第Ⅰ部

対応分析法とは（概要）

大隅　昇

１．データをどう考えるか　－質的データと量的データ－

伝統的な統計学（数理統計学など）では，通常，取得した測定値（実測値，データ）を**連続的**（continuous）か**離散的**（discrete）かに分けて考える．これはその背景に統計的分布（確率分布）を連続型変数の確率分布と考えるか離散型変数の確率分布と見るかという考え方があることと無関係ではない．前者の例が正規分布や指数分布であり，後者の例として二項分布やポアソン分布などがある．

具体的な測定例でいうと，身長や体重を測定したデータは連続量とみなし，また電話の呼数や車の台数などは離散量と考える．また，品質管理などの分野では，これを計量的，計数的と対応させて考えている．

どちらかというと，こうした数学的な区分は，数理的展開を行ううえでは便利であるが，現実の測定・調査や分析場面においてはこれだけでは十分ではない．そこで多くの場合，“**尺度**”（scale）によるデータの分類区分を併せて用いる．これはまず，“**質的データ**”（qualitative data）と“**量的データ**”（quantitative data）とに分けて考え，質的データはさらに**名義尺度**（nominal scale）と**順序尺度**（ordinal scale）に，また量的データは**区間尺度**（interval scale）と**比例尺度**（ratio scale）に分けて考える[1]．この考え方に従うと，定性情報，定量情報に関わりなく，多くの調査・測定データの解釈の自由度が拡がる．注意すべきことは，大まかな言い方ではあるが，いわゆる質的データは原則として四則演算（加減乗除）の適用が難しいということである（よって後述のように「数量化」や「尺度化」の発想が生まれる）．

なお最近は，データの様相が多様化し，こうした枠組みだけでは，必ずしも十分な説明ができなくなっている．つまり別の視点からのデータの分類区分も必要となってきた．たとえば，画像（イメージ：静止画，動画），音声など，そのままでは非数値的であって何らかの加工変換を必要とする場合もある．また，**テキスト型データ**（textual data）を含む文字情報も上のような枠組みでは必ずしもうまくは説明できないこともある．このようなことで，以下のような区分を考えておくことも時には必要である．

○数値的か非数値的か

数値的データ（numerical data）
　数量，数値，計数として表記されるもの
非数値的データ（non-numerical data）
　文字，記号，イメージ（静止画，動画），音声など

○構造的か非構造的か

構造化データ（structured data）
　カテゴリー化，タグ化，コード化などを行い，正規化されたデータ
　リレーショナル・データベースなどとして整備されたデータなど
非構造化データ（unstructured data）
　とくに何も手当や加工がなされない生（なま）データ（原データ）
　自由回答・自由記述のデータ
　日記形式，聴き取りやインタビューで収集のデータ
　ソーシャル・メディア関連（ツイッター，ブログなどに表れる非定型データ）
　場合によっては，画像（動画，静止画），音声などのデータ

どのような分類を行っても，それで一意的に区分して考えるのではなく（いつもそうできるとは限らず），状況に応じて解釈や分析上の操作に都合のよい形で用いることを考えることも肝要である（分析者が必要とする情報にとって，扱い易く意味ある形で用いるということ）．

たとえば，上の区分では自由回答質問で得たデータやツイッターやブログのデータを“非構造化”データとした．しかし，これもデータを集める操作・手順の観点から考えると，異

---

[1] 名義尺度は名目尺度，分類尺度ともいう．また順序尺度は順位尺度ともいう．これは名義尺度であって並びの順が意味を持つような場合をいう．また，区間尺度は間隔尺度と，比例尺度を比率尺度ともいう．比例尺度は，区間尺度であって比が意味を持つような場合をいう．

なる見方もあるだろう．調査で用いる自由回答質問は，調査設計者がある意図をもって（つまり，これこれのことを聞き出したい，として）質問文を作成し問いかける，いわば質問を一種のセンサーとして人の意見・態度を聞き出すという**“双方向的”**な（インタラクティヴな）データ収集行為といえる．自由回答質問文の作成者の意図が反映された自由回答が得られるという点で，完全な非構造的というよりも**“半構造的”**（semi-structured）というべきかもしれない．一方，ソーシャル・メディア上に流れる発言・発話などの情報は，発言者あるいは発言者同士の間で交わされる情報であり，かならずしもそれを用いた分析を考えるデータ収集者の意向・意図を反映したものではない．つまりこの種のデータを分析者・利用者からみると，発信者からの**“１方向的”**なデータであり典型的な“非構造的”なデータと考えられる．いわゆるビッグデータ・アナリティックスなどで，この種のデータを扱う場合には，データ特性の構造化の程度がどのようなものか，十分な注意が必要である．

２．調査におけるデータ収集環境の変化

調査環境の急速な変化，とくに調査環境の悪化が指摘され，調査の質の低下が深刻な問題とされるように，様々の原因で満足できる内容の調査がきわめて困難になってきた．とくに従来からの定量的調査の実施困難性やさまざまな問題，たとえば，回収率・回答率の低下，非標本誤差や調査不能・無回答の増大（さまざまな**“調査誤差”**（survey errors）の介入，大隅他（2011)），そして貴重な標本抽出枠（サンプリング・フレーム）であった住民基本台帳や選挙人名簿等の閲覧制限，個人情報保護法の実施等に関連した調査情報取得環境の変容がある．

一方，ウェブ調査（インターネット調査，オンライン調査）などの新たな**調査方式**（調査モード；survey mode）が登場し，従来の方法論の見直し，たとえばクオータ・サンプリング，エリア・サンプリング，郵送調査，電話調査，面接調査等のあり方が改めて問われている．とくに調査費用面の負荷が増大し，適切な調査を実施することが困難となり，いきおいウェブ調査やモバイル調査などの安易な方向に向かう傾向にある．しかも，ウェブ調査のような新しい調査方式の国内研究は欧米諸国に比べてかなり遅れている（Bethlehem and Biffignandi（2012），Couper（2008），大隅（2002）他）．

また，多くの研究分野では，定性型の**データ収集方式**（data collection mode）に移行する傾向にある（たとえば，Flick（2002））．理由はいろいろあるだろうが，最大の理由は，ここでも定量型の選択肢型設問形式やこれに類した形式によるデータ取得だけでは，調査対象（回答者など）から本当に知りたいこと（本音，実状・実態）を把握できないのではないかという懸念，それと調査環境の悪化から，いわゆる標本調査的な，とくに確率標本を基本とする確率的アプローチによる調査の実施が困難となってきたこと，そもそも標本の大きさが十分な定量的なデータの収集が困難な調査対象が多くなってきていること（たとえば，介護・福祉の問題，環境評価や食品衛生における測定環境）などが考えられる．

とくに社会調査においては，従来とは異なる意味で，あるいは従来にもまして質的・定性的調査への関心が高まっている．標本の大きさがそれなりに大きく，また伝統的な標本調査法に従ったサンプリング操作を経て得られる**確率標本**（probability sample）を用いた量的な調査（たとえば従来型調査の中心であった面接調査，留置（とめおき）調査[3]，郵送調査等）が，経済的にも労力の面からも負担が大きく，一方それに見合った成果が次第に期待できない状況にあることから（たとえば回収率・参加率・協力率の低下，無回答・回答拒否・調査不能の増加），質的調査や定性調査に関心が移行する傾向も見られる．いわゆる**非確率標本**（non-probability sample）や**便宜的標本**（convenence sample：コンビニエンス・サンプルやボランティア・ウェ

[3] 留置調査とは，調査員が調査対象者を訪問し，調査票を手渡して調査回答を依頼する調査方式．これを「留置」（drop-off）という．このあとの調査票の回収方法で，いくつかの方式が考えられる．①一定期間の時間経過のあと（例：1 週間後），調査員が回答済みの調査票の回収に出向く（drop-off and pick-up）．②回収を別の調査方式で行う．たとえば，郵送（drop-off and mail-back），インターネット（オンライン）（drop-off and invitation to a Web survey）など．③その回答方式の種類の選択を回答者に委ねる場合．そうではなく，あらかじめ回答方式を固定した場合などがある．後者は一種の「混合方式」（mixed-mode）となる．

ブ・パネル）を対象とする調査が増えている．

もっとも，市場調査分野等では早くから，グループ・インタビュー（GI）やフォーカス・グループ（FG），モチベーション・リサーチなどが利用されてきた．さらに，テキスト･マイニング手法の登場で，これらの方法による取得データの解析法も改めて注目されている．また，きわめて少数のサンプル（レア・サンプル）や，条件を限定した回答者を相手としたモニター調査や，ウェブ調査などの調査方式では，自由回答や自由記述の質問を多用し，ここで取得したデータの質的解析を試みることが多くなってきた．また，顧客の個人登録情報と購買行動を電子的に追跡し記録し，集めた大量のデータセットを利用することも可能になっている．これらは電子的なデータ取得がきわめて容易になったということ，一方では情報が多様化していることを意味している．

こうしたウェブ（WWW）やインターネットの技術進歩と普及により，**コンピュータ支援の調査情報取得**（CASIC：Computer Assisted Survey Information Collection）や**コンピュータ支援によるデータ収集**（CADAC：Computer Assisted Data Collection）の研究や実用化が進み，とくに自由回答に代表される**テキスト型データ**（textual data）の取得が，内容の質の適否に関わりなく，容易に，しかも大量取得が可能となった．このようなことで，いきおい自由回答質問を多用する調査（とくに消費者動向調査，顧客満足度調査，インターネット・マーケティング）も多くなった．

さらに，企業業務レベルでは CRM（Customer Relationship Management），顧客動向や顧客満足度（CS：customer satisfaction）の把握などとの関連で，企業のコール・センター，コンタクト･センターや顧客相談窓口における取得データの定性情報解析など多種多様な試みがあり，また具体的方法論や解析システムの開発への期待も高い．最近は，いわゆるソーシャル・メディアとして流通する情報，たとえば，ツイッター，ブログ，フェイスブックなどで拡がる個々人の発信する大量の情報を分析し有用な知見を得るための**ビッグデータ・アナリティックス**（big data analytics）などが，従来のデータ・マイニングやテキスト・マイニングと結びついて議論されるようになってきた．このように，今後は，調査環境の多様化，**情報通信技術**（ICT）の応用可能性の拡大に伴い，文章型・文字型のいわゆるテキスト型データ（textual data）の取得や解析の機会の増大が考えられる．ここでは，用いるデータセットの代表性（誰を調べているのか）や，データ量が多いことが､本当に質（データの質，調査の質）の保証につながるのか，といった，かなり深刻な課題も抱えている．

3. 対応分析とは

以上を前置きとして，多次元データ解析手法の一つである**対応分析法**（CA：correspondence analysis）について，ここで簡単に紹介する．また，できるだけ統計ソフトウェア（WordMiner，JMP，JMP スクリプト[4]）が出力する数値例を用いて説明する．若干の数式を用いるが．数理の詳細を知りたい場合は，本資料に続く資料として「第Ⅱ部 対応分析法の基本的な考え方（数理の要点）」を用意したので，それを参考にし，また参考文献に挙げた資料をみていただきたい．

3．1 対応分析法と数量化法 III 類

“対応分析法”あるいは対応分析（AFC：Analyse Factorielle des Correspondances，Analyse des Correspondances）はフランスの研究者，ベンゼクリ（J.-P. Benzécri）により，1960 年代初期（1962 年頃）に提唱された方法である．ベンゼクリは，いわゆるフランスならびに欧州圏におけるデータ解析（analyse des données）の提唱者，指導者として中心的な役割を果たしてきた．対応分析法はこうした研究活動の中で登場した手法の一つで，Correspondence Analysis（CA：コレスポンデンス分析）の名称で仏語圏から欧米圏（とくに英語圏）の研究者間に次第に知られるようになり，また多くの統計ソフトウェアに搭載されたことで急速に普及した[5]．

---

[4] JMP スクリプトとは，JMP 固有のプログラミング言語を用いて作られたプログラムのこと．ここでは，対応分析を行うスクリプト・モジュールを用いる．

[5] たとえば，Hill, M.O. (1974): Correspondence Analysis: A Neglected Multivariate Method, *J. Roy. Stat. Soc. Ser. C*, **23**,

一方，日本国内では，ベンゼクリよりはるかに早く（1952～1954 年頃），（故）林知己夫が数量化法・数量化理論として，一連のさまざまな手法を提案してきた（例：数量化法 I 類，II類，III類，IV類など）．その一つに広く利用されてきた“**林の数量化法III類**”（quantification method - type III，パターン分類法）」がある[6]．これも多くの統計ソフトウェアに質的データの分析手法として実装され広く利用されてきた．

じつは，数理的には対応分析法は数量化法 III 類と同等である．しかし林はいわゆる数量化理論全体の枠組みの中で総合的かつ体系的に“**質的データの数量化**”という視点から考察し，その一つの手法として数量化法III類を考えた（とくにスケーリング・尺度化と関連させた独自のアイディアを展開した）．一方，ベンゼクリは，クロス表（2 元クロス表）の独立性の検定に用いる**ピアソンのカイ二乗統計量**に注目し，2 元データ表に対する**“多次元の質的データの合成変数”**を作る（低次元の空間内に布置する）方法として，その 2 元データ表の項目間の関連性（対応）を測る方法として考えた（詳しくは後述）．

つまり，林・ベンゼクリ両氏の思想的な背景，数量化法III類・対応分析それぞれの発展の経緯，応用分野や理念，定式化にはかなり異なるものがある．しかも，彼らの執筆論文や著書では，両名の個性的かつ独特の論理展開が行われてきた．たとえば，ベンゼクリ（1976 ,1992）は、質量(mass)，プロファイル（profile），重心（center of gravity），慣性（intertia），座標（coordinates），ホイヘンスの定理（Huyghens’ formula，慣性モーメントの定理）といった，主に物理学分野における用語が次々と登場する（これはベンゼクリの出自に由来する）．こうした用語を使ってはいるが，統計的な用語では，重心が平均，慣性が分散，質量が周辺度数に対応している（後述）．こうした記述や符丁に不慣れな人々にとっては混乱をまねく．また，ベンゼクリの用いる数式記法も独特である．さらには，多くの関連原書がフランス語で書かれていることがある．こうしたことから，両手法はあたかも別の方法のように思われてきた時期もあったが，じつは数理的には同じ方法である（いまだに誤解している向きもある）．

さらにその後，対応分析法や数量化法 III 類に類似の手法が，さまざまな研究分野で登場したことで，それらの手法相互の関係も詳しく調べられるようになった．たとえば，同等あるいは類似の手法として，以下がある．

- 双対尺度法（dual scaling；西里静彦（1980））．
- 逆反復平均法・集群分析法[7]（reciprocal averaging method ; M. O. Hill（1973，1974）他）
- 等質性分析（homogeneity analysis；Gifi（1990），J. Meulman（1984）他）

また欧米，国内の研究にも，多くの関連手法が登場した．とくに，フランスを中心とする欧州圏では，さまざまなデータ表形式に対応する対応分析の変形手法がいろいろと考案されてきた．たとえば，以下がある．

- 多重対応分析法（多重クロス表の対応分析；MCA：Multiple Correspondence Analysis）
- 対数線形モデルとの関連研究，N.Lauro 他（1982），Hudon（1990），Choulakian（1988）
- 非対称対応分析法（NSCA：non symmetrical CA，N.C. Lauro（1994））
- 正準対応分析法（Canonical CA）
- 連関分析法（Association Analysis；L. A. Goodman（1986）ほか）
- その他：Subset CA，Joint CA など（Greenacre（1984，2007）ほか）

3. 2 対応分析法の要約 ―仕組み―

3. 2. 1 数量化の本質

対応分析法，数量化法 III 類とも，登場してからすでに数十年を経た方法論である．しかし，その本質的な意味や正しい理解が行き届いているとは言い難い面もある．テキスト型データ

340-354.

[6]数量化 I 類，数量化II類，数量化 III 類のように，ローマ数字を付けた呼称は，飽戸弘氏による命名．

[7] reciprocal averaging は，two-way averaging, cross-calibration, two-way successive calibration などの別称もある．

のマイニングのような定性型データに対して，なぜ利用可能なのか，またその適用可能性はいかほどか，といったことを含めて，とくに対応分析法について簡単に要約しよう．より詳しい説明は別の資料として用意したので，ここでは“対応分析法とはこんなことを行う方法”という入り口を，なるべく数値例，図表，グラフィカル表現を用いて説明する．

まず「**“数量化”**とは何か」を考える（このあとに例も挙げた）．林知己夫の考え方は，質的データに対しては，数量は与えられまた計量されるものとして，しかも数理的な（制約）条件のもとに作られた手法をデータに当てはめることが，そもそも無理があるのではないかとの主張である．つまり，「本来，数（数量，数値）はあらかじめそのものに内在するのではなく，目的を達成するために科学的に与えるものであり，そのための道具と“目的に応じてふさわしく与えるもの”である」という立場をとる．そもそも生(なま)の質的な測定データ（数値とは限らない定性・質的情報）の示す意味表現と，分析に用いるために必要とする数値とは峻別して考えるべきとの見方でもある．林（1993，2001）．

さらに数量化で重要なことは，諸事象はあらかじめ「線形的である」あるいは線形として説明できるものではなく，「線形にする，あるいはいかにして線形にできるか（できそうか）」，そのような数量の与え方があり得るのか，またそうあるような**データ収集方式**（data collection mode）はいかに工夫すべきか，調査であれば適切な**調査方式**（調査モード；survey mode）の設定と**調査票・質問文の作り方**はどうあるべきか，といったことにあるという観点からのアプローチである．この考え方は，いわゆる伝統的な多変量解析的な発想とはやや異なる方向である．元来は非線形の事象が多いのであるから，それをなるべく扱いやすい線形に近い形にすること（できるか，を考える），併せて実験の計画を工夫し，その現象解明に適したデータの取得法と解析法を通じて問題を解明する筋道を明らかにするという立場である．これが発展的に**“データの科学”**（data science）の概念につながる[8]．（林（2001））．

形式論的にいえば，たとえば数量化法Ⅰ類は線形重回帰モデルである．数量化法Ⅱ類も判別分析の変形として位置づけられる．さらに広く利用されている数量化法Ⅲ類は質的データの尺度化という言い方も可能であろう．一方，ベンゼクリが提案したように，対応分析法は 2 元データ表の多次元情報を，なるべく低次元のユークリッド空間内にうまく近似射影するという着想から得た定式化である．これは林による数量化・尺度化の発想とはだいぶ異なる．

そもそも“数量化”とは何か．あるいは，なぜ質的データに対して対応分析を用いるのか．たとえば，ある調査質問の選択肢（カテゴリー）として「非常に満足」に 4 を与え，以下同じように「満足」＝3，「あまり満足でない」＝2，「まったく満足でない」＝1 と付与したとしよう[9]．こうしたアプリオリに与えた，大きさに意味のない形式的な数値（ここでは順序尺度）を使った四則演算，たとえば平均値を出すとか，あるいはこの数値化の意味をよく考えずに多変量解析手法（たとえば主成分分析）を適用するなどの操作は正しいのだろうか．数量化法はこれに疑問があると考え，数値はアプリオリに与えるべきではない，ましてや線形性（線形モデル）をこうした名目的な数値にはいきなり想定はできない，むしろ「数量」（あるいは尺度）は現象を代表する（説明するであろう）データにもとづいて作られるもの，つまり数理的には新たな座標空間を作り出すことにあると考える．対応分析法と，この点で通底するものがある．

数量化法におけるもう一つの特徴は，「外的基準のある場合」と「外的基準のない場合」を分けて考えることにある．これは多次元データ解析を考えるうえで重要な要素である．この視点から数量化の各手法を整理することで，いわゆる数量化法が体系化される．たとえば，「外的基準のある場合」として数量化法Ⅰ類，Ⅱ類が，「外的基準のない場合」としてⅢ類，Ⅳ類，Ⅴ類，そしてⅥ類が位置づけられる．数量化法 III 類は外的基準のない場合の典型的な

---

8 最近，“データサイエンス”といった語句が登場しているが，林知己夫とそのグループは，すでに 1995 年あたりから，“データの科学”（data science）として独自の主張を展開してきた（林知己夫ほか）．

9 このような得点化の方式を**ライカート方式**という（提案者の R. Likert から，国内ではリッカートの呼称が普及しているがここはライカートとした）．典型的な態度尺度化の方法で，広く利用されてきたがこうした方式への一つの反論が数量化法である．林（1993），西里（2007）などをみるとよい．

手法であるが，ここらの詳しいことは文献を参照されたい（たとえば林（1993），岩坪（1987））．ここでもっとも重要なことは，数量化の核心は「数のないもの（その典型が質的データ）を測定で探査し，これに数量を与えてデータ解析（分析）し，その現象についての特有の知見を得て，かつ洞察（インサイト）する」ということある．そして，この点では，上述のように対応分析法の思想にも通じるものがある．

一方，ベンゼクリは，対応分析法を始めて適用した分野が，言語解析・語彙用語論の分野のデータ解析であった[10]．つまり典型的な非定型のあるいは定性的なデータを想定したアプローチであった．そもそものきっかけが 2 元データ表形式の処理，しかも従来の定量的アプローチとは異なる視点からデータ構造を観察することにあった．基本は（かなり条件を緩めた）**“2 元データ表”**（two-way table）の行と列との対応関係（構造）を測ることにある．ベンゼクリによると，“非負の要素からなり，行あるいは列のプロファイルが意味を持つようなデータ表”の構造探査ということになる（Benzécri（1982））．

以上の考え方は，定性情報の典型例である比較的規模の大きな，しかも疎な 2 元データ表の形をとることが多い**テキスト型データの解析**（textual data analysis）に対してそのまま応用できる．また，ソーシャル・メディアなどでもこうした 2 元の行列形式の大規模なデータ表を扱うことが多いであろう．

3. 2. 2.　簡単な例による数量化の確認

上で，言葉として，対応分析法あるいは数量化法 III 類が行うこと（考え方の基本）を述べたが，これだけではなかなか分かりにくい．そこで簡単な例で，“質的データを数量として評価する”とはどういうことかを調べよう．そもそも**“対応分析法とは何を行うのか”**，また**“数量化”**とはいったい何を行っているのか，これをある調査データからえた情報を用いて説明する．なお，ここで用いる記号や記法は，対応分析法の数理の要点を説明した別の資料「第Ⅱ部対応分析法の基本的な考え方」で用いた表記法に合わせてあるので，これを併読していただきたい．

［事例とする調査の概要］

筆者らが行った環境意識調査で用いた調査質問文と，得られた回答データを用いて説明しよう．まず，調査の概要を以下に記す．かなり昔に実施された調査であるが，環境意識調査としては標本の大きさも多く，調査内容も体系的にまとまっており，しっかりした“確率標本”を用いて行われた稀な調査である．

調査内容：「都市環境のすみやすさに関する調査」
調査設計：統計数理研究所（林知己夫，水野欽司，大隅昇他）／実査：輿論科学協会
目的：都市住民の環境への意識，住みやすさ感，満足感などの総合的調査
調査実施年：1983, 1984, 1985 年
標本抽出枠：選挙人名簿から単純無作為・系統抽出（層化なし），よって確率標本
調査方式：調査員による訪問留置，自記式調査法

この調査の全体を通じて用いた「共通の質問」のなかから，以下の 2 問を用いることにしよう．また，例示とする調査データは．上の表の 1983 年次の調査の結果を用いる．よって，分析結果の解釈には，この頃の（かなり昔の）状況であることを念頭に考える必要がある．

この調査の全体を通じて用いた「共通の質問」のなかから，以下の 2 問を用いることにする．また，例示とする調査データは．表の 1983 年次の調査の結果を用いる．よって，分析結果の解釈には，この頃の（かなり昔の）状況であることを念頭に考える必要がある．

回答総数＝1,973（名）のうち，「無回答（non-response）・その他」を除いて集計した 1,946 名のクロス表を示した．通常は回答拒否や DK（Don’t Know）などもありうるが（その理由付けが調査の質の評価の観点からは重要），ここでは説明を簡単にするために除外してある．

---

10 ベンゼクリがレンヌ（Rennes）で行った講義録（6 課程，1963）：“Statistique et structure des langues naturelles: Essai de synthese mathematique”．（「自然言語の統計学と構造：数学的総合化に関する小論」）．

表 1「都市環境のすみやすさに関する調査」の概要

調査年次	調査対象地域	計画標本の大きさ（人）	回収標本の大きさ（人）	回収率（%）
1983 年 （S58 年）	千里ﾆｭｰﾀｳﾝ	1,800	1,205	67.0
	千葉市市街部	1,440	768	53.3
	小計	**3,240**	**1,973**	**60.9**
1984 年 （S59 年）	箕面市の一部	870	635	67.6
	千葉市市街隣接部	1,440	1,074	74.6
	小計	**2,310**	**1,709**	**71.9**
1985 年 （S60 年）	東京都江東区	1,170	755	61.4
	三鷹市	900	593	64.5
	小計	**2,070**	**1,348**	**67.1**
総計		**7,620**	**5,030**	**66.0**

この調査の全体を通じて用いた「共通の質問」のなかから，以下の 2 問を用いることにする．また，例示とする調査データは．上の表の 1983 年次の調査の結果を用いる．よって，分析結果の解釈には，この頃の（かなり昔の）状況であることを念頭に考える必要がある．

回答総数＝1,973（名）のうち，「無回答（non-response）・その他」を除いて集計した 1,946 名のクロス表を示した．なお，通常は回答拒否や DK（Don't Know）などもありうるが（またその理由付けが調査の質の評価の観点からは重要であるが）ここでは説明を簡単にするために除外してある．

質問 *I*：あなたは，いま住んでいるまちが気に入っていますか．（一つ選ぶ）

1. たいへん気に入っている
2. まあ気に入っている
3. あまり気に入っていない
4. 気に入っていない

質問 *J*：あなたが住んでいる地区は，都市としては，“緑（みどり）が多い”と感じますか．それとも少ないと感じますか．（一つ選ぶ）

1. かなり多い
2. 多いほうである
3. ふつう
4. 少ない
5. 少ないほうである

また，元のデータ表（つまり回収した調査データ）の一部を表 2 に示した．これはいわゆる**「（サンプル・個体）×（変量・項目）」の多変量構造データ**である．実際のデータ表の寸法は（1,973 行×122 項目）である．この多変量（多数項目）の中から 2 つの質問 *I*，*J* を選んで集計した結果が表 3 のクロス表となる．この加工過程を知っておくことが対応分析法を理解するうえで重要である．じつは対応分析法はどちらのデータ表（表 2 の 2 項目を指定してえられるアイテム･カテゴリー型データ表）からも分析可能で，しかも結果は同等である（後述）．

この 2 つの質問文の選択肢はいずれも“名義尺度”である．とくに質問 *I* は典型的な“順序尺度”である．つまり，回答選択肢には便宜的に数字を付けてあるが，これがそのまま数値コードとして統計量の算出などに利用できるとはかぎらない（この数値コードを用いて平均値や分散，標準偏差などを求める意味があるかは慎重に考える必要がある）．つまりこの測定値，回答データは，あらかじめ用意された選択肢を選んだにすぎない．この質問 *I* の 4 つの選択肢，質問 *J* の 5 つの選択肢に対して，なにか数量的に扱えるような数値を与えることができるのか，というのが対応分析法あるいは数量化法 III 類の課題である．これを具体的に行ってみよう．実際にここで用いるデータセットは，表 2 のような文字データ（つまりテキスト型データ）で与えられている（図中でアミカケとした箇所）．つまり質的データであり数値ではないことに注意しよう．

表 2　データセットの例（JMP データテーブルとして表示）

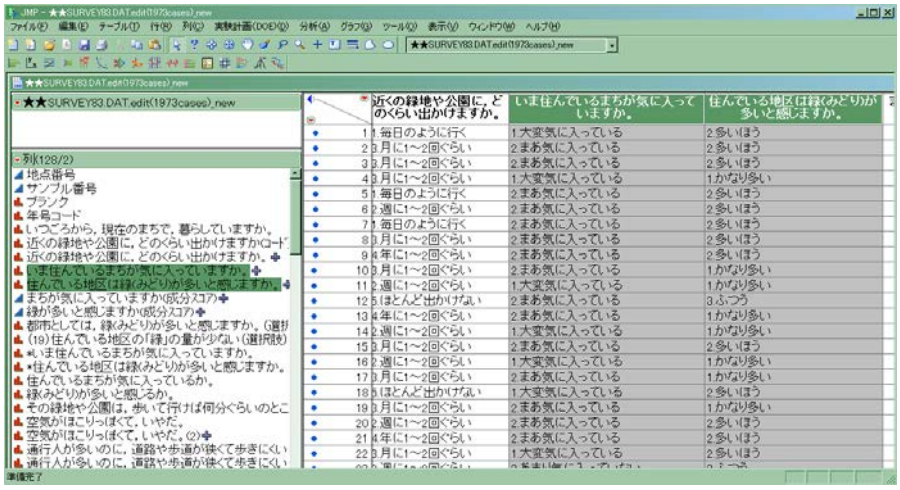


表 3　質問 I と質問 J との 2 元クロス表: $\mathbf{F}=\left(f_{ij}\right)$

	項目 J 質問 A：あなたは，いま住んでいるまちが気に入っていますか.						
項目 I 質問 B：あなたの住んでいる地区は，都市としては“緑（みどり）が多い”と感じますか.	選択肢（j） 選択肢（i）	1.かなり多い	2.多いほう	3.ふつう	4.少ないほう	5.少ない	行和 （f_{i+}）
	1.大変気に入っている	166	239	86	26	7	**524**
	2.まあ気に入っている	131	598	324	146	36	**1,235**
	3.あまり気に入っていない	6	40	55	51	20	**172**
	4.気に入っていない	2	2	0	5	6	**15**
	列和（f_{+j}）	**305**	**879**	**465**	**228**	**69**	**1,946** （N）

対応分析法の基本となる出発行列は“**2 元データ表**”であり，その典型例が“**2 元クロス表**”である．ここでも上の 2 つの質問についてこのクロス表を作る．上の課題を言い替えると，たとえばここで，「1．大変気に入っている」かつ「2．かなり多い」を選んだ人と「3.あまり気に入っていない」「4.少ないほう」を選んだ人の間には違いがあることはわかるが，その定量的な違いは数量として測れるのか，ということである．ここで，「1．大変気に入っている」は数値 1，「3.あまり気に入っていない」は数値 3 のコードが付与されているからその差は「2」だ，などと考えてよいのだろうか，ということである[11]．ではどうするか，それへの解答の 1 つを対応分析が与えてくれる．これを形式的に誘導してみよう．ここでは，統計ソフトウェア（JMP：ジャンプ）を用いて必要な諸量を求めてみる．また，最近はこういうことは統計ソフトウェアで簡単に処理が可能である．

ここでまず，のちの説明のために，寸法が（$m \times n$）の 2 元クロス表（表 4）を示す若干の記号・記法を用意しよう[12]．

$$\underset{m\times n}{\mathbf{F}}=\left(f_{ij}\right) \qquad \begin{pmatrix} f_{ij}\geq 0 \\ i\in I,\ j\in J \end{pmatrix} \tag{1}$$

ここで，I と J は，それぞれ行と列の項目の選択肢の集合を表わす．

$$I=\{1,2,\cdots,m\},\quad J=\{1,2,\cdots,n\} \tag{2}$$

11　こうしたコーディングを行って得点化する方式がライカート方式である．

12　詳細は，別資料「第Ⅱ部」に述べたので，そちらを参照．

表 4（項目 I× 項目 J）のクロス表 $\underset{m\times n}{\mathbf{F}}=\left(f_{ij}\right)$

		項　目　J						
	選択肢	1	2	…	j	…	n	行和
項目 I	1	f_{11}	f_{12}	…	f_{1j}	…	f_{1n}	f_{1+}
	2	f_{21}	f_{22}	…	f_{2j}	…	f_{2n}	f_{2+}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	i	f_{i1}	f_{i2}	…	f_{ij}	…	f_{in}	f_{i+}
	⋮	⋮	⋮		⋮		⋮	⋮
	m	f_{m1}	f_{m2}	…	f_{mj}	…	f_{mn}	f_{m+}
	列和	f_{+1}	f_{+2}	…	f_{+j}	…	f_{+n}	f_{++}

ここで，行に質問 I，列に質問 J を対応させたとして，それぞれの**選択肢**[13]に相当する．よって，この記法を用いると，表 3 の質問の選択肢の集まりは次のように書ける．

- 質問 I の 4 つの選択肢を $I=\{1,2,3,4\}$，つまり $m=4$．
- 質問 J の 5 つ選択肢を $J=\{1,2,3,4,5\}$，つまり，$n=5$．
- また，$f_{ij}\left(i\in I, j\in J\right)$ は，各セル（マス）の回答度数（頻度），たとえば，「3.あまり気に入っていない」「2.多いほう」を選んだ回答者は $f_{32}=40$（人）あったとなる．

さらに，以下の記法を用意する．

$$f_{i+}=\sum_{j=1}^{n} f_{ij} \quad \text{（行和）} \tag{3}$$

$$f_{+j}=\sum_{i=1}^{m} f_{ij} \quad \text{（列和）} \tag{4}$$

$$\sum_{i=1}^{m}\sum_{i=1}^{n} f_{ij}=\sum_{i=1}^{m} f_{i+}=\sum_{j=1}^{n} f_{+j}=N \quad \text{（総度数）} \tag{5}$$

ここで 2 元クロス表の**相対度数**つまり**確率分布**を考える．これに関連する以下の行列，ベクトルを用意する．

$$\mathbf{P}_{IJ}=\frac{1}{N}\mathbf{F}=\left(p_{ij}\right) \quad \left(i\in I, j\in J\right) \text{（同時確率分布）} \tag{6}$$

$$\mathbf{P}_{I}=diag\left(p_{i+}\right) \quad \left(i\in I\right) \text{（行の周辺確率分布）} \tag{7}$$

$$\mathbf{P}_{J}=diag\left(p_{+j}\right) \quad \left(j\in J\right) \text{（列の周辺確率分布）} \tag{8}$$

なおここで，

$$p_{ij}=\frac{f_{ij}}{N} \quad \text{ここで } N=\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} \ \left(\equiv f_{++}\right) \tag{9}$$

13 これを，“カテゴリー”ということが多いかもしれないが．ここでは選択肢とした．ほかに，オプション，モダリティ（modalité），フォルム（forme）などともいう．

$$p_{i+}=\frac{f_{i+}}{N}=\frac{\sum_{j=1}^{n} f_{ij}}{N},\ p_{+j}=\frac{f_{+j}}{N}=\frac{\sum_{i=1}^{m} f_{ij}}{N} \qquad (10)$$

である．また $diag(\bullet)$ は対角行列を意味する．

表 5 の中の，$\underset{m\times 1}{\mathbf{r}}=\left(p_{1+},p_{2+},\cdots,p_{i+},\cdots,p_{m+}\right)^t$，$\underset{1\times n}{\mathbf{c}}=\left(p_{+1},p_{+2},\cdots,p_{+j},\cdots,p_{+n}\right)^t$ はそれぞれ重心ベクトルであり，また対応分析法では（この個々の要素を）**“質量”**（mass）という．

$$\underset{m\times 1}{\mathbf{r}}=\left(p_{1+},p_{2+},\cdots,p_{i+},\cdots,p_{m+}\right)^t \quad（p_{i+}\text{を要素とする列ベクトル}） \qquad (11)$$

この各要素 p_{i+} を**“行の質量”**（row mass）とよび，$\underset{m\times 1}{\mathbf{r}}$ は**列プロファイルの平均ベクトル（重心）**に相当する．

$$\underset{1\times n}{\mathbf{c}}=\left(p_{+1},p_{+2},\cdots,p_{+j},\cdots,p_{+n}\right)^t \quad（p_{+j}\text{を要素とする列ベクトル}） \qquad (12)$$

一方，この各要素 p_{+j} を**“列の質量”**（column mass）とよび，$\underset{1\times n}{\mathbf{c}}$ は**行の平均ベクトル（重心）**となる．

これでクロス表 $\underset{m\times n}{\mathbf{F}}$ を説明する記号はほぼそろった．では，このクロス表 $\underset{m\times n}{\mathbf{F}}=\left(f_{ij}\right)$ から実際に得られた情報，つまり統計ソフトウェアが出力した情報を拾い出し，それぞれについて説明解釈を試みよう．

表 5　確率行列 $\mathbf{P}_{IJ}$ と行および列の相対確率

		項　目　J							
		1	2	…	j	…	n	行の確率（$\mathbf{P}_I$ の対角要素）	行の平均ベクトル
項目 I	1	p_{11}	p_{12}	…	p_{1j}	…	p_{1n}	p_{1+}	$\underset{m\times 1}{\mathbf{r}}=\begin{pmatrix}p_{1+}\\p_{2+}\\\vdots\\p_{i+}\\\vdots\\p_{m+}\end{pmatrix}$
	2	p_{21}	p_{22}	…	p_{2j}	…	p_{2n}	p_{2+}	
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
	i	p_{i1}	p_{i2}	…	p_{ij}	…	p_{in}	p_{i+}	
	⋮	⋮	⋮		⋮		⋮	⋮	
	m	p_{m1}	p_{m2}	…	p_{mj}	…	p_{mn}	p_{m+}	
	列の確率（$\mathbf{P}_J$ の対角要素）	p_{+1}	p_{+2}	…	p_{+j}	…	p_{+n}	1（総和）	↑ 列プロファイルの重心
	列の平均ベクトル	$\underset{n\times 1}{\mathbf{c}}=\left(p_{+1},p_{+2},\cdots,p_{+j},\cdots,p_{+n}\right)^t$						← 行プロファイルの重心	

まず，対応分析の結果を見る前の準備として，このクロス表の**“独立性の検定”**を行ってみる．つまり，2 つの質問 I,J は独立である（関連がない）を「帰無仮説」として，この 2

つの質問 I, J 間の関連性を調べる．このとき用いる検定統計量の 1 つが，次の**ピアソンのカイ二乗検定統計量**である（以下では「ピアソンのカイ二乗統計量」あるいは縮めて「カイ二乗統計量」ということがある）．じつは対応分析では，このカイ二乗統計量が重要な意味を持つので，ここで調べておく．

$$\chi_p^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{N\left(p_{ij} - p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\left(f_{ij} - \dfrac{f_{i+}f_{+j}}{N}\right)^2}{\dfrac{f_{i+}f_{+j}}{N}} \tag{13}$$

ここで，$\dfrac{f_{i+}f_{+j}}{N}$ は，クロス表の独立性の検定で，**独立モデル**を想定したときの第 (i, j) セルの**期待度数** e_{ij} で，$e_{ij} = Np_{i+}p_{+j} = N\dfrac{f_{i+}}{N}\dfrac{f_{+j}}{N} = \dfrac{f_{i+}f_{+j}}{N}$ となる（ここで，$p_{i+} = \dfrac{f_{i+}}{N}$，$p_{+j} = \dfrac{f_{+j}}{N}$）．つまり，上の式は次のようにも書ける．

$$\chi_p^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\left(f_{ij} - \dfrac{f_{i+}f_{+j}}{N}\right)^2}{\dfrac{f_{i+}f_{+j}}{N}} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\left(f_{ij} - e_{ij}\right)^2}{e_{ij}} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(\text{実現度数} - \text{期待度数})^2}{\text{期待度数}}$$

通常，クロス表（分割表）の独立性の検定では，このピアソンのカイ二乗統計量 χ_p^2 が自由度（*d.f.*: degree of freedom）$(m-1)(n-1)$ の χ^2 **分布に“近似する”**ことを使って χ^2 検定を行う[14].

$$\chi_p^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\left(f_{ij} - \dfrac{f_{i+}f_{+j}}{N}\right)^2}{\dfrac{f_{i+}f_{+j}}{N}} \simeq \chi_{(m-1)(n-1)}^2 \tag{14}$$

たとえば，表 3 のクロス表から得た統計ソフトウェア（JMP）の出力結果をそのまま引用すると，以下のようになる．JMP など，多くの統計ソフトウェアでは，検定統計量としてピアソンのカイ二乗統計量の他に尤度比カイ二乗統計量なども表示することが多いがここでは取り上げない（関連書を参照）．

住んでいる地区は緑(みどり)が多いと感じますか。といま住んでいるまちが気に入っていますか。の分割表に対する分析

検定

N	自由度	(-1)*対数尤度	R2乗(U)
1946	12	141.17134	0.0812

検定	カイ2乗	p値(Prob>ChiSq)
尤度比	282.343	<.0001*
Pearson	340.309	<.0001*

図 1　クロス表の独立性の検定の結果

14 初等統計の基本的な知識の 1 つであるから，大抵の統計学の本に説明がある．また，Everitt（1999）などを参照するとよい．なおここで，離散型のカイ二乗統計量を連続型確率分布である χ^2 分布で近似することに注意しよう．

ここでは，ピアソンのカイ 2 乗統計量は $\chi_p^2 = 340.309$ となり，統計的検定の結果は有意確率（< 0.0001*）とあるので“高度に有意”となる．つまり，2 つの質問 I, J の間には「何らか」の関連がありそうだ（ないとは言えない），となる．しかし，この検定結果だけからは，「2 つの質問 I, J の選択肢には具体的にどのような関連」があるかまでは分からない．

では，これを対応分析法ではどう読むのであろうか．再び統計ソフトウェアが出力する情報を形式的におってみる．ここでは，数式による細かい説明や数理展開はひとまず横において，対応分析法が何を行うかを表に要約した（表 7，表 8）．ベンゼクリ他の進めてきたフランスにおける対応分析の研究では，所与の 2 元データ表（たとえばクロス表）は一種の**多次元データ**と考えることからはじまる．このデータ表の寸法が示す次元数，つまり行数（m）次元あるいは列数（n）次元の空間内に分布するデータと考える．つまり，行の側からみれば n 次元空間 R^n 内にある m 個の点の分布であり，列の側からみれば m 次元空間 R^m 内での n 個の点の分布ということになる（表 6 に要約）．

ここで行の側を考えてみよう．この場合，m 個の行から作られる**“行プロファイル”**（row profile）間の距離として**“カイ二乗距離”**（Chi-square distance）を用いる（行プロファイルは表 6）．そして，このカイ二乗距離を良く近似するように，低次元のユークリッド空間に射影する．列に関しても，同じ考え方を適用する．列の場合，n 個の**“列プロファイル”**（column profile）の間の距離もカイ二乗距離を用いる．この列の側でも，このカイ二乗距離を良く近似するように，低次元のユークリッド空間に射影する．

要点は一般のユークリッド距離ではなく，重み付きのユークリッド距離である**“カイ二乗距離”**を用いることにある．このカイ二乗距離は表 6 にあるような式で与えられる．これが対応分析法でどのように機能するかについては，順次述べることにし，ここでは，いま取り上げた調査データについて，この一連の操作を形式的に適用し，対応分析法を行ってみよう．なおここでは，統計ソフトウェアが出力する情報を順に追ってみる．

① 固有値，特異値ほかの情報

まず，**特異値**，**固有値，寄与率**ほかが出力される（表 9）．これを読み替えると以下のようになる．

- **固有値**（λ_k）とは，次に示す**“成分スコア”**[16]の**分散**である．つまり対応分析で得た成分スコアの分散である．対応分析ではこれを**“慣性”**（inertia）という[17].
- **特異値**（α_k）とは固有値の正の平方根（$\sqrt{\lambda_k}$）である．これはまた，2 つの項目 $I.J$ のそれぞれの選択肢の各成分スコア間の相関係数に相当する（うしろに例で示す）．
- **寄与率**とは，各成分が**“全情報（総変動）”**つまり**“全慣性”**（total inertia）に占める寄与の程度を表す．ここでは，第 1 成分が約 71%，第 2 成分が約 25%の情報量がある．
- **固有値の総和**とは，このクロス表が多次元データとして示す情報の総量つまり**“全慣性”**であって，**カイ二乗統計量**（χ_p^2）を総度数（ここでは全回答者数）（N）で割った値に等しい．個々の固有値は，この総変動に占める各成分の説明力の大きさを示す情報にあたる．
- 対応分析の特性から，**固有値あるいは特異値の数**（K）は，$K = \min\{m,n\} - 1$ となる．つまり，所与のクロス表の行と列の数の小さいほうから 1 を引いた数となる．ここの例では，$K = \min\{m,n\} - 1 = \min\{4,5\} - 1 = 3$ であり，実際に 3 つの成分が得られる．

16；ここでは“coordonées”（仏），“coordinates”（英）に対して“成分スコア”とした．これを，主座標，座標，数量化得点，数量化スコア，最適スコア，得点，スコアなどともいう．

17 「慣性」とは本来は物理学や力学系で用いる用語の１つである．ベンゼクリは好んでこうした分野の用語を用いている．統計でいう分散は，いわゆる平均値の周りの２次の積率（モーメント）ともいうが，これに類した呼称である．

・　また，固有値の大きさは，$0 \le \lambda_k \le 1$，つまり非負で 1 を越えることはない．

② 2 つの質問文の各選択肢への成分スコア

成分スコアには，**“行の成分スコア”** $z_{ik}\left(i \in I; k = 1,2,\cdots,K\right)$ と **“列の成分スコア”** $z^*_{jk}\left(j \in J; k = 1,2,\cdots,K\right)$ がある．表 10 に，これの出力情報を整理した．元のクロス表の行，列の双方から同時的に観察することが対応分析法の特徴の 1 つである．

③ 成分スコアの平均値と分散の関係

ここで，成分スコアの**平均値**と**分散**を，求めた数値から実際に算出してみよう[18]．いま，行の側つまり質問 I の 4 つの選択肢に対する成分スコアと，列の側つまり質問 J の 5 つの選択肢に対する成分スコアを，それぞれ以下の記号で表す．

$$\text{行の選択肢}\, i \in I \,\text{に対する第}\, k \,\text{成分スコア}：\ z_{ik}\begin{pmatrix} i \in I \\ k = 1,2,\cdots,K \end{pmatrix} \qquad (15)$$

$$\text{列の選択肢}\ j \in J \ \text{に対する第}\, k \,\text{成分スコア}：\ z^*_{jk}\begin{pmatrix} j \in J \\ k = 1,2,\cdots,K \end{pmatrix} \qquad (16)$$

表 6　分析の方向

対応分析法の基本データ表	
① 寸法が $m \times n$ の 2 元データ表またはクロス表 $\underset{m\times n}{\mathbf{F}} = \left(f_{ij}\right) \quad \begin{pmatrix} f_{ij} \ge 0 \\ i \in I,\ j \in J \end{pmatrix}$ ② 確率行列とその周辺相対確率 $\mathbf{P}_{IJ} = \frac{1}{N}\mathbf{F} = \left(p_{ij}\right) \quad \left(i \in I, j \in J\right)$（同時確率分布） $\mathbf{P}_I = diag\left(p_{i+}\right) \quad \left(i \in I\right)$（行の周辺確率分布） $\mathbf{P}_J = diag\left(p_{+j}\right) \quad \left(j \in J\right)$（列の周辺確率分布）	
行の側から分析	列の側から分析
n 次元空間 R^n 内での分析	m 次元空間 R^m 内での分析
行和を 1 としたときの「行のプロファイル」で $\left(n-1\right)$ 次元内に分布する m 個の点 $\mathbf{N}_I = \mathbf{P}_I^{-1}\mathbf{P}_{IJ} = \left\{q_{ij} = \frac{p_{ij}}{p_{i+}} \middle\vert i \in I, j \in J\right\}$	列和を 1 としたときの「列のプロファイル」で $\left(m-1\right)$ 次元内に分布する n 個の点 $\mathbf{N}_J = \mathbf{P}_J^{-1}\mathbf{P}_{JI} = \left\{q^*_{ij} = \frac{p_{ij}}{p_{+j}} \middle\vert i \in I, j \in J\right\}$
行プロファイル間のカイ二乗距離 $d_B^2\left(i,i'\right) = \sum_{j=1}^{n}\frac{1}{p_{+j}}\left(q_{ij} - q_{i'j}\right)^2$ $= \sum_{j=1}^{n}\frac{1}{p_{+j}}\left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}}\right)^2$	列プロファイル間のカイ二乗距離 $d_B^2\left(j,j'\right) = \sum_{i=1}^{m}\frac{1}{p_{i+}}\left(q^*_{ij} - q^*_{ij'}\right)^2$ $= \sum_{i=1}^{m}\frac{1}{p_{i+}}\left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}}\right)^2$

18　この成分スコアについて，数量（連続変量）として記述的統計量を算出していることに注意しよう．

表 7　n 次元空間 R^n 内での分析

	その 1	その 2（対称化）
対象とするデータの形	$x_{ij}^{**}=x_{ij}-\bar{x}_j=\dfrac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}-\sqrt{p_{+j}}$ $\underset{m\times n}{\mathbf{X}}=\left(x_{ij}\right)\qquad \underset{m\times n}{\mathbf{X}^*}=\left(x_{ij}^{**}\right)$	$y_{ij}=\dfrac{p_{ij}}{\sqrt{p_{i+}p_{+j}}}$ $\underset{m\times n}{\mathbf{Q}}=\left(y_{ij}\right)$
対象とする行列（共分散行列）	$\underset{n\times n}{\mathbf{V}}=\left(s_{jj'}\right)=\mathbf{X}^t\mathbf{P}_I\mathbf{X}-\overline{\mathbf{x}}\overline{\mathbf{x}}^t$ $\underset{n\times n}{\mathbf{V}}=\left(s_{jj'}\right)=\left(\mathbf{X}^*\right)^t\mathbf{P}_I\mathbf{X}^*$	$\underset{n\times n}{\mathbf{V}^*}=\mathbf{Q}^t\mathbf{Q}$
固有値と特異値	上の行列の固有値は同じ値となる $\lambda_k\left(\begin{matrix}k=1,2,\cdots,K\\K=\min\{m,n\}-1\end{matrix}\right)$ $0\le\lambda_k\le 1$ 第 k 成分の特異値；$\alpha_k=\sqrt{\lambda_k}$	上の行列の固有値を（あえて）μ_k とおくと以下. $\mu_k\left(\begin{matrix}k=1,2,\cdots,K\\K=\min\{m,n\}-1\end{matrix}\right)$ $0\le\mu_k\le 1$ しかしここで，“$\mu_k=\lambda_k$”となる. 第 k 成分の特異値；$\alpha_k=\sqrt{\lambda_k}$
固有ベクトル	行列 $\underset{n\times n}{\mathbf{V}}$ の“固有値 $\lambda_0=0$”に対して自明の解として以下の固有ベクトル $\underset{1\times n}{\mathbf{l}_0^t}=\left(\sqrt{p_{+1}},\sqrt{p_{+2}},\cdots,\sqrt{p_{+j}},\cdots,\sqrt{p_{+n}}\right)$	行列 $\underset{m\times n}{\mathbf{Q}}$ の“固有値 $\mu_0=1$”に対して自明の解として以下の固有ベクトル $\underset{1\times n}{\mathbf{l}_0^t}=\left(\sqrt{p_{+1}},\sqrt{p_{+2}},\cdots,\sqrt{p_{+j}},\cdots,\sqrt{p_{+n}}\right)$
固有値の数	固有値の数，つまり得られる成分数は，どちらも $K=\min\{m,n\}-1$（個） たとえば，$m>n$ とすると，$m-n$ 個の固有値は「0」となる．つまり，行列の階数（ランク）の縮退がおこる.	
成分スコア	$x_{ij}^{**}=x_{ij}-\bar{x}_j=\dfrac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}-\sqrt{p_{+j}}$ $z_{ik}=\sum_{j=1}^{n}l_{jk}x_{ij}^{**}=\sum_{j=1}^{n}\left(\dfrac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}-\sqrt{p_{+j}}\right)l_{jk}$ $\left(i\in I;k=1,2,\cdots,K\right)$	$y_{ij}=\dfrac{p_{ij}}{\sqrt{p_{i+}p_{+j}}}$ $z_{ik}=\sum_{j=1}^{n}l_{jk}x_{ij}=\sum_{j=1}^{n}\left(\dfrac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}\right)l_{jk}$ $\left(i\in I;k=1,2,\cdots,K\right)$

表 8　m 次元空間 R^m 内での分析

	その 1	その 2（対称化）
対象とするデータ	$x_{ij}^{\dagger} = x_{ij} - \overline{x}_i = \dfrac{p_{ij}}{p_{+j}\sqrt{p_{i+}}} - \sqrt{p_{i+}}$ $\underset{m\times n}{\mathbf{X}} = \left(x_{ij}\right) \qquad \underset{m\times n}{\mathbf{X}}^{\dagger} = \left(x_{ij}^{\dagger}\right)$	$y_{ij} = \dfrac{p_{ij}}{\sqrt{p_{i+}p_{+j}}}$ $\underset{m\times n}{\mathbf{Q}} = \left(y_{ij}\right)$
対象とする行列	$\underset{m\times m}{\mathbf{S}} = \left(s_{ii'}\right) = \mathbf{X}\mathbf{P}_J\mathbf{X}^t - \overline{\mathbf{x}}\overline{\mathbf{x}}^t$ $\underset{m\times m}{\mathbf{S}} = \left(s_{ii'}\right) = \mathbf{X}^{\dagger}\mathbf{P}_J\left(\mathbf{X}^{\dagger}\right)^t$	$\underset{m\times m}{\mathbf{W}} = \mathbf{Q}\mathbf{Q}^t$ 注：上の $\underset{n\times n}{\mathbf{V}}^{*} = \mathbf{Q}^t\mathbf{Q}$ から得られる固有値と $\underset{m\times m}{\mathbf{W}}$ からえられる固有値は自明根を除き同じとなる.
固有値と特異値	上の行列の固有値は同じ値となる 第 k 成分の固有値： $\lambda_k \left(\begin{matrix} k = 1,2,\cdots,K \\ K = \min\{m,n\} - 1 \end{matrix} \right)$ $0 \le \lambda_k \le 1$ 第 k 成分の特異値；$\alpha_k = \sqrt{\lambda_k}$	上の行列の固有値を（あえて）μ_k とおくと以下. $\mu_k \left(\begin{matrix} k = 1,2,\cdots,K \\ K = \min\{m,n\} - 1 \end{matrix} \right)$ $0 \le \mu_k \le 1$ しかしここで，“$\mu_k = \lambda_k$”となる. 第 k 成分の特異値；$\alpha_k = \sqrt{\lambda_k}$
固有ベクトル	行列 $\underset{m\times m}{\mathbf{S}}$ の“固有値 $\lambda_0 = 0$”に対して自明の解として以下の固有ベクトル $\underset{1\times m}{\mathbf{u}_0^t} = \left(\sqrt{p_{1+}}, \sqrt{p_{2+}}, \cdots, \sqrt{p_{i+}}, \cdots, \sqrt{p_{im}}\right)$	行列 $\underset{m\times m}{\mathbf{W}}$ の“固有値 $\mu_0 = 1$”に対して自明の解として以下の同じ固有ベクトル $\underset{1\times m}{\mathbf{u}_0^t} = \left(\sqrt{p_{1+}}, \sqrt{p_{2+}}, \cdots, \sqrt{p_{i+}}, \cdots, \sqrt{p_{im}}\right)$
固有値の数	固有値の数，つまり得られる成分数は，どちらも $K = \min\{m,n\} - 1$（個）	
成分スコア	$x_{ij}^{\dagger} = x_{ij} - \overline{x}_i = \dfrac{p_{ij}}{p_{+j}\sqrt{p_{i+}}} - \sqrt{p_{i+}}$ $z_{jk}^{*} = \sum_{i=1}^{m} l_{jk} x_{ij}^{\dagger} = \sum_{j=1}^{m} \left(\dfrac{p_{ij}}{p_{+j}\sqrt{p_{i+}}} - \sqrt{p_{i+}} \right) u_{ik}$ $\left(j \in J; k = 1,2,\cdots,K\right)$	$y_{ij} = \dfrac{p_{ij}}{\sqrt{p_{i+}p_{+j}}}$ $z_{jk}^{*} = \sum_{i=1}^{m} l_{jk} x_{ij}^{**} = \sum_{j=1}^{m} \left(\dfrac{p_{ij}}{p_{+j}\sqrt{p_{i+}}} \right) u_{ik}$ $\left(j \in J; k = 1,2,\cdots,K\right)$

表 9　固有値，特異値，寄与率，累積寄与率

成分 k	特異値 α_k	固有値（慣性） λ_k	寄与率	累積寄与率	累積寄与率（%）
1	0.35288	0.12452	0.7121	0.7121	71.2
2	0.20959	0.04393	0.2512	0.9633	96.3
3	0.08014	0.00642	0.0367	1.0000	(100)
	固有値の総和	**0.17487**	—	—	—

表 10　成分スコアの要約

	対応する固有値：	$\lambda_1 = 0.12452$	$\lambda_2 = 0.04393$	$\lambda_3 = 0.00642$
質問文	質問選択肢	第 1 成分スコア	第 2 成分スコア	第 3 成分スコア
		z_{i1}	z_{i2}	z_{i3}
質問 A：あなたは，いま住んでいるまちが気に入っていますか.	1.大変気に入っている	-0.4442	0.2027	-0.0353
	2.まあ気に入っている	0.0623	-0.1315	0.0311
	3.あまり気に入っていない	0.7886	0.1907	-0.1698
	4.気に入っていない	1.3458	1.5567	0.6157
質問文	質問選択肢	z^*_{j1}	z^*_{j2}	z^*_{j3}
質問 B：あなたの住んでいる地区は，都市としては“緑（みどり）が多い”と感じますか.	1. かなり多い	-0.5403	0.3235	-0.0640
	2.多いほう	-0.1118	-0.1055	0.0657
	3.ふつう	0.1545	-0.1506	-0.0613
	4.少ないほう	0.5530	0.0750	-0.1069
	5.少ない	0.9438	0.6805	0.2119

ここで，固有値あるいは特異値の個数は，$K=\min\{m,n\}-1$ となるので，いまのクロス表の場合は，$K=\min\{m,n\}-1=\min\{4,5\}=3$ となる．つまりここでは，3 つの固有値に対応して第 1～第 3 成分スコアまでがえられる（表 10）.

表 11　成分スコアの平均値の算出例

		z_{ik} の平均値 $\overline{z_k}$ の算出		
	i \ k	第 1 成分	第 2 成分	第 3 成分
$f_{i+}z_{ik}$ の値	1	-232.7608	106.2148	-18.4972
	2	76.9405	-162.4025	38.4085
	3	135.6392	32.8004	-29.2056
	4	20.1870	23.3505	9.2355
平均値	$\overline{z_k}$	**0.0000**	**0.0000**	**0.0000**
		z^*_{jk} の平均値 $\overline{z}^*_k$ の算出		
	j \ k	第 1 成分	第 2 成分	第 3 成分
$f_{+j}z^*_{jk}$ の値	1	-164.7915	98.6675	-19.5200
	2	-98.2722	-92.7345	57.7503
	3	71.8425	-70.0290	-28.5045
	4	126.0840	17.1000	-24.3732
	5	65.1222	46.9545	14.6211
平均値	$\overline{z}^*_k$	**0.0000**	**0.0000**	**0.0000**

以上を準備として，この成分スコアの平均値と分散を求めてみよう．このとき，z_{ik} の加重 f_{i+}（あるいは p_{i+}）付の平均値となることに注意する．z^*_{jk} についても同様に，加重 f_{+j}（あるいは p_{+j}）付の平均値となる．元の $N=1{,}496$ 人の個々にこのスコアが付与され，これがクロス表に集約された各行和あるいは列和となるからである（うしろに挙げた，成分スコアと個々の回答の関係，表 13，表 14 を参照）.

$$\text{行の第}k\text{成分スコアの平均値：}\ \overline{z}_k=\frac{1}{N}\sum_{i=1}^{m}f_{i+}z_{ik}=\sum_{i=1}^{m}p_{i+}z_{ik}=0\ \left(k=1,2,\cdots,K\right)\quad(17)$$

列の第k成分スコアの平均値：$\bar{z}_k^* = \frac{1}{N}\sum_{j=1}^{n} f_{+j} z_{jk}^* \sum_{j=1}^{n} p_{+j} z_{jk}^* = 0 \left(k = 1,2,\cdots,K\right)$　(18)

つまり，**成分スコアの平均値**はすべて「0」で，$\bar{z}_k = 0, \bar{z}_k^* = 0$となる．言い替えると，平均値の周りで中心化したことになっている．

つぎに，成分スコアの分散を求めよう．

$$V\left[z_{ik}\right] = \frac{1}{N}\sum_{i=1}^{m} f_{i+}\left(z_{ik} - \bar{z}_k\right)^2$$

行の第k成分スコアの分散：
$$= \frac{1}{N}\sum_{i=1}^{m} f_{i+} {z_{ik}}^2 \qquad (19)$$
$$= \sum_{i=1}^{m} p_{i+} {z_{ik}}^2 \left(k = 1,2,\cdots,K\right)$$

$$V\left[z_{jk}^*\right] = \frac{1}{N} f_{+j}\left(z_{jk}^* - \bar{z}_k^*\right)^2$$

列の第k成分スコアの分散：
$$= \frac{1}{N}\sum_{j=1}^{n} f_{+j}\left(z_{jk}^*\right)^2 \qquad (20)$$
$$= \sum_{j=1}^{n} p_{+j}\left(z_{jk}^*\right)^2 \left(k = 1,2,\cdots,K\right)$$

表 12　成分スコアの分散の算出例

		z_{ik} の分散算出		
	j \ k	第 1 成分	第 2 成分	第 3 成分
$f_{i+}z_{ik}^2$ の値	1	103.39235	21.52974	0.65295
	2	4.79339	21.35593	1.19450
	3	106.96507	6.25504	4.95911
	4	27.16766	36.34972	5.68630
分散	$V\left[z_{ik}\right]$	**0.12452**	**0.04393**	**0.00642**
		z_{jk}^* の分散算出		
	j \ k	第 1 成分	第 2 成分	第 3 成分
$f_{+j}\left(z_{jk}^*\right)^2$ の値	1	89.03685	31.91894	1.24928
	2	10.98683	9.78349	3.79419
	3	11.09967	10.54637	1.74733
	4	69.72445	1.28250	2.60550
	5	61.46233	31.95254	3.09821
分散	$V\left[z_{jk}^*\right]$	**0.12452**	**0.04393**	**0.00642**

ここでまず，2 つの分散$V\left[z_{ik}\right]$，$V\left[z_{jk}^*\right]$が，いずれも 3 つの固有値に等しいことがわかる．つまり，ここでは，

$$\begin{cases} V\left[z_{i1}\right] = V\left[z_{j1}^*\right] = \lambda_1 = 0.12452 \\ V\left[z_{i2}\right] = V\left[z_{j2}^*\right] = \lambda_2 = 0.04393 \quad (\text{ここで，} 1 > \lambda_1 > \lambda_2 > \lambda_3 > 0\ ) \\ V\left[z_{i3}\right] = V\left[z_{j3}^*\right] = \lambda_3 = 0.00642 \end{cases} \qquad (21)$$

となる．また，以下の関係にある．

$$\sum_{k=1}^{3} V\left[z_{i1}\right] = \sum_{k=1}^{3} V\left[z_{jk}^{*}\right] = \sum_{k=1}^{3} \lambda_k = 0.17487$$
$$\Updownarrow \qquad (22)$$
$$\frac{\chi_p^2}{N} = \frac{340.309}{1946} = 0.174876\cdots \doteqdot 0.17488$$

ここで，χ_p^2は前に示した式（7）のカイ二乗統計量である．つまり，ここで次の重要な関係がなりたつ．

$$\sum_{k=1}^{K} \lambda_K = \frac{\chi_p^2}{N} \quad \text{（総変動，慣性の総和）} \quad \Leftrightarrow \quad \begin{cases} \sum_{k=1}^{K} \lambda_K = 0.174874 \\ \frac{\chi_p^2}{N} = \frac{340.309}{1946} = 0.174876\cdots = 0.17488 \end{cases} \qquad (23)$$

こうして，以下の重要な性質があることがわかる．

ⅰ）成分スコアの分散の和（総変動）は，固有値の総和（慣性の総和，全慣性）に等しい．

ⅱ）カイ二乗統計量を総サンプル数で割った量と固有値の総和は等しい．

④ 分析の対称性

行と列との成分スコアの分散が等しいということは，はじめの出発行列である表 3 のクロス表を転置したクロス表から出発しても解が同じであること示唆しているが，実際にそうなる．上の 2 つの要約表 11，12 からも明かである．この**“対称性”**があることが，対応分析法の特徴でもあり，利便性の高い理由の 1 つでもある[24]．

⑤ 第 1 成分スコアの観察（その 1）

ここでとくに，第 1 成分スコアに注目し，これを数直線上にプロットしてみよう．つまり，2 つの質問I, Jの各選択肢に付与された**“数量（スコア）”**がどのように分布するかを調べる．表 10 から，成分スコアを拾い，図を描くと下の 2 つの図となる（図 2）．

上が質問Iの 4 つの選択肢「1.大変気に入っている」「2.まあ気に入っている」「3.あまり気に入っていない」「4.気に入っていない」に，また下が質問Jの 5 つの選択肢「1．かなり多い」「2.多いほう」「3.ふつう」「4.少ないほう」「5.少ない」の成分スコアの並びである．

これは，選択肢という質的データである名義尺度（であり順序尺度）に対して，この調査回答者の意識・意見が，新たに数値として比較可能な（数値演算が可能な）“数量”に変換されたと考えることもできる（尺度化の観点から考える）．

ここで質問Iについては，4 つの選択肢は等間隔とはならないが，選択肢の並び（序列）は，「1.大変気に入っている」「2.まあ気に入っている」「3.あまり気に入っていない」「4.気に入っていない」の順序は保持されている．「4.気に入っていない」が離れており，「1.大変気に入っている」と「2.まあ気に入っている」は近いようだ，ということがわかる．

つぎに，質問Jについては，これも 5 つの選択肢の並び順はそろってはいるが，それぞれの**“距離”**には差違がある．「4.少ないほう」「5.少ない」がやや遠く，「1．かなり多い」も反対の意味で遠く，「2.多いほう」「3.ふつう」が近いとなっている．

こうした距離感がこの調査回答者の意識尺度の表れであると考えるのが**“数量化”の目標**

---

24 多変量解析，多次元データ解析の手法で，似たような関係を示す手法はあるが，質的データを扱う自由度が高いという点で対応分析法は使い易い．

である[25]．ここでは，さいわいにも，2 つの質問のいずれもが，元の名義尺度・順序尺度の選択肢の並びの順を保持しているが，**いつもこうなるわけではない**ことに注意しよう．言い替えると，そうであるから成分スコアという数量化により質問の選択肢がその質問文を作るときに意図したとおりに機能したか（たとえば，選択肢の序列が保持されたか否か），回答者が意図のとおりに回答する傾向がみられるのか（たとえば，同じ質問に類似の成分スコアの傾向が表れることがあるか），などを知ることが意味を持つのである．

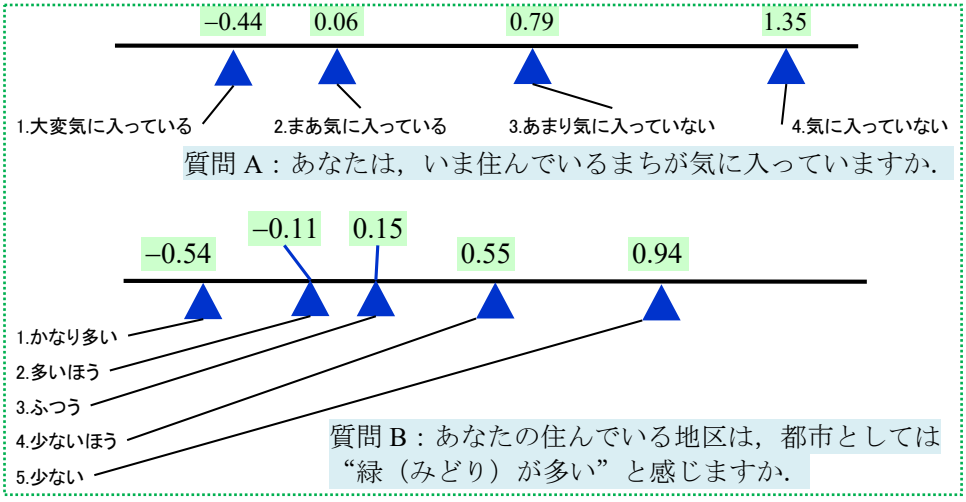


図 2　第 1 成分スコアの観察

⑥ 第 1 成分スコアの観察（その 2）：散布図とクロス表の観察

ここで，行と列，それぞれの第 1 成分スコアと，元のクロス表と合わせて観察してみよう．上では，第 1 成分スコアをそれぞれもとの選択肢との関連で比べたが，これの対応関係の意味（相関）をクロス表に対応させて観察しよう．

求めた行と列の第 1 成分スコアと元のクロス表との関係は，以下のように読み替えられる．たとえば，表 3 のクロス表で，質問 I から選択肢「1.大変気に入っている」を，質問 J から選択肢「1．かなり多い」を選んだ回答者は「166（人）」（$=f_{11}$）である．そしてこの選択肢に対しては，成分スコアが「1. かなり多い」$\Leftrightarrow z_{ik}=-0.4442$，「1. かなり多い」$\Leftrightarrow z^{*}_{jk}=-0.5403$ が対応する．つまり，この同じ成分スコアを持つ（回答した）回答者が 166（人）であった，ということである．以下同じように，クロス表のすべてのセルの組み合わせ，つまりセル (i,j) が，$(i,j)=(1,1)$（「1．大変気に入っている」かつ「2．かなり多い」）から，$(i,j)=(4,5)$（「4.気に入っていない」かつ「5．少ない」）までの，$4\times5=20$ のセル内の回答パターンに集約される．ただし，ここでは，$(i,j)=(4,3)$ の「4.気に入っていない」かつ「3．ふつう」の回答者はいないから「度数ゼロ」となる．これに成分スコアを付与して要約すると表 13，表 14 が得られる（ここは 20 通りのすべてのパターンで示した）．つまり，表 3 にみるように，回答者は 1,496 名であっても，実際の回答パターンは表 14 の 20 通りのどれかの組合せであり，また成分スコアとなる．いまこの表から，行と列のそれぞれの成分スコアを座標として，相対的な度数の大きさを円の面積としたバブルプロットを描いてみる（図 3[26]，図 4）．ここで，いくつかの組合せについて，クロス表のセル度数を書き入れてあるので，元の 2 元クロス表と比べてみるとよい．

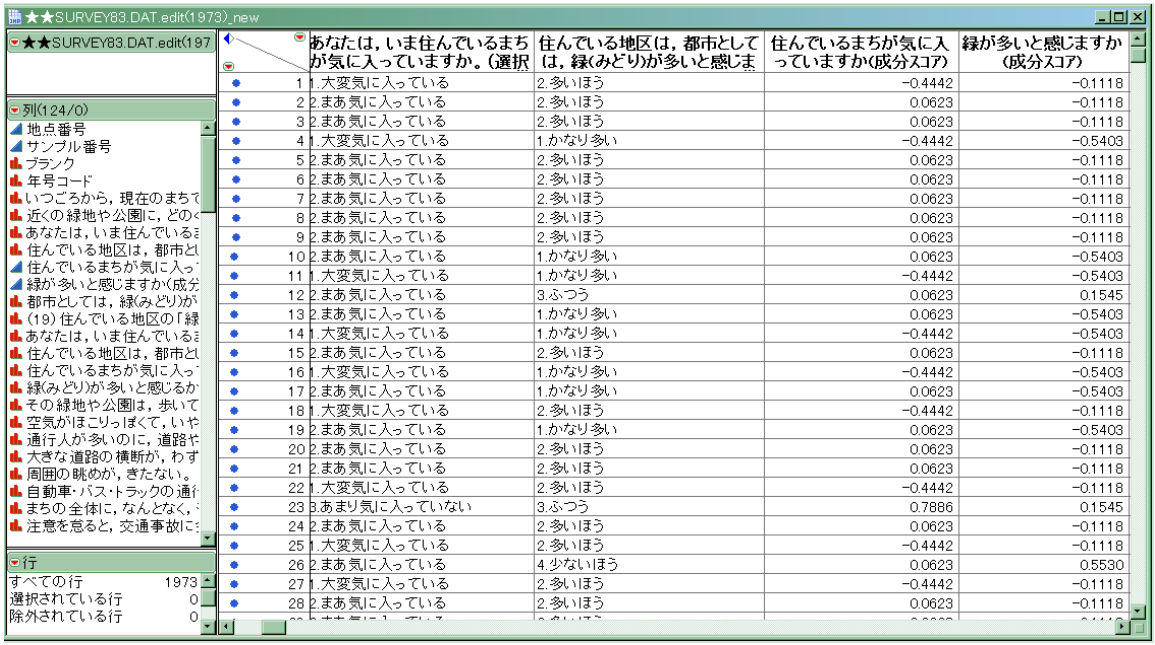
---

[25] じつは，この調査の 3 年次（6 地域）における，対応分析で得たこの 2 つの質問の回答傾向はほぼ同じであることがわかっている．ライカート方式の得点化の立場からいうと，この順序尺度へのコーディングによる数値化情報は，そのまま用いても（例：平均値を求めるなど）大きな齟齬は生じないかもしれないことを示唆している．

[26] 求めた成分スコアの単位で描いた散布図を“双対散布図”（twin-map）などという（Gaugh（1977））．

同時に**相関係数**を求めてみよう．ここで相関係数は $Cor\left(z_{i1}, z^*_{j1}\right) = 0.35288$ となる．これは，第 1 成分の固有値 $\lambda_1 = 0.12452$ の正の平方根，つまり**特異値** $\alpha_1 = 0.35288$ に相当する．以下，第 2 成分，第 3 成分についても同様の関係がなりたつ（そうなることが対応分析である）．つまり，$Cor\left(z_{i2}, z^*_{j2}\right) = 0.20959, Cor\left(z_{i3}, z^*_{j3}\right) = 0.08014$ となる．つまり，第 1 成分の情報（関連性）がもっとも大きく，第 2，第 3 成分と情報量が低減するにつれ相関が弱くなる（それが固有値の大きさの順序，つまり寄与率の大きさの順に対応するという意味）．

表 13　回答の選択肢に第1成分スコアを付与して要約した情報



	あなたは，いま住んでいるまちが気に入っていますか。(選択	住んでいる地区は，都市としては，緑(みどり)が多いと感じま	住んでいるまちが気に入っていますか(成分スコア)	緑が多いと感じますか(成分スコア)
1	1.大変気に入っている	2.多いほう	−0.4442	−0.1118
2	2.まあ気に入っている	2.多いほう	0.0623	−0.1118
3	2.まあ気に入っている	2.多いほう	0.0623	−0.1118
4	1.大変気に入っている	1.かなり多い	−0.4442	−0.5403
5	2.まあ気に入っている	2.多いほう	0.0623	−0.1118
6	2.まあ気に入っている	2.多いほう	0.0623	−0.1118
7	2.まあ気に入っている	2.多いほう	0.0623	−0.1118
8	2.まあ気に入っている	2.多いほう	0.0623	−0.1118
9	2.まあ気に入っている	2.多いほう	0.0623	−0.1118
10	2.まあ気に入っている	1.かなり多い	0.0623	−0.5403
11	1.大変気に入っている	1.かなり多い	−0.4442	−0.5403
12	2.まあ気に入っている	3.ふつう	0.0623	0.1545
13	2.まあ気に入っている	1.かなり多い	0.0623	−0.5403
14	1.大変気に入っている	1.かなり多い	−0.4442	−0.5403
15	2.まあ気に入っている	2.多いほう	0.0623	−0.1118
16	1.大変気に入っている	1.かなり多い	−0.4442	−0.5403
17	2.まあ気に入っている	1.かなり多い	0.0623	−0.5403
18	1.大変気に入っている	2.多いほう	−0.4442	−0.1118
19	2.まあ気に入っている	1.かなり多い	0.0623	−0.5403
20	2.まあ気に入っている	2.多いほう	0.0623	−0.1118
21	2.まあ気に入っている	2.多いほう	0.0623	−0.1118
22	1.大変気に入っている	2.多いほう	−0.4442	−0.1118
23	3.あまり気に入っていない	3.ふつう	0.7886	0.1545
24	2.まあ気に入っている	2.多いほう	0.0623	−0.1118
25	1.大変気に入っている	2.多いほう	−0.4442	−0.1118
26	2.まあ気に入っている	4.少ないほう	0.0623	0.5530
27	1.大変気に入っている	2.多いほう	−0.4442	−0.1118
28	2.まあ気に入っている	2.多いほう	0.0623	−0.1118

⑦ 第 1，第 2 成分スコアの布置図と同時布置図

複数の成分スコアを比べるために，散布図や散布図行列を**“布置図”**（representation あるいは configuration）として用いる．対応分析法では，この布置図による**視覚化**が重要である（これが分析目標の 1 つ）．質問文の選択肢の並び順や尺度化の様子，それと選択肢間の対応などを直感的に把握できるからである．このとき，情報の多い成分スコアから，つまり分散（であり固有値）の大きい成分スコアから観察する．よって，通常は寄与率を考慮し，固有値の大きい方からの数成分の組み合わせを**布置図**あるいは**散布図行列**として観察する．

通常は，k, k' 成分についての，

行成分スコア $\left(z_{ik}, z_{ik'}\right)$ の布置図

列成分スコア $\left(z^*_{jk}, z^*_{jk'}\right)$ の布置図

これを行と列の成分スコア $\left(z_{ik}, z_{ik'}\right)$ と $\left(z^*_{jk}, z^*_{jk'}\right)$ を重ねた同時布置図

を作る．グラフィカル･ツールとしては，これらを切り替えて観察できるような機能を備えていることが望ましい．実際に，多くのソフトにはこうした機能がある．

いまの例について，始めの 2 成分（第 1，第 2）の成分スコアの**“布置図”**（図 5）と**“同時布置図”**（simultaneous representation）（図 6）である．図の円の大きさ（バブル）は，元のクロス表のセル内度数，つまり回答数に合わせて描いてある．ここにみるように，2 つの質問の選択肢の対応関係がきれいに読み取れる．ここでは，いずれの成分スコアも分散が固有

値となるようにした．つまり，分散を 1 とする標準化を行ってはいない[27]．

表 14　クロス表と第 1 成分スコアの関係を調べる

回答パターン	質問 A	質問 B	回答数（人）	行の第 1 成分スコア z_{i1}	列の第 1 成分スコア z^*_{j1}
1	1.大変気に入っている	1.かなり多い	166	-0.4442	-0.5403
2	1.大変気に入っている	2.多いほう	239	-0.4442	-0.1118
3	1.大変気に入っている	3.ふつう	86	-0.4442	0.1545
4	1.大変気に入っている	4.少ないほう	26	-0.4442	0.5530
5	1.大変気に入っている	5.少ない	7	-0.4442	0.9438
6	2.まあ気に入っている	1.かなり多い	131	0.0623	-0.5403
7	2.まあ気に入っている	2.多いほう	598	0.0623	-0.1118
8	2.まあ気に入っている	3.ふつう	324	0.0623	0.1545
9	2.まあ気に入っている	4.少ないほう	146	0.0623	0.5530
10	2.まあ気に入っている	5.少ない	36	0.0623	0.9438
11	3.あまり気に入っていない	1.かなり多い	6	0.7886	-0.5403
12	3.あまり気に入っていない	2.多いほう	40	0.7886	-0.1118
13	3.あまり気に入っていない	3.ふつう	55	0.7886	0.1545
14	3.あまり気に入っていない	4.少ないほう	51	0.7886	0.5530
15	3.あまり気に入っていない	5.少ない	20	0.7886	0.9438
16	4.気に入っていない	1.かなり多い	2	1.3458	-0.5403
17	4.気に入っていない	2.多いほう	2	1.3458	-0.1118
18	4.気に入っていない	3.ふつう	0	—	—
19	4.気に入っていない	4.少ないほう	5	1.3458	0.5530
20	4.気に入っていない	5.少ない	6	1.3458	0.9438

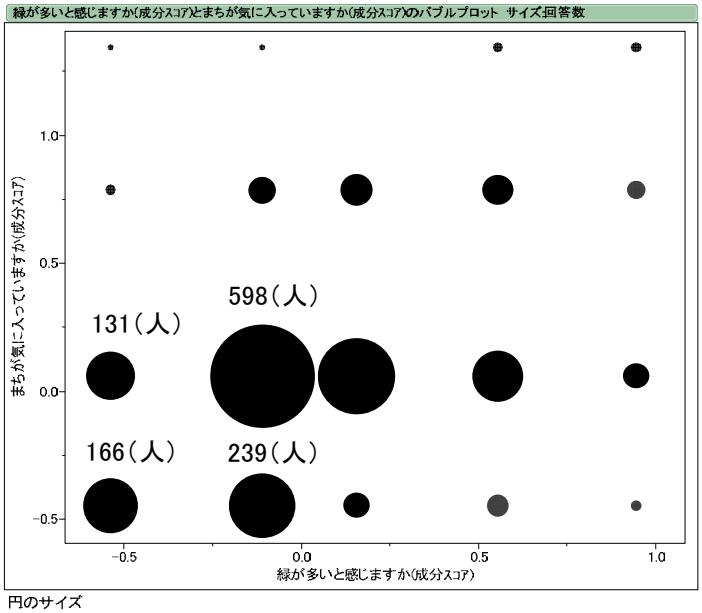


図 3　第 1 成分スコアと回答数を用いた双対散布図

バブル（●）の大きさがクロス表のセルの回答数のそれに対応する．目盛は成分スコアに合わせた

27 同時布置図の中で，列の成分スコアと行の成分スコアの相互の関係をどう読むかは，注意が必要である．これは双対性の性質の項で述べる．じつは同時布置図を描くとき，行・列それぞれの成分スコアの分散を固有値そのままとするか，標準化を行い分散 1 とするかがある（つまり，4 通りの組合せ）．ここでは，とくに断りがないかぎり，行・列いずれの成分スコアも標準化を行わず，分散（固有値）のままとする．これの違いについては「第Ⅱ部」に記した．

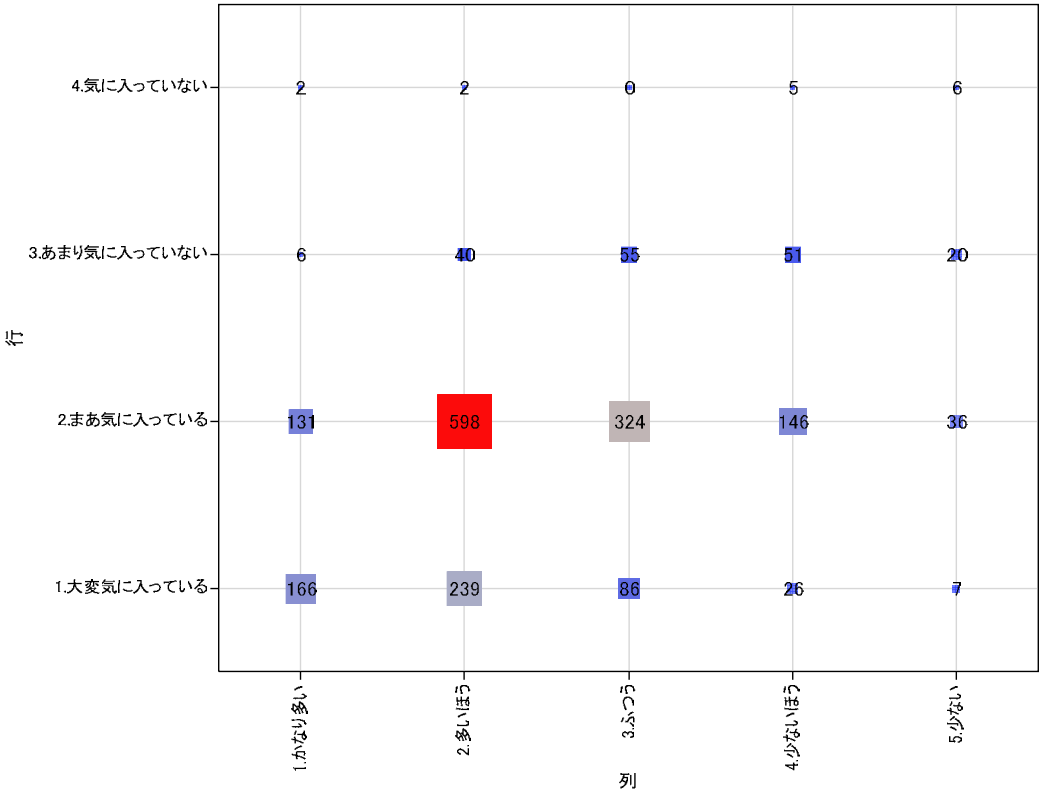


図 4　回答数を用いた双対散布図
■の大きさがクロス表のセルの回答数のそれに対応する
ここはクロス表のイメージに合わせて格子上に置いた（JMP 利用）

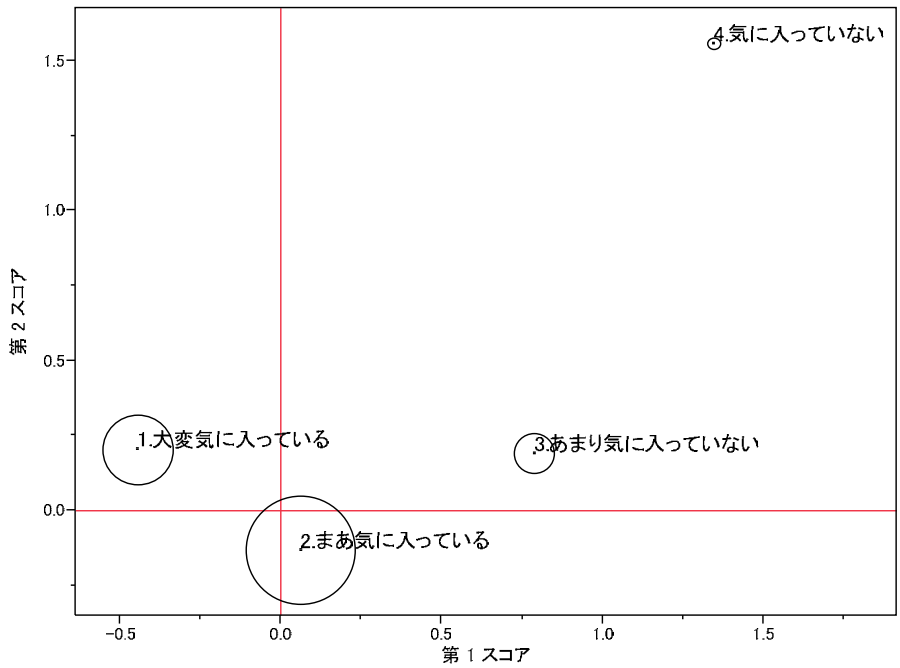


図 5－1　行（質問 I ）の成分スコア布置図

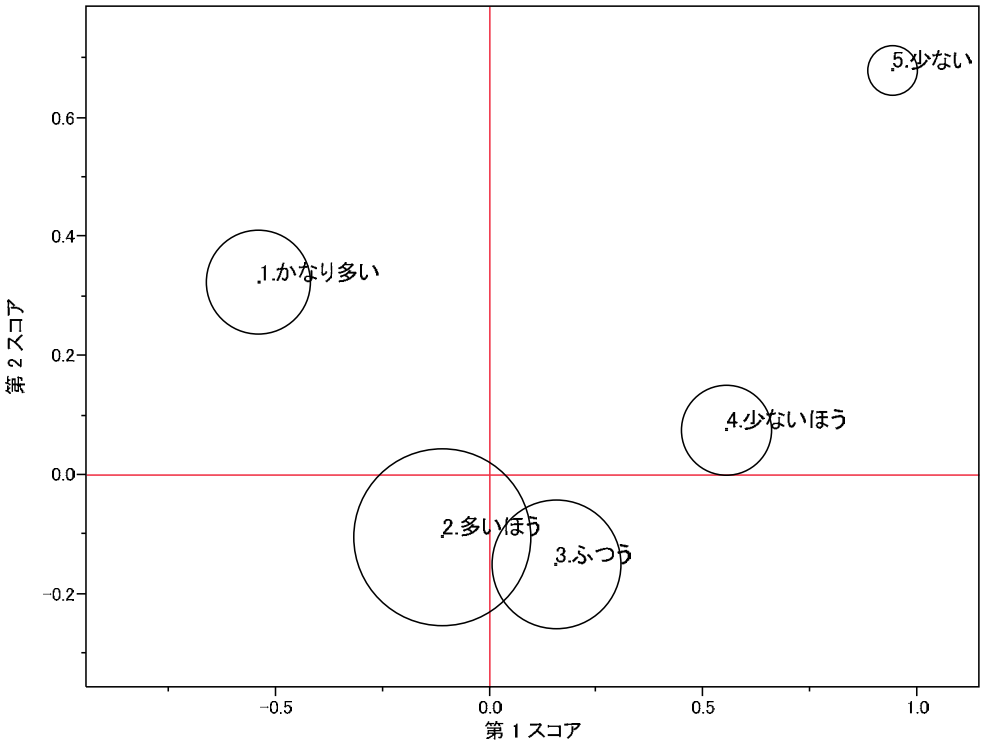


図 5－2　列（質問 J ）の成分スコア布置図

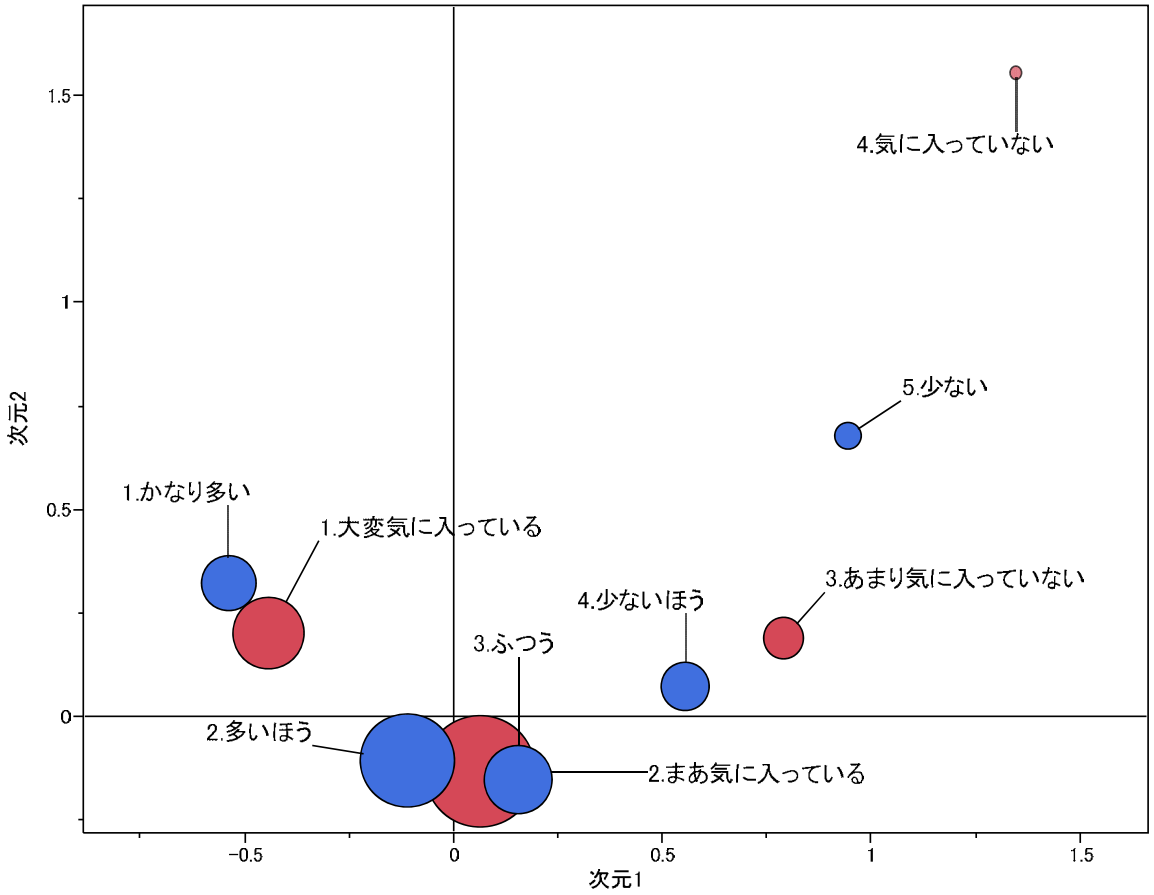


図 6　行と列の第1，第2成分スコアの同時布置図

以上が，対応分析法が提供する基本的な情報と，それの読み方，解釈である．考え方が明解でかつ実用的な手法であることが分かるであろう．

3. 2. 3.　分析対象とするデータ表（データ表をどう考えるか）

ここまで，対応分析法は **2 元データ表**であれば，比較的緩やかな条件で適用できると説明してきた．ここでは（意識的に）文字型データつまり**テキスト型データ**（textual data）の利用を念頭に，この 2 元データ表の特徴と適用範囲を考えてみよう．再度，対応分析法で扱う 2 元データ表の要件を列記すると，いくつかの種類に分けられる．

① 原則として“クロス表型”で表記される場合
② 二値の応答型データ（「yes」「no」型，0－1 型）である場合
③ 複数の要約表（集計表）などから再編集でえられた 2 元表形式のデータ
④ 各種の統計資料・統計情報から再編集して得られる 2 元データ表

ここで言う **2 元データ表**（two way data table）の特徴は，

- データ表の各要素（各セル内の値）が非負の数値であること
- 行または列の“**プロファイル**[29]が意味のある”データ（プロファイルについては後述）
- つまり，データ表の行または列の“**比率パターン**が意味を持つ”データ表

であればよい，ということである．これを満たせば，“ほぼ”どのようなデータ表でも利用できるということである．たとえばこれに含まれるデータ表として以下がある．

- 通常の 2 元クロス表あるいは分割表
- （0，1）型データ行列（2 元クロス表の特別な場合と考えられる；すでにみた例や，次の例 1，例 2 など），いわゆる**インシデント行列**（incident matrix）など．
- 多重クロス表（バート表）[30]
- いわゆる**インジケータ行列**（indicator matirix）あるいは**アイテム･カテゴリー型**
- 複数のクロス表を並置してえられる 2 元データ表
- そのほか，多くの統計表（数値が非負の集約データで，上の条件を満たすとき）

とくにここで，2 元クロス表，（0，1）型データ行列，アイテム･カテゴリー型データ，多重クロス表といったデータ（表）間には，密接な関係があって，じつはどれを考えるにも数理的には（ほぼ）同じように考えられることが知られている（後述する）．

換言すると，データ収集時の状況，あるいは事前のデータ取得計画（調査票設計など），既存のデータの再利用（2 次分析）など，状況に応じてデータ表のしかるべき形が決められても，事後の分析の自由度がかなりある（さまざまな 2 元データ表に加工が可能）ということである．これは簡単な例を見ることが理解を容易にするだろう．

例 1：2 値型応答データ

数量化法 III 類の説明で必ず登場するデータ表形式である．たとえば，つぎの表 15 のデータ表では，4 名のサンプル[31]（回答者）に 3 つの銘柄のどれが好きかを尋ね「好きな銘柄には 1」を，そうではない場合は 0 を選ぶといった場合を想定した「（サンプル）×（項目）」型の人工データ例である[32]．これはクロス表の特別な場合と考えられる（セル内度数が 1 か 0 のみ）．こうした 2 値応答の行列を**“インシデント行列”**（incident matrix）ということがある．

---

29 “プロフィル”の意味や性質については，うしろであらためて述べる．
30 多重クロス表と「多元クロス表」とは別のこと，両者は異なることに注意．多重クロス表の対応分析を多重対応分析という．
31 ここで「サンプル」としたが，正確には標本要素（調査対象者）のうち，回答してくれた人のこと（回答者）．あまり適切な言い方ではない（いずれ語句を調整する予定）．
32 対応分析法の数理を要約した「第Ⅱ部 対応分析法の基本的な考え方」で，これを例として説明している．

表 15　二値データ表の例

サンプル	銘柄 A	銘柄 B	銘柄 C
サンプル１	1	0	1
サンプル２	0	1	0
サンプル３	1	0	0
サンプル４	0	1	1

この 2 値型データ表は，**文字情報（テキスト型データ）**を用いて次のようなデータ表に書き替えても**情報の内容には変わりはない**．むしろ，調査などで回答を集めた時点では，表 3 の形式のほうが一般的である．かつては，これをソフトで集計処理するために，わざわざ“コーディング”して表 15 とするなどに変換することを行ってきた（必要であった）のだが，そういう必要もない．自由回答データのように，単語・語句の長さも，文字数も，前もってわからない，非構造的なデータなどを自由に扱えるコンピュータ環境が当たり前のことになっている．

表 16　表2を文字情報で表現

サンプル	サンプルが選んだ銘柄
サンプル１	銘柄 A，銘柄 C
サンプル２	銘柄 B
サンプル３	銘柄 A
サンプル４	銘柄 B，銘柄 C

このように**テキスト型データ**として書き方を変えてみると，これは**質的データ**であるということに他ならない．換言するとここでサンプルや銘柄は質的データであって，基本的にはこのままでは計量的な統計処理は難しいことを示している[33]．

例 2：好みの清涼飲料水の選択

表 17 は 30 名の調査対象者が 8 種の清涼飲料水のどれを「好む」か，その選んだ結果のデータ表である（ある論文[34]のデータ表を若干再編集した）．ここでは「好む＝1」，それを選ばなかったときを「0」と対応させてある[35]．サンプルの誰がどのような清涼飲料水を選ぶのか，飲料水の相互の類似・関連に関心があるといった場面を想定した例である．

このデータ表も，前の例にならうと表 18 のようにテキスト型データを用いて書き替えることができる．実際，大抵のソフトではこの表 18 の形式のデータ表を扱うことができる（例：WordMiner，JMP）．

これは，あらかじめ用意した銘柄を質問文の選択肢から選ばせるのではなく“自由記述”として「あなたの好きな飲み物（の商品名）を列記してください」「あなたの好きなハンドバッグのブランド名をいくつでも列記してください」などの質問を設ける場面に読み替えてみるとよい．

ところでここで，回答者番号が「2，3，5，21，22」「12，27」などは，同じ回答パターンであることがわかる．こういう場合，これらを併合して（頻度を積み上げて），つまり同じ回答行あるいは同じ回答列を併合して得られるデータ表の対応分析とこのデータ表の対応分析とは同等の結果となる（いわゆる**“分布の同等性”**の性質がある[36]）．

33　統計ソフトウェアによっては，たとえば JMP では，「カテゴリカル」プラットフォームを利用すると，カンマ区切りのデータから，このインシデント行列を作成できる．

34　D.L. Hoffman and G.R. Franke (1986): Correspondence Analysis: Grafical Representation of Categorical Data in Marketing Research, *Journal of Marketing Research*, Vol.XXIII, p213-227.

35　ここは形式的に「0，1」を対応させたが，じつは「好む＝1」としたことに対応させて「0」としてよいかは議論の余地がある．「好む」の余事象を形式的に「0」としただけで，これは「好まない」を意味してはいない．

36　“分布の同等性”（equivalence of distribution あるいは distributional equivalence）は，対応分析法の重要な性質の 1 つである．これについては，資料「第Ⅱ部」で説明している．

表 17　好きな清涼飲料水

ｻﾝﾌﾟﾙ番号	ｺｶｺｰﾗ	ﾀﾞｲｴｯﾄｺｰｸ	ﾀﾞｲｴｯﾄﾍﾟﾌﾟｼ	ﾀﾞｲｴｯﾄ 7 ｱｯﾌﾟ	ﾍﾟﾌﾟｼ	ｽﾌﾟﾗｲﾄ	Tab	7ｱｯﾌﾟ	ｻﾝﾌﾟﾙ番号	ｺｶｺｰﾗ	ﾀﾞｲｴｯﾄｺｰｸ	ﾀﾞｲｴｯﾄﾍﾟﾌﾟｼ	ﾀﾞｲｴｯﾄ 7 ｱｯﾌﾟ	ﾍﾟﾌﾟｼ	ｽﾌﾟﾗｲﾄ	Tab	7ｱｯﾌﾟ
1	1	0	0	0	1	1	0	1	16	0	0	0	0	1	1	0	0
2	1	0	0	0	1	0	0	0	17	0	1	0	0	0	1	0	0
3	1	0	0	0	1	0	0	0	18	1	1	0	0	1	0	0	0
4	0	1	0	1	0	0	1	0	19	1	0	0	0	0	0	0	1
5	1	0	0	0	1	0	0	0	20	1	1	1	0	1	0	0	0
6	1	0	0	0	1	1	0	0	21	1	0	0	0	1	0	0	0
7	0	1	1	1	0	0	1	0	22	1	0	0	0	1	0	0	0
8	1	1	0	0	1	1	0	1	23	0	1	0	1	0	0	1	0
9	1	1	0	0	0	1	1	1	24	1	1	0	0	1	0	0	0
10	1	0	0	0	1	0	0	1	25	0	1	1	1	0	0	0	0
11	1	0	0	0	1	1	0	0	26	0	1	0	1	0	0	1	0
12	0	1	0	0	0	0	1	0	27	0	1	0	0	0	0	1	0
13	0	0	1	1	0	1	0	1	28	1	0	0	0	0	1	0	1
14	1	0	0	0	0	1	0	0	29	1	0	0	0	0	1	0	0
15	0	1	1	0	0	0	1	0	30	0	1	1	0	0	0	1	0

表 18　好きな清涼飲料水の文字表現（表 17 を文字情報で表現したとき）

回答者番号	回答者が選んだ「好む」清涼飲料	回答者番号	回答者が選んだ「好む」清涼飲料
1	ｺｶｺｰﾗ, ﾍﾟﾌﾟｼｺｰﾗ, ｽﾌﾟﾗｲﾄ, 7 ｱｯﾌﾟ	16	ﾍﾟﾌﾟｼｺｰﾗ, ｽﾌﾟﾗｲﾄ
2	ｺｶｺｰﾗ, ﾍﾟﾌﾟｼｺｰﾗ	17	ﾀﾞｲｴｯﾄｺｰｸ, ｽﾌﾟﾗｲﾄ
3	ｺｶｺｰﾗ, ﾍﾟﾌﾟｼｺｰﾗ	18	ｺｶｺｰﾗ, ﾀﾞｲｴｯﾄｺｰｸ, ﾍﾟﾌﾟｼｺｰﾗ
4	ﾀﾞｲｴｯﾄｺｰｸ, ﾀﾞｲｴｯﾄ 7 ｱｯﾌﾟ, Tab	19	ｺｶｺｰﾗ, 7 ｱｯﾌﾟ
5	ｺｶｺｰﾗ, ﾍﾟﾌﾟｼｺｰﾗ	20	ｺｶｺｰﾗ, ﾀﾞｲｴｯﾄｺｰｸ, ﾀﾞｲｴｯﾄﾍﾟﾌﾟｼ, ﾍﾟﾌﾟｼｺｰﾗ
6	ｺｶｺｰﾗ, ﾍﾟﾌﾟｼｺｰﾗ, ｽﾌﾟﾗｲﾄ	21	ｺｶｺｰﾗ, ﾍﾟﾌﾟｼｺｰﾗ
7	ﾀﾞｲｴｯﾄｺｰｸ , ﾀﾞｲｴｯﾄﾍﾟﾌﾟｼ, ﾀﾞｲｴｯﾄ 7 ｱｯﾌﾟ, Tab	22	ｺｶｺｰﾗ, ﾍﾟﾌﾟｼｺｰﾗ
8	ｺｶｺｰﾗ , ﾀﾞｲｴｯﾄｺｰｸ , ﾍﾟﾌﾟｼｺｰﾗ, ｽﾌﾟﾗｲﾄ, 7 ｱｯﾌﾟ	23	ﾀﾞｲｴｯﾄｺｰｸ, ﾀﾞｲｴｯﾄ 7 ｱｯﾌﾟ, Tab
9	ｺｶｺｰﾗ , ﾀﾞｲｴｯﾄｺｰｸ , ｽﾌﾟﾗｲﾄ , Tab, 7 ｱｯﾌﾟ	24	ｺｶｺｰﾗ, ﾀﾞｲｴｯﾄｺｰｸ, ﾍﾟﾌﾟｼｺｰﾗ
10	ｺｶｺｰﾗ, ﾍﾟﾌﾟｼｺｰﾗ, 7 ｱｯﾌﾟ	25	ﾀﾞｲｴｯﾄｺｰｸ, ﾀﾞｲｴｯﾄﾍﾟﾌﾟｼ, ﾀﾞｲｴｯﾄ 7 ｱｯﾌﾟ
11	ｺｶｺｰﾗ, ﾍﾟﾌﾟｼｺｰﾗ, ｽﾌﾟﾗｲﾄ	26	ﾀﾞｲｴｯﾄｺｰｸ, ﾀﾞｲｴｯﾄ 7 ｱｯﾌﾟ, Tab
12	ﾀﾞｲｴｯﾄｺｰｸ, Tab	27	ﾀﾞｲｴｯﾄｺｰｸ, Tab
13	ﾀﾞｲｴｯﾄﾍﾟﾌﾟｼ, ﾀﾞｲｴｯﾄ 7 ｱｯﾌﾟ, ｽﾌﾟﾗｲﾄ, 7 ｱｯﾌﾟ	28	ｺｶｺｰﾗ, ｽﾌﾟﾗｲﾄ, 7 ｱｯﾌﾟ
14	ｺｶｺｰﾗ, ｽﾌﾟﾗｲﾄ	29	ｺｶｺｰﾗ, ｽﾌﾟﾗｲﾄ
15	ﾀﾞｲｴｯﾄｺｰｸ, ﾀﾞｲｴｯﾄﾍﾟﾌﾟｼ, Tab	30	ﾀﾞｲｴｯﾄｺｰｸ, ﾀﾞｲｴｯﾄﾍﾟﾌﾟｼ, Tab

例 3：ある調査データの集計表の相互の関係

ここで調査データの別の例をみよう．ある自治体で行った市民意識調査の例である．質問は「あなたは今の生活環境の中で日頃どのような過ごし方をしていますか．次の質問のどれか一つに○をつけてください．」としていくつか挙げた項目のうちから，次の 2 つを選んだ．

質問 I ：昔からの習慣をよく守っているか．
1．守っている　2．まあ守っている　3．あまり守っていない　4．守っていない

質問 J ：神社や、お寺詣りをよくするか．
1．お寺詣りをよくする　2．たまにお寺詣りをする　3．あまりお寺詣りをしない
4．お寺詣りをしない

ここで，各表の関係を説明しよう．

- まず，表 20 は調査で得たデータセットを電子化したファイル，つまり原始データセットである．
- ここで上の 2 つの質問文を指定しクロス表 $\mathbf{F}=\left(f_{ij}\right)$ を作る（表 19）．
- 表 21 は表 20 から，上の 2 つの質問文を指定し切り出した表である．通常は，この表で

集計分析などを行う．

- かりにこれを**コーディング**すると，表 22 が得られる．これを「**C**表」とする．
- この表を，**インジケータ行列**（または**完備排反型行列**[39]）の **A** 表に展開すると表 23 の右側の表となる．ここでは，「応答あり（1）」「応答なし（0）」の（0，1）型データ表となっていることに注意しよう．また，この場合，行和（サンプル別の 1 の数）は，2 つの“選択肢数の和”となることにも注意しよう．ここでは，無回答も含めて「10」となる[40]．
- さらに，アイテム･カテゴリー型データ表（インジケータ行列） **A** を転置して $\mathbf{A}^t$ とし，あらたな行列 $\mathbf{B}=\mathbf{A}^t\mathbf{A}$ を作ると表 24 となる．これを（2 つの質問から作る）**多重クロス表**あるいは**バート表（バート行列）**という[41]．多重クロス表は質問項目が多数あっても作れる．また，このバート表の対応分析法（多重クロス表の対応分析：MCA）を行うことは，さまざまな利点がある．

表 19 （質問 I × 質問 J ）のクロス表 $\mathbf{F}=\left(f_{ij}\right)$ の生成（度数のみ表示）

		質問 J				
	選択肢	お寺詣りをよくする	たまにお寺詣りをする	あまりお寺詣りをしない	お寺詣りをしない	無回答
質問 I	守っている	41	26	22	15	2
	まあ守っている	25	67	45	30	0
	あまり守っていない	6	13	34	31	0
	守っていない	1	6	7	27	0
	無回答	1	4	1	2	7

表 20　市民意識調査データの一（多変量構造のデータ表）

回収サンプル番号	地域コード	計画サンプル番号	この公園構想を知っていましたか	この公園構想を知っていましたか。(選択肢)	この公園構想を知っていましたか。(選択肢)	1_近くの緑地や公園等をよく散策している。(選択肢)	3_昔からの習慣をよく守っている。(選択肢)	6_神社や、お寺詣りをよくする。(選択肢)	8_自分のなすべき役割は積極的に果している。(選択肢)
1	11	181	1	はい	知っていた	あまり散策しない	まあ守っている	お寺詣りをしない	役割はあまり果たしていない
2	11	185	2	いいえ	知らなかった	あまり散策しない	あまり守っていない	たまにお寺詣りをする	まあ役割は果たしている
3	11	188	1	はい	知っていた	散策しない	守っている	あまりお寺詣りをしない	役割はあまり果たしていない
4	11	189	2	いいえ	知らなかった	散策しない	まあ守っている	お寺詣りをしない	まあ役割は果たしている
5	11	198	2	いいえ	知らなかった	あまり散策しない	まあ守っている	たまにお寺詣りをする	まあ役割は果たしている
6	12	199	2	いいえ	知らなかった	あまり散策しない	まあ守っている	たまにお寺詣りをする	まあ役割は果たしている
7	12	204	2	いいえ	知らなかった	あまり散策しない	まあ守っている	あまりお寺詣りをしない	役割は果たしている
8	12	205	2	いいえ	知らなかった	まあ散策している	あまり守っていない	お寺詣りをしない	役割は果たしている
9	12	207	1	はい	知っていた	まあ散策している	まあ守っている	たまにお寺詣りをする	まあ役割は果たしている
10	12	209	1	はい	知っていた	よく散策している	まあ守っている	たまにお寺詣りをする	まあ役割は果たしている
11	12	211	2	いいえ	知らなかった	散策しない	無回答	お寺詣りをしない	まあ役割は果たしている
12	12	212	2	いいえ	知らなかった	あまり散策しない	まあ守っている	あまりお寺詣りをしない	役割は果たしている
13	12	215	2	いいえ	知らなかった	散策しない	まあ守っている	お寺詣りをよくする	役割は果たしている
14	13	217	2	いいえ	知らなかった	まあ散策している	守っている	お寺詣りをよくする	まあ役割は果たしている
15	13	221	2	いいえ	知らなかった	あまり散策しない	まあ守っている	あまりお寺詣りをしない	役割は果たしている
16	13	223	1	はい	知っていた	まあ散策している	守っていない	あまりお寺詣りをしない	まあ役割は果たしている
17	13	227	1	はい	知っていた	あまり散策しない	まあ守っている	お寺詣りをしない	まあ役割は果たしている
18	13	228	2	いいえ	知らなかった	散策しない	守っている	お寺詣りをしない	まあ役割は果たしている
19	13	229	2	いいえ	知らなかった	まあ散策している	まあ守っている	あまりお寺詣りをしない	役割は果たしている
20	14	231	2	いいえ	知らなかった	散策しない	まあ守っている	あまりお寺詣りをしない	まあ役割は果たしている
21	13	232	1	はい	知っていた	よく散策している	守っていない	あまりお寺詣りをしない	まあ役割は果たしている
22	14	238	2	いいえ	知らなかった	あまり散策しない	まあ守っている	たまにお寺詣りをする	役割は果たしている
23	14	239	2	いいえ	知らなかった	散策しない	まあ守っている	お寺詣りをしない	役割は果たしている
24	14	242	1	はい	知っていた	無回答	守っている	お寺詣りをよくする	役割は果たしている
25	14	244	2	いいえ	知らなかった	まあ散策している	まあ守っている	あまりお寺詣りをしない	役割は果たしている
26	14	248	2	いいえ	知らなかった	あまり散策しない	まあ守っている	あまりお寺詣りをしない	まあ役割は果たしている
27	14	251	2	いいえ	知らなかった	まあ散策している	あまり守っていない	あまりお寺詣りをしない	役割はあまり果たしていない
28	15	253	1	はい	知っていた	散策しない	まあ守っている	たまにお寺詣りをする	役割は果たしている
29	15	254	2	いいえ	知らなかった	散策しない	守っていない	お寺詣りをしない	役割は果たしていない

[39]対応分析法では、これを完備排反型（complete disjunctive form；forme disjunctive complète）ともいう．数量化法 III 類などでは，**アイテム･カテゴリー型**という．本稿では，主にインジケータ行列とする．

[40] ここで，かりに無回答や DK があっても，選択肢数を増やせばよい，表 23 は無回答を含む例となっている．

[41] バート表（バート行列）とは，提唱者の C. Burt の名前を与えたものである．

表 21　表 20 から切り出した2つの質問

3_昔からの習慣をよく守っている。(選択肢)	6_神社や、お寺請りをよくする。(選択肢)
まあ守っている	お寺請りをしない
あまり守っていない	たまにお寺請りをする
守っている	あまりお寺請りをしない
まあ守っている	お寺請りをしない
まあ守っている	たまにお寺請りをする
まあ守っている	たまにお寺請りをする
まあ守っている	あまりお寺請りをしない
あまり守っていない	お寺請りをしない
まあ守っている	たまにお寺請りをする
まあ守っている	たまにお寺請りをする
無回答	お寺請りをしない
まあ守っている	あまりお寺請りをしない
まあ守っている	お寺請りをよくする
守っている	お寺請りをよくする
まあ守っている	あまりお寺請りをしない
守っていない	あまりお寺請りをしない
まあ守っている	お寺請りをしない
守っている	お寺請りをしない
まあ守っている	あまりお寺請りをしない
まあ守っている	あまりお寺請りをしない
守っていない	あまりお寺請りをしない
まあ守っている	たまにお寺請りをする
まあ守っている	お寺請りをしない
守っている	お寺請りをよくする
まあ守っている	あまりお寺請りをしない
まあ守っている	あまりお寺請りをしない
あまり守っていない	あまりお寺請りをしない
まあ守っている	たまにお寺請りをする
守っていない	お寺請りをしない

⇒

(左のデータ表を右のようにコーディング)

(注) 多くのソフトでは左側の表から処理可能

表 22　数値コードへの変換(C 表)

3_昔からの習慣をよく守っている。	6_神社や、お寺請りをよくする。
2	4
3	2
1	3
2	4
2	2
2	2
2	3
3	4
2	2
2	2
5	4
2	3
2	1
1	1
2	3
4	3
2	4
1	4
2	3
2	3
4	3
2	2
2	4
1	1
2	3
2	3
3	3
2	2
4	4

表 23　コード変換したデータ表(C 表)をアイテム・カテゴリー型データ表(A 表)に変換

3_昔からの習慣をよく守っている。	6_神社や、お寺請りをよくする。
2	4
3	2
1	3
2	4
2	2
2	2
2	3
3	4
2	2
2	2
5	4
2	3
2	1
1	1
2	3
4	3
2	4
1	4
2	3
2	3
4	3
2	2
2	4
1	1
2	3
2	3
3	3
2	2
4	4

⇒

1.守っている	2.まあ守っている	3.あまり守っていない	4.守っていない	5.無回答	1.お寺請りをよくする	2.たまにお寺請りをする	3.あまりお寺請りをしない	4.お寺請りをしない	5.無回答
	1							1	
		1				1			
1							1		
	1							1	
	1					1			
	1					1			
	1						1		
		1						1	
	1					1			
	1					1			
				1				1	
	1						1		
	1				1				
1							1		
	1						1		
			1				1		
	1							1	
1								1	
	1						1		
	1						1		
			1				1		
	1					1			
					1				
	1						1		
	1						1		
		1					1		
	1					1			
			1					1	

(注) 調査データの一部を切り取ったので，たまたま無回答に空欄がある．

表 24　多重クロス表（バート表）$\mathbf{B}=\mathbf{A}^t\mathbf{A}$ の例（表 23 から生成）

	質問	質問 I					質問 J				
質問	選択肢	守っている	まあ守っている	あまり守っていない	守っていない	無回答	お寺請りをよくする	たまにお寺詣りをする	あまりお寺詣りをしない	お寺詣りをしない	無回答
質問 I	守っている	106	0	0	0	0	41	26	22	15	2
	まあ守っている	0	167	0	0	0	25	67	45	30	0
	あまり守っていない	0	0	84	0	0	6	13	34	31	0
	守っていない	0	0	0	41	0	1	6	7	27	0
	無回答	0	0	0	0	15	1	4	1	2	7
質問 J	お寺詣りをよくする	41	25	6	1	1	74	0	0	0	0
	たまにお寺詣りをする	26	67	13	6	4	0	116	0	0	0
	あまりお寺詣りをしない	22	45	34	7	1	0	0	109	0	0
	お寺詣りをしない	15	30	31	27	2	0	0	0	105	0
	無回答	2	0	0	0	7	0	0	0	0	9

表 25　上のバート表の説明

	質問 I	質問 J
質問 I	（質問 I）×（質問 I）のクロス表 つまり質問 I の周辺度数が対角要素に入った対角行列	（質問 I）×（質問 J）のクロス表
質問 J	（質問 J）×（質問 I）のクロス表	（質問 J）×（質問 J）のクロス表 つまり質問 J の周辺度数が対角要素に入った対角行列

このようにデータ表を作ったとき，各表の間には重要な関係がある．対応分析法の数理的考察から，各表から出発した解析結果の間には，ある同等性や関連性があることが知られている．以下に主な関係を抜粋しよう[42]．

① 表 23 の**インジケータ行列 $\mathbf{A}$** の対応分析の結果は，じつは表 19（クロス表 $\mathbf{F}=\left(f_{ij}\right)$）の対応分析の結果に同等である．

② またそれらは「表 24 の**バート表 $\mathbf{B}=\mathbf{A}^t\mathbf{A}$** の対応分析の結果にも同等」となる（注：表 23 で得られる固有値を λ_k^A とし，バート表のそれを λ_k^B とすると，$\lambda_k^B=\left(\lambda_k^A\right)^2$ の関係がある）．

③ 質問項目が I,J の 2 つの場合，そのクロス表 $\mathbf{F}=\left(f_{ij}\right)$ に対する対応分析の固有値 λ_k^F と，バート表を 2 元データ表とみなしてこれに対応分析を適用して得た固有値 λ_k^B には，対応関係がある．上の関係と合わせて記すと，$\lambda_k^B=\left(\lambda_k^A\right)^2=\left(\dfrac{1\pm\sqrt{\lambda_k^F}}{2}\right)^2$ の関係がある．つまり，クロス表 $\mathbf{F}=\left(f_{ij}\right)$ と，表 24 の 2 項目のバート表との結果は解析的には同等である

④ 表 24 のバート表の分析を行うと，ここでは各項目の選択肢に付与される成分スコアを使って，表 23 のアイテム・カテゴリー型の回答側（サンプル側）の成分スコアを求めることができる．ここで，いわゆる“**追加処理**”（supplementary treatment）を行う[43]．

これらの関係はもちろん数理的に証明されている．ここで重要なことは，解析したいデー

42　さらに詳しい説明は「第Ⅱ部」に記した．

43　「追加処理」は，フランス流のデータ解析における典型的な手順の一例．「第Ⅱ部」に説明．

タ表を，目的に応じて使いわける自由度があるという点である．

例 4：クロス表，多重クロス表，インジケータ行列のデータ表の関係

ここでまとめとして，もう 1 つのデータ例で確かめよう[44]．ここでは 2 つの質問 I と質問 J について選択肢が以下のようであるとする．つまり，このデータ表の分析目的は，回答者があるレストランを選ぶときに，どのような評価基準で選ぶだろうか，その関連を調べたいという課題である．

質問 I: 次にあげるレストランのうち，あなたがお気に入りのレストランはどれですか．
（1 つ選ぶ）

1. いりふね　2. かりや　3. きくみ　4. さとみ　5. クラーク
6. コルシカ　7. バッハ　8. ムガール　9. ラ・マレ　10. ロゴスキー

質問 J: そのレストランを選択したときの評価基準は，次の 3 つのうちのどれでしょうか．
（1 つ選ぶ）

1. 工夫・サービス　2. 味　3. 量

この 2 つの質問に対して，回答者がそれぞれ 1 つだけ選択肢を選ぶものとする．このとき，N 人（＝1,284 人）の回答者からの回答は，表 26 のように寸法が（N 人）×（2 項目）のデータ表として集められる．なおここでは DK（Don’t Know :わからない）や NA（No Answer:無回答）などはなかったものとする（あってもよいが，説明を簡略にするために省く）．

表 26 （回答者）×（項目）のデータ表

項目 / 回答者	I（レストラン）	J（評価基準）
1	バッハ	味
2	ムガール	量
3	さとみ	量
4	ラ・マレ	工夫・サービス
5	きくみ	味
：	：	：
：	：	：
N	いりふね	量
N=1, 284（回答者数）		

表 27　項目 I（レストラン）×項目 J（評価基準）の 2 元クロス表 $\mathbf{F}=\left(f_{ij}\right)$

項目 I / 項目 J	工夫・サービス	味	量	行和
いりふね	98	25	32	155
かりや	105	35	38	178
きくみ	35	8	67	110
さとみ	42	46	7	95
クラーク	34	14	54	102
コルシカ	32	77	13	122
バッハ	48	76	18	142
ムガール	49	44	16	109
ラ・マレ	49	82	15	146
ロゴスキー	48	35	42	125
列和	540	442	302	1,284（＝N）

