WordMiner™における多次元データ解析

― ミニチュアデータによる対応分析法の仕組みの解読 ―

WordMiner活用セミナー
2006年6月30日
於：日本電子計算株式会社

テキスト･マイニング研究会
http://wordminer.comquest.co.jp/
統計数理研究所
大隅　昇
ohsumi@ss.iij4u.or.jp

トークの内容

- ここではとくに，対応分析法（数量化法III類，コレスポンデンス・アナリシス）の仕組みを概観する．
- ミニチュア・データで数値例を追うことで理解を深めること．
- 質的データの数量化・計量化の意味は？
- ピアソンのカイ二乗統計量によるクロス表の独立性の検定について（対応分析法と密接に関連する）．
- 対応分析法の基本的な仕組みについて．
- 出力される統計値をどう読むか（主要な統計値のみ）．
  - 固有値，寄与率
  - 成分スコア
  - 布置図，同時布置図
- ごく基本的な知識についてのみ触れる．WordMinerは，さらに多様な処理機能を備えている，
  - 有意性のテスト（頻度，距離）⇒別の資料を用意した
  - 追加処理機能

2

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.
- 統計学の初歩的な知識は必要である.
  - 質的データとは何か？
  - クロス表とは何か？
  - なぜ(クロス)集計で比率データを用いるのか？
  - クロス表の独立性の検定(ピアソンのカイ二乗統計量の利用)
- 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.
- データ表，用語の確認などはテキストで再確認のこと.

3

対応分析法による数量化とは？

- テキスト型データに限らず，質的データ，とくに「調査型データ」は多くの場合そのままでは計量化情報として使えない(数量化の意義がここにある).
- 質的データ(名義尺度，順序尺度)の数量化はどう行うか？
- 質的データの原点はクロス表型データにある.
- クロス表は質的データ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(対応，関連を分析)できる.
- 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(対応分析という用語の出所).

4

対応分析法の本質

- 対応分析法（ベンゼクリ氏提唱）は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）.
- 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.
- 数量化法Ⅲ類（林知己夫氏提唱）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.

5

クロス表の“独立性の検定”（古典的な手法）

- 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）.
- クロス表の表側と表頭の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（$p_{ij} = p_{i+}p_{+j}$）. を考える. ⇒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（断定はできない，隔靴掻痒の考え方）.
- そもそも「2つの項目I, J間に関係」がありそうだから，あるいは「そのように意図」してデータを集めたはずである.

6

ピアソンのカイ二乗統計量の役割

- 対応分析では，ピアソンのカイ二乗統計量が重要な役割を果たす．
- 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（「独立でない」ことだけを知っても利点が少ない）．
- 対応分析法は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，固有値の大きさ（＝相関の情報に相当）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に測れることになる（ここで質的データの数量化となる）．

7

対応分析の仕組みの要約（テキストで確認）

- 二元のクロス表（型）を基本のデータ表とする（表19，17ﾍﾟｰｼﾞ）
- このデータ表からプロフィルを作る（表20，図2，18～20ﾍﾟｰｼﾞ）
  - 行のプロフィル（行の相対比率のデータ）⇒式（8），19ﾍﾟｰｼﾞ
  - 列のプロフィル（列の相対比率のデータ）⇒式（9），19ﾍﾟｰｼﾞ
- 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3，21ﾍﾟｰｼﾞ）
- 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）．

8

例題に用いる「質問」と「選択肢」の確認

- まず用いる質問文を挙げる．［例5］，14ﾍﾟｰｼﾞから

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

9

原データ表のイメージ

- 質問を「項目」，選ぶカテゴリーを「選択肢」という．
- 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原データ表がある（表15，15ﾍﾟｰｼﾞ，右の表）．
- この表からクロス表を生成する（表16，16ﾍﾟｰｼﾞ）．
- この場合，項目*I*の選択肢数は*m*=10，項目*J*のそれは*n*=3．
- この質的情報である選択肢に数量を付与することが数量化である．
- 選択肢を構成要素（単語，語句）と読み替えれば一般のテキスト型データとなる．

	項目*I*	項目*J*
ｻﾝﾌﾟﾙ番号	評価基準	レストラン名
1	味	コルシカ
2	量	ムガール
3	味	ロゴスキー
4	味	さとみ
5	量	バッハ
6	「工夫・ｻｰﾋﾞｽ」	コルシカ
7	「工夫・ｻｰﾋﾞｽ」	ムガール
8	味	コルシカ
9	「工夫・ｻｰﾋﾞｽ」	バッハ
10	味	コルシカ
11	量	ロゴスキー
12	味	「ラ・マレ」
13	「工夫・ｻｰﾋﾞｽ」	バッハ
14	量	ロゴスキー
15	味	さとみ
16	「工夫・ｻｰﾋﾞｽ」	「ラ・マレ」
17	味	「ラ・マレ」
18	「工夫・ｻｰﾋﾞｽ」	クラーク
19	味	さとみ
20	味	ムガール
…	～ 省略 ～	…
1283	味	さとみ
1284	量	ロゴスキー

