対応分析法・数量化法III類の考え方

テキスト・マイニング研究会
第3回WordMiner活用セミナー
2005年5月19日－20日
於 統計数理研究所

大隅　昇
ohsumi@ss.iij4u.or.jp

本日のトークの内容

- WordMiner「多次元データ解析」の主要な機能
  - 対応分析法とその周辺の処理機能（有意性テスト他）
  - クラスター化機能（ハイブリッド方式），…
- 今回は「対応分析法」の基礎的な考え方を中心に以下を紹介
  - データとデータ表をどう考えるか
  - 定性情報の「数量化」とは
  - データ表の特徴（どのようなデータ表を扱うか，その理由）
  - 対応分析法・数量化法III類の数理（仕組みの簡単な紹介）
  - WordMinerに実装の機能
  - 数値例による確認
- テキストの内容の要点を抜粋して紹介する
- 次のセミナーへの入り口，他の知識を理解するための入門

2

◆データをどう考えるか

- 数学的分類
  - 連続的変数か離散的変数か
  - 統計学のテキストなどにある分類
- 「尺度（scale）」による分類
  - 質的データ（名義尺度，順序尺度）
  - 量的データ（区間尺度，比例尺度）
- 両者を勘案して使い分ければよい
- 定性調査においては尺度分類の方が説明しやすい
- とくにテキスト型データを質的データと考えること

3

尺度による分類と数学的分類の要約［表１］

		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 色度、光沢度	温度（℃） 硬度 比重	単位を持つ測定値データの大部分（長さ，重さなど）
離散量	多値	機械名 作業者名 工場名 原産地名，など	段階的評価の成績データ 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	ＴＶのチャンネル 体育館の利用日数 車の故障台数，など	車の走行台数 都市内人口 参加者数 家の戸数
	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（回数を考慮に入れれば多値データとなる）	瓶入りと缶入りのジュース単価（二値の分類区分で層化）

こうした分類区分を目安とすればよい．調査設計，データ取得時から意識すること．

4

◆定性情報・質的情報の数量化とは

- 定性情報の取得，あるいは定性調査におけるデータ取得方式（data collection mode）には様々な方法がある．
- 調査環境の変化，とくに電子的調査情報取得手法（CASIC，CADAC）の研究の進歩がある
- テキスト型データの取得が容易となった
  - インターネット調査，GI，FGなどの自動化（OFGなど）
  - コール・センター，コンタクト・センターなどでのデータ収集
  - ITスキル，技術改善が優先・重視されている
  - データ解析の本質が軽視，ソフトウェア依存となっている
- 様々な分野におけるテキスト型データを含む質的・定性的情報の分析への要求の高まり
  - 福祉研究，看護学研究，人事管理，企業組織研究，…
  - 定量的な選択肢型の情報収集では適切に対応しきれない

5

テキスト・マイニングとそのツールの登場

- 多数のツールが登場してきた
- 市場調査分野などで先行的に普及した（国内では）
- コストパーフォーマンスと性能・機能の乖離現象
- 開発側から考えるとソフト開発投資額も無視できない規模
- 研究者にとっては価格の問題がハードルであった
- 市場調査などではやや見直し・反省の気配（熱が醒める？）
  - 本当に有効か，何に役立ったと言えるのか，…
  - ソフトに投げ入れれば何かがすぐに得られるという幻想と幻滅（？）
  - 分析内容の暗箱化（何を行っているのかが曖昧）
  - クライアントの要求とのギャップ，…
- 一部の研究分野における質的研究への関心の高まり
- どのような研究，とくに定性調査（qualitative research）があるか
  ⇒テキスト・マイニング研究会ホームページを参照

6

何が問題か？

- 正しい理解が徹底しないこと
- 依然として過剰期待があること
- ソフトを十分に使いこなせていない
- 客観的に見直す必要があること（適用可能性の再評価）
- ソフト提供者の側としては，ノウハウの可能な範囲の開示
- 用いる方法論の内容の透明化（暗箱化の回避）
- 利用者の理解向上が必要（誤用・濫用の回避）

- このようなことで，微速前進ではあるが少しでもWordMinerの理解を徹底したいと考えての普及活動の一環

7

◆数量化法III類と対応分析法

- 数量化法III類（quantification method, type III）
  - 「III類」とは飽戸弘氏の命名による俗称
  - 林知己夫氏により提唱された手法（1955～1956年頃）
  - 正式には「パターン分類（法）」という（この年代に稀有な発想）
  - 一連の数量化法（quantification methods）の一つ（数量化Ⅰ類～Ⅵ類）
- 対応分析法（ Analyse Factorielle des Correspondances）
  - ベンゼクリ氏（J.-P. Benzécri）により提唱（1962年頃）
  - 正式にはAFC（Analyse Factorielle des Correspondances）
  - CA：Correspondence Analysisとして英語圏に紹介された（1974年，M.O.Hill）
  - 対応分析（法）と命名（大隅・林他⇒国内で初めて紹介，M.Rouxを招聘）
  - コレスポンデンス分析，コレスポンデンス・アナリシスなど
    （多分，その本質的な意味が分からなかったのでこんな名称が出た）
- その他，同等あるいは類似の手法が多数ある（テキスト参照）．

8

数量化のために

- そもそも定性情報は量的処理操作が難しい，あるいはできないことがある，よって数量化（quantification）が必要
- とくにテキスト型データの場合は，各種の事前処理も必要
- 大まかには「解析に適した形式のデータを生成するまで」の操作（不定形な情報を構造化するまでの操作）と，その後の「（詳細な）解析を行うための解析手順」とがある
- 事前の手当として，例えば調査であれば「調査計画，データ取得方式，調査方式（モード），調査票」の効果的な設計が必須
- 他の分野（研究，実務）でも，ほぼ同様の対応が必要

9

数量化とは　ー簡単な例示による確認ー［例3］

- ある調査における取得データを取り上げる．

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

※尺度の分類によればいずれも「順序尺度」である

10

ある調査データの一部[(回答者)×(項目)](表7)

					選択肢(テキスト)のまま			選択肢をコード化				
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。(コード)	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。(コード)	住んでいるまちが気に入っているか。(コード)	緑(みどり)が多いと感じるか。(コード)	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

11

「質問」の特徴と留意点

- いわゆる「質的データ」である
- とくにこの2問は「順序尺度」である(名義尺度であって選択肢の意味に序列の関係がある)

- さてここで，以下の問題を考える.

Q:このようなデータに対して以下の操作は可能だろうか

①四則演算を行うこと，例えば平均値や標準偏差を求めることが可能か.

②主成分分析，因子分析など，原則として「量的データ」を対象とした手法は適用できるのか.

③クロス表を作成し比率データを観察するのはなぜか.

