WordMiner におけるクラスター化法

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

0. はじめに

この資料は，WordMiner におけるクラスタリングが，何を行っているかの概要を説明することを目的として書かれている．ここでは，WordMiner のクラスター化でどのような情報が得られるのか，その出力情報や諸量（統計量や指標，グラフィカル情報）をどう解釈するのか，といったことを，簡単なデータセットを用いて説明する．なお，数理的な細かい内容を述べることが目的でないので，より詳しい情報あるいは数理的な説明や解説は，うしろに挙げた参考文献を参照していただきたい．

なお，以下の記述で，主要な語句には英語を並記する場合がある．これは参考文献と併せて読む場合を配慮してのことである．対応分析法ほかで用いる元来の英語あるいは仏語の日本語訳が，国内では必ずしも統一されていないことがある．たとえば，“Correspondence Analysis”，“Analyse des Correspondances”がその典型例だが，ここではこれに“対応分析”という用語をあてる．これをコレスポンデンス分析とか，そのままコレスポンデンス・アナリシスとしている場合もみられる．また，“慣性”（inertia），“カイ二乗距離”（Chi-square distance, χ^2 －distance），“プロフィル”（profiles），成分スコア（coordinate）など，対応分析に特有の用語句もある．そのようなことで，ここでは書き手の判断で必要に応じて英語も記すようにした．

1. WordMiner におけるクラスター化法（概要）

分析対象を分類する方法はさまざまである．統計的データ解析をはじめ，データ・マイニングやテキスト・マイニングでも分類手法は重要なツールとなっているし，その手法の呼称もさまざまである．かつては，自動分類法（automatic classification）と呼び，またパターン認識などでは，教師なし分類（unsupervised classification）などと呼ばれた．これらをクラスタリング，クラスター化法あるいはクラスター分析（cluster analysis）ということもある（しかし，自動分類法とクラスター分析の源流はまったく異なる）．また，教師あり分類（supervised classification）のことを多変量解析などでは“判別分析”と呼んで，いわゆる自動分類法とは異なる位置づけで考えてきた．

自動分類法あるいはクラスター化法は，その算法（アルゴリズム）によって，いくつかに分類される．1 つは階層的分類法（hierarchical classification）であり，もう一つは非階層的分類法（non-hierarchical classification）である．これらはさらに細分され，さまざまな方法が誕生し，提案されている．もちろん，こうした自動分類法に関する研究論文，文献も無数にある．

ここでは，WordMiner で用いている分類方式の要点を述べる．WordMiner では，階層的分類法のうち**凝集型階層的分類法**（AHC：agglomerative hierarchical classification）の 1 つである**ウォード法**（ウォードの基準：Ward’s criterion によるウィシャート方式の算法：Wishart’s algorithm）による分類法），と，非階層的分類法の代表的な手法である分散最小化基準を用いる，いわゆる**分割型分類法**（partitioning-type classification）の 1 つである ***k*-平均法**（*k*-means method）を用いている．なお，なるべく大規模データの分類が可能なように，WordMiner ではこれらを混用するハイブリッド方式（あるいは**混合方式** mixed clustering approaches）の 1 つを採用している．WordMiner で用いているクラスター化手順を以下に簡単に要約する．以上の詳細は，文末に挙げた参考文献を参照されたい．

[WordMiner におけるクラスター化の方法]

まずここで，WordMiner で用いている計算手順の概要を述べる．細かいことはアルゴリズムの流れ図などを用いた説明が必要なので，ここはおよそこのようなことということを，なるべく言葉で記すことにする．

（手順1） 上述のように，凝集型階層的分類法と分割化型分類法とを併用する（混合方式）．

（手順2） 分類には，対応分析法によって得た“**成分スコア**”（principal coordinates または coordinates）を用いる．［注：重要なこととして，“用いる成分数の指定とクラスタ

一化の関係”があるが，これについては例も用いて後述する］

（手順3） はじめに，分類対象を凝集型階層的分類法の 1 つであるウォード法を用いて分類する．つまり原則としてすべての対象をこの方法で分類する．ただしこのとき“**相互最近隣の規則**”（RNN: reciprocal nearest neighbours rule）を用いて，近い位置にある点（似ている成分スコア）の圧縮化処理を行う．これで階層的分類の手間が圧縮される（計算量が減る）という性質を用いる．

（手順4） 階層的分類法で得た情報，たとえば階層分類の結合水準の変化や図化したデンドログラム（樹形図）などの観察で，適切なクラスター数（群の数）を決める．あるいは利用者が希望するクラスター数を指定する．これは利用者が恣意的に決める．

（手順5） これを“**初期分類**”として，この階層的分類法で得られたクラスター情報を用いて，次に分割型分類法（*k*-平均法）で，各クラスターの重心ベクトルをガイドとしてクラスターのチューニングを行う．いわゆる重心移動アルゴリズム（moving center algorithms）による再配置法でクラスター内の各点の移動・調整（consolidation, refinement）を行う．

（手順6） 最終的に得られたクラスター化情報を要約表として出力する．出力の統計量の意味・解釈については後ろの分析例を参照．

（注1） 相互最近隣の規則”による圧縮とは，簡単にいえば，ある点（成分スコア）からみて一番近く（最近隣にあって），相手の点からみてももっとも近い（最近隣にある），つまり“相互に最近隣”（mutually nearest neighbour）にある点は，より近い関係にある（よく似ている）とみなせるので，こうした点（成分スコア）を先に集めることで階層化の作業量を低減させる方法．類似の方法に，nearest-neighbor chain アルゴリズムという方式もある．

（注2） クラスター化の段階では，対応分析で得た“**成分スコア**”（principal coordinate あるいは単に coordinate）を用いるが，これは同時に“カイ二乗距離”（Chi-square distance）を用いた分類を行うことに相当する（プロフィル間のカイ二乗距離）．こうすることで，階層分類の結合水準は，クラスター内変動（クラスター内分散）に比例しこれは（ピアソンの）カイ二乗統計量の分解・併合を利用することにも相当する．これらについては後述の分析例を参照のこと．

（注3） “**クラスター数**”をいくつとするか，最適クラスター数をどう決めるかという問題への解答は見つかっていない．多数の研究があるが，多くは経験則的であるか，あるいは特定の構造の検出に向いた方法であって，一般化されたクラスター数決定の方法は「ない」といってよい．この事情は，ヒストグラムの級数の決め方や，いわゆる最適層別問題に通底することである．

2. 分類対象として扱うデータとデータ表

WordMiner におけるクラスター化では，対応分析法で得た“**成分スコア**”（coordinate）を対象に分類操作を行う．対応分析法で扱うデータ表の形式は，大別して「（サンプル）×（構成要素変数）」，「（構成要素変数）×（質的変数）」がある．このことから，分類対象は，サンプル（個体，ケース），構成要素（用語，語句など単語群），質的変数（調査の質問・選択肢，人口統計学的変数，クラスター化で得たクラスター変数など）をそのときどきの状況に応じて任意に扱うことができる．また一般に，扱うデータ表の寸法は，行・列ともに非常に大きく，またセル内の度数が非常に少ない疎な行列となることが多い．

成分スコアを用いることから，利用時に指定する“成分数”とクラスター化の関係を知っておくことが重要である．また，対応分析で得た成分スコアを用いるクラスター化は，元のデータ表の“**カイ二乗距離**”（Chi-square distance）を用いることに同じである．ここでいくつかの重要な性質があるが，これらについては分析例で説明する．

3. 簡単な例によるクラスター化手順の説明

ここでは，電卓や筆算でも追跡確認できるような簡単な例を用いてクラスター化の手順を説明する．数理的な説明は参考文献に譲って，ここではおもに出力結果として得られる情報の読み方，解釈の方法について述べる．また，WordMiner が出力した情報だけでは説明に不足する場合は，補足の情報を加えるようにした．

3.1 対応分析法のためのデータセットとここでの目標

3.1.1 準備

対応分析法では，出発行列として“**二元データ表**”（two-way table）を扱う．ここでいま**寸法が（$m\times n$）の二元のデータ表**（m 行，n 列のデータ表）を以下の式（行列）で表す．二元のデータ表とは，原則として，表の各セル内の度数が非負の数値であって，また行あるいは列の比率のパターンを考えることが意味あるようなデータ表である．たとえばもっとも単純な例として“**クロス表**”（cross-classified table）がある．

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

ここで，f_{ij} は二元データ表の$(i,\ j)$セル内の度数である（よって非負の値）．また I と Jは，それぞれ行と列の項目とその要素の集合を表わし以下のように書いておく．つまりクロス表であれば 2 つの**質問項目** I と Jと，それぞれの**選択肢**（カテゴリー，オプション）に相当する．

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

表 1 （項目 I×項目 J）の二元データ表 $\underset{m\times n}{\mathbf{F}}=\left(f_{ij}\right)$

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

この二元データ表について，以下の式を用意する．

二元データ表の行和（項目 I の周辺度数）：$f_{i+}=\dfrac{\sum_{j=1}^{n} f_{ij}}{N}\quad (i=1,2,\cdots.m)$ (3)

二元データ表の列和（項目 J の周辺度数）：$f_{+j}=\dfrac{\sum_{i=1}^{m} f_{ij}}{N}\quad (j=1,2,\cdots.n)$ (4)

二元データ表の総和（総サンプル数）：$f_{++}=\sum_{i=1}^{m}\sum_{j=1}^{n}f_{ij}=\sum_{i=1}^{m}f_{i+}=\sum_{j=1}^{n}f_{+j}=N$　(5)

以上の準備のもとに，説明用の 1 組のミニチュア・データセットを用意する（架空の例）．ある調査で，2 つの質問 I と質問 J について，質問文と回答選択肢が以下のようであるとする．このデータの分析目的は，ある回答者がリストにあるレストランからどれか 1 つ選ぶときに，どのような選択基準（評価基準）で選ぶか，その関連を調べたいという課題を想定している．

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

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

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

1. 工夫・サービス　2. 味　3. 量

この 2 つの質問に対して，回答者がそれぞれ 1 つだけ選択肢を選ぶものとする．このときの回答者数を N 人（＝1,284 人）として，この回答者の回答分布は表 1 のように，寸法が（N 人）×（2 項目）のデータ表として集められたとしよう（表 2）．なおここでは DK（Don’t Know：わからない）や NA（No Answer:無回答）などはなかったものとする（あっても分析は可能だが，説明を簡潔にするために「なし」としただけ）．

表 2　調査データの例［レストランの評価］

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

表 3　（項目 I）×（項目 J）の2元クロス表　$\mathbf{F}_{m\times n}=\left(f_{ij}\right)$

項目 I ＼ 項目 J		評価基準			行　和
		工夫・サービス	味	量	
レストラン名	いりふね	98	25	32	155
	かりや	105	35	38	178
	きくみ	35	8	67	110
	さとみ	42	46	7	95
	クラーク	34	14	54	102
	コルシカ	32	77	13	122
	バッハ	48	76	18	142
	ムガール	49	44	16	109
	ラ・マレ	49	82	15	146
	ロゴスキー	48	35	42	125
列　和		540	442	302	1284

ここで表 2 から，質問 I（レストラン）と質問 J（評価基準）の**クロス表**を作成したところ，表 3 のようになったという．このクロス表を例とし，またここでは二元データ表の行の要素，つまりレストランをクラスター化で分類することを考える．なお，二元データ表の列の要素，つまり評価基準についてもクラスター化を考えることができるが，これはすべて「行」（レストラン）を「列」（評価基準）と読み替えて考えればよい．

また WordMiner では，「**（サンプル）×（構成要素変数）**」，「**（構成要素変数）×（質的変数）**または「**（構成要素変数）×（クラスター変数）**」と表記しているが，いずれも“二元のデータ表”を扱っていることには変わりがない．通常はサンプル数や構成要素数はかなり大きい．つまり，データ表の寸法が大きいので，ここで使う例のように，すべての成分数（そして成分スコア）を用いることはない．しかし，原理･仕組みはここで述べる簡単な例とまったく同じに考えてよい．

このクロス表から，行和を 100（または 1）とそろえた比率の表を作る．これを“**行のプロフィル**”（row profile）という．同じようにして，列和を 100（または 1）とする比率の表を作る．これを“**列のプロフィル**”（column profile）という．プロフィルとは，行または列の大きさをそろえて相対的にパターンを比べる手続きである．たとえば，行のプロフィルは，各レストランが評価基準に対してどのようなパターン（回答比率）となる傾向があるかを知る，というように使う．対応分析はこのプロフィルの関係を，行と列の双対的関係として調べることでもある（注：表のセル内の元の頻度の大小を比べているわけではない）．

（注4）　“**プロフィル**”（profile）とは，クロス表の行または列の**相対度数（相対確率）**のことである．対応分析ではこれをプロフィルを名付けている．次の式（6），式（7）にあるように，（相対）確率 $p_{ij,}\ q_{ij}$ として扱うこともある（どちらもプロフィル）．

プロフィルを式で下のように表す．

$$\text{行のプロフィル：}\ \underset{m\times n}{\mathbf{F}_I} = \left\{ p_{ij} = \frac{f_{ij}}{f_{i+}} \,\middle|\, i \in I, j \in J \right\} \tag{6}$$

$$\text{列のプロフィル：}\ \underset{m\times n}{\mathbf{F}_J} = \left\{ q_{ij} = \frac{f_{ij}}{f_{+j}} \,\middle|\, i \in I, j \in J \right\} \tag{7}$$

ここで，表 4 の**列和**は行の要素（ここでは 10 のレストラン）の 3 つの列要素（評価基準）の**平均比率**（行の平均プロフィル；つまり**平均ベクトル**あるいは**重心**）である．同じように，表 5 の**行和**は 3 つの列要素（評価基準）の平均ベクトル（重心）になっている．この見方があとの説明で重要になる．

このクロス表から次の式（8）にしたがい“**ピアソンのカイ二乗統計量**” χ_p^2（Chi-square statistic）を求める．これはいわゆる“クロス表の 2 つの項目の独立性”の検定を行う統計量として知られている．

$$\chi_p^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\left(f_{ij} - \dfrac{f_{i+} f_{+j}}{N} \right)^2}{\dfrac{f_{i+} f_{+j}}{N}} \left(= \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\left(\text{実現度数} - \text{期待度数} \right)^2}{\text{期待度数}} \right) \tag{8}$$

これを表 3 のクロス表について求めると次の値になる．

$$\chi_p^2 = 330.860 \tag{9}$$

対応分析法ではこのピアソンのカイ二乗統計量が重要な役割をはたすので，上にこれを求めておく．

3.1.2 分析の方針

ここで，これから行う分析，とくにクラスター化の内容を以下に要約しておこう．

① まず，**分類対象**をここで取り上げた 10 のレストランとする（行の要素の分類）．つまり，どのレストランへの回答傾向（プロフィル）が似ているかを分類で調べる．
② このデータ表から，直接なんらかの方法でレストランを分類することも可能だが，ここでは表 2 のクロス表に**対応分析**を適用し，得られた“**成分スコア**”を用いてクラスター化を行う．［注：成分スコアを用いることは，プロフィルのカイ二乗距離を用いることに関係］
③ データ表の寸法は行数：$m=10$（レストラン），列数：$n=3$（評価基準）である．よって，対応分析の性質から，ここで得られる成分スコアは 2 成分まで（得られる固有値は 2 個まで）となる．［固有値の個数：$K=\min\{m,n\}-1=3-1=2$］．
④ クロス表の行のプロフィル間の**カイ二乗距離**にもとづく凝集型階層的分類法と相互最近隣の規則を用いて，レストランの分類を考えることができる．
⑤ ここでは，クロス表の対応分析でえた成分スコアを用いた**平方ユークリッド距離**による分類（ウォード法）を用いる．しかし，これと④の操作とは実質的には同等であることが知られている（後述する）．
⑥ 通常は**クラスター数**を与えてクラスター化を行う．ここでは，クラスター化の履歴を調べるために，あえて 2 群から 10 群までを指定する．なお 1 群とはすべてのレストランを 1 群とみなすということ，一方，10 群とは個々のレストランを 1 つの群とすること，これは分類を行わない場合に相当する．
⑦ 階層的分類における“**結合水準の値**”と“**カイ二乗統計量**”，それと対応分析で得られる“**固有値**”のそれぞれの間に，ある関係がある．これらを調べる．
⑧ また，得られたクラスターと各種の統計量（クラスター内変動，クラスター間変動，総変動など）をどのように評価，解釈するかを調べる．