10

クロス表の生成

- クロス表を作成する（表16，16ﾍﾟｰｼﾞ）．ここでは「行（表側）に項目I（レストラン）」を「列（表頭）に項目J（評価基準）」を充てた．（表19参照，17ﾍﾟｰｼﾞ）
- クロス表は多次元データである（⇒後述，図2，図3；20～21ﾍﾟｰｼﾞ）．
- 対応分析法では，行と列を入れ替えても（転置しても）解は同じである．

項目 J / 項目 I	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

訂正

ピアソンのカイ二乗統計量

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

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

（自由度$=m+n-2$のχ^2分布に近似して検定）

ここで，ピアソンのカイ二乗統計量は離散確率変数であるが（クロス表のマスは離散的），それを連続確率分布のχ2分布で近似して検定するということ．

参考：ピアソンの χ^2 統計量の一般表記

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

① 期待度数とは一つのモデルから得られる理論値と考える．ここでは「独立モデル」を想定したときの期待度数に相当する．
② 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）．
③ 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（モデルが正しければ，ゼロに近くなるような指標）．
④ 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．
⑤ ピアソンの χ^2 統計量では関連度は直接は測れない．
⑥ 対応分析法ではこれを固有値により測ろうとしている．

ピアソンのカイ二乗統計量による検定を行う

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

$\frac{95\times442}{1284} = 32.702 \Leftrightarrow$ 46に近いか？
$\frac{95\times302}{1284} = 22.344 \Leftrightarrow$ 7に近いか？
$\Rightarrow$ （順次，各セル内の実現度数が近ければχ^2は距離として小さくなる）

$$= \frac{\left(46 - \frac{95\times442}{1284}\right)^2}{\frac{95\times442}{1284}} + \frac{\left(7 - \frac{95\times302}{1284}\right)^2}{\frac{95\times302}{1284}} + \cdots + \frac{\left(38 - \frac{178\times302}{1284}\right)^2}{\frac{178\times302}{1284}} + \frac{\left(105 - \frac{178\times540}{1284}\right)^2}{\frac{178\times540}{1284}}$$

$$= 5.4070 + 10.5372 + \cdots + 0.3570 + 12.1351 \fallingdotseq 330.860$$

$\chi^2_{m+n-2} = \chi^2_{11}(0.05) = 19.675$
$\chi^2_{11}(0.01) = 24.725$
（有意水準 α =0.01および0.05の値と比較）

有意確率： $P\left(\chi^2_{m+n-2} > \chi^2\right) \fallingdotseq 0$

結論：独立ということは，ほとんど起こりえないようだ（有意確率からみても）．
（かなりの確度で）2つの項目間は無関係であるとはいえないが，関係があると断定はできない．

14

クロス表の式による一般表記

「(項目I)×(項目J)」のクロス表(2元クロス表，分割表)を以下の式で表す(式(1)，(2)，17ページ)．

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

項目Iと項目Jのクロス表という，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の選択肢をこれで表す
- 項目Iの選択肢数はm個，項目Jの選択肢数はn個
- ここの例では，m=10(のレストラン)，n=3(の評価基準)となる．

15

項目Iと項目Jのクロス表のイメージ(表19，17ページ)

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

(項目I)×(項目J)のクロス表

		表頭 項目 J						
		選択肢	1	2	⋯	j	⋯	n	行和
表側 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_{++}

16

行と列のプロフィルを作る

（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100%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

行プロフィル$\mathbf{N}_I=(q_{ij})$を求める（表21，22ページ）

評価基準 / 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）．

これを図にすると，23ページの図3のようになる．

列プロフィル$\mathbf{N}_J=(q_{ij}^*)$を求める（表22，22ページ）

評価基準 / 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．

同様に，これも多次元空間内の図と考えることができる．

19

データ行列を作る（なぜこの形か，その理由が重要）

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

$$\Downarrow$$

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

（x_{ij}を要素とする行列$\underset{m\times n}{\mathbf{X}}$を データ行列と考える）

テキスト，24ページ，式（10），訂正
※式(10),(11)は「偏差」の形

レストラン	評価基準 1．味	2．量	3．工夫・サービス
1．さとみ	0.8253	0.1519	0.6817
2．バッハ	0.9122	0.2614	0.5212
3．ムガール	0.6880	0.3027	0.6932
4．いりふね	0.2749	0.4257	0.9749
5．コルシカ	1.0757	0.2197	0.4045
6．クラーク	0.2339	1.0916	0.5140
7．ロゴス	0.4772	0.6928	0.5921
8．きくみ	0.1240	1.2559	0.4906
9．ラ・マレ	0.9573	0.2118	0.5175
10．かりや	0.3351	0.4402	0.9096