12

これらの解答は，…

①四則演算を行うこと，例えば平均値や標準偏差を求めることが可能か．

答え：形式的には可能でも，演算の意味があるかは保証されない．

②主成分分析，因子分析など，原則として「量的データ」を対象とした手法は適用できるのか．

答え：これも形式的適用はあり得るが，実はいろいろと問題がある，とくに因子分析の利用には細心の注意が必要．

③クロス表を作成し比率データを観察するのはなぜか．

答え：質的データの情報のもっとも簡単な「計量化・数量化」の方法であるから（対応分析法に密接に関係）．

13

いわゆる「数量化」の原点は何か？

- 数量化法III類の誕生の経緯（故林知己夫氏）
- その考え方・思想は何かを簡潔に言えば，…
  - 質的データに対して，数値はアプリオリに与えるべきではない．
  - 線形性（線形モデル）をその名目的な数値にそのまま想定はできない．
  - 「数量」は現象を説明するであろうデータに基づいて作られるもの（別に作るもの，「数量の作り方」を考えるべきこと）．
  - 新たな（なるべく線形となるような）座標空間（スコア）を作り出すこと．
- この発想は，そのまま定性情報である「テキスト型データ」に当てはまるであろう（選択肢はこの一つではないが）．
- つまりテキスト型データは一旦計量化・数量化した後に，さらなる分析に進むべきであるという選択肢があるだろう．
- 換言すると，生のまま数値（コード）として扱う処理には問題があるのではないか？

14

◆対応分析法（AFC）とは？

- 質的データ，とくに「調査型データ」は多くの場合そのまま計量化して使えない（ここは数量化法と似た発想）．
- 質的データの原点はクロス表型データ表にある．
- つまり，質的データ（名義尺度，順序尺度）情報を集約化したデータ表である．
- クロス表（分割表）を基礎情報と考えると，これは比率データで観察を行うように，視点が表側の側と表頭の側と2つの方向から分析できる．
- このとき，それぞれ比率データは多次元空間内に布置する多次元データと見なすことができる（⇒後述）．
- この視点から，クロス表型データ表の表側，表頭の対応関係を測ることができるのではないか．

15

古典的な手法：クロス表の独立性の検定

- 既存の方法論として，分割表（クロス表）の「独立性の検定」がある（例：「ピアソンのカイ二乗統計量」を用いる）．
- この発想は，以下のような考え方である．
- 表側と表頭の2つの項目I，Jの間には「関係がない（独立）」という帰無仮説をたてる．
- つまり表側と表頭にある2つの項目は無関係という独立モデル（ $p_{ij} = p_{i+}p_{+j}$ ）．⇒テキスト，表19，図2などを参照
- これが統計的に棄却されれば帰無仮説を棄却，よって表側と表頭の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（関係がありそうと言えるだろう）とする検定法（隔靴掻痒の分析結果となる）．

16

対応分析法の本質

- ベンゼクリはこれを別の視点から謎解きした.
- クロス表の表側と表頭のそれぞれの項目の相対比率データを考える(「プロフィル」と名付けた).
- プロフィルのカイ二乗距離(加重付距離)を考え, これが近いものは近い位置にあるとする(加重化した比率データが似ているものは近いとする).
- つまりプロフィルを多次元空間内に布置する多次元データと考え, その空間内での加重平均指標を作り次元の縮約を行う(主成分分析のような合成指標化を考える).
- プロフィルは行と列との両方から観察できるから, 双対性を考慮して分析を行う.
- 一見すると, 数量化法III類と異なる定式化のように見えるが実は同等の手法である.

17

ピアソンのカイ二乗統計量との関係

- 定式化の結果として(そのようになるように定式化して), ピアソンのカイ二乗統計量と密接な関係にある.
- 本来, クロス表は2つの項目間に何らかの意味があるとして観測(測定)したはずなのに, 独立性の検定の帰無仮説(独立モデル)のような設定では情報の活用が十分ではない.
- よって, クロス表の行と列との2項目間の関連性を主成分分析型手法とすることで, 固有値(=相関の情報に相当)の大きさで測ることを可能とした.
- つまり, 2つの項目間の関連性と対応関係を計量的測れることになる(ここでも質的データの計量化となる).

18

対応分析法・数量化法III類が扱うデータ表

- 対応分析法・数量化法III類の原理から考えると，そこで扱えるデータ表の形式にはかなり自由度がある．
  - 一般に対応分析法・数量化法III類で扱うデータ表の形式
  - WordMinerで扱うデータ表の形式
- 扱うデータ表の相互の関係を理解することが肝要である．
- これについて概略を述べる．テキストも良く読んでいただきたい．
- 後述するように，数量化法III類と対応分析法が扱うデータ表は一見すると異なるようにみえる（ここに誤解がある）．

19

対応分析法で扱う「データ表」

- 原則として「二元のデータ表（クロス表型）」を基本
- 二値の応答型データ（「yes」「no」型，0－1型）である場合
- 「二元のデータ表」の特徴（条件）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
  - 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
- 以上を満たせば，ほぼ，どのようなデータ表でも利用できる．
- 同時に，さまざまなデータ表の変形が用意できる．
- WordMinerを利用するため，テキスト型データをこれに適合させる工夫が必要．

20

例えば，…

- 通常の二元クロス表（分割表）が基本となる
- （0，1）型データ行列（二元クロス表の特別な場合に相当）
- 多重クロス表・バート表（「多元クロス表」ではないことに注意）
- 多くの統計表（数値が非負の集約化データで上の条件を満たすとき）
- この「二元のデータ表」をどのように作成するか
- つまりデータ収集法（data collection mode）と取得計画（design）の問題である
- 何でも質的データ（⇒自由回答・自由記述，テキスト型データ）であればよいとはならない．

21

取り上げた調査データ（表7）の2項目からクロス表を生成

（表6に相当）

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

22

その他の表の観察

- テキストに，対応分析法で扱う典型的な例をいくつか示した
  - 例1：二値型データ表（セル内度数が1の二元データ表）（表2，3）
  - 例2：好みの清涼飲料水（「好き」かそうでないか）（表4，5）
  - 例3：ここで説明に用いた環境意識調査から得たデータ（表6，7）
  - 例4：ある自治体で行った意識調査の例（表8～13）⇒表14も参照
  - 例5：種々のデータ表の関係を知るために作った人工データ表（表15～18）
- 以上に登場するデータ表の名称と意味に気をつけよう
  - ①（回答・サンプル）×（多変量・多数項目）型の多次元構造
  - ②二元クロス表あるいはそれに相当のデータ表（①から生成）
  - ③アイテム・カテゴリー型データ表（インジケータ行列）
  - ④多重クロス表（バート表）
- それぞれ表側と表頭にどんな情報が置かれているかに注意