表 4 行のプロフィル（レストランの比較）［$\mathbf{F}_I$ の分布］

項目 I ＼ 項目 J		評価基準			行　和
		工夫・サービス	味	量	
レストラン名	いりふね	63.2	16.1	20.6	100.0
	かりや	59.0	19.7	21.3	100.0
	きくみ	31.8	7.3	60.9	100.0
	さとみ	44.2	48.4	7.4	100.0
	クラーク	33.3	13.7	52.9	100.0
	コルシカ	26.2	63.1	10.7	100.0
	バッハ	33.8	53.5	12.7	100.0
	ムガール	45.0	40.4	14.7	100.0
	ラ・マレ	33.6	56.2	10.3	100.0
	ロゴスキー	38.4	28.0	33.6	100.0
列の相対度数 $p_{+j}\times100$		**42.1**	**34.4**	**23.5**	
（列の重心）		$p_{+j}\,(j=1,2,3)$　（*）ここは $p_{+j}\times100$（%）			

表 5　列のプロフィル（評価基準の比較）[$\mathbf{F}_J$　の分布]

項目 I ＼ 項目 J		評価基準			行の相対度数（$p_{i+}\times 100$）
		工夫・サービス	味	量	
レストラン名	いりふね	18.1	5.7	10.6	**12.1**
	かりや	19.4	7.9	12.6	**13.9**
	きくみ	6.5	1.8	22.2	**8.6**
	さとみ	7.8	10.4	2.3	**7.4**
	クラーク	6.3	3.2	17.9	**7.9**
	コルシカ	5.9	17.4	4.3	**9.5**
	バッハ	8.9	17.2	6.0	**11.1**
	ムガール	9.1	10.0	5.3	**8.5**
	ラ・マレ	9.1	18.6	5.0	**11.4**
	ロゴスキー	8.9	7.9	13.9	**9.7**
列　和		100.0	100.0	100.0	
行の重心		$p_{i+}\ (i=1,2,\cdots,10)$　（*）ここは $p_{i+}\times 100$（%）			

4. 例による分析結果と内容の説明

4.1 観察（その 1）　対応分析の実行と結果の確認

表 1 のクロス表に対応分析を適用して得られる基本情報を，以下に順に示す．

4. 1. 1 固有値と寄与率，累積寄与率

まず，はじめに固有値，寄与率ほかを調べる．

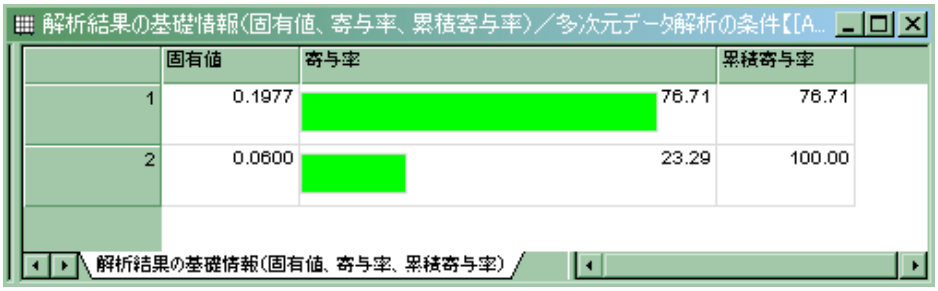


図 1　固有値，寄与率，累積寄与率

表 6　図 1 の情報の整理

α	λ_α	固有値	寄与率	累積寄与率 $\sum_\alpha \nu_\alpha$
第 1 固有値	λ_1	0.1977	76.71	76.71
第 2 固有値	λ_2	0.0600	23.29	100.00
	和	0.2577	100.00	—

ここで，ある確認を行う．対応分析の重要な性質の 1 つとして，“**固有値の和**”（総変動，つまり変動の総量であり情報の総量）と，上でクロス表から求めた“**ピアソンのカイ二乗統計量**”との間に次の関係がある．以後の分析と解釈で重要な性質なので，まずここで数値例として示す（確認する）．

［性質 1］固有値の和と総変動の関係

$$\sum_{\alpha=1}^{K}\lambda_\alpha=\frac{\chi_p^2}{N}\quad\left(\text{ここで，}\ K=\min\{m,n\}-1\right)\tag{10}$$

上の関係がなりたつ．これを例について確かめると以下のようになる．

$$\sum_{\alpha=1}^{K}\lambda_\alpha=\lambda_1+\lambda_2=0.2577\quad\Leftrightarrow\quad\frac{\chi_p^2}{N}=\frac{330.860}{1284}=0.257679\cdots\doteqdot 0.2577\tag{11}$$

つまり，

［固有値の和］
＝［（用いたクロス表のピアソンのカイ二乗統計量）÷（クロス表の総和）］(12)

の関係がある．

対応分析は，このカイ二乗統計量を**総変動**として，これを固有値で成分ごとにどう分けるか（**分解**するか）を行っていることになる．この点で主成分分析に類似している．

（注5） **総変動**：変動を対応分析では“inertia”（直訳すると“**慣性**”）という．総変動は“total inertia”となる．多変量解析などでいう**全分散**（total variance）に相当する．

（注6） **固有値と寄与率**：対応分析で得られる固有値 $\lambda_\alpha\left(\alpha=1,2,\cdots,K;K=\min\{m.n\}-1\right)$ から，以下の関係と寄与率が得られる．固有値の値は非負で 1 を越えることはない（つまり，$0\le\lambda_\alpha\le 1\ \left(\alpha=1,2,\cdots,K;K=\min\{m,n\}-1\right)$）．また**固有値の個数**は元の解析対象とした二元データ表（クロス表）の行と列の寸法の小さい方から 1 を引いた個数（$K=\min\{m,n\}-1$）となる（つまり，成分スコアの分布は，この次元数内の空間に入るということ）．また**寄与率**は以下の式となる．

$$\text{寄与率：}\ \nu_\alpha=\frac{\lambda_\alpha}{\sum_{\alpha=1}^{K}\lambda_\alpha}\times 100(\%)\quad\left(\begin{matrix}\alpha=1,2,\cdots,K\\ K=\min\{m,n\}-1\end{matrix}\right)\tag{13}$$

（第 α 成分の寄与率の式）

この ν_α を，α について累積すると**累積寄与率**となる（例：図 1，表 6 で確認）．

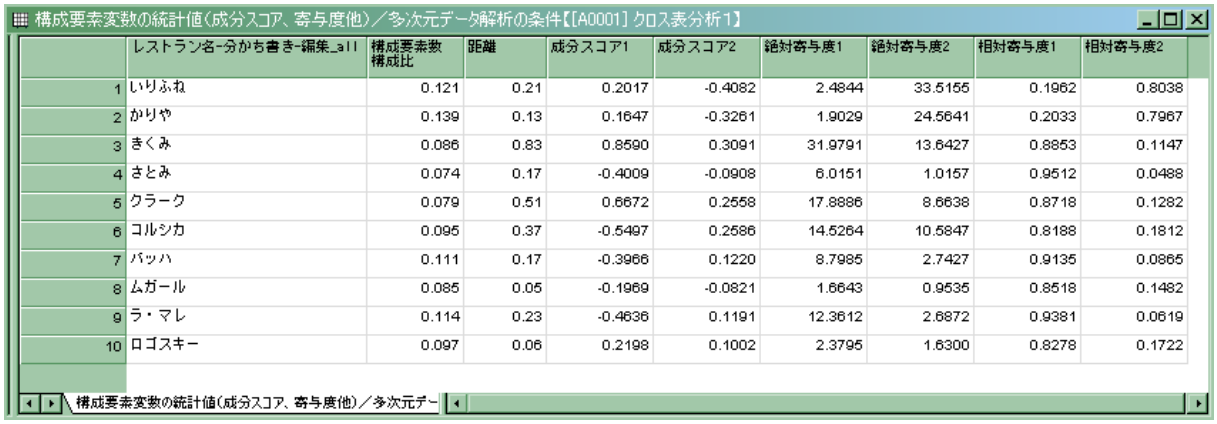
（注7） カイ二乗統計量に関連するクロス表の“**連関性の測度**”の 1 つとして，次の“**平均平方コンティンジェンシー係数**”（mean square contingency coefficient：平均二乗関連係数）がある．式（13）は，この係数に同じであることがわかる．

$$\phi^2=\frac{\chi_p^2}{N}$$

4．1．2　レストラン（行の要素）に対する成分スコアほか

つぎに成分スコアほかを要約する表が WordMiner の出力の図 2 である．ここで「構成要素数構成比」とは，ここでは表 4 の**行のプロフィル**（**周辺度数の割合**：p_{i+}）に相当する（対応分析では，これを“**行の質量**”（mass）ということがある）．また「距離」は，ここでは“**カイ二乗距離の二乗**”を示している．また，成分スコアは対応分析で得た値でプロフィルのあ

る種の加重平均と思えばよい（これについては，うしろに例を示した）．
ここで，**行のプロフィル**とカイ二乗距離，カイ二乗統計量，固有値の和の間にある性質があるのでこれを確認しておこう．（注：ここで，列のプロフィルと読み替えても同じことがなり立つ）．



構成要素変数の統計値(成分スコア、寄与度他)／多次元データ解析の条件【[A0001] クロス表分析1】

	レストラン名-分かち書き-編集_all	構成要素数構成比	距離	成分スコア1	成分スコア2	絶対寄与度1	絶対寄与度2	相対寄与度1	相対寄与度2
1	いりふね	0.121	0.21	0.2017	-0.4082	2.4844	33.5155	0.1962	0.8038
2	かりや	0.139	0.13	0.1647	-0.3261	1.9029	24.5641	0.2033	0.7967
3	きくみ	0.086	0.83	0.8590	0.3091	31.9791	13.6427	0.8853	0.1147
4	さとみ	0.074	0.17	-0.4009	-0.0908	6.0151	1.0157	0.9512	0.0488
5	クラーク	0.079	0.51	0.6672	0.2558	17.8886	8.6638	0.8718	0.1282
6	コルシカ	0.095	0.37	-0.5497	0.2586	14.5264	10.5847	0.8188	0.1812
7	バッハ	0.111	0.17	-0.3966	0.1220	8.7985	2.7427	0.9135	0.0865
8	ムガール	0.085	0.05	-0.1969	-0.0821	1.6643	0.9535	0.8518	0.1482
9	ラ・マレ	0.114	0.23	-0.4636	0.1191	12.3612	2.6872	0.9381	0.0619
10	ロゴスキー	0.097	0.06	0.2198	0.1002	2.3795	1.6300	0.8278	0.1722

構成要素変数の統計値(成分スコア、寄与度他)／多次元デー

図 2　成分スコア他の情報

まず図 2 の出力情報から，構成要素構成比，距離の部分を取り出し，さらに必要な情報を別の表に整理し以下の内容を確認する．

確認 1:
距離の確認：ここでは“**カイ二乗距離の二乗**”を示している．これの誘導を確認する．つまり，行の各要素（レストラン）の行のプロフィル**と列和**（つまり**行の平均プロフィル**，列の周辺度数の割合：$p_{+j}\,(j=1,2,3)$）との距離を調べること．

確認 2:
この距離とピアソンのカイ二乗統計量あるいは固有値の和との関係，つまり次の性質があることを調べること．

[性質 2]
ここで，総変動（つまり，固有値の和）は以下のように表せる．

$$\frac{\chi_p^2}{N}=\sum_{i=1}^{m}(\text{クロス表の第}i\text{行の比率})\times\left[\text{第}i\text{行プロフィルと行の平均プロフィルとの}\chi^2\text{ 距離}\right] \quad (14)$$

これを順に調べる．まず，行の要素（レストラン）のプロフィルとクロス表の**列の周辺度数の割合**（つまり行の平均プロフィル：p_{+j}）とから**カイ二乗距離**を求める操作を調べるために必要な作業表をいくつか作る（表 7，表 8）．このとき，数値の正確な確認のため図 2 の出力情報よりも桁数を増やしてある．
また表 7 は，レストラン「いりふね」を 1 つの例としてこれの計算表を作ったものである．この例では，“**カイ二乗距離の二乗**”**の和**は以下のようになる．

$$\frac{(0.6323-0.4206)^2}{0.4206}+\frac{(0.1613-0.3442)^2}{0.3442}+\frac{(0.2065-0.2352)^2}{0.2352} \quad (15)$$
$$=0.1066+0.0972+0.0035=0.2073\fallingdotseq 0.21$$

こうして得た「0.21」が図 2 の表内の「いりふね」の「距離」に欄の数値に相当する．よって，“**カイ二乗距離**”は上の（正の）平方根をとって次のようになる．

$$\sqrt{\frac{(0.6323-0.4206)^2}{0.4206}+\frac{(0.1613-0.3442)^2}{0.3442}+\frac{(0.2065-0.2352)^2}{0.2352}}=\sqrt{0.2073}\fallingdotseq 0.455 \quad (16)$$

表 7　例:「いりふね」についてカイ二乗距離の二乗を算出

	行のプロフィル（行の比率パターン）			和チェック
	工夫・サービス	味	量	
(a) いりふねのプロフィル	0.6323	0.1613	0.2065	1.000
(b) 行の平均プロフィル [表 4 の周辺確率： p_{+j}]	0.4206 （0.420560748）	0.3442 （0.34423676）	0.2352 （0.235202492）	1.000 1.000
$(c)=[(a)-(b)]^2/(b)$	0.1066 （0.106604126）	0.0972 （0.097217561）	0.0035 （0.003502654）	**0.2073≒0.21** （0.207324342）

ここで，いわゆる単純な“**ユークリッド距離**”との違いを見ておこう．「いりふね」の比率の行ベクトルと列和の周辺確率つまり行の平均プロフィル（重心）（ p_{+j} ）とのユークリッド距離であるから下のようになる．上と比べると値が小さくなるという特徴がある．つまり，ここのカイ二乗距離とは， p_{+j} を加重とする“**重み付き距離**”の 1 つである．ここでは行の要素について示したがこれをすべて列の要素の読み替えても同じ関係がなり立つ（列の平均プロフィル $p_{i+}\left(i=1,2,\cdots,10\right)$ を加重とする重み付き距離とする）．

＜単純なユークリッド距離の場合＞

$$\sqrt{(0.6323-0.4206)^2+(0.1613-0.3442)^2+(0.2065-0.2352)^2}=\sqrt{0.791232}\fallingdotseq 0.2813$$