実際に求めると左の表になる．
↓
このようなデータ行列を用いる理由がある．

20

行列表記すると，…

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

$$\Downarrow$$

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

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

ここで，$\mathbf{P}_J^{1/2} = diag\left(\frac{1}{\sqrt{p_{+j}}}\right)$ という対角行列

ここの各行列の記法，意味についてはテキスト，18ページを参照

このデータ行列Xの共分散行列の固有値問題を解くこと，あるいはスペクトル分解を行うこと．

21

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；“双対性”がある!!!）．

※図2と図3を確認

22

行と列のプロフィルの関係をイメージすると，…

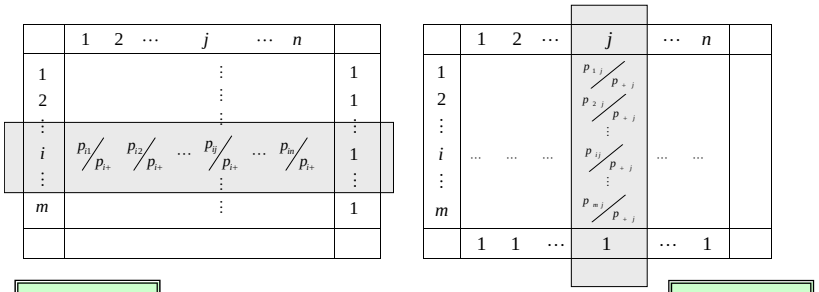


表21に相当

表22に相当

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

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

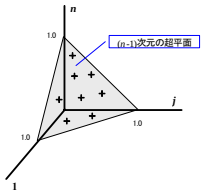
◆列和を 1 としたときの「列のプロフィル」で
（$m-1$）次元内に分布する n 個の点

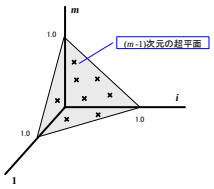
$$\mathbf{N}_J = \left\{ q^*_{ij} = \frac{p_{ij}}{p_{+j}} \middle| i \in I, j \in J \right\}$$

行のプロフィル　　列のプロフィル

23

成分スコアの布置をイメージで示すと，…

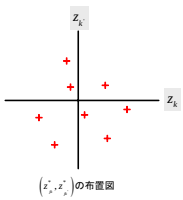


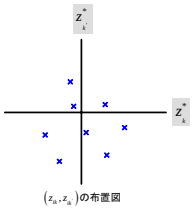


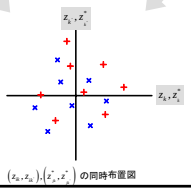
$\mathbf{N}_I = \left\{ q_{ij} = \frac{p_{ij}}{p_{i+}} \middle| i \in I, j \in J \right\}$ の分布

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

①重心座標系の空間内で次元縮約を行う
②データ表から得られた成分スコアを布置図とする（行成分スコア，列成分スコア）
③必要に応じて，行成分スコアと列成分スコアとの同時布置図とする







24

行プロフィル$\mathbf{N}_I=(q_{ij})$を再確認する（表21，22ページ）

評価基準 / 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，

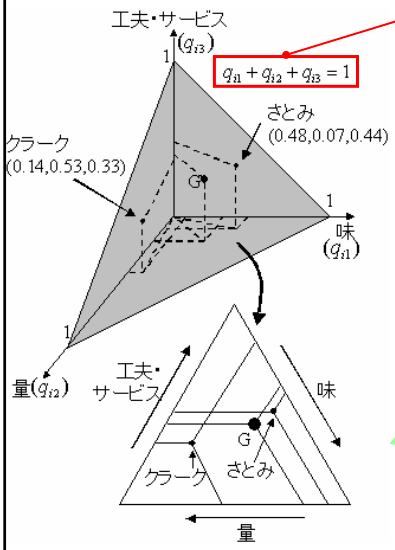
$\sum_{j=1}^{n} q_{ij}=1$（行和＝１）となっている．これを，$(n-1)=3-1=2$，つまり２次元の空間内に布置する 10 のレストランと考える．また列の平均ベクトル（平均比率）は重心に相当する（図 4 の中の G がそれに相当）．

これを図にすると，23ページの図4のようになる．

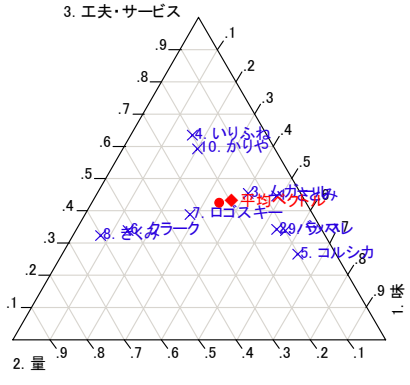
25

行プロフィルの意味を図で確認（図4，23ページ）

この制約（行和＝1）があるので，ここでは2次元になる



三角図



26

さらに，…

1）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（21,22ページ）の左側のパスを考える．

2）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）「行和＝1」という制約から，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．