23

データ表の相互の関連（重要）

- 国内では数量化法III類は，(0, 1)型データとすることが多い
  - （サンプル）×（もの，項目，カテゴリー）のデータ表（例：表2～5）
  - アイテム･カテゴリー型のデータ表（⇒後述）
  - こうしたデータ表しか扱えないと思われている節がある（誤解）
- 対応分析法ではより一般的に2元のデータ表を扱う
  - （回答・サンプル）×（多変量項目）型
  - アイテム･カテゴリー型（インジケータ行列）
  - 多重クロス表・バート表（Burt's tables, Burt's matrix）
  - この他の二元表タイプ
  - これらのデータ表の間の関係が重要（⇒実は同じことを考えている）

24

対応分析法におけるデータ表の関係

- データ表の関係を例でみる（例4：自治体の意識調査）
- 元となる「（回答・サンプル）×（項目）型」データ表（表8）
- ここで，質的データ（名義尺度，順序尺度）が多いことに注意
- コード変数，文字変数（テキスト型データ）と表記が混在
- ある2項目を切り出し（指定すると）表9を得る
- それをコード化すると（しなくてもよいが）表10となる．
- 表10をアイテム・カテゴリー型（インジケータ行列）に展開する（表11）
- 表10（表9）からクロス表を生成（表12）
- 表11（アイテム・カテゴリー型）から行列演算（行列の積）で，表13の多重クロス表（バート表）を生成
- 表12のクロス表は表13の多重クロス表内のブロック行列となる
- 以上の関係は多数項目（多変量）となっても同様になる（⇒演習問題2とその「補足」に要約した）．

25

多重クロス表の例［表13／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
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
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

①表12のクロス表がブロック行列で入っている
②対角ブロック行列がそれぞれの項目の周辺度数分布
③当然，対称行列となっている

質問Aと質問Bのクロス表

◆対応分析の仕組み

- 二元のクロス表（型）を基本のデータ表とする（表19）
- このデータ表からプロフィルを作る（表20，図2）
  - 行のプロフィル（行の相対比率のデータ）⇒式（8）
  - 列のプロフィル（列の相対比率のデータ）⇒式（9）
- プロフィルをそれぞれ行あるいは列の多次元空間内のデータと考え，この空間内での次元縮約を行う（加重平均による合成指標化）⇒つまり主成分（成分）を作る（図3）
- 要点は，
- 比率を考えることで，行と列との双方向からデータを観察することに注意する（双対性に関連する）（例えば式（19），（20））
- ここでいう「多次元データ，多次元空間」とは何かが重要（図3）．

27

数値例で見ることが理解を容易にする

- 数式の誘導やその意味付けをある程度知ることは重要
- 必要最小限の数理はテキストに記したのでこれを参照
- ここでは何より，対応分析法の仕組みを知ること
- これを実装するWordMinerの機能を理解すること
- 例示したデータ表のうちから「例5」のレストラン評価を用いる
- これは対応分析法の仕組みを理解するために簡略化した人工データである
- まずこの例5の「意味，内容」（データ表の意味）を理解する．
- 以下，それぞれの情報，データ表，用語の確認など順を追って説明する（（テキストで確認のこと）．

28

①「質問」と「選択肢」の確認

- まず用いる例を挙げる．［例5］

質問I：次に挙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. 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		

質問J：その選択時の評価基準は次の3つのうちのどれでしょうか？
1. 味　2. 量　3. 工夫・サービス

サンプル数（回答者数）がN=1,284（人），2項目（IとJ）の多変量データ構造のデータ表（表15）から，クロス表（表16）が得られる．
ここで質問を「項目」と呼び，選ぶカテゴリーを「選択肢」と名付ける．
この場合，項目Iの選択肢数はm=10，項目Jのそれはn=3となる．

29

②クロス表の生成

1）表16のクロス表を作成する．ここでは「行に項目I」を「列に項目J」を充てた．（表19参照）
2）クロス表は多次元データである（⇒後述，図2，図3）．
3）このクロス表を以下の式で書く（式（1），（2））．

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

（項目Iと項目Jのクロス表，ここで寸法は$m \times n$である）

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

（項目Iと項目Jの選択肢）
※項目Iの選択肢数はm個，項目Jの選択肢数はn個

「（項目I）×（項目J）」のクロス表　※表19を確認

30

クロス表：　$\underset{m\times n}{\mathbf{F}}=\left(f_{ij}\right)$　　$\left(f_{ij}\geq 0, i\in I, j\in J\right)$

表頭

表側		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_{++}

31

③プロフィルを作成する

（ｉ）行のプロフィル，つまり行の相対度数（相対確率）を求める．いわゆる行和を1と揃えた（行100%とした）表と思えばよい，これが表21である．［図2の左側の流れ］

（ii）列のプロフィル，列和を1とした列の相対度数（相対確率）を求める．これが表22である．［図2の右側の流れ］

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

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

※式（8），（9）と図2を確認

32

④多次元空間に布置することの意味[表21,22]

（ｉ）行のプロフィルとは「項目：評価基準の3つの選択肢」（＝3次元空間内）に「項目：レストランの10の選択肢」が布置するデータ空間と考える（行の向きに行和＝1と揃えたことに注意）．［表21］

（ii）同じく，列のプロフィルとは「項目：レストランの10の選択肢」（＝10次元空間内）に「項目：評価基準の3つの選択肢」が布置するデータ空間と見ることもできる（ここは列和＝1で揃えた）．［表22］

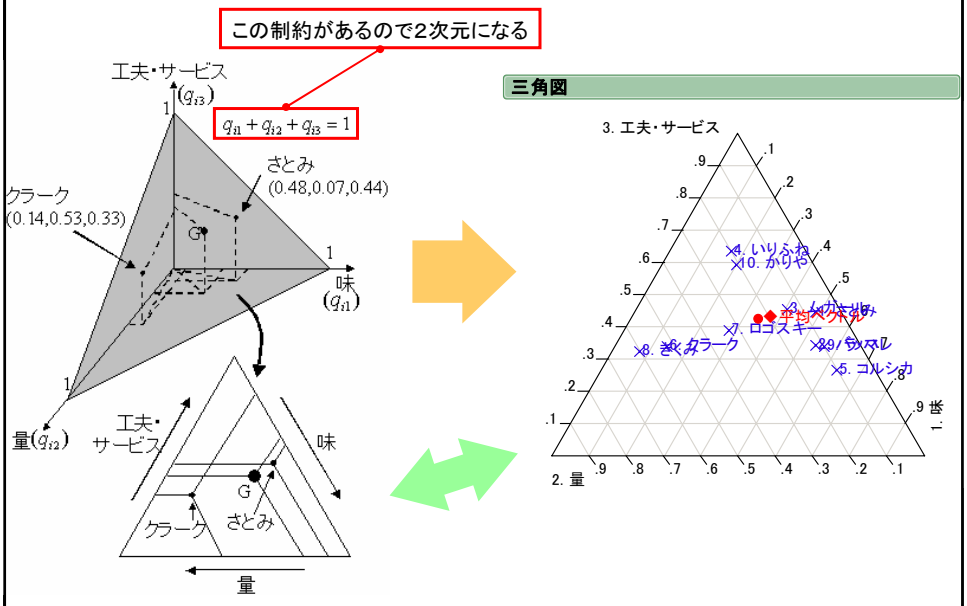
1）この一般的なクロス表（表19）から得られる行プロフィル，列プロフィルの関係を図式化したものが図2と図3である．

