } |\theta_i - \theta(l)| \leq \theta_0 \\ 0 & \text{otherwise} \end{cases}$$

この状況を模式的に図 4-4 に示す。半直線 $\theta(l)$ が、点 i の周囲で基準値 θ_0 の範囲に入る場合に $w_i = 1$ となり、その範囲に入らない場合は $w_i = 0$ となる。ただし、基準値 θ_0 以下である半直線の近傍の範囲内に布置された変数が 1 つしかない場合、 w_i の全てが 0 となるため式[4.3.1]の分母が 0 となり、 H_w が計算できない。その場合は、 $H_w = 0$ とする。

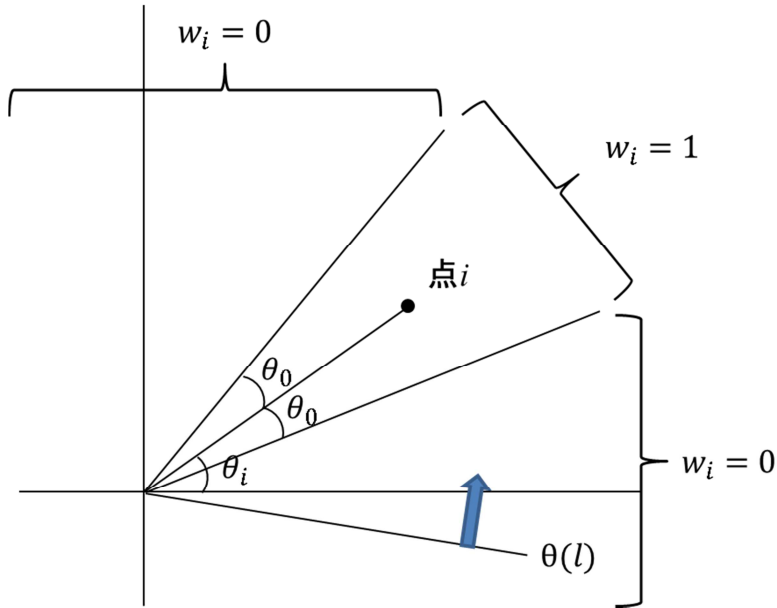


図 4-4 角度による重みづけの模式図

(2) 距離の逆数による重みづけ

この方法は、変数の点から半直線までの円弧による距離の逆数を利用する方法である。点 i から半直線までの距離は、点 i の動径 r_i 及び点 i の偏角 θ_i と半直線の偏角 $\theta(l)$ との差によって $r_i|\theta_i - \theta(l)|$ と表される。この距離を用いて、変数 i に対する重み w_i を考える。SSA の結果において、Modular Facet の同心円の中心を原点として、布置される全変数の点を含む円の最大半径を r_{\max} とする。この r_{\max} を用いて点 i の動径 r_i を基準化すると、

$$[4.3.3] \quad 0 \leq \frac{r_i}{r_{\max}} \leq 1$$

となる。また、点 i の偏角 θ_i と半直線の偏角 $\theta(l)$ との差は次式のように基準化が可能である。

$$[4.3.4] \quad 0 \leq \frac{|\theta_i - \theta(l)|}{\pi} \leq 1$$

これらから、 $r_i|\theta_i - \theta(l)|$ は次のように基準化ができる。

$$[4.3.5] \quad 0 \leq \frac{r_i|\theta_i - \theta(l)|}{\pi r_{\max}} \leq 1$$

この基準化された距離の逆数を利用して、次式のように重みづけを定義する。

$$[4.3.6] \quad w_i = \frac{1}{a \frac{r_i|\theta_i - \theta(l)|}{\pi r_{\max}} + 1}.$$

ただし、 a はパラメータであり、重みは $0 \leq w_i \leq 1$ であることから $a > 0$ とする。パラメータ a の値はどのような値でもよいが、基準化した距離の最大値 1 をとるときに $w_i = 0$ であることが望ましいため、 a の値は十分に大きい方がよい。

(3) ロジスティック関数を利用した重みづけ

パラメータを伴った関数を利用して重みづけを行う方法を提案する。重み関数として次のロジスティック関数を利用する方法がある。単純化するためにパラメータ数を最小にして、

$$[4.3.7] \quad w(x; a, b) = \frac{1}{1 + \exp(a(x-b))}.$$

ここで x は円弧による距離であり、項目 i に対しては $x = r_i|\theta_i - \theta(l)|$ を代入する。 x の値が減少するに伴い w の値が 1 に近づき、 x の値が増加するに伴い w の値が 0 に近づくことが望ましい。この条件を満たすようにするため、式[4.3.7]におけるパラメータ a は $a > 0$ とする。パラメータ a の値が大きいほど w の値の変化は急激になる一方、パラメータ a の値が小さいほど w の値の変化は穏やかとなる。このことは、パラメータ a の値を十分に大きく設定することによって、式[4.3.7]を用いた重みづけは式[4.3.2]による角度を利用した重みづけと等価になることを意味する。これは、このロジスティック関数の変曲点 $x = b$ の近傍で重み w が 1 から 0 に切り替わることを意味する。式[4.3.7]から、

$$[4.3.8] \quad \frac{dw}{dx} = \frac{-a \exp(a(x-b))}{(1 + \exp(a(x-b)))^2}$$

であることから、変曲点 $x = b$ における傾きは $-\frac{a}{4}$ であり、パラメータ a の値が十分に大きいとき、その傾きは

$$[4.3.9] \quad \lim_{x=b} \lim_{a \rightarrow \infty} \frac{dw}{dx} = \lim_{a \rightarrow \infty} -\frac{a}{4} = -\infty$$

となり、 x に対して急激に減少する。さらに任意の $\varepsilon (> 0)$ を考え、 $x = b \pm \varepsilon$ における重み w を考える。

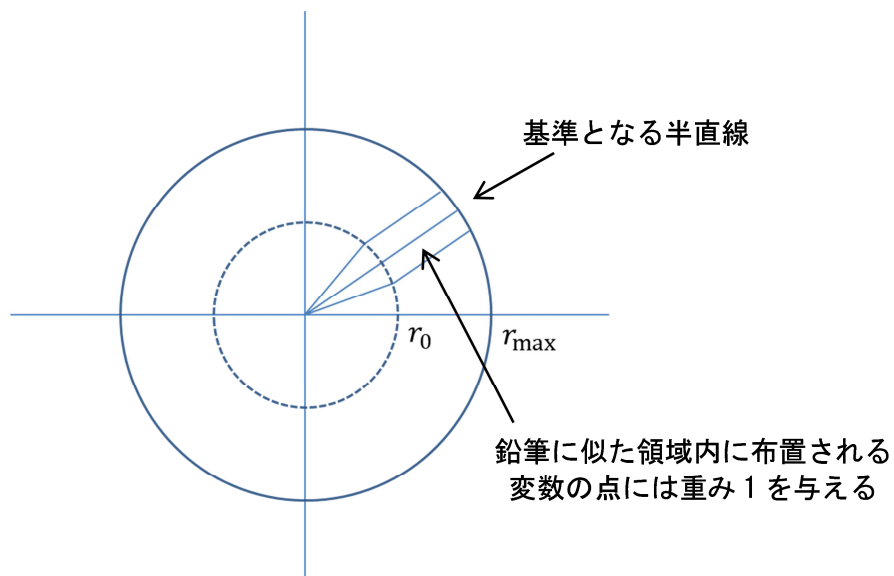
$$[4.3.10] \quad \begin{cases} \lim_{a \rightarrow \infty} w = \lim_{a \rightarrow \infty} \frac{1}{1+\exp(a\varepsilon)} = 1 & \text{if } x = b + \varepsilon \\ \lim_{a \rightarrow \infty} w = \lim_{a \rightarrow \infty} \frac{1}{1+\exp(-a\varepsilon)} = 0 & \text{if } x = b - \varepsilon \end{cases}$$

であることから、 $x = b$ の近傍で重み w が 1 から 0 に切り替わることになる。さらに、式 [4.3.7] におけるパラメータ b の値を変化させることによって半直線からの近傍の範囲を調節することが可能となり、 b の値を大きくすることによって半直線から離れた位置に布置された点に対してもより大きな重みづけを行うことができる。 $b = r_i \theta_0$ とし、 a を大きくとっておけば式 [4.3.7] の重みづけは式 [4.3.2] による角度を利用した重みづけとほぼ同じといえる。

ところで、このロジスティック関数を利用した重みづけにおいて $b = r_i \theta_0$ として a を大きくとった場合、及びそれとほぼ同じ範囲の点を重みづける角度を利用した重みづけ（式 [4.3.2]）を利用する場合では、布置において重みづけられる領域は Polar Facet によって区切られる楔形領域となる。この楔形領域の周縁部に布置している変数は、同心円の中心からの距離が大きい。そのため、円弧で求められる半直線からの距離が非常に大きくなってしまふという問題点がある。そこで、同心円の中心から離れた点に対して次のような工夫を行うことにする。

半直線との角度差が基準値 θ_0 の範囲内にあり、かつ半直線からの距離がある一定値以内に布置される変数の点に対して重みを与えることを考える。ロジスティック関数のパラメータ a の値を十分に大きく設定することで重みの変化が極端に変わるが、その状況を模式的に表したものを図 4-5 に示す。

(a)



(b)

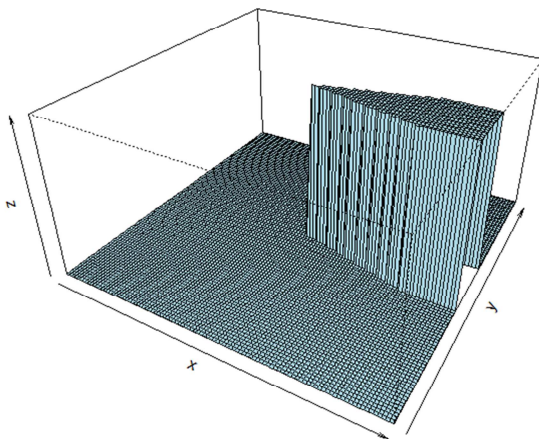


図 4-5 重み 1 を与える変数の点が布置される領域を定義する模式図。(a) SSA の結果の 2 次元平面、(b) SSA の結果の 2 次元平面に対して重みを z 軸にとった 3 次元の俯瞰図。

図 4-5 の状況をなす重みづけの方法を以下に示す。先の距離の逆数による重みづけにおけるのと同様に、SSA の結果において Modular Facet の同心円の中心を原点として、布置される全ての変数の点を含む円の最大半径を r_{\max} とする。これは、原点からもっとも離れて布置された変数の点の原点からの距離である。ここで、 λ ($0 \leq \lambda \leq 1$) を用いて、 r_0 を次式のように定義する。

$$[4.3.11] \quad r_0 = \lambda r_{\max}.$$

最大半径 r_{\max} の線分について、半径 r_0 の部分から半直線近傍の距離を一定にすることで、図 4-5 に示すような鉛筆型の領域を指定できる。これは、ロジスティック関数のパラメータ b を次のように定義することで可能となる。

$$[4.3.12] \quad b = \begin{cases} r\theta_0 & \text{if } 0 \leq r < r_0 \\ r_0\theta_0 & \text{if } r \geq r_0 \end{cases}.$$

以上のように定義されたロジスティック関数に対して、 θ_0 及び λ を指定することで、1 に近い値の重みを与える半直線近傍の範囲を自由に調整することができる。

ここまでに定義した 3 種類の重みづけの方法によって H_w を算出し、偏角 $\theta(l)$ ($-\pi \leq \theta(l) \leq \pi$) の半直線に対する H_w の値の高低を調べることで、Polar Facet の領域区分を検討することが可能となる。その際、 H_w の値の極小値が領域の区分線の候補となる。そこで、以下に第 3 章 5 節に示した人工データ及び調査データに対してこれまでに説明した手法を適用し、その有用性について検証する。

4.4 人工データへの適用

本節では、前節までの Radex モデルの探索支援手法を人工データに適用する。人工データは第 3 章 5 節で説明した 14 変数からなる人工データ 2 を利用する。変数 1～7 の間では Guttman エラーは生じないことから、 $H_{ij} = 1$ ($1 \leq i \leq 10, 1 \leq j \leq 10$) となる。これは変数 8～14 についても同様である。

犯罪行動の分析（例えば、Canter, 2000; Canter et al., 2003）を参考にして、変数間の連関性に Jaccard 係数（Jaccard, 1908）を使用して SSA を適用する。ただし、準単調性による順位写像変換を用いた。ここで、Jaccard 係数を利用する理由について言及する。犯罪行動データの収集方法はいろいろあるが、Canter (2000) のデータは犯罪者プロファイリングを目的としていることから、警察のデータを利用している。警察のデータは、あくまで事件の立件を目的とした捜査のために収集されるものであるため、被疑者に対して直接尋ねた結果として得られるものではない。また、被疑者に対して取調を実施するとしても、被疑者

は裁判の証拠となる可能性を考えて正直に全てを話すとは限らない。つまり、ある行動特徴が把握されれば「該当あり」となるが、把握されない特徴は「該当なし」とは言い切れない、という二値データである。そのため、2変数のクロス集計において非該当同士のセル（表 2-2 のクロス集計表におけるセル度数 d ）を使用しない連関係数として、Jaccard 係数がよく利用される（Alison et al., 2001）。

人工データ 2 に SSA を適用した結果は図 3-11 に示すとおりである。変数 1～7 及び変数 8～14 の点列が右と左それぞれに布置され、双方が中心から反対方向に伸びている様子がわかる。このとき、もっとも割合の高い変数 1 及び変数 8（割合 80%）が図の中心に布置され、割合の低い変数が外縁方向に並び、もっとも割合の低い変数 7 及び変数 14（割合 10%）が外縁部に布置されている。つまり、割合を反映した Modular Facet の存在が見られる。図 4-6 は同じ SSA の結果であるが、変数 1 及び変数 8 の中心付近を原点とするように変更されている。

(a) 変数 1 と変数 8 の中点を原点とした場合 (b) 最適化によって原点を求めた場合

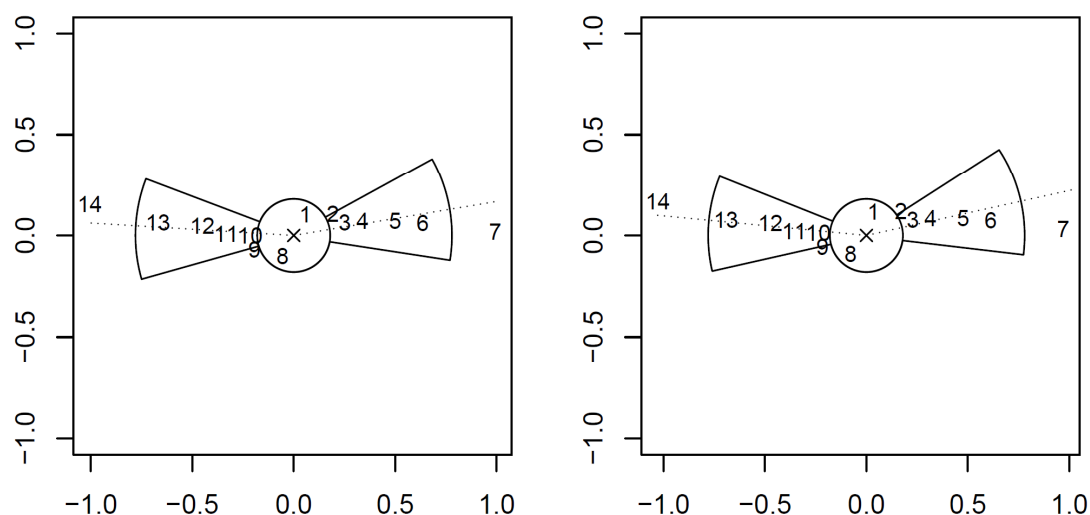


図 4-6 14 変数の二値データに SSA を適用した結果（CoA = 0.0056）。(a) 変数 1 と変数 8 の中点を原点とした場合。(b) 最適化によって原点を決めた場合。ただし、 H_w の産出には角度（基準値 $\theta_0 = \pm \pi/12$ ラジアン）による重みづけを用いた。

そこで、変数 1 及び変数 8 の 2 点の中心を暫定的に Modular Facet の同心円の中心とみなし、そこを原点とした場合（図 4-6 (a)）をみることにする。原点から各変数の点までの距離と各二値変数の割合との関係を表したグラフを図 4-7 (a) に示す。変数 1 と変数 8 の中点を暫定的に原点として、各二値変数の割合と原点からの距離との関係をプロットした点を実線で結んだ。Modular Facet の同心円の中心が適切にとられていれば、グラフは右下がりの単調減少を示すはずである。しかし、割合 60% の変数と 50% の変数の部分で、単調減少の関係が成立していない。そのため、変数 1 と変数 8 の中点を同心円の中心にとった場合は、Modular Facet の当てはめとしては問題があることがわかる。

そこで式[4.2.2]で定義される目的関数に対して、準ニュートン法による最適化によって目的関数を最小化する同心円の中心を求めた。式[4.2.2]では関数 h が具体的には定義されていない。図 4-7 に示される点のプロットの様子から、同心円の中心からの距離と各二値変数の割合との間に概ね単調減少の関係が認められることから、ここでは関数 h をパラメータ c と k を用いて、 $h(r; c, k) = \frac{c}{r^k}$ と定義し、次式によって定義された目的関数 L に対して最適化を行うことにする。

$$[4.4.1] \quad L = \sum \left(p_i - \frac{c}{(r_i)^k} \right)^2$$

ここで c と k は任意の値とする。さらに、同心円の座標を (x_0, y_0) 、変数 i の座標を (x_i, y_i) として、式[4.2.1]を再掲すると、

$$r_i = r_i(x_0, y_0) = \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2}$$

であった。よって、この目的関数 L の最小化は、 (x_0, y_0, c, k) をパラメータとして準ニュートン法による最適化を用いることで達成できる。

目的関数 L の最適化によって同心円の中心を求める手法を用いた結果を図 4-6 (b) に示す。目視ではほとんどわからないが、図 4-6 (a) の原点（SSA の結果における原点を基準としたとき、 $(x_0, y_0) = (0.006085, -0.895680)$ ）に比較して、図 4-6 (b) の原点（SSA の結果における原点を基準としたとき、 $(x_0, y_0) = (0.022918, -0.012629)$ ）の位置は x 軸正の方向及び y 軸正の方向に若干だが、ずれている。この最適化によって求められた原点を用いて、各二値変数の割合と原点から各変数までの距離との関係を図示したものを図 4-7 (b) に示す。これをみるとわかるように、最適化によって求めた原点を利用した場合、各二値変数の割合と原点から各変数までの距離との間には完全な単調減少の関係がみられる。このことから、最適化によって Modular Facet の同心円の中心を求めることで、適切な中心を設定できることが示された。

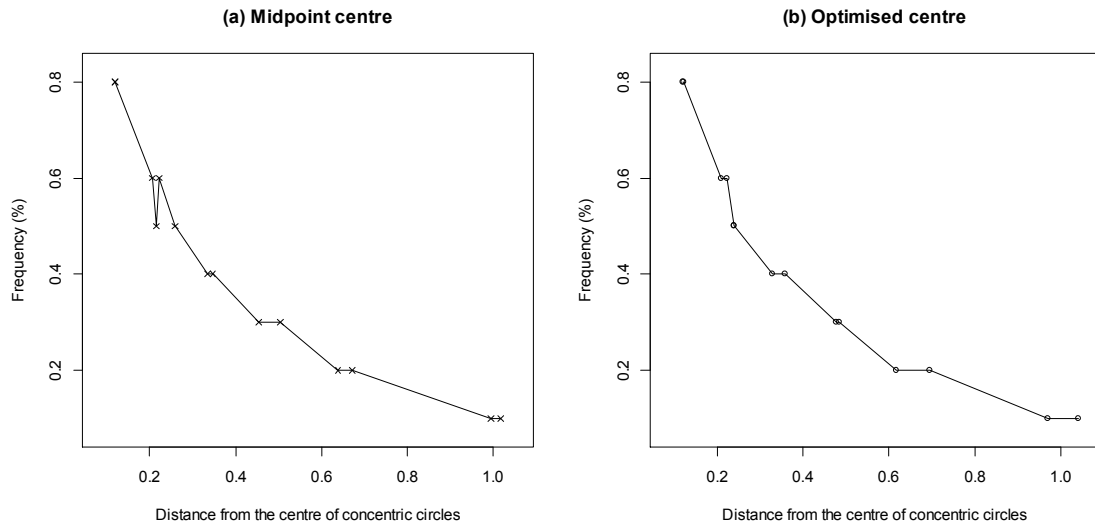


図 4-7 各二値変数中心からの距離と変数の割合の関係。(a) Midpoint centre : 変数 1 と変数 8 の中点を原点とした場合、(b) Optimised centre : 最適化によって求めた点を原点とした場合。

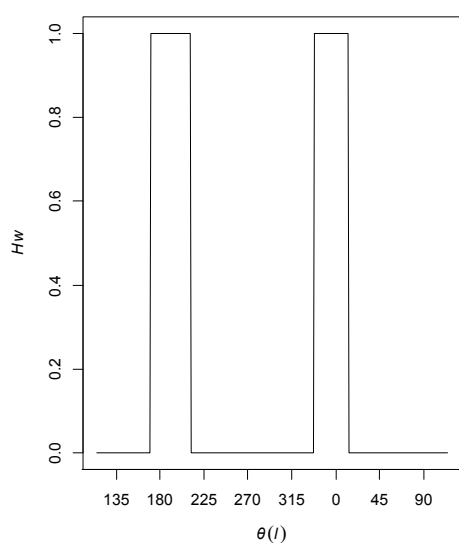
次に、最適化によって求めた原点を中心として、半直線を原点の周囲に反時計回りに 2π ラジアン回転させる。ただし、 $\theta(l) = 0$ を x 軸負の方向とする。偏角の値を変えた半直線のそれぞれについて、式[4.3.1]によって H_w を求める。先に示した 3 種類の重みづけの方法によって H_w を計算した結果を図 4-8 に示す。ただし、基準値 $\theta_0 = \pi/12$ ラジアンとし、パラメータについて距離の逆数による重みづけでは $a = 10,000$ 、ロジスティック関数を利用した重みづけでは $a = 1$ 及び $a = 10,000$ として計算した。

図 4-6 に示される SSA の結果から、変数 1~7 が布置される x 軸正の方向 ($\theta(l) = 190^\circ$ 付近)、変数 8~14 が布置される x 軸負の方向 ($\theta(l) = 360^\circ$ 付近)に半直線があるときに、 H_w は相対的に高い値を示すはずである。図 4-8 をみると、実際、半直線の偏角 θ_l について $\theta_l = 190^\circ$ 及び $\theta(l) = 350^\circ$ の部分で相対的に H_w が高い値を示している。このことは異なる 3 種類の重みづけのどの方法でも共通している。ただし、図 4-8 (c) ロジスティック関数を利用した重みづけにおいてパラメータ $a = 1$ のグラフでは、 H_w の値の高い部分が幅広く分布し、Homogeneity 係数の高い変数群の位置を適切に特定できていない。そのため、ロジスティック関数を利用した場合、パラメータ a は十分に大きい方が望ましいと思われる。以上から、半直線の偏角 $\theta(l)$ が 190° 及び 360° の付近で Guttman Scale を構成する変数群が存在することが示された。

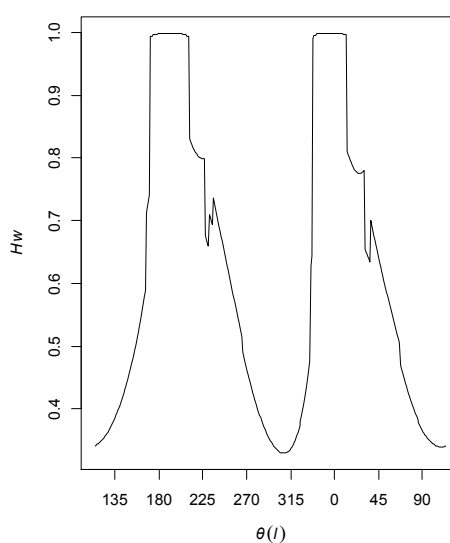
Guttman Scale を構成する変数の位置をより直感的に把握するには、 H_w の値を SSA の結果の原点の周囲に重ねて表示した方がよいと考えられる。そこで、 $H_w = 0$ の基準線を示す参照円を Modular Facet の同心円中心の周囲に描き、 H_w の値の増減を原点の周りに描いた

ものを図 4-6 (b) に示す。なお図 4-6 (a) は暫定的に変数 1 と変数 11 の中点を原点に設定したグラフであるが、同様の手法で H_w の増減を原点の周りに表示した。ここで、利用した重みづけの手法は角度を利用した重みづけ（基準値 $\theta_0 = \pm\pi/12$ ラジアン）である。図中において原点から伸びる波線の半直線は、 H_w の極大値を示す半直線の位置を示している。図 4-6 において、蝶ネクタイのような形状の部分が H_w の値が相対的に高い位置を示しており、 H_w の極大値を示す波線の半直線がその中心を通過して 2 本観測されることから、その近傍に Homogeneity 係数の高い変数群が布置されていることを確認できる。

(a) 角度を利用した重みづけ



(b) 距離の逆数による重みづけ ($a = 10,000$)



(c) ロジスティック関数を利用した重みづけ ($a = 1$ (左)、 $a = 10,000$ (右))

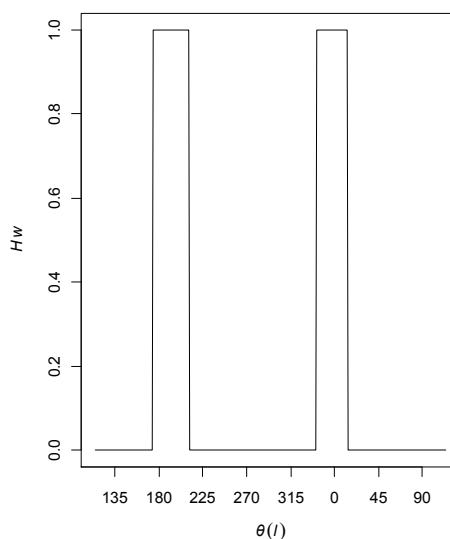
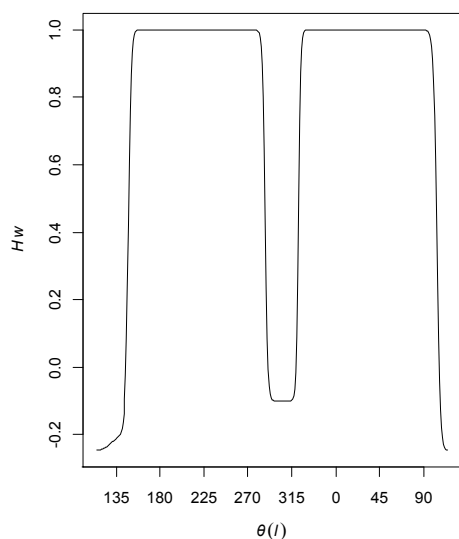


図 4-8 3 種類の重みづけの方法による H_w の値（各グラフの横軸は半直線の偏角 $\theta(l)$ で単位は度数による）

4.5 メディア暴力調査データへの適用

次に、SSA の結果に対する Radex モデルの探索支援手法を実際の調査データであるメディア暴力調査データ（17 変数）に適用した結果を検討する。

メディア暴力調査データに SSA を適用し、2 次元の布置を得た。その 2 次元の結果に対して Modular Facet の同心円の中心を式[4.4.1]に示す目的関数の最適化によって求めた。結果を図 4-9 に示す。このグラフに示された二値変数の割合を見ると、割合の高い変数ほど原点の近くに、割合の低い変数ほど原点から遠位に布置されていることがわかる。原点を中心とした同心円を描くことで、割合を反映する Modular Facet をみることができよう。

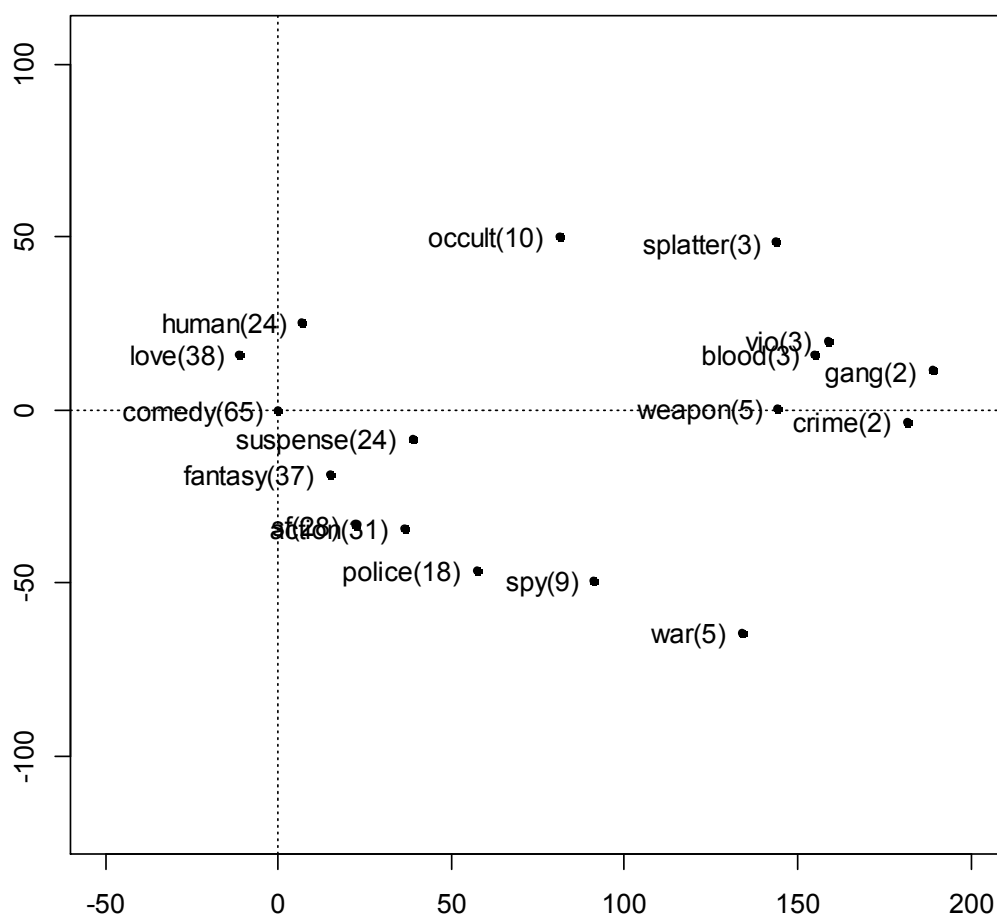
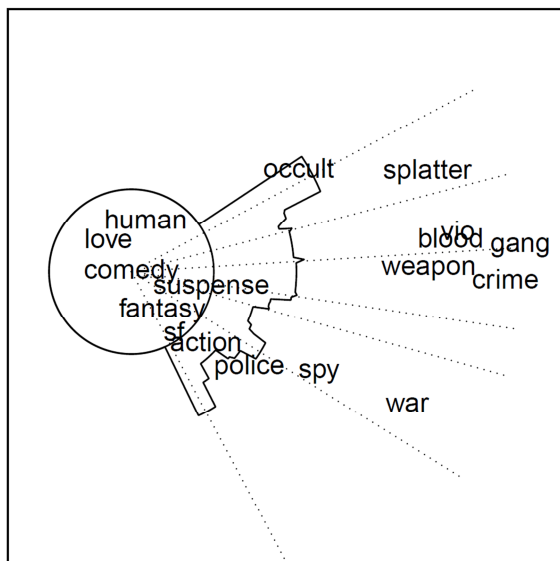