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．

7）例えばここで，評価基準の3つの選択肢はm－1＝9（次元）の空間内に布置すると考える．

27

次元数の縮約を行うこと

- 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）
- 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．⇒多次元情報を少数次元で説明する．
- 主成分分析と同じようなこと（加重和を作る）が考えられるのか？
- そのためのデータの構造は（作り方は）？
  前にみたデータ表$\mathbf{X}=(x_{ij})$を扱えばよい，またこの形でないと不都合を生じる．
- 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）

※21ページ，図3の布置図イメージを確認

28

データ行列$\mathbf{X}=(x_{ij})$の分解（固有値問題他）

- 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）.
- 固有値，固有ベクトルあるいは特異値を求めること.
- 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.
- 固有値，固有ベクトル，と寄与率が情報縮約の程度を知る指標となる（あとで要約）.
- 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する.
- 固有ベクトルが加重平均の式の係数に相当する.
  （注：式（10），（11）に固有ベクトルを加重とする一次結合式）
- 成分スコア（数量化スコア，数量化得点）の算出.
- 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.

29

行側（項目Iの選択肢i）の成分スコアz_{ik}を求める

$$\underset{n\times 1}{\mathbf{l}_k}=\begin{pmatrix} l_{1k}\\ l_{2k}\\ \vdots\\ l_{jk}\\ \vdots\\ l_{nk}\end{pmatrix}\quad \begin{pmatrix} k=1,2,\cdots,K\\ K=\min\{m,n\}-1\end{pmatrix}\quad \begin{pmatrix}\text{第}k\text{固有値}\lambda_k\text{に対する}\\ \text{固有ベクトル}\end{pmatrix}$$

$$\underset{n\times K}{\mathbf{L}}=(\mathbf{l}_1,\mathbf{l}_2,\cdots,\mathbf{l}_k,\cdots,\mathbf{l}_K)\quad \begin{pmatrix}\text{固有ベクトルを列ベクトル}\\ \text{とする行列}\end{pmatrix}$$

$$(\text{ここで},\ K=\min\{m,n\}-1)$$

$$\underset{m\times K}{\mathbf{Z}}=\underset{m\times n}{\mathbf{X}}\ \underset{n\times K}{\mathbf{L}}=\mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{1/2}\underset{n\times K}{\mathbf{L}}=(z_{ik})\quad \begin{pmatrix} i\in I\\ K=\min\{m,n\}-1\end{pmatrix}$$

$$\Updownarrow$$

$$z_{ik}=\sum_{j=1}^{n}l_{jk}x_{ij}=\sum_{j=1}^{n}l_{jk}\left(\frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}\right)=\sum_{j=1}^{n}l_{jk}\left(\frac{q_{ij}}{\sqrt{p_{+j}}}\right)\quad \begin{pmatrix} i\in I\\ K=\min\{m,n\}-1\end{pmatrix}$$

30

行側（項目Iの選択肢i）の成分スコアz_iとは

$$z_{ik}=\sum_{j=1}^{n} l_{jk}x_{ij}=\sum_{j=1}^{n} l_{jk}\left(\frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}}\right)=\sum_{j=1}^{n} l_{jk}\left(\frac{q_{ij}}{\sqrt{p_{+j}}}\right)\quad\begin{pmatrix} i\in I \\ K=\min\{m,n\}-1\end{pmatrix}$$

- 要するに，元のデータ行列の要素の固有ベクトル要素を加重とする一次結合（⇒**加重平均**になっている）．
- 元のn次元（正確には（$n-1$）次元）の成分が合成された指標であること．
- 主成分分析における合成指標化に類似している．
- よって，固有値の大きさと寄与率を目安に，固有値の影響が何成分あたりまで及ぶかを考えて成分スコアを採用する．
- 列側（項目Jの選択肢j）の成分スコアは双対性から得られる．あるいは元のクロス表を転置して解くことでも同じ解となる（双対性がある）．⇒後述，数値例も参照

31

この例の固有値，寄与率，固有ベクトル

固有値と寄与率

固有値はクロス表の行・列の寸法から$K=\min\{m,n\}-1=\min\{10,3\}-1=2$個得られる．

成分k	固有値 λ_k	寄与率（%）	累積寄与率（%）
1	0.19766	76.71	76.71
2	0.06002	23.29	100.0

固有値ベクトル

固有ベクトルは以下となった．

$$\underset{3\times 2}{\mathbf{L}}=(\mathbf{l}_1,\mathbf{l}_2)=\begin{pmatrix}-0.691 & -0.423\\ 0.718 & -0.500\\ 0.088 & 0.756\end{pmatrix}$$

（$\underset{10\times 3}{\mathbf{X}}$から出発した場合に得られた固有ベクトル）

32

例題について確認

$$\underset{10\times 2}{\mathbf{Z}} = \underset{10\times 3}{\mathbf{X}}\ \underset{3\times 2}{\mathbf{L}} = \mathbf{P}_I^{-1}\mathbf{P}_{IJ}\mathbf{P}_J^{1/2}\underset{3\times 2}{\mathbf{L}}$$

$$= \begin{pmatrix} 0.8253 & 0.1519 & 0.6817 \\ 0.9122 & 0.2614 & 0.5212 \\ \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots \\ 0.9573 & 0.2118 & 0.5175 \\ 0.3351 & 0.4402 & 0.9096 \end{pmatrix} \times \begin{pmatrix} -0.691 & -0.423 \\ 0.718 & -0.500 \\ 0.088 & 0.756 \end{pmatrix}$$