2）ここで“行と列の両方向”から見ていることに注意（行と列を転置しても情報は変わらない）．

※図2と図3を確認

33

行のプロフィルの意味を図で確認[図4]



さらに，…

1）例題についてさらに「行のプロフィル」側からの観察を続ける．つまり，図2，図3の左側のパスを考える．

2）このとき，$n=3$であるから行プロフィルを実際に「3次元空間内」に描いてみると図4の左側の図となるだろう．

3）ここで「行和＝1」という制約があるから，10のレストランの布置は実は$n-1=2$（次元）の平面内に入る（自由度が1だけ減る）．

4）実際に，図4の左の図の網かけ部の平面上に分布する．

5）これをそのまま（点の布置関係を保持したまま）射影すると図4の右側の図（三角座標系の図＝三角図）となるだろう．

6）この例は，視認できる3次元（2次元）の説明であるが，多次元になって次元数が上がっても考え方は同じである．

35

⑤次元数の縮約を行うこと

1）考えるデータ布置の空間が異なるが多次元空間内での次元縮約を考えることには違いがない．

重心座標系（barycentric coordinate system）

三角図：三角座標系（triangular coordinate system）

2）高次元の空間に布置されるデータを少数次元内に縮約せねばならない．

2）主成分分析と同じようなことが考えられるのか？

3）そのためのデータの構造は（作り方は）？

数式（10）～（14）のような変換を行ったデータを扱えばよい，またこの形でないと不都合を生じる．

4）こうする理由がある（例：ピアソンのカイ二乗統計量と関係）

※図3の布置図イメージを確認

36

⑥データ行列の分解（固有値問題他）

1）データ行列を作りその共分散行列の固有値問題に帰着する（あるいは元のデータ表の特異値分解：SVD）

2）データがある形であることを除けば多くの合成指標型手法（主成分分析など）と同じ解法となる.

3）固有値，固有ベクトルを求める

4）固有値と寄与率が情報縮約の程度を知る指標

5）プロフィル（を加工したある形）の加重平均（＝成分スコア）を求めることに帰着［成分スコア＝数量化スコア，数量化得点］

6）固有ベクトルが加重平均の式の係数に相当

（注：式（10），（11）に固有ベクトルを加重とする一次結合式）

7）成分スコア（数量化スコア，数量化得点）の算出

8）行と列との双方向から考えるから成分スコアも2組ある

37

⑦成分スコア，固有値の性質を確認

1）成分スコアは項目Iの選択肢i（$\in I$）と項目Jの選択肢j（$\in J$）のそれぞれに対して付与される. ［表25，図6参照］

2）両者の成分スコアの関係が重要（とくに双対性）

3）成分の数，つまり固有値の数はクロス表の行数（m）と列数（n）の少ない方から1を引いた数：$K=\min\{m, n\}-1$となる

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

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

※図3の布置図イメージを確認

38

成分スコアと元の確率行列の関係（表25）

		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							
	⋮	⋮	⋮	⋮	⋮	⋮	⋮								
	k	z^*_{1k}	z^*_{2k}	…	z^*_{jk}	…	z^*_{nk}	⇐ 列の項目 J の選択肢の成分スコア							
	⋮	⋮	⋮	⋮	⋮	⋮	⋮								
	k'	$z^*_{1k'}$	$z^*_{2k'}$	…	$z^*_{jk'}$	…	$z^*_{nk'}$								
	⋮	⋮	⋮	⋮	⋮	⋮	⋮								
	K	z^*_{1K}	z^*_{2K}	…	z^*_{iK}	…	z^*_{nK}								

補足１：固有値と寄与率

$$0 \le \lambda_k \le 1 \quad \left(k = 1, 2, \cdots, K; K = \min\{m, n\} - 1\right)$$ （第k成分の固有値）

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

式（24）

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

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

式（25）

40

補足2：固有値とピアソンのカイ二乗統計量の関係

固有値とカイ二乗統計量の関係

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

ピアソンのカイ二乗統計量

$$\chi^2=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{N\left(p_{ij}-p_{i+}p_{+j}\right)^2}{p_{i+}p_{+j}}=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(f_{ij}-\frac{f_{i+}f_{+j}}{N}\right)^2}{\frac{f_{i+}f_{+j}}{N}} \qquad \text{式(28)}$$

※この関係は対応分析法を考えるうえできわめて重要

41

参考：ピアソンの χ^2 統計量の一般型

$$\chi^2=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{\left(\text{実現度数}_{ij}-\text{期待度数}_{ij}\right)^2}{\text{期待度数}_{ij}}$$

①クロス表の表側と表頭という2つの分布の一種の距離となっている（乖離度を測る）.
②ただし，独立モデルのとき，もっとも小さくなるような距離.
③現実には，表側と表頭との間の相関・関連を知りたい，つまり独立でない程度を知りたいはずである.
④ピアソンの χ^2 統計量では直接はこれを測れない.
⑤式（26）はこれを固有値により測ろうとしていることに注意.

42

⑧双対性について

1）2項目I, Jの各選択肢に付与の成分スコア間の関係に注目

2）いわゆる「双対性（duality）」がある.（きわめて重要）

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

式（19）

（行の成分スコアは列のそれのプロフィルの加重平均）

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

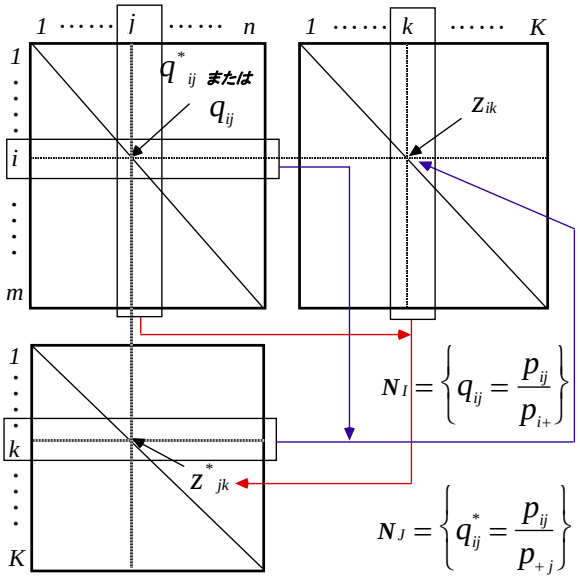
式（20）

（列の成分スコアは行のそれのプロフィルの加重平均）

※これら成分スコアの関係は図6と表25のように考える.

