対応分析法・数量化法 III 類の考え方

大隅　昇
テキスト・マイニング研究会代表
統計数理研究所・名誉教授

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，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とが多かった．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．

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十分な対応が可能とはかぎらない．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・名目尺度**(nominal scale）と**順序尺度・順位尺度**（ordinal scale）に，また量的データは**区間尺度・間隔尺度**（interval scale）と**比例尺度・比率尺度**（ratio scale）とに分けて考える．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．またこの量的データ，質的データの枠組みと，連続的，離散的と考える見方との対応は，表１のように，数学的分類と尺度による分類の関係を二元表の形で考えると分かり易いだろう．注意すべきことは，原則としていわゆる質的データは四則演算（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」の発想が生まれる）．

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，見かけ上は計量化せずに扱うことがある．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として必要である．

○数値的か非数値的か

数値的（numerical）
　数量，数値，計数として表記されるもの
非数値的（non-numerical）
　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

構造的データ（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・データベースとして整備されたデータなど
非構造的データ（unstructured data）
　とくに何も手当や加工がなされない生・裸のままのデータ
　自由回答・自由記述のデータ
　ソーシャル・メディア・データ（ツイッター，ブログなどで集まるデータ）
　場合によっては，画像（動画，静止画），音声などのデータ

なお，どのような区分分類を行っても，それで一意的に区分して考えるのではなく，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が肝要である（つまり，分析者が必要とする情報にとって扱い易い，意味ある形で用いるということ）．

表１　データの分類区分

		質的データ		量的データ	
		名義尺度	順序尺度	区間尺度	比例尺度
連続量		（この組み合わせは考えられない）	音の強さの段階的区分 水の濁度・色度 水の硬度（ｱﾒﾘｶ硬度のとき） （*）アナログ量の目盛などを順序化して考える	温度（℃） 金属の硬度（例：ｼｮｱ硬度） 比重 光学光度（ｶﾝﾃﾞﾗ） 光沢度（%） （*）単位に注意	単位を持つ測定値の大部分 長さ(㎝, m, mm)，重さ（kg, g，μg）など 所得金額
離散量	多値	機械名 作業者名 生産工場名 食品の原産地名 電器製品の銘柄	段階評価の成績データ（A, B, C, D） 調査票の選択式質問における選択肢（「満足」「やや満足」「満足でない」）など	体育館の利用日数 車の故障台数 年間の旅行回数 年間読書冊数	車の走行台数 都市内人口数 参加者数 家の戸数 不良品数と良品数
	二値	性別（男、女） 「あり，なし」(有，無) スイッチの状態（「入，切」） など	物体の大きさ（大きい，小さい） 濃度（濃い，薄い） 硬さ（硬い，柔らかい）など	旅行経験の有無の計数（「あり」の回数） PC の所有有無の計数（所有の数）	瓶入りと缶入りのジュース単価（二値分類で層化のとき）

（注）データあるいは測定値の特性をどう分類区分するかということと，それを何らかの加工を経て，別の情報に置き換えること（情報の変換）とは分けて考えた方がよい．ここでいうデータの分類とは，その操作上の“一つの目安”とする考え方である．

２．定性調査におけるテキスト型データや自由回答の役割

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，回収率低下，非標本誤差や調査不能・無回答の増大，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，郵送法，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．

また，諸研究分野でも，定量型のデータ収集方式（data collection mode）から，定性型に移行する傾向にある．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，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行われる量的な調査（たとえば従来型調査の中心であった面接調査，留置調査，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（たとえば回収率の低下），質的調査や定性調査に関心が移行する傾向が見られる．

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のだが，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が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（これ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を意味する）．

とくにインターネットの普及により，**電子的調査情報取得手法**（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，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，自由回答質問を多用する調査（とくに消費者動向調査，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大量情報をどのように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・文字型によるデータ取得や解析の機会の増大が考えられる．

3. 対応分析の数理

以上を前置きとして，WordMiner の多次元データ解析ツールの一つである**対応分析法**（コレスポンデンス分析）について，ここで簡単に紹介する．なるべく数値例により，また WordMiner が出力する数値例に照合させながら説明するが，必要な最小限の数式は用意することにする．数理の詳細を知りたい場合は，参考文献に挙げた資料をみるとよい．

3．1　対応分析法と数量化法 III 類

対応分析法（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，形式的には質的データの主成分分析と考えることもできる．ベンゼクリは，いわゆるフランス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（対応分析）はこうした研究活動の中で登場した手法の一つで，コレスポンデンス分析（CA：Correspondence Analysis）の名称で欧米圏（とくに英語圏）の研究者に紹介され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．

一方，日本国内では，ベンゼクリよりはるかに早く（1955～1956 年頃），（故）林知己夫が数量化法・数量化理論の一環として，さまざまな手法を提案した（例：数量化法Ⅰ類，Ⅱ類，Ⅲ類，Ⅳ類など）．その一つとして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）」がある．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．

実は，数理的には数量化法 III 類は対応分析法と同等である．しかし林はいわゆる数量化理論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，尺度化と関連させて展開した）．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，二元データ表という“多次元の質的データの主成分分析型手法”として，このピアソンのカイ二乗統計量とクロス表の項目間の関連性（対応）を測る方法を考えた．

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提案の経緯や理念にはかなり異なるものがある．彼らの執筆論文や著書の中に両名の個性的な論述として現れるために，一時期，両手法はあたかも別の方法のように思われてきたことがあるが，実は数理的には同じ方法である．

また，その後，数量化法 III 類，対応分析法に類似の手法が，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てきている．たとえば，同等あるいは類似の手法として，以下がある．たとえば，

- 双対尺度法（dual scaling；西里静彦）
- 逆反復平均法あるいは集群分析法（reciprocal averaging method；M. O. Hill 他）
- 等質性分析（homogeneity analysis；Gifi, J. Meulman 他）

などが知られている．

また欧米，国内の研究に，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多々考案されてきた．たとえば，

- 多重対応分析法（多重クロス表・バート表の対応分析）
  （MCA：Multiple Correspondence Analysis）
- 変形多重化クロス表への適用
  （N.C. Lauro の手法他，対数線形モデルとの関連研究など）
- 正準対応分析法（Canonical CA）
- 連関分析法（Association Analysis；L. A. Goodman）

その他，無数にある．また手法相互の関連性についても多くの研究報告がある．

3．2 対応分析法の要約 ―仕組み―

3．2．1 数量化の本質

以上のように，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．しかしながら，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．テキスト型データ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も含めて，対応分析法・数量化法 III 類の仕組みについて簡単に要約しよう．

まず「数量化とは何か」を考える．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，そもそもおかしい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）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．

さらに数量化で重要なことは，あらかじめ「線形である」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，いかにして線形にできるか」，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．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，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．林［1］）．

しかし形式論的にいえば，たとえば数量化法Ⅰ類は一般化重回帰分析モデルである．数量化法Ⅱ類も判別分析の変形として位置づけられる．さらにもっとも広く利用されている数量化法Ⅲ類も質的データの主成分分析型手法という言い方も可能であろう．フランスのベンゼクリ（Benzécri）が提案したように（後述），分割表という多次元データ表についてピアソンのカイ二乗統計量による独立性検定を別の視点から考察するという着想から得た対応分析による定式化もあるが，これは林による数量化・尺度化の発想とは異なる．

まず単純な例をみる．いま，ある調査質問の選択肢として「非常に満足」に５を与え，以下同じように「満足」＝４，「わからない」＝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．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因子分析を行う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」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．

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手法であるが，ここらの詳しいことは文献を参照していただきたい（たとえば林［1］，［5］）．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（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る」という視点にある．そして，この考え方は，定性情報の典型例である**テキスト型データの解析**（textual data analysis）に通底することとしてそのまま応用できるのである．

３．２．２　分析対象とするデータ表

これを要約するといくつかのパターンに分けられる．

① 原則として二元のデータ表（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

ここで言う**二元のデータ表**（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
- 行または列の“プロフィルが意味のある”データ（注：プロフィルについては後述）
- 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．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．

- 通常の二元クロス表あるいは分割表
- （0，1）型データ行列（二元クロス表の特別な場合と考えられる；次の例１，例２参照）
- 多重クロス表（バート表）（多重であって「多元」ではないことに注意）
- 多くの統計表（数値が非負の集約データで，上の条件を満たすようなとき）

ここで，二元クロス表，(0，1) 型データ行列，多重クロス表といったデータ（表）の間には，密接な関係があって，実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）．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１：二値型応答データ

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，表２のデータ表では，４名のサンプル（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．これは後述するように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）．

表2　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

この二値型データ表は，文字情報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だが，いままでは，これをソフトで処理するために，わざわざ“コーディング”して表 2 とするなどを行ってきた．

表3　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**文字変数（テキスト型データ）**として書き方を変えることができる，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このままでは計量的な処理は難しいことを示している．

例２：好みの清涼飲料水の選択

表 4 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（ある論文のデータ表を若干リメイクした）．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（注：ここは形式的に「0，1」を対応させたが，実は「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）．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 5 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，WordMiner ではこの表 5 の形式のデータ表を扱うことができる）．たとえば，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だろう．

表4　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5　好きな清涼飲料水の文字表現（表4を文字情報で表現したとき）

ｻﾝﾌﾟﾙ番号	サンプルが選んだ「好む」清涼飲料	ｻﾝﾌﾟﾙ番号	サンプルが選んだ「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：クロス表の例

ある調査（環境意識調査）で用いた２つの質問から作成したクロス表の例を示す．ここでは回答総数＝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）なども起こりうるが（またその理由付けが調査の質の評価の観点からは重要なのだが）ここでは除外してある．

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．（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 B：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

ところでここで，元のデータ表（つまり回収した調査データ）の一部を表 7 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．この多変量の中から 2 つの質問 A，B を選んで得られたのが上のクロス表 6 となる．この加工過程を良く知っておくことが，数量化法 III 類・対応分析法を理解するうえで重要である．実は数量化法 III 類・対応分析法はどちらのデータ表（表 6 および表 7 の 2 項目を指定した表）からも算出でき（分析可能で），しかも結果は同等である（後述）．

表6　クロス表の例（表側に質問 A，表頭に質問 B が対応）

度数 列% 行%	1. かなり多い	2. 多いほう	3. ふつう	4. 少ない	4. 少ないほう	行　和 行%
1. たいへん気に入っている	166 54. 43 31. 68	239 27. 19 45. 61	86 18. 49 16. 41	7 10. 14 1. 34	26 11. 40 4. 96	524 26. 93
2. まあ気に入っている	131 42. 95 10. 61	598 68. 03 48. 42	324 69. 68 26. 23	36 52. 17 2. 91	146 64. 04 11. 82	1, 235 63. 46
3. あまり気に入っていない	6 1. 97 3. 49	40 4. 55 23. 26	55 11. 83 31. 98	20 28. 99 11. 63	51 22. 37 29. 65	172 8. 84
4. 気に入っていない	2 0. 66 13. 33	2 0. 23 13. 33	0 0. 00 0. 00	6 8. 70 40. 00	5 2. 19 33. 33	15 0. 77
列　和 列%	305 15. 67	879 45. 17	465 23. 90	69 3. 55	228 11. 72	1, 946

（注）ここでは行比率，列比率データも並記してある．実は，集計操作で比率データを観察することは，後述するように対応分析の行うことに類似性がある（後述の「プロフィル」を参照）．

表7　元の調査データの一部

サンプル番号	地点番号	年号コード	いつごろから，現在のまちで，暮らしていますか。	近くの緑地や公園に，どのくらい出かけますか。	あなたは，いま住んでいるまちが気に入っていますか。(選択肢)	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)	(19)住んでいる地区の“緑”の量が少ない(選択肢)	あなたは，いま住んでいるまちが気に入っていますか。(ｺｰﾄﾞ)	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。(ｺｰﾄﾞ)	住んでいるまちが気に入っているか。(ｺｰﾄﾞ)	緑(みどり)が多いと感じるか。(ｺｰﾄﾞ)	その緑地や公園は，歩いて行けば何分ぐらいのところにありますか。
30	35	1	56	1	1. たいへん気に入っている	2. 多いほうである	3. あまりない	1	2	1	2	4
29	35	1	53	3	2. まあ気に入っている	2. 多いほうである	3. あまりない	2	2	2	2	6
27	35	1	42	3	2. まあ気に入っている	2. 多いほうである	3. あまりない	2	2	2	2	3
26	35	1	56	3	1. たいへん気に入っている	1. かなり多い	4. まったくない	1	1	1	1	5
25	35	1	53	1	2. まあ気に入っている	2. 多いほうである	3. あまりない	2	2	2	2	5
23	35	1	42	2	2. まあ気に入っている	2. 多いほうである	3. あまりない	2	2	2	2	5
22	35	1	54	1	2. まあ気に入っている	2. 多いほうである	4. まったくない	2	2	2	2	5
19	35	1	42	3	2. まあ気に入っている	2. 多いほうである	3. あまりない	2	2	2	2	5
17	35	1	47	4	2. まあ気に入っている	2. 多いほうである	3. あまりない	2	2	2	2	7
15	35	1	54	3	2. まあ気に入っている	1. かなり多い	3. あまりない	2	1	2	1	3
14	35	1	56	2	1. たいへん気に入っている	1. かなり多い	3. あまりない	1	1	1	1	1
13	35	1	42	5	2. まあ気に入っている	3. ふつう	3. あまりない	2	3	2	3	5
12	35	1	50	4	2. まあ気に入っている	1. かなり多い	3. あまりない	2	1	2	1	5
8	35	1	54	2	1. たいへん気に入っている	1. かなり多い	4. まったくない	1	1	1	1	1
7	35	1	54	3	2. まあ気に入っている	2. 多いほうである	3. あまりない	2	2	2	2	3
6	35	1	42	2	1. たいへん気に入っている	1. かなり多い	4. まったくない	1	1	1	1	3
2	35	1	57	3	2. まあ気に入っている	1. かなり多い	3. あまりない	2	1	2	1	5
1	35	1	44	5	1. たいへん気に入っている	2. 多いほうである	3. あまりない	1	2	1	2	1
4	35	1	42	3	2. まあ気に入っている	1. かなり多い	4. まったくない	2	1	2	1	5
11	35	1	46	2	2. まあ気に入っている	2. 多いほうである	4. まったくない	2	2	2	2	10
30	30	1	54	4	2. まあ気に入っている	2. 多いほうである	3. あまりない	2	2	2	2	15
28	30	1	99	3	1. たいへん気に入っている	2. 多いほうである	3. あまりない	1	2	1	2	5
29	30	1	54	2	3. あまり気に入っていない	3. ふつう	3. あまりない	3	3	3	3	5
27	30	1	54	1	2. まあ気に入っている	2. 多いほうである	4. まったくない	2	2	2	2	10
26	30	1	37	4	1. たいへん気に入っている	2. 多いほうである	4. まったくない	1	2	1	2	10
25	30	1	55	4	2. まあ気に入っている	4. 少ないほう	3. あまりない	2	4	2	4	5
23	30	1	37	5	1. たいへん気に入っている	2. 多いほうである	9. 無回答	1	2	1	2	20
22	30	1	56	5	2. まあ気に入っている	2. 多いほうである	3. あまりない	2	2	2	2	99
21	30	1	37	5	2. まあ気に入っている	2. 多いほうである	3. あまりない	2	2	2	2	8
19	30	1	37	4	1. たいへん気に入っている	2. 多いほうである	3. あまりない	1	2	1	2	5