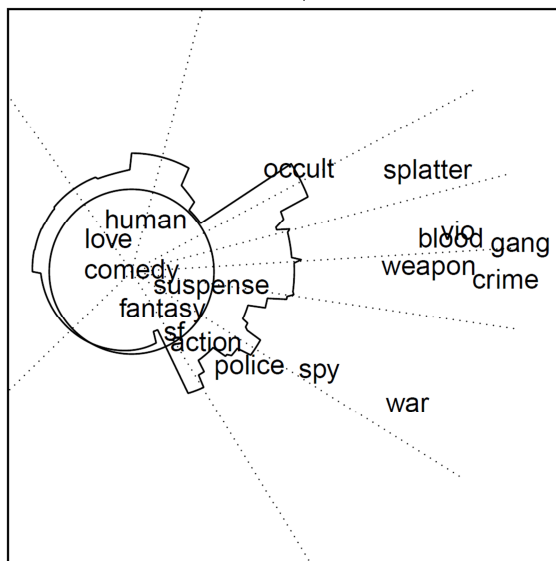
図 4-9 メディア暴力調査データ (17 変数) に適用した SSA の結果 (2 次元、CoA = 0.115130)

原点の周囲に半直線を回転させ、原点の周りに参照円を描き、それに沿って H_w の増減を示したグラフを図 4-10 に示す。重みづけの種類は 3 通りの方法それぞれを利用し、パラメータについて、距離の逆数による重みづけでは $a = 10,000$ 、ロジスティック関数を利用した重みづけでは、 $a = 1$ 及び $a = 10,000$ の 2 種類を実施した。

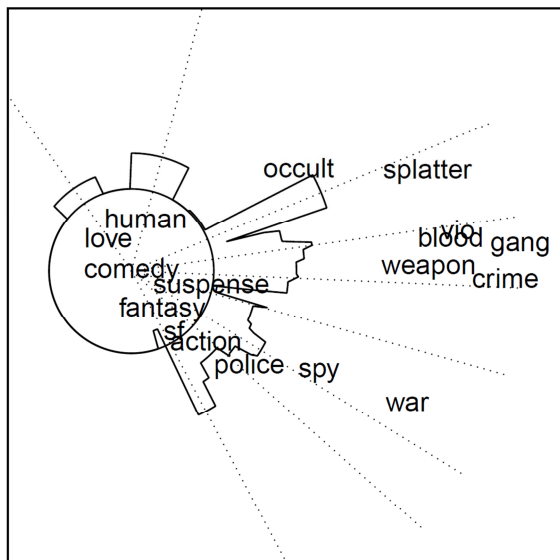
(a) 角度を利用した重みづけ



(b) 距離の逆数による重みづけ
(パラメータ $a = 10,000$)



(c) ロジスティック関数を利用した重みづけ (パラメータ $a = 10,000$)



(d) ロジスティック関数を利用した重みづけ
(パラメータ $a = 1$)

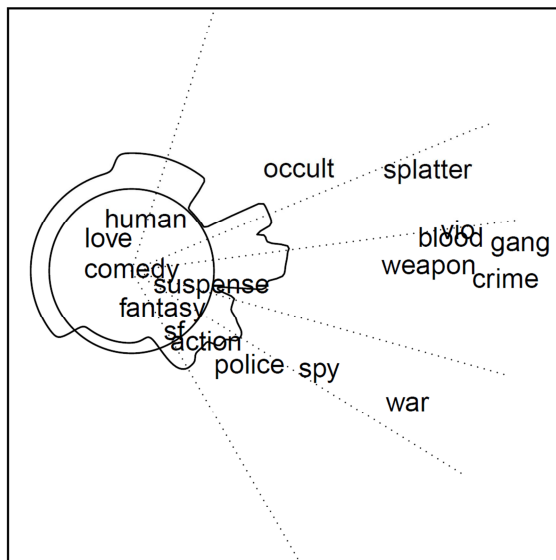


図 4-10 重みづけ Homogeneity 係数 (H_w) の値を重ねて表示した SSA の結果
(ただし、各変数の位置はラベルの中央となる)

布置された各変数の意味を見ながら、目視によって主観的に変数の分類を吟味すると、概ね 4 つの領域に分けられる。一つ目は、SF (sf)、アクションもの (action)、警察・スパイ (police、spy) ものである。二つ目は、暴力的メディアに関連する変数群で、具体的には武器が描写される映画 (weapon)、流血シーンのあるもの (blood)、暴力的なシーンを含むもの (vio)、犯罪や反社会的組織がえがかれたもの (crime、gang) である。三つ目は、オカルト (occult)、スプラッター (splatter) のホラーものである。ただし、この二つ目と三つ目の領域は併せて一つの領域を形作る可能性もある。これら 3 領域が主要な領域群である。最後の四つ目は、恋愛もの (love)、ヒューマンドラマ (human) の変数であり、前者 3 つの領域からは離れて布置されている。

そこで、算出された H_w の値を参照しながら、再度結果をみることにする。原点周囲の参照円の周りに、 H_w の値の水準が描かれている。 H_w の値が高いほど、その部分に沿って原点から伸びる半直線方向に布置された変数群は Guttman Scale の構成度が相対的に高くなり、Polar Facet の領域として等質性が高いとみなすことができる。

図 4-10 (a) では、 H_w の極大値の部分を示す波線が 7 本みられる。しかし、大まかに把握すれば H_w のピークは 4 つである。 y 軸負の方向にみられるピークから反時計回りの順に、各変数のクラスターは次のようになる。

- ・ ファンタジーもの (fantasy)、SF もの (sf)
- ・ 警察もの (police)、スパイもの (spy)、戦争もの (war)
- ・ 暴力的メディアの諸特徴 (武器の使用 (weapon)、流血 (blood)、暴力 (vio)、犯罪 (crime)、反社会的組織 (gang))
- ・ オカルトもの (occult)、スプラッターもの (splatter)

これらの 4 クラスターのうち、オカルトの変数群は H_w の値が最も高い。しかし、図 4-10 (a) では暴力的メディアの諸特徴に対応する H_w のピークが明瞭ではない。回転させる半直線の周囲で重みづけを行う幅を表す角 θ_0 を調整することで、 H_w の値の高低の形状をより明瞭にすることができるかもしれない。また、このグラフでは恋愛もの、ヒューマンドラマに対応する H_w のピークが観測されなかった。

そこで、回転させる半直線の周囲で重みづけを行う幅を表す角 θ_0 を 3 種類 ($\theta_0 = \pm 7.5^\circ, \pm 15^\circ, \pm 22.5^\circ$ 、すなわち半直線の周囲 $15^\circ, 30^\circ, 45^\circ$ の範囲) として重みづけ Homogeneity 係数 (H_w) の値を重ねて表示した SSA の結果を図 4-11 に示す。

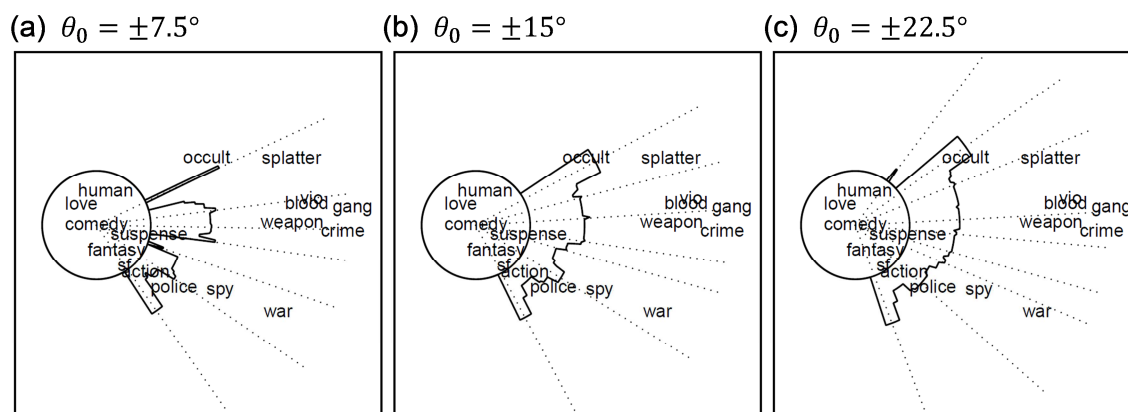


図 4-11 メディア暴力調査データ (17 変数) の SSA に対して、角度による重みづけによって H_w を計算する際に利用するパラメータ (θ_0) の値を変えて検討した結果 (ただし、各変数の位置はラベルの中央となる)

図 4-11 (a) を見るとわかるように、回転させる半直線で重みづけを行う幅を表す角 θ_0 を小さくすることによって、 H_w のピークがより峻別しやすくなり、Polar Facet の区分線を引きやすくなる。一方、(c) を見ると H_w のピークが一層曖昧になっており、角 θ_0 を大きくすると H_w の値の高低の形状が判別しにくくなり望ましくないことがわかる。

再び図 4-10 の結果に戻る。距離の逆数による重みづけの手法である図 4-10 (b) は、(a) に比べて参照円周囲の H_w ピークの形状はほぼ同じであり、さらにヒューマンドラマ (human) に対応する山が出現している。 H_w の形状は概ね 5 つのピークから構成される。

- ・ ファンタジーもの (fantasy)、SF (sf) (さらにアクションもの (action) を追加できる可能性もある)
- ・ アクションもの (action)、警察・スパイもの (police、spy)、戦争もの (war)
- ・ 暴力的メディア (weapon、blood、vio、crime、gang)
- ・ オカルトもの (occult)、スプラッターもの (splatter)
- ・ ヒューマンドラマ系 (human)

ロジスティック関数を利用した重みづけを用いて H_w の水準を描いた図 4-10 (c) は、パラメータ $a = 10000$ に関して、もっとも明瞭な 6 つのピークを描いている。ただし、些細な極大値を示す 3 つの波線は無視するものとする。 H_w の値の増減を示す形状の全体をみると、6 つの領域があるとみなすことが望ましい。そのうち、恋愛もの (love)、ヒューマンドラマ (human) を除いた 4 つの領域に着目すると、変数クラスターは SF 系 (fantasy、sf)、アクション系 (action、police、spy、war)、メディア暴力系 (weapon、blood、vio、crime、gang)、オカルト系 (occult、splatter) に分けられよう。

また、ロジスティック関数による重みづけについて、パラメータ a の値を小さくした場合、半直線からの各変数の距離に応じて重みづけられる変数が増えることにより、 H_w の値

の変化がより穏やかなものになっている。図 4-10 (d) はパラメータ $a = 1$ とすることで、パラメータ $a = 10000$ である (c) よりも H_w の値の変化が滑らかである。しかし、 H_w の値のピークを識別するのが却って難しくなっている。また恋愛もの (love) を示す山が (d) では消失している。このことは、ロジスティック関数の変曲点における傾きが急であるほうが重みづけとしては望ましいことを示唆していると考えられる。

先の角度による重みづけと同様に、ロジスティック関数を用いた重みづけの手法においてもパラメータの値を変えることで重みづけの仕方を調整することができる。そこで、重みづけ関数のパラメータの値をさまざまに変えたセットを準備し、その重みづけを調査データに適用した。これまでにロジスティック関数で利用したパラメータ値は、 $\theta_0 = \pm 15^\circ$ 、 $\lambda = 0.5$ である。そこで、次の 3 種類の θ_0 ($\theta_0 = \pm 7.5^\circ, \pm 15^\circ, \pm 22.5^\circ$ 、すなわち半直線の周囲 $15^\circ, 30^\circ, 45^\circ$ の範囲) 及び 3 種類の λ ($\lambda = 0.3, 0.5, 0.7$) を掛け合わせた 9 種類のパラメータ値の組み合わせによる重みづけの手法により、調査データを分析した。結果を図 4-12 (a) ~ (i) に示す。

9 つのグラフをみると、変数の実質的意味から示唆される領域を適切に反映したものはグラフ (c) と (e) である。

グラフ (a) にはオカルトの変数群 (occult、splatter) の部分に対応する H_w の山が全くみられない。その山はグラフ (b) にはみられるものの、非常に急峻なピークとなっている。このことから、オカルトの変数領域は不安定であるといえる。この領域にはオカルトに関する変数が 2 つしか含まれておらず、オカルトとスプラッターの変数がお互いに離れて布置されているため、 θ_0 を小さく設定した場合、 H_w の山が出現しなくなるためであると考えられる。このことは、オカルトに関連する変数の領域をみた場合に、グラフ (a)、(d) では対応する H_w の山が存在せず、グラフ (g) では急峻な山がみられることから確認できる。これらのグラフ (a)、(d)、(g) は全て狭い角度幅 $\theta_0 = \pm 7.5^\circ$ で計算されたものである。この角度幅 θ_0 の値が小さすぎる場合、Guttman Scale を構成する度合いが高い変数の組み合わせがあっても、適切に検出できない可能性がある。

グラフ (f) と (i) は、グラフ (c) に比べて H_w の値の谷間が明瞭ではない。特に暴力的メディア特徴の変数群 (weapon、blood、vio、crime、gang) に対応する H_w の山が見分けにくい。そのため、布置される変数のまとまりが Polar Facet の領域を決めるのに十分なレベルの明瞭さを持っていないといえる。グラフ (h) も領域を検出する上では適切ではない。メディア暴力の諸変数が布置される領域部分で、 H_w の示す山が不明瞭であるからである。

以上から、グラフ (c) 及び (e) が H_w の値の山や谷を把握する上でもっとも適切である。このことから、 H_w の値の高低をよりよく検出するためには、 λ として小さい値、 θ_0 として大きな値、あるいはその逆の組み合わせが必要であるが、 θ_0 として小さすぎる値は避けた方がよいことが示唆される。

なお、本章で利用したプログラムを付録 G.4 に示す。

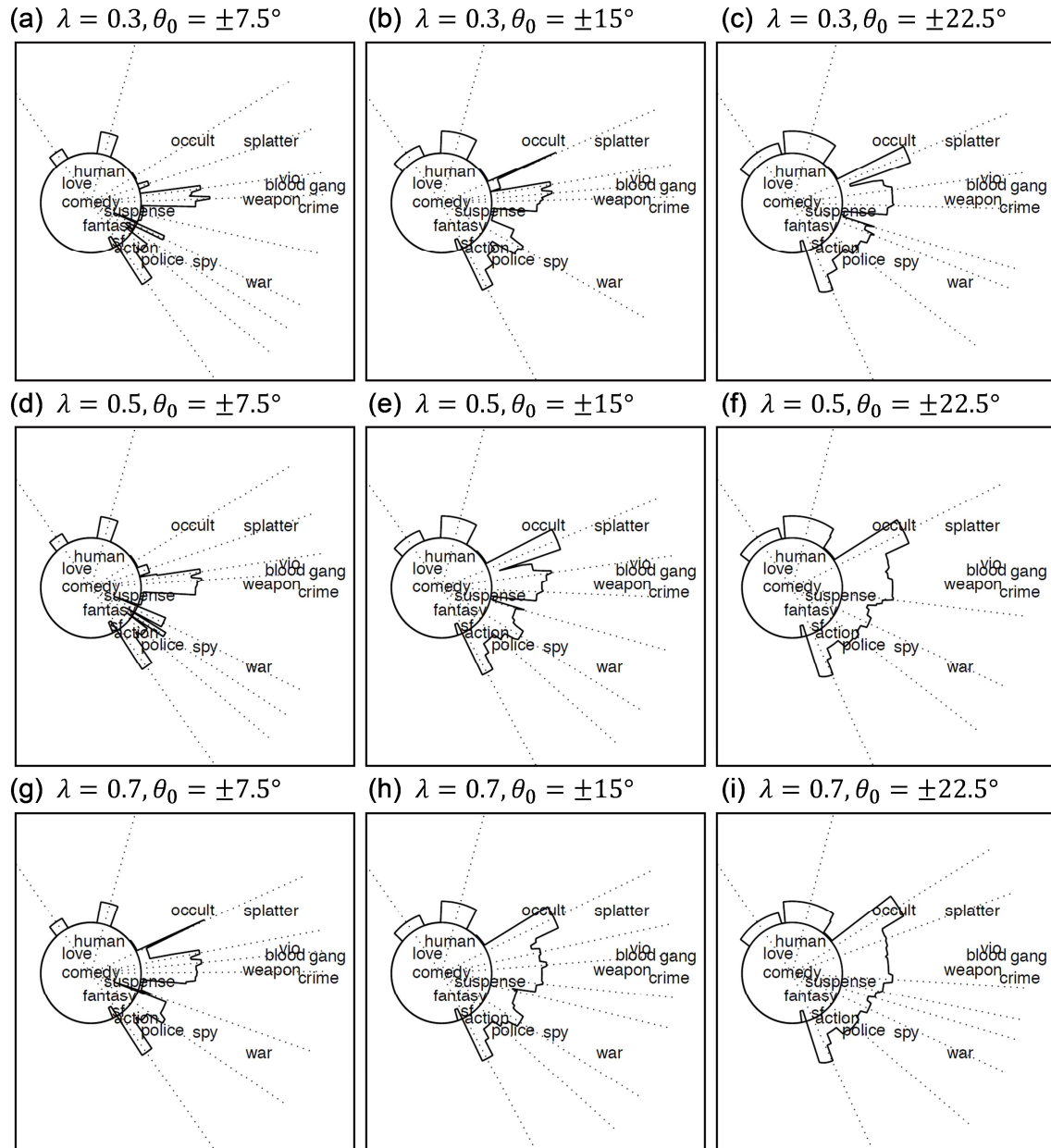


図 4-12 メディア暴力調査データ (17 変数) の SSA の結果に対して、 H_w を計算する際に利用するロジスティック関数のパラメータ (λ 及び θ_0) の値を変えて検討した結果 (ただし、各変数の位置はラベルの中央となる)

4.6 まとめ

本章では、二値で測定された行動科学的データに対して SSA を適用し、得られた 2 次元の結果について Radex モデルを探索する支援手法の提案を行った。人工データと調査データに適用した結果、人工データについては適切に Modular Facet を当てはめることができ、また重みづけ Homogeneity 係数の値を利用することによって適切に Guttman Scale の位置を特定し、Polar Facet を当てはめることができたことから、SSA の布置に対して適切に Radex モデルを当てはめることができたといえる。調査データについても、二値変数の割合を反映する Modular Facet を当てはめることができ、目視によって大まかに分類されていた変数群に対して、Guttman Scale の構成度に応じた適切な分類を示唆することが可能となった。さらに、Homogeneity 係数の重みづけの計算方法として、ロジスティック関数を利用した方法がパラメータの調整を検討する上で有効であり、Polar Facet を示唆する重みづけ Homogeneity 係数の値を算出する際に注意すべきパラメータ値の設定について、有益な示唆が得られたといえる。

データに対するモデルの当てはまりを考える場合、その当てはまりの程度を評価する方法があることが望ましい。Modular Facet については、同心円構造の中心となる点を便宜的に策定するため、SSA の結果に布置される各二値変数の割合と中心から各変数までの距離が単調減少関係になることが一つの指標であることを示した。一方、Polar Facet については重みづけされた Homogeneity 係数の値を用いて領域の区分線を引く場所を特定する手法を提案した。しかし、Polar Facet における領域区分の妥当性は、あくまで SSA の 2 次元平面上に布置される各変数群の組み合わせと変数の実質的意味から検討されるべきであり、当てはまりの程度を統計的に評価することは難しい。特定の連関係数(ここでは Jaccard 係数)を用いて SSA を適用した既存の結果に対して、Radex モデルがみられるかどうかを Homogeneity 係数という別の指標で吟味しているために生じている問題点であるともいえる。

各二値変数の割合及び Homogeneity 係数を用いて、直接的にデータが Radex モデルに当てはまるかどうかを数量的に定式化することによって、そのモデルに対するデータの当てはまりの度合いを客観的に評価することが可能となると考えられる。そこで、次章では Radex モデルへの当てはまりを定式化し、直接的に当てはまりの度合いを評価する手法を提案する。

第 5 章 Radex モデル多重尺度構成法 (RADXMS) の提案

第 5 章では Radex モデルへの当てはまりを定式化し、直接的に当てはまりの度合いを評価する手法を検討する。前章では SSA (最小空間分析) の結果に対して Radex モデルの探索を支援する手法を提案したが、本章では二値によって測定された行動科学的データを構成する変数について、Guttman Scale の構成度を利用して 2 次元平面上または 3 次元空間内に Radex モデルを直接視覚化する手法を提案する。以下、データから直接 Radex モデルを視覚化する本手法を Radex モデル多重尺度構成法 (Radex Multiple Scaling, RADXMS) と呼ぶことにする。

5.1 問題と目的

前章に引き続き、本章においても二値データの変数間の関係から二値変数の割合を考慮して複数の一次元尺度を抽出することが狙いである。その方法として、Guttman Scale の一次元尺度を視覚化することを目的とする。前章同様に、Facet 理論の領域モデルとして Radex に着目し、変数間の関係に Radex モデルを当てはめることを考える。

前章においては、特定の連関係数として Jaccard 係数を利用し、その値によって非類似度を算出し、SSA を適用した結果に対して Radex モデルを当てはめることを目指した。しかし、Radex モデルをデータに当てはめる場合、その適合度の指標が計算できることが望ましい。なぜならば、適合度の指標があれば、複数の適合結果のうちいずれの結果が適合がよいかを比較することができるからである。前章の手法では、Modular Facet の適合については客観的に評価することが可能であったが、Polar Facet の適合については客観的な評価はできなかった。

その一つの理由として、Polar Facet の当てはめが既存の SSA モデルに対する Homogeneity 係数という別の指標を事後的に用いていることが挙げられる。逆に、2 次元の布置を Homogeneity 係数を利用して表現することができれば、Homogeneity 係数を利用した指標によって、適切な Radex モデルの定式化が可能となると考えられる。

本章では、Homogeneity 係数を利用し、直接的に Radex モデルを表現することにより、Guttman Scale の新たな視覚化手法 RADXMS を提案することを目的とする。

5.2 Radex モデル多重尺度構成法の理論的背景

得られた二値データから変数間の関係を Radex モデルによって表現しようとするとき、まず Radex モデルが 2 つの Facet から構成されていることを考慮する必要がある。すなわち同心円の領域に区分される Modular Facet と中心から放射線状に伸びる半直線によって区分される Polar Facet である。そこで、これら 2 つの Facet の特徴を満たすように 2 次元平面上または 3 次元空間内の点の位置を決める方法を説明する。

ここで、2 次元平面上あるいは 3 次元空間内に変数をプロットする方法を考える場合、こ

れら 2 つの Facet を表現するために極座標を用いるのが都合がよい。なぜならば、Modular Facet と Polar Facet を数式で表現する際、それぞれ Radex モデルの中心からの距離、中心周りの偏角を利用することになるからである。以下、変数 i を表す 2 次元平面上の点 i ($i = 1, \dots, M$) の、中心からの距離（または動径）を r_i 、x 軸の正の方向を始線とする偏角を θ_i （ただし、単位はラジアンとする）として、極座標 (r_i, θ_i) で変数 i の位置を表現することとし、 r_i と θ_i の求め方を提案する。なお 2 次元平面上における Radex モデルの説明を行った後、3 次元空間の場合に拡張する方法を提示する。

ここで中心について付言する。第 4 章で記述した Radex モデル探索支援手法 (RADXEA) では、変数を表す点の所与の布置に対して、Radex モデルの中心を事後的に決定したが、本章の方法では Radex モデルの中心を原点とした 2 次元平面または 3 次元空間内に変数を表す点を布置するものであり、Radex モデルの中心は布置を作成する始めの段階から原点として決められている。

5.2.1 Radex モデルにおける Modular Facet の表現

初めに、Modular Facet を表現するための r_i の定め方を説明する。変数が Radex モデルとして表現される場合、割合が同じ変数を表す点は同一の同心円上に布置され、割合が高い変数を表す点はより中心に、割合が低い変数を表す点はより周縁に布置されることが望ましい。そこで各二値変数の割合によって各点の中心からの距離を決めることにする。

Radex モデルの中心を原点とする極座標 (r_i, θ_i) で表される変数 i ($i = 1, \dots, M$) について、その割合を $p_i (0 \leq p_i \leq 1)$ とする。ここで、サンプルサイズを n 、変数 i の頻度を n_i とすれば、 $p_i = n_i/n$ である。このとき、 p_i の値が大きいほど r_i の値が小さくなるようにする。つまり、 r_i が p_i の単調減少関数になるようにする。単調減少の関係を表す関数はいくつか考えられるが、ここでは次のような関係式を用いることにする。

$$[5.2.1] \quad r = r(p) = (1 - p)^\alpha \quad \text{ただし } \alpha > 0$$

これによって、割合 p_i が 1 となる変数は原点に、 p_i が 0 となる変数は半径 1 の単位円上に布置されることになる。ここで α はパラメータであり、 r_i が p_i に関する単調減少関数であることを保証するため、 $\alpha > 0$ とする。また、 α の値を調整することによって p_i と r_i の対応関係を変えることが可能となる。特に、式[5.2.1]を用いて r_i の値を定めることによって、データの分布あるいは二値変数の割合の分布に応じて、各変数の布置の差を拡大もしくは縮小するように各点を配置することが可能となる。

5.2.2 Radex モデルにおける Polar Facet の表現

次に、Polar Facet を表現するための点 i の偏角 θ_i の定め方について説明する。基本的な考え方は、各変数同士の Guttman Scale の構成度を利用して各点の位置を決めるというもの

である。2次元平面では Guttman Scale の構成度の高い変数同士が図の中心から一直線上に並んで配置されるように各点の位置を決める。つまり、完全な Guttman Scale を構成する変数同士は、図の中心から伸びる同一の半直線上に配置される。一方、Guttman Scale を構成する度合いが低い変数同士は、それぞれの点を通る中心からの半直線が互いに離れるように配置される。つまり、完全な Guttman Scale を構成する変数同士では極座標における偏角の差 $|\theta_i - \theta_j|$ が0となるように、一方 Guttman Scale を構成する度合いが低い変数同士では、偏角の差 $|\theta_i - \theta_j|$ が大きくなるように、各点の位置を決定する。2次元平面上で考える場合、もっとも離れた位置とするには、その最大値を π として $|\theta_i - \theta_j| \leq \pi$ とするのが一法であろう。これにより、Guttman Scale の構成度に対応するように各点の位置を決めることが可能となる。

第2章で論じたように、Guttman Scale の構成度を測るには各種の指標が利用できる。例えば、Reproducibility 係数、Scalability 係数、Loevinger の Homogeneity 係数があげられる。Guttman Scale の構成度を測定するために利用する指標としてはいずれでも利用可能であるが、ここでは第2章で説明したとおり、Loevinger の Homogeneity 係数を利用することとする。なぜならば、Reproducibility 係数や Scalability 係数に比べて、Loevinger の Homogeneity 係数は回答集計表の周辺度数に依存する度合いが低い点で優れているとされるからである (Mokken, 1971)。例えば、2項目間の Homogeneity 係数 H_{ij} ($i = 1, \dots, M, j = 1, \dots, M$) において $p_i > p_j$ であるとき、変数 i の正反応数が大きく変数 j の正反応数が小さい場合に、 H_{ij} は負で絶対値の大きい値を取る (付録A参照)。その意味で、Homogeneity 係数が周辺度数に全く影響を受けないとはいえない。しかし、正反応同士のセル度数から、周辺度数によって計算される期待度数を差し引いて係数を算出するため、Homogeneity 係数は Reproducibility 係数や Scalability 係数に比べて周辺度数から受ける影響の度合いが低いといえる。そこで、Guttman Scale の構成度を測る指標として、ここでは Loevinger の Homogeneity 係数 H_{ij} を利用することとする。

平面に布置された2つの変数 i 及び j を極座標 (r_i, θ_i) 及び (r_j, θ_j) で表したとき、それら2変数の間のユークリッド距離を d_{ij} とする。このとき、 H_{ij} の値の大小関係を反映するように d_{ij} の値を決めることによって、2次元平面におけるそれぞれの点の位置を決定する。 H_{ij} の値が大きい変数が Polar Facet における同一の領域内に布置されるようにするためには、 $|\theta_i - \theta_j|$ が小さくなるようにすればよい。特に、 $H_{ij} = 1$ を満たすときは、 $\theta_i - \theta_j = 0$ 、つまり2つの変数 i 及び j が原点から伸びる同一の半直線上に布置されることが望ましい。このように (r_i, θ_i) を決定することによって、2次元平面上に各変数間の Guttman Scale の構成度を反映する Polar Facet を描くことが可能となる。

5.2.3 Homogeneity 係数を変換する関数

続いて、 H_{ij} の値の大小関係に対応する変数間の距離 d_{ij} の取り得る範囲を議論する。変数 i と変数 j それぞれの動径 r_i 及び r_j は前節 5.2.1 の方法によって定まったものと

する。このとき 2 次元平面における 2 変数の各点は、それぞれ動径 r_i 及び r_j の円周上に位置する。この 2 変数が完全な Guttman Scale を構成する場合は $H_{ij} = 1$ となり、 $|\theta_i - \theta_j| = 0$ となるため、2 変数間の距離は $|r_i - r_j|$ となる。逆に、 H_{ij} が低い場合、2 変数 i 及び j は離れて布置されることが望ましく、特に H_{ij} が最小値をとる場合、2 点間の距離が最大となるように布置するには $|\theta_i - \theta_j| = \pi$ とし、このとき 2 変数間の距離は $r_i + r_j$ となる。つまり、2 点間の距離 d_{ij} について次式が成立する。

$$[5.2.2] \quad |r_i - r_j| \leq d_{ij} \leq r_i + r_j$$

このように d_{ij} の範囲は、変数 i と変数 j にそれぞれ対応する r_i 及び r_j によって定まる制約が伴う。

ところで、この場合、従来の多次元尺度法による方法を使って点の布置を決めることができないという問題がある。従来の多次元尺度法では、類似度を関数によって変換した値と 2 次元上の 2 点間の距離 d_{ij} の差の二乗和によって定義される目的関数の最小化を行うことで点の布置を決定する。しかし、二値変数の割合を考慮する本手法では、式[5.2.2]による制約が伴うため、類似度を変換する関数として変数の組み合わせ全てに対して共通する関数を利用することでは対応ができない。例えば、割合の異なる 4 つの変数 i, j, k, l について、 $H_{ij} = H_{kl}$ が成り立っているものとする。ここで、この Homogeneity 係数を変換する関数を g とする。通常の多次元尺度法では g として一つの関数のみが定義され、 $g(H_{ij}) - d_{ij}$ と $g(H_{kl}) - d_{kl}$ が比較されることから、計量的な関数を利用する場合、目的関数の最小化において $d_{ij} = d_{kl}$ が得られる場合もありうる。しかし、式[5.2.2]の制約が伴う場合、 d_{ij} 及び d_{kl} の範囲は $|r_i - r_j| \leq d_{ij} \leq r_i + r_j$ 及び $|r_k - r_l| \leq d_{kl} \leq r_k + r_l$ となる。二値変数の割合が異なることから、必ずしも $|r_i - r_j| = |r_k - r_l|$ かつ $r_i + r_j = r_k + r_l$ となるとは限らない。

そこで、変数の組み合わせごとに Homogeneity 係数を変換する単調減少関数を変更することによってこの問題を解決することとした。つまり、 H_{ij} を変換する関数を g_{ij} 、 H_{kl} を変換する関数を g_{kl} としたとき、必ずしも $g_{ij} = g_{kl}$ とはしないわけである。