43

（ｉ）双対性の考え方（図6）



（ii）成分スコアと元の確率行列の関係（表25）

		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}								
	⋮	⋮	⋮	⋮	⋮	⋮	⋮								
	k	z^*_{1k}	z^*_{2k}	…	z^*_{jk}	…	z^*_{nk}								
	⋮	⋮	⋮	⋮	⋮	⋮	⋮								
	k'	$z^*_{1k'}$	$z^*_{2k'}$	…	$z^*_{jk'}$	…	$z^*_{nk'}$								
	⋮	⋮	⋮	⋮	⋮	⋮	⋮								
	K	z^*_{1K}	z^*_{2K}	…	z^*_{iK}	…	z^*_{nK}								

⇑ 行の項目 I の選択肢の成分スコア

⇐ 列の項目 J の選択肢の成分スコア

⑨成分スコアの布置図と同時布置図

1）行の選択肢への成分スコア，列の選択肢への成分スコアの<u>ドットプロット図（1次元）</u>や<u>散布図（布置図）</u>を描く．

2）同じ成分軸について行と列の成分スコアを<u>重ねた図を同時布置図</u>という．

$$\left(z_{ik}, z_{ik'}\right) \quad \begin{pmatrix} i = 1,2,\cdots,m \\ k,k' = 1,2,\cdots,K \\ K = \min\{m,n\} - 1 \end{pmatrix}$$

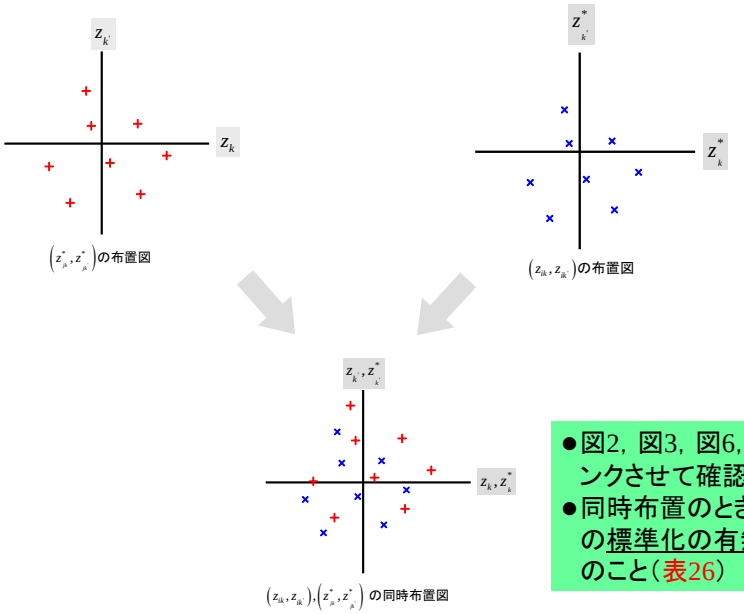
（行の選択肢への成分スコア）　式（21）

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

（列の選択肢への成分スコア）　式（22）

※図3の布置図イメージを確認

布置図と同時布置図（図3）



- 図2，図3，図6，表25をリンクさせて確認
- 同時布置のとき固有値の標準化の有無に注意のこと（表26）

「レストランと評価基準」の例で数値を確認

2項目への成分スコア

		成分スコア 第1成分スコア	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

固有値と寄与率

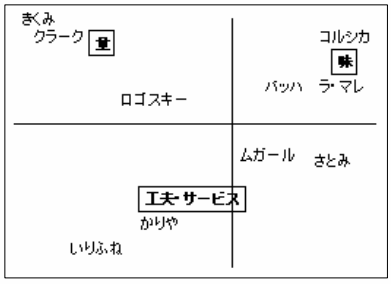
主成分 k	固有値 λ_k	寄与率(%)
1	0.19766	76.71
2	0.06002	23.29

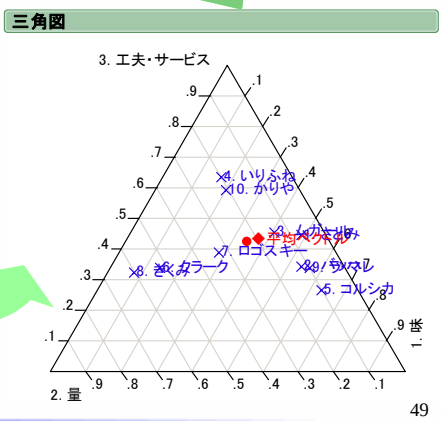
※成分スコアは表25，図6と対応させて確認のこと
①固有値の数はK=min{m, n}−1=2となった．
②2成分に対する成分スコアが算出される．

※表23，24に相当
※表25との対応に注意

48

得られた布置図と三角図の比較（図4，図5）





※図4と図5を比較のこと

元の2次元上の10のレストランの三角図布置が左の成分スコアとして再現されている（2成分スコアとして）

テキスト型データはなぜ質的データか？

- テキスト型データを質的データにどのように変換できるのかを考える．
- テキスト型データがまず分析処理単位に分けられていなければならない．
- つまり何らかの形で扱い単位を決めること⇒WordMinerではこれを「構成要素（fragments）」という．
- 構成要素はどんな単位であってもよい．
- WordMinerの標準装備の機能として分かち書き処理機能がある，これを行うと区切りのない日本語テキスト型データが分かち書き処理され構成要素の単位に分解される．
- よって何らかの「処理単位＝構成要素」が作れればよいので他の分かち書きツールを使ったデータを用いても良い．
  - 例：事例紹介，樋口耕一氏のトーク（KH Coder）
  - 例：茶筌などの形態素解析ソフトを使うなど

WordMinerで扱うデータ表形式（表27）

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う

51

WordMinerにおけるデータ表のイメージ図（図8）

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

①「（回答・サンプル）×（構成要素）」のデータ表の場合
②「（構成要素）×（質的変数）」のデータ表の例場合
③その他，必要に応じて変数指定を行い，このデータ表に合わせる

52

Web調査の質問の例

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。

53

「（回答・サンプル）×（構成要素）」のデータ表の例（表28）

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＞

54

生成した「(回答・サンプル)×(構成要素)」のクロス表(表29)

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

このタイプのデータ表は行列の寸法が大きく，かつ要素の度数が非常に疎になることが特徴
※固有値はきわめて小さい値となることが多い　⇒　固有ベクトルの向きの動きにも注意

55

「(構成要素)×(質的変数)」のデータ表の例(表30)

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

質的変数として指定

構成要素変数として指定

56

生成した「(構成要素)×(年齢区分)」のクロス表(表31)

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

一般には質的変数の選択肢数が大きくはないので，要素内の度数が比較的まとまるのが特徴
※固有値の値も，比較的大きい値が出る⇒よって寄与率も目安となる

57

再考：質的データの数量化の意味(数値例)

- 数量化の意味をより理解するために簡単な数値例をみる.
- 対応分析法の理解をより深める.
- この例を通じて林の数量化法の考え方を知る.
- 簡単な人工データを用意する.
- なるべく対応分析の機能と構造が見えるように作ってみる.
- 対応分析法と数量化法III類は“数理的には”同等であることを知る.