表 8　カイ二乗距離の算出ほか（桁数を増やしてリチェック:7 桁で確認）

レストラン名（ i ）	カイ二乗距離の要素			構成要素数 構成比	距離	$\frac{\chi_p^2}{N}=\sum_\alpha \lambda_\alpha$ の確認
	工夫・サービス	味	量	①列の平均プロフィル（行の質量）（ p_{i+} ）	②カイ二乗距離の二乗	③＝①×②
いりふね	0.1066041	0.0972176	0.0035027	0.1207165	0.2073243	0.0250275
かりや	0.0681846	0.0633187	0.0020025	0.1386293	0.1335058	0.0185078
きくみ	0.0249137	0.2141904	0.5943787	0.0856698	0.8334828	0.0714043
さとみ	0.0011031	0.0569077	0.1108962	0.0739875	0.1689070	0.0124970
クラーク	0.0181054	0.1243993	0.3679900	0.0794393	0.5104948	0.0405533
コルシカ	0.0595549	0.2390521	0.0703164	0.0950156	0.3689235	0.0350535
バッハ	0.0162076	0.1059357	0.0499616	0.1105919	0.1721049	0.0190334
ムガール	0.0019913	0.0102716	0.0332267	0.0848910	0.0454897	0.0038617
ラ・マレ	0.0171636	0.1372508	0.0746459	0.1137072	0.2290603	0.0260458
ロゴスキー	0.0031783	0.0119870	0.0431974	0.0973520	0.0583627	0.0056817
						↓（和） **0.2576660** （固有値の和）

他の行の要素（レストラン）についても同じようにカイ二乗距離の二乗を求めて表 8 のように要約する．さらに各レストランについて求めたすべての行について和を求める．表の丸数字で示すと，①欄が図 2 あるいは表 5 の行の周辺確率 p_{i+}（列の平均プロフィル）であり，②欄が各レストランのカイ二乗距離の二乗，それらの積が③欄である．ここで③欄の和を求めると「0.2576660（≒0.2577)」となり，これは式（10）の固有値の和（そして総変動）に相当する．これで上の[性質 2]の式（14）を数値例で確かめたことになる．

4. 1. 3 成分スコアの布置図の観察

ここでは，レストランの**成分スコア布置図**と，これと（列側要素の）3 つの評価基準に対する成分スコアとの**同時布置図**を示した．これらの観察から，各レストランと各評価基準の関連（対応）が読みとれる．

ここでは，レストランの分類を行うので，各レストランがどのような布置関係にあるかを覚えておこう．のちにクラスター化過程を確認する際にもこれを用いることにする．なおこの例では，最大次元数（成分数）が 2 であるから，データ表の全情報がこの 2 次元空間内に布置されていることに注意しよう（図 1，表 6 から，2 つの成分で寄与率が 100%）．一般には，かなり寸法の大きいデータ表を扱うので，大きな固有値（高い寄与率）とはならず，布置図に描ける情報の量は限られる（それを寄与率が示している）．

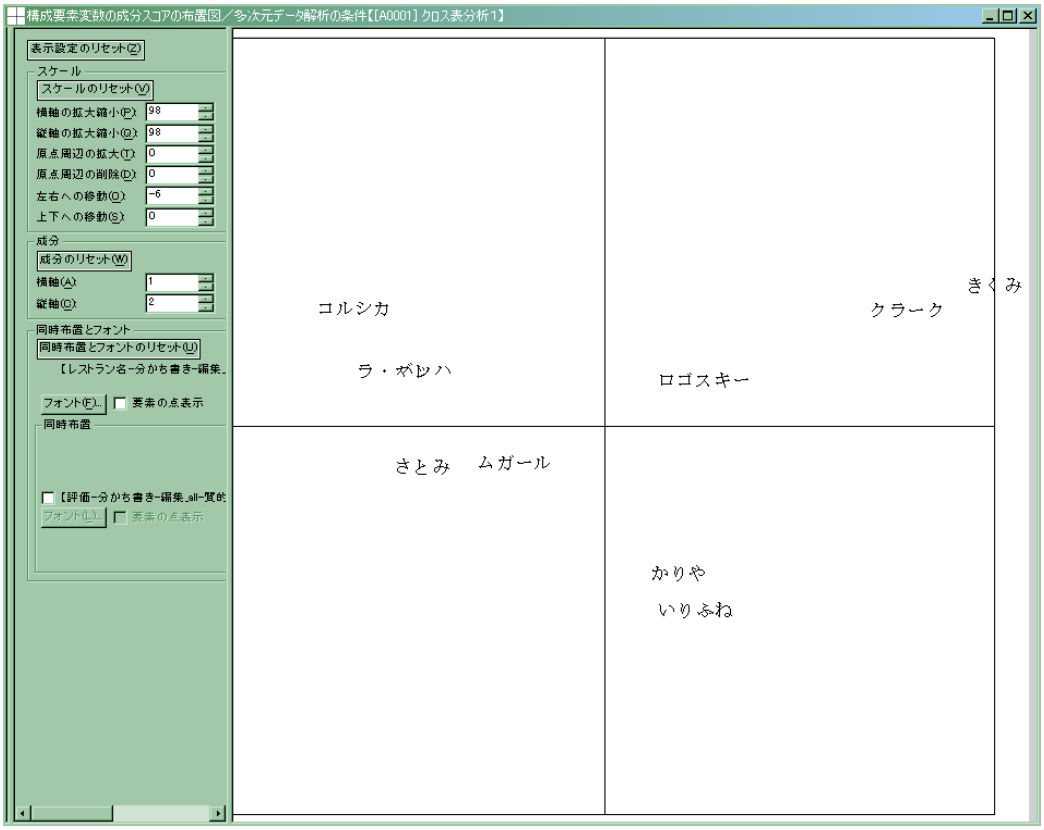


図 3 レストランの成分スコアの布置図

［$\lambda_1 = 0.1977(76.7\%), \lambda_2 = 0.060(23.3\%)$］（固有値と寄与率）

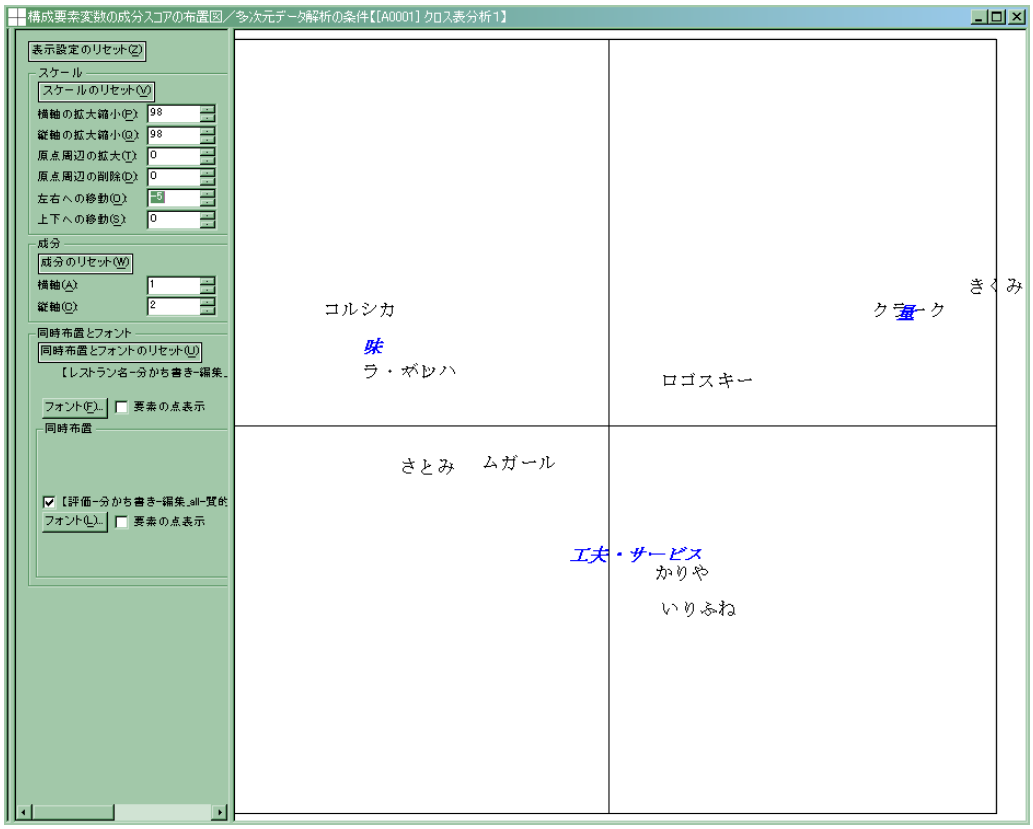


図 4　レストランと評価基準の成分スコア同時布置図

[性質 3]
対応分析で得た成分スコア間のユークリッド距離（平方ユークリッド距離）と，元のクロス表のプロフィル間のカイ二乗距離（平方カイ二乗距離）について，次の重要な関係がある．

［対応分析で得た（レストランの）成分スコア間のユークリッド距離］
＝［元のクロス表の（レストランの）プロフィル間のカイ二乗距離］　　(17)

これを例で示しておこう．ここでは，分類対象として考えているデータ表の行の要素“レストランの間の距離”について例を調べる．列の側の要素（評価項目）についても，「行」を「列」と読み替えて，同じ関係がなり立つ．

表 9　成分スコアとプロフィルの距離の一覧(確認用)

	対応分析から得た成分スコア (2 成分) (*) 図 2 から		行のプロフィル (3 つの評価基準のパターン) (*) 表 4 から：p_{ij}		
レストラン名	成分スコア 1	成分スコア 2	工夫・ｻｰﾋﾞｽ	味	量
バッハ	-0.3966	0.1220	0.3380	0.5352	0.1268
ラ・マレ	-0.4636	0.1191	0.3356	0.5616	0.1027
距離の比較	平方ユークリッド距離＝**0.004497** ユークリッド距離＝**0.06706273**		平方カイ二乗距離＝**0.00449669** カイ二乗距離＝**0.06705736**		
いりふね	0.2017	0.4082	0.6323	0.1613	0.2065
かりや	0.1647	0.3261	0.5899	0.1966	0.2135
距離の比較	平方ユークリッド距離＝**0.008109** ユークリッド距離＝**0.09005226**		平方カイ二乗距離＝**0.00810677** カイ二乗距離＝**0.09003759**		
			列和の比率（行の重心プロフィル）		
			0.4206	0.3442	0.2352

例 1:「バッハ」と「ラ・マレ」

ⅰ）成分スコアからユークリッド距離を求める

まず，バッハとラ・マレの“平方ユークリッド距離”は次にようになる．

$$d^2(\text{バッハ，ラ・マレ}) = [-0.3966-(-0.4636)]^2 + [0.1220-0.1191]^2 = 0.004497$$

よって，“ユークリッド距離”は次のようになる．

$$d(\text{バッハ，ラ・マレ}) = \sqrt{[-0.3966-(-0.4636)]^2 + [0.1220-0.1191]^2} = \sqrt{0.004497} = 0.06706273$$

ⅱ）次に，行のプロフィルと行の平均ベクトルから，同じレストランのプロフィル間の“平方カイ二乗距離”と“カイ二乗距離”を求める．

$$\chi_d^2(\text{バッハ，ラ・マレ}) = \frac{(0.3380-0.3356)^2}{0.4206} + \frac{(0.5352-0.5616)^2}{0.3442} + \frac{(0.1268-0.1027)^2}{0.2352}$$
$$= 0.00449669$$

よって，カイ二乗距離は以下となる．

$$\chi_d(\text{バッハ，ラ・マレ}) = \sqrt{\frac{(0.3380-0.3356)^2}{0.4206} + \frac{(0.5352-0.5616)^2}{0.3442} + \frac{(0.1268-0.1027)^2}{0.2352}}$$
$$= 0.06705736$$

例 2:「いりふね」と「かりや」

ⅰ）成分スコア間の“平方ユークリッド距離”は以下となる．

$$d^2(\text{いりふね，かりや}) = (0.2017-0.1647)^2 + (0.4082-0.3261)^2 = 0.008109$$

よって，“ユークリッド距離”は以下となる．

$$d^2(\text{いりふね，かりや}) = \sqrt{(0.2017-0.1647)^2 + (0.4082-0.3261)^2} = \sqrt{0.008109} = 0.09005226$$

ⅱ）行プロフィル間の“平方カイ二乗距離”は以下となる．

$$\chi_d^2(\text{いりふね，かりや}) = \frac{(0.6323-0.5899)^2}{0.4206} + \frac{(0.1613-0.1966)^2}{0.3442} + \frac{(0.2065-0.2135)^2}{0.2352} = 0.00810677$$

よって，“カイ二乗距離”は以下となる．

$$\chi_d(\text{いりふね，かりや}) = \sqrt{\frac{(0.6323-0.5899)^2}{0.4206} + \frac{(0.1613-0.1966)^2}{0.3442} + \frac{(0.2065-0.2135)^2}{0.2352}}$$
$$= \sqrt{0.00810677} = 0.09003759$$

以上の 2 つの例を表 9 に要約したが, 計算誤差の範囲で, 確かに両者の値は一致しており, 式（17）の[性質 3]がなり立つ．つまり，成分スコアのユークリッド距離とカイ二乗距離の間に見られる“距離の関係”を対応分析ではうまく使い分けていることになる．これは対応分析で得た「成分スコアによるクラスター化」は，元の「クロス表のカイ二乗距離」によるクラスター化に同じであることを示唆している．これを確かめることが次の課題である．

4.2 観察（その 2） 階層的分類の実行と結果の確認

つぎに，レストランの“**成分スコア**”を用いたクラスター化による分類結果と各種の統計

情報を示す．上に[性質 3]で示したように，“成分スコアを用いたユークリッド距離によるクラスター化”は，“カイ二乗距離によるプロフィルそのもののクラスター化”に同等である．

4. 2. 1 階層的分類の結合順と結合水準の情報

クラスター化を行うと，まず基本情報として，クラスター化過程を示す図 5 の情報が得られる．これは階層的に分類対象（レストラン）の併合を繰り返して得られる情報である．これを説明するためには，“**デンドログラム**”（樹形図；dendrogram）があると分かりやすいので，これも作ってみる（図 6，図 7）．

なお，WordMiner では，かなり規模の大きいデータセットの分類を想定しており，そのような場合にはデンドログラムの描画や観察が煩雑となるため，出力描画を行わない．ここではクラスター化過程を説明するために便宜的に利用してみる．

図 5 階層的分類の結合順とその結合水準他

表 10 クラスター生成情報：図 5 の再編集

クラスター数	階層水準に含まれる異なり構成要素数	階層水準に含まれる構成要素数	階層の結合水準値
9	2	288	0.00025
8	2	333	0.00052
7	2	212	0.00163
6	2	204	0.00165
5	3	410	0.00222
4	5	614	0.00973
3	3	458	0.01538
2	5	670	0.06788
1	10	1284	0.15842
			0.25768［結合水準の和＝固有値の和＝0.2577］

4. 2. 2 デンドログラム（樹形図）と階層水準

まず，以下の説明に用いるためにデンドログラムを用意した．ここでは 2 つのデンドログラムを作ってみた．図に見るように，これの書き方は一通りではない．つまり，ここで必要な情報はデンドログラムの見栄えではなく，クラスターが結合する順序とそのときの水準の大きさとに注意して観察する（レストラン名の並び順が異なるが，結合の順序関係が同じことに注意）．

```
rank  ind. iden  dendrogram   (indices as percentages of sum of indices :        .25768   min =   .10% / max = 61.48%)

  1   .64 1-Satomi      --+
  2  3.78 3-Mugale      --*-----+
  3   .10 2-Bach        --+     !
  4   .86 9-La_Male     --*+    !
  5 61.48 5-Corsa       ---*----*-------------------------------------------------------------------------------------------------+
  6  5.97 7-Rogosky     ----------+                                                                                              !
  7   .20 10-Kariya     --+       !                                                                                              !
  8 26.34 4-Irifune     --*-------*--------------------------------+                                                             !
  9   .63 6-Clark       --+                                       !                                                             !
 10 ----- 8-Kikumi      --*-----------------------------------------*-------------------------------------------------------------*
```

図 6　デンドログラムの例(1)