次に、この解決方法に関する具体的な手法を説明する。すなわち Homogeneity 係数を変換する単調減少関数 g_{ij} の定義について提案する。 $g_{ij}(h)$ の値域と d_{ij} の範囲を対応させて、

$$[5.2.3] \quad \eta^2 = \sum_i \sum_{i < j} \{g_{ij}(H_{ij}) - d_{ij}\}^2$$

として目的関数 η^2 を定義し、この η^2 を最小化することで d_{ij} を求める。ここで、 r_i 及び r_j の値が固定のもとで（すなわち、周辺度数が固定のもとで）Homogeneity 係数 H_{ij} の最大値 H_{ij}^{\max} は 1 である一方、最小値 $H_{ij|a,\beta}^{\min}$ は、

$$[5.2.4] \quad H_{ij|\alpha,\beta}^{\min} = \begin{cases} 1 - \frac{n}{n-\alpha} & (\alpha - \beta \leq n - \alpha \Leftrightarrow \frac{\alpha}{n} \leq 0.5 + \frac{\beta}{2n}) \\ 1 - \frac{n}{\alpha-\beta} & (n - \alpha < \alpha - \beta \Leftrightarrow \frac{\alpha}{n} > 0.5 + \frac{\beta}{2n}) \end{cases}$$

となる。ただし、変数 i と変数 j のクロス集計表におけるそれぞれの該当度数を α 、 $\alpha - \beta$ とする(付録 A を参照)。そのため、 $\alpha - \beta \leq n - \alpha$ の場合、 $H_{ij|\alpha,\beta}^{\min} < 0$ であり、 $n - \alpha < \alpha - \beta$ の場合、 $n - \alpha < \alpha - \beta < n$ であれば $H_{ij|\alpha,\beta}^{\min} < 0$ となり、負の値を取り得る。なお、 α と β を固定することにより、変数 i と変数 j の該当度数が固定されることから、各変数の該当割合 p_i 及び p_j が固定され、それによって式[5.2.1]によって動径 r_i 及び r_j も固定されることから、 $H_{ij|\alpha,\beta}^{\min} = H_{ij|p_i,p_j}^{\min} = H_{ij|r_i,r_j}^{\min}$ である。

$H_{ij} = 1$ のとき $g_{ij}(h)$ は最小値 $|r_i - r_j|$ をとり、 H_{ij} について定めた何らかの最小値を H_{ij}^{\min} で表すことにすると、 $H_{ij} = H_{ij}^{\min}$ のとき $g_{ij}(h)$ は最大値 $r_i + r_j$ をとることが望ましい。そのため、 $\min(g_{ij}(h)) = |r_i - r_j|$ とする一方、 $\max(g_{ij}(h)) = r_i + r_j$ となるような関数 g_{ij} を考える必要がある。その際、 H_{ij}^{\min} をどの値にするかによって、平面上に布置される点の広がり異なることから、その値をどのように設定するかが問題となる。

例えば、所与のデータにおける $\frac{1}{2}M(M-1)$ 個の $\{H_{ij}\}$ のうち最小の値である H_{data}^{\min} をとったとき、 $g_{ij}(H_{data}^{\min}) = r_i + r_j$ となるよう、関数 g_{ij} を定義するという考え方があろう。この方法は一見すると簡単ではあるが、関数 g_{ij} がデータに依存するという問題点がある。例えば、 $\{H_{ij}\}$ が比較的に高い値ばかりとなった場合、その中で相対的に低い H_{data}^{\min} を示す 2 変数がかもとも離れた位置に布置されることになる。これは H_{data}^{\min} が例えば 0.8 という比較的高い値であったとしても、 $\{H_{ij}\}$ の中で相対的に低い値であるという理由で $H_{ij} = H_{data}^{\min}$ なる変数 i と j を最大に引き離して布置することになるため、合理的ではない。そこで、 $\max(g_{ij}(h))$ に対応する定義域の下限 h_{ij}^{lower} の設定方法として、次の 4 通りを考える。

- ① $h_{ij}^{\text{lower}} = 1 - n = H_{\text{theory}}^{\min}$ (理論上取り得る H の下限値)
- ② $h_{ij}^{\text{lower}} = H_{ij|r_i,r_j}^{\min}$ (変数 i 、 j の正反応固定の下で理論上取り得る下限値)
- ③ $h_{ij}^{\text{lower}} = 0$
- ④ $h_{ij}^{\text{lower}} = \min(0, H_{data}^{\min})$

なお Homogeneity 係数 H_{ij} の最小値の記号について簡単に整理すると

- ・ H_{theory}^{\min} : 理論上の下限値
- ・ $H_{ij|\alpha,\beta}^{\min} = H_{ij|p_i,p_j}^{\min} = H_{ij|r_i,r_j}^{\min}$: 2 変数の正反応数をそれぞれ固定した際の理論上の下限値
- ・ H_{data}^{\min} : 所与のデータから計算される $\{H_{ij}\}$ のうちの最小値

となる。

①はサンプルサイズが大きくなるほど $H_{\text{theory}}^{\min} = 1 - n$ の絶対値も大きくなり、実データにおける $\{H_{ij}\}$ の最小値 H_{data}^{\min} と理論上の最小値 $H_{\text{theory}}^{\min} = 1 - n$ の差が非常に大きくな

る可能性がある。この場合、実際のデータにおける $\{H_{ij}\}$ の分布を十分に広げて 2 次元平面に点を布置することができず、望ましくない。

②は、①ほど実データにおける $\{H_{ij}\}$ の最小値 H_{data}^{\min} と理論上の最小値 H_{theory}^{\min} の差が大きくなる可能性がある。しかし、式[5.2.4]において、

$$[5.2.5] \quad \alpha = n - 1 \quad \text{かつ} \quad \beta \geq n - 2$$

または

$$[5.2.6] \quad \alpha = \beta + 1 > n - 1$$

のとき①と同じ状況が発生するという問題点がある。また、変数の組み合わせ (i, j) ごとに Homogeneity 係数の下限値が異なるため、 $\{H_{ij}\}$ における Homogeneity 係数の順序が $g_{ij}(H_{ij})$ によって変換した後に保たれないことから、慎重に検討する必要がある（付録 D）。

③は一つの解決法である。ただし、 $H_{\text{data}}^{\min} < 0$ が生じた場合の対処が必要になる。経験的に $H_{ij} < 0.3$ は Guttman Scale として意味をなさないという指摘（van Schuur, 2011）があることから、負の値をとる H_{ij} については全て $H_{ij} = 0$ とすることも可能ではある。しかし、負の値をとる H_{ij} の絶対値が非常に大きくなる場合、それらを全て $H_{ij} = 0$ とすることは、 H_{ij} の値の大小関係を適切に反映していないという懸念がある。

④は、負の値をとる H_{ij} が生じた場合のデータに対処する方法であり、いわば折衷案である。しかし、データに応じて適切に $g_{ij}(h)$ の最大値を取ることができ、それによって定まる d_{ij} の最大値を適切に選ぶことが可能となると考えられる。

h の下限値 h_{ij}^{lower} の設定方法は、以上に言及した方法以外にもあり得ると考えられるが、以下では、④の方法を用いることとする。そこで、与えられたデータにおける H_{ij} の最小値 H_{data}^{\min} が負の値をとらない場合と負の値をとる場合に分けて考える。 $H_{\text{data}}^{\min} \geq 0$ のとき、 h_{ij}^{lower} に対応する H_{ij} を 0 と考える。 $H_{ij} = 0$ は 2 変数 i 及び変数 j が独立であると仮定して周辺度数から計算される期待値と、観測されたガットマン・エラーの度数が同一となる場合である。一方、 $H_{\text{data}}^{\min} < 0$ のとき、その値を h_{ij}^{lower} と考える。これらは次式のように表される。

$$[5.2.7] \quad \begin{cases} 0 \leq h \leq 1 & \text{if } H_{\text{data}}^{\min} \geq 0 \\ H_{\text{data}}^{\min} \leq h \leq 1 & \text{if } H_{\text{data}}^{\min} < 0 \end{cases}$$

さて、この H_{ij} の取り得る値の範囲と、式[5.2.2]に示す 2 点間の距離 d_{ij} の取り得る範囲を対応させる関数 g_{ij} はいろいろと考えられる。例えば、2 点 i 及び j が極座標上に布置されている場合、各々の偏角 θ_i と θ_j の差を $\theta_{ij} = |\theta_i - \theta_j|$ としたとき、動径 r_i と r_j

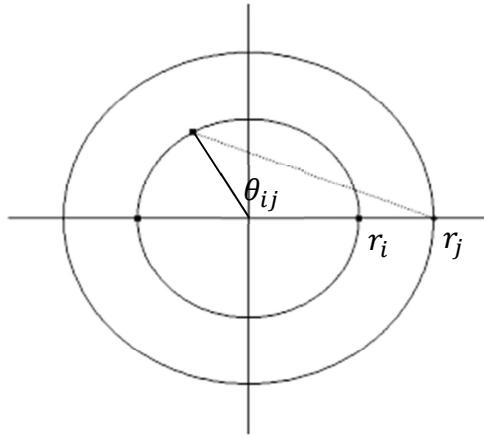
によって示される 2 次元上の 2 点間の距離 $\sqrt{(r_i \cos \theta_i - r_j \cos \theta_j)^2 + (r_i \sin \theta_i - r_j \sin \theta_j)^2}$ で表される θ_i 及び θ_j の関数は、偏角に対する動径 r_i 及び r_j が描く円周上の 2 点間の距離を表すため利用できる。ここで $\theta_j = 0$ とし、 $\theta_{ij} = \theta$ としても一般性は失われない。そこで、極座標表示による点 j ($r_j, 0$) に対する点 i (r_i, θ) が $0 \leq \theta \leq \pi$ の範囲で移動する際の 2 点間の距離 $\sqrt{(r_i \cos \theta - r_j)^2 + (r_i \sin \theta)^2}$ を $g_{ij}(\theta)$ とすることができる。横軸を θ 、縦軸を $g_{ij}(\theta) \equiv \sqrt{(r_i \cos \theta - r_j)^2 + (r_i \sin \theta)^2}$ とした 2 次元上の曲線を図 5-1(b) に示す。

この関数を用いる場合、 $H_{ij} = 1$ のとき最小の 2 点間の距離をとるように対応させることから、

$$[5.2.8] \quad g_{ij}(h) = \sqrt{\left[r_i \cos \left\{ \frac{(1-h)\pi}{1-H_{\text{data}}^{\min}} \right\} - r_j \right]^2 + \left[r_i \sin \left\{ \frac{(1-h)\pi}{1-H_{\text{data}}^{\min}} \right\} \right]^2}$$

とすることが考えられる。

(a) 2 点 i, j の移動しうる範囲



(b) 2 点 i, j 間の距離

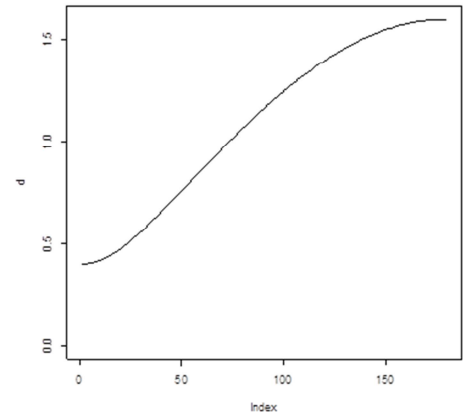


図 5-1 動径の異なる円周上の 2 点が移動しうる範囲と 2 点間の距離

しかし、ここでは、図 5-1(b) に描かれる曲線を近似した一次関数による直線を $g_{ij}(h)$ とする、より単純な方法を考える。横軸に h 、縦軸に $g_{ij}(h)$ をとった関数、及び横軸に $1-h$ 、縦軸に $g_{ij}(h)$ をとった関数による 2 次元上の直線を図 5-2 に示す。図の左側は、 $H_{\text{data}}^{\min} \geq 0$ の場合、右側は $H_{\text{data}}^{\min} < 0$ の場合である。ここでは、 H_{ij} が小さくなるほど d_{ij} が大きくなることから、図の下側、横軸に $1-h$ をとった関数を g_{ij} とする。

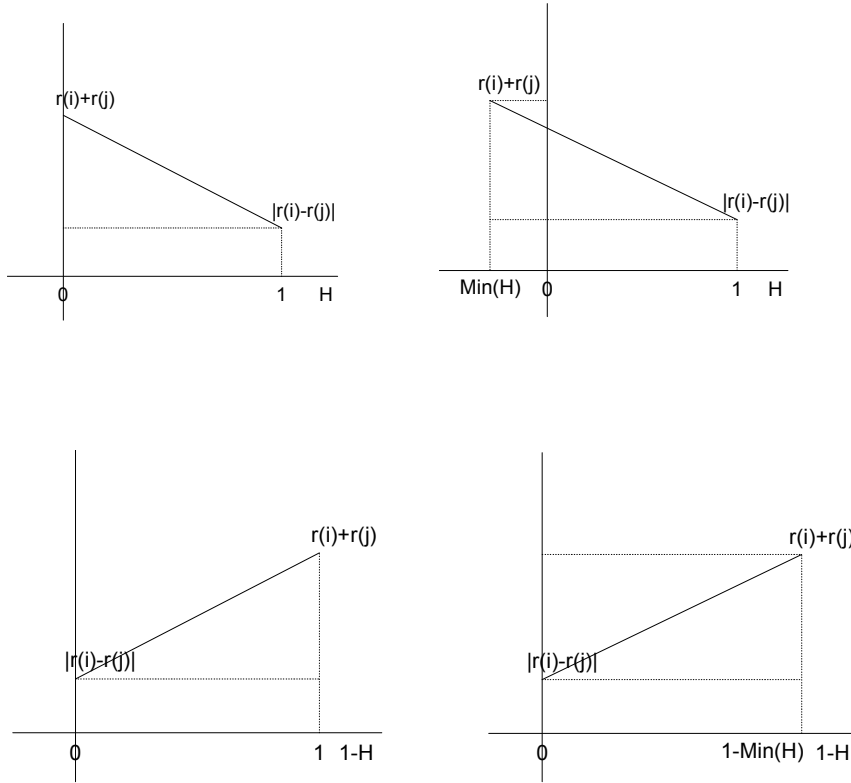


図 5-2 h と d_{ij} の関係及び $1-h$ と d_{ij} の関係を直線で示したグラフ

図 5-2 下を示された一次関数 g_{ij} によって変換された $g_{ij}(h)$ と d_{ij} の差の二乗和を目的関数 η^2 とすると、次式のように表される。

$$[5.2.9] \quad \eta^2 = \begin{cases} \sum_i \sum_{i < j} \left(\left((1-h)(r_i + r_j - |r_i - r_j|) + |r_i - r_j| \right) - d_{ij} \right)^2 & \text{if } H_{\text{data}}^{\min} \geq 0 \\ \sum_i \sum_{i < j} \left(\left((1-h) \frac{r_i + r_j - |r_i - r_j|}{1 - H_{\text{data}}^{\min}} + |r_i - r_j| \right) - d_{ij} \right)^2 & \text{if } H_{\text{data}}^{\min} < 0 \end{cases}$$

この目的関数 η^2 を最小化する θ を準ニュートン法による最適化を用いて求めることにより、各点の位置を決める θ が決まる。ただし、 θ とは各変数の極座標における偏角 θ_i ($i = 1, \dots, M$) を要素とするベクトルである ($\theta = (\theta_1, \dots, \theta_M)'$)。準ニュートン法を用いるにあたって利用する、 η^2 の θ_i による偏微分の式を付録 E に示す。

ここで、 η^2 の値は選択する変数の数やデータそのものに依存する。そのため、異なるデータ同士で Radex モデルに対するデータの適合の度合いを比較することができない。そこで、Kruskal の MDS における式[3.2.8]及び式[3.2.9]と同様に、式[5.2.9]の η^2 を次式のように定義した T で除した値の平方根 S をもって評価することを考える。

$$[5.2.10] \quad T = \begin{cases} \sum_i \sum_{i < j} \left((1 - H_{ij})(r_i + r_j - |r_i - r_j|) + |r_i - r_j| \right)^2 & \text{if } H_{\text{data}}^{\min} \geq 0 \\ \sum_i \sum_{i < j} \left((1 - H_{ij}) \frac{r_i + r_j - |r_i - r_j|}{1 - H_{\text{data}}^{\min}} + |r_i - r_j| \right)^2 & \text{if } H_{\text{data}}^{\min} < 0 \end{cases}$$

$$[5.2.11] \quad S = \sqrt{\frac{\eta^2}{T}}$$

この S は正の値を取り、最大値は $\eta^2 = T$ 、つまり全ての i, j について $d_{ij} = 0$ となるときに値 1 をとることから、 $0 \leq S \leq 1$ となる。この値が小さいほど、2 次元平面上の変数間の距離の大小関係がデータの大小関係をより正確に反映していることから、得られた布置のデータへの適合がよいといえる。

以上のようにして 2 つの Facet を構成するように点の位置をプロットすることにより、Radex モデルを満たす点の布置を求めることができる。

なお、最適化によって得られる解 θ ベクトルについて、任意の角度で回転させた結果も明らかに解になる。つまり、回転に対する不定性が存在する。そのため、実際に最適化を行う際は、 M 個の変数に対応する θ ベクトルの M 個の要素のうち、1 番目の要素をあらかじめ特定の値に固定し、残り $M - 1$ 個の θ ベクトルの要素を最適化によって求めることとした。これによって、回転に対する不定性の問題を回避することができる。

最後に、データへの適合の考え方を利用することで、分析に用いる変数の選択が可能となる点について言及する。式[5.2.9]における η^2 を変数ごとに分解し、変数 i に対する η_i^2 を次式のように定義する。

$$[5.2.12] \quad \eta_i^2 = \begin{cases} \sum_{j \neq i} \left(\left((1 - H_{ij})(r_i + r_j - |r_i - r_j|) + |r_i - r_j| \right) - d_{ij} \right)^2 & \text{if } H_{\text{data}}^{\min} \geq 0 \\ \sum_{j \neq i} \left(\left((1 - H_{ij}) \frac{r_i + r_j - |r_i - r_j|}{1 - H_{\text{data}}^{\min}} + |r_i - r_j| \right) - d_{ij} \right)^2 & \text{if } H_{\text{data}}^{\min} < 0 \end{cases}$$

この η_i^2 の値がもっとも大きい変数が、データの布置に対する適合を悪くしていることになる。実際には、 η_i^2 の値の大きい順に並べ、その値の減衰状況を見て、他の変数よりも η_i^2 の値が大きい変数は除外して再度分析する、という利用の仕方も可能である。特に、データに含まれる他の変数と Guttman Scale を構成しにくい変数は、このような形で検出することができる。

5.2.4 ブートストラップ法による布置の安定性

各変数間の Guttman Scale の構成度と二値変数の割合に応じて視覚化する本手法は、各変数の点を布置する際の極座標における偏角のパラメータを推定する。しかし、データが若干異なることによって Guttman Scale の構成度が異なった場合、各点の布置も当然異なってくる。そのため、結果の安定性を併せて検討することが役立つ可能性がある。

そこで、ブートストラップ法 (Efron & Tibshirani, 1993) を利用して、結果の安定性を考える。サンプルサイズ n の標本から重複を許してサイズ n の観測値の再抽出を B 回繰り返す。そして再抽出された観測値のセットごとに二値変数の割合と変数間の Homogeneity 係数を計算し、Homogeneity 係数に基づいて最適化を行い、各変数の極座標 (r_i, θ_i) を求める。

通常のブートストラップ法では、重複を許して B 回にわたってリサンプリングを行う場合、抽出における各個人の出現頻度はさまざまであり、場合によっては B 個の観測値のセットの中に 1 回も選択されない個人が出てくる可能性もある。ある先行研究 (Linting et al., 2007b) では、等質性分析の結果の安定性をブートストラップ法によって検討する際に全個人が同じ抽出回数、つまり B 回、出現するように配意した Balanced Bootstrap (Efron & Tibshirani, 1993) を用いる方が、 B 個の観測値データの平均を利用した統計量におけるバイアスを調整するという点で望ましい、としている。そこで、本研究においても Balanced Bootstrap の手法を用いて結果の安定性を検討する。

Balanced Bootstrap の具体的な方法は次のとおりである。得られている大きさ n の観測値のセットを B 回コピーしてつなげて大きさ nB の観測値のセットを作り、そのセットについて任意の並べ替えを行う。次に、そのセットの初めから n 個ずつを取り出して B 個の観測値のセットを作製する。これによって、抽出された全ての観測値のセットの中に、サイズ n に含まれる各観測値が等しく B 回出現するように抽出されることになる。

B 個のリサンプルについて、それぞれ目的関数の最適化によって解を求めるが、ここでは、得られた解の回転及び反転を行い、オリジナルサンプルのデータに対して最適化を行った解にもっとも近い布置を求め、それらをブートストラップによって得られた布置とする。具体的には次の手順で計算した。

フルサンプルを用いた M 個の変数の布置に対応する極座標を表す動径ベクトルを \mathbf{r}_o 、各要素を r_{oi} ($i = 1, \dots, M$) とし、偏角ベクトルを $\boldsymbol{\theta}_o$ 、各要素を θ_{oi} ($i = 1, \dots, M$) とする。また、ブートストラップ法で抽出した観測値セットを用いた M 個の変数の布置に対応する

極座標を表す動径ベクトルを \mathbf{r}_b 、各要素を r_{bi} ($i = 1, \dots, M$) とし、偏角ベクトルを $\boldsymbol{\theta}_b$ 、各要素を θ_{bi} ($i = 1, \dots, M$) とする。このとき、ブートストラップ法の観測値セットによる座標を θ_{rot} だけ回転させ、各変数の座標 (r_{oi}, θ_{oi}) と (r_{bi}, θ_{bi}) の対応する変数同士のユークリッド距離の二乗和を最小とする θ_{rot} を準ニュートン法による最適化によって求めた。すなわち、目的関数 $L(\theta_{\text{rot}})$ を次式によって定義する。

$$[5.2.13] \quad L(\theta_{\text{rot}}) = \sum_{i=1}^M \{(r_{oi} \cos \theta_{oi} - r_{bi} \cos(\theta_{bi} + \theta_{\text{rot}}))^2 + (r_{oi} \sin \theta_{oi} - r_{bi} \sin(\theta_{bi} + \theta_{\text{rot}}))^2\}.$$

そしてこの $L(\theta_{\text{rot}})$ を最小化する θ_{rot} を求めた。また、 (r_{bi}, θ_{bi}) ($i = 1, \dots, M$) を反転させた布置を回転させて、同様にオリジナルサンプルを用いた布置と比較し、各変数同士のユークリッド距離の二乗和を最小とする $\theta_{\text{rot}'}$ を準ニュートン法による最適化によって求めた。目的関数 $L(\theta_{\text{rot}'})$ は次のとおりである。

$$[5.2.14] \quad L(\theta_{\text{rot}'}) = \sum_{i=1}^M \{(r_{oi} \cos \theta_{oi} - r_{bi} \cos(-\theta_{bi} + \theta_{\text{rot}'}))^2 + (r_{oi} \sin \theta_{oi} - r_{bi} \sin(-\theta_{bi} + \theta_{\text{rot}'}))^2\}.$$

$L(\theta_{\text{rot}})$ 及び $L(\theta_{\text{rot}'})$ の値を最小とする回転角 θ_{rot} または反転後の回転角 $\theta_{\text{rot}'}$ を用いて、ブートストラップ法の観測値セットによる布置に対して回転、または反転及び回転を施し、それを最終的な解とした。なお、ブートストラップ法の各観測値セットに RADXMS を適用する際の初期値には、オリジナルサンプルの RADXMS の結果を用いた。

5.2.5 3次元への拡張

データが 2 次元では十分に η^2 を小さくすることができない場合もあり得る。そこで、同様の布置を 3 次元空間に拡張した場合について付言する。xyz 軸からなる 3 次元空間内に各点をプロットする場合、変数 i の極座標を次のように設定する。変数 i の動径を r_i 、変数 i と z 軸を含む平面において変数 i が z 軸と成す角度を θ_i 、変数 i と z 軸を含む平面が xy 平面と交わる直線の原点周りの偏角を φ_i とし、変数 i の極座標を $(r_i, \theta_i, \varphi_i)$ と表すことにする。

動径 r_i については、5.2.1 節で説明した 2 次元平面における議論がそのまま利用できる。そこで、 r_i を式[5.2.1]によって定義する。このとき変数 i 及び変数 j の 3 次元空間における距離 d_{ij} は次式で表される。

$$[5.2.15] \quad d_{ij} = \sqrt{(r_i \sin \theta_i \cos \varphi_i - r_j \sin \theta_j \cos \varphi_j)^2 + (r_i \sin \theta_i \sin \varphi_i - r_j \sin \theta_j \sin \varphi_j)^2 + (r_i \cos \theta_i - r_j \cos \theta_j)^2}$$

このように定義された d_{ij} を用いて、式[5.2.9]によって定義された η^2 を最小化する θ_i

及び φ_i が、3次元空間内における点の布置を示す偏角となる。 η^2 の最小化は、2次元における場合と同様に準ニュートン法によって求めることが可能であり、その結果、 η^2 を最小化するパラメータ $\hat{\theta}_i$ 及び $\hat{\varphi}_i$ が求まることから、3次元空間における布置を得ることができる。準ニュートン法を用いるにあたって利用する、 η^2 の θ_i 及び φ_i による偏微分の式を付録Fに示す。

また、2次元の場合と同様に η^2 からデータの布置に対する適合を示す指標として式[5.2.11]によって定義された S を算出することができる。以上から、3次元空間における点の布置にも拡張することができる。

5.3 人工データへの適用

5.2節で説明した手法を用いてデータを分析し、2次元平面への布置を求めた結果について検証する。まず、人工データ2に示す14変数からなる二値データに適用する。変数1～7の変数同士、及び変数8～14の変数同士では完全な Guttman Scale を構成しており、それらの変数間の Homogeneity 係数は1となる。一方、変数1～7の変数と変数8～14の変数との間の Homogeneity 係数は、表5-1に示すように $-0.50 \sim 0.25$ の値をとる。そこで、これら14変数を2次元に布置する場合、変数1～7のプロットによる一列、変数8～14のプロットによる一列が原点から配置され、それぞれの列は互いに離れて布置されることが望まれる。

表 5-1 人工データ2（14変数）における変数間の Homogeneity 係数

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13
v2	1.00												
v3	1.00	1.00											
v4	1.00	1.00	1.00										
v5	1.00	1.00	1.00	1.00									
v6	1.00	1.00	1.00	1.00	1.00								
v7	1.00	1.00	1.00	1.00	1.00	1.00							
v8	-0.25	-0.33	-0.20	-0.13	-0.33	0.25	-0.50						
v9	-0.33	-0.08	-0.05	-0.13	-0.25	0.00	-0.50	1.00					
v10	-0.10	0.05	0.04	0.05	-0.07	0.20	-0.20	1.00	1.00				
v11	-0.25	-0.06	0.00	0.04	-0.11	0.00	-0.17	1.00	1.00	1.00			
v12	-0.50	-0.08	-0.07	0.00	-0.05	0.07	0.00	1.00	1.00	1.00	1.00		
v13	-0.50	0.00	-0.10	-0.08	-0.21	-0.13	-0.13	1.00	1.00	1.00	1.00	1.00	
v14	-0.50	-0.25	-0.20	-0.17	-0.43	-0.25	-0.11	1.00	1.00	1.00	1.00	1.00	1.00

そこで、表5-1に示される各変数間の Homogeneity 係数の値を反映するように、2次元平面上に点をプロットした結果を図5-3に示す。ただし、式[5.2.1]における動径を決めるとき

の α の値を変えて、 $\alpha = 1, 2, 3, 4, 5, 10$ の 6 つの値について計算を実施した。

変数 1~7 の点列と変数 8~14 の点列の場所が α の値によって異なるものの、それぞれ離れて布置されていることがわかる。また、 α の値を大きくするほど、割合の高い変数が原点の近傍に圧縮されて布置される様子が観察される。つまり α の値を調整することによって図示の仕方を調整することが可能である。例えば、割合の高い変数よりも割合の低い変数間の識別に焦点を当てて視覚化を行いたい場合、 α の値を大きくすることが望ましい。しかし、割合の高い変数は圧縮されて中心に布置されることから、それらについては位置関係を弁別しにくいという点がその代償となる。

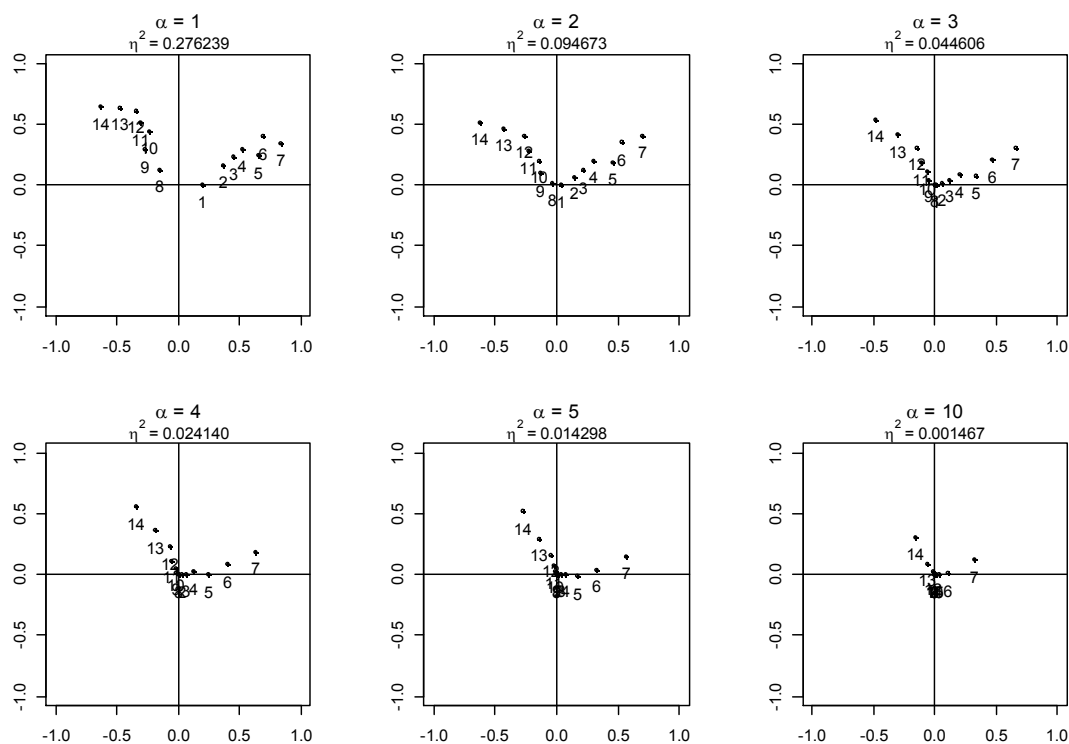


図 5-3 人工データ 2 の RADXMS による布置 ($\alpha = 1, 2, 3, 4, 5, 10$)