58

①用意したデータ表（表32）

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ここで「サンプル」と「銘柄」の2項目を選ぶ．
②前にみた例に合わせると項目*I*をサンプル，項目*J*を銘柄と対応させてみる．
③右側「次年度調査の銘柄」から「年齢区分」までは追加処理機能の説明（テキスト）．

59

②「（サンプル）×（銘柄）」のクロス表（表33）

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

項目「サンプル」の10の選択肢

項目「銘柄」の6選択肢

60

検討事項

- 「(サンプル)×(銘柄)」のクロス表としたが，前にみた“2項目のクロス表”とは少しイメージが異なる．これはクロス表になっているのか．
- 「サンプル」を表側項目Iとし「銘柄」を表頭項目Jとし，「好きな銘柄」に応答(回答)＝1としたのでデータ表の要素として「度数＝1」を充てたと考える．
- つまりこれもクロス表の特別なケースである．
- 林の数量化法III類では，これを(もの)×(項目・反応)の二値型データ表と考える．
- 同時に対応分析には「分布の同等性」という重要な性質がある(⇒後述，重要)．
- これが保持できるような二元表であればほとんど適用できる．

61

③固有値，寄与率と成分スコアの算出(表34，36)

k	固有値	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

①固有値の数：$K=\min\{m,n\}-1=6-1=5$となる．
②人工データの例であり，かつある構造をもつように作ったので始めの固有値の寄与率が高い．
③始めの2成分で全情報の約8割を占める．
④固有ベクトルを使って成分スコアを求めると表36のようになった．
⑤ここで行(サンプル)と列(銘柄)の双方の成分スコアがある(図3，表25を参照)．この関係を良く理解すること．
※成分スコア⇒　表36　⇒　表25も参照

62

④第１成分スコアの大きさで行，列を並べ替える（表35）

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		

①きれいに線形に並んでいる，この相関はどの程度？
②第2成分以上のスコアでも同じように並べ替えができる．

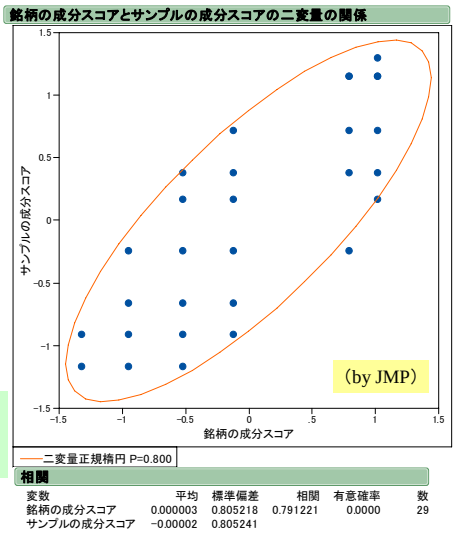
63

⑤第１成分スコアによる散布図（図9）

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

●相関係数：r=0.7912
●固有値の正の平方根
$\sqrt{\lambda_1} \doteqdot 0.7912$

（Twin-map，Dual-mapなどの呼称あり）



①スコアの実寸で図を描くと縦軸，横軸の間隔（スコアの間隔）が等間隔でない．
②つまり数量化により元の選択肢（質的データ）が量的データに変換されている（“数量化”された）．
③成分スコアの相関係数：r=0.7912となった．これが第１固有値の正の平方根に一致する．

要約：数量化の意味を再考する

- 元のデータ表（クロス表，表33）の行（サンプル）と列（銘柄）の各選択肢に新たな数量（成分スコア）を付与できた．
- それを成分別に観察すると，成分スコアの相関（係数）が固有値の正の平方根に相当する．
- つまり，選択肢に付与された成分スコアが数量（量的データ）として機能する．これが数量化の目標であった．
- 対応分析法では，プロフィルに注目したが，林の数量化法III類では，正にこれが付与した数量の相関を最大にするという最適化を行うことに相当する．
- データ表（クロス表）は多次元データであるが，これが少数の次元で説明できる（元の多次元の情報の損失がなるべく少ないような線形の加重平均を作った）．
- しかし，明らかに主成分分析のような量的データの分析とは異なる考え方である（質的データの数量化）．

65

分布の同等性（distributional equivalence）

- 例（表33）について，まずサンプルの側から考える．
- この例では，銘柄が6種であるから，サンプルの回答パターンの出現の組み合わせは実は「2^6=64通り」しかないはずである．
- よって，サンプルの回答パターンはこの64通りのどれかである．
- 仮に同じパターンがあったなら分析によって得る成分スコアも同じ数値（スコア）のはずである．
- 同じパターン（行パターン）となったサンプルを行方向に併合したとする．度数＝1の部分が併合により加算される．
- しかしプロフィル（比率データ）を考える限りはその値に差異はない（そのようになっている），

66

実験1：行側の同じ回答パターンの併合

ID	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+8	0	0	2	0	0	2	4
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

①この例（表33）では，サンプル3とサンプル8が同じ回答パターン
②これを行側で併合した（度数「1」⇒度数「2」となる）
③このとき表33とこの表の対応分析の結果は同等である．

67

実験2：さらに行と列に同じパターンがある例

ID	銘柄A	銘柄B	銘柄C	銘柄D	銘柄E	銘柄F	銘柄A*	銘柄C*	行和
サンプル1	0	1	0	0	1	1	0	0	3
サンプル2	0	0	0	0	0	1	0	0	1
サンプル3	0	0	1	0	0	1	0	1	3
サンプル4	0	1	1	0	1	1	0	1	5
サンプル5	0	1	1	0	0	1	0	1	4
サンプル6	1	1	1	0	1	0	1	1	6
サンプル7	1	1	0	1	1	0	1	0	5
サンプル8	0	0	1	0	0	1	0	1	3
サンプル9	1	1	0	0	1	0	1	0	4
サンプル10	1	0	0	1	1	0	1	0	4
サンプル6*	1	1	1	0	1	0	1	1	6
サンプル6*	1	1	1	0	1	0	1	1	6
列和	6	8	7	2	8	6	6	7	50

①サンプル3とサンプル8が同じ回答パターン
②さらに行側にサンプル6と同じパターンが2サンプルある（6*）
③列側にも銘柄Aと銘柄Cと同じパターンが2つある（A*，C*）
④ここでは行，列双方に同一パターンがあることに注意して圧縮する．