例４　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 A：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 B：神社や、お寺請りをよくするか.
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

表8　元の調査データの一部

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、お寺請りをよくする。(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	お寺請りをしない	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	たまにお寺請りをする	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	あまりお寺請りをしない	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	お寺請りをしない	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	たまにお寺請りをする	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	たまにお寺請りをする	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	あまりお寺請りをしない	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	お寺請りをしない	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	たまにお寺請りをする	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	たまにお寺請りをする	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	お寺請りをしない	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	あまりお寺請りをしない	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	お寺請りをよくする	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	お寺請りをよくする	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	あまりお寺請りをしない	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	あまりお寺請りをしない	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	お寺請りをしない	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	お寺請りをしない	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	あまりお寺請りをしない	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	あまりお寺請りをしない	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	あまりお寺請りをしない	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	たまにお寺請りをする	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	お寺請りをしない	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	お寺請りをよくする	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	あまりお寺請りをしない	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	あまりお寺請りをしない	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	あまりお寺請りをしない	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	たまにお寺請りをする	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	お寺請りをしない	役割は果たしていない

表9　表 8 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
コードに変換)

(注) WordMiner では
左側の表 9 から処理可能

表 10　数値コードの変換

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

表 11　コード変換したデータ表をアイテム・カテゴリー型データ表に変換

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					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 12　（質問 A × 質問 B）のクロス表の生成（度数のみ表示）

		質問 B				
	選択肢	お寺請りをよくする	たまにお寺請りをする	あまりお寺請りをしない	お寺請りをしない	無回答
質問 A	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 13　多重クロス表（バート表）の例（表 11 から生成）

	質問	質問 A					質問 B				
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	たまにお寺請りをする	あまりお寺請りをしない	お寺請りをしない	無回答
質問 A	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 B	お寺請りをよくする	41	25	6	1	1	74	0	0	0	0
	たまにお寺請りをする	26	67	13	6	4	0	116	0	0	0
	あまりお寺請りをしない	22	45	34	7	1	0	0	109	0	0
	お寺請りをしない	15	30	31	27	2	0	0	0	105	0
	無回答	2	0	0	0	7	0	0	0	0	9

表 14　上の表の説明

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

色々な表を作ったここで，各表の関係を要約しよう．

① 表 8 は，元の調査データ，つまり「(回答・サンプル）×（項目・多変量)」の多変量構造データである．
② 表 9 は，表 8 のある 2 つの質問（項目）に注目しこれを切り出した表である．これも「(回答・サンプル）×（項目・多変量＝2 項目)」のデータ表である．
③ 表 10 は，表 9 の各選択肢に数値コードを与えて数値化したデータ表である．
④ 表 11 は，表 9（表 10）**をアイテム・カテゴリー型**に展開したデータ表である（**インジケータ行列**ともいう；indicator matrix)．ここで，質問 A の選択肢は無回答を入れて 5 個，質問 B の選択肢は無回答を入れて 4 個であるから，展開後のデータ表の寸法は，行数は回答者数のまま，列数が 9 個の「(回答・サンプル）×（アイテム・カテゴリー変数)」のデータ表となる．またアイテム・カテゴリー型データ表の場合，各行の和がいずれも項目数（ここでは 2）となること，つまり行方向にみた「1 の総数＝項目数」となることに注意しよう．
⑤ 表 12 は，表 9 から作った二元のクロス表である．
⑥ 表 13 は，表 11 の右側のアイテム・カテゴリー型データ表と，それを転置して得られる行列との積から作られる，いわゆる**多重クロス表**（**バート表**；Burt's table, Burt matrix）である．これは明らかに対称行列であり，「(項目）×（項目)」型のデータ表となっている．また，行列の右上ブロックには表 12 のクロス表が入り，その対称に位置する左下ブロックにはこのクロス表を転置したクロス表が入っている．また，対角ブロックには，質問 A と質問 B の選択肢別の度数（周辺度数）が入っている．これ（表 13）を要約説明すると表 14 のようになる．

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，各表から出立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．すなわち，

① 表 11 のアイテム・カテゴリー型データ表の対応分析の結果は，実は表 12 の対応分析の結果に同等である．
② またそれらは「表 13 のバート表の対応分析の結果にも同等」となる（表 11 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 また，表 13 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 11 のアイテム・カテゴリー型の回答（サンプル）の成分スコアを求めることもできる．
④ 表 12 のクロス表と，表 13 の 2 項目のバート表との結果は解析的には同等である（ただし固有値が異なる形で現れる)．

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（大隅[8]，[9])．

例５　クロス表，多重クロス表，アイテム・カテゴリー型データ表の関係

次に例４に類似の例をみる（これも架空の人工データ）．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選択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．

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）の回答者の分布は表 15 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，説明を簡略にするため）．

表 15 （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 16 （項目 I）×（項目 J）の2元クロス表

項目 I ＼ 項目 J	1. 味	2. 量	3. 工夫･サービス	行和
1. さとみ	46	7	42	95
2. バッハ	76	18	48	142
3. ムガール	44	16	49	109
4. いりふね	25	32	98	155
5. コルシカ	77	13	32	122
6. クラーク	14	54	34	102
7. ロゴスキー	35	42	48	125
8. きくみ	8	67	35	110
9. ラ・マレ	82	15	49	146
10. かりや	35	38	105	178
列　和	442	302	540	1, 284

ここで得た各データ表の関係は，例４に同様に解釈すればよい．なお，例４，例５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（演習問題 2 を参照）．

ここでは，

① （回答・サンプル）×（多変量の項目）のデータ表（表 15）
② 多変量の項目について加工生成したアイテム・カテゴリー型のデータ表（表 17）
③ 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）を生成（表 18）
④ この多重クロス表の非対角部のブロック行列として表 16 のクロス表が得られる．

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特徴がある．

表 17 （回答者）×（アイテム・カテゴリー）のデータ表

項目 / 回答者	I 1 さとみ	2 バッハ	3 ムガール	4 いりふね	…	9 ラ・マレ	10 かりや	J 1 味	2 量	3 工夫	3 サービス
1	0	1	0	0	…	0	0	1	0	0	
2	0	0	1	0	…	0	0	0	1	0	
3	1	0	0	0	…	0	0	0	1	0	
4	0	0	0	0	…	1	0	0	0	1	
5	⋮	⋮	⋮	⋮	…	⋮	⋮	⋮	⋮	⋮	
⋮	⋮	⋮	⋮	⋮	…	⋮	⋮	⋮	⋮	⋮	
1, 284	0	0	0	1	…	1	1	0	1	0	

表 18　多重クロス表（バート表）の生成

項目	1. さとみ	2. バッハ	3. ムガール	4. いりふね	5. コルシカ	6. クラーク	7. ロゴスキー	8. きくみ	9. ラ・マレ	10. かりや	1. 味	2. 量	3. 工夫・サービス
1. さとみ	95										46	7	42
2. バッハ		142									76	18	48
3. ムガール			109								44	16	49
4. いりふね				155							25	32	98
5. コルシカ					122						77	13	32
6. クラーク						102					14	54	34
7. ロゴスキー							125				35	42	48
8. きくみ								110			8	67	35
9. ラ・マレ									146		82	15	49
10. かりや										176	35	38	105
1. 味	46	76	44	25	77	14	35	8	82	35	442		
2. 量	7	18	16	32	13	54	42	67	15	38		302	
3. 工夫・サービス	42	48	49	98	32	34	48	35	49	105			540

（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．

4．　対応分析法の数理

4．1　準備

以上を前置きとして，**対応分析法の考え方**に従ってその仕組みを簡単に記述する（つまりベンゼクリのいうフランス流の定性的な多次元データの主成分分析型アプローチ）．しかしすでに述べたように，これは数量化法Ⅲ類と同等である．結果として，数量化法Ⅲ類と同等であることは後述の例題の中で説明する．

対応分析法では，出発行列として“**二元のデータ表**”，たとえばもっとも単純には“クロス表（分割表）”を考えればよい．上に上げた例１～例５のいずれの表も２元のデータ表の形式であることに注意しよう．いま寸法が（$m\times n$）**の二元クロス表型データ表**を記号（式）で以下のように表す．ここで，f_{ij} はクロス表の $(i,\ j)$ セル内の度数である（前の例 4，例５などを思い出そう）．

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

ここで，I と J は，それぞれ行と列の項目の選択肢の集合を表わし以下のように書いておく．つまりクロス表でいえば質問の**選択肢**（カテゴリー，オプション）に相当する（表 19 参照，以下「項目」とその「選択肢」という表現を用いる）．

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

表 19（項目 I× 項目 J）のクロス表 $\mathbf{F}_{m\times n}=\left(f_{ij}\right)$

		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_{++}

次に寸法が（m×n）の二元のクロス表，つまり表 19 のクロス表から作られる**相対度数**つまり**確率分布**を考える．これを以下のように表す．

$$\mathbf{P}_{IJ}=\left(p_{ij}\right)\quad\left(i\in I,j\in J\right)\text{（同時確率分布）}\tag{3}$$

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

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

ここで，

$$p_{ij}=\frac{f_{ij}}{N},\ N=\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}\ \left(\equiv f_{++}\right)\tag{6}$$

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

である．また $diag(\cdot)$は対角行列を意味する．以上を表と模式図に表すと次のようになる（表 20，図 1）．

ここで，表 20 は，式（3），（4），（5）に対応する行列，ベクトルを表す．左の図 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．

表 20　確率行列 $\mathbf{P}_{IJ}$

		項　目　J						
		1	2	…	j	…	n	行の確率（$\mathbf{P}_I$ の対角要素）
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）	p_{+1}	p_{+2}	…	p_{+j}	…	p_{+n}	1

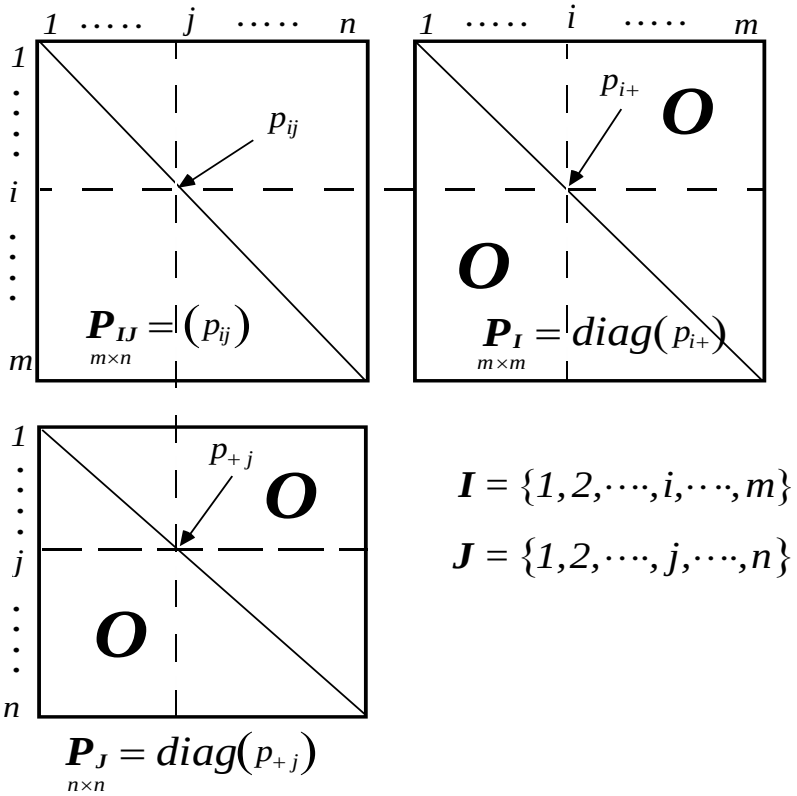


図1　確率行列の構成

4．2　プロフィルとは

対応分析法では“**プロフィル**”の概念が重要である．プロフィ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，行と列のプロフィルがある．

①**行のプロフィル**（つまり行の比率パターン）

$$\mathbf{N}_I = \left\{ q_{ij} = \frac{p_{ij}}{p_{i+}} \middle| i \in I, j \in J \right\} \quad \text{（行のプロフィル）} \tag{8}$$

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

③ **列のプロフィル**（つまり列の比率パターン）

$$\mathbf{N}_J = \left\{ q_{ij}^* = \frac{p_{ij}}{p_{+j}} \middle| i \in I, j \in J \right\} \quad \text{（列のプロフィル）} \tag{9}$$

ここでは，$\sum_{i=1}^{m} q_{ij}^* = 1$ の制約から，列のプロフィルは（$m-1$）次元空間に分布する n 個の点の集合である（図 2，3）．なお，$\mathbf{N}_I$，$\mathbf{N}_J$ のプロフィルの分布のことを（つまり成分スコアの布置のことを）フランス流には**“雲”**（nuage，英語の cloud）という．

ここに見るように，クロス表という多次元データ表を，行と列との両方から考えることが対応分析の特徴である（つまり行と列との関連性・対応を考察すること，よって**対応分析法**である）．

（**注1**） ここで，行の相対確率 p_{i+}，列の相対確率 p_{+j} は，それぞれ行の重心，列の重心と呼ぶこともある．

このプロフィルには次の性質がある．

（1） プロフィルとは行あるいは列の**比率のパターン**を考えることである．

（2） したがって，データ（測定値）の実質的な量・大きさを見ているわけではない（この点で主成分分析とは異なる）．

（3） たとえば，（学生・サンプル）×（科目・変量）のデータ表とし，測定値が試験成績（得点）を考えたとき，

- 実得点の特徴，科目間の関連性や総得点の序列，（成績点の）高低を見るなら，主成分分析を使うことになる．
- 科目の学生別パターンや均衡，（相対的に）どの科目で浮き沈みがあるのか，成績得点の傾向（パターン）を見るなら対応分析を使う．