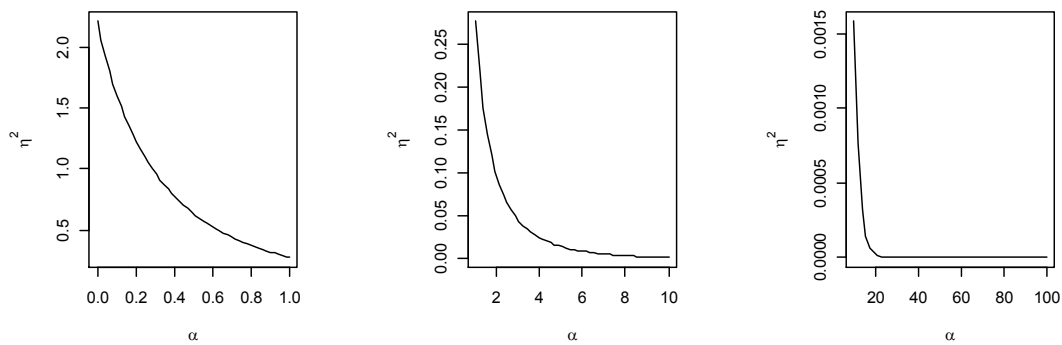


図 5-4 α 値に対する人工データ 2 の RADXMS の結果における η^2 の推移

ここで、 α の値を $[0,1]$ 、 $[1,10]$ 、 $[10,100]$ の範囲で変化させながら、それぞれについて人工データ 2 に対する RADXMS を実行し、最適化された η^2 の値の推移を描いた結果を図 5-4 に示す。図に示されるように、 α に対して η^2 の値が単調に減少していくことがわかる。このため、データから適切な α の値を導くことはできない。変数間の視覚化がしやすい度合いを測定する何らかの指標を導入することで、適切な α の値を導くことが可能となると考えられる。

また、先ほどの図 5-3 では、2 つの点列が互いに遠くに布置される、具体的には 180° 反対側に布置されるという結果になっていない。これは、一方の点列に含まれる点と他方の点列に含まれる点との Homogeneity 係数が $-0.5 \leq H_{ij} \leq 0.25$ の範囲に分布していることによる。各点列でもっとも該当割合の高い変数 1 と変数 8 に着目すると、変数 1 と変数 14 は $H_{ij} = -0.50$ でもっとも離れ、変数 8 と変数 7 も $H_{ij} = -0.50$ でもっとも離れることになる。しかし、変数 1 ともう一方の点列（変数 8～14）の変数の番号が小さくなるに従って H_{ij} の値は大きくなる。同様に変数 8 ともう一方の点列（変数 1～7）の変数の番号が小さくなるに従って、 H_{ij} の値は大きくなることから、点列同士は図示されている以上には広がらない。

そこで、負の値をとる Homogeneity 係数については Guttman Scale を構成する度合いは低く、負の値の間で大小関係を吟味する必要性はない、という立場で分析を行ってみる。つまり、負の値をとる Homogeneity 係数の値を全て 0 とし、その上で 5.2.3 節における③ $h_{ij}^{\text{lower}} = 0$ によって同じ人工データ 2 を分析した。結果を図 5-5 に示す。

想定されたとおり、それぞれの点列（変数 1～7 及び変数 8～14）が離れて布置されている。このことから、Guttman Scale を構成する異なる変数群がある場合、負の値をとる Homogeneity 係数を全て 0 とし、 $0 \leq H_{ij} \leq 1$ の範囲で d_{ij} を計算する ③ $h_{ij}^{\text{lower}} = 0$ の方法も Guttman Scale の弁別には有効であることがわかる。

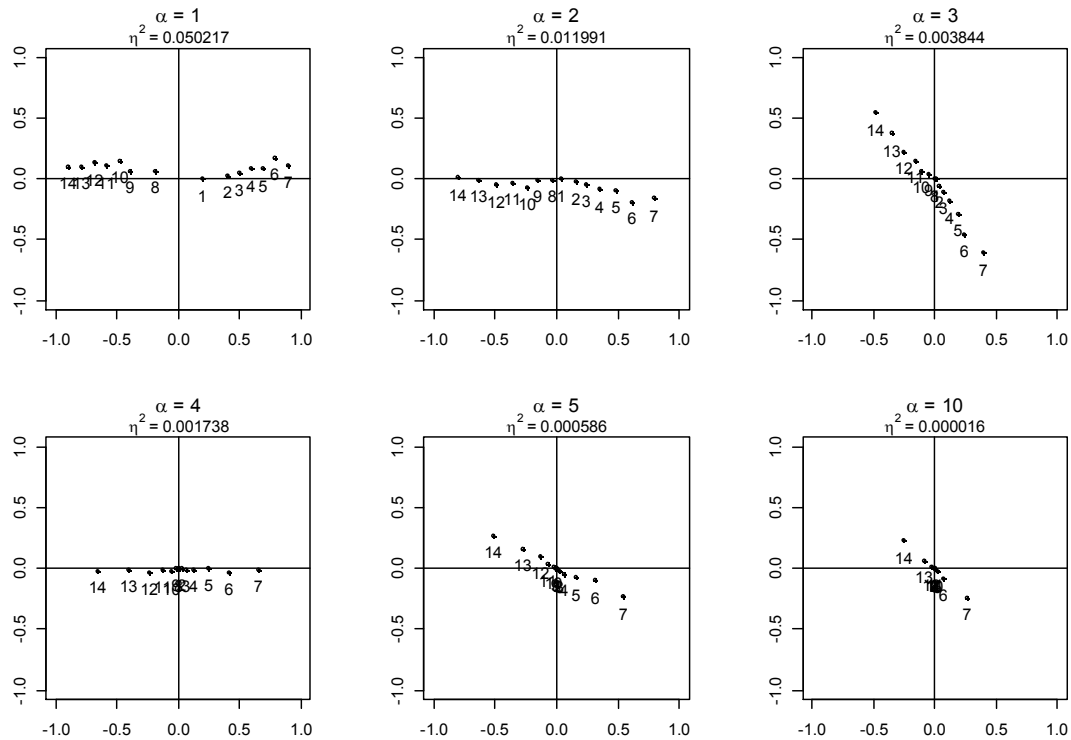
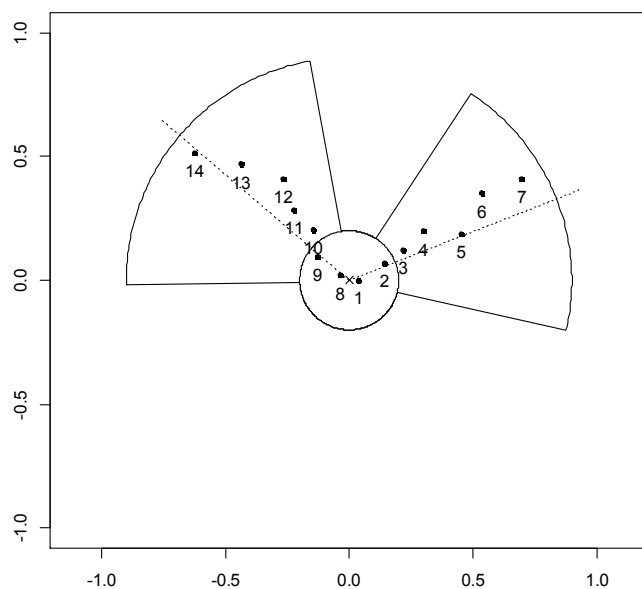


図 5-5 人工データ 2 において負の値をとる Homogeneity 係数を全て 0 とした RADXMS による布置 ($\alpha = 1, 2, 3, 4, 5, 10$)

次に、 $\alpha = 2$ として再度 RADXMS によって点を布置させるとともに、第 4 章で提案した重みづけ Homogeneity 係数の値を重ね合わせた結果 (重みづけにはロジスティック関数を利用し、 $a = 10000$ 、 $\lambda = 0.5$ 、 $\theta_0 = \pm \pi/12$ ラジアン) を図 5-6 に示す。ただし、上段 (a) は Homogeneity 係数に負の値があるときにその最小値を利用する場合 (④ $h_{ij}^{\text{lower}} = \min(0, H_{\text{data}}^{\text{min}})$)、下段 (b) は Homogeneity 係数の負の値を全て 0 とする場合 (③ $h_{ij}^{\text{lower}} = 0$) である。

これを見ると明らかなように、(a) (b) 2 つの図ともに、人工データについて Guttman Scale を構成する変数群を明確に識別できるようにデータの視覚化が行われたといえる。

(a) Homogeneity 係数に負の値があるときにその最小値を利用する場合



(b) Homogeneity 係数の負の値を全て 0 とする場合

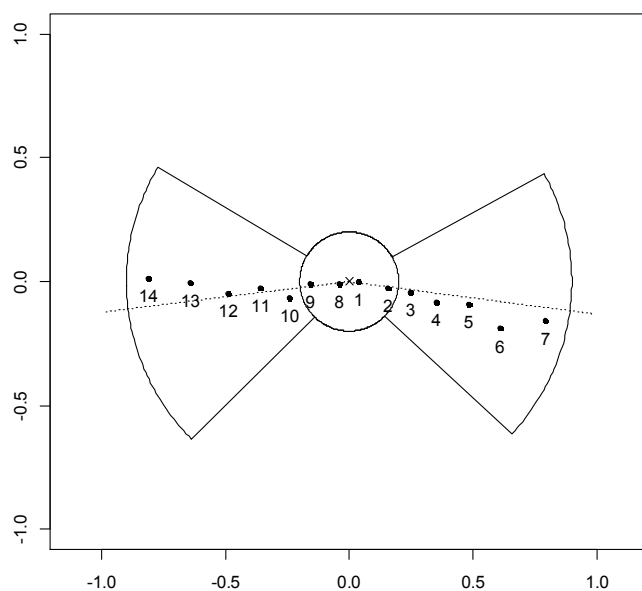


図 5-6 RADXMS に重みづけ Homogeneity 係数の水準（ロジスティック関数による重みづけ、 $a = 10000$ 、 $\lambda = 0.5$ 、 $\theta_0 = \pm 15^\circ$ ）を重ね合わせた表示

5.4 メディア暴力調査データへの適用

5.4.1 メディア暴力調査データによる布置の作成

続いて、3.5.2 で説明したメディア暴力調査データに対して、RADXMS 及び RADXEA 重みづけ Homogeneity 係数による視覚化手法（2 次元）を適用する。式[5.2.1]における α について、以下、 $\alpha = 2$ を適用する。

ここで、メディア暴力調査データを構成する 17 変数について、それらの変数間の Homogeneity 係数を表 5-2 に示す。

表 5-2 メディア暴力調査データ（17 変数）における変数間の Homogeneity 係数

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16
v2	0.22															
v3	0.11	0.34														
v4	-0.05	-0.02	-0.24													
v5	-0.25	0.14	-0.20	0.74												
v6	-0.06	0.07	-0.10	0.19	0.33											
v7	-0.09	0.04	-0.13	0.60	0.30	0.37										
v8	-0.14	0.28	0.27	0.71	0.37	0.51	0.56									
v9	0.13	0.31	0.23	0.50	0.36	0.57	0.43	0.44								
v10	0.03	0.09	0.03	0.11	0.15	0.27	0.35	0.41	0.60							
v11	0.01	0.14	0.13	0.08	0.09	0.01	0.09	0.33	0.32	0.23						
v12	0.10	0.13	0.22	0.26	0.18	0.23	0.28	0.38	0.37	0.36	0.75					
v13	-0.11	0.00	-0.23	0.26	0.25	0.07	0.43	0.45	0.58	0.10	0.17	0.21				
v14	-0.06	0.01	-0.05	0.17	0.13	0.03	0.30	0.60	0.51	0.15	0.38	0.50	0.46			
v15	0.17	0.36	0.32	0.54	0.22	0.37	0.29	0.44	0.38	0.24	0.43	0.34	0.51	0.49		
v16	0.02	0.16	-0.03	0.60	0.31	0.32	0.32	0.68	0.54	0.21	0.23	0.37	0.50	0.61	0.73	
v17	0.19	0.25	0.31	0.55	0.21	0.40	0.31	0.49	0.51	0.55	0.49	0.35	0.49	0.50	0.60	0.70

RADXMS を利用する場合、準ニュートン法による最適化に関して、初期値の設定が結果を左右している。つまり、最適化によって得られた解が局所解となっている可能性があるということである。そこで、ランダムな初期値を複数回設定してプログラムを実行し、式[5.2.9]における η^2 をもっとも小さくする結果を採用することにする。

そこで、準ニュートン法によって最適化するための 17 変数の平面上の布置に関する角度の初期値をランダムに 10000 回発生させ、その中で S 値が最小となる布置を採用することとした。その結果、 $\eta^2 = 9.035272$ 、 $S = 0.282788$ の布置を得た。結果を図 5-7 に示す。

メディア暴力とは離れた位置で、恋愛もの (love)、ヒューマンドラマ (human) が、 H_w を示す水準で一つの山を構成したが、山自体は高くなかった。一方、メディア暴力と関連する領域として、一つは警察 (police)、戦争もの (war)、一つはアクション (action)、スパイもの (spy)、一つは暴力 (vio)、流血シーン (blood)、武器 (weapon) が描かれたもの、一つはオカルト系 (occult、splatter) となると判断された。このほか、わずかなピークを示すものとして、犯罪 (crime)、ギャングもの (gang) がみられた。このうち、 H_w を示す水準はオカルト系の値が高く、警察、戦争ものの値は高くなかった。

これらの布置は、概ね第 4 章の SSA の図 4-9 の布置と似たような分類となった。S F (sf)

とアクション (action) の位置が若干異なっているものの、ギャングや犯罪がメディア暴力と関連する変数群の近くに位置し、アクションは犯罪と警察・戦争ものの間に位置するという点は共通している。また、ファンタジー (fantasy) がスプラッターに近い方に布置している点も類似している。

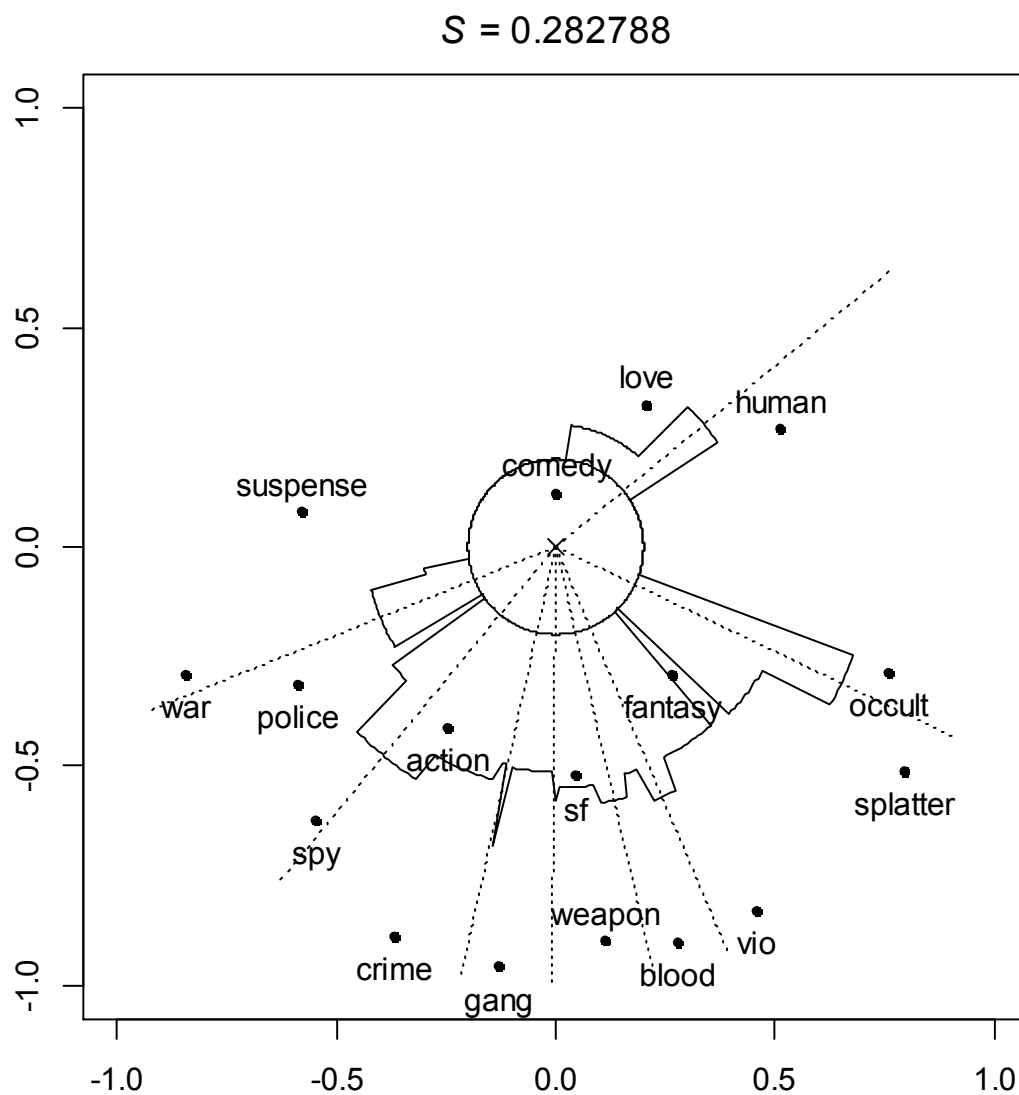


図 5-7 メディア暴力調査データ (17 変数) に初期値 10000 回のうち S 値がもっとも低い RADXMS 及び重みづけ Homogeneity 係数の水準による RADXEA を適用した結果 ($\alpha = 2$ 、ロジスティック曲線を利用した重み付け、 $\lambda = 0.5$ 、 $\theta_0 = \pm 1/12$ ラジアン)

5.4.2 ブートストラップ法による布置の安定性の検討

ここで $B = 200$ として $n = 1208$ の調査データに対して **Balanced Bootstrap** を実施し、抽出された各標本に準ニュートン法による最適化を行い、各変数の極座標について r_i を横軸、 θ_i を縦軸として散布図を描くとともに、 r_i と θ_i 各々のヒストグラムを図 5-8-1 から図 5-8-5 までに示す。

r_i 及び θ_i とともに、ある値を中心に山形の分布を示している。ここで、 θ_i は循環データであるため、ヒストグラムの横軸は $0 \sim 2\pi$ の値を示している。例えば、変数 **human** は $\theta = 0$ を中心に分布し、オカルト系の変数 (**occult**、**splatter**)、及び **S F** (**sf**) は $\theta = 2\pi$ を中心に分布しているが、その反対の境界値に連続して分布していることに注意したい。

そこで、これら r_i 及び θ_i の各分布について 90% をカバーするデータ範囲のうち、レンジが最小となる幅を 90% 信頼区間として r_i と θ_i のそれぞれのデータ範囲を特定し、それらの範囲を示す円弧の作る図形をオリジナルサンプルを用いて算出した布置に重ね合わせて (StataCorp, 2007)、図 5-9 に示す。なお、図の中心の周りに円が表示され、その円の周囲に山や谷として描かれているが、これは第 4 章で示した重みづけ **Homogeneity** 係数による **RADXE**A を適用した H_w の値の推移を示す。

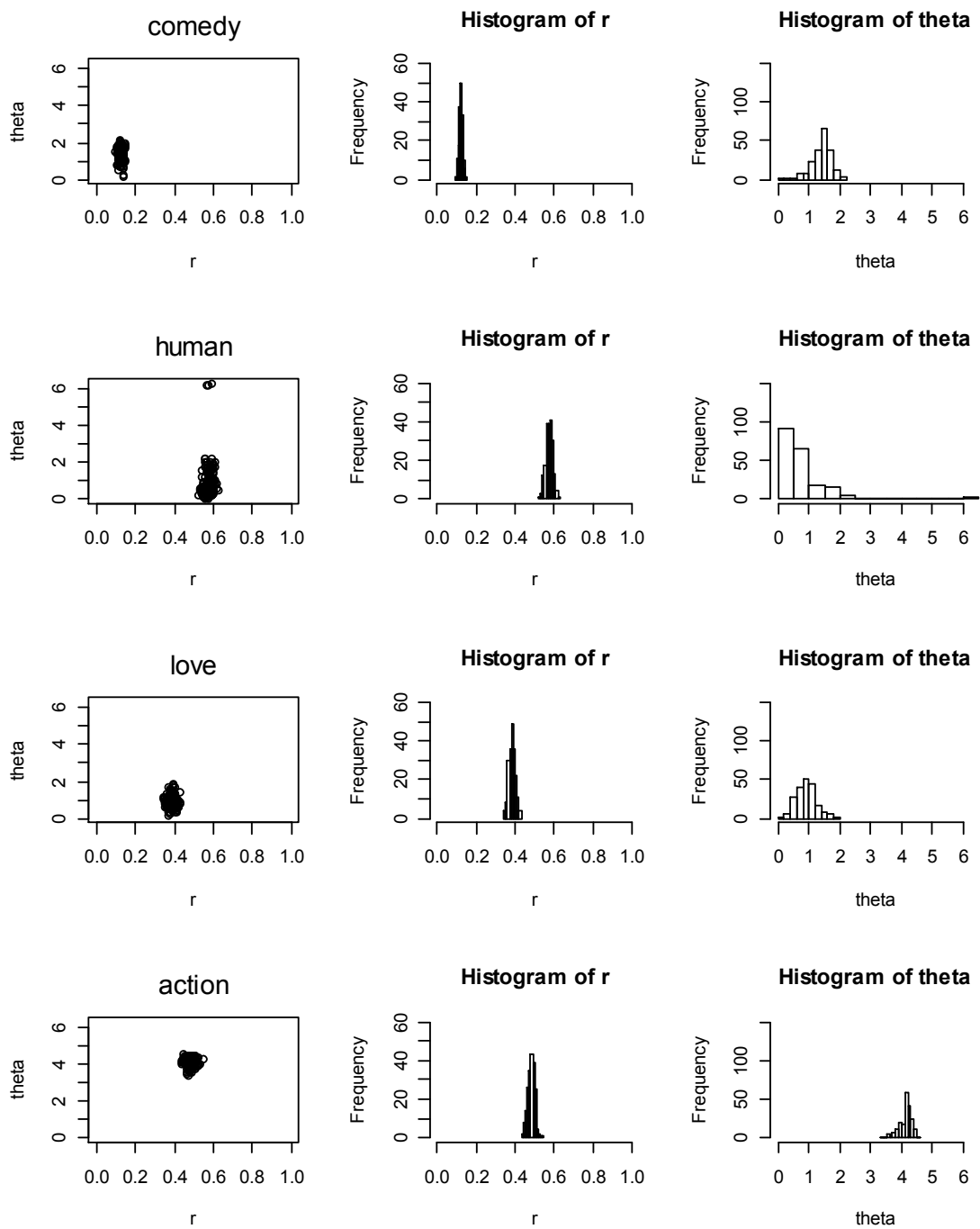


図 5-8-1 メディア暴力調査データについて Balanced Bootstrap 法により抽出された観測値 200 セットの各変数 (変数 1~4) の分布

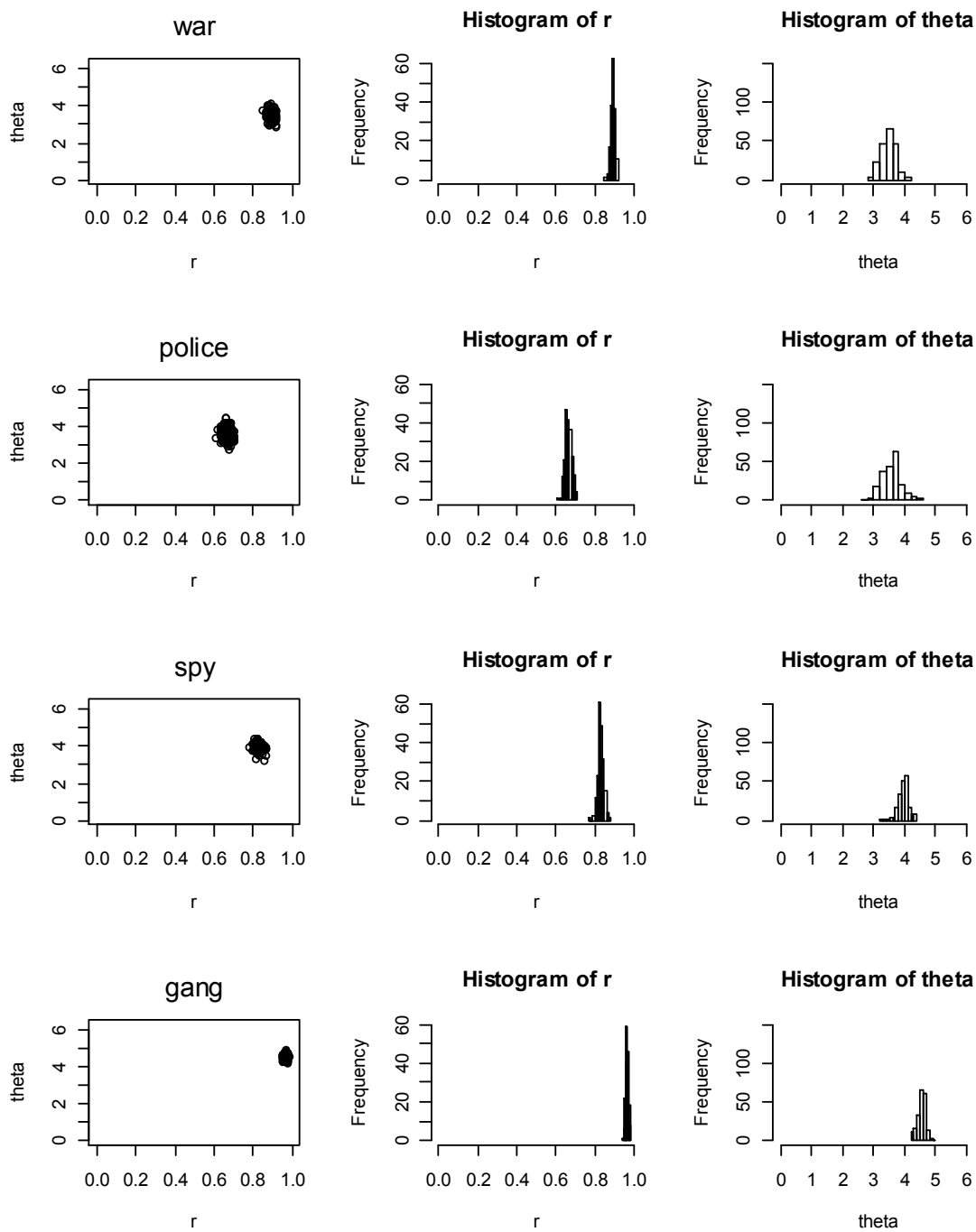


図 5-8-2 メディア暴力調査データについて Balanced Bootstrap 法により抽出された観測値 200 セットの各変数 (変数 5~8) の分布

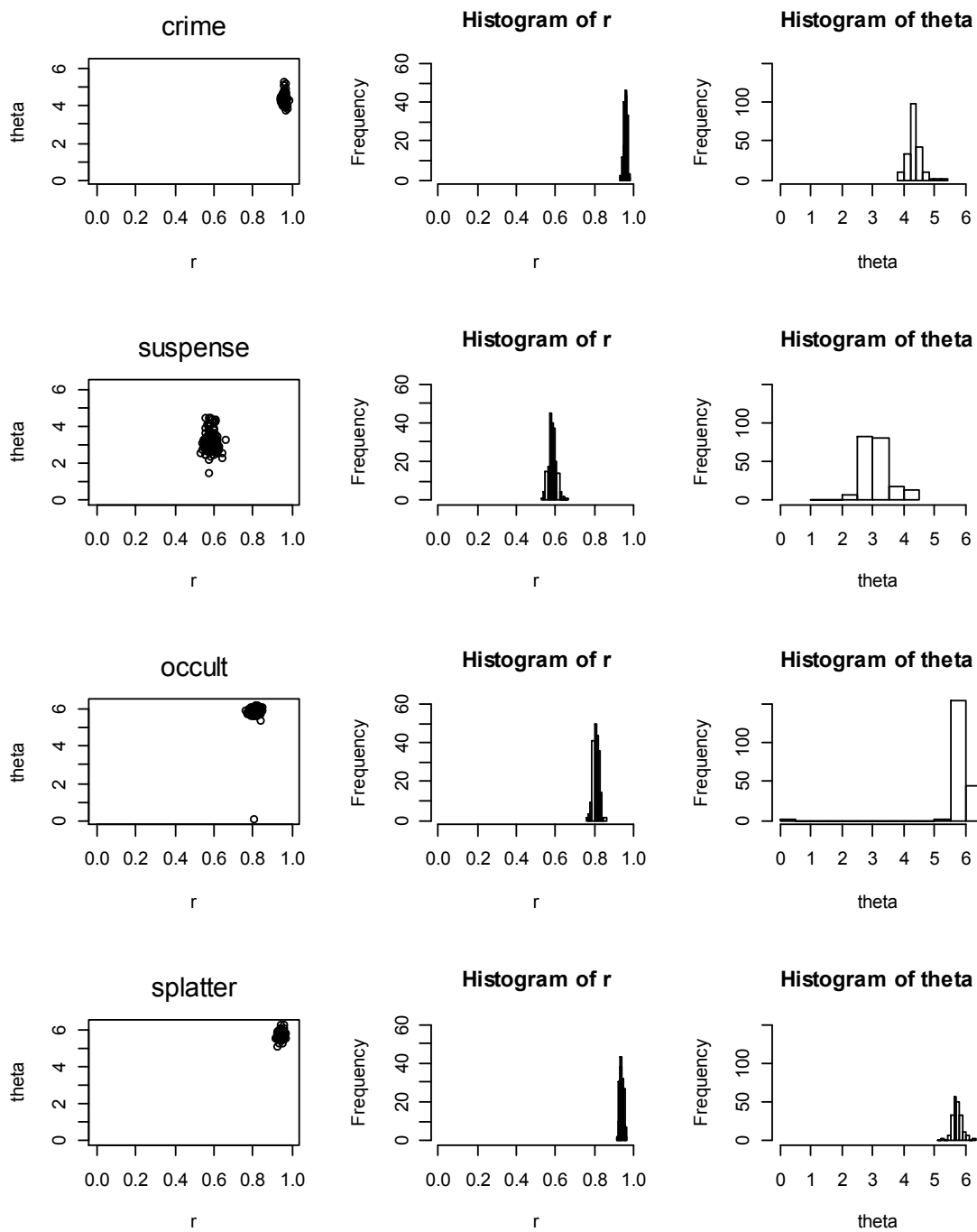


図 5-8-3 メディア暴力調査データについて Balanced Bootstrap 法により抽出された観測値 200 セットの各変数 (変数 9~12) の分布