68

実験3：実験2のプロフィルの同じ行と列を併合

ID	銘柄A+A*	銘柄B	銘柄C+C*	銘柄D	銘柄E	銘柄F
サンプル1	0	1	0	0	1	1
サンプル2	0	0	0	0	0	1
サンプル3+8	0	0	4	0	0	2
サンプル4	0	1	2	0	1	1
サンプル5	0	1	2	0	0	1
サンプル6+6*	6	3	6	0	3	0
サンプル7	2	1	0	1	1	0
サンプル9	2	1	0	0	1	0
サンプル10	2	0	0	1	1	0

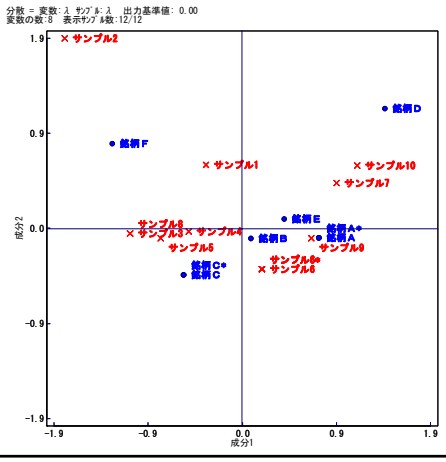
①行側でサンプル3とサンプル8を併合
②行側でサンプル6と同じパターンを2サンプルを併合（6*）
③列側を銘柄Aと銘柄A*を併合
④列側を銘柄Cと銘柄C*を併合
⑤以上を行うと大きさが（9×6）の上のデータ表となる.
⑥実験2のデータ表とこのデータ表の対応分析の結果は同等である.

69

確認1：実験2の表から分析（12サンプル×8銘柄）

ｻﾝﾌﾟﾙ数：12

No	固有値	寄与率	累積寄与率
1	0.539	0.580	0.580
2	0.202	0.218	0.798
3	0.127	0.136	0.935
4	0.044	0.048	0.982
5	0.017	0.018	1.000

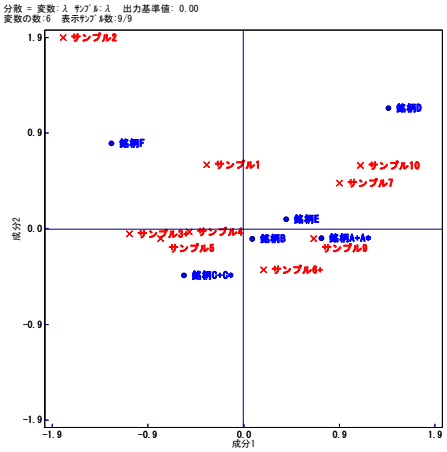


①このチェックにはJUSE/StatWorksを用いた.
②固有値は次元縮退（階数が減る）があって5個しか出ない.

確認2：実験3の表から分析（9サンプル×6銘柄）

ｻﾝﾌﾟﾙ数：9

No	固有値	寄与率	累積寄与率
1	0.539	0.580	0.580
2	0.202	0.218	0.798
3	0.127	0.136	0.935
4	0.044	0.048	0.982
5	0.017	0.018	1.000



①固有値はここでも5個しか出ない．
②布置図の点の重複関係の違いに注意のこと（サンプル数，銘柄数が異なる）．

以上から分かることは，…

- 実験1：（9×6）と表33（元のデータ表：10×6）は同じ結果となる（2サンプルを併合）．
- 実験2：（12×8）と実験3 ：（9×6）とは同じ結果となる（サンプルと銘柄の同じプロフィルをそれぞれ併合）．
- ただし，プロフィル（行あるいは列の比率のパターン）を変えるような併合はできない．
- 回答パターンの組み合わせを知っているとデータ表の圧縮化が可能である（同一パターンを併合する）．

72

◆「短文」による対応分析法の機能の確認

- WordMinerにおける対応分析法の機能を知るには，

①人工的に短い文章を作ってみる
②雑誌，小説・エッセイ，その他書籍に登場する短文を集める
③できるだけ，構造が分かるような文章を用意する
など，日頃から心がけるとよい．

- ここでは，ごく単純な文章を用意した．
  - （ｉ）『私が文章を書く』を基本にいくつか変形文を作る．
  - （ii）「否定語」を入れる．
  - （iii）単語「文章」がない文をいくつか入れる．
- このとき，対応分析の結果に文章の特徴がどう表れるかを観察する．
- 実際のテキスト型データにはさらに複雑になるからきめ細かい探査が重要となる．

73

①データ表，分かち書き，キーワード，他

サンプル番号		（ｉ）原文	（ii）分かち書きのみ	（iii）キーワード	（iv）分かち書き編集（助詞，句読点削除）
SEQ	ID	短文	短文-分かち書き	短文-キーワード	短文-分かち書き-編集 all
[00000001]	1	私は、書かない。	私 は 、 書かない 。	私	私 書かない
[00000002]	2	私が書いた文章である。	私 が 書いた 文章 で ある 。	私 文章	私 書いた 文章 ある
[00000003]	3	私に書いた文章です。	私 に 書いた 文章 です 。	私 文章	私 書いた 文章 です
[00000004]	4	私の書けない文章だ。	私 の 書けない 文章 だ 。	私 文章	私 書けない 文章 だ
[00000005]	5	私が書いた文章である。	私 が 書いた 文章 で ある 。	私 文章	私 書いた 文章 ある
[00000006]	6	私には書けない文章です。	私 には 書けない 文章 です 。	私 文章	私 には 書けない 文章 です
[00000007]	7	私と書いた文章である。	私 と 書いた 文章 で ある 。	私 文章	私 書いた 文章 ある
[00000008]	8	文章には書けない私のこと。	文章 には 書けない 私 の こと 。	文章 私	文章 には 書けない 私
[00000009]	9	私が書く。	私 が 書く 。	私	私 書く
[00000010]	10	私が書いた。	私 が 書いた 。	私	私 書いた
[00000011]	11	私と書いた。	私 と 書いた 。	私	私 書いた
[00000012]	12	私を書いた文章である。	私 を 書いた 文章 で ある 。	私 文章	私 書いた 文章 ある

①図8のイメージ図を参照のこと．
②「分かち書き」そのものを用いるとき（（ii）の欄）
③分かち書きから「助詞」を削除（（iv）の欄）
④キーワード抽出の結果（動詞系「書く，書いた…」が除外されてしまう）（（iii）の欄）