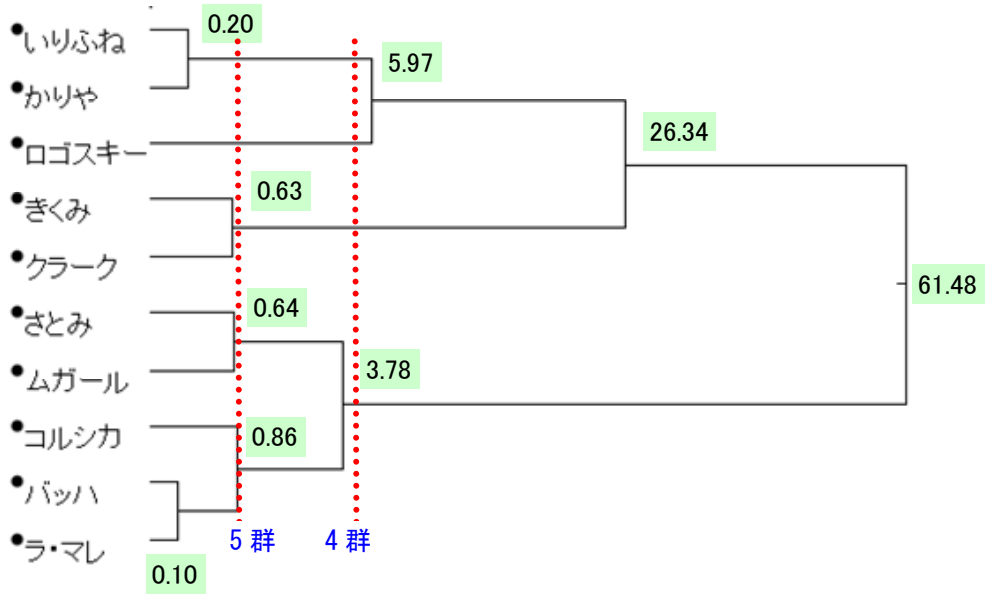
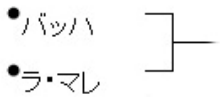


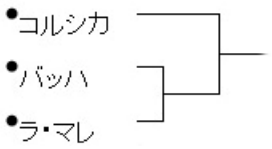
図 7　デンドログラムの例(2)
[表 17 の「割合」を結合レベルに書き入れたデンドログラム]

図 5 あるいは表 10 の“**階層の結合水準**”(hierarchical indices)と図 7 のデンドログラムは，階層的分類の各クラスターの結合段階の状況を説明している．たとえば，もっとも近い（似ている）レストランは｛バッハ｝と｛ラ・マレ｝であって，これをはじめに併合して（merge）大きさが「2」のクラスター｛バッハ，ラ・マレ｝を作る（下の図）．



つぎに似ている 2 つのレストランからクラスター｛いりふね，かりや｝を作る．以下同じように，分類対象（レストラン）あるいは生成したクラスターの併合を“階層的に”ボトムアップに繰り返す(よって凝集型階層的分類という)．この履歴がデンドログラムであり図 5，表 10 の結合水準の履歴である．また成分スコアの布置図に階層化過程を書き入れると図のような入れ子構造（階層）となっていることや，各レストランの関係がよく分かるだろう．

また，5 群のレベルでは，上で生成したクラスター｛バッハ，ラ・マレ｝に，新たなレストラン｛コルシカ｝が併合され，大きさが「3」の｛コルシカ，バッハ，ラ・マレ｝のクラスターができる（下の図）．



以下同じように，入れ子の関係，つまり階層構造としてクラスター化が進み，最後は 1 つの群，つまり 10 のレストランすべてを包含することになる（10－1=9 回の併合で完結）．デンドログラム上のこれらの併合の様子を，成分スコアの布置図に書き入れてみると図 8 のようになる（丸数字の順に結合が進む）．これで，階層・入れ子の関係，つまり併合過程が理解されるだろう．また図の中での点の離れ具合（距離）が反映された結合順となっていることもわかる．WordMiner は，このクラスター化履歴と結合水準の情報を図 5 のように出力する．

（注7） 前に“**相互最近隣の規則**”について触れた．図 8 で，①，②，③などのクラスター生成時にこの関係がなり立っている．たとえば，{バッハ} からみて {ラ・マレ} が一番近く，またこの逆がなり立つ．

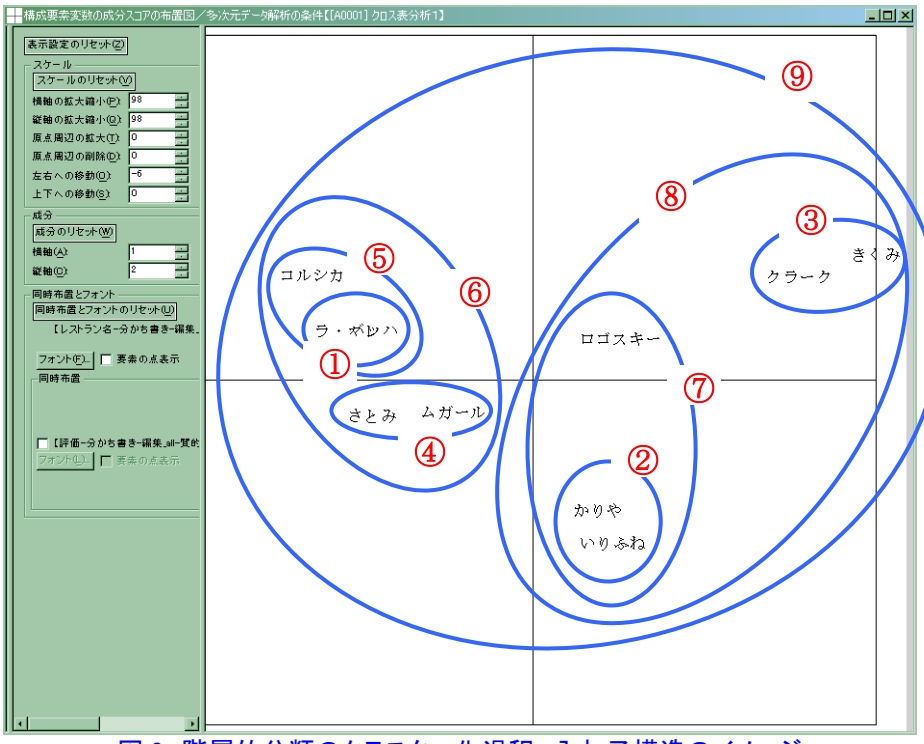


図 8　階層的分類のクラスター化過程，入れ子構造のイメージ

4. 2. 3　階層の結合水準の意味

このとき，“**階層の結合水準**”と“**クラスター内変動**”（within-cluster variances）との間には重要な関係があるので，これを調べる．4.3 項であらためて詳しく示すが，その前準備としてまず，“**クラスター内変動**”と“**クラスター間変動**”（between-cluster variances），それと“**総変動**”（total inertia）つまり“**固有値の和**”との相互の関係を示す．

記法の準備：

ここでまず，説明に必要な統計量とその記法を用意する．

クラスター数： k（これは階層の水準に対応することに注意）
総変動： S_T（あらためてこの記号で表す）

クラスター内変動：$S_W(k,l)$（クラスターをk群としたときの，そこに含まれるあるクラスターlの郡内変動，よって$l=1,2,\cdots,k$）

クラスター間変動：$S_B(k)$（クラスターをk群としたときの群間変動）

クラスター間変動比：$\eta_k=\dfrac{S_B(k)}{S_T}\times 100$（%）（$k$群のとき）

ある階層水準（クラスター）におけるカイ二乗統計量：$\chi_p^2(k)$（クラスターをk群としたときのピアソンのカイ二乗統計量をこれで表す）．$k=10$（群）としたときは，分類しないとき（もとのクロス表のまま）であるから，$\chi_p^2(10)=\chi_p^2$となる．後でみるようにクラスター数kを変えることは**圧縮したクロス表**を調べること，このカイ二乗統計量を分解することにつながる［（表 20）を参照］．

（注9）　ここで総変動S_Tは，いったん分析対象のクロス表が決まると値が確定する（ある一定値になる）．またクラスター内変動はクラスター数が増えるほど次第に（単調に）小さくなる．一方，クラスター間変動は（総変動が一定であるから）クラスター数が減ると次第に増える（クラスター内変動とクラスター間変動はトレードオフの関係）．したがって“**クラスター間変動比**”η_kの変化を調べることは，クラスター化（まとまり）の程度を知る指標の 1 つとなる（しかし，単調変化なので目安とはなってもクラスター数を決めるクリティカルな基準とはならない）．

［性質 4］

階層の結合水準について，次の性質がある．

いま，クラスター化で得られる全成分数（K）を指定したとき，以下の関係がある（ここで，$K=\min\{m,n\}-1$）．

［階層の結合水準値の和］＝［固有値の和］　（18）

つぎに，成分数を全成分数（K）より少ない成分数K^*（$<K$）としたとき、上の関係は以下のようになる．

［階層の結合水準値の和］＝［その指定した成分数K^*までの固有値の和］　（19）

例：表 10 に階層の結合水準値の和の欄を作ったが，ここにみるように式（18）の関係がなり立つ．式（19）については，うしろに例を示す．

［性質 5］

上で約束した記法によると，各統計量（総変動，クラスター間変動，クラスター内変動，ピアソンのカイ二乗統計量，固有値の和）の間に以下の関係がある．

［総変動］＝［クラスター間変動］＋［クラスター内変動の和］（＝固有値の和）　（20）

$$S_T=S_B(k)+\sum_{l=1}^{k}S_W(k,l)\quad\Leftrightarrow\quad S_T=\sum_{\alpha=1}^{K}\lambda_\alpha=\frac{\chi_p^2}{N}\qquad(20-1)$$

あるいは言い替えて，

［総分散］＝［クラスター間分散］＋［クラスター内分散の和］　　（20－2）

あるいは，これも言い替えて，

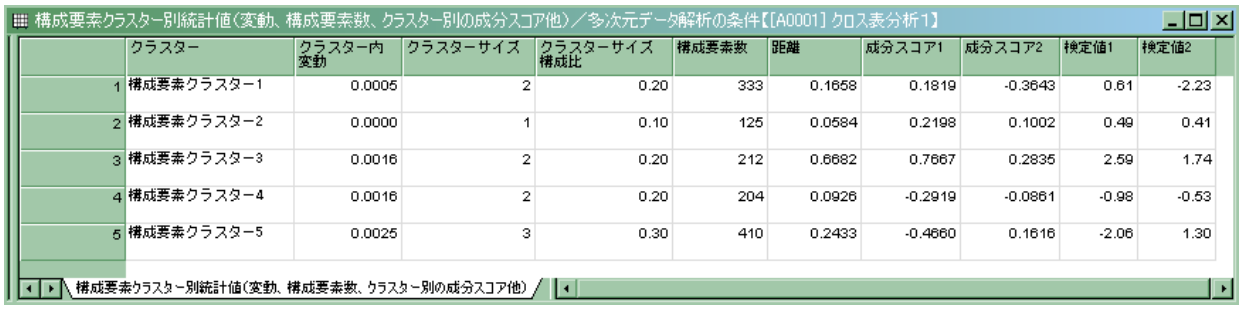
［Total inertia］
＝［between-clusters inertia］＋［within-clusters inertia］　　（20－3）

ここで，上の関係を例で確認する．その前準備として，上の分類で得たあるクラスターについて，まず，クラスター化における成分スコアの意味を調べる．続いて上に挙げた関係式を調べる．

4．2．4　クラスター数を $k=5$（群）としたときを例として

（確認1）　クラスターの成分スコアほか

ここでは「クラスター数＝5（$k=5$）」とした場合を例として，種々の関連を調べる．まず，WordMiner が出力する図 9 の情報とそれを説明用に書き替えた表 11 を作った．この表につけた丸番号の情報について順をおって説明する．


構成要素クラスター別統計値(変動、構成要素数、クラスター別の成分スコア他)／多次元データ解析の条件【[A0001] クロス表分析1】

	クラスター	クラスター内変動	クラスターサイズ	クラスターサイズ構成比	構成要素数	距離	成分スコア1	成分スコア2	検定値1	検定値2
1	構成要素クラスター1	0.0005	2	0.20	333	0.1658	0.1819	-0.3643	0.61	-2.23
2	構成要素クラスター2	0.0000	1	0.10	125	0.0584	0.2198	0.1002	0.49	0.41
3	構成要素クラスター3	0.0016	2	0.20	212	0.6682	0.7667	0.2835	2.59	1.74
4	構成要素クラスター4	0.0016	2	0.20	204	0.0926	-0.2919	-0.0861	-0.98	-0.53
5	構成要素クラスター5	0.0025	3	0.30	410	0.2433	-0.4660	0.1616	-2.06	1.30

構成要素クラスター別統計値(変動、構成要素数、クラスター別の成分スコア他)

図 9 「クラスター数：$k=5$（群）」としたときのクラスター要約情報

表 11　図 9 の書き替え（説明用）

①	②	③	④	⑤	⑥	⑦		⑧	
クラスター $k=5$（群） $l=1,2,\cdots,5$	クラスター内変動 $S_w(k,l)$	クラスターサイズ	クラスターサイズ構成比	構成要素数	距離	成分スコア 1	成分スコア 2	検定値 1	検定値 2
1	0.0005	2	0.2	**333**	0.1658	**0.1819**	-0.3643	0.61	**-2.23**
2	0.0000	1	0.1	125	0.0584	0.2198	0.1002	0.49	0.41
3	0.0016	2	0.2	212	0.6682	0.7667	0.2835	**2.59**	1.74
4	0.0016	2	0.2	204	0.0926	-0.2919	-0.0861	-0.98	-0.53
5	0.0025	3	0.3	**410**	0.2433	**-0.4660**	0.1616	**-2.06**	1.3

表 12　情報の要約（図 2 からの引用）

レストラン名	構成要素数 構成比	距離	λ_1 に対する 成分スコア 1	λ_2 に対する 成分スコア 2
いりふね	0.121	0.21	**0.2017**	-0.4082
かりや	0.139	0.13	**0.1647**	-0.3261
きくみ	0.086	0.83	0.8590	0.3091
さとみ	0.074	0.17	-0.4009	-0.0908
クラーク	0.079	0.51	0.6672	0.2558
コルシカ	0.095	0.37	**-0.5497**	0.2586
バッハ	0.111	0.17	**-0.3966**	0.1220
ムガール	0.085	0.05	-0.1969	-0.0821
ラ・マレ	0.114	0.23	**-0.4636**	0.1191

① **クラスター**：生成した 5 群のクラスターに変数名「構成要素クラスター1」，…と名前を付与し，これをあらたな変数（例：質的変数に転用）として利用できる．

② **クラスター内変動**：これは $S_W(k,l)$ に相当する統計量である．この例では $S_W(5,1)=0.005,\cdots,S_W(5,5)=0.0025$ と対応する．各クラスター内変動の大きさ，つまり**クラスター内分散**の大きさを表す．これが小さいほど，まとまりのよいクラスターとなる．なお，クラスターサイズが 1 個（シングルトン：singleton という）の場合，とうぜん変動はないので「0」となる．ここでは「クラスター2」がそれにあたる（$S_W(5,2)=0.0000$）．

また，クラスターサイズが複数であっても，それぞれの数値（成分スコア）がまったく同じであれば，同じくクラスター内変動は「0」となる．またこの例ではクラスター5 が，いちばんバラツキが大きい（$S_W(5,5)=0.0025$）．［図 11 をみるとわかるだろう］

③ **クラスターサイズ**：クラスター内に所属する分類対象（レストラン）の数．ここでは 5 つのクラスター内それぞれに入った分類対象の数（レストランの数）を示している．たとえば，クラスター5 は 3 つのレストランからなる．