44 これは架空のトイ・データである．これを用いて，対応分析法の性質など，うしろでさらに調べる．

ここで得た各データ表の関係は，例 3 と同様に解釈すればよい．なお，例 3，例 4 では選んだ項目を 2 項目としたが，これを一般に多数の項目としても類似の関係が成り立つことが分かっている[45]．ここでは，以下を確認する．

① （回答・サンプル）×（多変量の項目）のデータ表（表 26）
② 2 項目のクロス表（表 27）
③ 多変量の項目について加工生成したアイテム・カテゴリー型（インジケータ行列）のデータ表 **A**（表 17）
④ アイテム・カテゴリー型データ表を転置した行列と元のアイテム・カテゴリー型データ表の積から多重クロス表（バート表）$\mathbf{B} = \mathbf{A}^t\mathbf{A}$ を生成（表 29）
⑤ この多重クロス表の非対角部のブロック行列として表 27 のクロス表が得られる．

表 26 のような（回答・サンプル）×（多変量の項目）のデータ表の場合，林の数量化法では通常はアイテム・カテゴリー型データ表から出発する．一方，対応分析法では，上にみたようにさまざまなタイプのデータ表を用い，またそれら相互の数理的な関係が考察されており，これらを使い分けるという点に特徴がある[46]．

表 28 （回答者）×（アイテム・カテゴリー）のデータ表，インジケータ行列（**A** 表）

項目 / 回答者	*I*							*J*		
	1	2	3	4	…	9	10	1	2	3
	いりふね	かりや	きくみ	さとみ		クラーク	コルシカ	バッハ	工夫・サービス	味	量
1	0	0	0	0	…	0	1	0	1	0
2	0	0	0	0	…	0	0	0	0	1
3	0	0	0	1	…	0	0	0	0	1
4	0	0	0	0	…	0	0	0	1	0
5	0	0	1	0	…	0	0	0	1	0
⋮	⋮	⋮	⋮	⋮	…	⋮	⋮	⋮	⋮	⋮
1,284	1	0	0	0	…	0	0	0	0	1

表 29 2 項目の多重クロス表（バート表）$\mathbf{B} = \mathbf{A}^t\mathbf{A}$ の生成

項目	いりふね	かりや	きくみ	さとみ	クラーク	コルシカ	バッハ	ムガール	ラ・マレ	ロゴスキー	工夫・サービス	味	量
いりふね	155										98	25	32
かりや		176									105	35	38
きくみ			110								35	8	67
さとみ				95							42	46	7
クラーク					102						34	14	54
コルシカ						122					32	77	13
バッハ							142				48	76	18
ムガール								109			49	44	16
ラ・マレ									146		49	82	15
ロゴスキー										125	48	35	42
工夫・サービス	98	105	35	42	34	32	48	49	49	48	540		
味	25	35	8	46	14	77	76	44	82	35		442	
量	32	38	67	7	54	13	18	16	15	42			302

（注）この表で空白のセル（対角ブロック行列の非対角要素）はすべてゼロである．

45 「第Ⅱ部」で述べている．
46 ここらについては，Benzécri（1980），Greenacre（1984），ほかを参照．

4. 対応分析法の考え方（基本）

4．1 数値例による確認

以上のように様々な表に対する対応分析が考えられるのだが，ここでは上記の例 4 で用いたクロス表に対する対応分析を説明する（表 27）．

ここで登場する多くの用語は，対応分析法に特有のものが多い．つまり，提唱者のベンゼクリの考え方が反映された用語句が多数登場する．このことが，国内における対応分析法の利用者の理解を妨げているようにもみえる．たとえば，プロファイルとストレッチ・プロファイル，質量，重心と平均ベクトル，三角座標と重心座標系，成分スコア[47]などである．これらを用いて対応分析法について順をおって説明する．

① プロファイルと重心座標系

まず，プロファイルを作ってみる．**“プロファイル”**の考え方は，対応分析法を理解するうえで重要である．これを理解するには例を作るのが早い．このレストラン評価のデータ表では寸法は m=10, n=3 である．プロファイルとは行または列の比率（相対確率）の分布のことであり，それぞれを**“行プロファイル”, “列プロファイル”**という（表 30, 31）．これに相当する 2 種のプロファイル $q_{ij}, q_{ij}^{*}\left(i \in I, j \in J\right)$ を実際に作ってみる．

表 30 行プロファイル $\mathbf{N}_I = \mathbf{P}_I^{-1}\mathbf{P}_{IJ} = \left\{ q_{ij} = \dfrac{p_{ij}}{p_{i+}} \middle| i \in I, j \in J \right\}$ と行の重心（列の質量）

評価項目 / レストラン	工夫 サービス	味	量
いりふね	0.632	0.161	0.206
かりや	0.590	0.197	0.213
きくみ	0.318	0.073	0.609
さとみ	0.442	0.484	0.074
クラーク	0.333	0.137	0.529
コルシカ	0.262	0.631	0.107
バッハ	0.338	0.535	0.127
ムガール	0.450	0.404	0.147
ラ・マレ	0.336	0.562	0.103
ロゴスキー	0.384	0.280	0.336
p_{+j}	0.421	0.344	0.235
$\mathbf{c}$	$\mathbf{c} = (0.421, 0.344, 0.235)^t$		

まず，行プロファイルから考えてみる．**行プロファイル** $\mathbf{N}_I$ は q_{ij} を要素とする行列で，$\sum_{j=1}^{n} q_{ij} = 1$（行和＝１）となっている（この例では $\sum_{j=1}^{3} q_{ij} = 1$ つまり $q_{i1} + q_{i2} + q_{i3} = 1$）．これを，（$n$−1）＝3−1＝2，つまり 2 次元の空間内に布置する 10（個）のレストランと考える．また**“列の平均ベクトル”**（平均比率）を $\mathbf{c}$ で表すが，これは**“行プロファイルの重心”**に相当する（うしろの図 7 の中の G，図 8 の「＋重心」がそれに相当）．

47 ここでは成分スコアとしたが，座標（coordinates），スコア（scores），数量化得点などの呼称がある．

表 31　列プロファイル $\mathbf{N}_J = \mathbf{P}_J^{-1}\mathbf{P}_{JI} = \left\{ q_{ij}^* = \dfrac{p_{ij}}{p_{+j}} \middle| i \in I, j \in J \right\}$ と列の重心（行の質量）

評価項目 / レストラン	工夫 サービス	味	量	p_{i+}	$\mathbf{r}$
いりふね	0.181	0.057	0.106	0.121	$\mathbf{r} = \begin{pmatrix} 0.121 \\ 0.139 \\ 0.086 \\ 0.074 \\ 0.079 \\ 0.095 \\ 0.111 \\ 0.085 \\ 0.114 \\ 0.097 \end{pmatrix}$
かりや	0.194	0.079	0.126	0.137	
きくみ	0.065	0.018	0.222	0.086	
さとみ	0.078	0.104	0.023	0.074	
クラーク	0.063	0.032	0.179	0.079	
コルシカ	0.059	0.174	0.043	0.095	
バッハ	0.089	0.172	0.06	0.111	
ムガール	0.091	0.1	0.053	0.085	
ラ・マレ	0.091	0.186	0.05	0.114	
ロゴスキー	0.089	0.079	0.139	0.097	

つぎに**列プロファイル**を考える（表 31）．このとき，対応分析法の特徴として，行側とまったく同じように（対称に）定式化することがある．つまり，列プロファイル $\mathbf{N}_J$ は q_{ij}^* を要素とする行列で，$\sum_{i=1}^{m} q_{ij}^* = 1$（列和＝１）となる．これを，$(m-1) = 10-1 = 9$，つまり 9 次元の空間内に布置する 3 つの「評価基準」と考える．ここでもまた，**“行の平均ベクトル”**（平均比率）$\mathbf{r}$ は **“列プロファイルの重心”** に相当する．

ここでは行の側から観察し，表 30 の行プロファイル $\mathbf{N}_I = \mathbf{P}_I^{-1}\mathbf{P}_{IJ} = \left\{ q_{ij} = \dfrac{p_{ij}}{p_{i+}} \middle| i \in I, j \in J \right\}$ を図 7 に合わせて描いてみよう．ここでは，空間 $R^n = R^3$（3 次元空間）内で $\sum_{j=1}^{3} p_{+j} = 1$ の制約があるから（$n-1$）=3－1=2 次元空間内の点として，行プロファイルは表される．（点が自由に動ける次元が 2 次元であることを「自由度が 2 である」ともいう）．これが図 7 の上の図であり，図でアミをかけた部分にある行プロファイルの座標系を **“重心座標系”**（barycentric coordinate system）という．ここではこれが 2 次元平面となっている．この平面上に分布する 10 の「レストラン」の行プロファイルは図に表せるのでこれを実際に描いてみると図 8 のようになる（この三角図の作図は統計ソフト JMP を利用）．この場合を **“三角座標系”**（triangular coordinate system）という．こうした座標系内での行あるいは列のプロファイルの分布を，**“雲”**[48]と呼んでいる．

一方，列の側から観察すると，$(m-1) = 10-1 = 9$ 次元内の空間に布置する 3 つの「評価基準」分布を考えればよい．このようにプロファイルを行と列の双方向から同等に（対称に）考える（表 31）．またここでは，比率データあるいはそのような加工が意味あるデータとして扱っていることに注意しよう．

この例では，たまたま行プロファイルの分布を視認できるように 2 次元平面内の三角座標系として表される（そのようにデータを作った）．しかし列プロファイルは，より高次元の空間となるが (9 次元内の超平面と考えるので)，こうした場合は視覚的には観察できないが“重心座標系”内の分布を想像すればよい．

ではここで，対応分析法は何を行うのであろうか．たとえばここで，元の行プロファイルは 2 次元平面内に分布している 10 個の点（レストラン）であるから，これの直線上への射影（点の影の分布）を考え，その点の分布の**分散を最大化**することを考える．こうして得られる直線（軸）を **“主軸”**（principal axis）といいその上に分布する点の値を **“成分スコア”**（coordinates）とする．この行プロファイルと 2 つの主軸（第 1 成分，第 2 成分）と成分スコ

[48] この“雲”という用語もベンゼクリが用いており，フランス語の“nuage”（英語の cloud）にあたる．

アの関係を実際に示すと図 9 のようになる．

これを言い換えると，点と点とのあいだの距離を**“カイ 2 乗距離”**（Chi-square distance）で測り，それを低次元のユークリッド空間内に描画する（映す）という方法である．とくにこの例では，元の行プロファイルは 2 次元平面内で分布する点なので，2 つの成分スコア（を使えば，情報の損失なく，データの特徴のすべて（つまり元の 2 元データ表の全情報）を表すことができる[49]．

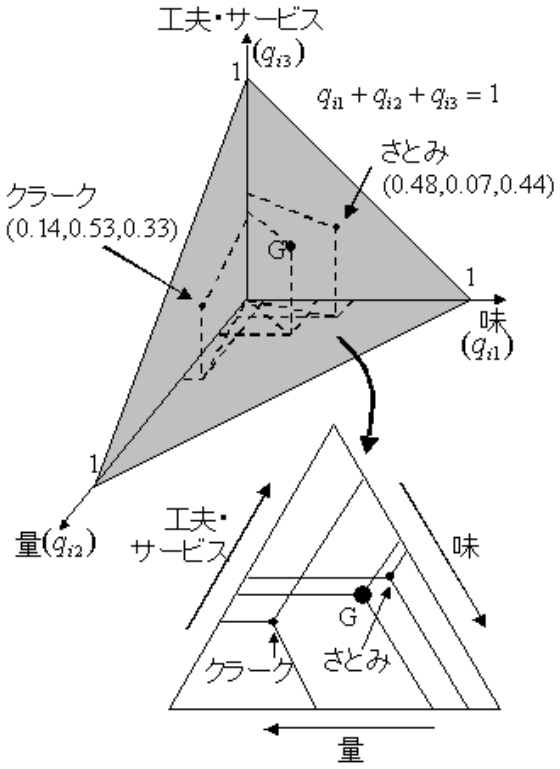


図 7　行「レストラン」の布置の考え方（プロファイルを重心座標系に射影）

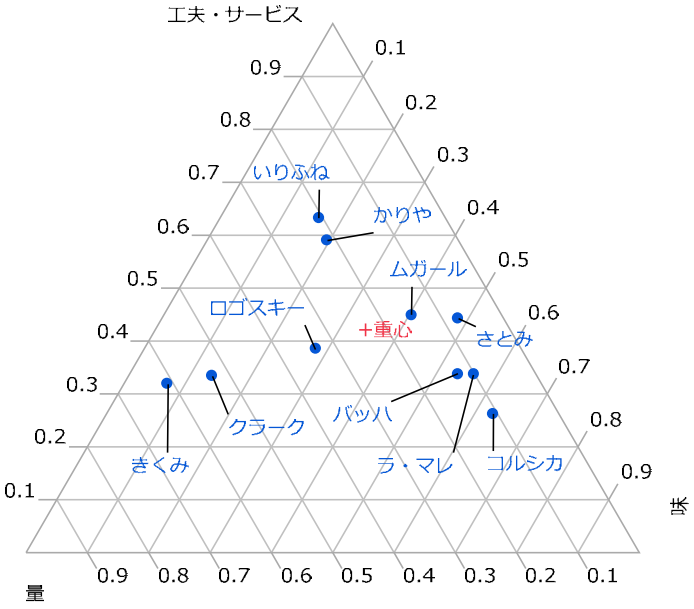


図 8　三角座標系（正三角形）上の行プロファイル（レストラン）の分布

[49] こういう状況を説明するために，このトイ・データを用意したということ．またこれについて，「第Ⅱ部」でも詳しく述べる．

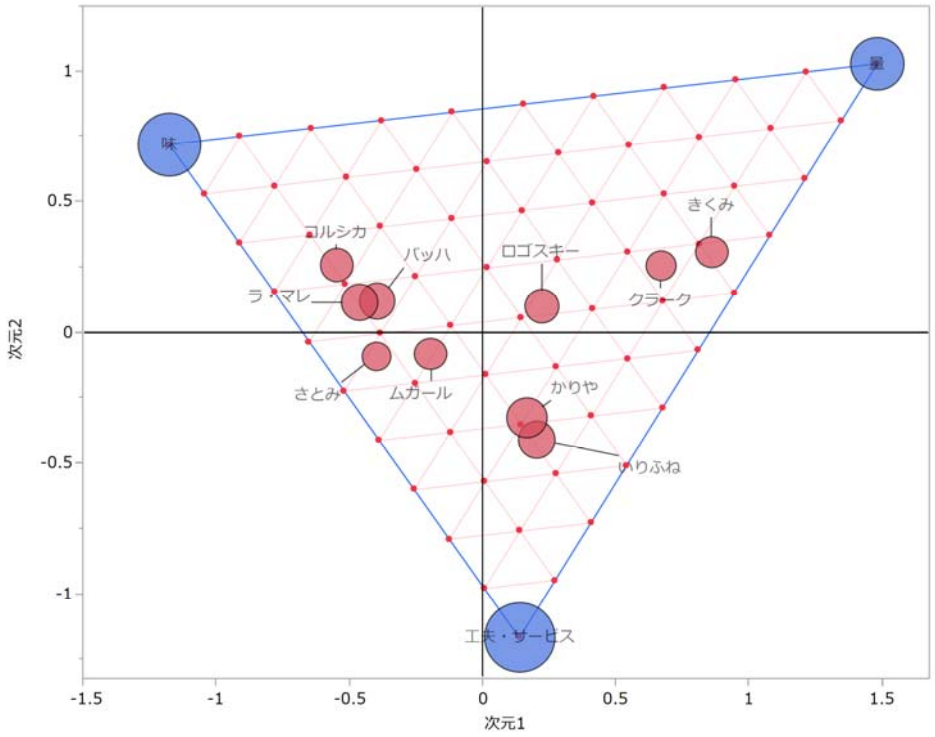


図 9　行プロファイル（ストレッチ・プロファイル）の重心座標系と主軸，成分スコアの対応

4. 2　対応分析法で行うこと

では，対応分析法とは何を行うのであろうか．目標を先に示すと，以下の操作を行うことである．ここからの説明はやや分かりにくいかもしれないが，おおよそこんなこと（このような考え方）ということを読み取ればよい[50]．上の例にみたように，2 次元に縮約できる場合には（それが全情報であるから），単に三角図を描いてその平面内での点つまり行（レストラン）の分布を分析することに同じである．しかし，一般には、高次元の重心座標系での分析となるので（しかし原理・考え方は同じであるが），これを数理的な符丁で言い換えることが必要である．とくに，より一般的に考える場合には，以下の事項に注意することが必要となる[51]．

① 元の 2 元データ表（ここではクロス表）をみると，各列和の大きさが異なることがわかる．そこで，正三角形で考える重心座標系とは異なる，あらたな**“ストレッチした”**重心座標系を考えること．

② つまり，単純な正三角形の座標系では，この列和の重みの違いが考慮されていない．そこで，**“ストレッチしたプロファイル”**とすることで，列和が小さい列ほど，（カイ二乗距離の意味で）プロファイルが中心からより離れるように，列和比率の平方根の逆数（質量の平方根の逆数，つまり $1/\sqrt{p_{+j}}$，表 30 のベクトル $\mathbf{c}$ の各要素の平方根の逆数）で三角図の各辺の長さを調整する（伸ばす）．これを**“ストレッチ・プロファイル”**（streched profile）という．図 9 はこれをイメージして描いてある．この三角形は等辺ではなくそれぞれの辺を伸ばしてある．

③ そして 1 次元目の成分に射影した（表 30 で $\dfrac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}$ とある）重み付きのプロファイルの分散が最大となるように（ここでは，成分 1 軸上に射影の点の分布），各辺がストレッチ

50　ここらは，詳しくは「第Ⅱ部」をあわせて確認すること．
51　ここらの考え方を，スライド資料に模式的に示したので，そちらも参照のこと．

された三角図を回転させてある．このことは，$\dfrac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}$ を要素とするデータ行列を考えることに相当する．

④ またこの回転のとき，原点が平均（重心）ベクトル$(0.421,0.344,0.3.235)$に対応するように，座標原点を調整する（重心でセンタリングするという）．

ここで，①の性質がとくに重要である．つまり，2 元データ表の列和の大きさの違いをどのように（行の）プロファイルに反映させるか，である．これを考慮して，列和が小さい列ほどカイ二乗距離がに大きく反映するようにする．たとえば，列和が 100 である列の度数が 1 つ変化するのと，列和が 10 しかない列の度数が 1 つ変化するのであれば，後者のほうが，カイ 2 乗距離としての影響が大きく，対応分析の結果に強く反映するようにする，と考える．

また，対応分析では，元の重心座標系で，カイ 2 乗距離で測るプロファイル間の距離を“カイ二乗距離”で測っていることを，高次元を低次元に縮約化する際に，（成分スコアを合成変数とすることで）なるべく低次元の“ユークリッド空間”内に表せるように，新たな成分スコアという座標を求めることに相当する．この例では，元の次元が 2 次元であるので，2 次元の平面によって，カイ 2 乗距離のすべてを表現でき，情報をロスすることはない．

[観察]

たとえば，「きくみ」と「クラーク」を比べた場合，「クラーク」のほうが，カイ二乗距離で見た場合に，より平均（重心）に近い．さらに，“「きくみ」と「クラーク」のカイ 2 乗距離”は，“「きくみ」と「ロゴスキー」のカイ 2 乗距離”の，（グラフの見た目でのユークリッド距離より）およそ 1/4～1/3 程度になっている（あまり顕著ではないのだがそうなっている）．

ところで，実際の対応分析の計算処理は，**“特異値分解”**や**“固有値問題”**を解くなど，行列演算の機能を実装したソフトウェアや数学ライブラリーを使えば，比較的，簡単に行える．

たとえば，ここまでに用意した行列，ベクトルなどの記法を使うと，$y_{ij}^{*}=\dfrac{p_{ij}-p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}}$ を要素とするつぎの行列$\mathbf{Y}^{*}=\left(y_{ij}^{*}\right)$を作り，次のようにこれの**特異値分解**（SVD：singular value decomposition）を行えばよい．

$$\mathbf{Y}^{*}=\left(y_{ij}^{*}\right)=\mathbf{P}_{I}^{-1/2}\left(\mathbf{P}_{IJ}-\mathbf{r}\mathbf{c}^{t}\right)\mathbf{P}_{J}^{-1/2} \qquad (24)$$

ところで，この要素 y_{ij}^{*} は，以下のように書き替えることができる．これは，クロス表の独立性の検定に用いるピアソンのカイ二乗統計量における，各セルのカイ 2 乗値の平方根に対応する．

$$y_{ij}^{*}=\frac{p_{ij}-p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}}=\frac{f_{ij}-\dfrac{p_{i+}p_{+j}}{N}}{\sqrt{\dfrac{p_{i+}p_{+j}}{N}}}=\frac{f_{ij}-e_{ij}}{\sqrt{e_{ij}}} \qquad (25)$$

さらに細かい話しになるが，ここで行列$\mathbf{Y}^{*}$の特異値分解を行うことは，行列$\underset{n\times n}{\mathbf{V}}=\left(\mathbf{Y}^{*}\right)^{t}\mathbf{Y}^{*}$の**固有値問題**を解くこと，つまりこれの固有値方程式を解くことに同じである．さらにここ

で，証明は省略するが，$y_{ij}=\dfrac{p_{ij}}{\sqrt{p_{i+}p_{+j}}}$ を要素とする行列 $\underset{m\times n}{\mathbf{Q}}=\left(y_{ij}\right)$ の特異値分解を行ってえも同じ結果がえられることが分かっている[52]．これはまた，要素 y_{ij} として，先ほどと同様に行列 $\underset{n\times n}{\mathbf{V}}=\mathbf{Q}^t\mathbf{Q}$ の固有値問題としても解くこともできる[53]．

［数値例で確認］

特異値分解の関係を例で確かめよう．なおここで，準備として，行列 $\left(\mathbf{Y}^*\right)^t\mathbf{Y}^*$ あるいは $\mathbf{Y}^*\left(\mathbf{Y}^*\right)^t$ の固有値（λ_k，ここで $k=1,2,\cdots,K;K=\min\left\{m,n\right\}-1$）を対角要素とする行列を $\mathbf{\Lambda}=diag\left(\lambda_k\right)$ とする．特異値は固有値の正の平方根であるので，特異値を対角要素とする対角行列は特異値を用いれば，$\mathbf{\Lambda}^{1/2}=diag\left(\sqrt{\lambda_k}\right)=diag\left(\alpha_k\right)$ である．さらに，行列 $\mathbf{Y}^*$ の左特異ベクトルからなる行列を $\mathbf{U}$，右特異ベクトルからなる行列を $\mathbf{L}$ とする．このとき，特異値分解の結果は以下のようになる．

$$\mathbf{U}=\begin{pmatrix} 0.158 & -0.579\\ 0.138 & -0.496\\ 0.566 & 0.369\\ -0.245 & -0.101\\ 0.423 & 0.294\\ -0.381 & 0.325\\ -0.297 & 0.166\\ -0.129 & -0.098\\ -0.352 & 0.164\\ 0.154 & 0.128 \end{pmatrix},\ \mathbf{L}=\begin{pmatrix} 0.088 & -0.756\\ -0.691 & 0.423\\ 0.718 & 0.500 \end{pmatrix} \qquad (26)$$

$$\mathbf{\Lambda}^{1/2}=\begin{pmatrix} \sqrt{\lambda_1} & 0\\ 0 & \sqrt{\lambda_2} \end{pmatrix}=\begin{pmatrix} \sqrt{0.198} & 0\\ 0 & \sqrt{0.060} \end{pmatrix}$$

この特異値分解で得た結果の左・右特異ベクトルと特異値には，$\underset{m\times n}{\mathbf{Y}^*}=\underset{m\times K}{\mathbf{U}}\ \underset{K\times K}{\mathbf{\Lambda}^{1/2}}\ \underset{K\times n}{\mathbf{L}^t}$ の関係がある．今，用いている数値例では，以下のようになる．

---

[52] この場合は得られる固有値の最大根（自明解）が「1」となり，これを除外すると $\mathbf{Y}^*$ を解いた場合に同じになる．これは $\mathbf{Y}^*$ を解いて得られる最小固有値（自明解）に対応している．こうした行列演算の性質がが知られているのでそれを用いる．

[53] 別資料「第Ⅱ部」も参照．

$$\underset{m\times n}{\mathbf{C}} = \underset{m\times K}{\mathbf{U}}\ \underset{K\times K}{\Lambda^{1/2}}\ \underset{K\times n}{\mathbf{L}^t} = \begin{pmatrix} 0.158 & -0.579 \\ 0.138 & -0.496 \\ 0.566 & 0.369 \\ -0.245 & -0.101 \\ 0.423 & 0.294 \\ -0.381 & 0.325 \\ -0.297 & 0.166 \\ -0.129 & -0.098 \\ -0.352 & 0.164 \\ 0.154 & 0.128 \end{pmatrix} \times \begin{pmatrix} \sqrt{0.198} & 0 \\ 0 & \sqrt{0.060} \end{pmatrix} \times \begin{pmatrix} 0.088 & -0.691 & 0.718 \\ -0.756 & 0.423 & 0.500 \end{pmatrix}$$

$$= \begin{pmatrix} 0.11341 & -0.10837 & -0.02055 \\ 0.09721 & -0.09370 & -0.01663 \\ -0.04618 & -0.13560 & 0.22580 \\ 0.00903 & 0.06496 & -0.09065 \\ -0.03789 & -0.09954 & 0.17109 \\ -0.07523 & 0.15083 & -0.08187 \\ -0.04233 & 0.10832 & -0.07444 \\ 0.01301 & 0.02955 & -0.05315 \\ -0.04418 & 0.12504 & -0.09220 \\ -0.01758 & -0.03420 & 0.06489 \end{pmatrix} \qquad (27)\ (\bigstar)$$

さらに，用意した記号に合わせて記すと以下のようになる．各自，確認するとよいだろう．

$$\mathbf{r} = \begin{pmatrix} 0.12072 \\ 0.13863 \\ 0.08567 \\ 0.07399 \\ 0.07944 \\ 0.09502 \\ 0.11059 \\ 0.08489 \\ 0.11371 \\ 0.09735 \end{pmatrix},\quad \mathbf{c} = \begin{pmatrix} 0.42056 \\ 0.34424 \\ 0.23520 \end{pmatrix},\quad \mathbf{r}\mathbf{c}^t = \begin{pmatrix} 0.05077 & 0.04156 & 0.02839 \\ 0.05830 & 0.04772 & 0.03261 \\ 0.03603 & 0.02949 & 0.02015 \\ 0.03112 & 0.02547 & 0.01740 \\ 0.03341 & 0.02735 & 0.01868 \\ 0.03996 & 0.03271 & 0.02235 \\ 0.04651 & 0.03807 & 0.02601 \\ 0.03570 & 0.02922 & 0.01997 \\ 0.04782 & 0.03914 & 0.02674 \\ 0.04094 & 0.03351 & 0.02290 \end{pmatrix} \qquad (28)$$

$$\mathbf{P}_{IJ} = \left(p_{ij}\right) = \begin{pmatrix} 0.07632 & 0.01947 & 0.02492 \\ 0.08178 & 0.02726 & 0.02960 \\ 0.02726 & 0.00623 & 0.05218 \\ 0.03271 & 0.03583 & 0.00545 \\ 0.02648 & 0.01090 & 0.04206 \\ 0.02492 & 0.05997 & 0.01012 \\ 0.03738 & 0.05919 & 0.01402 \\ 0.03816 & 0.03427 & 0.01246 \\ 0.03816 & 0.06386 & 0.01168 \\ 0.03738 & 0.02726 & 0.03271 \end{pmatrix} \tag{29}$$

$$\mathbf{P}_I^{-1/2} = \begin{pmatrix} 2.87817 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2.68579 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3.41654 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.67638 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.54799 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3.24416 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3.00703 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3.43217 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2.96556 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3.20500 \end{pmatrix} \tag{30}$$

$$\mathbf{P}_J^{-1/2} = \begin{pmatrix} 1.54200 & 0 & 0 \\ 0 & 1.70440 & 0 \\ 0 & 0 & 2.06195 \end{pmatrix} \tag{31}$$

$$\mathbf{Y}^* = \left(y_{ij}^*\right) = \mathbf{P}_I^{-1/2}\left(\mathbf{P}_{IJ} - \mathbf{r}\mathbf{c}^t\right)\mathbf{P}_J^{-1/2} = \begin{pmatrix} 0.11342 & -0.10834 & -0.02060 \\ 0.09722 & -0.09367 & -0.01667 \\ -0.04621 & -0.13545 & 0.22565 \\ 0.00904 & 0.06489 & -0.09059 \\ -0.03791 & -0.09943 & 0.17098 \\ -0.07523 & 0.15074 & -0.08177 \\ -0.04232 & 0.10825 & -0.07436 \\ 0.01302 & 0.02951 & -0.05312 \\ -0.04417 & 0.12495 & -0.09210 \\ -0.01759 & -0.03416 & 0.06485 \end{pmatrix} \tag{32}$$ （★★）

（★）の行列 $\underset{m\times n}{\mathbf{C}}$ と（★★）の行列 $\mathbf{Y}^*$ は等しくなるが，これは $\mathbf{Y}^*$ が $\underset{m\times K}{\mathbf{U}}\ \underset{K\times K}{\mathbf{\Lambda}^{1/2}}\ \underset{K\times n}{\mathbf{L}^t}$ のように特異値分解されたことを意味する．そして，対応分析の“成分スコア”は，この行列 $\mathbf{U}$ や $\mathbf{L}$ かの要素を係数（重み）とする**“合成変数”**として表される．

たとえば行の成分スコアは，

$$\underset{m\times n}{\mathbf{P}_I^{-1/2}}\ \underset{n\times K}{\mathbf{U}}\ \underset{K\times K}{\mathbf{\Lambda}^{1/2}} \quad （ここで \mathbf{P}_I^{-1/2} = diag\left(1/\sqrt{p_{i+}}\right)） \tag{33}$$

から得られる．一方，列の成分スコアは，

$$\underset{n\times m}{\mathbf{P}_J^{-1/2}}\ \underset{n\times K}{\mathbf{L}}\ \underset{K\times K}{\mathbf{\Lambda}}^{1/2} \quad （ここで \underset{n\times m}{\mathbf{P}_J^{-1/2}} = diag\left(1/\sqrt{p_{+j}}\right)） \tag{34}$$

から，それぞれ求められる[54].

こうした計算処理は通常はソフトウェアが行ってくれる[55]．実用上大切なことは，対応分析として，得られた固有値，固有ベクトル，そして成分スコアなどが，どのように機能し，何を意味しているかを知ることである．これを引き続いて例で追ってみる．

まず．表 27 のクロス表に対応分析を適用し，**固有値，成分スコア**を算出すると，次の結果を得る．前述のように，固有値はデータ表の行と列の寸法の小さい次元数から 1 を引いたもの，つまりここでは $\min\{m,n\}-1=3-1=2$ となるので固有値の数は 2 個となる（表 32）．従って，2 つの固有値と固有ベクトルに対応する 2 つの成分スコアが，行（項目 I）と列（項目 J）とのそれぞれの選択肢，つまり 10 のレストランと 3 個の評価基準に与えられる．これが，表 33 にある成分スコアの一覧である．

表 32　固有値と寄与率

成分 k	特異値 α_k	固有値 λ_k	寄与率(%) ν_k
1	0.4446	0.19766	76.71
2	0.2450	0.06002	23.29

表 33　レストランと評価基準への成分スコア

		成分スコア	
		第 1 成分スコア	第 2 成分スコア
項目と選択肢＼成分		z_{i1}	z_{i2}
項目 I	いりふね	0.20169	-0.40820
	かりや	0.16472	-0.32610
	きくみ	0.85898	0.30915
	さとみ	-0.40087	-0.09077
	クラーク	0.66717	0.25584
	コルシカ	-0.54972	0.25857
	バッハ	-0.39656	0.12200
	ムガール	-0.19686	-0.08210
	ラ・マレ	-0.46355	0.11909
	ロゴスキー	0.21980	0.10024
項目と選択肢＼成分		z^*_{j1}	z^*_{j2}
項目 J	工夫・サービス	0.06055	-0.28561
	味	-0.52347	0.17643
	量	0.65787	0.25247

54　「第Ⅱ部」に詳しい説明がある．

55　実際の計算処理は，さまざまな数値計算アルゴリズムが工夫されている．ここでみた筆算によるような手順は用いない．とくに，2 元データ表の寸法が大きくなって，しかも各セル内の度数がひどく疎であるような場合には，特別な解法が必要となる．統計ソフトによっては，高次元データ表の処理がむずかしい場合もある．

この得られた成分スコアの**同時布置図**が図 10 である．対応分析では，このグラフが重要である．ここでは，3 つの「評価基準」のどれにどのレストランが高く関係するかを視覚的に読み取れる（つまり，クロス表の行と列との対応関係を吟味している）．またこの布置図に，図 8，9 で示した三角図が再現されていることもわかるであろう．この例は，空間 R^3 内の分布の観察であって，たかだか 2 つの成分スコアであるから，情報は完全に（平面内に）再現される．

一般には，2 元データ表の寸法はかなり大きくなる．つまり，2 つの項目 I, J の選択肢数が非常に大きくなっても，得られた固有値の大きさに順に（つまり，変動の大きい成分から）観察することで，多次元情報をかなり少ない少数次元の成分で説明できるという可能性を示唆している（次元の圧縮化であり，いわゆる“**節約の原理**”[56]（principle of parsimony）の典型例である）．実際にそのような寸法の大きい 2 元データ表を，テキスト・マイニングなどでは扱うので，この次元縮小の効果は大きいのである．こうした例はあとでみよう．

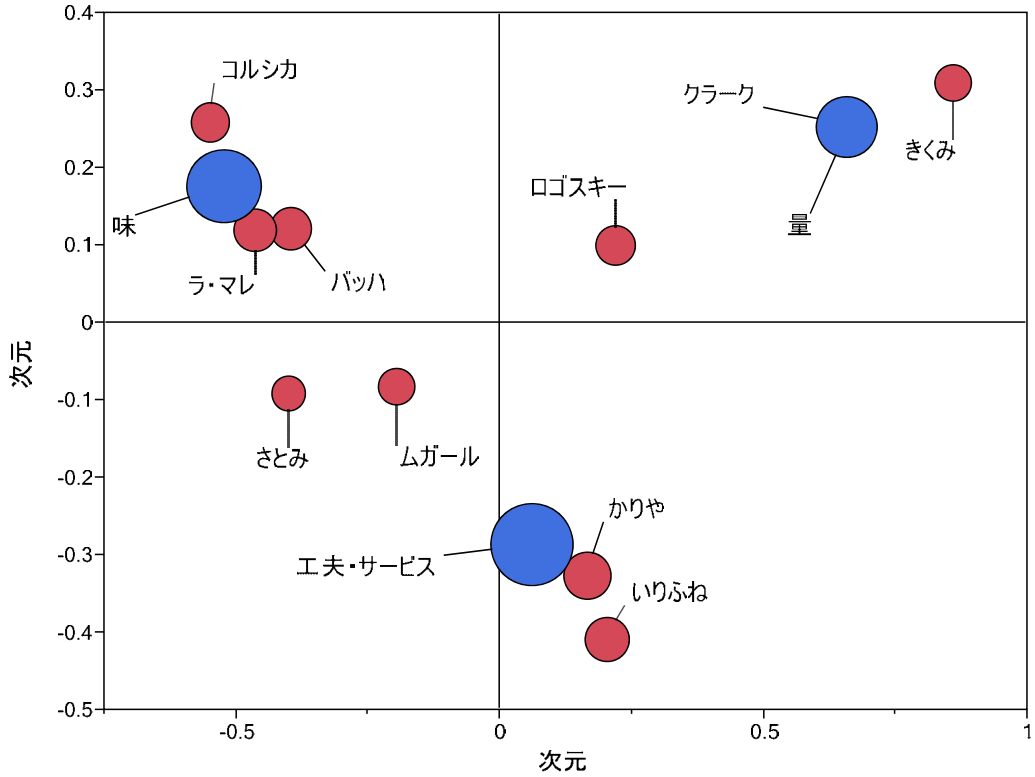


図 10　成分スコアから得た同時布置図
（注）成分スコアの図で，レストランの布置が図 4 あるいは図 9 の三角座標上の分布に類似していることに注意（三角図の布置が再現されている）

4. 3　成分スコアとその性質（とくに双対性）

対応分析法における，行と列の成分スコアには“**双対性**”（duality）の関係がある．ここで，

$$z_{ik}\left(i \in I, k = 1,2,\cdots,K\right)\quad（選択肢 i に対する第 k 成分の成分スコア）\qquad (35)$$

$$z^{*}_{jk}\left(j \in J, k = 1,2,\cdots,K\right)\quad（選択肢 j に対する第 k 成分の成分スコア）\qquad (36)$$

56　「ケチの原理」「思考節約の原理」などとも言う．類似の言葉として「オッカムの剃刀」がある．統計的モデルの例でいえば，モデルが複雑になるほど，所与のデータに対して敏感に応答しすぎるという現象が起こる（過剰適合）．重回帰モデルなどを想定したときに，必要な共変量・説明変数の決め方などで節約の原理が重要になる（とえば，赤池の情報量規準の適用などを想起してみるとよい）．ここでは，多変量・多次元の情報を少数次元の合成変数として表わせるか，というう意味で用いている．簡潔であり単純化することが必ずしも適切であるとは言えないという反論ももちろんある（情報の損失も起こりえるから）．

とすると，行と列との双対性とは，以下のように表される．また，この関係は図 11 のように表される．

$$z_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{n} \left(\frac{p_{ij}}{p_{i+}} \right) z^*_{jk} \left(i \in I, k = 1,2,\cdots,K \right) \tag{37}$$

$$z^*_{jk} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{m} \left(\frac{p_{ij}}{p_{+j}} \right) z_{ik} \left(j \in J, k = 1,2,\cdots,K \right) \tag{38}$$

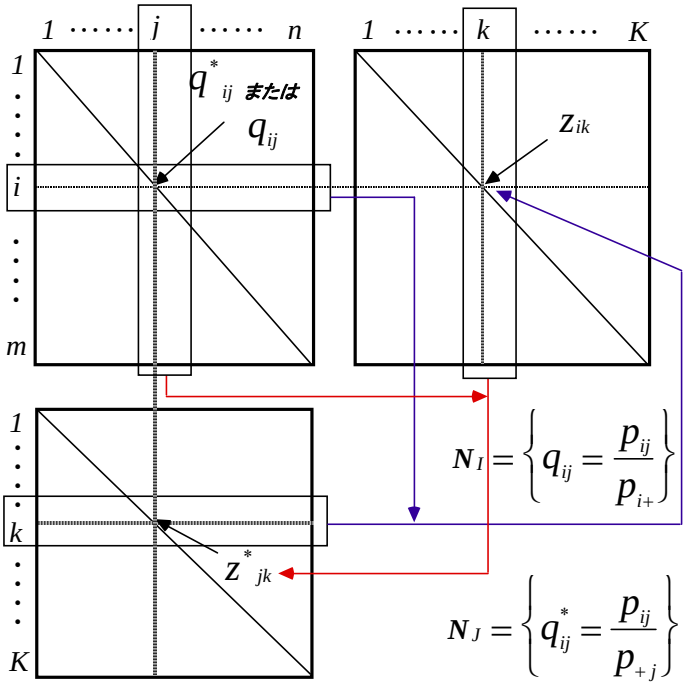


図 11　行成分スコアと列成分スコアの間の関係（双対性）

これを，いままで用いてきた行列，ベクトルの記号を用いると，以下のように書ける．

$$\underset{m\times 1}{\mathbf{z}_k} = \frac{1}{\sqrt{\lambda_k}} \underset{m\times m}{\mathbf{P}_I^{-1}} \underset{m\times n}{\mathbf{P}_{IJ}} \underset{n\times 1}{\mathbf{z}^*_k}, \quad \underset{m\times 1}{\mathbf{z}_k} = \begin{pmatrix} z_{1k} \\ z_{1k} \\ \vdots \\ z_{ik} \\ \vdots \\ z_{mk} \end{pmatrix} \tag{39}$$

$$\underset{n\times 1}{\mathbf{z}^*_k} = \frac{1}{\sqrt{\lambda_k}} \underset{n\times n}{\mathbf{P}_J^{-1}} \underset{n\times m}{\mathbf{P}_{JI}} \underset{m\times 1}{\mathbf{z}_k}, \quad \underset{n\times 1}{\mathbf{z}^*_k} = \begin{pmatrix} z^*_{1k} \\ z^*_{2k} \\ \vdots \\ z^*_{jk} \\ \vdots \\ z^*_{nk} \end{pmatrix}, \text{ ここで } \mathbf{P}_{JI} \text{ は } \mathbf{P}_{IJ} \text{ の転置行列，つまり } \mathbf{P}_{JI} = \mathbf{P}_{IJ}^t, \tag{40}$$

この式の意味は重要である．言葉で説明すると，以下のような関係を示している．

- 項目 I のある選択肢 i の成分スコアは，項目 J の標準化スコア（分散 1 に標準化した成分スコア）を列プロファイルで加重した平均である．
- また，項目 J のある選択肢 j の成分スコアは，項目 I の標準化スコア（分散 1 に標準化した成分スコア）を行プロファイルで加重した平均である，

そして，上の 2 つの式で z_{ik} ， z^*_{jk} がたすきがけになって左右の項に入っていることに注意しよう（上の図 11 と表 34 で確認）．列と行の成分スコアでこのような関係がある性質を，**“双対性”**(duality) という．また上の式 (31)，(32) を“互いに**推移関係** (transition relationship) にある”という．この性質は，成分スコアの同時布置図を観察するときに重要である．

表 34　項目 I, J の選択肢の成分スコアと確率行列の関係

		項　目　J						成分スコア							
		1	2	⋯	j	⋯	n	1	2	⋯	k	⋯	k'	⋯	K
項目 I	1	p_{11}	p_{12}	⋯	p_{1j}	⋯	p_{1n}	z_{11}	z_{12}	⋯	z_{1k}	⋯	$z_{1k'}$	⋯	z_{1K}
	2	p_{21}	p_{22}	⋯	p_{2j}	⋯	p_{2n}	z_{21}	z_{22}	⋯	z_{2k}	⋯	$z_{2k'}$	⋯	z_{2K}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	i	p_{i1}	p_{i2}	⋯	p_{ij}	⋯	p_{in}	z_{i1}	z_{i2}	⋯	z_{ik}	⋯	$z_{ik'}$	⋯	z_{iK}
	⋮	⋮	⋮		⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	m	p_{m1}	p_{m2}	⋯	p_{mj}	⋯	p_{mn}	z_{m1}	z_{m2}	⋯	z_{mk}	⋯	$z_{mk'}$	⋯	z_{mK}
成分スコア	1	z^*_{11}	z^*_{21}	⋯	z^*_{j1}	⋯	z^*_{n1}	⇑							
	2	z^*_{12}	z^*_{22}	⋯	z^*_{j2}	⋯	z^*_{n2}	行の項目 I の選択肢の成分スコア							
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	$\mathbf{Z}_{m\times K} = \underbrace{\mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2}}_{m\times n}\mathbf{L}_{n\times K}$							
	k	z^*_{1k}	z^*_{2k}	⋯	z^*_{jk}	⋯	z^*_{nk}	⇐ 列の項目 J の選択肢の成分スコア							
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	$\mathbf{Z}^*_{n\times K} = \underbrace{\mathbf{P}_J^{-1}\mathbf{P}_{JI}\mathbf{P}_I^{-1/2}}_{n\times m}\mathbf{U}_{m\times K}$							
	k'	$z^*_{1k'}$	$z^*_{2k'}$	⋯	$z^*_{jk'}$	⋯	$z^*_{nk'}$								
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	（ここで，$K = \min\{m,n\} - 1$）							
	K	z^*_{1K}	z^*_{2K}	⋯	z^*_{iK}	⋯	z^*_{nK}								

4. 4　成分スコアの解釈

得られた成分スコアについて，**“布置図”**や**“同時布置図”**描いて観察することが対応分析で得た結果解釈に有効である[57]．

①　成分スコアの布置図

行あるいは列の選択肢に対する成分スコア，つまり表 34 にある成分スコアのうち，観察したい 2 成分 k, k' を指定して布置図を描き成分スコアの分布を観察する．

[57] 対象とする 2 元データ表の寸法が大きくなると，とくにテキスト・マイニングで扱うようなデータ表では布置図での観察は面倒である．こうした場合の手当，工夫が必要である．

$$\left(z_{ik}, z_{ik'}\right) \quad \begin{pmatrix} i=1,2,\cdots,m \\ k,k'=1,2,\cdots,K \\ K=\min\{m,n\}-1 \end{pmatrix} \qquad \text{（行の選択肢への成分スコア）} \qquad (41)$$

$$\left(z^*_{jk}, z^*_{jk'}\right) \quad \begin{pmatrix} i=1,2,\cdots,m \\ k,k'=1,2,\cdots,K \\ K=\min\{m,n\}-1 \end{pmatrix} \qquad \text{（列の選択肢への成分スコア）} \qquad (42)$$

② 成分スコアの同時布置図

行, 列それぞれの選択肢への成分スコアを重ねた散布図を**同時布置図**という[58]. すなわち,

$$\left(z_{ik}, z_{ik'}\right), \left(z^*_{jk}, z^*_{jk'}\right) \quad \begin{pmatrix} i=1,2,\cdots,m \\ k,k'=1,2,\cdots,K \\ K=\min\{m,n\}-1 \end{pmatrix} \qquad (43)$$

を同じ散布図上にのプロットして図式化する.

これは前にみた表 34, 図 11 に相当するもので, 別の図として書き替えたに過ぎない. 要は行側の成分スコア（項目 I の選択肢への成分スコア）と列側の成分スコア（項目 J の選択肢への成分スコア）とをいつも対に考えることにある（式（33), (34)）.

いまの例について, これを作ってみよう. 図 11 は, 10 の「レストラン」の成分スコア, 図 12 は 3 つの「評価基準」の成分スコア, それぞれについての布置図である. また前に示した図 10 が同時布置図に相当する.

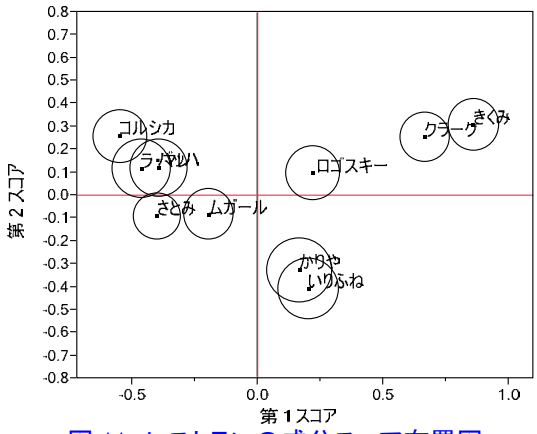


図 11　レストランの成分スコア布置図

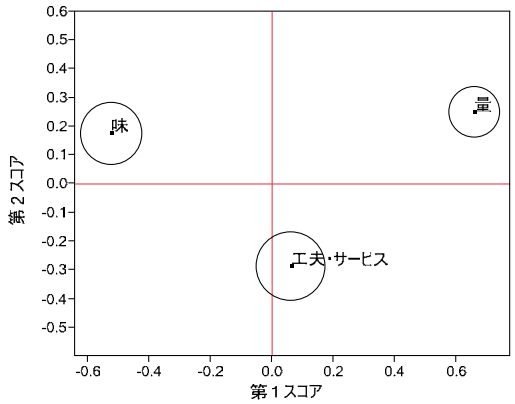


図12　評価基準の成分スコア布置図

58「同時布置図の考え方」については「第Ⅱ部」に示した. 同時布置図の解釈には注意が必要である. また, これを巡るさまざまな議論がある.

この例はクロス表の寸法も小さく，また意図的に分かりやすい例として作った人工データである．よって，実用上はさほど説得力のある例とはいえないのだが，対応分析法の仕組みはわかるであろう．

［成分スコアを観察する際の注意事項］

(1) まず，個々の成分スコアを 1 次元的に観察する（第 1 成分スコアに注目する）．とくに，第 1 固有値の寄与率が高いときにはこの操作が大切である．行と列との第 1 成分スコアを数直線上に並べて描いてみると良い（たとえば，図 2 はその一例である）．

(2) 固有値の先頭の方に，値の大きな固有値とくに「1」またはそれに近い値が現れるような場合は注意する．このとき，はずれ値的なパターンをもつプロファイルがある，または特徴的なプロファイルがなく，ほとんどまとまって分布することが考えられる．

(3) 次に，（固有値の大きさを念頭に）任意の 2 つの成分スコアに注目し，散布図（布置図）を描き各点の布置の相対的な位置関係に注目する．

(4) 軸の解釈は場合に応じて考慮する．また，軸に解釈を与えることよりも，成分スコアの相対的な遠近，位置関係を観察する（これを客観的に観察するために，後述の相対寄与度，絶対寄与度を用いる）．

(5) 「多重クロス表」から求めたサンプルの成分スコアの解釈は「元の変量・項目の選択肢のスコア」(つまりアイテム･カテゴリー型に展開した延べのカテゴリーへのスコア)であるから意味理解に注意する（とくに選択肢の並び順，順序関係に注意）．

(6) 固有値，寄与率は，多重クロス表から出発の場合は，大きくなることはほとんどないので解釈には注意する（見かけ上，寄与率が大きくはならない[59]）．

(7) 選択肢が順序尺度の場合には，図中の選択肢の並び順に注意する．

(8) この意味で成分スコアを用いたクラスター化操作には十分な注意が必要である．単純な *k*-平均法や階層的分類（例：ウォード法）ではうまく対応できないことがある，たとえば，WordMiner では階層的分類法の 1 つウォード法と分割化型分類法の *k*-平均法とを併用したハイブリッド法を用いている[60]．

(9) **“はずれ値”**の存在に注意する．はずれ値は元のデータ表の中の頻度分布の不均衡（プロファイルの不均衡）から生じる．これに敏感なことが対応分析の特徴でもある．

(10) 成分スコアの**“標準化”**：行あるいは列の成分スコアの同時布置を考えたとき，それらの標準化（平均値＝0，分散＝1 とすること）に際しては，それぞれを「分散が 1 になるように“標準化する”場合」と「分散が固有値になるように調整する場合」がある．つまり．少なくとも 4 通りの組み合わせがある（表 35）．元来の対応分析では，多くの場合，行・列ともに**“分散が固有値となるように調整した”**成分スコアを用いることが多い（つまり，成分スコアの分散は固有値 λ_k のままを用いる）．また，数量化法 III 類の場合には，両者の分散を 1 となるように標準化することが多いようである[61]．

(11) 同時布置に関するいくつかの議論があること．表 35 の「その 2」や「その 3」が望ましいとする意見もある[62]「その 2」や「その 3」の分散 1 に標準化された列もしくは行は，列と行との距離も，自然に解釈できるという利点がある．たとえば図 13 に，上の例を「その 2」「その 3」のオプションで描いてみた．しかしこの例にみるように，たいていは，点の分布，列成分スコアと行成分スコアの関係が不自然となり観察もしにくい．前述したように，対応分析法では，通常は表 35 の「その 1」，つまり成分スコアの標準化を行わないで布置図を描くことが多い(また，それについての理由もある[63])．

---

59 多重対応分析の場合の寄与率の考え方については，「第Ⅱ部」で説明している．

60 「第Ⅲ部」に説明がある．

61 多くの場合，標準化操作についての断りがないので布置図の観察を見誤ることがある．利用上注意すること．

62 たとえば，*Journal of Marketing Research* 誌の論文，Carroll, J.D., Green, P.E., and Schaffer, C.M. (1986, 1987, 1989)や Greenacre, M.J. (1989)を参照．

63 Lebart, L., Salem, A. and Berry, L. (1998): *Exploring Textual Data*, Kluwer Academic Publishers など参照．

表 35　成分スコアの分散の組み合わせ

	組合せ	項目 I の選択肢の成分スコア：z_{ik}	項目 J の選択肢の成分スコア：z^{*}_{jk}
分散の大きさ	その 1	λ_k	λ_k
	その 2	λ_k	1
	その 3	1	λ_k
	その 4	1	1

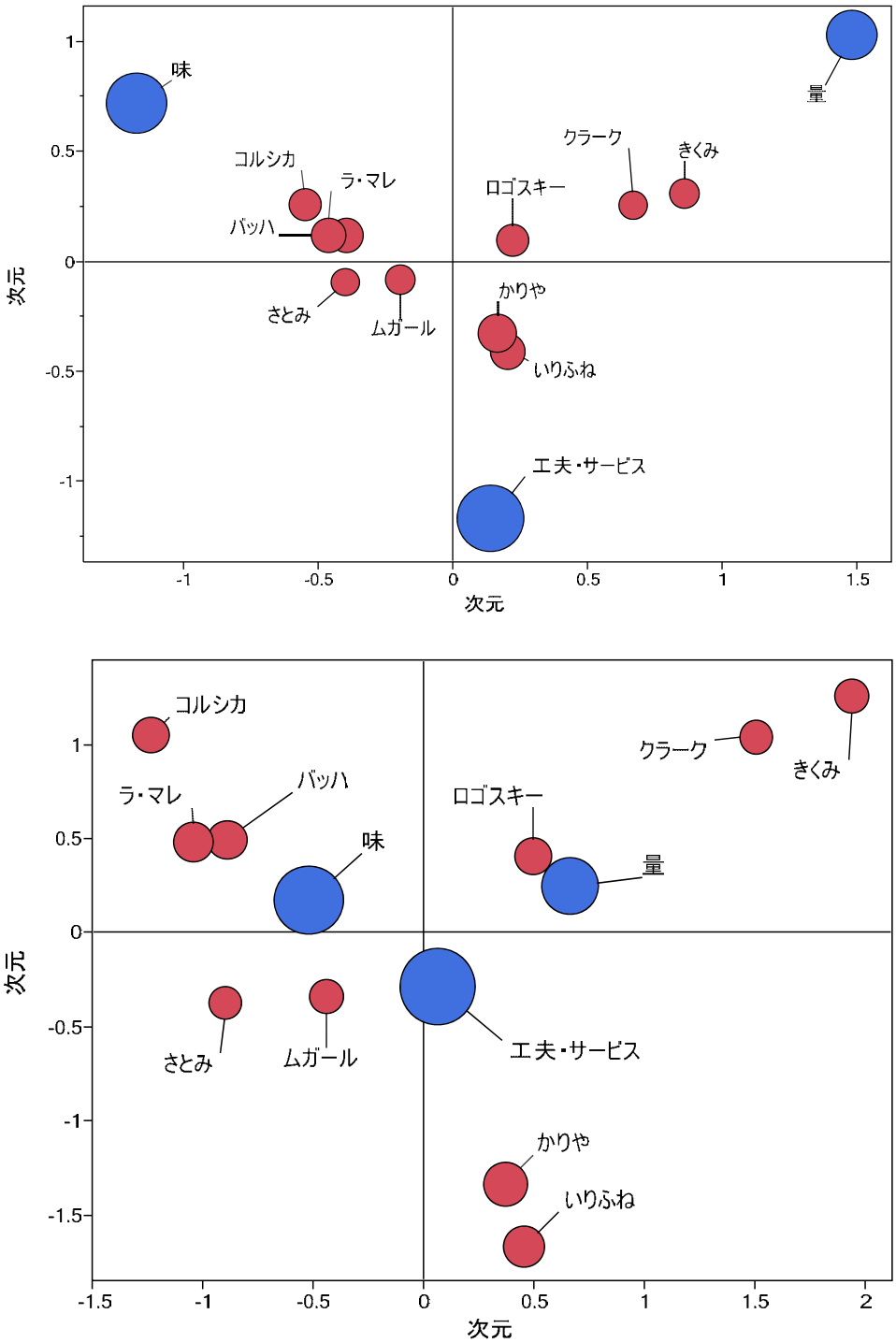


図 13 「その 2」(上)，「その 3」(下)の組合せで描いた図

[布置図・同時布置図の見方，解釈の要点]

布置図，同時布置図を探索的に観察しながら分析を進めることが大切である．たとえばWordMiner を利用して行う初動探査，探索的アプローチにおいては，以下の事項を念頭に対応するとよいだろう．なお以下に記すことについては，うしろに挙げた分析例の説明を読んだ後に，再度確認するのがよいだろう[64].

(1) まず，成分スコアの個々の布置図を観察する．固有値の大きさ（成分スコアの分散）を勘案しながら，なるべく多数の成分の組み合わせを観察する方がよい．
(2) 行のスコアと列のスコアについて，布置図内の位置が近いからといって，そのまま「類似している，距離が近い」と判断してはいけない．これは双対性の原理から明らかである（相互のプロファイルを加重とする平均になっている）．
(3) このことから，両者のスコアを（同時布置内で）同時的には括れない．たとえば，クラスター化を両者を併せたスコアについて同時的には行えない．
(4) しかし，相互の加重平均とした結果であるから，近い位置にある行と列との点（成分スコア）は，双対性を考慮したうえで親近性を評価すればよい．行と列ともに，分散が固有値に等しくなるように調整した成分スコア（表 35 の「その 1」）の場合，行間の距離，列間の距離といった解釈を，行と列との間の距離と解釈してはいけない[65].
(5) 布置図は高次元空間内のおよその情報（近似）を知るための手がかりとする．次元縮約を行ったことに注意しよう．
(6) つまり，成分スコアは元の 2 元データ表の選択肢の合成指標つまり加重平均であるから，じつは多次元情報の縮約でもあることに注意する．
(7) 実用上は，寸法が大きく，しかも非常に疎なデータ表を扱うことが多い．たとえば，テキスト・マイニングなどで扱うデータ表は行列の寸法が大きいだけでなく非常に疎となるので，少数次元内に布置することが難しい．よって布置図は一つの目安とし，かならず**他の情報**（たとえば，**寄与度，有意性テスト**など）を併用する[66].
(8) 無数の成分スコアの点を布置した図の視認には限界があるので（煩雑になる，視覚化の限界）別の方法と併用する（**有意性テストの結果**を検討）．
(9) 基本的には“視認できる情報の範囲，限界”をよく知ったうえで用いる．同時に，固有値（特異値）と寄与率，寄与度（相対寄与度，絶対寄与度），その他の情報も併用，吟味する．
(10) 布置図の観察は**“はずれ値の検出”**には有効である．布置図は布置図の分布の周辺から観察するのがよい．これと**寄与度（絶対寄与度，相対寄与度）**を併用するとよい（具体的にどの程度のはずれ具合かを知る）．
(11) テキスト型データの場合，2 元データ表の形式を使い分ける．たとえば，「(回答・サンプル) × (構成要素変数)」のデータ表[67]からは，回答・サンプルと構成要素の関係を知る．
    - どの回答にはどんな構成要素が使われているかなど
    - クラスター化で得た類型を特徴付ける構成要素を有意性テストで客観的に調べ，要約する．

---

64 この資料のうしろのほうに，いくつかの分析例を挙げた．これを確認してから，ここの記述を再読されたい．

65 バイプロットの提唱者である Gabriel (2002) によると，行と列との関係を“バイプロット”としてみたとき，原点からある点へのベクトルの内積は，それほど不正確ではないと指摘している．「間違っている (のではないか)」とされている解釈も，実用上は，それほどは間違っていないことを指摘している．Gabriel, K. R. (2002), Goodness of fit of biplots and correspondence analysis, *Biometrika*, **89**(2), pp.423-436.

66 うしろに分析例として説明してある．

67 (回答・サンプル) × (構成要素変数)，(構成要素変数) × (質的変数：属性やクラスター変数) などのデータ表の意味・説明については，別資料を参照のこと（例：http://wordminer.org/tips/63）．

(12) さらに「(構成要素変数) × (質的変数：選択肢型質問，属性やクラスター変数)」の 2 元データ表から，構成要素変数と質的変数や人口統計学的要因・属性，あるいはクラスター変数との関係を調べる．
- このときは比較的少数次元の空間内に布置できることが多いので，布置図をしっかりと観察する．
- どの質的変数がどの構成要素に強く関係かを，有意性テストで客観的に調べ要約する．
- 社会調査データの場合は，とくに人口統計学的要因（属性，ライフスタイルなど）が関連することが多いので，こうした変数から取り上げて探査するとよい．

(13) はずれ値となりやすい，まれな回答例や出現頻度の低い構成要素を探査するとき，**追加処理機能**[68]を使うとよいことがある．多くの場合，回答分布に偏りがあるのが常である．また追加処理機能を使って一時除去を行った解析から，その除去効果を知る．

(14) 構成要素変数の再編集を繰り返し，その編集の効果を知る．
- 構成要素変数の編集（自由回答・自由記述に登場した単語・語句）によって，その影響がどこにどう現れるかを知る．

4. 5 対応分析で用いる主な指標

対応分析の分析結果を適切に解釈するため，さまざまな指標が必要である．ここでは対応分析法の結果を読み解くうえで必要な主な指標を説明する．一見面倒にみえるが，後に述べる例題，分析例を参考にする，あるいは自分で人工的にミニチュアのトイ・データを作ってみて，諸指標がどのような挙動をするかを知るのもよいだろう．

対応分析法で必要となる統計量をはじめ，基本の要素を一覧とした（表 36）[69]．ここでは，2 元データ表の，行の側からの分析（n次元空間R^n内での分析）と，列の側からの分析（m次元空間R^m内での分析）に分けて示してある．対応分析法を使うためには，この表のそれぞれの要素がどのような意味を持ち，どのように使えるかを知れば，とりあえずは十分である．

表 36 分析の基本要素

	項目 I ：行の側から分析	項目 J ：列の側から分析
	n 次元空間 R^n 内での分析	m 次元空間 R^m 内での分析
	行和を 1 としたときの「行プロファイル」で $(n-1)$ 次元内に分布する m 個の点	列和を 1 としたときの「列プロファイル」で $(m-1)$ 次元内に分布する n 個の点
プロファイル	行プロファイル $\mathbf{N}_I = \mathbf{P}_I^{-1}\mathbf{P}_{IJ} = \left\{ q_{ij} = \dfrac{p_{ij}}{p_{i+}} \middle\| i \in I, j \in J \right\}$	列プロファイル $\mathbf{N}_J = \mathbf{P}_J^{-1}\mathbf{P}_{JI} = \left\{ q_{ij}^* = \dfrac{p_{ij}}{p_{+j}} \middle\| i \in I, j \in J \right\}$
プロファイル間の距離	行プロファイル間のカイ二乗距離 $d_B^2(i,i') = \sum_{j=1}^{n} \dfrac{1}{p_{+j}} \left(q_{ij} - q_{i'j} \right)^2$ $= \sum_{j=1}^{n} \dfrac{1}{p_{+j}} \left(\dfrac{p_{ij}}{p_{i+}} - \dfrac{p_{i'j}}{p_{i'+}} \right)^2$	列プロファイル間のカイ二乗距離 $d_B^2(j,j') = \sum_{i=1}^{m} \dfrac{1}{p_{i+}} \left(q_{ij}^* - q_{ij'}^* \right)^2$ $= \sum_{i=1}^{m} \dfrac{1}{p_{i+}} \left(\dfrac{p_{ij}}{p_{+j}} - \dfrac{p_{ij'}}{p_{+j'}} \right)^2$
固有値 (慣性) 寄与率 累積寄与率	固有値：$\lambda_k \left(\begin{matrix} k = 1,2,\cdots,K \\ K = \min\{m,n\} - 1 \end{matrix} \right)$　ここで $0 \le \lambda_k \le 1$ 寄与率：$\nu_k = \dfrac{\lambda_k}{\sum_{k=1}^{K} \lambda_k} \times 100(\%) \left(\begin{matrix} k = 1,2,\cdots,K \\ K = \min\{m,n\} - 1 \end{matrix} \right)$　（第 k 成分の寄与率）	

68 追加処理機能については「第Ⅱ部」も参照．
69 「第Ⅱ部」から引用．

	このν_kをkについて累積すれば**累積寄与率**$\sum_{k=1}^{K^*}\nu_k$となる.	
総変動・総分散 慣性の総量 （全慣性）	$\chi_p^2=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{N\left(p_{ij}-p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}}$　　または　$\phi^2=\frac{\chi_p^2}{N}=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(p_{ij}-p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}}$ ここで，固有値との間に，$\phi^2=\frac{\chi_p^2}{N}=\sum_{k=1}^{K}\lambda_k\left(K=\min\{m,n\}-1\right)$の関係がある.	
総変動・総分散 慣性の関係	$\phi^2=\frac{\chi_p^2}{N}=\sum_{i=1}^{m}p_{i+}\sum_{j=1}^{n}\frac{1}{p_{+j}}\left(\frac{p_{ij}}{p_{i+}}-p_{+j}\right)^2$ $=\sum_{i=1}^{m}p_{i+}\sum_{j=1}^{n}\left(\frac{q_{ij}}{\sqrt{p_{+j}}}-\sqrt{p_{+j}}\right)^2$	$\phi^2=\frac{\chi_p^2}{N}=\sum_{j=1}^{n}p_{+j}\sum_{i=1}^{m}\frac{1}{p_{i+}}\left(\frac{p_{ij}}{p_{+j}}-p_{i+}\right)^2$ $=\sum_{j=1}^{n}p_{+j}\sum_{i=1}^{m}\left(\frac{q_{ij}^*}{\sqrt{p_{i+}}}-\sqrt{p_{i+}}\right)^2$
絶対寄与度 (*)100倍して%として用いることがある.	第k成分における選択肢$i(\in I)$の絶対寄与度 $C_k(i)=\frac{p_{i+}\left(z_{ik}\right)^2}{\lambda_k}\quad\begin{pmatrix}i\in I,k=1,2,\cdots,K\\K=\min\{m,n\}-1\end{pmatrix}$ $\sum_{i=1}^{m}C_k(i)=1$	第k成分における選択肢$j(\in J)$の絶対寄与度 $C_k(j)=\frac{p_{+j}\left(z_{jk}^*\right)^2}{\lambda_k}\quad\begin{pmatrix}j\in J,k=1,2,\cdots,K\\K=\min\{m,n\}-1\end{pmatrix}$ $\sum_{j=1}^{n}C_k(j)=1$
相対寄与度 あるいは 平方相関 (*)100倍して%として用いることがある.	選択肢$i(\in I)$に対する相対寄与度 $C_k^*(i)=\frac{d_K^2(i,G)}{d^2(i,G)}=\frac{z_{ik}^2}{\sum_{j=1}^{n}\frac{1}{p_{+j}}\left(\frac{p_{ij}}{p_{i+}}-p_{+j}\right)^2}$ $\begin{pmatrix}i\in I,k=1,2,\cdots,K\\K=\min\{m,n\}-1\end{pmatrix}$ $d^2(i,G)$：点$i\in I$から重心Gまでの平方カイ二乗距離	選択肢$j(\in J)$に対する相対寄与度 $C_k^*(j)=\frac{d_K^2(j,G)}{d^2(j,G)}=\frac{\left(z_{jk}^*\right)^2}{\sum_{i=1}^{m}\frac{1}{p_{i+}}\left(\frac{p_{ij}}{p_{+j}}-p_{i+}\right)^2}$ $\begin{pmatrix}j\in J,k=1,2,\cdots,K\\K=\min\{m,n\}-1\end{pmatrix}$ $d^2(j,G)$：点$j\in J$から重心Gまでの平方カイ二乗距離
双対性	$\underset{m\times1}{\mathbf{z}_k}=\frac{1}{\sqrt{\lambda_k}}\underset{m\times m}{\mathbf{P}_I^{-1}}\underset{m\times n}{\mathbf{P}_{IJ}}\underset{n\times1}{\mathbf{z}_k^*}$ $z_{ik}=\frac{1}{\sqrt{\lambda_k}}\sum_{j=1}^{n}\left(\frac{p_{ij}}{p_{i+}}\right)z_{jk}^*\left(i\in I;k=1,2,\cdots,K\right)$ 列の選択肢$j\in J$の成分スコアの加重和が$i\in I$の成分スコア	$\underset{n\times1}{\mathbf{z}_k^*}=\frac{1}{\sqrt{\lambda_k}}\underset{n\times n}{\mathbf{P}_J^{-1}}\underset{n\times m}{\mathbf{P}_{JI}}\underset{m\times1}{\mathbf{z}_k}$ $z_{jk}^*=\frac{1}{\sqrt{\lambda_k}}\sum_{i=1}^{m}\left(\frac{p_{ij}}{p_{+j}}\right)z_{ik}\left(j\in J;k=1,2,\cdots,K\right)$ 列の選択肢$i\in I$の成分スコアの加重和が$j\in J$の成分スコア
再生公式 または 推移公式	$p_{ij}=p_{i+}p_{+j}\left(1+\sum_{k=1}^{K}\frac{1}{\sqrt{\lambda_k}}z_{ik}z_{jk}^*\right)\quad\left(i\in I,j\in J,K=\min\{m,n\}-1\right)$	

5. テキスト型データの解析への応用

5. 1　データ表の基本的な組み合わせ

以上に述べたことから明かなように，対応分析は，（クロス表に限らず）より一般的な 2 元データ表を分析する手法である．ここで，前に述べたように，テキスト型データを質的データと考えて，データ表を構成する見方を思い出そう．選択肢型質問のように，限られた用語句を用いるのではなく，これを自由回答や自由記述でえた情報から作った単語群や語句類に当てはめることは容易に想像できる．つまり，自由回答質問や自由記述で得た**“テキスト型データ”**（textual data）の分析に適したデータ表を用意するには，どのように考えればよいだろう．基本の考え方は，2 元データ表の，列側と行側にどのような要素を対応させればよいか，を工夫することである．ここで 1 つの課題は，自由回答・自由記述のような文章化された情報から，どのように単語や語句を取り出すか，である．多くのテキスト・マイニング（TM）ツールでは，以下のような操作を行う．つまり，裸のテキスト型データ（生データ）そのものをいきなり対応分析やクラスター化で分析することは，ほぼ無理である．言い替えると，TM では，事前処理加工が，どのように行われたか，あるいは行うべきか，その手順の明示化が重要な鍵となる．しかもその多くの手順がアドホックでありヒューリスティックとなることも多い．残念ながら多くの TM ツールでは，この重要な処理過程についての説明や記述があまり明らかではないことが問題である．

- 日本語で書かれた文章には，句読点などの区切り記号での区切りはあっても，基本的には切れ目がない．そこで，データ解析上の扱い単位としたいある要素単位に“文章を区分する”必要がある．これを**“分かち書き”**（segmentation）といい，分かち書きを行う過程を**分かち書き処理**という．
- この分かち書きを行うツール（ソフト）は，フリーウェア，シェアウェア，有料の商品まで，といろいろある[70]．いずれにしても，日本語のテキスト型データを扱うには，この分かち書き処理が必要となる．
- 分かち書きの結果は，“単語”とは限らない．また，言語学的になにか規則的にあるいは厳密な定義にしたがって，文章がきちんと分かち書きができるとは限らない．用いるソフトによって差違があるし，そもそも言語学的，文法的にも厳密な分かち書きが可能とも限らない．また非常に重要な点は，自由回答・自由記述，あるいは一般に，人により書かれた文章は，よほどのことがないかぎりは，一定の規則，たとえば文法通りには書かれてはいないことが多い．つまり，かなり“ノイズの入った”不確かで**“非構構造的”**あるいは**“半構造的”，“非定型”**なデータである，ということである[71]．これを前提に分析を進める必要がある．
- 分かち書き処理と同時に，分かち書き単位の品詞特定などを行う，いわゆる**“形態素解析”**がある．これをどのように行うか，どこまで行うか，などもいろいろ検討事項あるが，ここでは触れない[72]．また，用いる形態素解析ツールによって，処理結果・内容が同じにはならないことにも注意する必要がある．
- 分かち書き処理の結果は**一意には決まらない**ので，分析目的に応じて，必要な範囲で**“辞書編集”**などを行うこともある．たとえば，「情報処理技術」とあると，「情報」「処理」「技術」もあれば，「情報処理」「技術」もあり，また「情報」「処理技術」とすることもあるかもしれない．ここらは，分かち書きツールのオプションを使い分けて，また必要に応じて，当該分析のための辞書を作るなどが必要となる．

---

70 茶筌（Chasen），mecab，juman，Wordbreaker など多数ある．「茶筌」を利用する例が多い．WordMiner は独自のツール（Happiness）を用いている．

71 テキスト型データはそういう性格のデータであることを前提に分析を考えねばならない．換言すると，たとえば社会調査であれば，どのような自由回答質問の作り方が適切であるか，という調査票設計の適否が，のちの分析に影響する，ということである．

72 欧米語（とくに 26 文字の組合せで語句を構成する）には，原則分かち書きは不要である．日本語は欧米語に比べて構造が複雑とされ，形態素解析の結果はばらつきが多い．欧米のテキスト・マイニング・ツールがそのまま利用することの難しさがここにある．

・ ここで必要な情報は，文章の分かち書きで得られる結果は一意ではなく，また上の例でみるように分割（分かち書き）の最小単位が単語とも限らず，複数の語句の集まりであることもある（ゴミも混じっている）．そこでここでは，データ解析上で扱う単位を**“構成要素”**（components）と呼ぶことにする．上の例であれば，「情報」「処理」「技術」「情報処理」「処理技術」「情報処理技術」とどれもが利用目的によって使い分ける必要があり，これをここでは構成要素と呼ぶことにする．

さて，以上を準備すると，ある文章を分かち書きしてえられる“構成要素”は，いままでの議論の中で登場した“選択肢”と同じように扱える質的データである．調査で用いる選択肢型質問の質的変数と，自由回答からえられる構成要素の出現度数との関係を表したイメージが図 14 である．ここで，「$w_j^{(i)}$」は「第 i 番目の回答者から得たテキスト型データの分かち書きで得た**構成要素**の出現度数を表す．行側が人口統計学的変数（たとえば年齢区分）であれば．その選択肢区分ごとの構成要素の度数を表している．

		分かち書きで得られる構成要素（単語，語句，キーワード…）				
「回答者・サンプル」あるいは「質的変数・人口統計学的変数」あるいはクラスター化でえたクラスター変数	1	$w_1^{(1)}$	$w_2^{(1)}$	…	$w_j^{(1)}$	…
	2	$w_1^{(2)}$	$w_2^{(2)}$	…	$w_j^{(2)}$	…
	⋮	⋮	⋮	⋮	⋮	…
	i	$w_1^{(i)}$	$w_2^{(i)}$	…	$w_j^{(i)}$	…
	⋮	⋮	⋮	⋮	⋮	…
	n	$w_1^{(n)}$	$w_2^{(n)}$	…	$w_j^{(n)}$	…

図 14 （回答者）×（構成要素），（質的変数）×（構成要素）のイメージ図

このように分類集計すれば，テキスト型データも違和感なく質的データとして扱うことができる．そもそも，テキスト型データとは構造的にそのような性格を備えている．つまり，対応分析法との相性がよいともいえる．

表 36 WordMiner におけるデータ表の関係

項目：J	項目：I
● 構成要素変数 分かち書き結果 キーワード抽出の結果	● 回答（サンプル），個体
	● 質的変数 （選択肢型設問・人口統計学的変数等）
● 別のソフトで生成した語句群 たとえば「茶筌」などを利用して抽出した語句群	● クラスター変数 ※）クラスター・メンバーシップ情報から得られるクラスター変数は質的変数に変換して名義尺度データとして使える
● 独自に辞書編集した語句群	● 他の外部情報源から生成した質的変数 （*）分析対象とするデータセットの項目の一部と関連付けができる情報があるとき

（注）すでに述べてきたように，ここで行と列（I と J ）を入れ替えても同じである．

5. 2 数値例による確認 —WordMiner と JMP スクリプトを用いた簡単な分析例—

ここで，2 つのソフトウェア（WordMiner，JMP）を用いて，実際にどのような分析が行え

るのか，いくつかの例により説明する．まず，表 36 に要約したデータ表が，ソフトでは実際にどのように出力されるかを例で示そう．たとえば WordMiner では，「多次元データ解析」のモジュールの中で，2 種の 2 元データ表「(回答・サンプル) × (構成要素変数)」および「(構成要素変数) × (質的変数)」を生成し分析する．どちらを用いるかは表 36 に挙げた組み合わせと分析目的に応じて指定する．

WordMiner には，専用の分かち書きツール[73]が組み入れられているので，これにより分かち書き処理が行われる．これを使いたくない，あるいは他の分かち書き処理や形態素解析を行った結果をインポートして用いてもよい[74]．

いずれにしても，基本的な構成は図 14 のように考えればよい．ここで，構成要素とは，前述のように，単語，語句，キーワードなど，“分析で扱う単位”のことをいう．すなわち「(回答・サンプル) × (構成要素)」あるいは「(質的変数・クラスター変数) × (構成要素)」の 2 元データ表が基本となる．また，すでに例でもみたように，この形式に当てはめること（読み替えること）ができるデータ表はすべて解析対象とできる．

5. 2. 1 ウェブ調査の例 —テキスト型データのデータ表の生成 —

ここで，ある調査でえた自由回答質問のテキスト型データを例として説明しよう．この例は，調査方式（調査モード）としてウェブ調査を用いて，あるウェブ・パネルを標本抽出枠として行った調査で得られたデータセットである．まず，ここで取り上げた質問文を示し，つぎに実際の回収データセットの一部をみる．そしてこの自由回答質問の実際の回答データ，それの分かち書き結果，実際に得られる 2 元データ表の形と内容，…と順をおって調べよう．なおここでは，対応分析法を適用する“2 元データ表”がどのように生成されるかを示す．これを用いた分析内容は，うしろの 5．4．5 節であらためて取り上げる．

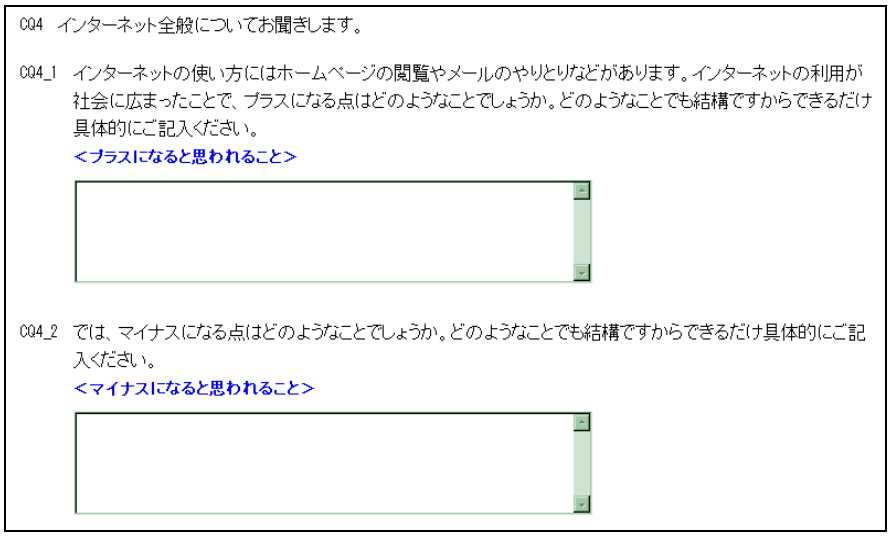
［調査の概要］

調査課題名	普段の生活やインターネットなどについて
調査対象地域	首都 40km 圏及び近畿 20km 圏
調査対象者 （標本抽出枠相当）	あるウェブ・パネル（非公募型[75]）に登録の首都 40 ㎞圏・近畿 20 ㎞圏に在住の 12 歳以上 65 歳未満の男女（パネル構成の詳細は省く）
調査方式	ウェブ調査
計画標本の大きさ	857 人
有効回収標本の大きさ	529 人
参加率[76]	61.7%
主な調査内容	生活意識 （普段の生活やインターネットなどについて） ①普段の生活での気持ち ②政治，政党支持，革新か保守か ③社会の移り変わり・IT の進歩 ④インターネットの利用状況 ⑤自分の性格やものの見方 ⑥商品，ブランド（お茶飲料）について ⑦情報源への接触や情報の内容 ⑧人口統計学的変数 （性別，年齢，職業，インターネット利用環境などの基本属性）
調査実施期間	2005 年 3 月 9 日～2005 年 3 月 16 日

73 富士通エフ・アイ・ピー株式会社（FIPS）開発の Happinedss を利用．
74 たとえば，茶筌や KH Coder などの出力を用いることもできる．
75 ウェブ・パネルをなるべく確率的パネルに近づけるよう配慮して設計した場合をいう．これに対して，一般のウェブ・パネルは，公募型のいわゆるボランティア・パネルである．
76 ウェブ調査では，厳密な意味での回収率を定義できない．そこで，用いた計画標本の大きさに対する回収標本の大きさの占める割合を「参加率」(participation rate) として用いる．

この調査で用いた多数の質問文の中から，つぎにあげる 2 つの質問を取り上げる．ここでは，「インターネットの利用が社会に広まったこと」についての「プラスになると思われる点」と「マイナスになると思われる点」と分けて尋ねていること，調査時点が 2005 年であることに注意しよう．

［ここで用いる自由回答質問］



① 構成要素変数つまり語句群の変数の生成

簡単な例として表 37 を挙げる．これは，このウェブ調査で得た「(回答・サンプル) × (多変量項目)」のデータ表の一部である．引用した質問は上にあげた 2 つの自由回答質問（Q4－1，Q4－2）を用いた．

ここでまず，元の自由回答データ（自由回答原文）から，分かち書き処理機能で「分かち書き」「キーワード」の 2 種の変数を作る．たとえば，分かち書き処理の例が表 38 にある．ここで記号「▲」は，実際には「半角空白」が入っている．これがここでいう構成要素を分析単位とする**構成要素変数**（つまり質的データ）となる．また，表 39 は，さらにキーワードに絞り込んで抽出した場合の例である．分かち書き，キーワードとも，表 37 にある原文と比べてみるとよい．ここで，明らかに**“情報の変換操作”**があることを忘れてはならない．

② 「(回答・サンプル) × (構成要素)」のデータ表の例

個々のサンプル（回答者）が自由回答として記述した語句（の一部）である構成要素を，図 14 のイメージに合わせて 2 元データ表として構築する．これで得られるデータ表の例が表 40 である．これが，**「(回答・サンプル) × (構成要素)」**の場合の解析対象のデータ表となる．

通常は，回答者数もある程度の大きさであり，また構成要素は用いる語句数によっては，かなり大きな数となる．ある回答者が 1 回しか用いなかった語句も含めすべての語句を拾い出すと，ちょっとした回答者数でも，数千から数万になることは日常茶飯である．そこで必要に応じて，ある語句の出現度数をどこかで切り捨てるという操作を行うことが多い[77]．図 15 に構成要素の分布の例をあげた(左側がここで取り上げたウェブ調査データから得た情報)．このように，かならず出現頻度 1 がもっとも多い**指数的に低減**する分布となる[78]．

③ (構成要素) × (質的変数)」のデータ表の例

次に，表 41 のような例をみよう．同じウェブ調査の質問で得たデータ表から一部を切り出

---

77 ある“閾値”を設けて出現語句数を調整する．自由回答だけでなく，ツイッターやブログなどで集めたデータの多くは出現頻度 1 が圧倒的に多い（峰のない典型的な**ロングテイル**な分布）．よって，大抵は“特徴語句をスクリーニングした”として，出現頻度・利用頻度の多い語句を拾い出す．こうしたフィルタリング処理がどのように行われているかを，ソフトの中で明確に示すべきであるが，たいていは曖昧である．

78 ジフの法則（Zipf’s law；ジフ分布）あるいはパレートの法則（Pareto’s law；パレート分布）などで説明されることが多い．

したものである．行側を「サンプル（回答者)」とし，列側に，自由回答質問（Q4－1，Q4－2 の 2 問)，選択肢型質問，人口統計学的変数である「性別」「年齢区分」「性年齢区分」「未既婚」などがある．

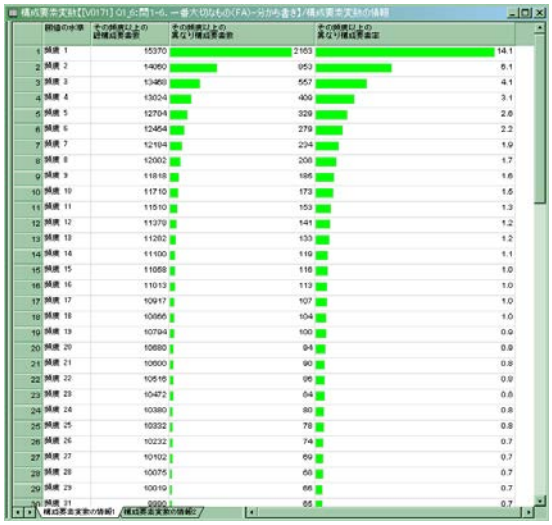


図 15　構成要素(語句)の頻度分布(2 例)

また「年齢区分」とこの分かち書きでえた「構成要素（分かち書き)」を（WordMiner の多次元データ解析の）オプションとして指示すると表 41 のような 2 元データ表が得られ，これが解析対象のデータ表となる．ここでは，ある閾値以上の出現頻度のキーワードを選び，その出現頻度の行和の大きさでソート（降順）してある．こうすると出現した構成要素と属性の間の度数の傾向を観察しやすい（WordMiner にはそのような機能もある)．

さて，ここまでの説明で，対応分析法が何を行い，どのような情報を提供してくれるのか，おおよそのことを説明した．また，テキスト・マイニングへの適用も違和感なく対応できることもみえるであろう．ここで総仕上げとして，人工データ，実際データを用いて，具体的なデータ探査の手順をおってみよう．ここでは，以下のデータセットを用いる．

分析例 1：人工データ

前にみた表 15 のデータ表の寸法を大きくしたトイ・データ．これを**“追加処理”**（supplementary treatment）の簡単な例示としても扱う．

分析例 2：好みの清涼飲料水

対応分析法が扱うデータ表の例とした表 17 にある「好みの清涼飲料水」の例を取り上げる．

分析例 3：ウェブ調査による調査データ

あるウェブ・パネルを対象に行った「情報に関する調査」でえた集約化データ表の例．調査設計時に事後の分析手法として対応分析法の利用を想定して集めたデータである．

分析例 4：ウェブ調査による意識調査データ

自由回答質問の例として，上の 5．2．1 節で述べた「普段の生活やインターネットなどについてのおうかがい」調査を用いる．この自由回答データの一部を用いて，実際に対応分析法で分析を行ってみる．

分析例 1，分析例 2 は，数量化法 III 類の説明で用いられる典型的な 2 値型データ表（インシデント行列）である．分析例 3 は，一般に集約データとしてよくみられるデータ表である．分析例 4 は，対応分析法がかなり有効に機能する自由回答質問によるテキスト型データの解析，つまりテキスト･マイニングの小さな応用例である．なお以下の分析では，主に WordMiner と JMP スクリプトを用いる．また，データ編集などで，エクセルを一部利用する．

表 37 ウェブ調査のデータ表の例(データセットの一部を切り出し)

	自由回答質問		選択肢型質問(質的変数)	人口統計学的変数				
ｻﾝﾌﾟﾙ ID	CQ4-1. ｲﾝﾀｰﾈｯﾄのﾌﾟﾗｽ点	CQ4-2. ｲﾝﾀｰﾈｯﾄのﾏｲﾅｽ点	AQ1.現在の生活満足度	HQ1. 性別	HQ2. 年齢(5 歳区分)	HQ2. 年齢(10 歳区分)	HQ3. 未既婚	HQ5. 職業(中区分)
2	調べたい事柄が、すぐに調べられる。//また、海外へも、メールですぐに連絡がとれる。/電話より安い。	不必要な情報を、判断力のない子ども達が得てしまう危険がある。	2.一応満足している	2.女性	40 ～ 44 歳	40 ～ 49 歳	2.結婚している	04.家業手伝い
3	情報がすぐに手に入る。世界が縮まる。	感情がなくなると思う。	3.やや不満である	1.男性	20 ～ 24 歳	20 ～ 29 歳	1.未婚	02.会社員,公務員など(管理職.事務職,営業・販売)
4	色々な情報を、得られる。	相手に情報を与えてしまいそう。	2.一応満足している	1.男性	45 ～ 49 歳	40 ～ 49 歳	2.結婚している	01.事業主,役員,自由業,専門職
5	真実の情報がリアルタイムに地域差が無く知ることができる。	情報量が多すぎて、真実を見極めるのに手間取る。	2.一応満足している	1.男性	40 ～ 44 歳	40 ～ 49 歳	2.結婚している	01.事業主,役員,自由業,専門職
8	メールは相手の時間や状況を気にせず、こちらの意思を手早く伝える事が出来る。郵便と異なり写真も簡単に(写真やさんで現像しなくても)送れる。/インターネットは自宅にいながら買い物をしたり、さまざまな情報を入手できる。	メールで気軽に会話していると、人と対面して話をする際の緊張感が薄れてしまう。人とのつながりが希薄になる。/インターネットはいらない情報もたくさんあるので、本当に必要な情報を見極めることが大切だと思う。	2.一応満足している	2.女性	30 ～ 34 歳	30 ～ 39 歳	2.結婚している	07.専業主婦
16	物事を広く知る事ができる。	人との接触が希薄になる。	2.一応満足している	2.女性	60 ～ 64 歳	60 ～ 69 歳	3.離婚または死別し、現在は独身	08.無職
19	家に居ながら仕事、買い物ができる。/知りたい情報がすぐに調べられる。	目に悪い。	2.一応満足している	2.女性	20 ～ 24 歳	20 ～ 29 歳	1.未婚	05.パート・アルバイト
20	いろんな情報が得やすくなった	問題あるサイトが教育上悪い影響を与えている ほんの一部ではあるが	1.十分満足している	1.男性	55 ～ 59 歳	50 ～ 59 歳	2.結婚している	01.事業主,役員,自由業,専門職
23	出向かずに居ながらにして短時間で、意思の疎通ができる。	言葉足らずで、誤解が生じやすい。	2.一応満足している	2.女性	50 ～ 54 歳	50 ～ 59 歳	2.結婚している	05.パート・アルバイト
24	取引などわざわざ、その場に行かなくてもメールでやり取りができる。	相手が画面となってしまい、人との交流がすくなくなると思う。	3.やや不満である	2.女性	20 ～ 24 歳	20 ～ 29 歳	1.未婚	08.無職
25	情報の伝達が早くなった。	プライバシーが侵されつつある。	2.一応満足している	1.男性	50 ～ 54 歳	50 ～ 59 歳	2.結婚している	01.事業主,役員,自由業,専門職

28	欲しいと思った情報が簡単に手に入る。	個人情報などが漏れたり、悪用されたりする危険性が増えた。	2.一応満足している	1.男性	35 ～ 39 歳	30 ～ 39 歳	2.結婚している	02.会社員，公務員など（管理職．事務職，営業・販売）
29	知りたい情報がすぐに調べられる。	個人情報の流出や、ネット犯罪の増加。/外出が減ること。	2.一応満足している	2.女性	30 ～ 34 歳	30 ～ 39 歳	2.結婚している	04.家業手伝い
31	日常生活の中で旅行、本、レストラン、等情報が容易に且つ早く掴めるようになった。/予約についても、インターネットでの申し込みは割引があり、便利になった。	メールは時間の制約がなく、利点ともいえるが、生の声ではない為相手の感情がつかめず、声の持つ特徴や色といった感覚が衰えると思う。/現在は子供も色々な情報にアクセスでき、年齢以上の悪い知識も得てしまう。	2.一応満足している	2.女性	60 ～ 64 歳	60 ～ 69 歳	4.無回答	07.専業主婦
33	知りたいことがすぐに調べられるし、その答えがすぐわかる。/電話など面倒なことが文字を打つだけですぐに連絡が出来る。	特に思い浮かばない。	3.やや不満である	2.女性	15 ～ 19 歳	12 ～ 19 歳	1.未婚	06.学生
34	知りたい事が瞬時に検索でき、自己満足が得られる。興味のある事をとことん調べるこニが出来る、パソコンで出来ないことはない	便利すぎて外に出かける機会が減った。運動不足になる。	4.きわめて不満である	1.男性	55 ～ 59 歳	50 ～ 59 歳	2.結婚している	02.会社員，公務員など（技術職．製造，労務）
35	電話と違い、お互い時間を気にせず、メールにてやりとりできる。/時間がない時に家にいて、買い物もできる点。	メールでのやりとりがしやすい反面、人とのコミュニケーションが本当の意味で薄くなる。//	2.一応満足している	2.女性	35 ～ 39 歳	30 ～ 39 歳	2.結婚している	07.専業主婦
36	調べ物に役立つ。/メールでコミュニケーションが容易にとれる。	不必要な情報が入ってくる。無駄な時間を費やしがち。	1.十分満足している	2.女性	35 ～ 39 歳	30 ～ 39 歳	2.結婚している	07.専業主婦
37	ホームページで事業のアピールや仕事の受注/	個人情報の流出	2.一応満足している	1.男性	35 ～ 39 歳	30 ～ 39 歳	2.結婚している	02.会社員，公務員など（管理職．事務職，営業・販売）
38	知りたいことが簡単に早急に調べられ、面識のない人との交流等、個人の行動範囲以外の所まで手が届くこと。	最近ニュースで多い自殺サイトで知り合ったとみられる人々の集団自殺や個人情報の流出等、悪用しようとする側にも便利なこと。	3.やや不満である	2.女性	35 ～ 39 歳	30 ～ 39 歳	2.結婚している	07.専業主婦

表 38　2 つの質問（Q4－1，Q4－2）の分かち書きの例　※実際は「▲」は空白（半角スペース）が入る

	自由回答の分かち書き	
ｻﾝﾌﾟﾙ ID	CQ4-1.ｲﾝﾀｰﾈｯﾄのﾌﾟﾗｽ点	CQ4-2.ｲﾝﾀｰﾈｯﾄのﾏｲﾅｽ点
2	調べたい▲事柄▲が▲、▲すぐに▲調べられる▲。▲／▲／▲また▲、▲海外▲へ▲も▲、▲メール▲で▲すぐに▲連絡▲が▲とれる▲。▲／▲電話▲より▲安い▲。	不必要▲な▲情報▲を▲、▲判断力▲の▲ない▲子ども▲達▲が▲得て▲しまう▲危険▲が▲ある▲。
3	情報▲が▲すぐに▲手▲に▲入る▲。▲世界▲が▲縮まる▲。	感情▲が▲なくなる▲と▲思う▲。
4	色々▲な▲情報▲を▲、▲得られる▲。	相手▲に▲情報▲を▲与えて▲しまい▲そう▲。
5	真実▲の▲情報▲が▲リアルタイム▲に▲地域差▲が▲無く▲知る▲こと▲が▲できる▲。	情報量▲が▲多すぎて▲、▲真実▲を▲見極める▲のに▲手間取る▲。
8	メール▲は▲相手▲の▲時間▲や▲状況▲を▲気▲にせず▲、▲こちら▲の▲意思▲を▲手早く▲伝える▲事▲が▲出来る▲。▲郵便▲と▲異なり▲写真▲も▲簡単▲に▲（▲写真▲や▲さん▲で▲現像▲しなくても▲）▲送れる▲。▲／▲インターネット▲は▲自宅▲に▲いながら▲買い物▲を▲したり▲、▲さまざま▲な▲情報▲を▲入手▲できる▲。	メール▲で▲気軽▲に▲会話▲して▲いる▲と▲、▲人▲と▲対面▲して▲話▲を▲する▲際▲の▲緊張感▲が▲薄れて▲しまう▲。▲人▲と▲の▲つながり▲が▲希薄▲に▲なる▲。▲／▲インターネット▲はいらない▲情報▲も▲たくさん▲ある▲ので▲、▲本当▲に▲必要▲な▲情報▲を▲見極める▲こと▲が▲大切▲だ▲と▲思う▲。
16	物事▲を▲広く▲知る▲事▲が▲できる▲。	人▲と▲の▲接触▲が▲希薄▲に▲なる▲。
19	家▲に▲居▲ながら▲仕事▲、▲買い物▲が▲できる▲。▲／▲知りたい▲情報▲が▲すぐに▲調べられる▲。	目▲に▲悪い▲。
20	いろんな▲情報▲が▲得▲やすく▲なった	問題▲ある▲サイト▲が▲教育上▲悪い▲影響▲を▲与えて▲いる▲ほんの▲一部▲では▲ある▲が
23	出向かず▲に▲居▲ながら▲に▲して▲短時間▲で▲、▲意思▲の▲疎通▲が▲できる▲。	言葉足らず▲で▲、▲誤解▲が▲生じ▲やすい▲。
24	取引▲など▲わざわざ▲、▲その▲場▲に▲行かなくても▲メール▲で▲やり▲取り▲が▲できる▲。	相手▲が▲画面▲と▲なって▲しまい▲、▲人▲と▲の▲交流▲が▲すく▲なくなる▲と▲思う▲。
25	情報▲の▲伝達▲が▲早く▲なった▲。	プライバシー▲が▲侵▲され▲つつ▲ある▲。
28	欲しい▲と▲思った▲情報▲が▲簡単▲に▲手▲に▲入る▲。	個人情報▲など▲が▲漏れたり▲、▲悪用▲されたり▲する▲危険性▲が▲増えた▲。
29	知りたい▲情報▲が▲すぐに▲調べられる▲。	個人情報▲の▲流出▲や▲、▲ネット犯罪▲の▲増加▲。▲／▲外出▲が▲減る▲こと▲。
31	日常生活▲の▲中▲で▲旅行▲、▲本▲、▲レストラン▲、▲等▲情報▲が▲容易▲に▲且つ▲早く▲掴める▲ように▲なった▲。▲／▲予約▲に▲ついて▲も▲、▲インターネット▲で▲の▲申し込み▲は▲割引▲が▲あり▲、▲便利▲に▲なった▲。	メール▲は▲時間▲の▲制約▲が▲なく▲、▲利点▲とも▲いえる▲が▲、▲生▲の▲声▲では▲ない▲為相手▲の▲感情▲が▲つかめず▲、▲声▲の▲持つ▲特徴▲や▲色▲と▲いった▲感覚▲が▲衰える▲と▲思う▲。▲／▲現在▲は▲子供▲も▲色々▲な▲情報▲に▲アクセス▲でき▲、▲年齢▲以上▲の▲悪い▲知識▲も▲得て▲しまう▲。
33	知りたい▲こと▲が▲すぐに▲調べられる▲し▲、▲その▲答え▲が▲すぐ▲わかる▲。▲／▲電話▲など▲面倒▲な▲こと▲が▲文字▲を▲打つ▲だけ▲で▲すぐに▲連絡▲が▲出来る▲。	特▲に▲思い浮かばない▲。
34	知りたい▲事▲が▲瞬時▲に▲検索▲でき▲、▲自己満足▲が▲得られる▲。▲興味▲の▲ある▲事▲を▲と▲ことん調べる▲こニ▲が▲出来る▲、▲パソコン▲で▲出来▲ない▲こと▲は▲ない	便利▲すぎて▲外▲に▲出かける▲機会▲が▲減った▲。▲運動不足▲に▲なる▲。
35	電話▲と▲違い▲、▲お互い▲時間▲を▲気▲に▲せず▲、▲メール▲にて▲やりとり▲できる▲。▲／▲時間▲が▲ない▲時▲に▲家▲に▲いて▲、▲買い物▲も▲できる▲点▲。	メール▲で▲の▲やりとり▲が▲しやすい▲反面▲、▲人▲と▲の▲コミュニケーション▲が▲本当▲の▲意味▲で▲薄くなる▲。▲／▲／
36	調べ物▲に▲役立つ▲。▲／▲メール▲で▲コミュニケーション▲が▲容易▲に▲とれる▲。	不必要▲な▲情報▲が▲入って▲くる▲。▲無駄▲な▲時間▲を▲費やしがち▲。
37	ホームページ▲で▲事業▲の▲アピール▲や▲仕事▲の▲受注▲／	個人情報▲の▲流出
38	知りたい▲こと▲が▲簡単▲に▲早急▲に▲調べられ▲、▲面識▲の▲ない▲人▲と▲の▲交流▲等▲、▲個人▲の▲行動範囲▲以外▲の▲所▲まで▲手▲が▲届く▲こと▲。	最近▲ニュース▲で▲多い▲自殺サイト▲で▲知り合った▲と▲みられる▲人々▲の▲集団自殺▲や▲個人情報▲の▲流出▲等▲、▲悪用しよう▲と▲する▲側▲に▲も▲便利▲な▲こと▲。

表 39　キーワード抽出の例　※空白（半角スペース）で区切られている

	自由回答分かち書きからキーワード抽出	
ｻﾝﾌﾟﾙ ID	CQ4-1.ｲﾝﾀｰﾈｯﾄのﾌﾟﾗｽ点／キーワード	CQ4-2.ｲﾝﾀｰﾈｯﾄのﾏｲﾅｽ点／キーワード
2	事柄 海外 メール 連絡 電話	不必要 情報 判断力 子ども 達 危険
3	情報 手 世界	感情
4	色々 情報	相手 情報
5	真実 情報 リアルタイム 地域差 無く	情報量 多すぎて 真実
8	メール 相手 時間 状況 気 こちら 意思 事 郵便 写真 簡単 現像 インターネット 自宅 買い物 情報 入手	メール 気軽 会話 人 対面 話 際 緊張感 つながり 希薄 インターネット 情報 たくさん 本当 必要 大切
16	物事 事	人 接触 希薄
19	家 居 仕事 買い物 情報	目
20	いろんな 情報 得	問題 サイト 教育上 影響 ほんの 一部
23	出向かず 居 短時間 意思 疎通	誤解
24	取引 わざわざ 場 メール やり	相手 画面 人 交流 すく
25	情報 伝達	プライバシー 侵
28	情報 簡単 手	個人情報 悪用 危険性
29	情報	個人情報 流出 ネット犯罪 増加 外出
31	日常生活 中 旅行 本 レストラン 等 情報 容易 且つ 予約 インターネット 申し込み 割引 便利	メール 時間 制約 利点 生 声 為相手 感情 特徴 色 感覚 現在 子供 色々 情報 アクセス 年齢 以上 知識
33	電話 面倒 文字 連絡	特
34	事 瞬時 検索 自己満足 興味 こニ パソコン 出来	便利 すぎて 外 機会 運動不足
35	電話 お互い 時間 気 メール にて やりとり 時 家 買い物 点	メール やりとり 反面 人 コミュニケーション 本当 意味
36	調べ物 メール コミュニケーション 容易	不必要 情報 無駄 時間
37	ホームページ 事業 アピール 仕事 受注	個人情報 流出
38	簡単 早急 面識 人 交流 等 個人 行動範囲 以外 所手	最近 ニュース 自殺サイト 人々 集団自殺 個人情報 流出等 側 便利

(*)ここで,「サンプル ID」は表 38 のそれに対応する. 各回答者の元の意見（の分かち書き）と比べて，意見の特徴，重要語句が強調されている.

表 40 （サンプル・回答者）×（構成要素）のデータ表（一部を切り出し） ※合わせて 3, 485 の語句（構成要素）がある

ｻﾝﾌﾟﾙ ID	行和（出現頻度）	インターネット	オークション	コミュニケーション	ショッピング	タイムリー	ニュース	ネット	ホームページ	メール	意見	閲覧	遠く	遠くに	遠方	何	何でも	価格	可能	家	海外	外出	楽
行和	**3485**	20	4	15	4	4	4	8	10	47	12	5	8	4	5	4	4	4	11	33	7	6	8
482	29	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	3
49	27	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
371	27	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
453	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
435	26	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
357	25	0	0	0	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	1
430	25	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
235	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
446	22	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
124	20	0	0	1	0	0	0	0	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0
172	20	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	19	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
46	18	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0
280	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
303	18	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
304	18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
448	18	0	0	0	0	0	0	0	0	2	0	0	1	1	0	0	0	0	0	0	0	0	0
118	17	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
168	17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
198	17	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
205	17	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
309	17	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
22	16	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0
273	16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
361	16	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
377	16	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
444	16	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0
500	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

表 41 （構成要素）×（性年齢区分）のデータ表（寸法 229 語句×14 区分から一部を切り出し）

SEQ	構成要素	行和	01. 男性 12～19 歳	02. 男性 20～29 歳	03. 男性 30～39 歳	04. 男性 40～49 歳	05. 男性 50～59 歳	06. 男性 60 歳～69 歳	08.男性, 無回答	09. 女性 12～19 歳	10. 女性 20～29 歳	11. 女性 30～39 歳	12. 女性 40～49 歳	13. 女性 50～59 歳	14. 女性 60～69 歳	無回答（性別不明）
142	情報	300	5	28	43	29	17	27	0	7	18	64	36	17	9	0
32	できる	232	5	19	13	20	14	14	0	5	33	62	29	14	4	0
133	出来る	76	0	3	6	2	5	2	0	6	9	21	8	8	6	0
113	事	74	1	3	4	3	7	1	0	1	10	24	5	11	4	0
116	時間	72	1	2	5	5	3	3	0	0	2	30	15	5	1	0
79	簡単	61	1	6	3	2	6	6	0	1	7	21	5	2	1	0
175	知りたい	60	0	3	2	4	2	5	0	3	7	24	5	5	0	0
42	なった	59	2	3	11	4	3	3	0	1	4	16	5	4	3	0
150	人	57	2	5	2	1	1	1	0	5	14	15	8	2	1	0
22	すぐに	57	2	4	4	3	1	0	0	3	6	22	9	3	0	0
125	手	47	1	6	6	3	0	0	0	1	8	11	9	1	1	0
64	メール	47	1	1	4	2	1	0	0	1	7	16	4	7	3	0
200	入手	44	2	2	4	6	6	6	0	0	2	8	5	2	1	0
23	する	42	0	7	4	3	0	5	0	0	4	12	4	3	0	0
196	得られる	40	0	2	2	5	1	5	0	0	3	6	7	6	3	0
19	して	38	0	6	5	2	1	2	0	0	2	7	5	4	4	0
2	ある	35	0	5	3	3	2	1	0	0	6	6	6	1	2	0
213	便利	34	0	0	2	1	0	1	0	3	2	12	5	7	1	0
183	調べられる	34	0	1	1	0	2	1	0	2	3	19	4	1	0	0
74	家	33	0	2	3	1	0	1	0	1	4	15	1	4	1	0
53	ように	33	1	2	5	2	1	2	0	1	2	11	4	1	1	0
44	なる	33	1	1	3	4	6	4	0	2	4	5	2	1	0	0
12	いろいろな	31	1	2	0	1	1	2	0	1	3	13	5	2	0	0
198	入る	30	1	4	5	3	0	0	0	0	5	6	5	1	0	0
197	得る	30	0	3	4	2	1	3	0	0	0	9	2	4	2	0
119	自分	30	1	3	1	1	1	2	0	1	7	9	4	0	0	0
202	買い物	29	0	1	0	2	0	0	0	0	4	12	6	4	0	0
184	調べる	29	0	0	2	0	2	0	0	1	5	12	5	1	1	0
135	瞬時	27	0	2	0	2	3	1	0	0	1	4	9	5	0	0
115	時	27	0	2	2	2	2	0	0	0	1	12	4	2	0	0

5. 2. 2 分析例 1:トイ・データによる分析

ここでは，表 42 のミニチュアなトイ・データを用いて対応分析法を行う．ここで得られる各統計量，情報の解読を試みる．ここの目標は，データ表を意図的に構造化することで対応分析がどのように機能するかを体験的に知ることにある．データ表は表 42 のような構成であり，以下のような場面を想定している．

- 10 名の回答者（サンプル）に対して，ある商品の「好きな銘柄」を列記（自由記述）してもらうという場面を考える．表の「銘柄」欄がこれに相当する．
- 同じ調査を，時点を変えて調べた結果が「次年度調査の銘柄」欄にある．ここで識別のために先頭に「●」を付けた．
- 「銘柄」を問うときに，併せて「では，その選んだ銘柄のうちで一番好きなものを“ひとつだけ”選んでください」と質問して得られた結果が「一番好きな銘柄」欄にある．ここでも識別のため銘柄名の前に記号「◆」を付けた（布置図の観察の識別用のため）．
- この他，属性として「性別，年齢区分」も項目として用意した（性別に▼，年齢区分に★の識別記号も付けた）．

最近のソフトはこのような文字情報となった調査データをかなり自由に扱える[79]．エクセルやエディタを用いて上の形式のデータ表を事前に作成すればよい．

表 42 質的データとして表現したデータ表

回答者	銘柄	次年度調査の銘柄	一番好きな銘柄	性別	年齢区分
回答者 1	銘柄 B, 銘柄 E, 銘柄 F	●銘柄 E, ●銘柄 F	◆銘柄 B	▼男性	★30 代
回答者 2	銘柄 F	●銘柄 F, ●銘柄 B	◆銘柄 F	▼男性	★40 代
回答者 3	銘柄 C, 銘柄 F	●銘柄 F	◆銘柄 C	▼男性	★30 代
回答者 4	銘柄 B, 銘柄 C, 銘柄 E, 銘柄 F	●銘柄 C, ●銘柄 B	◆銘柄 E	▼男性	★30 代
回答者 5	銘柄 B, 銘柄 C, 銘柄 F	●銘柄 B, ●銘柄 C, ●銘柄 F	◆銘柄 C	▼男性	★30 代
回答者 6	銘柄 A, 銘柄 B, 銘柄 C, 銘柄 E	●銘柄 A, ●銘柄 B	◆銘柄 A	▼女性	★30 代
回答者 7	銘柄 A, 銘柄 B, 銘柄 D, 銘柄 E	●銘柄 D, ●銘柄 E	◆銘柄 B	▼女性	★20 代
回答者 8	銘柄 C, 銘柄 F	●銘柄 C, ●銘柄 F	◆銘柄 F	▼男性	★40 代
回答者 9	銘柄 A, 銘柄 B, 銘柄 E	●銘柄 B, ●銘柄 E	◆銘柄 E	▼女性	★30 代
回答者 10	銘柄 A, 銘柄 D, 銘柄 E	●銘柄 A, ●銘柄 E	◆銘柄 D	▼女性	★30 代

表 43 （回答者・サンプル）×（銘柄）のクロス表

銘柄 / サンプル	銘柄Ａ	銘柄Ｂ	銘柄Ｃ	銘柄Ｄ	銘柄Ｅ	銘柄Ｆ	行和
回答者 1	0	1	0	0	1	1	3
回答者 2	0	0	0	0	0	1	1
回答者 3	0	0	1	0	0	1	2
回答者 4	0	1	1	0	1	1	4
回答者 5	0	1	1	0	0	1	3
回答者 6	1	1	1	0	1	0	4
回答者 7	1	1	0	1	1	0	4
回答者 8	0	0	1	0	0	1	2
回答者 9	1	1	0	0	1	0	3
回答者 10	1	0	0	1	1	0	3
列和	4	6	5	2	6	6	29

実際にこのデータ表を WordMiner にインポートし分析を進める．またその得られた結果の一部をエクスポートし，それを JMP により再分析する．データ表入力のあと「銘柄」を構成要素変数に指定し分かち書き処理を行うと「分かち書き」と「キーワード」がそれぞれ構成

79 たとえば，WordMiner では扱える文字数の制限もない．JMP も文字情報の扱いにかなり自由度がある．

要素変数として生成される．

次に（回答者・サンプル）×（構成要素変数）のデータ表を指定し「多次元データ解析」を行うと表 44 のクロス表が得られ，これが対応分析の対象データ表となる．ここでは，データ表の寸法は，回答者数=10（名），銘柄＝6（A～F までの 6 選択肢）となる．

① 特異値・固有値と寄与率，累積寄与率

始めに特異値，固有値，寄与度，累積寄与度を観察する．この例では以下の値が得られた（表 44）．なおここで，固有値の個数は K=min{10, 6}－1=5 個のはずで，確かにそのようになっている．

表 44 固有値，寄与率の表

成分 k	特異値 α_k	固有値 λ_k	寄与率（%） ν_k	累積寄与率（%） $\sum_k \nu_k$
1	0.7912	0.6260	61.41	61.41
2	0.4332	0.1877	18.41	79.82
3	0.3667	0.1345	13.19	93.01
4	0.2126	0.0452	4.43	97.45
5	0.1612	0.0260	2.55	100.00

② 成分スコアの観察

ここで得られた成分スコア，寄与度（絶対寄与度，相対寄与度）他を一括して統計量の要約表とした．ここでは，このうちの成分スコア，寄与度を表 46 としてあげた．ここで，すでにいくつかの例でみたように，まずこの表にある第 1 成分スコアに注目する．この行（回答者）と列（銘柄）それぞれの第 1 成分スコアを大きさの順にソートして，つまり行と列とを入れ替えてみると表 45 が得られる．ここで，数値「1」（つまり「好む」として選んだ銘柄の度数）の並びが対角にきれいに並んでいるのが分かる（線形化されている）．しかしこれだけでは“何が**数量化**されたか”がよく見えない．

さらに一歩進めて，対応分析でえた固有値（あるいは特異値）がどのような意味を持つのかを調べる．これは次のような**双対散布図**を用意すると理解しやすい．「回答者」「銘柄」それぞれに与えられた第 1 成分スコアを，前に「都市環境の住みやすさ」の 2 つの質問の分析で表 14 のように行った手順で，ここでも成分スコアの並べ替えを行う．これで得られた双対散布図が図 16 である．同時にもとのクロス表の行（回答者）と列（銘柄）の並べ替えも行う．これが表 45 となる．つまり，「都市環境の住みやすさ」調査データとの違いは，出発行列がクロス表であるか，あるいはここでみたように“二値型データ表”（インシデント行列）であるかの違いだけである．しかしここで分析対象とした“二値型データ”は，クロス表内の出現頻度が「1」と考えたクロス表と考えれば，特別なことを行ったわけでないことがわかるであろう．

さてここで得た図 16 の意味・解釈は重要である．なんども繰り返すが，データ表の「回答者」，「銘柄」のいずれも質的データであることに注意しよう．もとの表の「回答者」という 10 の選択肢，「銘柄」という 6 つの選択肢は名義尺度であり，このままでは数量として扱えない．ここで，数量化 III 類の発想に従うと，かりにこれらの選択肢に新たな数量を付与して，これの相関係数を最大化するとした．これが**“数量化”**と言われる所以である．またこの方法が質的データの線形化となっていることも上でみたように分かる（表 45，図 16）．明らかに図 9 は表 35 に対応するものである．

ここで対応分析法を用いて得た成分スコアとしてこの数量を観察すると，もとの名目的なコード（選択肢に付与のコード）ではない，新たな点間の分布，距離関係が意味のある別の数量空間を作ったことになる．またこの成分スコアは，大小が意味を持ち区間尺度データとして演算（加減乗除）も可能な数値として扱える．一方，テキスト型データの分析をこの視点から行っているので，それを越えた情報取得には別の視点からのアプローチが必要である．

」

表 45　サンプルとと銘柄の第1成分スコアで並べ替えたデータ表

ID	銘柄D	銘柄A	銘柄E	銘柄B	銘柄C	銘柄F	行和	「回答者」の第１成分スコア
回答者 2	0	0	0	0	0	1	1	1.63912
回答者 3	0	0	0	0	1	1	2	1.45824
回答者 8	0	0	0	0	1	1	2	1.45824
回答者 5	0	0	0	1	1	1	3	0.91076
回答者 4	0	0	1	1	1	1	4	0.47840
回答者 1	0	0	1	1	0	1	3	0.21208
回答者 6	0	1	1	1	1	0	4	-0.3073
回答者 9	0	1	1	1	0	0	3	-0.8355
回答者 7	1	1	1	1	0	0	4	-1.1503
回答者 10	1	1	1	0	0	0	3	-1.4724
列和	2	4	6	6	5	6	29	
「銘柄」の第１成分スコア	-1.6574	-1.18980	-0.64770	-0.14570	1.01067	1.29691		

サンプルの番号	銘柄の成分スコア	サンプルの成分スコア
1	−0.513	0.168
1	−0.115	0.168
1	1.026	0.168
2	1.026	1.297
3	0.800	1.154
3	1.026	1.154
4	−0.513	0.379
4	−0.115	0.379
4	0.800	0.379
4	1.026	0.379
5	−0.115	0.721
5	0.800	0.721
5	1.026	0.721
6	−0.941	−0.243
6	−0.513	−0.243
6	−0.115	−0.243
6	0.800	−0.243
7	−1.311	−0.910
7	−0.941	−0.910
7	−0.513	−0.910
7	−0.115	−0.910
8	0.800	1.154
8	1.026	1.154
9	−0.941	−0.661
9	−0.513	−0.661
9	−0.115	−0.661
10	−1.311	−1.165
10	−0.941	−1.165
10	−0.513	−1.165

⇒

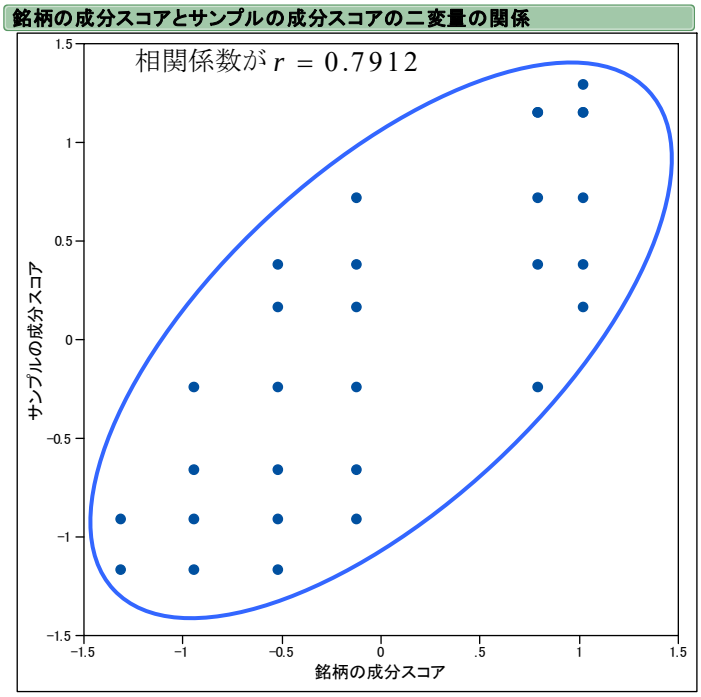


図 16　第 1 成分スコアの回答者と銘柄の散布図

表 46　成分スコアと寄与度の一覧

（ｉ）選択肢 $i(\in I)$，つまり「回答者（サンプル）」への成分スコアと絶対寄与度，相対寄与度

サンプル	成分スコア1	成分スコア2	成分スコア3	成分スコア4	成分スコア5	絶対寄与度1	絶対寄与度2	絶対寄与度3	絶対寄与度4	絶対寄与度5	相対寄与度1	相対寄与度2	相対寄与度3	相対寄与度4	相対寄与度5
サンプル1	0.168	0.297	-0.673	0.121	-0.165	0.465	4.852	34.839	3.337	10.868	0.046	0.144	0.741	0.024	0.045
サンプル2	1.297	-0.813	-1.090	-0.494	0.239	9.265	12.154	30.481	18.595	7.570	0.439	0.173	0.310	0.064	0.015
サンプル3	1.154	-0.393	0.399	-0.101	-0.057	14.665	5.669	8.174	1.556	0.855	0.803	0.093	0.096	0.006	0.002
サンプル4	0.379	0.230	-0.033	0.164	-0.212	3.157	3.870	0.109	8.156	23.857	0.533	0.196	0.004	0.100	0.168
サンプル5	0.721	0.157	0.117	0.278	0.290	8.581	1.357	1.060	17.632	33.294	0.723	0.034	0.019	0.107	0.117
サンプル6	-0.243	0.425	0.387	-0.173	-0.010	1.303	13.290	15.358	9.148	0.049	0.141	0.431	0.357	0.071	0.000
サンプル7	-0.910	-0.252	-0.045	0.220	0.146	18.252	4.650	0.211	14.775	11.345	0.860	0.066	0.002	0.050	0.022
サンプル8	1.154	-0.393	0.399	-0.101	-0.057	14.665	5.669	8.174	1.556	0.855	0.803	0.093	0.096	0.006	0.002
サンプル9	-0.661	0.558	-0.114	-0.328	0.105	7.222	17.146	0.994	24.638	4.362	0.497	0.354	0.015	0.122	0.013
サンプル10	-1.165	-0.754	0.088	-0.052	-0.132	22.426	31.345	0.599	0.607	6.947	0.695	0.291	0.004	0.001	0.009

（ⅱ）選択肢 $j(\in J)$，つまり「銘柄」への成分スコアと絶対寄与度，相対寄与度

銘柄	成分スコア1	成分スコア2	成分スコア3	成分スコア4	成分スコア5	絶対寄与度1	絶対寄与度2	絶対寄与度3	絶対寄与度4	絶対寄与度5	相対寄与度1	相対寄与度2	相対寄与度3	相対寄与度4	相対寄与度5
銘柄A	-0.941	-0.013	0.216	-0.391	0.169	19.525	0.013	4.765	46.708	15.196	0.795	0.000	0.042	0.137	0.026
銘柄B	-0.115	0.544	-0.164	0.220	0.159	0.440	32.648	4.121	22.148	19.954	0.033	0.723	0.065	0.118	0.061
銘柄C	0.800	0.012	0.693	0.062	-0.057	17.611	0.013	61.522	1.467	2.145	0.568	0.000	0.426	0.003	0.003
銘柄D	-1.311	-1.161	0.059	0.396	0.044	18.944	49.505	0.175	23.973	0.507	0.533	0.417	0.001	0.049	0.001
銘柄E	-0.513	0.194	-0.177	-0.038	-0.277	8.681	4.137	4.815	0.661	61.017	0.641	0.092	0.076	0.004	0.187
銘柄F	1.026	-0.352	-0.400	-0.105	0.039	34.799	13.685	24.600	5.043	1.183	0.780	0.092	0.119	0.008	0.001