（4） 従って，対応分析はデータ表のセル内の数値・頻度の大きさとプロフィルの分布のバランスに敏感である（少数頻度のセル，はずれ値などの影響が大きい）

ここで，確率行列，行と列とのプロフィル，布置図，同時布置図などの関係を模式的に示しておこう（図 2，図 3）．

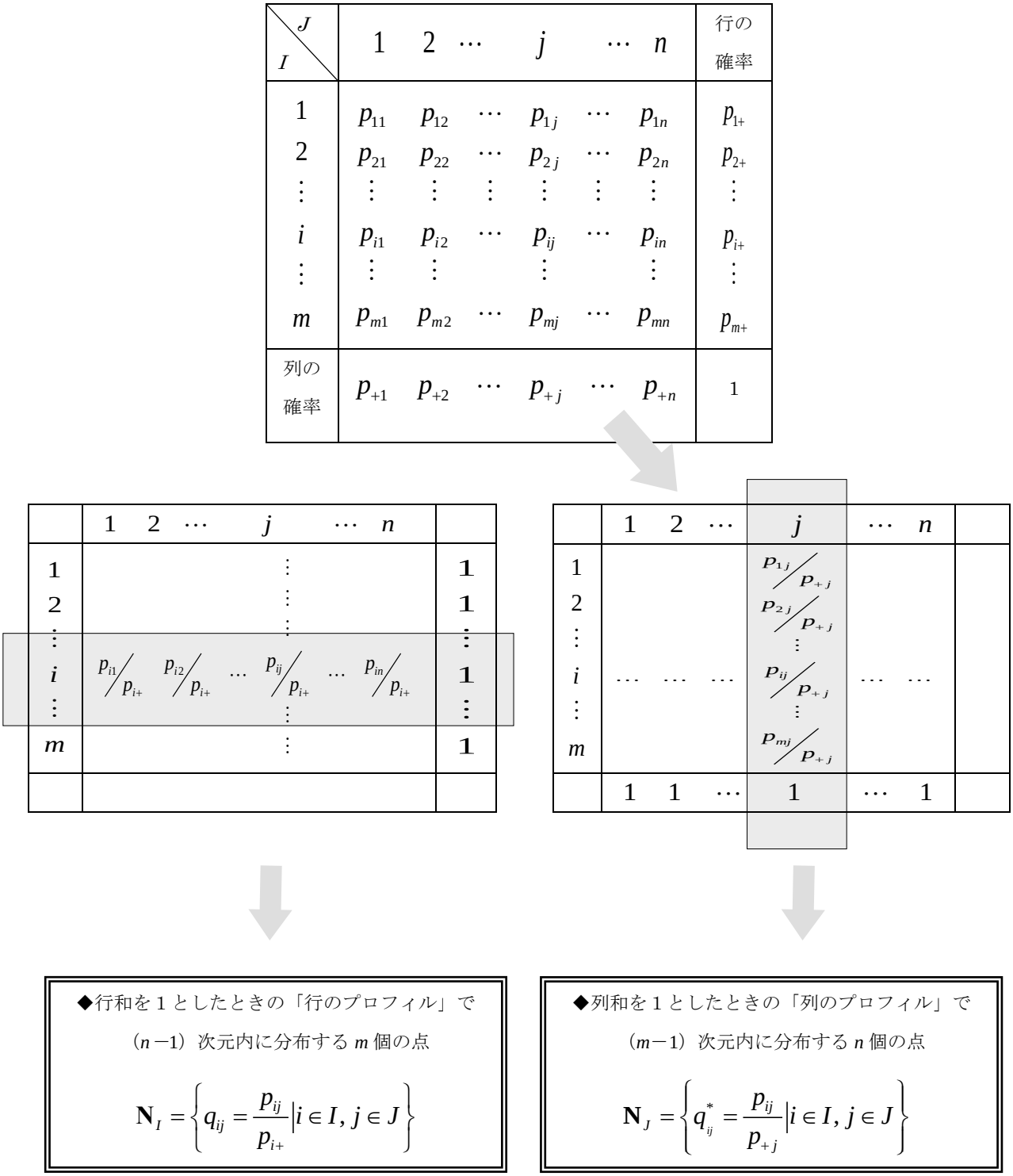
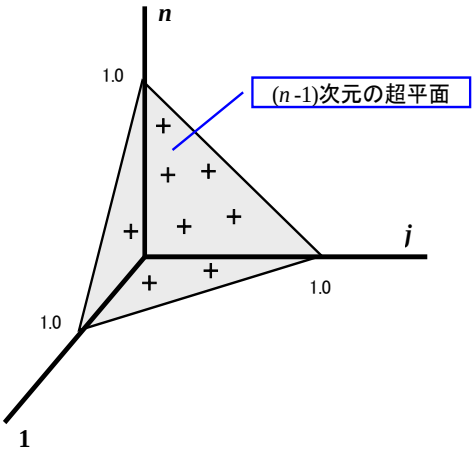
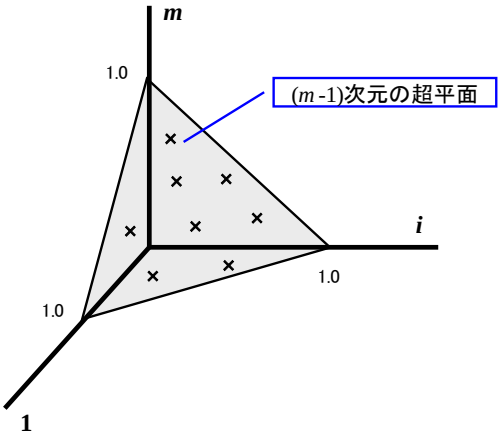


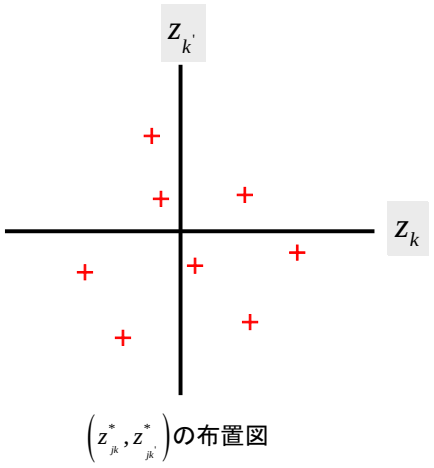
図 2　行のプロフィルと列のプロフィルの関係

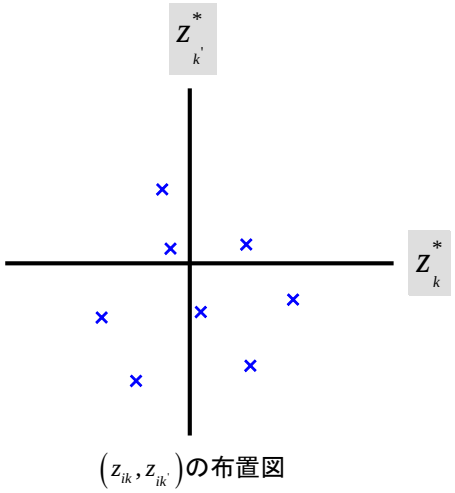




$$\mathbf{N}_I = \left\{ q_{ij} = \frac{p_{ij}}{p_{i+}} \middle| i \in I, j \in J \right\} \text{の分布}$$

$$\mathbf{N}_J = \left\{ q^*_{ij} = \frac{p_{ij}}{p_{+j}} \middle| i \in I, j \in J \right\} \text{の分布}$$





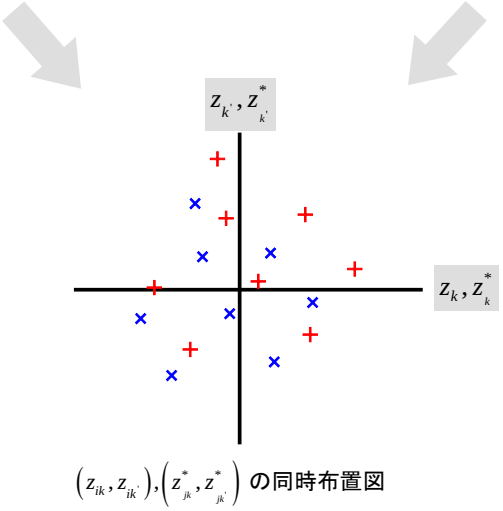


図 3　行，列のプロフィルと成分スコアの布置図の関係

（†）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，WordMiner のデフォルトは（1, 2）成分を設定

[プロフィルと重心座標系の例]

プロフィルの考え方は，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見るのが早い. 前に例 5 としたレストラン評価のデータ表を考えよう（寸法は m=10, n=3）. プロフィルとは行または列の比率（相対比率）の分布であるから，これに相当するプロフィル $q_{ij}, q_{ij}^{*}\left(i \in I, j \in J\right)$ を実際に作ってみる.

表 21　行のプロフィルの分布

評価項目 / レストラン	1. 味	2. 量	3. 工夫 サービス	行　和
1. さとみ	0. 484	0. 074	0. 442	1. 000
2. バッハ	0. 535	0. 127	0. 338	1. 000
3. ムガール	0. 404	0. 147	0. 450	1. 000
4. いりふね	0. 161	0. 206	0. 632	1. 000
5. コルシカ	0. 631	0. 107	0. 262	1. 000
6. クラーク	0. 137	0. 529	0. 333	1. 000
7. ロゴスキー	0. 280	0. 336	0. 384	1. 000
8. きくみ	0. 073	0. 609	0. 318	1. 000
9. ラ・マレ	0. 562	0. 103	0. 336	1. 000
10. かりや	0. 197	0. 213	0. 590	1. 000
列の平均ベクトル	0. 344	0. 235	0. 421	1. 000

$$\mathbf{N}_I = \left\{ q_{ij} = \frac{p_{ij}}{p_{i+}} \middle| i \in I, j \in J \right\}$$

これは 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. これを，$(n-1)=3-1=2$, つまり 2 次元の空間内に布置する 10 のレストランと考える．また列の平均ベクトル（平均比率）は重心に相当する（図 4 の中の G がそれに相当）.

表 22　列のプロフィルの分布

評価項目 / レストラン	1. 味	2. 量	3. 工夫 サービス	行の平均 ベクトル
1. さとみ	0. 104	0. 023	0. 078	0. 078
2. バッハ	0. 172	0. 060	0. 089	0. 089
3. ムガール	0. 100	0. 053	0. 091	0. 091
4. いりふね	0. 057	0. 106	0. 181	0. 181
5. コルシカ	0. 174	0. 043	0. 059	0. 059
6. クラーク	0. 032	0. 179	0. 063	0. 063
7. ロゴスキー	0. 079	0. 139	0. 089	0. 089
8. きくみ	0. 018	0. 222	0. 065	0. 065
9. ラ・マレ	0. 186	0. 050	0. 091	0. 091
10. かりや	0. 079	0. 126	0. 194	0. 194
列　和	1. 000	1. 000	1. 000	1. 000

$$\mathbf{N}_J = \left\{ q_{ij}^{*} = \frac{p_{ij}}{p_{+j}} \middle| i \in I, j \in J \right\}$$

これは q_{ij}^{*} を要素とする行列で，

$\sum_{i=1}^{m} q_{ij}^{*} = 1$（列和＝１）となっている．これを，（$m-1$）＝10－1＝９, つまり 9 次元の空間内に布置する 3 つの評価項目と考える．また行の平均ベクトル（平均比率）は重心に相当する.

ここで，表 21 のプロフィル $\mathbf{N}_I = \left\{ q_{ij} = \dfrac{p_{ij}}{p_{i+}} \middle| i \in I, j \in J \right\}$ を図 3 に合わせて描くと，

$\sum_{j=1}^{3} p_{+j} = 1$ の制約があるから（$n-1$）=3−1=2 次元空間内の点として布置される（自由度が 2 となる）．これがつぎの図 4 の左図であり，ここで 2 次元平面となった（図でアミをかけた部分の）**重心座標系**（barycentric coordinate system）という．ここでは布置が 2 次元となったことで三角座標系（triangular coordinate system）で表せるのでこれを実際に描いてみると右図のようになる（この三角図の作図は統計ソフト JMP を利用した）．

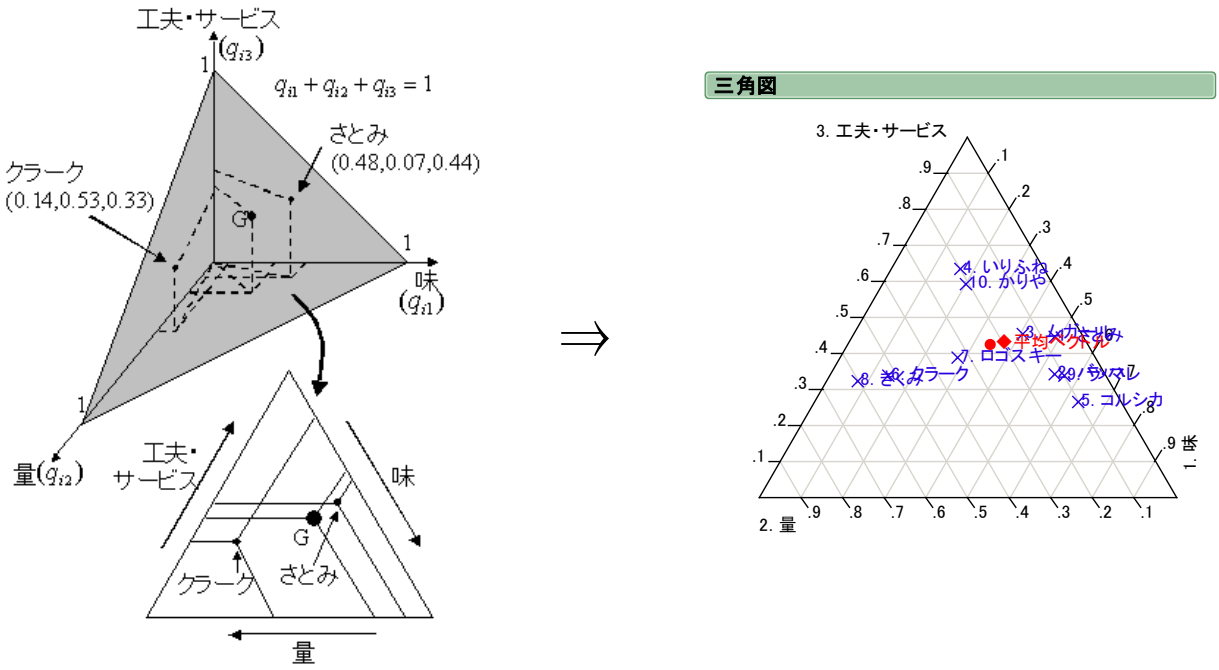


図4　行「レストラン」の布置の考え方（プロフィルを重心座標系に射影）

そして，この図上に分布する点（いまは 10 のレストラン）を考え，この空間内である種の主成分分析を行うというのが対応分析である．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 つの評価項目の分布を考えればよい．このようにプロフィルを行と列の双方向から考える．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．

そして表 21 ついて対応分析を適用し，実際に**固有値，成分スコア**を算出すると，次の結果を得る．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（$n-1$）=3−1=2 となるので固有値の数は 2 個となる．従って，2 つの固有値に対応する 2 つの成分スコアが，行と列とのそれぞれの選択肢，つまり 10 のレストランと 3 個の評価項目に与えられる．これが，表 23 にある成分スコアの一覧である（後述する表 25 に相当）．これらの諸統計量については以下の節で述べる．

表 23　レストランと評価項目への成分スコア

		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	さとみ	0.40067	−0.09077
	バッハ	0.39656	0.12200
	ムガール	0.19686	−0.08210
	いりふね	−0.20169	−0.40820
	コルシカ	0.54972	0.25857
	クラーク	−0.66717	0.25584
	ロゴスキー	−0.21980	0.10024
	きくみ	−0.85898	0.30915
	ラ・マレ	0.46355	0.11909
	かりや	−0.16472	−0.32610
項目と選択肢＼成分		z^*_{j1}	z^*_{j2}
項目 J	味	0.52347	0.17643
	量	−0.65787	0.25247
	工夫・サービス	−0.06055	−0.28561