④ **クラスターサイズ構成比**：各クラスター内に占める分類対象の割合．

⑤ **構成要素数**：この場合は，各クラスター内に属する回答者数（サンプル数）となる．

⑥ **距離**：各クラスターの重心（セントロイド）から（布置座標の）原点までの距離（ここは**カイ二乗距離の二乗**となる）．前にレストランについて説明したことを，「クラスター」と読み替えればよい．

⑦ **成分スコア**：対応分析で得られた成分スコアから得たクラスターの成分スコア．ここは固有値に合わせて 2 つの成分スコアがある．ここの成分スコアはクラスター内の構成要素数の**加重平均**となる（つまりクラスターの重心）．うしろに算出例を示した．

⑧ **検定値**：成分スコアの有意性を検定した結果．**正規近似**で評価する．値（絶対値）が大きいほど，そのクラスターの特徴があると考える．検定値 1 は成分スコア 1 に，検定値 2 は成分スコア 2 にそれぞれ対応する．有意水準を 5%としたとき，この検定値の（絶対値が）1.96 より大きいとこの有意水準で有意と考える．たとえば，検定値 1 ではクラスター3 の成分スコアが有意となる．同じくクラスター5 の成分スコア 1 も有意となる．これについても，簡単な例をうしろに示した．

成分スコアの算出方法:

ここで，この例のクラスター数が 5 群の場合の成分スコアの算出方法を調べる．このため表 11 と表 12 を用いる．また確認の計算表を作る（表 13）．

例 1:クラスター5 の成分スコア 1（第 1 固有値に対応）の値「-0.4660」を調べる（表 11 の⑦欄，最下段の太字）．

- このクラスターは大きさが 3 でその所属要素（メンバーシップ）は{バッハ，コルシカ，ラ・マレ}である．
- この 2 つのはじめの成分スコアは図 2（表 12 に引用）で得られているのでこれを用いる．

$$\frac{-0.5497\times122+(-0.3966)\times142+(-0.4636)\times146}{122+142+146}=\frac{-191.0622}{410}\doteqdot-0.4660$$

これで該当する成分スコアを得た．これを表 13 の例 1 の計算表にまとめた．

例 2:クラスター1 の{いりふね，かりや}を調べる．ここでは表 13 から表 11 のクラスター1 の成分スコアは，以下のようになる．

$$\frac{0.2017\times155+0.1647\times178}{155+178}=\frac{60.5801}{333}\doteqdot 0.1819$$

なおここで，表 12 の「構成要素数構成比」＝「(クロス表の）行の相対確率 p_{i+}」であることに注意して，[（構成要素数構成比×成分スコア）の和]×1284÷[行和]としても同じである．例 1 ならば，

$$\frac{\{0.095\times(-0.5497)+0.111\times(-0.3966)+0.114\times(-0.4636)\}\times1284}{122+142+146}=\frac{-191.44196}{410}\doteqdot -0.4669$$

となる．同じように例 2 についても求められる．これらの数値の前者とのズレは計算誤差の範囲である．

表 13　クラスター5, クラスター1 の成分スコア 1 の算出

例 1:クラスター5 について

レストラン名	⑤構成要素数構成比	距離	①λ_1に対する成分スコア 1	②行和（表 3 の行和）	①×②
コルシカ	0.095	0.37	-0.5497	122	-67.0634
バッハ	0.111	0.17	-0.3966	142	-56.3172
ラ・マレ	0.114	0.23	-0.4636	146	-67.6856
↓　併合により下のクラスターになる				↓	↓
{バッハ，コルシカ，ラ・マレ}				410	-191.0662
				（表 11）→	**-0.4660**

例 2:クラスター1 について

レストラン名	構成要素数構成比	距離	①λ_1に対する成分スコア 1	②行和（表 3 の行和）	①　②
いりふね	0.121	0.21	0.2017	155	31.2635
かりや	0.139	0.13	0.1647	178	29.3166
↓　併合により下のクラスターになる				↓	↓
{いりふね，かりや}				333	60.5801
				（表 11）→	**0.1819**

成分スコアの布置図の確認:

ここで，元の成分スコアとクラスター化で得たクラスターの成分スコアとの関係を実際に布置図の上で観察してみよう．はじめのクロス表からえた各レストランの成分スコアの布置と（つまり図 3，表 12 に同じ），ここで 5 群の場合に求めた表 11（⑦欄）のクラスターの成分スコア（つまり**クラスターの重心**）の関係を布置図とすると図 10 のようになる．

図 10 で「構成要素クラスター1」「構成要素クラスター2」…が，各クラスターの成分スコア，つまり上に表 11 の⑦欄で確認し求めた成分スコアである（文字の中央がクラスターの重心の位置）．

たとえば，上で確かめたクラスター{バッハ，コルシカ，ラ・マレ}の中に「構成要素クラスター5」がある．他のクラスターについても，同じように観察される．また，この階層レベル（$k=5$ 群）では{ロゴスキー}はサイズが 1 個であるから元と同じ成分スコアのままとなっている．

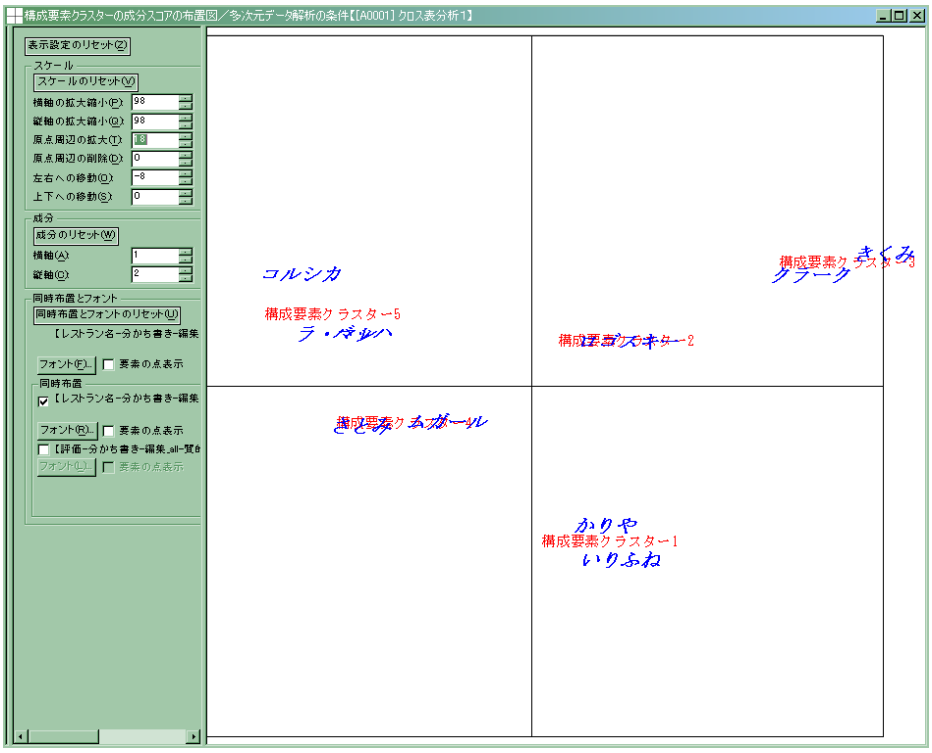


図 10　もとの成分スコアと 5 群の場合のクラスターの成分スコア（重心）の布置図

検定値の算出方法：

ここで，クラスター化で得た結果（表 11）にある“**検定値**”について，この例を用いて説明する．ここで“**検定統計量の実現値**”のことを検定値と呼ぶことにしで，この検定統計量をどのように考えているかを簡単に示し，それを数値例で確認する．

ｉ）検定統計量の考え方

まずここでは，成分スコアを用いるので，これの性質を確認するために，以下の記号を準備する．

準備 1：

x_α ：これを第 α 番目固有値に対応する成分スコア（coordinate）とする（“成分スコア α ”ということ）．このとき，対応分析で得た成分スコアの平均値は「0」である（対応分析を行った結果として，そのように調整されているということ，つまり $\bar{x}_\alpha = 0$）．

λ_α ：すでに約束したように，所与のデータ表の対応分析で得た第 α 番目の固有値．つまり，これが第 α 成分スコアの“**分散**”である．

準備 2：

以上を前提に（確認として），ここでは，以下のような条件を想定している．

- （母集団からの）**単純無作為抽出**（SRS：simple random sampling）を前提とする．
- かりに，大きさ N の母集団から，無作為抽出で大きさ n の標本を抽出したとする．
- ここで標本抽出にあたって“**復元抽出**”（WR：with replacement）と“**非復元抽出**”（WOR：without replacement）がある．ここでは一般の社会調査のように“非復元抽出”とする．
- つまり，単純無作為非復元抽出（SRSWOR）を適用したとする．

このとき，大きさが n の標本から得た“**標本平均**”は，“平均値が「0」で，分散（Var）

が次の式で与えられることは，標本抽出の原理からよく知られたことである（つまり，標本平均という統計量の標本分布の平均値と分散の関係）．

$$Var = \frac{1}{n}\frac{N-n}{N-1}\lambda_\alpha \tag{21}$$

ここで，各記号は以下を意味する．

N ：母集団の大きさ（ここではデータセット全体を母集団とみなす）
λ_α ：これが母集団の分散に相当と考える（第α成分の分散）
n ：標本の大きさ
［注：ここで$\frac{N-n}{N-1}$を**“有限修正項”**という］

なおここで，Nが，nに比べて非常に大きい場合，また復元抽出としたときは，上の式は以下のように近似される．

$$Var = \frac{1}{n}\frac{N-n}{N-1}\lambda_\alpha \approx \frac{\lambda_\alpha}{n} \tag{22}$$

ここで得られた第α成分の成分スコア（x_α）について，以下の“**標準化**”を行い，この統計量が，平均は「0」，分散は「1」の“**正規分布**” $N\left(0,1^2\right)$**に近似する**として検定を行う．

成分スコアの標準化（第α成分について）：

$$z_\alpha^* = \frac{x_\alpha - \bar{x}_\alpha}{\sqrt{Var}} = \sqrt{n\frac{N-1}{N-n}}\frac{x_\alpha}{\sqrt{\lambda_\alpha}} \simeq N\left(0,1^2\right) \tag{23}$$

同様に，ここでNが，nに比べて非常に大きい場合，あるいは復元抽出とすると，標準化変数は以下となる．

$$z_\alpha = \frac{x_\alpha - \bar{x}_\alpha}{\sqrt{Var}} = \frac{\sqrt{n}x_\alpha}{\sqrt{\lambda_\alpha}} \simeq N\left(0,1^2\right) \tag{24}$$

ここで，以下のように考える．かりに複数回の無作為非復元抽出を行ったとすると，それぞれは上の分布に従うはずである．またそれぞれの標本は母集団の縮図となっているはずである．

一方，クラスター化で得た複数のグループが，かりに元の母集団にクラスター的な分かれた構造がなく，ランダムに分布しているならば，個々のグループの傾向は似ているはずである．クラスター化の結果，個々のクラスター内の測定値（ここでは成分スコア）は類似し，一方，クラスター間がよく分かれているならば，個々のクラスターで得た成分スコアに上の検定を適用すると，クラスター化が顕著であれば（分類がうまく機能すれば）大きくずれるはずである（つまり**“有意になる”**だろう，ということ）．この状況を各クラスターについて検証することで，個々のクラスター化の程度を知る1つの**“目安”**とする．（正規近似とする）検定統計量のここでの使い方は，おおよそ上のようなことである．“目安”としたように，また上の説明から明らかだが，この指標は，相対的に個々のクラスター化の様子を知る実用的

かつ発見的・探索的に使うツールで，検定値の示す数値のわずかな違いを議論することではない．ここで，注意すべき点として以下がある．

- 母集団と見立てるデータセットのサイズ（N）はそれなりに大きいことが必要．
- また，標本に相当するクラスターサイズもある程度の大きさ（n）が必要であること．
- 正規近似を用いていることには限界があること．
- 言い替えると，この母集団の大きさ N や標本の大きさ n が小さいとき，とくに n が小さいときには，近似の程度があまりよくないこと．

検定値の算出例:

ここで引き続いて，いまみている 5 群の例について，検定値の求め方を説明する．またそのあとに，この検定値をどのように用いるかを説明する．再び，図 10，表 11 から必要部分を取り出し，検定値の算出に必要な作業欄を加える．

表 14　図 10, 表 11 の書き替え(説明用)

成分 1 について

①	②	③	④	⑤	⑥	⑦	⑧
クラスター $k=5$ $l=1,2,\cdots,5$	クラスターサイズ（n）	クラスター内変動 $S_w(k,l)$	成分スコア 1	WR から推定	WOR から推定	検定値 1	相対誤差（%）（WOR）
1	2	0.0005	0.1819	0.57855	0.57878	0.61	-5.11800
2	1	0.0000	0.2198	0.49434	0.49434	0.49	0.88539
3	2	0.0016	0.7667	2.43858	2.43953	**2.59**	-5.80959
4	2	0.0016	-0.2919	-0.92842	-0.92878	-0.98	-5.22605
5	3	0.0025	-0.4660	-1.81528	-1.81669	**-2.06**	-11.81093
	(10)						

成分 2 について

クラスター $k=5$ $l=1,2,\cdots,5$	クラスターサイズ（n）	クラスター内変動 $S_w(k,l)$	成分スコア 2	WR から推定	WOR から推定	検定値 2	相対誤差（%）（WOR）
1	2	0.0005	-0.3643	-2.10329	-2.10411	**-2.23**	-5.6454177
2	1	0.0000	0.1002	0.40906	0.40906	0.41	-0.2281007
3	2	0.0016	0.2835	1.63679	1.63743	1.74	-5.8950424
4	2	0.0016	-0.0861	-0.49710	-0.49729	-0.53	-6.1712415
5	3	0.0025	0.1616	1.14268	1.14358	1.30	-12.032597
	(10)						

ここで，表の丸数字にしたがって、簡単に説明する．

① クラスター：ここで用いる例のクラスター数（5 群）
② クラスターサイズ：分類対象とした 10 のレストランを 5 群のクラスターに分けた結果．それぞれがレストラン数．これを検定統計量で用いる標本の大きさ（n）に対応させる．
③ クラスター内変動：つまりクラスター内分散 $S_w(k,l)$ のこと．
④ 対応分析で得られた成分スコア．上の表が成分スコア 1，下の表が成分スコア 2.
⑤ 無作為復元抽出とした，式（24）から得た（理論上の推定した）検定値.
⑥ 無作為非復元抽出とした，式（23）から得た（理論上の推定した）検定値.
⑦ 検定値：プログラムが出力した値（あるアルゴリズムにより求めた近似値）.
⑧ 相対誤差：ここは，［（⑥の絶対値）－（⑦の絶対値）］÷［⑦の絶対値］×100（%）とした．つまり，計算で得た検定値からみた相対的な誤差を評価する量.