(*) 以上の表における「絶対寄与度」や「相対寄与度」は，後述する．

（数学的な証明は省くが）ここで得た第 1 固有値 $\lambda_1 = 0.6260$ の正の平方根，つまり特異値は図 16 の双対散布図の相関係数に等しい．実際に左の表の成分スコアの数値から相関係数を求めると約「0.79121」となる．一方，特異値は，$\sqrt{\lambda_1} \doteqdot 0.7912$ であり両者は一致する．

なお，林の数量化法 III 類では，「このある数量が回答者と銘柄（の各選択肢）に付与できたとして，その相関を最大化する」という問題を考えている（尺度化の発想）．また，通常は，数量化法 III 類を適用すると，（成分スコアに対応する）得られる数量化得点の分散は「1」となるように標準化されている（ことが多い）．一方，対応分析法のアプローチに従い 2 元データ表の一つとして出発したとして分析を進めると，成分スコアの分散は固有値（λ_k）とした（表 35 参照）．しかしこれも，対応分析法で得た成分スコアを分散 1 となるように標準化を行えば，同じ結果となる．

ここで同じ結論を得たとはいえ，このスコアの処理の違いには注意せねばならない．とくに，同時布置図の上での距離の解釈や，得られたスコアを用いた二次分析，たとえばクラスター化などを行う際には要注意である．

いずれにしても，両者の考え方（定式化），理念･思想はまったく異なるのだが，行っている数理的な操作は同等と考えてよいことを意味している．時代も経緯も異なるものの，同じような結果となる分析手法が異なる国で誕生したことは興味あることである．

③布置図と同時布置図の観察

布置図，同時布置図をみよう．これは回答者と銘柄それぞれの選択肢に対する成分スコアを図に表せばよい．まず，布置図の横軸，縦軸に観察したい成分を選ぶ．たとえば，第 1 成分，第 2 成分を選べばその布置図となる．

（ｉ）回答者の成分スコアの布置図と銘柄の成分スコアの布置図

まず，「回答者」「銘柄」それぞれの布置図を描く（図 17）．ここではデフォルト値の横軸が 1 軸，縦軸を 2 軸とした（1, 2）軸について出力した．

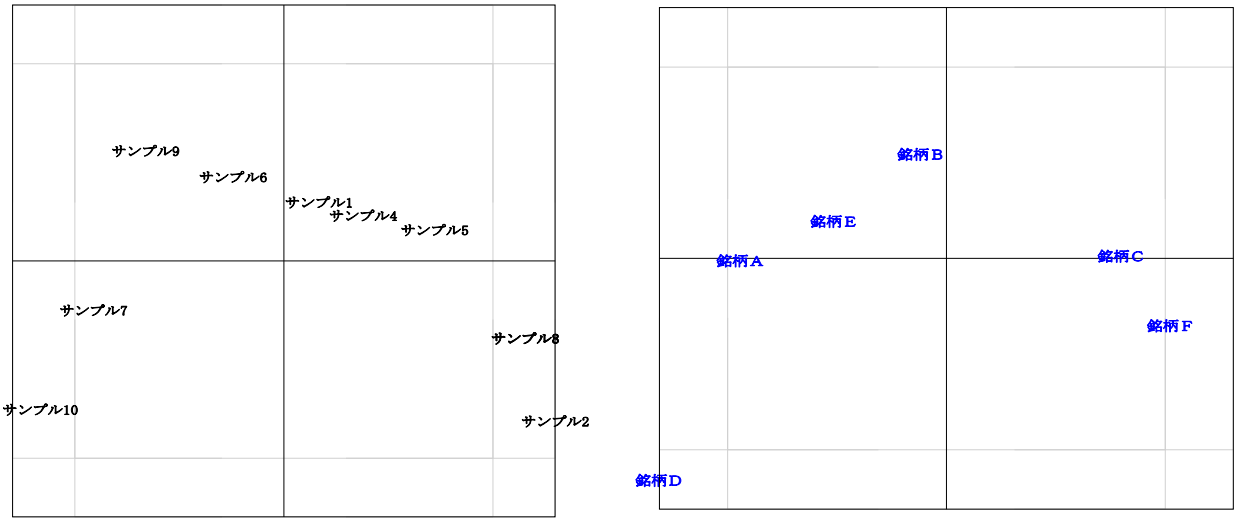


図 17　第 1, 第 2 成分スコアの布置図
（左が「回答者」, 右が「銘柄」に対応）

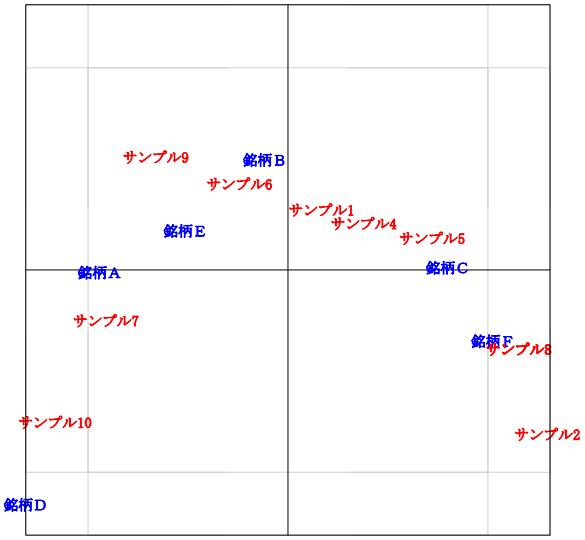


図 18 「回答者」と「銘柄」の同時布置図

（ⅱ）同時布置図

つぎに同時布置図を作ると図 18 が得られる．ここで確かに上の図 17 の 2 つの布置図を重ねた形となり，回答者と，銘柄の（各選択肢の）相互の関係が図 9 や表 35 でみたように対応していることが分かる．前に示したように 2 つの項目の関係はかなり高い相関関係にあり対応があることを示している．

この図のような放物線のような形状は“**馬蹄形効果**”（horseshoe effect）という．このような形状が見られた場合，「データ表の構造は線形的である」こと（主要な構造はおもに 1 次元で説明されること」を示唆している．この例でも，データ表の構造はかなり線形的であり，第 1 次元だけで十分に度数表の特徴をとらえている．このことを表 45 の並べ替えデータと図 16 が示している[80]．

④絶対寄与度と相対寄与度

ここで，**寄与度**（絶対寄与度，相対寄与度）について説明する．表 46 の絶対寄与度の欄は縦方向に和をとると「100」（%）となる．これは各点（行の選択肢）が，各次元にどの程度寄与しているかを表す指標で**“絶対寄与度”**という．たとえば，表 46 において，第 1 次元目の「回答者 10」の「絶対寄与度 1 」は，22.426（%）とある．また，「回答者 1」の絶対寄与度は，0.4653（%）である．つまり，「回答者 10」は第 1 次元を構成するのに，約 1/4 ほども寄与しているが，「回答者 1」はほとんど寄与していない．これを求めるには，[各次元の成分スコアの 2 乗]×度数を求め，つぎに次元ごとにこれらの和を求めて，最後にこれで割ればよい[81]．この絶対寄与度は，かりに各次元に何らかの名称を与えるとき（各成分軸が何かを表していることを解釈するとき）に参考にするとよい．

たとえば，成分 1 の列方向に第 1 成分に対応する「絶対寄与度 1 」をみると，回答者 7，10，8，3 等の値が他に比べて大きい（回答者 3 と 8 は同じ回答パターンであるから成分スコアが同値で点が重なる）．しかし第 2 成分をみると，回答者 9，10，それにつづき 2，6 などの値がやや大きい．回答者 1，2 はむしろ第 3 成分内での寄与が高い（とくに，回答者 2 に注意，図では第 1 軸の右端に位置するが回答者 8 ほど第 1 成分内での寄与は高くない）．

---

[80]馬蹄型効果の数理的証明は複雑である．エルミート直交多項式などの知識が必要とされる．岩坪［2］にも詳しい説明がある．

[81] 式による記述と説明は「第Ⅱ部」にある．

次に，“**相対寄与度”（平方相関）**について説明する．

「相対的寄与度」は，ここでは横方向（行方向）の和が「１」（100%）となるような指標となっている．これは，点が各次元によって，どれほど良く**近似されているか**を示す指標である．たとえば，「回答者 7」の第 1 次元における値は，85.97%となっている．これは，第 1 次元だけで，「回答者 7」の変動の 85.97%を説明できることを意味している．

「銘柄」の側も同様の観察を行えばよい．第１成分の中に占める「絶対寄与度」が高いのは銘柄 F がかなり大きく，34.80%である．続いて銘柄 A，C，D などがあるがこれらの値は，だいたい 17%～20%程度である．これが小さい銘柄 B は成分１への寄与は少なく，成分２や４に関係し，銘柄 C は成分 3，銘柄 D は成分 2，また銘柄 E は成分５でそれぞれ寄与が高い．

実用においては，先ほどの「回答者」に対する結果と，これらの特徴を踏まえて，まずはじめの 2 成分である第１軸と，第 2 軸を解釈すればよい．

また「相対寄与度」は，ここでも銘柄 B を除いてすべての銘柄は第１成分によって，50%～80%は近似できていることを示している．

ここで留意すべきことは，2 次元上に描かれた散布図の情報は“そこでいま眺めている 2 つの成分軸へ射影した図”であり，**全情報ではない**，ということである．よって，布置図の観察に併せて，“どの点がどの成分にどれだけ寄与しているか？”，そして，各点が“各成分によってどれほど近似されているか？”をこれらの寄与度から知ることが必要である．

⑤ 追加処理あるいは追加要素

次のような場面で，“**追加処理”**（supplementary treatment）を行うと効果的な場合がある[82]．追加処理の対象とする項目（変量），選択肢を**“追加要素”**（supplementary elements）という．ここではどのような場面での適用が考えられるかを述べる．さらに関心のある読者は参考文献にある書を参照していただきたい[83]．

i)　はずれ値の一時除去と再配置を行いたいとき

・一時的に除去した「はずれ値」を一旦除外して再分析した元のデータ表の中に再布置する[84]．
・そしてはずれ値を含まないデータ構造からみたはずれ値の影響を知る．

ii)　判別分析的，グループ間類似や差異を見る

- 層別変数や属性などで，複数のグループに分けられるデータセットを，層やグループ単位で「追加処理」する．
- たとえば「男性グループ」のデータ表があって，この構造から始めの分析を行い，次に「女性グループ」は相対的にどこに位置するかを知りたい（男女一緒に分析したこととは内容が異なる）．
- あるブランドの認知度・好感度をある年度に調べ，別の年度で再度同じような調査を行ったとする．このとき，両年度間に類似・差異があるかを，ある年度の方向から知りたい（たとえば，X 年度の分析結果・構造からみた Y 年度の位置づけ）．

iii)　データ表の，行の追加と列の追加，あるいは，一時除去と再配置を行うこと

- 「(回答）×（構成要素)」に，別の構成要素変数を追加して違いをみる
- 「(構成要素）×（質的変数)」に，構成要素群を追加，あるいは別の質的変数を追加

ここでそのすべてを示すことは困難だが，上の人工データ例を使ってその仕組みを簡単に検証する．

---

82 追加処理の手順の説明については．「第Ⅱ部」を参照．

83 たとえば，ベンゼクリ（1976，1980），大隅・ルバールほか（1994）

84 出現頻度の不釣り合いな部分を一時除去する，サブセット分析（subset analysis）という方法もある．たとえば，Greenacre（2007）を参照．

[追加処理の例]

ここで用意した表 42 のデータ表について，追加処理の例を考える．いま次のような場面を想定しよう．

i) 表 42 で，上のように「(回答者) × (銘柄)」の関係を分析し，相互の関係を上にみたように知ったとする．
ii) この（架空の）調査で，選んだ銘柄のうち，さらに「一番好きな銘柄」を選んでもらったとする．
iii) このとき，もとの「(回答者) × (銘柄)」のデータ構造（関係）からみて，この「一番好きな銘柄」はそれぞれどのような関係にあるか，どこにポジショニングできるかを知りたい．これがここでの第 1 の課題である．
iv) さらに，属性情報も取得してあるので（表 42 の性別，年齢区分），これが「(回答者) × (銘柄)」の関係からみてどこに位置するかを知りたいとする．これを次の課題とする．

ここで注意することは，追加処理を考えるということは，すでに何らかのデータ表についての吟味分析がある程度進んだ中で（元のデータ表の構造は保持したまま），さらに，求めた空間のなかに別の項目を射影したい場面で利用することにある．あるいは，上に列記したように，データ表の特定の行あるいは列の挙動が不自然，たとえばはずれ値があるなどの現象が見られたとき，この該当する行（あるいは列）を一時除去することで，その影響を除去できる．しかし後になって，やはり除外した行（あるいは列）のプロファイルが，求めた空間でどこに位置するのかを知りたいということもある．こうした場面で追加処理を用いる．

なお，「行プロファイル」「列プロファイル」の両者の追加処理がありうる．

さて，ここでみる例題を再度整理しよう．

i) 表 42 のデータ表について「(回答者) × (銘柄)」についての分析結果はすでに得た．
ii) たとえば，このときの追加処理を「構成要素変数」として行うか，「質的変数」として行うかの 2 つのオプションがある．
iii) ここで「一番好きな銘柄」を構成要素変数として追加処理することを考える．このとき，表 43 で得たクロス表の右側に「一番好きな銘柄」を (0, 1) 化して並置した行列を加えたことを考えればよい．

まず，構成要素変数を追加処理した結果を示そう．つまり，「(回答者) × (銘柄)」の構造に対して「一番好きな銘柄」という追加構成要素変数は（相対的に）どこに位置するかを知るということである．ここでは，

- まず，もとの「(回答者) × (銘柄)」で得た同時布置図を出力し，
- 次にここで追加処理した構成要素変数「一番好きな銘柄」をさらに重ねて布置する，

とした（ここは WordMiner を利用）．この図だけでなく，それぞれ個別の布置図を観察することもあり得る．実際に追加処理によ計算した結果は表 47 の成分スコアと「相対寄与度」が得られる（なお，「絶対寄与度」は，算出する意味があまりないので出力していない）．なおここで元の銘柄と区別するため追加処理とした銘柄には「◆」を付けた．こうして得られた追加処理による成分スコアと，元の成分スコアの同時布置図を描くと図 17 が得られる．

表 47 「一番好きな銘柄」の追加処理で得た成分スコアと相対寄与度の一覧

一番好きな銘柄	成分スコア1	成分スコア2	成分スコア3	成分スコア4	成分スコア5	相対寄与度1	相対寄与度2	相対寄与度3	相対寄与度4	相対寄与度5
◆銘柄A	-0.307	0.982	1.055	-0.814	-0.059	0.011	0.107	0.124	0.074	0.000
◆銘柄B	-0.469	0.052	-0.979	0.802	-0.059	0.055	0.001	0.240	0.161	0.001
◆銘柄C	1.185	-0.272	0.704	0.415	0.721	0.351	0.019	0.124	0.043	0.130
◆銘柄D	-1.472	-1.741	0.241	-0.242	-0.820	0.241	0.337	0.006	0.007	0.075
◆銘柄E	-0.179	0.909	-0.199	-0.387	-0.333	0.008	0.206	0.010	0.038	0.028
◆銘柄F	1.549	-1.392	-0.942	-1.399	0.565	0.600	0.484	0.222	0.489	0.080

ここでも，図と各成分の成分スコア，相対寄与度を併せて観察すると，ほとんど説明の必要がないくらい，追加処理とした「一番好きな銘柄」の関係（位置づけ，ポジショニング）が分かるであろう（図 17，図 18 と比べて見ると追加処理の効果が分かる）．つまり，このような使いかたが可能ということである．

次にここで「性別と年齢区分」の属性変数（質的データ）があるので，これも併せて追加処理としてみよう．結果は図 18 のようになった．ここでは，男性と女性が左右に分かれ，また年齢区分と銘柄の間に，ある種の対応関係があることが見えるであろう．

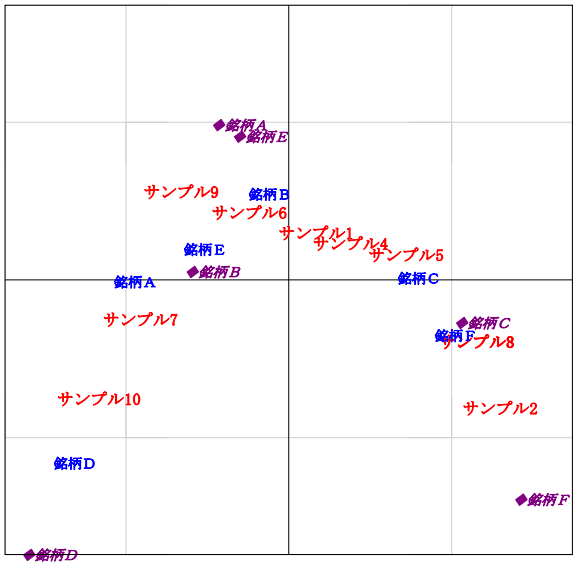


図 17 「一番好きな銘柄」の追加処理で得た布置図

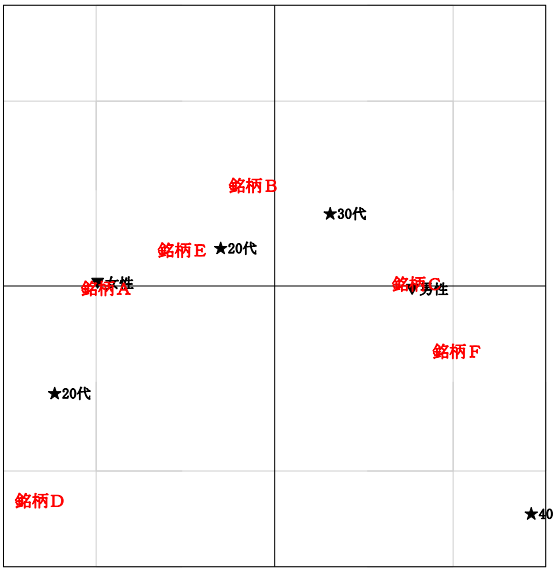


図 18 「性別」「年齢区分」の追加処理で得た布置

ここで，性別や年齢区分を質的変数の追加処理要素として用いたが，これは当然（性別）×（銘柄），（年齢区分）×（銘柄）といったデータ表から直接，計算した対応分析とは異なる分析であることに注意しよう（つまり分析目的が異なる）．

分析を行うにあたって，データ表の行と列の意味内容（対応関係）の何を知りたいかという分析方針や，自分の分析目的に合ったデータ収集方式と事後のデータ表の作成方法とを事前に考えて取り組むことが重要である．

5. 3. 2 分析例 2:好みの清涼飲料水

この例は（表 17），分析例 1 と同様に，2 値型のデータ表である．表側は回答者，表頭は清涼飲料水の銘柄となっている．回答者は提示された銘柄から自分の好むものを選べば「1」とし，それ以外は「0」とした**インシデント行列**となっている．数量化法 III 類では典型的な例であるが，対応分析法では，すでに述べたようにクロス表の 1 種と考えればよい．これで得られた結果を順に示そう．

① 固有値，特異値，寄与率など

まず，固有値，特異値，寄与率，累積寄与率を表 48 に一覧とした．

この 2 元データ表の寸法は $m=30, n=8$ であった．よって，固有値は $K=\min\{m,n\}-1=7$ 成分までである．このうち，始めの 2 成分で総変動に占める寄与率は約 61%，3 成分まで考えると約 74%占める．よって，始めの 3 成分を観察すれば情報の大半は解釈できそうである．また，特異値をみると，どの成分についてもかなりの相関があることがわかる．

表 48　固有値，特異値，寄与率など

成分 k	特異値 α_k	固有値 λ_k	寄与率(%) ν_k	累積寄与率(%) $\eta_k = \sum_k \lambda_k$
1	0.86191	0.74289	43.8	43.8
2	0.54406	0.296	17.4	61.2
3	0.45672	0.2086	12.3	73.5
4	0.42654	0.18194	10.7	84.2
5	0.39113	0.15298	9.0	93.2
6	0.24708	0.06105	3.6	96.8
7	0.233	0.05429	3.2	100
固有値総和 (総変動)	—	1.697747	—	—

②　成分スコアの布置図，同時布置図

ではここで布置図を描いてみる．上の情報から，はじめの 3 成分あたりの観察で十分にもみえるので，まず第 1，第 2 成分スコアについて布置図を描いてみる．回答者側の布置図が図 19，銘柄側の布置図が図 20 である．とくに銘柄側の布置図では，8 種の清涼飲料がおよそ 3 つのグループ{セブンアップ，スプライト}，{コカコーラ，ペプシコーラ}，{7 アップ，ダイエットコーク，ダイエットペプシ，ダイエット 7 アップ}に分かれることがわかる．とくにダイエット系が 1 つのグループを構成している．

これに対して，30 名の回答者はどう対応しているだろうか．これをみるためには図 21 の同時布置図を観察する．併せて，表 49，表 50 に示した寄与度（相対寄与度，絶対寄与度）も観察する．はじめの 2 成分に注目しているならば，対応する成分の寄与度を調べる．表 49，表 50 には，特徴的な回答者と銘柄にアミカケを付けたので，この数値と布置図とを比較してみよう．

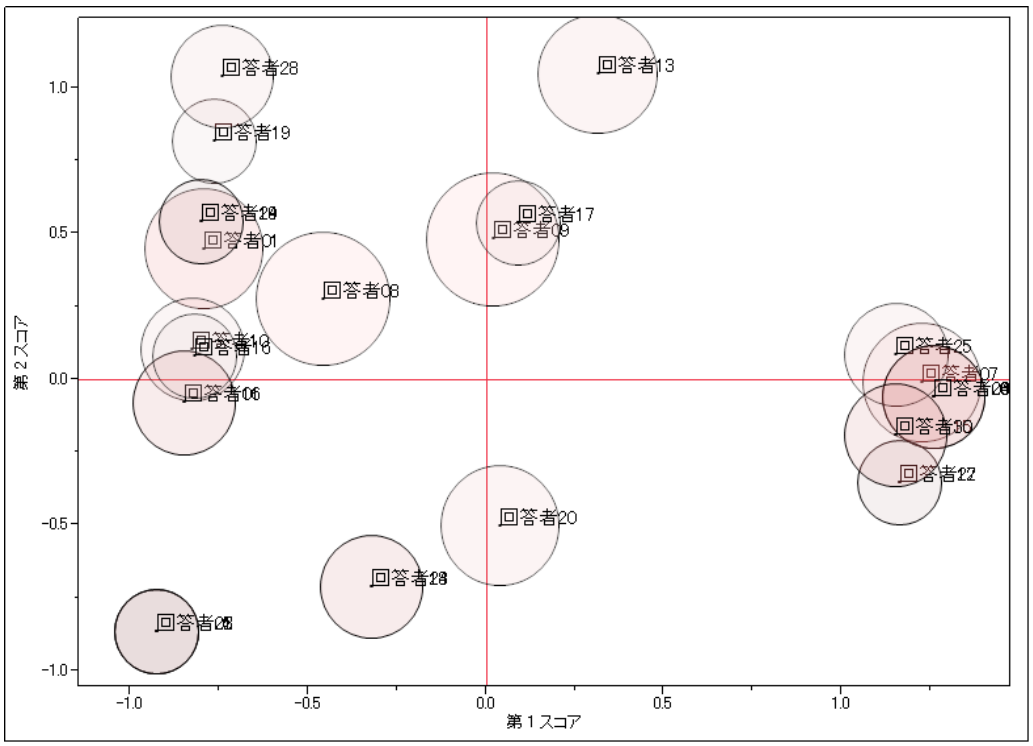


図 19　回答者の布置図（第 1，第 2 成分スコア）

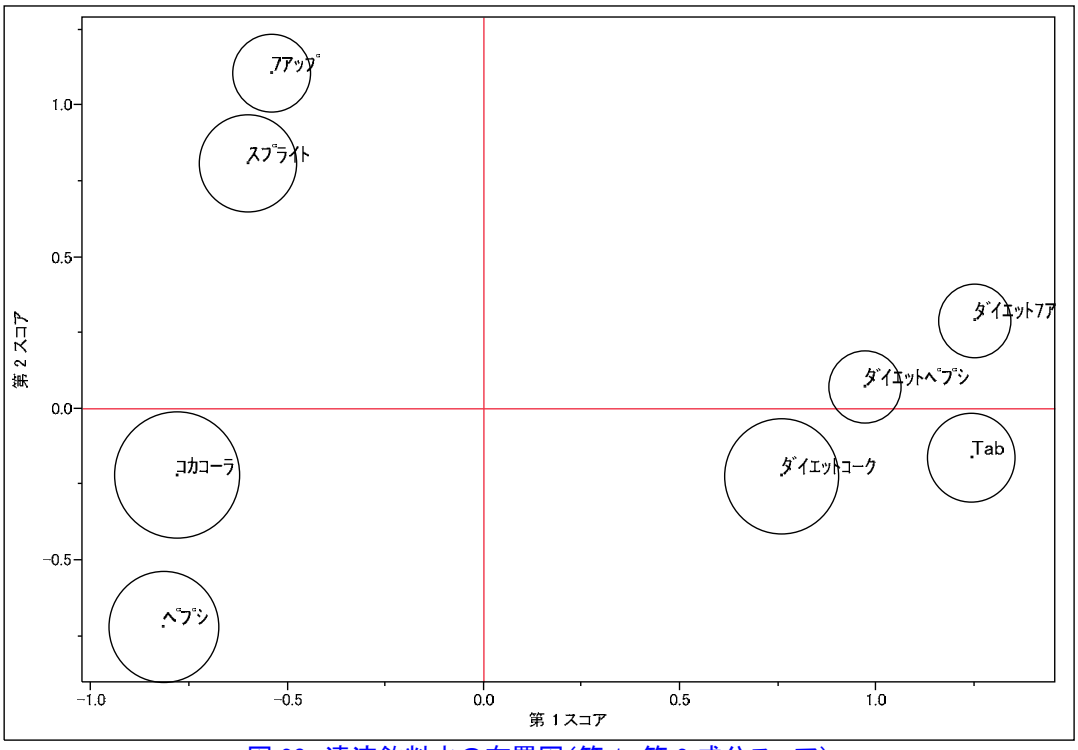


図 20　清涼飲料水の布置図（第 1，第 2 成分スコア）

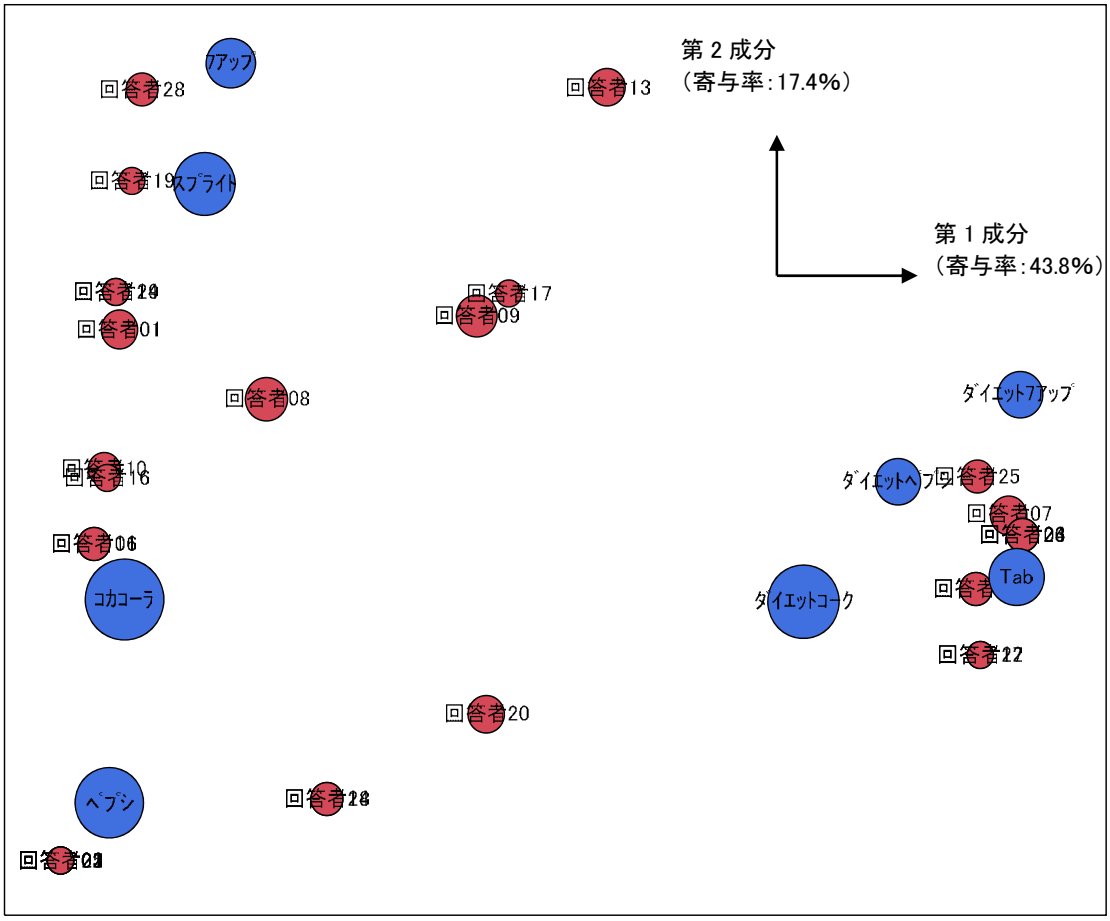


図 21　回答者・清涼飲料の同時布置図（第 1，第 2 成分スコア）

③ 寄与度の観察

前の例にならって，ここでも寄与度を調べてみる．寄与度には，「相対寄与度」と「絶対寄与度」があることはすでに述べた[85]．これらの指標がどのように機能しているか，また布置図の観察にどのように用いるかを調べる．

ⅰ）絶対寄与度の観察

「絶対寄与度」は，各成分軸の解釈に用いる指標である．各成分軸において、行や列の各選択肢がどれほどその成分に寄与しているかを示す指標である．例について，行側（回答者）と列側（銘柄）のそれぞれについて，これを一覧とした（表 49，表 50）．

まず，回答者の傾向をながめる．はじめの 2 成分について，特徴的な値のセルをアミカケとした．成分 1 では，回答者 04，07，15，23，26，30 などの値が大きい．布置図（図 19）をみると，これらの回答者は，第 1 軸（第 1 成分）の右端のほうにまとまっている．しかも，第 2 成分方向へのチラバリがちいさいことに注意しよう．

第 2 成分スコアが比較的大きい回答者ではあるが，図の左下の回答者 02，22 は第 2 成分スコアの値も大きい．しかしこれらは第 2 成分の「絶対寄与度」が大きいことに注意しよう．左上の回答者 28，19 なども類似の傾向にある（しかし，回答者 19 は第 4 成分で寄与度が大きい）．

同じように，清涼飲料水の「銘柄」の側の「絶対寄与度」を観察する．8 つの銘柄がほぼ 3 つのグループに分かれている．成分 1 への寄与度が高い銘柄は，{ダイエットコーク（＋方向），ダイエット 7 アップ（＋方向），Tab（＋方向），コカコーラ（－方向），ペプシ（－方向）}である．このことより第 1 成分は，おそらく「ダイエット飲料か否か」を表す軸と解釈できる．一方，成分 2 への寄与度が高い銘柄は，{ペプシ（－方向），スプライト（＋方向），7 アップ（＋方向）}である．これから，第 2 成分は，「コーラか否か」を表す軸と解釈できるだろう．「寄与度」はこのように，各成分の解釈を助けてくれる指標である．

ⅱ）相対寄与度の観察

各成分によって，行（回答者）や列（銘柄）がどれぐらい近似良く表されいてるかを示す指標である．つまり，各選択肢が各成分軸によって，どの程度説明されているかを示す指標である．これは言い替えると，各点と成分スコアの重心までの距離（ユークリッド距離）が，各成分軸によって，どれぐらい説明されているかを示す指標である．

まず，第 1 成分，第 2 成分についてみると，行（回答者）の側からは，表 51 のアミカケ部分の回答者が，第 1 成分や第 2 成分によって，よく近似されている．たとえば，回答者 01 は，67% + 22% = 89%が，第 1 成分と第 2 成分によって説明されることが分かる．また，列（清涼飲料水銘柄）については，表 52 の同じくアミカケ部分の銘柄が寄与していると解釈する．

85 「相対寄与度」「絶対寄与度」については，資料「第Ⅱ部」に詳しい説明がある．

表 49　回答者に関する絶対寄与度(%)

回答者	周辺度数	周辺割合 (%)	成分 1	成分 2	成分 3	成分 4	成分 5	成分 6	成分 7
回答者 01	4	4.7	3.95227	3.17431	0.00251	0.89657	1.27e-6	0.33833	5.66586
回答者 02	2	2.3	2.688	5.85599	0.41902	0.18904	0.56705	1.37612	0.01461
回答者 03	2	2.3	2.688	5.85599	0.41902	0.18904	0.56705	1.37612	0.01461
回答者 04	3	3.5	7.41782	0.04108	1.67818	0.60193	13.1088	0.08348	0.00377
回答者 05	2	2.3	2.688	5.85599	0.41902	0.18904	0.56705	1.37612	0.01461
回答者 06	3	3.5	3.39573	0.07401	0.41751	5.19568	0.40712	1.38386	0.93946
回答者 07	4	4.7	9.39072	0.00177	4.14206	0.359	0.33139	2.70637	0.08341
回答者 08	5	5.8	1.65301	1.51312	0.41471	0.64589	0.35388	12.9217	0.21164
回答者 09	5	5.8	0.00252	4.57217	6.19425	2.89485	2.14357	0.38065	0.05264
回答者 10	3	3.5	3.20942	0.12495	1.5192	13.8962	0.0006	0.79247	3.52674
回答者 11	3	3.5	3.39573	0.07401	0.41751	5.19568	0.40712	1.38386	0.93946
回答者 12	2	2.3	4.21614	0.98951	11.7491	1.34687	2.61369	0.04061	0.73868
回答者 13	4	4.7	0.61433	17.2568	16.4134	0.60897	1.4048	0.00836	3.90428
回答者 14	2	2.3	2.01447	2.31889	1.8762	6.76584	0.00719	9.22555	11.5696
回答者 15	3	3.5	6.20738	0.43306	0.71707	0.27051	18.2834	3.15139	1.20256
回答者 16	2	2.3	2.11694	0.05324	0.74014	11.6762	0.52575	1.76602	31.9761
回答者 17	2	2.3	0.026	2.2862	6.38972	10.856	0.8291	21.0994	2.47314
回答者 18	3	3.5	0.49407	5.97849	0.13012	0.08951	0.02305	9.0676	4.22691
回答者 19	2	2.3	1.84017	5.23913	0.86713	22.6207	0.52807	0.25636	5.23714
回答者 20	4	4.7	0.00941	3.9364	8.09232	1.02405	7.24345	1.47119	2.06939
回答者 21	2	2.3	2.688	5.85599	0.41902	0.18904	0.56705	1.37612	0.01461
回答者 22	2	2.3	2.688	5.85599	0.41902	0.18904	0.56705	1.37612	0.01461
回答者 23	3	3.5	7.41782	0.04108	1.67818	0.60193	13.1088	0.08348	0.00377
回答者 24	3	3.5	0.49407	5.97849	0.13012	0.08951	0.02305	9.0676	4.22691
回答者 25	3	3.5	6.23863	0.08667	18.1433	2.87389	1.48369	4.99072	6.36494
回答者 26	3	3.5	7.41782	0.04108	1.67818	0.60193	13.1088	0.08348	0.00377
回答者 27	2	2.3	4.21614	0.98951	11.7491	1.34687	2.61369	0.04061	0.73868
回答者 28	3	3.5	2.59757	12.7641	0.1717	1.55987	0.32503	0.39938	0.99592
回答者 29	2	2.3	2.01447	2.31889	1.8762	6.76584	0.00719	9.22555	11.5696
回答者 30	3	3.5	6.20738	0.43306	0.71707	0.27051	18.2834	3.15139	1.20256

表 50　銘柄に関する絶対寄与度(%)

清涼飲料水	周辺度数	周辺割合 (%)	成分 1	成分 2	成分 3	成分 4	成分 5	成分 6	成分 7
ｺｶｺｰﾗ	18	20.9	17.199	3.38878	0.03569	2.56144	0.0175	16.8591	39.0083
ﾀﾞｲｴｯﾄｺｰｸ	15	17.4	13.5121	2.9329	7.42277	0.02551	4.0465	43.2774	11.341
ﾀﾞｲｴｯﾄﾍﾟﾌﾟｼ	6	7.0	8.88056	0.12657	49.8988	7.05455	24.0596	2.85078	0.15236
ﾀﾞｲｴｯﾄ 7 ｱｯﾌﾟ	6	7.0	14.6596	1.9794	8.48474	0.0966	67.2168	0.017	0.56907
ﾍﾟﾌﾟｼ	14	16.3	14.5928	28.5496	1.95379	0.185	2.07888	4.35734	32.0035
ｽﾌﾟﾗｲﾄ	11	12.8	6.22971	28.3559	9.49855	41.6684	0.00272	0.09624	1.35777
Tab	9	10.5	21.7272	0.94004	20.5344	4.94584	1.26471	28.1639	11.9588
7ｱｯﾌﾟ	7	8.1	3.199	33.7269	2.17122	43.4626	1.31328	4.37832	3.60917

表 51　回答者に関する「相対寄与度」(%)

回答者	周辺度数	周辺割合 (%)	成分 1	成分 2	成分 3	成分 4	成分 5	成分 6	成分 7
回答者 01	4	4.7	67.225	21.5129	0.012	3.73479	4.46e-6	0.47291	7.04239
回答者 02	2	2.3	49.6294	43.0799	2.17228	0.85482	2.15598	2.0879	0.01971
回答者 03	2	2.3	49.6294	43.0799	2.17228	0.85482	2.15598	2.0879	0.01971
回答者 04	3	3.5	68.9426	0.15213	4.37953	1.37011	25.0893	0.06376	0.00256
回答者 05	2	2.3	49.6294	43.0799	2.17228	0.85482	2.15598	2.0879	0.01971
回答者 06	3	3.5	66.8305	0.58038	2.3072	25.0428	1.64999	2.2381	1.3511
回答者 07	4	4.7	85.8453	0.00644	10.6319	0.80372	0.62384	2.03306	0.05572
回答者 08	5	5.8	44.9099	16.3797	3.16365	4.29757	1.97985	28.8491	0.42019
回答者 09	5	5.8	0.05311	38.3603	36.6234	14.9286	9.29491	0.65866	0.081
回答者 10	3	3.5	43.3002	0.67169	5.75514	45.9156	0.00167	0.87861	3.47702
回答者 11	3	3.5	66.8305	0.58038	2.3072	25.0428	1.64999	2.2381	1.3511
回答者 12	2	2.3	47.722	4.46261	37.3408	3.73361	6.09215	0.03777	0.61098
回答者 13	4	4.7	4.79072	53.62	35.94	1.16305	2.25594	0.00536	2.22491
回答者 14	2	2.3	29.9448	13.7343	7.83098	24.631	0.02202	11.2694	12.5675
回答者 15	3	3.5	57.6925	1.60372	1.87134	0.61575	34.993	2.40691	0.81675
回答者 16	2	2.3	27.1556	0.2721	2.66587	36.6819	1.38882	1.86162	29.9741
回答者 17	2	2.3	0.34785	12.186	24.0017	35.5674	2.28404	23.1952	2.4177
回答者 18	3	3.5	12.3721	59.6508	0.91492	0.54897	0.11888	18.6594	7.73486
回答者 19	2	2.3	17.9992	20.4183	2.38153	54.1878	1.06365	0.20606	3.74335
回答者 20	4	4.7	0.16051	26.744	38.7447	4.27646	25.4343	2.06146	2.57854
回答者 21	2	2.3	49.6294	43.0799	2.17228	0.85482	2.15598	2.0879	0.01971
回答者 22	2	2.3	49.6294	43.0799	2.17228	0.85482	2.15598	2.0879	0.01971
回答者 23	3	3.5	68.9426	0.15213	4.37953	1.37011	25.0893	0.06376	0.00256
回答者 24	3	3.5	12.3721	59.6508	0.91492	0.54897	0.11888	18.6594	7.73486
回答者 25	3	3.5	47.0763	0.26059	38.4421	5.31108	2.30552	3.09472	3.50976
回答者 26	3	3.5	68.9426	0.15213	4.37953	1.37011	25.0893	0.06376	0.00256
回答者 27	2	2.3	47.722	4.46261	37.3408	3.73361	6.09215	0.03777	0.61098
回答者 28	3	3.5	31.3486	61.3771	0.58182	4.6104	0.80778	0.39608	0.8783
回答者 29	2	2.3	29.9448	13.7343	7.83098	24.631	0.02202	11.2694	12.5675
回答者 30	3	3.5	57.6925	1.60372	1.87134	0.61575	34.993	2.40691	0.81675

表 52　銘柄に関する「相対寄与度」(%)

清涼飲料水	周辺度数	周辺割合 (%)	成分 1	成分 2	成分 3	成分 4	成分 5	成分 6	成分 7
ｺｶｺｰﾗ	18	20.9	73.418	5.76379	0.04278	2.67783	0.01539	5.91398	12.1682
ﾀﾞｲｴｯﾄｺｰｸ	15	17.4	61.4479	5.31429	9.47821	0.02841	3.78945	16.173	3.76881
ﾀﾞｲｴｯﾄﾍﾟﾌﾟｼ	6	7.0	29.7311	0.16883	46.9071	5.78416	16.5872	0.7843	0.03727
ﾀﾞｲｴｯﾄ 7 ｱｯﾌﾟ	6	7.0	46.188	2.48488	7.50621	0.07454	43.611	0.0044	0.13102
ﾍﾟﾌﾟｼ	14	16.3	49.1558	38.3177	1.84795	0.15261	1.44204	1.20615	7.87773
ｽﾌﾟﾗｲﾄ	11	12.8	20.4203	37.0341	8.74234	33.4503	0.00184	0.02592	0.32523
Tab	9	10.5	66.7963	1.15149	17.7259	3.72382	0.80067	7.1152	2.68663
7 ｱｯﾌﾟ	7	8.1	11.1134	46.6846	2.11794	36.9784	0.93951	1.24993	0.91624

5. 3. 3 分析例 3:ウェブ調査による調査データ

応用例として，あるウェブ・パネルを用いて行った「情報に関する調査」でえた集約化データ表の分析を挙げる．これは，調査設計時に，事後の分析手法として対応分析法の利用を想定して質問文を設計し，集めたデータである．このように，調査後に行う分析で対応分析法がどのように使えそうかを念頭に，調査前の質問文設計時に工夫することも大切なことである．

ここでは，多数の質問文から「情報接触とその情報源の評価」に関連して用意した，つぎの質問を取り上げる．

[用いた質問文の一部]

Q17. 現在私たちは、情報を入手できる手段として数多くの情報源に囲まれており、
それらの情報源についていろいろな意見が言われています。
さて、以下でAからDの4つのことばがあてはまる情報源にはどのようなものがあるでしょうか。
あなたが「あてはまる」と思われるものをすべてお選びください。(それぞれいくつでも)

	A. 情報が正確 ↓	B. 情報が詳しい ↓	C. 情報量が多い ↓	D. 信頼できる ↓
1. テレビの番組	□	□	□	□
2. ケーブルテレビ・衛星放送の番組	□	□	□	□
3. ラジオの番組	□	□	□	□
4. 新聞の記事(電子版を含む)	□	□	□	□
5. 新聞の紙面広告(電子版を含む)	□	□	□	□
6. 書籍(漫画・コミック以外)	□	□	□	□
7. 一般の雑誌・週刊誌(漫画・コミック以外)の記事	□	□	□	□
8. 各分野専門の情報誌の記事	□	□	□	□
9. パンフレット・カタログ・ダイレクトメール	□	□	□	□
10. 都・県や市・区など自治体の広報誌紙	□	□	□	□
11. 所属する会や組織の会報・同人誌・ニュースレター	□	□	□	□

	情報が正確	情報が詳しい	情報量が多い	信頼できる
12. パソコンでみるインターネットサイト	□	□	□	□
13. 携帯電話・PHS、スマートフォンでみるインターネットサイト	□	□	□	□
14. インターネットブログ、ブログ	□	□	□	□
15. ツイッター(Twitter)	□	□	□	□
16. 電子書籍(電子書籍端末や電子ブックリーダーで読む)	□	□	□	□
17. ミクシィ(mixi)	□	□	□	□
18. フェイスブック(Facebook)	□	□	□	□
19. グリー(GREE)	□	□	□	□
20. モバゲータウン	□	□	□	□
21. YouTube	□	□	□	□
22. ニコニコ動画	□	□	□	□
23. 1～22の中にはひとつもない	□	□	□	□

図 22 調査票の一部(用いた質問文の一部)

調査票の一部を示したが，ここにあるようにマトリクス（グリッド）形式を用いた質問文である．表側（ひょうそく）にある“23 の情報源”に対して，表頭（ひょうとう）にあるような“評価項目”を 9 項目用意した．これを下に要約する．

なお，マトリクスが大きくなるので，調査票ではいくつかに区分して画面表示するようにした．こうした調査票仕様（フォーマット）が，回答者の回答行動に影響を及ぼすことも考

えられる．とくに，調査誤差でいう**“非標本誤差”**（測定誤差，無回答誤差など）の介入が無視できない．非常に大きなマトリクスを画面一杯に展開・表示することも，回答者への威圧感や辟易感・回答拒否につながるおそれもある（とくに，無回答の発生など）．いずれにしても，こうした画面設計の“効果測定”に留意のうえ，調査票を設計することが重要である．見栄えよりも“回答者が，いかに抵抗なく，また誤りなく回答できるか”が重要な検討要素である[87].

●情報源（23 選択肢）

1. テレビの番組
2. ケーブルテレビ・衛星放送の番組
3. ラジオの番組
4. 新聞の記事（電子版を含む）
5. 新聞の紙面広告（電子版を含む）
6. 書籍（漫画・コミック以外）
7. 一般の雑誌・週刊誌（漫画・コミック以外）の記事
8. 各分野専門の情報誌の記事
9. パンフレット・カタログ・ダイレクトメール
10. 都・県や市・区など自治体の広報誌紙
11. 所属する会や組織の会報・同人誌・ニュースレター
12. パソコンでみるインターネットサイト
13. 携帯電話・PHS、スマートフォンでみるインターネットサイト
14. インターネットブログ、ブログ
15. ツイッター（Twitter）
16. 電子書籍（電子書籍端末や電子ブックリーダーで読む）
17. ミクシィ(mixi)
18. フェイスブック（Facebook）
19. グリー（GREE）
20. モバゲータウン
21. YouTube
22. ニコニコ動画
23. 1～22 の中にはひとつもない

●評価項目（9 選択肢）

情報が正確
情報が詳しい
情報量が多い
信頼できる
生活に欠かせない
役に立つ
世間の話題や流行を知る
商品を選び購入する
古くさい

回答者は上の調査票の該当箇所に「レ」を入れる．なお，以降の対応分析では，「1～22 の中には 1 つもない」および「無回答」は，まったく除外してから，分析を行っている．こうして得られる回答データから，“［情報源（22 選択肢)］×［評価項目（9 項目)］の 2 元データ表が得られる．もちろん，こうした取得情報は，人口統計学的変数（属性など）が影響することがあるだろう．実際，ここで得たデータでも，性別，年齢，職業などが関連することは分かっている．しかし，ここでは，まず，全体像として，この 2 元データ表からどのような情報が読み取れるのか，1 つの分析例として，対応分析法により探査分析する．

87 これについては，さまざまな研究がある．たとえば，Couper, M.P. (2008): *Designing Effective Web Surveys*, Cambridge University Press.などを参照．これの紹介記事が以下にある．
http://wordminer.org/smr/documents/documents-oubun/237

表 53 ［情報源(24 選択肢)］×［評価項目(9 項目)］の 2 元データ表

質問項目	Q17_A 情報が正確	Q17_B- 情報が詳しい	Q17_C- 情報量が多い	Q17_D – 信頼できる
全回答者数（回答者数）	347	347	347	347
1. テレビの番組	100	114	235	90
2. ケーブルテレビ・衛星放送の番組	43	91	103	44
3. ラジオの番組	65	68	86	54
4. 新聞の記事（電子版を含む）	140	153	131	131
5. 新聞の紙面広告（電子版を含む）	36	59	90	28
6. 書籍（漫画・コミック以外）	41	98	108	39
7. 一般の雑誌・週刊誌（漫画・コミック以外）の記事	19	84	139	12
8. 各分野専門の情報誌の記事	88	163	89	86
9. パンフレット・カタログ・ダイレクトメール	31	90	90	18
10. 都・県や市・区など自治体の広報誌紙	125	70	38	132
11. 所属する会や組織の会報・同人誌・ニュースレ	45	73	50	46
12. パソコンでみるインターネットサイト	26	118	274	24
13. 携帯電話・PHS、スマートフォンでみるインターネットサイト	18	63	166	15
14. インターネットブログ、ブログ	6	59	150	5
15. ツイッター（Twitter）	6	31	143	9
16. 電子書籍（電子書籍端末や電子ブックリーダー	22	39	103	19
17. ミクシィ(mixi)	9	36	128	11
18. フェイスブック（Facebook）	12	17	115	12
19. グリー（GREE）	6	20	99	5
20. モバゲータウン	8	20	100	4
21, YouTube	16	42	135	10
22. ニコニコ動画	7	22	122	6
23. 1～22 の中にはひとつもない	54	23	11	62
99. 無回答	0	0	0	0

［情報源(24 選択肢)］×［評価項目(9 項目)］の 2 元データ表(つづき)

質問項目	Q18_A- 生活に欠かせない	Q18_B- 役に立つ	Q18_C- 世間の話題や流行を知る	Q18_D- 商品を選び購入する	Q18_E – 古くさい
回答者数	347	347	347	347	347
1. テレビの番組	226	166	248	51	20
2. ケーブルテレビ・衛星放送の番組	42	94	74	30	17
3. ラジオの番組	50	103	79	9	82
4. 新聞の記事（電子版を含む）	135	167	124	17	24
5. 新聞の紙面広告（電子版を含む）	29	62	74	59	16
6. 書籍（漫画・コミック以外）	57	100	71	33	16
7. 一般の雑誌・週刊誌（漫画・コミック以外）の記事	27	74	136	46	12
8. 各分野専門の情報誌の記事	24	137	63	47	10
9. パンフレット・カタログ・ダイレクトメール	12	76	59	155	29
10. 都・県や市・区など自治体の広報誌紙	41	148	25	7	79
11. 所属する会や組織の会報・同人誌・ニュースレ	11	102	25	5	67
12. パソコンでみるインターネットサイト	186	204	200	182	0
13. 携帯電話・PHS、スマートフォンでみるインターネットサイト	77	105	107	44	0
14. インターネットブログ、ブログ	34	67	135	14	1
15. ツイッター（Twitter）	21	43	117	5	1
16. 電子書籍（電子書籍端末や電子ブックリーダー	9	53	61	7	1
17. ミクシィ(mixi)	25	43	107	8	8
18. フェイスブック（Facebook）	14	35	94	7	9
19. グリー（GREE）	6	18	79	3	6
20. モバゲータウン	7	20	76	2	6
21, YouTube	27	88	120	4	1
22. ニコニコ動画	10	46	94	2	5
23. 1～22 の中にはひとつもない	25	12	13	38	123
99. 無回答	0	0	0	1	0

① 固有値，特異値，寄与率など

まず，基本情報として，固有値．特異値，寄与率などを調べる．なお，繰り返しになるが，上の表から「1～22 の中にはひとつもない」，および，「無回答」は分析から除外している[88]．この結果から，はじめの 2 成分で全体の総変動の約 76%，第 3 成分まで考えると，約 89%の情報がある．とくに，第 1 成分だけで約 50%も占める．つまり，はじめの数成分を観察すればほぼ十分であろう．

88 これらを含めたデータ表について対応分析を行うとどうなるであろうか．

表 54　固有値，特異値，寄与率，累積寄与率

成分 k	特異値 α_k	固有値 λ_k	寄与率(%) ν_k	累積寄与率(%) $\sum_k \nu_k$
1	0.40244	0.16195	**50.4**	**50.4**
2	0.28439	0.08088	**25.2**	**75.6**
3	0.20595	0.04242	**13.2**	**88.8**
4	0.14903	0.02221	6.9	95.7
5	0.08845	0.00782	2.4	98.2
6	0.06512	0.00424	1.3	99.5
7	0.03572	0.00128	0.4	99.9
8	0.018	0.00032	0.1	100
固有値総和 (総変動)	—	0.321122	—	—

② 成分スコアの観察

情報源，評価項目それぞれの成分スコアの布置図と，同時布置図を描き観察する．すでにこの程度の選択肢数となると，図の判読がかなりむずかしくなる（とくに図 24）．

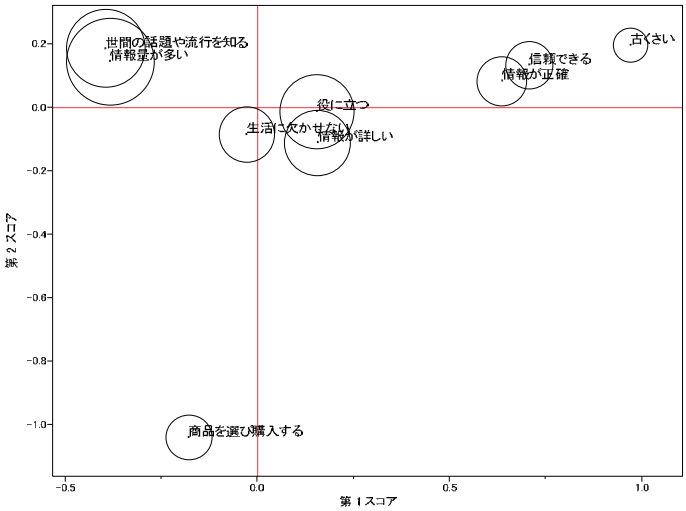


図 23 「評価項目」の布置図（第 1，第 2 成分スコア）

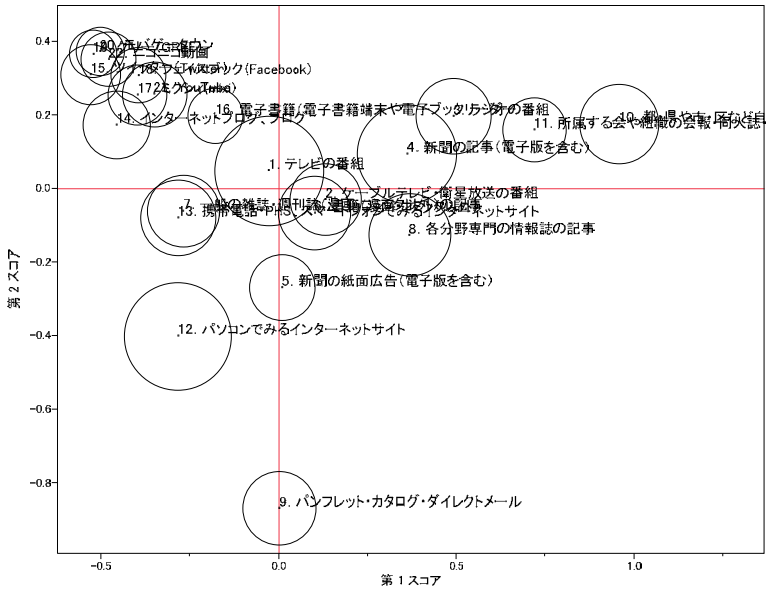


図 24 「情報源」の布置図（第 1，第 2 成分スコア）

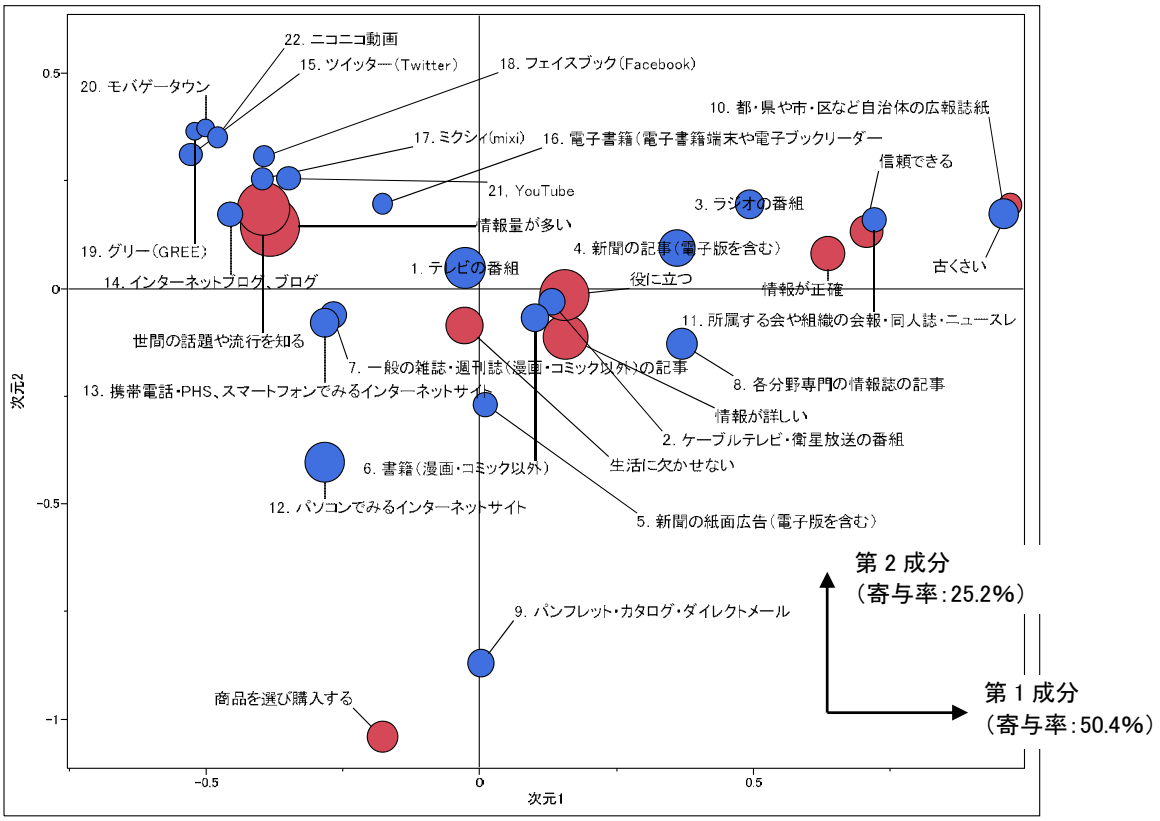


図 25 「情報源」と「評価項目」の同時布置図(はじめの 2 成分)

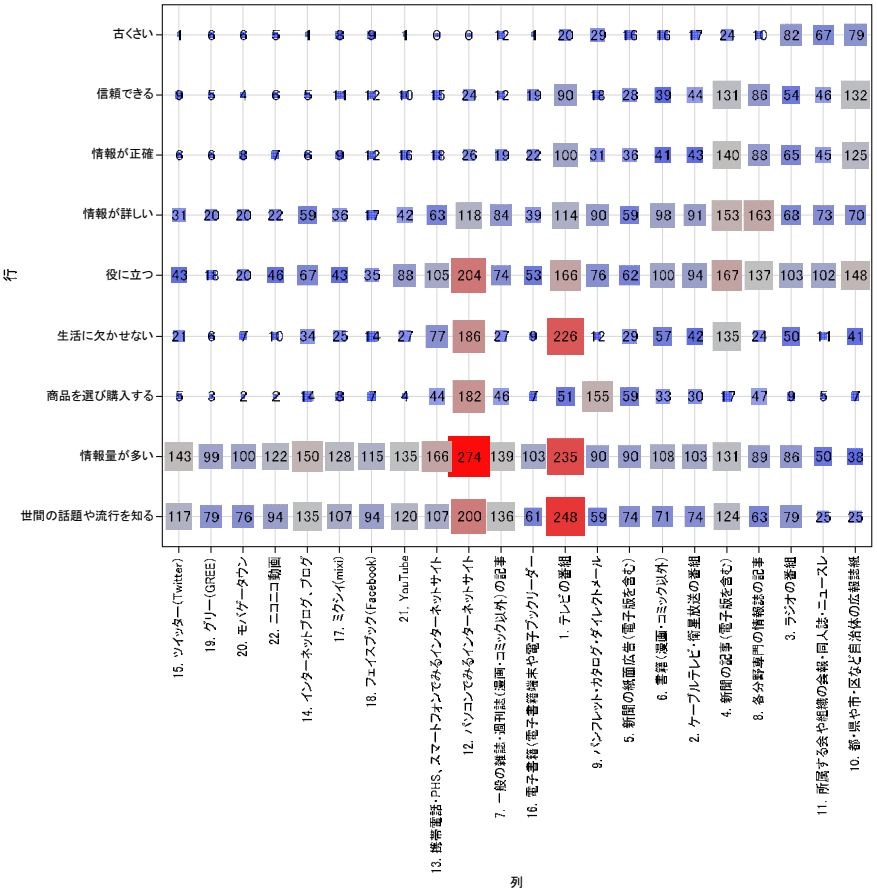


図 26 第1成分スコアについて行と列の並べ替え

表 55 「情報源」に関する「絶対寄与度」(%)と「相対寄与度」(%)（はじめの 3 成分）

情報源	周辺度数	周辺割合(%)	「絶対寄与度」			「相対寄与度」		
			成分 1	成分 2	成分 3	成分 1	成分 2	成分 3
1. テレビの番組	1250	10.2	0.04902	0.30068	25.5231	0.56428	1.72833	76.939
2. ケーブルテレビ・衛星放送の番組	538	4.4	0.46087	0.04994	0.02134	46.7512	2.52991	0.5669
3. ラジオの番組	596	4.9	7.30175	2.33441	5.63751	56.8836	9.08183	11.5021
4. 新聞の記事（電子版を含む）	1022	8.4	6.69741	0.92806	17.3106	54.122	3.74525	36.6362
5. 新聞の紙面広告（電子版を含む）	453	3.7	0.00205	3.28035	0.95877	0.09368	75.0044	11.4967
6. 書籍（漫画・コミック以外）	563	4.6	0.28661	0.25636	0.51745	23.4978	10.496	11.1106
7. 一般の雑誌・週刊誌（漫画・コミック以外）の記事	549	4.5	2.00916	0.2052	2.53683	60.0597	3.0633	19.8605
8. 各分野専門の情報誌の記事	707	5.8	4.83881	1.14317	0.10688	43.2235	5.09949	0.25005
9. パンフレット・カタログ・ダイレクトメール	560	4.6	6.12e-5	42.7316	15.9014	0.00024	82.2936	16.0601
10. 都・県や市・区など自治体の広報誌紙	665	5.4	30.7575	2.07992	0.28615	92.7585	3.13246	0.22601
11. 所属する会や組織の会報・同人誌・ニュースレ	424	3.5	11.1032	1.10576	13.0852	66.8684	3.3256	20.6388
12. パソコンでみるインターネットサイト	1214	9.9	4.95215	19.7394	5.17362	27.8126	55.3628	7.60981
13. 携帯電話・PHS、スマートフォンでみるインターネットサイト	595	4.9	2.42885	0.36819	1.89995	64.5289	4.88499	13.2199
14. インターネットブログ、ブログ	471	3.9	4.96848	1.43435	0.17576	82.1205	11.8391	0.7608
15. ツイッター（Twitter）	376	3.1	5.3181	3.68791	0.56952	71.7905	24.8615	2.01349
16. 電子書籍（電子書籍端末や電子ブックリーダー	314	2.6	0.50519	1.24889	0.3446	21.798	26.9105	3.89412
17. ミクシィ(mixi)	375	3.1	3.00622	2.4599	0.71268	66.889	27.333	4.15297
18. フェイスブック（Facebook)	315	2.6	2.48633	3.03165	1.78124	51.1357	31.1373	9.59445
19. グリー（GREE）	242	2.0	3.33982	3.28979	2.86626	55.4258	27.2643	12.4577
20. モバゲータウン	243	2.0	3.09163	3.41978	2.50942	54.1558	29.9152	11.5123
21, YouTube	443	3.6	2.74802	2.96016	0.04328	54.6678	29.4078	0.22549
22. ニコニコ動画	314	2.6	3.64872	3.94457	2.03845	57.9811	31.3027	8.48355

③ 寄与度の観察

この例でも，絶対寄与度と相対寄与度を求め，一覧とした（表 55）．ただしここでは，固有値の寄与率の情報を勘案して初めの 3 成分について示した．前の例と同様に，各成分で特徴的な「情報源」について「アミカケ」としてある．ここではとくに解説を行わないのが，成分スコアの布置図，同時布置図と合わせて観察すれば，ある傾向が見えてくるであろう．つぎに述べるクラスター化の結果も合わせて吟味するとより効果的である．

④ クラスター化情報の観察

2 元データ表の寸法が大きくなると，布置図や同時布置図は点の重なりが多くなって，判読がむずかしくなる．そこで，寄与度（絶対寄与度，相対寄与度）を観察して，各成分における行，列の要素の影響度を観察することが大事であることは上でみた．これに加えて，**クラスター化**を行うことも効果的である[89]．とくに次に挙げる分析例 4 の自由回答データのように，データ表の寸法が相当大きくなると，クラスター化による分類は非常に有用である．

図 27 は，「22 の情報源」について，クラスター化を行いその結果を**デンドログラム**（樹形図）で示したものである．また，図 28 は「評価項目」の分類結果である．これを布置図・同時布置図と併用して，また寄与度を確認しながら，情報源の間の相互の類似性を観察する．たとえば，この例では，

グループ 1 :「テレビ番組」「書籍」「新聞や雑誌」それに「PC やスマートフォンでみるサイト」といった，情報を一方向で“受け手”として利用する場面と，

グループ 2 :「パンフレット・カタログ・ダイレクト・メール」（これは 1 つのグループを形成）

89 対応分析法とその特性を利用したクラスター化法の詳細については，別資料「第Ⅲ部」に記した．

グループ 3：「ツイッター」「ニコニコ動画」「ミクシィ」「フェイスブック」のように，双方向で情報を授受する場面，

グループ 4：そして，「ラジオ番組」「同人誌やニュースレター」「新聞記事」「情報誌記事」といった，従来からのややクラシックな情報源から取得の場面，

といったように，おおよそ分かれている．また，この 4 つのグループに対して，おおまかではあるが，評価項目が，ほぼ以下のように対応していることも読み取れる[90]．

表 56　情報源と評価項目との関連（抜粋）

グループ	情報源	評価項目
グループ 1	1. テレビの番組 2. ケーブルテレビ・衛星放送の番組 5. 新聞の紙面広告（電子版を含む） 6. 書籍（漫画・コミック以外） 7. 一般の雑誌・週刊誌（漫画・コミック以外）の記事	情報が詳しい 役に立つ 生活に欠かせない
グループ 2	9. パンフレット・カタログ・ダイレクトメール	商品を選び購入する
グループ 3	15. ツイッター（Twitter） 17. ミクシィ(mixi) 18. フェイスブック（Facebook) 19. グリー（GREE） 20. モバゲータウン 22. ニコニコ動画	情報量が多い 世間の話題や流行を知る
グループ 4	3. ラジオの番組 4. 新聞の記事（電子版を含む） 8. 各分野専門の情報誌の記事 10. 都・県や市・区など自治体の広報誌紙 11. 所属する会や組織の会報・同人誌・ニュースレター	情報が正確 信頼できる 古くさい

とくに，注目したい特徴として，以下がある．

① 「グループ 1」に含まれる情報源は，布置図のほぼ中央に位置している．このことは，“平均的な”印象を持たれているメディアということである．

② 一方「グループ 3」は，双方向性という特性のある，最近のソーシャル・メディア系のメディアである．

③ また，「グループ 4」は，旧来からの典型的な情報源であり「古くさい」との判断もあるが，「情報が正確」「信頼出来る」という意味で，他の情報源とは異なる特性を持っていると判断されたことが読み取れる．

布置図，同時布置図の観察で注意することは，各情報源と評価項目とは，つまりデータ表の行と列の選択肢の成分スコアは，図の上で（距離が）近いことが，そのまま類似性が高いと解釈してはいけない，ということである．これは，すでに述べた“双対性”を考えれば明らかである．

たとえば，「9. パンフレット・カタログ・ダイレクトメール」と「商品を選び購入する」とは，非常に近い位置に布置されている．しかし，この布置図の上で直接，2 点間の距離を考えてはいけない，ということである．ではどうしてこの 2 点は近接しているのか．じつは，「9. パンフレット・カタログ・ダイレクトメール」に付与された成分スコアは，「商品を選

90 いつも，このように排反的に重なりなく分けることがよいわけではない．むしろ通常はグループ内あるいはグループ間の関係は曖昧なことが多い．それにも関わらず（やや強引に）グループ化を行いその特徴を観察することに意味がある．

び購入する」の成分スコアだけと比べるのではなく，“9 つの評価項目”すべての成分スコアの（プロファイルを加重とする）加重平均であり，結果として「9. パンフレット・カタログ・ダイレクトメール」の近傍にその成分スコアが位置したと考えているのである．また，この逆の関係も**双対性の関係**からなり立つ[91]. 対応分析法の結果はそのように仕組みで機能しており，その特性を利用して同時布置図を解釈するのである．

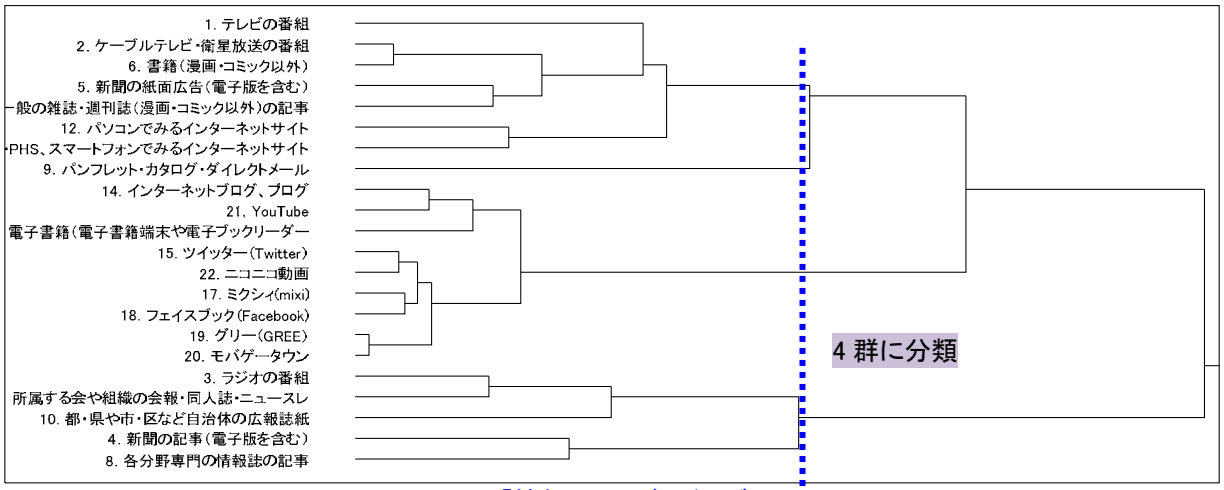


図 27 「情報源」のデンドログラム

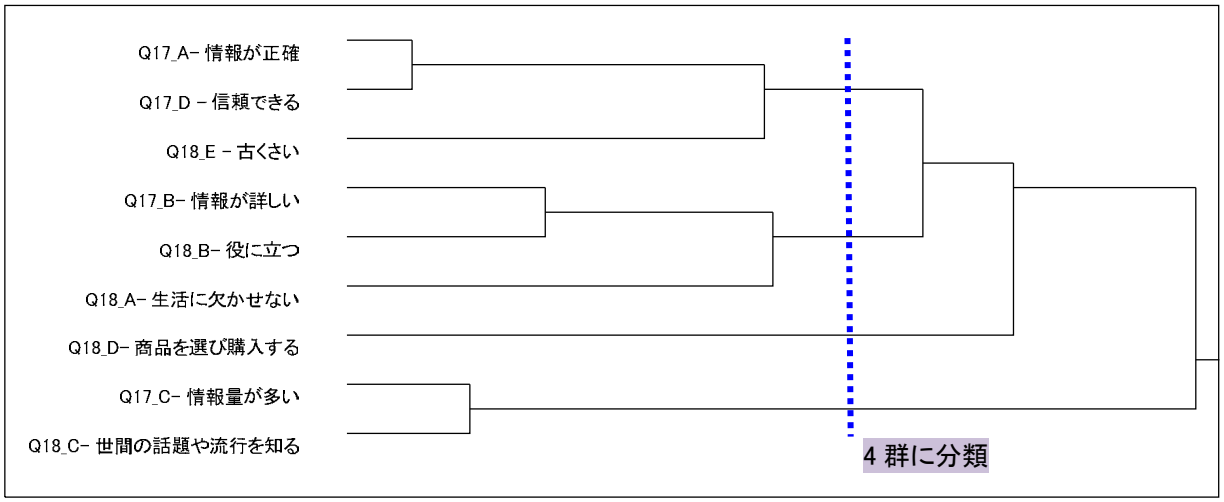


図 28 「評価項目」のデンドログラム

5. 3. 4 分析例 4:ウェブ調査による意識調査データ

最後の例として，前に 5．2 節で述べた「普段の生活やインターネットなどについてのおうかがい」調査の自由回答質問からえたデータセットを用いる．この自由回答データの一部を用いて，実際に対応分析法を行ってみる．まず，この種のデータ表の特徴を知っておくことは重要である．

- 2 元データ表の寸法は**“非常に大きくなる”**こと．行・列の大きさが，数千から数万に及ぶことも珍しくないこと．
- しかも，2 元データ表のセル内の度数が非常に少ない，いわゆる“疎な行列”（sparse matrix）であり，ときとして条件の悪い行列（ill-posed matirix）となることもあること[92].
- 自由回答データは，分かち書き処理や辞書編集を繰り返すなどが頻発し，出発行列とす

91 双対性については，4.3 節、4.4 節に述べた．「参考文献」にあげたフランス語圏の書にも必ず説明がある．

92 たとえば，プロファイルが極端に偏っている少数のグループがあるとき．自由回答であると記述語句が特定の内容に偏っているなどの現象がみられるときなど．

る 2 元データ表の**“寸法が確実に決まらないこと”**があること.
- こうしたあまり条件のよくない 2 元データ表でも,（ある程度の精度で）実用に耐える解を求めねばならず，場合によっては，特別な計算アルゴリズムを必要とすること[93].

さいわい，最近の PC 環境は計算処理能力の向上もあり，また実用に耐えるアルゴリズムも登場して，かなり大きな寸法のデータ表でも対応分析法の処理が可能となっている[94].

さてここで例に戻って説明しよう．まず，ここで用いる調査の概要は，すでに前に述べた（5. 2 節）．そこでは元の多変量型データ行列から，どのようにして対応分析法の対象とする 2 元データ表を生成するか，について述べた．ここでは，そのときにみた表 41 にあるような，「（構成要素変数）×（属性）型」の 2 元データ表を対象としてみよう．より，具体的には，構成要素変数として，分かち書き処理あとに簡単な辞書編集を行って[95]生成した単語･語句群の“構成要素変数”と，“性年齢区分”（ここでは 10 歳区分）という属性を質的変数とする“構成要素変数×質的変数”の 2 元データ表を分析対象とする. ちなみに, このデータ表の寸法は「229 語句×性年齢区分（14 選択肢)」である（実際には「無回答」を除いたので，性年齢区分は 12 区分，つまり，寸法が 229 行×12 列の 2 元データ表).

① 固有値，特異値，寄与率の観察

ここでもまず．固有値，特異値，寄与率などを要約する．データ表の寸法が大きくなったことで，情報がややあいまいになり，各成分の情報量（変動，慣性）が小さめとなったことに注意しよう．たとえば，はじめの 4 成分で寄与率が約 51%程度である．また特異値をみると，そう高い相関ではないがデータ表の寸法を考えると，それなりに相関があるとみてよさそうである[96].

表 58 固有値，特異値，寄与率ほか

成分 k	特異値 α_k	固有値 λ_k	寄与率(%) ν_k	累積寄与率(%) $\sum_k \nu_k$
1	0.3927	0.1542	17.7	17.7
2	0.3227	0.1041	11.9	29.6
3	0.3164	0.1001	11.5	41.0
4	0.2916	0.0850	9.7	50.8
5	0.2724	0.0742	8.5	59.3
6	0.2678	0.0717	8.2	67.5
7	0.2622	0.0688	7.9	75.4
8	0.2447	0.0599	6.9	82.2
9	0.2430	0.0590	6.8	89.0
10	0.2314	0.0536	6.1	95.1
11	0.2063	0.0426	4.9	100
固有値総和（総変動）	—	0.873193	—	—

93 たとえば，Berry（1992）の論文にあるような大規模で疎なデータ表を解くアルゴリズムなど.

94 「第Ⅱ部」も参照.

95 ここの辞書編集は「句読点，記号・符号」程度の除去をいう．実際には，類語や同義語の置換，誤記の訂正や削除など，複数の辞書を作って丁寧に編集を行うことで，分析結果がより鮮明となる．ここでは，あえてほとんど辞書編集を行わずに，いきなり対応分析法を適用してみた.

96 データ表の寸法が数千行・列となると，かなり小さい特異値となりその変化もなだらかで寄与率も小さいことが通例である．こうした場合の特異値や寄与率の特徴をどう解釈・吟味するかの検証方法が必要だろう.

	固有値	寄与率	累積寄与率
1	0.1542	17.66	17.66
2	0.1041	11.92	29.58
3	0.1001	11.47	41.05
4	0.0850	9.74	50.78
5	0.0742	8.50	59.28
6	0.0717	8.21	67.49
7	0.0688	7.88	75.37
8	0.0599	6.86	82.23
9	0.0590	6.76	88.99
10	0.0536	6.14	95.12
11	0.0426	4.88	100.00

図 29　固有値他の表示例

② カイ二乗統計量との関係と解釈

ここでちなみに，カイ二乗統計量 χ_p^2 を求めてみると，$\chi_p^2/N=\sum_{i=1}^{K}\lambda_k$ の関係から，$\chi_p^2=N\times\sum_{i=1}^{K}\lambda_k=3485\times0.873193=3043.078$ となる（注：ここで N はここで扱っているデータ表の総度数，つまり総構成要素数である．表 40 を参照）．データ表の寸法は，$m=229, n=12$ であるから，自由度は $d.f.=(m-1)(n-1)=228\times11=2508$ となるが，こういう計算を行うまでもなく（ほとんど意味がない），これだけ寸法の大きなデータ表となると，統計的には有意となる．

各成分の特異値，とくにはじめのほうのいくつかは，ほとんど大きさに変化がない．第 1 成分の特異値は $\alpha_1=0.39267$，つまり，回答者の自由回答でえた 229 語の用語群（構成要素群）と「性年齢区分」の間には，やはりゆるやかな相関があると読めるのである．

ここで，図 30 の性年齢区分の成分スコアの布置図をみよう．これをみるとまず，「男女の性差」があることが読み取れる．また，年齢区分については布置図だけからは，以下の傾向があるようにみえるが，実は，性別，年齢の違いによって意見に違いがあることが後述の有意性テストの結果と合わせて観察することで見えてくる（性差，年齢差はたしかに観察されるのだが，個々の発言は 2 次元という少数次元内には情報が収まらないということである）．

- 若年層では男女差はないようにみえる（しかし，実は差違がある）．
- 逆に，加齢の伴い年齢が上がると男女の差が大きくなる傾向がある．とくに，50 歳以上はその傾向が顕著である．
- 男女とも，20 代～40 代あたりは，あまり差違が顕著ではなさそうである（ここも実は加齢に伴い意見に違いがみえる）．

③ 性年齢区分で特徴的な語句の観察

ここで用いたデータ表は表 41 にある「(構成要素変数) × (性年齢区分)」で，ここから「無回答」を除いた寸法が 229 行×12 列の 2 元データ表であった．そして自由回答で登場しここで構成要素として選ばれた語句と性年齢区分のおおよその対応関係は観察できた．では次に，各性年齢区分に属する回答者がどのような語句を用いて（正確には構成要素を用いて）意見を述べたのか，どのような傾向にあるのかを調べよう．布置図に合わせて言えば，各性年齢区分の点と構成要素の点との関係である（この言い方には少し注意が必要だが，おおまかに言うと，という意味でこう記しておく）．

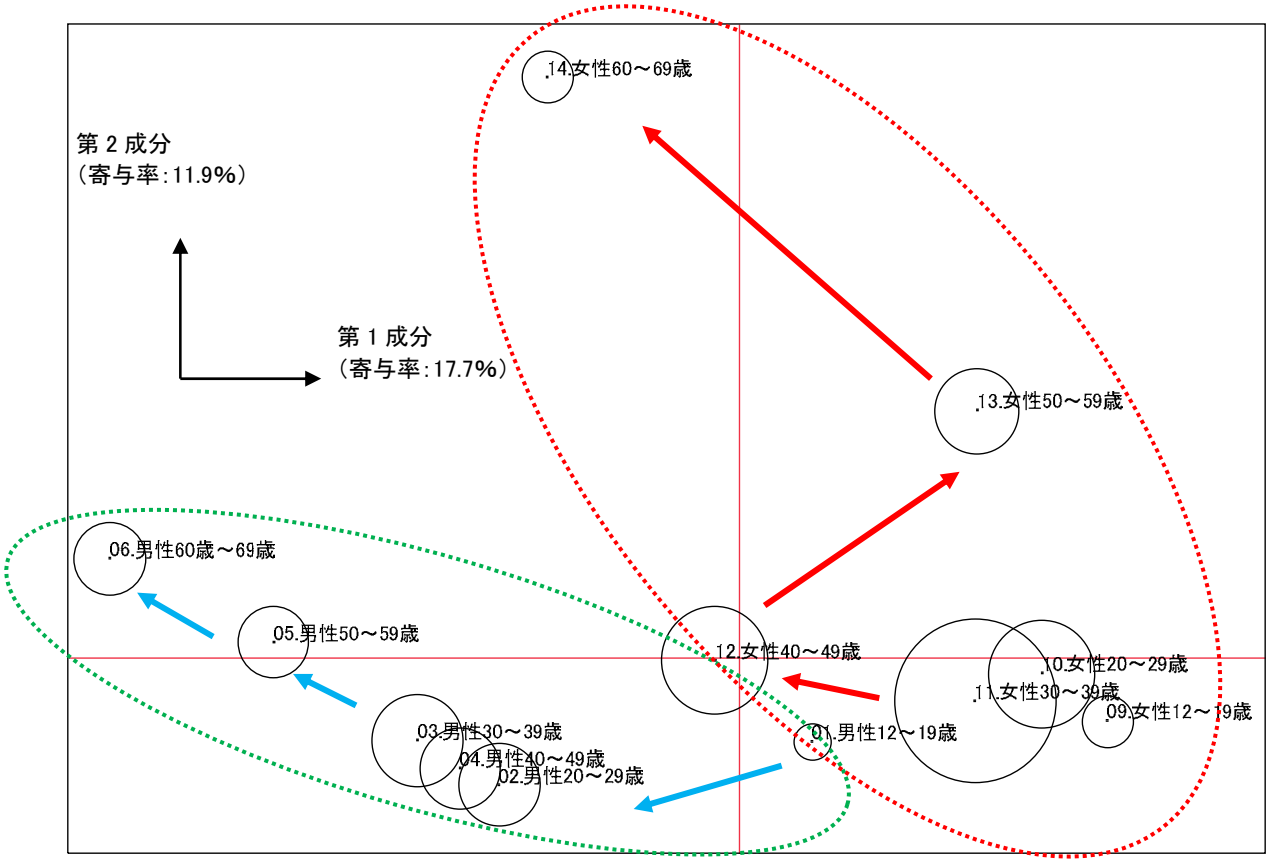


図 30　性年齢区分（12 選択肢）の成分スコア布置図

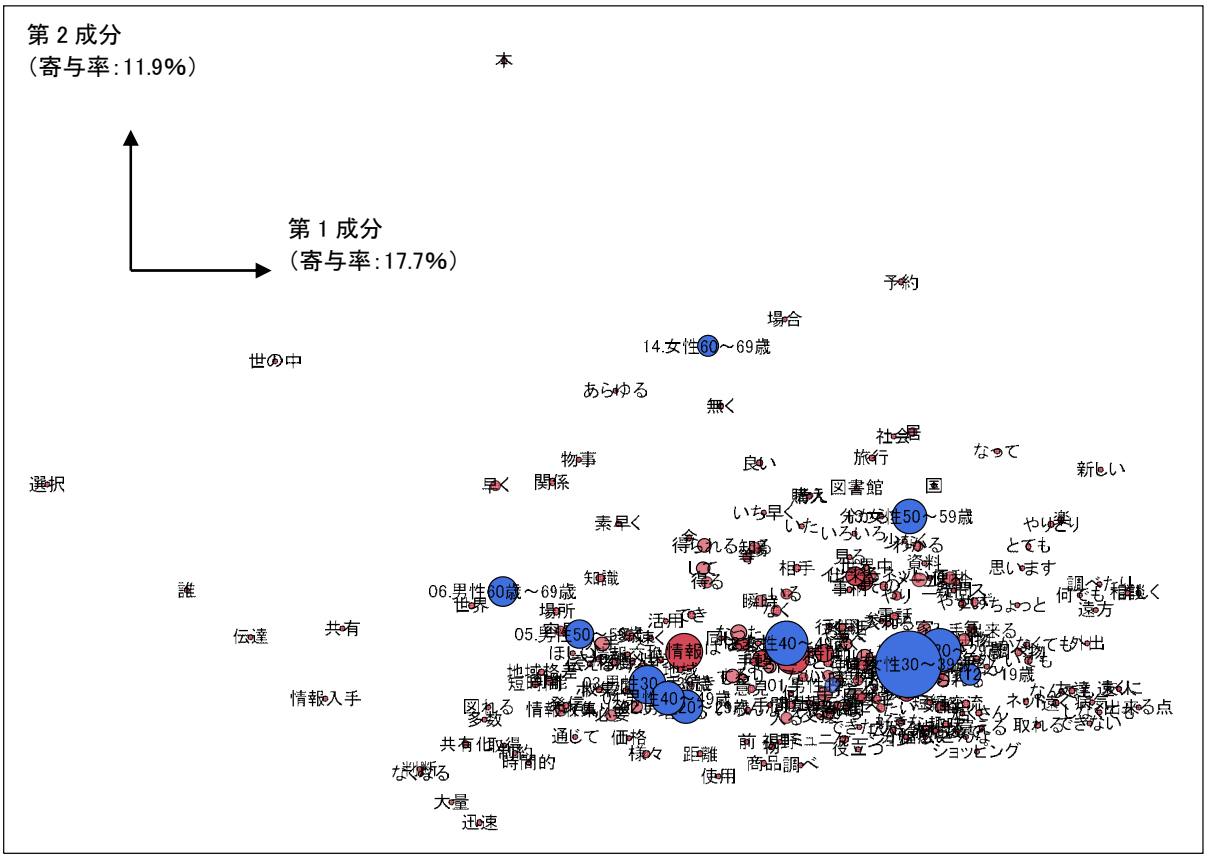


図 31　構成要素（229 の語句）と性年齢区分の同時布置図

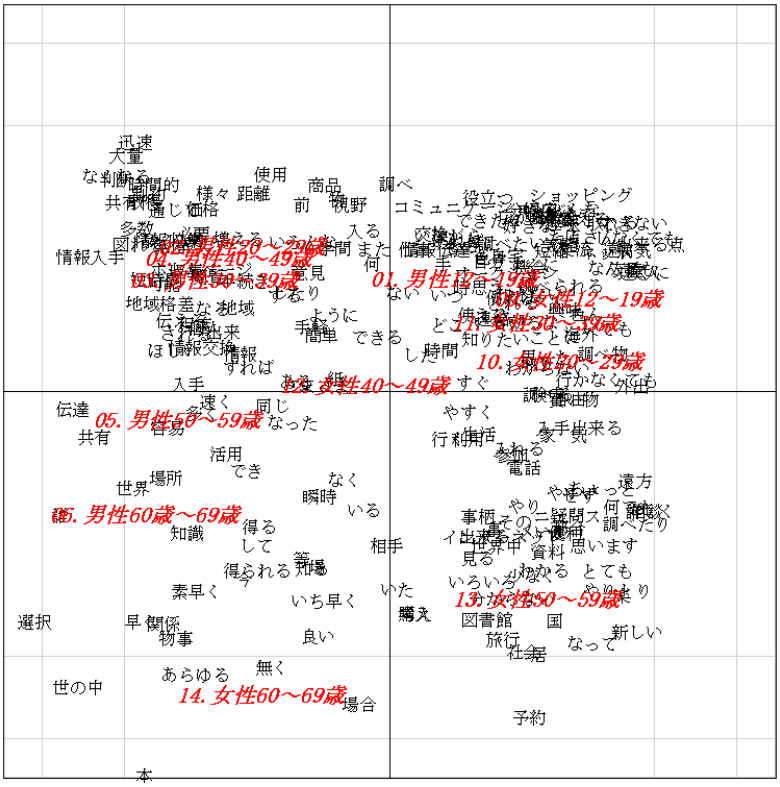


図 32　同時布置図の別例（図 31 の尺度を意図的に変形して描いた図）

表 59 は，WordMiner が出力した情報の一部である．WordMiner には「頻度による有意性テスト」という機能がある[97]．これを用いると，図 33 あるいは表 59 のような情報が得られる．これは，各性年齢区分に含まれる回答者が，自由回答の中で（全体の平均的傾向からみて）“どのような語句を積極的に用いたか”あるいは反対に“あまり用いていないか”を，それぞれ「上位」「下位」に有意かどうかを判定して，要約した情報である．

図 33 は WordMiner の出力そのものであるが，これをもう一度見易く整理した情報が表 59 である．この表で，たとえば区分（選択肢）「01.男性 12～19 歳」に属する回答者は，ここの上位にある「できない」「情報伝達」「交換」「しなくても」「外出」「距離」「無く」「わからない」「機会」「遠く」「入手」といった語句が前向きに用いられたと読む．「01.男性 12～19 歳」ではどうであろうか．ここでは有意となった語句数が多く，上位には「情報」「共有化」「必要」「物事」，…，「コミュニケーション」，…，「タイムリー」，…，「情報入手」…と続く．また下位には「人」「知りたい」「瞬時」「買い物」…とある．

では女性はどうであろうか．区分「09.女性 12～19 歳」では，「交流」「色々」「気軽」「出来る」「地域」…とあって男性とはすこし異なる語句が上位にある．男性同様，もっとも構成要素数が多い区分「11.女性 30～39 歳」の上位，下位の登場語句も男性のそれとは傾向が異なる．じつはこうした観察は，図 31，図 32 の同時布置図での観察をより客観的に行っていることになる（寄与度の傾向も考慮して）．各年齢区分の上位，下位に登場の語句を観察すると，性年齢区分と自由回答の意見の違いが見えてくるであろう．

しかし分析者にとっては，これでもまだ不満が残るだろう．ここでは省略するが，この段階からもう一歩踏み込んで，個々の回答者がここで登場した語句をそれぞれの自由回答文の中で，どのように用いたかを知りたい．じつはこの情報も WordMiner は出力してくれるが，ここでは紙幅の都合で（文字の量が多い！），省略したが，別の事例紹介記事にこれの一部をまとめたのでそれを参照していただきたい[98]．

97　これについては，別に用意した資料や分析事例を参照のこと．
98　資料「質的データのマイニングと対応分析法事例［2015］.pdf」として用意した．

	01.男性12～19歳 サンプル数：36 異なり構成要素数：36	02.男性20～29歳 サンプル数：28 異なり構成要素数：111	03.男性30～39歳 サンプル数：55 異なり構成要素数：112	04.男性40～49歳 サンプル数：49 異なり構成要素数：115	05.男性50～59歳 サンプル数：33 異なり構成要素数：94	06.男性60歳～69歳 サンプル数：28 異なり構成要素数：85	09.女性12～19歳 サンプル数：20 異なり構成要素数：60	10.女性20～29歳 サンプル数：56 異なり構成要素数：160	11.女性30～39歳 サンプル数：119 異なり構成要素数：144	12.女性40～49歳 サンプル数：61 異なり構成要素数：144	13.女性50～59歳 サンプル数：36 異なり構成要素数：112	14.女性60～69歳 サンプル数：11 異なり構成要素数：57
上位 1	できない	様々	情報	情報収集	知識	選択	交流	参加	調べられる	瞬時	なって	本
上位 2	情報伝達	迅速	共有化	必要	なる	世界	色々	人	病気	ほしい	居	場合
上位 3	交換	商品	必要	物	世の中	早く	気軽	相談	時間	時間	便利	予約
上位 4	しなくても	知識	物事	なくなる	閲覧	情報	出来る	思った	趣味	関係	やりとり	出来る
上位 5	外出	する	なった	時間的	共有	誰	地域	友人	好きな	なく	楽	して
上位 6	距離	判断	欲しい	情報	図れる	容易	人	機会	家	入れる	事	関係
上位 7	無く	使用	好きな	迅速	入手	入手	調べ物	興味	どんな	タイムリー	国	知識
上位 8	わからない	して	コミュニケーション	広がる	伝達	伝達	便利	いても	知りたい	何	新しい	居
上位 9	機会	取得	容易	程度	情報交換	得られる	できない	やすい	友達	価格	旅行	等
上位 10	遠く	調べ	タイムリー	点	増える	個人	なんでも	ネット	たくさん	活用	瞬時	早く
上位 11	入手	情報	制約	入手	地域	する	ショッピング	行かなくても	取れる	資料	見る	得られる
上位 12		もの	前	欲しい	地域格差	多く	出来る点	自分	せず	手間	少なく	インターネット
上位 13		発信	大量	さまざま	共有化	場所	詳しく	検索	時	手	メール	あらゆる
上位 14		ある	すぐ	出来	事	なる	図書館	すれば	すぐに	したり	わかる	購入
上位 15		手	情報入手	短時間	ホームページ	簡単	速く	オークション	気軽	利用	せず	社会
上位 16		短時間	通じて		簡単	思う	人たち	ニュース	思う	交換	とても	図書館
上位 17			可能		容易	可能	離れた	遠くに	いろいろな	手続き	無く	世の中
上位 18			情報収集					疑問	何でも	色ん	得られる	いた
上位 19			入る					取り	幅広い	通じて	いる	いろいろ
上位 20			手軽					ときに	気	買い物	電話	物事
上位 21			いる					友達	調べる		予約	
上位 22								例えば	買い物		すぐ	
上位 23								場合	いて		知る	
上位 24								色ん	いながら		良い	
上位 25								思う			気	
上位 26								やり				
上位 27								調べたい				
下位 20									欲しい			
下位 19									して			
下位 18									意見			
下位 17									いつ			
下位 16									必要			
下位 15									ある			
下位 14									機会			
下位 13									程度			
下位 12									入れる			
下位 11									入手			
下位 10									瞬時			
下位 9									地域			
下位 8									良い			
下位 7									なる			
下位 6			人					早く	得られる	気		
下位 5			知りたい					入手	関係	事		
下位 4			瞬時	調べる		事		いながら	世界	検索	情報収集	
下位 3		調べる	買い物	出来る		メール		得る	多く	家	情報	
下位 2		時間	いろいろな	調べられる	する	手		時間	情報	容易	手	
下位 1		便利	できる	人	手	すぐに	時間	情報	早く	すぐ	自分	

図 33　表 59 の引用元の表（WordMiner の出力情報）

表 59　性年齢区分別にみた自由回答における特徴的な発語（男性：上位，下位 20 位まで）［有意性テストの結果から抜粋］

性年齢区分	01.男性 12～19 歳	02.男性 20～29 歳	03.男性 30～39 歳	04.男性 40～49 歳	05.男性 50～59 歳	06.男性 60 歳～69 歳
異なり構成要素数	36	111	112	115	94	85
上位 1	できない	様々	情報	情報収集	知識	選択
上位 2	情報伝達	迅速	共有化	必要	なる	世界
上位 3	交換	商品	必要	物	世の中	早く
上位 4	しなくても	知識	物事	なくなる	閲覧	情報
上位 5	外出	する	なった	時間的	共有	誰
上位 6	距離	判断	欲しい	情報	図れる	容易
上位 7	無く	使用	好きな	迅速	入手	入手
上位 8	わからない	して	コミュニケーション	広がる	伝達	伝達
上位 9	機会	取得	容易	程度	情報交換	得られる
上位 10	遠く	調べ	タイムリー	点	増える	個人
上位 11	入手	情報	制約	入手	地域	する
上位 12		もの	前	欲しい	地域格差	多く
上位 13		発信	大量	さまざま	共有化	場所
上位 14		ある	すぐ	出来	事	なる
上位 15		手	情報入手	短時間	ホームページ	簡単
上位 16		短時間	通じて		簡単	思う
上位 17			可能		容易	可能
上位 18			情報収集			
上位 19			入る			
上位 20			手軽			
下位 6			人			
下位 5			知りたい			
下位 4			瞬時	調べる		事
下位 3		調べる	買い物	出来る		メール
下位 2		時間	いろいろな	調べられる	する	手
下位 1		便利	できる	人	手	すぐに

表 59（つづき）

性年齢区分	09.女性 12～19 歳	10.女性 20～29 歳	11.女性 30～39 歳	12.女性 40～49 歳	13.女性 50～59 歳	14.女性 60～69 歳
異なり構成要素数	60	160	187	144	112	57
上位 1	交流	参加	調べられる	瞬時	なって	本
上位 2	色々	人	病気	ほしい	居	場合
上位 3	気軽	相談	時間	時間	便利	予約
上位 4	出来る	思った	趣味	関係	やりとり	出来る
上位 5	地域	友人	好きな	なく	楽	して
上位 6	人	機会	家	入れる	事	関係
上位 7	調べ物	興味	どんな	タイムリー	国	知識
上位 8	便利	いても	知りたい	何	新しい	居
上位 9	できない	やすい	友達	価格	旅行	等
上位 10	なんでも	ネット	たくさん	活用	瞬時	早く
上位 11	ショッピング	行かなくても	取れる	資料	見る	得られる
上位 12	出来る点	自分	せず	手間	少なく	インターネット
上位 13	詳しく	検索	時	手	メール	あらゆる
上位 14	図書館	すれば	すぐに	したり	わかる	購入
上位 15	速く	オークション	気軽	利用	せず	社会
上位 16	人たち	ニュース	思う	交換	とても	図書館
上位 17	離れた	遠くに	いろいろな	手続き	無く	世の中
上位 18		疑問	何でも	色ん	得られる	いた
上位 19		取り	幅広い	通じて	いる	いろいろ
上位 20		ときに	気	買い物	電話	物事
下位 20			欲しい			
下位 19			して			
下位 18			意見			
下位 17			いつ			
下位 16			必要			
下位 15			ある			
下位 14			機会			
下位 13			程度			
下位 12			入れる			
下位 11			入手			
下位 10			瞬時			
下位 9			地域			
下位 8			良い			
下位 7			なる			
下位 6		早く	得られる	気		
下位 5		入手	関係	事		
下位 4		いながら	世界	検索	情報収集	
下位 3		得る	多く	家	情報	
下位 2		時間	情報	容易	手	
下位 1	時間	情報	早く	すぐ	自分	

④ 別の課題，分析の方向

ここまでは，表 41 にある「(構成要素変数) × (性年齢区分)」のデータ表の構造探査を進めてきた．これで，選んだ構成要素，つまり自由回答内で用いられた語句群（の一部）が，性年齢区分という質的変数とどう関連するかの傾向は見えてきた．しかし，分析者にとっては，この自由回答質問には，“さらに別の構造が潜在”するかもしれない，と考えるかもしれない（多くの場合，さまざまな方向から探査的，発見的に分析を進めるであろう）．

たとえば，5.2.1 節でこの調査データの概要を説明したときに，別に用意した選択肢型質問があると述べた．ここで「社会の移り変わり」や「IT の進歩」などに関する選択肢型質問があるのだが，上の性年齢区分に変えて，こうした質問文を用いるとどういう結果となるので

あろうか．あるいはまた，こうした質的変数を用いずに，「(回答者・サンプル）×（構成要素)」の形式の 2 元データ表から出発し，たとえば，回答者を分類し，個々の自由回答意見を“類型化”することが可能だろうか．これに対する解答は「可」である．これについても順をおって述べる．

6. むすび

この稿（第 I 部）では，主に対応分析法の基本的な考え方，概念，適用例について述べた．この対応分析法を基本に，さまざまな「**多次元データ解析の機能**」を実装化することは可能である．実際に，WordMiner や JMP スクリプトは，ここでみたようなさまざな処理機能を備えている．

とくに，テキスト・マイニングの観点からは，対応分析法の基本機能の他にさまざまな実用上の機能が必要となる．たとえば以下のような機能がある．

i） 分かち書きで得た構成要素変数（単語・語句，キーワード）が，質的変数，属性変数，クラスター化変数などとの関係でどのような構造を持つものかを有意性テストする機能[99]［たとえば，上の分析例 4 でみた表 59］.

ii） 回答・回答者のクラスター化，構成要素のクラスター化を行い，またここでも有意性テストを行うこと．

iii） 独自のクラスター化処理の方法を用いること（階層的分類法と非階層的分類法をハイブリッドした方式)．このとき，対応分析法の特性と階層的分類法・分割化型分類法のハイブリッドを行うなどの工夫．

ソフトウェアとしては，単に多次元データ解析手法を組み入れただけでは不十分であり，それらの特性を活かした多くの機能が必要である．WordMiner や JMP スクリプトによる処理は，こうした周辺機能にも配慮がなされていることが特徴である．

さらに，「第Ⅱ部」で対応分析法の基本数理（仕組み）について述べ，「第Ⅲ部」では，対応分析法の特性を活かした「クラスター化法」について述べる．

99 ここでは超幾何分布の正規分布近似を用いたテストを行う．詳しくはテキスト･マイニング研究会のホームページにアップロードの資料を参照．Lebart 他（1998）も参照のこと．

付録:演習問題

演習問題として，いくつかの人工データ表を用意した，表計算ソフトウェアや統計ソフトウェア（例：WordMiner や JMP スクリプト）への入力形式としてどのようなデータ表をつくればよいか，また実際にこれらのソフトで計算を行ったときに出力される諸統計量を観察して理解するヒントとしてほしい．

（ｉ）演習問題 1

次のような回答者数が 10，項目数が 2 のデータ表が得られた．このデータ表から，以下の問に答えよ．

問１：インジケータ行列のデータ表を作成せよ．

問２：多重クロス表（バート表）を作成せよ．また，（項目 *I*）×（項目 *J*）の 2 元クロス表を確認せよ．

表1　データ表

個体＼項目	*I*	*J*
1	1	1
2	1	2
3	2	2
4	1	1
5	1	2
6	2	3
7	1	1
8	2	3
9	2	3
10	1	3

個体＼項目	*I*	*J*
1	はい	満足
2	はい	ふつう
3	いいえ	ふつう
4	はい	満足
5	はい	ふつう
6	いいえ	満足でない
7	はい	満足
8	いいえ	満足でない
9	いいえ	満足でない
10	はい	満足でない

表 2　多重クロス表（バート表）

項目と選択肢		項目 *I*		項目 *J*		
		1	2	1	2	3
項目 *I*	1	6	0	3	2	1
	2	0	4	0	1	3
項目 *J*	1	3	0	3	0	0
	2	2	1	0	3	0
	3	1	3	0	0	4

（ii）演習問題 2

次のような，回答者数が 20，3 項目の質問からなる（行列 $\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$ で表す）データ表がある．これについて，次の問に答えよ．

問 1：このデータ表から，**インジケータ行列**のデータ表を生成せよ．
問 2：**多重クロス表**を作成せよ．また，各項目のクロス表がどのように現れるかを観察せよ．

表3　元のデータ表

回答者	項目 A_1	項目 A_2	項目 A_3
1	1	1	1
2	1	1	2
3	1	2	1
4	2	1	2
5	2	1	2
6	2	1	3
7	2	2	1
8	2	2	2
9	3	1	2
10	3	1	3
11	3	1	3
12	3	2	1
13	3	2	1
14	3	2	2
15	3	2	2
16	3	2	3
17	4	1	3
18	4	2	2
19	4	2	2
20	4	2	3

表4　インジケータ行列のデータ表

項目と選択肢 / 回答者	項目 A_1			項目 A_2			項目 A_3		
	1	2	3	4	1	2	1	2	3
1	1	0	0	0	1	0	1	0	0
2	1	0	0	0	1	0	0	1	0
3	1	0	0	0	0	1	1	0	0
4	0	1	0	0	1	0	0	1	0
5	0	1	0	0	1	0	0	1	0
6	0	1	0	0	1	0	0	0	1
7	0	1	0	0	0	1	1	0	0
8	0	1	0	0	0	1	0	1	0
9	0	0	1	0	1	0	0	1	0
10	0	0	1	0	1	0	0	0	1
11	0	0	1	0	1	0	0	0	1
12	0	0	1	0	0	1	1	0	0
13	0	0	1	0	0	1	1	0	0
14	0	0	1	0	0	1	0	1	0
15	0	0	1	0	0	1	0	1	0
16	0	0	1	0	0	1	0	0	1
17	0	0	0	1	1	0	0	0	1
18	0	0	0	1	0	1	0	1	0
19	0	0	0	1	0	1	0	1	0
20	0	0	0	1	0	1	0	0	1

インジケータ行列のデータ表と多重クロス表の関係は，次のように 2 つのデータ表を結合して考えると分かり易いだろう．

表5　インジケータ行列のデータ表と多重クロス表の関係

項目と選択肢 / 回答者	項目 A_1				項目 A_2		項目 A_3		
	1	2	3	4	1	2	1	2	3
1	1	0	0	0	1	0	1	0	0
2	1	0	0	0	1	0	0	1	0
3	1	0	0	0	0	1	1	0	0
4	0	1	0	0	1	0	0	1	0
5	0	1	0	0	1	0	0	1	0
6	0	1	0	0	1	0	0	0	1
7	0	1	0	0	0	1	1	0	0
8	0	1	0	0	0	1	0	1	0
9	0	0	1	0	1	0	0	1	0
10	0	0	1	0	1	0	0	0	1
11	0	0	1	0	1	0	0	0	1
12	0	0	1	0	0	1	1	0	0
13	0	0	1	0	0	1	1	0	0
14	0	0	1	0	0	1	0	1	0
15	0	0	1	0	0	1	0	1	0
16	0	0	1	0	0	1	0	0	1
17	0	0	0	1	1	0	0	0	1
18	0	0	0	1	0	1	0	1	0
19	0	0	0	1	0	1	0	1	0
20	0	0	0	1	0	1	0	0	1

インジケータ行列のデータ表

項目	選択肢	1	2	3	4	1	2	1	2	3
項目 A_1	1	3	0	0	0	2	1	2	1	0
	2	0	5	0	0	3	2	1	3	1
	3	0	0	8	0	3	5	2	3	3
	4	0	0	0	4	1	3	0	2	2
項目 A_2	1	2	3	3	1	9	0	1	4	4
	2	1	2	5	3	0	11	4	5	2
項目 A_3	1	2	1	2	0	1	4	5	0	0
	2	1	3	3	2	4	5	0	9	0
	3	0	1	3	2	4	2	0	0	6

多重クロス表（バート表）（表 6）に同じ

表 6　多重クロス表（バート表）

	項目	項目 A_1				項目 A_2		項目 A_3		
項目	選択肢	1	2	3	4	1	2	1	2	3
項目 A_1	1	3	0	0	0	2	1	2	1	0
	2	0	5	0	0	3	2	1	3	1
	3	0	0	8	0	3	5	2	3	3
	4	0	0	0	4	1	3	0	2	2
項目 A_2	1	2	3	3	1	9	0	1	4	4
	2	1	2	5	3	0	11	4	5	2
項目 A_3	1	2	1	2	0	1	4	5	0	0
	2	1	3	3	2	4	5	0	9	0
	3	0	1	3	2	4	2	0	0	6

＜補足＞

ここで，表 6 の多重クロス表（バート表）に対応分析を適用して得られる結果と，インジケータ行列のデータ表から得た結果は同等であることが知られている．また，表 6 のバート表で得た結果に対して，表 5 の上部にあるインジケータ行列のデータ表を（回答者側の）追加処理要素として扱うと回答者の成分スコアを求めることができる．すなわち，多重クロス表の寸法の行列の演算処理が可能であればかなり大規模な回答者であっても成分スコアの算出が可能ということである．

日本国内における数量化法 III 類の利用環境では，大抵の場合，インジケータ行列のデータ表を扱うことが多い．またそのように数量化法 III 類を使うとの記述も散見する（若干の誤解があるかもしれない）．しかし，いったんバート表の形で集計分析し（コンピュータ内でインコア処理し）得られた成分スコアの式を用いて，回答者の成分スコアを外部メモリー（ハードディス内）で処理するほうが効率的であり，回答者数の制約を受けずに済むという利点がある[100]．

ここで，インジケータ行列から出発した場合と，（バート表）から出発した場合とにどのような関係にあるかを付表に要約した[101]．とくに，2 項目の場合には，バート表と 2 元クロス表とから得た結果の間には，かなりはっきりした関係が分かっている．これらの情報を知って，2 元データ表を使い分けることは対応分析の長所である．

つまり，この要約表に共通することは，いずれもが**“2 元データ表”**であるということである，クロス表，インジケータ行列のデータ表，そして多重クロス表を使った対応分析の結果には，お互いにある関係があること，とくにインジケータ行列のデータ表と多重クロス表との結果はじつは同じ内容となっていることを示している．このデータ表間の関係を知って分析を進めることは重要である．また，いままでにみてきたように，構成要素変数，質的変数をうまく組み合わせることでインジケータ行列とクロス表データに対応させることも可能である．

2 元データ表は，さらに順位データ（ランキング・データ）や反対変数（counter variables）あるいは重複化変数（doubling variables）といった使い方もある．こうした使い方や意味については関連書を参照するとよいだろう（たとえば，Greenacre（2007），Jambu（1989），Volle（1985）など）．

100　実際にこうした手順を組み込むことで，実質的にはバート表の寸法だけに依存する処理量で済み，サンプル数の大きさに依存しないで成分スコアの算出も可能になる．

101　こうした関係が得られることは「第Ⅱ部」で説明した．そこからの引用．

付表　2 元データ表の相互の関係

タイプ	データ表の形	データ表の次元数 （寸法）	固有値の関係
タイプ 1	2 項目 I と J のクロス表 $\underset{n_i \times n_j}{\mathbf{F}} = \mathbf{A}_i^t \mathbf{A}_j$ または $\underset{n_j \times n_i}{\mathbf{B}^*} = \mathbf{A}_j^t \mathbf{A}_i$	$n_i \times n_j$ (*) クロス表の寸法を表す前に用いた記号では， $m=n_i, n=n_j$ と対応 (*) 固有値の個数は， $K = \min\{n_i, n_j\} - 1$	ここで得る固有値を λ_k^F で表す
タイプ 2	2 項目 I と J のインジケータ行列のデータ表 $\underset{N \times n^*}{\mathbf{A}} = \left[\underset{N \times n_i}{\mathbf{A}_i}, \underset{N \times n_j}{\mathbf{A}_j} \right]$ ここで $\left(n^* = n_i + n_j\right)$	$N \times n^* \ \left(n^* = n_i + n_j\right)$	$\lambda_k^A = \dfrac{1 \pm \sqrt{\lambda_k^F}}{2}$
	タイプ 1 とタイプ 2 の固有値の関係 i）値の大きい方から $n_i - 1$ 個 ⇒ $\lambda_k^A = \dfrac{1+\sqrt{\lambda_k^F}}{2}$ ii）値の小さい方から $n_j - 1$ 個 ⇒ $\lambda_k^A = \dfrac{1-\sqrt{\lambda_k^F}}{2}$ iii）間に含まれる $n_j - n_i$ 個 ⇒ 1/2 = 0.5 となる．(*) $n_i = n_j$ のときにはこれは現れない．		
タイプ 3	2 項目の多重クロス表（バート表） $\underset{n^* \times n^*}{\mathbf{B}} = \underset{n^* \times N}{\mathbf{A}^t} \underset{N \times n^*}{\mathbf{A}} \quad \left(n^* = n_i + n_j\right)$	$n^* \times n^* \ \left(n^* = n_i + n_j\right)$	$\lambda_k^B = \left(\lambda_k^A\right)^2 = \left(\dfrac{1 \pm \sqrt{\lambda_k^F}}{2}\right)^2$
タイプ 4	一般の M 項目のインジケータ行列 $\underset{N \times n^*}{\mathbf{A}} = \left[\underset{N \times n_1}{\mathbf{A}_1}, \underset{N \times n_2}{\mathbf{A}_2}, \cdots, \underset{N \times n_i}{\mathbf{A}_i}, \cdots, \underset{N \times n_j}{\mathbf{A}_j}, \cdots, \underset{N \times n_M}{\mathbf{A}_M} \right]$ ここで $n^* = \sum_{j=1}^{M} n_j$	$N \times n^* \ \left(n^* = \sum_{j=1}^{M} n_j\right)$ (*) 固有値の個数は， $K^* = \sum_{j=1}^{M} \left(n_j - 1\right) = n^* - M$	λ_k^A $\lambda_k^A = \sqrt{\lambda_k^B}$
タイプ 5	M 項目の多重クロス表（バート表） $\underset{n^* \times n^*}{\mathbf{B}} = \underset{n^* \times N}{\mathbf{A}^t} \underset{N \times n^*}{\mathbf{A}}$ ここで $n^* = \sum_{j=1}^{M} n_j$	$n^* \times n^* \ \left(n^* = \sum_{j=1}^{M} n_j\right)$ (*) 固有値の個数は， $K^* = \sum_{j=1}^{M} \left(n_j - 1\right) = n^* - M$	λ_k^B $\lambda_k^B = \left(\lambda_k^A\right)^2$

【参考文献】

[1] ウヴェ・フリック著，小田博志，山本則子，春日常，宮地尚子訳（2004），質的研究入門－＜人間の科学＞のための方法論，春秋社．（*）下の［37］の翻訳版

[2] 岩坪秀一（1987）：数量化法の基礎，朝倉書店（*）数量化法について丁寧に解説された書．

[3] 舟島なおみ（2000），質的研究への挑戦，医学書院．

[4] 森本栄一（2005），戦後日本の統計学の発達－数量化理論の形成から定着へ－，行動計量学，第32巻，第1号（通巻62号），45-67．

[5] 西里静彦（2007）:データ解析への洞察，K・G りぶれっと（No.18），関西学院大学出版会．
（*）小冊子ではあるが，計量心理学の立場から尺度化の立場から数量化のあり方について平易かつ的確に述べられている．

[6] 大隅昇(2004)，「調査環境の変化に対応した新たな調査法の研究」報告．

[7] 大隅昇，L. Lebart，他（1994），記述的多変量解析法，日科技連出版社．

[8] 大隅昇（1989），統計的データ解析とソフトウェア，日本放送出版協会．

[9] 大隅昇（監訳）（2011），調査法ハンドブック，朝倉書店．（*）下の［46］の全訳版．

[10] 樋口耕一（2004），計量テキスト分析の方法と実践，大阪大学大学院人間科学研究科，博士論文．

[11] 樋口耕一（2014），社会調査のための計量テキスト分析―内容分析の継承と発展を目指して，ナカニシヤ出版．

[12] 林知己夫（1993），数量化－理論と方法，朝倉書店．

[13] 林知己夫（1996），データ解析からデータサイエンスへ－科学としてのデータを語る，データウェアハウスがビジネスを変える，日経 BP ムック．

[14] 林知己夫（2000），これからの国民性研究－人間研究の立場と地域研究・国際比較研究から－，統計数理，第48巻，第1号，33-66．［http://artemis.ism.ac.jp/proc/pdf/48-1-033.pdf］

[15] 林知己夫（2000），反時代的考察，市場調査 No. 244，（2000年7月）4-17．

[16] 林知己夫（2001），データの科学，シリーズ＜データの科学＞1，朝倉書店．

[17] 萱間真美（2013），質的研究のピットフォール－知らないために，抜け出るために，医学書院．

[18] Benzécri, J.-P. (1976) : *L'Analyse de Données, Tome 1*: *Taxinomie*, *Tome II*: *L'Analyse des Correspondances*, Dunod (second edtion).

[19] Benzécri, J.-P. (1980) : *Pratique de L'Analyse des Données – Analyse des Correspondances Exposé Elémentaire -, Tome I*, Dunod.

[20] Benzécri, J.-P. (1980) : *Pratique de L'Analyse des Données – Linguistique et Lexicologie -, Tome III*, Dunod.（*）この *Pratique de L'Analyse des Données* は3巻からなる．とくにこの第Ⅲ巻では，言語学・語彙研究の事例が紹介されている．

[21] Benzécri, J.-P. (1982), *Historie et Préhistoire de i'Analyse des Données*, Dunod.

[22] Benzécri, J.-P. (1992), *Correspondence Analysis Handbook*, Marcel Dekker.

[23] Berry, M.W. (1992) : Large-Scale Sparse Singular Value Computations, *The International Journal of Supercomputer Applications*, **6**, 1, 13-49.

[24] Bethlehem, J., and Biffignand, S. (2011): *Handbook of Web Surveys*, John Wiley and Sons.

[25] Bruynooghe, M. (1978): Classification Ascendante Hiérarchique de Grands Ensembles de Données, *Cahiers de l'Analyse des Données*, Vol. 3, No. 1, 35-46.

[26] Carroll, J.D., Green, P.E., and Schaffer, C.M. (1986): Interpoint distance comparisons in correspondence analysis, *Journal of Marketing Research*, vol. 23, 271-280.

[27] Carroll, J.D., Green, P.E., and Schaffer, C.M. (1987): Interpoint distance comparisons in correspondence analysis: A clarification, *Journal of Marketing Research*, vol. 24, 445-450.

[28] Carroll, J.D., Green, P.E., and Schaffer, C.M. (1989): Reply to Greenacre's commentary on the Carroll-Green-Schaffer scaling of two-way correspondence analysis solutions, *Journal of Marketing Research*, vol. 26, 366-368.

[29] Chatfield, C. (1995), *Problem Solving - A Statistical Guide*, second edition, Chapman & Hall.

[30] Clausen, Sten-Erik (1998): *Applied Correspondence Analysis*, Series: Quantitative Applications in the Social Sciences No.121, Sage Publications, Inc.

[31] Deewester, S. Dumais, S.T., Fumas, G.W., Landauer, T.K., and Harshman, R. (1990): Indexining by Layent Semantic Analysis, *Jornal of the American Society for Information Science*, **41**(6), 391-407.

[32] Douglas, J., Green, F.E., and Schaffer, C.M. (1986): Interpoint Distance Comparisons in Correspondence Analysis, *Jornal of Marketing Research*, Vol. XXIII, August, 271-280.

[33] Douglas, J., Green, F.E., and Schaffer, C.M. (1987): Comparing Interpoint Distances in Corepondence Analysis : A Clarification, *Jornal of Marketing Research*, Vol. XXIV, November, 445-450.

[34] Escofier, B. And Pages, J. (1990) : *Analyses Factorielles Simples et Multiples*, Dunod.

[35] Everitt, B.S. (1977, 1992), *The Analysis of Contingency Tables*, second edition, Chapman and Hall.