- ここで左側の行列**X**は，前に作ってある．
- 右側の行列は固有ベクトルを列とする行列**L**になる．
- 加重和を作る．

$$= \begin{pmatrix} -0.401 & 0.091 \\ -0.397 & -0.122 \\ -0.197 & 0.082 \\ 0.202 & 0.408 \\ -0.550 & -0.259 \\ 0.667 & -0.256 \\ 0.220 & -0.100 \\ 0.859 & -0.309 \\ -0.464 & -0.119 \\ 0.165 & 0.326 \end{pmatrix} = \left(z_{ik}\right) \quad \begin{pmatrix} i \in I = \{1,2,\cdots,10\} \\ k = 1,2 \end{pmatrix}$$

ここでは，固有ベクトルの要素の符合の向きがテキストと逆になった．⇒固有ベクトルの符合は一意に決まらない．±を反転させればよい．
24ページ，表と比較せよ．

列側（項目Jの選択肢j）の成分スコアz_{jk}*も同様

$$x_{ij}^* = \frac{p_{ij}}{p_{j+}\sqrt{p_{i+}}} = \frac{q_{ij}^*}{\sqrt{p_{i+}}} \quad (i \in I, j \in J) \quad \Leftrightarrow \quad \underset{3\times 10}{\mathbf{X}^*} = \mathbf{P}_J^{-1}\mathbf{P}_{JI}\mathbf{P}_I^{1/2} = \left(x_{ij}^*\right)$$

$$\underset{3\times 2}{\mathbf{Z}^*} = \underset{3\times 10}{\mathbf{X}^*}\ \underset{10\times 2}{\mathbf{U}} = \mathbf{P}_J^{-1}\mathbf{P}_{JI}\mathbf{P}_I^{1/2}\underset{10\times 2}{\mathbf{U}} = \left(z_{jk}^*\right)$$

ここで，$\underset{10\times 2}{\mathbf{U}}$は$\underset{3\times 10}{\mathbf{X}^*}$の共分散行列から得られる固有ベクトル行列，$\mathbf{P}_{JI}$は$\mathbf{P}_{IJ}$の転置行列，

$\mathbf{P}_I^{1/2} = diag\left(\frac{1}{\sqrt{p_{i+}}}\right)$は対角行列

$$= \mathbf{P}_J^{-1}\mathbf{P}_{JI}\mathbf{P}_I^{1/2} \times \begin{pmatrix} -0.245 & 0.101 \\ -0.297 & -0.166 \\ -0.129 & 0.098 \\ \cdots & \cdots \\ \cdots & \cdots \\ \cdots & \cdots \\ \cdots & \cdots \\ \cdots & \cdots \\ -0.352 & -0.164 \\ 0.138 & 0.496 \end{pmatrix} \Rightarrow \quad Z^* = \left(z_{jk}^*\right) = \begin{pmatrix} -0.523 & -0.176 \\ 0.658 & -0.252 \\ 0.061 & 0.286 \end{pmatrix}$$

（固有ベクトルの行列）

例について数値を要約する

2項目への成分スコア

項目と選択肢		成分スコア 第1成分スコア z_{i1}	第2成分スコア 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}	成分 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

確認：デモを見る

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

元のデータ表$\mathbf{X}=(x_{ij})$と成分スコアの関係

レストラン	評価基準 1．味	2．量	3．工夫・サービス	z_{i1}	z_{i2}
1．さとみ	0.8253	0.1519	0.6817	−0.401	0.090
2．バッハ	0.9122	0.2614	0.5212	−0.397	−0.122
3．ムガール	0.6880	0.3027	0.6932	−0.197	0.082
4．いりふね	0.2749	0.4257	0.9749	0.201	0.408
5．コルシカ	1.0757	0.2197	0.4045	−0.550	−0.259
6．クラーク	0.2339	1.0916	0.5140	0.667	−0.256
7．ロゴス	0.4772	0.6928	0.5921	0.220	−0.101
8．きくみ	0.1240	1.2559	0.4906	0.859	−0.309
9．ラ・マレ	0.9573	0.2118	0.5175	−0.464	−0.120
10．かりや	0.3351	0.4402	0.9096	0.165	0.326
z^*_{j1}	−0.523	0.658	0.061		
z^*_{j2}	−0.176	−0.252	0.286		