表 24　固有値と寄与率

主成分 k	固有値 λ_k	寄与率(%)
1	0.19766	76.71
2	0.06002	23.29

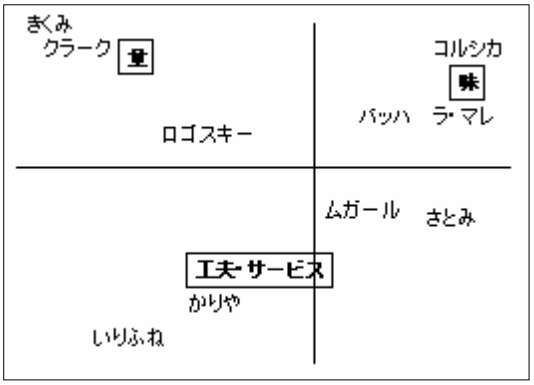


図 5　成分スコアから得た同時布置図

（注）成分スコアの図で，レストランの布置が図 4 の三角座標のそれに類似していることに注意（図 4 の布置が再現されている）．

4，3　データ行列の生成と解法

このように（二元の）クロス表を出発行列とし，上に準備した情報を用いて，

$$x_{ij}=\frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}=\frac{q_{ij}}{\sqrt{p_{+j}}}\ \left(i\in I, j\in J\right) \tag{10}$$

または

$$x^*_{ij}=\frac{p_{ij}}{p_{j+}\sqrt{p_{i+}}}=\frac{q^*_{ij}}{\sqrt{p_{i+}}}\ \left(i\in I, j\in J\right) \tag{11}$$

を行列要素とする行列を作る．たとえば，x_{ij} を要素とする行列を次のように作る．

$$\underset{m\times n}{\mathbf{X}}=\left(x_{ij}\right)\ \left(i\in I, j\in J\right) \tag{12}$$

これが対応分析法における基本のデータ行列となる．ここで，プロフィル（比率データ）そのものを用いずに，$1/p_{+j}$，$1/p_{i+}$ を加重とした要素とすることに大切な意味があるのだが，ここでは形式的に記しておく（理由のいくつかについては後述する）．なおこの座標系を前にも示したように**重心座標系**（barycentric coordinate system）という．とくに，項目が 3 項目のときには，次元数が 2 となりいわゆる**三角図**（三角座標系：triangular coordinate system）となる（上でみたレストランの例を参照）．

あるいは，これを次のように書き替えて，

$$y_{ij}=\frac{p_{ij}}{\sqrt{p_{i+}p_{+j}}}=\frac{f_{ij}}{\sqrt{f_{i+}f_{+j}}}\quad\left(p_{i+}\neq 0,\,p_{+j}\neq 0;\,f_{i+}\neq 0,\,f_{+j}\neq 0\right)\tag{13}$$

を要素とする次の行列を作っても結果が同等であることが知られている．

（注2）かりにここで行和，列和がゼロとなったときには該当の列あるいは行のスクイズを行う．

（注3） 標準的な対応分析とは， y_{ij} を要素とする次の行列 $\underset{m\times n}{\mathbf{Q}}=\left(y_{ij}\right)$ の**固有値問題**あるいは**特異値分解**を行うことである．

$$\underset{m\times n}{\mathbf{Q}}=\left(y_{ij}\right)\ \left(i\in I,\,j\in J\right)\tag{14}$$

これは前に用意した各行列を用いると，以下のように書き替えられる．

$$\underset{m\times n}{\mathbf{Q}}=\mathbf{P}_I^{-1/2}\mathbf{P}_{IJ}\mathbf{P}_J^{1/2}\tag{15}$$

従って，この行列 $\mathbf{Q}$ から次の分散共分散行列を作り，これの**固有値問題**として処理すればよい．

$$\underset{n\times n}{\mathbf{V}}=\mathbf{Q}^t\mathbf{Q}=\mathbf{P}_J^{-1/2}\mathbf{P}_{JI}\mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{1/2}\ \text{（}x_{ij}\text{ の}\textbf{分散共分散行列}\text{に相当する）}\tag{16}$$

ここで，$\mathbf{Q}^t$ は $\mathbf{Q}$ の転置行列を，$\mathbf{P}_{JI}$ は $\mathbf{P}_{IJ}$ の転置行列を，それぞれ表す．対応分析とは，この行列 $\underset{n\times n}{\mathbf{V}}$（分散共分散行列に相当）の固有値問題（あるいは、特異値分解，スペクトル分解）を考えることに帰着する．つまり，行列 $\underset{m\times n}{\mathbf{X}}$ あるいは $\underset{m\times n}{\mathbf{Q}}$ をデータ行列と見立てたときの主成分分析に他ならない．数量化法 III 類では，この $\underset{n\times n}{\mathbf{V}}$ の形が非対称行列となって現れるだけで，実は解は同じことになることが知られている．

以上のように求めた**固有値**（**特異値**の二乗）と**固有ベクトル**を用いて，いわゆる“**成分スコア**”（数量化得点，数量化スコアなどともいう，1 種の加重和・合成指標）を求める．WordMiner を用いるとこの成分スコアが得られるので，ここでは具体的な算出式は省略する．これが WordMiner ではどのように出力され，またどのように解釈するかが重要であるのでこれについて必要最小限の情報を示すにとどめる．

4．4 成分スコアとその性質（とくに双対性）

対応分析法では**成分スコア**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付与される．そして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分析する）．また，成分スコアともとのデータ表（二元データ表）との関係は図式で模式的に眺めることが理解を容易にするので，これを以下に示す．

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次のように表す（WordMiner が自動的に算出してくれる．後述の数値例を参照）．

$$z_{ik}\left(i \in I, k = 1,2,\cdots,K\right)$$ （選択肢 i に対する第 k 成分の成分スコア） (17)

$$z^{*}_{jk}\left(j \in J, k = 1,2,\cdots,K\right)$$ （選択肢 j に対する第 k 成分の成分スコア） (18)

これと，元の二元データ表との関係を模式的に示す（図 6）．

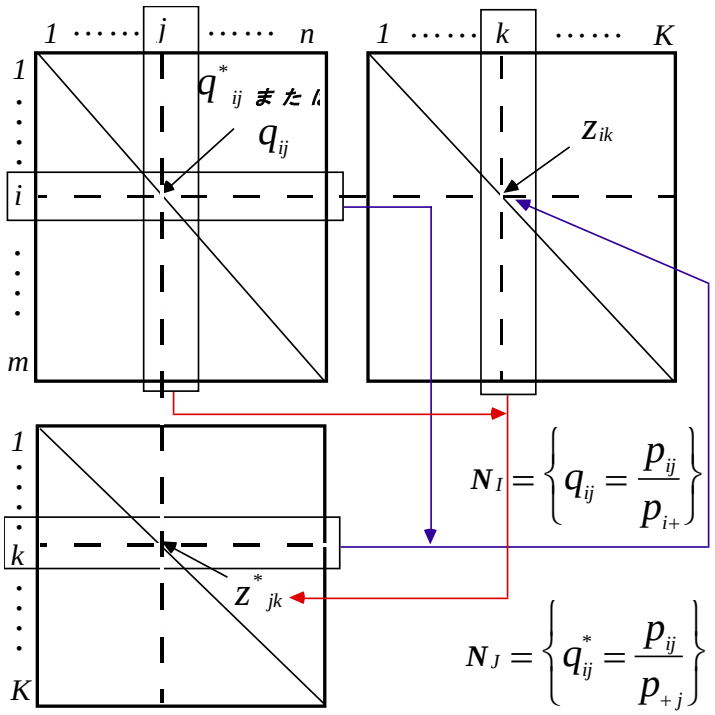


図 6　成分スコアとプロフィルの関係（双対性）を示す模式図

ここでは，２項目 I，J の各選択肢に付与された成分スコア間の関係も同時に示しており，これは以下のように書き表すことができる．

$$z_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{n} \left(\frac{p_{ij}}{p_{i+}} \right) z^{*}_{jk} \left(i \in I, k = 1,2,\cdots,K\right) \tag{19}$$

$$z^{*}_{jk} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{m} \left(\frac{p_{ij}}{p_{+j}} \right) z_{ik} \left(j \in J, k = 1,2,\cdots,K\right) \tag{20}$$

この式の意味は重要である．これを読み解くと「項目 I のある選択肢 i の成分スコアは，項目 J の選択肢への成分スコアのプロフィルの加重平均となる」こと，一方，「項目 J のある選択肢 j の成分スコアは，項目 J の選択肢への成分スコアのプロフィルの加重平均となる」という重要な性質がある．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 6 と表 25 で確認）．つまり**双対性**（duality）の関係にあり，また上の式（19），（20）を互いに**推移関係**（transition relationship）にあるという．

表 25　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 2 ⋮ i ⋮ m	p_{11}	p_{12}	⋯	p_{1j}	⋯	p_{1n}	z_{11}	z_{12}	⋯	z_{1k}	⋯	$z_{1k'}$	⋯	z_{1K}
		p_{21}	p_{22}	⋯	p_{2j}	⋯	p_{2n}	z_{21}	z_{22}	⋯	z_{2k}	⋯	$z_{2k'}$	⋯	z_{2K}
		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
		p_{i1}	p_{i2}	⋯	p_{ij}	⋯	p_{in}	z_{i1}	z_{i2}	⋯	z_{ik}	⋯	$z_{ik'}$	⋯	z_{iK}
		⋮	⋮		⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
		p_{m1}	p_{m2}	⋯	p_{mj}	⋯	p_{mn}	z_{m1}	z_{m2}	⋯	z_{mk}	⋯	$z_{mk'}$	⋯	z_{mK}
成分スコア	1 2 ⋮ k ⋮ k' ⋮ K	z^*_{11}	z^*_{21}	⋯	z^*_{j1}	⋯	z^*_{n1}	⇑ 行の**項目 I の選択肢の成分スコア**							
		z^*_{12}	z^*_{22}	⋯	z^*_{j2}	⋯	z^*_{n2}								
		⋮	⋮	⋮	⋮	⋮	⋮								
		z^*_{1k}	z^*_{2k}	⋯	z^*_{jk}	⋯	z^*_{nk}	⇐ 列の**項目 J の選択肢の成分スコア**							
		⋮	⋮	⋮	⋮	⋮	⋮								
		$z^*_{1k'}$	$z^*_{2k'}$	⋯	$z^*_{jk'}$	⋯	$z^*_{nk'}$								
		⋮	⋮	⋮	⋮	⋮	⋮								
		z^*_{1K}	z^*_{2K}	⋯	z^*_{iK}	⋯	z^*_{nK}								

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．

4．5　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」を描いて観察する．

①　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 25 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．

$$\left(z_{ik}, z_{ik'}\right) \quad \begin{pmatrix} i = 1,2,\cdots,m \\ k,k' = 1,2,\cdots,K \\ K = \min\{m,n\} - 1 \end{pmatrix} \qquad \text{（行の選択肢への成分スコア）} \qquad (21)$$

$$\left(z^{*}_{jk}, z^{*}_{jk'}\right) \quad \begin{pmatrix} i=1,2,\cdots,m \\ k,k'=1,2,\cdots,K \\ K=\min\{m,n\}-1 \end{pmatrix} \qquad \text{（列の選択肢への成分スコア）} \qquad (22)$$

［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．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．

(2) 次に，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．

(3) 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．

(4) 「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となる）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）．

(5) 固有値，寄与率の解釈は，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注意する（高い寄与率は数理的に現れることがない，付録の［補足］参照）．

(6) 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．

(7) 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*-means 法や階層的分類ではうまく対応できないことがある，よって WordMiner では階層的分類法の 1 つウォード法と *k*-means 法とを併用したハイブリッド法を用いる）．

(8) **“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（プロフィルの不均衡）から生じる．これは対応分析の特徴でもある．

(9) 行あるいは列の各選択肢に付与された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「標準化する場合」と「標準化しない場合」があるので，4 通りの組み合わせがある（下の表）．WordMiner では，**いずれも標準化しない**（成分スコアの分散は固有値 λ_k のままを用いている）．その理由については参考文献を参照のこと（大隅他［8］，Lebart 他［24］）．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ている（注：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，元のクロス表の行和，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）．

表 26　成分スコアの分散の組み合わせ

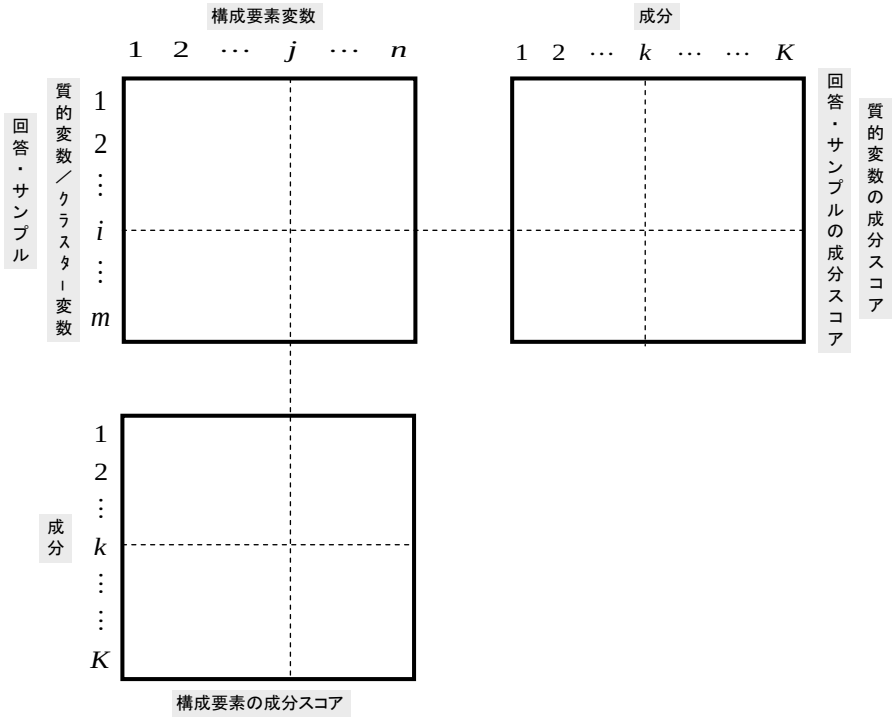
	項目 I の選択肢の 成分スコア：z_{ik}	項目 J の選択肢の 成分スコア：z^*_{jk}
分散の大きさ	λ_k λ_k 1 1	λ_k 1 λ_k 1

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**という．すなわち，

$$\left(z_{ik}, z_{ik'}\right), \left(z^*_{jk}, z^*_{jk'}\right) \qquad \begin{pmatrix} i = 1,2,\cdots,m \\ k,k' = 1,2,\cdots,K \\ K = \min\{m,n\} - 1 \end{pmatrix} \tag{23}$$