[36] Everitt, B.S. and Dunn, G.(2001): *Applied Multivariate Data Analysis*, second edition, Arnold.
[37] Faberge, J.-M. (1970-71): L'Analyse des Profils, Cours de J.M. Faverge, Bruxelles.（講義録）
[38] Faberge, J.-M. (1977): Correspondance Entre L'Analyse Binaire Classique et L'Analyse de Benzécri, Année Psycol., 149-160.
[39] Flick, U.(2002), *An Introduction to Qualitative Research* , second edition, Sage Publications.
[40] Gabriel, K. R. (2002), Goodness of fit of biplots and correspondence analysis, *Biometrika,* **89**(2), pp.423-436.
[41] Gauch, H.G. et al. (1977): A Comparative Study of Reciprocal Averaging and Other Ordination, Techniques, *J. Ecol.* **65**, 157-174.
[42] Goodman, L.A. (1986): Some Useful Extensions of Usual Correspondence Analysis Approach and the Usual Log-Linear Models Approach in the Analysis of Contingency Tables, *International Statistical Review*, **54**, 3, 243-309.
[43] Greenacre, M.J. (1984): *Theory and Applications of Correspondence Analysis*, Academic Press.
[44] Greenacre, M.J. (1989): the Carroll-Green-Schaffer scaling in correspondence analysis: A theoretical and empirical appraisal, Journal of Marketing Research, vol. 26, 358-365.
[45] Greenacre, M.J. (2007): *Correspondence Analysis in Practice* (second edition), Academic Press.
[46] Groves, R.M. and others (2004), *Survey Methodology*, John-Wiley.
[47] Hill, M.O. (1973): Reciprocal Averaging –An Eigen Vector Method of Ordination, *J. Ecol*. **61**, 237-249.
[48] Hill, M.O. (1974): Correspondence Analysis: A Neglected Multivariate Method, *J. Roy. Stat. Soc. Ser. C*, **23**, 340-354.
[49] Hudon, G. (1990): Une Comparison des Resultats de Modeles Lo—lineaires et de Generalisations de l'Analyse des Correspondances, *Rev. Statistique Appliquee*, XXXVIII (2), 43-53.
[50] Israels, A. (1987): *Eigenvalue Techniques for Qualitative Data*, DSWO Press, Leiden.
[51] Jackson, J.E. (1991, 2003): *A User's Guide to Principal Components*, John Wiley & Sons.
[52] Jambu, M. (1989): *Exploration Informatique et Statistique des Données*, Dunod.
[53] Juan, Programme de Classification Hierarchique par l'Algorithme de la Recherche en Chaine des Voisins Reciproques, *Le Cahiers de l'Analyse des Données*, VII, 2, 219-226.
[54] Lauro, N.C. and Decarli, A. (1982): Correspondence Analysis and Log-linear Models in Multiway Contingency Tables Study Some Remarks on Experimental Data, *Metron*, **40**, 213-234.
[55] Le Roux, B. and Rouanet, H. (2004): *Geometrical Data Analysis – From Correspondence Analysis to Structural Data*, Dordrecht Kluwer.
[56] Le Roux, B. and Rouanet, H. (2010): *Multiple Correspondence Analysis*, Series: Quantitative Applications in the Social Sciences No.163, Sage Publications, Inc.
[57] Lebart, L., Salem, A. and Berry, L. (1998), *Exploring Textual Data,* Kluwer Academic Publishers.
[58] Ling, R. and Pratt, J.W. (1984): The Accuracy of Peizer Approximations to the Hypergeometric Distribution, with Comparisons Some Other Approximations, *Journal of the American Statistical Asociation,* **79**, 385, 49-60.
[59] Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979): *Multivariate Analysis*, Academic Press.
[60] Maulman, J. (1982): *Homogeneity Analysis of Incomplete Data*, DSWO Press, Leiden.
[61] Murtagh, F.D. (1985): A Survey of Algorithms for Contiguity-constraned Clustering and Related Problems, *The Computer Journal*, **28**, 82-88.
[62] Murtagh, F.D. (1985): *Multidimensional Clustering Algorithms*, COMPSTAT Lectures 4, Physica Verlag.
[63] Nishisato, S. (1980): *Analysis of Categorical Data; Dual Scaling and its Applications,* University of Toronto Press.
[64] Rizzi, A. (ed.) (1995): *Some Relations between Matrices and Structures of Multidimensional Data Analysis*, Giardini Editore e Stampatori in Pisa.
[65] Volle, M. (1985): *Analyse des Données*, 3eme édition, Economica.

（※）以下のテキスト・マイニング研究会のホームページに，対応分析法を巡る話題，WordMiner を使った事例文献，日本語関連やテキスト・マイニング関連の文献，そして WordMiner 利用上のヒント，活用セミナー・テキスト情報，調査方法論関連文献などが掲載されているので参考にするとよい.

テキスト・マイニング研究会　URL：http://wordminer.org/

第Ⅱ部

対応分析法の基本的な考え方（数理の要点）

大隅　昇

1. 対応分析の仕組み

対応分析法（AFC：Analyse Factorielle des Correspondances，Analyse des Correspondances）はフランスの研究者であるベンゼクリ（J.-P. Benzécri）により提唱された多次元データ解析の手法の１つである[1]．これは，それより早くに林知己夫により提唱されたパターン分類（数量化法 III 類）と同等手法である．また，類似の手法も多数ある．しかしここでは**対応分析法の考え方**に従ってその仕組みを簡単に述べる．また，数量化法Ⅲ類との関係は後述の例題の中で説明する．

1.1 記号と記法の準備

対応分析法では，出発行列として“**2 元データ表**”（two-way data table）を想定する[2]．ここでいう 2 元データ表とは，以下のような比較的ゆるやかな条件を満たすデータ表をいう．

- **2 元**（two-way）の行列形式となっていること
- 各要素（セル）内の数値は**“非負の値”**であること
- 行あるいは列の比率（割合）パターン（これを“**プロファイル**”という）を考える意味があるような場合
- あるいはそれに相当する場面を想定できる 2 元データ表

われわれが日常的にさまざまな場面で遭遇するデータ表は，こうした要件を満たすことが多いので，それだけ対応分析法の利用の自由度が高い．典型的な例として，**“クロス表”**（分割表）を考えればよい[3]．ここでもこれを基本のデータ表として説明する．いま寸法が$(m\times n)$の **2 元クロス表型データ表**を以下のように表す．ここで，f_{ij} はクロス表の(i,j)セル内の度数（ないしはそれに相当の非負の数値）である．

$$\underset{m\times n}{\mathbf{F}}=\left(f_{ij}\right)\qquad\begin{pmatrix}f_{ij}\geq 0\\ i\in I,\ j\in J\end{pmatrix}\tag{1}$$

ここで，I と J は，それぞれ行と列の項目の各要素（標識）の集合[4]を表わし，これを以下のように表す．

$$I=\{1,2,\cdots,i,\cdots,m\},\quad J=\{1,2,\cdots,j,\cdots,n\}\tag{2}$$

これらは，クロス表でいえば**質問項目**とその**選択肢**に相当する（表 1 参照）．ここでは，「項目」とその「選択肢」という表現を用いる．

つぎに寸法が（$m\times n$）の 2 元クロス表（表 1）から作られる**相対度数**（経験確率分布）を考える．そしてこれに関連する以下の行列，ベクトルを用意する．

$$\underset{m\times n}{\mathbf{P}_{IJ}}=\frac{1}{N}\underset{m\times n}{\mathbf{F}}=\left(p_{ij}\right)\quad\left(i\in I,\ j\in J\right)\text{（同時確率分布）}\tag{3}$$

---

[1] J.-P. Benzécri は，1962 年頃に始めて、この対応分析法を言語学研究のデータ解析に使ってみせた．一方．林知己夫は，それよりずっと早い 1952 年にパターン分類法として提唱し，市場調査分野で実際に使ってみせた（デザインの評価）．両者の考え方，アプローチが異なることで，かつては同じ手法ということがよく理解されなかった．Benzécri と林知己夫の交誼などもあって，両国の研究交流が始まり，いまも続いている．

[2] 対応分析法の研究では，この 2 元データ表を基礎として，さまざまなデータ表やデータ構造に適した無数の手法が提案されてきた．ここでは，もっとも基本となる上の条件を満たすような場合を想定して説明する．

[3] クロス表だけではなく，さまざまな 2 元データ表への適用を想定していることが特徴．

[4] フランス流の用語でいうと“フォルム（forme）あるいはモダリティ（modalité）の集合”のことをいう．モダリティの原義は「様式，様態」だが，ここではいわゆる**選択肢**あるいは**カテゴリー，オプション**のこと．

$$\underset{m\times m}{\mathbf{P}_I} = diag\left(p_{i+}\right) \quad \left(i \in I\right) \text{（行の周辺確率分布）} \tag{4}$$

$$\underset{n\times n}{\mathbf{P}_J} = diag\left(p_{+j}\right) \quad \left(j \in J\right) \text{（列の周辺確率分布）} \tag{5}$$

ここで，

$$p_{ij} = \frac{f_{ij}}{N} \tag{6}$$

$$p_{i+} = \frac{f_{i+}}{N} = \frac{\sum_{j=1}^{n} f_{ij}}{N} \tag{7}$$

$$p_{+j} = \frac{f_{+j}}{N} = \frac{\sum_{i=1}^{m} f_{ij}}{N} \tag{8}$$

である．また，$N = \sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} \left(\equiv f_{++}\right)$（総度数）である．

$diag(\bullet)$は対角行列を意味する．たとえば，$\underset{m\times m}{\mathbf{P}_I} = diag\left(p_{i+}\right)$は対角要素が p_{i+} で非対角の要素はすべて「0」の，寸法が$m\times m$の正方行列であり，$\underset{n\times n}{\mathbf{P}_J} = diag\left(p_{+j}\right)$ は対角要素が p_{i+} で非対角の要素はすべて「0」の，寸法が$n\times n$の正方行列である．

$$\underset{m\times m}{\mathbf{P}_I} = diag\left(p_{i+}\right) = \begin{pmatrix} p_{1+} & & & \mathbf{O} & \\ & p_{2+} & & & \\ & & \ddots & & \\ & & & p_{i+} & \\ & & & & \ddots \\ \mathbf{O} & & & & p_{m+} \end{pmatrix} \text{（}m\times m\text{の対角行列）} \tag{9}$$

$$\underset{n\times n}{\mathbf{P}_J} = diag\left(p_{+j}\right) = \begin{pmatrix} p_{+1} & & & \mathbf{O} & \\ & p_{+2} & & & \\ & & \ddots & & \\ & & & p_{+j} & \\ & & & & \ddots \\ \mathbf{O} & & & & p_{+n} \end{pmatrix} \text{（}n\times n\text{の対角行列）} \tag{10}$$

表 1（項目 I ×項目 J ）の 2 元クロス表 $\underset{m\times n}{\mathbf{F}}=\left(f_{ij}\right)$

		項　目　J						
	選択肢	1	2	…	j	…	n	行和
項目 I	1	f_{11}	f_{12}	…	f_{1j}	…	f_{1n}	f_{1+}
	2	f_{21}	f_{22}	…	f_{2j}	…	f_{2n}	f_{2+}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	i	f_{i1}	f_{i2}	…	f_{ij}	…	f_{in}	f_{i+}
	⋮	⋮	⋮		⋮		⋮	⋮
	m	f_{m1}	f_{m2}	…	f_{mj}	…	f_{mn}	f_{m+}
	列和	f_{+1}	f_{+2}	…	f_{+j}	…	f_{+n}	$f_{++}=N$

さらに，行の平均ベクトルと列の平均ベクトルをつぎのように表す．

（ｉ）行の平均ベクトル

$$\underset{m\times 1}{\mathbf{r}}=\left(p_{1+},p_{2+},\cdots,p_{i+},\cdots,p_{m+}\right)^{t}$$ （p_{i+} を要素とする列ベクトル）　　(11)

この各要素 p_{i+} を**行の質量**（mass）といい，この $\underset{m\times 1}{\mathbf{r}}$ はまた，**列プロファイルの平均ベクトル（重心）**に相当する．

（ⅱ）列の平均ベクトル

$$\underset{n\times 1}{\mathbf{c}}=\left(p_{+1},p_{+2},\cdots,p_{+j},\cdots,p_{+n}\right)^{t}$$ （p_{+j} を要素とする列ベクトル）　　(12)

一方，この各要素 p_{+j} を**列の質量**（mass）とし，$\underset{1\times n}{\mathbf{c}}$ は**行プロファイルの平均ベクトル（重心）**となる．

このとき，単位ベクトル $\mathbf{1}_n$, $\mathbf{1}_m$ を使うと，$\mathbf{r},\mathbf{c}$ は以下のように表される．

$$\underset{m\times 1}{\mathbf{r}}=\mathbf{P}_{IJ}\mathbf{1}_n,\quad \underset{n\times 1}{\mathbf{c}}=\mathbf{P}_{JI}\mathbf{1}_m \qquad (13)$$

$$\text{ここで,}\quad \mathbf{1}_n=\begin{pmatrix}1\\1\\\vdots\\1\\\vdots\\1\end{pmatrix}\Big\}(n\text{個の}1),\ \mathbf{1}_m=\begin{pmatrix}1\\1\\\vdots\\1\\\vdots\\1\end{pmatrix}\Big\}(m\text{個の}1)$$

また，$\mathbf{P}_{JI}=\mathbf{P}_{IJ}^{t}$，つまり，$\mathbf{P}_{IJ}$ の転置行列を表す．

以上を模式図に表すと表 2，図 1 のようになる．図 1 は，各記号を表に対応させて描いた模式図である．それぞれの対応をここで確認するとよい．

1．2 プロファイルとは

対応分析法では“**プロファイル**”（profile）の概念が重要である．**プロファイル**とはクロス表の行あるいは列の**相対比率（相対確率）のパターン**のことをいう．つまり，行と列のプロファイルがある．

（ｉ）行プロファイル（行の比率パターン）

$$\underset{m\times n}{\mathbf{N}_I}=\underset{m\times m}{\mathbf{P}_I^{-1}}\,\underset{m\times n}{\mathbf{P}_{IJ}}=\left\{q_{ij}=\frac{p_{ij}}{p_{i+}}\middle|\, i\in I, j\in J\right\}\quad\text{（行プロファイル）}\qquad(14)$$

ここで，$\sum_{j=1}^{n} q_{ij}=1$の制約があるから，行のプロファイルは$(n-1)$次元の空間に分布するm個の点の集合である（図 2，3）．

（ⅱ）列プロファイル（列の比率パターン）

$$\underset{n\times m}{\mathbf{N}_J}=\underset{n\times n}{\mathbf{P}_J^{-1}}\,\underset{n\times m}{\mathbf{P}_{JI}}=\left\{q_{ij}^{*}=\frac{p_{ij}}{p_{+j}}\middle|\, i\in I, j\in J\right\}\quad\text{（列プロファイル）}\qquad(15)$$

（ここで，$\mathbf{P}_{JI}=\mathbf{P}_{IJ}^{t}$は$\mathbf{P}_{IJ}$の転置行列を表す）

ここでは，$\sum_{i=1}^{m} q_{ij}^{*}=1$の制約から，列のプロファイルは$(m-1)$次元空間に分布する$n$個の点の集合である（図 2，3）．なお，$\mathbf{N}_I$，$\mathbf{N}_J$の**プロファイルの分布**のことを，フランス流には**“雲”**（nuage，英語の cloud）という[5]．

このプロファイルにはつぎの特徴がある．

- プロファイルとは行あるいは列の**比率のパターン**である．
- したがって，生のデータ（測定値，度数）そのものではない．
- 対応分析では，列や行の周辺和によって基準化する（長さをそろえる）．この基準化操作は，一般に，周辺和が小さいものほど，度数の変化に敏感に反応するようになっている[6]．
- そのため，列和や行和が小さい場合や，はずれ値などの影響を受けやすい[7]．

---

[5] ここで，プロファイルは，$q_{ij}=\frac{p_{ij}}{p_{i+}}=\frac{f_{ij}}{f_{i+}}$，$q_{ij}^{*}=\frac{p_{ij}}{p_{+j}}=\frac{f_{ij}}{f_{+j}}$ となることに注意しよう.つまり相対度数（相対確率）を考えていることに同じである．

[6] たとえば，クロス表の分析で，周辺和を 100（%）として，そろえて観測することを思い出そう．

[7] これに対する手当として，サブセット対応分析（SCA）や追加処理といった方法がある．

表 2　確率行列 $\mathbf{P}_{IJ} = \frac{1}{N}\mathbf{F}$

		項　目　J						
		1	2	…	j	…	n	行の確率（$\mathbf{P}_I$ の対角要素） $\mathbf{r}_{m\times 1}$（列プロファイルの重心）
項目 I	1	p_{11}	p_{12}	…	p_{1j}	…	p_{1n}	p_{1+}
	2	p_{21}	p_{22}	…	p_{2j}	…	p_{2n}	p_{2+}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	i	p_{i1}	p_{i2}	…	p_{ij}	…	p_{in}	p_{i+}
	⋮	⋮	⋮		⋮		⋮	⋮
	m	p_{m1}	p_{m2}	…	p_{mj}	…	p_{mn}	p_{m+}
	列の確率（$\mathbf{P}_J$ の対角要素） $\mathbf{c}^t_{1\times n}$（行プロファイル重心）	p_{+1}	p_{+2}	…	p_{+j}	…	p_{+n}	1
		（列の質量）						

（行の質量）

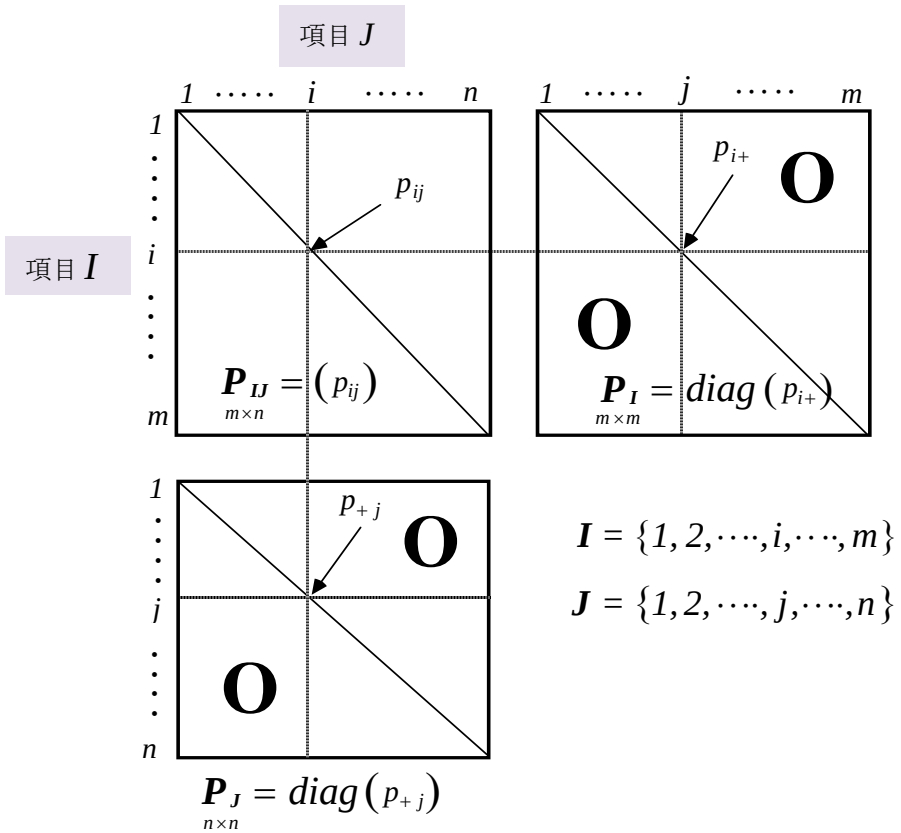


図1　確率行列の構成: $\mathbf{P}_{IJ}, \mathbf{P}_I, \mathbf{P}_J$ の関係（模式図）

I \ J	1	2	…	j	…	n	行の確率
1	p_{11}	p_{12}	…	p_{1j}	…	p_{1n}	p_{1+}
2	p_{21}	p_{22}	…	p_{2j}	…	p_{2n}	p_{2+}
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	p_{i1}	p_{i2}	…	p_{ij}	…	p_{in}	p_{i+}
⋮	⋮	⋮		⋮		⋮	⋮
m	p_{m1}	p_{m2}	…	p_{mj}	…	p_{mn}	p_{m+}
列の確率	p_{+1}	p_{+2}	…	p_{+j}	…	p_{+n}	1

＜行プロファイルの行列＞

	1	2	…	j	…	n	
1				⋮			1
2				⋮			1
⋮				⋮			⋮
i	p_{i1}/p_{i+}	p_{i2}/p_{i+}	…	p_{ij}/p_{i+}	…	p_{in}/p_{i+}	1
⋮				⋮			⋮
m				⋮			1

＜列プロファイルの行列＞

	1	2	…	j	…	n	
1				p_{1j}/p_{+j}			
2				p_{2j}/p_{+j}			
⋮				⋮			
i	…	…	…	p_{ij}/p_{+j}	…	…	
⋮				⋮			
m				p_{mj}/p_{+j}			
	1	1	…	1	…	1	

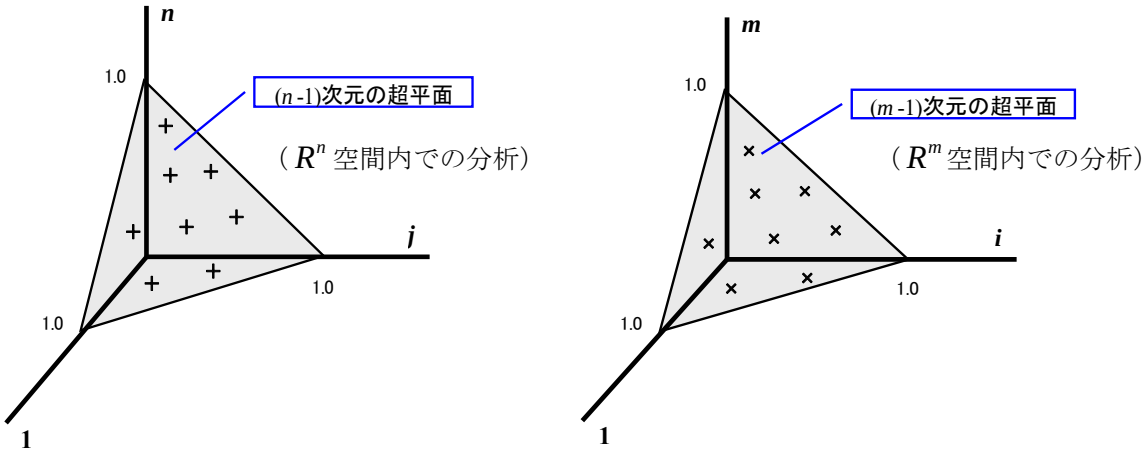
◆行和を 1 としたときの「行プロフィル」で
（$n-1$）次元内に分布する m 個の点

$$\underset{m\times n}{\mathbf{N}_I} = \underset{m\times m}{\mathbf{P}_I^{-1}}\,\underset{m\times n}{\mathbf{P}_{IJ}} = \left\{ q_{ij} = \frac{p_{ij}}{p_{i+}} \middle| i \in I, j \in J \right\}$$

◆列和を 1 としたときの「列プロフィル」で
（$m-1$）次元内に分布する n 個の点

$$\underset{n\times m}{\mathbf{N}_J} = \underset{n\times n}{\mathbf{P}_J^{-1}}\,\underset{n\times m}{\mathbf{P}_{JI}} = \left\{ q_{ij}^{*} = \frac{p_{ij}}{p_{+j}} \middle| i \in I, j \in J \right\}$$

図 2　行プロフィルと列プロフィルの関係



$\underset{m\times n}{\mathbf{N}_I} = \underset{m\times m}{\mathbf{P}_I^{-1}}\,\underset{m\times n}{\mathbf{P}_{IJ}} = \left\{ q_{ij} = \frac{p_{ij}}{p_{i+}} \middle| i \in I, j \in J \right\}$ の分布　$\underset{n\times m}{\mathbf{N}_J} = \underset{n\times n}{\mathbf{P}_J^{-1}}\,\underset{n\times m}{\mathbf{P}_{JI}} = \left\{ q_{ij}^* = \frac{p_{ij}}{p_{+j}} \middle| i \in I, j \in J \right\}$ の分布

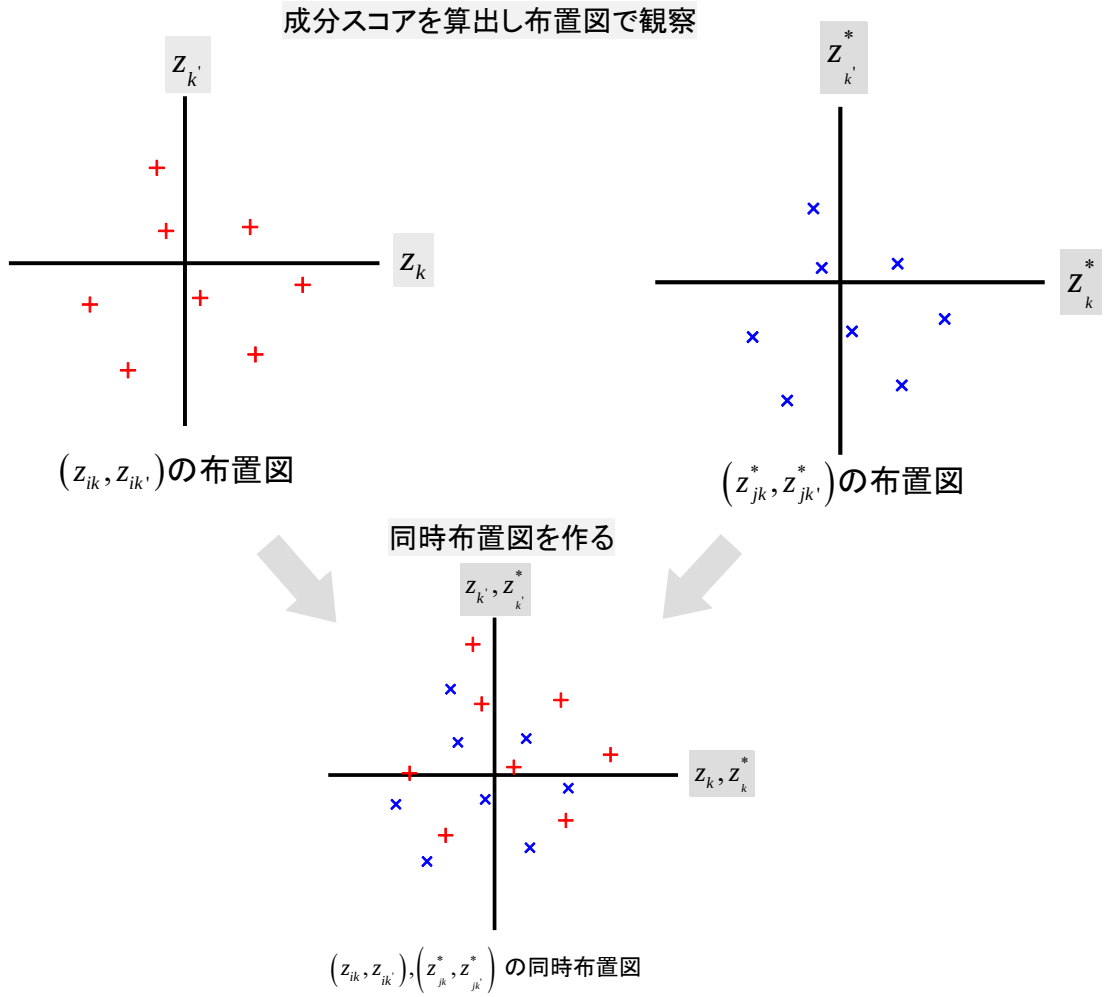


図 3　行，列のプロフィルと成分スコアの布置図の関係

（†）ここで，z_{ik}，z^*_{jk} は**成分スコア**である，これの詳細は後述する．

（‡）ここでは，（k, k'）成分を指定したとした．

いま，小さな架空のトイ・データを用意して，それぞれの関係を確認しよう．このデータ表（表 3）は，4 名の「回答者」に 3 つの「銘柄」のどれが好きかを尋ね「好きな銘柄には 1」を，そうではない場合は 0 とするという場面を想定している．これは，「(回答者) × (項目)」型であり，同時に応答数が 1 のクロス表とも考えられる[8].

表 3　2 元データ表の例

回答者	銘柄 A	銘柄 B	銘柄 C	行和 f_{i+}
回答者 1	1	0	1	2
回答者 2	0	1	0	1
回答者 3	1	0	0	1
回答者 4	0	1	1	2
列和 f_{+j}	2	2	2	6

表 4　確率行列: $\mathbf{P}_{IJ}=\left(p_{ij}\right)$ と周辺度数 p_{i+}, p_{+j}

回答者	銘柄 A	銘柄 B	銘柄 C	行和 p_{i+}	$\mathbf{r}$ の要素 (行の質量)
回答者 1	0.1667	0.0000	0.1667	0.3333 $(2/6)$	p_{1+}
回答者 2	0.0000	0.1667	0.0000	0.1667 $(1/6)$	p_{2+}
回答者 3	0.1667	0.0000	0.0000	0.1667 $(1/6)$	p_{3+}
回答者 4	0.0000	0.1667	0.1667	0.3333 $(2/6)$	p_{4+}
列和 p_{+j}	0.3333 $(2/6)$	0.3333 $(2/6)$	0.3333 $(2/6)$	1.0 (全確率)	
$\mathbf{c}^t$ の要素 (列の質量)	p_{+1}	p_{+2}	p_{+3}		

このデータ表を，大きさが $m\times n=4\times 3$ の $\underset{4\times 3}{\mathbf{F}}$ とし，この $\mathbf{F}$ から確率行列 $\mathbf{P}_{IJ}=\left(p_{ij}\right)$ と行，列それぞれの周辺確率分布 p_{i+}, p_{+j} および重心ベクトル $\mathbf{r},\mathbf{c}$ を作る（表 4）.

1. 3　プロファイルと重心座標系

対応分析法の基本的なアイデアは，プロファイル間の距離を**“カイ二乗距離”**で表し，かつ，高次元に分布しているそのプロファイルを，低次元の空間に射影することにある（図 3）．これを理解するには例を見るのが早い．この例について，**行プロファイル**と**列プロファイル**を作ってみよう．

いま行の側，つまり回答者側の 4 名からみた行プロファイル $\mathbf{N}_I$ を考えよう（つまり空間 R^n 内での分析を考える）．ここでは列つまり「銘柄」側が 3 つの選択肢であるから行プロファイルとその重心を**三角図**として描いてみた．このような座標系を**三角座標系**（triangular coordinate system）という．これについては，別のレストランのデータ例も見てほしい[9]．三角座標は上の図 3 の $(n-1)$ 次元のアミカケ部分の平面に相当する．対応分析では，このアミカケ部分の中の点の分布（つまり行プロファイルの分布）を考える．なおこの例では，2 次元

[8] この例は「第 I 部」で用いた例の 1 つに同じ.
[9] 「第 I 部」の例として詳しい説明がある．こちらの例のほうが三角図の意味が分かりやすいだろう.

の平面で表されるが，通常は，2 元データ表の次元数（行列の寸法）は（ここでは$(n-1)$次元）高い次元数となるので次元縮約を行うことに意味がある（“節約の原理”[10]あるいは“オッカムの剃刀”の達成）．

表 5　行プロファイル $\mathbf{N}_I = \mathbf{P}_I^{-1}\mathbf{P}_{IJ} = \left\{ q_{ij} = \dfrac{p_{ij}}{p_{i+}} \middle| i \in I, j \in J \right\}$ とその重心

回答者	銘柄 A	銘柄 B	銘柄 C
回答者 1	0.5	0	0.5
回答者 2	0	1	0
回答者 3	1	0	0
回答者 4	0	0.5	0.5
p_{+j}	0.3333	0.3333	0.3333
$\mathbf{c}$	$\mathbf{c} = (0.3333, 0.3333, 0.3333)^t$		

表 6　列プロファイル $\mathbf{N}_J = \mathbf{P}_J^{-1}\mathbf{P}_{JI} = \left\{ q_{ij}^* = \dfrac{p_{ij}}{p_{+j}} \middle| i \in I, j \in J \right\}$ とその重心

回答者	銘柄 A	銘柄 B	銘柄 C	p_{i+}	$\mathbf{r}$
回答者 1	0.5	0	0.5	0.3333	$\mathbf{r} = \begin{pmatrix} 0.3333 \\ 0.1667 \\ 0.1667 \\ 0.3333 \end{pmatrix}$
回答者 2	0	0.5	0	0.1667	
回答者 3	0.5	0	0	0.1667	
回答者 4	0	0.5	0.5	0.3333	

三角図

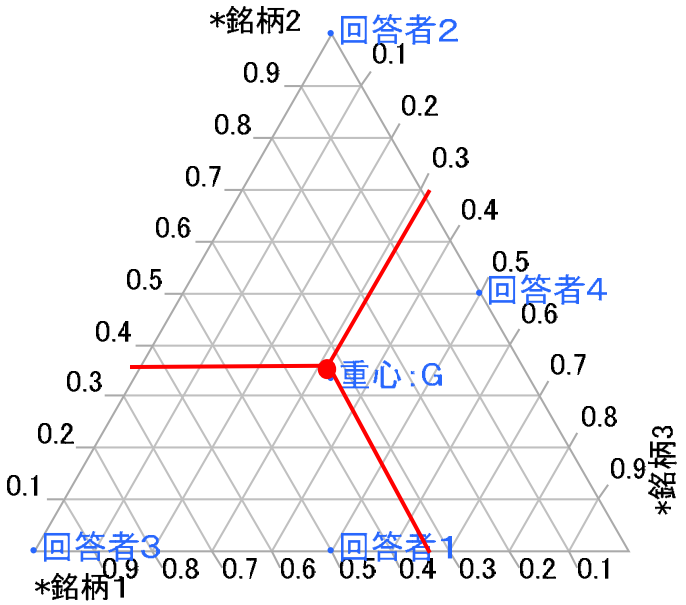


図　三角座標系に布置した行プロファイル $\mathbf{N}_I$ の分布

（4 名の回答者の布置と重心 G ：$\mathbf{c} = (0.3333, 0.3333, 0.3333)^t$ ）

[10] 節約の原理（priciple of parsimony），オッカムの剃刀（Occam's razor）とは，いわゆる「ケチの原理」のこと．ここでは，多次元情報のデータ表の構造探査をより少ない次元数内に移し替えて効率よく観察できるような操作を喩えている．

1．4 カイ二乗距離とユークリッド距離

ここで，対応分析で重要な意味をもつ**カイ二乗距離**（chi-square distance）あるいは**平方カイ二乗距離**（squared chi-square distance）についてふれる．カイ二乗距離とはストレッチ・プロファイルの**重み付きのユークリッド距離**のことである．上に用意した記号を用いると，2元データ表の行と列それぞれについて，以下のように表せる．

（ｉ）項目 I の 2 つの選択肢 i と i' の間の平方カイ二乗距離

$$d_B^2(i,i') = \sum_{j=1}^{n} \frac{1}{p_{+j}} \left(q_{ij} - q_{i'j}\right)^2 = \sum_{j=1}^{n} \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2 \quad \left(i, i' \in I\right) \tag{16}$$
$$= \sum_{j=1}^{n} \left(\frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}} - \frac{p_{i'j}}{p_{i'+}\sqrt{p_{+j}}} \right)^2$$

（ⅱ）項目 J の 2 つの選択肢 j と j' の間の平方カイ二乗距離

$$d_B^2(j,j') = \sum_{i=1}^{m} \frac{1}{p_{i+}} \left(q_{ij}^* - q_{ij'}^*\right)^2 = \sum_{i=1}^{m} \frac{1}{p_{i+}} \left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}} \right)^2 \quad \left(j, j' \in J\right) \tag{17}$$
$$= \sum_{i=1}^{m} \left(\frac{p_{ij}}{p_{+j}\sqrt{p_{i+}}} - \frac{p_{ij'}}{p_{+j'}\sqrt{p_{i+}}} \right)^2$$

ここで，列の重み p_{+j}，行の重み p_{i+} を，対応分析法ではそれぞれ，“**列の質量**”（column mass），“**行の質量**”（row mass）と呼ぶ．この式からわかるように，カイ二乗距離とは，質量を重みとした重み付き距離になっている．かりに，重み p_{+j}, p_{i+} がない場合は，単なる**平方ユークリッド距離**となる．このような平方カイ二乗距離を考える理由のうち，もっとも重要な性質は**“分布の同等性”**（equivalence of distribution）である[11]．

ここまでを表に要約すると，以下のようになる（表 7）．

表 7　分析の方向

行の側から分析	列の側から分析
n 次元空間 R^n 内での分析	m 次元空間 R^m 内での分析
行和を 1 としたときの「行プロファイル」で（$n-1$）次元内に分布する m 個の点	列和を 1 としたときの「列プロファイル」で（$m-1$）次元内に分布する n 個の点
$\mathbf{N}_I = \mathbf{P}_I^{-1}\mathbf{P}_{IJ} = \left\{ q_{ij} = \dfrac{p_{ij}}{p_{i+}} \middle\| i \in I, j \in J \right\}$	$\mathbf{N}_J = \mathbf{P}_J^{-1}\mathbf{P}_{JI} = \left\{ q_{ij}^* = \dfrac{p_{ij}}{p_{+j}} \middle\| i \in I, j \in J \right\}$
行のプロファイル間の平方カイ二乗距離 $d_B^2(i,i') = \sum_{j=1}^{n} \frac{1}{p_{+j}} \left(q_{ij} - q_{i'j}\right)^2$ $= \sum_{j=1}^{n} \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2$	列のプロファイル間の平方カイ二乗距離 $d_B^2(j,j') = \sum_{i=1}^{m} \frac{1}{p_{i+}} \left(q_{ij}^* - q_{ij'}^*\right)^2$ $= \sum_{i=1}^{m} \frac{1}{p_{i+}} \left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}} \right)^2$

[11] これについては，うしろで述べる．

2. データ行列の生成と解法

2.1 成分スコアの算出とその性質

では，対応分析法の計算はどのように行われるのであろうか．ここでは数理的証明は省略し，計算方法だけを示す．上に準備した記法を用いて，その計算方法を説明する．ここでは行プロファイルについて説明するが，列プロファイルについても，同様の議論が展開できる．

まず，

$$x_{ij} = \frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}} = \frac{q_{ij}}{\sqrt{p_{+j}}} \quad (i \in I, j \in J) \tag{18}$$

を行列要素とする行列を作る．これは，行プロファイルを，列の質量の平方根 $\sqrt{p_{+j}}$ の逆数で“ストレッチした”ことになる．たとえば，これを図 4 の正三角形の例でいうと，各辺を $\sqrt{p_{+j}}$ の逆数で伸ばしたと思えばよい．これを，**“ストレッチ・プロファイル”**（stretched profile）という．この x_{ij} を要素とする行列をつぎのように作る．

$$\underset{m\times n}{\mathbf{X}} = \mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2} = \left(x_{ij}\right) \quad (i \in I, j \in J) \tag{19}$$

ここで，$\mathbf{P}_I^{-1} = diag\left(\frac{1}{p_{i+}}\right)$，および，$\mathbf{P}_J^{-1/2} = diag\left(\frac{1}{\sqrt{p_{+j}}}\right)$ である．

これが対応分析法における基本の**データ行列**となる[12]．ここで，行プロファイルそのものを用いずに，$\mathbf{P}_J^{-1/2} = diag\left(\frac{1}{\sqrt{p_{+j}}}\right)$ を加重とした要素としている点に注意しよう．その理由は，上に述べた**カイ二乗距離**を用いることに関係する．この質量でストレッチした三角座標系を，**重心座標系**（barycentric coordinate system）ともいう[13]．

この，x_{ij} について，列の側の項目 J の選択肢 j について，平均値と分散，共分散を求めてみよう．このとき，行和の加重（p_{i+}）を行うことに注意する[14]．まず，平均値は以下のように計算される．

$$\bar{x}_j = \frac{1}{N}\sum_{i=1}^{m} f_{i+}x_{ij} = \sum_{i=1}^{m} p_{i+}x_{ij} = \sum_{i=1}^{m} p_{i+}\left(\frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}\right) = \sqrt{p_{+j}} \quad (j \in J) \tag{20}$$

x_{ij} の各列からこの平均値 $\bar{x}_j$ を引くと（つまり，平均値の周りの**偏差**を作ると），以下のようになる[15]．

$$x_{ij}^{*} = x_{ij} - \bar{x}_j = \frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}} - \sqrt{p_{+j}} \tag{21}$$

よって，**分散と共分散**は以下のようになる．

---

[12] 出発行列にはいろいろな表記があるが，ここではこうしたということ．このあとにも別の形で登場する．
[13] ここらは，第 I 部で，簡単な数値例を用いて説明した．
[14] ここは，（行側を測定対象，列側を変量とみた主成分分析を想定してみればよいだろう．
[15] こうした操作を，平均値の周りで**“中心化する”（センタリング）**という．

$$s_{jj'} = \frac{1}{N}\sum_{i=1}^{m} f_{i+}\left(x_{ij} - \overline{x}_j\right)\left(x_{ij'} - \overline{x}_{j'}\right) = \sum_{i=1}^{m} p_{i+}\left(\frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}} - \sqrt{p_{+j}}\right)\left(\frac{p_{ij'}}{p_{i+}\sqrt{p_{+j'}}} - \sqrt{p_{+j'}}\right)$$
$$= \sum_{i=1}^{m}\left(\frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}}\right)\left(\frac{p_{ij'} - p_{i+}p_{+j'}}{\sqrt{p_{i+}p_{+j'}}}\right) \tag{22}$$

（ここで，$j = j'$のとき分散，$j \neq j'$のときは共分散となる）

この$s_{jj'}$を要素とする行列がx_{ij}^*の**分散共分散行列** $\underset{n\times n}{\mathbf{V}}$ である．ここでいま，

$$\mathbf{X}^* = \left(x_{ij}^*\right)$$

と表すと，以下のように書ける．

$$\underset{n\times n}{\mathbf{V}} = \left(s_{jj'}\right) = \left(\mathbf{X}^*\right)^t \mathbf{P}_I \mathbf{X}^* \tag{23}$$

形式的には，この**分散共分散行列** $\underset{n\times n}{\mathbf{V}}$ **の固有値問題**として扱うことになる．ところでここで，式（22）の要素$s_{jj'}$に注目すると，うしろに示す**ピアソンのカイ二乗統計量**の要素を表す式を用いて，

$$y_{ij}^* = \frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}} = \frac{f_{ij} - \dfrac{f_{i+}f_{+j}}{N}}{\sqrt{\dfrac{f_{i+}f_{+j}}{N}}} = \frac{f_{ij} - e_{ij}}{\sqrt{e_{ij}}} \tag{24}$$

を要素とする行列$\mathbf{Y}^* = \left(y_{ij}^*\right)$の**特異値分解**を行うことと，と言い替えてもよい[16]．

上の式（24）に注目し，この分散共分散行列 $\underset{n\times n}{\mathbf{V}}$ の対角和（跡和あるいはトレース）を求めると，以下のようになる．

$$tr\left(\underset{n\times n}{\mathbf{V}}\right) = \sum_{j=1}^{n} s_{jj} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\left(p_{ij} - p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}} \tag{25}$$

これは，2 元クロス表の“**分散の総量**”であり“**総変動**”を表している．対応分析では，これを“**慣性の合計**”（全慣性；total inertia）あるいは単に“**慣性**”（inertia）と呼ぶ．

ところで，よく知られた 2 元クロス表の“**独立性の検定**”に用いる**ピアソンのカイ二乗統計量**χ_p^2は，以下のように表される（あらためてうしろで述べる）．

[16] こう書き替える必要もないのだが，これを考える意味を，うしろに数値例で示した．

$$\chi_p^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{N\left(p_{ij} - p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\left(f_{ij} - \dfrac{f_{i+}f_{+j}}{N}\right)^2}{\dfrac{f_{i+}f_{+j}}{N}} \tag{26}$$

このことから，

$$tr\left(\underset{n\times n}{\mathbf{V}}\right) = \frac{\chi_p^2}{N} \tag{27}$$

となる．つまり，対応分析における**“慣性の合計”**または**“慣性”**とは，ピアソンのカイ 2 乗検定統計量を 2 元クロス表の総度数 N で割った値となっている．

ここまでに用いた記号を，もう一度整理しよう．

列の質量 $\sqrt{p_{+j}}$ でストレッチした行プロファイルを要素とする行列を考える．

$$\underset{m\times n}{\mathbf{X}} = \left(x_{ij}\right) = \left(\frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}\right) = \left(\frac{q_{ij}}{\sqrt{p_{+j}}}\right) \tag{28}$$

ここで，第 i 行目の行プロファイルを，

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{ij} \\ \vdots \\ x_{in} \end{pmatrix}$$

とすると，行列全体は，

$$\underset{m\times n}{\mathbf{X}} = \begin{pmatrix} \mathbf{x}_1^t \\ \mathbf{x}_2^t \\ \vdots \\ \mathbf{x}_i^t \\ \vdots \\ \mathbf{x}_m^t \end{pmatrix} \tag{29}$$

である（ここで，t は転置を表す，以下同じ）．

つぎに，$\overline{x}_j$ を要素とする平均ベクトルと，全要素が 1 であるベクトルを，それぞれ，

$$\underset{n\times 1}{\overline{\mathbf{x}}}=\begin{pmatrix}\overline{x}_1\\ \overline{x}_2\\ \vdots\\ \overline{x}_j\\ \vdots\\ \overline{x}_n\end{pmatrix}=\begin{pmatrix}\sqrt{p_{+1}}\\ \sqrt{p_{+2}}\\ \vdots\\ \sqrt{p_{+j}}\\ \vdots\\ \sqrt{p_{+n}}\end{pmatrix},\ \underset{m\times 1}{\mathbf{1}_m}=\begin{pmatrix}1\\ 1\\ \vdots\\ 1\\ \vdots\\ 1\end{pmatrix} \tag{30}$$

で表すと，**平均ベクトル**は以下となる．

$$\underset{n\times 1}{\overline{\mathbf{x}}}=\frac{1}{N}\mathbf{X}^t\mathbf{1}_m \tag{31}$$

さらに，**分散共分散行列**は，下式のように表される[17]．

$$\begin{aligned}\underset{n\times n}{\mathbf{V}}&=\left(s_{jj'}\right)=\frac{1}{N}\sum_{i=1}^{m}f_{i+}\left(\mathbf{x}_i-\overline{\mathbf{x}}\right)\left(\mathbf{x}_i-\overline{\mathbf{x}}\right)^t=\sum_{i=1}^{m}p_{i+}\left(\mathbf{x}_i-\overline{\mathbf{x}}\right)\left(\mathbf{x}_i-\overline{\mathbf{x}}\right)^t\\ &=\mathbf{X}^t\mathbf{P}_I\mathbf{X}-\overline{\mathbf{x}}\overline{\mathbf{x}}^t=\left(\mathbf{X}^*\right)^t\mathbf{P}_I\mathbf{X}^*\end{aligned} \tag{32}$$

そして，形式的にいえば，この行列 $\underset{n\times n}{\mathbf{V}}=\left(s_{jj'}\right)$ の固有値問題を解くことが目標である．

またここで，$y_{ij}^*=\dfrac{p_{ij}-p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}}$ を（また $s_{jj'}$ でもある）要素とするつぎの行列，

$$\mathbf{Y}^*=\left(y_{ij}^*\right)=\mathbf{P}_I^{-1/2}\left(\mathbf{P}_{IJ}-\mathbf{r}\mathbf{c}^t\right)\mathbf{P}_J^{-1/2} \tag{33}$$

を作り[18]，この行列 $\mathbf{Y}^*$ の特異値分解を行うことでも同じ結果が得られる．

ところで，この行列 $\underset{n\times n}{\mathbf{V}}=\left(s_{jj'}\right)$ の固有値問題とは，固有方程式を $\left|\mathbf{V}-\lambda\mathbf{I}_n\right|=0$（ここで，$\mathbf{I}_n$ は寸法が $n\times n$ の単位行列）を解くことである．しかし，この行列は非対称であり，そこでこれを対称化した，$y_{ij}=\dfrac{p_{ij}}{\sqrt{p_{i+}p_{+j}}}=\dfrac{f_{ij}}{\sqrt{f_{i+}f_{+j}}}\left(p_{i+}\neq 0,p_{+j}\neq 0;f_{i+}\neq 0,f_{+j}\neq 0\right)$ をあらたなデータ行列の要素とする行列 $\mathbf{Q}=\left(y_{ij}\right)\left(i\in I,j\in J\right)$ を作る[19]．この行列 $\mathbf{Q}$ から得た分散共分散行列の固有値問題として解く（この結果は上の行列 $\mathbf{V}$ を解くことに同じことがわかっている[20]）．これをいままでに用意した行列を用いて表すと以下のように書ける．

$$\underset{m\times n}{\mathbf{Q}}=\underset{m\times m}{\mathbf{P}_I^{-1/2}}\ \underset{m\times n}{\mathbf{P}_{IJ}}\ \underset{n\times n}{\mathbf{P}_J^{-1/2}} \tag{34}$$

---

[17] ここで，加重 $\mathbf{P}_I$ を考えないで，$\left(\mathbf{X}^*\right)^t\mathbf{X}^*$ とすると，主成分分析（PCA）に相当すると考えられる．

[1818] つまり，式（33）と（34）は，記法を変えて言い換えただけである．

[19] 対称化することで，たとえば上のように頻度 f_{ij}，f_{i+},f_{+j} などで扱える．

[20] 固有値のうちの自明根の出方が異なるだけである．

そして，この行列$\mathbf{Q}$から得た分散共分散行列の固有値問題として扱う．つまり，

$$\underset{n\times n}{\mathbf{V}^*}=\mathbf{Q}^t\mathbf{Q}=\mathbf{P}_J^{-1/2}\mathbf{P}_{JI}\mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2} \tag{35}$$

の分散共分散行列の固有方程式$\left|\mathbf{V}^*-\lambda\mathbf{I}_n\right|=0$（ここで$\mathbf{I}_n$は寸法が$n\times n$の単位行列），を解くことである．この$\underset{n\times n}{\mathbf{V}^*}$の固有値問題は（あるいは$\underset{m\times n}{\mathbf{Q}}$の特異値分解は），はじめの固有値$\lambda_0$に対する固有ベクトル$\mathbf{l}_0$の要素は以下のようになる．これは自明な解となる[21]．

$$\underset{1\times n}{\mathbf{l}_0^t}=\left(\sqrt{p_{+1}},\sqrt{p_{+2}},\cdots,\sqrt{p_{+j}},\cdots,\sqrt{p_{+n}}\right) \tag{36}$$

統計ソフトウェアを用いるときは，この自明な解を除いて，2 番目以降の固有値を用いる．こうして得られる**固有値と固有ベクトル**を以下のように表す．

$$\text{固有値：}\ \lambda_k\left(\begin{array}{l}0\le\lambda_k\le 1\\ k=1,2,\cdots,K\\ K=\min\left\{m,n\right\}-1\end{array}\right)$$

$$\text{固有ベクトル：}\ \mathbf{l}_k\left(\begin{array}{l}k=1,2,\cdots,K\\ K=\min\left\{m,n\right\}-1\end{array}\right)$$

このとき，固有値の個数は 2 元データ表の行と列のそれぞれの小さい方から 1 を引いた$K=\min\left\{m,n\right\}-1$となる（プロファイルの分布の空間を思い出そう，次元の縮退があるということ）．また，固有値は非負で値は 1 を越えることはない[22]（$0\le\lambda_k\le 1$）．

ここでは，これ以上数理的な特性や原理の細かい説明は行わずに，得られる諸量を対応分析法としてどのような意味があり，どのように用いるかという観点から話しを進めよう．

ここでまず，行の成分スコアを考えよう．いま，$\mathbf{P}_J^{-1/2}\mathbf{L}$を加重とする行プロファイルの加重和を作ると（つまり合成変数を作ると），それが**行の成分スコア**である．この行の各点（つまり選択肢$i\in I$）に与えられる**成分スコア**[23]（coordinates）の行列は，以下のようになる．

$$\underset{m\times K}{\mathbf{Z}}=\underbrace{\mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2}}_{m\times n}\underset{n\times K}{\mathbf{L}} \tag{37}$$

ここで，

$$\underset{n\times K}{\mathbf{L}}=\left(\mathbf{l}_1,\mathbf{l}_2,\cdots,\mathbf{l}_k,\cdots,\mathbf{l}_K\right)$$

$$\underset{1\times n}{\mathbf{l}_k^t}=\left(l_{1k},l_{2k},\cdots,l_{jk},\cdots,l_{nk}\right)$$

---

[21] この自明解は成分スコア他の算出には直接は用いられないが，つまり通常は出力表示されないが，実はあとで述べるピアソンのカイ二乗統計量や再生公式の説明で重要な意味を持つ．

[22] 証明が必要であるが略す（たとえば，Volle（1985），Jambu（1989），大隅他（1994）などを参照）．

[23] 呼称がいろいろある．**主座標**（principal coordinates），**座標**（coordinates），**得点**（scores）などということがある．また，数量化法では，数量化得点あるいは数量化スコアなどという．ここでは**成分スコア**と記すことにする．

である．固有ベクトルは$\mathbf{l}_k^t\mathbf{l}_k=1,\mathbf{l}_k^t\mathbf{l}_{k'}=0\left(k\neq k'\right)$であり，互いに直交している（このような条件のもとに固有方程式を解いた）．この成分スコアの行列$\underset{m\times K}{\mathbf{Z}}$の，第$k$成分のスコアは，ベクトルを用いて以下のように書ける．

$$\underset{m\times 1}{\mathbf{z}_k}=\underbrace{\mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2}}_{m\times n}\underset{n\times 1}{\mathbf{l}_k}\left(k=1,2,\cdots,K\right) \tag{38}$$

さらに，成分スコアの行列$\underset{m\times K}{\mathbf{Z}}$の$\left(i,k\right)$要素，つまり“**項目$I$の第$i$選択肢の第$k$成分のスコア**”$z_{ik}\left(i\in I;k=1,2,\cdots,K\right)$は，以下のようになる．

（ｉ）項目Iの第i選択肢の第k成分スコア

$$z_{ik}=\sum_{j=1}^{n}l_{jk}x_{ij}=\sum_{j=1}^{n}\left(\frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}\right)l_{jk}\left(i\in I;k=1,2,\cdots,K\right) \tag{39}$$

つまり，**行の成分スコア**は，列の標準化成分スコア（列の標準座標）を係数とした行のストレッチ・プロファイルの加重和となっている．利用する上では，成分スコアがこうした加重和とした**“合成変数”**（合成指標）であることを知っておくことが重要である[24].

つぎにここで，この第k成分スコアの平均値$\overline{z}_k$と分散$V\left[\overline{z}_k\right]$を調べておく．

$$\begin{aligned}\overline{z}_k&=\frac{1}{N}\sum_{i=1}^{m}f_{i+}z_{ik}=\sum_{i=1}^{m}p_{i+}z_{ik}=\sum_{i=1}^{m}p_{i+}\left\{\sum_{j=1}^{n}l_{jk}\left(\frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}\right)\right\}\\&=\sum_{j=1}^{n}l_{jk}\sum_{i=1}^{m}\frac{p_{ij}}{\sqrt{p_{+j}}}=\sum_{j=1}^{n}l_{jk}\sqrt{p_{+j}}=0\end{aligned} \tag{40}$$

最後の等式では，各固有ベクトルが（自明解に対応する）最初の固有ベクトルと直交すること，すなわち，$\mathbf{l}_k^t\mathbf{l}_0=0\left(k\neq 0\right)$であることを用いた（$\mathbf{l}_0$は式（36）から）．このようにして，**“成分スコアの平均値は0”**ということになる．

また，分散は，

$$\begin{aligned}V\left[z_k\right]&=\frac{1}{N}\sum_{i=1}^{m}f_{i+}\left(z_{ik}-\overline{z}_k\right)^2=\sum_{i=1}^{m}p_{i+}{z_{ik}}^2=\sum_{i=1}^{m}\left\{p_{i+}\sum_{j=1}^{n}l_{jk}\left(\frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}\right)\right\}^2\\&=\sum_{i=1}^{m}\left\{\sum_{j=1}^{n}l_{jk}\left(\frac{p_{ij}}{\sqrt{p_{i+}p_{+j}}}\right)\right\}^2=\lambda_k\end{aligned} \tag{41}$$

となる．ここでは以下を用いた．

$$V\left[z_k\right]=\mathbf{z}_k^t\mathbf{P}_I\mathbf{z}_k=\left(\mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2}\mathbf{l}_k\right)^t\mathbf{P}_I\left(\mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2}\mathbf{l}_k\right)=\mathbf{l}_k^t\underbrace{\mathbf{Q}^t\mathbf{Q}\mathbf{l}_k}_{=\lambda_k\mathbf{l}_k}=\lambda_k\underbrace{\mathbf{l}_k^t\mathbf{l}_k}_{=1}=\lambda_k$$

[24] ある成分スコアという1つの数値となっているが，実は，（元の多次元の）選択肢の加重和であることに注意しよう．

これは，**“第k成分スコアの分散は対応する第k固有値となる”**ということである．したがって，成分kの成分スコアの標準偏差は$\sqrt{V\left[z_k\right]}=\sqrt{\lambda_k}$ となり，これは特異値分解における**特異値**に等しい（ここではα_k で表した）．

ところでここで，もとの$\mathbf{Q}$ を転置して得られるつぎの行列 $\underset{m\times m}{\mathbf{W}}$ について，上と同じように固有値問題として解いても（行列の性質から）解は同じとなる．

$$\underset{m\times m}{\mathbf{W}}=\underset{m\times n}{\mathbf{Q}}\,\underset{n\times m}{\mathbf{Q}^t} \tag{42}$$

つまり，もとのデータ表を転置した行列に対して対応分析を行っても，同じ固有値や特異値が得られることを示している．ここで，固有方程式$\left|\mathbf{W}-\lambda\mathbf{I}_m\right|=0$（ここで，$\mathbf{I}_m$ は大きさが $m\times m$ の単位行列）を解くと，以下のようになる．

① このとき，固有値は前に同じで$\lambda_k\left(\begin{matrix}k=1,2,\cdots,K\\K=\min\left\{m,n\right\}-1\end{matrix}\right)$ となる．

② 行列$\mathbf{W}$を解いて得られる固有ベクトルを$\mathbf{u}_k$ で表すと，上の行列$\mathbf{V}^*$から得た固有ベクトル$\mathbf{l}_k$ との間に以下の双対の関係がある[25]．

$$\underset{m\times 1}{\mathbf{u}_k}=\frac{1}{\sqrt{\lambda_k}}\underset{m\times n}{\mathbf{Q}}\,\underset{n\times 1}{\mathbf{l}_k}\quad\text{あるいは}\quad\underset{n\times 1}{\mathbf{l}_k}=\frac{1}{\sqrt{\lambda_k}}\underset{n\times m}{\mathbf{Q}^t}\,\underset{m\times 1}{\mathbf{u}_k} \tag{43}$$

そして，このときの**列の成分スコア**の行列は，以下となる．

$$\underset{n\times K}{\mathbf{Z}^*}=\underbrace{\mathbf{P}_J^{-1}\mathbf{P}_{JI}\mathbf{P}_I^{-1/2}}_{n\times m}\underset{m\times K}{\mathbf{U}} \tag{44}$$

ここで，

$$\underset{m\times K}{\mathbf{U}}=\left(\mathbf{u}_1,\mathbf{u}_2,\cdots,\mathbf{u}_k,\cdots,\mathbf{u}_K\right)$$

$$\underset{1\times m}{\mathbf{u}_k^t}=\left(u_{1k},u_{2k},\cdots,u_{ik},\cdots,l_{mk}\right)$$

$$\underset{n\times 1}{\mathbf{z}_k^*}=\underbrace{\mathbf{P}_J^{-1}\mathbf{P}_{JI}\mathbf{P}_I^{-1/2}}_{n\times m}\underset{m\times 1}{\mathbf{u}_k}\left(k=1,2,\cdots,K\right)$$

つまり，“項目Jの第j選択肢の第k成分のスコア”$z_{jk}^*\left(j\in J;k=1,2,\cdots,K\right)$は，以下のようになる．

（ⅱ）項目Jの第j選択肢の第k成分スコア

$$z_{jk}^*=\sum_{i=1}^{m}u_{ik}x_{ij}^*=\sum_{i=1}^{m}\left(\frac{p_{ij}}{p_{+j}\sqrt{p_{i+}}}\right)u_{ik}\left(j\in J;k=1,2,\cdots,K\right) \tag{45}$$

---

[25] これは，$\underset{m\times n}{\mathbf{Q}}=\sum_{k=1}^{n}\sqrt{\lambda_k}\,\underset{m\times 1}{\mathbf{u}_k}\,\underset{1\times n}{\mathbf{l}_k^t}$ とかける．つまり特異値分解に相当する．

以上で，行と列，それぞれの選択肢について，**固有値**（**特異値**の二乗）と**固有ベクトル**を用いて，**“成分スコア”**を求めることができた．また，固有値とは成分スコアの分散であり，固有ベクトルはその成分スコアという合成変数の係数に関係することが分かった．つまり．上の式（39），（45）の形からみてわかるように，これら**成分スコア**は一種の加重和であり**合成変数**である．そして多くの場合，統計ソフトウェアを用いるとこれらの諸統計量が算出・出力される．

なお, 特異値分解とは以下のような関係で, ある行列を右辺のように分解することを指す.

$$\underset{m\times n}{\mathbf{Y}^{*}}=\sum_{k=1}^{K}\sqrt{\lambda_k}\,\mathbf{u}_k\mathbf{l}_k^{t}=\underset{m\times K}{\mathbf{U}}\;\underset{K\times K}{\Lambda^{1/2}}\;\underset{K\times n}{\mathbf{L}^{t}} \tag{46}$$

ここで，$\underset{K\times K}{\Lambda}=diag\left(\lambda_k\right)$として，$\underset{K\times K}{\Lambda^{1/2}}=diag\left(\sqrt{\lambda_k}\right)=diag\left(\alpha_k\right)$は$\sqrt{\lambda_k}=\alpha_k\left(k=1,2,\cdots,K\right)$を対角要素とする対角行列である．$\alpha_k=\sqrt{\lambda_k}\left(k=1,2,\cdots,K\right)$が**特異値**であり，固有値の正の平方根である．後述するように，特異値は行と列の**成分スコア間の相関係数**に相当する．以上のことから，数理的には，対応分析や数量化 III 類の解を求めるには,

$$y_{ij}^{*}=\frac{p_{ij}-p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}}\quad\text{あるいは}\quad y_{ij}=\frac{p_{ij}}{\sqrt{p_{i+}p_{+j}}}=\frac{f_{ij}}{\sqrt{f_{i+}f_{+j}}}$$

を要素とする行列の特異値分解を行い，そこで得られた左特異ベクトル，右特異ベクトル，特異値を計算する．よって，行列演算が備わっているソフトウェアを用いれば，これらの計算結果が得られる．

これを，数量化 III 類では，どのように考えているのだろうか．ここでは，質的データから，いまの場合は，行あるいは列の選択肢に，かりにある**数量**が与えられたとして，これらの相関が最大になるように最適化して得られる値をスコア（数量化得点，最適スコア）とすることである．いわゆる**尺度化**（スケーリング）の発想から定式化されている．

一方，対応分析法では，行あるいは列のストレッチ・プロファイルを多次元データと見なして，これをより少数次元の空間に成分スコアとして射影することを行う．このとき，成分スコアは，もとのプロファイル間のカイ二乗距離を成分スコア間のユークリッド距離に変換したものなので，クラスター化法などの“距離”にもとづく二次分析を行うときの幾何学的な解釈・操作に矛盾が生じないという利点がある．

ここまでに述べた導出方法（合成変数の係数の固有ベクトルのノルムを 1 の条件下で，その合成変数の分散を最大化するという導出方法）について，**n 次元空間 R^n 内の分析**と，**m 次元空間 R^m 内の分析**とに分けて表に要約しておこう（表 8，表 9）．じつはこれは，すでに図 3 に模式的に示したことを整理したことに相当する．またいままでの議論から，所与の 2 元データ表と，それを転置してえられるデータ表のいずれから出発しても，得られる結果は同じである（これも，対応分析法の特徴）.

表 8　n 次元空間 R^n 内での分析

	その 1	その 2（対称化）
対象とするデータの形	$x_{ij}^{*}=x_{ij}-\overline{x}_j=\dfrac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}-\sqrt{p_{+j}}$ $\underset{m\times n}{\mathbf{X}}=\left(x_{ij}\right)\qquad\underset{m\times n}{\mathbf{X}^{*}}=\left(x_{ij}^{*}\right)$	$y_{ij}=\dfrac{p_{ij}}{\sqrt{p_{i+}p_{+j}}}$ $\underset{m\times n}{\mathbf{Q}}=\left(y_{ij}\right)$
対象とする行列（共分散行列）	$\underset{n\times n}{\mathbf{V}}=\left(s_{jj'}\right)=\mathbf{X}^{t}\mathbf{P}_I\mathbf{X}-\overline{\mathbf{x}}\overline{\mathbf{x}}^{t}$ $\underset{n\times n}{\mathbf{V}}=\left(s_{jj'}\right)=\left(\mathbf{X}^{*}\right)^{t}\mathbf{P}_I\mathbf{X}^{*}$	$\underset{n\times n}{\mathbf{V}^{*}}=\mathbf{Q}^{t}\mathbf{Q}$
固有値と特異値	第 k 成分の固有値： $\lambda_k\left(\begin{matrix}k=1,2,\cdots,K\\K=\min\{m,n\}-1\end{matrix}\right)$ $0\le\lambda_k\le 1$ 第 k 成分の特異値；$\alpha_k=\sqrt{\lambda_k}$	上の行列の固有値を（あえて）μ_k とおくと以下. 第 k 成分の固有値： $\mu_k\left(\begin{matrix}k=1,2,\cdots,K\\K=\min\{m,n\}-1\end{matrix}\right)$ $0\le\mu_k\le 1$ しかしここで，“$\mu_k=\lambda_k$”となる. 第 k 成分の特異値；$\alpha_k=\sqrt{\lambda_k}$
固有ベクトル	行列 $\underset{n\times n}{\mathbf{V}}$ の“固有値 $\lambda_0=0$”に対して自明の解として以下の固有ベクトル $\underset{1\times n}{\mathbf{l}_0^{t}}=\left(\sqrt{p_{+1}},\sqrt{p_{+2}},\cdots,\sqrt{p_{+j}},\cdots,\sqrt{p_{+n}}\right)$	行列 $\underset{m\times n}{\mathbf{Q}}$ の“固有値 $\mu_0=1$”に対して自明の解として以下の固有ベクトル $\underset{1\times n}{\mathbf{l}_0^{t}}=\left(\sqrt{p_{+1}},\sqrt{p_{+2}},\cdots,\sqrt{p_{+j}},\cdots,\sqrt{p_{+n}}\right)$
固有値の数	固有値の数，つまり得られる成分数は，どちらも $K=\min\{m,n\}-1$（個） たとえば，$m>n$ とすると，$m-n$ 個の固有値は「0」となる．つまり，行列の階数（ランク）の縮退がおこる.	
成分スコア	$x_{ij}^{*}=x_{ij}-\overline{x}_j=\dfrac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}-\sqrt{p_{+j}}$ $z_{ik}=\sum_{j=1}^{n}l_{jk}x_{ij}^{*}=\sum_{j=1}^{n}\left(\dfrac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}-\sqrt{p_{+j}}\right)l_{jk}$ $\left(i\in I;k=1,2,\cdots,K\right)$	$y_{ij}=\dfrac{p_{ij}}{\sqrt{p_{i+}p_{+j}}}$ $z_{ik}=\sum_{j=1}^{n}l_{jk}x_{ij}=\sum_{j=1}^{n}\left(\dfrac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}\right)l_{jk}$ $\left(i\in I;k=1,2,\cdots,K\right)$
成分スコアの統計量	平均値：$\overline{z}_k=0$ 分散：$Var\left(z_k\right)=\lambda_k\left(k=1,2,\cdots,K;K=\min\{m,n\}-1\right)$	

表 9　m 次元空間 R^m 内での分析

	その 1	その 2（対称化）
対象とするデータ	$x_{ij}^{\dagger}=x_{ij}^{+}-\overline{x}_{i}^{+}=\dfrac{p_{ij}}{p_{+j}\sqrt{p_{i+}}}-\sqrt{p_{i+}}$ $\underset{m\times n}{\mathbf{X}}=\left(x_{ij}\right)\qquad \underset{m\times n}{\mathbf{X}}^{\dagger}=\left(x_{ij}^{\dagger}\right)$	$y_{ij}=\dfrac{p_{ij}}{\sqrt{p_{i+}p_{+j}}}$ $\underset{m\times n}{\mathbf{Q}}=\left(y_{ij}\right)$
対象とする行列	$\underset{m\times m}{\mathbf{S}}=\left(s_{ii'}\right)=\mathbf{X}\mathbf{P}_{J}\mathbf{X}^{t}-\overline{\mathbf{x}}\overline{\mathbf{x}}^{t}$ $\underset{m\times m}{\mathbf{S}}=\left(s_{ii'}\right)=\mathbf{X}^{\dagger}\mathbf{P}_{J}\left(\mathbf{X}^{\dagger}\right)^{t}$	$\underset{m\times m}{\mathbf{W}}=\mathbf{Q}\mathbf{Q}^{t}$
固有値と特異値	第 k 成分の固有値： $\lambda_{k}\left(\begin{array}{l}k=1,2,\cdots,K\\K=\min\left\{m,n\right\}-1\end{array}\right)$ $0\le\lambda_{k}\le 1$ 第 k 成分の特異値；$\alpha_{k}=\sqrt{\lambda_{k}}$	上の行列の固有値を（あえて）μ_{k} とおくと以下. 第 k 成分の固有値： $\mu_{k}\left(\begin{array}{l}k=1,2,\cdots,K\\K=\min\left\{m,n\right\}-1\end{array}\right)$ $0\le\mu_{k}\le 1$ しかしここで，“$\mu_{k}=\lambda_{k}$”となる. 第 k 成分の特異値；$\alpha_{k}=\sqrt{\lambda_{k}}$
固有ベクトル	行列 $\underset{m\times m}{\mathbf{S}}$ の“固有値 $\lambda_{0}=0$”に対して自明の解として以下の固有ベクトル $\underset{1\times m}{\mathbf{u}_{0}^{t}}=\left(\sqrt{p_{1+}},\sqrt{p_{2+}},\cdots,\sqrt{p_{i+}},\cdots,\sqrt{p_{im}}\right)$	行列 $\underset{m\times m}{\mathbf{W}}$ の“固有値 $\mu_{0}=1$”に対して自明の解として以下の同じ固有ベクトル $\underset{1\times m}{\mathbf{u}_{0}^{t}}=\left(\sqrt{p_{1+}},\sqrt{p_{2+}},\cdots,\sqrt{p_{i+}},\cdots,\sqrt{p_{im}}\right)$
固有値の数	固有値の数，つまり得られる成分数は，どちらも $K=\min\left\{m,n\right\}-1$（個） たとえば，$m>n$ とすると，$m-n$ 個の固有値は「0」となる．つまり，行列の階数（ランク）の縮退がおこる.	
成分スコア	$x_{ij}^{\dagger}=x_{ij}^{+}-\overline{x}_{i}^{+}=\dfrac{p_{ij}}{p_{+j}\sqrt{p_{i+}}}-\sqrt{p_{i+}}$ $z_{jk}^{*}=\sum_{i=1}^{m}u_{ik}x_{ij}^{\dagger}=\sum_{i=1}^{m}\left(\dfrac{p_{ij}}{p_{+j}\sqrt{p_{i+}}}-\sqrt{p_{i+}}\right)u_{ik}$ $\left(j\in J;k=1,2,\cdots,K\right)$	$y_{ij}=\dfrac{p_{ij}}{\sqrt{p_{i+}p_{+j}}}$ $z_{jk}^{*}=\sum_{i=1}^{m}u_{ik}x_{ij}=\sum_{i=1}^{m}\left(\dfrac{p_{ij}}{p_{+j}\sqrt{p_{i+}}}\right)u_{ik}$ $\left(j\in J;k=1,2,\cdots,K\right)$
成分スコアの統計量	平均値：$\overline{z}_{k}^{*}=0$ 分散：$Var\left(z_{k}^{*}\right)=\lambda_{k}\left(k=1,2,\cdots,K;K=\min\left\{m,n\right\}-1\right)$	

2. 2 簡単な例題による確認

ここで，上に述べたことを同じトイ・データで確かめよう．これはクロス表の特別な場合と考えられる（セル内の度数が 1 か 0 のみ）．このクロス表 $\mathbf{F}=\left(f_{ij}\right)$ をもう一度引用しよう．

表 10　2 元データ表 $\mathbf{F}=\left(f_{ij}\right)$ の例

回答者	銘柄 A	銘柄 B	銘柄 C	行和 f_{i+}
回答者 1	1	0	1	2
回答者 2	0	1	0	1
回答者 3	1	0	0	1
回答者 4	0	1	1	2
列和 f_{+j}	2	2	2	6

ちなみにいま，この 2 元データ表について，**カイ二乗統計量** χ_p^2 を求めてみる．ここでいうカイ二乗統計量は以下の式で与えられる．カイ二乗統計量を 2 元データ表の総度数（N）で割った値は，対応分析では非常に重要な指標であり，前述のようにこれを**慣性の合計**（全慣性；total inertia）あるいは**慣性**という．すでに述べたことから，これはいわば所与の 2 元データ表の保有する総変動（総分散）に相当する．ピアソンのカイ二乗統計量は，下式で求められる．

$$\chi_p^2=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{N\left(p_{ij}-p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}}=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(f_{ij}-\dfrac{f_{i+}f_{+j}}{N}\right)^2}{\dfrac{f_{i+}f_{+j}}{N}} \tag{47}$$

上のデータ表からこれを求めると以下の値となる．なおここで，期待度数は $e_{ij}=Np_{i+}p_{+j}=\dfrac{f_{i+}f_{+j}}{N}$ である．

$$\chi_p^2=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(f_{ij}-\dfrac{f_{i+}f_{+j}}{N}\right)^2}{\dfrac{f_{i+}f_{+j}}{N}}=\frac{(1-0.66667)^2}{0.66667}+\frac{(0-0.66667)^2}{0.66667}+\cdots+\frac{(1-0.66667)^2}{0.66667}+\frac{(1-0.66667)^2}{0.66667}$$

$$=6.000$$

表 11　実現度数（回答頻度）と期待度数 $e_{ij}=Np_{i+}p_{+j}=\dfrac{f_{i+}f_{+j}}{N}$ の表

実現度数（上段）：f_{ij} 期待度数（下段）：e_{ij}	銘柄 1	銘柄 2	銘柄 3	行和
回答者 1	1 0.66667	0 0.66667	1 0.66667	2
回答者 2	0 0.33333	1 0.33333	0 0.33333	1
回答者 3	1 0.33333	0 0.33333	0 0.33333	1
回答者 4	0 0.66667	1 0.66667	1 0.66667	2
列和	2	2	2	6 （総和）

つぎに，上で用いた記法に従って，順をおってそれぞれ必要な情報を作ってみよう．なおここでは，あえて“筆算”により数値確認を行ってみる．

① 2 元データ表から行列 $\mathbf{F}=\left(f_{ij}\right)$ の用意

$$\mathbf{F}=\begin{pmatrix}1&0&1\\0&1&0\\1&0&0\\0&1&1\end{pmatrix}$$

ここで，回答者：$I=\{1,2,3,4\}$，また銘柄：$J=\{1,2,3\}$

② 行列 $\mathbf{P}_{IJ},\mathbf{P}_I,\mathbf{P}_J$ を $\mathbf{F}=\left(f_{ij}\right)$ から作る．

$$\mathbf{P}_{IJ}=\begin{pmatrix}\frac{1}{6}&0&\frac{1}{6}\\0&\frac{1}{6}&0\\\frac{1}{6}&0&0\\0&\frac{1}{6}&\frac{1}{6}\end{pmatrix},\ \mathbf{P}_I=\begin{pmatrix}\frac{2}{6}&0&0&0\\0&\frac{1}{6}&0&0\\0&0&\frac{1}{6}&0\\0&0&0&\frac{2}{6}\end{pmatrix},\ \mathbf{P}_J=\begin{pmatrix}\frac{2}{6}&0&0\\0&\frac{2}{6}&0\\0&0&\frac{2}{6}\end{pmatrix}$$

③ よって行列 $\mathbf{Q}$ とそれから作られる $\mathbf{V}^*$ は以下のようになる．

$$\underset{4\times3}{\mathbf{Q}}=\underset{4\times4}{\mathbf{P}_I^{-1/2}}\underset{4\times3}{\mathbf{P}_{IJ}}\underset{3\times3}{\mathbf{P}_J^{-1/2}}=\begin{pmatrix}\frac{1}{2}&0&\frac{1}{2}\\0&\frac{1}{\sqrt{2}}&0\\\frac{1}{\sqrt{2}}&0&0\\0&\frac{1}{2}&\frac{1}{2}\end{pmatrix}\Rightarrow\ \mathbf{V}^*=\mathbf{Q}^t\mathbf{Q}=\begin{pmatrix}\frac{3}{4}&0&\frac{1}{4}\\0&\frac{3}{4}&\frac{1}{4}\\\frac{1}{4}&\frac{1}{4}&\frac{1}{2}\end{pmatrix}$$

④ 固有方程式を作り，固有値，固有ベクトルを求める．

$$\left|\mathbf{V}^*-\lambda\mathbf{I}\right|=\begin{vmatrix}\frac{3}{4}-\lambda&0&\frac{1}{4}\\0&\frac{3}{4}-\lambda&\frac{1}{4}\\\frac{1}{4}&\frac{1}{4}&\frac{1}{2}-\lambda\end{vmatrix}=0$$

これは行列式で，この式から固有値を求める代数方程式は $\left(4\lambda-3\right)\left(4\lambda-1\right)\left(\lambda-1\right)=0$ となる[26]．これを解くと，3 つの固有値は，値の大きい方から，$\lambda_0=1,\lambda_1=\frac{3}{4},\lambda_2=\frac{1}{4}$ を得る．ここ

[26] ここでは，λ の 3 次方程式となったので，3 つの根がある．

で，最大根 $\lambda_0 = 1$ に対する固有ベクトルを求めると，

$$\begin{pmatrix} \sqrt{p_{+1}} \\ \sqrt{p_{+2}} \\ \sqrt{p_{+3}} \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{2}{6}} \\ \sqrt{\frac{2}{6}} \\ \sqrt{\frac{2}{6}} \end{pmatrix} = \begin{pmatrix} \sqrt{\frac{1}{3}} \\ \sqrt{\frac{1}{3}} \\ \sqrt{\frac{1}{3}} \end{pmatrix}$$ となる．これは，いわゆる**自明解**である[27].

これを除く 2 つの固有値 $\lambda_1 = \frac{3}{4} = 0.75, \lambda_2 = \frac{1}{4} = 0.25$ に対する固有ベクトルを求めると，それぞれ，

$$\mathbf{l}_1 = \begin{pmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \\ 0 \end{pmatrix}, \mathbf{l}_2 = \begin{pmatrix} \frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \\ -\frac{2}{\sqrt{3}} \end{pmatrix}$$

がえられる．これを改めて行列 $\mathbf{L} = (\mathbf{l}_1, \mathbf{l}_2)$ とする．つまり，

$$\mathbf{L} = \begin{pmatrix} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ 0 & \frac{2}{\sqrt{3}} \end{pmatrix} = \begin{pmatrix} -0.707 & -0.408 \\ 0.707 & -0.408 \\ 0.000 & 0.816 \end{pmatrix}$$

である．

⑤ つぎに回答者 $i \in I$ に対する成分スコアを求める．
式（37）に上で求めた各数値を代入すると，以下となる．

[27] ここでは行列 $\mathbf{V}^*$ を用いた．これを式（23）あるいは表 8 にある行列 $\mathbf{V}$ を用いると，$\lambda_0 = 0$ が自明解に対応する固有値となる．あとの固有値は同じ値となる．

$$\underset{m\times K}{\mathbf{Z}} = \underbrace{\mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2}}_{m\times n}\underset{n\times K}{\mathbf{L}} = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 3 \end{pmatrix} \times \begin{pmatrix} \frac{1}{6} & 0 & \frac{1}{6} \\ 0 & \frac{1}{6} & 0 \\ \frac{1}{6} & 0 & 0 \\ 0 & \frac{1}{6} & \frac{1}{6} \end{pmatrix} \times \begin{pmatrix} \sqrt{3} & 0 & 0 \\ 0 & \sqrt{3} & 0 \\ 0 & 0 & \sqrt{3} \end{pmatrix} \times \begin{pmatrix} -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ 0 & \frac{2}{\sqrt{3}} \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{\sqrt{3}}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} \\ \sqrt{\frac{3}{2}} & \frac{1}{\sqrt{2}} \\ -\sqrt{\frac{3}{2}} & \frac{1}{\sqrt{2}} \\ \frac{\sqrt{3}}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} \end{pmatrix} = \begin{pmatrix} -0.6123 & -0.3536 \\ 1.2247 & 0.7071 \\ -1.2247 & 0.7071 \\ 0.6123 & -0.3536 \end{pmatrix}$$

⑥ さらに銘柄 $j \in J$ に対する成分スコアを求める．

ここでは，うしろに示した双対性の式（50）を用いよう[28]．これから，以下がえられる．

$$\mathbf{Z}^* = \mathbf{P}_I^{-1}\mathbf{P}_{JI}\mathbf{Z}\mathbf{\Lambda}^{-1/2} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 3 \end{pmatrix} \times \begin{pmatrix} \frac{1}{6} & 0 & \frac{1}{6} & 0 \\ 0 & \frac{1}{6} & 0 & \frac{1}{6} \\ \frac{1}{6} & 0 & 0 & \frac{1}{6} \end{pmatrix} \times \begin{pmatrix} -\frac{\sqrt{3}}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} \\ \sqrt{\frac{3}{2}} & \frac{1}{\sqrt{2}} \\ -\sqrt{\frac{3}{2}} & \frac{1}{\sqrt{2}} \\ \frac{\sqrt{3}}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} \end{pmatrix} \times \begin{pmatrix} \frac{2}{\sqrt{3}} & 0 \\ 0 & 2 \end{pmatrix}$$

$$= \begin{pmatrix} -\frac{3}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ \frac{3}{2\sqrt{2}} & \frac{1}{2\sqrt{2}} \\ 0 & -\frac{1}{\sqrt{2}} \end{pmatrix} \fallingdotseq \begin{pmatrix} -1.0606 & 0.3536 \\ 1.0606 & 0.3536 \\ 0 & -0.7071 \end{pmatrix}$$

2．3　統計ソフトウェアによる確認

ここでは電卓を使った筆算で解いてみたが，これを統計ソフトウェア（JMP）により解いてみよう．その結果は以下となった．なお，コンピュータ内での計算は固有の数値計算アルゴリズムがあって，上に筆算で確認したような手順は用いてはいない．とくに，テキスト型データなどで扱う，データ表の寸法が大きく，しかもセル内の度数が非常に疎な行列の場合には，特別な手当が必要となることがある[29]．

---

[28] うしろに示した特異値分解によれば，特異ベクトルの行列を用いれば得られる．

[29] たとえば，M.W. Berry (1992), Large-Scale Sparse Singular Value Computation, *The International Journal of Supercomputer Applications*, **6**, 1, 13-49.この他にもいろいろある．WordMiner でも特別な方法を使っている．

表 12　得られた固有値と寄与率

固有値	特異値	寄与率
$\lambda_1 = 0.75$	0.8660	75%
$\lambda_2 = 0.25$	0.5	25%

表 13　固有ベクトル（R^n から解いたとき）

	λ_1 に対して	λ_2 に対して
	$\mathbf{l}_1$	$\mathbf{l}_2$
銘柄 1	-0.707	-0.408
銘柄 2	0.707	-0.408
銘柄 3	0	0.816

なおここで，以下が確認される．まず，カイ二乗統計量と固有値との間に以下の関係がある[30]．

$$\sum_{k=1}^{K}\lambda_K = \frac{\chi_p^2}{N} \;\Rightarrow\; \sum_{k=1}^{2}\lambda_K = \lambda_1+\lambda_2 = 0.75+0.25 = \frac{\chi_p^2}{6} = 1$$

また，行列 $\mathbf{V}^* = \mathbf{Q}^t\mathbf{Q}$ の対角の和（跡和あるいはトレース）と固有値，カイ二乗統計量の間にはつぎの関係がある[31]．

$$tr\left(\mathbf{V}^*\right)-1 = \sum_{k=1}^{K}\lambda_K = \frac{\chi_p^2}{N} \;\Rightarrow\; tr\left(\mathbf{V}^*\right)-1 = \left(\underbrace{\frac{3}{4}+\frac{3}{4}+\frac{1}{2}}_{=2}\right)-1 = \sum_{k=1}^{K}\lambda_K = \frac{\chi_p^2}{6} = 1$$

表 14　行側（回答者）の行成分スコア（分散＝ λ）

項目：I	回答者	第 1 成分スコア $\mathbf{z}_1$	第 2 成分スコア $\mathbf{z}_2$
1	回答者 1	0.6124	-0.3536
2	回答者 2	-1.2247	0.7071
3	回答者 3	1.2247	0.7071
4	回答者 4	-0.6124	-0.3536

表 15　列側（銘柄）の列成分スコア（分散＝ λ）

項目：J	銘柄	第 1 成分スコア $\mathbf{z}_1^*$	第 2 成分スコア $\mathbf{z}_2^*$
1	銘柄 1	1.0607	0.3536
2	銘柄 2	-1.0607	0.3536
3	銘柄 3	0.0000	-0.7071

こうして，当たり前のことであるが，ここで得た結果は，前にみた筆算による結果に一致する．

---

[30] これは，クロス表の連関性の測度の 1 つである**平均平方関連係数** ϕ^2 でもある（35 ページ脚注参照）．

[31] 前に示した，式（27）の $tr\left(\mathbf{V}\right) = \sum_{k=1}^{K}\lambda_K = \frac{\chi_p^2}{N}$ に対応する．ここで「1」が固有値 $\lambda_0 = 1$ に対応する．

［求めた成分スコアの観察と特徴］

ここでえられた成分スコアはどのようなことを示しているのだろうか．その主な性質をすこし調べる．

① 成分スコアの意味

対応分析で得られた**成分スコア**は，項目 I の 4 つの選択肢（4 人の「回答者」）と項目 J の 3 つの選択肢（3 つの「銘柄」）という質的データを数量として表したことになる．たとえばこれが，調査における質問文の選択肢「満足，やや満足，あまり満足でない，満足でない」つまり順位尺度の選択肢に数量として解釈できるスコアを与えることができることを意味する．この点で，数量化法 III 類に通底する．この数量化 III 類的に考えた場合における成分スコアの相関係数と特異値の関係については後述する．

② プロファイル間のカイ二乗距離と成分スコアのユークリッド距離

対応分析では，プロファイル間の距離をカイ二乗距離として定義する．また，得られた成分スコア間のユークリッド距離は，もとの 2 元データ表について得られるカイ二乗距離に対応している．

まず，表 7 から，回答者間のカイ二乗距離は以下となる．これに従って 4 名の回答者間のもとのクロス表から得た**平方カイ二乗距離**を求めると表 16 となった．

$$d_B^2(i,i') = \sum_{j=1}^{n} \frac{1}{p_{+j}} \left(q_{ij} - q_{i'j}\right)^2 = \sum_{j=1}^{n} \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2 = \sum_{j=1}^{n} \left(\frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}} - \frac{p_{i'j}}{p_{i'+}\sqrt{p_{+j}}} \right)^2 \tag{48}$$

たとえば，回答者 1 と回答者 3 との間のカイ二乗統計量を求めてみよう．

$$d_B^2\left(i=\text{回答者}1, i'=\text{回答者}3\right) = \sum_{j=1}^{n} \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2$$

$$= \frac{1}{p_{+1}} \left(\frac{p_{i1}}{p_{i+}} - \frac{p_{i'1}}{p_{i'+}} \right)^2 + \frac{1}{p_{+2}} \left(\frac{p_{i2}}{p_{i+}} - \frac{p_{i'2}}{p_{i'+}} \right)^2 + \frac{1}{p_{+3}} \left(\frac{p_{i3}}{p_{i+}} - \frac{p_{i'3}}{p_{i'+}} \right)^2$$

$$= \frac{1}{0.3333}(0.5-1)^2 + \frac{1}{0.3333}(0-0)^2 + \frac{1}{0.3333}(0.5-0)^2 = 1.5$$

よって，**カイ二乗距離**は，$d_B\left(i=\text{回答者}1, i'=\text{回答者}3\right) = \sqrt{1.5} \fallingdotseq 1.22474$ となる．以下，すべての組合せを調べると表 16 のようになる．

表 16 プロファイルから得た回答者間のカイ二乗距離の表

	回答者 1	回答者 2	回答者 3	回答者 4
回答者 1	0	2.121320344	1.224744871	1.224744871
回答者 2		0	2.449489743	1.224744871
回答者 3			0	2.121320344
回答者 4				0

つぎに，行の成分スコアについて，同じように**平方ユークリッド距離**を求めると表 17 となる．たとえば，上と同じ回答者 1 と 3 の成分スコア間の平方ユークリッド距離は，

$$d_E^2\left(i=\text{回答者}1, i'=\text{回答者}3\right)=\sum_{k=1}^{K}\left(z_{ik}-z_{i'k}\right)^2=\left(z_{i1}-z_{i'1}\right)^2+\left(z_{i2}-z_{i'2}\right)^2$$

$$=\left(0.6124-1.2247\right)^2+\left(-0.3536-0.7071\right)^2=0.37491129+1.12508449\fallingdotseq 1.5000$$

となる．よって，ユークリッド距離は，$d_E\left(i=\text{回答者}1, i'=\text{回答者}3\right)=1.22474$ となり上に一致する．さらにここでも，すべての回答者間のユークリッド距離を求めると表 17 となる．ここで数値のわずかの違いは計算上の誤差の範囲で，両者は一致している．

表 17　成分スコアから得た回答者間のユークリッド距離

	回答者 1	回答者 2	回答者 3	回答者 4
回答者 1	0	2.121320343	1.224744872	1.224744872
回答者 2		0	2.449489742	1.224744871
回答者 3			0	2.121320344
回答者 4				0

もちろん，この関係は列（銘柄）のプロファイル間のカイ二乗距離と，列成分スコアのユークリッド距離についてもなり立つ．このことは，行成分スコア，列成分スコアを用いたさまざまな**幾何学的な性質を**前提して行う**事後の分析**にとって，非常に便利な特性である．たとえば，成分スコアに対して，クラスター化（自動分類，クラスタリング）を行うことも幾何学的には矛盾なく適用される．とくに，大量のテキスト型データ，自由回答・自由記述データなどを扱うときには非常に有効である[33]．

③ 行と列の成分スコア間の相関係数と特異値

行あるいは列の同じ成分（k）に対応する成分スコア間の相関係数は，固有値の（正の）平方根 $\sqrt{\lambda_k}$ あるいは特異値 α_k に相当する．これを上の例で確認しよう．

まず，第 1 成分について得られた成分スコアを行，列について大きさの順に並べ替えて得られる情報を作る．これが表 17 である[34]．表をみて明らかなように，行列要素「1」が右上から左下に向けて対角に並んでいることがわかる（ほぼ線型の関係にある）．ここで項目 I と項目 J のそれぞれに与えられた第 1 成分スコアを散布図として表すと図 6 が得られる[35]．この質的データ（回答者，銘柄）に付与された数量得点である第 1 成分スコアの相関係数を求めると，$r=0.866025$ となるが，これは第 1 特異値（α_1）に等しい．そして確かに第 1 固有値の正の平方根でもある（$\lambda_1=\alpha_1^2=0.866025^2=0.7499993\fallingdotseq 0.75$）．同様に，第 2 成分スコアについても，回答者と銘柄の相関係数が第 2 特異値 $\alpha_2=0.5$ よって第 2 固有値 $\lambda_2=\alpha_2^2=0.25$ となる．

表 17　第1成分スコアで並べ替えたデータ表と第 1 成分スコアの関係

F から	銘柄 2	銘柄 3	銘柄 1	第 1 成分スコア： z_{i1}
回答者 3	0	0	1	1.2247
回答者 1	0	1	1	0.6124
回答者 4	1	1	0	-0.6124
回答者 2	1	0	0	-1.2247
第 1 成分スコア: z_{j1}^*	-1.0607	0.0000	1.0607	

[33] つまり，質的データを成分スコアとして数量化することの利点が最大限に活かされる．クラスター化を実際にどう行うか，またテキスト・マイニングでどう用いるかについては，別の資料「第Ⅲ部」で説明する．

[34] これを**序列化**（ordination）ということがある．

[35] これを**対の散布図**（twin-map）あるいは**双対散布図**ということがある（Gauch, H.G. et al., 1977）．

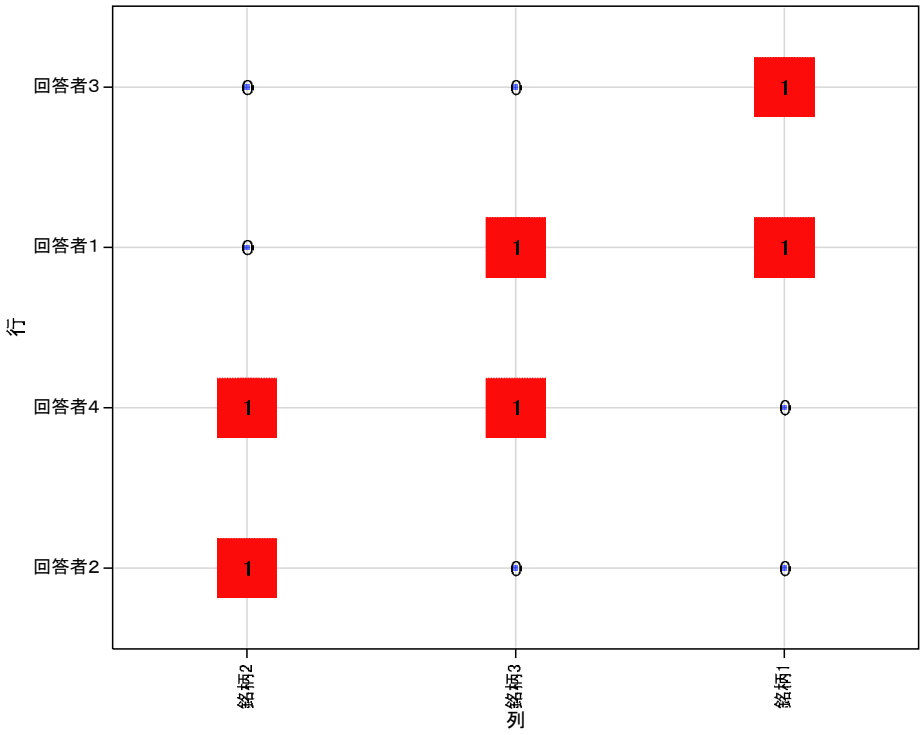
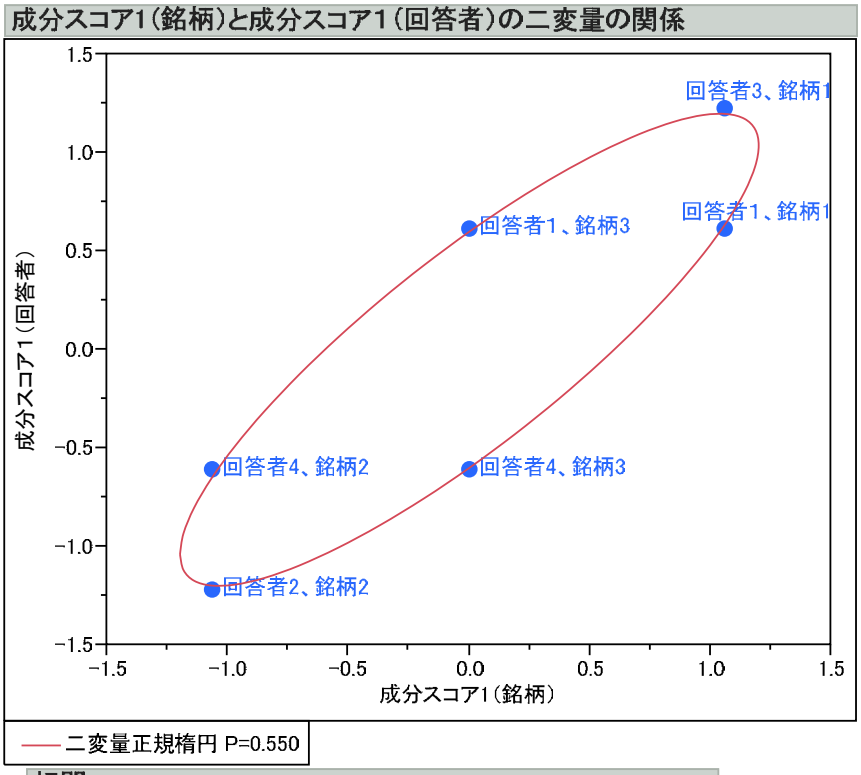


図 5　対の散布図をバブルプロットとした例
（第 1 成分スコアで並べ替えたクロス表の度数に対応させた）



相関

変数	平均	標準偏差	相関	p値	数
成分スコア1（銘柄）	−4.8e−16	0.948683	0.866025	0.0257*	6
成分スコア1（回答者）	1.11e−16	0.948683			

図 6　第 1 成分スコアについて描いた対の散布図（twin－map）の例
成分スコアを数量データ（量的データ）とみて相関係数を算出

④ 特異値分解の確認

ここで，上の例について，**特異値分解**から内容を確かめておこう．いま，

$$\underset{m\times n}{\mathbf{C}} = \sum_{k=1}^{K} \sqrt{\lambda_k}\,\mathbf{u}_k \mathbf{l}_k^t = \underset{m\times K}{\mathbf{U}}\ \underset{K\times K}{\Lambda^{1/2}}\ \underset{K\times n}{\mathbf{L}^t} \tag{49}$$

と書く．前に得た固有ベクトルと特異値（固有値の平方根）から，

$$\mathbf{U} = \begin{pmatrix} 0.408 & -0.408 \\ -0.577 & 0.577 \\ 0.577 & 0.577 \\ -0.408 & -0.408 \end{pmatrix},\ \mathbf{L} = \begin{pmatrix} -0.707 & 0.408 \\ 0.707 & 0.408 \\ 0.000 & -0.816 \end{pmatrix},\ \mathbf{\Lambda}^{1/2} = \begin{pmatrix} \sqrt{0.75} & 0 \\ 0 & \sqrt{0.25} \end{pmatrix}$$

であるから，下の行列を得る．

$$\underset{m\times K}{\mathbf{U}}\ \underset{K\times K}{\Lambda^{1/2}}\ \underset{K\times n}{\mathbf{L}^t} = \begin{pmatrix} 0.408 & -0.408 \\ -0.577 & 0.577 \\ 0.577 & 0.577 \\ -0.408 & -0.408 \end{pmatrix} \times \begin{pmatrix} \sqrt{0.75} & 0 \\ 0 & \sqrt{0.25} \end{pmatrix} \times \begin{pmatrix} -0.707 & 0.707 & 0 \\ 0.408 & 0.408 & -0.816 \end{pmatrix}$$

$$= \begin{pmatrix} 0.1666 & -0.3330 & 0.1665 \\ -0.2356 & 0.4710 & -0.2354 \\ 0.4710 & -0.2356 & -0.2354 \\ -0.3330 & 0.1666 & 0.1665 \end{pmatrix}$$

一方，

$$y_{ij}^* = \frac{p_{ij} - p_{i+}p_{+j}}{\sqrt{p_{i+}p_{+j}}}\ \left(i \in I, j \in J\right)$$

を要素とする行列 $\mathbf{Y}^* = \left(y_{ij}^*\right)$ を作ってみよう．

$$\mathbf{Y}^* = \left(y_{ij}^*\right) = \mathbf{P}_I^{-1/2}\left(\mathbf{P}_{IJ} - \mathbf{r}\mathbf{c}^t\right)\mathbf{P}_J^{-1/2} \tag{50}$$

$$\mathbf{P}_{IJ} = \left(p_{ij}\right) = \begin{pmatrix} \frac{1}{6} & 0 & \frac{1}{6} \\ 0 & \frac{1}{6} & 0 \\ \frac{1}{6} & 0 & 0 \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \end{pmatrix} = \begin{pmatrix} 0.1667 & 0.0000 & 0.1667 \\ 0.0000 & 0.1667 & 0.0000 \\ 0.1667 & 0.0000 & 0.0000 \\ 0.0000 & 0.1667 & 0.1667 \end{pmatrix}$$

$$\mathbf{P}_I^{-1/2}=\begin{pmatrix}1.732051 & 0 & 0 & 0\\ 0 & 2.44949 & 0 & 0\\ 0 & 0 & 2.44949 & 0\\ 0 & 0 & 0 & 1.732051\end{pmatrix},$$

$$\mathbf{P}_J^{-1/2}=\begin{pmatrix}1.732051 & 0 & 0\\ 0 & 1.732051 & 0\\ 0 & 0 & 1.732051\end{pmatrix}$$

$$\mathbf{r}=\begin{pmatrix}\frac{2}{6}\\ \frac{1}{6}\\ \frac{1}{6}\\ \frac{2}{6}\end{pmatrix},\mathbf{c}=\begin{pmatrix}\frac{2}{6}\\ \frac{2}{6}\\ \frac{2}{6}\end{pmatrix},\mathbf{rc}^t=\begin{pmatrix}0.111111 & 0.111111 & 0.111111\\ 0.055556 & 0.055556 & 0.055556\\ 0.055556 & 0.055556 & 0.055556\\ 0.111111 & 0.111111 & 0.111111\end{pmatrix}$$

$$\mathbf{Y}^*=\left(y_{ij}^*\right)=\mathbf{P}_I^{-1/2}\left(\mathbf{P}_{IJ}-\mathbf{rc}^t\right)\mathbf{P}_J^{-1/2}=\begin{pmatrix}0.16667 & -0.33333 & 0.16667\\ -0.23570 & 0.47140 & -0.23570\\ 0.47140 & -0.23570 & -0.23570\\ -0.33333 & 0.16667 & 0.16667\end{pmatrix}$$

これは，上の行列$\mathbf{C}$に一致している（数値のわずかのずれは数値計算上の誤差）.

なおここで，固有ベクトルとその行列について，以下の関係がある，これを確認しておこう.

$$\mathbf{U}=\begin{pmatrix}0.408 & -0.408\\ -0.577 & 0.577\\ 0.577 & 0.577\\ -0.408 & -0.408\end{pmatrix},\ \mathbf{U}^t=\begin{pmatrix}-0.408 & 0.577 & 0.408 & -0.577\\ -0.408 & 0.577 & -0.408 & 0.577\end{pmatrix}$$

$$\Rightarrow \mathbf{U}^t\mathbf{U}=\begin{pmatrix}0.999698 & 0\\ 0 & 0.998784\end{pmatrix}=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$$

$$\mathbf{L}=\begin{pmatrix}-0.707 & 0.408\\ 0.707 & 0.408\\ 0.000 & -0.816\end{pmatrix},\ \mathbf{L}^t=\begin{pmatrix}-0.707 & 0.707 & 0\\ 0.408 & 0.408 & -0.816\end{pmatrix}$$

$$\Rightarrow \mathbf{L}\mathbf{L}^t=\begin{pmatrix}0.998786 & 5.55112E-17\\ 5.55112E-17 & 0.998786\end{pmatrix}=\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$$

（*）ここで「E-17」とは 10^{-17} つまりほぼ 0 ということ.

つまり，ここで，行列$\mathbf{U}$は，行列$\mathbf{Y}^*\left(\mathbf{Y}^*\right)^t$の固有ベクトルを列ベクトルとする行列であり，行列$\mathbf{L}$は行列$\left(\mathbf{Y}^*\right)^t\mathbf{Y}^*$の固有ベクトルを列ベクトルとする行列である．また，$\mathbf{U}^t\mathbf{U}$，$\mathbf{L}\mathbf{L}^t$ともに単位行列となり，その行列の階数は$K=\min\{m,n\}-1$であり，固有ベクトルが互いに直

交していることを示している（そうなるように固有ベクトルを求めた）．

ここまでで，対応分析法の導出方法や計算方法の要点を説明した．他の導出方法や，線形代数などの数理的な背景については参考文献を参照されたい．

3. 成分スコアとその性質

3. 1 双対性について

対応分析法では，項目 I の選択肢と，項目 J の選択肢のそれぞれに対して，**成分スコア**が付与される．このとき，所与のデータ表（2 元データ表）の，行側（空間 R^n 内）の分析とまったく同じ方法で，列側（空間 R^m 内）に対しても分析を行える[36]．つまり行と列との**“対称性”**（symmetry）があり，これが対応分析法の特徴である．利用上はそれら両者の成分スコアの相互の関係を知ることが重要である（つまり，データ表の行と列の双方向から考察する）．また，成分スコアと，もとのデータ表との関係は模式図で眺めることが理解を容易にするので，これを以下に示す（**“布置図による視覚化”**も対応分析法の重要な特性）．

まず，項目 I の選択肢，項目 J の選択肢それぞれに与えられる成分スコアを，前に用意の記法に従いつぎのように表す．

$z_{ik}\left(i \in I, k = 1,2,\cdots,K\right)$（選択肢 $i \in I$ に対する第 k 成分の行成分スコア）

$z^*_{jk}\left(j \in J, k = 1,2,\cdots,K\right)$（選択肢 $j \in J$ に対する第 k 成分の列成分スコア）

これと，もとの 2 元データ表のプロファイルとの関係を模式的に示す（図 7）．

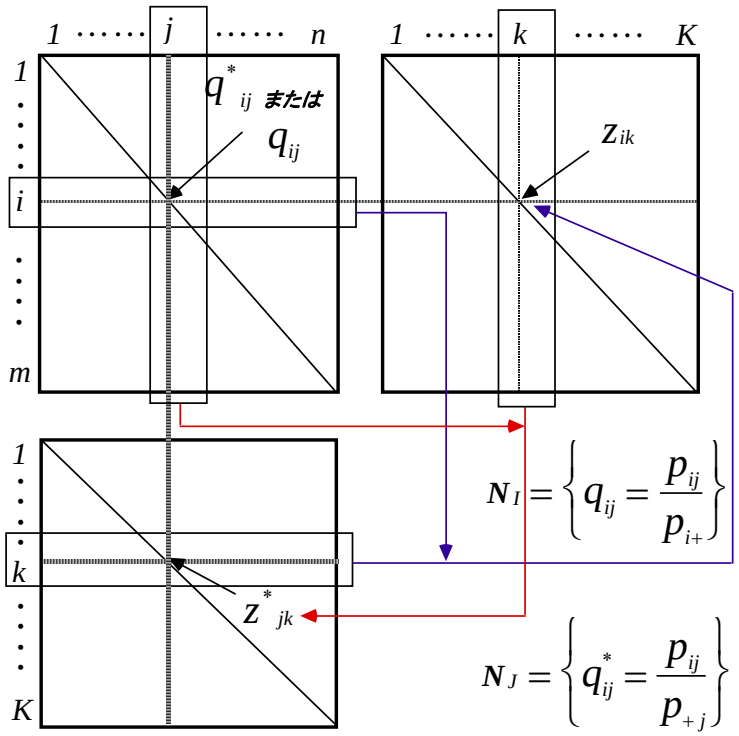


図 7 成分スコアとプロファイルの関係（双対性）を示す模式図

この 2 項目 I，J の成分スコア間には，つぎのような関係がある．このような関係を**“双対性”**（duality）という．

[36] こうではない，非対称の方法も考えられている．

$$\underset{m\times 1}{\mathbf{z}_k} = \frac{1}{\sqrt{\lambda_k}} \underset{m\times m}{\mathbf{P}_I^{-1}} \underset{m\times n}{\mathbf{P}_{IJ}} \underset{n\times 1}{\mathbf{z}_k^*} \tag{51}$$

$$z_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{n} \left(\frac{p_{ij}}{p_{i+}} \right) z_{jk}^* \left(i \in I; k = 1,2,\cdots,K \right) \tag{52}$$

$$\underset{m\times K}{\mathbf{Z}} = \mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{Z}^*\mathbf{\Lambda}^{-1/2} \quad （ここで \mathbf{\Lambda}^{-1/2} = diag\left(1/\sqrt{\lambda_k}\right)） \tag{53}$$

$$\underset{n\times 1}{\mathbf{z}_k^*} = \frac{1}{\sqrt{\lambda_k}} \underset{n\times n}{\mathbf{P}_J^{-1}} \underset{n\times m}{\mathbf{P}_{JI}} \underset{m\times 1}{\mathbf{z}_k} \tag{54}$$

$$z_{jk}^* = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{m} \left(\frac{p_{ij}}{p_{+j}} \right) z_{ik} \left(j \in J; k = 1,2,\cdots,K \right) \tag{55}$$

$$\underset{n\times K}{\mathbf{Z}^*} = \mathbf{P}_J^{-1}\mathbf{P}_{JI}\mathbf{Z}\mathbf{\Lambda}^{-1/2} \quad （ここで \mathbf{\Lambda}^{-1/2} = diag\left(1/\sqrt{\lambda_k}\right)） \tag{56}$$

この式の意味は重要である．これを読み解くと，つぎの重要な性質がある．

- “行成分スコアは，列成分スコアを重みとした，行プロファイルの加重平均となる”こと．
- 一方，“列成分スコアは，行成分スコアを重みとした，列プロファイルの加重平均となる”こと

上の 2 つの式で，z_{ik}，z_{jk}^* が，たすきがけになって左右の項に入っていることに注意しよう（上の図 7 と表 16 で確認）．このように，上の式（52），（55）は互いに**推移関係**（transition relationship）にあることから，**“推移公式”**（transition equation）あるいは**“遷移方程式”**ともいう[37]．これ以上の数理的定式化については他の参考文献に譲って，ここでは具体的に利用上の主な性質について要約する．

3．2 成分スコアの解釈

得られた成分スコアは図に描いて観察する．前述のように，視覚化は対応分析法の有用な機能の 1 つである．これを“**布置図**”（representation）あるいは“**同時布置図**”（simultaneous representation）という．

① 成分スコアの散布図（布置図）

行あるいは列の選択肢に対する成分スコア，つまり表 16 にある成分スコアのうち，作図に必要な（観察したい）2 成分 k,k' を指定して散布図を描き成分スコアの分布を観察する．また多数の成分を同時に比較するには，**散布図行列（多変量連関図）**なども用いる．

$$\left(z_{ik}, z_{ik'} \right) \quad \begin{pmatrix} i = 1,2,\cdots,m \\ k,k' = 1,2,\cdots,K \\ K = \min\{m,n\} - 1 \end{pmatrix} \quad （行成分スコア） \tag{57}$$

[37] formule de transition（仏）． transition formula ともいう．訳語もいろいろある（例：遷移方程式）．

$$\left(z^{*}_{jk}, z^{*}_{jk'}\right) \quad \begin{pmatrix} i=1,2,\cdots,m \\ k,k'=1,2,\cdots,K \\ K=\min\{m,n\}-1 \end{pmatrix} \qquad \text{（列成分スコア）} \qquad (58)$$

表 16　項目 I, J の選択肢の成分スコアと確率行列の関係

		項　目　J						行成分スコア							
		1	2	⋯	j	⋯	n	1	2	⋯	k	⋯	k'	⋯	K
項目 I	1	p_{11}	p_{12}	⋯	p_{1j}	⋯	p_{1n}	z_{11}	z_{12}	⋯	z_{1k}	⋯	$z_{1k'}$	⋯	z_{1K}
	2	p_{21}	p_{22}	⋯	p_{2j}	⋯	p_{2n}	z_{21}	z_{22}	⋯	z_{2k}	⋯	$z_{2k'}$	⋯	z_{2K}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	i	p_{i1}	p_{i2}	⋯	p_{ij}	⋯	p_{in}	z_{i1}	z_{i2}	⋯	z_{ik}	⋯	$z_{ik'}$	⋯	z_{iK}
	⋮	⋮	⋮		⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	m	p_{m1}	p_{m2}	⋯	p_{mj}	⋯	p_{mn}	z_{m1}	z_{m2}	⋯	z_{mk}	⋯	$z_{mk'}$	⋯	z_{mK}
列成分スコア	1	z^{*}_{11}	z^{*}_{21}	⋯	z^{*}_{j1}	⋯	z^{*}_{n1}	⇑							
	2	z^{*}_{12}	z^{*}_{22}	⋯	z^{*}_{j2}	⋯	z^{*}_{n2}	行の項目 I の選択肢の成分スコア							
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	$\mathbf{Z}_{m\times K} = \underbrace{\mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2}}_{m\times n}\mathbf{L}_{n\times K}$							
	k	z^{*}_{1k}	z^{*}_{2k}	⋯	z^{*}_{jk}	⋯	z^{*}_{nk}	⇐ 列の項目 J の選択肢の成分スコア							
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	$\mathbf{Z}^{*}_{n\times K} = \underbrace{\mathbf{P}_J^{-1}\mathbf{P}_{JI}\mathbf{P}_I^{-1/2}}_{n\times m}\mathbf{U}_{m\times K}$							
	k'	$z^{*}_{1k'}$	$z^{*}_{2k'}$	⋯	$z^{*}_{jk'}$	⋯	$z^{*}_{nk'}$	（ここで $K=\min\{m,n\}-1$）							
	⋮	⋮	⋮	⋮	⋮	⋮	⋮								
	K	z^{*}_{1K}	z^{*}_{2K}	⋯	z^{*}_{iK}	⋯	z^{*}_{nK}								

［成分スコアを観察する際の注意事項］

(1)　まず，個々の成分スコアを 1 次元的に観察する．もっとも情報が多い（最大固有値つまり最大の分散 λ_1 に対する），行と列との第 1 成分スコアを**数直線上に並べて**描いてみると良い[38]．第 1 固有値の寄与率が非常に大きいときにはこの操作が大切である．とくに，“はずれ値”は，初めの方の成分，とくに第 1 成分に現れやすい．これは，はずれ値が成分スコアの分散を大きくするからである（原理からあきらか）．

(2)　つぎに，（固有値が大きい方から順に）2〜3 の成分スコアに注目し，布置図を描き各点の布置の相対的な位置関係に注目する．たとえば，固有値の大きい方から，第 1 成分スコアと第 2 成分スコア，第 3 成分スコアなどを比べる．同時に比較するために，**散布図行列**などを用いるのもよい．

(3)　場合に応じて，後述の“**寄与度**”（**相対寄与度，絶対寄与度**）を目安に，成分スコアを観察する．たとえば，“**絶対寄与度**”をもとに成分軸を解釈する．また，成分軸に解釈を与えるだけではなく，成分スコアの布置図の中での相対的な遠近，位置関係を観察する．また，“**相対寄与度**”により，項目の選択肢がどの成分でよりよく代表されるか（その選択肢にとって説明力が高い成分か）を調べる．

(4)　“多重クロス表”（バート表）から求めたサンプルの成分スコアの解釈は「もとの変量・

[38] たとえば，資料の第 I 部「対応分析法とは（概要）」の，19 ページ，図 2 を参照．

項目の選択肢のスコア」(つまりアイテム・カテゴリー型に展開した延べのカテゴリーに付与の成分スコア）であるから意味理解に注意する（とくに選択肢の並び順，順序・序列関係に注意）.

(5) **固有値，寄与率**の解釈は，多重クロス表，とくにインジケータ行列から出発した場合は，その特性上大きくなることはほとんどないので注意する．見かけ上，高い寄与率は数理的に現れることがない[39].

(6) このとき，インジケータ行列，バート表他の多重対応分析で扱う各データ表の固有値と寄与率について**“調整済み慣性や寄与率”**も提案されているので，それも参考にする．うしろに数値例で示す.

(7) 質問文（項目）の選択肢が**順序尺度**の場合には図中の選択肢の並び順に注意する．並び順がもとの選択肢のそれと異なり崩れたときには，その質問文の選択肢の作り方を再吟味することがよい.

(8) 成分スコアを用いたクラスター化操作には十分な注意が必要である．単純な k-平均法や階層的分類ではうまく対応できないことがある．カイ二乗統計量（つまり総変動）の分解を利用したクラスター化，つまり成分スコアによるクラスター化の工夫が必要である[40].

(9) 布置図の上では，成分軸の端のほうにある点から観察する（重心から遠い位置にある点から観察）．そして，どの成分軸に近いかを観察する．寄与度との併用がよい.

(10) とくに**“はずれ値”**の存在に注意する．はずれ値はもとのデータ表の中の頻度分布の不均衡つまりプロファイルの不均衡から生じる[41]．これは対応分析の特徴でもある.

(11) 行と列の成分スコアの**同時布置**を考えたとき，それらの標準化（平均値＝0，分散＝1とすること）の有無に注意する．それぞれを標準偏差で（つまり，特異値 α_k あるいは固有値 λ_k の正の平方根で）「標準化する場合」と「標準化しない場合」があるので，少なくとも 4 通りの組み合わせがある[42]（下の表 20）．通常の対応分析法では，**いずれも標準化しない**ことが多い（組合せの「その 1」，成分スコアの分散は固有値 λ_k のままを用いる）．その理由については参考文献（大隅他（1994），Lebart 他（1998））を参照のこと.

(12) なお，平均値は標準化の有無に関係なく，常にゼロとなるように調整される[43].