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

（x_{ij}を要素とする行列$\underset{m\times n}{\mathbf{X}}$をデータ行列と考える）

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

成分スコアとクロス表$\mathbf{F}=(f_{ij})$の対応関係

		項　目　J						成分スコア									
		1	2	…	j	…	n	1	2	…	k	…	k'	…	K
項目 I	1	f_{11}	f_{12}	…	f_{1j}	…	f_{1n}	z_{11}	z_{12}	…	z_{1k}	…	$z_{1k'}$	…	z_{1K}
	2	f_{21}	f_{22}	…	f_{2j}	…	f_{2n}	z_{21}	z_{22}	…	z_{2k}	…	$z_{2k'}$	…	z_{2K}
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	i	f_{i1}	f_{i2}	…	f_{ij}	…	f_{in}	z_{i1}	z_{i2}	…	z_{ik}	…	$z_{ik'}$	…	z_{iK}
	⋮	⋮	⋮		⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	m	f_{m1}	f_{m2}	…	f_{mj}	…	f_{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^*_{jK}	…	z^*_{nK}								

37

主要な性質を調べる

◎以下で，とりあえず必要な対応分析法の特性，性質を概観する．

- 固有値と寄与率，累積寄与率
- 固有値とピアソンのカイ二乗統計量の関係（重要）
- 成分スコアの双対性
- 成分スコアの布置図，同時布置図

38

固有値と寄与率，累積寄与率（テキスト31，32ﾍﾟｰｼﾞ）

i) $0 \le \lambda_k \le 1 \quad (k = 1, 2, \cdots, K; K = \min\{m, n\} - 1)$

第k成分の固有値
（固有値は非負で1を越えない）

ii) $\begin{cases} v_k = \dfrac{\lambda_k}{\sum_{k=1}^{K} \lambda_k} \times 100(\%) & \begin{pmatrix} k = 1, 2, \cdots, K \\ K = \min\{m, n\} - 1 \end{pmatrix} \\ \sum_{k=1}^{t} v_k : \text{第}t\text{成分までの累積寄与率}(\%) \end{cases}$

第k成分の寄与率
と累積寄与率

iii) $\dfrac{\chi^2}{N} = \sum_{k=1}^{K} \lambda_k \quad (K = \min\{m, n\} - 1)$ $\begin{pmatrix} \text{固有値の総和} \\ \Updownarrow \\ \text{ピアソンのカイ二乗統計量} \end{pmatrix}$

- 固有値の総和とピアソンのカイ二乗統計量には上の関係がある．
- つまり，固有値がクロス表の項目間の関連性を測っていることになる．
- この関係は対応分析法を考えるうえできわめて重要である．

39

◆成分スコア，固有値の性質（再確認）

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，26ﾍﾟｰｼﾞ）

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．

⇒ 例では，$K=\min\{10, 3\}-1=2$となる．

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

$z^*_{jk} \quad (j \in J, k = 1, 2, \cdots, K)$ （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する．
- 算出式は，既に例で示した通り．

40

双対性について（27ページ）

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）

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

（行の成分スコアは列のそれのプロフィルの加重和）

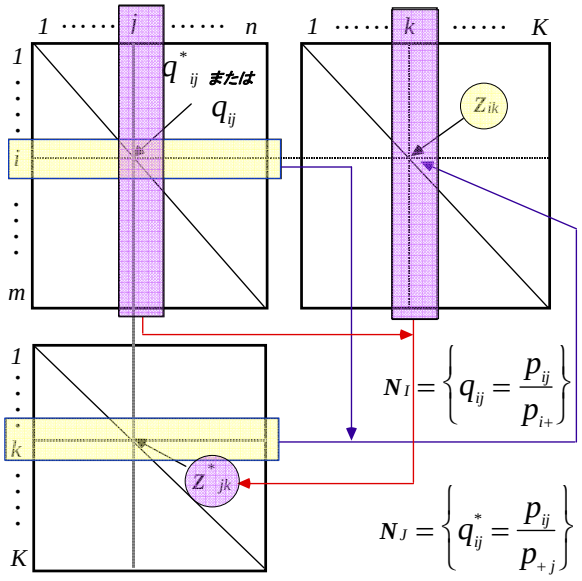
$$z_{jk}^{*}=\frac{1}{\sqrt{\lambda_k}}\sum_{i=1}^{m}\left(\frac{p_{ij}}{p_{+j}}\right)z_{ik}=\frac{1}{\sqrt{\lambda_k}}\sum_{i=1}^{m}q_{ij}^{*}z_{ik}\quad\left(j\in J,k=1,2,\cdots,K\right)$$ 式（20）