を同じ散布図のプロット図として図式化する．よって，元のデータ表と行の選択肢の成分スコア，列の選択肢の成分スコアの関係は図 7 のように考えればよい．



（注）ここで，（回答・サンプル）×（構成要素変数），（質的変数）×（構成要素変数）のいずれも二元クロス表形式であることに注意する．行と列との，同じ成分に対応する成分スコアを取り出して同時布置図として図式化する．

図 7　元のデータ表と成分スコアの関係（表 25 に対応する）

これは前にみた表 25，図 6 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（23））．

③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．とくに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．

(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，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（相互のプロフィルを加重とする平均になっている）．（注：実は，同時布置図の解釈を巡っては，いろいろな議論がある．参考文献を参照のこと）

(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のスコアについて同時的には行えない．

(4) しかし，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．

(5) 布置図は高次元空間内のおよその情報を知るための手がかりとする．

- 非常に疎なデータ表を扱うことが多いので，とくに（回答・サンプル）×（構成要素変数）データ表は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**構成要素の有意性テスト**の一覧などと併用する．
- 無数の構成要素（単語，語句など）を布置した図の視認には限界がある（煩雑になる）ので別の方法と併用する（**有意性テストの結果**を検討）
- 布置図の観察は**“はずれ値の検出”**に有効である．つまり，布置図は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）．

(6) 「(回答・サンプル) ×（構成要素変数)」のデータ表から，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．

(7) 「(構成要素変数) ×（質的変数：属性やクラスター変数)」のデータ表から，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見る．
- どの質的変数が，構成要素に強く関係する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）を取り上げて探査するとよい．

(8) 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**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．

- 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．

(9) 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．

(10) 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．

4．6 対応分析のいくつかの性質

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 WordMiner が出力表示する主要な指標をいくつか説明する．一見難しそうであるが，後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，諸指標がどのような挙動をするものかを知るのもよいだろう．

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．ここで，**固有値の個数**は元の解析対象としたデータ表（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）．

$$tr(\mathbf{V})-1=\sum_{k=1}^{K}\lambda_k \quad \left(K=\min\{m,n\}-1\right) \quad \text{（固有値の和）} \qquad (24)$$

（ここで，$tr(\mathbf{V})$ は行列 $\mathbf{V}$ のトレース＝対角要素の和・跡和，を示す）

$$\textbf{寄与率}：\nu_k=\frac{\lambda_k}{\sum_{k=1}^{K}\lambda_k}\times 100(\%) \quad \begin{pmatrix} k=1,2,\cdots,K \\ K=\min\{m,n\}-1 \end{pmatrix} \quad \text{（第 } k \text{ 成分の寄与率の式）} \qquad (25)$$

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 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）．

③ クロス表の独立性の検定との関係

クロス表の表側と表頭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．これは，表側と表頭の 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**，$p_{ij}=p_{i+}p_{+j}$），これが統計的に棄却され

れば，帰無仮説を棄却，つまり表側と表頭の 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）とする検定法である（ある意味，隔靴掻痒である）．

ところで，このような説明をここで引用した理由は，対応分析法はこれの見方を変えて，実際に表側と表頭の 2 つの項目 I, J の間にどの程度の関係があるのかを具体的な量として示すことにある．たとえば，固有値（の和）とピアソンのカイ二乗統計量との間には，前述のように次の関係がある．

$$tr(\mathbf{V})-1=\frac{\chi_p^2}{N}=\sum_{k=1}^{K}\lambda_k \quad \left(K=\min\{m,n\}-1\right) \tag{26}$$

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，これはここで約束した記号を用いると次のように書ける．

$$\chi_p^2=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{N\left(p_{ij}-p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}} \tag{27}$$

$$\chi_p^2=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{N\left(p_{ij}-p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}}=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(f_{ij}-\dfrac{f_{i+}f_{+j}}{N}\right)^2}{\dfrac{f_{i+}f_{+j}}{N}} \tag{28}$$

（注）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**することを使って検定を行う．

$$\chi_p^2=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(f_{ij}-\dfrac{f_{i+}f_{+j}}{N}\right)^2}{\dfrac{f_{i+}f_{+j}}{N}}\sim\chi_{(m-1)(n-1)}^2 \tag{29}$$

③再生公式

上の関係に関連して以下の公式が知られている．つまり，p_{ij}，p_{i+}，p_{+j} と成分スコアとの間に成り立つ公式である（**再生公式**：reconstitution formula；Fisher’s identity とも言う）．ここで p_{ij} は，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ということを示している．

$$p_{ij}=p_{i+}p_{+j}\left\{1+\sum_{k=1}^{K}\frac{1}{\sqrt{\lambda_k}}z_{ik}z_{jk}^{*}\right\}=p_{i+}p_{+j}+p_{i+}p_{+j}\left\{\sum_{k=1}^{K}\frac{1}{\sqrt{\lambda_k}}z_{ik}z_{jk}^{*}\right\} \tag{30}$$
$$\left(i\in I,\ j\in J,K=\min\{m,n\}-1\right)$$

この数式の右辺の第 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帰無仮説（つまり独立モデル：$p_{ij}=p_{i+}p_{+j}$）となっている

ことに注意しよう．また，第 2 項に成分スコアが含まれ，この項が２つの項目間の関連性を測っていることになる．このようにここでも，ピアソンの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ィルやデータ行列を考える理由がここらにある（行列 $\underset{m\times n}{\mathbf{X}}=\left(x_{ij}\right)$ や $\underset{m\times n}{\mathbf{Q}}=\left(y_{ij}\right)$ のように設定することにより，上のような各関係が成り立つ）．

④絶対寄与度

絶対寄与度（あるいは**単に寄与度**：absolute contributions）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．

（ｉ）第 k 成分における選択肢 $i\left(\in I\right)$ の絶対寄与度

$$C_k\left(i\right)=\frac{p_{i+}\left(z_{ik}\right)^2}{\lambda_k}\quad\left(\begin{matrix}i\in I,k=1,2,\cdots,K\\K=\min\left\{m,n\right\}-1\end{matrix}\right),\ \sum_{i=1}^{m}C_k\left(i\right)=1 \tag{31}$$

（ⅱ）第 k 成分における選択肢 $j\left(\in J\right)$ の絶対寄与度

$$C_k\left(j\right)=\frac{p_{+j}\left(z_{jk}^*\right)^2}{\lambda_k}\quad\left(\begin{matrix}j\in J,k=1,2,\cdots,K\\K=\min\left\{m,n\right\}-1\end{matrix}\right),\ \sum_{j=1}^{n}C_k\left(j\right)=1 \tag{32}$$

⑤相対寄与度

相対寄与度（relative contributions）あるいは**平方相関**（squared correlations）とは，ある選択肢 $i\left(\in I\right)$ または選択肢 $j\left(\in J\right)$ が，どの成分に対してどの程度寄与しているかを知る指標である．たとえば，ある選択肢 $i\left(\in I\right)$ に注目し，その選択肢が各成分 $k\left(k=1,2,\cdots,K\right)$ のどれにどの程度寄与するかを知りたいときに用いる．

（ｉ）選択肢 $i\left(\in I\right)$ に対する相対寄与度

$$C_k^*\left(i\right)=\frac{z_{ik}^2}{\sum_{j=1}^{n}p_{+j}\left(\frac{p_{ij}-p_{i+}p_{+j}}{p_{i+}p_{+j}}\right)^2}\quad\left(\begin{matrix}i\in I,k=1,2,\cdots,K\\K=\min\left\{m,n\right\}-1\end{matrix}\right) \tag{33}$$

(ⅱ)**選択肢 $j(\in J)$ に対する相対寄与度**

$$C_k^*(j)=\frac{\left(z_{jk}^*\right)^2}{\sum_{i=1}^{m} p_{i+}\left(\dfrac{p_{ij}-p_{i+}p_{+j}}{p_{i+}p_{+j}}\right)^2}\quad\begin{pmatrix} j\in J, k=1,2,\cdots,K \\ K=\min\{m,n\}-1\end{pmatrix} \tag{34}$$

数式で表すとやや煩雑にみえるが，これを覚える必要はない．WordMiner はこれらの指標を成分スコアと共に算出するので，利用上はこれらの寄与度の情報の読み方・解釈を理解すればよい．これは後述の例題で実際に WordMiner が出力する情報を使って説明する．

⑥平方カイ二乗距離を用いること（加重付きの距離とすること）

(ⅰ)選択肢 i と l との間の平方カイ二乗距離（プロフィル間のカイ二乗距離）

$$\begin{aligned} d^2(i,l)&=\sum_{j=1}^{n}\frac{1}{p_{+j}}\left(\frac{p_{ij}}{p_{i+}}-\frac{p_{il}}{p_{l+}}\right)^2=\sum_{j=1}^{n}\left(\frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}-\frac{p_{il}}{p_{l+}\sqrt{p_{+j}}}\right)^2 \\ &=\sum_{j=1}^{n}\left(x_{ij}-x_{lj}\right)^2 \end{aligned} \tag{35}$$

(ⅱ)選択肢 j と t との間の平方カイ二乗距離（プロフィル間のカイ二乗距離）

$$\begin{aligned} d^2(j,t)&=\sum_{i=1}^{m}\frac{1}{p_{i+}}\left(\frac{p_{ij}}{p_{+j}}-\frac{p_{it}}{p_{+t}}\right)^2=\sum_{i=1}^{m}\left(\frac{p_{ij}}{p_{+j}\sqrt{p_{i+}}}-\frac{p_{it}}{p_{+t}\sqrt{p_{i+}}}\right)^2 \\ &=\sum_{i=1}^{m}\left(x_{ij}-x_{it}\right)^2 \end{aligned} \tag{36}$$

ここで選択肢間の**プロフィルの距離**を，いわゆるユーリッド距離を用いずに，上のように $1/p_{+j}$, $1/p_{i+}$ を加重とする**平方カイ二乗距離**（chi-square distance）として扱う（注：この加重がないとプロフィルの単純な**平方ユーリッド距**離となる）．実はこうしなければならない理由がある．$1/p_{+j}$, $1/p_{i+}$ の加重付きとすることで，すでに述べたピアソンのカイ二乗統計量との種々の関係を保持することが可能となる．もう一つの理由として**「分布の同等性」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) 等値プロフィル（比率パターンが同じ）となる行（あるいは列）の併合は，列（あるいは行）の距離に影響を与えない．

(2) 等値プロフィルの併合は（分析）結果に影響を与えない（あるいは結果が変わらない）．

(3) 換言すると，対応分析はこうした考え方が当てはまる（そう考えてもよい）データに対して有効な方法である．

この性質を簡単な例になぞらえると以下のようなことである．

例１：

たとえば，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，対応分析法では行和を 1 とした場合にいずれも同じ比率となるので，実質的には同じパターン（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．

例２：清涼飲料水の例

前出の清涼飲料水の好み選択の例を考える．このデータ表の中で（表 5)，たとえば次の 2 サンプルはそれぞれ回答パターンがまったく同じである（選んだ飲料水が同じ)．

21：{ｺｶｺｰﾗ,ﾍﾟﾌﾟｼｺｰﾗ} と 22：{ｺｶｺｰﾗ,ﾍﾟﾌﾟｼｺｰﾗ} は同じ
12：{ﾀﾞｲｴｯﾄｺｰｸ,Tab} と 27：{ﾀﾞｲｴｯﾄｺｰｸ,Tab} は同じ

このとき，この２サンプルの行を併合しても（つまり表中で行を併合し度数「１」を「２」と増やしても）結果は変わらない．列の側についても，同様のことが成り立つ．換言すると，このように比率のパターン（つまりプロフィル）の視点から考えていることが対応分析の特徴である（この点で主成分分析とは異なる)．

つまりクロス表の行（または列）のパターンが同じ数値，つまり行比率（または列比率）が同じ行（または列）は併合して加えても解析結果には影響しない（同等である)．

⑦成分スコアの（平方）ユークリッド距離とプロフィル間の（平方）カイ二乗距離の間の重要な性質

実はこの**両者は相互に等しい**という性質がある．このことから，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，元の二元データ表の行（あるいは列）のプロフィ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が知られている．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．WordMiner では，階層的分類法の１つであるウォード法と非階層的分類法の１つである k-平均法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．

（注）このことについては，別の資料「WordMiner におけるクラスター化法」に記したのでそれを参

照のこと．

5. データ表の基本的な組み合わせ

以上に述べたことを予備知識として，WordMiner で用いる機能を二元データ表の，表頭（列）と表側（行）に以下の表 27 のように対応させることで，各種のデータ表の解析を行うことが可能となる．また，双対性からデータ表の行と列を転置しても（項目 I と項目 J とを入れ替えても）解析結果は変わらない．この性質は，元のデータ表の作成時に覚えておくとよい．このように組み合わせを変えながら，どれが有意で意味があるかを“探査的”に調べる．

表 27　WordMiner におけるデータ表の関係

表側項目：I	表頭項目：J
● 構成要素変数 （分かち書き，キーワード）	● 回答（サンプル），個体
● 構成要素変数，キーワード変数 （分かち書き，キーワード）	● 質的変数 （選択肢型設問・属性項目等）
● 構成要素変数 （分かち書き，キーワード）	● 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

[WordMiner における処理操作]

表 27 に要約したデータ表が，WordMiner で実際にどのように出力されるかを例で示そう．WordMiner では「多次元データ解析」のモジュールの中で，2 種のデータ表「(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 27 に挙げた組み合わせと分析目的に応じて指定する．

たとえば，基本的な構成は図 8 のように考えればよい．なお，図 8 の中で，「$w_j^{(i)}$」は「第 i 番サンプルの回答の自由記述データの分かち書きで得た**構成要素**（つまり単語，語句，キーワードなど，分析で扱う単位）の系列を表す．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)」のデータ表が基本となる．また，今ま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．

		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・属性」	1 2 ⋮ i ⋮ n	$w_1^{(1)}$ $w_2^{(1)}$ … $w_j^{(1)}$ … $w_k^{(1)}$ … … $w_1^{(2)}$ $w_2^{(2)}$ … $w_j^{(2)}$ … $w_k^{(2)}$ … ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ $w_1^{(i)}$ $w_2^{(i)}$ … $w_j^{(i)}$ … … $w_l^{(i)}$ … ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ ⋮ $w_1^{(n)}$ $w_2^{(n)}$ … $w_j^{(n)}$ …

図 8 （回答者）×（構成要素），（質的変数）×（構成要素）のイメージ図

① 「（回答・サンプル）×（構成要素）」のデータ表の例

簡単な例として表 28 を挙げる．これはあるウェブ調査で得た「（回答・サンプル）×（多変量項目）」のデータ表の一部である．引用した質問は次に挙げるように「回答者のインターネットとのかかわり」を問う 2 つの質問の一つ「3－2.あなたご自身にとって「インターネット」は、どのようなことがらに活用できるか」を用いた．

まず，元の自由回答データ（回答原文）から WordMiner の分かち書き処理機能で「分かち書き」「キーワード」を抽出する．次に，このうちのキーワードを選び「（サンプル）×（構成要素，ここではキーワード）」のデータ表とした．WordMiner では，このデータ表から回答パターンのクロス表を生成し（表 29），それを解析対象のデータ表とする．これの生成の仕組みはすでに述べた通りである．