表 20　成分スコアの分散の組み合わせ

	組合せ	項目 I の選択肢の成分スコア：z_{ik}	項目 J の選択肢の成分スコア：z^*_{jk}
分散の大きさ	その 1	λ_k	λ_k
	その 2	λ_k	1
	その 3	1	λ_k
	その 4	1	1

② スコアの同時布置図

行，列それぞれの選択肢への成分スコアを重ねた散布図を**“同時布置図”**（simultaneous

[39] これについての性質は，付録の[補足]で若干説明した．**調整済みの固有値（慣性）や寄与率**なども使う.
[40] たとえば，WordMiner では階層的分類法の 1 つウォード法と k-平均法とを併用した**ハイブリッド法**を用いる.
[41] これを調整する 1 つの方法として，追加処理がある（後述）.
[42] この組合せの中で，同時布置図として幾何学的な描画に意味があるのは，おそらく「成分スコア vs 成分スコア」，「成分スコア vs 標準化成分スコア」，「標準化成分スコア vs 成分スコア」の 3 つである.
[43] これは成分スコア z_{ik}，z^*_{jk} をそのまま平均するということではなく，もとのクロス表の行和，列和ベクトルを加重とする加重平均値であることに注意．これについては，前に数値例で示した.

representation）という．すなわち，

$$\left(z_{ik}, z_{ik'}\right), \left(z^*_{jk}, z^*_{jk'}\right) \qquad \begin{pmatrix} i = 1,2,\cdots,m \\ k,k' = 1,2,\cdots,K \\ K = \min\{m,n\} - 1 \end{pmatrix} \tag{59}$$

を同じ散布図上に図として描く．この同時布置図に関しては，こうした表示方法が適切かを巡っての議論がある[44]．

4. 対応分析のいくつかの性質

4. 1 主要な指標とその使い方

対応分析の分析結果を適切に解釈するため，さまざまな指標が必要である．ここでは，多くの統計ソフトウェアが出力表示する主要な指標をいくつか説明する（例：WordMiner, JMP, JMP スクリプトなど）．一見すると面倒にみえるが，あとに述べる例題を参考にする，あるいは自分で人工的にミニチュアなデータセットを作ってみて，諸指標がどのように機能するものか，何を測っているかを知るのがよいだろう．

① 固有値と寄与率

行列 $\mathbf{V}^*$ から得られる**固有値**の系列 $\lambda_k \left(k = 1,2,\cdots,K; K = \min\{m.n\} - 1\right)$ から，以下の関係と**寄与率**が得られる．なお，すでに述べたように，**固有値の個数**はもとの解析対象とした 2 元データ表（クロス表）の行と列の寸法の小さい方から 1 を引いた個数（$K = \min\{m,n\} - 1$）となる（つまり，成分スコアの分布は，この次元数内の空間に入るということ）．

$$tr\left(\mathbf{V}^*\right) - 1 = \sum_{k=1}^{K} \lambda_k \quad \left(K = \min\{m,n\} - 1\right) \quad \text{（固有値の和であり全慣性）} \tag{60}$$

ここで，$tr\left(\mathbf{V}^*\right)$ は行列 $\mathbf{V}^*$ の対角要素の和（跡和トレース），を示す[45]．

$$\textbf{寄与率}：\nu_k = \frac{\lambda_k}{\sum_{k=1}^{K} \lambda_k} \times 100(\%) \quad \begin{pmatrix} k = 1,2,\cdots,K \\ K = \min\{m,n\} - 1 \end{pmatrix} \quad \text{（第 } k \text{ 成分の寄与率の式）} \tag{61}$$

この ν_k を k について累積すれば**累積寄与率**となる．なお，固有値の値は非負で，かつ 1 を越えることはない（つまり，$0 \le \lambda_k \le 1 \left(k = 1,2,\cdots,K; K = \min\{m,n\} - 1\right)$ である）．

② クロス表の独立性の検定との関係

クロス表の行と列の関係を統計的検定として評価する一つのモデルとして K. ピアソンの考えた“**独立性の検定**”があることは，すでに触れた．これは，行と列の 2 つの項目 I, J の間には関係がないという**帰無仮説**をたてて（つまり**独立モデル**，$p_{ij} = p_{i+}p_{+j}$ を仮説として），これが統計的に棄却されれば，帰無仮説を棄却，つまり行と列の 2 つの項目 I, J の間には何らかの関係がないとはいえない（つまり関係がありそうと言えるだろう）とする検定法であ

---

[44] 実は，同時布置については，さまざまな議論がある．ここでは，対応分析法でとられている方式を示したが，これとは別の方式を提案している例もある．たとえば，J.D. Carroll et al.(1986, 1987)，西里（1980），朝野（2008）．

[45] ここで，行列 $\mathbf{V}$ を用いると $tr\left(\mathbf{V}\right) = \sum_{k=1}^{K} \lambda_k$ となる．

る[46]（背理法的な二重否定的肯定となり，ある意味，隔靴掻痒である）．

ところで，すでに述べたように，対応分析法はこれの見方を変えて，データ表の行と列の 2 つの項目 I，J の間にどの程度の関係があるのかを成分スコアとそれらの相関係数という具体的な数量として示すことにある．たとえば，固有値の和（総変動，全慣性）とピアソンのカイ二乗統計量との間には，前述のようにつぎの関係がある[47]．

$$tr\left(\mathbf{V}^{*}\right)-1=\frac{\chi_p^2}{N}=\sum_{k=1}^{K}\lambda_k \quad \left(K=\min\left\{m,n\right\}-1\right) \tag{62}$$

ここで χ_p^2 はいわゆる**ピアソンのカイ二乗統計量**であり，これはいままでに用意した記号・記法を用いるとつぎのように書ける．

$$\chi_p^2=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{N\left(p_{ij}-p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}} \quad \text{または}[48] \quad \phi^2=\frac{\chi_p^2}{N}=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(p_{ij}-p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}} \tag{63}$$

$$\chi_p^2=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{N\left(p_{ij}-p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}}=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(f_{ij}-\dfrac{f_{i+}f_{+j}}{N}\right)^2}{\dfrac{f_{i+}f_{+j}}{N}} \tag{64}$$

ここで，$\dfrac{f_{i+}f_{+j}}{N}$ は，クロス表の独立性の検定で，**独立モデル**（$p_{ij}=p_{i+}p_{+j}$）を仮定したときの第 (i,j) セルの**期待度数**で，$e_{ij}=Np_{i+}p_{+j}=N\dfrac{f_{i+}}{N}\dfrac{f_{+j}}{N}=\dfrac{f_{i+}f_{+j}}{N}$ となる．つまり，上の式はつぎのようにも書ける．

$$\chi_p^2=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(f_{ij}-\dfrac{f_{i+}f_{+j}}{N}\right)^2}{\dfrac{f_{i+}f_{+j}}{N}}=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(f_{ij}-e_{ij}\right)^2}{e_{ij}}=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(\text{実現度数}-\text{期待度数}\right)^2}{\text{期待度数}}$$

クロス表（分割表）の**独立性の検定**では，このピアソンのカイ二乗統計量 χ_p^2 が自由度 $(m-1)(n-1)$ の χ^2 **分布に“近似する”**ことを使って χ^2 検定を行う．

---

[46] 2 元データ表の寸法が大きくなり，また総度数 N が大きくなると，検定結果はほとんど有意となる．

[47] なんども指摘のように行列 $\mathbf{V}$ の跡和と，$tr\left(\mathbf{V}\right)=tr\left(\mathbf{V}^{*}\right)-1$ の関係にある．よって $tr\left(\mathbf{V}\right)=\dfrac{\chi_p^2}{N}=\sum_{k=1}^{K}\lambda_k$．

[48] $\phi^2=\chi_p^2/N$ はクロス表の**連関性の測度**（measures of association）の 1 つで**平均平方関連係数**（mean square contingency coefficient）という（Everitt（1977）など）．またその正の平方根（ϕ）をファイ係数という．

$$\chi_p^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\left(f_{ij} - \dfrac{f_{i+} f_{+j}}{N} \right)^2}{\dfrac{f_{i+} f_{+j}}{N}} \sim \chi^2_{(m-1)(n-1)} \tag{65}$$

③ プロファイルの総変動の関係

行のプロファイルあるいは列のプロファイルの各点の分布の**全慣性**（total inertia）つまり**総変動**には以下の関係がある．

（ｉ）行プロファイルについて

$$\left(\text{全慣性}\right) = \sum_{i=1}^{m} \left(i\text{番目の質量}\right) \times \begin{pmatrix} i\text{番目のプロファイルから} \\ \text{その重心までのカイ二乗距離} \end{pmatrix} \tag{66}$$

これを式で書くと，以下となる．

$$\phi^2 = \frac{\chi_p^2}{N} = \sum_{i=1}^{m} p_{i+} \sum_{j=1}^{n} \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - p_{+j} \right)^2 = \sum_{i=1}^{m} p_{i+} \sum_{j=1}^{n} \left(\frac{q_{ij}}{\sqrt{p_{+j}}} - \sqrt{p_{+j}} \right)^2 \tag{67}$$

（ⅱ）列プロファイルについて

$$\left(\text{全慣性}\right) = \sum_{j=1}^{n} \left(j\text{番目の質量}\right) \times \begin{pmatrix} j\text{番目のプロファイルから} \\ \text{その重心までのカイ二乗距離} \end{pmatrix} \tag{68}$$

これは以下のように書ける．

$$\phi^2 = \frac{\chi_p^2}{N} = \sum_{j=1}^{n} p_{+j} \sum_{i=1}^{m} \frac{1}{p_{i+}} \left(\frac{p_{ij}}{p_{+j}} - p_{i+} \right)^2 = \sum_{j=1}^{n} p_{+j} \sum_{i=1}^{m} \left(\frac{q_{ij}^*}{\sqrt{p_{i+}}} - \sqrt{p_{i+}} \right)^2 \tag{69}$$

ここでも**カイ二乗距離**が重要な意味をもって機能していることがわかる．要するに，対応分析法で重要なことは（すでに数値例でもみたように）所与の 2 元データ表の総変動であり，行または列のプロファイルの変動でもあるカイ二乗統計量を，主成分（主軸でありもとのプロファイルの合成変数）に対応する成分スコアの分散（つまり固有値）の大きさにしたがって，切り分けることにある．主成分分析でも，“高次元（多変量）のデータ表を，少数次元の主成分という合成変数で説明する”という視点から導出できる．対応分析も，前述のように，ほぼ同様の方法で合成変数を導出できる．この点で 2 つの手法は，典型的な**“節約の原理”**（principle of parsimony）に則るアプローチである．

④ 再生公式

上の関係に関連して以下の重要な公式が知られている．この，p_{ij}，p_{i+}，p_{+j} と成分スコアとの間に成り立つ式を，**“再生公式”**（reconstitution formula）という[49]．これは，p_{ij} が右辺のように p_{i+}，p_{+j} と成分スコアの合成式で復元できることを示している[50]．

---

[49] formules de reconstituion（仏）

[50] ここで，$1 + \sum_{k=1}^{K} \frac{1}{\sqrt{\lambda_k}} z_{ik} z_{jk}^*$ の始めの「1」は，自明の固有値 λ_0 に対応すると考える．

$$p_{ij} = p_{i+}p_{+j}\left\{1+\sum_{k=1}^{K}\frac{1}{\sqrt{\lambda_k}}z_{ik}z_{jk}^{*}\right\} = p_{i+}p_{+j} + p_{i+}p_{+j}\left\{\sum_{k=1}^{K}\frac{1}{\sqrt{\lambda_k}}z_{ik}z_{jk}^{*}\right\} \quad (70)$$
$$\left(i\in I, j\in J, K=\min\{m,n\}-1\right)$$

あるいは，以下のようにも書ける．

$$\frac{p_{ij}}{p_{i+}p_{+j}} = 1+\sum_{k=1}^{K}\frac{1}{\sqrt{\lambda_k}}z_{ik}z_{jk}^{*} \quad \left(i\in I, j\in J, K=\min\{m,n\}-1\right) \quad (71)$$

さらに，度数 f_{ij}, f_{i+}, f_{+j} を用いて，以下のように表すこともできる．

$$f_{ij} = \left(\frac{f_{i+}f_{+j}}{N}\right)\left\{1+\sum_{k=1}^{K}\frac{1}{\sqrt{\lambda_k}}z_{ik}z_{jk}^{*}\right\} \quad \left(i\in I, j\in J, K=\min\{m,n\}-1\right) \quad (72)$$

この数式の右辺の括弧内の第 2 項を除外すると，ピアソンのカイ二乗統計量を使ってクロス表の独立性の検定を行う際に設定する独立モデル（$p_{ij} = p_{i+}p_{+j}$）となっていることに注意しよう．つまり，第 2 項に固有値と成分スコアが含まれ，この項が独立モデルからの乖離の程度を測っていることになる．このようにここでも，カイ二乗統計量との関係が表れる．つまり，前に約束したようなプロファイルやデータ行列から出発した理由の 1 つがここにある（行列 $\underset{m\times n}{\mathbf{X}} = \left(x_{ij}\right)$ や $\underset{m\times n}{\mathbf{Q}} = \left(y_{ij}\right)$，あるいは $\mathbf{Y}^{*} = \left(y_{ij}^{*}\right)$ と設定することで，上のような各関係が成り立つ）．

⑤ 絶対寄与度

絶対寄与度（absolute contributions）あるいは**単に寄与度**とは，第 k 成分の中に選択肢 $i\left(\in I\right)$ または選択肢 $j\left(\in J\right)$ が占める寄与の程度を表す指標である．つまり，ある成分 k に注目したとき，その成分の中で選択肢 $i\left(\in I\right)$ または選択肢 $j\left(\in J\right)$ がどの程度意味を持って働いているかを知るときに用いる．言い換えると，**ある成分 k の軸の解釈**を行う目安となる．またこれを 100 倍して%として示す場合もある．

（ⅰ）第 k 成分における選択肢 $i\left(\in I\right)$ の絶対寄与度

$$C_k\left(i\right) = \frac{p_{i+}\left(z_{ik}\right)^2}{\lambda_k} \quad \left(\begin{matrix} i\in I, k=1,2,\cdots,K \\ K=\min\{m,n\}-1 \end{matrix}\right), \ \sum_{i=1}^{m}C_k\left(i\right)=1 \quad (73)$$

（ⅱ）第 k 成分における選択肢 $j\left(\in J\right)$ の絶対寄与度

$$C_k\left(j\right) = \frac{p_{+j}\left(z_{jk}^{*}\right)^2}{\lambda_k} \quad \left(\begin{matrix} j\in J, k=1,2,\cdots,K \\ K=\min\{m,n\}-1 \end{matrix}\right), \ \sum_{j=1}^{n}C_k\left(j\right)=1 \quad (74)$$

⑥ 相対寄与度

相対寄与度（relative contributions）あるいは**平方相関**（squared correlations）とは，該当する点（選択肢）が各成分軸によって，どれほど良く**近似されているか**（説明できるか）を示す指標である．これは，点（選択肢）によって表されるベクトルと成分軸との角度を θ とする

と，図 8 にみるように，$\cos\theta$ の 2 乗，つまり，相関の 2 乗で計算される．

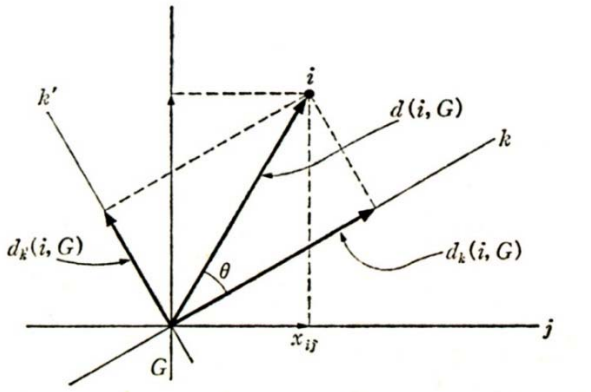


図 8　相対寄与度の考え方（模式図）

(ⅰ)選択肢 $i(\in I)$ に対する相対寄与度

重心 G から点 i（選択肢 i，空間 R^n 内の点 i）までの距離を考える．これは，選択肢 i と，全体の平均（重心）とのユークリッド距離である．この距離を 2 乗した平方距離は，つぎの式で求められる．

$$d^2(i,G)=\sum_{k=1}^{K} d_k^2(i,G) \tag{75}$$

この平方距離に対して，第 k 成分 $d_k^2(i,G)$ が占める割合を**“相対寄与度”**とする．上の図 8 で考えると，分かりやすいであろう．図 8 には，もとのプロファイルの座標空間（j と j'；$j,j'=1,2,\cdots,n$）に対して，得られた空間（直交する k と k'；$k,k'=1,2,\cdots,K$）のイメージを示してある．ここから $d_k^2(i,G)=z_{ik}^2$ でもあるから，相対寄与度は下の式で与えられる[51]．

$$C_k^*(i)=\frac{d_k^2(i,G)}{d^2(i,G)}=\frac{z_{ik}^2}{\sum_{j=1}^{n}\frac{1}{p_{+j}}\left(\frac{p_{ij}}{p_{i+}}-p_{+j}\right)^2}\quad\begin{pmatrix} i\in I, k=1,2,\cdots,K \\ K=\min\{m,n\}-1 \end{pmatrix} \tag{76}$$

図からわかるように，$\cos_k^2\theta=C_k^*(i)$ であるから，この指標を**平方相関**（squared correlation）と呼ぶこともある．また，$\sum_{k=1}^{K}\cos_k^2\theta=\sum_{k=1}^{K}C_k^*(i)=1$ となるので，上の $C_k^*(i)$ を 100 倍して，% として用いることもある．

(ⅱ)選択肢 $j(\in J)$ に対する相対寄与度

同じように，選択肢 i（空間 R^n 内の点 i）と，その成分軸 k への射影を考えて，**選択肢 $j(\in J)$**

---

[51] 式（76）で，$\sum_{j=1}^{n}\frac{1}{p_{+j}}\left(\frac{p_{ij}}{p_{i+}}-p_{+j}\right)^2=\sum_{j=1}^{n}\left(\frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}-\sqrt{p_{+j}}\right)^2=\sum_{j=1}^{n}\left(x_{ij}-\bar{x}_j\right)^2$ となる．18 ページの式（18），(20)を参照．つまり，重心（G）からの平方距離に相当する．式（77）も $i\in I$ を $j\in J$ と読み替えれば同様．

に対する相対寄与度を作れば以下となる．

$$C_k^*(j)=\frac{d_k^2(j,G)}{d^2(j,G)}=\frac{\left(z_{jk}^*\right)^2}{\sum_{i=1}^{m}\frac{1}{p_{i+}}\left(\frac{p_{ij}}{p_{+j}}-p_{i+}\right)^2}\quad\left(\begin{array}{l}j\in J,k=1,2,\cdots,K\\K=\min\{m,n\}-1\end{array}\right)\tag{77}$$

なお，ここでも上と同様に，$C_k^*(j)$を 100 倍して%として用いることもある．

数式で表すとやや煩雑にみえるが，これらを覚える必要はない．利用上は，ソフトウェアが出力するこれらの寄与度の情報の読み方・解釈を理解すればよい．これは別に例題として実際にソフトウェア（例：WordMiner や JMP スクリプト）が出力する情報を使って説明する[52]．また，この観察の仕方，解釈が大切である．

⑦ カイ二乗距離を用いることと「分布の同等性」

前に示した，カイ二乗距離をふたたび考えよう．

（ｉ）項目 I の 2 つの選択肢 i と i' の間の平方カイ二乗距離

$$d_B^2(i,i')=\sum_{j=1}^{n}\frac{1}{p_{+j}}\left(q_{ij}-q_{i'j}\right)^2=\sum_{j=1}^{n}\frac{1}{p_{+j}}\left(\frac{p_{ij}}{p_{i+}}-\frac{p_{i'j}}{p_{i'+}}\right)^2$$
$$=\sum_{j=1}^{n}\left(\frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}-\frac{p_{i'j}}{p_{i'+}\sqrt{p_{+j}}}\right)^2\tag{78}$$

（ⅱ）項目 J の 2 つの選択肢 j と j' の間の平方カイ二乗距離

$$d_B^2(j,j')=\sum_{i=1}^{m}\frac{1}{p_{i+}}\left(q_{ij}^*-q_{ij'}^*\right)^2=\sum_{i=1}^{m}\frac{1}{p_{i+}}\left(\frac{p_{ij}}{p_{+j}}-\frac{p_{ij'}}{p_{+j'}}\right)^2$$
$$=\sum_{i=1}^{m}\left(\frac{p_{ij}}{p_{+j}\sqrt{p_{i+}}}-\frac{p_{ij'}}{p_{+j'}\sqrt{p_{i+}}}\right)^2\tag{79}$$

ここで選択肢間の**プロファイルの距離**として，いわゆるユーリッド距離を用いずに，上のように質量（p_{i+},p_{+j}）を加重とする**カイ二乗距離**（Chi-square distance）として扱った．この理由の一つは，**“分布の同等性”**（distributional equivalency あるいは equivalence of distribution）[53]を保証するためである．分布の同等性とは以下のように要約される．

[分布の同等性（distributional equivalency）]

① プロファイルが同じ列（つまり，比率パターンが同じ列）を併合しても，行間のカイ二乗距離は変化しない．
② 同じように，プロファイルが同じ行（つまり，比率パターンが同じ行）を併合しても，列間のカイ二乗距離は変化しない．
③ つまり，プロファイルが同じ行同士を，あるいは列同士を併合しても，対応分析法の分析結果は変わらない．

---

52 「第Ⅰ部」にも例題を付けた．
53 principe d’equivalence distributionnelle（仏）

この性質の証明は他の文献に譲り[54]，ここではこの性質を簡単な数値例で説明しよう．

表 20　15 人の回答データとそれを併合したデータ表

回答者	銘柄 A	銘柄 B	銘柄 C
回答者 1	1	0	1
回答者 2	1	0	1
回答者 3	1	0	1
回答者 4	1	0	1
回答者 5	1	0	1
回答者 6	0	1	0
回答者 7	0	1	0
回答者 8	0	1	0
回答者 9	1	0	0
回答者 10	1	0	0
回答者 11	1	0	0
回答者 12	0	1	1
回答者 13	0	1	1
回答者 14	0	1	1
回答者 15	0	1	1
列和 f_{+j}	8	7	9

回答者	銘柄 A	銘柄 B	銘柄 C	行和 f_{i+}
回答者 1 ～5	5	0	5	10
回答者 6～8	0	3	0	3
回答者 9～11	3	0	0	3
回答者 12～15	0	4	4	8
列和 f_{+j}	8	7	9	24

例 1：

上で用いた表 10 のトイ・データと同じ質問への回答を，15 人の回答者から得たとしよう．その結果，表 22 のデータ表が得られたとする．ここでは意図的に（分かりやすくするため），同じ回答パターンとなった人をそろえて並べてある．つまり**“行のプロファイルが同じ人”**が何人かいるということである．たとえば，始めの 5 名はいずれも「銘柄 1」「銘柄 3」を選んでいる．そこでこれら 5 名は同じ回答パターンということで併合することにする（行で積み重ねる）．これを他の同じ回答パターンの回答者にも行い表を整理（併合圧縮）すると表 20 の右の表にように書ける．

ここで，この 2 つのデータ表にそれぞれ対応分析を適用すると，同じ結果が得られる，また，そうなるためには，前に約束した行プロファイル間の距離としてカイ二乗距離を用いることが必要である．列の側についても同様で，同じ列プロファイルを併合しても，結果は変わらない．

この性質は，扱う 2 元データ表の寸法が**大きく**，しかも表の要素が非常に**疎である**ような場合に効果を発揮する．典型例は，自由回答・自由記述やウェブサイト情報で扱うようなボリュームが非常に大きくしかもデータ表の要素は少ないテキスト型データの処理などでみられる．

i)　ほとんどの要素の度数が少数，とくに「1」が多く，しかも特定の語句が頻出するような場合．

ii)　要素の度数と語句の関係に若干の揺らぎはあるものの，内容が類似した語句（例：同義語）つまり類似の行プロファイルあるいは列プロファイルが多いような場合．

表 21　得られた固有値と寄与率

固有値	特異値	寄与率
$\lambda_1 = 0.70273$	$\alpha_1 = 0.83829$	77.9%
$\lambda_2 = 0.19905$	$\alpha_2 = 0.44615$	22.1%

[54] たとえば，Jambu（1989），大隅（1989）など．

表 22　回答者と銘柄の成分スコア（分散＝λ）

項目		成分スコア 1	成分スコア 2
項目 I		z_{i1}	z_{i2}
回答者	回答者 1	-0.6022	-0.2323
	回答者 2	-0.6022	-0.2323
	回答者 3	-0.6022	-0.2323
	回答者 4	-0.6022	-0.2323
	回答者 5	-0.6022	-0.2323
	回答者 6	1.38457	0.71523
	回答者 7	1.38457	0.71523
	回答者 8	1.38457	0.71523
	回答者 9	-1.1485	0.82518
	回答者 10	-1.1485	0.82518
	回答者 11	-1.1485	0.82518
	回答者 12	0.66429	-0.2873
	回答者 13	0.66429	-0.2873
	回答者 14	0.66429	-0.2873
	回答者 15	0.66429	-0.2873
項目 J		z^*_{j1}	z^*_{j2}
銘柄	銘柄 1	-0.9628	0.36816
	銘柄 2	1.16067	0.3191
	銘柄 3	-0.0469	-0.5754

表 23　回答者と銘柄の成分スコア（分散＝λ）

項目		成分スコア 1	成分スコア 2
項目 I		z_{i1}	z_{i2}
回答者	回答者 1 ～5	-0.6022	-0.2323
	回答者 6～8	1.38457	0.71523
	回答者 9～11	-1.1485	0.82518
	回答者 12～15	0.66429	-0.2873
項目 J		z^*_{j1}	z^*_{j2}
銘柄	銘柄 1	-0.9628	0.36816
	銘柄 2	1.16067	0.3191
	銘柄 3	-0.0469	-0.5754

こういう場面では，2 元データ表の行プロファイルあるいは列プロファイルが同じか，あるいはほとんど同じプロファイルが頻出する．こういうときに，同等あるいは類似プロファイルを併合処理して扱えるという利点がある[55]．

例 2：

たとえば，（100 点法で得た）4 つの科目の成績得点データで，

- 生徒 A : 15, 12, 10, 15
- 生徒 B : 90, 72, 60, 90(生徒 A の 6 倍)
- 生徒 C : 30, 24, 20, 30(生徒 A の 2 倍)

となった 3 人に対しては，対応分析法では行プロファイルが同じとなるので，実質的には同じ成分スコアを示すことになる．しかし通常の主成分分析を適用するとこの点数の比例倍の影響が分散を変えることになるので，3 名の生徒は異なるパターンとみなされる．つまり，総変動の見方が主成分分析などとは異なるのである．

[55] 現実のデータ解析では，こうした編集作業が必要になる．確かな結果を得るには手間がかかるのである．

⑧ 成分スコアのユークリッド距離とプロファイル間のカイ 2 乗距離

すでに上で数値例として調べたように，**“成分スコアの点間の（平方）ユークリッド距離は，プロファイル間の（平方）カイ二乗距離に等しい”**という性質がある（証明は下）．この性質は，成分スコアを別の分析（2 次分析）に用いるときに有効である．とくに，成分スコア間の平方ユークリッド距離を用いたクラスター化は，もとの 2 元データ表の行（あるいは列）のプロファイルの平方カイ二乗距離を用いたクラスター化に同じとなることは，対応分析法とクラスター化の明確な関係を示す重要な特性である．

この性質があることで，ユークリッド距離あるいは“平方”ユークリッド距離（つまり，平方和あるいは分散）を用いるさまざまなクラスター化手法を，対応分析で得た成分スコアに対してそのまま利用できる．たとえば，WordMiner では，階層的分類法の 1 つである**ウォード法**（Ward’s method）と，非階層的分類法（分割化型分類法）の 1 つである ***k*-平均法**（*k*-means method）を併用するハイブリッド方式のクラスター化法を用いている[56]．

この関係は，以下のように示すことができる．まず，行の側に注目して，行プロファイル間の**カイ二乗距離**を考える．これはつぎの式であった．

$$\begin{aligned} d_B^2(i,i') &= \sum_{j=1}^{n} \frac{1}{p_{+j}} \left(q_{ij} - q_{i'j} \right)^2 = \sum_{j=1}^{n} \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2 \\ &= \sum_{j=1}^{n} \left(\frac{p_{ij}}{p_{i+}\sqrt{q_{+j}}} - \frac{p_{i'j}}{p_{i'+}\sqrt{q_{+j}}} \right)^2 \end{aligned} \tag{80}$$

一方，再生公式から，以下がなり立つ．

$$p_{ij} = p_{i+} p_{+j} \left\{ 1 + \sum_{k=1}^{K} \frac{1}{\sqrt{\lambda_k}} z_{ik} z_{jk}^* \right\} \quad \left(i \in I, j \in J, K = \min\{m,n\} - 1 \right) \tag{81}$$

ここで，(81) を (80) に代入すると，以下となる．

$$\begin{aligned} d_B^2(i,i') &= \sum_{j=1}^{n} \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2 \\ &= \sum_{j=1}^{n} \frac{1}{p_{+j}} \left\{ p_{+j} \left(1 + \sum_{k=1}^{K} \frac{1}{\sqrt{\lambda_k}} z_{ik} z_{jk}^* \right) - p_{+j} \left(1 + \sum_{k=1}^{K} \frac{1}{\sqrt{\lambda_k}} z_{i'k} z_{jk}^* \right) \right\}^2 \\ &= \sum_{j=1}^{n} \sum_{k=1}^{K} \frac{p_{+j}}{\lambda_k} \left(z_{ik} z_{jk}^* - z_{i'k} z_{jk}^* \right)^2 = \sum_{j=1}^{n} \sum_{k=1}^{K} \frac{p_{+j}}{p_{+j} \left(z_{jk}^* \right)^2} \left(z_{ik} z_{jk}^* - z_{i'k} z_{jk}^* \right)^2 \\ &= \sum_{k=1}^{K} \left(z_{ik} - z_{i'k} \right)^2 = d_E^2(i',i) \end{aligned} \tag{82}$$

ここで，（z_{kj}^* の分散が）$\sum_{j=1}^{n} p_{+j} \left(z_{kj}^* \right)^2 = \lambda_k$ となることを用いた．

---

[56] このことについては，上に簡単な例で示した．また別の資料「第Ⅲ部」に記したのでそれを参照のこと．

同様に，列の側の選択肢 j, j' についても，以下がなり立つ．

$$d_B^2(j,j') = \sum_{i=1}^{m} \frac{1}{p_{i+}} \left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}} \right)^2 = \sum_{k=1}^{K} \left(z_{kj}^* - z_{kj'}^* \right)^2 = d_E^2(j,j') \qquad (83)$$

つまり，プロファイル間の**平方カイ二乗距離** d_B^2 は，成分スコア間の**平方ユークリッド距離** d_E^2 に等しいことが示される．

4. 2 対応分析法の分析手順の要約

ここで，いままでに得た対応分析法を行ううえで必要な情報を，表に要約しておこう．いままで述べたことから，2 元データ表の，行の側からの分析（n 次元空間 R^n 内での分析）と，列の側からの分析（m 次元空間 R^m 内での分析）に分けて示してある．とりあえず対応分析法を使うためには，この表のそれぞれの要素がどのような意味を持ち，どのように使えるかを知れば十分であろう．

表 24 分析の基本要素

	項目 I ：行の側から分析	項目 J ：列の側から分析
	n 次元空間 R^n 内での分析	m 次元空間 R^m 内での分析
	行和を 1 としたときの「行のプロファイル」で $(n-1)$ 次元内に分布する m 個の点	列和を 1 としたときの「列のプロファイル」で $(m-1)$ 次元内に分布する n 個の点
プロファイル	行のプロファイル $\mathbf{N}_I = \mathbf{P}_I^{-1}\mathbf{P}_{IJ} = \left\{ q_{ij} = \frac{p_{ij}}{p_{i+}} \middle\| i \in I, j \in J \right\}$	列のプロファイル $\mathbf{N}_J = \mathbf{P}_J^{-1}\mathbf{P}_{JI} = \left\{ q_{ij}^* = \frac{p_{ij}}{p_{+j}} \middle\| i \in I, j \in J \right\}$
プロファイル間の距離	行のプロファイル間のカイ二乗距離 $d_B^2(i,i') = \sum_{j=1}^{n} \frac{1}{p_{+j}} \left(q_{ij} - q_{i'j} \right)^2$ $= \sum_{j=1}^{n} \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2$	列のプロファイル間のカイ二乗距離 $d_B^2(j,j') = \sum_{i=1}^{m} \frac{1}{p_{i+}} \left(q_{ij}^* - q_{ij'}^* \right)^2$ $= \sum_{i=1}^{m} \frac{1}{p_{i+}} \left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}} \right)^2$
２つの距離の関係	プロファイル間の平方カイ二乗距離は成分スコア間の平方ユークリッド距離に等しい．	
	$d_B^2(i,i') = \sum_{j=1}^{n} \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2$ は， $d_E^2(i,i') = \sum_{k=1}^{K} \left(z_{ik} - z_{i'k} \right)^2$ に等しい	$d_B^2(j,j') = \sum_{i=1}^{m} \frac{1}{p_{i+}} \left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}} \right)^2$ は． $d_E^2(j,j') = \sum_{k=1}^{K} \left(z_{kj}^* - z_{kj'}^* \right)^2$ に等しい
固有値 寄与率 累積寄与率	固有値：$\lambda_k \left(\begin{matrix} k = 1,2,\cdots,K \\ K = \min\{m,n\} - 1 \end{matrix} \right)$ ここで $0 \le \lambda_k \le 1$ 寄与率：$\nu_k = \frac{\lambda_k}{\sum_{k=1}^{K} \lambda_k} \times 100(\%) \left(\begin{matrix} k = 1,2,\cdots,K \\ K = \min\{m,n\} - 1 \end{matrix} \right)$ （第 k 成分の寄与率） λ_k を k について累積すれば**累積寄与率** η_k となる．	

<table>
<tr><td>総変動・総分散
全慣性</td><td colspan="2">$\chi_p^2=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{N\left(p_{ij}-p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}}$　あるいは　$\phi^2=\frac{\chi_p^2}{N}=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(p_{ij}-p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}}$
ここで，固有値との間に，$\phi^2=\frac{\chi_p^2}{N}=\sum_{k=1}^{K}\lambda_k\left(K=\min\left\{m,n\right\}-1\right)$の関係がある．</td></tr>
<tr><td>総変動・総分散
全慣性</td><td>$\phi^2=\frac{\chi_p^2}{N}=\sum_{i=1}^{m}p_{i+}\sum_{j=1}^{n}\frac{1}{p_{+j}}\left(\frac{p_{ij}}{p_{i+}}-p_{+j}\right)^2$
$=\sum_{i=1}^{m}p_{i+}\sum_{j=1}^{n}\left(\frac{q_{ij}}{\sqrt{p_{+j}}}-\sqrt{p_{+j}}\right)^2$</td><td>$\phi^2=\frac{\chi_p^2}{N}=\sum_{j=1}^{n}p_{+j}\sum_{i=1}^{m}\frac{1}{p_{i+}}\left(\frac{p_{ij}}{p_{+j}}-p_{i+}\right)^2$
$=\sum_{j=1}^{n}p_{+j}\sum_{i=1}^{m}\left(\frac{q_{ij}^*}{\sqrt{p_{i+}}}-\sqrt{p_{i+}}\right)^2$</td></tr>
<tr><td>絶対寄与度</td><td>第 k 成分における選択肢$i\left(\in I\right)$の絶対寄与度
$C_k\left(i\right)=\frac{p_{i+}\left(z_{ik}\right)^2}{\lambda_k}\quad\left(\begin{matrix}i\in I,k=1,2,\cdots,K\\K=\min\left\{m,n\right\}-1\end{matrix}\right)$
$\sum_{i=1}^{m}C_k\left(i\right)=1$</td><td>第 k 成分における選択肢 $j\left(\in J\right)$ の絶対寄与度
$C_k\left(j\right)=\frac{p_{+j}\left(z_{jk}^*\right)^2}{\lambda_k}\quad\left(\begin{matrix}j\in J,k=1,2,\cdots,K\\K=\min\left\{m,n\right\}-1\end{matrix}\right)$
$\sum_{j=1}^{n}C_k\left(j\right)=1$</td></tr>
<tr><td>相対寄与度
平方相関
(*) 100 倍して用いることがある．</td><td>選択肢$i\left(\in I\right)$が，第k成分によって，どれぐらい近似されているかを示す指標
選択肢$i\left(\in I\right)$に対する相対寄与度
$C_k^*\left(i\right)=\frac{d_k^2\left(i,G\right)}{d^2\left(i,G\right)}=\frac{z_{ik}^2}{\sum_{j=1}^{n}\frac{1}{p_{+j}}\left(\frac{p_{ij}}{p_{i+}}-p_{+j}\right)^2}\quad\left(\begin{matrix}i\in I,k=1,2,\cdots,K\\K=\min\left\{m,n\right\}-1\end{matrix}\right)$
$d^2\left(i,G\right)$：点$i\in I$から重心Gまでのカイ二乗距離</td><td>選択肢 $j\left(\in J\right)$ が，第k成分によって，どれぐらい近似されているかを示す指標
選択肢 $j\left(\in J\right)$ に対する相対寄与度
$C_k^*\left(j\right)=\frac{d_k^2\left(j,G\right)}{d^2\left(j,G\right)}=\frac{\left(z_{jk}^*\right)^2}{\sum_{i=1}^{m}\frac{1}{p_{i+}}\left(\frac{p_{ij}}{p_{+j}}-p_{i+}\right)^2}\quad\left(\begin{matrix}j\in J,k=1,2,\cdots,K\\K=\min\left\{m,n\right\}-1\end{matrix}\right)$
$d^2\left(j,G\right)$：点 $j\in J$ から重心Gまでのカイ二乗距離</td></tr>
<tr><td>双対性
推移公式</td><td>$\underset{m\times1}{\mathbf{z}_k}=\frac{1}{\sqrt{\lambda_k}}\underset{m\times m}{\mathbf{P}_I^{-1}}\underset{m\times n}{\mathbf{P}_{IJ}}\underset{n\times1}{\mathbf{z}_k^*}$
$z_{ik}=\frac{1}{\sqrt{\lambda_k}}\sum_{j=1}^{n}\left(\frac{p_{ij}}{p_{i+}}\right)z_{jk}^*\left(i\in I;k=1,2,\cdots,K\right)$
列の選択肢 $j\in J$ の成分スコアの加重和が$i\in I$の成分スコア</td><td>$\underset{n\times1}{\mathbf{z}_k^*}=\frac{1}{\sqrt{\lambda_k}}\underset{n\times n}{\mathbf{P}_J^{-1}}\underset{n\times m}{\mathbf{P}_{JI}}\underset{m\times1}{\mathbf{z}_k}$
$z_{jk}^*=\frac{1}{\sqrt{\lambda_k}}\sum_{i=1}^{m}\left(\frac{p_{ij}}{p_{+j}}\right)z_{ik}\left(j\in J;k=1,2,\cdots,K\right)$
列の選択肢$i\in I$の成分スコアの加重和が $j\in J$ の成分スコア</td></tr>
<tr><td>再生公式</td><td colspan="2">$p_{ij}=p_{i+}p_{+j}\left(1+\sum_{k=1}^{K}\frac{1}{\sqrt{\lambda_k}}z_{ik}z_{jk}^*\right)\quad\left(i\in I,j\in J,K=\min\left\{m,n\right\}-1\right)$
または度数を用いると以下のように書ける．
$f_{ij}=\frac{f_{i+}f_{+j}}{N}\left(1+\sum_{k=1}^{K}\frac{1}{\sqrt{\lambda_k}}z_{ik}z_{jk}^*\right)\quad\left(i\in I,j\in J,K=\min\left\{m,n\right\}-1\right)$</td></tr>
</table>

4. 3 追加処理[57]

実際に対応分析法を利用すると，つぎのような現象が見られることがある．

i) 2 元データ表のある選択肢のプロファイルの数値が，他のプロファイルに比べて極端に小さい（あるいは偏りがある）とき，布置図の中でその選択肢に対する成分スコアの位置が，かなり外れたところに布置される．あるいは，その選択肢以外の他の選択肢の成分スコアが一部に集まり，偏った分布となる．
ii) 同時布置図において，行和や列和が少ないプロファイルが，かなり外れたところにプロットされて，全体の特徴がうまく把握できない場合が起こる．
iii) このようなとき，たとえば成分スコアの布置は図 9 のようなパターンが観測されることがある．
iv) 度数が非常に小さいセルを含む行や列があるとか[58]，行（あるいは列）の中の特定のセルだけに度数が集中しているようなときにもやはり布置図の解釈が困難となる場合がある．

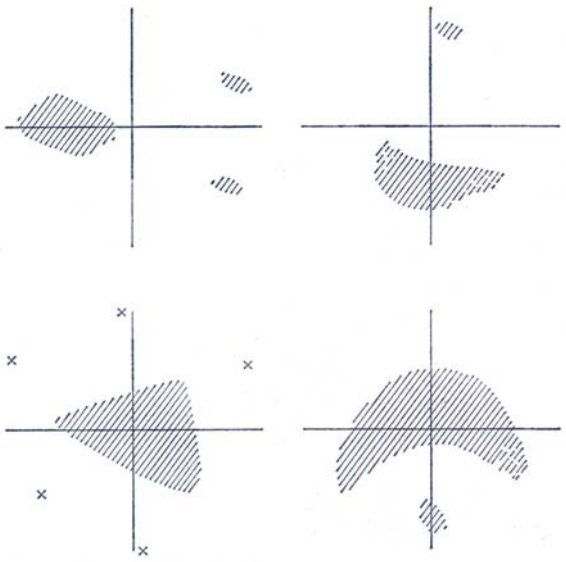


図 9 成分スコアの分布の例

このようなときに，問題となりそうな行（あるいは列）を計算からいったん除外して，残りのデータ表について成分を求め，その求めた成分式を用いて，先に除いた行（あるいは列）の成分スコアを推定するという方式が考えられている．

つまり，歪める要因である要素を一時的に除き，そのデータ表の本来の構造をはっきりと強調し引き立て，観測しようという意図がある．こうした操作を**“追加処理”**（supplementary treatment）という．この処理によりある特異なパターンがデータ表全体の本来の構造をゆがめていることを回避するという効果が期待できる．この追加処理を行うには前述の**双対性**を利用する．

［補足］

はずれ値の除外する別の方法として，“subset analysis”がある．これは，データ表の周辺和や期待値，残差などの計算時には，はずれ値を含めるが，実際に成分軸を計算する時には，はずれ値を含めないで行うという方法である．たとえば，JMP スクリプトを用いると，この

[57] 追加処理の簡単な数値例については，すでに第 I 部で示した．
[58] テキスト型データ，テキスト・マイニングで扱うような寸法が大きいデータ表ではよくある現象．

subset analysis の分析が可能である[59].

追加処理は行の側と列の側，それぞれについて行える．またその同時利用もある．ここでは，**行の追加処理**について，その手順を例で説明する．

［行の追加処理と追加要素］

追加処理の候補とする行の選択肢の集合を以下のように書く．

$$I^{+}=\{1,2,3,\cdots,i',\cdots,m'\}$$

この追加の候補とする行を**“追加の要素”**（supplementary elements）という．これに対して，実際に成分スコアを求めるために用いるもとのデータ表を**“実際データ”**（アクティブ・データ；active data）という．とくに，その行や列を**“実際変数”**（アクティブ変数；active variables）という[60]. これらの関係を模式図で表すと図 10 のようになる．このとき，I^{+} の各行（選択肢）に対するプロファイルは，

$$N_I^{+}=\left\{\frac{p_{ij}^{*}}{p_{i+}^{*}}\middle| i'\in I^{+},\ j\in J\right\},\ \text{ここで}J=\{1,2,\cdots,j,\cdots,n\} \tag{84}$$

であるから，このプロファイルを加重として，もとのデータ表から求めたつぎの成分スコアの関係式を用いて I^{+} の各行（選択肢）i' に対する成分スコアを推定する．つまり，つぎの推移公式

$$z_{ik}=\frac{1}{\sqrt{\lambda_k}}\sum_{j=1}^{n}\left(\frac{p_{ij}}{p_{i+}}\right)z_{jk}^{*}\ \left(i\in I;k=1,2,\cdots,K\right) \tag{85}$$

を用いて，このプロファイルを追加要素のそれに置き換え，

$$\varphi_{i'k}=\frac{1}{\sqrt{\lambda_k}}\sum_{j=1}^{n}\left(\frac{p_{ij}^{*}}{p_{i+}^{*}}\right)z_{jk}^{*}\ \begin{pmatrix} i'\in I^{+},k=1,2,\cdots,K \\ K=\min\{m,n\}-1 \end{pmatrix} \tag{86}$$

とする．なお，これをベクトルと行列で表すと，次式のようになる．

$$\boldsymbol{\varphi}_k=\frac{1}{\sqrt{\lambda_k}}\left(\mathbf{P}_I^{*}\right)^{-1}\mathbf{P}_{IJ}^{*}\mathbf{z}_k^{*},\ \text{ここで}\boldsymbol{\varphi}_k=\left(\varphi_{1k},\varphi_{2k},\cdots,\varphi_{i'k},\cdots,\varphi_{m'k}\right) \tag{87}$$

また，$\mathbf{P}_I^{*}=diag\left(p_{i'j}^{*}\right)$ である．

[59] subset analysis については，Greenacre (2007), *Correspondence Analysis in Practice*, second edition, Chapman &Hall/CRC を参照するとよい．（注：これの改訂版が刊行されるという）

[60] 実際データ，実際変数は，あまり適切な訳語とはいえないが，とりあえずここではこれらを用いる．

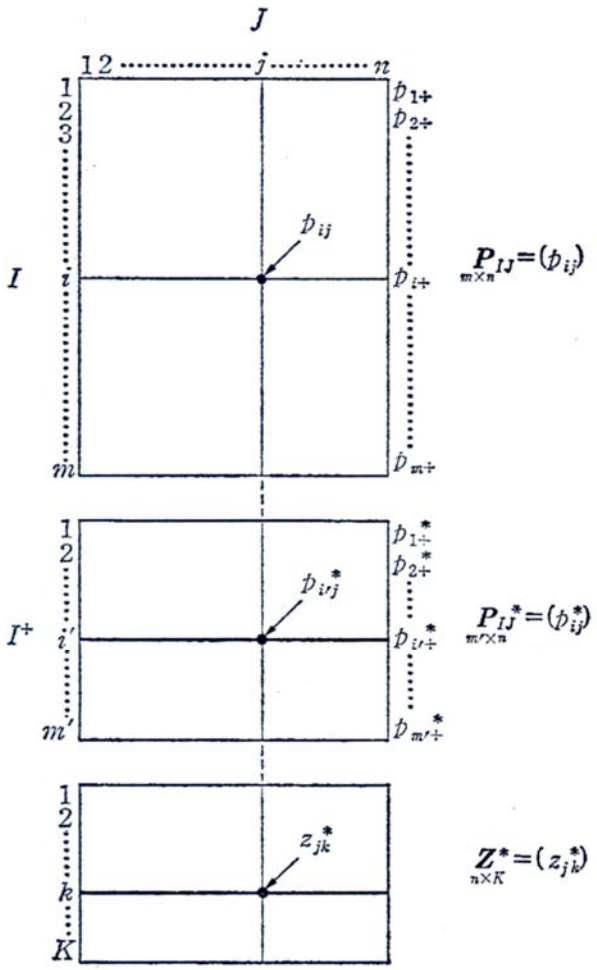


図 10　行の追加処理のイメージ

［列の追加処理］

まったく同じようにして，列の追加要素を以下のように記す．

$$J^+=\{1,2,3,\cdots,j',\cdots,n'\}$$

さらに，追加要素のプロファイルを作る．

$$\mathbf{N}_J^+=\left\{\frac{p_{ij}^{**}}{p_{+j}^{**}}\middle|\, j'\in J^+, i\in J\right\},\ \text{ここで}I=\{1,2,\cdots,i,\cdots,m\} \tag{88}$$

これを模式図で表すと図 11 となる．

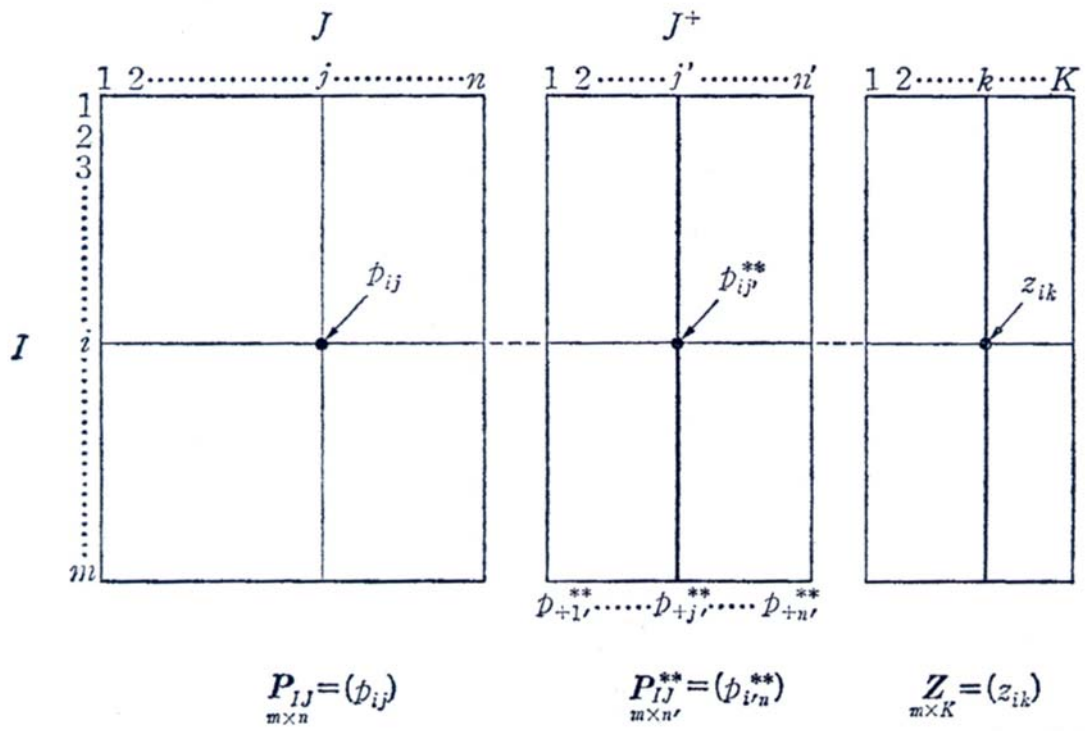


図 11　列の追加処理のイメージ

上に同じようにして，次式で表される推移公式

$$z_{jk}^{*}=\frac{1}{\sqrt{\lambda_k}}\sum_{i=1}^{m}\left(\frac{p_{ij}}{p_{+j}}\right)z_{ik}\ \left(j\in J;k=1,2,\cdots,K\right) \tag{89}$$

を用いて，以下のように成分スコア $\zeta_{j'k}$ を推定する．

$$\zeta_{j'k}=\frac{1}{\sqrt{\lambda_k}}\sum_{i=1}^{m}\left(\frac{p_{ij}^{**}}{p_{+j}^{**}}\right)z_{ik}\quad\begin{pmatrix}j'\in J^{+},k=1,2,\cdots,K\\ K=\min\{m,n\}-1\end{pmatrix} \tag{90}$$

ベクトル，行列を用いた場合は次式のようになる．

$$\boldsymbol{\zeta}_k=\frac{1}{\sqrt{\lambda_k}}\left(\mathbf{P}_J^{**}\right)^{-1}\mathbf{P}_{JI}^{**}\mathbf{z}_k \tag{91}$$

また，$\boldsymbol{\varphi}_k=\left(\varphi_{1k},\varphi_{2k},\cdots,\varphi_{i'k},\cdots,\varphi_{m'k}\right)$，$\mathbf{P}_J^{*}=diag\left(p_{ij'}^{*}\right)$

これから，成分スコアを推定する．以上で，λ_k，z_{ik}，z_{jk}^{*} は，それぞれ“**実際データ表**”（$I\times J$ の 2 元データ表）を出発行列として得た固有値，成分スコアである．

なお，この追加処理の手順は，特異なデータ，はずれ値的なデータの一時除去と再配置といった利用法のほかに，つぎのような場面で用いることが考えられる．

① 複数のグループの対比分析を行いたい．たとえば，男子と女子についてそれぞれデータ

表が与えられているとき，これらを一括して分析する場合と，男子を実際データとして，女子を追加処理の候補として（あるいはその逆もある）分析する場合とでは意味が異なるであろう．後者の場合は，男子のデータに見られる特徴や傾向を保持しながら，女子は（男子からみて）どのように位置づけられるかを検討するものであり，全体で分析した場合とは分析の意味や目的が異なる．このように，対比分析のさまざまな様相にあわせて追加処理を使い分けることが考えられる．

② したがって，複数のグループの“判別分析的な利用法”も考えられる．つまり，群 A のデータからみて，群 B のそれはどのように位置づけられるであろうか，といったことに用いる．

③ 項目側（データ表の列側）の追加処理項目の追加処理では，実際変数〈項目〉の関連性に対して，追加処理の候補となった項目がどのように類似しているのか，あるいは差異はあるのかといったことを調べることに利用できる．たとえば，調査回や調査年次の異なる質問文の比較などに適用する．

④ 追加処理の性質を利用した大量データの数量化得点の算出．多重クロス表（バート表）の関連分析法の数理的な性質を用いると，大量データの数量化得点の算出方式が考えられる（バート表で得た情報に，インジケータ行列の要素を追加する，これについては後述する）．

4. 4 簡単な数値例

ここで簡単な人工データを用いて，追加処理の手順を確かめる．用いるデータは「第 I 部」で数値例とした表 25 の寸法が 10×6 のデータ表（クロス表 $\mathbf{F}$）である．また，これから得た固有値，寄与率も挙げておこう（表 26）．

表 25 （回答者）×（銘柄）のクロス表

銘柄 / 回答者番号	銘柄A	銘柄B	銘柄C	銘柄D	銘柄E	銘柄F	行和
回答者 1	0	1	0	0	1	1	3
回答者 2	0	0	0	0	0	1	1
回答者 3	0	0	1	0	0	1	2
回答者 4	0	1	1	0	1	1	4
回答者 5	0	1	1	0	0	1	3
回答者 6	1	1	1	0	1	0	4
回答者 7	1	1	0	1	1	0	4
回答者 8	0	0	1	0	0	1	2
回答者 9	1	1	0	0	1	0	3
回答者 10	1	0	0	1	1	0	3
列和	4	6	5	2	6	6	29

表 26 固有値，寄与率の表

k	固有値 λ_k	寄与率（%）	累積寄与率（%）
1	0.6260	61.41	61.41
2	0.1877	18.41	79.82
3	0.1345	13.19	93.01
4	0.0452	4.43	97.45
5	0.0260	2.55	100.00

これを実際データとし，これにつぎの寸法が 3×6 の行列 $\mathbf{F}^{+}$ の**行要素を追加処理**することを考える．

$$I^{+}=\{1',2',3'\}$$

$$\mathbf{F}^{+}=\begin{pmatrix}1&0&0&1&0&1\\0&0&1&1&1&0\\1&1&1&1&1&1\end{pmatrix}$$

ここで，式（86）を用いて成分スコアを求めるが，式の中で，$\dfrac{p_{ij}^{*}}{p_{i+}^{*}}=\dfrac{f_{ij}^{*}}{f_{i+}^{*}}$ であることに注意すると，プロファイルを使わずとも，これは上の追加処理のクロス表 $\mathbf{F}^{+}$ から直接えられる．たとえば，表 27 にあるような計算表を作って成分スコアを推定すればよい．

表 27　行要素の追加処理の手順の例

	i' \ j	1	2	3	4	5	6	行　和
$\mathbf{F}^{+}\Rightarrow$	1'	1	0	0	1	0	1	3
	2'	0	0	1	1	1	0	3
	3'	1	1	1	1	1	1	6
$\mathbf{P}_{IJ}^{*}\Rightarrow$	1'	1/3	0	0	1/3	0	1/3	1
	2'	0	0	1/3	1/3	1/3	0	1
	3'	1/6	1/6	1/6	1/6	1/6	1/6	1
$\varphi_{1'k}$		−0.941	−0.115	0.8	−1.31	−0.513	1.03	λ_1=0.62603
$\left(\dfrac{p_{i'j}^{*}}{p_{i'+}}\right)z_{j1}^{*}$		$\dfrac{-0.941}{3}$	0	0	$\dfrac{-1.31}{3}$	0	$\dfrac{1.03}{3}$	−0.4070

たとえば，

$$\varphi_{1'k}=\frac{1}{\sqrt{\lambda_k}}\sum_{j=1}^{n}\left(\frac{p_{ij}^{*}}{p_{i+}^{*}}\right)z_{jk}^{*}=\frac{1}{\sqrt{\lambda_k}}\sum_{j=1}^{n}\left(\frac{f_{ij}^{*}}{f_{i+}^{*}}\right)z_{jk}^{*}$$

$$=\frac{1}{\sqrt{0.62603}}\times\left\{(-0.941)\times\frac{1}{3}-0,115\times0+0.8\times0-1.31\times\frac{1}{3}-0.513\times0+1.03\times\frac{1}{3}\right\}$$

$$=\frac{1}{\sqrt{0.62603}}\times(-0.40886)\doteq-0.5168$$

となる．以下同じように $\varphi_{2'k}=-0.4310,\varphi_{3'k}=-0.2210$ となる．こうして行の 3 つの追加要素の成分スコアが得られる．こうして得られたもとのデータ表の成分スコアと，追加行の成分スコアとの同時布置図を描くと図 12 のようになる．

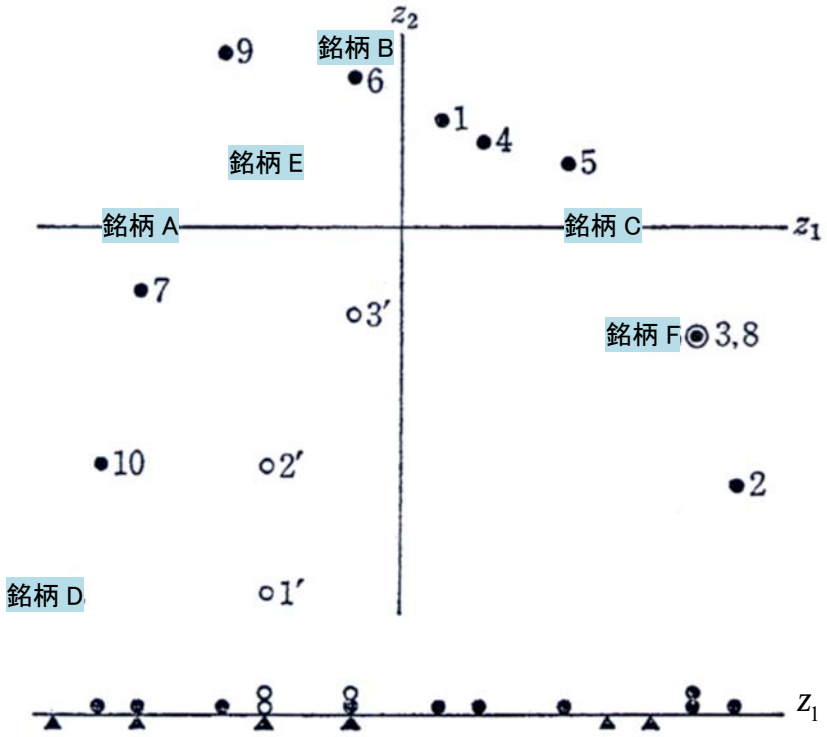


図 12　行の追加要素を加えた同時布置図

［結果の観察］

トイ・データによる簡単な数値例であるから，追加処理の効果があまり顕著には読み取れないかもしれない．実際には，はずれ値や特異なプロファイルとなる例が頻発するのであるが，データ表の寸法が大きいことなどもあり，解釈もすこし面倒となる．

いま用いた例では，たとえば行の追加要素「3'」をみよう．これは，要素がすべて「1」である．つまり，どれにも反応ありということでほぼ図の中心に位置する（平均的である）．また，「2'」がもとのサンプル 10 に近いことに注意しよう（両者を比べてみよう）．

なお，この例は，第 1 成分の固有値 $\lambda_1 = 0.62603$ の寄与率が約 61%となるので，しかも $\alpha_1 = \sqrt{\lambda_1} = \sqrt{0.62603} \doteqdot 0.7912$（行と列の第 1 成分スコア間の相関係数）となり第 1 成分スコア間の関連性も大きい（散布図とあわせて観察するとかなり線型関係にあるということ）．とくにこの成分に注目して行（回答者）と列（銘柄）の第 1 成分スコアだけを数直線上に布置すると図 12 の下のほうにある数直線上の分布となる（記号▲）．ここでとくに銘柄が左下から「銘柄 D」→「銘柄 A」→「銘柄 E」→「銘柄 B」→「銘柄 C」→「銘柄 F」と並んでいることに注意しよう．これに対してそれぞれのサンプル（回答者）がどう応答したかが読み取れる．つまり回答者の銘柄の選び方にはある傾向があることがわかる．ある質問文で銘柄という名義尺度にこの順の大きさに数量が付与されたということをも意味する．この例のように放物線上に点が並ぶことを**馬蹄形効果**[61]（horseshoe effect）という．

5.　多重クロス表と多重対応分析

多重クロス表（バート表）については「第 I 部」では，例を挙げて説明した．この多重クロス表を用いた対応分析法を**多重対応分析**（MCA : multiple correspondence analysis）という[62].

5. 1　2 元クロス表から得た多重クロス表

[61] こうした構造的な特徴を詳しく知るには岩坪秀一（1987）が詳しい.

[62] ACM：analyse des correspondances multiples（仏）

通常の調査では複数の質問項目についてのデータ表が得られる．前にみた例（レストラン・データ）をここでも引用し，2 元クロス表とこれを含む多重クロス表をふたたび確認しよう．もとの調査データ表から（表 28），**アイテム･カテゴリー型行列**にコード変換する．これが表 29 であり，**インジケータ行列**あるいは**完備排反型行列**[63]ともいう（以下では，インジケータ行列と呼ぶ）．これを行列 $\mathbf{A}$ で表そう．これを転置し $\mathbf{A}^t$ とし，行列 $\mathbf{A}$ との積の行列 $\mathbf{B} = \mathbf{A}^t\mathbf{A}$ を作ると表 30 が得られる．これが**多重クロス表**であり，**バート表**[64]（あるいはバート行列）である．これはあきらかに対称行列であり，また，対角部には 2 つの項目の周辺度数を対角要素とする対角行列が入る形になる．これをより一般的に示すと，表 30 の形式になる．

表 28 （回答者）×（項目）のデータ表（調査データのイメージ）

項目 / 回答者	I （レストラン）	J （評価基準）
1	バッハ	味
2	ムガール	量
3	さとみ	量
4	ラ・マレ	工夫・サービス
5	きくみ	味
：	：	：
：	：	：
N	いりふね	量

ここで，N=1, 284（回答者数）

表 29 インジケータ行列: $\mathbf{A}$ 表

回答者 \ 項目	I									
	1	2	3	4	…	9	10	1	2	3
	いりふね	かりや	きくみ	さとみ	クラーク	コルシカ	バッハ	工夫	サービス	味	量
1	0	0	0	0	…	0	1	0	1	0
2	0	0	0	0	…	0	0	0	0	1
3	0	0	0	1	…	0	0	0	0	1
4	0	0	0	0	…	0	0	0	1	0
5	0	0	1	0	…	0	0	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	…	⋮	⋮	⋮	⋮	⋮
	⋮	⋮	⋮	⋮	…	⋮	⋮	⋮	⋮	⋮
1, 284	1	0	0	0	…	0	0	0	0	1

[63] complete disjunctive form（forme disjunctive complete）を**完備排反型**と訳した．対応分析では，**インジケータ行列**と呼ぶことが多い．

[64] この表形式を考えた応用心理学者，C. Burt（シリル・バート；Cyril Burt）の名をとってこう呼ぶ．

表 30　多重クロス表（バート表）$\mathbf{B}=\mathbf{A}^t\mathbf{A}$ の生成

項目	いりふね	かりや	きくみ	さとみ	クラーク	コルシカ	バッハ	ムガール	ラ･マレ	ロゴスキー	工夫･サービス	味	量
いりふね	155										98	25	32
かりや		176									105	35	38
きくみ			110								35	8	67
さとみ				95							42	46	7
クラーク					102						34	14	54
コルシカ						122					32	77	13
バッハ							142				48	76	18
ムガール								109			49	44	16
ラ･マレ									146		49	82	15
ロゴスキー										125	48	35	42
工夫･サービス	98	105	35	42	34	32	48	49	49	48	540		
味	25	35	8	46	14	77	76	44	82	35		442	
量	32	38	67	7	54	13	18	16	15	42			302

(*) この表で空白のセル（対角ブロック行列の非対角要素）はすべてゼロである．

表 31　上のバート表の説明

	質問 I	質問 J
質問 I	（質問 I ）×（質問 I ）のクロス表 つまり質問 J の周辺度数が対角要素に入った対角行列	（質問 I ）×（質問 J ）のクロス表
質問 J	（質問 J ）×（質問 I ）のクロス表	（質問 J ）×（質問 J ）のクロス表 つまり質問 J の周辺度数が対角要素に入った対角行列

5. 2　複数の質問項目から得た多重クロス表

ではここで，複数の質問項目を扱う場合を考えよう．いま，表 32 にあるような M 個の項目からなるデータ表とする[65]（これを **C** 表と呼ぼう）．ここでは ${}_M C_2 = M\left(M+1\right)/2$ 通りの 2 元クロス表があるから，これらの 2 元クロス表を 2 元データ表形式に並置すると**多重クロス表**となる．多重クロス表は多元（multi-way）とは異なり 2 元データ表である．

このとき，**インジケータ行列**は，2 項目の場合と同様，表 33 のように表せる．このデータ表全体を行列 **A** で表すと，以下のようになる．

表 32　一般的な調査データ: **C** 表（コーディング行列：大きさが $N\times n^*$ ）

項目 / 回答者	Q_1	Q_2	…	Q_M
1	2	3	…	1
2	3	2	…	2
3	3	1	⋮	3
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
i	3	3	…	1
⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮
N	1	2	…	3
最大カテゴリー数	n_1	n_2	…	n_M

[65] たとえば，表 28 をコーディングによりこのような形式に変換したことを考えればよい．

表 33　インジケータ行列: **A** 表

項目 / 回答者	Q_1	Q_2	…	Q_M	行和
1	0 1 0 0 ⋯0	0 0 1 0 ⋯0	…	1 0 0 0 ⋯0	M
2	⋮	⋮	…	⋮	M
3	⋮	⋮	⋮	⋮	M
⋮	⋮	⋮	⋮	⋮	M
⋮	⋮	⋮	⋮	⋮	M
i	0 0 1 0 ⋯0	0 0 1 0 ⋯0	…	1 0 0 0 ⋯0	M
⋮	⋮	⋮	⋮	⋮	M
⋮	⋮	⋮	⋮	⋮	M
N	1 0 0 0⋯0	0 1 0 0 ⋯0	…	0 0 1 0 ⋯0	M
最大カテゴリー数	n_1	n_2	…	n_M	NM（総和）
総カテゴリー数	$n^* = n_1 + n_2 + \cdots + n_M = \sum_{j=1}^{M} n_j$				

このインジケータ行列を下のように表す.

$$\underset{N\times n^*}{\mathbf{A}} = \left[\underset{N\times n_1}{\mathbf{A}_1}, \underset{N\times n_2}{\mathbf{A}_2}, \cdots, \underset{N\times n_i}{\mathbf{A}_i}, \cdots, \underset{N\times n_j}{\mathbf{A}_j}, \cdots, \underset{N\times n_M}{\mathbf{A}_M} \right] \quad \left(\text{ここで, } n^* = \sum_{j=1}^{M} n_j \right) \tag{92}$$

ここで，転置行列を $\mathbf{A}^t$ とし，行列 $\mathbf{A}$ との積の行列 $\mathbf{B} = \mathbf{A}^t\mathbf{A}$ を作ると，2 項目の場合を一般化したつぎのバート行列が得られる.

$$\underset{n^*\times n^*}{\mathbf{B}} = \underset{n^*\times N}{\mathbf{A}^t}\ \underset{N\times n^*}{\mathbf{A}} = \begin{pmatrix} \mathbf{A}_1^t\mathbf{A}_1 & \mathbf{A}_1^t\mathbf{A}_2 & \cdots & \mathbf{A}_1^t\mathbf{A}_i & \cdots & \mathbf{A}_1^t\mathbf{A}_j & \cdots & \mathbf{A}_1^t\mathbf{A}_M \\ \mathbf{A}_2^t\mathbf{A}_1 & \mathbf{A}_2^t\mathbf{A}_2 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{A}_i^t\mathbf{A}_1 & \cdots & \cdots & \mathbf{A}_i^t\mathbf{A}_i & \cdots & \mathbf{A}_i^t\mathbf{A}_j & \cdots & \mathbf{A}_i^t\mathbf{A}_M \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{A}_j^t\mathbf{A}_1 & \cdots & \cdots & \mathbf{A}_j^t\mathbf{A}_i & \cdots & \mathbf{A}_j^t\mathbf{A}_j & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_M^t\mathbf{A}_1 & \cdots & \cdots & \mathbf{A}_M^t\mathbf{A}_1 & \cdots & \cdots & \cdots & \mathbf{A}_M^t\mathbf{A}_M \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{D}_1 & \mathbf{A}_1^t\mathbf{A}_2 & \cdots & \mathbf{A}_1^t\mathbf{A}_i & \cdots & \mathbf{A}_1^t\mathbf{A}_j & \cdots & \mathbf{A}_1^t\mathbf{A}_M \\ \mathbf{A}_2^t\mathbf{A}_1 & \mathbf{D}_2 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mathbf{A}_i^t\mathbf{A}_1 & \cdots & \cdots & \mathbf{D}_i & \cdots & \mathbf{A}_i^t\mathbf{A}_j & \cdots & \mathbf{A}_i^t\mathbf{A}_M \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{A}_j^t\mathbf{A}_1 & \cdots & \cdots & \mathbf{A}_j^t\mathbf{A}_i & \cdots & \mathbf{D}_j & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_M^t\mathbf{A}_1 & \cdots & \cdots & \mathbf{A}_M^t\mathbf{A}_1 & \cdots & \cdots & \cdots & \mathbf{D}_M \end{pmatrix} \tag{93}$$

林の数量化法では，通常はインジケータ行列つまり**アイテム・カテゴリー型データ表**から出発する．よって，回答者数（サンプル数）や質問項目が多いと，データ表の寸法が大きく

なる．一方，多重対応分析では，扱う多重クロス表（バート表）の寸法は，質問項目数（M）と総選択肢数の大きさ（n^*）で済む．後述するように，インジケータ行列とバート表との対応分析法の相互の数理的な関係が考察されており，これらを使い分けて，回答者の成分スコアも追加処理で得られるという利点がある（後述）．

バート表は対称行列であり，その各ブロック $\mathbf{A}_i^t\mathbf{A}_j$ は対応する 2 つの項目の 2 元クロス表となっている．つまり，非対角部分に置かれた $\mathbf{A}_i^t\mathbf{A}_j$ は，項目 Q_i と項目 Q_j（あるいは，前の記法によれば項目 I と項目 J）との 2 元クロス表を表す．さらにここで，対角部分に置かれた行列，

$$\underset{n_q\times n_q}{\mathbf{D}_q} = \mathbf{A}_q{}^t\mathbf{A}_q \quad (q=1,2,\cdots,M) \tag{94}$$

は，寸法が $n_q\times n_q$ の対角行列であり，その対角要素は各項目 Q_q の周辺度数に相当する．さらに，式（94）で表される対角行列を対角ブロックに置いた対角行列，

$$\underset{n^*\times n^*}{\mathbf{D}} = \begin{pmatrix} \mathbf{D}_1 & \mathbf{O} & \cdots & \mathbf{O} & \cdots & \mathbf{O} \\ \mathbf{O} & \mathbf{D}_2 & & & & \mathbf{O} \\ \vdots & & \ddots & & & \vdots \\ \mathbf{O} & & & \mathbf{D}_q & & \mathbf{O} \\ \vdots & & & & \ddots & \vdots \\ \underbrace{\mathbf{O}}_{n_1} & \underbrace{\mathbf{O}}_{n_2} & \cdots & \underbrace{\mathbf{O}}_{n_q} & \cdots & \underbrace{\mathbf{D}_M}_{n_M} \end{pmatrix} \tag{95}$$

を用意しておこう．これはつぎにインジケータ行列のデータ表 $\mathbf{A}$ およびバート表 $\mathbf{B}$ それぞれの対応分析と，それら相互の関係を調べるときに用いる．

以上にみるように，多重クロス表は，基本的には 2 元クロス表の一般化である[66]．ここでインジケータ行列のデータ表の場合と，バート表の場合の対応分析法について述べる．

5. 3 アイテム・カテゴリー型データ表の場合

すでに述べたように，これは $\underset{N\times n^*}{\mathbf{A}}$ で表され，行和はいずれも M である．ここで，このデータ表に対する対応分析を考えるために，M を対角要素とする以下の行列を作る．なおここで，$\mathbf{I}_N$ とは対角成分が 1，非対角成分が 0 の N 次正方行列（単位行列）を示す．

$$M\times\mathbf{I}_N = \begin{bmatrix} M & 0 & \cdots & 0 \\ 0 & M & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & M \end{bmatrix} \tag{96}$$

この行列の対角和は $tr\left(M\times\mathbf{I}_N\right)=NM$ となり，これはデータ表 $\mathbf{A}$ の総度数に等しい（表 31）．ここでまた，行列 $\underset{n^*\times n^*}{\mathbf{D}}$ の対角和は，$tr\left(\mathbf{D}_q\right)=N\left(q=1,2,\cdots,M\right)$ であるから，

[66] 繰り返すが多元クロス表（multiway tables）とは異なることに注意．

$tr\left(\underset{n^*\times n^*}{\mathbf{D}}\right)=\sum_{q=1}^{M} tr\left(\mathbf{D}_q\right)=NM$ となる．

ここで，データ表 $\mathbf{A}$ をクロス表 $\mathbf{F}$ と読み替えると，確率行列 $\mathbf{P}_{IJ}$ に相当する行列は，

$$\mathbf{P}_{IJ}=\frac{1}{NM}\mathbf{A} \tag{97}$$

となる．また， $\underset{n^*\times n^*}{\mathbf{D}}$ からそれぞれ $\mathbf{P}_I,\mathbf{P}_J$ に相当する対角行列を作ると，

$$\mathbf{P}_I=\frac{1}{NM}\times M\mathbf{I}_N=\frac{1}{N}\mathbf{I}_N \tag{98}$$

$$\mathbf{P}_J=\frac{1}{NM}\underset{n^*\times n^*}{\mathbf{D}} \tag{99}$$

が得られる．

以上から，行列 $\mathbf{A}$ の対応分析は，上に用意した，$\mathbf{P}_{IJ},\mathbf{P}_I,\mathbf{P}_J$ を使って行列 $\mathbf{P}_J^{-1/2}\mathbf{P}_{JI}\mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2}$ を作りこれの固有値問題を解くことに帰着する[67]．実際にこれを求めるとつぎのようになる．

$$\mathbf{V}=\mathbf{P}_J^{-1/2}\mathbf{P}_{JI}\mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2}=\left[\sqrt{NM}\,\underset{n^*\times n^*}{\mathbf{D}^{-1/2}}\right]\times\left[\frac{1}{NM}\mathbf{A}\right]^t\times\left[N\mathbf{I}_N^{-1}\right]\times\left[\frac{1}{NM}\mathbf{A}\right]\times\left[\sqrt{NM}\,\underset{n^*\times n^*}{\mathbf{D}^{-1/2}}\right]$$

$$=\frac{1}{M}\mathbf{D}^{-1/2}\underbrace{\mathbf{A}^t\mathbf{A}}_{\mathbf{B}}\mathbf{D}^{-1/2}=\frac{1}{M}\mathbf{D}^{-1/2}\mathbf{B}\mathbf{D}^{-1/2}$$

つまり固有方程式，$\mathbf{V}\mathbf{l}=\lambda^A\mathbf{l}$

$$\left(\frac{1}{M}\mathbf{D}^{-1/2}\mathbf{B}\mathbf{D}^{-1/2}\right)\mathbf{l}=\lambda^A\mathbf{l} \tag{100}$$

を解けばよい（固有値 λ^A および固有ベクトル $\mathbf{l}$ を求める）．これがインジケータ行列（かつアイテム･カテゴリー型行列）の対応分析となる[68]．

5. 4 バート表の場合

バート表 $\mathbf{B}$ の対応分析についても，上と同様の考え方で導くことができる．バート表 $\underset{n^*\times n^*}{\mathbf{B}}=\underset{n^*\times N}{\mathbf{A}^t}\underset{N\times n^*}{\mathbf{A}}$ はすでに述べたように**対称行列**であり，また非対角部分に置かれているブロック行列の 2 元クロス表 $\mathbf{A}_q{}^t\mathbf{A}_{q'}\ \ (q\neq q';q,q'=1,2,\cdots,M)$ は行和 1，総度数 N であることに注意すると，$tr\left(\mathbf{D}_q\right)=N\left(q=1,2,\cdots,M\right)$ から $\mathbf{B}$ 表の行和は NM，総度数は $NM\times M=NM^2$ となる．したがって，$\mathbf{P}_{IJ}$ に相当の行列は，

---

[67]前に示した､対応分析の基本行列となる式（35）を思い出そう．分散共分散行列：$\mathbf{V}^*=\mathbf{P}_J^{-1/2}\mathbf{P}_{JI}\mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2}$．

[68]ここで，対応分析（CA）の定式化にならって、平均の周りでセンタリング（中心化）を行うことがある．とくに subset CA（サブセット分析）ではこれが必要となる．これについては，たとえば Greenacre and Blasius (eds.)（2006）Greenacre（1993）を参照．

$$\mathbf{P}_{IJ} = \frac{1}{NM^2}\mathbf{B} \tag{101}$$

となる．また，$tr\left(\underset{n^*\times n^*}{\mathbf{D}}\right) = \sum_{q=1}^{M} tr\left(\mathbf{D}_q\right) = NM$ から $\mathbf{P}_I, \mathbf{P}_J$ は等しく，

$$\mathbf{P}_I = \mathbf{P}_J = \frac{1}{NM}\mathbf{D} \tag{102}$$

とおくことができる．よってここでも，これから行列 $\mathbf{P}_J^{-1/2}\mathbf{P}_{JI}\mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2}$ を作ると

$$\begin{aligned}\mathbf{V} &= \mathbf{P}_J^{-1/2}\mathbf{P}_{JI}\mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2} \\ &= \left[\sqrt{NM}\,\mathbf{D}^{-1/2}\right]\times\left[\frac{1}{NM^2}\mathbf{B}\right]^t\times\left[NM\mathbf{D}^{-1}\right]\times\left[\frac{1}{NM^2}\mathbf{B}\right]\times\left[\sqrt{NM}\,\mathbf{D}^{-1/2}\right] \\ &= \left[\frac{1}{M}\mathbf{D}^{-1/2}\mathbf{B}\mathbf{D}^{-1/2}\right]^t\left[\frac{1}{M}\mathbf{D}^{-1/2}\mathbf{B}\mathbf{D}^{-1/2}\right]\end{aligned} \tag{103}$$

となるので，ここでもまた固有方程式

$$\mathbf{V}\mathbf{l} = \lambda^B\mathbf{l} \tag{104}$$

$$\left[\frac{1}{M^2}\left(\mathbf{D}^{-1/2}\mathbf{B}\mathbf{D}^{-1/2}\right)^t\left(\mathbf{D}^{-1/2}\mathbf{B}\mathbf{D}^{-1/2}\right)\right]\mathbf{l} = \lambda^B\mathbf{l}$$

を解けばよい（固有値 λ^B および固有ベクトル $\mathbf{l}$ を求める）．

ここで 2 つの式（100），（104）を比べると，両者の固有値に関して，

$$\lambda^B = (\lambda^A)^2 \tag{105}$$

となり，**B** 表の固有値は **A** 表のそれのちょうど 2 乗になることがわかる．つまり，両者の解は，1 対 1 に対応がとれる解となっている．こうして，2 つの 2 元データ表，**A** 表と **B** 表の対応分析法の関連が示された．

つぎにこれらの関係をもとに，成分スコアの算出式や双対性の性質について述べるが，ここでそのために必要な式を再度整理しておこう．

［成分スコアを求める式］

$$\mathbf{z}_k = \mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2}\mathbf{l}_k \quad （行成分スコア） \tag{105}$$

$$\mathbf{z}_k^* = \mathbf{P}_J^{-1}\mathbf{P}_{JI}\mathbf{P}_I^{-1/2}\mathbf{l}_k \quad （列成分スコア） \tag{106}$$

［双対性の関係式］

$$\mathbf{z}_k = \frac{1}{\sqrt{\lambda_k}}\mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{z}_k^* \tag{107}$$

$$\mathbf{z}_k^* = \frac{1}{\sqrt{\lambda_k}}\mathbf{P}_J^{-1}\mathbf{P}_{JI}\mathbf{z}_k \tag{108}$$

上記の各式を利用して **A** 表，**B** 表それぞれの諸量を求める．

[A 表の成分スコア]

ここでは，**A** 表における成分スコアを算出する．

まず，行側の成分スコアを考えよう．これは，もとのデータ表（表 30）の行側つまり個体の成分スコアに相当する．これの誘導は，式 (107) を用いる．これに式 (97)，(98)，(99) で示される $\mathbf{P}_{IJ},\mathbf{P}_I,\mathbf{P}_J$ を，それぞれ代入し，また行側の第 k 成分スコアのベクトルを $\mathbf{z}_k^A$ と書くと，以下となる．

$$\begin{aligned}\mathbf{z}_k^A &= \mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2}\mathbf{l}_k \\ &= \left[\frac{1}{N}\mathbf{I}_N\right]^{-1}\times\left[\frac{1}{NM}\mathbf{A}\right]\times\left[\frac{1}{NM}\mathbf{D}\right]^{-1/2}\mathbf{l}_k \\ &= N\times\frac{1}{NM}\mathbf{A}\times\sqrt{NM}\,\mathbf{D}^{-1/2}\mathbf{l}_k \\ &= \sqrt{\frac{N}{M}}\mathbf{A}\mathbf{D}^{-1/2}\mathbf{l}_k\end{aligned} \tag{113}$$

すなわち，

$$\mathbf{z}_k^A = \sqrt{\frac{N}{M}}\mathbf{A}\mathbf{D}^{-1/2}\mathbf{l}_k \tag{114}$$

が得られる．

つぎに列側の成分スコアを考えよう．これは **A** 表をアイテム・カテゴリーに展開して得られる各質問への成分スコアに相当する．式 (108) に，式 (101) (102) で与えられる $\mathbf{P}_{IJ},\mathbf{P}_I,\mathbf{P}_J$ を，それぞれ代入する．ここで列側の第 k 成分スコアのベクトルを $(\mathbf{z}_k^*)^A$ と書くと，

$$\begin{aligned}(\mathbf{z}_k^*)^A &= \mathbf{P}_J^{-1}\mathbf{P}_{JI}\mathbf{P}_I^{-1/2}\mathbf{l}_k \\ &= \left[\frac{1}{NM}\mathbf{D}\right]^{-1}\times\left[\frac{1}{NM}\mathbf{A}\right]^t\times\left[\frac{1}{N}\mathbf{I}_N\right]^{-1/2}\mathbf{l}_k \\ &= NM\mathbf{D}^{-1}\times\left[\frac{1}{NM}\mathbf{A}\right]^t\times\sqrt{N}\mathbf{l}_k \\ &= \sqrt{N}\mathbf{D}^{-1}\mathbf{A}^t\mathbf{l}_k\end{aligned} \tag{115}$$

となる．よって，

$$(\mathbf{z}_k^*)^A = \sqrt{N}\mathbf{D}^{-1}\mathbf{A}^t\mathbf{l}_k \tag{116}$$

が得られる．

[B 表の成分スコア]

B 表は対称行列であるので，また行側・列側とも A 表の列側に対応することに注意すれば，第 k 成分に対する成分スコアベクトル $\mathbf{z}_k^B$ は，以下となる．

$$
\begin{aligned}
\mathbf{z}_k^B &= \mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{-1/2}\mathbf{l}_k \\
&= \left[\frac{1}{NM}\mathbf{D}\right]^{-1} \times \left[\frac{1}{NM^2}\mathbf{B}\right] \times \left[\frac{1}{NM}\mathbf{D}\right]^{-1/2}\mathbf{l}_k \\
&= NM\mathbf{D}^{-1} \times \frac{1}{NM^2}\mathbf{B} \times \sqrt{NM}\,\mathbf{D}^{-1/2}\mathbf{l}_k \\
&= \sqrt{\frac{N}{M}}\mathbf{D}^{-1}\mathbf{B}\mathbf{D}^{-1/2}\mathbf{l}_k
\end{aligned}
$$

すなわち，

$$
\mathbf{z}_k^B = \sqrt{\frac{N}{M}}\mathbf{D}^{-1}\mathbf{B}\mathbf{D}^{-1/2}\mathbf{l}_k \tag{117}
$$

が得られる．

5.5 双対性の関係から得られる性質

A 表では，行と列の成分スコアの $\mathbf{z}_k^A$ と $(\mathbf{z}_k^*)^A$ の間に

$$
\mathbf{z}_k^A = \frac{1}{\sqrt{\lambda_k^A}}\left(\frac{1}{M}\mathbf{A}\right)(\mathbf{z}_k^*)^A \tag{118}
$$

$$
(\mathbf{z}_k^*)^A = \frac{1}{\sqrt{\lambda_k^A}}\left(\frac{1}{M}\mathbf{D}^{-1}\mathbf{B}^t\right)\mathbf{z}_k^A
$$

という双対関係が成り立つ．一方 **B** 表では，これが対称行列（$\mathbf{B}^t = \mathbf{B}$）であることから，

$$
\mathbf{z}_k^B = \frac{1}{\sqrt{\lambda_k^B}}\left(\frac{1}{M}\mathbf{D}^{-1}\mathbf{B}\right)(\mathbf{z}_k^*)^B \tag{119}
$$

$$
(\mathbf{z}_k^*)^B = \frac{1}{\sqrt{\lambda_k^B}}\left(\frac{1}{M}\mathbf{D}^{-1}\mathbf{B}^t\right)\mathbf{z}_k^B \tag{120}
$$

となるので両式は等しい．これらより，以下の重要な性質が得られる．

性質 1

A 表の列側（質問項目の側）の成分スコアと，**B** 表の成分スコアとの間に，以下の関係がなり立つ．

$$
(\mathbf{z}_k^*)^B = \sqrt{\lambda_k^A}(\mathbf{z}_k^*)^A \tag{121}
$$

これは以下のようにして示すことができる．

まず，式（105）より $\lambda^B = (\lambda^A)^2$ であったから，$\lambda^A = \sqrt{\lambda^B}$ となる．さらに $(\mathbf{z}_k^*)^A$，$(\mathbf{z}_k^*)^B$ を，それぞれ対応する固有値 λ_k^A，λ_k^B の正の平方根（標準偏差）で割って標準化スコアに変換すると，その両者は等しいから，

$$
\frac{1}{\sqrt{\lambda_k^A}}(\mathbf{z}_k^*)^A = \frac{1}{\sqrt{\lambda_k^B}}(\mathbf{z}_k^*)^B \tag{122}
$$

となる．これに $\lambda^A = \sqrt{\lambda^B}$ を代入して整理すると上の式（121）が得られる．

性質 2

A 表の行側，すなわち個体（サンプル側）の成分スコアについては，

$$\mathbf{z}_k^A = \frac{1}{\sqrt{\lambda_k^B}}\left(\frac{1}{M}\mathbf{A}\right)\mathbf{z}_k^B \qquad (123)$$

という関係が成り立つ．これは以下のようにして示すことができる．

まず，式（121）より $(\mathbf{z}_k^*)^A = (\mathbf{z}_k^*)^B \Big/ \sqrt{\lambda_k^A}$ であるから，これを式（123）の左辺に代入すると，

$$\begin{aligned}
\mathbf{z}_k^A &= \frac{1}{\sqrt{\lambda_k^A}}\left(\frac{1}{M}\mathbf{A}\right)\times\frac{1}{\sqrt{\lambda_k^A}}(\mathbf{z}_k^*)^B \\
&= \frac{1}{\lambda_k^A}\left(\frac{1}{M}A\right)(\mathbf{z}_k^*)^B \\
&= \frac{1}{\sqrt{\lambda_k^B}}\left(\frac{1}{M}\mathbf{A}\right)(\mathbf{z}_k^*)^B \qquad (\text{ここで，} \lambda_k^A = \sqrt{\lambda_k^B}) \\
&= \frac{1}{\sqrt{\lambda_k^B}}\left(\frac{1}{M}\mathbf{A}\right)\mathbf{z}_k^B \qquad (\text{ここで，} (\mathbf{z}_k^*)^B = \mathbf{z}_k^B)
\end{aligned} \qquad (124)$$

となり式（123）が示される．これは式の右辺における $\mathbf{D}^{-1}\mathbf{B}^t$ を **A** で置き換えたもの，すなわち **B** 表から求めた式の $\mathbf{z}_k^B$ の成分式（合成変数）に **A** 表を**追加処理**したことに等しい．

これらの性質は，質問項目数に比べて個体数が非常に大きなデータ表の分析にきわめて有効である．たとえば，以下のような計算手順を考えればよい．

i) まずもとのデータ表（**C** 表）から **A** 表を作りこれをファイルに格納する．

ii) つぎに **B** 表を作り，これの対応分析から成分スコアを求め，項目の成分スコアとその成分式を算出する．

iii) この成分式を用いて i)で格納した **A** 表の個体のデータの成分スコアを**“追加処理”**により求め，必要に応じてこれもファイルに格納する．

こうして，大規模データの成分スコアの算出を，さほど大きな主記憶を用いずに実行することが可能になる．最近では主記憶容量が大型化し演算速度も高速化した計算機が増えているとはいえ，これらの性質を理解することで，計算機への負荷をかけることなく，大規模データを高速・簡便に処理できる．

以上で得られた関係を表 32 に要約した．要約表の意味することは，いずれも**“2 元データ表”**である，という点にあるこれら 2 元データ表，つまり 2 元クロス表，インジケータ行列，多重クロス表を使った対応分析の結果には，相互にある関係があること，とくにインジケータ行列と多重クロス表との結果は実は同じ内容となっていることを示している．こうしたデータ表間の関係を知って分析を進めることは重要である[69]．

[69] たとえば，WordMiner では，構成要素変数（語句の変数），質的変数をうまく組み合わせることでインジケータ行列とクロス表データに対応できる．

表 34　2 元データ表の相互の基本的な関係

タイプ	データ表の形	データ表の次元数 （寸法）	固有値の関係
タイプ 1	2 項目 $I\times J$ または $J\times I$ のクロス表 $\underset{n_i\times n_j}{\mathbf{F}}=\mathbf{A}_i^t\mathbf{A}_j$ または $\underset{n_j\times n_i}{\mathbf{F}^t}=\mathbf{A}_j^t\mathbf{A}_i$	$n_i\times n_j$ （*）前に用いたクロス表の寸法を表す記号によると，$m=n_i, n=n_j$ と対応 （*）固有値の個数は， $K=\min\{n_i,n_j\}-1$	λ_k^F
タイプ 2	2 項目 I と J のアイテム・カテゴリー型データ表 $\underset{N\times n^*}{\mathbf{A}}=\left[\underset{N\times n_i}{\mathbf{A}_i},\underset{N\times n_j}{\mathbf{A}_j}\right]$ ここで $\left(n^*=n_i+n_j\right)$	$N\times n^*\ \left(n^*=n_i+n_j\right)$	$\lambda_k^A=\dfrac{1\pm\sqrt{\lambda_k^F}}{2}$
	タイプ 1 とタイプ 2 の固有値の関係：　ここで，$n_i\geq n_j$ とする． ⅰ）値の大きい方から n_j-1 個　⇒　$\lambda_k^A=\dfrac{1+\sqrt{\lambda_k^F}}{2}$ ⅱ）値の小さい方から n_j-1 個　⇒　$\lambda_k^A=\dfrac{1-\sqrt{\lambda_k^F}}{2}$ ⅲ）間に含まれる n_i-n_j 個　⇒　$1/2=0.5$ となる．（*）$n_i=n_j$ のときにはこれは現れない．		
タイプ 3 （2 項目の場合）	2 項目の多重クロス表（バート表） $\underset{n^*\times n^*}{\mathbf{B}}=\underset{n^*\times N}{\mathbf{A}^t}\ \underset{N\times n^*}{\mathbf{A}}\quad\left(n^*=n_i+n_j\right)$	$n^*\times n^*\ \left(n^*=n_i+n_j\right)$	$\lambda_k^B=\left(\lambda_k^A\right)^2=\left(\dfrac{1\pm\sqrt{\lambda_k^F}}{2}\right)^2$
タイプ 4	一般の M 項目のアイテム・カテゴリー型データ表 $\underset{N\times n^*}{\mathbf{A}}=\left[\underset{N\times n_1}{\mathbf{A}_1},\underset{N\times n_2}{\mathbf{A}_2},\cdots,\underset{N\times n_i}{\mathbf{A}_i},\cdots,\underset{N\times n_j}{\mathbf{A}_j},\cdots,\underset{N\times n_M}{\mathbf{A}_M}\right]$ ここで $n^*=\sum_{j=1}^{M}n_j$	$N\times n^*\ \left(n^*=\sum_{j=1}^{M}n_j\right)$ （*）固有値の個数は， $K^*=\sum_{j=1}^{M}\left(n_j-1\right)=n^*-M$	λ_k^A $\lambda_k^A=\sqrt{\lambda_k^B}$
タイプ 5	M 項目の多重クロス表（バート表） $\underset{n^*\times n^*}{\mathbf{B}}=\underset{n^*\times N}{\mathbf{A}^t}\ \underset{N\times n^*}{\mathbf{A}}$ ここで $n^*=\sum_{j=1}^{M}n_j$	$n^*\times n^*\ \left(n^*=\sum_{j=1}^{M}n_j\right)$ （*）固有値の個数は， $K^*=\sum_{j=1}^{M}\left(n_j-1\right)=n^*-M$	λ_k^B $\lambda_k^B=\left(\lambda_k^A\right)^2$

5. 6 多重対応分析における固有値に関する重要な性質

5. 6. 1 固有値と寄与率

インジケータ行列の対応分析で得られる固有値（従って多重クロス表またはバート表の固有値）については，つぎの重要な性質がある．一般にこの種のデータ表の対応分析で得られる固有値（とその寄与率）は，値が小さくあたかも寄与が低いように見えるが，それはデータ表の構造的な制約から生じるものであることを示している（これについての詳細は Greenacre（1984），大隅他（1994）を参照）．また，これを改善するためのいくつかの提案，たとえば**“調整済み寄与率”**（adjusted contribution）などもある．この固有値と寄与率の関係について，簡単に述べる．

まずここで，今までに得た情報から，インジケータ行列から出発した場合とバート表から出発した場合の対応分析で得られる固有値と総変動（全慣性）の関係を整理する．ここで用いる記法，記号は，前に用意したそれに従う．

① 固有値の総和（その1）

インジケータ行列（ $\underset{N\times n^*}{\mathbf{A}}$ ）の対応分析で得られる総変動（全慣性）つまり固有値の総和は，得られる固有値を $\lambda_k^A\left(k=1,2,\cdots,K^*;K^*=n^*-M\right)$ とすると，以下のように表される．

$$inertia\left(\underset{N\times n^*}{\mathbf{A}}\right)=\frac{n^*}{M}-1=\frac{n^*-M}{M}\quad（ここで n^*=\sum_{j=1}^{M}n_j）\qquad(125)$$

これは，次のように書き替えられる．

$$固有値の総和（全慣性）=\sum_{k}^{K^*}\lambda_k^A=\frac{n^*}{M}-1\qquad(126)$$

つまり，インジケータ行列に展開したときの延べの次元数（n^*）の（項目数 M に対する）平均から 1 を引いた数に相当する．

もうすこし，具体的に調べると以下のようになる．いま，この固有値 λ_k^A の平均を $\bar{\lambda}^A$ とし，つまりインジケータ行列の対応分析で得られる固有値の数（次元数）である $K^*=n^*-M$ で割ると，以下となる．

$$固有値\lambda_k^A の平均：\quad \bar{\lambda}^A=\frac{1}{n^*-M}\sum_{k}^{K^*}\lambda_k^A=\frac{1}{n^*-M}\times\left(\frac{n^*}{M}-1\right)=\frac{1}{M}\qquad(127)$$

したがって，もっとも一般的な，インジケータ行列から出発したときの寄与率は，以下の式から得られる．これを目安にして，λ_k^A が $1/M$ より大きいか，小さいかを調べる[70]．

$$第k固有値の寄与率：\ \nu_k=\frac{\lambda_k^A}{\sum_{k=1}^{K^*}\lambda_k^A}=\frac{\lambda_k^A}{\frac{n^*}{M}-1}=\lambda_k^A\times\frac{M}{n^*-M}\left(k=1,2,\cdots,K^*\right)\qquad(128)$$

---

[70] λ_k^A で，$k=1$ のとき，つまり λ_1^A から次の式を作る．$\alpha_C=\frac{M}{M-1}\left(1-\frac{1}{M\lambda_1^A}\right)$，これは**“Cronbach（クロンバック）のα係数”**に相当する．いわゆる尺度測定において，その尺度が安定した結果を与えるものかどうかの**“信頼性”**（reliability）の測度の 1 つがこのα係数である．

通常，これを 100 倍して，割合として用いる．

実はここで，簡単に以下の性質があることがわかる．すでに $\lambda_k^A \le 1$ であることは知っているから，これと上の式を合わせると，以下となることが容易にわかる．

$$\text{第}k\text{固有値の寄与率：}\ \nu_k = \frac{\lambda_k^A}{\frac{n^*}{M}-1} = \frac{M}{n^*-M}\lambda_k^A \le \frac{M}{n^*-M} \qquad (129)$$

（注）この寄与率の和が上に求めた λ_k^A の平均 $\bar{\lambda}^A$ に相当する．

このことは，この寄与率には上限があるということに他ならない．つまり，インジケータ行列の場合，寄与率は上の式の右辺の値（$\frac{M}{n^*-M}$）を越えることはないのである．

②固有値の総和（その 2）

バート表（行列 $\underset{K\times K}{\mathbf{B}}$）からえた固有値 $\lambda_k^B\left(k=1,2,\cdots K^*;K^*=n^*-M\right)$ の総和（総変動，全慣性）は以下となる．

$$\text{固有値の 2 乗和（バート表の固有値の総和）：}\ inertia\left(\mathbf{B}\right)=\sum_k^{K^*}\left(\lambda_k^A\right)^2=\sum_k^{K^*}\lambda_k^B \qquad (130)$$

② 固有値の総和（その 3）

Benzécri は，上のインジケータ行列の場合にみられる不都合を改善するために，以下のような指標を提案している．これを**“調整済みの総変動（慣性）”**（調整済み固有値の和に相当）という[71]．これは以下のように表される．この調整済み総変動の値は区間［0, 1］に入る．

$$\lambda_k^{adj}=\left(\frac{M}{M-1}\right)^2\times\left(\lambda_k^A-\frac{1}{M}\right)^2\left(k=1,2,\cdots,K^*\right) \qquad (131)$$

なおここで，前でみたように $\lambda_k^A=\sqrt{\lambda_k^B}$ でもある．

ここで，上に求めた λ_k^A の平均 $\bar{\lambda}^A$ を使うと，この式は，以下となる．

$$\lambda_k^{adj}=\left(\frac{M}{M-1}\right)^2\times\left(\lambda_k^A-\bar{\lambda}^A\right)^2\left(k=1,2,\cdots,K^*\right) \qquad (132)$$

ここで，項目数が 2（$M=2$）のとき，つまり単純な 2 元クロス表のときには，これらの関係は以下のようになる．

$$\lambda_k^{adj}=4\left(\sqrt{\lambda_k^B}-\frac{1}{2}\right)^2=4\left(\lambda_k^A-\frac{1}{2}\right)^2\left(\equiv\lambda_k^F\right) \qquad (133)$$

これを，λ_k^A について解くと，以下となる．

---

[71] Benzécri’s adjusted inertias あるいは modified rates という．

$$\lambda_k^A = \frac{1}{2}\left(1 \pm \sqrt{\lambda_k^{adj}}\right)\left[\equiv \frac{1}{2}\left(1 \pm \sqrt{\lambda_k^F}\right)\right] \qquad (134)$$

これは表 34 に照らすと，上の 2 つの式の右辺の括弧内に示したように，$\lambda_k^{adj} = \lambda_k^F$ と読み替えられる．つまりこの調整済み寄与率 λ_k^{adj} とは，2 元クロス表（**F**）の場合の多重クロス表（バート表）への自然な拡張となっている．

⑤ 固有値の総和（その4）

Greenacre は，さらに形を変えた別の“調整済み総変動”を提案している[72]．これを**“非対角の慣性（2 乗和）の平均”**（average off-diagonal inertia）という[73]．

$$\text{非対角の慣性（２乗和）の平均} = \frac{M}{M-1} \times \left(inertia\left(\mathbf{B}\right) - \frac{n^* - M}{M^2} \right) \qquad (135)$$

なお，この式の要素を分けて，それぞれを以下のように非対角部の情報と対角部の情報として用いる．

$$\text{非対角の 2 乗和（off-diagonal inertia）：} \ inertia\left(\mathbf{B}\right) - \frac{n^* - M}{M^2} \qquad (136)$$

$$\text{対角の 2 乗和（diagonal inertia）：} \ \frac{n^* - M}{M^2} \qquad (137)$$

以上の性質はアイテム・カテゴリー型データ表あるいは多重クロス表の分析を扱うときに知っておくと便利である．

5. 6. 2 数値例による固有値と寄与率の関係の確認

ここで，上に述べた関係，性質を数値例で示そう．

数値例 1:

いま，ある 7 つの質問を考える（$M = 7$）．各質問にはそれぞれ 5 つの選択肢があるものとする．これをインジケータ行列に展開すると，延べで 7×5＝35 の総選択肢（カテゴリー）となる（$n^* = 7 \times 5 = 35$）．上の記号に合わせ，各数値を確認すると以下のようになる．

$$M = 7$$

$$n^* = 5 \times 7 = 35$$

$$K^* = n^* - M \ \Rightarrow \ 35 - 7 = 28$$

$$\sum_{k=1}^{K^*} \lambda_k^A = \frac{n^*}{M} - 1 \left(= \frac{n^* - M}{M} \right) \ \Rightarrow \ \sum_{k=1}^{K^*} \lambda_k^A = \frac{n^*}{M} - 1 = \frac{28}{7} = 4$$

ここで，各固有値と寄与率は以下のようであった．

---

[72] Greenacre et al. (2006)，Greenacre(1998)などを参照．

[73] 適当な訳語が見つからないので直訳的に訳してある．

$$\lambda_1 = 0.608, \lambda_2 = 0.506, \cdots, \lambda_{28} = 0.018$$

$$\sum_{k=1}^{K^*} \lambda_k^A = \lambda_1 + \lambda_2 + \cdots + \lambda_{27} + \lambda_{28} = 0.608 + 0.506 + \cdots + 0.025 + 0.018 = 4$$

$$\lambda_1^Z \times \frac{M}{n^* - M} \times 100 = 0.608 \times \frac{1}{4} = 15.2(\%)$$

$$\lambda_2^Z \times \frac{M}{n^* - M} \times 100 = 0.506 \times \frac{1}{4} = 12.6(\%)$$

このように，この例ではインジケータ行列を用いると，各固有値の寄与率は，$\frac{M}{n^* - M} = \frac{M}{K^*} = \frac{1}{4} = 0.25(25\%)$ を越えることはない（式（129））．よって固有値と寄与率の大きさの解釈に注意せねばならない．

また M（質問の総数）にくらべて K^*（全質問の延べの総選択肢数から項目数を引いた総次元数）が多くなると次第に寄与率が小さくなることもわかる．これは，項目数に比べて，選択肢数を増やすほど情報が曖昧になる傾向にあるということを示唆している．

数値例 2：固有値の関係

ここで，サンプル数 $N = 30$，項目が 2 項目（Q_1, Q_2），$n_i = 6, n_j = 5$，つまり $n^* = n_i + n_j = 11$ の例をトイ・データで確かめよう．上で示した関係，とくに，表 34 に要約した 2 つの質問項目の場合の，固有値の関係が確認できる．

表 35　もとのデータ表（C 表）

個体＼項目	Q_1	Q_2	個体＼項目	Q_1	Q_2
1	5	1	16	3	3
2	6	1	17	3	3
3	6	1	18	4	3
4	6	1	19	4	3
5	3	2	20	5	3
6	4	2	21	1	4
7	4	2	22	2	4
8	4	2	23	2	4
9	5	2	24	3	4
10	5	2	25	4	4
11	6	2	26	1	5
12	2	3	27	1	5
13	2	3	28	1	5
14	3	3	29	1	5
15	3	3	30	2	5

① バート表（B表）を作る

表 36　もとのデータ表から得たバート表 B

項目＼カテゴリー		Q_1						Q_2				
項目		1	2	3	4	5	6	1	2	3	4	5
Q_1	1	5									1	4
	2		5							2	2	1
	3			6					1	4	1	
	4				6				3	2	1	
	5					4		1	2	1		
	6						4	3	1			
Q_2	1					1	3	4				
	2			1	3	2	1		7			
	3		2	4	2	1				9		
	4	1	2	1	1						5	
	5	4	1									5

（*）数値の記入のない要素はゼロである．

② インジケータ行列（A 表）とバート表（B表）から得た固有値

ここでは，固有値数は，$K^* = n^* - M = (n_i + n_j) - M = 11 - 2 = 9$（個）となる．これを求めると表 37 のようになった．ここで，$\lambda_k^B = \left(\lambda_k^A\right)^2$ の関係が確認される．

表 37　インジケータ行列（A 表）とバート表（B表）の固有値の関係：$\lambda_k^B = \left(\lambda_k^A\right)^2$

k	$\lambda_k{}^A$（A表から）	$\lambda_k{}^B$（B表から）
1	0.93678	0.87755
2	0.86601	0.74998
3	0.68821	0.47364
4	0.59313	0.35180
5	0.50000	0.25000
6	0.40687	0.16554
7	0.31178	0.097208
8	0.13399	0.017953
9	0.063220	0.0039997

③ 2 元クロス表（F 表）から得た固有値

上のバート表の非対角ブロックに位置する 2 元クロス表（$Q_1 \times Q_2$ または $Q_2 \times Q_1$）から出発すると，固有値数は $K = \min\{m,n\} - 1 = \min\{6,5\} - 1 = 4$（個）となる．この得られた固有値，ここでは λ_k^F と記したが，これは表 38 となる．これとインジケータ行列の対応分析法からえた固有値 λ_k^A との関係を確かめると表 39 となる．ここでは，$n_i < n_j$ なので，

$n_j - n_i = 6-5 = 1$（個）の固有値が，$\lambda_5^A = 1/2 = 0.5$ となる．他の 8 個の固有値は表 39 のようになる．

表 38　2 元クロス表の対応分析の結果
（固有値，特異値，寄与率，累積寄与率）

成分（k）	特異値 $\sqrt{\lambda_k^F}$	固有値 λ_k^F	寄与率(%)	累積寄与率(%)
1	0.87356	0.76311	51.7	51.7
2	0.73203	0.53586	36.3	88.0
3	0.37643	0.1417	9.6	97.6
4	0.18626	0.03469	2.4	100
	固有値和	1.47537		

表 39　2 元クロス表から得た固有値 λ_k^F との比較：$\lambda_k^A = \left(1 \pm \sqrt{\lambda_k^F}\right)/2$

λ_k^F	$\lambda_k^A = \left(1+\sqrt{\lambda_k^F}\right)/2$	$\lambda_k^A = \left(1-\sqrt{\lambda_k^F}\right)/2$
λ_1^F	0.93678（$=\lambda_1^A$）	0.06322（$=\lambda_9^A$）
λ_2^F	0.86601（$=\lambda_2^A$）	0.13399（$=\lambda_8^A$）
λ_3^F	0.68821（$=\lambda_3^A$）	0.31178（$=\lambda_7^A$）
λ_4^F	0.59313（$=\lambda_4^A$）	0.40687（$=\lambda_6^A$）
$K = \min\{n_I, n_J\} - 1$ $= \min\{6,5\} - 1 = 4$（個）の固有値	（λ_k^A の値の大きい方から 4 個）	（λ_k^A の値の小さい方から 4 個）

これを，JMP スクリプト（MCA スクリプト）を用いて計算すると，次の固有値や寄与率が得られる（図 13）．この棒グラフの中央に引かれた縦線は，平均値 $1/M$ に相当する．これを寄与率の判断の 1 つの目安とする．これからは，はじめの 3〜4 成分あたりを観察すればよさそうである．

次元	固有値	割合(%)	累積(%)	固有値 2乗	割合(%)
1	0.937	20.82	20.82	0.878	29.37
2	0.866	19.24	40.06	0.750	25.10
3	0.688	15.29	55.36	0.474	15.85
4	0.593	13.18	68.54	0.352	11.78
5	0.500	11.11	79.65	0.250	8.37
6	0.407	9.04	88.69	0.166	5.54
7	0.312	6.93	95.62	0.097	3.25
8	0.134	2.98	98.60	0.018	0.60
9	0.063	1.40	100.00	0.004	0.13

図 13　インジケータ行列とバート表から得た固有値，寄与率，累積寄与率

(*) 左側が λ_k^A，右側が $\lambda_k^B = \left(\lambda_k^A\right)^2$ に対応する．

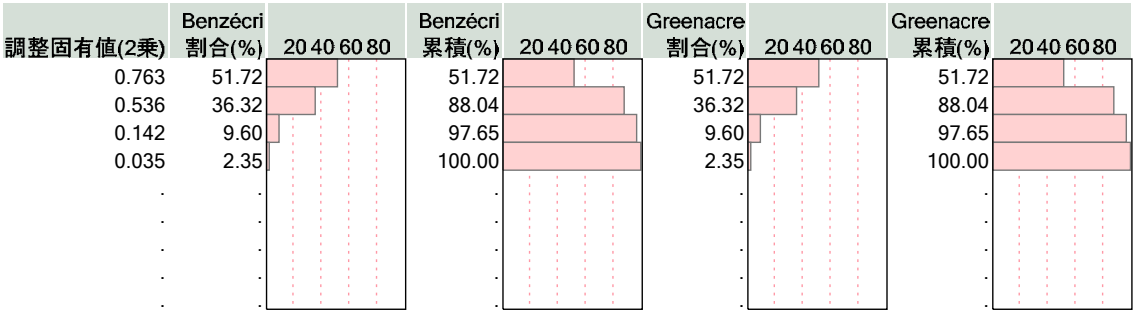


図 14　Benzécri と Greenacre の調整済み固有値と寄与率，累積寄与率
(*)ここは，2 項目であるので，両者の値は同じ

図 14 は，調整済みの固有値と寄与率である．ここでは，はじめの 2 成分あたりに注目すればよさそうだと示唆される．またここでは 2 つの質問項目であるから，Benzécri の基準と Greenacre のそれとは値が同じになっている．

数値例 3:

ある市民意識調査（農業公園に関する市民意識調査）を例として分析した．ここでは，4 つの質問と 2 つの人口統計学的変数（性別，年齢区分），合わせて 6 項目（$M=6$）を用いた（細かいことは省略）．用いた 6 項目とその選択肢数は以下のとおりである．

性別記入（選択肢数：2）
年齢区分（選択肢数：6）
近くの緑地や公園等をよく散策している（選択肢数：4）
昔からの習慣をよく守っている（選択肢数：4）
神社や，お寺詣りをよくする（選択肢数：4）
自分のなすべき役割は積極的に果している（選択肢数：4）

以上から，$n^*=24, K^*=24-6=18$ となる．計算で得られた固有値の和＝3 であったが，これを上の式から計算すると，以下のようになり一致する．

$$\frac{n^*-M}{M}=\frac{K^*}{M}=\frac{18}{6}=3$$

さらに，計算で得た $K^*=18$（個）の固有値の和を求めると，以下となり，確かに上に一致する．

$$\sum_{k=1}^{K^*}\lambda_k^A=\lambda_1+\lambda_2+\cdots+\lambda_{18}+\lambda_{19}=0.341+0.279+\cdots+0.093+0.085=3.00$$

さらに，各固有値の 2 乗和を求めると，以下となる．

$$\text{固有値の 2 乗和：}\sum_{k}^{K^*}\left(\lambda_k^A\right)^2=0.1165+0.0779+\cdots 0.0087+0.0073=0.5738$$

一方，（同じ質問項目から）バート表を作って，これに通常の対応分析を適用したところ，以下を得た．

$$inertia\left(\mathbf{B}\right)=\sum_{k}^{K^{*}}\lambda_{k}^{B}=0.573838\cdots=0.5738$$

これは，上に求めた値に一致することが確かめられる．

数値例 4:

ところで，上の数値例 3 については，多重対応分析の JMP スクリプトを用いて得られた結果もあわせて挙げておこう．このスクリプトは，まだ試用段階のプロトタイプ版であり，ここではいくつかの出力情報の抜粋表示にとどめる．

① 各種の固有値，その寄与率の性質の観察

数値例 3 の説明に用いた内容の，数値計算結果を示した．用いたデータの情報，固有値，寄与率などが出力される．

多重対応分析

データ情報

データ名: ◂▸柏市農業公園調査

データ中の人数	413
分析に使われた人数	391
質問数	6
全カテゴリー数	24
分析に使われた全カテゴリー数	24

固有値の和

固有値の和	3
固有値の2乗	0.57384
Benzecri和	0.07157
Greenacre和	0.08861
非対角の2乗	0.07384
対角の2乗	0.5

2乗和において対角ブロック部分が占める割合 : 87.13%

図 15　設定した条件と，各種の固有値，寄与率

次元	固有値	割合(%)	累積(%)	固有値 2乗	割合(%)	累積(%)
1	0.3413	11.38	11.38	0.1165	20.29	20.29
2	0.2791	9.30	20.68	0.0779	13.58	33.87
3	0.2319	7.73	28.41	0.0538	9.37	43.24
4	0.2003	6.68	35.09	0.0401	6.99	50.24
5	0.1951	6.50	41.59	0.0381	6.64	56.87
6	0.1847	6.16	47.75	0.0341	5.95	62.82
7	0.1730	5.77	53.52	0.0299	5.22	68.03
8	0.1688	5.63	59.14	0.0285	4.96	73.00
9	0.1549	5.16	64.30	0.0240	4.18	77.18
10	0.1516	5.05	69.36	0.0230	4.01	81.18
11	0.1345	4.48	73.84	0.0181	3.15	84.34
12	0.1319	4.40	78.24	0.0174	3.03	87.37
13	0.1278	4.26	82.50	0.0163	2.84	90.22
14	0.1221	4.07	86.57	0.0149	2.60	92.81
15	0.1190	3.97	90.54	0.0142	2.47	95.28
16	0.1053	3.51	94.05	0.0111	1.93	97.22
17	0.0931	3.10	97.15	0.0087	1.51	98.73
18	0.0854	2.85	100.00	0.0073	1.27	100.00

図 16　固有値 λ_k^A と $\lambda_k^B=\left(\lambda_k^A\right)^2$ と各寄与率，累積寄与率

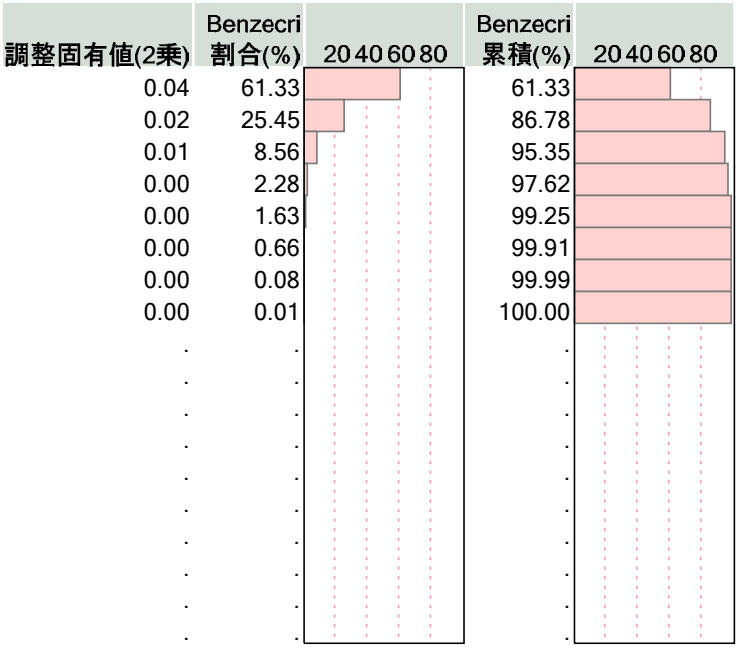


図 17　調整済み固有値とその寄与率（Benzécri の指標）

前と同様に，図 16 の棒グラフの中央に引かれた縦線は，平均値$1/M$ に相当する．これを寄与率の判断の 1 つの目安とする．たとえばこの例では，はじめの 3～4 成分あたりまでを重視すれば良さそうである．さらに，図 15 の調整済み固有値と寄与率によれば（ここでは Benzécri の指標），はじめの 2 成分で十分のようだ．

③ 質問項目とその選択肢に対する成分スコアの布置図

用いた 6 項目とその選択肢，つまり所与のデータ表（バート表）の列（であり行である）成分スコアの布置図を描くと図 18 のようになる．

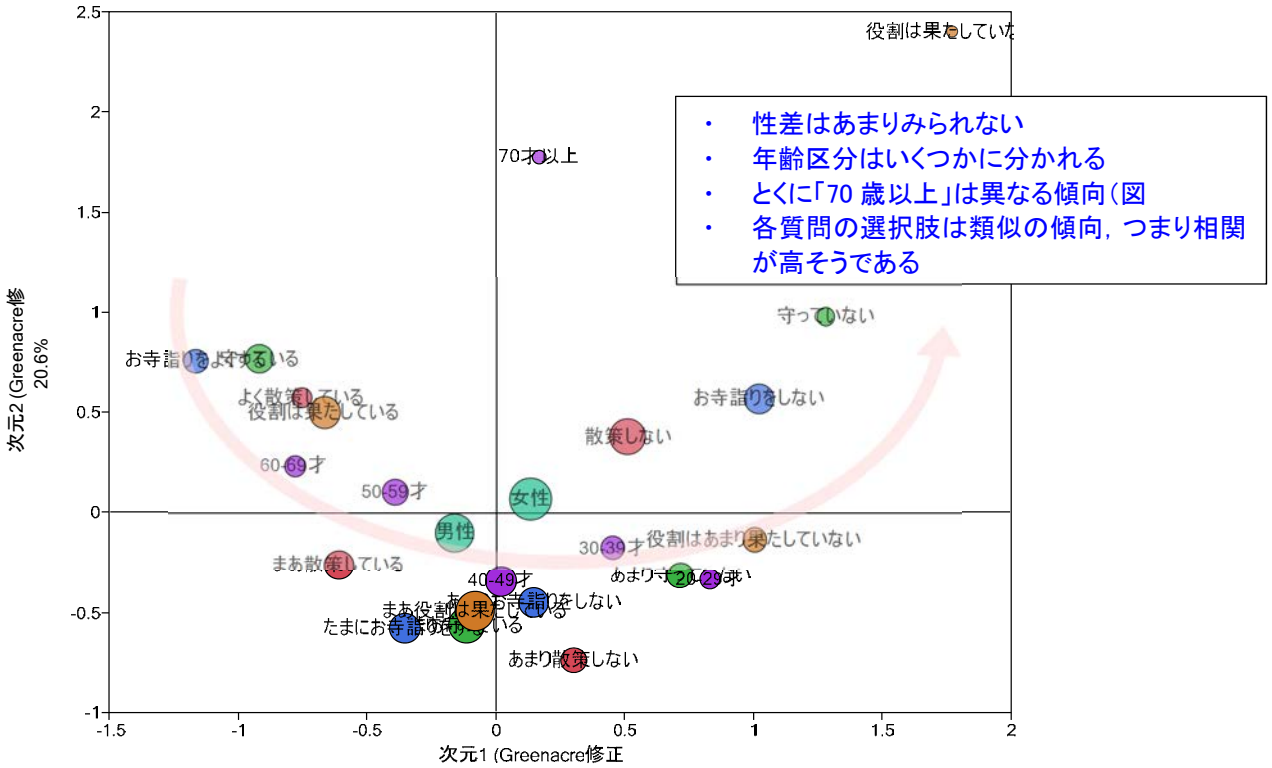


図 18　質問項目とその選択肢に対する成分スコア

④ 回答者の成分スコア

さらに，用いた項目別に回答者（サンプル）の成分スコアを布置図とし，これに**“集中楕円”**[74]（concentration ellpses）も書き入れてみた．ここでは，性別，2 つの質問について描いてみた（図 19，図 20）．性差，質問項目の選択肢間の差異の有無が観察される．

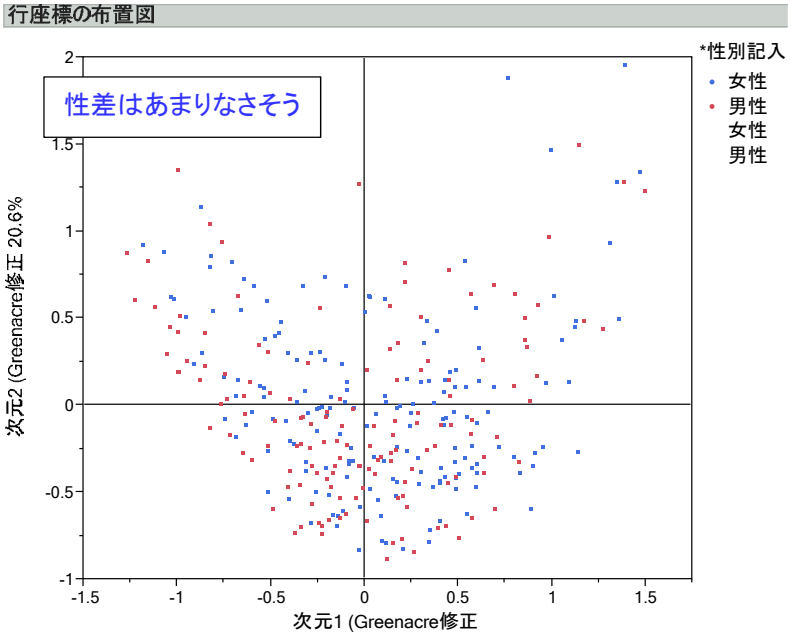


図 19　回答者の成分スコアの「性別」でみた集中楕円

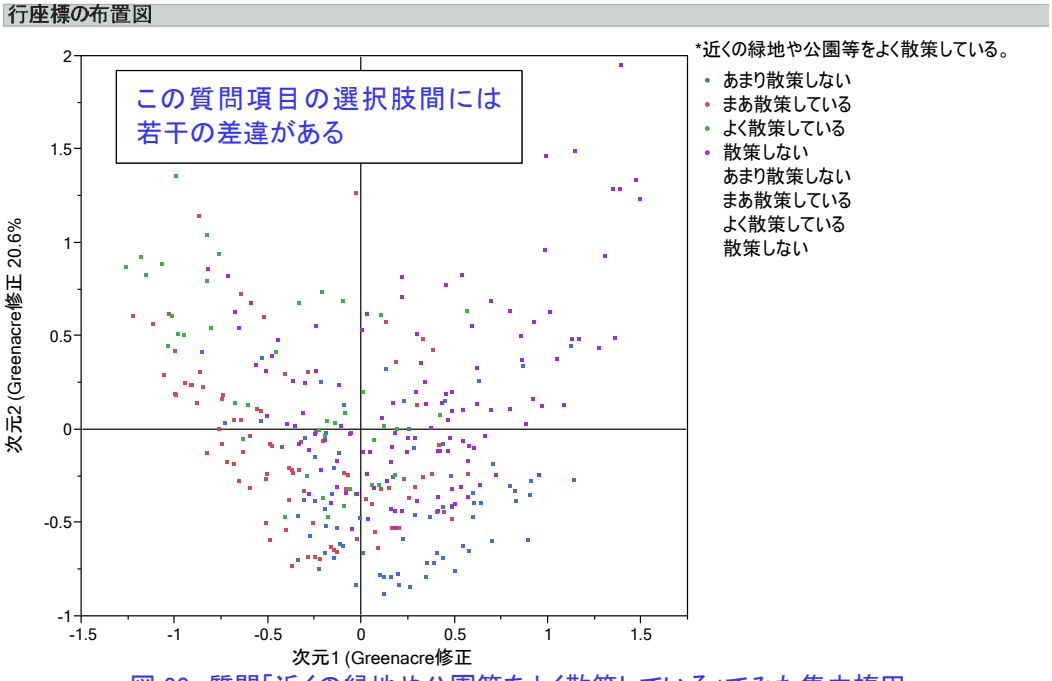


図 20　質問「近くの緑地や公園等をよく散策している」でみた集中楕円

[74] ここでいう集中楕円は，個々の点（回答者）の分布する範囲の目安である（選択肢別の成分スコアの分布の主軸の向きと相関を観察するツール）．これについては，たとえば Le Roux, B. and Rouanet,H. (2010)を参照．

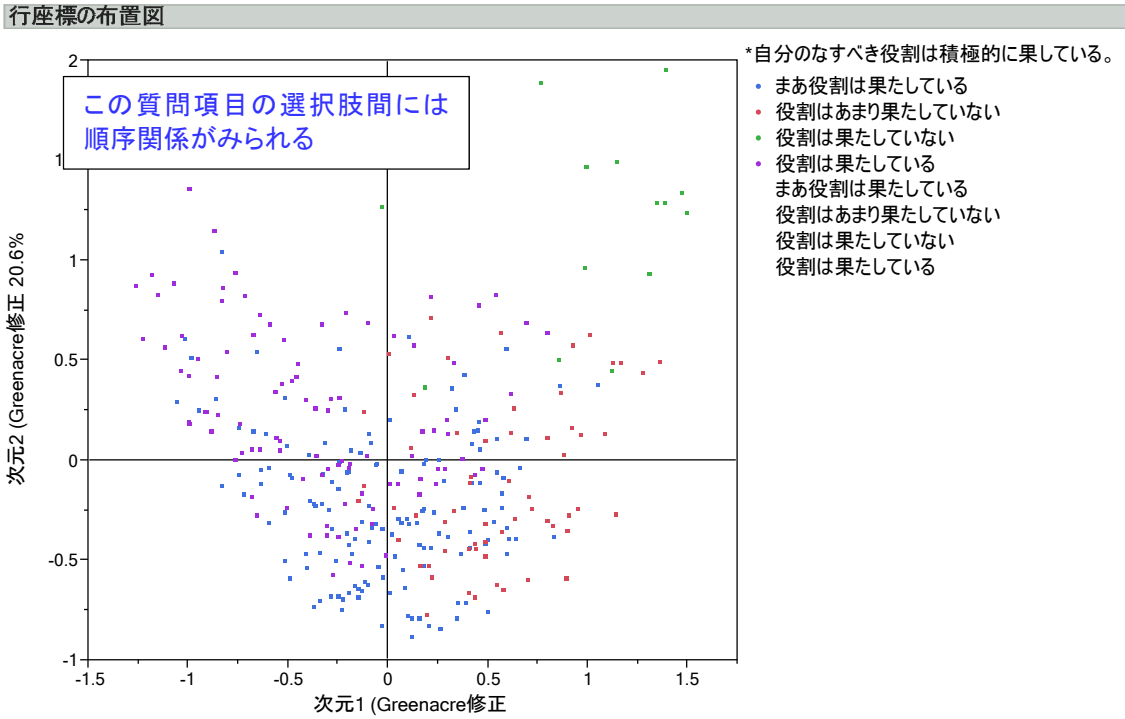


図 21 質問「自分のなすべき役割は積極的に果している」でみた集中楕円

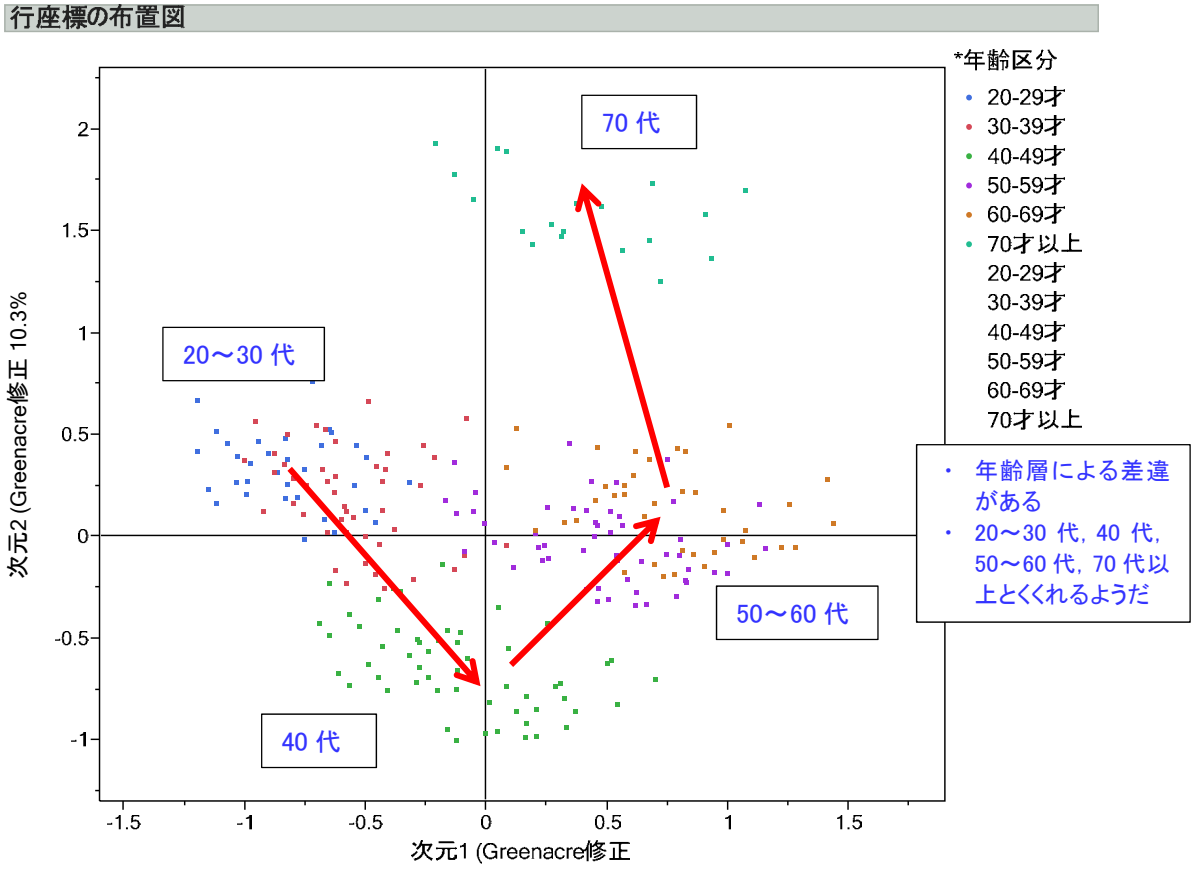


図 22 「年齢区分」でみた集中楕円

⑤ 絶対寄与度の観察

はじめの 3 成分について，絶対寄与度をグラフにした．多重対応分析では各項目（質問文）の各選択肢に対してそれぞれの寄与の程度が示される．棒グラフの色が各成分に対応する．

ここでは，第 1 成分（次元 1）が青，第 2 成分（次元 2）が赤，第 3 成分（次元 3）が緑に，それぞれ対応する．性別，年齢区分が，70 代を除いてはさほど軸の説明には寄与していないことが分かる（この場合，上の布置図の観察にも合っている）．

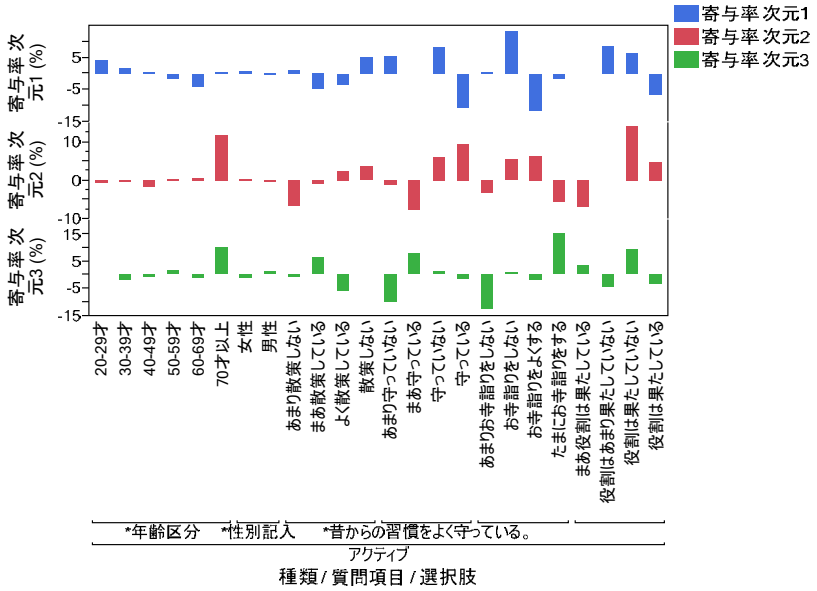


図 23　絶対寄与度の比較

⑥ 相対寄与度の観察

最後に，相対寄与度（平方相関）を示す．ここでも，初めの 3 成分について，累積寄与度を示した．つまり，グラフの棒の長さが長いほど，初めの 3 成分の説明力が高いことになる．また，グラフの色は，下から順に第 1 成分～第 3 成分と対応する．たとえば，年齢区分では「70 才以上」が特徴的で（はじめの 3 成分の説明力が高く），これは（この図の形式的な観察では）第 2 成分（赤）と第 3 成分（緑）の成分を説明している（図 22 も観察してみよう）．「昔からの習慣を守っているか」の選択肢「守っている」は，初めの 3 成分でかなりの説明がつけられ（50%を越えている），また第 1 成分（青）と第 2 成分（赤）の説明力が高いようだ，となる．

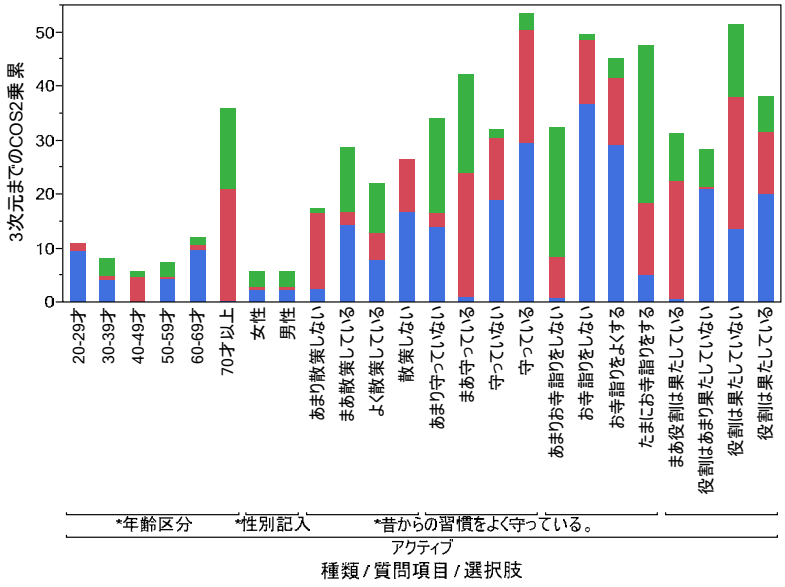


図 24　相対寄与度の比較

おもなキーワード

対応分析法（CA：Correspondence Analysis/AFC: Analyse Factorielle des Correspondances），数量化法Ⅲ類（パターン分類），クロス表・2 元データ表の分析，総変動と慣性，慣性の全体（全慣性），多重対応分析法（MCA：Multiple Correspondence Analysis），バート表とバート行列，アイテム･カテゴリー型，インジケータ行列，完備排反型行列，非定型・非構造化データ，自由回答・自由記述，テキスト型データ（textual data），テキスト・マイニング，大規模で疎なデータセットの処理，双対性，プロファイルとその同等性，（平方）カイ二乗距離と（平方）ユークリッド距離，布置図・同時布置図，追加処理と追加要素，実際データと追加データ

【参考文献】　※第Ⅰ部，第Ⅲ部と重複あり

[1] 岩坪秀一（1987）：数量化法の基礎，朝倉書店.