※これらを比べると“同じ原文を用いても情報が変容”することが分かる．
つまり分析目的に応じた編集加工が必要となる（客観性を失わない範囲で）．

74

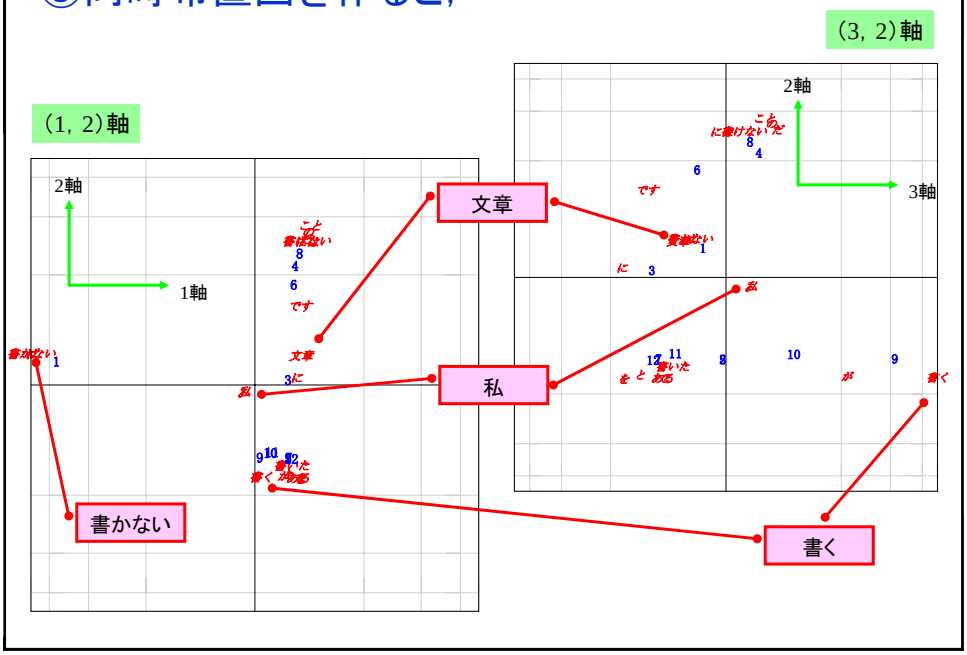
②「分かち書き」の構成要素の分布

構成要素	構成要素数	サンプル度数
私	12	12
文章	8	8
書いた	7	7
ある	4	4
が	4	4
で	4	4
書けない	3	3
です	2	2
と	2	2
には	2	2
の	2	2
こと	1	1
だ	1	1
に	1	1
は	1	1
を	1	1
書かない	1	1
書く	1	1

①利用頻度の高い語は「私，文章，書いた」など⇒共通に使われている
②利用頻度の少ない単語のうち「書く，書かない」
③これらの単語が布置図のどこに位置するか
④元の語句の並びはどう現れるか

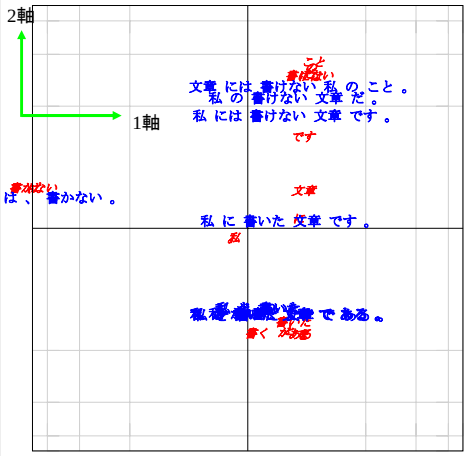
75

③同時布置図を作ると，…

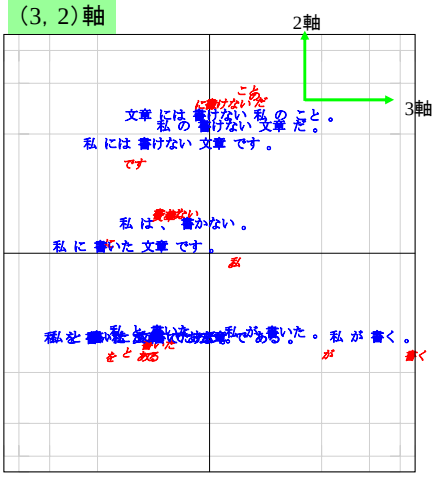


④「分かち書き文との同時布置図を作ると，…

(1，2)軸



(3，2)軸

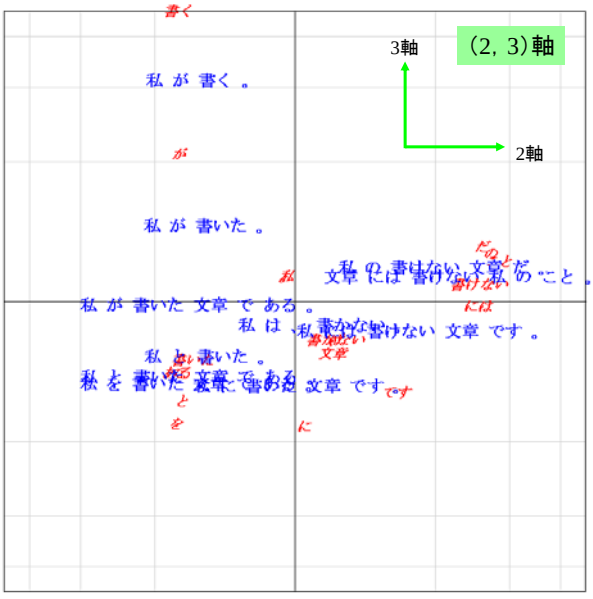


前の図を，分かち書きをラベル変数として布置

①単語の使われ方（文のつながり）と出現頻度の関係がみえる
②否定語の有無で分かれる
③利用頻度の少ない単語の位置
④共通語（「私」「文章」）つまり誰もが使った単語は重心近くに布置（平均ベクトルに近い）

⑤前の図の軸の向きを変えると視点も変わる

(2，3)軸



①文字情報が布置図内に入りので，軸の向きを変えても見える情報に違いがある．
②固有値，寄与率などを考慮しながら，高次成分の軸の組み合わせを観察すること．
③併せて寄与度（絶対寄与度，相対寄与度）の観察も行う．

◆その他の補足事項

- 寄与度の利用
  - 絶対寄与度（absolute contribution）あるいは寄与度（contribution）⇒ある成分に占めるある選択肢の寄与の程度
  - 相対寄与度（relative contribution）あるいは平方相関（squared correlation）⇒ある選択肢がどの成分で寄与が高いかの測度
  - 参考：主成分分析でいうと，絶対寄与度が成分負荷量（因子負荷量）の観察，相対寄与度が変数の重相関係数（の二乗）に類似した利用・解釈になる（⇒主成分分析の数理を確認のこと）
- 追加処理機能（supplementary treatment）と追加要素（supplementary elements）
  - 追加要素として「構成要素変数」を追加処理
  - 追加要素として「質的変数・属性」を追加処理
  - これについては別の項ならびに資料の「6.5節，6.6節」の数値例を確認

79