問３．次に、あなたと「インターネット」とのかかわりについてお伺いします。

3－1.　あなたご自身にとって「インターネット」は、どのようなことがらに活用できると思いますか。どんなことでも結構ですので、以下になるべく具体的にご記入ください。
3－2.　では、一般的に「インターネット」は、どのようなことがらに活用できる と思いますか。なるべく、他にはないような活用法を、どんなことでも結構ですので、以下になるべく具体的にご記入ください。

表 28 （サンプル）×（構成要素）の例

サンプル	構成要素（キーワードを用いたとき）
1	為，しらべ，利用，家族，遊園地，公園，食べ物屋，情報収集，調査
2	あまり，セキュリティ，必要，ミーティング，世間話，仕事
3	新製品，スペック，価格，お店
4	役所，証明書発行，受け取り
5	旅行，計画，観光地，チェック，お店，情報収集
6	情報収集，調査，メール，座席予約，航空機，列車，オークション
7	地図検索，鉄道，乗り換え，検索，その他，時々，必要，情報検索
8	通信販売，申し込み，旅行，情報収集
9	情報ツール
10	あまり，ふつう，店舗，販売，商品，販売店，ショッピング，建築図面作成用，ＣＡＤデータ，ダウンロード
11	自分，興味，事柄，容易，公式，専門家，情報
12	日常生活，中，帰省時，飛行機，時刻表，育児，経験談，アドバイス，仕事，必要，情報，特定人物，活動，著書
13	情報収集
14	電話，手紙，かわり
15	仕事上，事，出張，際，ホテル，情報，等
16	パソコン，周辺機器，仕様，価格，懸賞，応募，ドライバ，ダウンロード，ゲーム
17	掲示板，一つ，場所，みんな，話
18	調べ物，ショッピング，オークション
19	情報，収集，自己，ＰＲ
20	ニュース，天気，行楽情報，仕事，情報
21	映画，書籍，情報入手，求人検索，単語，等，検索，メール
22	専門的，事柄，情報収集
23	メール，一番，仕事，不明瞭，確認，美術館，博物館，映画，その他，催し物，情報収集，たまに，オークション，お食事，電車，時刻表，経路
24	調べ物，ホームページ，サイト
25	友人，知人，連絡
26	百科事典
27	趣味，人，交流，勉強，場所，交通機関，時間
28	天気予報，道路状況，宿泊情報，等，行楽，情報収集，辞書，新聞
29	自分，知識，情報，時間，辞書，新聞，地図，最近，ネット，使用，事
	＜以下，省略＞

表 29 （サンプル）×（構成要素）のデータ表(一部を切り出し)

サンプルID	SEQ	行和	ＨＰ	いろいろ	いろいろな	いろんな	お店	その他	ときに	やり	やりとり	アーティスト	イベント	インターネット	オークション	オンラインショッピング	ゲーム
	列和	3378	18	6	18	10	19	7	9	17	24	7	7	20	28	7	9
36	[00000042]	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
778	[00000846]	24	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
716	[00000773]	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
34	[00000040]	17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
558	[00000602]	15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
50	[00000058]	14	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
509	[00000548]	14	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0
759	[00000824]	14	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
98	[00000107]	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
139	[00000154]	13	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0
310	[00000338]	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
401	[00000432]	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
407	[00000438]	13	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
370	[00000400]	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
502	[00000540]	12	0	0	1	0	0	0	1	0	0	0	0	0	0	1	0
515	[00000554]	12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
639	[00000688]	12	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
51	[00000059]	11	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
52	[00000060]	11	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
89	[00000098]	11	1	0	0	0	0	0	0	1	1	0	0	0	0	0	0
157	[00000174]	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
303	[00000330]	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
484	[00000520]	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
564	[00000608]	11	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
676	[00000728]	11	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
801	[00000873]	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
27	[00000030]	10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

（注）この種のデータ表は構成要素数（分かち書き，キーワード）が多くなり，サンプル数もそれなりの数となるので，個々のセル（要素）の頻度がきわめて疎となるのが特徴，よってこれを見ただけでは傾向がみえない．

表 30　複数の項目を含むデータ表の例

サンプル	性別	年齢区分	性年齢区分	未既婚	職業	構成要素（ここではキーワード）
1	男性	4_35才～39才	男性/4_35才～39才	既婚	営業職	為 しらべ 利用 家族 遊園地 公園 食べ物屋 情報収集 調査
2	男性	5_40才～44才	男性/5_40才～44才	既婚	研究開発職	あまり セキュリティ 必要 ミーティング 世間話 仕事
3	女性	5_40才～44才	女性/5_40才～44才	既婚	主婦専業	新製品 スペック 価格 お店
4	男性	5_40才～44才	男性/5_40才～44才	既婚	労務職	役所 証明書発行 受け取り
5	女性	2_25才～29才	女性/2_25才～29才	既婚	主婦専業	旅行 計画 観光地 チェック お店 情報収集
6	男性	5_40才～44才	男性/5_40才～44才	既婚	研究開発職	情報収集 調査 メール 座席予約 航空機 列車 オークション
7	男性	4_35才～39才	男性/4_35才～39才	既婚	無職・その他	地図検索 鉄道 乗り換え 検索 その他 時々 必要 情報検索
8	女性	2_25才～29才	女性/2_25才～29才	既婚	主婦専業	通信販売 申し込み 旅行 情報収集
9	男性	3_30才～34才	男性/3_30才～34才	既婚	自営業とその家族	情報ツール
10	男性	6_45才～49才	男性/6_45才～49才	既婚	専門職	あまり ふつう 店舗 販売 商品 販売店 ショッピング 建築図面作成用 ＣＡＤデータ ダウンロード
11	男性	3_30才～34才	男性/3_30才～34才	未婚	無職・その他	自分 興味 事柄 容易 公式 専門家 情報
12	女性	3_30才～34才	女性/3_30才～34才	既婚	専門職	日常生活 中 帰省時 飛行機 時刻表 育児 経験談 アドバイス 仕事 必要 情報 特定人物 活動 著書
13	男性	9_60才～64才	男性/9_60才～64才	既婚	無職・その他	情報収集
14	男性	8_55才～59才	男性/8_55才～59才	既婚	管理職	電話 手紙 かわり
15	男性	7_50才～54才	男性/7_50才～54才	既婚	販売・保安・ｻｰﾋﾞｽ	仕事上 事 出張 際 ホテル 情報 等
16	男性	5_40才～44才	男性/5_40才～44才	既婚	営業職	パソコン 周辺機器 仕様 価格 懸賞 応募 ドライバ ダウンロード ゲーム
17	男性	5_40才～44才	男性/5_40才～44才	既婚	技能職	掲示板 一つ 場所 みんな 話
18	女性	3_30才～34才	女性/3_30才～34才	既婚	ﾊﾟｰﾄ・ｱﾙﾊﾞｲﾄ	調べ物 ショッピング オークション
19	女性	1_25才未満	女性/1_25才未満	未婚	自由業	情報 収集 自己 ＰＲ
20	女性	3_30才～34才	女性/3_30才～34才	既婚	技術職	ニュース 天気 行楽情報 仕事 情報

表 31 （年齢区分）×（構成要素）のデータ表の例（一部を切り取り）

通番		行和	1_25才未満	2_25才～29才	3_30才～34才	4_35才～39才	5_40才～44才	6_45才～49才	7_50才～54才	8_55才～59才	9_60才～64才
	列和	3378	438	510	570	517	514	260	264	104	121
117	情報	270	39	42	41	36	45	26	19	7	8
121	情報収集	130	11	19	21	27	20	14	10	2	5
109	趣味	99	15	14	19	15	17	6	7	1	3
33	メール	95	12	13	11	19	17	4	10	5	4
66	検索	79	12	9	14	11	11	5	9	2	5
84	仕事	74	8	5	14	14	11	9	8	3	1
162	友人	60	7	6	9	12	9	9	4	1	2
145	等	58	6	11	10	8	5	6	5	1	4
149	入手	58	7	8	4	10	12	4	6	2	3
166	旅行	56	1	8	9	10	9	3	7	2	2
55	活用	55	11	8	11	9	7	3	3	0	1
91	事	54	11	10	16	5	1	5	2	1	2
99	自分	49	9	6	15	3	6	3	5	1	0
18	ショッピング	48	3	7	12	8	3	7	3	3	1
150	買い物	46	3	11	9	9	7	2	2	0	1
170	連絡	46	4	5	6	6	9	5	3	3	3
24	ニュース	43	7	10	3	4	8	3	3	0	4
94	時	43	2	5	9	10	6	6	3	1	0
164	予約	42	2	8	6	7	6	7	1	0	5
135	調べ物	40	8	0	13	4	10	2	2	0	1
110	収集	36	3	6	4	6	6	8	2	0	0
31	ホームページ	34	6	6	5	6	3	1	6	0	1
128	人	34	11	10	6	4	2	0	1	0	0
93	事柄	33	6	3	7	4	4	2	2	4	1
165	利用	33	1	4	7	4	8	5	1	2	1
16	コミュニケーション	29	7	4	5	8	3	1	1	0	0
76	購入	29	3	5	2	4	5	4	3	0	3
13	オークション	28	3	5	5	6	5	0	2	0	2
113	商品	28	3	5	5	4	5	1	2	0	2
153	必要	28	3	4	3	5	5	5	1	1	1

（注）「情報」「情報収集」「メール」「検索」といったキーワードの頻度が多い，とく若年層のその傾向があることがこの表だけでも見えるだろう．

②「（構成要素）×（質的変数）」のデータ表の例

次に，表 30 のような例をみよう．前と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したものである．表側を「サンプル」とし，表頭に属性「性別」「年齢区分」「性年齢区分」「未既婚」，そして質問 3−2. の分かち書き処理で得た「キーワード」を項目としてある．

この表で，「年齢区分」と「構成要素（キーワード）」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 31 のようなクロス表が得られ，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．こうすると出現キーワード（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）．

6. 数値例による説明

6．1　データ表の準備

ここで人工的なデータ表（表 32）を用いて，WordMiner で得られる諸統計量，情報の解読を試みる．データ表を意図的に構造化することで対応分析がどう機能するかを見易くしておく．データ表は表 32 のような構成であり，以下のような場面を想定してみた．

表 32　質的データとして表現したデータ表

サンプル	銘柄	次年度調査の銘柄	一番好きな銘柄	性別	年齢区分
サンプル 1	銘柄 B, 銘柄 E, 銘柄 F	●銘柄 E, ●銘柄 F	◆銘柄 B	▼男性	★30 代
サンプル 2	銘柄 F	●銘柄 F, ●銘柄 B	◆銘柄 F	▼男性	★40 代
サンプル 3	銘柄 C, 銘柄 F	●銘柄 F	◆銘柄 C	▼男性	★30 代
サンプル 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
サンプル 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
サンプル 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
サンプル 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
サンプル 8	銘柄 C, 銘柄 F	●銘柄 C, ●銘柄 F	◆銘柄 F	▼男性	★40 代
サンプル 9	銘柄 A, 銘柄 B, 銘柄 E	●銘柄 B, ●銘柄 E	◆銘柄 E	▼女性	★30 代
サンプル 10	銘柄 A, 銘柄 D, 銘柄 E	●銘柄 A, ●銘柄 E	◆銘柄 D	▼女性	★30 代

① 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．
② 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．
③ 「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（後の布置図の観察の識別用のため）．
④ 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）．

WordMiner 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（扱える文字数の制限もない）．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．

実際にこのデータ表を WordMiner にインポートし分析を進める．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」がそれぞれ構成要素変数として生成される．［注：「分かち書き」と「キーワード」では，生成する構成要素の内容が異なるので注意．マニュアルを参照］

次に（回答・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，次の表 33 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，サンプル数=10（名），銘柄＝6（A～F までの 6 選択肢）となる．

表 33 （サンプル）×（銘柄）のクロス表

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

6．2 固有値と寄与率，累積寄与率

まず始めに固有値，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 34）．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 34 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

6．3 成分スコアの観察

WordMiner では，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て出力する．ここでは，このうちの成分スコア，寄与度を表 36 として挙げた．いまここで，この表にある第 1 固有値に対する第 1 成分スコアに注目する．行（サンプル）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（ソートして）みると表 35 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，固有値（あるいは特異値）と成分スコアの関係をこの表と関連づけて考える．対応分析で現れる固有値がどのような意味を持つのかを調べる．これは次のような図を用意すると理解しやすい．表 35 を作るときに用いたデータ表の表頭項目「銘柄」と表側項目「サンプル」の各選択肢に対して求めた第 1 固有値（λ_1）に対する第 1 成分スコア（表の「成分スコア 1」）に再び注目する．この両成分スコアを「横軸に銘柄，縦軸にサンプル」をとって散布図とする．図 9 の左の表が「銘柄」と「サンプル」への成分スコアであり，これを元に描いた散布図が図 9 の右側の図である．この図の意味・解釈は重要である．もとのデータ表では，表側のサンプル，表頭の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「サンプル」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ってこのままでは数量として扱えない．しかし対応分析法を用いて新たな数量を付与したことで量的データとして扱えるようになる．これが**“数量化”**と言われる所以である（数量化法 III 類の考え方）．同時にこの方法が質的データの線形化となっていることも分かる（表 35，図 9）．明らかに図 9 は表 35 に対応するものである．表 35 では行と列の入れ替えを行っただけであるが，この散布図のように「銘柄」と「サンプル」という名目的な名義尺度データの選択肢に対して付与されたある数量（成分スコア）を観察すると，もとの名目的コード（選択肢）ではない「新たな（点の間の分布，距離関係が）意味のある」別の数量空間が作られたことになる．この数量＝成分スコアは大小が意味を持ち区間尺度データとして加減乗除も可能な数値として扱える．たとえば，この成分スコアに対してクラスター化を行うことが可能となる（実際にそうしている）．ここに数量化を行う重要な意味があり，質的データであるテキスト型データを対応分析法で解析する目的がここにある．いきなりこの種の二元データ表に主成分分析を適用してはならないことも示唆している．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，それを越えた情報の取得には別の視点からのアプローチが必要である．

表 35　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	「サンプル」の第 1 成分スコア
サンプル 2	0	0	0	0	0	1	1	1.2969
サンプル 3	0	0	0	0	1	1	2	1.1538
サンプル 8	0	0	0	0	1	1	2	1.1538
サンプル 5	0	0	0	1	1	1	3	0.7206
サンプル 4	0	0	1	1	1	1	4	0.3785
サンプル 1	0	0	1	1	0	1	3	0.1678
サンプル 6	0	1	1	1	1	0	4	-0.2432
サンプル 9	0	1	1	1	0	0	3	-0.6611
サンプル 7	1	1	1	1	0	0	4	-0.9102
サンプル 10	1	1	1	0	0	0	3	-1.1650
列和	2	4	6	6	5	6	29	
「銘柄」の第 1 成分スコア	-1.3113	-0.9414	-0.5125	-0.1153	0.7997	1.0261		