[2] 林知己夫（1993）：数量化－理論と方法，朝倉書店.

[3] 西里静彦（2007）:データ解析への洞察，K・G りぶれっと（No.18），関西学院大学出版会.
（*）小冊子ではあるが，計量心理学の立場から尺度化の立場から数量化のあり方について平易かつ的確に述べられている.

[4] 大隅昇，L. Lebart，他（1994）：記述的多変量解析法，日科技連出版社.

[5] 大隅昇（1989）：統計的データ解析とソフトウェア，日本放送出版協会.

[6] Benzécri, J.-P. (1976) : *L'Analyse de Données, Tome 1*: *Taxinomie*, *Tome* 2: *L'Analyse des Correspondances*, Dunod (second edtion).

[7] Benzécri, J.-P. (1980) : *Pratique de L'Analyse des Données – Analyse des Correspondances Exposé Elémentaire -, Tome I*, Dunod.

[8] Benzécri, J.-P. (1980) : *Pratique de L'Analyse des Données – Linguistique et Lexicologie -, Tome III*, Dunod.
（*）この *Pratique de L'Analyse des Données* は 3 巻からなる．とくにこの第Ⅲ巻では，言語学・語彙研究の事例が紹介されている.

[9] Berry, M.W. (1992): Large-Scale Sparse Singular Value Computations, *The International Journal of Supercomputer Applications*, **6**, 1, 13-49.

[10] Clausen, Sten-Erik (1998): *Applied Correspondence Analysis*, Series: Quantitative Applications in the Social Sciences No.121, Sage Publications, Inc.
（*）藤本一男（訳），「対応分析入門 原理から応用まで」，オーム社．全訳に補足として訳者が加えた R 言語を用いた説明がある.

[11] Deewester, S. Dumais, S.T., Fumas, G.W., Landauer, T.K., and Harshman, R. (1990) : Indexining by Latent Semantic Analysis, *Journal of the American Society for Information Science*, **41**(6), 391-407.

[12] Douglas, J., Green, F.E., and Schaffer, C.M. (1986): Interpoint Distance Comparisons in Correspondence Analysis, *Journal of Marketing Research*, Vol. XXIII, August, 271-280.

[13] Douglas, J., Green, F.E., and Schaffer, C.M. (1987): Comparing Interpoint Distances in Correspondence Analysis: A Clarification, *Journal of Marketing Research*, Vol. XXIV, November, 445-450.

[14] Escofier, B. And Pages, J. (1990) : *Analyses Factorielles Simples et Multiples*, Dunod.

[15] Everitt, B.S. (1977): *The Analysis of Contingency Tables*, second edition, Chapman & Hall.

[16] Everitt, B.S. and Dunn, G. (2001): *Applied Multivariate Data Analysis*, second edition, Arnold.

[17] Gauch, H.G. et al. (1977): A Comparative Study of Reciprocal Averaging and Other Ordination, Techniques, *J. Ecol.* **65**, 157-174.

[18] Goodman, L.A. (1986): Some Useful Extensions of Usual Correspondence Analysis Approach and the Usual Log-Linear Models Approach in the Analysis of Contingency Tables, *International Statistical Review*, **54**, 3, 243-309.

[19] Greenacre, M.J. and Blasius, J. (eds.) (2006): *Multiple Correspondence Analysis and Related Methods*, Chapman and Hall/CRC.

[20] Greenacre, M.J. (2007): *Correspondence Analysis in Practice* (second edition), Chapman and Hall/CRC..

[21] Greenacre, M.J. (1984): *Theory and Applications of Correspondence Analysis*, Academic Press.

[22] Hill, M.O. (1973): Reciprocal Averaging –An Eigen Vector Method of Ordination, *J. Ecol.* **61**, 237-249.

[23] Hill, M.O. (1974): Correspondence Analysis-Neglected Multivariate Method-, *J. Roy. Stat. Soc.*,

Series C, **23**, 340-354.
[24] Israels, A. (1987): *Eigenvalue Techniques for Qualitative Data*, DSWO Press, Leiden.
[25] Jackson, J.E. (1991, 2003): *A User's Guide to Principal Components*, John Wiley & Sons.
[26] Jambu, M. (1989): *Exploration Informatique et Statistique des Données*, Dunod.
[27] Lebart, L., Salem, A. and Berry, L. (1998): *Exploring Textual Data,* Kluwer Academic Publishers.
[28] Le Roux, B. and Rouanet, H. (2010): *Multiple Correspondence Analysis*, Series: Quantitative Applications in the Social Sciences No.163, Sage Publications, Inc.
[29] Le Roux, B. and Rouanet, H. (2010): *Geometric Data Analysis – From Correspondence Analysis to Structured Data Analysis*, Kluwer Academic Publishers.
[30] Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979): *Multivariate Analysis*, Academic Press.
[31] Maulman, J. (1982): *Homogeneity Analysis of Incomplete Data*, DSWO Press, Leiden.
[32] Nishisato, S. (1980): *Analysis of Categorical Data; Dual Scaling and its Applications,* University of Toronto Press.
[33] Rizzi, A. (ed.) (1995): *Some Relations between Matrices and Structures of Multidimensional Data Analysis*, Giardini Editore e Stampatori in Pisa.
[34] Volle, M. (1985): *Analyse des Données*, 3eme édition, Economica

注：一部はすでに絶版となった本もある．Amazon.com などから中古本で購入可能なものもある．

第Ⅲ部

対応分析法とクラスター化法

―WordMiner，JMP スクリプトによる分析―

［暫定版］

大隅　昇

0. はじめに

この資料では，おもに対応分析法の特性を活かしたクラスター化法について述べる．ここでは，対応分析法で得られた成分スコアのクラスター化でどのような情報が得られるのか，その出力情報（統計量や指標，グラフィカル情報）をどう解釈するのか，といったことを，簡単なデータセットと事例データを用いて説明する．このため，ここでは，WordMiner と JMP（スクリプト）を用いた例示にそって述べる．しかし，これらのツールがなければ内容が分からない，ということではなく，ツールの利用ははあくまでも数値確認の助けとするためである．また，数理的な細かい内容を述べることが目的でないので，より詳しい情報あるいは数理的な説明や解説は，うしろに挙げた参考文献を参照していただきたい．

対応分析法と同様に，ここでも特有の用語句が登場する．すでにその多くは，第Ⅰ部，第Ⅱ部で紹介したのであるが，ここではさらに以下の記述で，主要な語句には英語を併記する場合がある．これは参考文献と併せて読む場合を配慮してのことである．対応分析法ほかで用いる元来の英語あるいは仏語の日本語訳が，国内では必ずしも統一されていないことがある．たとえば，“Correspondence Analysis”，“Analyse des Correspondances”がその典型例だが，ここではこれに“対応分析”という用語をあてる．これをコレスポンデンス分析とか，そのままコレスポンデンス・アナリシスとしている場合もみられる．また，すでにみた，“慣性”（inertia）あるいは慣性モーメント，“カイ二乗距離”（Chi-square distance, χ^2－distance），“プロファイル”（profiles）と“雲”（nuage, cloud），ストレッチ・プロファイル（streched profiles），成分スコア（coordinates），質量（mass），重心（barycentre，centroid），重心座標系（barycentric coordinate system）など，対応分析に特有の用語句もある．そのようなことで，ここでは書き手の判断で必要に応じて英語も記すようにした．また，対応分析法については，別に用意の「第 1 部」第Ⅱ部」の資料に述べてあるので，それらを併せて参考情報としていただきたい．

1. WordMiner におけるクラスター化法（概要）

分析対象を分類する方法はさまざまである．統計的データ解析をはじめ，データ・マイニングやテキスト・マイニングでも分類手法は重要なツールとなっている．また，その手法の呼称もさまざまである．かつては，**自動分類法**（automatic classification），**数値分類法**（nimerical taxonomy）といい，またパターン認識などでは，教師なし分類（unsupervised classification）などと呼ばれた．これらをクラスタリングあるいは**クラスター分析**（cluster analysis）ということもある（しかし，自動分類法・数値分類法とクラスター分析の源流はまったく異なる）．また，教師あり分類（supervised classification）のことを多変量解析などでは“判別分析”と呼んで，いわゆる自動分類法とは異なる位置づけで考えてきた．

自動分類法あるいはクラスター化法は，その**算法**（アルゴリズム）によって，いくつかに分類される．1 つは**階層的分類法**（hierarchical classification）であり，もう一つは**非階層的分類法**（non-hierarchical classification）である．これらはさらに細分され，さまざまな方法が誕生し，提案されている（大隅（1989））．もちろん，こうした自動分類法に関する研究論文，文献は無数にある[1]．

ここでは，WordMiner で用いている分類方式の要点を述べる．WordMiner では，階層的分類法のうち，**凝集型階層的分類法**（AHC：agglomerative hierarchical classification）の 1 つである**ウォード法**（正確には，ウォードの基準：Ward’s criterion によるウィシャート方式の算法：Wishart’s algorithm）による分類法）と，非階層的分類法の代表的な手法である分散最小化基準を用いる，いわゆる**分割型分類法**（partitioning-type classification）の 1 つである ***k*-平均法**（*k*-means method）を用いている．なお，なるべく大規模データの分類が可能なように，WordMiner ではこれらを混用するハイブリッド方式（あるいは**混合方式** mixed clustering approaches）の 1 つを採用している[2]．WordMiner で用いているクラスター化手順を以下に簡単

[1] 自動分類法，クラスター化法の基本的な考え方は，Everitt(1993)，Gordon（1999）などを参照するとよい．
[2] 実は，これの変形手法も無数にある．ここで用いる方法は Benzécri（1982），Murtaugh（1985）などを参照．

に要約する．以上の詳細は，文末にあげた参考文献を参照されたい．

[クラスター化の基本的な手順]

ここでは，対応分析法で得た結果を用いて，どのようなクラスター化を行うかを概観する．細かいことはアルゴリズムの流れ図などを用いた説明が必要なので，ここは“およそこのようなことを行う”ということを，なるべく言葉で記すことにする[3]．

なおすでに第Ⅰ部，第Ⅱ部で調べたように，対応分析の結果から得た成分スコア間の標準的な距離として**平方ユークリッド距離**を用いればよいことを知っている．このことは，対応分析の結果に対して，量的データに適用可能な**ほとんどの分類手法が利用できる**ことを意味する．したがって，個々の分類手法の利点・特長をうまくつなぎ合わせることで（良いとこ取りを行うことで），より実用に適したた分析方法として柔軟に対応できるという特長がある．

（手順1） 上述のように，凝集型階層的分類法と分割化型分類法とを併用する（混合方式）．
（手順2） 分類には，対応分析法によって得た“**成分スコア**”を用いる[4]．
（手順3） はじめに，分類対象を凝集型階層的分類法の 1 つである**ウォード法**を用いて分類する．つまり原則としてすべての対象をこの方法で分類する．ただしこのとき“**相互最近隣の規則**”（RNN: reciprocal nearest neighbours rule）を用いて，近い位置にある点（似ている成分スコア）の圧縮化処理を行う．これで階層的分類の手間が圧縮される（計算量が減る）という性質を用いる[5]．
（手順4） 階層的分類法で得た情報，たとえば階層分類の結合水準の変化や図化したデンドログラム（樹形図）などの観察で，適切なクラスター数（群の数）を決める．あるいは利用者が希望するクラスター数を指定する．これは利用者が恣意的に決める．
（手順5） これを“**初期分類**”として，この階層的分類法で得られたクラスター情報を用いて，次に分割型分類法（*k*-平均法）で，各クラスターの重心ベクトルをガイドとしてクラスターのチューニングを行う．いわゆる重心移動アルゴリズム（moving center algorithms）による再配置法でクラスター内の各点の移動・調整（consolidation, refinement）を行う．
（手順6） 最終的に得られたクラスター化情報を要約表として出力する．出力の統計量の意味・解釈については後ろの分析例を参照．

ここで，**“相互最近隣の規則”**による圧縮とは，簡単にいえば，ある点（成分スコア）からみて一番近く（最近隣にあって），相手の点からみてももっとも近い（最近隣にある），つまり“相互に最近隣”（mutually nearest neighbour）にある点は，より近い関係にある（よく似ている）とみなせるので，こうした点（成分スコア）を先に集めることで階層化の作業量を低減させる方法である．類似の方法に，nearest-neighbor chain アルゴリズムという方式もある．

また，クラスター化の段階では，対応分析で得た“**成分スコア**”を用いるが，すでに述べたように，これは同時にプロファイル間の“カイ二乗距離”（chi-square distance）を用いた分類を行うことに相当する．こうすることで，階層分類の結合水準は，クラスター内変動に比例しこれは（ピアソンの）カイ二乗統計量の分解・併合（加法性）を利用することにも相当する．これらについては後述の分析例を参照されたい．

クラスター化の課題の 1 つに，最適な“**クラスター数**”をいくつとするかがある．最適クラスター数をどう決めるかという問題への解答は見つかっていない．多数の研究があるが，

---

[3] WordMiner で用いるクラスター化は，ここに示す手順に準拠している．
[4] 重要なこととして，“用いる成分数の指定とクラスター化の関係”がある．これについては例も用いて後述する．
[5] とくに，初期の 2 元データ表の寸法が大きく，しかも各セル（要素）内の度数が少ない，つまり疎なデータ表の場合には，行プロファイルあるいは列プロファイルが非常に似たパターンが多いことがある．こういうときに，無数の類似パターン（つまり距離が近いプロファイル）をこの規則にしたがうアルゴリズムで圧縮化することは効率的である．

多くは経験則的であるか，あるいは特定の構造の検出に向いた方法であって，一般化されたクラスター数決定の方法は「ない」といってよい．この事情は，ヒストグラムの級数の決め方や，いわゆる最適層別問題にも通底することである．

2. 分類対象として扱うデータとデータ表

ここで用いるクラスター化では，対応分析法で得た“**成分スコア**”（coordinate）を対象に分類操作を行う．対応分析法で扱うデータ表の形式は，原則として**非負の要素からなる 2 元データ表**であればよい．また一般に，扱うデータ表の寸法は，行・列ともに非常に大きく，またセル内の度数が非常に少ない疎な行列となることが多い．とくにテキスト型データの場合はこの傾向が顕著である．たとえば，WordMiner ではこれを大別すると「(サンプル) × (構成要素変数)」，「(構成要素変数) × (質的変数)」がある．このように，分析対象は，サンプル（回答者，個体，ケース），構成要素（用語，語句など単語群），質的変数（調査の質問・選択肢，人口統計学的変数，クラスター化で得たクラスター変数など）をそのときどきの状況に応じて任意に扱えることが必要となる．

成分スコアを用いることから，利用時に指定する対応分析法で得た“成分数”（固有値数）とクラスター化の関係を知っておくことが重要である．1 つは，**カイ二乗統計量の加法性あるいは分解可能性**を用いること，もう 1 つは，対応分析で得た成分スコアを用いる**平方ユークリッド距離**にもとづくクラスター化は，元のデータ表の**平方カイ二乗距離**[6]（squared Chi-square distance）を用いる分類に同じであるという関係をうまく利用することである．このときに，いくつかの重要な性質がある．これらについては例を用いて順をおって説明する．

3. 簡単な例によるクラスター化手順の説明

ここでは，電卓と筆算で追跡確認できるような簡単な例を用いてクラスター化の手順を説明する．数理的な細かい説明は参考文献に譲って，ここではおもに出力結果として得られる情報の読み方，解釈の方法について述べる．また，WordMiner や JMP が出力した情報だけでは説明に不十分あるいは不足とみられる場合は，補足の情報を加えるようにした．

3.1 対応分析法のためのデータセットとここでの目標

3.1.1 準備

すでに第 I 部，第 II 部で述べたことであるが，対応分析法にかかわる基本情報をあらためて要約しておこう．対応分析法では，出発行列として“**2 元データ表**”（two-way table）を扱う．ここでいま，**寸法が（$m\times n$）の 2 元のデータ表**（m 行，n 列のデータ表）を以下の式（行列）で表す．2 元データ表とは，原則として，表の各セル内の度数が非負の数値であって，また行あるいは列の比率のパターンを考えることが意味あるようなデータ表である．たとえばもっとも単純な例として“**クロス表**”がある．

$$\underset{m\times n}{\mathbf{F}}=\left(f_{ij}\right)\qquad\left(f_{ij}\geq 0, i\in I, j\in J\right)\tag{1}$$

ここで，f_{ij} は 2 元データ表の (i,j) セル内の度数である（よって非負の値）．また I と J は，それぞれ行と列の項目とその要素の集合を表わし以下のように書いておく．つまりクロス表であれば 2 つの**質問項目** I と J と，それぞれの**選択肢**（カテゴリー，オプション）に相当する．

$$I=\{1,2,\cdots,i,\cdots m\},\quad J=\{1,2,\cdots,j,\cdots n\}\tag{2}$$

---

[6] 本稿では“平方カイ二乗距離”とした．第 I 部，第 II 部では，これを単に“カイ二乗距離”とした箇所もある．平方カイ二乗距離とカイ二乗距離については後述．

この 2 元データ表について，以下の式を用意する．

2 元データ表の行和（項目 I の周辺度数）：$f_{i+}=\sum_{j=1}^{n} f_{ij}\ (i=1,2,\cdots.m)$ (3)

2 元データ表の列和（項目 J の周辺度数）：$f_{+j}=\sum_{i=1}^{m} f_{ij}\ \ (j=1,2,\cdots.n)$ (4)

2 元データ表の総和（総サンプル数）：$f_{++}=\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}=\sum_{i=1}^{m} f_{i+}=\sum_{j=1}^{n} f_{+j}=N$ (5)

表 1 （項目 I×項目 J）の 2 元データ表 $\underset{m\times n}{\mathbf{F}}=\left(f_{ij}\right)$

		項　目　J						
	選択肢	1	2	…	j	…	n	行和
項目 I	1	f_{11}	f_{12}	…	f_{1j}	…	f_{1n}	f_{1+}
	2	f_{21}	f_{22}	…	f_{2j}	…	f_{2n}	f_{2+}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	i	f_{i1}	f_{i2}	…	f_{ij}	…	f_{in}	f_{i+}
	⋮	⋮	⋮		⋮		⋮	⋮
	m	f_{m1}	f_{m2}	…	f_{mj}	…	f_{mn}	f_{m+}
	列和	f_{+1}	f_{+2}	…	f_{+j}	…	f_{+n}	f_{++} (=N)

以上の準備のもとに，説明用の 1 組のミニチュア・データセットを用意する[7]（架空のトイ・データ）．ある調査で，2 つの質問 I と質問 J について，質問文と回答選択肢が以下のようであるとする．このデータの分析目的は，ある回答者がリストにあるレストランからどれか 1 つ選ぶときに，どのような選択基準（評価基準）で選ぶか，その関連を調べたいという課題を想定している．

質問 I：次にあげるレストランのうち，あなたがお気に入りのレストランはどれですか．（1 つ選ぶ）

1．いりふね　2．かりや　3．きくみ　4．さとみ　5．クラーク
6．コルシカ　7．バッハ　8．ムガール　9．ラ・マレ　10．ロゴスキー

質問 J：そのレストランを選択したときの評価基準は次の 3 つのうちのどれでしょうか．（1 つ選ぶ）

1．工夫・サービス　2．味　3．量

この 2 つの質問に対して，回答者がそれぞれ 1 つだけ選択肢を選ぶものとする．このときの回答者数を N 人（＝1,284 人）として，この回答者の回答分布は表 1 のように，寸法が（N 人）×（2 つの項目 I と J ）のデータ表として集められたとしよう（表 2）．なおここでは DK（Don’t Know :わからない）や NA（No Answer:無回答）などはなかったものとする（あっても分析は可能だが，説明を簡潔にするために「なし」としただけ）．

[7] この例では，第Ⅰ部，第Ⅱ部で用いたデータを再び使う．

表 2 調査データの例［レストランの評価］

項目 / 回答者	*I*（レストラン）	*J*（評価基準）
1	バッハ	味
2	ムガール	量
3	さとみ	量
4	ラ・マレ	工夫・サービス
5	きくみ	味
：	：	：
：	：	：
N	いりふね	量
N=1,284（回答者数）		

表 3 （項目 *I*）×（項目 *J*）の2元クロス表 $\mathbf{F}_{m\times n}=\left(f_{ij}\right)$

項目 *I* ＼ 項目 *J*		評価基準			行　和
		工夫・サービス	味	量	
レストラン名	いりふね	98	25	32	155
	かりや	105	35	38	178
	きくみ	35	8	67	110
	さとみ	42	46	7	95
	クラーク	34	14	54	102
	コルシカ	32	77	13	122
	バッハ	48	76	18	142
	ムガール	49	44	16	109
	ラ・マレ	49	82	15	146
	ロゴスキー	48	35	42	125
列　和		540	442	302	1284

ここで表 2 から，質問 *I*（レストラン）と質問 *J*（評価基準）の**クロス表**を作成したところ，表 3 のようになったという．このクロス表を例とし，またここでは 2 元データ表の行の要素，つまりレストランをクラスター化で分類することを考える．なお，2 元データ表の列の要素，つまり評価基準についてもクラスター化を考えることができるが，これはすべて「行」（レストラン）を「列」（評価基準）と読み替えて考えればよい[8]．

たとえば，WordMiner では，「**（サンプル）×（構成要素変数）**」，「**（構成要素変数）×（質的変数）**または「**（構成要素変数）×（クラスター変数）**」と表記しているが，いずれも“2 元データ表”を扱っていることには変わりがない．通常はサンプル数や構成要素数（単語・語句数）はかなり大きい．つまり，データ表の寸法が大きいので，ここで使う例のように，すべての成分数（そして成分スコア）を用いることはない．しかし，原理･仕組みはここで述べる簡単な例とまったく同じに考えてよい．

このクロス表から，行和を 100（または 1）とそろえた比率の表を作る．これを“**行プロファイル**”（row profile）という．同じようにして，列和を 100（または 1）とする比率の表を作る．これを“**列プロファイル**”（column profile）という．**プロファイル**とは，行または列の大きさをそろえて相対的にパターンを比べる手続きである．たとえば，行のプロファイルは，各レストランが評価基準に対してどのようなパターン（回答比率）となる傾向があるかを知る，というように使う．対応分析はこのプロファイルの関係を，行と列の双対的関係として調べることでもある（注：表のセル内の元の頻度の大小を比べているわけではない）．

ここで，プロファイルを下のように表す．

[8] この意味で，対称性があることが対応分析法の特徴である．なお，非対称を扱う方法もある．

$$\text{行プロファイル：}\ \underset{m\times n}{\mathbf{N}_I}=\left\{q_{ij}=\frac{p_{ij}}{p_{i+}}=\frac{f_{ij}}{f_{i+}}\middle|i\in I,j\in J\right\} \tag{6}$$

$$\text{列プロファイル：}\ \underset{n\times m}{\mathbf{N}_J}=\left\{q^{*}_{ij}=\frac{p_{ij}}{p_{+j}}=\frac{f_{ij}}{f_{+j}}\middle|i\in I,j\in J\right\} \tag{7}$$

表 4 の**列和**は行の要素（ここでは 10 のレストラン）の 3 つの列要素（評価基準）の**平均比率**（行の平均プロファイル；つまり**平均ベクトル**あるいは**重心**）である．同じように，表 5 の**行和**は 3 つの列要素（評価基準）の平均ベクトル（重心）になっている．この見方があとの説明で重要になる．なおここでは，プロファイルを割合で示してあるが，実際は（計算を行ううえでは）上の式にしたがう．

このクロス表から次の式（8）にしたがい“**ピアソンのカイ二乗統計量**” χ^2_p（Chi-square statistic）を求める．これはいわゆる“クロス表の 2 つの項目の独立性”の検定を行う統計量として知られている．

$$\chi^2_p=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(f_{ij}-\dfrac{f_{i+}f_{+j}}{N}\right)^2}{\dfrac{f_{i+}f_{+j}}{N}}\left(=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{(\text{実現度数}-\text{期待度数})^2}{\text{期待度数}}\right) \tag{8}$$

これを表 3 のクロス表について求めると次の値になる．

$$\chi^2_p=330.860 \tag{9}$$

すでに指摘してきたように，対応分析法ではこのピアソンのカイ二乗統計量が重要な役割をはたすので，これを求めておく[9]．

3.1.2 分析の方針

ここで，これから行う分析，とくにクラスター化の内容を以下に要約しておこう．

① まず，**分類対象**は，ここで取り上げた 10 のレストランとする（行の選択肢の分類）．つまり，どのレストランへの回答傾向（プロファイル）が似ているかを分類で調べる[10]．
② このデータ表から，直接なんらかの方法でレストランを分類することも可能だが，ここでは表 3 のクロス表に**対応分析**を適用し，得られた“**成分スコア**”を用いてクラスター化を行う[11]．
③ データ表の寸法は行数：$m=10$（レストラン），列数：$n=3$（評価基準）である．よって，対応分析の性質から，ここで得られる成分スコアは 2 成分まで（得られる固有値は 2 個まで）となる．［固有値の個数：$K=\min\{m,n\}-1=3-1=2$］．
④ クロス表の行のプロファイル間の**平方カイ二乗距離**と，カイ二乗統計量の加法性を用いた分類を用いる．
⑤ また同時に，クロス表の対応分析でえた成分スコアを用いた**平方ユークリッド距離**[12]によ

---

[9] ここらは，「第Ⅰ部」「第Ⅱ部」を参照のこと．
[10] もちろん，列の側，ここでは「評価基準」の分類も，同じ要領で考えればよい．
[11] 成分スコアのユークリッド距離を用いることは，プロファイルのカイ二乗距離を用いることに同等．
[12] 平方ユークリッド距離を用いるということは，データの分布の分散（の和）あるいは平方和（の和）を考えることに同じである．

る凝集型階層的分類法（ウォード法）と相互最近隣の規則を用いて，レストランの分類を行う．また，これと④の操作とは実質的には同等であることがわかる（後述する）．

⑥ 通常は**クラスター数**（g）を与えてクラスター化を行う．ここでは，クラスター化の履歴を調べるために，あえて 2 群から 10 群までを指定する．なお 1 群とはすべてのレストランを 1 群とみなすということ，一方，10 群とは個々のレストランを 1 つの群とすること，つまりこれは分類を行わない場合に相当する．

⑦ 階層的分類における“**結合水準の値**”と“**カイ二乗統計量**”，それと対応分析で得られる“**固有値**”つまり**成分スコアの分散**，のそれぞれの間にはある関係がある．これを調べる．

⑧ また，得られたクラスターと各種の統計量（クラスター内変動，クラスター間変動，総変動など）をどのように評価，解釈するかを調べる．

表 4　行のプロファイル（レストランの比較）［$\mathbf{N}_{I}$ の分布］
$m\times n$

項目 I ＼ 項目 J		評価基準			行　和
		工夫・サービス	味	量	
レストラン名	いりふね	63.2	16.1	20.6	100.0
	かりや	59.0	19.7	21.3	100.0
	きくみ	31.8	7.3	60.9	100.0
	さとみ	44.2	48.4	7.4	100.0
	クラーク	33.3	13.7	52.9	100.0
	コルシカ	26.2	63.1	10.7	100.0
	バッハ	33.8	53.5	12.7	100.0
	ムガール	45.0	40.4	14.7	100.0
	ラ・マレ	33.6	56.2	10.3	100.0
	ロゴスキー	38.4	28.0	33.6	100.0
列の相対度数（列の質量） （行の重心）$p_{+j}\times 100$		**42.1**	**34.4**	**23.5**	
		$p_{+j}\,(j=1,2,3)$　（*）ここは $p_{+j}\times 100$（%）			

表 5　列のプロファイル（評価基準の比較）［$\mathbf{N}_{J}$ の分布］
$n\times m$

項目 I ＼ 項目 J		評価基準			行の相対度数（行の質量）（$p_{i+}\times 100$）
		工夫・サービス	味	量	
レストラン名	いりふね	18.1	5.7	10.6	**12.1**
	かりや	19.4	7.9	12.6	**13.9**
	きくみ	6.5	1.8	22.2	**8.6**
	さとみ	7.8	10.4	2.3	**7.4**
	クラーク	6.3	3.2	17.9	**7.9**
	コルシカ	5.9	17.4	4.3	**9.5**
	バッハ	8.9	17.2	6.0	**11.1**
	ムガール	9.1	10.0	5.3	**8.5**
	ラ・マレ	9.1	18.6	5.0	**11.4**
	ロゴスキー	8.9	7.9	13.9	**9.7**
列　和		100.0	100.0	100.0	
列の重心		$p_{i+}\,(i=1,2,\cdots,10)$　（*）ここは $p_{i+}\times 100$（%）			

4. 例による分析結果と内容の説明 ― WordMiner と JMP による探査 ―

4.1 観察（その 1） 対応分析の実行と結果の確認

表 1 のクロス表に対応分析を適用して得られる基本情報を，以下に順に示す．また各指標について，その意味と解釈を簡単に説明する．

4. 1. 1 固有値と寄与率，累積寄与率

すでに，対応分析法の数理の概要は示した．ここでははじめに，WordMiner からえた固有値，寄与率ほかを調べる．

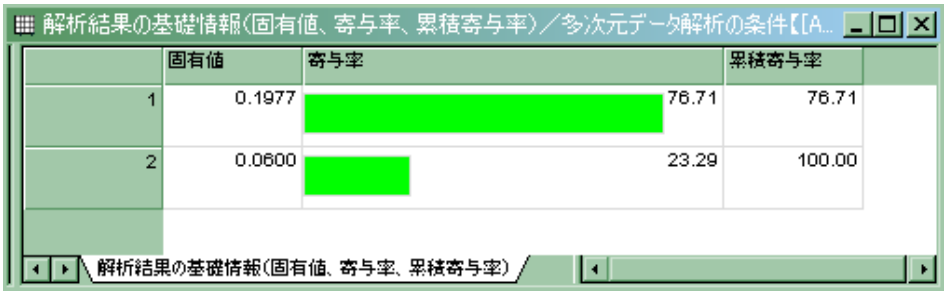


図 1 固有値，寄与率，累積寄与率

表 6 図 1 の情報の整理

k	固有値 λ_k		寄与率 ν_k	累積寄与率 $\sum_k \nu_k$
1	λ_1	0.1977	76.71	76.71
2	λ_2	0.0600	23.29	100.00
	和	0.2577	100.00	―

ここで，すでに示したある性質の確認を行う．対応分析の重要な性質の 1 つとして，“**固有値の和**”（総変動，つまり変動の総量であり情報の総量）と，上でクロス表から求めた“**ピアソンのカイ二乗統計量**”との間に次の関係がある．以後の分析と解釈で重要な性質なので，ここでは数値例として確認する．

[性質 1] 固有値の和と総変動の関係

$$\phi^2 = \sum_{k=1}^{K} \lambda_k = \frac{\chi_p^2}{N} \quad \left(\text{ここで，}\ K = \min\{m,n\} - 1\right) \tag{10}$$

上の関係がなりたつ[13]．これを例について確かめると以下のようになる．

$$\sum_{k=1}^{K} \lambda_k = \lambda_1 + \lambda_2 = 0.2577 \quad \Leftrightarrow \quad \frac{\chi_p^2}{N} = \frac{330.860}{1284} = 0.257679\cdots \doteqdot 0.2577 \tag{11}$$

つまり，

［固有値の和］
　＝［（用いたクロス表のピアソンのカイ二乗統計量）÷（クロス表の総度数）］(12)
の関係がある．

---

[13] ここで $\phi^2 = \frac{\chi_p^2}{N}$ と記したが，これは，クロス表の行と列との関係を測る**連関性の測度**の 1 つである“**平均平方関連係数**”（mean square contingency coefficient）に相当する．その表記の慣習にならってこう書いた．

対応分析は，このカイ二乗統計量を**総変動**として，これを固有値で成分ごとにどう分けるか（**分解**するか）を行っていることになる．この点で主成分分析に類似している．なおここで登場した主な性質や用語については，「第Ⅰ部」「第Ⅱ部」の対応分析法の説明で述べたことであるが，簡単に要約しておこう．

① **総変動：**これを対応分析では“total inertia”（直訳すると“**全慣性**”または**“慣性”**）という．この総変動は多変量解析などでいう**全分散**（total variance）に相当する．

② **固有値と寄与率**：対応分析で得られる固有値つまり成分スコアの分散 $\lambda_k\left(k=1,2,\cdots,K;K=\min\{m,n\}-1\right)$ から，以下の関係と寄与率が得られる．また固有値の値は非負で 1 を越えることはない（$0\le\lambda_k\le 1\ \left(k=1,2,\cdots,K;K=\min\{m,n\}-1\right)$）．**固有値の個数**は元の解析対象とした 2 元データ表の行と列の寸法の小さい方から 1 を引いた個数（$K=\min\{m,n\}-1$）となる（つまり，成分スコアの分布は，この次元数内の空間に入るということ）．また**寄与率**は以下の式となる．この ν_k を，k について累積すると**累積寄与率**となる（例：図 1，表 6 で確認）．

$$\text{寄与率：}\ \nu_k=\frac{\lambda_k}{\sum_{k=1}^{K}\lambda_k}\times 100(\%)\quad\left(\begin{matrix}k=1,2,\cdots,K\\K=\min\{m,n\}-1\end{matrix}\right)\qquad(13)$$

（第 k 成分の寄与率の式）

4. 1. 2 レストラン（行の選択肢）に対する成分スコアほか

つぎに，ここでも，WordMiner から得られる成分スコア，寄与度（相対寄与度，絶対寄与度）ほかの統計量の要約表を調べる（図 2）．ここで「構成要素数構成比」とは，表 5 の**行の平均プロファイル**（**周辺度数の割合**：p_{i+}）に相当する．対応分析では，これを“**行の質量**”（row mass）という．また「距離」とは，重心からの“**平方カイ二乗距離**”[14]を示している．また，成分スコアは対応分析で得た値でプロファイルの合成変数と思えばよい[15]．また，2 種の寄与度（絶対寄与度，相対寄与度）により，成分（k）に対するレストランの影響度を知ることができる[16]．絶対寄与度によって第 k 成分の軸の解釈を，つまり各レストランの成分軸への影響度を知ることができる．また，相対寄与度によって，第 k 成分によって，どれだけ行の選択肢，つまりここでは各レストランが近似されているか（説明できるか）を知ることができる．

ここで，**行プロファイル**とカイ二乗距離，カイ二乗統計量，固有値の和の間にある性質があるのでこれを確認しておこう[17]．

図 2 の出力情報から，構成要素構成比，距離の部分を取り出し，さらに必要な情報を別の表に整理し，以下の内容を数値例で確認する．

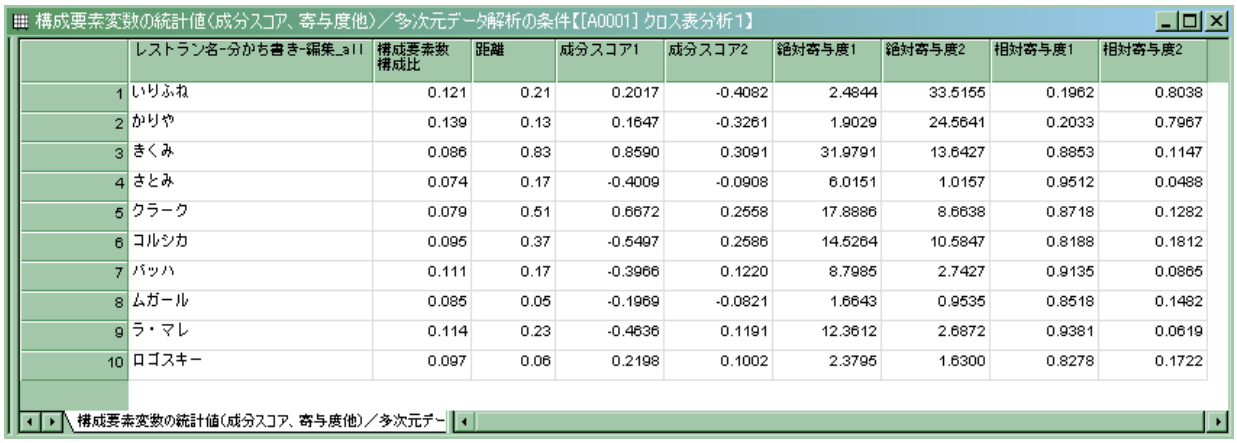
---

[14]カイ二乗距離の二乗の和の形をしているので，「平方カイ二乗距離」した．第Ⅰ部，第Ⅱ部では単に“カイ二乗距離”とも記した．ここでは，平方ユークリッド距離との対比の意味で，説明の都合でこうする．

[15] 成分スコアをはじめ，算出の方法や性質は，すでに第Ⅰ部，Ⅱ部で述べた，

[16] 2 種の寄与度についての説明はここでは述べないが，データ構造の探査のためには重要な指標である．「第Ⅱ部」に説明がある．

[17] ここで，すべてを列プロファイルと読み替えても同じことがなり立つ．



構成要素変数の統計値(成分スコア、寄与度他)／多次元データ解析の条件【[A0001] クロス表分析1】

	レストラン名-分かち書き-編集_all	構成要素変数構成比	距離	成分スコア1	成分スコア2	絶対寄与度1	絶対寄与度2	相対寄与度1	相対寄与度2
1	いりふね	0.121	0.21	0.2017	-0.4082	2.4844	33.5155	0.1962	0.8038
2	かりや	0.139	0.13	0.1647	-0.3261	1.9029	24.5641	0.2033	0.7967
3	きくみ	0.086	0.83	0.8590	0.3091	31.9791	13.6427	0.8853	0.1147
4	さとみ	0.074	0.17	-0.4009	-0.0908	6.0151	1.0157	0.9512	0.0488
5	クラーク	0.079	0.51	0.6672	0.2558	17.8886	8.6638	0.8718	0.1282
6	コルシカ	0.095	0.37	-0.5497	0.2586	14.5264	10.5847	0.8188	0.1812
7	パッハ	0.111	0.17	-0.3966	0.1220	8.7985	2.7427	0.9135	0.0865
8	ムガール	0.085	0.05	-0.1969	-0.0821	1.6643	0.9535	0.8518	0.1482
9	ラ・マレ	0.114	0.23	-0.4636	0.1191	12.3612	2.6872	0.9381	0.0619
10	ロゴスキー	0.097	0.06	0.2198	0.1002	2.3795	1.6300	0.8278	0.1722

構成要素変数の統計値(成分スコア、寄与度他)／多次元デー

図 2　成分スコア他の情報

確認 1:

距離の確認：ここで「距離」とは，重心からの**カイ二乗距離**を示している．以下で，この関係を数値的に求めて確認する．つまり，行の各選択肢（レストラン）の行プロファイルと列和つまり**行の平均プロファイル（重心）**（列の周辺度数の割合：$p_{+j}\ (j=1,2,3)$）との距離を２乗した値を求める．

確認 2:

このカイ二乗距離とピアソンのカイ二乗統計量あるいは固有値の和との関係，つまり次の性質があることを数値的に確認する．

[性質 2]

ここで，総変動（つまり，固有値の和）は以下のように表せる．

まず，クロス表の項目 I の第 $i\ (\in I)$ 選択肢について，以下の関係がある．

$$\phi^2=\frac{\chi_p^2}{N} \qquad (14-1)$$

$$=\sum_{i=1}^{m}(\text{クロス表の第}i\text{行の質量})\times\left[\text{第}i\text{行プロファイルと行の平均プロファイル（重心）との}\chi^2\text{ 距離}\right]$$

同じように，クロス表の項目 J の第 $j\ (\in J)$ 選択肢について，以下の関係がある．

$$\phi^2=\frac{\chi_p^2}{N} \qquad (14-2)$$

$$=\sum_{j=1}^{n}(\text{クロス表の第}j\text{行の質量})\times\left[\text{第}j\text{行プロファイルと列の平均プロファイル（重心）との}\chi^2\text{ 距離}\right]$$

これらの関係を順に調べる．まず，行の要素（レストラン）のプロファイルとクロス表の**列の周辺度数の割合**（つまり行の平均プロファイル：p_{+j}）を用いて，**各レストランの，重心からのカイ二乗距離を求めてみる**（表 7，表 8）．**なおここで，**数値の正確な確認のため図 2 の出力情報よりも桁数を増やしてある．

まず，ここで，行の 2 つのプロファイル i, i' 間の**平方カイ二乗距離**，およびその成分スコア間の**平方ユークリッド距離**は以下の式で表される[18]．

[18] これについては，別資料「第Ⅱ部」などを参照．ここで２つの距離の同等性を示してある．

選択肢 $i, i' \in I$ の行プロファイル間の平方カイ二乗距離

$$d_B^2(i,i') = \sum_{j=1}^{n} \frac{1}{p_{+j}} \left(q_{ij} - q_{i'j}\right)^2 = \sum_{j=1}^{n} \frac{1}{p_{+j}} \left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}} \right)^2$$

選択肢 $i, i' \in I$ に対する成分スコア間の平方ユークリッド距離

$$d_E^2(i,i') = \sum_{k=1}^{K} \left(z_{ik} - z_{i'k}\right)^2$$

この 2 つの距離，$d_B^2(i,i')$ と $d_E^2(i,i')$ とは，実は等しいことはすでに示した[19]．それをここでは，数値例で確かめる．

表 7　例:「いりふね」についてカイ二乗距離を算出

	行のプロファイル（行の比率パターン）			和チェック
	工夫・サービス	味	量	
(a) いりふねのプロファイル	0.6323	0.1613	0.2065	1.000
(b) 行の平均プロファイル [表 4 の周辺確率： p_{+j}]	0.4206 (0.420560748)	0.3442 (0.34423676)	0.2352 (0.235202492)	1.000 1.000
$(c)=\left[(a)-(b)\right]^2/(b)$	0.1066 (0.106604126)	0.0972 (0.097217561)	0.0035 (0.003502654)	**0.2073≒0.21** (0.207324342)

表 7 は，レストラン「いりふね」を 1 つの例としてこれの計算表を作ったものである．この場合は，重心からの“**平方カイ二乗距離”の和**は以下のようになる．

$$\frac{(0.6323-0.4206)^2}{0.4206} + \frac{(0.1613-0.3442)^2}{0.3442} + \frac{(0.2065-0.2352)^2}{0.2352} \qquad (15)$$
$$= 0.1066 + 0.0972 + 0.0035 = 0.2073 \fallingdotseq 0.21$$

こうして得た「0.21」が図 2 の表内の「いりふね」の「距離」に欄の数値に相当する．よって，この“**カイ二乗距離**”は次のようになる．

$$\sqrt{\frac{(0.6323-0.4206)^2}{0.4206} + \frac{(0.1613-0.3442)^2}{0.3442} + \frac{(0.2065-0.2352)^2}{0.2352}} = \sqrt{0.2073} \fallingdotseq 0.455 \qquad (16)$$

ここで，いわゆる単純な“**ユークリッド距離**”との違いを見ておこう．「いりふね」の比率の行ベクトルと列和の周辺確率つまり行の平均プロファイル（重心）（ p_{+j} ）とのユークリッド距離を求めると，「いりふね」の重心からのユークリッド距離は，以下のようになる．

ユークリッド距離の場合

$$\sqrt{(0.6323-0.4206)^2 + (0.1613-0.3442)^2 + (0.2065-0.2352)^2} = \sqrt{0.791232} \fallingdotseq 0.2813$$

上と比べると値が小さくなるという特徴がある．このカイ二乗距離はユークリッド距離とは異なり， p_{+j} を加重とする“**重み付き距離**”である．ここでは行の選択肢について示したが，式（14－2）のように，これをすべて列の選択肢に読み替えても同じ関係がなり立つ（つ

[19] 第Ⅱ部で，これを証明した．

まり，列の平均プロファイル $p_{i+}\left(i=1,2,\cdots,10\right)$ を加重とする重み付き距離とすること）．

表 8　カイ二乗距離の算出ほか（桁数を増やしてリチェック：7 桁で確認）

レストラン名（i）	カイ二乗距離の要素			構成要素数 構成比	距離	$\frac{\chi_p^2}{N}=\sum_\alpha \lambda_\alpha$ の確認
	工夫・サービス	味	量	①列の平均プロファイル（行の質量）（p_{i+}）	②平方カイ二乗距離	③＝①×②
いりふね	0.1066041	0.0972176	0.0035027	0.1207165	0.2073243	0.0250275
かりや	0.0681846	0.0633187	0.0020025	0.1386293	0.1335058	0.0185078
きくみ	0.0249137	0.2141904	0.5943787	0.0856698	0.8334828	0.0714043
さとみ	0.0011031	0.0569077	0.1108962	0.0739875	0.1689070	0.0124970
クラーク	0.0181054	0.1243993	0.3679900	0.0794393	0.5104948	0.0405533
コルシカ	0.0595549	0.2390521	0.0703164	0.0950156	0.3689235	0.0350535
バッハ	0.0162076	0.1059357	0.0499616	0.1105919	0.1721049	0.0190334
ムガール	0.0019913	0.0102716	0.0332267	0.0848910	0.0454897	0.0038617
ラ・マレ	0.0171636	0.1372508	0.0746459	0.1137072	0.2290603	0.0260458
ロゴスキー	0.0031783	0.0119870	0.0431974	0.0973520	0.0583627	0.0056817
				↑ （表 5）参照		↓ **0.2576660** （固有値の和）

他の行の選択肢（レストラン）についても同じように平方カイ二乗距離を求めて表 8 のように要約する．さらに各レストランについて求めたすべての行について和を求める．表の丸数字で示すと，表の①欄における数値は，表 5 に記載の周辺確率 p_{i+}（行の質量）である．②欄は，各レストランの重心からの平方カイ二乗距離である．また，③欄は，①と②の積である．ここで③欄の和を求めると「0.2576660（≒0.2577）」となり，これは式（10）の固有値の和（そして総変動，全慣性）に相当する．これで上の[性質 2]の式（14－1）を数値例で確かめたことになる．またここで，欄③に求めた個々のレストランの値は，そのレストランの重心からの離れ具合（距離）に相当するから，そのレストランがどういう変動をもって分布するか（重心から近いか，遠いか）の目安になる．たとえば，「きくみ」は重心から遠く，「ロゴスキー」「ムガール」などは，比較的重心に近いことが分かる（図 3 の布置図で確認しよう）．

4．1．3 成分スコアの布置図の観察

ここで，図 3 にレストランの**成分スコアの布置図**を．また図 4 に，列側選択肢の 3 つの評価基準に対する成分スコアとの**同時布置図**を示した．この図から，各レストランと各評価基準の関連（対応）が読みとれる．

いまここでは，行の側に注目し**“レストランのクラスター化”**を行うので，各レストランがどのような布置関係にあるかを覚えておこう．のちにクラスター化過程を確認する際にもこれを用いることにする．なおこの例では，最大次元数（成分数）が $K=2$ であるから，データ表の全情報がこの 2 次元空間内に布置されることに注意しよう（図 1，表 6 から，2 つの成分で寄与率が 100%）．一般には，かなり寸法の大きいデータ表を扱うので，大きな固有値（高い寄与率）とはならず，布置図に描ける情報の量は限られる（それを寄与率が示している）．

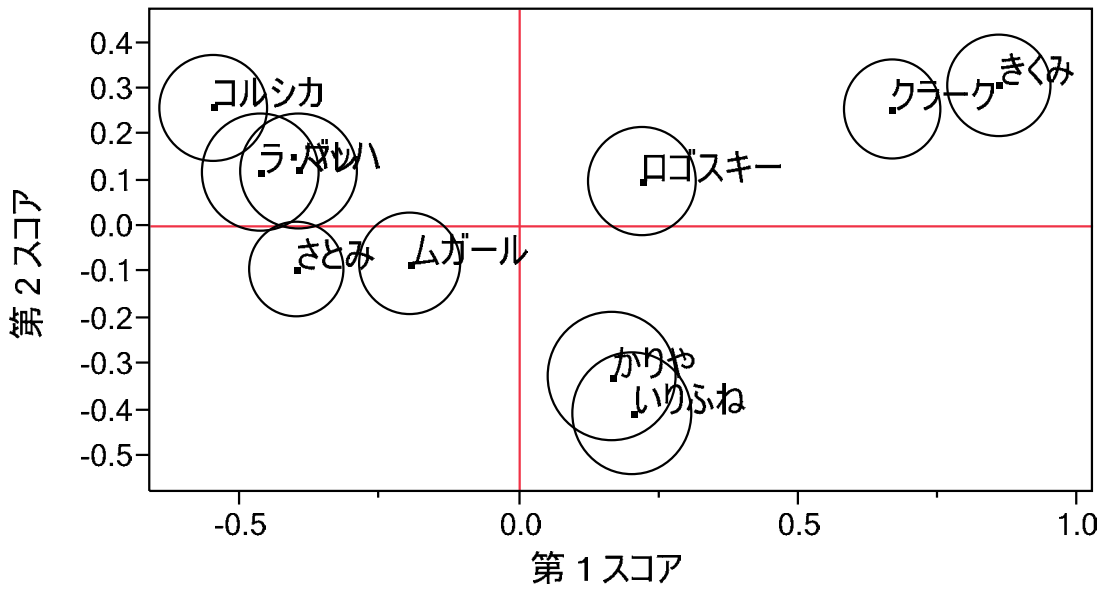


図 3　レストランの成分スコアの布置図

[$\lambda_1 = 0.1977(76.7\%), \lambda_2 = 0.060(23.3\%)$]（固有値と寄与率）

(*)ここで，図の縦横比を成分スコアの分散（固有値）の大きさに合わせた

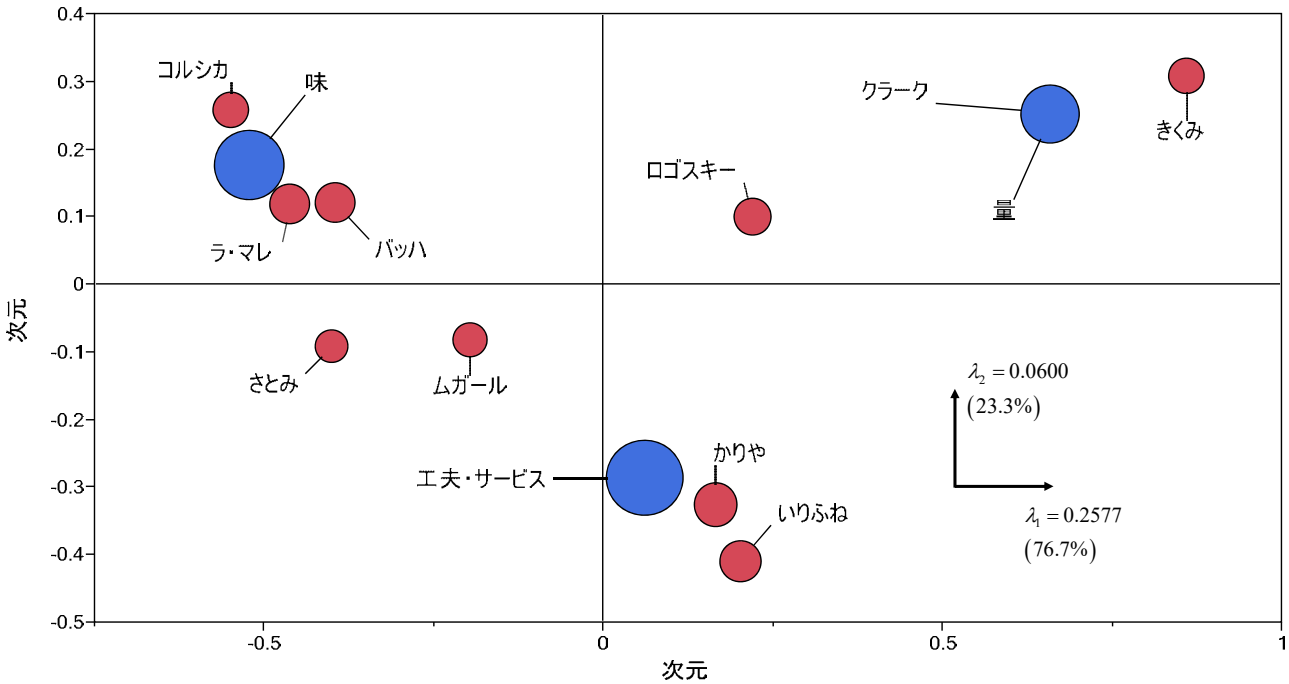


図 4　レストランと評価基準の成分スコア同時布置図

(*)ここでも，縦横比を成分スコアの分散（固有値）の大きさに合わせた

ところで，すでに述べたように，元のクロス表におけるプロファイル間の平方カイ二乗距離（ストレッチ・プロファイル間の平方距離）は，対応分析で得られた成分スコアのユークリッド距離に同等である．これについて「第Ⅱ部」では式で示したので，ここでは以下の[性質 3]として要約し，レストランのデータで数値的に確認しよう．

[性質 3]

対応分析で得た成分スコア間のユークリッド距離（平方ユークリッド距離）と，元のクロス表のプロファイル間の（平方）カイ二乗距離について，次の重要な関係がある．

［対応分析で得た（レストランの）**成分スコア間の平方ユークリッド距離**］
＝［元のクロス表の（レストランの）**プロファイル間の平方カイ二乗距離**］（17）

つまり，

$$d_E^2(i,i') = \sum_{k=1}^{K}\left(z_{ik} - z_{i'k}\right)^2 \text{ は，} \quad d_B^2(i,i') = \sum_{j=1}^{n}\frac{1}{p_{+j}}\left(\frac{p_{ij}}{p_{i+}} - \frac{p_{i'j}}{p_{i'+}}\right)^2 \text{ に等しい．} \quad (17-1)$$

同様に，列の側についても以下がなり立つ．

$$d_E^2(j,j') = \sum_{k=1}^{K}\left(z_{kj}^* - z_{kj'}^*\right)^2 \text{ は，} \quad d_B^2(j,j') = \sum_{i=1}^{m}\frac{1}{p_{i+}}\left(\frac{p_{ij}}{p_{+j}} - \frac{p_{ij'}}{p_{+j'}}\right)^2 \text{ に等しい．} \quad (17-2)$$

表 9　成分スコアとプロファイルの距離の一覧（確認用）

	対応分析から得た成分スコア（2 成分）（*）図 2 から		行のプロファイル（3 つの評価基準のパターン）（*）表 4 から：p_{ij}		
レストラン名	成分スコア 1	成分スコア 2	工夫・ｻｰﾋﾞｽ	味	量
バッハ	-0.3966	0.1220	0.3380	0.5352	0.1268
ラ・マレ	-0.4636	0.1191	0.3356	0.5616	0.1027
距離の比較	平方ユークリッド距離＝**0.004497** ユークリッド距離＝**0.06706273**		平方カイ二乗距離＝**0.00449669** カイ二乗距離＝**0.06705736**		
いりふね	0.2017	0.4082	0.6323	0.1613	0.2065
かりや	0.1647	0.3261	0.5899	0.1966	0.2135
距離の比較	平方ユークリッド距離＝**0.008109** ユークリッド距離＝**0.09005226**		平方カイ二乗距離＝**0.00810677** カイ二乗距離＝**0.09003759**		
			列和の比率（行の重心プロファイル）		
			0.4206	0.3442	0.2352

例 1:「バッハ」と「ラ・マレ」

ｉ）成分スコアからユークリッド距離を求める

まず，バッハとラ・マレの“平方ユークリッド距離”は次にようになる．

$$d_E^2(\text{バッハ，ラ・マレ}) = \left[-0.3966-(-0.4636)\right]^2 + \left[0.1220-0.1191\right]^2 = 0.004497$$

よって，“ユークリッド距離”は次のようになる．

$$d_E(\text{バッハ，ラ・マレ}) = \sqrt{\left[-0.3966-(-0.4636)\right]^2 + \left[0.1220-0.1191\right]^2} = \sqrt{0.004497} = 0.06706273$$

ⅱ）次に，行のプロファイルと行の平均ベクトルから，同じレストランのプロファイル間の“平方カイ二乗距離”と“カイ二乗距離”を求める．

$$d_B^2(\text{バッハ，ラ・マレ}) = \frac{(0.3380-0.3356)^2}{0.4206} + \frac{(0.5352-0.5616)^2}{0.3442} + \frac{(0.1268-0.1027)^2}{0.2352}$$
$$= 0.00449669$$

よって，カイ二乗距離は以下となる．

$$d_B(\text{バッハ，ラ・マレ}) = \sqrt{\frac{(0.3380-0.3356)^2}{0.4206} + \frac{(0.5352-0.5616)^2}{0.3442} + \frac{(0.1268-0.1027)^2}{0.2352}}$$
$$= 0.06705736$$

例 2:「いりふね」と「かりや」

ｉ）成分スコア間の“平方ユークリッド距離”は以下となる．

$$d_E^2(\text{いりふね，かりや}) = (0.2017-0.1647)^2 + (0.4082-0.3261)^2 = 0.008109$$

よって，“ユークリッド距離”は以下となる．

$$d_E(\text{いりふね，かりや}) = \sqrt{(0.2017-0.1647)^2 + (0.4082-0.3261)^2} = \sqrt{0.008109} = 0.09005226$$

ⅱ）行プロファイル間の“平方カイ二乗距離”は以下となる．

$$d_B^2(\text{いりふね，かりや}) = \frac{(0.6323-0.5899)^2}{0.4206} + \frac{(0.1613-0.1966)^2}{0.3442} + \frac{(0.2065-0.2135)^2}{0.2352} = 0.00810677$$

よって，“カイ二乗距離”は以下となる．

$$d_B(\text{いりふね，かりや}) = \sqrt{\frac{(0.6323-0.5899)^2}{0.4206} + \frac{(0.1613-0.1966)^2}{0.3442} + \frac{(0.2065-0.2135)^2}{0.2352}}$$
$$= \sqrt{0.00810677} = 0.09003759$$

以上の 2 つの例を表 9 に要約した．計算誤差の範囲で，両者の値は一致しており，式（17）の[性質 3]がなり立つ．つまり，成分スコアのユークリッド距離は，プロファイルのカイ二乗距離と等しい．対応分析では，この成分スコアのユークリッド距離とカイ二乗距離の間に見られる“距離の関係”をうまく使い分けていることになる．

これは対応分析で得た「成分スコアによるクラスター化」は，元の「クロス表のカイ二乗距離」によるクラスター化を考えればよいことを示している．これを確かめることが次の課題である．

4.2 観察（その 2） 階層的分類の実行と結果の確認

つぎに，レストランの“**成分スコア**”を用いたクラスター化による分類結果と各種の統計情報を示す．[性質 3]で示したように，成分スコアにおけるユークリッド距離は，カイ二乗距離と等しい．このことから，成分スコアを用いたクラスター化は，カイ二乗距離にもとづくクラスター化に相当する．

4. 2. 1　階層的分類の結合順と結合水準の情報

無数の階層的分類手法があるが，ここでは**ウォード法**[20]を用いている．ウォード法の場合は，分類対象間の距離として，原則として**“平方ユークリッド距離”**を用いる[21]．クラスター化を行うと，まず基本情報として，クラスター化過程を示す図 5 の情報が得られる．これは階層的に分類対象（レストラン）の併合を繰り返して得られる情報である．これを説明するためには，**“デンドログラム”**（樹形図あるいは階層樹木図；dendrogram）があると分かりやすいので，これも作ってみる（図 6，図 7）．

なお，通常はかなり規模の大きいデータセットの分類を想定せねばならない．そのような場合には，デンドログラムの描画や観察が煩雑となるため，またときには計算量も多くなるので，出力描画を抑制することになる．ここではクラスター化過程を説明するために利用してみる．またうしろに，JMP スクリプトを用いて，やや規模の大きな実際の調査データを分析したデンドログラムなどの例を示した．

	クラスター数	階層水準に含まれる異なり構成要素数	階層水準に含まれる構成要素数	階層の結合水準値
9	9	2	288	0.00025
8	8	2	333	0.00052
7	7	2	212	0.00163
6	6	2	204	0.00165
5	5	3	410	0.00222
4	4	5	614	0.00973
3	3	3	458	0.01538
2	2	5	670	0.06788
1	1	10	1284	0.15842

図 5　階層的分類の結合順とその結合水準他

表 10　クラスター生成情報：図 5 の情報の再編集

結合のステップ (r)	クラスター数 (g)	階層水準に含まれるレストラン数（†）	階層水準に含まれる回答者数（‡）	階層の結合水準値 $h(r)$	階層の結合水準値の累積和 $\sum_r h(r)$
❶	9	2	288	0.0003	0.0003
❷	8	2	333	0.0005	0.0008
❸	7	2	212	0.0016	0.0024
❹	6	2	204	0.0017	0.0041
❺	5	3	410	0.0022	0.0063
❻	4	5	614	0.0097	0.0160
❼	3	3	458	0.0154	0.0314
❽	2	5	670	0.0679	0.0993
❾	1	10	1,284	0.1584	**0.2577**
				0.2577［結合水準の和＝固有値の和＝0.2577］	

（†）クラスター内に入った**レストラン数**（最後のセルが 10 クラスター）
（‡）もとのデータ表からみた，つまり**回答者が選んだレストラン数**による計数（最後のセルが 1,284 名の回答者に相当），表 3 を参照．

[20] J. Ward の提案した，クラスターの等質性基準として平方和あるいは分散を用いる階層分類法．
[21] 平方ユークリッド距離を用いるということは，クラスターの等質性基準として平方和の和あるいは分散の和を用いることに同じである．さらに，一般には組み合わせ的階層化アルゴリズムを適用するために，ウィシャート（D. Wishart）が“組み合わせ的手法”で表せることを示したアルゴリズムを用いる．通常の統計ソフトウェアに実装のクラスター化では，ユークリッド距離を用いることが多いようだ．

4. 2. 2 デンドログラム(樹形図)と階層水準

まず，以下の説明に用いるためにデンドログラムを用意した．ここではあえて 2 つのデンドログラムを作ってみた．図に見るように，これの書き方は一通りではない．つまり，ここで必要な情報はデンドログラムの見栄えではなく，クラスターが結合する順序とそのときの水準の大きさ（結合水準）とに注意し観察することである（2 つの図は，レストラン名の並び順が異なるが，結合の順序関係が同じことに注意）．

```
rank  ind. iden  dendrogram   (indices as percentages of sum of indices :        .25768   min =   .10% / max = 61.48%)

   1   .64 1-Satomi      --+
   2  3.78 3-Mugale      --*-----+
   3   .10 2-Bach        --+     !
   4   .86 9-La_Male     --*+    !
   5 61.48 5-Corsa       ---*----*-------------------------------------------------------------------------------------------+
   6  5.97 7-Rogosky     ----------+                                                                                          !
   7   .20 10-Kariya     --+       !                                                                                          !
   8 26.34 4-Irifune     --*-------*---------------------------------+                                                        !
   9   .63 6-Clark       --+                                         !                                                        !
  10 ----- 8-Kikumi      --*-----------------------------------------*--------------------------------------------------------*
```

図 6　デンドログラムの例(1)

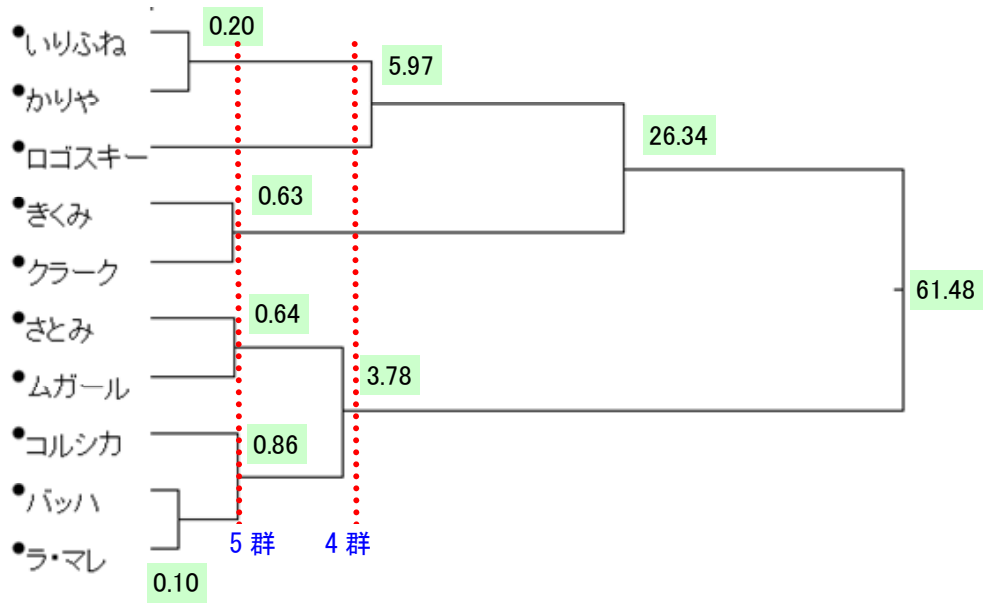
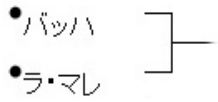


図 7　デンドログラムの例(2)
[表 17 の「割合」を結合レベルに書き入れたデンドログラム]

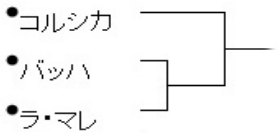
ここで重要なことは，クラスター化過程の階層の結合水準と対応分析でえた，ある情報（固有値の総和，つまりカイ二乗統計量）との間の関係である．これを調べるために，まず階層的分類が何を行うかを例で説明しよう．

図 5 あるいは表 10 の“**階層の結合水準**”（hierarchical indices）と図 7 のデンドログラムは，階層的分類の各クラスターの結合段階の状況を説明している．たとえば，もっとも近い（似ている）レストランは｛バッハ｝と｛ラ・マレ｝であって，これをはじめに併合して（merge）大きさ（クラスター・サイズ）が「2」のクラスター｛バッハ，ラ・マレ｝を作る（下の図）．



つぎに似ている 2 つのレストランからクラスター｛いりふね，かりや｝を作る．以下同じように，分類対象（レストラン）あるいは生成したクラスターの併合を“階層的に”ボトムアップに繰り返す（よって**“凝集型階層的分類”**[22]という）．この履歴が**デンドログラム**（樹形図）であり，図 5，表 10 の結合水準の履歴である．また成分スコアの布置図に階層化過程を書き入れると図のような入れ子構造（階層）となっていることや，各レストランの関係がよく分かるだろう．

また，5 群のレベルでは，上で生成したクラスター｛バッハ，ラ・マレ｝に，新たなレストラン｛コルシカ｝が併合され，大きさが「3」の｛コルシカ，バッハ，ラ・マレ｝のクラスターができる（下の図）．



以下同じように，入れ子の関係，つまり階層構造としてクラスター化が進み，最後は 1 つの群，つまり 10 のレストランすべてを包含することになる（10－1=9 回の併合で完結）．デンドログラム上のこれらの併合の様子を，成分スコアの布置図に書き入れてみると図 8 のようになる（丸数字の順に結合が進む）．これで，階層・入れ子の関係，つまり併合過程が理解されるだろう．また図の中での点の離れ具合（距離）が反映された結合順となっていることもわかる．

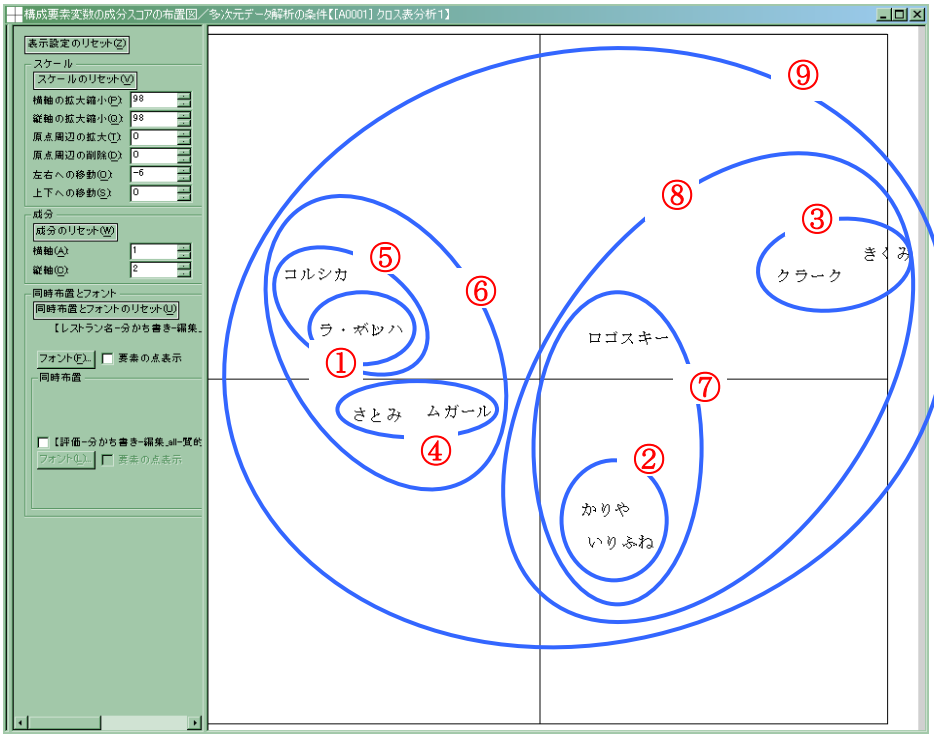


図 8　階層的分類のクラスター化過程，入れ子構造のイメージ

ところで前に“**相互最近隣の規則**”について触れた．これの簡単な例がここにみられる．いま，図 8 で，①，②，③などのクラスター生成時にこの関係がなり立っている．たとえば，｛バッハ｝からみて｛ラ・マレ｝が一番近く，またこの逆がなり立つ．つまり，こうした相互最近隣の関係にある分類対象から先にまとめれば，階層化の操作が減ることになり

[22]デンドログラムの書き方は一通りではないので，見栄えだけで判断するには注意すること．

（reducing），つまり処理速度が加速化されることが期待される[23]．とくに，もとのデータ表が疎であってしかもプロファイル間の距離が似ている，つまり成分スコアが似ているよう場合が多いようなとき，より効率的である．いまの例は小さなデータであるから，あまり効果は期待できないが，データ表の寸法が大きいときには，この相互最近隣の関係にある対象を先に併合を繰り返すことで，階層化のステップごとに次第に計算量は減ることが期待される．

4. 2. 3 階層の結合水準の意味

このとき，“**階層の結合水準**”と“**クラスター内変動**”（within-cluster variances）との間には重要な関係があるので，これを調べる．4.3 項であらためて詳しく示すが，その前準備として，まず“**クラスター内変動**”と“**クラスター間変動**”（between-cluster variances），それと“**総変動**”（全慣性：total inertia）つまり“**固有値の和**”との間の関係を示す．

記法の準備：

ここでまず，説明に必要な統計量とその記法を用意する[24]．

クラスター数： g （これは階層の水準に対応することに注意）

総変動： S_T（あらためてこの記号で表す）

クラスター内変動： $S_W(g,l)$（クラスターを g 群としたときの，そこに含まれるあるクラスター l の郡内変動，よって $l=1,2,\cdots,g$ ）

クラスター内変動の和： $\sum_{l=1}^{g} S_W(g,l)$

クラスター間変動： $S_B(g)$（クラスターを g 群としたときの群間変動）

クラスター間変動比： $\eta_g = \dfrac{S_B(g)}{S_T}\times 100$（%）（ g 群のとき）

ある階層水準（クラスター）におけるカイ二乗統計量： $\chi_p^2(g)$（クラスターを g 群としたときのピアソンのカイ二乗統計量をこれで表す）． $g=10$ （群）としたときは，分類しないとき（もとのクロス表のまま）であるから，$\chi_p^2(10)=\chi_p^2$ となる．後でみるようにクラスター数 g を変えることは**圧縮したクロス表**を調べることであり，これは**カイ二乗統計量を分解あるいは併合**することを意味する［（表 20）を参照］．

ここで総変動 S_T は，クラスター内変動とクラスター間変動の和との和であり，いったん分析対象のクロス表が与えられると値が確定する（ある一定値になる）．そして対応分析で扱うクラスター分析においては，この総変動 S_T が，カイ 2 乗値を総度数で割った値（$\phi^2=\chi_p^2/N$ ）となる．またクラスター内変動はクラスター数が増えるにつれて単調に減少する．一方，クラスター間変動は（総変動が一定であるから）クラスター数が増えるにつれて単調に増える．このように，クラスター内変動とクラスター間変動はトレードオフの関係がある．したがって“**クラスター間変動比**” η_g の変化を調べることは，クラスター化（まとまり）の程度を知る指標の 1 つとなるが，単調変化なので目安とはなってもクラスター数を決めるクリティカ

---

[23] この相互最近隣の関係を用いた階層的分類の加速化アルゴリズムが，Bruynooghe（1978），Juan（1982），Murtaugh（1985）などにより提案や検証がなされている．WordMiner でもウォード法とこれを併用している．

[24] ここでは，式の詳しい記述は省略した．

ルな基準とはならない[25].

数値例:表 10 に階層の結合水準値の和，つまり各結合におけるクラスター間変動を示した．これらの和について，式（18）の関係がなりたつ．また，式（19）は扱うデータ表の寸法が大きいときに考慮せねばならない性質である．つまり，全成分数（K）を用いずに，ある特定の成分数（$K^* < K$）までしか用いない場合には[26]，式（19）のような関係となる．

[性質 4]

階層の結合水準について，次の性質がある．

いま，クラスター化で，対応分析で得られる全成分数（K）を指定したとき，以下の関係がある（ここで，$K = \min\{m,n\} - 1$）．

［階層の結合水準値の和］＝［固有値の和］　（18）

つぎに，成分数を全成分数（K）より少ない成分数$K^*(<K)$としたとき、上の関係は以下のようになる．

［階層の結合水準値の和］＝［その指定した成分数K^*までの固有値の和］　（19）

[性質 5]

上で約束した記法によると，各統計量（総変動，クラスター間変動，クラスター内変動，ピアソンのカイ二乗統計量，固有値の和）の間に以下の関係がある．

［総変動］＝［クラスター間変動］＋［クラスター内変動の和（＝固有値の和）］　（20）

$$S_T = S_B(g) + \sum_{l=1}^{g} S_W(g,l) \quad \Leftrightarrow \quad S_T = \sum_{k=1}^{K} \lambda_k = \frac{\chi_p^2}{N} \qquad (20-1)$$

あるいは言い替えて，

［総分散］＝［クラスター間分散］＋［クラスター内分散の和］　（20－2）

あるいは，これを**慣性**により言い替えて，以下とする．

［全慣性：Total inertia］
＝［群間慣性：between-clusters inertia］＋［群内慣性：within-clusters inertia］　（20－3）

ここで，上の関係を数値例で確認する．その前準備として，上の分類で得たあるクラスターについて，まず，クラスター化における成分スコアの意味を調べる．続いて上にあげた関係式を調べる．

4. 2. 4 クラスター数を $g = 5$（群）としたときを例として

（確認 1） クラスターの成分スコアほか

ここでは「クラスター数＝5（$g = 5$）」とした場合を例として，種々の関連を調べる．ここでは，WordMiner が出力する図 9 の情報とそれを説明用に書き替えた表 11 を作った．この表

[25] これを巡るさまざまな指標の提案もあるが，決定的なものはないと思う．
[26] クラスター化に用いる成分数をいくつとするか，明確なルールがあるわけではない．固有値の大きさと寄与率などを参考に決めるのも 1 つの案である．

につけた丸番号の情報について順をおって説明する．


構成要素クラスター別統計値(変動、構成要素数、クラスター別の成分スコア他)／多次元データ解析の条件【[A0001] クロス表分析1】

	クラスター	クラスター内変動	クラスターサイズ	クラスターサイズ構成比	構成要素数	距離	成分スコア1	成分スコア2	検定値1	検定値2
1	構成要素クラスター1	0.0005	2	0.20	333	0.1658	0.1819	-0.3643	0.61	-2.23
2	構成要素クラスター2	0.0000	1	0.10	125	0.0584	0.2198	0.1002	0.49	0.41
3	構成要素クラスター3	0.0016	2	0.20	212	0.6682	0.7667	0.2835	2.59	1.74
4	構成要素クラスター4	0.0016	2	0.20	204	0.0926	-0.2919	-0.0861	-0.98	-0.53
5	構成要素クラスター5	0.0025	3	0.30	410	0.2433	-0.4660	0.1616	-2.06	1.30

構成要素クラスター別統計値(変動、構成要素数、クラスター別の成分スコア他)

図 9 「クラスター数：$g=5$（群）」としたときのクラスター要約情報

表 11　図 9 の書き替え（説明用）

①	②	③	④	⑤	⑥	⑦		⑧	
クラスター $g=5$（群） $l=1,2,\cdots,5$	クラスター内変動 $S_w(g,l)$	ｸﾗｽﾀｰ内のﾚｽﾄﾗﾝ数とその割合		ｸﾗｽﾀｰ内の回答者数	ｸﾗｽﾀｰ重心から原点までの距離	クラスターの重心		成分スコアを使った判定	
		クラスター・サイズ（レストラン）	クラスター・サイズ構成比	構成要素数（回答者）	距離（平方カイ二乗距離）	成分スコア 1	成分スコア 2	検定値 1	検定値 2
1	0.0005	2	0.2	**333**	0.1658	**0.1819**	-0.3643	0.61	**-2.23**
2	0.0000	1	0.1	125	0.0584	0.2198	0.1002	0.49	0.41
3	0.0016	2	0.2	212	0.6682	0.7667	0.2835	**2.59**	1.74
4	0.0016	2	0.2	204	0.0926	-0.2919	-0.0861	-0.98	-0.53
5	0.0025	3	0.3	**410**	0.2433	**-0.4660**	0.1616	**-2.06**	1.3
	0.0062	(10)		(1,284)					

注：ここで「「クラスター・サイズ」とは，ここでの分類対象，つまりレストランがクラスター化で分類されたクラスター内に所属したレストラン数に相当する．また構成要素」とは，元のデータ表で，個々の回答者が回答した（選んだ）「レストラン」に相当する．よって，ここで「構成要素数」とは，あるクラスターに所属する回答者数となる．

① **クラスター**：生成した 5 群のクラスターに変数名「構成要素クラスター1」，…と名前を付与し，これをあらたな変数（例：質的変数に変換したことに同じ）として利用できる[27]．

② **クラスター内変動**：これは $S_w(g,l)$ に相当する統計量である．この例では $S_w(5,1)=0.005,\cdots,S_w(5,5)=0.0025$ と対応する．つまり，**クラスター内変動**の大きさを表す．具体的には，群内の誤差平方和を，総度数（この例では，$N=1,284$）で割った値である．これが小さいほど，まとまりのよいクラスターとなる[28]．なお，クラスター・サイズが 1 個（シングルトン：singleton という）の場合，とうぜん変動はないので「0」となる．ここでは「クラスター2」がそれにあたる（$S_w(5,2)=0.0000$）．また，クラスター・サイズが複数であっても，それぞれの数値（成分スコア）がまったく同じであれば，同じくクラスター内変動は「0」となる．ちなみに，この例ではクラスター5 が，いちばんバラツキが大きい（$S_w(5,5)=0.0025$）．［図 11 をみるとわかるだろう］

③ **クラスター・サイズ**：クラスター内に所属する分類対象（ここではレストラン）の数．ここでは 5 つのクラスター内それぞれに入った分類対象の数（レストランの数）を示している．たとえば，クラスター5 は 3 つのレストランからなる．

[27] 大量の分類対象があるようなとき，クラスター化により類似した対象をグループとしてまとめて，その内容を吟味することは意味がある．たとえば，自由回答で類似した発語・語句をまとめて傾向を調べるなどが可能．

[28] もちろんここでは，平方和が小さい，つまりクラスター内の点のまとまり（等質性基準）を平方和で評価したら，という意味．つまり，クラスター評価基準は平方和だけではない，ということ．

④ **クラスター・サイズ構成比**：各クラスター内に占める分類対象の割合．クラスター・サイズがどのような分布となっているか，偏りがないか，などを観察する目安とする．
⑤ **構成要素数**：この場合は，各クラスター内に属する回答者数（サンプル数）となる．
⑥ **距離**：各クラスターの重心（セントロイド）から（布置座標の）原点までの距離（ここは**“平方カイ二乗距離”**となる）．前に「レストラン」について説明したことを，「クラスター」と読み替えればよい[29]．
⑦ **成分スコア**：対応分析で得られた成分スコアから得たクラスターの成分スコア．ここは固有値に合わせて 2 つの成分スコアがある．ここの成分スコアはクラスター内の構成要素数の**加重平均**となる（つまりクラスターの重心）．うしろに算出例を示した．
⑧ **検定値**：ある検定統計量の実現値をこう呼ぶことにする．これについてはうしろに算出方法を示した．成分スコアの有意性を検定した結果．**正規近似**で評価する．値（絶対値）が大きいほど，そのクラスターの特徴があると考える．検定値 1 は成分スコア 1 に，検定値 2 は成分スコア 2 にそれぞれ対応する．有意水準を 5%としたとき，この検定値（の絶対値）が 1.96 より大きいとこの有意水準で有意と考える．たとえば，検定値 1 ではクラスター3 の成分スコアが有意となる．同じくクラスター5 の成分スコア 1 も有意となる．

表 12　情報の要約（図 2 からの引用）

レストラン名	構成要素数 構成比	距離	λ_1 に対する 成分スコア 1	λ_2 に対する 成分スコア 2
いりふね	0.121	0.21	**0.2017**	-0.4082
かりや	0.139	0.13	**0.1647**	-0.3261
きくみ	0.086	0.83	0.8590	0.3091
さとみ	0.074	0.17	-0.4009	-0.0908
クラーク	0.079	0.51	0.6672	0.2558
コルシカ	0.095	0.37	**-0.5497**	0.2586
バッハ	0.111	0.17	**-0.3966**	0.1220
ムガール	0.085	0.05	-0.1969	-0.0821
ラ・マレ	0.114	0.23	**-0.4636**	0.1191
ロゴスキー	0.097	0.06	0.2198	0.1002

成分スコアの算出方法:
ここで，この例のクラスター数が 5 群の場合のクラスターに付与される成分スコアの算出方法を調べる．このため表 11 と表 12 を用いる．また確認の計算表を作る（表 13）．なおここで，クラスターの「成分スコア」として付与の値は，各クラスターの平均（つまり，重心）に相当する．以下でこのことを数値的に確認する（表 11，表 12，表 13）．

例 1:クラスター5 の成分スコア 1（第 1 固有値に対応）の値「-0.4660」を調べる（表 11 の⑦欄，最下段の太字）．
- このクラスターは大きさが 3 でその所属要素（メンバーシップ）は{バッハ，コルシカ，ラ・マレ}である．
- この 2 つのはじめの成分スコアは図 2（表 12 に引用）で得られているのでこれを用いる．

$$\frac{-0.5497\times122+(-0.3966)\times142+(-0.4636)\times146}{122+142+146}=\frac{-191.0622}{410}\doteqdot-0.4660$$ [30]

これで該当する成分スコア（ここでは，クラスターの重心）を得た．これを表 13 の例 1 の計算表にまとめた．

例 2:クラスター1 の{いりふね，かりや}を調べる．ここでは表 13 から表 11 のクラスター1

[29] 式（14－1），（14－2）を参照．
[30] ここは，質量を用いた加重平均という見方が分かりやすいかもしれない．

の「成分スコア」（クラスターの重心）は，以下のようになる．

$$\frac{0.2017\times155+0.1647\times178}{155+178}=\frac{60.5801}{333}\doteqdot0.1819$$

なおここで，表 12 の「構成要素数構成比」＝「（クロス表の）行の相対確率 p_{i+}」（また質量）であることに注意して，［（構成要素数構成比×成分スコア）の和］× N ÷［行和］（ここで， $N=1,284$）としても同じである．例 1 ならば，

$$\frac{\{0.095\times(-0.5497)+0.111\times(-0.3966)+0.114\times(-0.4636)\}\times1284}{122+142+146}=\frac{-191.44196}{410}\doteqdot-0.4669$$

となる．同じように例 2 についても求められる．これらの数値の前者とのズレは計算誤差の範囲である．

表 13　クラスター1，クラスター5 の成分スコア 1 の算出

例 1：クラスター1 について

レストラン名	構成要素数構成比	距離	①　λ_1 に対する成分スコア 1	②行和（表 3 の行和）	①×②
いりふね	0.121	0.21	0.2017	155	31.2635
かりや	0.139	0.13	0.1647	178	29.3166
↓　併合により下のクラスターになる				↓	↓
{いりふね，かりや}				333	60.5801
				（表 11）→	**0.1819**

例 2：クラスター5 について

レストラン名	⑤構成要素数構成比	距離	①　λ_1 に対する成分スコア 1	②行和（表 3 の行和）	①×②
コルシカ	0.095	0.37	-0.5497	122	-67.0634
バッハ	0.111	0.17	-0.3966	142	-56.3172
ラ・マレ	0.114	0.23	-0.4636	146	-67.6856
↓　併合により下のクラスターになる				↓	↓
{バッハ，コルシカ，ラ・マレ}				410	-191.0662
				（表 11）→	**-0.4660**

成分スコアの布置図の確認：

ここで，元の成分スコアとクラスター化で得たクラスターの成分スコア（クラスター重心）との関係を布置図の上で観察してみよう．はじめのクロス表からえた各レストランの成分スコアの布置と（つまり図 3，表 12 に同じ），ここで 5 群の場合に求めた表 11（⑦欄）のクラスターの成分スコア（つまり**クラスター重心**）の関係を布置図とすると図 10 のようになる．

図 10 で「構成要素クラスター1」「構成要素クラスター2」…が，各クラスターの成分スコア，つまり表 11 の⑦欄で確認した成分スコアである（文字の中央がクラスターの重心の位置）．

たとえば，上で確かめたクラスター5{バッハ，コルシカ，ラ・マレ}の中に「構成要素クラスター5」がある．他のクラスターについても，同じように観察される．また，この階層レベル（$k=5$ 群）では{ロゴスキー}はサイズが 1 個であるから元と同じ成分スコアのままとなっている．

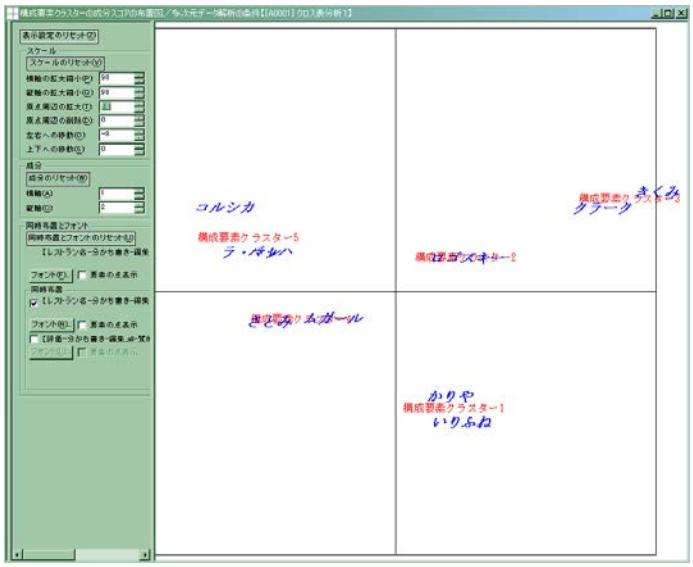


図 10　もとの成分スコアと 5 群の場合のクラスターの成分スコア（重心）の布置図

検定値の算出方法：

ここで，クラスター化で得た結果（表 11）にある“**検定値**”について，この例を用いて説明する．ここでは“**検定統計量の実現値**”のことを検定値（test values）と呼ぶことにする．この検定統計量をどのように考えているかを簡単に示し，それを数値例で確認する．

［検定統計量の考え方］

まずここでは，成分スコアを用いるので，これの性質を確認するために，以下の記号を準備する．

準備 1：

x_k：これを第k番目固有値に対応する成分スコアとする（“第k成分スコア”ということ）．このとき，対応分析で得た成分スコアの平均値は「0」である（対応分析を行った結果として，そのように調整されているということ，つまり$\bar{x}_k = 0$）[31]．

λ_k：すでに定義したように，所与のデータ表の対応分析で得た第k番目の固有値．つまり，これが第k成分スコアの“**分散**”である．

準備 2：

以上を前提に（確認として），ここでは，以下のような条件を想定している．

この検定では，分析に用いた所与のデータセットを母集団とみたて，また各クラスターを標本とみなす．そして，以下を想定する．

- （母集団からの）**単純無作為抽出**（SRS：simple random sampling）を前提とする．
- かりに，大きさ N の母集団から，無作為抽出で大きさ n の標本を抽出したとする．
- ここで標本抽出にあたって“**復元抽出**”（WR : with replacement）と“**非復元抽出**”（WOR : without replacement）がある．ここでは一般の社会調査のように“非復元抽出”とする．
- つまり，単純無作為非復元抽出（SRSWOR）を適用したとする．

このとき，母集団の平均が「0」，分散をλ_kと対応させて，ここから得た大きさがnの標本から求めた“**標本平均**”は，“平均値が「0」で，分散（Var）が次の式で与えられることは，

[31] ここで，成分スコアの平均値が「0」であり，分散が固有値 λ となることは「第Ⅱ部」で示した．

標本抽出の原理からよく知られたことである（つまり，標本平均という統計量の標本分布の平均値と分散の関係）.

$$Var = \frac{1}{n}\frac{N-n}{N-1}\lambda_k \tag{21}$$

ここで，各記号は以下を意味する[32].

N ：母集団の大きさ（ここではデータセット全体を母集団とみなす）
λ_k ：これが母集団の分散に相当すると考える（第k成分の分散）
n : 標本の大きさ

なおここで，N が，nに比べて非常に大きい場合，また復元抽出としたときは，上の式は以下のように近似される.

$$Var = \frac{1}{n}\frac{N-n}{N-1}\lambda_k \approx \frac{\lambda_k}{n} \tag{22}$$

ここで得られた第k成分の成分スコア（x_k）について，以下の“**標準化**”を行い，この統計量が，平均が「0」，分散が「1」の“**正規分布**” $N\left(0,1^2\right)$**に近似する**として検定を行う[33].

成分スコアの標準化（第k成分について）:

$$z_k^* = \frac{x_k - \bar{x}_k}{\sqrt{Var}} = \sqrt{n\frac{N-1}{N-n}}\frac{x_k}{\sqrt{\lambda_k}} \simeq N\left(0,1^2\right) \tag{23}$$

同様に，ここでN が，n に比べて非常に大きい場合，あるいは復元抽出とすると，標準化変数は以下となる.

$$z_k = \frac{x_k - \bar{x}_k}{\sqrt{Var}} = \frac{\sqrt{n}x_k}{\sqrt{\lambda_k}} \simeq N\left(0,1^2\right) \tag{24}$$

ここで，以下のように考える．かりに複数回の無作為非復元抽出を行ったとすると，それぞれは上の分布に従う．またそれぞれの標本は母集団の縮図となっているはずである．この検定統計量は，以下のような考えで導出されている．いま，母集団としたデータ全体から，非復元無作為抽出を行えば，そこから求められた検定統計量は式 (24) のような分布に従う.

一方，クラスター化で得た複数のグループは，非復元無作為抽出ではない（クラスター化という操作でえた似たものの集団）．かりにもとの母集団に特徴的なクラスターに分かれた構造がなく，ランダムに分布しているならば，個々のグループの傾向は似ているはずである．クラスター化の結果，個々のクラスター内の測定値（ここでは成分スコア）は類似し，一方，クラスター間がよく分かれているならば，個々のクラスターで得た成分スコアに，上の検定を適用してえられる検定統計量が大きな値を示していれば，それは，全体の平均からずれていることを示唆する．クラスター化が顕著であれば（分類がうまく機能すれば）大きくずれるはずである（つまり**“有意になる”**だろう）．この状況を各クラスターについて検証するこ

---

[32] ここで$\frac{N-n}{N-1}$を**“有限修正項”**という.

[33] この式をそのまま用いずに数値計算アルゴリズムによる近似計算を行うこともある (例 : WordMiner の計算法). これについてはたとえば，Ling, R. and Pratt, J.W. (1984)を参照.

とで，個々のクラスター化の程度を知る 1 つの**“目安”**とする．言い換えれば，この検定統計量は，探索的な分析において，全体の平均から外れていることを示す指標となるだろう．“目安”としたように，また上の説明から明らかだが，この指標は，相対的に個々のクラスター化の様子を知る実用的かつ**発見的に使う**ツールで，検定値の示す数値のわずかな違いを議論することではない．なお，この検定統計量を使う際に，注意すべき点として以下がある．

- 母集団と見立てるデータセットのサイズ（N）はそれなりに大きいことが必要．
- また，標本に相当するクラスター・サイズもある程度の大きさ（n）が必要であること．
- 正規近似を用いていることには限界があること．
- 言い替えると，この母集団の大きさ N や標本の大きさ n が小さいとき，とくに n が小さいときには，近似の程度があまりよくないこと[34]．
- この検定が有意であるからといって，クラスタリングが何らかの統計的意味をもつとは限らない．また，クラスターの存在を保証するものではない．かりに，まったくランダムなデータに対するクラスター化でも，その分類結果に依存して，場合によってはこの検定が有意となることがある（そういう傾向がある）．

検定値の算出例:

ここで引き続いて，いまみている 5 群の例について，検定値の求め方を説明する．またそのあとに，この検定値をどのように用いるかを説明する．再び，図 10，表 11 から必要部分を取り出し，検定値の算出に必要な作業欄を加える．

ここで，表に書き入れた丸数字にしたがって、簡単に説明する．

① **クラスター**：ここで用いる例のクラスター数（5 群，$g=5$）
② **クラスター・サイズ**:分類対象とした 10 のレストランを 5 群のクラスターに分けた結果．それぞれのレストラン数．これを検定統計量で用いる標本の大きさ（n）に対応させる．
③ **クラスター内変動**：つまりクラスター内分散 $S_w(g,l)$ のこと．
④ **成分スコア**：対応分析で得られた成分スコア．表 14 の上の表が成分スコア 1，下の表が成分スコア 2 に対応．
⑤ **無作為復元抽出**とした，式（24）から得た（理論上の推定した）検定値．
⑥ **無作為非復元抽出**とした，式（23）から得た（理論上の推定した）検定値．
⑦ **検定値**：プログラムが出力した統計量の実現値[35]．
⑧ **相対誤差**：これは参考情報で，実際には算出しない．ここは，［（⑥の絶対値）－（⑦の絶対値）］÷［⑦の絶対値］×100（%）とした．つまり，計算で得た検定値からみた相対的な誤差を評価する量．

観察のポイント:

- ここでは，n の大きさがかなり小さいから，当てはまりの程度はさほど良くない．
- 正規近似というおおまかな情報として観察する．有意水準を 5%とすると，（標準化した）検定値の絶対値が 1.96（あるいはおよそ 2）よりも大きければ“有意”と考える［$|z_k^*| \geq 1.96$（あるいは $|z_k^*| \geq 2.0$）かどうかを判定する］．
- この例で，“**形式的にこのルールを適用**”すると，成分 1 については「クラスター3」「クラスター5」あたりが有意，成分 2 については「クラスター1」が有意となる．

---

[34] 全体のデータ数（N）に比べて，クラスター数（g）を大きく与えて，あまり細かく分けると不安定になることがあるということ．クラスター数の決め方と合わせて，ここらのバランスをどう取るかの数理的な説明はむずかしい．所与のデータセットから試行錯誤的に探査するということだろう．

[35] ここで説明した式のとおりではなく，ある数値計算アルゴリズムにより求めた近似値．Ling（1984）ほか．

＜成分 1 で有意のクラスター＞
　　クラスター3＝{クラーク，きくみ}
　　クラスター5＝{バッハ，コルシカ，ラ・マレ}

＜成分 2 で有意のクラスター＞
　　クラスター1＝{いりふね，かりや}

ここで図 11 を観察しよう．10 のレストランを 5 群に分けたときにえられたクラスターのそれぞれの位置関係を観察すると，これらの関係が意味することが分かるであろう．「クラスター3 と 5」は第 1 成分の左右に位置し，「クラスター1」は，第 2 成分の下のほうに位置している（うしろに詳しく説明）．

表 14　図 10, 表 11 の書き替え（説明用）

成分 1 について

①	②	③	④	⑤	⑥	⑦	⑧
クラスター $g=5$ $l=1,2,\cdots,5$	クラスター・サイズ（n）	クラスター内変動 $S_w(g,l)$	成分スコア 1	WR から推定	WOR から推定	検定値 1	相対誤差（%）（WOR）
1	2	0.0005	0.1819	0.57855	0.57878	0.61	-5.11800
2	1	0.0000	0.2198	0.49434	0.49434	0.49	0.88539
3	2	0.0016	0.7667	2.43858	2.43953	**2.59**	-5.80959
4	2	0.0016	-0.2919	-0.92842	-0.92878	-0.98	-5.22605
5	3	0.0025	-0.4660	-1.81528	-1.81669	**-2.06**	-11.81093
	(10)						

成分 2 について

クラスター $g=5$ $l=1,2,\cdots,5$	クラスター・サイズ（n）	クラスター内変動 $S_w(g,l)$	成分スコア 2	WR から推定	WOR から推定	検定値 2	相対誤差（%）（WOR）
1	2	0.0005	-0.3643	-2.10329	-2.10411	**-2.23**	-5.6454177
2	1	0.0000	0.1002	0.40906	0.40906	0.41	-0.2281007
3	2	0.0016	0.2835	1.63679	1.63743	1.74	-5.8950424
4	2	0.0016	-0.0861	-0.49710	-0.49729	-0.53	-6.1712415
5	3	0.0025	0.1616	1.14268	1.14358	1.30	-12.032597
	(10)						

また，クラスター2 はシングルトン，つまり要素が{ロゴスキー}のみで（クラスター内変動は当然 0 であり）確かに検定値は小さくなる（元の全体の集団，母集団と差違がないという，当たり前の情報を示している）．図 10 にクラスターを書き入れた図を作り，もう一度下にあげた（図 11）．ここで上のクラスターが，どういう位置関係にあるかを検定値の評価結果と比べると，この操作がどういうことを調べたかがわかる．

たとえば，クラスター3 とクラスター5 は第 1 成分の左右の端にある．つまり“成分 1 に対する説明力”が高いということを示している（全体の平均・重心から第 1 軸にそって遠い位置にある）．一方，クラスター 1 は，第 2 軸（成分 2）の下の方に位置している．つまりそちらに向かって重心から遠く，説明力があるということになる．

さらに，残りのクラスター4 とクラスター2 は，他のクラスターに比べて中央に近く位置し（つまり全体の布置の平均・重心に近く），他のクラスターよりも相対的に説明力が弱いことを示している．

ここで注意することは，この例は成分数が少なく，全情報が 2 次元空間内に入っている，あえてそのような例を作ってある，ということである．多くの場合は，扱うデータ表の寸法は非常に大きいから，ここでみたようには明確にクラスター化と成分の関係をグラフィカルには観察できない．こういう場合には，成分軸を変えて布置図を観察し，同時に図 10，表 11 の“**成分スコアと検定値の一覧を観察**”してその傾向を慎重に探査することが有効である．

（確認 2） 各統計量の関係：各図，各表の見方

つづいて，式（18）～（20）にあげた関係がなり立つことを総合的に確認する．説明に必要な情報を，WordMiner の出力から拾い出して表 16 に要約した．この表は，各統計量（クラスター間変動，クラスター内変動，総変動）の関係を調べるための要約表である．ここにも丸数字を付けたので，これに対応させて説明する．

表 16　総変動，クラスター間変動，クラスター内変動の関係

項目 ＼ 統計量	クラスター $\begin{pmatrix} g=5 \\ l=1,2,\cdots,5 \end{pmatrix}$	⑨ 変動の大きさ	⑩ クラスター・サイズ（クラスター内のレストラン数）	⑪ 構成要素数（クラスター内のサンプル数）表 11 の⑤に同じ	⑫ 距離（重心からのカイ二乗距離の二乗）表 11 の⑥の同じ
クラスター間変動 $S_B(g)$	—	(0.2514)			
クラスター内変動 $S_W(g,l)$	1	0.0005	2	333	0.1658
	2	0.0000	1	125	0.0584
	3	0.0016	2	212	0.6682
	4	0.0016	2	204	0.0926
	5	0.0025	3	410	0.2433
クラスター内変動の和 $\sum_{l=1}^{q} S_W(g,l)$	$\left.\begin{matrix}\sum_{l=1}^{5} S_W(5,l) \\ =0.0062\end{matrix}\right\}\Rightarrow$	(0.0062) うしろの表 17 の④の $g=5$	(10) （行の要素数）	(1,284) （総和）	
総変動（全分散） S_T （固有値の和）	—	0.2577 (0.2576)			
変動比 [クラスター間変動/総変動] $\eta_g=\frac{S_B(g)}{S_T}\times 100$		0.9756 （97．6%）			

⑨ **変動の大きさ**：この欄の 5 つのクラスター内変動は表 11 の②欄に同じ数値である．これらを加えたものが“**クラスター内変動の和**”（$\sum_{l=1}^{5} S_W(g,l)=0.0062$）となる．一方，“**クラスター間変動**”として得られた値がクラスター間変動の欄にある $S_B(g)=0.2514$ である．これらを加えると式（18）～（20）のように“**総変動（全分散）**” S_T となる．これはすでに述べたように固有値の和（総変動，全慣性）に等しい．またここで，“**変動比**”（[変動比：η_g]＝[クラスター間変動/総変動]）は総変動に占める g 群のクラスター間変動の割合で，クラスターのある種の説明力を表す．

⑩ **クラスター・サイズ**：ここは各クラスター内に入る要素数（ここではレストランの数）．

⑪ **構成要素数**：これはクラスター内のサンプル数（ここでは回答者数）．表 11 の⑤欄に同じ．

⑫ **距離**：重心からの平方カイ二乗距離．表 11 の⑥に同じ．

こうして，式（19），（20）にあげた，つぎの性質が確認される．

［総変動］＝［クラスター間変動］＋［クラスター内変動の和］（＝固有値の和）

$$S_T = S_B(g) + \sum_{l=1}^{g} S_W(g,l) \quad \Leftrightarrow \quad S_T = \sum_{k=1}^{K} \lambda_k = \frac{\chi_p^2}{N}$$

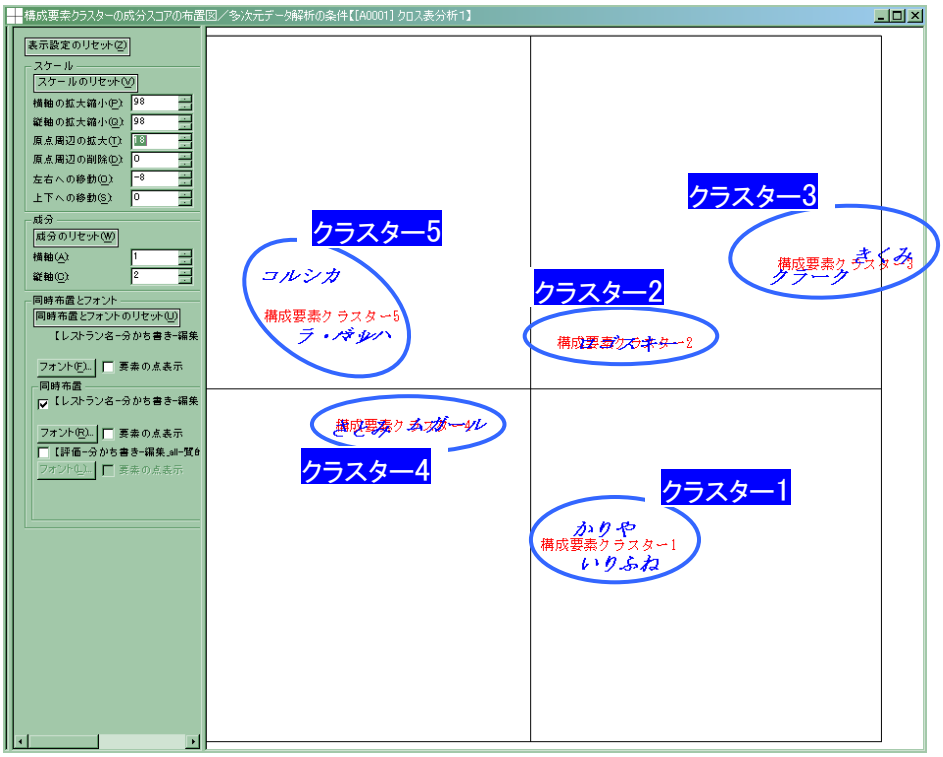


図 11　もとの成分スコアと 5 群の場合のクラスターの布置図

表 15　5 群の場合のクラスター構成

ｸﾗｽﾀｰ	クラスター・サイズ （クラスター内のレストラン数）	クラスター化履歴	工夫・ サービス	味	量
1	2	{いりふね，かりや}	203	60	70
2	1	{ロゴスキー}	48	35	42
3	2	{クラーク，きくみ}	69	22	121
4	2	{さとみ，ムガール}	91	90	23
5	3	{バッハ，コルシカ，ラ・マレ}	129	235	46

（10）

レストランを 5 群にクラスター化した場合の“**圧縮化したクロス表**”．つまり「5（群）×3（評価基準）」のクロス表．

なお，表 15 は，元のクロス表を圧縮した表である．つまり“（5 クラスター）×（3 評価基準）のクロス表”である．この圧縮したクロス表から求められたカイ二乗統計量は，クラスター間変動に総度数（N）を掛けたものに等しい．ここで得た圧縮したクロス表の総変動 $\chi_p^2(5) = 322.799$，$S_T = \sum_k \lambda_k = 0.2514 = 322.799/1284$ は，クラスター間変動と等しくなる．クラスター数を変えたとき，つまりクラスター化過程とこれら諸量との関係についてはうしろで例を用いてあらためて要約する［（表 20）を参照］．

4.3 あらためて階層的分類の階層化の意味を調べること

ここであらためてクラスター化の階層的分類の履歴，各結合水準の値の意味を説明する．ここで表 10 の出力情報から，あらためて表 17 のように情報を要約する（これは図 5，表 10 にさらに情報を追加して詳しく示したもの）．上では 5 群を例に説明したが，ここでは 1 群～10 群までの階層的分類の全履歴を観察し，そこで得られる各情報の意味を述べる．ここでもまた，表の各欄に丸数字を付与しこれにそって説明する．

① **階層水準に含まれる異なり構成要素数**：この例の場合，各クラスターレベルにおけるレストランの数となる．
② **階層水準に含まれる構成要素数**：ここでは，クロス表の総和である全回答者（1,284 名）が，併合の過程で各クラスターにどう配分されたか（併合・吸収されたか）を示す．
③ **階層の結合水準**：上で 5 群の場合について調べたように，その群における（併合・生成されたときの）“**クラスター内変動**”に相当する．つまり分類なし（クラスター数 $g=10$）のときが「0」，$g=9$ で「0.00025」，…以下分類の併合が進むにつれて単調に増えて，最後は 1 群（$g=1$）で総変動つまり元のクロス表の対応分析で得た固有値総和（慣性）となる．
④ **結合水準の累積和**：結合水準の累積履歴がこの欄の情報．よって最後のセルの値が固有値の和（0.25768）つまり慣性となる．
⑤ **総変動に占める割合**：総変動を 100 としたときの結合水準の割合（%）．図 7 のデンドログラムに書き入れた数字がこれに相当する．結合（水準）の強さのような意味をもつ．

表 17　階層の結合水準ほかの再確認

	①	②	③	④	⑤	⑥
ステップ (r)	クラスター数（g）	階層水準に含まれる異なり構成要素数	階層水準に含まれる構成要素数	階層の結合水準 $h(r)$	結合水準の累積和 $\sum_r h(r)$	総変動に占める割合(%)
	クラスターの遷移	(a) クラスターに含まれるレストランの数	(b) クラスター内のサンプル数	(c) デンドログラムで確認	(d) 各水準のクラスター内変動の和	④÷総変動（固有値総和）×100（%）
❶	9	2	288	0.00025	0.00025	0.10
❷	8	2	333	0.00052	0.00077	0.20
❸	7	2	212	0.00163	0.00240	0.63
❹	6	2	204	0.00165	0.00405	0.64
❺	5	3	410	0.00222	0.00627	0.86
❻	4	5	614	0.00973	0.01600	3.78
❼	3	3	458	0.01538	0.03138	5.97
❽	2	5	670	0.06788	0.09926	26.34
❾	1	10	1284	0.15842	**0.25768** [固有値の和]	61.48
—	—	—	—	**0.25768** [結合水準の和]	↑ ←($\lambda_1+\lambda_2$)	100.00

クラスター数の目安を得ること:

ここで，③の結合水準の変化と⑤の割合の変化を追跡して“**クラスター数 g を決める目安**”とする．たとえばこの例の場合は，4 群→5 群，あるいは 3 群→2 群の間での変化量が大きいので，クラスター数はおおよそ「4 群」あるいは「2 群」がよさそうと判断する．図 5 の出力情報から（棒グラフから）これの見当をつける．図 5 の棒グラフは階層の結合水準の変化をグラフ化したもので，おおよそ以下のように観察する．たとえば，棒グラフの変化が急に階段状に変化する位置，つまり表 17 の④（**結合水準の累積和**）あるいは⑤（**総変動に占める割**

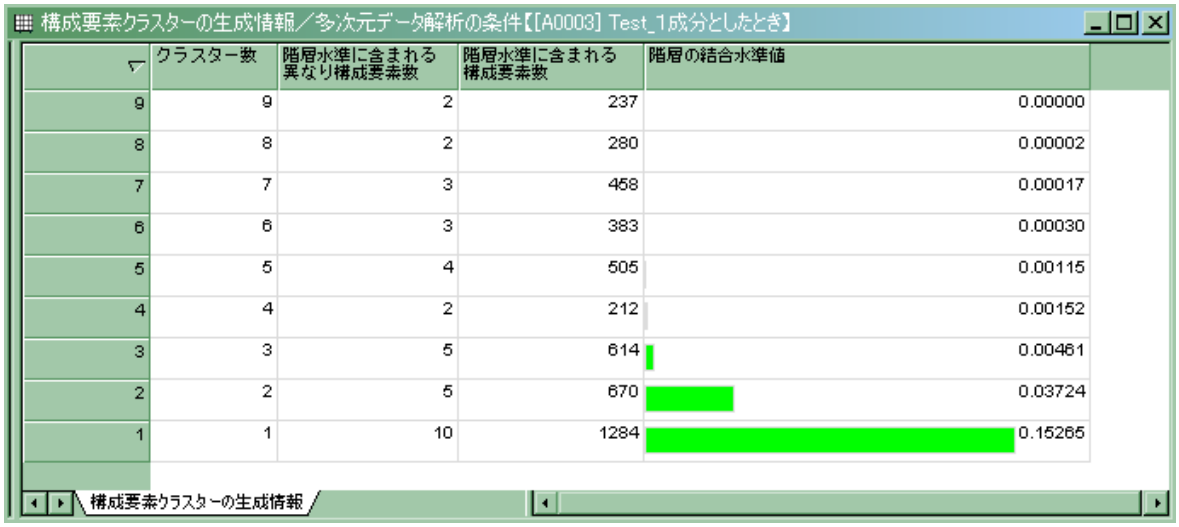
合）が大きく変化する位置を目安とする．

なおここで，③の「階層の結合水準」の関係がやや分かりにくいので，例で示そう．

例 1：5 群に結合の時点での場合を調べる（表 17 で $g=5$ に対応する行をみる）
①欄の「3」は，階層の（デンドログラム上の）第 5 回目の併合で｛バッハ，ラ・マレ｝＋｛コルシカ｝がくくられ，そのリンク数が「2＋1=3」となったことを示す．つまり，②欄のバッハ＝142，ラ・マレ＝146，コルシカ＝122 の（行和）の和が 142＋146＋122＝410（人）となったということ．

例 2：同様に，4 群と結合するときを考える（表 17 で $g=4$ に対応する行をみる）
｛バッハ，ラ・マレ｝＋｛コルシカ｝＝｛バッハ，ラ・マレ，コルシカ｝で（3）となった②「410」と，さらに①の｛さとみ，ムガール｝（2）の「204」とが併合して，（5）となり，「410＋204＝614（人）」のサイズのクラスターとなる．

重要な性質の確認（その 1）：
前に $g=5$（群）のクラスター化で得た圧縮化したクロス表（表 15）に対応分析を適用したときに得られる固有値の和，つまり**このクロス表の総変動**は，いま調べているクラスター化履歴の 5 群の“**クラスター間変動**”となる．つまり，$S_T=S_B(g)+\sum_{l=1}^{g}S_W(g,l)$ の関係から，$g=5$ に対して $S_T=S_B(5)+\sum_{l=1}^{5}S_W(5,l)$ で，表 16 から $S_B(5)=0.2514, S_T=0.2577$（一定）である．表 17 でいうと，④欄の $g=5$ に対応する値「0.00627」は，クラスター化でえた **5 群のときのクラスター内変動** $\sum_{l=1}^{5}S_W(5,l)=0.00627$ のことである．うしろの表 18 で $g=5$ に対応する欄でも確認できる．


構成要素クラスターの生成情報／多次元データ解析の条件【[A0003] Test_1成分としたとき】

	クラスター数	階層水準に含まれる異なり構成要素数	階層水準に含まれる構成要素数	階層の結合水準値
9	9	2	237	0.00000
8	8	2	280	0.00002
7	7	3	458	0.00017
6	6	3	383	0.00030
5	5	4	505	0.00115
4	4	2	212	0.00152
3	3	5	614	0.00461
2	2	5	670	0.03724
1	1	10	1284	0.15265

構成要素クラスターの生成情報

図 12　クラスター生成情報（1 成分のみを指定 $K^*=1$ のとき）

重要な性質の確認（その 2）：
ここでは，求めた成分数のすべてに対応する成分スコアを用いて分析した（つまり $K=2$ とした）．しかしここで，**“全成分数 K を指定せず”**に，これより少ないある成分数（$K^*<K$）を指定すると，その成分数（K^*）までの固有値の和が“階層の結合水準の総和”（総変動）となる（注：[性質 4]の式（19）の確認）．このことに注意しよう．

一般には，出発時の 2 元データ表の寸法が“大きい”ので，そのデータ表の対応分析の後に，サンプルのクラスター化を行うようなときには，**全成分数を用いずに**（情報量の多い）始めの方のいくつかの成分数（$K^* < K$）を指定するだろう[36]．このときは，階層の結合水準の総和（合計）は，**“その指定した成分数（K^*）までの固有値の和”**となる．

たとえば，この例では 2 つの固有値（$K=2$）があるが，クラスター化処理で「成分数=1（$K^*=1$）」と指定すると第 1 固有値の大きさ $\lambda_1=0.1977$ が**“階層の結合水準の総和”**となる．これを実際に行ってみると確かに図 12，表 18 のようになる．これは，全成分数 $K=2$ を指定して得られた表 17 に対応する情報である．かりにこのときのデンドログラムを描くと，（図 7 に対応する）この場合の**“結合レベル”**は表 17 の⑤欄となる．

表 18　図 12 の情報の要約

クラスター数（g）	①階層水準に含まれる異なり構成要素数	②階層水準に含まれる構成要素数	③階層の結合水準	④水準の累積和	⑤全変動を 100 としたときの水準の割合（%）
9	2	237	0.00000	0.00000	0.00
8	2	280	0.00002	0.00002	0.01
7	3	458	0.00017	0.00019	0.09
6	3	383	0.00030	0.00049	0.15
5	4	505	0.00115	0.00164	0.58
4	2	212	0.00152	0.00316	0.77
3	5	614	0.00461	0.00777	2.33
2	5	670	0.03724	0.04501	18.84
1	10	1284	0.15265	**0.19766**（$=\lambda_1$）	77.23（0.15265/0.19766）*100
—	（レストランに対応）	（回答者に対応）	**0.19766**（$=\lambda_1$）［結合水準の和］	—	—

4.4 クラスター化過程の総合考察

いままで説明した情報を総括し，それぞれがクラスター化の過程でどう使われているかを総合的に整理してみる．とくに，クラスター内変動，クラスター間変動，カイ二乗統計量（総変動であり固有値の和に関連），クラスター化の結合水準のそれぞれの関係が，クラスター化の中でどう利用され，何を示しているか，相互の関係はどうなるかを総合的にまとめてみる．

確認 1:

まず表 19 の情報を読み解く．すでに示した[性質 4]すなわち式（18），（19），[性質 5]すなわち式（20）ほかの関係が成立することを数値例として確認する．ここでも表内に付与した丸数字の番号に合わせて説明する．

① **クラスター間変動 $S_B(g)$**：式（20）で示したように，「総変動＝クラスター間変動＋クラスター内変動」の関係がある．クラスター間変動は，クラスター分析によって形成されたクラスターが，どの程度データの変動（クラスターの乖離度）を説明しているかを示している．前に示した「クラスター間変動÷全変動」で求められる「変動比」なども参考にして，現在のクラスターがどの程度，データの変動を説明しているかを解釈すると良いだろう．ただしこの指標は，クラスター数の変化につれて，単調に減少する．

[36] 通常のデータ解析では，データ表の寸法は，（数千～数万行）×（数千列以上）となることも珍しくはない．とくに，テキスト・マイニングなどで扱うデータ表の寸法は大きくなる．こういうときは，寄与率などを目安として，少ない成分数（$K^* < K$）を用いる．例：WordMiner のデフォルト値は $K^*=15$ 成分としてある．

② **カイ二乗統計量のチェック** $\chi_p^2(g)$：クラスター間変動にクロス表の総和，ここでは全回答者数（$N=1,284$）を乗ずると（そのクラスターの階層水準での）“カイ二乗統計量”（$\chi_p^2(g)$）となる．［クラスター間変動］×［1, 284］＝［そのクラスター階層水準での圧縮化したクロス表のカイ二乗統計量］，これを求めた欄が③である．クラスター間変動は，カイ 2 乗統計量を総度数(N=1,284)で割っている．逆に述べると，クラスター間変動に総度数を掛けると，カイ 2 乗統計量となる．

③ **カイ二乗統計量**：②の計算式で求めたカイ二乗統計量は，階層分類のある水準における圧縮したクロス表のカイ 2 乗統計量に等しい．これの確認．クラスター化の階層水準は“カイ二乗統計量の**分解**（あるいは併合）”に対応していることがわかる[37]．

④ **クラスター内変動** $S_W(g,l)$：クラスター内変動は，クラスターごとに計算される指標．たとえば 5 群であれば，5 つのクラスター内変動がある．この欄に示した数値は，それらの合計（クラスター内変動の和）である．クラスター数が 10 個の場合は（つまり未分類の場合），いずれのクラスターも，クラスター・サイズ 1 のシングルトン(singleton)であるから，クラスター内変動は 0 となる．クラスター内変動は，クラスター数の減少（増加）に伴い，単調に増加（現象）する．

⑤ **チェック**：式（20）の確認を行った結果．つまり，［クラスター間変動］＋［クラスター内変動の和］＝［固有値の和］つまり［$S_T = S_B(g) + \sum_{l=1}^{g} S_W(g,l)$］がなり立つ．

表 19　クラスター間変動，クラスター内変動，ピアソンのカイ二乗統計量の関係

		①	②＝①×1,284（s）	③クロス表から算出のとき	④	⑤チェック
ステップ（r）	クラスター数（g）	クラスター間変動 $S_B(g)$	カイ二乗統計量に相当 $\chi_p^2(g)$	カイ二乗統計量 $\chi_p^2(g)$	クラスター内変動の和 $\sum_{l=1}^{g} S_w(g,l)$	①＋④ S_T［総変動＝固有値の和］
❽	2	0.1584	203.4049	203.405	0.0992	0.2576
❼	3	0.2263	290.5692	290.564	0.0313	0.2576
❻	4	0.2417	310.3043	310.308	0.0159	0.2576
❺	5	0.2514	322.7976	322.799	0.0062	0.2576
❹	6	0.2536	325.6481	325.648	0.0040	0.2576
❸	7	0.2553	327.7667	327.767	0.0024	0.2577
❷	8	0.2569	329.8596	329.860	0.0008	0.2577
❶	9	0.2574	330.5401	330.540	0.0003	0.2577
初期	10	0.2577	330.8598	330.860	0.0000	0.2577

確認 2:

表 20 にクラスター化過程におけるこれらの各統計量の関係を要約した．ここで所与のクロス表（表 3）から出発したあと，$g=5$（群）から $g=1$（群）までの履歴を一覧にした．$g=9\sim6$（群）までは省略したが，とうぜん同じような関係がなり立つ．表 19 の下方の行から上に向かって，カイ二乗統計量，クラスター間変動，クラスター内変動がどう変化するかがわかるであろう．またこれら表 19，表 20 で，クラスター化の進行に伴う統計量（カイ二乗統計量，クラスター間変動，固有値とその和）の関係も読み取れる．

はじめに述べたように，ここで用いた 2 元データ表の寸法は小さい．しかしデータ表の寸法に関係なく，2 元表に対応分析とクラスター化を適用する際は上に述べた仕組みで統一的

[37] つまり，成分スコアの平方ユークリッド距離を用いた分類は，もとのクロス表の平方カイ二乗距離を用いた分類に同じ結果となる．対応分析は．所与のデータ表の行と列に成分スコアを付与するから，このことは行と列の同時分類を行うことにも対応する．

に処理が行われる. よって得られた統計量, 結果の解釈はここで述べた考え方が適用される.