観察のポイント:

- ここでは，n の大きさがかなり小さいから，当てはまりの程度はさほど良くない．
- 正規近似のおおまかな情報として観察する．有意水準を 5%とすると，（標準化した）検定値の絶対値が 1.96 よりも大きければ“有意”と考える［$\left|z_\alpha^*\right| \geq 1.96$ かどうかを判定する］．
- この例で，“**形式的にこのルールを適用**”すると，成分 1 については「クラスター3」「クラスター5」あたりが有意，成分 2 については「クラスター1」が有意となる．

＜成分 1 で有意のクラスター＞
　クラスター3＝{クラーク，きくみ}
　クラスター5＝{バッハ，コルシカ，ラ・マレ}

＜成分 2 で有意のクラスター＞
　クラスター1＝{いりふね，かりや}

また，クラスター2 はシングルトン，つまり要素が{ロゴスキー}のみで（クラスター内分散は当然 0 であり）確かに検定値は小さくなる（元の全体の集団，母集団と差違がないという，当たり前の情報を示している）．図 10 にクラスターを書き入れた図を作り，もう一度下にあげた（図 11）．ここで上のクラスターが，どういう位置関係にあるかを検定値の評価結果と比べると，この操作がどういうことを調べたかがわかる．

たとえば，クラスター3 とクラスター5 は第 1 成分の左右の端にある．つまり“成分 1 に対する説明力”が高いということを示している（全体の平均・重心から第 1 軸にそって遠い位置にある）．一方，クラスター 1 は，第 2 軸（成分 2）の下の方に位置している．つまりそちらに向かって重心から遠く，説明力があるということになる．

さらに，残りのクラスター4 とクラスター2 は，他のクラスターに比べて中央に近く位置し（つまり全体の布置の平均・重心に近く），他のクラスターよりも相対的に説明力が弱いことを示している．

ここで注意することは，この例は（意図的に）成分数が少ない例，とくに全情報が 2 次元空間内に入るような例を作ってあることである．多くの場合は，扱うデータ表の寸法は非常に大きいから，ここでみたようにはクリアにはクラスター化と成分の関係をグラフィカルには観察できない．こういう場合にこそ図 10，表 11 の“**成分スコアと検定値の一覧を観察**”してその傾向を慎重に探査することが有効である．

（確認 2） 各統計量の関係:各図，各表の見方

つづいて，式（18）～（20）にあげた関係がなり立つことを総合的に確認する．説明に必要な情報を，WordMiner の出力から拾い出して表 16 に要約した．この表は，各統計量（クラスター間変動，クラスター内変動，総変動）の関係を調べるために整理した表である．ここにも丸数字を付けたので，これに対応させて説明する．

⑨ **変動の大きさ**：この欄の 5 つのクラスター内変動は表 11 の②欄に同じ数値である．これらを加えたものが“**クラスター内変動の和**”（$\sum_{l=1}^{5} S_W(k,l) = 0.0062$）となる．一方，“**クラスター間変動**”として得られた値がクラスター間変動の欄にある $S_B(k) = 0.2514$ である．これらを加えると式（18）～（20）のように“**総変動（全分散）**” S_T となる．これはすでに述べたように固有値の和に等しい．また，“**変動比**”（［変動比：η_k ］＝［クラスター間変動/総変動］）は総変動に占める k 群のクラスター間変動の割合で，クラスターのある種の説明力を表す．

⑩ **クラスターサイズ**：ここは各クラスター内に入る要素数（ここではレストランの数）.
⑪ **構成要素数**：これはクラスター内のサンプル数．表 11 の⑤欄に同じ.
⑫ **距離**：重心からのカイ二乗距離の二乗．表 11 の⑥に同じ.

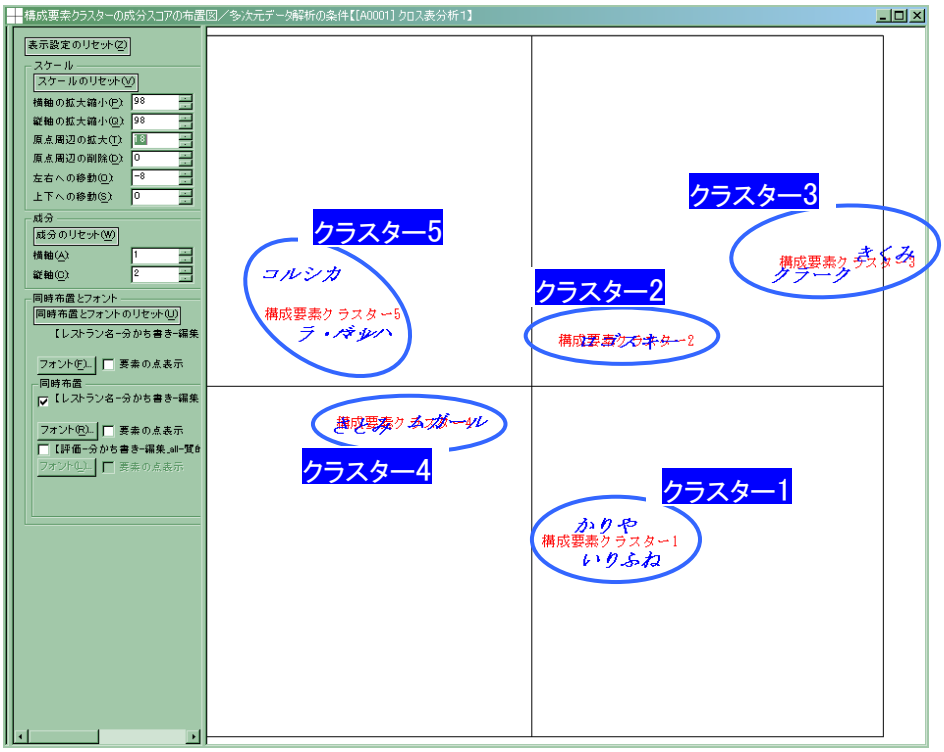


図 11　もとの成分スコアと 5 群の場合のクラスターの布置図

表 15　5 群の場合のクラスター構成

ｸﾗｽﾀｰ	クラスターサイズ （クラスター内のレストラン数）	クラスター化履歴	工夫・ サービス	味	量
1	2	{いりふね，かりや}	203	60	70
2	1	{ロゴスキー}	48	35	42
3	2	{クラーク，きくみ}	69	22	121
4	2	{さとみ，ムガール}	91	90	23
5	3	{バッハ，コルシカ，ラ・マレ}	129	235	46

（10）

レストランを 5 群にクラスター化した場合の“**圧縮化したクロス表**”．つまり「5 群×3（評価基準）」のクロス表.

なお，表 15 は，次に説明する階層的分類で得た各クラスター間の関連を調べるための補足情報である．この表は，いままでに得た情報に，さらに 5 群にクラスター化した場合の“**圧縮化したクロス表**”の情報である.

つまり「“5 群のクラスター”と“3 つの評価基準”」から作られる大きさが“（5 クラスター）×（3 評価基準）のクロス表”である．このクロス表の“総変動”S_T，つまりこのクロス表に対応分析を適用して得られる“ピアソンのカイ二乗統計量”（$\chi_p^2(k)$であり“固有値の和”と階層水準の間には，式（20）と同じような関係がなり立つ．たとえばここでは，$\chi_p^2(5)=322.799$，$S_T=\sum_\alpha \lambda_\alpha = 0.2514 = \dfrac{322.799}{1284}$ となる．クラスター数を変えたとき，つまりクラスター化過程とこれら諸量との関係についてはうしろで例を用いてあらためて要約する［（表 20）を参照］.

表 16　総変動，クラスター間変動，クラスター内変動の関係

項目 / 統計量	クラスター $\begin{pmatrix} k=5 \\ l=1,2,\cdots,5 \end{pmatrix}$	⑨ 変動の大きさ	⑩ クラスターサイズ（クラスター内のレストラン数）	⑪ 構成要素数（クラスター内のサンプル数）上の⑤に同じ	⑫ 距離（重心からのカイ二乗距離の二乗）上の⑥の同じ
クラスター間変動 $S_B(k)$	—	(0.2514)			
クラスター内変動 $S_W(k,l)$	1	0.0005	2	333	0.1658
	2	0.0000	1	125	0.0584
	3	0.0016	2	212	0.6682
	4	0.0016	2	204	0.0926
	5	0.0025	3	410	0.2433
クラスター内変動の和 $\sum_{l=1}^{k} S_W(k,l)$	—	(0.0062)	（10）（行の要素数）	（1,284）（総和）	
総変動（全分散） S_T（固有値の和）	—	0.2577 (0.2576)			
変動比 [クラスター間変動/総変動] $\eta_k = \frac{S_B(k)}{S_T} \times 100$		0.9756 （97．6%）			

4.3 あらためて階層的分類の階層化の意味を調べること

ここであらためてクラスター化の階層的分類の履歴，各結合水準の値の解釈を説明する．ここで表 10 の出力情報から，あらためて表 17 のように情報を要約する（これは図 5，表 10 にさらに情報を追加して詳しく示したもの）．上では主に 5 群の場合に注目したが，ここでは 1 群～10 群までの階層的分類の全履歴を観察し，そこで得られる各情報の意味を説明する．ここでもまた，表の各欄に丸数字を付与し説明する．

① **階層水準に含まれる異なり構成要素数**：この例の場合，各クラスターレベルにおける初期のクロス表のレストランの数となる．
② **階層水準に含まれる構成要素数**：ここでは，クロス表の総和である全回答者（1,284 名）が併合の過程で各クラスターにどう配分されたか（併合・吸収されたか）を示す．
③ **階層の結合水準**：上で 5 群の場合について調べたように，その群における（併合・生成されたときの）“**クラスター内変動の和**”に相当する．つまり分類なし（クラスター数 $k=10$）のときが「0」，$k=9$ で「0.00025」，…以下分類の併合が進むにつれて単調に増えて，最後は（$k=1$）で総変動つまり元のクロス表の対応分析で得た固有値の和となる．
④ **結合水準の累積和**：結合水準の累積履歴がこの欄の情報．よって最後のセルの値が固有値の和となる．
⑤ **総変動に占める割合**：総変動を 100 としたときの水準の割合（%）．図 7 のデンドログラムに書き入れた数字がこれに相当する．

表 17　階層の結合水準ほかの再確認

	①	②	③	④	⑤
クラスター数（k）	階層水準に含まれる異なり構成要素数	階層水準に含まれる構成要素数	階層の結合水準	結合水準の累積和	総変動に占める割合(%)
クラスターの遷移	(a) クラスターに含まれるレストランの数	(b) クラスター内のサンプル数	(c) デンドログラムで確認	(d) 各水準のクラスター内変動の和	③÷総変動（固有値総和）×100（%）
9	2	288	0.00025	0.00025	0.10
8	2	333	0.00052	0.00077	0.20
7	2	212	0.00163	0.00240	0.63
6	2	204	0.00165	0.00405	0.64
5	3	410	0.00222	0.00627	0.86
4	5	614	0.00973	0.01600	3.78
3	3	458	0.01538	0.03138	5.97
2	5	670	0.06788	0.09926	26.34
1	10	1284	0.15842	**0.25768** ［固有値の和］	61.48
—	—	—	**0.25768** ［結合水準の和］	—	100.00

クラスター数の目安を得ること:

③の結合水準の変化と⑤の割合の変化を追跡して“**クラスター数を決める目安**”とする．たとえばこの例の場合は，4 群→5 群，あるいは 3 群→2 群の間での変化量が大きいので，クラスター数はおおよそ「4 群」あるいは「2 群」がよさそうと判断する．図 5 の出力情報から（棒グラフ）からこれの見当をつける．図 5 の棒グラフは階層の結合水準の変化をグラフ化したもので，おおよそ以下のように観察する．たとえば，棒グラフの変化が急に階段状に変化する位置，つまり表 17 の④（結合水準の累積和）あるいは⑤（総変動に占める割合）が大きく変化する位置を目安とする．

なおここで，③の「結合の水準」の関係がやや分かりにくいので，例で示そう．

例 1: 5 群に結合の時点での場合を調べる（表 17 で $k=5$ に対応する行をみる）
①欄の「3」は，階層の（デンドログラム上の）第 5 回目の併合で｛バッハ，ラ・マレ｝＋｛コルシカ｝がくくられ，そのリンク数が「2＋1=3」となったことを示す．つまり，②欄のバッハ＝142，ラ・マレ＝146，コルシカ＝122 の（行和）の和が 142＋146＋122＝410（人）となったということ．

例 2: 同様に，4 群と結合するときを考える（表 17 で $k=4$ に対応する行をみる）
｛バッハ，ラ・マレ｝＋｛コルシカ｝＝｛バッハ，ラ・マレ，コルシカ｝で（3）となった②「410」と，さらに①の｛さとみ，ムガール｝（2）の「204」とが併合して，（5）となり，「410＋204＝614（人）」のサイズのクラスターとなる．

重要な性質の確認（その 1）:

前に $k=5$（群）のクラスター化で得た圧縮化したクロス表（表 15）に対応分析を適用したときに得られる固有値の和，つまりこのクロス表の総変動は，いま調べているクラスター化履歴の 5 群までの“**結合水準の累積和**”となり，これは“**クラスター内変動の和**”となっている．表 17 でいうと，④欄の $k=5$ に対応する値「0.00627」がそれに相当する．つまりこれが表 15 に作った 5 群の圧縮化したクロス表の対応分析で得た**総変動**であると同時に，元のクロス表の対応分析とクラスター化でえた **5 群のときのクラスター内変動**でもある．

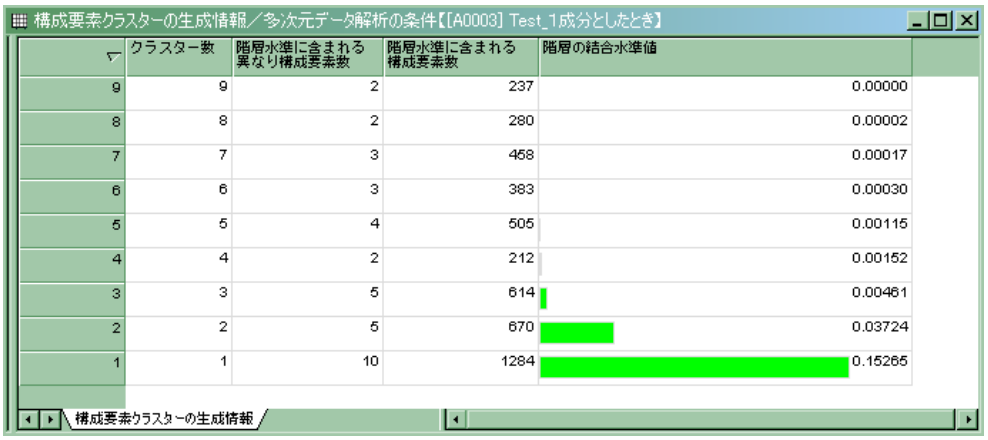
重要な性質の確認（その 2）:

ここでは求めた成分数のすべてに対応する成分スコアを用いて分析した（つまり $K=2$ とした）．しかしここで，“全成分数 K を指定せず”に，これより少ないある成分数（$K^{*}<K$）

を指定すると，その成分数（K^*）までの固有値の和が「合計」（総変動）となる（注：[性質4]の式（19）の確認）．

一般には，出発時の 2 元表（データ表）の寸法が“大きい”ので，寸法の大きい（サンプル）×（構成要素変数）のデータ表の対応分析の後に，サンプルのクラスター化を行うようなときには，全成分数を用いずに始めの方のある成分数を指定するだろう（たとえば，WordMiner の**デフォルト値はこれを 15 成分**としてある）．このときは，階層の結合水準の総和（合計）は，**“その指定した成分数（K^*）までの固有値の和”**となる．

たとえば，この例では 2 つの固有値（$K=2$）があるが，クラスター化処理で「成分数=1（$K^*=1$）」と指定すると第 1 固有値の大きさ $\lambda_1=0.1977$ が**“階層の結合水準の総和”**となる．これを実際に行ってみると確かに図 12，表 18 のようになる．



	クラスター数	階層水準に含まれる異なり構成要素数	階層水準に含まれる構成要素数	階層の結合水準値
9	9	2	237	0.00000
8	8	2	280	0.00002
7	7	3	458	0.00017
6	6	3	383	0.00030
5	5	4	505	0.00115
4	4	2	212	0.00152
3	3	5	614	0.00461
2	2	5	670	0.03724
1	1	10	1284	0.15265