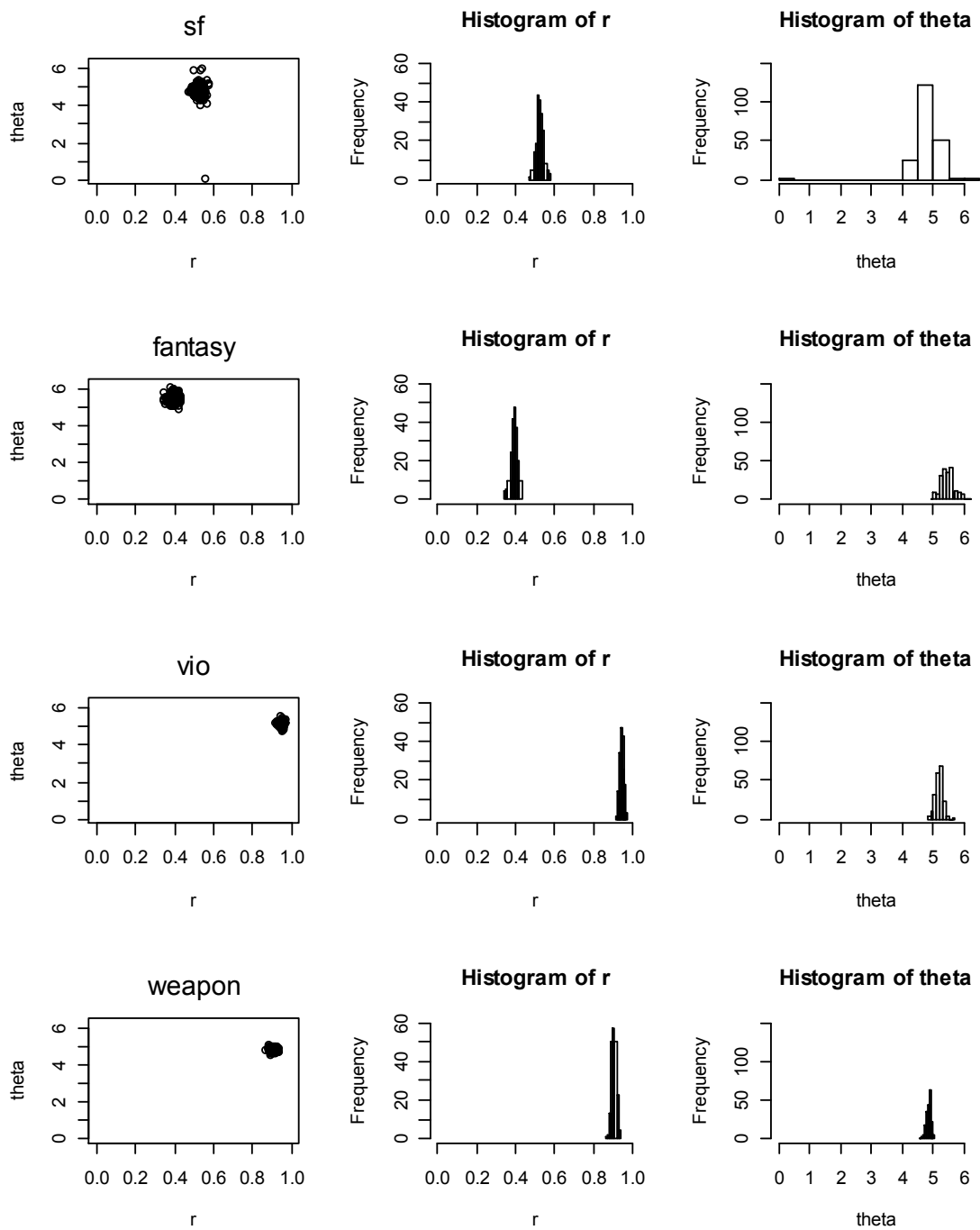


図 5-8-4 メディア暴力調査データについて Balanced Bootstrap 法により抽出された観測値 200 セットの各変数 (変数 13~16) の分布

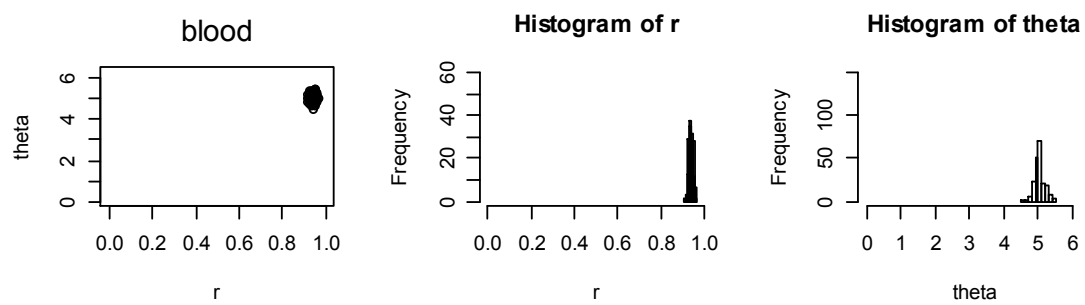


図 5-8-5 メディア暴力調査データについて Balanced Bootstrap 法により抽出された観測値 200 セットの各変数 (変数 17) の分布

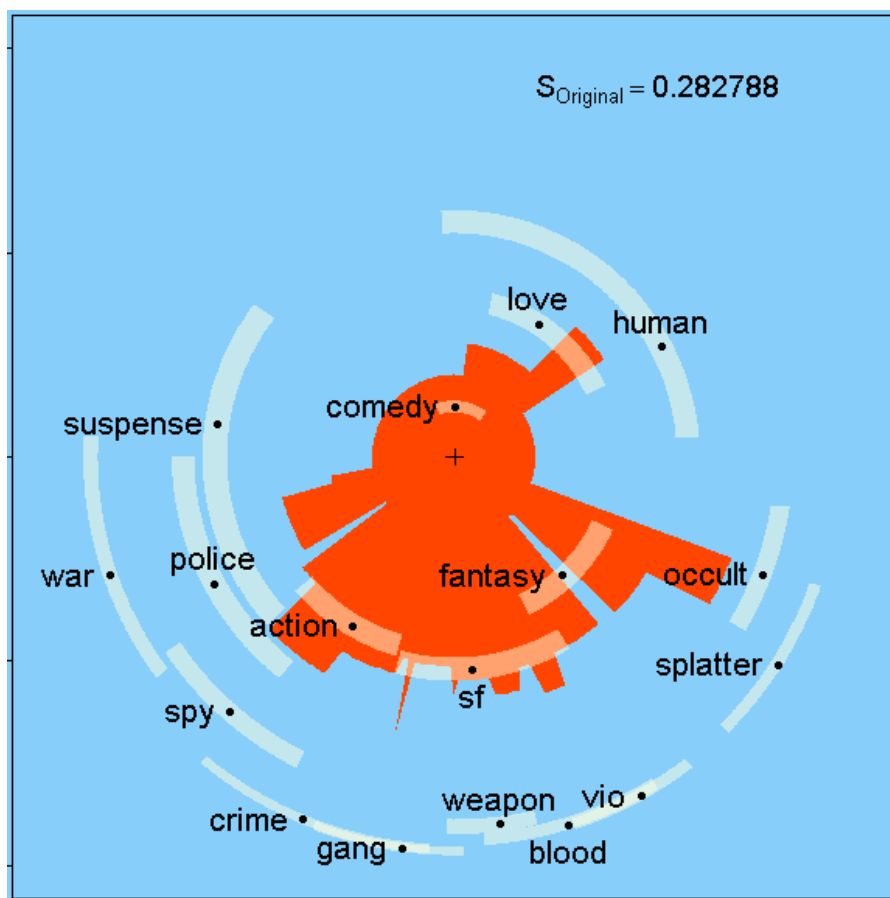


図 5-9 Balanced Bootstrap ($B = 200$ 回) による RADXMS 及び重みづけ Homogeneity 係数の水準による RADXEA を適用した結果($\alpha = 2$ 、ロジスティック関数による重み付け、 $a = 10000$ 、 $\lambda = 0.5$ 、 $\theta_0 = \pm \pi/12$ ラジアン)

S_{Original} : オリジナルサンプルにおける S の値

重みづけ Homogeneity 係数の水準を参照すると、図の上部から時計回りに、

- ①恋愛・ヒューマンドラマのクラスター (love、human)
- ②オカルトもののクラスター (occult、splatter)
- ③メディア暴力のクラスター(重みづけ Homogeneity 係数の水準の山は 2 つあるが近接している) (sf、weapon、blood、vio)
- ④犯罪もののクラスター (crime、gang)
- ⑤アクション・スパイもののクラスター (action、spy)
- ⑥ミリタリー (警察・戦争) もののクラスター (police、war)

と分かれることが見て取れる。このうち、重みづけ Homogeneity 係数の水準のレベルが高い

②オカルトもの、③メディア暴力、④犯罪もの、⑤アクションものの4つの変数クラスターが注目される。

さらに、各変数の90%信頼区間の範囲をみると、極座標における θ の値が重ならない変数同士は、Homogeneity係数に基づくクラスター化において同一クラスターには属しないと見なすことができる。このことを利用すると、①恋愛・ヒューマンドラマ系 (love、human)、②オカルト系 (occult、splatter)、③メディア暴力系 (sf、weapon、blood、vio)、⑥ミリタリー系 (police、war) の大きく4つの変数クラスターは互いに分離されることが確認できる。このことから、この4つの変数クラスターは安定的に存在していると考えられる。また、④犯罪系 (crime、gang) が③メディア暴力系と⑥ミリタリー系の間位置し、それらにまたがる形で⑤アクション系 (action、spy) が含まれてくる様子がうかがえる。あるいは、犯罪は警察もの、スパイものに近く、ギャングはメディア暴力のクラスターに近いことも示唆されよう。さらに、ファンタジー (fantasy) の変数もオカルト系とメディア暴力系の間位置することも読み取れる。

このように、重みづけ Homogeneity 係数の結果に対して **Balanced Bootstrap** によって算出した信頼区間領域の結果を補助的に参照することによって、Homogeneity 係数の平均が高い変数群を特定することが可能となり、それらが同一の領域に属していると思なすことによって、Radex モデルにおける **Polar Facet** の境界線を決定することが可能となる。ここでは、重みづけ Homogeneity 係数の水準の極大部分 (山の部分) に着目し、その極大部分同士を区切る谷の部分を目視によって見定めることによって **Polar Facet** の境界線を策定して図示した (図 5-10)。

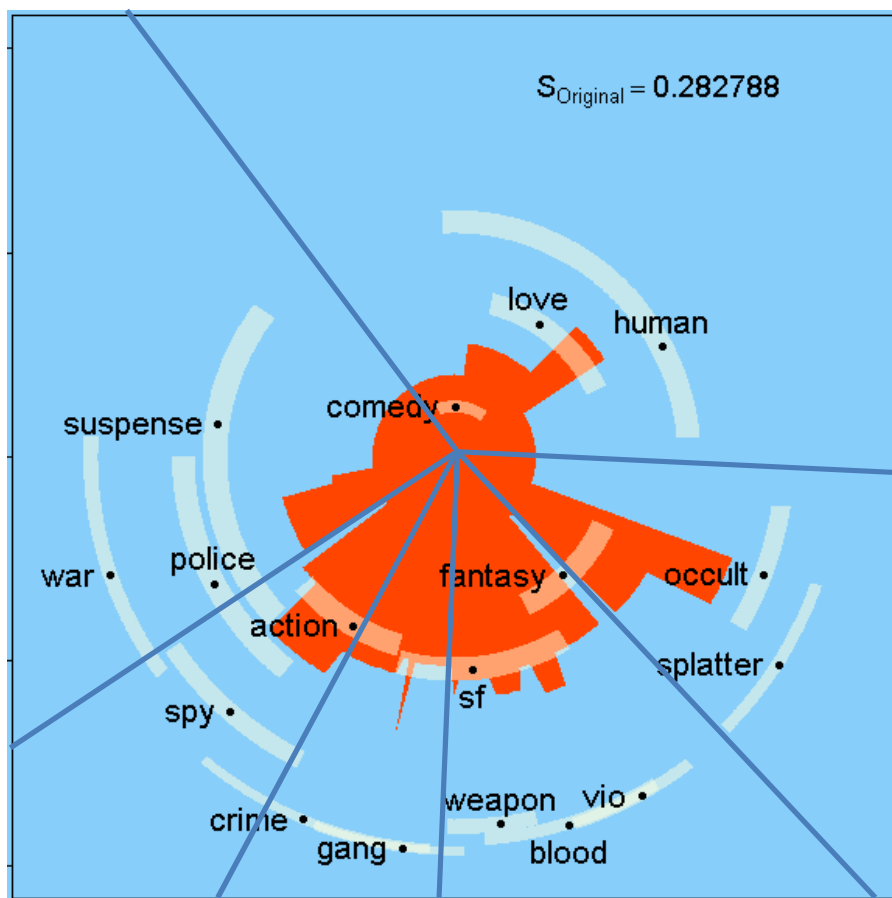


図 5-10 Polar Facet の境界線の一例

S_{Original} : オリジナルサンプルにおける S の値

最後に、表 5-2 で与えられた Homogeneity 係数に対して、その値を 1 から引いた値を非類似度とみなし、これをユークリッド距離行列に換算し、クラスター分析（Ward 法）を適用した結果を図 5-11 に示す。

デンドログラムを図に示すように横線で切った場合、大きく 4 つのクラスターに分類される。図 5-10 と概ね一致した分類となっているものの、図 5-10 では犯罪もの（Crime）とギャング・暴力団（Gang）が武器（Weapon）、暴力シーン（Vio）、流血シーン（Blood）等のクラスターから分離されている点、さらに警察（police）・戦争（war）がアクション（action）・スパイ（spy）の変数から分離され、サスペンス（suspense）は 1 つ離れている点が、クラスター分析の結果と異なることになる。RADXMS の結果を解釈する際に、こうしたクラスター分析の結果を補助的に利用することを考慮することも一法であると考えられる。

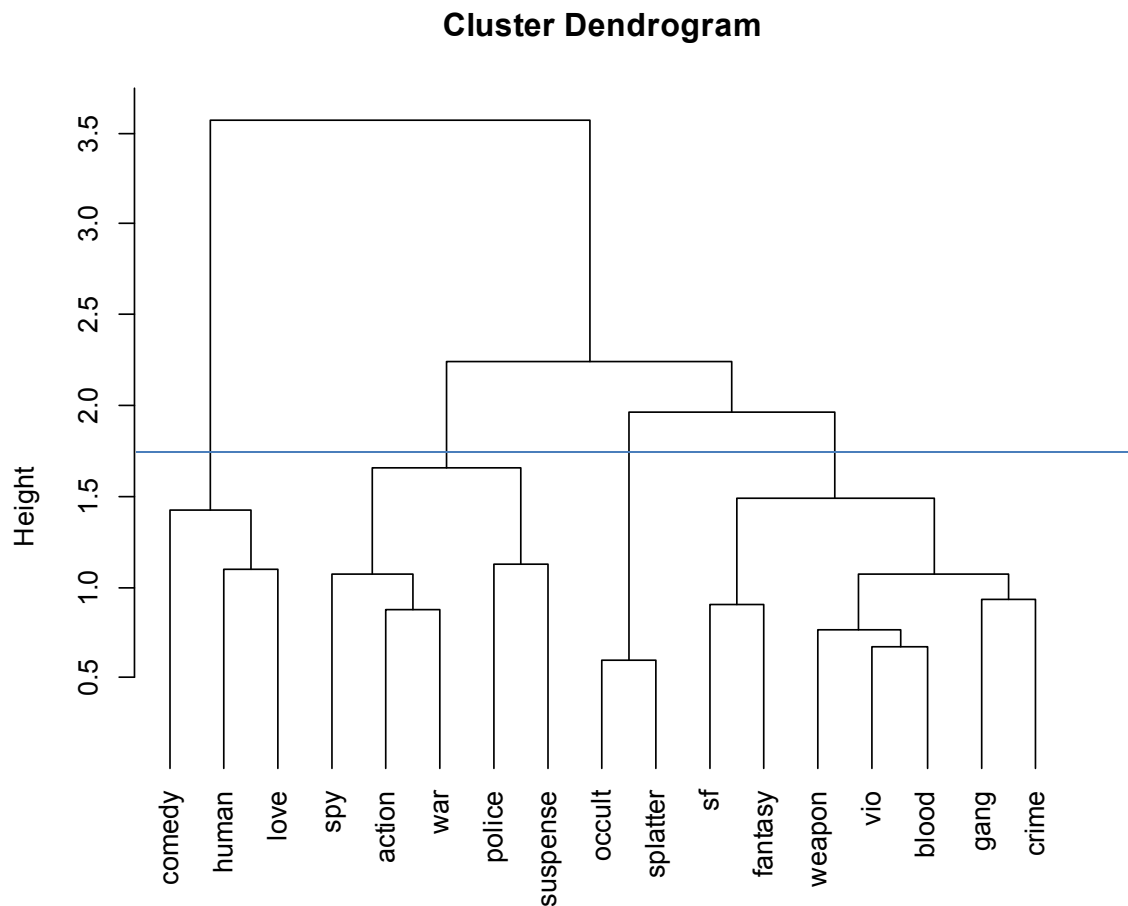


図 5-11 メディア暴力調査データ（17 変数）から計算した等質性係数を 1 から引いた値を非類似度として、ユークリッド距離に変換しクラスター分析（Ward 法）を適用した結果

5.4.3 他の視覚化手法との比較

次に、第 3 章で言及した視覚化手法である等質性分析（図 3-6）、Kruscal の MDS（図 3-9）、SSA（図 3-12）の結果について、変数の実質的意味を参照して、通常の見視による布置の解釈から得られると思われる変数クラスターを点線の円で囲んだ。さらに、変数の実質的意味と割合を参照して Radex モデルを想定し、RADXEA を援用して Polar Facet の境界線を描いた。そして、これらの結果と RADXMS の結果とを比較し、RADXMS の有用性について言及する。

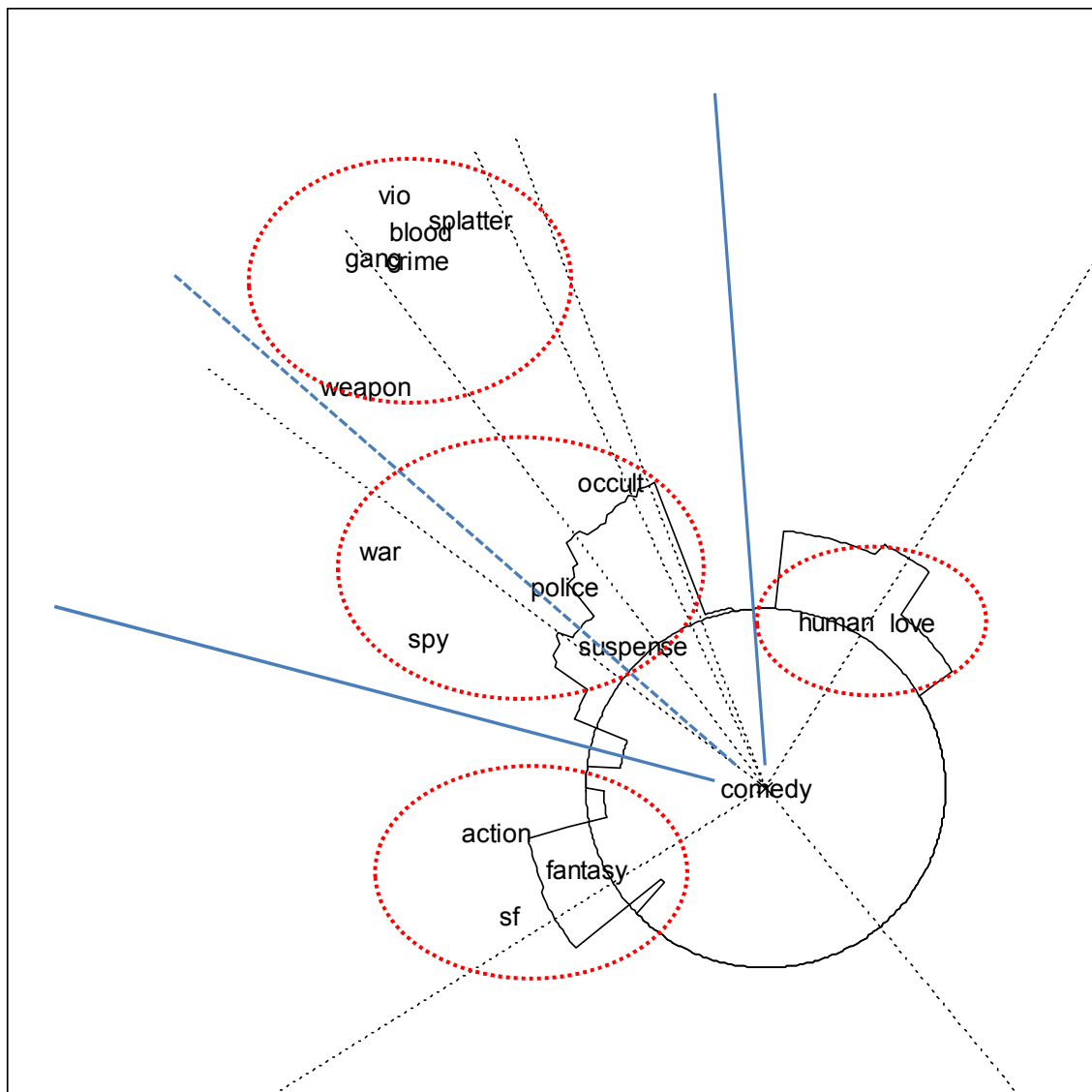


図 5-12 等質性分析の結果（図 3-6 の再掲）に Polar Facet の領域区分を策定した場合
（なお、点の布置はラベルの中央の位置となる）

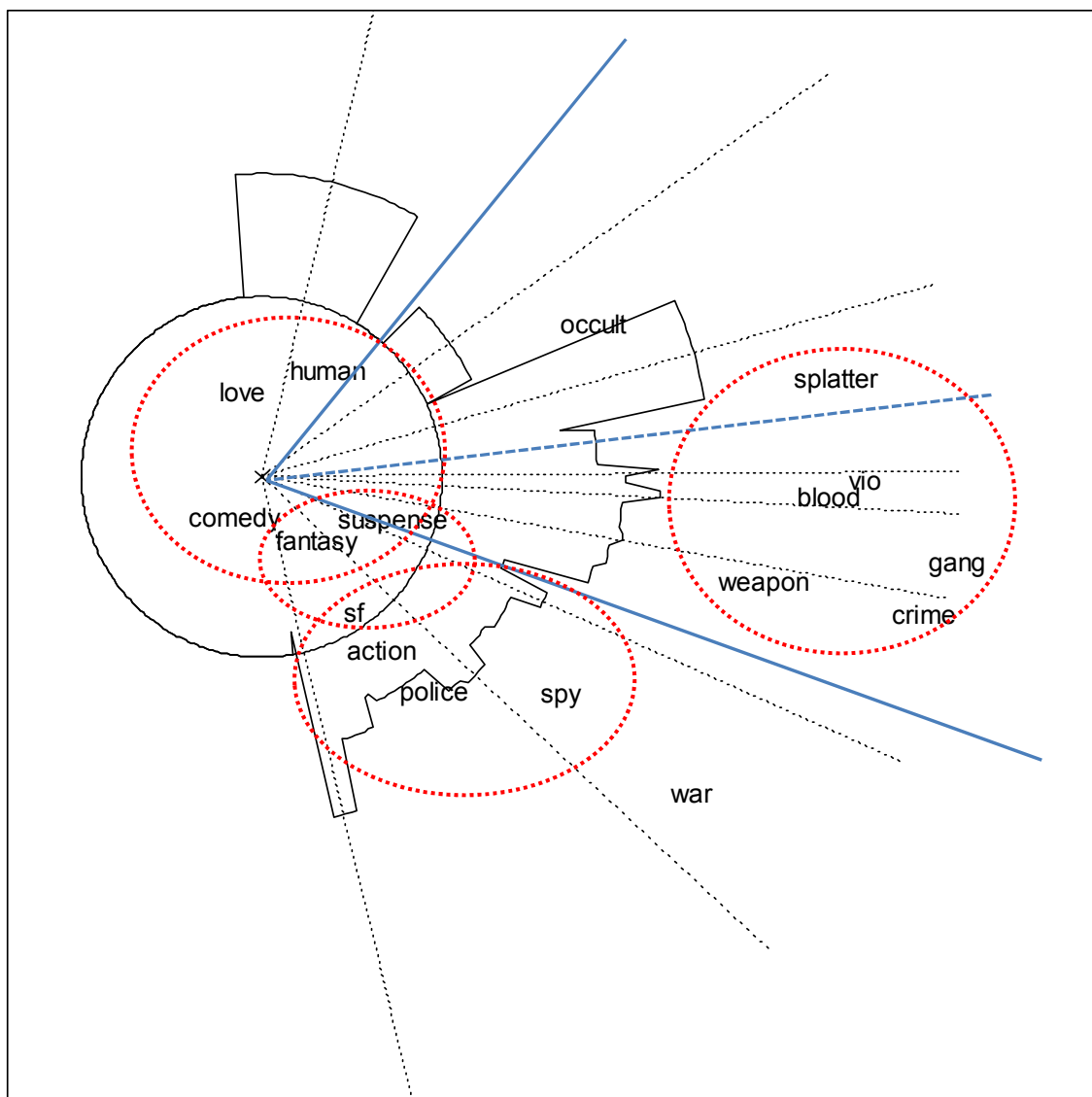


図 5-13 Kruskal の MDS の結果（図 3-9 の再掲）に Polar Facet の領域区分を策定した場合
（なお、点の布置はラベルの中央の位置となる）

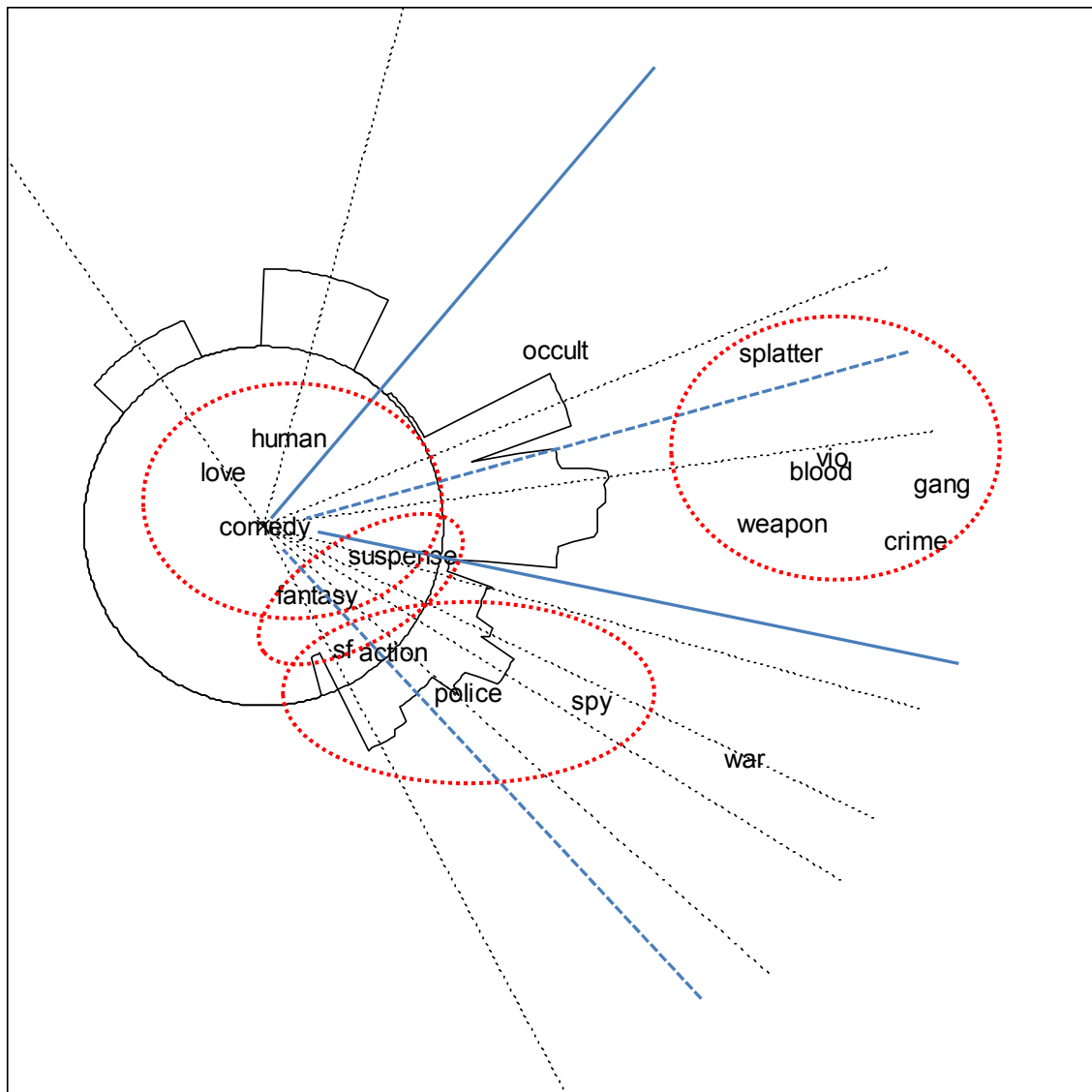


図 5-14 SSA の結果（図 3-12 の再掲）に Polar Facet の領域区分を策定した場合（図 4-9 と同一の図）（なお、点の布置はラベルの中央の位置となる）

まず、類似度の高い変数が互いの近傍に布置されるという前提に基づいた解釈から、通常の見方によって 3 つの結果を判断すると、各図において点線の円で囲まれた変数クラスターが確認される。等質性分析の結果（図 5-12）からは

- ①恋愛・ヒューマンドラマ系 (love、human)
- ②サスペンス・警察・スパイ・戦争・オカルト系 (suspense、police、spy、war、occult)
- ③犯罪・メディア暴力系 (weapon、crime、gang、blood、splatter、vio)
- ④ファンタジー・SF・アクション系 (fantasy、sf、action)

の 4 つのクラスターが観察される。

一方、Kruskal の MDS（図 5-13）及び SSA（図 5-14）の結果からは共通して、

- ①恋愛・ヒューマンドラマ・コメディ・ファンタジー・サスペンスもの（非暴力系）(love、human、comedy、fantasy、suspense)
- ②犯罪・メディア暴力・スプラッター系 (crime、gang、weapon、blood、vio、splatter)
- ③SF・アクション・警察・スパイ系 (sf、action、police、spy)

の概ね 3 つのクラスターが観察される。さらに、場合によっては、

- ④ファンタジー・SF・サスペンス系 (fantasy、sf、suspense)

をひとまとまりとして抽出できる可能性もある。

等質性分析と非計量的 MDS では、非暴力系の変数群の分離において、等質性分析の方の情報量が多いことがうかがえる。

次に、それぞれの結果について Radex モデルとして解釈した場合を記述する。目視による判断ではいずれも大きく 3 つの領域に区分できることがわかる。ただし、その区分の在り方は、やはり等質性分析の結果と非計量的 MDS で若干異なる。等質性分析の結果(図 5-12)では、

- ①恋愛・ヒューマンドラマ系 (love、human)
- ②オカルト・メディア暴力・犯罪・警察・スパイ・戦争系 (occult、splatter、blood、vio、crime、gang、police、weapon、spy、war)
- ③ファンタジー・SF・アクション系 (fantasy、sf、action)

に分けられる。一方、非計量的 MDS である Kruskal の MDS（図 5-13）及び Guttman の SSA（図 5-14）では、概ね

- ①恋愛・ヒューマンドラマ系 (love、human)
- ②オカルト・メディア暴力・犯罪系 (occult、splatter、blood、vio、weapon、crime、gang)
- ③アクション・警察・スパイ・戦争・SF系 (action、police、spy、war、sf、fantasy)

に分けられる。違いは、ファンタジー・SF系が等質性分析では独立した領域に布置されている一方、非計量的MDSでは警察・スパイものと同じ領域に布置されている点にみられる。

それぞれの図では、これらの目視によるおおまかな3領域は、実線で示された区分線に分けられている。KruskalのMDS及びSSAでは、場合によってはオカルト系とメディア暴力・犯罪ギャング系を分けることも可能であるが、目視の場合、その判断の根拠は恣意的とならざるを得ない。