表 20　クラスター化の履歴の要約情報

	＜はじめのクロス表＞	＜統計量と生成される圧縮化クロス表の履歴＞		
	(*) 10 群としたことに相当	工夫・サービス	味	量
10 群	いりふね	98	25	32
	かりや	105	35	38
	きくみ	35	8	67
	さとみ	42	46	7
	クラーク	34	14	54
	コルシカ	32	77	13
	バッハ	48	76	18
	ムガール	49	44	16
	ラ・マレ	49	82	15
	ロゴスキー	48	35	42
	⇓	⇓		
	$\chi_p^2 = \chi_p^2(10) = 330.860$ (クロス表から得たカイ二乗統計量)	総変動（分類前） 0.2577	0.257679×1284＝ 330.860	ｸﾗｽﾀｰ内変動の 和＝0

＜5 群に分類後の併合を以下で追跡＞

	クラスター化履歴	工夫・サービス	味	量
5 群	{さとみ，ムガール}	91	90	23
	{バッハ，コルシカ，ラ・マレ}	129	235	46
	{ロゴスキー}	48	35	42
	{いりふね，かりや}	203	60	70
	{クラーク，きくみ}	69	22	121
	⇓	⇓		
	$\chi_p^2(5) = 322.799$	ｸﾗｽﾀｰ間変動 $S_B(5) = 0.2514$	0.251401×1284 ＝322.799	ｸﾗｽﾀｰ内変動の 和＝0.0062
4 群	{さとみ，バッハ，ムガール，コルシカ，ラ・マレ}	220	325	69
	{ロゴスキー}	48	35	42
	{いりふね，かりや}	203	60	70
	{クラーク，きくみ}	69	22	121
	⇓	⇓		
	$\chi_p^2(4) = 310.308$	ｸﾗｽﾀｰ-間変動 $S_B(4) = 0.2417$	0.241673×1284 ＝310.308	ｸﾗｽﾀｰ内変動の 和＝0.0159
3 群	{さとみ，バッハ，ムガール，コルシカ，ラ・マレ}	220	325	69
	{いりふね，ロゴスキー，かりや}	251	95	112
	{クラーク，きくみ}	69	22	121
	⇓	⇓		
	$\chi_p^2(3) = 290.564$	ｸﾗｽﾀｰ間変動 $S_B(3) = 0.2263$	0.226296×1284 ＝290.564	ｸﾗｽﾀｰ内変動の 和＝0.0313
2 群	{さとみ，バッハ，ムガール，コルシカ，ラ・マレ}	220	325	69
	{いりふね，クラーク，ロゴスキー，きくみ，かりや}	320	117	233
	⇓	⇓		
	$\chi_p^2(2) = 203.405$	ｸﾗｽﾀｰ間変動 $S_B(2) = 0.1584$	0.158415×1284 ＝203.405	ｸﾗｽﾀｰ内変動の 和＝0.0992
1 群	{さとみ，バッハ，ムガール，コルシカ，ラ・マレ} {いりふね，クラーク，ロゴスキー，きくみ，かりや}	540	442	302
	$\chi_p^2(1) = 0.0$			

5. 応用事例の紹介

ここでは，ある調査研究に関連して行ったウェブ調査でえた意識調査データの一部を用いる．これを例として，おもにクラスター化法の利用方法とそれに関連のことがらを述べる．この調査研究の課題は，やや漠然としたことで，「ひとは，いわゆる“情報”をどのように捉えているだろうか」といった内容である（詳細は省略）．ここでは，この調査データを用いて，2 つの分析例を用いて，対応分析法とクラスター化法の利用方法について述べる．

5.1 調査の概要

はじめにこの調査の概要を簡単に記す．

調査テーマ：「情報に関する調査」（実験調査）
調査方式：ウェブ調査
実施期間： 2011 年 09 月 09 日 17:00 ～ 2011 年 09 月 13 日 09:00 まで
ウェブ・パネル：非公募型パネル（部分的に確率的パネル）
予想回答所要時間：約 20 分
計画標本数：766（人）［男性（412），女性（354）］
回収標本数：347（人）［男性（175），女性（172）］
有効回答率（参加率）：45.3（%）

ここで，人口統計学的変数のうち，回答者の年齢分布だけを示すと，以下のようになっている．

	サンプル数	15～19歳	20～24歳	25～29歳	30～34歳	35～39歳	40～44歳	45～49歳	50～54歳	55～59歳	60～64歳	65～69歳	この中にはない
合計	347	22	25	32	39	39	44	25	26	29	39	27	-
		6.3	7.2	9.2	11.2	11.2	12.7	7.2	7.5	8.4	11.2	7.8	-

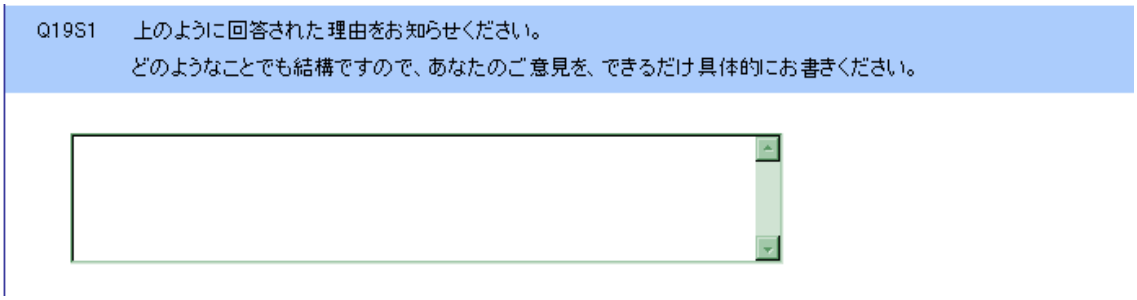
5.2 分析例（その 1）

この調査の電子調査票から，分析例（その 1）で用いる質問文につき，その選択肢型質問と自由回答質問のレイアウトをあげておく．

Q19 「情報」の考え方はいろいろありますが、「情報の送り手・発信者」と「情報の受け手・受信者」に関して、
下にあげたそれぞれの意見について、あなたはどう思われますか。
あなたのお考えにあてはまるものを、それぞれひとつずつお選びください。（ひとつずつ）

		1 非常にそう思う	2 まあそう思う	3 あまりそうは思わない	4 まったくそうは思わない
1 情報の値打ちや価値を、「送り手・発信者」の判断や考えにゆだねる時代だ	→	○	○	○	○
2 情報の値打ちや価値は、「送り手・発信者」が情報の確からしさや根拠などの裏づけを求められる時代だ	→	○	○	○	○
3 情報の値打ちや価値を、「受け手・受信者」自身が見極める能力を必要とされる時代だ	→	○	○	○	○
4 情報の値打ちや価値は、情報の「受け手・受信者」が個々人の関心や好みによって自由に価値付けすればよい時代だ	→	○	○	○	○

（1）選択肢型質問の例
図 13 電子調査票の一部（ここで用いる質問文）



(2)自由回答質問の例

図 13 電子調査票の一部(ここで用いる質問文)

＜分析に用いた選択肢型質問＞（質的変数）

Q19.「情報」の考え方はいろいろありますが、「情報の送り手・発信者」と「情報の受け手・受信者」に関して、下にあげたそれぞれの意見について、あなたはどう思われますか。
　あなたのお考えにあてはまるものを、それぞれひとつずつお選びください。(ひとつずつ)

Q19_1：情報の値打ちや価値を、「送り手・発信者」の判断や考えにゆだねる時代だ
Q19_2：情報の値打ちや価値は、「送り手・発信者」が情報の確からしさや根拠などの裏づけを求められる時代だ
Q19_3：情報の値打ちや価値を、「受け手・受信者」自身が見極める能力を必要とされる時代だ
Q19_4：情報の値打ちや価値は、情報の「受け手・受信者」が個々人の関心や好みによって自由に価値付けすればよい時代だ

＜分析に用いた自由回答質問＞（構成要素変数の元とする自由回答）

Q19S1：上のように回答された理由をお知らせください。
どのようなことでも結構ですので、あなたのご意見を、できるだけ具体的にお書きください。

ここでの自由回答質問のワーディングは，かなり漠然とした問い方に思えるかもしれない．質問 Q19 の 4 問は，回答者にとって，やや解釈・理解がむずかしいだろうと考え，続いて設けた 4 つの質問文の内容を手がかりとして回答を導く，つまり回答者に対して意図的にある情報提供を先に行ったうえで，自由回答を書いてもらうように設計してある[38]．

(1)集計結果の観察(一部)

ここで，上の 4 つの選択肢型質問への回答頻度と回答比率を要約した（表 21）．これらをみると，4 つの質問への回答傾向にそれぞれ特徴があることがみえるだろう（回答比率が 1 位と 2 位のセルをボールド表記とした）[39]．これらの回答傾向と，Q19S1 とした自由回答質問で得た内容との関連を調べることがここの目標である．

たとえば，「“Q19_4：情報の値打ちや価値は、情報の「受け手・受信者”が個々人の関心や好みによって自由に価値付けすればよい時代だ」についてみると，約 70%（21%＋49.3%）の人は「そう思う」と考え，残りの約 30%（23.6%＋6.1%）は「そうは思わない」という回答傾向にある．一方，「Q19_1：情報の値打ちや価値を、「送り手・発信者」の判断や考えにゆだねる時代だ」については，これとは逆に，約 35%（26.2%＋8.4%）の人は「そう思う」と考え，残りの約 65%（45.8%＋19.6%）は「そうは思わない」という回答している．

ではこれに回答のあと，それに続く自由回答質問への傾向がどのようなものかを，「そう思

[38] 実際，うしろで述べる「応用事例（その 2)」の回答傾向とかなり異なる．
[39] このクロス表の対応分析の結果はどうなるだろう．

う」人たちと「そうは思わない」という人たちの自由回答データと合わせて考えてみよう．

表 21　集計表　　[有効回収数：n=347（人）]

質問文	非常にそう思う	まあそう思う	あまりそうは思わない	まったくそうは思わない	合計
Q19_1．判断や考えにゆだねる時代	29	91	159	68	347
	8.4	**26.2**	**45.8**	19.6	(%)
Q19_2．情報の確からしさや根拠などの裏づけを求められる時代	117	165	47	18	347
	33.7	**47.6**	13.5	5.2	(%)
Q19_3．自身が見極める能力を必要とされる時代	211	109	21	6	347
	60.8	**31.4**	6.1	1.7	(%)
Q19_4．個々人の関心や好みによって自由に価値付けする時代	73	171	82	21	347
	21.0	**49.3**	**23.6**	6.1	(%)

(2) 用いる変数

用いる構成要素変数：

ここでは「Q19S1」の自由回答からえた**“構成要素”**[40]**（単語・語句）**を用いる．自由回答文を**分かち書き処理**のあと，簡単な語句の編集を行う．ここでは，記号，句読点，助詞，それとごく少数の語句の削除を行った（例：「とくになし」「特にない」「とくにありません」などの削除）．つまりここでは“ほとんど単語・語句の**編集を行わず**”に分析を行う．またここでは，これまでに述べたさまざまの機能をどう使うか，クラスター化の結果分析に的を絞って説明を進める．実は，日本語の特性を考えると“句読点や助詞が重要”な役割をはたすことは分かっているが，ここでは，主な利用語・発語に注目して分析を行う，ということである．また，一度しか使われない語句あるいは利用頻度が少ない語句が多いのであるが，またそれがこの種のテキスト型データの特徴であるが，ここでは，「4 語以上」登場の構成要素数を用いる（構成要素数を選ぶ**閾値**を 4 以上と指定する）．こうして確定した構成要素数は「213 語」（異なり構成要素数）である[41]．これら構成要素の分布については，下にあらためて述べる．またここで登場する構成要素の一部を，図に示した（図 15，図 16）．

用いる質的変数：

ここではまず，上に挙げた質問文のうち“Q19_4：情報の値打ちや価値は、情報の「受け手・受信者」が個々人の関心や好みによって自由に価値付けすればよい時代だ”を用いる（以下で，「Q19_4：個々人の関心や好みによって自由に価値付けすればよい時代」とする）．この質問文の選択肢は上にあるように「非常にそう思う」「まあそう思う」「あまりそうは思わない」「まったくそうは思わない」である（順序尺度）．

ここでかりに，さらに分析を進めて，構成要素変数に対応させる質的変数を，他の 3 つの質問文（Q19_1～Q19_3）と替えてみることで，4 つの質問文が，自由回答の内容とどう関連

[40]文章となった電子化テキスト型データは，分かち書き処理のあと，単語や語句の単位に分けられる．ここの要素単位は，単語（意味をなす最小単位の文字列）とは限らずいくつかがつながった複合語であることもある．そこで，分析対象とする要素単位を“構成要素”（component）と名付け，この単位で分析を行う．ここでは構成要素と語句を，同じような漠然とした意味で使いわける．

[41] この例では，総構成要素数（全語句数）は 4,702 語，同じ語句を 1 と計数した異なり構成要素数（異なる語句数）は 1,293 語（27.4%）である（図 14）．このように圧倒的に 1 語が多く，峰のないロングテイルの分布となるのが特長である．

するかを観察することが分析の目標である（事後のそのような分析場面を想定してこれら 4 つの質問文の設けた）．ここでは，“Q19_2：情報の値打ちや価値は、「送り手・発信者」が情報の確からしさや根拠などの裏づけを求められる時代だ”を使ってみよう（以下，「Q19_2：情報の確からしさや根拠などの裏づけを求められる時代」とする）．前述のように，「Q19_4」と「Q19_2」とは，4 つの選択肢に対する回答傾向がほぼ逆であることに注意しよう．

観察 1:構成要素変数の観察

ここで図 14 の**“構成要素数の頻度分布”**を観察する．ここにある 2 つの図のうち，図 14－1 は，全構成要素の分布である．ここでは，自由回答に登場した構成要素（語句，単語）の出現頻度が「1」以上，つまり**全構成要素数**が 4,720（語）あり，**異なり構成要素数**が 1,293（語）ある．これが全体に占める割合（異なり構成要素率）が，1,293÷4,720＝27.4（%）ということである[42]．以下，同じように，頻度 2 以上，頻度 3 以上，…と続く．この構成要素の頻度分布は，かならず峰のない指数的に低減する，しかも大抵は**裾の長い（ロングテイルの）分布**となる．

つぎに，図 14－2 を観察する．これは，閾値を 4 と指定して，つまり出現頻度が 3 以下の構成要素は除外して得られた頻度分布である．ここでは頻度 4 以上の全構成要素数（3,323 語）に対して**“異なり構成要素数”**が 213 語であり，その割合が 6.4%にあたる．ここで上の表（図 14－1）の同じ頻度 4 に相当の，3,350 語と 217 語（6.5%）と合っていない．この理由は，閾値で篩にかけたことで該当する回答者数（サンプル数）に違いが生じたためである．また，ここでは，この程度に絞り込んだ非常に少ない構成要素を対象に分析を行なうことに注意しよう．

また図 15 に，実際に選ばれた 213 語の構成要素の一覧，つまり回答者が記述した自由回答文から抽出・選出した語句の一覧（一部）を挙げた．これが**閾値** 4 でスクリーニングした「4 語以上」，つまり「213 語」の単語語句である．これを出現頻度の大きさに並べかえてある．図 15－1 は，構成要素数の昇順に，また図 15－2 は降順，つまり頻度数が多い方から順に選ぶと以下のようになる．

情報	する	ある	思う	**必要**	いる	**判断**	**自分**	だ	して	**受け手**
時代	**発信**	ない	その	**送り手**	もの	**正確**	では	**価値**	**見極める**	人
できる	べき	多い	**能力**	中	**発信者**	には	**正しい**	なって	思います	**責任**
それ	なる	ように	ため	なく	ならない	事	側	どれ	**氾濫**	される
しない	なので	**取捨選択**	受信者	いい	いけない	です	よって	**個人**	とは	
インターネット		**自由**	**大事**	だから	**メディア**	今	**選択**	しまう	それぞれ	どうか
よく	**意見**	何	**価値観**	**見極め**	**自分自身**	**信頼**	**力**	あり	いく	いろいろな
した	どの	**鵜呑み**	**感じる**	**受けて**	**受け手側**	重要	色々	**不正確**	（以下，続く）．	

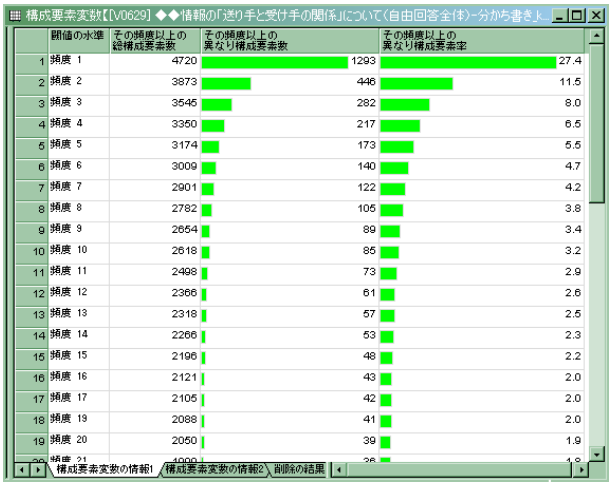
上から読むと，「情報」が 365（回），「する」が 122（回），…となり，下から読むと「あふれて」「いかなければ」「いて」「います」…と頻度 4 が続く．つまりここにみるように，ここでは 4 回以上利用された 231 種の異なる構成要素（語句）が，合わせて延べで 3,323（語）ある，ということである．

細かい分析はさておき，登場語句には，用意した質問文に含まれる語句が多数登場することに気付くであろう[43]（そうなることを想定して質問文を用意した）．実は，回答者の回答全部を並べて観察すると，これらの語句がどう結合されて発語となったかも読み取れる．しかしそれがなくても，ここに挙げた語句をつなげてみると，おおよその意見がどういう傾向に

[42] 異なり構成要素率は，回答者が記述した語句の重複の程度を示す 1 つの指標と考えられる，この値が小さいほど，同じ語句が繰り返し使われたことになる．うしろで述べる「応用事例（その 2）」の場合よりも，こちらのほうがこの割合が小さいことに注意する．これをどう考えるかは，うしろで述べる．

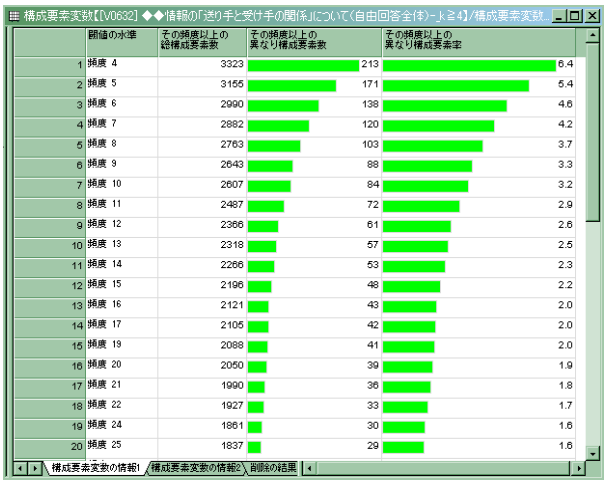
[43] この調査の他の質問文で用いた語句も多数登場する．自由回答には，こうした傾向がある．

あるかがみえてくる．ここでの観察はここまでとしよう．



	閾値の水準	その頻度以上の総構成要素数	その頻度以上の異なり構成要素数	その頻度以上の異なり構成要素率
1	頻度 1	4720	1293	27.4
2	頻度 2	3873	446	11.5
3	頻度 3	3545	282	8.0
4	頻度 4	3350	217	6.5
5	頻度 5	3174	173	5.5
6	頻度 6	3009	140	4.7
7	頻度 7	2901	122	4.2
8	頻度 8	2782	105	3.8
9	頻度 9	2654	89	3.4
10	頻度 10	2618	85	3.2
11	頻度 11	2498	73	2.9
12	頻度 12	2366	61	2.6
13	頻度 13	2318	57	2.5
14	頻度 14	2266	53	2.3
15	頻度 15	2196	48	2.2
16	頻度 16	2121	43	2.0
17	頻度 17	2105	42	2.0
18	頻度 19	2088	41	2.0
19	頻度 20	2050	39	1.9

図 14－1　1,293 語の構成要素の頻度分布（1語以上）



	閾値の水準	その頻度以上の総構成要素数	その頻度以上の異なり構成要素数	その頻度以上の異なり構成要素率
1	頻度 4	3323	213	6.4
2	頻度 5	3155	171	5.4
3	頻度 6	2990	138	4.6
4	頻度 7	2882	120	4.2
5	頻度 8	2763	103	3.7
6	頻度 9	2643	88	3.3
7	頻度 10	2607	84	3.2
8	頻度 11	2487	72	2.9
9	頻度 12	2366	61	2.6
10	頻度 13	2318	57	2.5
11	頻度 14	2266	53	2.3
12	頻度 15	2196	48	2.2
13	頻度 16	2121	43	2.0
14	頻度 17	2105	42	2.0
15	頻度 19	2088	41	2.0
16	頻度 20	2050	39	1.9
17	頻度 21	1990	36	1.8
18	頻度 22	1927	33	1.7
19	頻度 24	1861	30	1.6
20	頻度 25	1837	29	1.6

図 14－2　213 語の構成要素の頻度分布（4 語以上）

構成要素の一覧と検索

検索する構成要素変数名(V):

[V0632] ◆◆情報の「送り手と受け手の関係」について(自由回答全体)－k≧

検索文字列(I):　AND条件

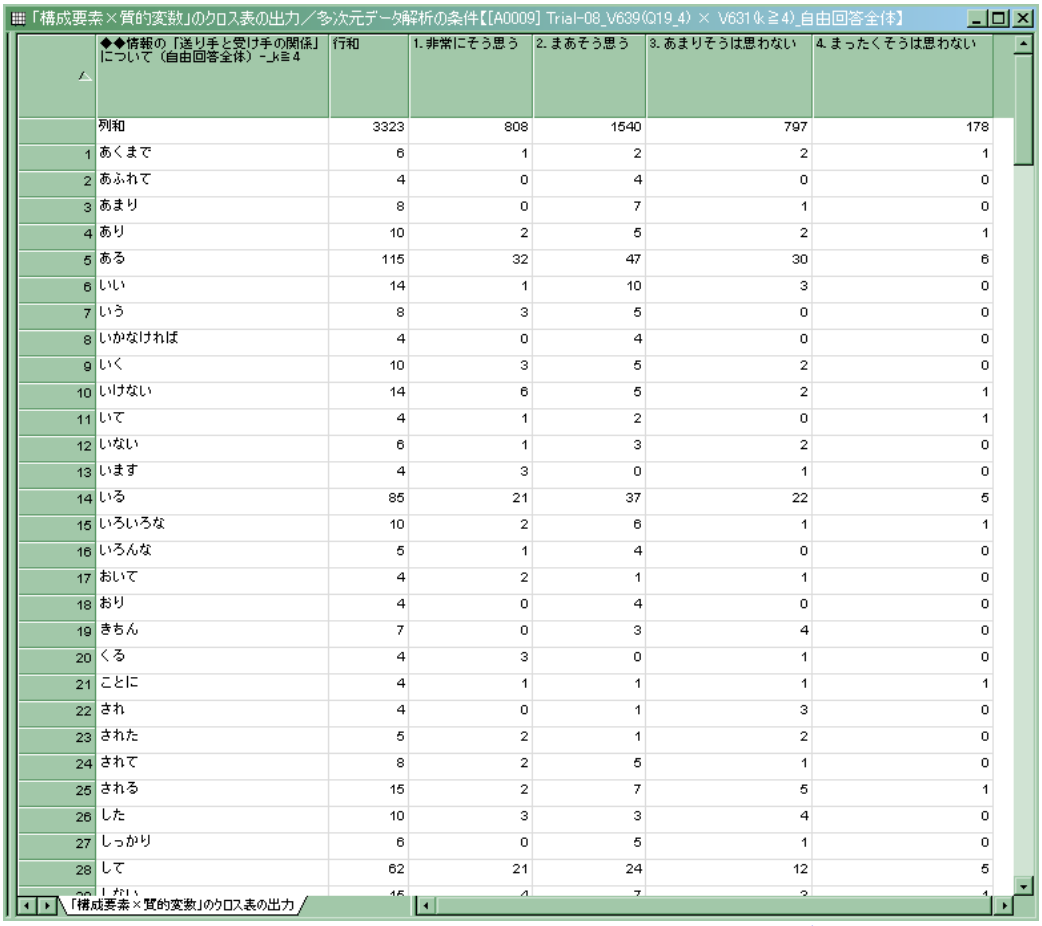
該当数:213/213 件　検索(X)　前方一致(F)　後方一致(S)

構成要素番号	構成要素	文字列長	構成要素数	サンプル度数
2	あふれて	4	4	4
8	いかなければ	6	4	4
11	いて	2	4	4
13	います	3	4	4
17	おいて	3	4	4
18	おり	2	4	4
20	くる	2	4	4
21	ことに	3	4	4
22	され	2	4	4
32	しも	2	4	4
52	とても	3	4	4
64	なら	2	4	4
67	にくい	3	4	4
71	ほしい	3	4	4
84	マスコミ	4	4	4
87	為	1	4	4
94	確かな	3	4	4
98	間違って	4	4	4
99	関して	3	4	4
100	関心	2	4	4
106	現在	2	4	4
108	限らない	4	4	4
110	個々人	3	4	4
112	好み	2	4	4
119	混乱	2	4	4
124	思った	3	4	4
134	主観	2	4	3

図 15－1　分析に用いた 213 語の構成要素一覧（頻度 4 語以上の語句），検索機能で昇順にソート

構成要素番号	構成要素	文字列長	構成要素数	サンプル度数
148	情報	2	356	213
34	する	2	122	89
5	ある	2	115	90
123	思う	2	104	94
196	必要	2	102	87
14	いる	2	85	68
192	判断	2	81	75
131	自分	2	77	66
43	た	1	63	59
28	して	2	62	49
141	受け手	3	58	53
128	時代	2	50	43
189	発信	2	46	37
57	ない	2	42	36
38	その	2	41	35
169	送り手	3	38	37
73	もの	2	37	32
163	正確	2	35	31
50	では	2	34	30
92	価値	2	34	30
105	見極める	4	33	32
158	人	1	32	28
48	できる	3	30	26
70	べき	2	28	25
175	多い	2	28	28
188	能力	2	27	26
183	中	1	26	25

図 15－2　分析に用いた 213 語の構成要素一覧，検索機能で降順にソート



	◆◆情報の「送り手と受け手の関係」について（自由回答全体）-k≧4	行和	1. 非常にそう思う	2. まあそう思う	3. あまりそうは思わない	4. まったくそうは思わない
	列和	3323	808	1540	797	178
1	あくまで	6	1	2	2	1
2	あふれて	4	0	4	0	0
3	あまり	8	0	7	1	0
4	あり	10	2	5	2	1
5	ある	115	32	47	30	6
6	いい	14	1	10	3	0
7	いう	8	3	5	0	0
8	いかなければ	4	0	4	0	0
9	いく	10	3	5	2	0
10	いけない	14	6	5	2	1
11	いて	4	1	2	0	1
12	いない	6	1	3	2	0
13	います	4	3	0	1	0
14	いる	85	21	37	22	5
15	いろいろな	10	2	6	1	1
16	いろんな	5	1	4	0	0
17	おいて	4	2	1	1	0
18	おり	4	0	4	0	0
19	きちん	7	0	3	4	0
20	くる	4	3	0	1	0
21	ことに	4	1	1	1	1
22	され	4	0	1	3	0
23	された	5	2	1	2	0
24	されて	8	2	5	1	0
25	される	15	2	7	5	1
26	した	10	3	3	4	0
27	しっかり	6	0	5	1	0
28	して	62	21	24	12	5

図 16　構成要素（213 語）×質問「Q19_4」（4 つの選択肢）の 2 元データ表（一部）

(3) 対応分析による基本情報

観察 2：データ表の確認

分析対象とする“2 元データ表”は，上に用意した「Q19_2」「Q19_4」の 2 つの質的変数と，上で整理した構成要素変数を用いる．これは，「（構成要素変数）×（質的変数）」＝「（213 語の構成要素）×（Q19_4 の 4 つの選択肢）または（Q19_2 の 4 つの選択肢）」，つまり寸法

が（213×4）の 2 元データ表となる．これの一部が図 16 である．そしてこの 2 元データ表に対応分析法を適用する．以下の説明では，WordMiner と JMP スクリプトの出力情報を引用しながら説明する．

観察 3:固有値，寄与率，累積寄与率の確認

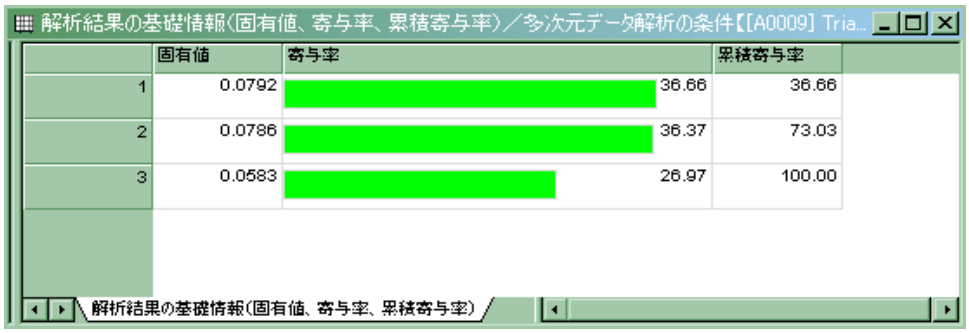


図 17　固有値，寄与率，累積寄与率の情報

ここで固有値の数は $K = \min\{213 \text{ 語}, 4 \text{ つの選択肢}\} - 1 = 3$（個）まで得られる．これを図 17 が示している．また，累積寄与率から，（形式的に読めば）はじめの 2 成分で全情報（総変動）の約 73%を占める．

観察 4:成分スコアとその布置図の観察

データ表の行側と列側の“成分スコア”を要約しよう．基本は次の 2 つの要約表である（図 18，図 19）．

図 18　構成要素(用いた単語群)の成分スコア，寄与度ほか(一部)

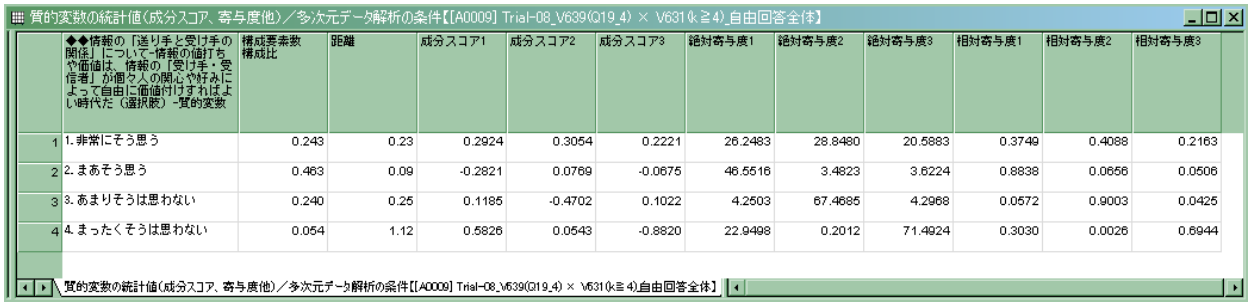


図 19　質的変数(質問「Q19_4」の 4 つの選択肢)の成分スコア，寄与度ほか

ここで，はじめの 2 成分の同時布置図をみる（寄与率を目安とすると全情報の約 73%がこの 2 次元空間内で説明できる）．ここでは 2 種の同時布置図として図 20－1，図 20－2 を用意した[44]．図 20－1，図 20－2 ともに，横軸を第 1 成分，縦軸が第 2 成分と指定した．またここでは，構成要素と質問文の 4 つの選択肢との成分スコアを**“同時布置図”**としてある．これで，質問文の 4 つの選択肢と，ここで用いた 213 語の語句とのおおまかな関連が見えてくる．細かい単語・語句の拾い出しはここでは行わないが，たとえば，「そう思う」側に特徴的な語句と，「そうは思わない」側に分布する語句があることが見える．とくに，図 20－2 のバブルプロットでは，バブルの大きさでQ19_4の4つの選択肢の回答頻度と語句の出現頻度の関係，つまり原点のあたりに「まあそう思う」「そう思う」に対応する語句が集中している．言い換えると，図の外側に分布する語句類がどのような傾向にあるかを観察することがコツである．

このとき，構成要素（単語・語句）と質問文選択肢との同時布置図で，これら両者が同じ空間にあるものとして，互いの布置の点の関係を，両者で距離として近いあるいは遠いという見方は適切ではない．ここでは成分スコアの双対性を勘案したうえで，観察することが必要である[45]．

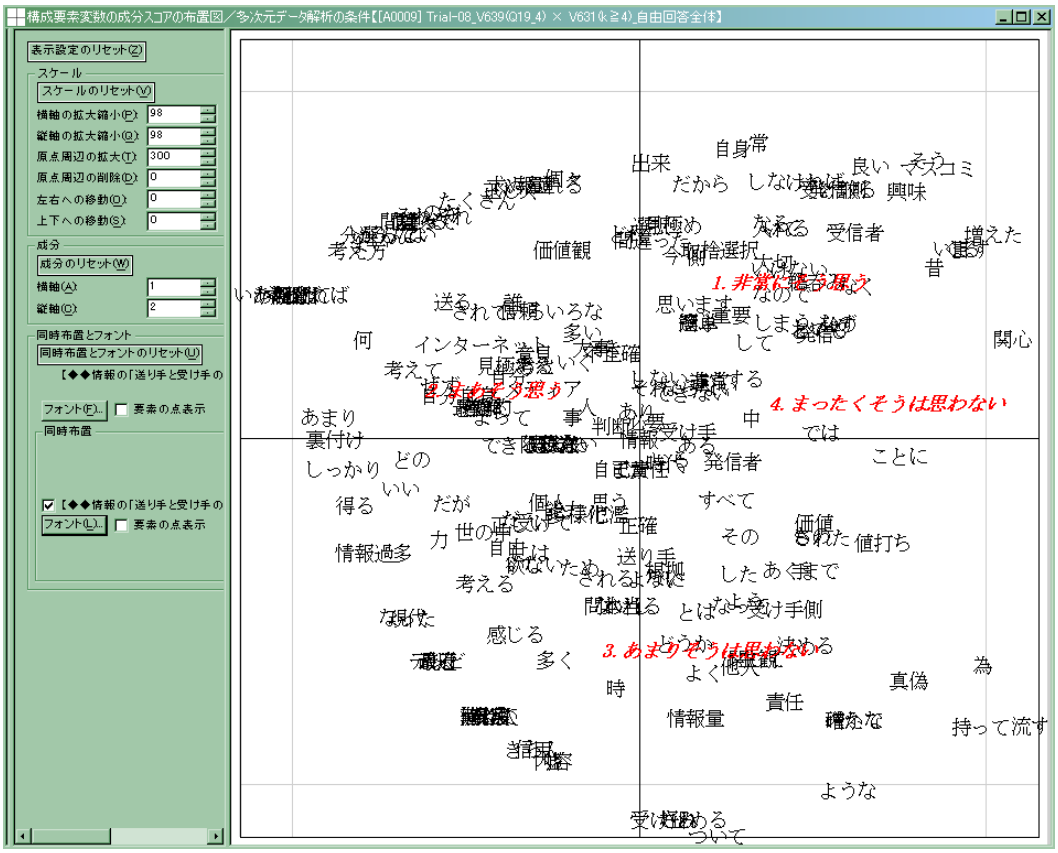


図 20－1　構成要素群と質問「Q19_4」の 4 つの選択肢の成分スコアの同時布置図

[44] 2 つの図で，成分 1（1 軸）方向の向きが逆になっている．これは，対応分析を行ってえられた固有ベクトルの符号は一意に定まらず，ここではたまたま符号が入れ替わったということである．

[45] 「第Ⅰ部」「第Ⅱ部」で同時布置図の観察の要領について述べた．

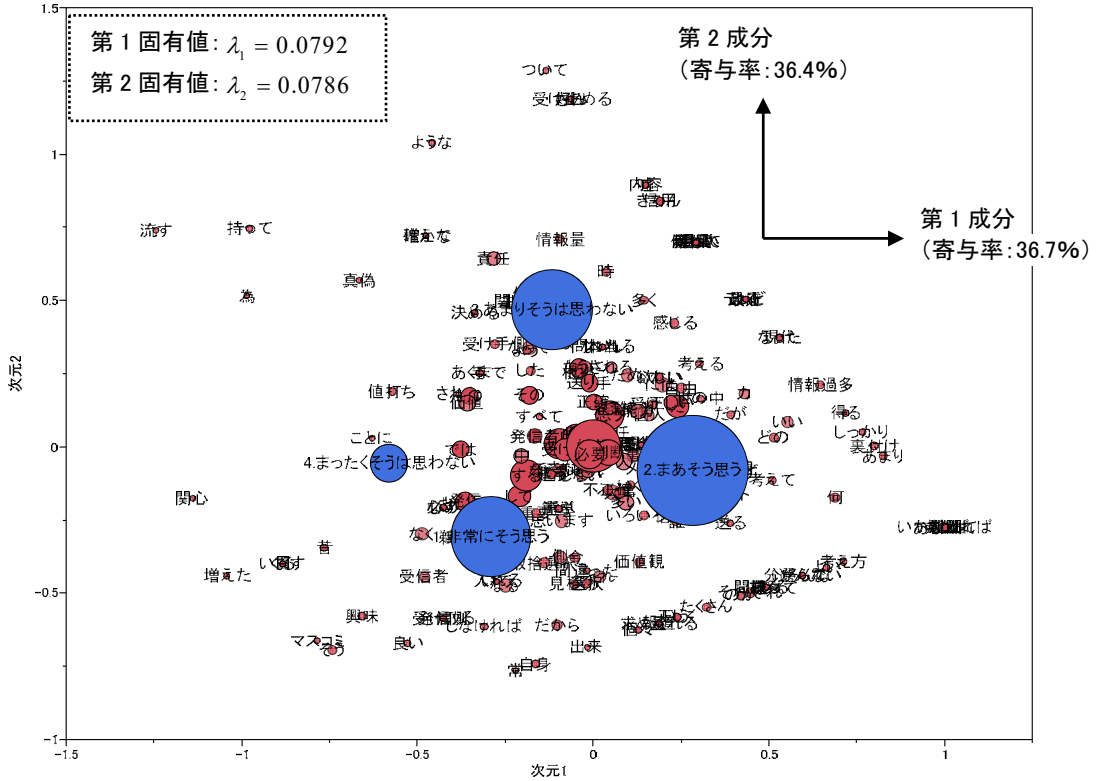


図 20－2　バブルプロットによる同時布置図（質問「Q19_4」のとき）

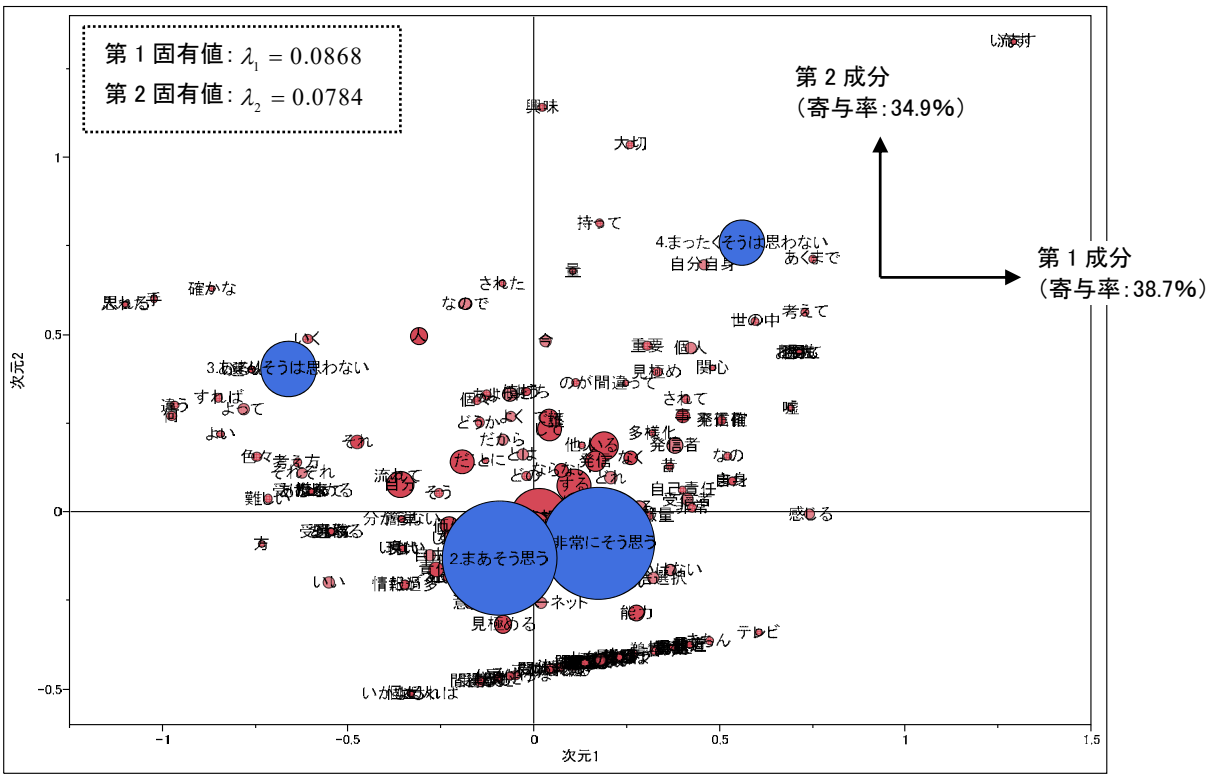


図 20－3　バブルプロットによる同時布置図（質問「Q19_2」のとき）

ではここで，もう 1 つの質問「Q19_2 : 情報の確からしさや根拠などの裏づけを求められる時代」を使うとどうであろうか．つまり，「（同じ 213 語の構成要素）×（別の Q19_2 の 4 つの選択肢）」の 2 元データ表の対応分析の結果と比べてよう．自由回答から得た 213 語の構成

要素は同じとし，それに対比させる質的変数を変えてみる．得られた同時布置図が図 20－3 である．

ここでも．Q19_4 と同様に，しかし 4 つの選択肢への回答頻度の傾向は異なるので，つまり「非常にそう思う」「まあそう思う」の頻度が多く（バブルが大きい），そのまわりに語句群が分布している．ここでは見にくいが，原点あたり（平均的な回答）に「情報」「氾濫」「正しい」「必要」「ある」「時代」…といった語句の頻度が大きいようだ（バブルが大きい）．

ここで別の視覚化情報として，**“行・列の並べ替えバブルプロット図”**を作ってみよう[46]．はじめの 2 つの成分スコアについて，（元の 2 元クロス表の）行と列との並べかえを行い，これを作ると図 20－4 となる．ここで左の図が第 1 成分スコアについて（つまり図 20－3 の横軸方向に対応），右が第 2 成分スコアについて（図 20－3 の縦軸方向に対応）の並べ替えに対応する．行側に語句の並びがある．列側の 4 つの選択肢に対して語句がどう分布するかがよくわかる（図中のバブルが大きい語句）．ここで上に指摘の語句（「情報」「氾濫」「正しい」「必要」「ある」「時代」…などが関与している様子がみえる．

ところで，第 1 成分の固有値は $\lambda_1 = 0.0792$ あるいは特異値は $\alpha_1 = \sqrt{0.0792} = 0.2814$，第 2 成分の固有値、特異値はそれぞれ $\lambda_2 = 0.0786$，$\alpha_2 = \sqrt{0.0786} = 0.2803$ となる．かりに“対の散布図”を描いたとすると，図 20－4 の左側が相関係数 $\alpha_1 = 0.2814$ に対応し，図の度数のバブルプロットの傾向は緩やかに左下から右上に向かって（線形的に）分布している様子がわかる．同じく，図 20－4 の右側もほぼ似たような傾向にある（相関係数は $\alpha_2 = 0.2803$）．この相関の大きさは，もとのデータ表の寸法（あるいは次元数）を考えるとそう小さくもなく，相応の関連があるといってよい．これは，Q19_2 についても，ほぼ似たような傾向にある．

さらにここで，語句群の分類後のデンドログラムも描いてみよう（図 20－4）．これも布置図と同様に，この程度の語句数になると結合の関係が読み取りにくい．もっとも利用頻度の多い「情報」の周辺を右の方に拡大してみたが，これでやっとこの周辺の語句の関係が読みとれるようだ（「情報」「情報過多」「氾濫」「判断」「必要」「裏付け」…）．

いずれの図（視覚化情報）も，この程度の点の数（213 語の語句と 4 つの選択肢）で，すでにかなりの単語が重なり視認がむずかしい．一般にもとの 2 元データ表の寸法はさらに大きいから，布置図や双対散布図での観察には限界がある．しかし，こうした視覚化は**初動探査の基本ツール**としては有効であり，とくに，布置図の周辺に位置した語句，つまり重心に近く平均的なプロファイルよりも，周辺の点（語句）を読み取ることは重要である．しかし，こうした視覚化ツールを基本として，さらに詳しい吟味を行う必要がある．またそのための分析ツールが必要である．ここでは別の視点から観察ツールを用意してみよう．

[46] この図は，成分スコアの並べ替えを行い元の 2 元データ表に対応させたもの，つまり行と列の並べ替えを行った図で，成分スコアのスケールに合わせた散布図とはなっていないのでバブルプロット図とした．

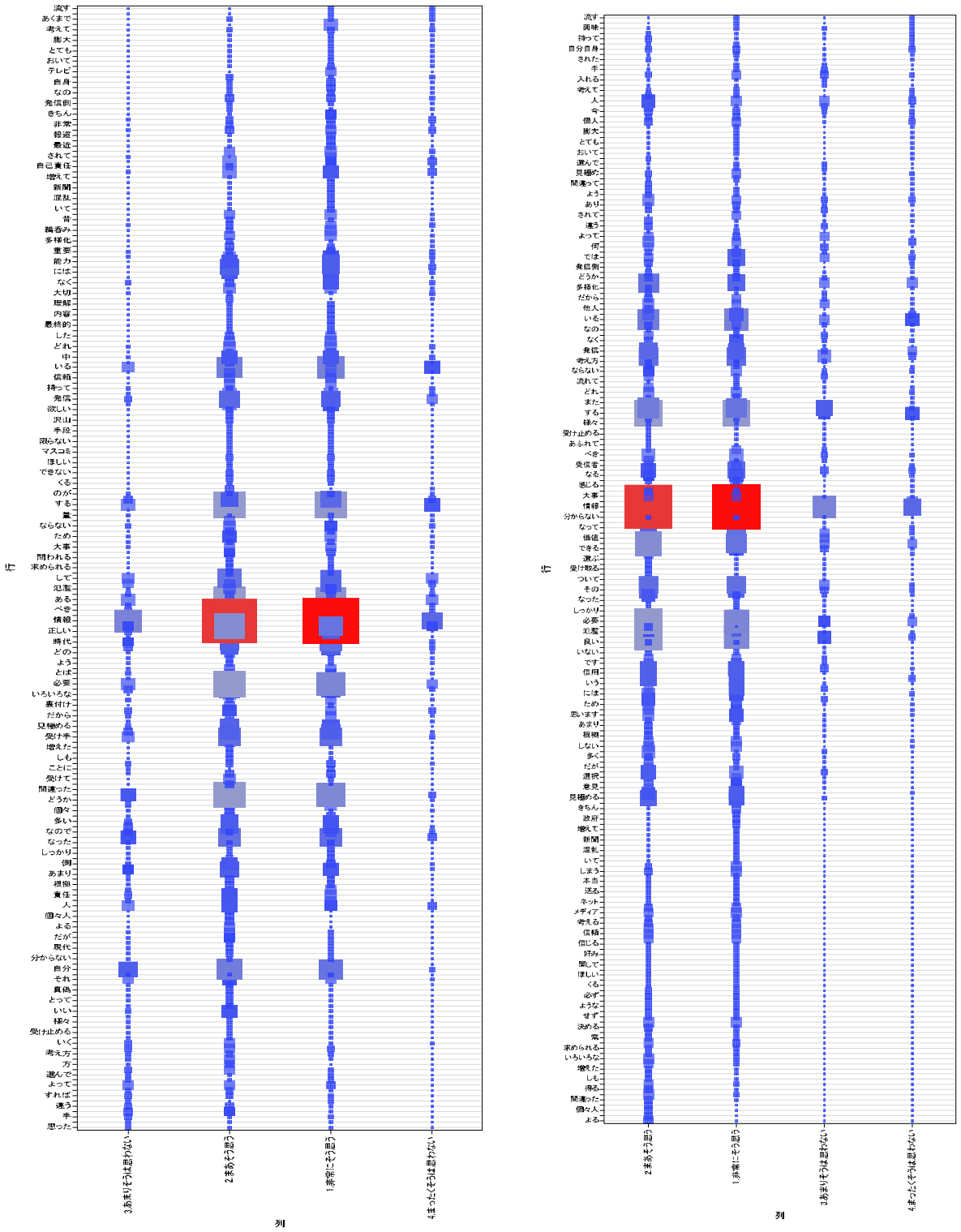


図 20－4　成分スコアの行と列の並べ替え情報
（左が第1成分，右が第2成分）

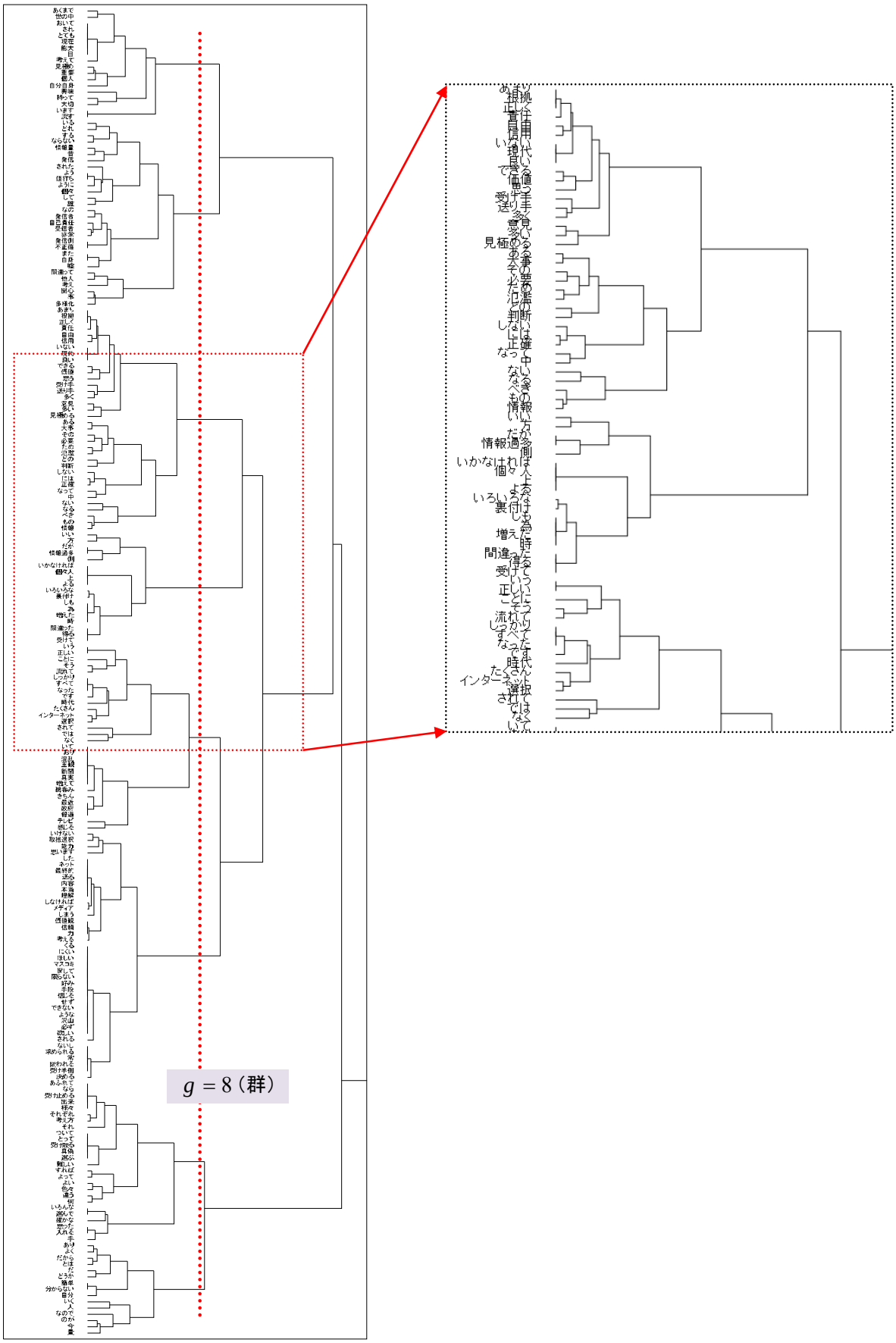


図 20－5　デンドログラムの例（一部を右に拡大）

観察 5:構成要素のクラスター化

すでに前のレストランの例でみたように，対応分析の特性を利用して，2 元データ表の行と列とのいずれについてもクラスター化を行うことができる．実際，質問 Q19_2 とクロスした 2 元データ表の「語句群」については，上の図 20－5 デンドログラムを描いてみた．

ここでは，質問 Q19_4 と構成要素（語句群）の 2 元データ表に戻って，構成要素のクラスター化過程を調べる．これは**「構成要素のクラスター生成情報」**として以下の出力情報が得られる（図 21）．ここで図の何カ所かに矢印を入れてみた．すでに述べたように，**“階層の結合水準”**とカイ二乗統計量あるいは変動（慣性）の大きさに注目して，かりにクラスターらしい構造が予想されると，クラスター化過程でクラスター間変動の変化，つまり図内の棒グラフに大きな変化（ギャップ）があるだろうと考える．厳密なルールではないが，クラスター間変動や変動比なども参考にして，1 つの目安にする．

この例のように，結合水準にやや大きく変化する部分があればよいが，かりにこの棒グラフの変化が滑らかで変化がない場合には[47]，顕著なクラスター構造が存在しないかもしれないと考える．ここらはいずれも，発見的かつ経験則的であって，理論的に厳密な考えではない．

ここでは，クラスター化履歴の図の観察と，必要ならば前に述べたクラスター間変動，クラスター内変動の和，総変動との比（変動比）．などを参考にしてクラスター数を決める．この例では，たとえば「8 群」（クラスター数：$g = 8$）としてみよう．もちろん探索的にクラスター数を替えて吟味することが望ましい．その意味で，計算を行う際に，この図を参考に複数のクラスター数を指定しておくのもよいだろう（この例では，このグラフの観察のあと．2 群から 18 群までを一括指定して再計算した）．

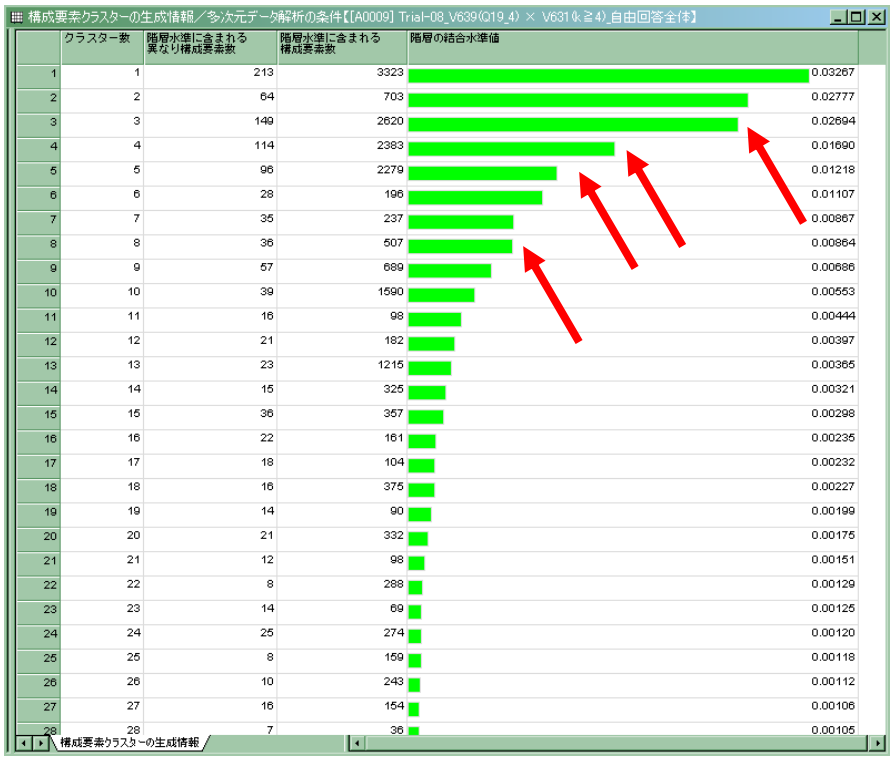


図 21　クラスター化過程の履歴の観察

[47] 大抵の場合，この変化が単調かつ滑らかに変化することが多い．つまりクラスター数の目安が付けにくいことが多い．多くの場合，（塊状，房状の）クラスターがはっきりと認められるケースは少ないということ．一方，はずれ値の影響も受けやすいので，クラスター化の手順に何らかの工夫が必要である．WordMiner で，（相互近隣関係の規則を用いる）階層的分類法のウォード法と分割化型分類法の *k*-平均法を併用するハイブリッド法を用いる理由は，はずれ値への手当の 1 つとなるからである．

観察 6:成分スコアの観察，検定値の吟味など

すでに説明した知識の助けを借りて（4．2．2 節，図 9），どのクラスターがどの成分に関連があり，またクラスター相互の関係はどうなっているか，特徴的なクラスターはどれか，といった情報を観察する．ここでの留意点は，レストランの例と違って，データ表の次元数が増えていることである．よって，各成分軸を変えて図を観察することがコツである．（図 22～24）．

たとえば，「構成要素クラスター1」のクラスター・サイズは 14，つまりここには異なり構成要素数の 14（語）の語句が所属しており，延べの構成要素数は 83（語）である．そしてそれらの成分スコアの重心がここでの成分スコアである．また，検定値をみると，いずれの成分でも値が大きいことが読み取れる．「構成要素クラスター2」については，クラスター・サイズは 18 で，第 3 成分にもっとも関与し，つぎに第 2 成分となる．クラスター・サイズがもっと大きい「構成要素クラスター5」は，その異なり構成要素数が「47」であるが，延べの構成要素数は 599 で，これは「構成要素クラスター6」の 1,741（語）よりも少ない．またその「構成要素クラスター6」は，検定値はいずれも小さく，布置図の重心あたりに位置する．このように，各統計量を目安に観察し，各成分に特徴的なクラスターを楕円で囲ってみた．

構成要素クラスター別統計値(変動、構成要素数、クラスター別の成分スコア他)／多次元データ解析の条件[[A0009] Trial-08_V639(Q19_4) × V632(k≧4)_自由回答全体]

	クラスター	クラスター内変動	クラスターサイズ	クラスターサイズ構成比	構成要素数	距離	成分スコア1	成分スコア2	成分スコア3	検定値1	検定値2	検定値3
1	構成要素クラスター1	0.0065	14	0.07	83	0.8432	0.4534	-0.5209	-0.6052	6.22	-7.18	-9.68
2	構成要素クラスター2	0.0076	18	0.08	134	0.5202	0.1053	0.3124	-0.6415	1.66	4.93	-11.76
3	構成要素クラスター3	0.0178	31	0.15	397	0.2097	0.3303	0.2951	0.1161	7.05	6.33	2.89
4	構成要素クラスター4	0.0052	21	0.10	132	0.5701	0.0015	-0.7383	0.1580	0.03	-12.68	3.15
5	構成要素クラスター5	0.0113	47	0.22	599	0.0743	-0.2659	-0.0597	0.0005	-7.32	-1.65	0.01
6	構成要素クラスター6	0.0211	46	0.22	1741	0.0063	0.0517	-0.0426	0.0425	1.40	-1.16	1.34
7	構成要素クラスター7	0.0013	15	0.07	89	0.8051	-0.8535	0.1920	-0.1997	-12.15	2.74	-3.31
8	構成要素クラスター8	0.0029	21	0.10	148	0.3264	-0.2561	0.5033	0.0863	-4.38	8.65	1.72

構成要素クラスター別統計値(変動、構成要素数、クラスター別の成分スコア他)

図 22　構成要素クラスター別の情報（成分スコア，検定値ほか）

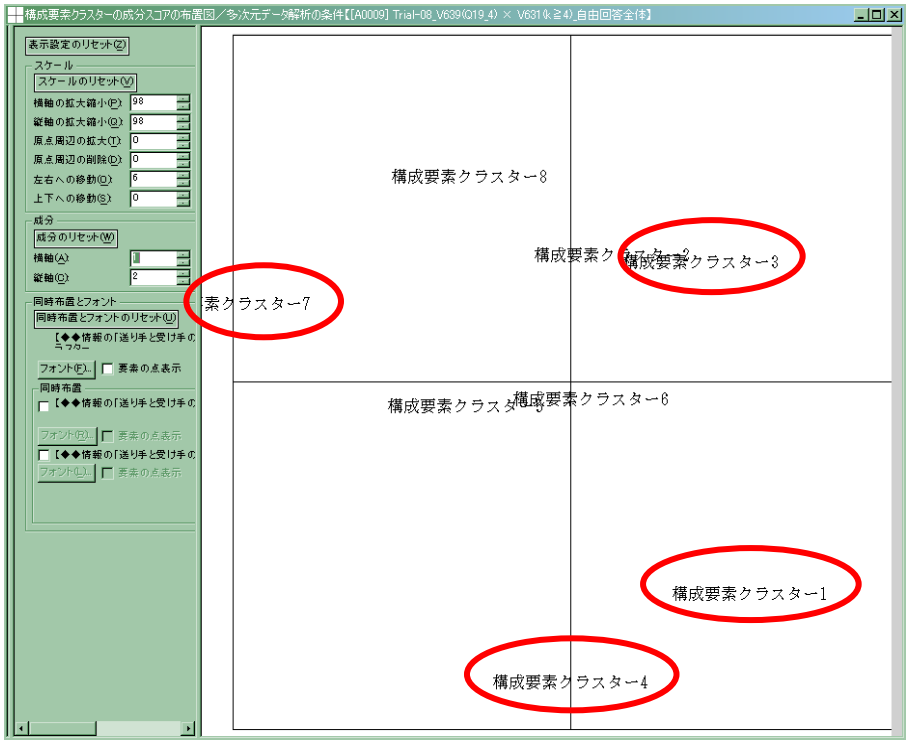


図 23　第 1 成分と第 2 成分の観察

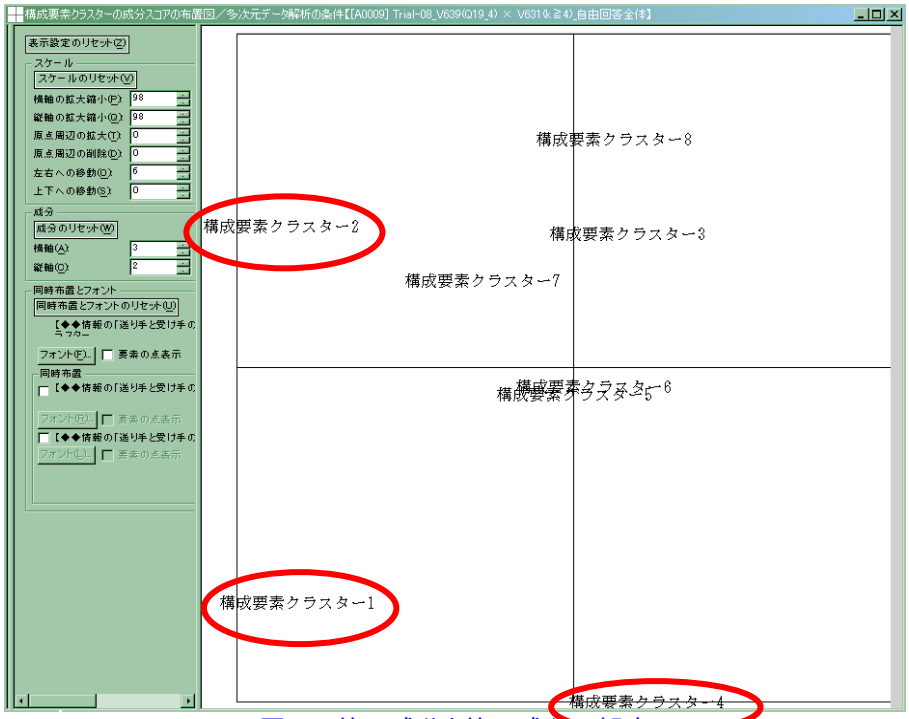
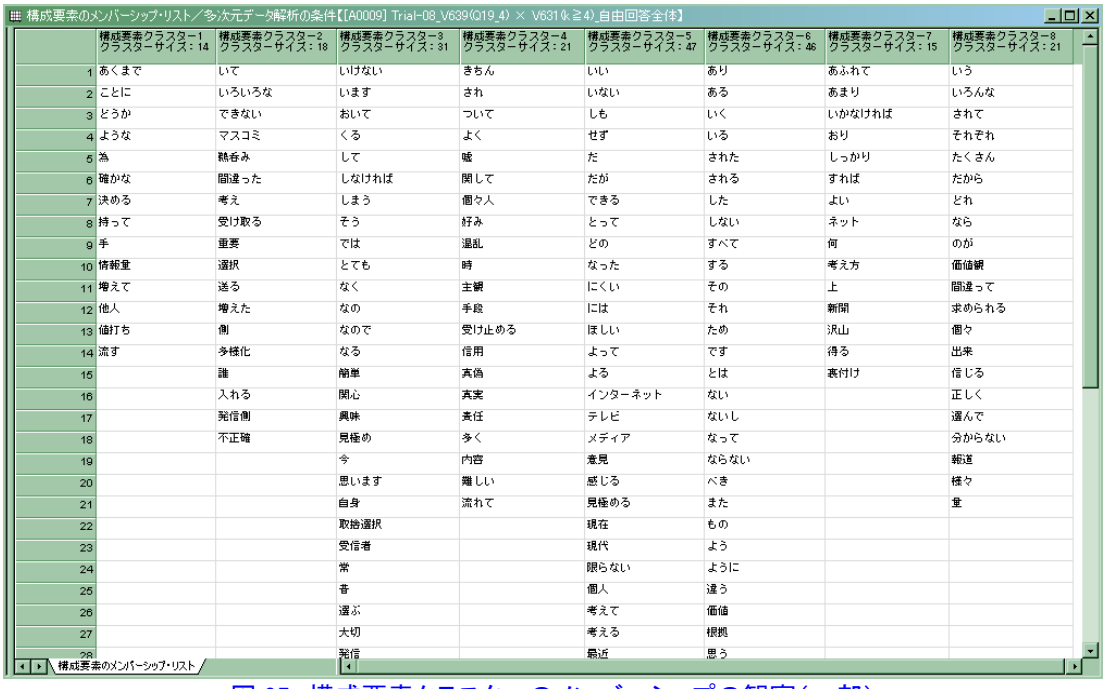


図 24　第 2 成分と第 3 成分の観察

観察 7:各クラスターのメンバーシップの確認

さらに，求めた 8 群（$g=8$）の各クラスターにはどのような構成要素（単語，語句）が含まれるのか，これを**“構成要素のメンバーシップリスト”**で観察する（図 25）．これは，デンドログラムを指定のクラスター数でカットしてえられる群に同じである．図 20−5 をみると，8 群の位置で，階層の水準が大きく変化していることもわかる．

この各クラスターのクラスター・サイズをみて，クラスター間の分布の様子と図 22 の表と併せて観察する．たとえば，クラスター1 には 14 個の語句が含まれ，その内容は図 25 の左のはじめの欄のようになる．



	構成要素クラスター1 クラスターサイズ：14	構成要素クラスター2 クラスターサイズ：18	構成要素クラスター3 クラスターサイズ：31	構成要素クラスター4 クラスターサイズ：21	構成要素クラスター5 クラスターサイズ：47	構成要素クラスター6 クラスターサイズ：46	構成要素クラスター7 クラスターサイズ：15	構成要素クラスター8 クラスターサイズ：21
1	あくまで	いて	いけない	きちん	いい	あり	あふれて	いう
2	ことに	いろいろな	います	され	いない	ある	あまり	いろんな
3	どうか	できない	おいて	ついて	しも	いく	いかなければ	されて
4	ような	マスコミ	くる	よく	せず	いる	おり	それぞれ
5	為	鵜呑み	して	嘘	だ	された	しっかり	たくさん
6	確かな	間違った	しなければ	関して	だが	される	すれば	だから
7	決める	考え	しまう	個々人	できる	した	よい	どれ
8	持って	受け取る	そう	好み	とって	しない	ネット	なら
9	手	重要	では	混乱	どの	すべて	何	のが
10	情報量	選択	とても	時	なった	する	考え方	価値観
11	増えて	送る	なく	主観	にくい	その	上	間違って
12	他人	増えた	なの	手段	には	それ	新聞	求められる
13	値打ち	側	なので	受け止める	ほしい	ため	沢山	個々
14	流す	多様化	なる	信用	よって	です	得る	出来
15		誰	簡単	真偽	よる	とは	裏付け	信じる
16		入れる	関心	真実	インターネット	ない		正しく
17		発信側	興味	責任	テレビ	ないし		選んで
18		不正確	見極め	多く	メディア	なって		分からない
19			今	内容	意見	ならない		報道
20			思います	難しい	感じる	べき		様々
21			自身	流れて	見極める	また		皇
22			取捨選択		現在	もの		
23			受信者		現代	よう		
24			常		限らない	ように		
25			音		個人	違う		
26			選ぶ		考えて	価値		
27			大切		考える	根拠		
28			発信		最近	思う		

図 25　構成要素クラスターのメンバーシップの観察（一部）

(4)構成要素と質問文の関係ほか ― WordMiner における機能 ―

ここまでに，対応分析法とその特性を用いたクラスター化法が提供する分類に関連する主な情報の観察の手順を示した．いきなり大きな寸法のデータ表の分析に取り組むのではなく，データ表の構造や分析結果がある程度みえるようなミニチュア・データ，あるいはせいぜいここでみた程度の規模のテキスト型データを用意して，クラスター化が何を行っているのかを体験するのがよいだろう．ここではさらに，WordMiner と JMP スクリプトに組み入れた独自の機能を用いた分析を試みる．

[補足]WordMiner の場合の分析機能

対応分析とその特性を活かしたクラスター化を巡るさまざまな応用機能をうまく使いこなすこと，それらを含めて総合的に分析を進めるとよい．たとえば，WordMiner では，対応分析とクラスター化のあと，処理結果がフローティング・フレーム内にすべて表示される．いまの例であると，図 26 のようなフレームが表示される．とくにここで，**「質的変数の構成要素の有意性テスト」**の項には，WordMiner の備える有用なツール群で，ここの利用方法を理解することが，対応分析法とクラスター化法を用いる上での 1 つの鍵である．

i) カテゴリー別の情報要約：ここで用いた質的変数つまり質問文の内容確認
ii) 頻度による有意性テスト要約：有意なサンプルの要約
iii) 頻度による有意性テスト要約：有意な構成要素の要約
iv) 距離による有意性テスト要約：有意なサンプルの要約
v) 頻度による有意性テスト要約：サンプル別一覧
vi) 頻度による有意性テスト要約：構成要素別一覧
vii) 距離による有意性テスト要約：サンプル別一覧

これらの情報の見方，解釈については，別の資料も用意したので，必要に応じてそれを参照していただきたい（参考文献の最後にあげたテキスト･マイニング研究会ホームページかあら）．ここでは ⅰ）と ⅲ），vi）の簡単な例を示そう．

観察 8:「カテゴリー別の情報要約」の観察

ここでまず，ⅰ）の**「カテゴリー別の情報要約」**の出力を調べる（図 27）．ここにみるように，分析に用いたサンプル数は，はじめに集計でみた（表 21 の）結果とは異なる．これは，構成要素（単語，語句）の編集や，出現構成要素数の選別，条件を満たさないサンプルの除外などの理由から，サンプル数が目減りしているからである．また，質問文の選択肢別（カテゴリー別）に，編集前後の構成要素数や異なり構成要素数，構成比率などが集計されている．全体の構成要素のうちのどの程度の構成要素を分析に用いたのか，それらが質問文にどう反映されたかなどが分かる．またこれを知っておくことはあとに続く分析結果の解釈のうえで必要である．

たとえばここで，対象となった「サンプル数」（点線枠の丸数字①）に，ある構成要素つまり回答者の記述語句（点線枠の丸数字②）は，それぞれの選択肢に対してどのような回答内容であったのか，これを次に調べよう[48]．

[48] ここで得た集計のサンプル数は（図 27 の①），表 21 の集計結果とずれている．その理由は，構成要素（単語，語句）の編集などを行ったことによるもの．

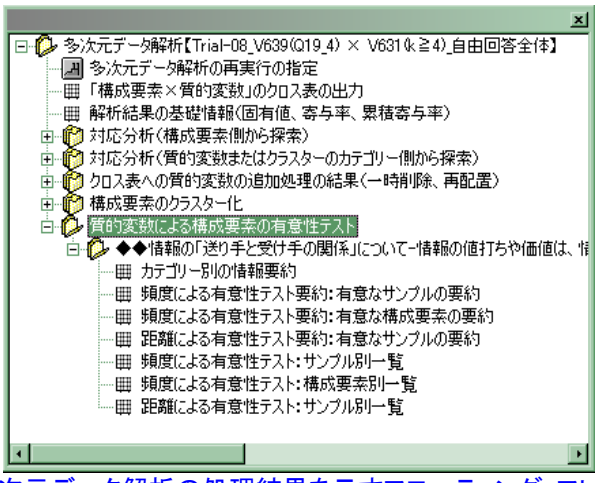


図 26　多次元データ解析の処理結果を示すフローティング・フレームの例

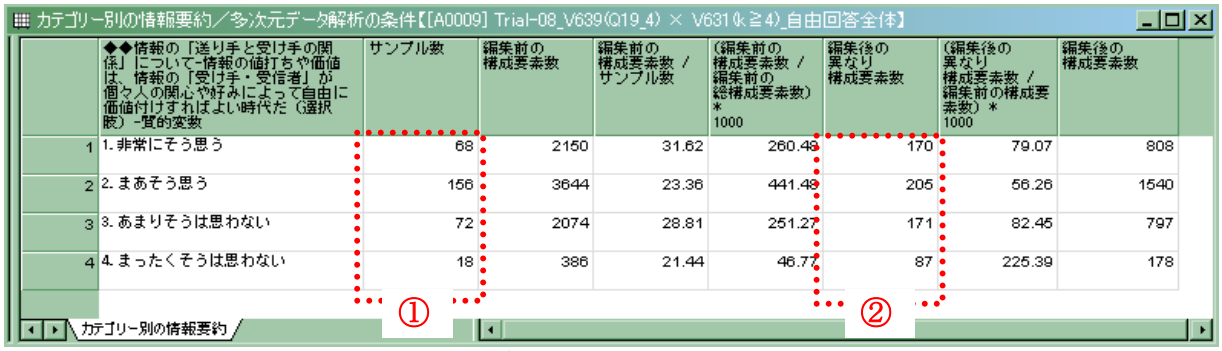


図 27 「カテゴリー別の情報要約」の出力情報

観察 9:質問文の観察(質的変数の観察)

すでに述べたように，所与のデータ表の“行側のクラスター化”（ここでは構成要素）と“列側のクラスター化”のいずれも同時的に分類される（対応分析の仕組みから当たり前のこと）．この例では，質的変数の質問文の選択肢は少ないので分類する意味はあまりない．一方，構成要素の回答分布（発言の内容）の構成を観測することは有効である．

たとえば，図 27 でみた集計結果の具体的な内容を**「質的変数による構成要素の有意性テスト」**の中で得られる**「頻度による有意性テストの要約」**情報から拾ってみる．質問「Q19_4」の**4 つの選択肢**のそれぞれで意味のある（有意となりそうな）**“上位の語句群”**と反対にあまり寄与しないと思われる**“下位の語句群”**とを観察する．

WordMiner では，たとえばこれを表 22 のように要約情報として提供する（前述のように他の探査ツールもいろいろあるが，まずこの要約から観察する）．

これに関連の情報の解釈についてはここでは述べない[49]．ここでは，クラスター化と関連させてこれら情報の観察も必要であることだけを指摘しておこう．この情報を表 22 として再編集した．ここから，質問「Q19_4」の 4 つの選択肢（「非常にそう思う」「まあそう思う」「あまりそうは思わない」「まったくそうは思わない」）の**それぞれに特徴的な語句の傾向**がみえてくる．

さらにここで，比較のために，質問「Q19_2」を用いた場合にえられる情報も要約する．これが表 23 である．つまりここでは，用いる自由回答データは上と同じ内容として，それから得た語句群と，その自由回答質問の前に置いた 4 つの選択肢からなる質問（Q19_1～Q19_4）との関連を探査することで，両者の関係を調べようという意図がある．その 4 つの質問文のうちの，回答傾向が異なる 2 つの質問「Q19_2」と「Q19_4」を比べてみるわけである．表 22 と表 23 を対比すると，確かに語句の使われ方の傾向は 2 つの質問間で異なるようである．

[49] 別に資料が用意されているのでそれを参照（参考文献のうしろにあげた）．

さらに一歩踏み込んで，では，それぞれ特徴があるとされた語句を，回答者は**実際にどのように記述していたか**を，調べよう．ここではとくに，選択肢「非常にそう思う」に対して有意とされた語句群が，元の自由回答ではどのような記述であったかを（WordMiner の）出力情報から整理要約した．表 24 が質問「Q19_2」の「非常にそう思う」と関連があるとされた元の自由回答である．同じく表 25 が質問「Q19_4」に対応する．

ここでもう一度，ここでみた 2 つの質問を書き出し確認する．

Q19_4：情報の値打ちや価値は、情報の**「受け手・受信者」が個々人の関心や好みによって自由に価値付け**すればよい時代だ”（「Q19_4：個々人の関心や好みによって自由に価値付けすればよい時代」）
Q19_2：情報の値打ちや価値は、**「送り手・発信者」が情報の確からしさや根拠などの裏づけを求められる**時代だ”（「Q19_2：情報の確からしさや根拠などの裏づけを求められる時代」）

これを確認のうえ，とくにあらためて太字・下線とした文言に注目して，再度，2 つの表（表 24，25）にある元の自由回答を観察しよう．ここで曖昧さは混ざるものの，それぞれが上で回答者が「非常にそう思う」を選んだことへの“意味付け”となっていることがみえる（読める）であろう．ここで重要なことは，ここに至る一連の分析は，**ほぼ自動的**に行ったという点にある．一方，手順の中で，分析者の判断（主観）に委ねられる部分も多々ある．

ここで“ほぼ自動的”とは，ここでは，簡単な統計的手法を用いて有意性テストを行った結果にもとづいて，自由回答を選んでいるからである．表 22，表 23 の「上位，下位」の選び方と，表 24，表 25 で「検定値」[50]の欄に示した数値が，この選ぶときに目安である．実はここでは，結果はほとんどが有意ではなく，しかし検定値の大きいものから順に選んでみた（検定値が 1.96 あるいは 2.0 を越える例はほとんどない）．つぎに述べる「応用事例（その 2）」に挙げた結果と対照的である．このうしろの例では，表 28，表 31 に見るように，高い有意となる回答例が多数ある．この両者の違いは，結果解釈のうえで重要であり，うしろでまた述べる．1 つだけ指摘するならば，この差違は**“質問文の難易度の違い”**あるいは**“回答者の質問文の理解度，解釈の差違”**に関係することを示唆している，ということである．つまり，ここの例でいえば，ここで用いた 4 つの質問文「Q19_1～Q19_4」は，つぎの例で用いる“ソーシャル・メディアの利用や普及で「Q13_A：プラスと思うこと」「Q13_B：マイナスと思うこと」”という 2 つの質問文に比べて，内容がわかりにくい，あるいは解釈がむずかしいのではないか，という，設計時当初の推測が当たっていると思われる．こうした知見が得られることが，自由回答分析ではきわめて重要である．これについては，うしろで再び触れることにしよう．

- どの質問が自由回答質問で得た語句群と関連ありそうか，その質問の選び方．さらに遡れば，（分析者の意図にあった）自由回答を得るための“適切な質問文の作り方”．
- 自由回答データには，若干意味不明の語句や，類似の語句（類語や同義語に近い語句）の言い替えがあること．
- つまり，単語・語句を整えること，言い換えると辞書編集やシソーラスなどの利用が必要ではないか，と思われること．
- ここで用いた選択肢型質問とは別に，自由回答質問と関連のある（説明力のある）別の質問文をどのように探すか（探せるのか）．
- 自由回答質問同士の分析はどのように行えるのか．

[50] ここでいう「検定値」および「検定値（無調整）」とは，ある検定統計量の実現値（ここでいう検定値）を使って行う簡易の統計的検定であり，正規近似を用いている．この検定値の絶対値が 1.96（丸めて約 2.0）を目安にする．これについての説明は別資料にある（参考文献を参照）．また，表には示さなかったが，この検定値に対する**“有意確率”**も算出・表示する．

・ ここでは，出現頻度が多い 3 語以下の単語・語句は分析では**用いていないこと**．閾値を設けて単語を選ぶ理由，その閾値の決め方にはなにか理由があるのだろうか．このようなある種の網をかけた（篩（ふるい）にかけた）情報からデータの特徴探査・抽出を行ったことになる．これをどう考えるか．

ここまでの分析でみたように，より細かい説明は省くが，対応分析法とクラスター化法の特性をうまく活かした探査的なアプローチを行う（マイニングを行う）**仕組み作りが課題**であるということである．

表 22 質問「Q19_4」の 4 つの選択肢について意味ある構成要素（単語群）の内容［図 28 から］
「Q19_4：個々人の関心や好みによって自由に価値付けすればよい時代」

有意の順位	1.非常にそう思う サンプル数：68 異なり構成要素数：170	2.まあそう思う サンプル数：156 異なり構成要素数：205	3.あまりそうは思わない サンプル数：72 異なり構成要素数：171	4.まったくそうは思わない サンプル数：18 異なり構成要素数：87
上位 1	そう	すれば	責任	増えた
上位 2	興味	沢山	ついて	流す
上位 3	受信者	自分	ような	持って
上位 4	なる	何	され	値打ち
上位 5	する	ネット	好み	誰
上位 6	発信	あまり	受け止める	側
上位 7	います	考え方	ない	なく
上位 8	くる	裏付け	なって	鵜呑み
上位 9	目	あふれて	きちん	重要
上位 10	取捨選択	いかなければ	持って	不正確
上位 11	では	おり	信用	発信
上位 12	して	上	もの	価値
上位 13	思います	新聞	その	どうか
上位 14	自身	いい	嘘	選択
上位 15	なく	見極める	真偽	
上位 16	必要	それぞれ	内容	
上位 17	常	しっかり	よく	
上位 18	いけない	よい	時	
上位 19	見極め	できる	情報量	
上位 20	だから	情報過多	難しい	
上位 21	中	誰	ように	
上位 22	昔	どの	多く	
上位 23	大切	だ		
上位 24	良い	いろんな		
上位 25		選んで		
上位 26		送る		
上位 27		得る		
上位 28		分からない		

下位 25		中		
下位 24		ような		
下位 23		昔		
下位 22		して		
下位 21		ある		
下位 20		なって		
下位 19		なく		
下位 18		受信者		
下位 17		興味	しなければ	
下位 16	送り手	います	のが	
下位 15	きちん	くる	個々	
下位 14	信用	為	考え方	
下位 13	裏付け	関心	自身	
下位 12	だ	目	いう	
下位 11	できる	流す	すれば	
下位 10	いい	その	正しく	
下位 9	あまり	する	誰	

下位 8	すれば	真偽	なる	
下位 7	情報過多	発信	自分	
下位 6	情報量	価値	たくさん	
下位 5	誰	そう	どれ	
下位 4	難しい	責任	それぞれ	
下位 3	どうか	では	側	
下位 2	正しい	持って	だから	だ
下位 1	には	もの	選択	必要

(*) 表頭先頭行のアミかけセル内に各選択肢に含まれる回答者数と異なり構成要素数の情報がある．図 27 の要約表と比較しよう．

表 23 質問「Q19_2」の 4 つの選択肢について意味ある構成要素（単語群）の内容
「Q19_2：情報の確からしさや根拠などの裏づけを求められる時代」

有意の順位	1.非常にそう思う サンプル数：110 異なり構成要素数：206	2.まあそう思う サンプル数：149 異なり構成要素数：206	3.あまりそうは思わない サンプル数：38 異なり構成要素数：116	4.まったくそうは思わない サンプル数：17 異なり構成要素数：95
上位 1	感じる	いい	自分	いる
上位 2	テレビ	よる	よって	います
上位 3	きちん	受けて	人	流す
上位 4	正しい	いかなければ	手	自分自身
上位 5	では	個々人	何	事
上位 6	されて	上	すれば	して
上位 7	鵜呑み	判断	違う	あくまで
上位 8	なく	側	だ	大切
上位 9	能力	考え	それ	個人
上位 10	最近	時	確かな	興味
上位 11	政府	いろいろな	思った	考えて
上位 12	報道	間違った	入れる	持って
上位 13	情報	他人	なので	世の中
上位 14	しまう	得る	いく	発信
上位 15	時代	方	色々	発信者
上位 16	ない	必要	いろんな	する
上位 17		見極める	選んで	重要
上位 18		裏付け	それぞれ	不正確
上位 19		思う	よい	人
上位 20			では	ように
上位 21			興味	見極め
上位 22			考え方	

下位 24	受けて			
下位 23	色々			
下位 22	個々			
下位 21	持って			
下位 20	事			
下位 19	いかなければ			
下位 18	間違って			
下位 17	個々人			
下位 16	上	情報		
下位 15	して	ない		
下位 14	判断	時代		
下位 13	違う	きちん		
下位 12	考え	興味		
下位 11	難しい	います		
下位 10	それ	ことに		
下位 9	ように	確かな		
下位 8	よる	流す		
下位 7	他人	テレビ		
下位 6	方	なので		
下位 5	側	正しい		
下位 4	人	なく	事	多い
下位 3	いく	されて	には	自分
下位 2	何	感じる	能力	見極める
下位 1	いい	では	正確	思う

表 24　Q19_4:個々人の関心や好みによって自由に価値付けすればよい時代の「非常にそう思う」の回答例

回答者 No.	検定値	検定値 (無調整)	自由回答（原文）
[00000325]	0.4925	0.5137	自身で学習して見極める必要がある。
[00000328]	0.3933	0.5487	個人的ではあるが興味の無い項目が多い
[00000049]	0.3885	0.3885	取捨選択は取り手がすべきことだから
[00000167]	0.3547	0.2792	自己の判断が責任となる時代だから
[00000341]	0.3445	0.3236	ネット上でトラブル情報を耳にすると、受信者が的確に判断する必要を感じているため。
[00000014]	0.3336	0.3347	現在、単なる“情報”の単位では膨大な量があるので、受信者が取捨選択する能力が大切だと思う。
[00000346]	0.324	0.3506	各種の情報が氾濫する時代、今後益々情報量が増大するだろう。、その中での情報価値は受け手が取捨選択する、しなければならない時代になる。
[00000260]	0.3237	0.2939	情報を鵜呑みにするのではなく、真偽を自分の責任で判断しなければいけない。
[00000041]	0.3156	0.3472	情報が多い中で、欲しい情報は自身で判断する必要がある。、得られるものの信憑性を自身で判断することが必要と考えます。
[00000110]	0.3124	0.2789	３と４ は今も昔もそう変わっていないと思う
[00000294]	0.279	0.2632	発信者は常に正確な情報を発信し、受信者がほしいと思う情報と一致したときに、初めて価値のある情報となると思います。
[00000043]	0.2738	0.2551	情報が溢れかえってるからこそ、受信者の見極めは大事だと思います
[00000087]	0.2697	0.2479	情報を発信する側がどのような人物なのかわからないので、自分自身で情報を見極め、判断していく必要があるため。
[00000163]	0.2647	0.1627	情報を必要としている人が自分で調査すべき
[00000276]	0.2507	0.2481	あらゆる情報が氾濫している今､各々が自分に必要なもの だけ取り込み後は排除する勇気とか努力がますます必要になってくる
[00000076]	0.2226	0.2555	受け手の興味・関心が細分化しているので、そのときの流行というものは確かにあるにしても、自分の興味によって情報の取捨選択が必要とされていると考えるから。
[00000215]	0.2171	0.2252	発信者の顔が見えないし､必ずしも的確な情報を発信しているとは限らない。悪意目的で発信している場合もあるので、受信者は個々に見極める必要がとても大事だと思う。
[00000318]	0.2068	0.1973	情報は発信の性格さが求められる。受け取る側はどうしても鵜呑みにしてしまうので、どちらもそれを正確に見極める力が必要になると思う
[00000273]	0.2033	0.1688	多くの情報の中から、自分の必要なもの、信じられるものを、賢く選ぶ目を持たなければいけないと思う。
[00000065]	0.2027	0.2302	twitter を例にコメントしますが、フォローしている人の TL を見ている時点で自身のフィルターが情報にかかっています。その中でさらに興味あるものだけ、目に留まったりするので、情報洪水の中では受信者の価値判断が重要だと思います。送り手にそれを求めてもらちが開かない。その点、かなり限定されたコミュニティーに所属していることも忘れてはいけないので、twitter ばかりにハマらないで、社会と繋がる事も大切。
[00000298]	0.197	0.2539	一方的な情報に惑わされることなく、個人の関心のあること以外、は受け流します。、、興味のあることはみますがあとはみないようにしています。
[00000213]	0.1931	0.2379	情報氾濫の中では、取捨選択ができる能力が必要。特に日本人は右ならえの民族なので、正しく情報を租借する必要がある。、庇護流言に日本人は昔から非常に動揺し弱い。、他者の価値観を認めながら、最終的には自己の価値観・判断を大切にすることがこれからますます必要になる。
[00000315]	0.1896	0.1965	情報社会ですので 振り回される事なく 、自分自身正しく判断出来る様、常に心がけたい
[00000131]	0.1866	0.1976	情報の多い中で振り回されるのではなく、必要な情報だけを取り入れればいいと思っているから。
[00000070]	0.1862	0.1052	自分で考えて判断するのが今の世の中必須、

表 25　Q19_2:情報の確からしさや根拠などの裏づけを求められる時代の「非常にそう思う」の回答

回答者 No.	検定値	検定値(無調整)	もとの自由回答
[00000345]	0.5771	0.5771	情報に惑わされない
[00000218]	0.304	0.2256	情報が多すぎて自分で選択する時代
[00000012]	0.2997	0.3016	最近、捏造されたニュースなどをテレビでよく見かけるから。
[00000264]	0.2852	0.2762	情報を操作されている場合もあるので。
[00000121]	0.2688	0.2469	受け手は送り手の情報を鵜呑みにするのではなく、自分自身でその価値を判断すべきである。
[00000054]	0.268	0.2796	配信される情報を鵜呑みにし、流されてしまうことほど愚かしいことはないと思う。面倒ではあるが、多種多様な情報から、真実を見出す賢さが求められている時代なのでは。
[00000246]	0.2511	0.2643	情報がすべて正しく無い
[00000260]	0.2435	0.2341	情報を鵜呑みにするのではなく、真偽を自分の責任で判断しなければいけない。
[00000073]	0.2353	0.1428	選ぶ時代になっている
[00000159]	0.2294	0.2334	個人個人で必要な情報が変わってきており、情報に流されない、惑わされず正しい情報を、つかむスキルや責任が求められる時代だと思います。
[00000210]	0.2223	0.2191	情報はオープンで公平であるべき。記者クラブ報道が情報を歪めているように感じる。
[00000111]	0.2126	0.1469	情報を発信する側が一方向からではなく、中立の立場で発信する必要がある。、受け手側も情報を鵜呑みにするのではなく、その情報が正しいものかどうかを自分で判断するべきである。、
[00000021]	0.2092	0.2092	情報の信頼性が不十分だから
[00000177]	0.2092	0.2223	正確な情報が少ないから
[00000035]	0.2056	0.2187	個人での情報発信が容易に出来る世の中になり、送り手も受け手も情報に対する価値観が非常にこの数年で多様化、多岐にわたっていると感じる。、その中で、受け手の情報受諾に置いてはきちんと整理して受け止めることが必要であると感じる。、送り手はきちんと信頼できる情報を流すべきである。、簡単なことに感じるが、出来てない世の中になっていると感じる。
[00000254]	0.2007	0.1827	あくまでも情報は正しいものをきちんと流すべき。デマやウソを平気で流すのは良くない事と思う。
[00000280]	0.1924	0.1688	必要のない情報とかもとても多いので選択が大変
[00000050]	0.19	0.1631	情報過多な現代は受け手が情報を選択し、正しいかどうかを自身で見極める時代だと思う。
[00000100]	0.185	0.2117	あまりにも沢山のメディアで、それぞれ多くの情報を発信しているので、全てを聞き入れていては、頭が混乱してしまうから。、また、最近、テレビの報道でもやらせや過剰な演出など、正確さに欠ける情報が多いと感じるため、情報を吟味する力が必要であると感じるから。
[00000283]	0.1793	0.2088	インターネット等情報が多い中、情報の見極めが特に必要
[00000018]	0.178	0.1778	情報量が増えているなかで、手に入れる情報がそもそも正しい情報であるかどうかが重要であると思う
[00000242]	0.1773	0.2047	送り手・発信者は情報伝達に誠実、誠意と責任感を持つべきである。受け手・受信者が鵜呑みをするではなく、自身は的確な情報を見極める能力を身に着けるべきである。

5. 3 分析例(その 2)

クラスター化を用いる別の分析を試みる．前節で用いた同じ意識調査の別の質問について，検討する．調査の概要はすでに述べたので省略し，まず用意した自由回答質問を確認しよう．またこの例では，（その 1）での結果とは異なる傾向，とくに自由回答質問の解釈に対する回答者の反応がより具体的に比較対照できることに注意しよう．

[用いた自由回答質問文]

ここでは，つぎの自由回答質問文を用意した．前の「応用事例（その 1)」で用いた質問文に比べて，内容がかなり具体的かつ分かりやすいことに注意しよう．

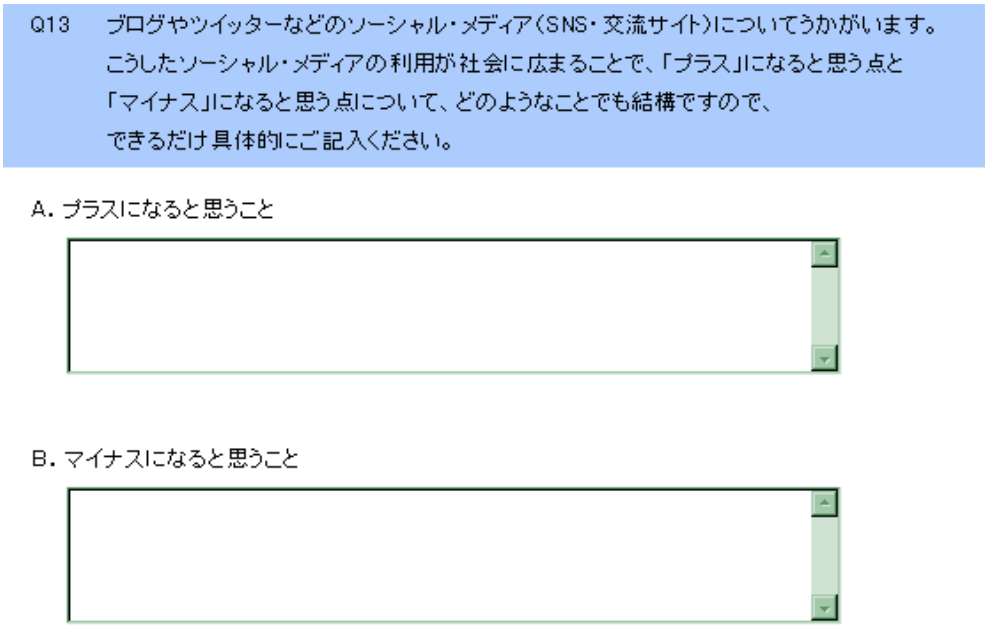


図 29 ソーシャル・メディアに関する自由回答質問

さらにこの調査では，この自由回答質問の前後に，「情報利用」に関する多数の質問文（選択肢型）を設けた．ごくいくつかを，とくにソーシャル・メディアなどの関連するものをいくつかあげると以下のような質問がある．また，上の分析で用いた質問「Q19」は，この質問文のかなりうしろに置かれていることを指摘しておこう[51].