表 36　成分スコアと寄与度の一覧

①選択肢$i(\in I)$，つまり「サンプル」への成分スコア，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

②選択肢$j(\in J)$，つまり「銘柄」への成分スコア，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

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

⇒

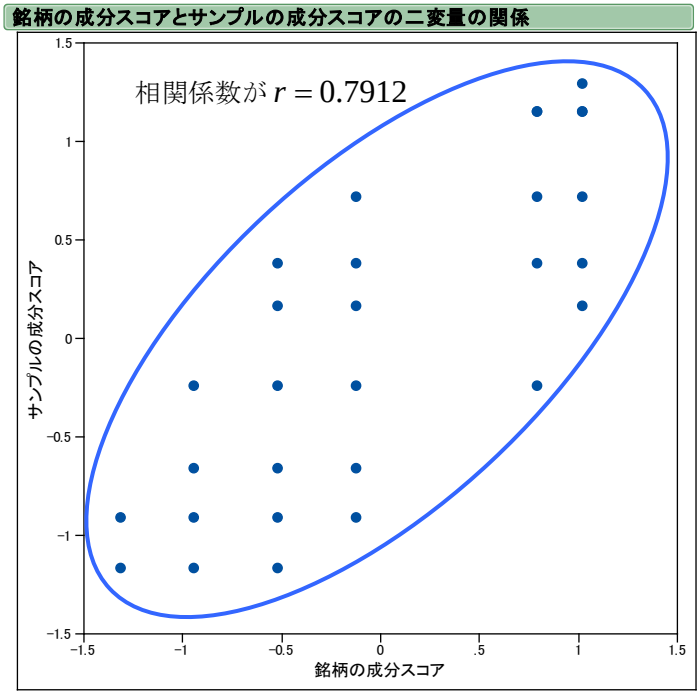


図 9　第 1 成分スコアのサンプルと銘柄の散布図

（注）この表では，サンプル番号を識別コードとして入れてみたが，もちろん銘柄の選択肢も各数値に対応する．右側の図が表 36 の（0，1）型データ表の行（縦軸⇒サンプル）と列（横軸⇒銘柄）にそれぞれ対応していることに注意しよう（同じ数量があるので，その点の重なりはあるので注意）．

実は，ここで得た第 1 固有値 $\lambda_1 = 0.6260$ の（正の）平方根は（特異値に相当），上の図 9 の散布図の（ピアソンの）相関係数に等しい．実際に散布図（つまり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，$\sqrt{\lambda_1} \doteqdot 0.7912$ であり両者は一致する．つまり，ここでいう数量化とは，「銘柄」と「サンプル」という名目的な特性（項目であり選択肢）に対して“**ある（量的な）数量を付与できたとして**”これを求めたところ，表のような数値（成分スコア）が得られ，実際にその相関（線形の意味での関係）をかなり大きい値（数量）とできたということである．

実は，林の数量化法 III 類では，「このある数量がサンプルと銘柄（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．一方ここでは，対応分析法のアプローチに従い二元データ表（例：クロス表）から出発したとして定式化し，同じ結論を得たということである．つまり両者の考え方（定式化），理念･思想は異なるが，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（もちろん，数理的にも対応分析法と数量化法 III 類は同等であることは示される）．

6．4　布置図と同時布置図の観察

次に布置図，同時布置図をみよう．WordMiner では布置図，同時布置図ともにインタラクティブに画面上で成分を指摘することにより観察できる．これは式（21），（22），（23）に従って，サンプル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，式（23）で$k=1,\ k'=2$とすれば第 1 成分スコアと第 2 成分スコアとの布置図となる．WordMiner では，デフォルトとしてこの（1, 2）軸（成分）の図を始めに表示するので，必要に応じて軸を選べば自在に成分の組み合わせに合わせた布置図が得られる．

① サンプルの成分スコアの布置図と銘柄の成分スコアの布置図

これを WordMiner では次の図のように出力する．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．

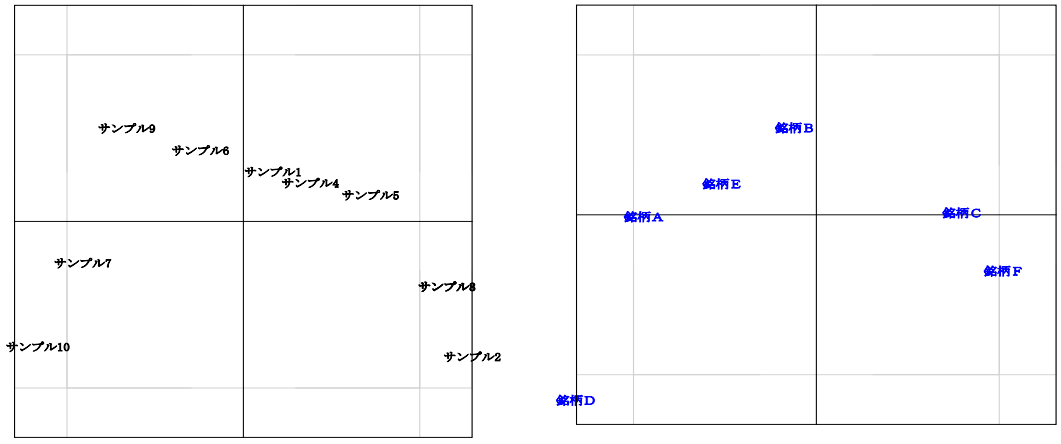


図 10　成分スコアの布置図

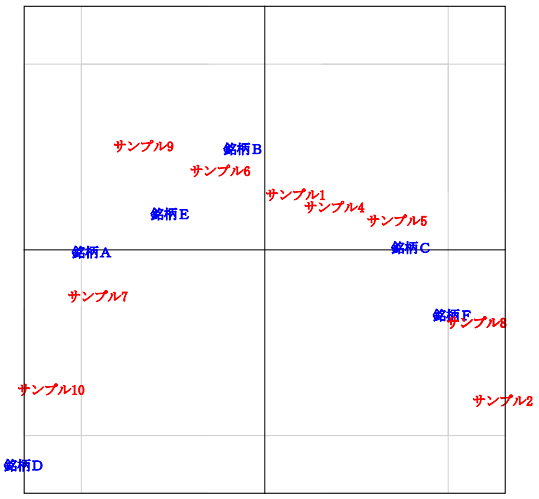


図 11　同時布置図

② 同時布置図

同時布置図を作ると図 11 が得られる．ここで確かに上の図 10 の 2 つの布置図を重ねた形となり，サンプルと，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．

とくにこの図のような放物線のような形状は“**馬蹄形効果**”といい，実はデータ表の構造がかなり線形的であることを意味している（表 35 の並べ替えデータと図 9 を参照）．

６．５　絶対寄与度と相対寄与度

これの数理的な説明はそう難しいものではないが，若干の予備知識を必要とするので，専門書に譲り（ベンゼクリ[14, 15]，大隅他[9]など），ここでは WordMiner が出力した数値の読み方・解釈を試みる．WordMiner では，寄与度（絶対寄与度，相対寄与度）を，成分スコアと併せて表 36 のような形で一覧とする（この他の情報もあるがここでは省略する）．

[絶対寄与度と相対寄与度の読み方]

前述のように絶対寄与度は「ある成分に占めるある選択肢の占める割合（寄与）」を知る指標である．具体的に例でみる．

たとえば，「サンプル」の側に注目し表 36 の①（上の表）の**絶対寄与度**を見る．ここで「絶対寄与度１」～「絶対寄与度５」が，第１成分から第５成分に対応する絶対寄与度を示す．またここでは，列の方向（縦方向）の和が「100」となるように表記してある（式（31）の絶対寄与度参照）．つまり，ある成分に注目したとき，その成分内での各サンプルの寄与の大きさを示している．

たとえば，成分１の列方向に絶対寄与度１をみると，サンプル７，１０，８，３，７等の値が他に比べて大きい（サンプル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，サンプル９，１０，それにつづき２，６などの値がやや大きい．サンプル１，２はむしろ第３成分内での寄与が高い（とくに，サンプル２に注意，図では第１軸の右端に位置するがサンプル８ほど第１成分内での寄与は高くないということ）．

次に**相対寄与度**をみる．こちらは，ある選択肢がどの成分内で寄与するかを知る指標である．ここでは横方向（行方向）の和が「１」となるような指標となっている．サンプル１，６を除いて総じて値が大きい．とくに，サンプル１は相対寄与度３（成分３に対応）で値が大きく，またサンプル６は相対寄与度２（成分２）で値がやや大きい．つまり，サンプル６は第１成分よりはこれらの成分への寄与が高いということである．相対寄与度１で値が大きい，つまりここでは第１成分で大方の説明が付くということであり，図を観察するとどの位置にあるかとその意味が分かるであろう．

「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，続いて銘柄 A，C，D などがあるがこれらの値はかなり小さい．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．これと固有値の大きさとその寄与率を考慮すると，始めの２成分内で寄与の高い銘柄に注目すべき，という構造が見えてくる．

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成分への寄与が大きい（銘柄 B は第２成分に寄与する，図からも明らか）．

ここで留意すべきことは，散布図の情報は“そこでいま眺めている成分軸の組み合わせの中での射影図”であり多次元データとしての**全情報ではない**，ということである．よって，絶対寄与度，相対寄与度も図と併せて観察し，どの成分でどう寄与するかを多次元的に知る

ことが必要である．

なお一般には出発時のデータ表の各セル内の頻度が，かなり疎であるとか，とくに（サンプル）×（構成要素変数）の場合，出現度数が偏ったデータ表であるとき（はずれ値があるようなとき）には，寄与度をよく観察して，どれがデータ構造をゆがめる原因となっているかを“探索的に”チェックすることが必要である（つまり，それなりの手間がかかる）．ここでみたような簡便な例は実用場面ではあまり登場しないということである．

６．６ 追加処理あるいは追加要素

次のような場面で，いわゆる**追加処理**（supplementary treatment）を行うと効果的な場合がある．追加処理の対象とする項目（変量），選択肢を**追加処理要素**（supplementary elements）という．追加処理の数理的な説明はやや面倒なのでここでは省略するが，次のような場面での適用が考えられる．関心のある読者は参考文献をみていただきたい（たとえば，ベンゼクリ[14, 15]，Lebart[24]，大隅[9]）．

①はずれ値の一時除去と再配置を行いたいとき

(1)　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．

(2)　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．

②判別分析的，グループ間類似や差異を見る

(1)　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．

(2)　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）．

(3)　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）．

③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

(1)　（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

(2)　（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．

[追加処理の例]

ここで使った表 33 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．

① 表 33 で，（サンプル）×（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．

② 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．

③ このとき，もとの（サンプル）×（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．

④ さらに，属性情報も取得してあるので（表 32 の性別，年齢区分），これが（サンプル）×（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．これを別の課題とする．

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），さらに別の項目の影響度を知りたいといった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）のプロフィルの影響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のである（参考：ジャックナイフ法，ブートストラップ的なアプローチに類似する）．

したがって機能的には，元のデータ表において「行のプロフィル」「列のプロフィル」の両者の追加処理がありうるが，WordMiner では（原則として）元のデータ表の列の側への追加処理機能を，「質的変数の追加」「構成要素変数の追加」として処理するようになっている．

さて，ここでみる例題を再度整理しよう．

① 表 33 のデータ表について「(サンプル) ×（銘柄)」についての分析結果はすでに得た．

② なお WordMiner では，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．

③ たとえば「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．

④ このとき，表 33 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，（サンプル）×（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（サンプル）×（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．WordMiner ではこの図だけでなく，それぞれ個別の布置図を観察することも可能である．実際に計算した結果は表 37 のような成分スコアと相対寄与度が得られる（なお，絶対寄与度は式の定義上算出できない，[22]，[24]）．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．

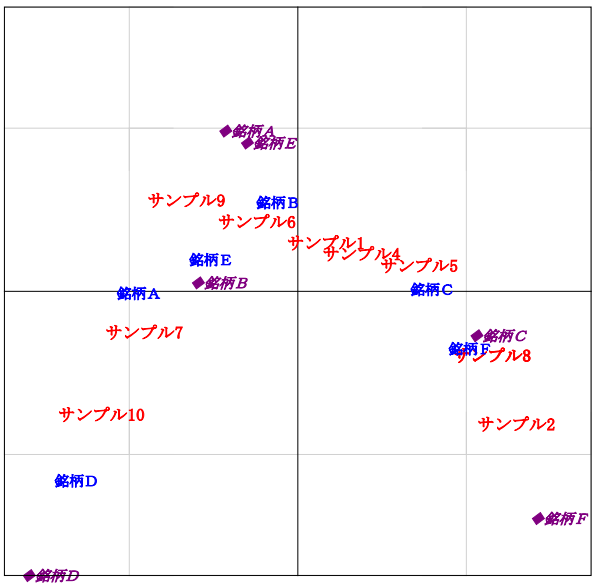
ここで得た追加処理による成分スコアと，元の成分スコア（の同時布置図）を重ねて描くと次の図 12 が得られる．

表 37 「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 11 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 13 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，ある対応関係があることが見えるであろう．



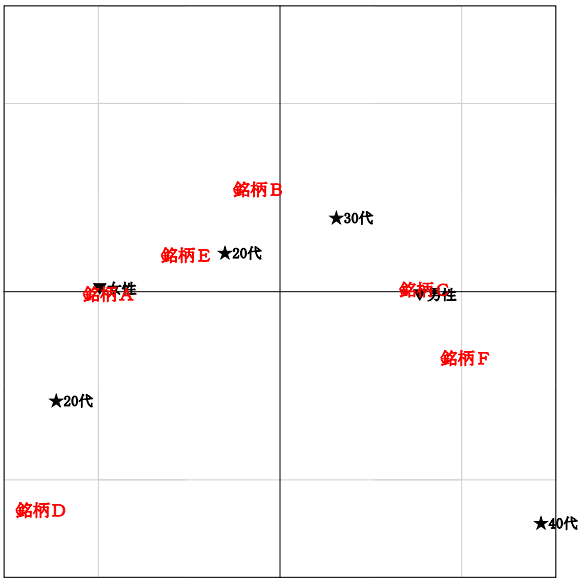


図 12 「一番好きな銘柄」の追加処理で得た布置図　図 13 「性別」「年齢区分」の追加処理で得た布置

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出立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）．

これは出発時の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上の仮説設定が必要であり，また自分の分析目的に合ったデータ収集方式とデータ表の作成方法とを事前に考えて取り組むことが肝要ということに他ならない．

7. むすび

この稿では，主に対応分析法の基礎知識について述べた．WordMiner の「**多次元データ解析の機能**」には，この対応分析法の基本機能の他にさまざまな実用的な機能が含まれている．たとえば以下のような機能もある．

① 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［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］．
② 回答・サンプルのクラスター化，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．
③ 独自のクラスター化処理の方法を用いていること（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）．

いわゆる多次元データ解析手法を組み入れただけでなく，それをうまく活用した多くの機能があることが WordMiner の特徴である．これらについては，今後も本稿と同様になるべく事例や人工データを交えた紹介説明を行う予定である．

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，WordMiner 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 WordMiner 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