そこで、第4章で提案したRADXEAを参照すると、太い点線でさらにPolar Facetの領域を区分できる。等質性分析の結果(図5-12)では、②オカルト・メディア暴力・犯罪ギャング・警察・スパイ・戦争系の領域から、サスペンス・警察・スパイ・戦争・武器の変数群を分けることが提案できる。また、KruskalのMDSの結果(図5-13)では、②オカルト・メディア暴力・犯罪ギャング系の領域から、オカルト・スプラッターの変数を分けることが提案できる。さらにGuttmanのSSAの結果(図5-14)でも、②オカルト・メディア暴力・犯罪ギャング系の領域から、オカルト・スプラッターの変数を分けることが提案できるほか、③アクション・警察・スパイ・戦争・SF系の領域からSF・ファンタジー系の領域を峻別することが提案できる。これらの結果からも、RADXEAが領域の区分に役立つことが示唆される。

なお、等質性分析の結果はKruskalのMDSやSSAの結果とは異なる。特に、オカルト系がメディア暴力の変数群から分離しにくい点に見出される。2次元平面において馬蹄形の歪みがある平面に、中心点から放射状のPolar Facetを重ねていることが問題である可能性もある。等質性分析の結果に対して、RADXEAをそのまま導入して解釈を進めることには注意を要するかもしれない。

一方、これらの3つの分析手法と比較することによって、Homogeneity係数によるRADXMSの利点は次のような点に見出される。

- (1) メディア暴力のクラスターとオカルト系のクラスターを明確に分けることができる
- (2) 犯罪の変数は、アクション、警察、スパイの変数との関連性があり、メディア暴力と警察・スパイものとの中間に位置する可能性を検討できる
- (3) 比較的割合の高いファンタジー、SFの変数が、必ずしもアクション・警察・スパイ・戦争系のクラスターに属する訳ではなく、オカルト系、メディア暴力系とGuttman Scaleを構成する可能性があることを検討できる

このように Homogeneity 係数を用いた提案手法を用いることで、Radex モデルに対するデータのあてはまりを検討する際、Polar Facet の領域の区分線について、データに基づいた検討が可能となるだけでなく、各領域にまたがる変数の存在についても洞察を得ることができる。以上の変数群の分類を表 5-3 に要約する。

表 5-3 分析手法ごとに解釈された変数群の分類結果のまとめ

分析手法	変数群(非暴力系)		変数群(アクション系)		変数群(メディア暴力系)		
① 等質性分析	恋愛・ヒューマン ドラマ		ファンタジー・SF・ アクション		サスペンス・ミタ ラー・オカルト	メディア暴力・犯 罪	
② Kruskal MDS	非暴力	ファンタジー・SF・ サスペンス	アクション・SF・警 察・スパイ			犯罪・メディア暴 力・スプラッター	
③ SSA	非暴力	ファンタジー・SF・ サスペンス	アクション・SF・警 察・スパイ			犯罪・メディア暴 力・スプラッター	
④ 等質性分析 + RADXEA	恋愛・ヒューマン ドラマ		ファンタジー・SF・ アクション		サスペンス・ミタ ラー	オカルト・メディア 暴力	
⑤ Kruskal MDS + RADXEA	恋愛・ヒューマン ドラマ		アクション・ミタ ラー・SF			メディア暴力・犯 罪	オカルト
⑥ SSA + RADXEA	恋愛・ヒューマン ドラマ		SF・ファンタジー	サスペンス・ミ タラー		メディア暴力・犯 罪	オカルト
⑦ RADXMS + RADXEA	恋愛・ヒューマン ドラマ		アクション・スパイ	ミタラー	犯罪	メディア暴力	オカルト

注) 変数群を大まかに非暴力系、アクション系、メディア暴力系の 3 つに分類しているが、テーマによってはこの 3 分類にまたがるものもあり、各テーマの配置はあくまで暫定的なものである。

この結果をみるとわかるように、Radex モデルを想定せずに①等質性分析、②Kruskal の MDS、③SSA を適用した場合、非暴力の変数群やオカルト・犯罪・メディア暴力など、変数が混在していることがわかる。RADXEA を併用した場合、⑥SSA+RADXEA ではアクション系の変数群からミタラーが分かれてきていることがわかる。また、メディア暴力系の変数群についても、④等質性分析+RADXEA ではミタラーが、⑤Kruskal の MDS+RADXEA と⑥SSA+RADXEA ではオカルトが分かれてきている。本論文で提案する⑦RADXMS+RADXEA では、アクション系ではアクション・スパイとミタラーに変数群が分離し、メディア暴力系の変数群についてもメディア暴力のほか、犯罪、オカルトの変数群に分かれており、メディア暴力の内容による下位分類が抽出されている。

さらに、分類された各変数群について Guttman Scale を構成している程度を検証するため、各変数群の Homogeneity 係数を調べた。2 変数の組み合わせから算出される Homogeneity 係数の平均を「平均 H 」、複数の変数から式[2.3.8]によって求められる Homogeneity 係数を「一括 H 」として、表 5-4 に示す。ただし、Homogeneity 係数の値に応じて色分けを行っており、値 0.3 未満は白色、0.3 以上 0.5 未満は黄色、0.5 以上 0.7 未満は緑色、0.7 以上は青色とした。「平均 H 」または「一括 H 」の値のうち、どちらか大きい方の値に基づいて、変数につい

でも同じ色を付している。また、変数群を分ける際に、区分線付近に存在する変数は括弧内に示し、その変数を含めた場合の Homogeneity 係数を示すと同時に、「w/o」のセルで括弧内の変数を除外した場合の Homogeneity 係数の値を併せて示した。

この結果をみると、Radex モデルを想定しない①等質性係数、②Kruskal の MDS、③SSA ではメディア暴力系の変数群で Homogeneity 係数の値が 0.4~0.5 を示すものの、概して値は高くない。Guttman Scale を構成する度合いの高い変数群の分類がうまくできていないことがわかる。RADXEA を利用した場合、利用しない場合よりも Homogeneity 係数の値の高い変数群の組み合わせが増えている。特に、⑤Kruskal の MDS + RADXEA と⑥SSA + RADXEA では、メディア暴力の変数群からオカルトのテーマが分離され、メディア暴力・犯罪の変数群の Homogeneity 係数の値もより高くなっている。しかし、④等質性分析+RADXEA では、メディア暴力の変数群から Homogeneity 係数の値の高い変数群が分離されているわけではない。このことから、RADXEA の利用は、等質性分析よりも非計量的 MDS の方が有効である可能性が示唆される。

今回提案する⑦RADXMS + RADXEA では、アクション系とメディア暴力系において青色や緑色で網掛けされた変数群が①~⑥の結果よりも多くみられ、Homogeneity 係数の値の高い変数群がもっとも多く得られていることがわかる。特に、アクション系やメディア暴力のテーマを見ると、アクション・スパイ (action, spy)、ミリタリー (police, war)、オカルト (occult, splatter)、犯罪 (crime, gang)、メディア暴力 (weapon, blood, vio) と変数群が分かれている。RADXEA による H_w の水準線をみると、例えば犯罪のテーマは、犯罪もの (crime) がアクションもの (action) に近接していたり、ギャング・暴力団 (gang) がメディア暴力のテーマに近接していたりするなど、分類の仕方には恣意的な判断が必要になるという面はあるが、これらの分類された変数群は、Egan et al. (1999) が示唆する暴力的コンテンツへの関心の分類であるミリタリズム、オカルティズム、犯罪同一性と類似したものとなっている。このことから、⑦RADXMS + RADXEA の手法は、従来の視覚化手法に比べて Homogeneity 係数の値の高い変数群の組み合わせを抽出することに成功している。すなわち、Guttman Scale を構成する変数群の分離がもっともよく行われていると考えられる。

表 5-4 各尺度の Homogeneity 係数

	非暴力系				アクション系				メディア暴力系				
	テーマ	変数	平均H	一括H	テーマ	変数	平均H	一括H	テーマ	変数	平均H	一括H	
①等質性分析	恋愛・ヒューマンドラマ	love, human	0.34	0.34	ファンタジー・SF・アクション	action, sf, fantasy	0.30	0.29	サスペンス・ミステリー・オカルト	suspense, police, spy, war, occult	0.22	0.22	
②Kruskal MDS	非暴力	love, human, comedy, fantasy, suspense, (sf) w/o (sf)	0.07 0.09	0.07 0.08	アクション・SF・警察・スパイ	sf, action, police, spy	0.32	0.27	犯罪・メディア暴力・スプラッター	vio, splatter, blood, gang, crime, (weapon) w/o (weapon)	0.49 0.43	0.49 0.43	
		ファンタジー・SF・サスペンス suspense, fantasy, sf	0.24	0.24									
③SSA	非暴力	human, love, comedy, (suspense, fantasy) w/o (suspense, fantasy) w/o (suspense) w/o (fantasy)	0.09 0.22 0.10 0.14	0.08 0.23 0.08 0.14	アクション・SF・警察・スパイ	sf, action, police, spy	0.32	0.27	犯罪・メディア暴力・スプラッター	weapon, blood, vio, crime, gang, splatter	0.49	0.49	
		ファンタジー・SF・サスペンス suspense, fantasy, sf	0.24	0.24									
④等質性分析 + RADXEA	恋愛・ヒューマンドラマ	love, human	0.34	0.34	ファンタジー・SF・アクション	fantasy, action, sf	0.30	0.29	オカルト・メディア暴力・犯罪・ミステリー	suspense, police, spy, occult, war, weapon, crime, gang, blood, splatter, vio	0.38	0.32	
										↓			
										サスペンス・ミステリー	suspense, police, spy, war, weapon	0.29	0.30
										occult, crime, gang, blood, splatter, vio, (suspense, police, weapon) w/o (suspense, police, weapon) w/o (suspense) w/o (police) w/o (weapon)	0.42 0.44 0.44 0.45 0.41	0.36 0.45 0.39 0.42 0.34	
										オカルト・メディア暴力			
⑤Kruskal MDS + RADXEA	恋愛・ヒューマンドラマ	love, human	0.34	0.34	アクション・ミステリー・SF	(comedy), fanatsy, (suspense), sf, action, police, spy, war w/o (comedy, suspense) w/o (comedy) w/o (suspense)	0.19 0.31 0.28 0.19	0.18 0.27 0.24 0.19	オカルト・メディア暴力・犯罪	occult, weapon, blood, splatter, vio, crime, gang	0.47	0.47	
										↓			
										オカルト	occult, splatter	0.75	0.75
										メディア暴力・犯罪	weapon, blood, vio, crime, gang	0.55	0.57
⑥SSA + RADXEA	恋愛・ヒューマンドラマ	love, human, (comedy) w/o (comedy)	0.22 0.34	0.23 0.34	アクション・ミステリー・SF	(comedy), fantasy, suspense, sf, action, police, spy, war w/o (comedy)	0.19 0.28	0.18 0.24	オカルト・メディア暴力・犯罪	occult, weapon, splatter, blood, vio, crime, gang	0.47	0.47	
						↓				↓			
						ファンタジー	fantasy, sf, (action) w/o (action)	0.30 0.46	0.29 0.46	オカルト	occult, splatter	0.75	0.75
									メディア暴力・犯罪	weapon, blood, vio, crime, gang	0.55	0.57	
⑦RADXMS + RADXEA	恋愛・ヒューマンドラマ	(comedy), love, human w/o (comedy)	0.22 0.34	0.23 0.34	アクション・スパイ・ミステリー	action, (police), spy, (crime) w/o (police, crime) w/o (police) w/o (crime) (suspense), police, war w/o (suspense)	0.44 0.60 0.51 0.39 0.25 0.33	0.36 0.60 0.55 0.34 0.26 0.33	オカルト	(fantasy), occult, splatter w/o (fantasy)	0.54 0.75	0.49 0.75	
										(fantasy), (sf), weapon, vio, blood w/o (fantasy, sf) w/o (fantasy) w/o (sf)	0.56 0.68 0.56 0.58	0.53 0.68 0.57 0.60	
											(action), (sf), crime, gang w/o (action, sf) w/o (action) w/o (sf)	0.49 0.44 0.49 0.55	0.33 0.44 0.49 0.54

注) 色分けは Homogeneity 係数の値 0.3 未満 (白)、0.3 以上 0.5 未満 (黄)、0.5 以上 0.7 未満 (緑)、0.7 以上 (青) である。w/o は括弧内の変数を除外した場合の値を示す。

表 5-5 RADXMS + RADXEA によって抽出された Guttman Scale の候補

テーマ	変数	平均 H	一括 H
アクション系	action (31), spy (9)	0.60	0.60
	action (31), spy (9), crime (2)	0.51	0.55
オカルト系	occult (10), splatter (3)	0.75	0.75
	fantasy (37), occult (10), splatter (3)	0.54	0.49
メディア暴力系	weapon (5), vio (3), blood (3)	0.68	0.68
	fantasy (37), sf (28), weapon (5), vio (3), blood (3)	0.56	0.53
犯罪系	action (31), crime (2), gang (2)	0.55	0.54

さらに、⑦RADXMS + RADXEA から得られた変数群のうち、Homogeneity 係数の値が比較的高かったものを表 5-5 に示す。ただし、変数の後ろの括弧内に二値変数の該当割合(%)を示す。

このように、該当割合の異なる変数群から、Homogeneity 係数の値が比較的高い変数の組み合わせが示唆された。これらの変数群を Guttman Scale の候補と考えることができよう。いくつかの変数群には、複数のテーマに出現する変数がある（例えば、action や fantasy）。これらは、該当割合が変数群の中では相対的に高い変数であり、RADXMS の結果では、より中央に近い部分に布置されるものである。該当割合の高い変数は、複数の Guttman Scale の候補の中に共通して出現することは十分考えられる。逆に言えば、共通して現れる変数群は、RADXMS において近接して布置されていることから、これらの Guttman Scale は、より類似性が高い変数群である、ということができよう。

なお、本章で利用したプログラムを付録 G.5 に示す。

5.5 まとめ

本章では、Guttman Scale を構成する構成度の指標として Loevinger (1947, 1948) による Homogeneity 係数を用いて、2次元平面において Radex モデルを視覚化する手法 (RADXMS) を提案した。変数の位置を極座標で表し、二値変数の割合を利用して、割合が高い変数ほど図の中心に、割合が低い変数ほど図の周縁に布置されるように、各変数の原点からの距離を定義した。Homogeneity 係数の高い変数同士が、図の中心から一直線に並ぶように、各変数間の距離の取り得る範囲を考慮し、Homogeneity 係数から定義した非類似度と平面上の変数間の距離の残差平方和によって目的関数を定義し、最適化を行って変数の布置を求めるとともに、目的関数によってデータの Radex モデルへのあてはまりの程度を評価する指標を提案した。続いて、Balanced Bootstrap によって布置の安定性を検討し、変数のクラスターが識別できることを確認した。さらに、布置した図に対して第 4 章で提案した Radex モデル探索支援手法 (RADXEA) を用いて、重みづけされた Homogeneity 係数の値の高低を原点の周りに描いた円に沿って表示し、Balanced Bootstrap による 90%信頼領域を表示することで、Polar Facet の領域区分をより客観的なデータによって検討することができることを示した。

また、二値データの視覚化に関わる従来の手法として、等質性分析、連関に Jaccard 係数を用いた非計量的多次元尺度法の Kruscal の MDS 及び Guttman の SSA をメディア暴力調査データに適用した結果を RADXMS の結果と比較した。これらの従来の視覚化手法に対する従来の解釈の方法を用いた場合では、Homogeneity 係数の値の高い変数群の抽出は困難であった。次にこれら従来の視覚化手法に RADXEA を併用した場合、本章で提案した手法 RADXMS と似たような変数群を導出した。しかし、提案した RADXMS はより Homogeneity 係数の値の高い変数群を導出するとともに、変数群の中間に位置する変数の存在を示唆した。特に、メディア暴力調査データに RADXMS を適用した結果、メディア暴力の下位分類を示唆する変数群の抽出がもっともよくなされていた。これらの結果から、RADXMS と RADXEA の利用は、視覚化によって Guttman Scale の候補を抽出する上で、より有効な手法であることが確認された。

第 6 章 総括

本研究では、二値データにおける変数間の関係を Guttman Scale の構成度を用いて分析し、2 次元平面における Radex モデルとして視覚化する手法の提案を行った。第 1 章では本論文の意義を簡潔に説明し、第 2 章及び第 3 章では本論文の鍵となる概念である Guttman Scale と Radex モデルについて説明し、従来の視覚化手法である等質性分析や非計量的多次元尺度法について説明した。新手法の提案は第 4 章及び第 5 章に記述した。第 4 章では、特定の連関係数を用いて非計量的多次元尺度法の一つである SSA（最小空間分析）を適用して得られた 2 次元の布置に対して、Radex モデルの探索を支援する手法を提案した。また第 5 章では、Guttman Scale の構成度の指標を用いて 2 次元平面に直接 Radex モデルを描き、Guttman Scale を構成する変数群を見出す方法を提案した。

本研究が提案する 2 つの手法、第 4 章で提案した Radex モデル探索支援手法 (RADXEA) 及び第 5 章で提案した Radex モデル多重尺度法構成法 (RADXMS) には、次のような共通した特徴がある。

- ・ 従来の手法とは異なり、変数間の関係を表現する際に二値変数の割合による順序性を明示的に視覚化する
- ・ Radex モデルを直接的に視覚化する
- ・ Guttman Scale の構成度を変数間の関係に利用する

続いて、次節以降において第 4 章及び第 5 章それぞれで提案した各手法の特徴を整理するとともに、今後の発展的な方向について言及する。

6.1 Radex モデル探索支援手法 (RADXEA)

この手法の有効性は、次のように整理できる。

- ・ Facet 理論における領域の区分の仕方をデータに基づいて支援することが可能である
- ・ 二値変数の割合と平面上の距離の関係を利用して、最適化によって Modular Facet の中心を決めることが可能である
- ・ Loevinger の Homogeneity 係数に重みづけを行うことにより、Guttman Scale の構成度を直接的に図示することが可能である
- ・ Guttman Scale の構成度を利用して、Polar Facet の領域区分の仕方についてデータによる指標を提示する
- ・ 重みづけの計算方法やパラメータの変更によって、自由に図を調整することが可能であるため、データに応じて分析者が適切に調整することができる

また、本分析手法について、次のような発展的な方向性が期待される。

- ・ 既存の非計量的 MDS の結果に Radex モデルを当てはめる場合、Radex モデルの中心が複数あるような場合に対応する手法を検討する必要がある
- ・ 二値変数の割合と Radex モデルの中心からの距離が単調減少の関係にならない場合、単調減少関係からの逸脱度に関する指標の作成も検討する必要がある

変数の布置が常に Radex モデルを形成するとは限らない。実際のさまざまな調査データを扱う中で、Radex モデルの中心が複数生じるようなデータも存在することが考えられる。その場合、二値変数の割合による等高線を描く方法が一つの解決法であろう。ただし、Guttman Scale を構成する点列間の関係をどのように定義するか、については新たに考えていく必要がある。そうした視覚化の手法も今後の課題となろう。

また、Radex モデルにおいて二値変数の割合と中心からの距離が完全な単調減少の関係を満たさないことも想定される。そのような場合に、単調減少の関係から逸脱する度合いを定義することにより、その度合いのもっとも小さい中心を選んで Radex モデルを検討することが可能となるだろう。また、割合に誤差を仮定する場合、割合に基づく順序関係に SSA における弱単調性のような緩い順序関係を適用することで、割合のわずかの差を同一の順序関係として扱う、ということも一つの解決方法となり得る。そのため、Radex モデルにおいて割合を反映する Modular Facet を数量化する際に非計量的な手法を利用することも、意義ある今後の課題となろう。

6.2 Radex モデル多重尺度構成法 (RADXMS)

この手法の有効性は、次のように整理できる。

- ・ Guttman Scale を構成する変数群を Radex モデルの中心から一つの直線状に並ぶ点列として視覚的に把握することが可能となる
- ・ Guttman Scale を構成する変数による点列が複数ある場合、それら同士の関係を中心周りの角度差によって表現することが可能である
- ・ データの Radex モデルに対する当てはまりの度合いを表す指標を定義し、複数の布置について当てはまりの良さを比較することが可能である
- ・ 第 4 章の RADXEA を併用することにより、Polar Facet の領域区分を検討することができる

また、本分析手法について、次のような発展的な方向性が期待される。

- ・ 二値変数の割合及び Guttman Scale の構成度の指標における順序関係によって布置を

決める非計量的な手法の検討

- ・ Guttman Scale の構成に寄与しない変数を除外し、Guttman Scale の構成度の高い複数の変数列のみを抽出して布置する手法の検討

第 5 章で提案した Radex モデル多重尺度構成法は、二値変数の割合、Guttman Scale の構成度を関数によって変換し、その変換された値の大小関係を反映するように空間内に布置して Radex モデルを作成する。二値変数の割合や Guttman Scale の構成度におけるわずかの差を誤差として扱う場合、SSA における弱単調性関数による変換を行い、その順序関係を反映するように空間内の距離を決めて布置させるのも一つの解決方法である。そのため、非計量的な手法を検討することも今後の課題となる。

課題の 2 点目については、空間内の点ごとに式[5.2.12]の値を計算できるため、その値の大きな点から除外していけば、与えられたデータの中で Guttman Scale 構成度の高い変数の組み合わせのみを選択できる。しかし、どの値までを除外すればよいか、という基準を考える必要がある。例えば、全ての変数 i について計算した η_i^2 をその値の大きい順にプロットし、変化の大きい部分に着目して基準を設定する、といった方法などが考えられよう。この点についても、今後の研究課題となりうる。

6.3 本研究における提案手法の意義

本研究で提案した Radex 探索支援手法 (RADXEA) 及び Radex 多重尺度構成法 (RADXMS) の意義は、次のようにまとめられる。

二値データとして測定された行動科学的変数に対して、二値変数の該当割合を考慮することによって各項目の頻度を踏まえ、Guttman Scale の構成度の高い変数群を視覚化することによって、Guttman Scale の抽出を可能にしたことである。Facet 理論で示唆される領域モデルの一つである Radex モデルは、こうしたデータの視覚化に有用であり、その Radex モデルを直接的に視覚化することを可能とした点も提案手法の有用性を示すものであるといえる。これは、既に検証すべき明確な仮説がある研究よりも、データの中から探索的に仮説を構築する際の尺度構成に有用な手法である。

RADXEA や RADXMS では、いくつかのパラメータを調節して結果を求めている。データから最適化等を行うことで自動的にパラメータを決める方が、手法を適用する際に便利であり、結果の報告も統一的に行うことができよう。しかし、本手法が焦点を当てる探索的な研究の場合、データの特性に応じて分析を少しずつ変えながら、データの特性を明らかにする、という側面も重要である。今回の RADXEA や RADXMS では、パラメータの変更は布置されるデータの密度と特に関連すると思われる。得られたデータがこういった特性を示すのかを考えながら、パラメータを調整する本手法は、探索的な研究に向いているといえる。

最後に、現象として出現頻度の非常に低い事象を研究対象とする場合、該当割合の低い

変数が現象を理解する鍵となることもある。これら出現頻度の低い変数と他の変数との関係が反映された結果を検討することで、重要な知見を得ることができる。RADXEA や RADXMS では該当割合の非常に低い変数は周縁部に布置されるものの、分析から切り捨てられることがない。この特徴は、サンプル自体が少なくなりがちな発生頻度の低い現象を研究の対象とする場合でも適用することが可能である。この点も、本研究の提案手法の意義として付け加えることができよう。

6.4 多値データへの適用

本研究では、もっぱら二値データを分析の対象として議論してきた。しかし、三値以上のカテゴリからなる多値データについても、本研究の提案手法を適用することは可能である。RADXEA においては、多値データに相関係数等の連関係数を用いることによって MDS の布置を得ることができるが、その後に Guttman Scale の構成度が計算できれば、RADXEA が利用できる。

今回、Guttman Scale の構成度として利用した Homogeneity 係数については、実際に多値データに適用する方法が提案されている (Molenaar, 1991)。多値データのカテゴリ数を k とすると、2 変数のクロス集計表には k^2 個のセルが存在する。また、変数のカテゴリによってその変数を二値化する方法は $k-1$ 通りある。それらの二値変数は項目段階変数 (item step variable) と呼ばれる (van Schuur, 2003)。そうすると 2 変数のクロス集計表において、項目段階変数は $2k-2$ 個存在する。このことから、これらの項目段階変数の順序に従って区分される領域は $2k-1$ 個存在することになる。つまり、クロス集計表のセルのうち、この項目段階変数の順序に従って区分される領域以外のセル数が Guttman Error となるセル数となり、その数は k^2-2k+1 個となる。これらのセルに対して、そのセルが Guttman Error となる項目段階変数によるクロス集計表の数を重みとして、実際のエラー度数の重みづけによる総和 $\text{Err}(\text{obs}, \text{weighted})$ と、周辺度数から算出したエラーセルの期待度数に関する重みづけによる総和 $\text{Err}(\text{exp}, \text{weighted})$ によって、多値データ変数間の Homogeneity 係数は、

$$[6.3.1] \quad H = 1 - \frac{\text{Err}(\text{obs}, \text{weighted})}{\text{Err}(\text{exp}, \text{weighted})}$$

によって算出できる (van Schuur, 2003; Sijtsma & Molenaar, 2002)。

このようにして算出された Homogeneity 係数を利用すれば、RADXEA は適用が可能といえる。ただし、Radex モデル多重尺度構成法については、各変数の割合をどのように視覚化に結び付けるかについて考える必要がある。二値データの場合、ある変数の割合 π_i は「該当あり」の割合を π_1 、「該当なし」の割合を $\pi_0 = 1 - \pi_1$ とするとき、 $\pi_i = \pi_1$ とすることにより変数の同心円中心からの距離を算出した。このとき値 π_1 は、「該当あり」に値 1、

「該当なし」に値 0 を割り当て、各カテゴリの重み付き平均 ($\pi_i = \pi_1 = \pi_0 \times 0 + \pi_1 \times 1$) として計算される。そこで、多値データについても同様に平均を用いて変数の「割合」を定義することができよう。例えば 3 件法によって測定された三値データについて、「該当あり②」(割合 π_2) に値 2、「該当あり①」(割合 π_1) に値 1、「該当なし」(割合 $\pi_0 = 1 - \pi_2 - \pi_1$) に値 0 を割り当てるとき、変数の「割合」 π_i は、

$$[6.3.2] \quad \pi_i = \pi_0 \times 0 + \pi_1 \times 1 + \pi_2 \times 2$$

として定義されよう。ただし、各カテゴリに割り当てる数値をどのように決めるべきか、については検討の余地があろう。

これとは別に、先ほど言及した項目段階変数ごとに Homogeneity 係数を算出でき、二値変数である項目段階変数の場合、その割合はそのまま利用できる。同一変数に属する項目段階変数同士の Homogeneity 係数は常に 1 となるため、Radex モデル多重尺度法に適用すれば、それらの項目段階変数は常に中心からの半直線上にプロットされる。そこで、多値データ変数を項目段階変数に分解して、RADXMS を適用するという方法も考えられる。このとき、2 変数間の関係が項目段階変数同士の関係それぞれに同一の重みをつけて最適化がなされることの是非を検討していく必要がある。

6.5 データの視覚化の意義

本研究で提案した 2 つの分析手法について、その主たる目的であるデータの視覚化という観点から振り返ることとしたい。データの視覚化は、データ空間 (Data Space) をユークリッド空間などの表現空間 (Representation Space) へとマッピングする (de Leeuw, 1998)。このデータの視覚化を行う技法 (Technique) としてどういった性質を備えていることが望ましいか、という観点から、de Leeuw (1998) は次の 5 つの特徴を挙げている。

視覚化の技法の持つべき 5 つの特徴

- ① データに対して最大の縮約を行う
- ② 科学による事前知識をより多く取り込む
- ③ データの背景やノイズから、安定した興味深い効果を取り出す
- ④ データのもっとも重要な側面を示す
- ⑤ データ空間から表現空間への変換が微分可能であり (Differential) 滑らか (Smooth) である

そこで、今回の提案手法が、これら 5 つの条件を満たしている、ということを見ていくこととする。

特徴①について。データの視覚化は、データに含まれる特定の関係を単純化して示すこ

とによって観察者が把握しやすくするように支援する。そのため、データの情報の縮約を可能とすべきである。本研究における MDS や RADXMS は、データ間の関係を 2 次元平面上（場合によっては 3 次元空間内）に布置させることで次元の縮約を行っており、特徴①に当てはまる。

特徴②について。科学による事前知識を取り込むという点に関して、本研究では二値変数の割合に関する知識を視覚化に取り込んだ部分が該当する。二値変数の割合は、その変数が示す特徴の出現率の高低を示し、その出現率の高低が行動特徴の特殊化といった概念の中で意味をもつことにより、二値変数の割合の違いが変数群（尺度）のもつ特徴を理解する鍵となる。そして、それら尺度同士の関係は Radex モデルによって示すことができる。その意味で、こうした実質科学的な知識を視覚化に取り入れることを可能にしたといえる。

特徴③について。一見すると雑然と配置されているように見える 2 次元平面上の変数に対して、Radex モデルを通して眺めた時、Guttman scale という一次元構造を満たす変数間の関係が浮き上がる、という意味で、本研究の提案手法は視覚化技法の特徴③を満たしている。

特徴④について。本研究が取り扱うデータの重要な側面として、割合の大きく異なる変数同士においても一次元性が成立する、という Guttman Scale の性質が挙げられる。本研究の提案手法である RADXEA 及び RADXMS は、Guttman Scale を構成する変数群を Radex モデルの中心から周縁へと向かう直線上に並ぶ点列として視覚化することを可能とした。その意味で、特徴④を満たしている。

特徴⑤について。この特徴は安定したデータ表現を意味し、具体的にはデータ空間と表現空間の間の連続性について言及している。RADXEA 及び RADXMS についていえば、変数間の類似度（もしくは非類似度）、変数間の Guttman Scale の構成度、さらに RADXEA においては所与の MDS 等の空間布置をこの文脈におけるデータ空間と考えると、データ空間から表現空間へと変換する関数として連続し微分可能な関数を利用することによって、与えられたデータに対応した表現空間における変換値を計算する。そのため、データ空間のわずかな変化に応じて、わずかに変化した表現空間の位置が決まる。つまり双方の空間の関係は微分可能な関数で定義されており、滑らかな関係（de Leeuw, 1998）である。ただし、RADXMS については、二値データから Guttman Scale 構成度の指標を計算するため、二値データをデータ空間と考える場合、データ空間と表現空間の間の微分可能で滑らかな関係には該当しない、といえる。

以上から、本研究で提案した RADXEA や RADXMS は de Leeuw（1998）が視覚化の技法として備えるべき 5 つの特徴にほぼ合致していると考えられる。その意味で、本研究で提案した手法は、データの視覚化手法として望ましい性質を持っているといえよう。

しかし、この最後の特徴⑤については、データ空間と表現空間をつなぐ変換として連続的でない技法が有用である場合についても、検討の余地がある。RADXMS に関する前節でも言及したように、例えば、データのサンプルサイズが大きい場合、各項目の該当割

合は連続的に変化するとは言いがたく、さらにクロス集計表から算出される Guttman Scale の構成度 H_{ij} の値も連続的に変化することが期待しにくい。そのため H_{ij} のデータ空間と表現空間の関係を計量的に表現せず、 H_{ij} データの順序関係のみを保持するかたちで関連づける非計量的な手法の利用も考えられる。それは、データ表現における頑健性と関連する。

データによっては、データの値の変化に敏感でない方がよい場合もある。例えば、行動科学的データがそれに該当すると思われる。人間の行動を題材とする心理学等の行動科学では、客観的な指標によって測定できない事象に対して、何らかの数値を割り当てることによって、計量的なものとして取り扱うことができるようにしている。そのため、数量化されたデータはせいぜい順序情報のみであるということも少なくない。その場合、その順序情報のみを利用する非計量的な手法の活用が考えられる。非計量的な手法は、数値の異なるデータセットであっても順序が同じである限り、同じ結果を返す。つまり、データに誤差を含んだ変動があっても、順序情報が同一である限り、同じ結果を返すことから、データの変動に対して「鈍感」な手法といえるわけである。データのサンプルサイズが小さい場合、こうした「鈍感」な手法の方が頑健な結果が得られると考えられる。