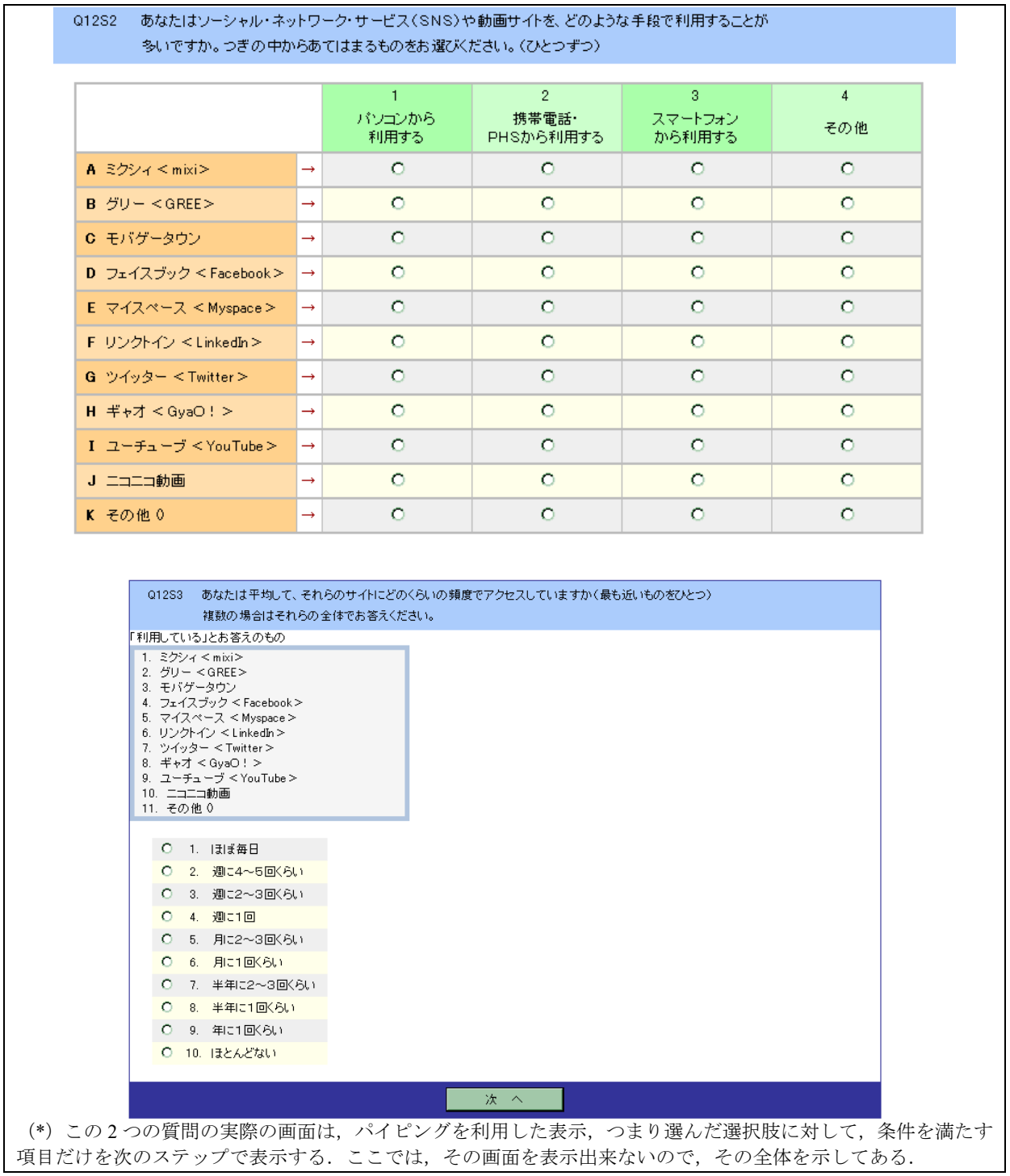
Q12 まず、ソーシャル・ネットワーク・サービス(SNS)や動画サイトについておうかがいします。
あなたが「ご存知のもの」をすべてお選びください。(いくつでも)

＜チェックボックス形式の選択肢は省略＞

Q12S1 あなたがご存知のソーシャル・ネットワーク・サービス(SNS)や動画サイトの中で、
あなたが「利用しているもの」をすべてお選びください。(いくつでも)

＜チェックボックス形式の選択肢は省略＞

[51] もしかすると，“文脈効果”などがあることも予想される，ということ．後述．



Q12S2　あなたはソーシャル・ネットワーク・サービス（SNS）や動画サイトを、どのような手段で利用することが多いですか。つぎの中からあてはまるものをお選びください。（ひとつずつ）

	1 パソコンから利用する	2 携帯電話・PHSから利用する	3 スマートフォンから利用する	4 その他
A ミクシィ＜mixi＞ →	○	○	○	○
B グリー ＜GREE＞ →	○	○	○	○
C モバゲータウン →	○	○	○	○
D フェイスブック ＜Facebook＞ →	○	○	○	○
E マイスペース ＜Myspace＞ →	○	○	○	○
F リンクトイン ＜LinkedIn＞ →	○	○	○	○
G ツイッター ＜Twitter＞ →	○	○	○	○
H ギャオ ＜GyaO！＞ →	○	○	○	○
I ユーチューブ ＜YouTube＞ →	○	○	○	○
J ニコニコ動画 →	○	○	○	○
K その他 0 →	○	○	○	○

Q12S3　あなたは平均して、それらのサイトにどのくらいの頻度でアクセスしていますか（最も近いものをひとつ）
複数の場合はそれらの全体でお答えください。

「利用している」とお答えのもの

1. ミクシィ＜mixi＞
2. グリー ＜GREE＞
3. モバゲータウン
4. フェイスブック ＜Facebook＞
5. マイスペース ＜Myspace＞
6. リンクトイン ＜LinkedIn＞
7. ツイッター ＜Twitter＞
8. ギャオ ＜GyaO！＞
9. ユーチューブ ＜YouTube＞
10. ニコニコ動画
11. その他 0

○ 1. ほぼ毎日
○ 2. 週に4～5回くらい
○ 3. 週に2～3回くらい
○ 4. 週に1回
○ 5. 月に2～3回くらい
○ 6. 月に1回くらい
○ 7. 半年に2～3回くらい
○ 8. 半年に1回くらい
○ 9. 年に1回くらい
○ 10. ほとんどない

次 へ

（*）この 2 つの質問の実際の画面は，パイピングを利用した表示，つまり選んだ選択肢に対して，条件を満たす項目だけを次のステップで表示する．ここでは，その画面を表示出来ないので，その全体を示してある．

図 30　用意したソーシャル・メディアに関するいくつかの質問文

ここで「Q12」でまず「交流サイト」の「認知度」を調べ，つぎに「Q12S1」で「利用しているサイト」を確認，さらに「Q12S2」でその「利用の仕方」を問い，そしてこのあとに「利用頻度」をたずねている．

実は，これから用いる上の自由回答質問（図 29）はこのブロックの最後に置いてある．このような場合，一種の“**文脈効果**”[52]（context effect），つまり自由回答の記述の中に，これらの質問文のワーディングが影響することも考えられる．ここでは，こうした効果も想定のう

[52] 先行する（あるいは時にはうしろに続く）質問文の文言・文脈が，その自由回答質問への回答に影響することが予想される，ということ．

えで，上の自由回答質問を用意した．また，ソーシャル・メディアの「プラスになると思うこと」「マイナスになると思うこと」と分けて回答の記入を求めるように設定してある．

この自由回答質問の内容と，前節でみた「Q19．情報の考え方（情報の送り手と受け手）…」との内容を比べると，すでに上で少し触れたように．後者「Q19」のほうが，4 つの選択肢型質問文の内容がやや**“難しい”**と感じるであろう．こうした質問文の記述内容（ワーディング）の影響ははたして自由回答に影響するのだろうか．また，影響するとしたら，具体的にどのような現象として観察されるのだろうか．こういうことも念頭に，この自由回答質問の分析を進めてみよう．

観察 1:構成要素の分布の観察

ここでもまず，自由回答データ（回答原文）を分かち書き処理し，それに“簡単な辞書編集”を行った結果からえられた構成要素変数（語句群）の分布の観察から始めよう．この例でも，多くの場合に分かち書き後にみられるさまざまな不具合，たとえば，分かち書きの不備，ゴミや誤記入の訂正（編集）や削除（削除辞書の適用），同義語・類語の置き換え（置換辞書編集の利用）などは，一切行わずに，その分かち書き情報（構成要素とする語句群）を用いる．つまりここでいう“簡単な辞書編集”とは，**“句読点除去程度”の操作**をいう．こうして得た**全構成要素**（分析に用いる語句群）の頻度分布が図 31 である．

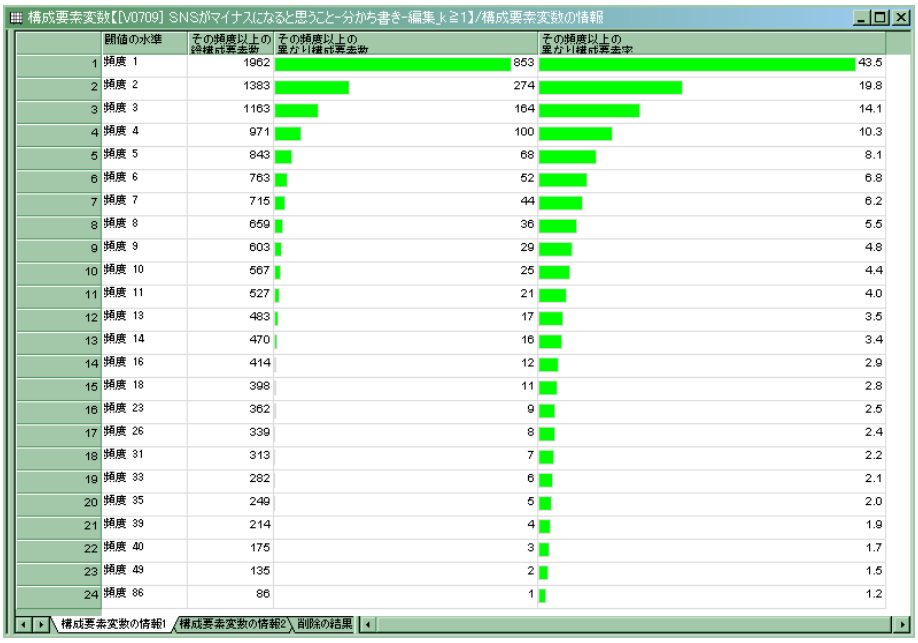
構成要素変数【V0706】SNSがプラスになると思うこと-分かち書き-編集_k≧1】/構成要素変数の情報

	閾値の水準	その頻度以上の総構成要素数	その頻度以上の異なり構成要素数	その頻度以上の異なり構成要素率
1	頻度 1	1877	718	38.3
2	頻度 2	1389	230	16.6
3	頻度 3	1207	139	11.5
4	頻度 4	1081	97	9.0
5	頻度 5	985	73	7.4
6	頻度 6	880	52	5.9
7	頻度 7	844	46	5.5
8	頻度 8	795	39	4.9
9	頻度 9	771	36	4.7
10	頻度 10	699	28	4.0
11	頻度 11	659	24	3.6
12	頻度 12	637	22	3.5
13	頻度 13	601	19	3.2
14	頻度 14	588	18	3.1
15	頻度 15	560	16	2.9
16	頻度 16	545	15	2.8
17	頻度 18	513	13	2.5
18	頻度 20	477	11	2.3
19	頻度 21	437	9	2.1
20	頻度 24	416	8	1.9
21	頻度 25	392	7	1.8
22	頻度 27	367	6	1.6
23	頻度 29	340	5	1.5
24	頻度 31	311	4	1.3
25	頻度 63	280	3	1.1
26	頻度 80	217	2	0.9
27	頻度 137	137	1	0.7

構成要素変数の情報1 | 構成要素変数の情報2 | 削除の結果

図 31 「プラスになると思うこと」の構成要素分布

まず「Q12_A：プラスになると思うこと」について調べる．ここで，頻度 1，つまり 1 語以上の語句（全構成要素数）が延べで 1,977 語あり，重複を行わずに計数した異なり語句数（異なり構成要素数），つまり 1 語（1 回）のみ登場の語句が 718 語となったこと，その全構成要素数に占める割合が 38.3%であったことがわかる．また，頻度 2 語以上の異なり語句数は急に減って 230 語（16.6%）となることもわかる．この語句数の減衰状況に注目しよう．


構成要素変数[[V0709] SNSがマイナスになると思うこと-分かち書き-編集ﾚﾍﾞﾙ1]/構成要素変数の情報

	閾値の水準	その頻度以上の総構成要素数	その頻度以上の異なり構成要素数	その頻度以上の異なり構成要素率
1	頻度 1	1962	853	43.5
2	頻度 2	1383	274	19.8
3	頻度 3	1163	164	14.1
4	頻度 4	971	100	10.3
5	頻度 5	843	68	8.1
6	頻度 6	763	52	6.8
7	頻度 7	715	44	6.2
8	頻度 8	659	36	5.5
9	頻度 9	603	29	4.8
10	頻度 10	567	25	4.4
11	頻度 11	527	21	4.0
12	頻度 13	483	17	3.5
13	頻度 14	470	16	3.4
14	頻度 16	414	12	2.9
15	頻度 18	398	11	2.8
16	頻度 23	362	9	2.5
17	頻度 26	339	8	2.4
18	頻度 31	313	7	2.2
19	頻度 33	282	6	2.1
20	頻度 35	249	5	2.0
21	頻度 39	214	4	1.9
22	頻度 40	175	3	1.7
23	頻度 49	135	2	1.5
24	頻度 86	86	1	1.2

構成要素変数の情報1 | 構成要素変数の情報2 | 削除の結果

図 32 「マイナスになると思うこと」の構成要素分布

つぎに「Q13_B：マイナスになると思うこと」について調べる．ここで，頻度 1，つまり 1 語以上の語句（構成要素数）が延べで 1,962 語あり，重複を行わずに計数した異なり語句数（異なり構成要素数），つまり 1 語につき 1 回と計数した語句が 853 語となったこと，その全構成要素数に占める割合が 43.5%であったことがわかる．またここで，頻度 2 語以上の異なり語句数は急に減って 274 語（19.9%）となることもわかる．ここでも語句数の減衰状況に注目しよう．そしてこの 2 つ「プラス…」「マイナス…」については，ほぼ類似した分布となっているようにみえる．

ところで，前にみた「Q19.情報の考え方」についての構成要素の分布は（図 14），この 2 つの分布とかなり異なる傾向にある．とくに，頻度 1 と頻度 2 の構成比率とそのあとの低減の様子がかなり異なる．たとえば上位 3 位までの構成比率をみると（表 25），「Q19.情報の考え方」は約 46.9%，「プラス…」が約 66.4%，「マイナス…」が約 77.4%となり，かなり異なる．これは単純な指標であるが，テキスト型データの観察では重要な操作である[53]．

表 25 構成要素の分布の比較

		Q19：情報の考え方			Q13_A：プラスと思うこと			Q13_B：マイナスと思うこと		
順位	閾値の水準	総構成要素数	その頻度以上の異なり構成要素数	その頻度以上の異なり構成要素率	総構成要素数	その頻度以上の異なり構成要素数	その頻度以上の異なり構成要素率	総構成要素数	その頻度以上の異なり構成要素数	その頻度以上の異なり構成要素率
1	頻度 1	4720	1293	**27.4**	1877	718	**38.3**	1962	853	**43.5**
2	頻度 2	3873	446	**11.5**	1389	230	**16.6**	1383	274	**19.8**
3	頻度 3	3545	282	**8.0**	1207	139	**11.5**	1163	164	**14.1**
4	頻度 4	3350	217	**6.5**	1081	97	**9.0**	971	100	**10.3**
5	頻度 5	3174	173	**5.5**	985	73	**7.4**	843	68	**8.1**
6	頻度 6	3009	140	**4.7**	880	52	**5.9**	763	52	**6.8**
7	頻度 7	2901	122	**4.2**	844	46	**5.5**	715	44	**6.2**
8	頻度 8	2782	105	**3.8**	795	39	**4.9**	659	36	**5.5**
9	頻度 9	2654	89	**3.4**	771	36	**4.7**	603	29	**4.8**
10	頻度 10	2618	85	**3.2**	699	28	**4.0**	567	25	**4.4**

[53] ここでは，同一の調査内の同じ回答者から得た異なる自由回答質問を比べていることに注意する．異なる調査，異なる回答者群からえた構成要素の分布の比較は，このようには行えない．つまり，回答者の負担にならない範囲で，1 つの調査に内容の異なる複数の自由回答を設けることは事後の分析にとって意味がある．

観察 2:登場する語句の観察(構成要素の内容)

つぎに，2 つの質問「Q13_A：プラスと思うこと」「Q13_B：マイナスと思うこと」の利用語句の内容を観察しよう．図 33，図 34 が，その語句の頻度分布である．ここでは約「8 回以上」登場した語句をソートして上位から並べてある．2 つを比べると「情報」「人」「コミュニケーション」「なる」など，両者に共通に登場する語句がある．とくに「情報」は，「プラス…」では 137（回），132（人）が記述，「マイナス…」では，86（回），76（人）が用いている．一方，「プラス…」「マイナス…」にそれぞれ関連ありそうな異なる語句が並んでいる（とくに説明を必要としないだろう）．つまり，「プラス…」「マイナス…」と 2 つに分けて質問を用意したことの効果があったとみてよさそうである．

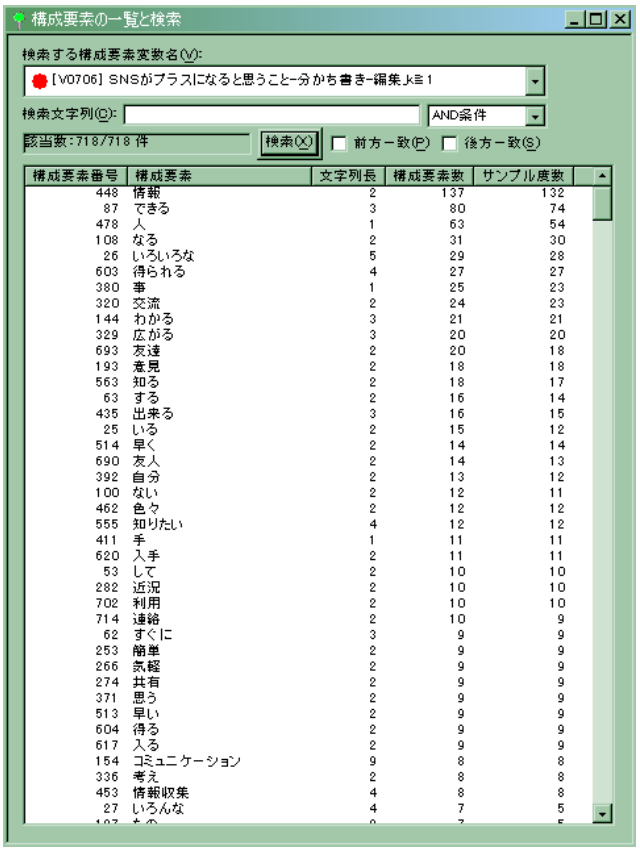


図 33 「Q13_A:プラスと思うこと」の語句の頻度

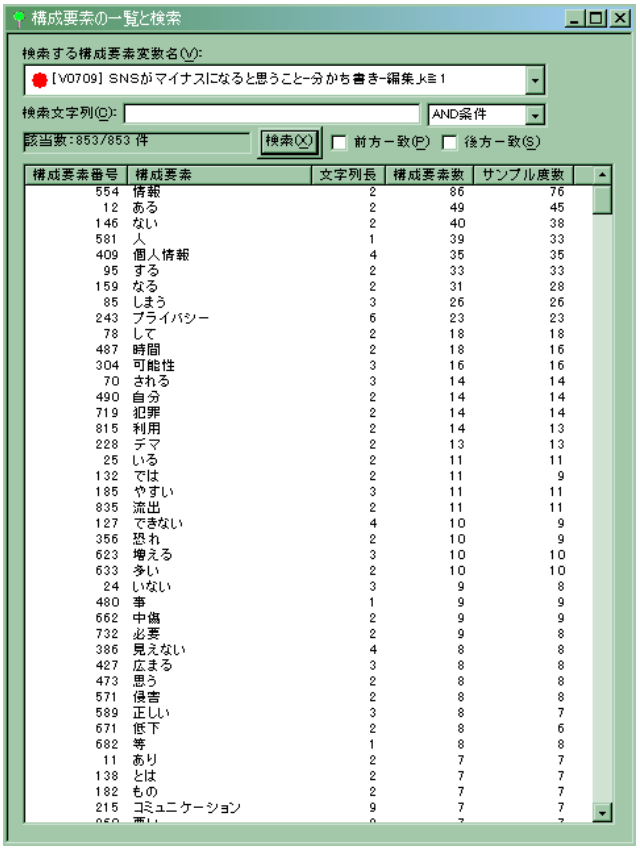


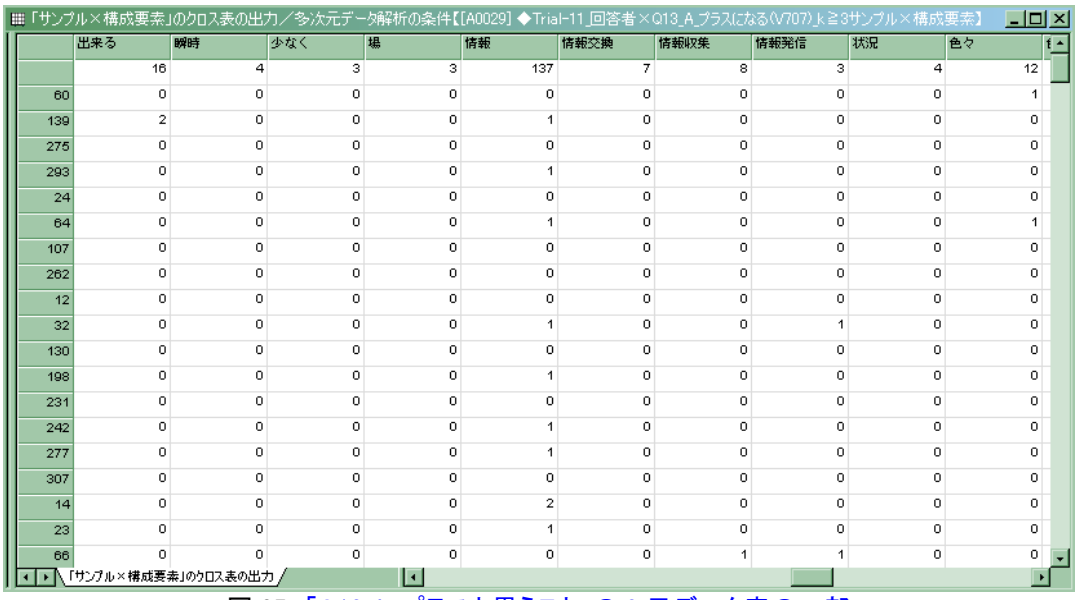
図 34　Q13_B:マイナスと思うことの語句の頻度

観察 3:クラスター化による自由回答の類型化

質問「Q19：情報の考え方」では，自由回答の情報を，他に設けた選択肢型質問（質的変数）との関連を探査することを試みた．ここでも同じような分析を行うことも可能であるが，少し視点を変えて，取り上げた 2 つの質問「プラス…」「マイナス…」の間で，意見の違いがどう見えるのかを“回答者をその自由回答だけから類型化し調べる，つまり**“意見の傾向を探査”**することを行ってみる[54]．このためには，図 35 にあるような「(回答者・サンプル) ×（構成要素変数)」の“2 元データ表”を分析対象として，回答者と構成要素群（語句群）とをそれぞれ同時的に対応分析法を行い，また得られた成分スコアを用いてクラスター化を行う．一般にここで注意することは，回答者数も多く，また選んだ語句数もかなりの数となるので，

[54] パターン認識的な言い方をすると，いわゆる“教師なし分類”を行うことに相当する．これに対して，選択肢型質問を用いた分析は，自由回答に対比させる参照指標とする分類情報が質的変数として与えられた分類ということもできる．

この種のデータ表の寸法はかなり大きくなることである．また，図 35 のように，大抵は各セル内の度数は少なくきわめて疎なデータ表となる（図 35 は一番出現頻度の大きい「情報」のあたりを切り取ってみたが，それでもこのような疎な状態）．



「サンプル×構成要素」のクロス表の出力／多次元データ解析の条件【[A0029] ◆Trial-11_回答者×Q13_A_プラスになる(V707)_k≧3サンプル×構成要素】

	出来る	瞬時	少なく	場	情報	情報交換	情報収集	情報発信	状況	色々
	16	4	3	3	137	7	8	3	4	12
60	0	0	0	0	0	0	0	0	0	1
139	2	0	0	0	1	0	0	0	0	0
275	0	0	0	0	0	0	0	0	0	0
293	0	0	0	0	1	0	0	0	0	0
24	0	0	0	0	0	0	0	0	0	0
64	0	0	0	0	1	0	0	0	0	1
107	0	0	0	0	0	0	0	0	0	0
262	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0
32	0	0	0	0	1	0	0	1	0	0
130	0	0	0	0	0	0	0	0	0	0
198	0	0	0	0	1	0	0	0	0	0
231	0	0	0	0	0	0	0	0	0	0
242	0	0	0	0	1	0	0	0	0	0
277	0	0	0	0	1	0	0	0	0	0
307	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	2	0	0	0	0	0
23	0	0	0	0	1	0	0	0	0	0
66	0	0	0	0	0	0	1	1	0	0

「サンプル×構成要素」のクロス表の出力

図 35 「Q13_A:プラスと思うこと」の 2 元データ表の一部

この例では，回答者数（サンプル数）は，337（人）とさほど大きくはないが，もし全語句を用いたとすると，つまり登場した全語句（総構成要素）を用いると，図 31、図 32 にあるように「プラス…」が 718（語），「マイナス…」が 853（語）ある[55]．このまま対応分析，クラスター化と進んでもよいが，すこし語句を絞り込んで，とくに 1 回しか登場しない頻度 1 の語句を除外する[56]．このあと，「プラス…」については“頻度 2 以上と頻度 3 以上”を，「マイナス…」については“2 以上の語句”を使って分析しよう．ここで閾値を変える理由については，分析の段階で述べる．この場合「プラス…」が 230（語）および 139（語），「マイナス…」が 274（語）となる．

以下，まず「プラス…」のデータ表について，一連の分析結果をおってみよう．

［解析の条件］

はじめに，ここで用いたデータ表の条件を整理しておこう．

- 解析対象とするサンプル数＝347（人）［全回答者数］
- 解析対象から除外のサンプル数＝24（人），よって計算に用いたサンプル数＝323（人）
- 解析対象とする構成要素変数名＝SNS がプラスになると思うこと
  （*）分かち書き後に簡単な編集を行い選出する．
- 出現頻度 2 以上の解析対象とする「異なり構成要素数」＝230（語）
- 出現頻度 3 以上の解析対象とする「異なり構成要素数」＝139（語）
- 「解析対象とする総構成要素数」＝1,389（語）

対応分析の処理条件と結果

指定した成分数（成分軸の数）＝228

---

[55] この語句数も，テキスト・マイニングで扱う大きさとして，さほど大きくはない．数千～数万語となることもよくある．ここでは寸法が最大でも「347×718」あるいは「347×853」の 2 元データ表で済むということ．

[56] この時点で，1 回しか登場しない語句も重要である，あるいはそうした語句の中に重要な特徴がある，と考えるならここのアプローチは使えない．また，1 語を含む全構成要素を用いる分析を行うべきである．

（*）ここではデータ表の寸法が $K = \min\{m,n\} - 1 = \min\{323,230\} - 1 = 229$ あるいは $K = \min\{m,n\} - 1 = \min\{323,139\} - 1 = 138$ となる．かりにここで $K^* < K$ であるようなある K^* を指定すると，固有値の総和（総変動）が異なることに注意しよう[57]．
クラスター数＝サンプル（回答者），構成要素とも 15 群（これを選ぶ理由は後述）

観察 4:固有値，寄与率，累積寄与率の観察

まず，$K = 228$ として得られた固有値と寄与率，累積寄与率は図 36 のようになった．この結果は，いままでにみた固有値の傾向とはかなり異なる．とくに，以下のようなことに注意が必要である．経験的には，また対応分析の特性から，このような特徴がみえる．

- 固有値の大きさがなだらかに変化すること．
- はじめの 2 つが大きな値で，しかも 1 である．（固有値の大きさは 1 を越えないから，もっとも大きい値となったということ）．
- 第 1 成分スコア，第 2 成分スコアの分散がこの値ということは，実は，データ表のどれかのプロファイルに**はずれ値的なパターン**が含まれる可能性があることを示唆していること．
- こうした場合，そのような行（つまり回答者）あるいは列（語句）を探すこと，場合によってはそれらを除外して再計算を行うことが，データ表の構造をしっかりと把握するために必要かもしれないこと．寄与度も探査の役にたつだろう．
- 寄与率の変化もなだらかであり，始めの 10 成分でも累積で 14.5%，20 成分までで 25.8%程度である．のちに，クラスター化に用いる成分数を指示する場合にこのことを念頭にして決めること．データ表の寸法が大きくなるほど，こういう傾向にある．

図 36　固有値，寄与率，累積寄与率の一覧(2 語以上のとき)
(*)ここで，総変動(固有値の総和)つまり全慣性＝53.8645

[57] これを変えたときの総変動の分解の変化については，すでに述べてある．

	固有値	寄与率	累積寄与率
1	0.7548	2.13	2.13
2	0.7351	2.07	4.20
3	0.7127	2.01	6.21
4	0.6841	1.93	8.14
5	0.6589	1.86	10.00
6	0.6189	1.75	11.75
7	0.6060	1.71	13.46
8	0.5955	1.68	15.14
9	0.5854	1.65	16.79
10	0.5746	1.62	18.41
11	0.5541	1.56	19.97
12	0.5413	1.53	21.50
13	0.5389	1.52	23.02
14	0.5364	1.51	24.53
15	0.5268	1.49	26.02
16	0.5158	1.46	27.47
17	0.5143	1.45	28.93
18	0.5097	1.44	30.36
19	0.4989	1.41	31.77
20	0.4840	1.37	33.14
21	0.4804	1.36	34.49
22	0.4711	1.33	35.82
23	0.4675	1.32	37.14
24	0.4604	1.30	38.44
25	0.4545	1.28	39.72
26	0.4392	1.24	40.96
27	0.4371	1.23	42.19
28	0.4298	1.21	43.40

図 37　固有値，寄与率，累積寄与率の一覧(3 語以上のとき)

ここで，以下のような対応が考えられる．

① この結果から，個別にはずれ値となった，あるいはその候補となりそうな，回答者（サンプル）あるいは構成要素（語句）を，細かく探して除外してから再分析を行う場合（当然手間が面倒だが，より確かな傾向はわかる可能性は大きい）．
② 単純に閾値を変えて，構成要素数を絞り込み，再分析するとき．

ここでは，単純に②を採用してみよう．こうしたことで，以下のような**情報の損失**が生じることに注意して，分析を先に進める．

- 用いる構成要素（異なり構成要素数）が，230（語）から 139（語）に激減すること．
- つまり，丁度 2 回登場する語句が $230-139=91$（語）もあるが，これに該当する語句は以下の分析には参加しないこと．

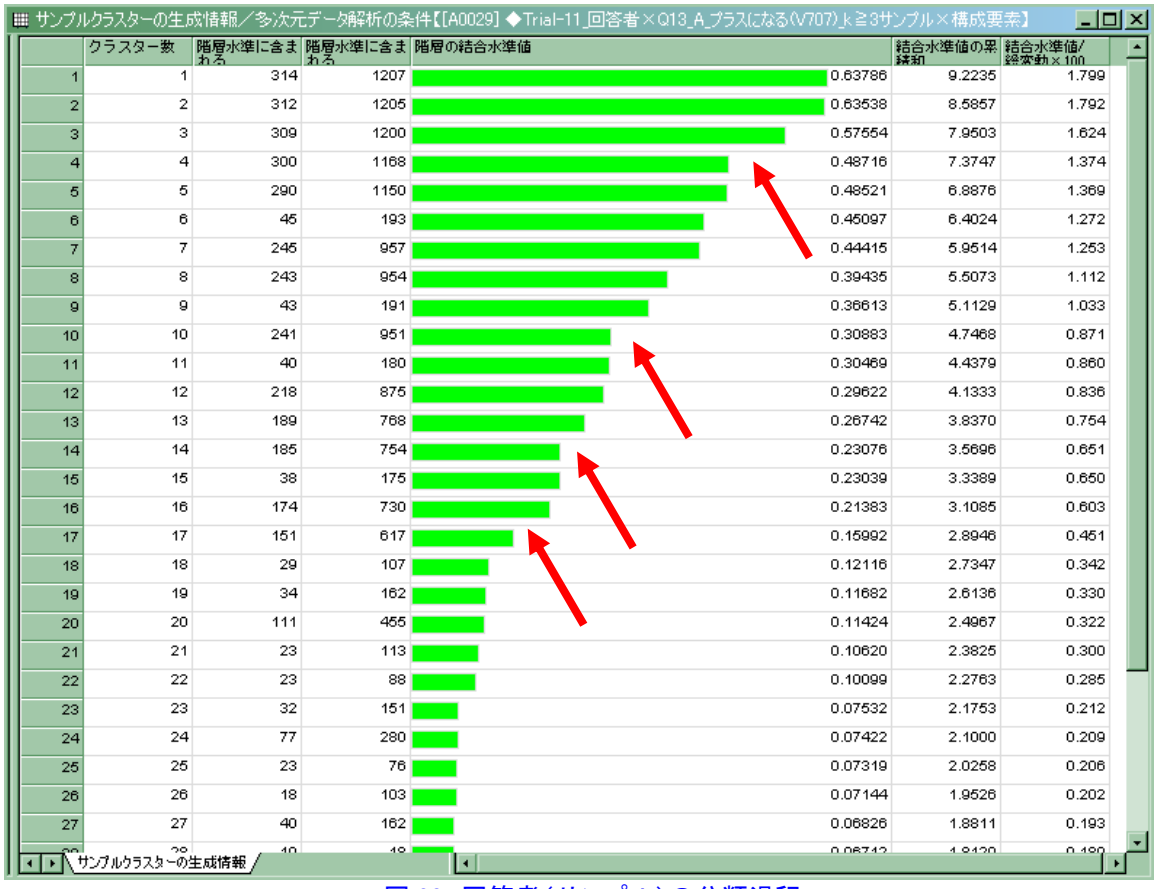
こうして得られた固有値ほかの情報が図 37 である．前の結果と比べると，1 という特異な固有値は無くなったものの，固有値の低減の様子は，依然としてなだらかである．寄与率の変化は，始めの 10 成分で累積 18.4%，20 成分までで 33.1%程度となり，前よりは若干多めとなる．しかし，データ表の寸法を考えると，この例にかぎらずおおむねこうした傾向にあることが多い．しかしここで注意することは，得られる成分スコアは行プロファイルあるいは列プロファイルの固有ベクトル要素を加重とする**合成変数**となっている，という点である．寄与率が小さいとはいえ，総変動に占める固有値（つまり各成分スコアの分散）の関与の程度は，この上位（大きいほう）の成分に情報が圧縮されていると考えていることである．寄与率は小さいが，対応分析や主成分分析といった次元縮約化型の分析手法は，変動の大きさを考慮して，固有値（成分スコアの分散）がなるべく大きいほうから，次元 1，次元 2，…と

順に成分スコアを構成してゆくので，初めの成分軸を採用することが意味を持つ．ここでは，このことを念頭に，閾値 3 以上の結果を用いて，つぎのクラスター化を検討する．

観察 5：クラスター化とクラスター数の検討

すでにみてきたように，対応分析法とその特性を利用したクラスター化は，所与のデータ表の行と列について同時的に分類を行えることにある（行と列に対して，同じ処理を行う）．

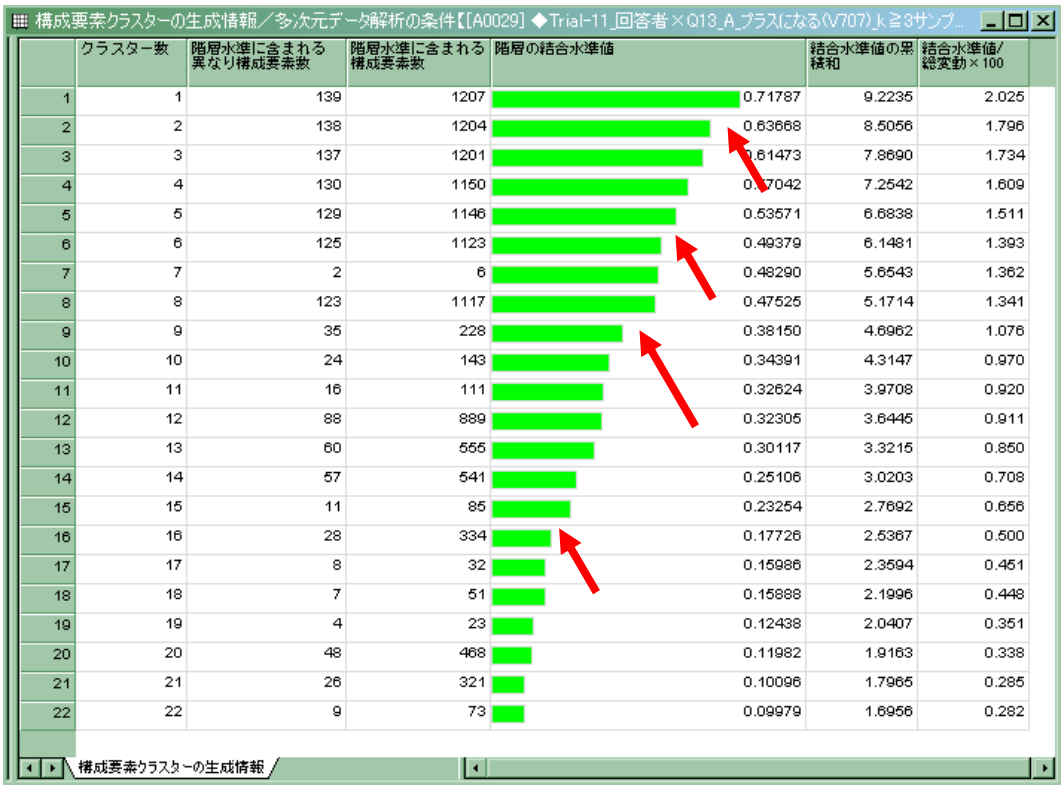
まず回答者（サンプル）のクラスター化履歴を観察しよう．図 38 から，**階層の結合水準**の変化が大きく変わるクラスター数の位置は「4～5 群」「10～11 群」「14～15 群」そして「17～18 群」あたりである．一方，構成要素（語句）のクラスター化履歴では，図 39 から，やはり「2 群」「6 群」そして「9～10 群」あたりだろう．ここでは，クラスター数として，回答者側を「15 群」に，構成要素側を「10 群」と設定してみよう[58]．



	クラスター数	階層水準に含まれる	階層水準に含まれる	階層の結合水準値	結合水準値の累積和	結合水準値/総変動×100
1	1	314	1207	0.63786	9.2235	1.799
2	2	312	1205	0.63538	8.5857	1.792
3	3	309	1200	0.57554	7.9503	1.624
4	4	300	1168	0.48716	7.3747	1.374
5	5	290	1150	0.48521	6.8876	1.369
6	6	45	193	0.45097	6.4024	1.272
7	7	245	957	0.44415	5.9514	1.253
8	8	243	954	0.39435	5.5073	1.112
9	9	43	191	0.36613	5.1129	1.033
10	10	241	951	0.30883	4.7468	0.871
11	11	40	180	0.30469	4.4379	0.860
12	12	218	875	0.29622	4.1333	0.836
13	13	189	768	0.26742	3.8370	0.754
14	14	185	754	0.23076	3.5696	0.651
15	15	38	175	0.23039	3.3389	0.650
16	16	174	730	0.21383	3.1085	0.603
17	17	151	617	0.15992	2.8946	0.451
18	18	29	107	0.12116	2.7347	0.342
19	19	34	162	0.11682	2.6136	0.330
20	20	111	455	0.11424	2.4967	0.322
21	21	23	113	0.10620	2.3825	0.300
22	22	23	88	0.10099	2.2763	0.285
23	23	32	151	0.07532	2.1753	0.212
24	24	77	280	0.07422	2.1000	0.209
25	25	23	76	0.07319	2.0258	0.206
26	26	18	103	0.07144	1.9526	0.202
27	27	40	162	0.06826	1.8811	0.193

図 38　回答者（サンプル）の分類過程

[58] 繰り返すが，クラスター数を決める厳密な指標はない．ここでも，結合水準の変化量と変化点に注目してこのように決める．


構成要素クラスターの生成情報／多次元データ解析の条件【[A0029] ◆Trial-11_回答者×Q13_A_プラスになる(V707)_k≧3サンプ...

	クラスター数	階層水準に含まれる異なり構成要素数	階層水準に含まれる構成要素数	階層の結合水準値	結合水準値の累積和	結合水準値/総変動×100
1	1	139	1207	0.71787	9.2235	2.025
2	2	138	1204	0.63668	8.5056	1.796
3	3	137	1201	0.61473	7.8690	1.734
4	4	130	1150	0.57042	7.2542	1.609
5	5	129	1146	0.53571	6.6838	1.511
6	6	125	1123	0.49379	6.1481	1.393
7	7	2	6	0.48290	5.6543	1.362
8	8	123	1117	0.47525	5.1714	1.341
9	9	35	228	0.38150	4.6962	1.076
10	10	24	143	0.34391	4.3147	0.970
11	11	16	111	0.32624	3.9708	0.920
12	12	88	889	0.32305	3.6445	0.911
13	13	60	555	0.30117	3.3215	0.850
14	14	57	541	0.25106	3.0203	0.708
15	15	11	85	0.23254	2.7692	0.656
16	16	28	334	0.17726	2.5367	0.500
17	17	8	32	0.15986	2.3594	0.451
18	18	7	51	0.15888	2.1996	0.448
19	19	4	23	0.12438	2.0407	0.351
20	20	48	468	0.11982	1.9163	0.338
21	21	26	321	0.10096	1.7965	0.285
22	22	9	73	0.09979	1.6956	0.282

構成要素クラスターの生成情報

図 39　構成要素(語句)の分類過程

観察 6:クラスター化の情報(基本情報のみ)の観察

さて，この回答者と構成要素とのクラスター化の構成（分類結果）を要約しよう．これが表 26 である．これをみると，クラスター・サイズが 10（人）以上のクラスターは，太字数字で指摘した 7 個ほど，サイズを 30（人）以上と絞ると 4 個ほどである．あとはいわば，はずれ値的なサイズの小さなクラスターである．意見の大勢はこれらのクラスターにまとまり，他は個別的な意見であろうと予想される．

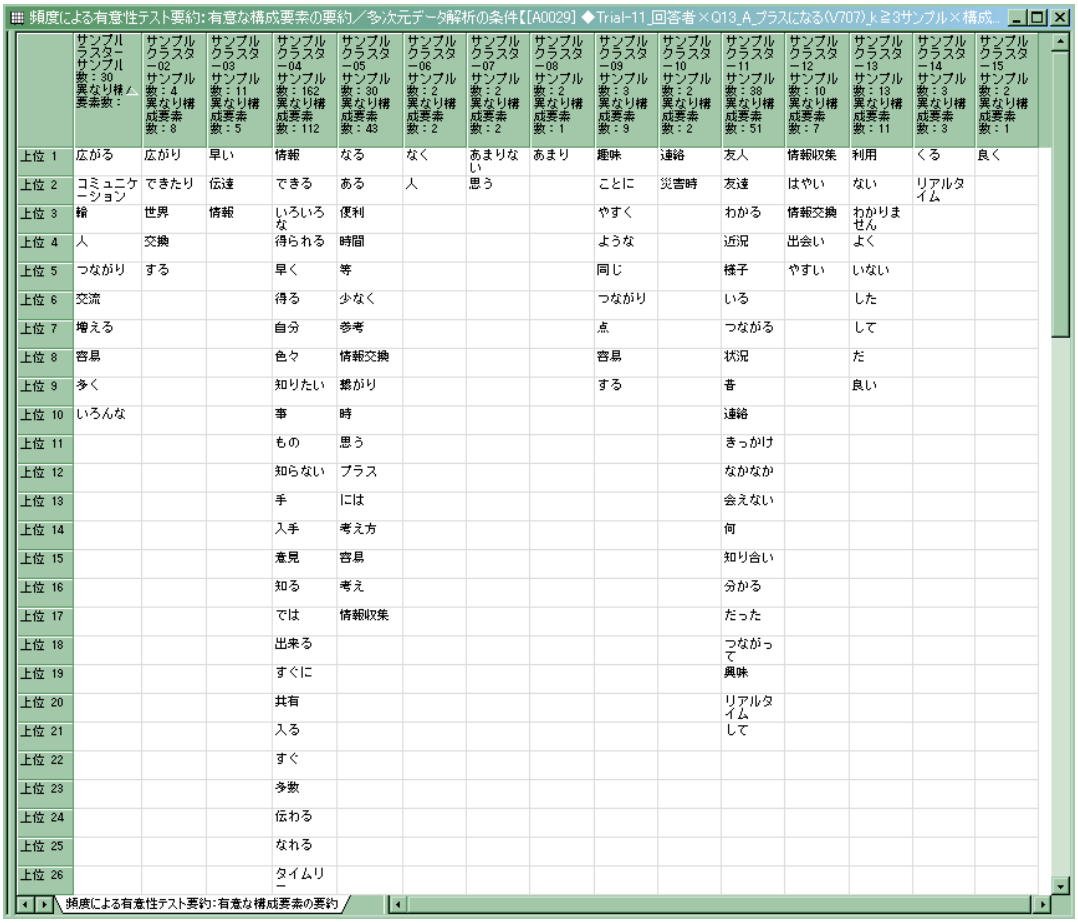
つぎに，構成要素の側のクラスター化情報を観察する．こちらは，数個（2〜3 個）のクラスター（クラスター05，06），せいぜいクラスター05 があり，その他にサイズの小さなクラスターがある．

表 26 「Q13_A:プラスと思うこと」の回答者分類と構成要素(語句)の分類

回答者(サンプル)のクラスター	サンプル数
01	**30**
02	4
03	**11**
04	**162**
05	**30**
06	2
07	2
08	2
09	3
10	2
11	**38**
12	**10**
13	**13**
14	3
15	2

構成要素(語句)のクラスター	構成要素数(語句数)
01	1
02	4
03	1
04	1
05	**86**
06	**25**
07	**12**
08	7
09	1
10	1

ここで，前の例と同じように，分析に用いた構成要素（語句）が，各クラスターでどのように有意となったかを要約した．これが図 40 である（上が上位から，下が下位から表示）．



頻度による有意性テスト要約:有意な構成要素の要約／多次元データ解析の条件【[A0029] ◆Trial-11_回答者×Q13_A_プラスになる(V707)_k≧3サンプル×構成...

	サンプルクラスター サンプル数：30 異なり構成要素数：	サンプルクラスター02 サンプル数：4 異なり構成要素数：8	サンプルクラスター03 サンプル数：11 異なり構成要素数：5	サンプルクラスター04 サンプル数：162 異なり構成要素数：112	サンプルクラスター05 サンプル数：30 異なり構成要素数：43	サンプルクラスター06 サンプル数：2 異なり構成要素数：2	サンプルクラスター07 サンプル数：2 異なり構成要素数：2	サンプルクラスター08 サンプル数：2 異なり構成要素数：1	サンプルクラスター09 サンプル数：3 異なり構成要素数：9	サンプルクラスター10 サンプル数：2 異なり構成要素数：2	サンプルクラスター11 サンプル数：38 異なり構成要素数：51	サンプルクラスター12 サンプル数：10 異なり構成要素数：7	サンプルクラスター13 サンプル数：13 異なり構成要素数：11	サンプルクラスター14 サンプル数：3 異なり構成要素数：3	サンプルクラスター15 サンプル数：2 異なり構成要素数：1
上位 1	広がる	広がり	早い	情報	なる	なく	あまりない	あまり	趣味	連絡	友人	情報収集	利用	くる	良く
上位 2	コミュニケーション	できたり	伝達	できる	ある	人	思う		ことに	災害時	友達	はやい	ない	リアルタイム	
上位 3	輪	世界	情報	いろいろな	便利				やすく		わかる	情報交換	わかりません		
上位 4	人	交換		得られる	時間				ような		近況	出会い	よく		
上位 5	つながり	する		早く	等				同じ		様子	やすい	いない		
上位 6	交流			得る	少なく				つながり		いる		した		
上位 7	増える			自分	参考				点		つながる		して		
上位 8	容易			色々	情報交換				容易		状況		だ		
上位 9	多く			知りたい	繋がり				する		音		良い		
上位 10	いろんな			事	時						連絡				
上位 11				もの	思う						きっかけ				
上位 12				知らない	プラス						なかなか				
上位 13				手	には						会えない				
上位 14				入手	考え方						何				
上位 15				意見	容易						知り合い				
上位 16				知る	考え						分かる				
上位 17				では	情報収集						だった				
上位 18				出来る							つながって				
上位 19				すぐに							興味				
上位 20				共有							リアルタイム				
上位 21				入る							して				
上位 22				すぐ											
上位 23				多数											
上位 24				伝わる											
上位 25				なれる											
上位 26				タイムリー											

頻度による有意性テスト要約:有意な構成要素の要約

頻度による有意性テスト要約:有意な構成要素の要約／多次元データ解析の条件【[A0029] ◆Trial-11_回答者×Q13_A_プラスになる(V707)_k≧3サンプル×構成要...

	サンプルクラスター サンプル数：30 異なり構成要素数：	サンプルクラスター02 サンプル数：4 異なり構成要素数：8	サンプルクラスター03 サンプル数：11 異なり構成要素数：5	サンプルクラスター04 サンプル数：162 異なり構成要素数：112	サンプルクラスター05 サンプル数：30 異なり構成要素数：43	サンプルクラスター06 サンプル数：2 異なり構成要素数：2	サンプルクラスター07 サンプル数：2 異なり構成要素数：2	サンプルクラスター08 サンプル数：2 異なり構成要素数：1	サンプルクラスター09 サンプル数：3 異なり構成要素数：9	サンプルクラスター10 サンプル数：2 異なり構成要素数：2	サンプルクラスター11 サンプル数：38 異なり構成要素数：51	サンプルクラスター12 サンプル数：10 異なり構成要素数：7	サンプルクラスター13 サンプル数：13 異なり構成要素数：11	サンプルクラスター14 サンプル数：3 異なり構成要素数：3	サンプルクラスター15 サンプル数：2 異なり構成要素数：1
下位 1	いろいろな			友達	人						情報		情報		
下位 2	情報			友人	できる						いろいろな		できる		
下位 3	事			広がる	情報						交流		人		
下位 4	わかる			なる	わかる						広がる				
下位 5				早い	友達						意見				
下位 6				情報収集	広がる						得られる				
下位 7				コミュニケーショ							人				
下位 8				輪							早く				
下位 9				情報交換							自分				
下位 10				連絡							色々				
下位 11				近況							ない				
下位 12				わかる							なる				
下位 13				様子											
下位 14				容易											
下位 15				わかりません											
下位 16				つながり											
下位 17				利用											

頻度による有意性テスト要約:有意な構成要素の要約

図 40　構成要素（語句）の分類結果（上位，下位の一部を切り取り）

表 27　分類結果の統計量の要約

設定条件	回答者（サンプル）のクラスター化	構成要素（語句）のクラスター化
指定した成分数（成分軸の数）	138	
総変動（固有値の総和）（カイ二乗統計量）	35.4488	
クラスター化で用いる成分数	15	15
指定したクラスター数	15	10
クラスター内変動の和	3.2618	4.2844
クラスター間変動	5.9618	4.9391
クラスターの総変動（指定した成分数までのカイ二乗統計量）	9.2235	9.2235
クラスター間変動比	0.6464	0.5355
クラスター間変動比(%)	**64.64**	**53.55**

観察 7:回答者のクラスター化情報の吟味

ここで得た，回答者（サンプル）の 15 群の内容，つまり**元の自由回答（原文）**ではどう記述しているのだろうか．この種の情報は紙数をとるのが一般的であるが，この質問「Q13_A：プラスと思うこと」は，実は回答記入量は**さほど多くはない**．それでもなお，かなりのボリュームとなるので，すべての回答者の回答を載せることはせずに，ある（統計的な）有意性テストを行い，クラスター内で意味のある回答の順にソートして一覧としてみる．表 28 がこの要約である[59]．たとえば，クラスター1 には 30 人の回答者が含まれるので，すべての回答者を有意となった検定値の大きさの順に特徴的な回答を示した，ということである．クラスター4 のサイズは 162（名）となるが，こういう場合は有意の順に上からある人数（ここでは 90 名）を示した．他のクラスターについても同様に要約した．また併せて「性年齢区分」の情報（質的変数）も表の右欄に加えた．

なお，前に述べたように，自由回答原文の分かち書きのあと，ほとんど辞書編集などは**行っていない**ことに注意しよう．また，この質問文の問い方（文言，ワーディング）によっては，回答記入量（つまり文字数）が少なくなる[60]．前の「Q19.情報の考え方」の場合は，構成要素の頻度分布も異なり，また書き込みの量も総じて多い．「Q19.…」のほうが，（質問の文言がやや概念的で）回答時に考えさせる要素が多く，回答者はかなり記入作業に気配りをしている様子もみえる．これは，**質問文の問い方**によって回答内容に差が出る，ということを示唆している[61]．

一方，「Q13…」の 2 つの自由回答質問は，回答者の日常体験や日ごろ見聞きしている情報からの類推や連想がただちに回答に影響することが予想される．そして記入量も短くまとまる傾向にある．このように質問文が自由回答に影響を及ぼすことは，調査票設計当初からある程度は予想されていたが，このように具体的な分析で比べると，（この 1 つのウェブ調査という限られた中での事象ではあるが，また過去の経験則として），**“自由回答質問はワーディングにより自由記述の内容は影響を受ける”**と言われている現象の I 部が観察されたといえる[62]（少なくとも，質問文の設計の適否が必要条件ではある）．

得られた各クラスターについて，その内容をここで個別に吟味するまでもなく，それぞれ

---

[59] WordMiner が出力した情報を再整理した．WordMiner は，クラスター別に 1 つのシートとして出力し，またこれらの情報をテキスト・ファイルとしてエクスポートする．それを若干編集したものである．

[60] 過去の例：「あなたにとって一番大切なものは何ですか」「一番好きな食べ物は」などは非常に短い回答となる．

[61] このことは，調査方法論における質問文作成に関わる重要な基本要素である．たとえば，大隅監訳（2011），Fowler（1995）などを参照．

[62] この調査に限らず，一般に自由回答は質問文のワーディングやその前後の質問文の影響を受けることが指摘されている．

がある特徴を示していることがわかるであろう．ここでよく行われる操作に，得られたクラスターの内容をみて（類推して），そのクラスターの特徴を意味付けること（いわゆるネーミングあるいはラベリング）を行うことがある[63]．しかし，クラスターを排反的に分けることは，実はあまり現実的ではない（実態にそぐわない）．つまり，多くの場合，クラスター間の差違はそう顕著ではなく（クラスターの存在は曖昧で），はっきりとクラスター状（房状，塊状）に分かれる（分類する）ことを期待することはむずかしいと考えるのが妥当である[64]．

表 28 「Q13_A:プラスと思うこと」への自由回答の類型化（15 群）

クラスター01	検定値	検定値(無調整)	「Q13_A_ＳＮＳがプラスになると思うこと」の自由回答（原文）	性年齢区分
1	5.8970	5.8970	コミュニケーション	10.40 代女性
2	3.2892	3.2892	交流の輪が広がる	07.10 代女性
3	3.0534	3.0534	人とのコミュニケーションが広がる	03.30 代男性
4	2.9035	3.0828	価値観が広がる	04.40 代男性
5	2.8344	2.9816	友達の輪が広がる	01.10 代男性
6	2.1221	2.1221	人と人とのつながりが容易に広がる	02.20 代男性
7	1.9927	1.9927	情報共有、人とのつながりが広がる	03.30 代男性
8	1.7421	1.7421	コミュニティーが広がること。	07.10 代女性
9	1.7421	2.0366	情報発信の場が広がる	04.40 代男性
10	1.6554	1.6554	人の輪が広がる。情報が早い。	09.30 代女性
11	1.4759	1.4759	知識や情報が広がる	09.30 代女性
12	1.4743	1.7516	コミュニケーションが取りやすい	02.20 代男性
13	1.3728	1.3072	気軽に人とコミュニケーションができる	07.10 代女性
14	1.2884	1.2884	共通の話題の交友の輪が広がると思う	08.20 代女性
15	1.2800	1.2800	交流が増える	05.50 代男性
16	1.2518	1.2518	人と人との新たな交流が増える	04.40 代男性
17	1.2510	1.1840	人と人とのつながりが容易になる	01.10 代男性
18	1.2284	1.2284	多くの人が交流しあえる	10.40 代女性
19	1.1794	1.1794	コミュニケーションが気楽の取れる	11.50 代女性
20	1.1367	1.1125	人の輪が広がる。、他の人の知恵を借りることができる。、	05.50 代男性
21	1.1178	1.1398	いろんな人と出会い、いろんな情報も手に入るし、輪が広がる。	10.40 代女性
22	1.0701	1.1506	新しい情報がいちはやく分かる。友達の輪が広がる。	09.30 代女性
23	0.9197	0.8817	より多くの情報を得られる。、交流が広がる。	09.30 代女性
24	0.8796	0.7648	つながりができることころ	02.20 代男性
25	0.8711	0.8711	情報量が多くなる、知識のはばが広がる	06.60 代男性
26	0.7466	0.7802	時事的な出来事の情報が早く知れる･情報を共有してコミュニケーションが広がる	07.10 代女性
27	0.7371	0.8291	常に友達とのコミュニケーションがとれる	02.20 代男性
28	0.6188	0.6188	たくさんの人と知り合える。	09.30 代女性
29	0.6086	0.6107	世界が広がる。現代の孤独感を紛らわせることができる？、どんな過疎地でもコミュニケーションができる。など	04.40 代男性
30	0.1895	0.1895	コミュニケーション手段の多様化､特定の趣味関係者との交流を増やせる	04.40 代男性
クラスター02				
1	1.5696	1.5830	情報の広がり	09.30 代女性
2	1.2693	1.2693	世界が広がり視野も広くなる	12.60 代女性
3	0.5273	0.5273	希少な悩みなど相談できたりします。	11.50 代女性
4	0.4549	0.5010	世界が広がり、知らなかった地域の人と友達になったり、、情報を交換できたり、励ましてもらったり、、心が上向きになり、その日元気に過ごすことができたりする。	11.50 代女性
クラスター03				
1	3.6931	3.6931	情報が早い	02.20 代男性
2	3.6931	3.6931	情報が早い	04.40 代男性
3	3.6931	3.6931	情報が早い	11.50 代女性

63 この操作を無定見に行うことはあまり薦められない．またこの操作は付与した標識をあらたな“質的変数”（名義尺度的に）用いることを意味する．

64 用意したクラスター評価基準（最適化基準）にしたがって，あるアルゴリズムで分類した，しかも排反的に重なりのないクラスターを作った（クラスター化）ということであって，このことが厳密な意味での分類を意味してはいないからである．そういう意味で，クラスター化はきわめてヒューリスティックな操作である．

4	3.2487	3.2487	情報の伝達が早い	03.30 代男性
5	2.7699	2.7699	情報が早い。	03.30 代男性
6	2.7699	2.7699	情報が早い。	11.50 代女性
7	2.2159	2.2159	情報の周りが早い	04.40 代男性
8	2.2066	2.2066	情報の速い伝達	04.40 代男性
9	1.2609	1.2609	情報が高速に伝達される。	04.40 代男性
10	0.6455	0.7198	会話の如く意志の伝達が出来る	06.60 代男性
11	0.6155	0.6155	情報が早い。、距離的、時間的に離れた人との縁が切れにくい。	04.40 代男性
クラスター04				
1	4.7176	4.7176	情報	02.20 代男性
2	2.8460	2.8460	いろいろな情報が得られる	12.60 代女性
3	2.6431	2.6431	情報が得られる	03.30 代男性
4	2.6431	2.6431	情報が得られる	12.60 代女性
5	2.6346	2.6346	早く情報が入手できる	04.40 代男性
6	2.5659	2.5659	情報が共有できる	03.30 代男性
7	2.3701	2.3701	情報を得る	02.20 代男性
8	2.0867	2.0867	情報の共有	01.10 代男性
9	2.0867	2.0867	情報の共有	10.40 代女性
10	2.0867	2.0867	情報の共有	06.60 代男性
11	2.0149	2.0149	知りたい情報が得られる。	03.30 代男性
12	1.9823	1.9823	幅広い情報が得られる	03.30 代男性
13	1.9642	1.9642	情報の収集	06.60 代男性
14	1.9597	2.0110	いろいろな人と情報を共有できる	10.40 代女性
15	1.8973	1.8973	いろいろな情報が直ぐ得られる。	05.50 代男性
16	1.8802	1.8802	情報がすぐ得られる。	05.50 代男性
17	1.8209	1.8209	タイムリーな情報を得られる	04.40 代男性
18	1.7796	1.9135	情報が簡単に入手できる	10.40 代女性
19	1.7665	1.7665	いろいろな視点からの情報が入手できる	05.50 代男性
20	1.7465	1.7465	自分が知らない情報を得られる。	05.50 代男性
21	1.7429	1.7429	自分の知らない情報を得る事ができる。	11.50 代女性
22	1.7394	1.7551	みんなでいろいろな情報を交換できる	01.10 代男性
23	1.6441	1.6441	いろいろな意見が聞ける	11.50 代女性
24	1.6217	1.6217	いろいろな意見を知ることができる	10.40 代女性
25	1.6028	1.6028	知りたい情報を継続的に入手できる。	06.60 代男性
26	1.5859	1.7513	幅広く情報が得られる。	03.30 代男性
27	1.5859	1.5859	さまざまな情報が得られる	05.50 代男性
28	1.5361	1.5361	タイムリーな情報を得ることができる	04.40 代男性
29	1.5253	1.6435	思いもかけない情報が入手できる	10.40 代女性
30	1.5253	1.5253	オンタイムで情報を入手できる。	05.50 代男性
31	1.4916	1.7613	いろいろな人と交流できる	11.50 代女性
32	1.4558	1.4558	いろいろな情報が入って来る事。	05.50 代男性
33	1.4207	1.4207	新聞などよりも早く情報を得る事ができる	02.20 代男性
34	1.3774	1.3774	普段接する事の無い情報を得る事ができる。	11.50 代女性
35	1.3694	1.5073	すぐに情報を手に入れられる	09.30 代女性
36	1.3348	1.2503	リアルタイムな情報を入手	08.20 代女性
37	1.3239	1.3239	各種情報が広く知れ渡る事ができる。	06.60 代男性
38	1.3215	1.3215	重要な情報が得られる。	04.40 代男性
39	1.2912	1.3609	いろいろな人の意見や考え、いろいろな情報を得られるところ。	09.30 代女性
40	1.2649	1.3047	いろいろな人と知り合え、情報も得られる。	10.40 代女性
41	1.2631	1.2753	不特定多数から、情報を交換できる。	06.60 代男性
42	1.2593	1.2065	リアルタイムに色々な情報を得られる。	11.50 代女性
43	1.2520	1.2520	情報の共有される。　⇒「情報の共有がなされる。」の誤記	03.30 代男性
44	1.2309	1.2309	色々な情報・意見が手に入る。	08.20 代女性
45	1.2013	1.2487	自分に必要な情報がすぐ手に入る	05.50 代男性
46	1.1852	1.2116	今現在の情報をタイムリーに入手できる。	04.40 代男性
47	1.1794	1.3862	幅広く情報が集められる	04.40 代男性
48	1.1794	1.1794	情報の早期習得	05.50 代男性
49	1.1768	1.1768	知りたいことを早く知ることができる。	09.30 代女性
50	1.1513	1.1513	色々なタイプの情報が手に入る	08.20 代女性
51	1.1448	1.1448	知りたい情報がすぐわかる	02.20 代男性
52	1.1352	1.1711	いろいろな人の意見を知ることができる。	10.40 代女性
53	1.1333	1.1333	最新の情報を得る事が出来る、すぐに欲しい情報が見られる、	09.30 代女性
54	1.1328	1.1328	いつでもどこでも情報が得られる	10.40 代女性
55	1.1254	1.3092	いろいろな情報を手軽に手に入れられる。	03.30 代男性

56	1.1069	1.1395	いろいろな人、国を問わず情報が素早く取得できる	06.60 代男性
57	1.1034	1.1865	いろいろな情報が簡単に知る事がが良い。	12.60 代女性
58	1.0947	1.0947	いろいろな事をしれる。	04.40 代男性
59	1.0856	1.0856	災害などのとき情報が早く伝わる	10.40 代女性
60	1.0803	1.0803	いろいろな意見がうかがえる。	05.50 代男性
61	1.0780	1.0357	情報をより早く入手できたり、調べたり出来る。	04.40 代男性
62	1.0625	1.1239	今、自分に必要な情報を早く得られる。、	05.50 代男性
63	1.0485	1.0485	情報の収集が早く、多くなる。	12.60 代女性
64	1.0316	1.0499	良い情報がすぐ広まる。	11.50 代女性
65	1.0271	1.0271	マスメディアが隠す情報が手に入る	02.20 代男性
66	1.0097	1.0097	情報が瞬時に広く伝えられる。	06.60 代男性
67	1.0009	1.2483	確実に交流できる	02.20 代男性
68	0.9959	1.2206	知らない人と交流できること	09.30 代女性
69	0.9917	1.1296	誰でも意見が発信できる	06.60 代男性
70	0.9811	1.0845	タイムリーな情報を手に入れられる。	03.30 代男性
71	0.9740	0.9740	・情報が早く伝わる。、・知りたいことの答えがすぐに得られる。	04.40 代男性
72	0.9707	1.0459	手軽に情報を手に入れることができる。	12.60 代女性
73	0.9690	0.9690	１つの情報を多角的に認識できる。	03.30 代男性
74	0.9690	0.9690	率直な情報を得ることが出来る。	11.50 代女性
75	0.9658	0.9658	色々な情報が すぐにわかる	09.30 代女性
76	0.9652	1.0335	色々な考えを知る事が出来る	12.60 代女性
77	0.9435	0.9435	情報の流通が早くなる	04.40 代男性
78	0.9435	0.9435	情報の広まり方が早くなる	05.50 代男性
79	0.9358	0.9358	情報が共有しやすい、新たな事を知ることが出来る	02.20 代男性
80	0.8943	1.0210	いろんな人から いろんな情報が入る	09.30 代女性
81	0.8890	0.9972	色んな人の情報を気軽に知ることができる	08.20 代女性
82	0.8769	0.9075	不特定多数と情報の収集や交換ができたりする､知らない人と色々と話せたりできる	02.20 代男性
83	0.8637	0.8637	いろいろなことが判る	08.20 代女性
84	0.8418	0.7185	世界的な情報が瞬時に分かる	06.60 代男性
85	0.8398	0.8398	すぐに知りたい情報がはいるので便利。	03.30 代男性
86	0.8323	0.8424	今必要と思う情報をすぐに、調べることが出来るし、信頼できる情報も多い	11.50 代女性
87	0.8261	0.8261	情報を共有できて、オープンになれる。	08.20 代女性
88	0.8197	0.9611	いままででは知り合えなかったひとと交流できる	03.30 代男性
89	0.8000	0.8000	情報を共有できる、ニュースに対する意見がわかる。	03.30 代男性
90	0.7928	0.8546	情報が入りやすい。、気軽に交流できる。	07.10 代女性
クラスター05				
1	4.0291	4.0291	便利	10.40 代女性
2	2.8252	2.8252	タイムラグが少なくなる	05.50 代男性
3	2.7794	2.7794	便利になると思う	06.60 代男性
4	2.0222	2.0222	繋がりが簡易になる	03.30 代男性
5	1.8835	1.8835	一人悩む事が少なくなる。	12.60 代女性
6	1.6195	1.6195	お店の情報源になる	08.20 代女性
7	1.5157	1.5212	いろいろな人間の考え方が参考になる	06.60 代男性
8	1.3616	1.3810	いろいろな情報が得られる、調べる時間が少なくなる	12.60 代女性
9	1.3496	1.3496	情報量が豊富になること	10.40 代女性
10	1.3122	1.4514	情報を得やすくなる	02.20 代男性
11	1.1965	1.1965	情報交換が容易になる。、時間の概念が変わる。	05.50 代男性
12	1.1792	1.2691	一般人の意見が聞けて参考になる。	10.40 代女性
13	1.0994	1.0994	時間がある時いつでも利用が出来る	05.50 代男性
14	1.0122	1.0122	日々のストレス解消や楽しみになる。	09.30 代女性
15	0.9034	0.9483	時には害となる項目もあるが､いろいろな意見、考え方など見聞きできよい。	06.60 代男性
16	0.8997	1.1660	意見のやり取りが簡単になる。	08.20 代女性
17	0.8311	0.9161	他人の考えが分かり、良くも悪くも参考になる。	06.60 代男性
18	0.8128	0.8742	個人による情報発信が容易になる。、情報収集の際にも選択肢が増える。、人間関係が豊かになる可能性がある。	02.20 代男性
19	0.8127	0.8813	簡単に情報収集できることはプラスになると思う。、	12.60 代女性
20	0.8058	0.8058	娯楽の一つである	10.40 代女性
21	0.6748	0.7328	過去つながりのあった方との再会の機会となる。	05.50 代男性
22	0.6748	0.6748	みんなで協力すれば形に成り、人助けになる。	12.60 代女性
23	0.4977	0.5142	必要な情報の入手において､従来の図書館等での調査・調査会社等からの資料購入等に比べ時間・経費等が大幅に低減・節約でき､、社会	06.60 代男性

			経済的な効果が非常に大きい。、特に外国の資料等を考えると大いに社会経済的にプラスになる。	
24	0.4426	0.4426	余り無いと思う	12.60 代女性
25	0.4155	0.4155	知りたかった情報が分って便利。	10.40 代女性
26	0.2770	0.2770	情報が何処でも確認できて便利なところ	10.40 代女性
27	0.2266	0.2807	色々な情報が出回る。有益な情報もある。、災害時などは便利だというイメージ。	10.40 代女性
28	0.2024	0.2024	個人個人の考えが表せていいし、社会との繋がりが持てる。	06.60 代男性
29	0.1802	0.1619	モラルを守って使う分には、旧友との交流やワールドワイドな情報交換が可能な事。	09.30 代女性
30	0.0034	0.0529	同じ境遇の方々と情報交換ができること。	08.20 代女性
クラスター06				
1	1.4058	1.4058	利用経験なく判らない	06.60 代男性
2	0.4781	0.4781	場所に関係なく、世界中の人と繋がれること	02.20 代男性
クラスター07				
1	2.1087	2.1087	あまりない。	05.50 代男性
2	2.0758	2.0758	あまりないと思う	03.30 代男性
クラスター08				
1	2.1539	2.1539	あまり無い	10.40 代女性
2	2.1539	2.1539	あまり感じない	10.40 代女性
クラスター09				
1	0.9124	0.9124	同じような趣味を持った仲間を探すことが容易な点	02.20 代男性
2	0.8113	0.8113	趣味などの知人を増やせる	03.30 代男性
3	0.5754	0.5754	共通の趣味を持つ者通しがつながりやすく、オフ会などで集まることにより経済が活性化する	08.20 代女性
クラスター10				
1	1.8372	1.8372	災害時の連絡、海外時の連絡	03.30 代男性
2	1.3999	1.3999	地震などの災害時の連絡	09.30 代女性
クラスター11				
1	2.6267	2.6267	知り合いの近況がわかる	11.50 代女性
2	2.3781	2.1084	友人ができる	03.30 代男性
3	2.3763	2.3763	友達とつながること	03.30 代男性
4	2.1768	2.1768	友達の近況が分かること	01.10 代男性
5	1.9931	2.0104	あまり会わない友達の様子がわかる。	09.30 代女性
6	1.9677	1.9677	相手の様子がわかる	11.50 代女性
7	1.7144	1.7505	離れている友人の近況がわかる。、今まで疎遠だった友達と連絡がとれる。	10.40 代女性
8	1.5842	1.5842	昔の友達とつながれる事	09.30 代女性
9	1.5429	1.5429	ネットをとおして友達の状況が、わかる。	11.50 代女性
10	1.5266	1.5266	古い友達とつながれる	08.20 代女性
11	1.4269	1.2650	友人と近況報告ができる	03.30 代男性
12	1.3961	1.4223	普段なかなか会えない友人とつながっていられる。お互いの近況を伝えられる。	09.30 代女性
13	1.3859	1.3859	周りの友人たちの近況がわかるところ	08.20 代女性
14	1.3056	1.3056	友人の最近の様子が、日記などでわかること。	09.30 代女性
15	1.2895	1.2592	なかなか会えない友人の近況などがわかる。同じような興味を持つ人と知り合うきっかけになる。	03.30 代男性
16	1.2319	1.2319	参考になったり、近況がわかる。	10.40 代女性
17	1.1882	1.1882	昔の友達などとまたつながれる。	04.40 代男性
18	1.1817	1.1898	友達のリアルタイムな状況が知れる。、昔の友達とメッセージをかわせる。	07.10 代女性
19	1.1212	1.1395	友人が日々どんな事をしているか会わなくても様子が分かるのでつながっている気になります。	09.30 代女性
20	1.1011	1.1354	友達とミクシィをやっているので、つぶやきで今どこで何をしているかわかる。	01.10 代男性
21	1.0607	1.0179	ご無沙汰している友達を探す事が出来る。	11.50 代女性
22	0.9873	0.9873	知りありの動向がわかる	04.40 代男性
23	0.9836	0.9739	遠方に住んでいる友人や忙しくてなかなか会えない友人の様子を知ることができるので、つぶやきや日記の内容をきっかけに実際に会うことにつながること。	08.20 代女性
24	0.9076	0.9076	昔の知り合いに再開しやすい、	02.20 代男性
25	0.8752	0.9199	知りたいことがリアルタイムに直ぐにわかる	04.40 代男性
26	0.8580	0.9101	友人とメールで連絡するより手っ取り早い気がする。	09.30 代女性
27	0.7348	0.7348	人とつながる	04.40 代男性

28	0.7240	0.6768	友達や家族に気軽に近況を報告したりできる。、友達や家族のことも知ることができる。	09.30 代女性
29	0.6837	0.6986	情報がすぐにいっぱい得られる、友達の近況がわかる、もしスマホをなくしたりしても連絡がとれたりする	02.20 代男性
30	0.6018	0.6120	友人や知り合いと連絡をとらなくても、相手の状況がわかる。メールが携帯で直接よりも、ＳＮＳを通しての方が機関があいていても取りやすい。	09.30 代女性
31	0.5492	0.5630	他人が何を思っているかが解る。	06.60 代男性
32	0.5488	0.5153	海外の友人とも気軽に連絡がちれたり、近況報告ができる。⇒「…連絡がとれたり…」の誤記.	08.20 代女性
33	0.5370	0.5149	異業種の人と情報交換ができる。、音信不通だった学生時代の友人に連絡が取れるようになった	04.40 代男性
34	0.5367	0.4557	連絡の取れてなかった人ともつながることができる	02.20 代男性
35	0.5263	0.5263	興味が無いので何とも思わない	03.30 代男性
36	0.4994	0.5912	見たい人の現在の状況が分かる	08.20 代女性
37	0.4082	0.4082	情報がわかる	12.60 代女性
38	0.2382	0.2018	しばらく会っていない友人と再び出会うきっかけになる。、イベントなどの情報を知るのに役立つ。、	09.30 代女性
クラスター12				
1	6.3328	6.3328	情報収集	05.50 代男性
2	6.3328	6.3328	情報収集	06.60 代男性
3	3.7345	3.7345	情報交換	04.40 代男性
4	3.3558	3.3558	情報交換、情報収集	05.50 代男性
5	2.1109	2.1109	情報収集ができる	02.20 代男性
6	1.5339	1.5339	情報収集、仲間との出会い	08.20 代女性
7	1.2448	1.2448	ストレス解消、情報交換	09.30 代女性
8	1.1354	1.1354	情報がかなりはやい	09.30 代女性
9	0.8221	0.8221	情報収集がはやいこと。イベント関係の告知がしやすい、広まりやすいこと。	02.20 代男性
10	0.6177	0.6177	出会いや発見が増えて、情報の取得もはやいこと	08.20 代女性
クラスター13				
1	5.5114	5.5114	わかりません	12.60 代女性
2	5.1726	5.1726	よくわかりません	12.60 代女性
3	3.2289	3.2289	利用していないのでよくわかりません。	04.40 代男性
4	2.4919	2.4919	利用したことがないのでよくわからない	12.60 代女性
5	1.9743	1.9743	申し訳ありません。、利用していないためわかりません	06.60 代男性
6	1.8877	1.8877	利用したことがないのでわからない。	11.50 代女性
7	1.8824	1.8824	特にない	09.30 代女性
8	1.8824	1.8824	特にない	04.40 代男性
9	1.8824	1.8824	特にない	12.60 代女性
10	1.5321	1.5321	利用していないので、特に無し	05.50 代男性
11	1.4891	1.4891	利用しないのでわからない	12.60 代女性
12	1.4118	1.4118	別にない。	07.10 代女性
13	0.6632	0.6820	ブログやツイッターなどは利用したことがないので、よくわかりませんが、テレビでタレントさんたちが自分の考えていることを、ファンの人々に知ってもらうのは良いことだと思います。	12.60 代女性
クラスター14				
1	1.2396	1.2396	リアルタイムで返事が返ってくる	04.40 代男性
2	0.8992	0.8992	世の中の流れが伝わってくる	02.20 代男性
3	0.6744	0.7645	人との接点が変わってくること	03.30 代男性
クラスター15				
1	2.2294	2.2294	良く分からない	02.20 代男性
2	2.2294	2.2294	良く分りません	06.60 代男性

それでもなお，それぞれの**クラスター化**の内容を観察することが必要だろうから，ステレオタイプとなることを承知のうえで，その特徴を表 29 に要約した．

語句の編集をほとんど行っていないことから，あきらかに分類内容には若干の矛盾や不具合も観察される（これが，自由回答の日常的な現象である）．もちろん，誤記や日本語としての表記のゆらぎなどはそのままである．さらに再分析（であり細分析）を行うことで，類型化の内容がより洗練されたものとなるのだが，大抵は，とくに実用上はそう手間暇をかける余裕もないだろうから，ここで示した程度で解釈を行うことでもよいだろう．かりにもう少

し踏み込んで，**細分析を行う**ならば，たとえば以下のようなことに留意するとよい．

- クラスター7，8，そしてクラスター13，15 のように，「意見がない」「わからない」など，質問に対して前向きに発言しない，あるいは協力参加の気持ちがみえない回答は，除外する，という考え方．
- しかもこの回答の影響を除外して（これらの意見を伏せて）他の回答者の傾向をさらに分析したい（強調したい），と考えるかもしれない．このようなとき，これらの語句を「削除辞書」として登録し，辞書編集を行うという手段もある．
- ここでの注意点は，質問内容によっては，「わからない」が重要な鍵となる場合がある，という点である．つまり，自由回答質問の意図が回答者に“正しく理解されずに”，結果として「わからない」などが増える場合である．つまり，こうした回答（記述）がどの程度の頻度で登場するか，に注意することである．
- かりにここで，個々の文言（記述内容）には，微妙に意味の違いが感じられる，あるいは読み取れるから，それを知ることが必要だ，と考えるならば，これはクラスター化の限界を越えた議論となる（別のアプローチの検討が必要である）．
- また，意味解釈が付けにくいクラスターも当然起こりえる．これには無理な解釈を付けずに，そのまま素直に記述内容を読み取ればよいだろう．
- ここで注意することは，分析時には，ここに挙げた自由回答文（原文）を“すべて用いて行ったわけではない”ということである．ここでいうクラスター化では，その判断まではできない，ということである（文章を“内容分析”的に子細に分析を行うこととは異なる）．
- 言い換えると，繰り返しとなるが，調査であれば**“自由回答の質問文の作り方”**と，それ以外に**“適切な選択肢型質問”**を用意し，両者を併用するということである．
- さらに言うまでもなく，用いる**“調査方式”**（調査モード）も非常に関連がある．

表 29　クラスターの類型化と特徴

クラスター	クラスターサイズ	主な特徴
クラスター1	30	「コミュニケーション」という語句に代表されるグループ．「人との交流」「友だち，人と人とのつながり」など．
クラスター2	4	「情報の広がり」「交流」など　(*) 記述内容がやや曖昧
クラスター3	11	情報の伝達，その迅速性
クラスター4	162	「情報」をキーワードにそれにつながる「情報収集」「情報情報入手の容易性」「情報共有」…とさまざまな側面 (*) クラスター・サイズのもっとも大きなクラスター
クラスター5	30	「便利」をキーワードに「利便性」「情報の多様化」「開放性」など
クラスター6	2	利用経験，場所に依存しない　(*) 曖昧なクラスター
クラスター7	2	あまりない　(*) 意見がない
クラスター8	2	あまり無い　(*) 上のクラスター7 の言い換え，置換辞書で括れる例
クラスター9	3	興味，趣味などの共有化，つながり
クラスター10	2	地震など災害時対応に利点
クラスター11	38	知己，友人，友達などとの交流，情報取得，接触の機会のあること
クラスター12	10	情報収集，情報交換など
クラスター13	13	「わからない」に代表される意見 (*) 自由回答ではこのパターンが多い．必要に応じて削除辞書を作成処理 (*) クラスター7 とは異なる表現
クラスター14	3	即時性，人との接点など　(*) やや曖昧なクラスター
クラスター15	2	ほぼクラスター13 に類似のクラスター

いずれにしても，こうした自由回答の分析では，かなりヒューリスティック（発見的）であり探査的，帰納的な操作が求められる，ということである．このことは，何かの分析手法を適用して“1 つの厳密な解を求めること”とは，やや距離をおいた考え方である．所与のデータ表に対応分析法を適用し，クラスター化を行い，…といった操作はコンピュータが行っても，ここで説明したような人（分析者）の判断に委ねられる要素が多いこと，つまりあ

る意味での“さじ加減”を勘案することが必要である（それが実用のデータ解析である）.

[補足]検定値による観察

ここで，表 28 内に示した「検定値」について，簡単に触れよう．前の事例分析（その 1）で，若干これについてふれたことを思い出そう．表 24 に同じような検定値が表示されている．単純に，表 28 と表 24 の検定値をくらべると，かなり様子が異なることに気付くであろう．

- 検定値の数値が，表 24 にくらべ，表 28 のほうが大きい値が多い．
- とくに，表 24 では，「1」を越える数値はほとんどみられない．
- 表 28 の特徴として，文章（回答）の長さが短いものほど，値が大きめで，文章が長くなるほど値は低減する．
- また，表 24 のほうが，表 23 よりも文章が長目であるようにみえる．

実は，このどれもがこの検定の仕組みを反映した結果なのである．文章が長目になるということは，用いた構成要素（語句）を用いるチャンスが減ること，つまり情報がそれだけ曖昧になるということである．一方，回答が構成要素として選ばれた語句のどれかに集中して（それを中心に）用いているならば，検定値は大きくなる可能性が高いのである．見方を変えると，これらの検定値の差違は，前にも指摘したように，「Q19」と「Q13」とは，自由回答質問の問い方，あるいは回答者の受け取り方が違っていることを示唆している．前者は質問「Q19_2」や「Q19_4」への回答意向との関係で，回答者にとってやや理解しにくい質問であって，回答も長目となったことが読み取れる．一方，「Q13」は質問文の内容が比較的理解しやすく，回答も簡潔に短く記述されている，ということである．

しかしここで，2 つの分析にはある違いがある．「Q19」の場合には，「(回答者) × [構成要素（語句)]」の 2 元データ表の分析は行っていないことである．ここでは示さないが，この結果がどうなるか，予想してみてほしい．

確認 8:性年齢区分との比較

ところで表 28 には参考情報として「性年齢区分」（質的変数）も入れた．かりにこの人口統計学的変数が回答に関連する（相関がある）ならば，なにか傾向がみえるはずだからである．しかし，少なくともこの表に挙げた範囲の情報からは，性別や年齢は大きくは関与していないようにみえる．この相関の有無を確かめるには，「性年齢区分」を質的変数とし，これと選んだ語句群つまり構成要素変数との 2 元データ表の対応分析とクラスター化を行ってみればよい．これを実際に行って，えられた同時布置図が図 41 である．

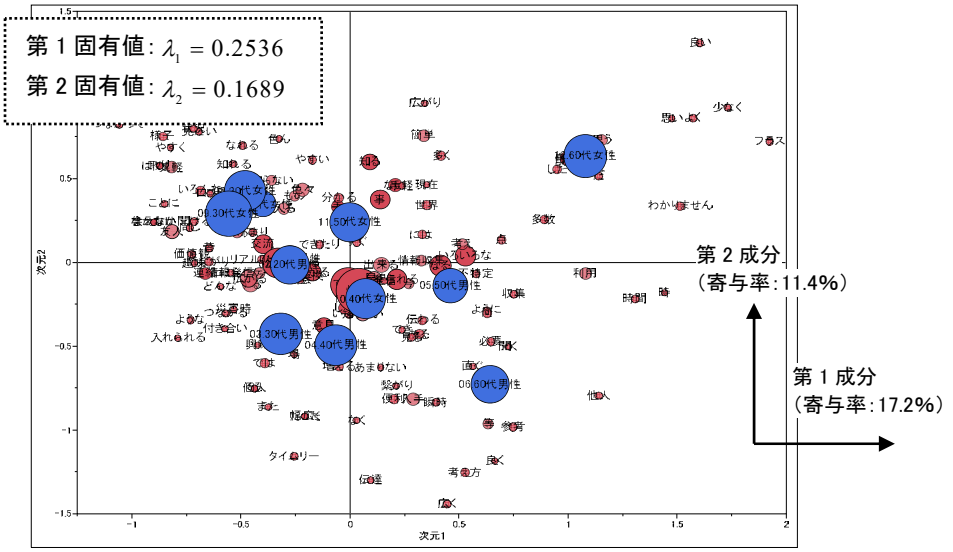


図 41　構成要素(139 語句)×性年齢区分(質的変数)の対応分析

この同時布置図から，（図が見にくいのだが），1 軸にそって左にある「若年層」から右方向の「60 代」へと移る傾向がみえる．また，性差は右側に位置する「60 代の男女」の差違を除くと，さほど大きくなく，図の左の方に混在している．また，重心（中心）あたりに男女とも 40 代，50 代が位置しているが（つまり回答者数が多い），全体に年齢のシフトはそう明らかではない．図に書き入れた固有値から，特異値は $\alpha_1=\sqrt{\lambda_1}=0.5036, \alpha_2=\sqrt{\lambda_2}=0.4110$ となりかなりの相関がある．しかし寄与率をみると，17.2（%），11.4（%）であり，そう大きくはないことに注意しよう．

ここで，性年齢区分別のすべての回答の記述はやめて，特徴的な 60 代の男女と，40 代の男女について．要約しよう．これらを観察すると，検定値が大きい回答者は少ない．そして自由回答（原文）は，特徴的なものもあるが，性年齢区分の中でのまとまりがなく，どちらかというと内容が発散している．上でみた表 28 の分類結果が，かなり明確であったことと対照的である．

表 30　性年齢区分別にみた自由回答（原文）

＜60 代男性＞

回答者 No.	検定値	検定値 (無調整)	ＳＮＳがプラスになると思うこと（原文）
[00000323]	1.0423	1.0423	良く分りません
[00000313]	0.6851	0.7093	いろいろな人間の考え方が参考になる
[00000305]	0.5633	0.5633	他人の考えが分かり、良くも悪くも参考になる。
[00000286]	0.5172	0.6007	情報が瞬時に広く伝えられる。
[00000331]	0.434	0.3106	人の考え方がわかる。
[00000324]	0.3312	0.3774	必要な情報の入手において、従来の図書館等での調査・調査会社等からの資料購入等に比べ時間・経費等が大幅に低減・節約でき、、社会経済的な効果が非常に大きい。、特に外国の資料等を考えると大いに社会経済的にプラスになる。
[00000333]	0.2586	0.2706	世界的な情報が瞬時に分かる
[00000343]	0.2566	0.3078	便利になると思う
[00000290]	0.2046	0.3383	時には害となる項目もあるが、いろいろな意見、考え方など見聞きできよい。
[00000346]	0.2011	0.2011	各種情報が広く知れ渡る事ができる。
[00000328]	0.181	0.2207	他人が何を思っているかが解る。
[00000291]	0.0743	0.1248	個人個人の考えが表せていいし、社会との繋がりが持てる。
[00000282]	0	0.1295	情報の収集
[00000284]	0	0.2015	不特定多数から、情報を交換できる。
[00000287]	0	0.0838	Youtube は映画・音楽など最新のものに触れることができるほか、息子の嫁が孫の動画をアップしてくれるので高画質の動画を見ることができる、facebook は友人関係が利用する度合いが期待したほどではない
[00000288]	0	0.2151	知りたい情報を継続的に入手できる。
[00000289]	0	0.049	自分が知りたい情報が即座にわかること。
[00000295]	0	0.3603	誰でも意見が発信できる
[00000296]	0	0.0172	情報収集
[00000304]	0	0.1579	人の意見を聞くことができる。
[00000322]	0	-0.0228	いろいろな人、国を問わず情報が素早く取得できる
[00000326]	0	0.2524	利用経験なく判らない
[00000330]	0	0.07	会話の如く意志の伝達が出来る
[00000332]	0	0	情報量が多くなる、知識のはばが広がる
[00000341]	0	0.2818	申し訳ありません。、利用していないためわかりません
[00000347]	0	-0.0573	情報の共有

＜60 代女性＞

回答者 No.	検定値	検定値 (無調整)	ＳＮＳがプラスになると思うこと（原文）
[00000315]	2.5024	2.5024	よくわかりません
[00000335]	2.3541	2.3541	わかりません
[00000301]	0.8285	0.8285	利用したことがないのでよくわからない
[00000308]	0.8003	0.7567	いろいろな情報が簡単に知る事がが良い。
[00000345]	0.7735	0.7735	余り無いと思う
[00000327]	0.7415	0.7415	一人悩む事が少なくなる。
[00000306]	0.6653	0.6456	簡単に情報収集できることはプラスになると思う。、
[00000314]	0.4681	0.5672	いろいろな情報が得られる、調べる時間が少なくなる

[00000302]	0.4423	0.438	いろいろな情報が早く・詳細に・身近に手に入れることができる点はとてもプラスだと思う
[00000318]	0.4228	0.3877	自分がわからないこと、疑問に思うこと、いろいろな分野の情報がみられる。知ることができる
[00000320]	0.4195	0.4233	自分の為になる情報が得られる。
[00000311]	0.393	0.393	特にない
[00000297]	0.3879	0.4944	色々な考えを知る事が出来る
[00000340]	0.3732	0.375	ブログやツイッターなどは利用したことがないので、よくわかりませんが、テレビでタレントさんたちが自分の考えていることを、ファンの人々に知ってもらうのは良いことだと思います。
[00000317]	0.3677	0.3763	いろいろな情報が得られる
[00000293]	0.3587	0.3587	利用しないのでわからない
[00000339]	0.3102	0.3102	今どのような事が起きているのか知る事が出来る。
[00000300]	0.3071	0.3226	自分のプライベートなことを不特定多数の人たちに知らせること自体危険なことと思っている。プラスになると思う点は何か？と考えても思いつかない。ただ、大統領、首相など政治家が自分の信条を伝えるには手っ取り早いかも知れない。
[00000344]	0.1754	0.1754	普段、参加しにくい人達が簡単に参加出来る。
[00000336]	0.1634	0.1006	いろいろなことが分かる。交流がひろまる。
[00000292]	0.1243	0.1632	現在の社会情勢の民間レベルでの受け止め方、自社の評判を知る手立て
[00000299]	0.1205	0.104	世の中の良い物も悪い物も知りたいことを他人に知られずに情報を得ることができる。
[00000337]	0.1105	0.1105	みんなで協力すれば形に成り、人助けになる。
[00000294]	0.1069	0.1526	多数の人の考え、思いを直接知ることができる。
[00000334]	0.0166	0.054	多くの情報が短時間に多くの人に伝わるから、３月の震災の時のように、支援などが迅速に行われやすい。
[00000285]	0	0.0074	情報の収集が早く、多くなる。
[00000298]	0	-0.295	情報がわかる
[00000310]	0	-0.0112	ユーチューブでニュースの画像を見て、関心が深くなることあり、懐かしい音楽なども聴くことができる。
[00000319]	0	0.0291	手軽に情報を手に入れることができる。
[00000325]	0	0.0116	情報が得られる
[00000342]	0	0.1048	世界が広がり視野も広くなる

＜40 代男性＞ 「情報」をキーとする記述;「情報交換」「情報迅速性」など

回答者 No.	検定値	検定値(無調整)	ＳＮＳがプラスになると思うこと（原文）
[00000199]	1.0817	1.0817	情報交換
[00000216]	1.0484	1.1203	早く情報が入手できる
[00000170]	0.9957	0.9957	情報が早い
[00000169]	0.8078	0.9384	タイムリーな情報を得られる
[00000190]	0.7441	0.7441	情報の速い伝達
[00000195]	0.6191	0.7189	今現在の情報をタイムリーに入手できる。
[00000163]	0.5974	0.5974	情報の周りが早い
[00000159]	0.5242	0.5591	情報をより早く入手できたり、調べたり出来る。
[00000203]	0.5049	0.5498	タイムリーな情報を得ることができる
[00000212]	0.4382	0.4745	欲しい情報のヒントが得られる点、情報の信憑性に注意するようになる点
[00000210]	0.4286	0.4601	情報を早く入手できる。マスメディアでは取り上げられない情報入手。人的交流が広がる。
[00000158]	0.4252	0.4252	情報が高速に伝達される。
[00000198]	0.3956	0.5397	価値観が広がる
[00000206]	0.3287	0.4882	幅広く情報が集められる
[00000209]	0.263	0.263	情報の流通が早くなる
[00000166]	0.2393	0.3457	リアルタイムで返事が返ってくる
[00000217]	0.2374	0.4926	情報発信の場が広がる
[00000179]	0.2191	0.328	重要な情報が得られる。
[00000196]	0.2191	0.3634	情報の即時性が向上する
[00000207]	0.1931	0.2789	・情報が早く伝わる。、・知りたいことの答えがすぐに得られる。
[00000178]	0.1795	0.3166	知りたいことがリアルタイムに直ぐにわかる
[00000221]	0.1659	0.1553	情報が早い。、距離的、時間的に離れた人との縁が切れにくい。
[00000162]	0.1514	0.1918	異業種の人と情報交換ができる。、音信不通だった学生時代の友人に連絡が取れるようになった
[00000213]	0.0495	0.1107	世界が広がる。現代の孤独感を紛らわせることができる？、どんな過疎地でもコミュニケーションができる。など
[00000161]	0	0.0263	昔の友達などとまたつながれる。

[00000164]	0	0	知りありの動向がわかる
[00000176]	0	0.0802	人とつながる
[00000177]	0	0	特にない
[00000180]	0	0.1093	「浅い」が、「広く」付き合いができるところ。、限りなく、生活水準が充実する。
[00000184]	0	0.0561	気軽に色々な画像を見たり、他人の意見を見ることが出来る。
[00000189]	0	-0.2693	いろいろな事をしれる。
[00000201]	0	0.1109	個人の意見や、作品が発表できる
[00000202]	0	0.1362	利用していないのでよくわかりません。
[00000208]	0	-0.0136	人と人との新たな交流が増える
[00000211]	0	0	コミュニケーション手段の多様化、特定の趣味関係者との交流を増やせる

＜40 代女性＞ 「情報」をキーとする記述；上位のほうに「利便性」「情報伝達，共有」など

回答者 No.	検定値	検定値 (無調整)	ＳＮＳがプラスになると思うこと（原文）
[00000186]	3.159	3.159	便利
[00000192]	0.8627	0.8627	あまり無い
[00000224]	0.8627	0.8627	あまり感じない
[00000204]	0.7477	0.7477	知りたかった情報が分って便利。
[00000220]	0.499	0.499	災害などのとき情報が早く伝わる
[00000223]	0.4985	0.4985	情報が何処でも確認できて便利なところ
[00000197]	0.4424	0.7362	情報の共有
[00000225]	0.4026	0.4339	色々な情報が出回る。有益な情報もある。、災害時などは便利だというイメージ。
[00000182]	0.365	0.4909	いろいろな人と情報を共有できる
[00000194]	0.2839	0.3708	いろいろな人と知り合え、情報も得られる。
[00000185]	0.2809	0.3331	情報が早く得られる、災害などの情報通信にも役だつ、
[00000181]	0.2636	0.2636	娯楽の一つである
[00000171]	0.2582	0.3008	いろいろな情報が多くの人とやり取りできる。、知り合いが増える。
[00000172]	0.2218	0.2984	緊急時に情報が早く伝わる。また、口コミなどの商品販売などに利用できる。
[00000187]	0.2212	0.3296	情報が簡単に入手できる
[00000200]	0.2197	0.2197	多くの人が交流しあえる
[00000215]	0.2141	0.2286	いろんな人と出会い、いろんな情報も手に入るし、輪が広がる。
[00000174]	0.1896	0.381	思いもかけない情報が入手できる
[00000191]	0.1896	0.3014	いつでもどこでも情報が得られる
[00000183]	0.1754	0.2929	いろいろな意見を知ることができる
[00000168]	0.1671	0.2517	参考になったり、近況がわかる。
[00000219]	0.1671	0.2456	一般人の意見が聞けて参考になる。
[00000205]	0.1228	0.205	いろいろな人の意見を知ることができる。
[00000222]	0.0664	0.1353	幅広く情報を得ることができる、日常で知り合うことがあまりない方たちと、交流が出来る
[00000214]	0.0369	0.0986	必要な情報が瞬時に得られること。、今回の震災や原発事故など非常時には、公共機関が秘している事も、ソーシャルメディアのお陰でかなり真実が収集できた。
[00000160]	0	0.0867	離れている友人の近況がわかる。、今まで疎遠だった友達と連絡がとれる。
[00000167]	0	0	コミュニケーション
[00000188]	0	-0.0147	情報量が豊富になること

(3)「Q13_B：マイナスと思うこと」への自由回答の分析

続いて，「Q13_B：マイナスと思うこと」について，同じような手順を用いて，対応分析そしてクラスター化を行い得られた結果を順に観察しよう．

［解析の条件］

ここでも，上とほぼ似たような条件を設定した．

- 有効サンプル数（回答者数）＝275（人）
- 異なり構成要素数＝274（語），構成要素数＝1383（語）
- 最大成分数＝273；2 元データ表の寸法＝275（人）×274（語）から出発
- 使用した成分数＝回答者，構成要素いずれも「10 成分」
- クラスター数＝回答者，構成要素いずれも「15 群」

この条件で得られた結果を，順をおって説明する．

観察 9:固有値，寄与率ほかの観察

ここで得られた固有値，寄与率ほかは図 41 のようになった．固有値数は最大 273（個）あるので，個々の固有値の寄与率は，前と同様のかなり小さく，また変化もなだらかである．初めの 10 成分で，累積寄与率は約 11.9%，20 成分まで拾っても約 22.3%程度である．

ここでも，初めの方の固有値の値は大きく，はずれ値的なプロファイルの存在が予想されるのであるが，ここはこのままで先に進むことにする．

図 41 「Q13_B:マイナスと思うこと」，固有値ほか

観察 10:クラスター数の検討

ここでも，クラスターの階層水準の変化を観察し，回答者（サンプル）および構成要素（語句）のいずれも「15 群」とした．この結果得られたクラスター構成が表 31 である．ここまでの手順は，前の分析例（その 2）に同じである．

表 31 をみると，回答者の分類結果は，クラスター・サイズの大きなクラスターが数個（6 個程度），構成要素の分類結果もクラスター・サイズのまとまった 6～7 群，あとはサイズの小さなクラスターがある，となっている．

表 31 「Q13_B:マイナスと思うこと」の回答者分類と構成要素（語句）の分類

回答者（サンプル）のクラスター	サンプル数
01	**23**
02	**92**
03	**106**
04	**14**
05	1
06	2
07	8
08	**22**
09	**31**
10	2
11	1
12	1
13	2
14	3
15	3

構成要素（語句）のクラスター	構成要素数（語句数）
01	**156**
02	**34**
03	**15**
04	6
05	1
06	**13**
07	**15**
08	1
09	**17**
10	**10**
11	1
12	1
13	1
14	2
15	1

観察 11:回答者のクラスター化情報

これも前の要領にならって，クラスター別に検定値の大きい順に回答者の記述した原文を要約してみる．これが表 32 である．これを眺めると，「プラス…」の場合の同種の情報（表 28）とはかなり様子が異なることがわかるであろう．各クラスターの検定値の大きい語句から順に目で追ってみると，たしかに「マイナスと思うこと」に関連する語句類がつぎつぎと登場する．これはあらためてここで説明するまでもなく，かなり明瞭に意見が分かれている．前にならって，各クラスターの内容を表 33 のように要約した．なおここで，クラスター・サイズの小さいクラスターは一部を除外した．また，その一部は，他のサイズの大きいクラスターと類似した内容であるので，そのクラスターに説明を加えた．

もっともクラスター・サイズの大きい「クラスター3」は，内容がやや曖昧にみえる．しかし検定値をみると，はじめの 9 人あたりまで「ない」「とくに）ない」「わかりません」「特にない」…といった，いわば意見が無い人たちである．もし，こうした記述には意味がなさそうということであれば，関連する構成要素（語句）を削除辞書に登録し，再分析（であり細分析）を行えばよい．

一方，その他のサイズの大きいクラスターは，表 33 に整理したように，それぞれ特徴的な語句がまとまっているとみてよいだろう．またここでも，（検定の仕組みから）語句数が多い回答が（検定値は小さくなるにつれて）回答文の長さが次第に長くなる（次第に曖昧になる）．

ここで内容がやや曖昧にみえたクラスター3 は再吟味が必要かもしれない．また，クラスター12，15，14 のように，サイズが小さく他のクラスターとほぼ類似のものもある．こうしたことから，クラスター数の条件を変える（もう少し少なめにする）などして再調整が必要かもしれない．各種の統計指標を観察しながら，探索的，試行錯誤的要素は，クラスター化処理においてまぬがれない手順である．

ここで，併せて（表 31 の）各クラスター・サイズが，「Q13_A：プラスと思うこと」に比べて（表 26），きわだって明確に分かれてはいないことに注意しよう．つまり，クラスターの分離の程度（クラスター化の度合い）が「Q13_B」がゆるいということを示唆している．

しかし，「Q13_A」と「Q13_B」の 2 つの自由回答の傾向は，たしかに顕著な違いがあり，回答者は質問文の意図を汲んで回答してくれたと考えてよさそうである．つまり，この質問文は調査設計時に意図したように機能したと考えてよさそうである．

表 32 「Q13_B:マイナスと思うこと」への自由回答の類型化（15 群）

クラスター01	検定値	検定値(無調整)	「Q13_B_ＳＮＳがマイナスになると思うこと」の自由回答（原文）	性年齢区分
1	4.6357	4.6357	時間の無駄	02.20 代男性
2	4.6357	4.6357	時間の無駄	04.40 代男性
3	3.0545	3.0545	時間が取られる	03.30 代男性
4	3.0545	3.0545	時間がかかる	09.30 代女性
5	2.6363	2.6363	プライバシ	02.20 代男性
6	2.3178	2.3178	時間を無駄に使ってしまう	08.20 代女性
7	2.1044	2.2551	くだらない話で時間がつぶれる	05.50 代男性
8	1.8327	1.9211	時間を浪費している	03.30 代男性
9	1.604	1.616	依存してしまって時間がもったいないことがある	07.10 代女性
10	1.5921	1.5921	時間ばかりが過ぎていってしまう。本を読む時間がなくなる。	03.30 代男性
11	1.5377	1.4683	他の事をする時間がない	04.40 代男性
12	1.4885	1.4885	偏りがちになる気がする	03.30 代男性
13	1.3446	1.3446	不要な情報まで取り入れることになる。、時間を無駄に消費することになる。	09.30 代女性
14	1.333	1.3322	具合が悪い時でも一通り見ないと気が済まないので時間がもったいない気がする時がある。	09.30 代女性
15	1.1986	1.2639	重要性の低い話や間違った話が多い。、これに反応すると時間の無駄。	06.60 代男性
16	1.1534	1.1534	ネットに依存すること。	07.10 代女性
17	1.1184	1.1184	せわしない。ゆっくり自分で考えたりする時間が減る。	10.40 代女性
18	0.965	1.0097	必要と思われる様々な資料が比較的簡単に入手できるので、、その取捨選別等に時間がとられ、資料を深く考察する、時間がなくなる。、	06.60 代男性
19	0.7166	0.7166	情報が氾濫過ぎる、時よりプライバシのことは気になる。	11.50 代女性

20	0.5634	0.5634	まとまりがつかなくなる	05.50 代男性
21	0.5137	0.7397	内容に気をつけないといけない。	05.50 代男性
22	0.4209	0.873	会話が少なくなる	10.40 代女性
23	0	0.1392	直接的なコミュにケーションスキルの低下。	09.30 代女性
クラスター02				
1	4.0723	4.0723	嘘の情報	04.40 代男性
2	3.9761	3.9761	情報の氾濫	11.50 代女性
3	3.1239	3.1239	間違った情報に惑わされる	05.50 代男性
4	2.9417	2.9417	誤った情報に惑わされる。	12.60 代女性
5	2.8605	2.8605	不確かな情報が出回る	09.30 代女性
6	2.5664	2.5664	情報の正確性に欠ける	08.20 代女性
7	2.3689	2.5093	余計な情報に惑わされやすい	12.60 代女性
8	2.364	2.364	情報が流れすぎる。	09.30 代女性
9	2.3263	2.5115	正しい情報がわかりにくくなる	08.20 代女性
10	2.1204	2.1204	誤った情報が広まること。	05.50 代男性
11	1.9858	1.9858	正確性	03.30 代男性
12	1.8912	2.0511	情報が錯綜する。	02.20 代男性
13	1.8912	1.8912	情報の漏えい、、	02.20 代男性
14	1.8912	1.8912	情報の正解度が低くなる	04.40 代男性
15	1.8912	1.8912	情報が多過ぎて振り回される。	11.50 代女性
16	1.8912	1.8953	情報の確かさが不明	12.60 代女性
17	1.8912	1.8912	情報の混戦と真ぴょう性	06.60 代男性
18	1.8175	1.8822	誤った情報も瞬く間に広がる。	08.20 代女性
19	1.8077	1.9481	余計な情報が多すぎる。	06.60 代男性
20	1.7748	1.7748	情報が多すぎるので、必要な情報の精査が必要。	04.40 代男性
21	1.7669	1.794	情報が正しいかすぐには判断できない。	05.50 代男性
22	1.7453	1.8099	情報の流失、嘘に踊らされる	03.30 代男性
23	1.6628	1.6844	不確かな情報と正しい情報の判別が難しく、情報に惑わされやすい。、個人情報が容易に漏れてしまう。	02.20 代男性
24	1.576	1.6982	情報が漏れちゃうかも知れない。	07.10 代女性
25	1.576	1.576	しらないあいだに情報がながれる	02.20 代男性
26	1.576	1.576	どの情報が真実かわからなくなる	02.20 代男性
27	1.576	1.8654	情報が間違いであるかもしれない	09.30 代女性
28	1.576	1.576	いろいろな情報が野放し状態。	05.50 代男性
29	1.576	1.576	情報に左右されすぎる。、	11.50 代女性
30	1.5266	1.6408	悪い情報や書込みが増えたりする	10.40 代女性
31	1.4136	1.5212	誤った情報を入手する可能性がある。	12.60 代女性
32	1.3765	1.3795	正しい情報かどうかあてにならない。、悪意を持った情報の危険性。	04.40 代男性
33	1.3723	1.4242	匿名性の情報が氾濫し、真偽の判断がつかないまま情報が一人歩きする。	09.30 代女性
34	1.3527	1.3661	情報の正確性に問題がある。デマを流しやすい。	10.40 代女性
35	1.3473	1.385	情報の取捨選択が難しいので、デマに踊らされやすい。	03.30 代男性
36	1.3422	1.3422	デマなどが出回りやすい	03.30 代男性
37	1.2713	1.3899	情報がどこまで広がっているのか不安	12.60 代女性
38	1.2689	1.3985	情報が正しくないとき、デマ、風評被害が起こりやすい。	12.60 代女性
39	1.2605	1.3671	嘘の情報が流れていると、それを信じてしまう。	04.40 代男性
40	1.2168	1.3572	知らない人に情報が流れる	12.60 代女性
41	1.1928	1.2836	ガセの情報などが氾濫しすぎていること。	09.30 代女性
42	1.1147	1.1147	デマが飛び交う	10.40 代女性
43	1.0846	1.0867	正確な情報かの判別がつきにくいこと	03.30 代男性
44	1.0817	1.1181	、、悪い情報が流れた場合どこまで信じていいかわからない	12.60 代女性
45	1.0507	1.2379	情報が流れる先が果てしないので怖い。	12.60 代女性
46	1.0372	1.0372	コミュニケーション能力の低下	03.30 代男性
47	1.0256	1.0449	ルールやマナーを守れない人達による炎上等の問題、誤った情報が広がる可能性が高い（間違った情報のリツイート等）	05.50 代男性
48	1.0038	1.0376	情報によっては社会全体に悪い影響を簡単に与えやすい。	06.60 代男性
49	0.9779	1.0256	必要のない情報の量も増えてしまう。	03.30 代男性
50	0.9035	0.9282	・情報の散乱による信憑性の低下、・情報を扱うことの、リスク・セキュリティに対する意識の低下、・ネット上の交流になりがち。対人スキルの低下	08.20 代女性
クラスター03				
1	2.9526	2.9526	ない	02.20 代男性
2	1.4883	1.4883	わかりません	12.60 代女性
3	1.4763	1.4763	とくにない	01.10 代男性
4	1.3805	1.3805	利用していないのでよくわかりません。	04.40 代男性
5	1.0738	1.1714	申し訳ありません。、利用していないためわかりません	06.60 代男性

6	1.0219	1.0219	利用していないので、特に無し	05.50 代男性
7	0.9842	0.9842	特にない	09.30 代女性
8	0.9842	0.9842	特にない	03.30 代男性
9	0.9842	0.9842	特にない	04.40 代男性
10	0.8806	1.068	人の意見に左右される。	06.60 代男性
11	0.8733	1.1688	中傷が広まる	06.60 代男性
12	0.8671	0.8671	利用の仕方次第	10.40 代女性
13	0.8472	1.008	他人に見られるのが嫌だ	02.20 代男性
14	0.7382	0.7382	別にない。	07.10 代女性
15	0.7011	0.9212	知りたくないものを知ってしまう	11.50 代女性
16	0.6942	0.8701	良くない噂話が広まる事。	10.40 代女性
17	0.6842	0.7668	犯罪に利用されるかも。	06.60 代男性
18	0.6801	0.7909	誹謗、中傷、犯罪等のきっかけ	05.50 代男性
19	0.655	0.655	今のところ感じていない	05.50 代男性
20	0.6493	0.7871	変なうわさがたつ様な事がないとは、限らない！	07.10 代女性
21	0.6385	0.8255	うわさなどが飛び交い人を傷つける	12.60 代女性
22	0.6377	0.9593	人間関係が悪くなる	04.40 代男性
23	0.6257	0.6707	言葉の影響力を理解していない人には想像もつかない様な大きな事が起こる恐れがあるのではないかと不安に思う、、	09.30 代女性
24	0.594	0.81	まったく面識のない人とのやりとりなので何だか不安が残る	10.40 代女性
25	0.5905	0.5905	特にマイナスはない	06.60 代男性
26	0.5486	0.6347	信憑性が把握できない	12.60 代女性
27	0.5417	0.6067	人から人へ、噂が段々に違う方向になってしまうのが嫌な感じがする。	11.50 代女性
28	0.5416	0.5416	人とのつながりを大切にしなくなりそう。	09.30 代女性
29	0.5367	0.6148	人は現実とは違ったパーソナリティを演じているケースが多く、誤解を招く。	05.50 代男性
30	0.5291	0.5455	卑猥なことを書き込んだり、人を中傷したりしているのを見聞きするとソーシャルメディアの存在中止すべきだと思う。	12.60 代女性
31	0.5278	0.5278	現実でのコミュニケーションが薄くなる	01.10 代男性
32	0.5254	0.6205	対面のコミュニケーションではないので、ずれがおこりやすく、ずれたときの関係の修復が容易ではない。	03.30 代男性
33	0.4921	0.6399	根拠がないことが広まる	06.60 代男性
34	0.4901	0.7392	常に人の目を気にして発言しなければならない。	08.20 代女性
35	0.4851	0.6229	変な事は、言えない。	04.40 代男性
36	0.4639	0.5785	人の悪口を書いてる人がいるので、気分が悪くなることがある。	08.20 代女性
37	0.4487	0.5057	自分の言いたいことを読む人によって誤解されることが多々あるのではないかと心配です。	12.60 代女性
38	0.4306	0.5098	直接の対話ではないため、コミュニケーションをとるのが難しい。誤解が生じることが多いようなところ。	09.30 代女性
39	0.4237	0.4237	くだらない意見が多い	06.60 代男性
40	0.4121	0.4121	誹謗・中傷の恐れがある	04.40 代男性
41	0.403	0.403	悪用されたり、犯罪につながったり。	09.30 代女性
42	0.397	0.4826	他人を安易に誹謗、中傷することができる。	12.60 代女性
43	0.376	0.6172	犯罪に使われたりしそう	09.30 代女性
44	0.376	0.6193	犯罪が増えるかも	04.40 代男性
45	0.373	0.5605	如何わしいコンテンツが存在する為、取捨選択できない未成年者等にこれらコンテンツを利用して社会問題化に至る事。	06.60 代男性
46	0.3724	0.437	悪意の人も参加し易く、影響が大きい。	12.60 代女性
47	0.3724	0.4508	変なことを書いて責任問題になる人が増える、出会い系状態になる	02.20 代男性
48	0.3693	0.6549	ついつい、長時間利用してしまうこと。	09.30 代女性
49	0.3617	0.432	個人の誹謗・中傷など悪い方向で使われることは非常に困る	12.60 代女性
50	0.3521	0.3521	人を知らないうちに傷つけたりする事がある。	11.50 代女性
クラスター04				
1	3.7117	3.7117	相手が見えない	08.20 代女性
2	2.6383	2.6383	顔の見えないコミュニケーション	05.50 代男性
3	1.6266	1.6266	不特定多数が相手というのは何となく恐ろしい	12.60 代女性
4	1.3012	1.3012	不特定多数の人間と関わる可能性があり、相手の素性がわからない	10.40 代女性
5	1.2614	1.1401	不特定多数に情報がいきわたる。	03.30 代男性
6	1.1821	1.135	不特定多数に自分の情報を開示すること。、相手が見えないこと。、	03.30 代男性
7	1.0822	1.1183	顔の見えない相手との関わりが増えるので、犯罪などに巻き込まれる恐れがある	08.20 代女性
8	0.8982	0.9436	偏った考え方になる。相手が見えない中でのコミュニケーションでトラブルが考えられる。	03.30 代男性
9	0.8802	0.8802	1 日中パソコンから離れられなくなり、引きこもりが増える。	04.40 代男性

10	0.8227	0.8852	顔を合わせないため、相手を傷つけるような書き込みや、ブログの炎上などがあり、引きこもりなどを増やす恐れがある	04.40 代男性
11	0.7058	0.7058	仮想空間と現実空間の分別がつかない人間が増える	03.30 代男性
12	0.7043	0.7043	余計なしがらみが増える。	02.20 代男性
13	0.4377	0.4377	真偽が疑わしい	03.30 代男性
14	0.3377	0.3682	「顔」が見えないので、付き合いが「浅く」「不透明」。、限りなく、匿名性が広がって、「猜疑心」が助長される。	04.40 代男性
クラスター05				
1	0.331	0.331	手軽すぎて、気密性の線が緩む。	10.40 代女性
クラスター06				
1	1.2989	1.2989	情報量が多すぎて見たいものを選ぶのが大変	09.30 代女性
2	1.2908	1.2908	情報量がとても多いので絞込みが大変です	05.50 代男性
クラスター07				
1	1.4754	1.4754	自分の考えを無責任に言える。感情に立体感が無くなる。	12.60 代女性
2	1.0591	1.0591	極端、過激な考えに影響されそう。	10.40 代女性
3	0.7847	0.8116	自分の考えを持っていないと「扇動」とは言わないが、無責任に同じように考え、結果、極端な行動に同調しかねない。	06.60 代男性
4	0.7745	0.7745	文面上のやり取りのみで感情が読み取れない	07.10 代女性
5	0.7101	0.7101	自分のことをソーシャル・メディアに流そうとは思わない	06.60 代男性
6	0.7087	0.7087	良く分りません	06.60 代男性
7	0.589	0.589	個人情報の流失、失言に対する異常とも言えるバッシングが一人歩きすることで、社会全体が同方向への感情をもつことの危険性	12.60 代女性
8	0.5349	0.6218	実際に直接会って話す必要が無くなること。	01.10 代男性
クラスター08				
1	9.3319	9.3319	個人情報	03.30 代男性
2	5.5219	5.5219	個人情報の流出	02.20 代男性
3	5.5219	5.5219	個人情報の流出	08.20 代女性
4	5.5219	5.5219	個人情報の流出	08.20 代女性
5	5.5219	5.5219	個人情報の流出	04.40 代男性
6	3.5566	3.5566	個人情報が守れない	09.30 代女性
7	3.1145	3.1145	個人情報が漏れること	09.30 代女性
8	3.1145	3.1145	個人情報が漏れる。	11.50 代女性
9	3.0483	3.0483	個人情報などの漏洩	09.30 代女性
10	2.5758	2.5758	個人情報の流出の可能性があること	01.10 代男性
11	2.5758	2.5758	個人情報の流出の可能性がある。	06.60 代男性
12	2.5757	2.62	個人情報が流出する恐れがある	04.40 代男性
13	2.4616	2.4616	個人情報の漏洩の可能性。	03.30 代男性
14	2.333	2.333	個人情報が出すぎてしまう	01.10 代男性
15	2.333	2.333	個人情報の管理。	11.50 代女性
16	2.333	2.333	個人情報がでる、	06.60 代男性
17	2.307	2.307	個人情報が漏れる可能性を懸念。	12.60 代女性
18	1.7086	1.7815	個人情報の流出。、直接的な会話がなくなる。	01.10 代男性
19	1.492	1.5488	個人情報の流出が激しい（犯罪利用の危険性が高まる）	01.10 代男性
20	1.2895	1.3419	個人情報の流出の危険 ホントのことか、作り事かわかりにくい所がある。	11.50 代女性
21	1.2245	1.2245	個人情報がもれる危険がある、使いすぎて高額な請求額がくる	12.60 代女性
22	0.313	0.2981	人との接点が変わってくること	03.30 代男性
クラスター09				
1	4.7232	4.7232	プライバシーの侵害	03.30 代男性
2	4.7232	4.7232	プライバシーの侵害	09.30 代女性
3	4.1595	4.1595	プライバシーが侵害される	09.30 代女性
4	3.6915	3.6915	プライバシーが保護されない	03.30 代男性
5	3.5656	3.5656	プライバシーが漏れる	09.30 代女性
6	3.1005	3.1005	プライバシーの侵害が懸念される	02.20 代男性
7	3.08	3.08	プライバシーがまもれない	12.60 代女性
8	2.5046	2.5046	プライバシーが守られない可能性がある	05.50 代男性
9	2.2392	2.2392	プライバシーが守られない恐れがある	04.40 代男性
10	1.6707	1.8783	個人情報が多いのでプライバシーが心配	08.20 代女性
11	1.5744	1.6837	誹謗・中傷が多い。プライバシーの侵害	05.50 代男性
12	1.5281	1.5897	知らない人にプライバシーが漏れる、	11.50 代女性
13	1.4728	1.5077	間違った情報が広まる、プライバシーの侵害	04.40 代男性
14	1.4001	1.4001	モラルが守られるか心配である	04.40 代男性
15	1.2963	1.3295	プライバシーがあまり無い。知られたくないプライベートの部分もばれてしまう。	04.40 代男性
16	1.289	1.32	プライバシーがなくなる。、犯罪に巻き込まれる。	09.30 代女性

17	1.2639	1.2639	プライベートがばれる	03.30 代男性
18	1.2305	1.2805	個人情報をツイッターに流すことで、プライバシーが侵害されることがある	04.40 代男性
19	1.155	1.2112	プライバシーの守る範囲が難しいと思う	08.20 代女性
20	1.1265	1.3372	個人情報の流失やプライバシーの保護が出来ない。	05.50 代男性
21	0.9826	0.9826	あまり感じない	10.40 代女性
22	0.8239	0.8923	自分のプライバシーがオープンになりすぎる危険性がある。、悪質な勧誘やいたずらメールに利用される。	09.30 代女性
23	0.8205	0.8787	知られたくないことを知られる可能性がある	04.40 代男性
24	0.7894	0.8313	犯罪への利用が懸念される	03.30 代男性
25	0.7583	0.7583	プライベートなことまで知れ渡る	04.40 代男性
26	0.615	0.6478	プライバシーが守られるか心配。、(芸能人がお店にきた、など)、モラルが守られなかったり、金儲け業者が入り込んできたり。	08.20 代女性
27	0.5938	0.5654	自分のプライベートが必要以上にさらされる危険性がある。、依存症になる危険性がある。	10.40 代女性
28	0.4942	0.7238	事件や犯罪に巻き込まれる可能性が高まる	03.30 代男性
29	0.4736	0.5181	・モラルのない使われ方をする人が出てきて、、 プライバシー（個人情報）が保護されない。、・不確実な情報でも多くの人に伝わり、信じられてしまう、 可能性がある。	04.40 代男性
30	0.3393	0.4463	個人情報が公になりすぎる事	09.30 代女性
31	0.1969	0.3408	匿名でネットで話していたら、知られたくないことを友人にバレテしまうこと。	08.20 代女性
クラスター10				
1	1.0638	1.1385	嘘の情報が流れるのも早い	09.30 代女性
2	1.0637	1.0637	嫌なうわさも早い。	09.30 代女性
クラスター11				
1	0.8092	0.8092	監視されてる感じがする。情報漏洩が起こりやすい	02.20 代男性
クラスター12				
1	2.979	2.979	情報漏洩	04.40 代男性
クラスター13				
1	1.4183	1.4183	利用経験なく判らない	06.60 代男性
2	1.0409	1.0409	情報量過多で的確な判断が出来なくなりやすい	06.60 代男性
クラスター14				
1	5.3844	5.3844	誹謗中傷	04.40 代男性
2	2.5881	2.5881	個人の誹謗中傷	04.40 代男性
3	1.3461	1.3461	誹謗中傷をたまにみる	02.20 代男性
クラスター15				
1	5.6285	5.6285	個人情報流出	03.30 代男性
2	2.8142	2.8142	個人情報流出、	08.20 代女性
3	1.1441	1.1441	リスクが大きい(個人情報流出など)	07.10 代女性

この自由回答は，質問「Q13_A：プラスと思うこと」に比べて，各クラスターの意味づけが明確なものばかりではない．とくに，「クラスター3」は，内容が微妙に異なる表現が混在しているようにみえる．一方，(ステロタイプの意味付けとなるが)表 33 に要約したような特徴もみえる．

表 33　クラスターの類型化とネーミング

クラスター	クラスターサイズ	主な特徴
クラスター1	23	時間の使い方，無駄，浪費，情報氾濫の影響
クラスター2	92	嘘，誤った情報，情報の正確さ・不正確，情報の判断・判別，信憑性・真偽
クラスター3	106	ソーシャル・メディアの諸事象の混在　(*) やや曖昧なクラスター
クラスター4	14	不特定多数の対応，相手が見えないこと，匿名性など
クラスター7	8	情報への責任・無責任，先導，感情など　(*) 他とはやや異なる意見
クラスター8	22	個人情報，情報流出，情報漏洩，…／クラスター12，15 に同じ
クラスター9	31	プライバシー，誹謗中傷など／クラスター14 はほぼ同じ

5.5 ここまでのまとめ

おわりに，ここまでに述べた説明と分析例を通じて，クラスター化における留意点と，自由回答質問の考え方について，簡単に要約しておこう．

（1）クラスター化処理における要点と課題

① いわゆる（房状，塊状にまとまって，しかも互いによく分離した）“クラスター”と認識できるようなはっきりしたクラスターは（well-separated clusters），現実にはほとんどない．さらに，各クラスター間を，排反的に厳密に境界をつけて線引きするようなクラスターは，現実にはほとんどない，あるいは現実的ではない，と考えるべきである[65]．

② むしろ，漠然としてはいるが，なにかの類似性があるグループを作り出すこと，つまり**“クラスター化”**を行っている，と考えるほうが現実的である．このときに，いわゆる**“クラスター化（最適化）基準”**つまりどのような形のクラスターを生成するかの基準を用意することが多い．これを設けるということは，その基準に従った（その条件を満たすような）クラスターを生成することを意味する．たとえば，ウォード法や *k*-平均法では，クラスター内分散の最小化を基準とするので，大きさのそろった（クラスター内分散が似通った）クラスターを作りやすい．

③ よって，クラスター化後に，こうした状況を判断する別の**“目安となる指標”**の準備が必要となる．たとえば，上でみた有意性テストなど．

④ とくに，**“はずれ値の影響”**に十分な注意が必要である．多くのクラスター化手法，とくに分散最小化基準，平方ユークリッド距離など使うウォード法や *k*-平均法は，はずれ値の影響を受けやすい．見当外れの位置（座標）にクラスター重心があると判断するなど，いわゆる**“的外れ”**（ワイルド・ショット；wild shot）現象がおこる[66]．

⑤ はずれ値は，クラスター数の設定や結果として得たクラスター構造にも影響を及ぼす．こうした場合，はずれ値候補をいったん除外して再計算を行い，その除外したはずれ値を**追加処理で再配置**する，という操作が有効なことがある．

⑥ また，クラスター数は一意に定めずに，何通りかの分類を行って探査的に決めることがよい．既述のように，クラスター数を決める最適な方法はない，しかしこのときに，その見当をつけるための指標，対応分析法の特性を利用するならば，階層的分類時の**“階層の結合水準”**などを目安とすることが妥当であろう[67]．

⑦ 対応分析は，偏ったプロファイルの影響を受けやすい．とくに，寸法が大きくて疎なデータ表を扱う場合は，はずれ値的なプロファイルがあると，これが変動の大きさに影響し，成分スコアの大きさに影響する，結果としてクラスター化にも影響する．１つの目安として，所与の 2 元データ表の行和あるいは列和に注目して，あまりに度数が小さい行あるいは列を一時除去する，という方法もある[68]．

⑧ 対応分析でえた結果を用いるクラスター化で，成分数の指定や，クラスター数の設定を，探査的に調べる必要がある．得られる結果は厳密に一意に決まるものではないこと（つまり，きわめて探査的になる）．

⑨ クラスター化という自動分類の役割は，個々の測定対象の観察よりも全体としてみたときの傾向探査，特徴抽出を行うことである．いわば類似した群を複数作り，それらの相互の関連を評価することで，個々の群の特徴を強調すること（差違をみること）にある．

---

[65] いわゆる“ファジイ・クラスタリング”という考え方がある．たとえば，*k*-平均法のファジイ版として“ファジイ *c*-平均法”なども提案されている．ここでは，分類対象がクラスターに所属する確からしさを確率で与えるなどを行う．

[66] WordMiner で，相互近隣関係の規則を使った階層的分類法で初期化を行い，*k*-平均法で細分類を行う理由の１つはこの現象をすこしでも回避するための手当である．

[67] 階層の結合水準を追うことは，（成分スコアからみた）クラスターの変動の大きさを評価することに同じ．

[68] Greenacre によると，これを“subset correspondence analysis”と呼んでいる．

(2) 自由回答質問の設計上の留意点

自由回答質問の設計の適否が，自由回答の記述内容に影響することは，上の 2 つの分析例をみただけでも，わかるであろう．これの要点を挙げておく．

① 社会調査であれば，分析の細かい操作手順を知る前に，まずはうまい適切な質問文を作成すること（事後の策より事前の配慮，転ばぬ先の杖）．また，状況に応じて（余力があれば）“**予備調査**”や，さらに可能ならば“**パイロット調査**”などを行うこともよいだろう[69]．
② つまり，自由回答質問の分析では，調査票設計が重要な要素の 1 つとなること．
③ 自由回答とは，回答者から自由に，あるいは勝手な意見を聞き出す（引き出す）ことではない．むしろ，何を聞き出すかを，どのように構造化できるのか，あるいはできそうか，に留意すべきこと．俗な言い方をすれば，偏りのない「本音を聞き出す」，抵抗なく記入してもらえるようなうまい質問文が作れるだろうか，ということである．
④ とくに“問い方は単純かつ明解に”しかも“複雑にはならないこと”，つまり（概念的で考えさせるような）解釈が面倒な質問文は避けるべきである．また“正確な日本語で質問文を記述”することも重要である．
⑤ おおくの場合，英語の“open-ended”question に，「自由回答（質問）」の訳をあてている[70]．“open-ended”には「制約がない，変更可能な，決まった答えのない」…という含みもある．これは，書き方という操作に制約がないのであって，書いてもらう内容を，自由勝手に書くこととは，すこし意味が違うと考えるのが妥当である．

喩えていうならば，あまり些末なこと，あるいは厳密さを追うことに気を取られて，肝心の目標を見失わないこと，「木を見て森を見ず」とならぬことである．「闇夜の烏は見つけにくい」が，これをどれだけ「霧の中の烏」（≒半構造化の可能性）へ，あるいはさらに「晴れた日差しの中の烏」（≒構造化し輪郭を明確に掴む）として認識できるか，そのような手続きがあるのだろうか，ということである．多くの場合，「薄霧の中の烏」であっても，データ構造がそこまで見通せるようにすることのほうが，はるかに有効なのである．

---

[69] プリテスト（事前テスト : pretest），予備調査（preliminary test, pretest, field pretest），パイロット調査（pilot survey, pilot study）の区別はさほど明らかではないこともある．一般に「プリテスト」とは，調査票や質問文，調査用具を，調査あるいは実験の前にテストすること．例 : プリテストとして便宜的標本を使うなどがある．「パイロット調査」とは，“本調査と（ほぼ）同じ調査環境，条件下”で，その調査が設計通りに機能するかを確認する小規模の事前調査のこと．

[70] 選択肢を用意した選択肢型質問を“closed question”という．英語表記のほうが，質問文の特徴をより明示的に表しているようだ．いわゆるフリーアンサー（free answer）という言い方は，欧米ではあまり一般的ではないだろう．

【キーワード】　※第Ⅰ部，第Ⅱ部と重複あり

対応分析法（CA: Correspondence Analysis，AFC: Analyse des Correspondances），慣性（inertia），カイ二乗距離（Chi-square distance），ピアソンのカイ二乗統計量（Chi-square statistic），自動分類（automatic classification），クラスター分析（cluster analysis），凝集型階層的分類法（AHC : agglomerative hierarchical classification），分割型分類法（partitioning-type classification），*k*-平均法（*k*-means method），混合方式（mixed clustering approaches），成分スコア”（principal coordinates, coordinates），相互最近隣の規則（RNN：reciprocal nearest neighbours rule），2 元データ表（two-way data table），クロス表，構成要素，構成要素変数，質的変数，プロファイル（profile），行プロファイル（row profile），列プロファイル（column profile），固有値，寄与率，累積寄与率，総変動・全分散（total inertia, total variance），デンドログラム・樹形図（dendrogram），階層の結合水準（hierarchical indices），クラスター内変動・クラスター内分散（within-cluster variances），クラスター間変動・クラスター間分散（between-cluster variances），クラスター間変動比，検定統計量（test statistic）と検定値（test value），単純無作為抽出（SRS : simple random sampling），復元抽出（SWR : sampling with replacement）と非復元抽出（SWOR : sampling without replacement），標本平均の分布，有限母集団と標本，標準化，正規分布，正規近似，クラスター数を決める目安，同時布置図，布置図

【参考文献】※第Ⅰ部，第Ⅱ部と重複あり

[1] Everitt, B.S. (1993): *Cluster Analysis*, Arnold.
[2] Gordon, A.D. (1999): *Classification,* second edition, Chapman & Hall.
[3] Greenacre, M.J. (1984): *Theory and Applications of Correspondence Analysis*, Academic Press.
[4] Greenacre, M.J. (2007): *Correspondence Analysis in Practice* (second edition), Academic Press.
[5] Greenacre, M.J. (ed.) (2006): *Multiple Correspondence Analysis and Related Methods*, Chapman & Hall/CRC.
[6] Jambu, M. (1989): *Exploration Informatique et Statistique des Données*, Dunod.
[7] Lance, G.N. and Williams (1967) : A General Theory of Classificatory Sorting Strategies – Hierarchical System, *Computer Journal*, vol. 9, no. 4, 373-380.
[8] Lance, G.N. and Williams (1977) : Hierarchical Classificatory Methods, in *Statistical Methods for Digital Computers, Volume III*, Enslein K., Ralston, A., and Wilf, H.S. (eds.), John Wiley and Sons.
[9] Le Roux, B. and Rouanet, H. (2004): *Geometrical Data Analysis – From Correspondence Analysis to Structural Data*, Dordrecht Kluwer.
[10] Le Roux, B. and Rouanet, H. (2010): *Multiple Correspondence Analysis*, Series: Quantitative Applications in the Social Sciences No.163, Sage Publications, Inc.
[11] Lebart, L., Salem, A. and Berry, L. (1998): *Exploring Textual Data,* Kluwer Academic Publishers.
[12] 大隅昇（1989）：統計的データ解析とソフトウェア，日本放送出版協会.
[13] 大隅昇，Ludovic Lebart 他（1994）：記述的多変量解析法，日科技連出版社.
[14] 岩坪秀一（1987）：数量化法の基礎，朝倉書店.

（*）この他，テキスト・マイニング研究会ホームページから提供される各種の情報がある．とくに，対応分析法については，ホームページの「技術解説」の項から，解説文の pdf 形式のファイルがダウンロードできる．

◆テキスト・マイニング研究会ホームページ：
http://wordminer.org/
◆技術解説： http://wordminer.org/tips/63
「よくある質問へのヒント」 http://www.wordminer.org/wp-content/uploads/2013/04/63_9.pdf
・構成要素，異なり構成要素の分布の特性
・有意性テスト（とくに頻度による有意性テスト）
◆レシピ： http://wordminer.org/tips/37