次のようなサンプル数が 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：アイテム・カテゴリー型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）の二元クロス表を確認せよ．

表１　データ表

項目 / 個体	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

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，サンプル数が 20，3 項目の質問（$\mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3$）からなるデータ表がある．これについて，次の問に答えよ．

問１：このデータ表から**アイテム・カテゴリー型**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

サンプル	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　アイテム･カテゴリー型データ表

項目と選択肢 / サンプル	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

アイテム・カテゴリー型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 アイテム･カテゴリー型データ表と多重クロス表の関係

項目と選択肢 / サンプル	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

アイテム・カテゴリー型データ表

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

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，アイテム・カテゴリー型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 の上部にあるアイテム・カテゴリー型データ表を（サンプル側の）追加処理要素として扱うとサンプル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サンプル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，アイテム・カテゴリー型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が（誤解があるようだが），対応分析的なアプローチから考えればより一般的に大量サンプルの場合の処理が十分に可能である．ここはフランス流のアプローチを採用して，上のように多重クロス表とアイテム・カテゴリー型データ表を用いた処理の方が効率的である．

[補足]

いま，演習問題 2 を例として，各データ表に対応分析法を適用したときに得られる各データ表の関係を整理しておこう．どのようなデータ表から出発したときに，何が得られたかを知っておくことが大切と考えるからである．

1. データ表の関係，その固有値

まず元となる「(回答・サンプル) × (多変量・多数項目)」からなるデータ表から得られるアイテム・カテゴリー型データ表（インジケータ行列），そしてアイテム・カテゴリー型データ表から得られる多重クロス表（バート表）をそれぞれ以下のように表す．

① サンプル数が N，項目数（質問数）が M 個の「(回答・サンプル) × (多変量・多数項目)」型データ表をアイテム･カテゴリー型データに変換して得られるデータ表を以下の式で表す．上の例で言えば，N=20，M=3 のデータ表 3 を，表 4 のアイテム・カテゴリー型に変換するということに相当する．

ここで，$\underset{N\times n_j}{\mathbf{A}_j}$ は寸法が N 行，列数（つまり 0, 1 に展開後の延べカテゴリー数）が n_j 個の項目 j のインジケータ行列である（各行の選択肢のどれかに 1 が 1 個のみある）．表 2 の項目 3 であれば $\mathbf{A}_3$ が寸法が N=20（サンプル），（項目）$n_j=3$ のインジケータ行列となっている．

従って，これを一般に M 個（M 項目）の分割行列で表すと，次の式のようになる．

$$\underset{N\times n^*}{\mathbf{A}}=\left[\underset{N\times n_1}{\mathbf{A}_1},\underset{N\times n_2}{\mathbf{A}_2},\cdots,\underset{N\times n_i}{\mathbf{A}_i},\cdots,\underset{N\times n_j}{\mathbf{A}_j},\cdots,\underset{N\times n_M}{\mathbf{A}_M}\right]$$

$$\left(\begin{array}{l}\text{ここで，}\ n^*=\sum_{j=1}^{M}n_j\\ K^*=\sum_{j=1}^{M}\left(n_j-1\right)=n^*-M\end{array}\right)$$

② 次に，このデータ行列 $\underset{N\times n^*}{\mathbf{A}}$ と，これを転置した $\underset{n^*\times N}{\mathbf{A}^t}$ との積の行列を作ると，次のいわゆる多重クロス表（バート表） $\underset{n^*\times n^*}{\mathbf{B}}$ が得られる．この行列は当然対称行列となり，その寸法は $n^*\times n^*$ となる．

$$\underset{n^*\times n^*}{\mathbf{B}}=\underset{n^*\times N}{\mathbf{A}^t}\ \underset{N\times n^*}{\mathbf{A}}=\begin{pmatrix}\mathbf{A}_1^t\mathbf{A}_1 & \mathbf{A}_1^t\mathbf{A}_2 & \cdots & \mathbf{A}_1^t\mathbf{A}_i & \cdots & \mathbf{A}_1^t\mathbf{A}_j & \cdots & \mathbf{A}_1^t\mathbf{A}_M\\ \mathbf{A}_2^t\mathbf{A}_1 & \mathbf{A}_2^t\mathbf{A}_2 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots\\ \vdots & \vdots & \mathbf{A}_1^t\mathbf{A}_1 & \vdots & \vdots & \vdots & \vdots & \vdots\\ \mathbf{A}_i^t\mathbf{A}_1 & \cdots & \cdots & \mathbf{A}_i^t\mathbf{A}_i & \cdots & \mathbf{A}_i^t\mathbf{A}_j & \cdots & \mathbf{A}_i^t\mathbf{A}_M\\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots\\ \mathbf{A}_j^t\mathbf{A}_1 & \cdots & \cdots & \mathbf{A}_j^t\mathbf{A}_i & \cdots & \mathbf{A}_j^t\mathbf{A}_j & \cdots & \cdots\\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots\\ \mathbf{A}_M^t\mathbf{A}_1 & \cdots & \cdots & \mathbf{A}_M^t\mathbf{A}_1 & \cdots & \cdots & \cdots & \mathbf{A}_M^t\mathbf{A}_M\end{pmatrix}$$

例を挙げれば表 6 がこれに相当する（ここでは，$n^*=n_1+n_2+n_3=4+2+3=9$ となっている）．表 6 の多重クロス表は，上の表記によると以下のようになる．

$$\underset{9\times 9}{\mathbf{B}}=\underset{9\times 20}{\mathbf{A}^t}\ \underset{20\times 9}{\mathbf{A}}=\begin{pmatrix}\mathbf{A}_1^t\mathbf{A}_1 & \mathbf{A}_1^t\mathbf{A}_2 & \mathbf{A}_1^t\mathbf{A}_3\\ \mathbf{A}_2^t\mathbf{A}_1 & \mathbf{A}_2^t\mathbf{A}_2 & \mathbf{A}_2^t\mathbf{A}_3\\ \mathbf{A}_3^t\mathbf{A}_1 & \mathbf{A}_3^t\mathbf{A}_1 & \mathbf{A}_3^t\mathbf{A}_3\end{pmatrix}$$

ここのブロック行列は表 9 のそれに対応している．たとえば，対角ブロック行列の $\mathbf{A}_1^t\mathbf{A}_1$ は，

$$\mathbf{A}_1^t\mathbf{A}_1=\begin{pmatrix}3 & 0 & 0 & 0\\ 0 & 5 & 0 & 0\\ 0 & 0 & 8 & 0\\ 0 & 0 & 0 & 4\end{pmatrix}$$

となり（項目 $\mathbf{A}_1$ の周辺度数を対角要素とする対角行列），ブロック行列 $\mathbf{A}_2^t\mathbf{A}_3$ は項目 $\mathbf{A}_2,\mathbf{A}_3$ のクロス表に相当する．その対称の位置に，このクロス表の行と列を転置した $\mathbf{A}_3^t\mathbf{A}_2$ が位置する．

たとえば表 9 と付き合わせると，$\mathbf{A}_2^t\mathbf{A}_3$ および $\mathbf{A}_3^t\mathbf{A}_2$ は次のようになる（表 9 も参照）.

$$\mathbf{A}_2^t\mathbf{A}_3=\begin{pmatrix}1&4&4\\4&5&2\end{pmatrix},\ \mathbf{A}_3^t\mathbf{A}_2=\begin{pmatrix}1&4\\4&5\\4&2\end{pmatrix}$$

このような関係は他の各ブロック行列についても同様である.

③ 以上のように各データ表を用意すると，それぞれに対して対応分析法を適用して得られる結果にはある規則性がある．ここでは固有値についてのみ関係を一覧とするが，当然固有ベクトルについてもある規則性がある.

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 $\underset{n_i\times n_j}{\mathbf{E}}=\mathbf{A}_i^t\mathbf{A}_j$ または $\underset{n_j\times n_i}{\mathbf{E}^*}=\mathbf{A}_j^t\mathbf{A}_i$	$n_i\times n_j$	λ_k^E
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 $\left(n^*=n_i+n_j\right)$	$N\times n^*\ \left(n^*=n_i+n_j\right)$	$\lambda_k^A=\dfrac{1\pm\sqrt{\lambda_k^E}}{2}$
タイプ３	２項目の多重クロス表（バート表） $\underset{n^*\times n^*}{\mathbf{B}}=\underset{n^*\times N}{\mathbf{A}^t}\ \underset{N\times n^*}{\mathbf{A}}\quad\left(n^*=n_i+n_j\right)$	$n^*\times n^*\ \left(n^*=n_i+n_j\right)$	$\lambda_k^B=\left(\lambda_k^A\right)^2=\left(\dfrac{1\pm\sqrt{\lambda_k^E}}{2}\right)^2$
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 $n^*=\sum_{j=1}^{M}n_j$	$N\times n^*\ \left(n^*=\sum_{j=1}^{M}n_j\right)$	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 $n^*=\sum_{j=1}^{M}n_j$	$n^*\times n^*\ \left(n^*=\sum_{j=1}^{M}n_j\right)$	λ_k^B $\lambda_k^B=\left(\lambda_k^A\right)^2$

この要約表の意味するところは，クロス表，アイテム・カテゴリー型データ表，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，とくにアイテム・カテゴリー型データ表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．なお，WordMiner では，すでに今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アイテム・カテゴリー型とクロス表データに対応できる．

2. 固有値に関する重要な性質

アイテム・カテゴリー型データ表の固有値（従って多重クロス表・バート表の固有値）については次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ということを示している（ここらの詳細は Greenacre[22]，大隅他[8]を参照）．

① アイテム・カテゴリー型データ表から得た固有値を λ_k^A とする．

② このとき，固有値の総和は以下のようになる．

$$\sum_{k=1}^{K^*}\lambda_k^A=\frac{n^*}{M}-1\left(=\frac{n^*-M}{M}\right),\quad \text{ここで}n^*=\sum_{j=1}^{M}n_j$$

③ 固有値の個数：$K^*=\sum_{j=1}^{M}\left(n_j-1\right)=n^*-M$，ここで M は項目の総数

④ ある固有値 λ_k^A の寄与率は，

$$\frac{\lambda_k^A}{\sum_{k=1}^{K^*}\lambda_k^A}\times 100=\lambda_k^A\times\frac{M}{n^*-M}\times 100(\%)$$

⑤ ここで，$0\le\lambda_k^A\le 1$ であるから，各成分の寄与率が $\dfrac{M}{n^*-M}$ を越えることはない．

⑥ とくにすべての項目が 2 項選択，つまり $n_j=2\ (j=1,2,\cdots,M)$ であるとき $n^*=\sum_{j=1}^{M}n_j=2M$ であるから，$\dfrac{M}{n^*-M}=\dfrac{M}{2M-M}=1$ となる．このときがもっとも寄与率が高いということになる（換言すると項目の選択肢数・カテゴリー数が増えるほど固有値と寄与率は小さくなるということ）．

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を扱うときに知っておくと便利である．

数値例：

ある 7 つの質問を考える。各質問には 5 つの選択肢があるものとする．これをアイテム・カテゴリー型に展開すると、延べで 7×5＝35 の総選択肢（カテゴリー）となる。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。

$M = 7\left(\text{質問の総数}\right)$

$n_1 = n_2 = \cdots = n_6 = n_7 = 5\left(\text{7つの質問のそれぞれの選択肢数}\right)$

$n^* = \sum_{j=1}^{M} n_j \;\Rightarrow\; n^* = n_1 + n_2 + \cdots + n_6 + n_7 = 5\times 7 = 35\left(\text{総選択肢数}\right)$

$K^* = \sum_{j=1}^{M}\left(n_j - 1\right) = n^* - M \;\Rightarrow\; 35 - 7 = 28$

$\sum_{k=1}^{K^*} \lambda_k^A = \frac{n^*}{M} - 1 \left(= \frac{n^* - M}{M}\right) \;\Rightarrow\; \sum_{k=1}^{K^*} \lambda_k^A = \frac{n^*}{M} - 1 = \frac{28}{7} = 4\left(\text{固有値の総和}\right)$

$\sum_{k=1}^{K^*} \lambda_k^A = \lambda_1 + \lambda_2 + \cdots + \lambda_{27} + \lambda_{28} = 0.608 + 0.506 + \cdots + 0.025 + 0.018 = 4\left(\text{数値例で確認}\right)$

$\lambda_1^A \times \frac{M}{n^* - M} \times 100 = 0.608 \times \frac{1}{4} = 15.2\left(\%\right)\left(\text{第1成分の寄与率}\right)$

$\lambda_2^A \times \frac{M}{n^* - M} \times 100 = 0.506 \times \frac{1}{4} = 12.6\left(\%\right)\left(\text{第2成分の寄与率}\right)$

……

……

以上にみるように, この例では各固有値の寄与率は $\frac{M}{n^* - M} = \frac{M}{K^*} = \frac{1}{4} = 0.25\left(25\%\right)$ を越えることはない．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 n^*（全質問の延べの総選択肢数）が多くなると次第に寄与率が小さくなることも分かる（つまり、選択肢数を増やすほど情報が曖昧になるということ）.

【参考文献】(※順不同)

[1] 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.

[2] 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］

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

[4] 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.

[5] 林知己夫（1993），数量化－理論と方法，朝倉書店.

[6] 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.

[7] 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,(CD-ROM のみ).

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

[9] 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.

[10] 大隅昇（監訳）（2011），調査法ハンドブック，朝倉書店.（*）下の［26］の全訳版.

[11] 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.

[12] 舟島なおみ（2000），質的研究への挑戦，医学書院.

[13] 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.

[14] P.G.ホーエル著，浅井晃，村上正康訳（1990），入門数理統計学，培風館.

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

[16] Benzécri, J.-P. (1976), *L'Analyse de Données, Tome* 1: *Taxinomie*, *Tome* 2: *L'Analyse des Correspondances*, Dunod (second edtion).

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

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

[19] Brigitte Le Roux and Henry Rouanet (2004): *Geometrical Data Analysis – From Correspondence Analysis to Structural Data*, Dordrecht Kluwer.

[20] Brigitte Le Roux and Henry Rouanet (2010): *Multiple Correspondence Analysis*, Series: Quantitative Applications in the Social Sciences No.163, Sage Publications, Inc.

[21] Everitt, B.S. and Dunn, G.(2001), *Applied Multivariate Data Analysis*, second edition, Arnold.

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

[23] Flick, U.(2002), *An Introduction to Qualitative Research* , second edition, Sage Publications.

[24] Greenacre, M.J. (2007), *Correspondence Analysis in Practice* (second edition), Academic Press.

[25] Greenacre, M.J. (1984), *Theory and Applications of Correspondence Analysis*, Academic Press.

[26] Groves, R.M., Fowler, F.J. Couper, M.P. and others (2004), *Survey Methodology*, John-Wiley.

[27] Lebart, L., Salem, A. and Berry, L. (1998), *Exploring Textual Data*, Kluwer Academic Publishers.

（※）テキスト・マイニング研究会のホームページに，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・テキスト情報などが掲載されているので参考にするとよいだろう.

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

2013 年 1 月　更新