一方、サンプルサイズが十分に大きい場合、順序情報のみを利用する非計量的な手法は、データに含まれる情報量を切り捨ててしまうことになる。その場合は、むしろデータの情報を積極的に活用する方向性を検討すべきであろう。各二値変数の割合は比率尺度であることから、例えば、その変数の割合を所与として、クロス集計表における 2 変数ともに該当する割合をパラメータとして扱うことによって尤度を計算することにより、尤度を最大化する布置を最適解とする、という方法が考えられる。これはまさしく de Leeuw (1998) のいうデータ空間と表現空間の連続性であり、モデルから計算される Guttman Scale の構成度、さらには変数の該当割合の推定値を求めることも可能となる。今後検討すべき課題であると思われる。

最後に、本研究の中心的テーマであるデータの視覚化について、改めてその意義について言及しておきたい。データの視覚化は、観察の対象となる現象や、そこに含まれる現象間の関係を新たな観点で見ることを可能にする (Friendly, 2006)。それによって、データの視覚化は観察者に新たな洞察をもたらし、新たな仮説の生成へとつながってゆく可能性がある (Keim, 2001)。

今回提出する分析の結果である 2 次元平面上のデータの布置は、従来の目視のみによる把握では解釈が難しかったり、場合によっては恣意的になったりするという問題があった。さらに、観察者が仮説に基づいて眺めるとき、その仮説に対する意識が働くばかりに、仮説による解釈以外の解釈可能性は見落とされがちとなる。視覚化された結果をみるときに、何らかの枠組みが提示されることによって、解釈が容易になったり、あるいは想定していたものとは違った解釈が得られたりすることもあり得る。これはまさしく、新たな仮説の生成につながる可能性を提供しているわけである。こうした機能を果たすことで、視覚化の意義は高まる。

本研究は特に、二値データの中から一次元性をもつ尺度を構成する変数群を探し出し、**Guttman Scale** としての尺度の候補となりうる変数群を特定するとともに、それら尺度を構成する変数群同士の関連性を把握することが目的であった。本研究で提案された手法によって、二値データの変数間の関係から変数群を見出すことが可能となる。それにより、現象の構造や下位分類の存在を把握できるだけでなく、その下位分類同士の関係を直感的に把握できる。そのため、現象の構造について一步進んだ仮説の構築が可能となる。今後、行動科学分野において既存の尺度が存在しない現象について、特に出現率の違いを利用した現象理解を試みる場合、本研究における提案手法は、現象の構造的理解を深めることにつながり、新たな仮説の生成に寄与する手法として活用されることが期待できよう。

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—. 東京：朝倉書店.
- Alison, L. J., Snook, B., Stein, K. L. (2001). Unobtrusive measurement: using police information for forensic research. *Qualitative Research*, 1, 241-254.
- Anderson, C. A., Berkowitz, L., Donnerstein, E., Huesmann, L. R., Johnson, J. D., Linz, D., Malamuth, N. M., & Wartella, E. (2003). The influence of media violence on youth. *Psychological Science in the Public Interest*, 4, 81-110.
- 東美晴. (1999). 暴力観とメディアの嗜好：中学2年生の意識調査から. 長寿社会研究所・家庭問題研究所研究年報 (5), 89-97.
- Baroni-Urbani, C., & Buser, M. W. (1976). Similarity of binary data. *Systematic Zoology*, 25, 251-259.
- Borg, I., & Groenen, P. J. F. (1998). Regional interpretations in multidimensional scaling. In J. Blasius & M. Greenacre (Eds.), *Visualization of categorical data* (pp. 347-364). San Diego, CA: Academic Press.
- Borg, I., & Shye, S. (1995). *Facet theory: Form and content*. Thousand Oaks, CA: Sage.
- Canter, D. (Ed.). (1985). *Facet theory: Approaches to social research*. New York : Springer-Verlag.
- Canter, D. (2000). Offender profiling and criminal differentiation. *Legal and Criminological Psychology*, 5, 23-46.
- Canter, D. V., Bennell, C., Alison, L. J., & Reddy, S. (2003). Differentiating sex offences: A behaviorally based thematic classification of stranger rapes. *Behavioral Sciences & the Law*, 21, 157-174.
- Charles, K. E., & Egan, V. (2009). Sensational interests are not a simple predictor of adolescent offending: Evidence from a large normal British sample. *Personality and Individual Differences*, 47, 235-240.
- Chen, L., & Buja, A. (2009). Local multidimensional scaling for nonlinear dimension reduction, graph drawing and proximity analysis. *Journal of the American Statistical Association*, 104, 485, 209-219.
- Clogg, C. C., & Sawyer, D. O. (1981). A comparison of alternative models for analyzing the scalability of response patterns. *Sociological Methodology*, 12, 240-280.
- Comstock, G. (2008). A sociological perspective on television violence and aggression. *American Behavioral Scientist*, 51, 1184-1211.
- Cox, T. F., & Cox, M. A. A. (1994). *Multidimensional scaling*. London: Chapman & Hall.
- Coxon, A. P. M. (1982). *User's guide to multidimensional scaling*. London: Heinemann.
- de Leeuw, J. (1998). Here's looking at multivariables. In J. Blasius & M. Greenacre (Eds.), *Visualization of categorical data* (pp. 1-11). San Diego, CA: Academic Press.

- Dice, L. R. (1945). Measure of the amount of ecologic association between species. *Ecology*, 26, 297-302.
- Driver, H. E., & Kroeber, A. L. (1932). Quantitative expression of cultural relationship. *The University of California Publications in American Archaeology and Ethnology*, 31, 211-256.
- Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hill.
- Egan, V., Auty, J., Miller, R., Ahmadi, S., Richardson, C., & Gargan, I. (1999). Sensational interests and general personality traits. *Journal of Forensic Psychiatry*, 10, 567-582.
- Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, 7, 117-140.
- Friendly, M. (2008). A brief history of data visualization. In C.-h. Chen, W. K. Härdle, A. Unwin (Eds.), *Handbook of data visualization* (pp. 15-56). Berlin: Springer-Verlag Berlin Heidelberg.
- Funk, J. B., Hagan, J., Schimming, J., Bullock, W. A., Buchman, D. D., & Myers, M. (2002). Aggression and psychopathology in adolescents with a preference for violent electronic games. *Aggressive Behavior*, 28, 134-144.
- Gifi, A. (1990). *Nonlinear multivariate Analysis*. Chichester: John Wiley & Sons.
- Gleason, H. A. (1920). Some applications of the quadrat method. *Bulletin of the Torrey Botanical Club*, 47, 21-33.
- Goodenough, W. M. (1944). A technique for scale analysis. *Journal of Educational and Psychological Measurement*, 4, 179-190.
- Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49, 732-764.
- Green, B. F. (1956). A method of scalogram analysis using summary statistics. *Psychometrika*, 21, 79-88.
- Greenacre, M. (2007). *Correspondence analysis in practice* (Second edition). Boca Raton, FL: Chapman & Hall/CRC.
- Guttman, L. (1944). A basis for scaling qualitative data. *American Sociological Review*, 9, 139-150.
- Guttman, L. (1950). The basis for scalogram analysis. In S. A. Stouffer and others (Eds.), *Measurement and prediction: Studies in social psychology in World War II* (Vol. 4). Princeton: Princeton University Press.
- Guttman, L. (1954). A new approach to factor analysis: The Radex. In: Lazarsfeld, P. F. (ed.) *Mathematical thinking in the social sciences* (pp. 258-348). New York: The Free Press.
- Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, 33, 469-506.
- Guttman, L. (1982). Facet theory, smallest space analysis, and factor analysis. *Perceptual and Motor Skills*, 54, 491-493.
- Guttman, R., & Greenbaum, C. W. (1998). Facet theory: Its development and current status.

- European Psychologist*, 3, 13-36.
- Hakkanen, H., Hagelstam, C., Santtila, P. (2003). Stalking actions, prior offender-victim relationships and issuing of restraining orders in a Finnish sample of stalkers. *Legal and Criminological Psychology*, 8, 189-206.
- Hamman, U. (1961). Merkmalbestand und Verwandtschaftsbeziehungen den Farinosae: Ein Betrag zum System der Monokotyledonen. *Willdenowia*, 2, 639-768.
- Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychology Measurement*, 9, 139-164.
- 林知己夫. (1974). 数量化の方法. 東京：東洋経済新報社.
- 林知己夫・鮑戸弘. (1976). 多次元尺度解析法. 東京：サイエンス社.
- 岩坪秀一. (1987). 数量化法の基礎. 東京：朝倉書店.
- Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Société vaudoise des sciences naturelles*, 44, 223-271.
- Jackson, J. M. (1949). A simple and more rigorous technique for Scale Analysis. In *A manual of scale analysis* (Part II). Montreal: McGill University.
- Johnston, D. D. (1995). Adolescents' motivations for viewing graphic horror. *Human Communication Research*, 21, 522-552.
- 川喜田二郎. (1967). 発想法—創造性開発のために. 東京：中央公論社
- Keim, D. A. (2001). Visual exploration of large data sets. *Communications of the ACM*, 44, 38-44.
- 警察庁・科学警察研究所. (2000). 最近の少年による特異凶悪事件の前兆等に関する緊急調査報告書. 東京：警察庁.
- 木村通治・真鍋一史・安永幸子・横田賀英子. (2002). ファセット理論と解析事例：行動科学における仮説検証・探索型分析手法. 京都：ナカニシヤ出版.
- Kirsh, S. J. (2006). *Children, adolescents, and media violence: A critical look at the research*. Thousand Oaks, CA: Sage.
- Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1-28.
- Linting M., Meulman J. J., Groenen P. J. F., & Kooij A. J. van der. (2007a). Nonlinear principal components analysis: Introduction and application. *Psychological Methods*, 12, 336-358.
- Linting M., Meulman J. J., Groenen P. J. F., & Kooij A. J. van der. (2007b). Stability of nonlinear principal components analysis: An empirical study using the balanced bootstrap. *Psychological Methods*, 12, 359-379.
- Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. *Psychological Monographs*, 61(4), (Whole No. 285).
- Loevinger, J. (1948). The technique of homogeneous tests compared with some aspects of "scale analysis" and factor analysis. *Psychological Bulletin*, 45, 507-529.

- McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, 34, 100-117.
- McGee, J. P., & DeBernardo, C. R. (1999). The classroom avenger. *The Forensic Examiner*, 8, 1-16.
- Menzel, H. (1953). A new coefficient for scalogram analysis. *Public Opinion Quarterly*, 17, 268-280.
- Michailidis, G., & de Leeuw, J. (1998). The Gifi system of descriptive multivariate analysis. *Statistical Science*, 13, 307-336.
- Miyadera, T. (2005). An exploratory study of juvenile homicide in Japan (Unpublished master's thesis). University of Birmingham, Birmingham, UK.
- Miyadera, T. (2014). An examination of regionality in a configuration of smallest space analysis using Loevinger's homogeneity coefficient. *Quality and Quantity*, 49, 1203-1218.
- Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Netherlands: Mouton & Co.
- Molenaar, I. W. (1991). A weighted Loevinger H-coefficient extending Mokken scaling to multicategory items. *Kwantitatieve Methoden*, 12(37), 97-117.
- 西里静彦. (2007). データ解析への洞察—数量化の存在理由(K. G. りぶれつと). 西宮：関西学院大学出版会.
- 西里静彦. (2010). 行動科学のためのデータ解析：情報把握に適した方法の利用. 東京：培風館.
- Ochiai, A. (1957). Zoogeographic studies on the soleoid fishes found in Japan and its neighboring regions. *Bulletin of the Japanese Society for Fish Science*, 22, 526-530.
- 岡太彬訓・今泉 忠. (1994). パソコン多次元尺度構成法. 東京：共立出版株式会社.
- Rand, W. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66, 846-850.
- Rogers, D. J., & Tanimoto, T. T. (1960). A computer program for classifying plants. *Science*, 132, 1115-1118.
- Russel P. F., & Rao, T. R. (1940). On habitat and association of species of anopheline larvae in southeastern Madras. *Journal of the Malaria Institute of India*, 3, 153-178.
- Sagi, P. C. (1959). A statistical test for the significance of a coefficient of reproducibility. *Psychometrika*, 24, 19-27.
- Santtila, P., Hakkanen, H., Alison, L., & Whyte, C. (2003). Juvenile firesetters: Crime scene actions and offender characteristics. *Legal and Crimiological Psycholpgy*, 8, 1-20.
- Shye, S., Elizur, D., & Hoffman, M. (1994). *Introduction to facet theory: Content design and intrinsic data analysis in behavioural research*. Thousand Oaks, CA: Sage Publications, Inc.
- Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand Oaks, CA: Sage Publications, Inc.
- Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships.

- The University of Kansas Science Bulletin*, 38, 1409-1438.
- Sokal, R. R., & Sneath, P. H. (1963). *Principles of numerical taxonomy*. San Francisco: W. H. Freeman and Company.
- Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Philosophical Transactions of the Royal Society of London*, 5, 1-34.
- StataCorp. (2007). *Stata base reference manual* (Release 10). College Station, TX: StataCorp LP.
- Stouffer, S. A., Guttman, L., Suchman, E. A., & Lazarsfeld, P. F. (1950). *Measurement and prediction: Studies in social psychology in World War II*. Princeton: Princeton University Press.
- Strizhakova, Y., & Krcmar, M. (2007). Mood management and video rental choices. *Media Psychology*, 10, 91-112.
- Suchman, E. A. (1950). The scalogram board technique for scale analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, & P. F. Lazarsfeld, *Measurement and prediction: Studies in social psychology in World War II* (pp. 91-121). Princeton: Princeton University Press.
- 高根芳雄. (1980). 多次元尺度法. 東京：東京大学出版会.
- 竹内啓編. (1989). 統計学辞典. 東京：東洋経済新報社.
- 豊田秀樹. (2002). 項目反応理論 [入門編] —テストと測定の科学—. 東京：朝倉書店.
- Van Schuur, W. H. (2003). Mokken scale analysis: Between the Guttman scale and parametric item response theory. *Political Analysis*, 11, 139-163.
- van Schuur, W. H. (2011). *Ordinal item response theory: Mokken scale analysis*. Thousand Oaks, CA: Sage Publications, Inc.
- Warrens, M. J. (2008). On association Coefficients for 2×2 tables and properties that do not depend on the marginal distributions. *Psychometrika*, 73, 777-789.
- Weiss, A., Egan, V., & Figueredo, A. J. (2004). Sensational interests as a form of intrasexual competition. *Personality and Individual Differences*, 36, 563-573.
- White, B. W., & Saltz, E. (1957). Measurement of Reproducibility. *Psychological Bulletin*, 54, 81-99.
- Yokota, K., & Canter, D. (2004). Burglars' specialisation: Development of a thematic approach in investigative psychology. *Behaviormetrika*, 31, 153-167.
- Yule, G. U. (1912). On the methods of measuring the association between two attributes. *Journal of the Royal Statistical Society*, 75, 579-652.
- Zuckerman, M. (1996). Sensation seeking and the taste for vicarious horror. In J. B. Weaver, III, & R. Tamborini (Eds.), *Horror films: Current research on audience preferences and reactions* (pp. 147-160). Mahwah, NJ: Lawrence Erlbaum Associates, Inc. Publishers.

謝辞

本研究は、著者が総合研究大学院大学複合科学研究科統計科学専攻後期博士課程在学中に行った結果をまとめたものである。

総合研究大学院大学及び統計数理研究所准教授の土屋隆裕先生には、主任指導教官を快く引き受けていただき、本研究を続ける中で、統計科学に関する様々な示唆をご教示いただくとともに、在学の間、途切れることなく暖かいご指導をいただきました。また、指導教官を引き受けていただいた総合研究大学院大学及び統計数理研究所教授の吉野諒三先生には、調査科学や数量化理論についてご教示いただき、折に触れて研究のアイデアについて多くの有益なご助言をいただきました。心より感謝申し上げます。

総合研究大学院大学及び統計数理研究所教授の中村隆先生には、調査データ解析についてご指導いただくとともに、本論文について数多くの有益なご助言をいただきました。総合研究大学院大学及び統計数理研究所助教の朴堯星先生には、本論文の審査を通して重要なご指摘をいただきました。さらに統計数理研究所客員教授であり東洋英和女学院大学名誉教授の林文先生には、本論文の審査を通して示唆に富むご助言をいただきました。厚く御礼申し上げます。

さらに総合研究大学院大学及び統計数理研究所准教授の前田忠彦先生には、調査科学セミナーでの発表の機会をいただき、本研究について有用なアドバイスをいただきました。青山学院大学教授の真鍋一史先生には、Guttman の Facet 理論に関する長年のご研究から、本研究について有益なご示唆をいただきました。深く感謝いたします。

最後に、著者が博士課程在学中に統計数理研究所で過ごす中で、セミナーや講義等においてご指導いただいた先生方、様々な交流の機会をもつことのできた大学院の学生仲間の皆様、研究活動を支援いただいた統計数理研究所図書室や企画グループ研究支援担当の皆様をはじめ、関係の皆様にご心より感謝申し上げます。

付録 A 2 変数間の Homogeneity 係数の理論上の最小値

変数 i 及び変数 j のクロス表の度数を下表のように a, b, c, d とおく。いずれも 0 以上の整数であり、サンプルサイズは $n = a + b + c + d$ である。項目 i と項目 j の該当度数をそれぞれ $\alpha (\geq 0)$ と $\alpha - \beta (\geq 0)$ とする。該当割合はそれぞれ $p_i = \alpha/n$ と $p_j = (\alpha - \beta)/n$ であるから $p_i \geq p_j$ となり、 c が Guttman Error の度数（網掛け部分）となる。

		変数 j		
		該当	非該当	
変数 i	該当	a	b	α
	非該当	c	d	$n - \alpha$
		$\alpha - \beta$	$n - (\alpha - \beta)$	n

Homogeneity 係数 H_{ij} ([2.3.12]) は

$$[A.1] \quad H_{ij} = H_{ij}(\alpha, \beta, c) = 1 - \frac{E_{\text{obs}}}{E_{\text{exp}}} = 1 - \frac{c}{n \cdot \frac{n-\alpha}{n} \cdot \frac{\alpha-\beta}{n}} = 1 - \frac{nc}{(n-\alpha)(\alpha-\beta)}$$

で与えられるが、 $n - \alpha > 0$ 、 $\alpha - \beta > 0$ でないと定義できないので、

$$n - \alpha > 0 \Rightarrow \alpha \leq n - 1, \quad \alpha - \beta > 0 \Rightarrow \alpha \geq \beta + 1$$

でなければならない。これより、整数 α と β の範囲が

$$[A.2] \quad 1 \leq \alpha \leq n - 1, \quad 0 \leq \beta \leq \alpha - 1$$

のように得られる。

c は

$$\begin{aligned} a = (\alpha - \beta) - c \geq 0 &\Rightarrow c \leq \alpha - \beta \\ d = (n - \alpha) - c \geq 0 &\Rightarrow c \leq n - \alpha \end{aligned}$$

より

$$[A.3] \quad 0 \leq c \leq \min(\alpha - \beta, n - \alpha)$$

の範囲を動くこととなる。

2 項目間の Homogeneity 係数の理論的最小値

まず α 、 c を固定すると、 H_{ij} は β に対して単調減少であるから、 $\beta = \alpha - 1$ のとき最小値をとり、

$$[A.4] \quad H_{ij}(\alpha, \alpha - 1, c) = 1 - \frac{nc}{n - \alpha}$$

となる。これは、さらに c に対して単調減少であるが、 $c \leq \min(1, n - \alpha)$ より $c = 1$ のとき最小となり、

$$[A.5] \quad H_{ij}(\alpha, \alpha - 1, 1) = 1 - \frac{n}{n - \alpha}$$

となる。これは α に対して単調減少であるから、 $\alpha = n - 1$ のとき最小となり、理論上の最小値

$$[A.6] \quad H_{\text{theory}}^{\min} \equiv \min(H_{ij}) = 1 - n$$

を得る。このときクロス表は

		変数 j		
		該当	非該当	
変数 i	該当	0	$n - 1$	$n - 1$
	非該当	1	0	1
		1	$n - 1$	n

となっている。

周辺度数を固定したときの Homogeneity 係数 H_{ij} の理論的最小値

項目 i と項目 j の該当度数 α と $\alpha - \beta$ を固定したときの、 H_{ij} の最小値 $H_{ij|\alpha, \beta}^{\min}$ を考える。式[A.1] から H_{ij} は c に対して単調減少であることから、式[A.2] より次のような場合分けを行って最小値を求める。

$$\textcircled{1} \quad \alpha - \beta \leq n - \alpha \Leftrightarrow \alpha \leq \frac{1}{2}(n + \beta) \Leftrightarrow \frac{\alpha}{n} \leq 0.5 + \frac{\beta}{2n} \text{ のとき}$$

$c = \alpha - \beta$ で最小となることから、

$$[\text{A.7}] \quad H_{ij|\alpha,\beta}^{\min} = 1 - \frac{n}{n-\alpha}$$

となる。

$$\textcircled{2} \quad n - \alpha < \alpha - \beta \Leftrightarrow \alpha > \frac{1}{2}(n + \beta) \Leftrightarrow \frac{\alpha}{n} > 0.5 + \frac{\beta}{2n} \text{ のとき}$$

$c = n - \alpha$ で最小となることから、

$$[\text{A.8}] \quad H_{ij|\alpha,\beta}^{\min} = 1 - \frac{n}{\alpha - \beta}$$

となる。

付録 B よく見る映画・ドラマの特徴に関する質問文及び質問項目

あなたは、どのような映画やドラマ（アニメをふくむ）をよく見ますか。つぎのジャンルや特徴のうち、あてはまるものを全てえらんでください。（いくつでも）

※あまり見ない人は「映画やドラマ（アニメをふくむ）はあまり見ない」をえらんで、つぎに進んでください。

1. コメディーもの
2. ヒューマンドラマ系
3. 恋愛もの
4. アクションもの
5. 戦争や軍隊もの
6. 警察や刑事もの
7. スパイもの
8. ギャング、暴力団もの
9. 犯罪者が主人公の話
10. サスペンス、推理もの
11. ホラー（幽霊などオカルト系）
12. ホラー（残虐な殺害シーンのあるスプラッター系）
13. S F もの
14. ファンタジー（空想世界）もの
15. 暴力的な場面が出てくるもの
16. 武器（銃やナイフなど）を使った攻撃場面がでてくるもの
17. 血が流れるなどのリアルな場面がでてくるもの
18. その他の映画やドラマ（アニメをふくむ）
19. 映画やドラマ（アニメをふくむ）はあまり見ない

付録 C Facetted SSA の数理モデル

ここに M 個の項目からなる集合 A があり、その集合は k 個の互いに背反な下位集合に分類されるとし、それらを A_1, \dots, A_k とする。 M 個の項目は、これら k 個のいずれか 1 つの分類に属するものとする。この下位集合が、対象が属するカテゴリである。

(1) Axial モデル

この領域モデルは、 k 個の領域を平行する $k-1$ 個の直線によって区分する。 j ($j = 1, \dots, k-1$) 番目の直線はパラメータ θ を用いて、次式によって定義される。

$$[C.1] \quad x \sin \theta - y \cos \theta + c_j = 0.$$

この直線は、原点を中心に半径 c_j の円の接線であり、 θ を変えることによって、直線の傾きを自由に調整できる。また、各カテゴリに対応する帯状の領域の幅は、 c_j の値を変えることによって変更できる。点 (x_i, y_i) から j 番目の直線[C.1]までの距離を符号付き距離として

$$[C.2] \quad d_i^{\text{AX}}(j) = x_i \sin \theta - y_i \cos \theta + c_j$$

と表すとき、目的関数 $L_{\text{AXIAL}}(\theta, c_1, \dots, c_{k-1})$ は次式によって定義される。

$$[C.3] \quad L_{\text{AXIAL}}(\theta, c_1, \dots, c_{k-1}) = \sum_j \left[\sum_{\substack{(x_i, y_i) \in A_j \\ d_i^{\text{AX}}(j) > 0}} d_i^{\text{AX}}(j) - \sum_{\substack{(x_i, y_i) \in A_{j+1} \\ d_i^{\text{AX}}(j) < 0}} d_i^{\text{AX}}(j) \right].$$

これは、 j 番目の領域に布置されていないカテゴリ j の点から j 番目の領域までの距離の総和に当たる。これを最小化するようなパラメータ $\theta, c_1, \dots, c_{k-1}$ を求めることで、点の布置に対して最適な領域モデルが選択される。

(2) Radial モデル

この領域モデルは、 k 個の領域を $k-1$ 個の同心円によって区分する。 j ($j = 1, \dots, k-1$) 番目の円は、半径 r_j 、円の中心 (X, Y) として次式により定義される。

$$[C.4] \quad (x - X)^2 + (y - Y)^2 = r_j^2.$$

この円は、カテゴリ A_j に割り当てられた領域とカテゴリ A_{j+1} に割り当てられた領域とを

分割する線となる。点 (x_i, y_i) からこの j 番目の円[C.4]までの距離 $d_i^{\text{RD}}(j)$ を符号付き距離として

$$[\text{C.5}] \quad d_i^{\text{RD}}(j) = \sqrt{(x_i - X)^2 + (y_i - Y)^2} - r_j$$

と表すとき、目的関数 $L_{\text{RADIAL}}(X, Y, r_1, \dots, r_{k-1})$ は次式によって定義される。

$$[\text{C.6}] \quad L_{\text{RADIAL}}(X, Y, r_1, \dots, r_{k-1}) = \sum_j \left[\sum_{\substack{(x_i, y_i) \in A_j \\ d_i^{\text{RD}}(j) > 0}} d_i^{\text{RD}}(j) - \sum_{\substack{(x_i, y_i) \in A_{j+1} \\ d_i^{\text{RD}}(j) < 0}} d_i^{\text{RD}}(j) \right].$$

これは、 j 番目の領域に布置されていないカテゴリ j の点から、同心円に挟まれた j 番目の領域までの距離の総和に当たる。これを最小化するようなパラメータ $X, Y, r_1, \dots, r_{k-1}$ を求めることで、点の布置に対して最適な領域モデルが選択される。

(3) Angular モデル

この領域モデルは、 k 個の領域をある点から放射状に伸びる k 個の半直線によって区分する。 j ($j = 1, \dots, k$) 番目の半直線は端点 (X, Y) 及びパラメータ θ_j を用いて次式によって定義される。

$$[\text{C.7}] \quad \left\{ (x, y) \left| \begin{array}{l} x \sin \theta_j - y \cos \theta_j = X \sin \theta_j - Y \cos \theta_j \\ \text{and} \\ x \cos \theta_j - y \sin \theta_j > X \cos \theta_j - Y \sin \theta_j \end{array} \right. \right\}.$$

これは、端点 (X, Y) から放射上に伸びる傾き $\tan \theta_j$ の半直線を表す。これに対して、点 (x_i, y_i) を点 (X, Y) の周りの偏角を θ_i 、動径を r_i として

$$[\text{C.8}] \quad \begin{cases} x_i = X + r_i \cos \theta_i \\ y_i = Y + r_i \sin \theta_i \end{cases}$$

とすると、点 (x_i, y_i) から j 番目の半直線までの円弧による距離 $d_i^{\text{AN}}(j)$ を次式のように定義する。

$$[\text{C.9}] \quad d_i^{\text{AN}}(j) = \begin{cases} r_i |\theta_i - \theta_j| & \text{if } |\theta_i - \theta_j| < \frac{\pi}{2} \\ r_i |2\pi + \theta_i - \theta_j| & \text{if } |2\pi + \theta_i - \theta_j| < \frac{\pi}{2} \\ r_i |\theta_i - \theta_j - 2\pi| & \text{if } |\theta_i - \theta_j - 2\pi| < \frac{\pi}{2} \\ \frac{1}{2}\pi r_i & \text{otherwise.} \end{cases}$$

これは、 $0 \leq \theta_i \leq 2\pi$ 、 $0 \leq \theta_j \leq 2\pi$ としたとき、 $-2\pi \leq \theta_i - \theta_j \leq 2\pi$ に対して、 j 番目の半直線の周囲に偏角 θ_i が $\pm \frac{\pi}{2}$ の範囲内にあるときは点 (x_i, y_i) から半直線までの円弧による距離になり、偏角 θ_i が $\pm \frac{\pi}{2}$ の範囲を超えたときは $\frac{1}{2}\pi r_i$ とするものである。このとき、目的関数 $L_{\text{ANGULAR}}(X, Y, \theta_1, \dots, \theta_k)$ は次式によって定義される。

$$[\text{C.10}] \quad L_{\text{ANGULAR}}(X, Y, \theta_1, \dots, \theta_k) = \sum_j \sum_{\substack{(x_i, y_i) \in A_j \\ (x_i, y_i) \notin S_j}} \min[d_i^{\text{AN}}(j), d_i^{\text{AN}}(j+1)].$$

ただし、 S_j は A_j に割り当てられた、 j 番目と $j+1$ 番目の半直線に挟まれた楔形の領域を指す。つまり、この目的関数はカテゴリ j に属する点のうち、その属すべき領域 S_j に含まれなかった点の S_j までの円弧による距離の総和に当たる。これを最小化するようなパラメータ $X, Y, \theta_1, \dots, \theta_k$ を求めることで、点の布置に対して最適な領域モデルが選択される。

付録 D Homogeneity 係数を変換する関数 g について

Homogeneity 係数を変換する関数 g として、② $h_{ij}^{\text{lower}} = H_{ij|r_i, r_j}^{\min}$ を利用する場合、変数 i 、 j の正反応固定の下で理論上取り得る下限値を利用することから、 (i, j) の組み合わせごとに Homogeneity 係数の下限値 h_{ij}^{lower} が異なる。

人工データ 2 について、関数 g として

$$\textcircled{2} \ h_{ij}^{\text{lower}} = H_{ij|r_i, r_j}^{\min}$$

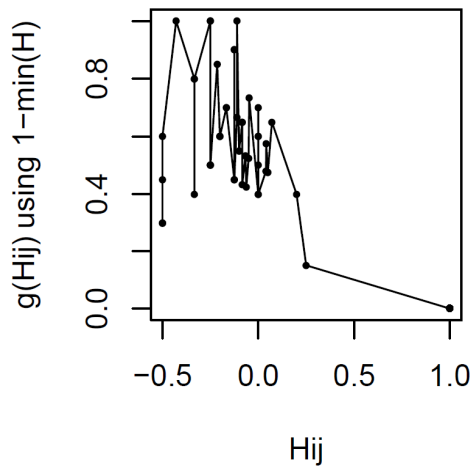
を用いた場合と

$$\textcircled{4} \ h_{ij}^{\text{lower}} = \min(0, H_{\text{data}}^{\min}) \quad (\text{人工データ 2 の場合、} \min(0, H_{\text{data}}^{\min}) = -0.5)$$

を用いた場合について、Homogeneity 係数の順序と関数 g による変換後の値の順序を比較した。Homogeneity 係数を横軸に、関数 g による変換後の値を縦軸としてプロットした図を下に示す。

図 (a) を見ると、Homogeneity 係数の順序関係が関数 g による変換後で保たれていないことがわかる。図の右側は関数 g の全てについて下限値を共通としているため、Homogeneity 係数の順序関係が変換後も保たれている。Homogeneity 係数が負の値をとる場合に、その値をどう評価するのかが関係してくる。Homogeneity 係数は正の方向のみに関心がある係数であるため、負の値には意味が無いとされる。負の値をとる Homogeneity 係数の順序関係について、どういった方法が望ましいかについては慎重な検討が必要であろう。