（列の成分スコアは行のそれのプロフィルの加重和）

※これら成分スコアの関係は図6と表25のように考える．

41

双対性の考え方（図6，26ページ）



42

確認：行プロフィル$\mathbf{N}_I=(q_{ij})$と成分スコアの関係

レストラン	評価基準 1. 味	2. 量	3. 工夫・サービス	z_{i1}	z_{i2}
1. さとみ	0.484	0.074	0.442	−0.401	0.090
2. バッハ	0.535	0.127	0.338	−0.397	−0.122
3. ムガール	0.404	0.147	0.450	−0.197	0.082
4. いりふね	0.161	0.206	0.632	0.201	0.408
5. コルシカ	0.631	0.107	0.262	−0.550	−0.259
6. クラーク	0.137	0.529	0.333	0.667	−0.256
7. ロゴス	0.280	0.336	0.384	0.220	−0.101
8. きくみ	0.073	0.609	0.318	0.859	−0.309
9. ラ・マレ	0.562	0.103	0.336	−0.464	−0.120
10. かりや	0.197	0.213	0.590	0.165	0.326
z^*_{j1}	−0.523	0.658	0.061		
z^*_{j2}	−0.176	−0.252	0.286		

$$z_{21}=\frac{1}{\sqrt{\lambda_1}}\sum_{j=1}^{3}\left(\frac{p_{2j}}{p_{2+}}\right)z^*_{j1}=\frac{1}{\sqrt{\lambda_1}}\sum_{j=1}^{3}q_{2j}z^*_{j1}=\frac{1}{\sqrt{0.19766}}\left(q_{21}z^*_{11}+q_{22}z^*_{21}+q_{23}z^*_{31}\right)$$

2つの要素の計算例

$$=\frac{1}{\sqrt{0.19766}}\left\{0.535\times(-0.523)+0.127\times0.658+0.338\times0.061\right\}\fallingdotseq-0.395$$

$$z_{62}=\frac{1}{\sqrt{\lambda_2}}\sum_{j=1}^{3}\left(\frac{p_{6j}}{p_{6+}}\right)z^*_{j2}=\frac{1}{\sqrt{\lambda_2}}\sum_{j=1}^{3}q_{6j}z^*_{j2}=\frac{1}{\sqrt{0.06002}}\left(q_{61}z^*_{12}+q_{62}z^*_{22}+q_{63}z^*_{32}\right)$$

$$=\frac{1}{\sqrt{0.06002}}\left\{0.137\times(-0.176)+0.529\times(-0.252)+0.333\times0.286\right\}\fallingdotseq-0.254$$

43

確認：列プロフィル$\mathbf{N}_J=(q_{ij}{}^*)$と成分スコアの関係

レストラン	評価基準 1. 味	2. 量	3. 工夫・サービス	z_{i1}	z_{i2}
1. さとみ	0.104	0.023	0.078	−0.401	0.090
2. バッハ	0.172	0.060	0.089	−0.397	−0.122
3. ムガール	0.100	0.053	0.091	−0.197	0.082
4. いりふね	0.057	0.106	0.181	0.201	0.408
5. コルシカ	0.174	0.043	0.059	−0.550	−0.259
6. クラーク	0.032	0.179	0.063	0.667	−0.256
7. ロゴス	0.079	0.139	0.089	0.220	−0.101
8. きくみ	0.018	0.222	0.065	0.859	−0.309
9. ラ・マレ	0.186	0.050	0.091	−0.464	−0.120
10. かりや	0.079	0.126	0.194	0.165	0.326
z^*_{j1}	−0.523	0.658	0.061		
z^*_{j2}	−0.176	−0.252	0.286		

1つの要素の計算例

$$z^*_{32}=\frac{1}{\sqrt{\lambda_2}}\sum_{i=1}^{10}\left(\frac{p_{i3}}{p_{+3}}\right)z_{i2}=\frac{1}{\sqrt{\lambda_2}}\sum_{i=1}^{10}q^*_{i3}z_{i2}$$

$$=\frac{1}{\sqrt{0.06002}}\left(q^*_{13}z_{12}+q^*_{23}z_{22}+\cdots+q^*_{93}z_{92}+q^*_{10,3}z_{10,2}\right)$$

$$=\frac{1}{\sqrt{0.06002}}\left\{0.078\times0.090+0.089\times(-0.122)+\cdots+0.091\times(-0.120)+0.194\times0.326\right\}=0.286$$