図 12　クラスター生成情報（1 成分のみを指定のとき）

表 18　図 12 の情報の要約

クラスター数	①階層水準に含まれる異なり構成要素数	②階層水準に含まれる構成要素数	③階層の結合水準	④水準の累積和	③全変動を 100 としたときの水準の割合（%）
9	2	237	0.00000	0.00000	0.00
8	2	280	0.00002	0.00002	0.01
7	3	458	0.00017	0.00019	0.10
6	3	383	0.00030	0.00049	0.25
5	4	505	0.00115	0.00164	0.83
4	2	212	0.00152	0.00316	1.60
3	5	614	0.00461	0.00777	3.93
2	5	670	0.03724	0.04501	22.77
1	10	1284	0.15265	**0.19766**（$=\lambda_1$）	100.00
—	—	—	**0.19766**（$=\lambda_1$）［結合水準の和］	—	—

4.4 クラスター化過程の総合考察

いままで説明した情報を総括し，それぞれがクラスター化の過程でどう使われているかを総合的に整理してみる．とくに，クラスター内変動，クラスター間変動，ピアソンのカイ二乗統計量（つまり総変動であり固有値の和である），クラスター化の結合の水準のそれぞれの関係が，クラスター化の中でどう利用され，何を示しているか，相互の関係はどうなるか，を総合的にまとめてみる．

確認 1:

まず表 19 の情報を読み解く．すでに示した[性質 4]すなわち式（18），（19），[性質 5]すなわち式（20）ほかの関係が成立することを数値例として確認する．ここでも表内に付与した丸数字の番号に合わせて説明する．

① **クラスター間変動** $S_B(k)$：式（20）にみるように総変動に占めるクラスター間変動の推移を示している．この値は同時に階層的分類のまとまりの程度の指標でもあるので変動比を使うとよい（しかし，単調に変化する）．

② **カイ二乗統計量のチェック** $\chi_p^2(k)$：変動にクロス表の総和，ここでは全回答者数（N=1,284）を乗ずると（そのクラスターの階層水準での）“カイ二乗統計量”（$\chi_p^2(k)$）となる．[クラスター間変動]×[1284]＝[そのクラスター階層水準での圧縮化したクロス表のカイ二乗統計量]，これを求めた欄．

③ **カイ二乗統計量**:上の情報を各クラスターの階層水準で得た**圧縮化したクロス表**から求めた“カイ二乗統計量”がこの欄の数値である．これが②に一致していることが確認できる．つまり，クラスター化の階層水準とは“カイ二乗統計量の**分解**（あるいは併合）”に対応していることがわかる．

④ **クラスター内変動** $S_w(k,l)$：5 群の例でみたように，それぞれのクラスター数だけクラスター内変動があるが，これの和がこの欄の数値である．10 群のとき，個々の分類対象（レストラン）が個々のクラスター（singleton）であるから個々のクラスター内変動=0 となる．分類対象あるいはクラスターの併合が進むにつれて単調にこの値は増える．

⑤ **チェック**：式（20）の確認を行った結果．[クラスター間変動]＋[クラスター内変動の和]＝[固有値の和]がなり立つ．[$S_T = S_B(k) + \sum_{l=1}^{k} S_w(k,l)$]

表 19　クラスター間変動，クラスター内変動，ピアソンのカイ二乗統計量の関係

	①	②＝①×1,284（s）	③クロス表から算出のとき	④	⑤チェック
クラスター数（k）	クラスター間変動 $S_B(k)$	カイ二乗統計量に相当 $\chi_p^2(k)$	カイ二乗統計量 $\chi_p^2(k)$	クラスター内変動の和 $\sum_{l=1}^{k} S_w(k,l)$	①＋④ S_T [総変動＝固有値の和]
2	0.1584	203.4049	203.405	0.0992	0.2576
3	0.2263	290.5641	290.564	0.0313	0.2576
4	0.2417	310.3081	310.308	0.0159	0.2576
5	0.2514	322.7989	322.799	0.0062	0.2576
6	0.2536	325.6506	省略	0.0040	0.2576
7	0.2553	327.7667	省略	0.0024	0.2577
8	0.2569	329.8647	省略	0.0008	0.2577
9	0.2574	330.5363	省略	0.0003	0.2577
10	0.2577	330.8598	330.860	0.0000	0.2577

確認 2:

表 20 に，クラスター化過程におけるこれらの各統計量の関係を要約した．ここで所与のクロス表（表 3）から出発したあと，$k=5$（群）から $k=1$（群）までの履歴を一覧にしてある．なお，$k=9\sim6$（群）までは省略したが，とうぜん同じような関係がなり立つ．表 18 の下方の行から上に向かって，カイ二乗統計量，クラスター間変動，クラスター内変動がどう変化するかがわかるであろう．またこれら表 19，表 20 で，クラスター化の進行に伴う統計量（カイ二乗統計量，クラスター間変動，固有値とその和）の関係が読み取れるであろう．

はじめに述べたように，ここで用いた二元データ表の寸法は小さい．しかしデータ表の寸

法に関係なく，2 元表に対して対応分析とクラスター化を適用する際には上に述べた仕組みで統一的に処理が行われる．よって得られた統計量，結果の解釈はここで述べた考え方が適用される．

表 20　クラスター化の履歴の要約情報

＜はじめのクロス表＞　　＜統計量と生成される圧縮化クロス表の履歴＞

	(*) 10 群としたことに相当	工夫・サービス	味	量
10 群	さとみ	42	46	7
	ムガール	49	44	16
	バッハ	48	76	18
	ラ・マレ	49	82	15
	コルシカ	32	77	13
	ロゴスキー	48	35	42
	かりや	105	35	38
	いりふね	98	25	32
	クラーク	34	14	54
	きくみ	35	8	67

⇓　⇓

$\chi_p^2 = \chi_p^2(10) = 330.860$

（クロス表から得たカイ二乗統計量）

総変動（分類前）
0.2577

0.257679×1284＝
330.860

（総変動＝固有値の和）

＜5 群に分類後の併合を以下で追跡＞

	クラスター化履歴	工夫・サービス	味	量
5 群	{さとみ，ムガール}	91	90	23
	{バッハ，コルシカ，ラ・マレ}	129	235	46
	{ロゴスキー}	48	35	42
	{いりふね，かりや}	203	60	70
	{クラーク，きくみ}	69	22	121

⇓　⇓

$\chi_p^2(5) = 322.799$

クラスター間変動
$S_B(5) = 0.2514$

0.251401×1284＝
322.799

4 群	{さとみ，バッハ，ムガール，コルシカ，ラ・マレ}	220	325	69
	{ロゴスキー}	48	35	42
	{いりふね，かりや}	203	60	70
	{クラーク，きくみ}	69	22	121

⇓　⇓

$\chi_p^2(4) = 310.308$

クラスター間変動
$S_B(4) = 0.2417$

0.241673×1284＝
310.308

3 群	{さとみ，バッハ，ムガール，コルシカ，ラ・マレ}	220	325	69
	{いりふね，ロゴスキー，かりや}	251	95	112
	{クラーク，きくみ}	69	22	121

⇓　⇓

$\chi_p^2(3) = 290.564$

クラスター間変動
$S_B(3) = 0.2263$

0.226296×1284＝
290.564

2 群	{さとみ，バッハ，ムガール，コルシカ，ラ・マレ}	220	325	69
	{いりふね，クラーク，ロゴスキー，きくみ，かりや}	320	117	233

⇓　⇓

$\chi_p^2(2) = 203.405$

クラスター間変動
$S_B(2) = 0.1584$

0.158415×1284＝
203.405

1 群	{さとみ，バッハ，ムガール，コルシカ，ラ・マレ} {いりふね，クラーク，ロゴスキー，きくみ，かりや}	540	442	302

$\chi_p^2(1) = 0.0$

5. 応用事例

ここでは，ある調査研究に関連して行ったウェブ調査でえた意識調査データの一部を用いる．これを例として，おもにクラスター化法の利用方法とそれに関連のことがらを述べる．研究の課題は，やや漠然としたことで，「ひとは，いわゆる“情報”をどのように捉えているだろうか」といった内容である（詳細は省略）．

5.1 調査の概要

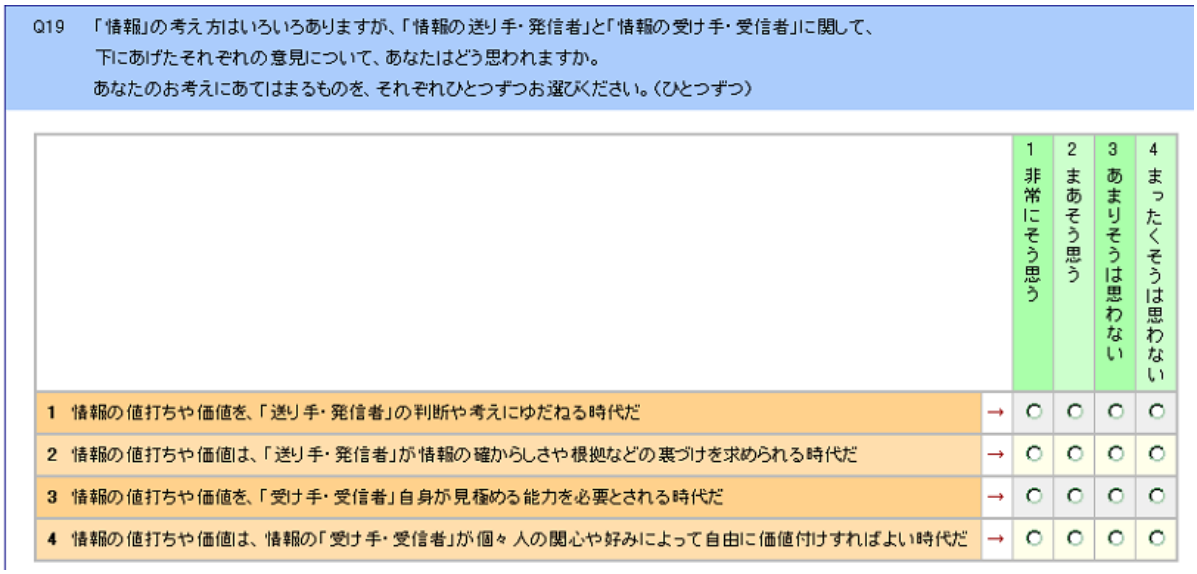
はじめにこの調査の概要を簡単に記す．

調査テーマ：「情報に関する調査」（実験調査）
調査方式：ウェブ調査
実施期間： 2011 年 09 月 09 日 17:00 ～ 2011 年 09 月 13 日 09:00 まで
ウェブ・パネル：非公募型パネル（部分的に確率的パネル）
予想回答所要時間：約 20 分
計画標本数：766（人）［男性（412），女性（354）］
回収標本数：347（人）［男性（175），女性（172）］
有効回収率：45.3（%）

ここで，回答者の年齢分布だけを示すと，以下のようになっている．

サンプル数	15～19歳	20～24歳	25～29歳	30～34歳	35～39歳	40～44歳	45～49歳	50～54歳	55～59歳	60～64歳	65～69歳
347	22 6.3	25 7.2	32 9.2	39 11.2	39 11.2	44 12.7	25 7.2	26 7.5	29 8.4	39 11.2	27 7.8

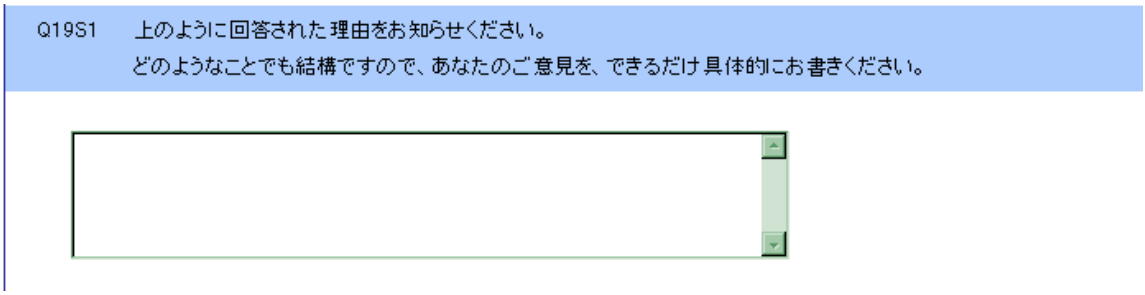
この調査の電子調査票から，ここで分析に用いる選択肢型質問と自由回答質問のレイアウトをあげておく．



Q19 「情報」の考え方はいろいろありますが、「情報の送り手・発信者」と「情報の受け手・受信者」に関して、
下にあげたそれぞれの意見について、あなたはどう思われますか。
あなたのお考えにあてはまるものを、それぞれひとつずつお選びください。（ひとつずつ）

		1 非常にそう思う	2 まあそう思う	3 あまりそうは思わない	4 まったくそうは思わない
1 情報の値打ちや価値を、「送り手・発信者」の判断や考えにゆだねる時代だ	→	○	○	○	○
2 情報の値打ちや価値は、「送り手・発信者」が情報の確からしさや根拠などの裏づけを求められる時代だ	→	○	○	○	○
3 情報の値打ちや価値を、「受け手・受信者」自身が見極める能力を必要とされる時代だ	→	○	○	○	○
4 情報の値打ちや価値は、情報の「受け手・受信者」が個々人の関心や好みによって自由に価値付けすればよい時代だ	→	○	○	○	○

（1）選択肢型質問の例
図 13 電子調査票の一部（ここで用いる質問文）



(2)自由回答質問の例
図 13 電子調査票の一部(ここで用いる質問文)

＜分析に用いた選択肢型質問＞ (質的変数)

Q19.「情報」の考え方はいろいろありますが、「情報の送り手・発信者」と「情報の受け手・受信者」に関して、下にあげたそれぞれの意見について、あなたはどう思われますか。
あなたのお考えにあてはまるものを、それぞれひとつずつお選びください。(ひとつずつ)

Q19_1：情報の値打ちや価値を、「送り手・発信者」の判断や考えにゆだねる時代だ
Q19_2：情報の値打ちや価値は、「送り手・発信者」が情報の確からしさや根拠などの裏づけを求められる時代だ
Q19_3：情報の値打ちや価値を、「受け手・受信者」自身が見極める能力を必要とされる時代だ
Q19_4：情報の値打ちや価値は、情報の「受け手・受信者」が個々人の関心や好みによって自由に価値付けすればよい時代だ

＜分析に用いた自由回答質問＞ (構成要素変数の元とする自由回答)

Q19S1：上のように回答された理由をお知らせください。
どのようなことでも結構ですので、あなたのご意見を、できるだけ具体的にお書きください。

(注10)　ここでの自由回答質問のワーディングは，かなり漠然とした問い方に思えるかもしれない．質問 Q19 は，回答者にとって，やや解釈・理解がむずかしいだろうと考え，続いて設けた 4 つの質問文のワーディングをきっかけとして回答を導く，つまり回答者に対して意図的に情報提供を行ったうえで，自由回答を書いてもらうように設計してある．

5.2 集計結果の観察(一部)

ここで，上の 4 つの選択肢型質問への回答頻度と回答比率を要約した（表 21）．これらをみると，4 つの質問への回答傾向にそれぞれ特徴があることがみえるだろう（回答比率が 1 位と 2 位のセルをボールド表記とした）．これらの回答傾向と，Q19S1 とした自由回答質問で得た内容との関連を調べる．

たとえば，「“Q19_4：情報の値打ちや価値は、情報の「受け手・受信者”が個々人の関心や好みによって自由に価値付けすればよい時代だ」についてみると，約 70%（21%＋49.3%）の人は「そう思う」と考え，残りの約 30%（23.6%＋6.1%）は「そうは思わない」という回答傾向にある．