(a) ② $h_{ij}^{\text{lower}} = H_{ij|r_i, r_j}^{\min}$



(b) ④ $h_{ij}^{\text{lower}} = \min(0, H_{\text{data}}^{\min})$

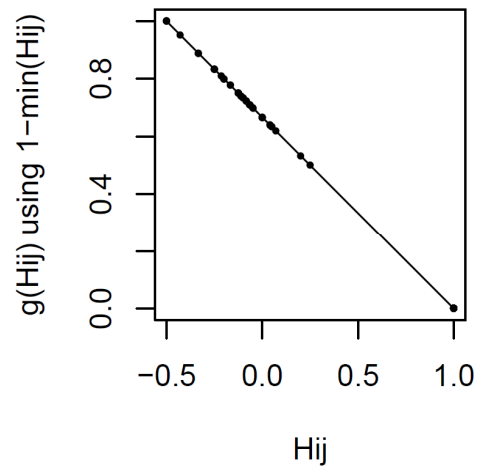


図 D-1 Homogeneity 係数と関数 g による変換後の値の順序関係

付録 E η^2 の θ_i による偏微分 (2 次元)

$$[E.1] \quad \eta^2 = \begin{cases} \left(\sum_i \sum_{i < j} \left((1-h)(r_i + r_j - |r_i - r_j|) + |r_i - r_j| \right) - d_{ij} \right)^2 & \text{if } H_{\text{data}}^{\min} \geq 0 \\ \left(\sum_i \sum_{i < j} \left((1-h) \frac{r_i + r_j - |r_i - r_j|}{1 - H_{\text{data}}^{\min}} + |r_i - r_j| \right) - d_{ij} \right)^2 & \text{if } H_{\text{data}}^{\min} < 0 \end{cases}$$

データにおける Homogeneity 係数 $\{H_{ij}\}$ の最小値を H_{data}^{\min} として、 $H_{\text{data}}^{\min} \geq 0$ のとき $H_{\text{data}}^{\min} = 0$ として、 η^2 の下段の式のみについて θ_i による偏微分を記載する。

$$[E.2] \quad d_{ij} = \sqrt{(r_i \cos \theta_i - r_j \cos \theta_j)^2 + (r_i \sin \theta_i - r_j \sin \theta_j)^2}$$

のもとで

$$[E.3] \quad \begin{aligned} \frac{\partial \eta^2}{\partial \theta_i} &= \sum_{j(\neq i)} \frac{-2}{d_{ij}} \left((1 - H_{ij}) \frac{r_i + r_j - |r_i - r_j|}{1 - H_{\text{data}}^{\min}} + |r_i - r_j| - d_{ij} \right) (-r_i \sin \theta_i (r_i \cos \theta_i - r_j \cos \theta_j) + \\ &\quad r_i \cos \theta_i (r_i \sin \theta_i - r_j \sin \theta_j)) \\ &= -2r_i \sum_{j(\neq i)} \frac{1}{d_{ij}} \left((1 - H_{ij}) \frac{r_i + r_j - |r_i - r_j|}{1 - H_{\text{data}}^{\min}} + |r_i - r_j| - d_{ij} \right) (-\sin \theta_i (r_i \cos \theta_i - r_j \cos \theta_j) \\ &\quad + \cos \theta_i (r_i \sin \theta_i - r_j \sin \theta_j)) \end{aligned}$$

付録 F η^2 の θ_i による偏微分 (3 次元)

$$[F.1] \quad \eta^2 = \begin{cases} \left(\sum_i \sum_{i < j} \left((1-h)(r_i + r_j - |r_i - r_j|) + |r_i - r_j| \right) - d_{ij} \right)^2 & \text{if } H_{\text{data}}^{\min} \geq 0 \\ \left(\sum_i \sum_{i < j} \left((1-h) \frac{r_i + r_j - |r_i - r_j|}{1-H_{\text{data}}^{\min}} + |r_i - r_j| \right) - d_{ij} \right)^2 & \text{if } H_{\text{data}}^{\min} < 0 \end{cases}$$

データにおける Homogeneity 係数 $\{H_{ij}\}$ の最小値を H_{data}^{\min} として、 $H_{\text{data}}^{\min} \geq 0$ のとき $H_{\text{data}}^{\min} = 0$ として、 η^2 の下段の式のみについて θ_i 及び φ_i による偏微分を記載する。

$$[F.2] \quad R_{ij} = (1 - H_{ij}) \frac{r_i + r_j - |r_i - r_j|}{1 - H_{\text{data}}^{\min}} + |r_i - r_j|$$

とすると、

$$[F.3] \quad d_{ij} = \sqrt{(r_i \sin \theta_i \cos \varphi_i - r_j \sin \theta_j \cos \varphi_j)^2 + (r_i \sin \theta_i \sin \varphi_i - r_j \sin \theta_j \sin \varphi_j)^2 + (r_i \cos \theta_i - r_j \cos \theta_j)^2}$$

のもとで、

$$[F.4] \quad \frac{\partial \eta^2}{\partial \theta_i} = - \sum_{j(\neq i)} \frac{R_{ij} - d_{ij}}{d_{ij}} \frac{\partial d_{ij}^2}{\partial \theta_i} \\ = -2 \sum_{j(\neq i)} \frac{r_i r_j (R_{ij} - d_{ij})}{d_{ij}} (-\cos \theta_i \cos \varphi_i \sin \theta_j \cos \varphi_j - \cos \theta_i \sin \varphi_i \sin \theta_j \sin \varphi_j + \sin \theta_i \cos \theta_j)$$

$$[F.5] \quad \frac{\partial \eta^2}{\partial \varphi_i} = - \sum_{j(\neq i)} \frac{R_{ij} - d_{ij}}{d_{ij}} \frac{\partial d_{ij}^2}{\partial \varphi_i} = -2 \sum_{j(\neq i)} \frac{r_i r_j (R_{ij} - d_{ij})}{d_{ij}} \sin \theta_i \sin \theta_j \sin(\varphi_i - \varphi_j)$$

付録G R プログラム

G.1 人工データ2の作成

```
z <- 5; loop <-1

while (loop>0) {
  set.seed(z)

  n <- 100
  p <- c(0.8,0.6,0.5,0.4,0.3,0.2,0.1)
  x <- matrix(9,n,length(p))

  for (i in 1:length(p)) {
    x[,i] <- c(rep(1,(n*p[i])),rep(0,(n-n*p[i])))
  }

  xx1 <- x; xx2 <- x
  k <- sample(1:n,size=n,replace=FALSE)
  xx2 <- xx2[k,]

  rs1 <- rowSums(xx1); rs2 <- rowSums(xx2)

  rtype <- rep(0,length=length(rs1))
  for(i in 1:length(rs1)) {
    if (rs1[i]>0 & rs2[i]>0) rtype[i] <-1
    if (rs1[i]>0 & rs2[i]==0) rtype[i] <-2
    if (rs1[i]==0 & rs2[i]>0) rtype[i] <-3
    if (rs1[i]==0 & rs2[i]==0) rtype[i] <-4
  }

  rtype4 <- which(rtype==4); rtype1 <- which(rtype==1)

  if (length(rtype4)>length(rtype1)) {
    z <- z + 1
  } else {
    rtype1 <-
    rtype1[sample(1:length(rtype1),size=length(rtype1),replace=FALSE)]
    rem1 <- length(rtype1)-length(rtype4)
    rtype1 <- rtype1[-c((length(rtype1)-rem1+1):length(rtype1))]
    nisi <- xx2[rtype4,]
    xx2[rtype4,] <- xx2[rtype1,]
    xx2[rtype1,] <- nisi

    loop <- 0
  }
}
```

```
data <- cbind(xx1,xx2)
```

G.2 等質性分析の実施

データ行列を `data` とする。

```
library(MASS)

res <- corresp(data,nf=2)
plot(res$cscore[,1],res$cscore[,2],type="n")
lab1 <- c("comedy(65)","human(24)","love(38)","action(31)","war(6)",
  "police(19)","spy(9)","gang(2)","crime(2)","suspense(24)",
  "occult(10)","splatter(3)","sf(28)","fantasy(37)","vio(3)",
  "weapon(5)","blood")
text(res$cscore[,1],res$cscore[,2],lab=lab1)
```

G.3 Kruskal の MDS の実施

データ行列を `data` とする。

```
library(MASS)
library(vegan)

f <- colMeans(data)
k3.dist <- vegdist(t(data),method="jaccard",binary=TRUE)
k3.mds <- isoMDS(k3.dist)
lab1 <- c("comedy(65)","human(24)","love(38)","action(31)","war(5)",
  "police(18)","spy(9)","gang(2)","crime(2)","suspense(24)",
  "occult(10)","splatter(3)","sf(28)","fantasy(37)","vio(3)",
  "weapon(5)","blood(3)")
plot(k3.mds$points, type="n", xlab="Coordinate 1",
  ylab="Coordinate 2",xlim=c(-0.6,0.7))
text(k3.mds$points, labels=lab1)
```

G.4 Radex モデル探索支援手法

重みづけ Homogeneity 係数の算出にロジスティック関数 (パラメータ $a = 10000$) を用いた場合

```
data <- read.table("film_a3.sco")
dataxy <- data[,2:3]
datax <- (dataxy[,1]-dataxy[1,1])*(-1)/100
datay <- (dataxy[,2]-dataxy[1,2])/100

nm <- read.table("film_a3.dat")

f <- colMeans(nm)*100
theta <- pi/12

L1 <- function(t) {
  L <- 0
  for (i in 1:ncol(nm)) {
    L <- L + (f[i] - t[3])
      / (((datax[i]-t[1])^2+(datay[i]-t[2])^2)^(t[4]/2)))^2
  }
  return(L)
}

L2 <- function(t) {
  LX <- 0
  LY <- 0
  LC <- 0
  LK <- 0
  for (i in 1:ncol(nm)) {
    LX <- LX + 2*(f[i]-t[3])
      / (((datax[i]-t[1])^2+(datay[i]-t[2])^2)^(t[4]/2)) * (-1)*t[3]*t[4]
      * (((datax[i]-t[1])^2+(datay[i]-t[2])^2)^(t[4]/2-1)) * (datax[i]-t[1])
      / (((datax[i]-t[1])^2+(datay[i]-t[2])^2)^t[4])
    LY <- LY + 2*(f[i]-t[3])
      / (((datax[i]-t[1])^2+(datay[i]-t[2])^2)^(t[4]/2)) * (-1)*t[3]*t[4]
      * (((datax[i]-t[1])^2+(datay[i]-t[2])^2)^(t[4]/2-1)) * (datay[i]-t[2])
      / (((datax[i]-t[1])^2+(datay[i]-t[2])^2)^t[4])
    LC <- LC + 2*(f[i]-t[3])
      / (((datax[i]-t[1])^2+(datay[i]-t[2])^2)^(t[4]/2)) *
      (-1/((datax[i]-t[1])^2+(datay[i]-t[2])^2)^(t[4]/2))
    LK <- LK + 2*(f[i]-t[3])
      / (((datax[i]-t[1])^2+(datay[i]-t[2])^2)^(t[4]/2)) * t[3]
      * (((datax[i]-t[1])^2+(datay[i]-t[2])^2)^(t[4]/2)
      * log(((datax[i]-t[1])^2+(datay[i]-t[2])^2)^(1/2))
      / (((datax[i]-t[1])^2+(datay[i]-t[2])^2)^t[4])
  }
  return(c(LX,LY,LC,LK))
}

init <- seq(0,0.01,length=4)
s <- optim(init,L1,gr=L2,method="BFGS")
(s)
```

```

org1 <- s$par
lab1 <- c("comedy","human","love","action","war","police","spy","gang","crime",
"suspense",
"occult","splatter","sf","fantasy","vio","weapon","blood")

datax <- datax-org1[1]
datay <- datay-org1[2]
dataxy <- data.frame(datax,datay)
datart <- data.frame(r=((dataxy[,1]^2+dataxy[,2]^2)^(1/2)),
t=atan2(dataxy[,2],dataxy[,1]))

varnum <- ncol(nm)
ho <- matrix(0, varnum, varnum)

for (i in 1:varnum) {
  for (j in 1:varnum) {
    z <- addmargins(table(factor(nm[,i],
levels=c(0,1)),factor(nm[,j], levels=c(0,1))),
margin=1:2)
    if (z[2,3]>z[3,2]) {
      ho[i,j] <- (z[2,2]/z[3,2]-z[2,3]/z[3,3])/(1-z[2,3]/z[3,3])
    }
    else {
      ho[i,j] <- (z[2,2]/z[2,3]-z[3,2]/z[3,3])/(1-z[3,2]/z[3,3])
    }
  }
}

wmat <- matrix(1, varnum, varnum)
aa <- 360
a <- seq(pi*(-1), pi,length=aa)

weight <- function(a, r, t) {
  c <- 10000
  if (abs(a-t)<=pi) {
    los <- r*abs(a-t)
  } else {
    los <- r*abs(a-(t-pi*2))
  }
  if (r/max(datart$r)<0.5) {
    inf <- r*theta
  } else {
    inf <- max(datart$r)/2*theta
  }
  d <- 1/(1+exp(c*(los-inf)))
  return(d)
}

hcal <- function(a) {
  w <- mapply(weight, a, datart$r, datart$t)
  wmat <- wmat*(w %o% w)
  bunshi <- matrix(0,varnum,varnum)
  bunbo <- matrix(0,varnum,varnum)
  for (i in 1:varnum) {

```

```

        for (j in 1:varnum) {
            bunshi[i,j] <- ho[i,j]*wmat[i,j]
            bunbo[i,j] <- wmat[i,j]
        }
    }
    bunshi[upper.tri(bunshi, diag=TRUE)] <- 0
    bunbo[upper.tri(bunbo, diag=TRUE)] <- 0
    hom <- sum(bunshi)/sum(bunbo)
    if (is.na(hom)==T) hom <- 0
    return(hom)
}

h <- rep(0.5, length=aa)

for (t in 1:aa) h[t] <- hcal(a[t])

h2 <- rep(0.5, length=(aa-1))
for (i in 1:(aa-1)) h2[i] <- h[i+1]-h[i]

##### 極大値探索
plus <- 0
minus <- 0
minuslst <- 0
minuslsthead <- c()
plustail <- 0
minushead <- 0
locmax <- c()
ips <- 0.001

for (i in 1:(aa-1)){
    if ((plus==0)&&(h2[i]>ips)) plus <- 1
    if (h2[i]>ips) plustail <- i
    if ((plus==0)&&(h2[i]< -ips)&&(minuslst==0)){
        minuslst <- 1
        minuslsthead <- i
    }
    if ((minus==0)&&(h2[i]< -ips)&&(plus==1)) minus <- 1
    if (h2[i]< -ips) minushead <- i
    if ((plus==1)&&(minus==1)){
        locmax <- c(locmax, (plustail+minushead)/2)
        plus <- 0
        minus <- 0
    }
    if ((i==(aa-1))&&(minuslst==1)&&(plus==1)&&(minus==0)){
        locmax <- c(locmax, ((plustail+minuslsthead+aa-1)/2)%(aa-1))
    }
}

par(mai=c(0.04,0.04,0.04,0.04))

plot(datax, datay, asp=1, type="n", xlab="", ylab="", xlim=c(-50,200), xax
t="n", yaxt="n")
text(datax, datay, labels=lab1, asp=1, cex=1.5)

```

```

lim <- par('usr')
par(new=T)
polygon((h+0.5)*80*cos(2*pi*c(1:aa)/360+pi),(h+0.5)*80*sin(2*pi*c(1:aa)/360+pi))
polygon((0.5)*80*cos(2*pi*c(1:aa)/360+pi),(0.5)*80*sin(2*pi*c(1:aa)/360+pi))
points(0,0,pch=4)
par(lty=3)

for (i in 1:length(locmax)) {
segments(0,0,max(datart[,1])*cos(locmax[i]/180*pi+pi),max(datart[,1])*sin(locmax[i]/180*pi+pi))
}

```


G.5 Radex モデル多重尺度構成法

```

windows(width = 6, height = 6)
par(mai=c(0.04,0.04,0.04,0.04))
par(bg="lightskyblue")

alpha <- 2
beta <- 1

data <- read.table("film_a3.dat")
nc <- ncol(data)
nr <- nrow(data)

r <- colMeans(data)
rmax <- max(r)

r <- (1-r)^alpha
r1 <- r

h0 <- matrix(0, nc, nc)

for (i in 1:nc) {
  for (j in 1:nc) {
    z <- addmargins(table(factor(data[,i],
levels=c(0,1)),factor(data[,j], levels=c(0,1))),margin=1:2)
    if (z[2,3]>z[3,2]) {
      h0[i,j] <- (z[2,2]/z[3,2]-z[2,3]/z[3,3])/(1-z[2,3]/z[3,3])
    }
    else {
      h0[i,j] <- (z[2,2]/z[2,3]-z[3,2]/z[3,3])/(1-z[3,2]/z[3,3])
    }
  }
}

ptn <- 0
if (min(h0)>0) ptn <- 1
if (min(h0)<=0) ptn <- 2

h1 <- 1-h0

etat <- 0

if (ptn==1) {
  for (i in 1:nc) {
    for (j in 1:nc) {
      if (i<j) etat <- etat +
(beta*(h1[i,j]*(r[i]+r[j]-abs(r[i]-r[j]))+abs(r[i]-r[j]))))^2
    }
  }
}

if (ptn==2) {
  for (i in 1:nc) {
    for (j in 1:nc) {

```

```

      if      (i<j)      etat      <-      etat      +
(beta*(h1[i,j]*(r[i]+r[j]-abs(r[i]-r[j]))/(1-min(h0))+abs(r[i]-r[j]
)))^2
    }
  }
}

set.seed(7100)

thetac <- pi/2
theta0 <- runif(nc,thetac,(thetac+pi*2))
theta0 <- theta0[-1]

d <- matrix(0,nc,nc)
rmat <- matrix(0,nc,nc)

eta2 <- function(theta) {

eta <- 0
theta <- c(thetac,theta)

dis <- function(p,q) {
  dist
  sqrt((r[p]*cos(theta[p])-r[q]*cos(theta[q]))^2+(r[p]*sin(theta[p])-
r[q]*sin(theta[q]))^2)
  return(dist)
}

if (ptn==1) {
  for (i in 1:nc) {
    for (j in 1:nc) {
      if      (i<j)      eta      <-      eta      +
(beta*(h1[i,j]*(r[i]+r[j]-abs(r[i]-r[j]))+abs(r[i]-r[j]))-dis(i,j))
^2
    }
  }
}

if (ptn==2) {
  for (i in 1:nc) {
    for (j in 1:nc) {
      if      (i<j)      eta      <-      eta      +
(beta*(h1[i,j]*(r[i]+r[j]-abs(r[i]-r[j]))/(1-min(h0))+abs(r[i]-r[j]
))-dis(i,j))^2
    }
  }
}

if (ptn==3) {
  for (i in 1:nc) {
    for (j in 1:nc) {
      if      (i<j)      eta      <-      eta      +
(beta*(h1[i,j]*(r[i]+r[j]-abs(r[i]-r[j]))/(1-minh(i,j))+abs(r[i]-r[

```

```

j]))-dis(i,j))^2
    }
  }
}

return(eta)
}

gr2 <- function(theta) {
  theta <- c(thetac,theta)

  dis <- function(p,q) {
    dist
    sqrt((r[p]*cos(theta[p]))-r[q]*cos(theta[q]))^2+(r[p]*sin(theta[p]))-
    r[q]*sin(theta[q]))^2)
    return(dist)
  }

  gr1 <- rep(0,length=nc)

  if (ptn==1) {
    for (k in 2:nc) {
      der <- 0
      for (j in 1:nc) {
        if (j!=k) der <- der - 2 *
(beta*(h1[j,k]*(r[j]+r[k]-abs(r[j]-r[k]))+abs(r[j]-r[k]))-dis(j,k))
*
      (-r[k] * sin(theta[k]) * (r[k] *
cos(theta[k]) - r[j] * cos(theta[j])) +
      r[k] * cos(theta[k]) * (r[k] * sin(theta[k])
- r[j] * sin(theta[j])) )
      )/dis(k,j)
      }
      gr1[k] <- der
    }
  }

  if (ptn==2) {
    for (k in 2:nc) {
      der <- 0
      for (j in 1:nc) {
        if (j!=k) der <- der - 2 *
(beta*(h1[j,k]*(r[j]+r[k]-abs(r[j]-r[k]))/(1-min(h0))+abs(r[j]-r[k]
))-dis(j,k)) *
      (-r[k] * sin(theta[k]) * (r[k] *
cos(theta[k]) - r[j] * cos(theta[j])) +
      r[k] * cos(theta[k]) * (r[k] * sin(theta[k])
- r[j] * sin(theta[j])) )
      )/dis(k,j)
      }
      gr1[k] <- der
    }
  }
}

```

```

}

if (ptn==3) {
  for (k in 2:nc) {
    der <- 0
    for (j in 1:nc) {
      if (j!=k) der <- der - 2 *
(beta*(h1[j,k]*(r[j]+r[k]-abs(r[j]-r[k]))/(1-minh(k,j))+abs(r[j]-r[
k]))-dis(j,k))*(-r[k] * sin(theta[k]) * (r[k] * cos(theta[k]) - r[j]
* cos(theta[j])) + r[k] * cos(theta[k]) * (r[k] * sin(theta[k]) - r[j]
* sin(theta[j])) ))/dis(k,j)
    }
    gr1[k] <- der
  }
}

return(gr1[-1])
}

s <- optim(theta0,eta2,gr=gr2,method="BFGS")
(s)

thetal <- c(thetac,s$par)
etas <- sqrt(s$value/etat)

rad <- r

lab
c("Comedy","Human","Love","Action","War","Police","Spy","Gang","Cri
me","Suspense","Occult","Splatter","Sci-Fi","Fantasy","Vio","Weapon
","Blood")
lab1 <- lab
thetal <- thetal%%(2*pi)

datart <- data.frame(r=rad,t=thetal)
datax <- rad*cos(thetal)
datay <- rad*sin(thetal)
dataxy <- data.frame(datax,datay)

##### パラメータ
pencil <- pi/12
rd <- 0.5
wmat <- matrix(1, nc, nc)
aa <- 360
a <- seq(pi*(-1), pi,length=aa)

weight <- function(a, radi, t) {
  c <- 10000
  if (abs(a-t)<=pi) {
    los <- radi*abs(a-t)
  } else {
    los <- radi*abs(a-(t-pi*2))
  }
  if (radi<rd) {

```

```

        inf <- radi*pencil
    } else {
        inf <- rd*pencil
    }
    d <- 1/(1+exp(c*(los-inf)))
    return(d)
}

hcal <- function(a) {
w <- mapply(weight, a, datart$r, datart$t)
wmat <- wmat*(w %o% w)
bunshi <- matrix(0,nc,nc)
bunbo <- matrix(0,nc,nc)
for (i in 1:nc) {
    for (j in 1:nc) {
        bunshi[i,j] <- h0[i,j]*wmat[i,j]
        bunbo[i,j] <- wmat[i,j]
    }
}
bunshi[upper.tri(bunshi, diag=TRUE)] <- 0
bunbo[upper.tri(bunbo, diag=TRUE)] <- 0
hom <- sum(bunshi)/sum(bunbo)
if (is.na(hom)==T) hom <- 0
return(hom)
}

h <- rep(0.5, length=aa)

for (t in 1:aa) h[t] <- hcal(a[t])
plot(h, type="l", ylim=c(0,max(h)*1.1))

h2 <- rep(0.5, length=(aa-1))
for (i in 1:(aa-1)) h2[i] <- h[i+1]-h[i]

##### 極大値探索
plus <- 0
minus <- 0
minus1st <- 0
minus1sthead <- c()
plustail <- 0
minushead <- 0
locmax <- c()

ips <- 0.001

for (i in 1:(aa-1)){
    if ((plus==0)&&(h2[i]>ips)) plus <- 1
    if (h2[i]>ips) plustail <- i
    if ((plus==0)&&(h2[i]< -ips)&&(minus1st==0)){
        minus1st <- 1
        minus1sthead <- i
    }
    if ((minus==0)&&(h2[i]< -ips)&&(plus==1)) minus <- 1
    if (h2[i]< -ips) minushead <- i
}

```

```

    if ((plus==1)&&(minus==1)){
      locmax <- c(locmax, (plustail+minushead)/2)
      plus <- 0
      minus <- 0
    }
    if ((i==(aa-1))&&(minus1st==1) && (plus==1) && (minus==0)){
      locmax <- c(locmax, ((plustail+minus1sthead+aa-1)/2)%%(aa-1))
    }
  }
}

plot(datax, datay, asp=1, pch=20, xlab="", ylab="", xlim=c(-1,1), ylim=c(-1,1))

lim <- par('usr')
par(new=T)

polygon((h*0.7+0.2)*cos(2*pi*c(1:aa)/360+pi), (h*0.7+0.2)*sin(2*pi*c(1:aa)/360+pi), asp=1, col="orangered", border=NA)
polygon((0.2)*cos(2*pi*c(1:aa)/360+pi), (0.2)*sin(2*pi*c(1:aa)/360+pi), asp=1, border=NA)

points(0,0,pch=3)

par(lty=3)

##### Bootstrap

dif <- function(tt1,tt2,rr1,rr2) {
  ntt <- length(tt1)
  phi0 <- 0.001
  L <- function(phi) {
    Lsq <- 0
    for (i in 1:ntt) {
      Lsq <- Lsq +
      (rr1[i]*cos(tt1[i])-rr2[i]*cos(tt2[i]+phi))^2+(rr1[i]*sin(tt1[i])-rr2[i]*sin(tt2[i]+phi))^2
    }
    return(Lsq)
  }

  dL <- function(phi) {
    dLsq <- 0
    for (i in 1:ntt) {
      dLsq <- dLsq +
      2*rr2[i]*sin(tt2[i]+phi)*(rr1[i]*cos(tt1[i])-rr2[i]*cos(tt2[i]+phi)) -
      2*rr2[i]*cos(tt2[i]+phi)*(rr1[i]*sin(tt1[i])-rr2[i]*sin(tt2[i]+phi))
    }
    return(dLsq)
  }

  op1 <- optim(phi0,L,gr=dL,method="BFGS")

```

```

    tt2 <- (pi-tt2)%%(2*pi)
    op2 <- optim(phi0,L,gr=dL,method="BFGS")
    if (op1$value<=op2$value) {
      return(op1$par)
    } else {
      return(op2$par+pi-tt2*2)
    }
  }

thetamat <- theta1
rr <- r

thetax0 <- r*cos(theta1)
thetay0 <- r*sin(theta1)

B <- 200
veck1 <- rep(1:nr,B)
veck2 <- sample(veck1,length(veck1),replace=FALSE)
veck <- matrix(veck2,B,byrow=T)

for (bt in 1:B) {
  i <- veck[bt,]
  data1 <- data[i,]

  r2 <- colMeans(data1)
  r2 <- (1-r2)^alpha

  h2 <- matrix(0, nc, nc)

  for (i in 1:nc) {
    for (j in 1:nc) {
      z <- addmargins(table(factor(data1[,i],
levels=c(0,1)),factor(data1[,j], levels=c(0,1))),margin=1:2)
      if (z[2,3]>z[3,2]) {
        h2[i,j] <- (z[2,2]/z[3,2]-z[2,3]/z[3,3])/(1-z[2,3]/z[3,3])
      }
      else {
        h2[i,j] <- (z[2,2]/z[2,3]-z[3,2]/z[3,3])/(1-z[3,2]/z[3,3])
      }
    }
  }
}

ptn <- 0
if (min(h2)>0) ptn <- 1
if (min(h2)<=0) ptn <- 2

h1 <- 1-h2
r <- r2
s1 <- optim(theta1[-1],eta2,gr=gr2,method="BFGS")
theta2 <- c(thetac,s1$par)%(2*pi)
phi1 <- dif(theta1,theta2,r1,r2)%(2*pi)
theta2 <- (theta2+phi1)%(2*pi)
thetamat <- rbind(thetamat,theta2)
rr <- rbind(rr,r2)

```

```

}

al <- 90
it <- 200

cir <- function(r) {
  nr <- length(r)
  sr <- sort(r)
  rdif <- numeric(nr)
  for (i in 1:(nr-1)) {
    rdif[i] <- sr[i+1]-sr[i]
  }

  alpha <- 1 - al/100
  kr <- round(nr*alpha)

  sr3 <- matrix(0,kr+1,nr-kr)
  for (i in 1:(kr+1)) {
    for (j in 1:(nr-kr)) sr3[i,j] <- sr[i+j-1]
  }

  sr3width <- numeric(kr+1)
  for (i in 1:(kr+1)) {
    sr3width[i] <- sr3[i,nr-kr]-sr3[i,1]
  }

  sr4 <- sr3[which.min(sr3width),]
  #sr4head <- sr4[1]
  #sr4tail <- sr4[nr-kr]
  sr4ht <- c(sr4[1],sr4[nr-kr])

  return(sr4ht)
}

cit <- function(theta) {
  nt <- length(theta)
  st <- sort(theta)
  alpha <- 1 - al/100
  kt <- round(nt*alpha)

  st3 <- numeric(nt)

  for (i in 1:nt) {
    st3[i] <- (st[(i+nt-kt-2)%%nt+1]+pi*2-st[i])%(pi*2)
  }
  st4ht
  c(st[which.min(st3)],st[(which.min(st3)+nt-kt-2)%%nt+1])
  return(st4ht)
}

for (m in 1:nc) {
  DX <- cbind(rr[-1,m],thetamat[-1,m]%(pi*2))
  rlow <- cir(DX[,1])[1]
  rhigh <- cir(DX[,1])[2]
}

```



```

thead <- cit(DX[,2])[1]
ttail <- cit(DX[,2])[2]
tdif <- (ttail-thead+pi*2)%(pi*2)

x1 <- rlow*cos(thead);      y1 <- rlow*sin(thead)
x2 <- rhigh*cos(thead);     y2 <- rhigh*sin(thead)
x3 <- rhigh*cos(ttail);     y3 <- rhigh*sin(ttail)
x4 <- rlow*cos(ttail);      y4 <- rlow*sin(ttail)

archx <- c(); archy <- c(); arclx <- c(); arcly <- c()
for (ii in 1:(it-1)) {
  archx <- c(archx,rhigh*cos(thead+tdif*ii/it))
  archy <- c(archy,rhigh*sin(thead+tdif*ii/it))
}
for (ii in 1:(it-1)) {
  arclx <- c(arclx,rlow*cos(ttail-tdif*ii/it))
  arcly <- c(arcly,rlow*sin(ttail-tdif*ii/it))
}
xcord      <-      c(x1,x2,archx,x3,x4,arclx);      ycord      <-
c(y1,y2,archy,y3,y4,arcly)
par(new=T)
polygon(xcord,ycord,asp=1,col=rgb(1,1,224/255,0.5),border=NA)
}

par(new=T)
plot(datax,datay,asp=1,pch=20,xlab="",ylab="",xlim=c(-1,1),ylim=c(-
1,1))
text(datax,datay,cex=1.3,asp=1,labels=lab,asp=1,pos=c(2,3,3,2,2,3,2
,2,2,2,2,2,1,2,2,3,1))

vall <- sprintf("%8.6f",etas)
text(0.5,0.9,cex=1.2,bquote(italic(S[Full] == .(vall))))

```