44

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）や散布図（布置図）</u>を描く．

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

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

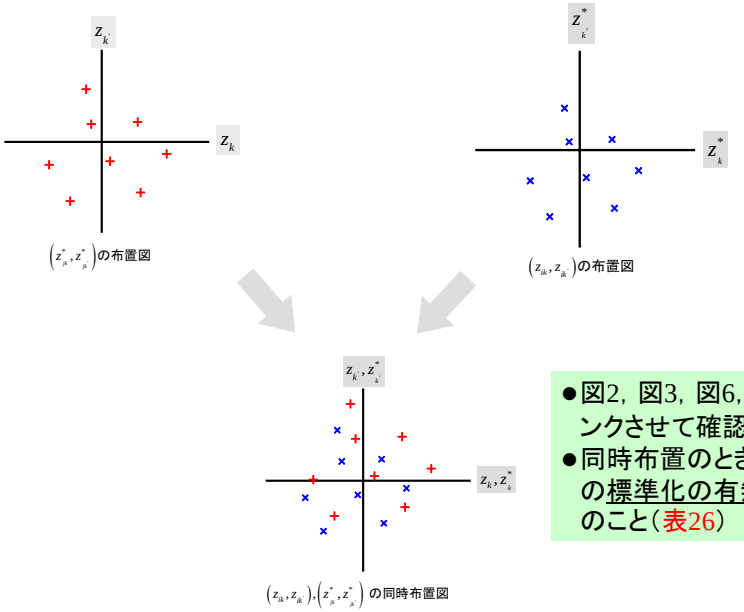
（行の選択肢への成分スコア） 式（21）

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

（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<u>標準化の有無</u>に注意のこと（表26）

例の数値と布置図，同時布置図を再確認

2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

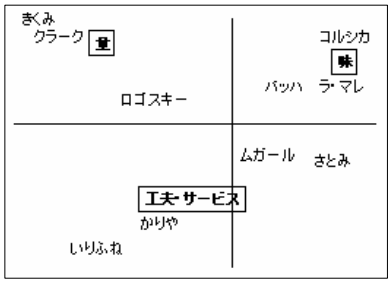
固有値と寄与率

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

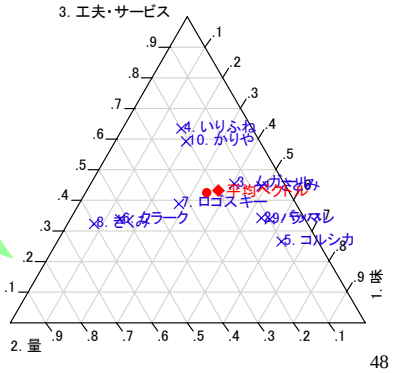
※成分スコアは表25，図6と対応させて確認のこと
①固有値の数は$K=\min\{m, n\}-1=2$となった．
②2成分に対する成分スコアが算出される．

※表23，24に相当
※表25との対応に注意

得られた同時布置図と三角図の比較（図4，図5）



三角図



※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）

（サンプル）×（構成要素）のデータ表の例

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

（サンプル）×（構成要素）のデータ表（一部を切り取り）．テキスト，37ページ

49

（質的変数）×（構成要素）のデータ表の例

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

（年齢区分）×（構成要素）のデータ表（一部を切り取り）．テキスト，38ページ

50

観察の目安(1)

- 成分スコアは，元の質的データ(クロス表の選択肢)のある種の加重平均である．つまり，多次元情報の要約(圧縮化情報)となっている．
- まず，固有値と寄与率を確認・観察する．
- 多くの場合大きい固有値，高い寄与率は期待できない．
  - 「(回答・サンプル)×(構成要素)」から出発⇒行列が疎のため固有値，寄与率は小さい．
  - 「(質的変数)×(構成要素)」から出発⇒質的変数の選択肢数にもよるが大抵は固有値．寄与率がほどほどの大きさとなる．
  - 一般に対応分析ではそう大きな固有値，寄与率とはならないことが多い．
- しかし，主要な特徴は，始めの方(大きい方)の固有値，成分に情報がある．

51

観察の目安(2)

- はずれ値的な構成要素の影響を受けやすい．
  - 例1:「とくにない」「なし」「ありません」「分かりません」など
  - 例2:出現頻度の少ない構成要素の存在
  - 初期の分析で閾値を大き目に設定し低い出現頻度の構成要素を切り捨てる⇒詳細分析の前の初動探査として
- はずれ値的な要素があるとき，見かけ上大きな固有値(高い寄与率)となることがある．
- これは成分スコアのはずれ値として現れるので，布置図や成分スコアの並びを観察して，必要に応じて一括除去する(削除辞書編集)．
  - 上の例1，例2などの削除辞書をあらかじめ作っておく
  - プロジェクトにより，これらを使い回しする(一度作れば済む)

52

観察の目安(3)

- 一般的な成分スコアの観察手順
  - 固有値の大きい始めの数成分(1成分～10成分あたりまで)
  - 成分軸の組合せを変えて，布置図を観察する
  - 布置図はその見ている2次元内のみの情報である
- 軸の解釈はあまり意味がない(誤解がある).
  - 因子分析とは異なること
  - 主成分分析に近い感覚で利用すること
  - 成分スコアの相対的な位置関係が重要である
  - 数理的な理屈からこうなる
- 選択肢型質問，つまり(WordMinerでいう)質的変数の作り方が重要である．これは調査方法論のスキルに関わることである(WordMinerの機能の問題ではない).

53

観察の目安(4)

- 同時布置図では，行成分スコアと列成分スコアとは同時に近いと括って考えることには注意する.
- これは双対性で述べたように，行・列の成分スコアの相互の関係を考えると明らかである(お互いの加重和となっていること).
- もちろん，行の成分スコアと列の成分スコアを同時にクラスタリングするなどを行ってはならない.
- 成分スコアの観察には，寄与度(絶対寄与度，相対寄与度)も参考情報とすること(テキスト，45ページあたり).
- その他，テキストの中に観察上の留意事項の記述があるので参考にするとよい(29～30ページなど).

54