ではこれに回答のあと，それに続く自由回答質問への傾向がどのようなものかを，「そう思う」人たちと「そうは思わない」という人たちの自由回答データと合わせて考えてみよう．

WordMiner ではこれを「(構成要素変数) ×（質的変数)」のデータ表から出発することで

分析を行うことに相当する．

表 21　集計表　　[有効回収数：n=347（人）]

質問文	非常にそう思う	まあそう思う	あまりそうは思わない	まったくそうは思わない	合計
Q19_1. 判断や考えにゆだねる時代	29	91	159	68	347
	8.4	**26.2**	**45.8**	19.6	(%)
Q19_2. 情報の確からしさや根拠などの裏づけ	117	165	47	18	347
	33.7	**47.6**	13.5	5.2	(%)
Q19_3. 自身が見極める能力を必要とされる時代	211	109	21	6	347
	60.8	**31.4**	6.1	1.7	(%)
Q19_4. 個々人の関心や好みによって自由に価値付けする時代	73	171	82	21	347
	21.0	**49.3**	**23.6**	6.1	(%)

5.3　分析

(1)用いる変数

用いる構成要素変数：

これは「Q19S1」の自由回答からえた構成要素（単語・語句）を用いる．自由回答文を分かち書き処理のあと，簡単な語句の編集を行う．ここでは，記号，句読点，助詞，それとごく少数の語句の削除を行った（例：「とくになし」「特にない」「とくにありません」などの削除）．つまりここでは“ほとんど単語・語句の**編集を行わず**”に分析を行う．またここでは，これまでに述べたさまざまの機能をどう使うか，クラスター化の結果分析に的を絞って説明を進める．実は，日本語の特性を考えると“句読点や助詞が重要”な役割をはたすことは分かっているが，ここでは，主な利用語・発語に注目して分析を行う，ということである．また，一度しか使われない語句あるいは利用頻度が少ない語句が多いのであるが（それがこの種のテキスト型データの特徴でもあるが），ここでは，「4 語以上」登場の構成要素数を用いる（構成要素数を選ぶ**閾値**を 4 以上と指定）．こうして確定した構成要素数は「213 語」（異なり構成要素数）である．ここで登場する構成要素の一部を，図に示した（図 15，図 16）．

用いる質的変数：

ここでは，上に挙げた質問文のうち“Q19_4：情報の値打ちや価値は、情報の「受け手・受信者」が個々人の関心や好みによって自由に価値付けすればよい時代だ”を用いる．この質問文の選択肢は上にあるように「非常にそう思う」「まあそう思う」「あまりそうは思わない」「まったくそうは思わない」である（順序尺度）．

ここでかりに，さらに分析を進めて，構成要素変数に対応させる質的変数を，他の 3 つの質問文（Q19_1～Q19_3）と替えてみることで，各質問文が，自由回答の内容とどう関連するかを観察することが可能である（実は，事後のそのような分析場面を想定してこれらの質問文の設計を行っている）．

構成要素変数の観察：

ここで図 14 の構成要素数の頻度分布を観察する．全構成要素数（3,323 語）に対して，分析に用いる**異なり構成要素数**が 213 語あり，それが 6.4%にあたる（ここでは，この程度の絞り込んだ少ない構成要素を対象に分析を行っていることに注意しよう）．また図 15 にある構成要素一覧，つまり回答者が記述した自由回答文から抽出・選出した語句の一覧をみる．

ここに閾値でスクリーニングした「4 語以上」，つまり「213 語」の単語語句がある．これを出現頻度の大きさに並べかえて，つまり図 15（左）のようにして，頻度数が多い方から 10 頻度までを順に選ぶと以下のようになる．

情報	する	ある	思う	**必要**	いる	**判断**	**自分**	だ	して	**受け手**
時代	**発信**	ない	その	**送り手**	もの	**正確**	では	**価値**	**見極める**	人
できる	べき	多い	**能力**	中	**発信者**	には	**正しい**	なって	思います	**責任**
それ	なる	ように	ため	なく	ならない	事	側	どれ	**氾濫**	される
しない	なので	**取捨選択**	受信者	いい	いけない	です	よって	**個人**	とは	
インターネット		**自由**	**大事**	だから	**メディア**	今	**選択**	しまう	それぞれ	どうか
よく	**意見**	何	**価値観**	**見極め**	**自分自身**	**信頼**	**力**	あり	いく	いろいろな
	した	どの	**鵜呑み**	**感じる**	**受けて**	**受け手側**	重要	色々	**不正確**	（以下，続く）．

細かい分析はさておいて，登場語句には，用意した質問文に含まれる語句が多数登場していることに気付くであろう（そうなることを想定している）．実はここで，回答者の回答全部を並べて観察すると，これらの語句がどう結合されて発語となったかも読み取れる．しかしそれがなくても，ここにある語句をつなげてみることでおおよその意見がどういう傾向にあるかがみえてくる．ここでの観察はここまでとしよう．

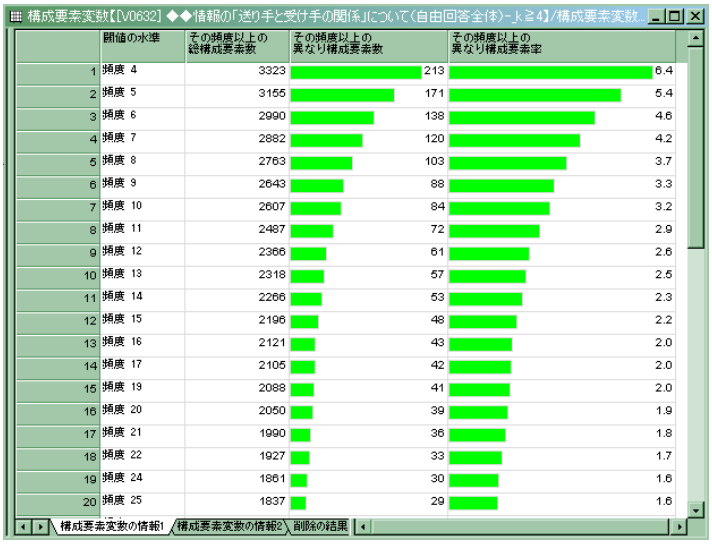


図 14　213 語の構成要素の頻度分布（4 語以上）

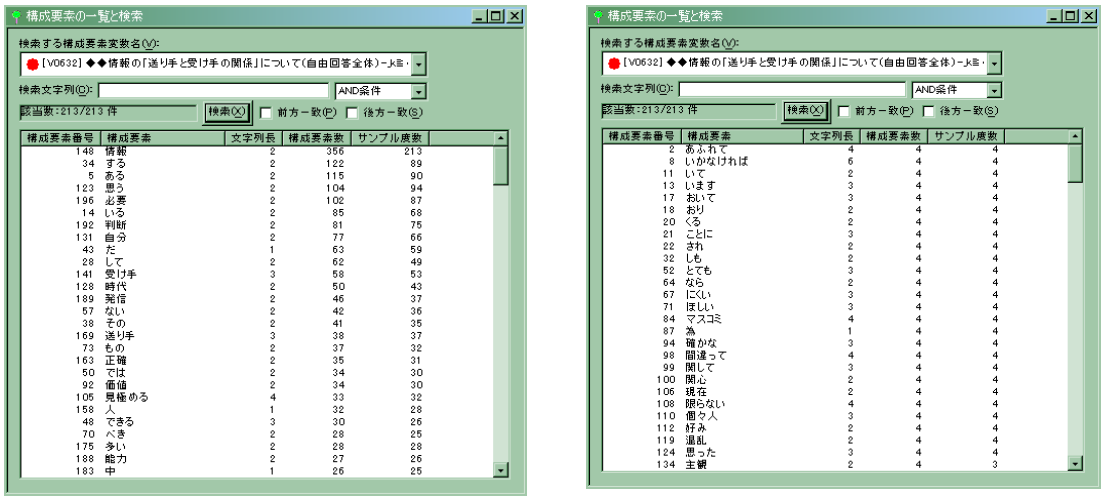
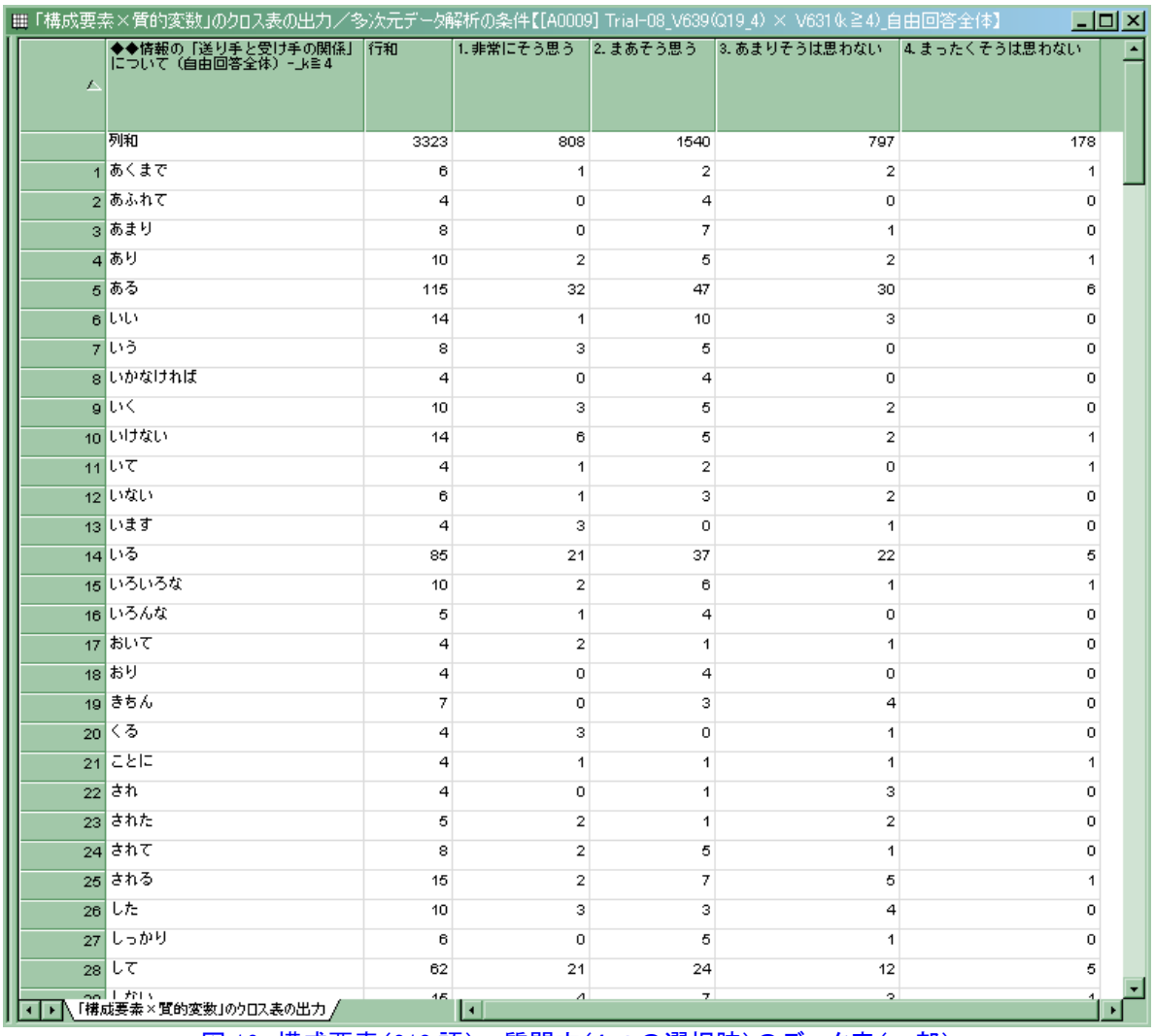


図 15　分析に用いた 213 語の構成要素一覧（頻度 4 語以上），検索機能で降順，昇順にソート閲覧

(2)対応分析による基本情報

分析対象とする“二元データ表”は，上に用意した質的変数と構成要素変数を用いる．これは，「(構成要素変数)×(質的変数)」＝「(213 語の構成要素)×(Q19_4 の 4 つの選択肢)」つまり**寸法が（213×4）の二元データ表**となるが，これが図 16 である．これに対応分析法を適用する．以下の説明では，なるべく WordMiner の出力情報を引用しながら説明する．

観察 1:データ表の確認



「構成要素×質的変数」のクロス表の出力／多次元データ解析の条件【[A0009] Trial-08_V639(Q19_4) × V631(k≧4)_自由回答全体】

	◆◆情報の「送り手と受け手の関係」について（自由回答全体）-_k≧4	行和	1.非常にそう思う	2.まあそう思う	3.あまりそうは思わない	4.まったくそうは思わない
	列和	3323	808	1540	797	178
1	あくまで	6	1	2	2	1
2	あふれて	4	0	4	0	0
3	あまり	8	0	7	1	0
4	あり	10	2	5	2	1
5	ある	115	32	47	30	6
6	いい	14	1	10	3	0
7	いう	8	3	5	0	0
8	いかなければ	4	0	4	0	0
9	いく	10	3	5	2	0
10	いけない	14	6	5	2	1
11	いて	4	1	2	0	1
12	いない	6	1	3	2	0
13	います	4	3	0	1	0
14	いる	85	21	37	22	5
15	いろいろな	10	2	6	1	1
16	いろんな	5	1	4	0	0
17	おいて	4	2	1	1	0
18	おり	4	0	4	0	0
19	きちん	7	0	3	4	0
20	くる	4	3	0	1	0
21	ことに	4	1	1	1	1
22	され	4	0	1	3	0
23	された	5	2	1	2	0
24	されて	8	2	5	1	0
25	される	15	2	7	5	1
26	した	10	3	3	4	0
27	しっかり	6	0	5	1	0
28	して	62	21	24	12	5

「構成要素×質的変数」のクロス表の出力

図 16　構成要素(213 語)×質問文(4 つの選択肢)のデータ表(一部)

観察 2:固有値，寄与率，累積寄与率の確認

解析結果の基礎情報(固有値、寄与率、累積寄与率)／多次元データ解析の条件【[A0009] Tria...

	固有値	寄与率	累積寄与率
1	0.0792	36.66	36.66
2	0.0786	36.37	73.03
3	0.0583	26.97	100.00

解析結果の基礎情報(固有値、寄与率、累積寄与率)

図 17　固有値，寄与率，累積寄与率の情報

ここで固有値の数は min{213 語，4 つの選択肢}－1＝3 個（3 根）まで得られる．これを図 17 が示している．また，累積寄与率から，はじめの 2 成分で全情報（総変動）の約 73%を占める．

観察 3:成分スコアとその布置図の観察

データ表の行側と列側の“成分スコア”に関する詳しい情報を WordMiner は出力する．基本は次の 2 つの要約表である（図 18，図 19）．

構成要素変数の統計値(成分スコア、寄与度他)／多次元データ解析の条件【[A0009] Trial-08_V639(Q19_4) × V631(k≧4)_自由回答全体】

	◆◆情報の「送り手と受け手の関係」について（自由回答全体）-_k≧4	構成要素数構成比	距離	成分スコア1	成分スコア2	成分スコア3	絶対寄与度1	絶対寄与度2	絶対寄与度3	相対寄与度1	相対寄与度2	相対寄与度3
1	あくまで	0.002	0.34	0.3244	-0.2539	-0.4076	0.2398	0.1480	0.5148	0.3133	0.1919	0.4948
2	あふれて	0.001	1.16	-1.0022	0.2741	-0.2796	1.5264	0.1151	0.1614	0.8676	0.0649	0.0675
3	あまり	0.002	0.72	-0.8243	0.0302	-0.1917	2.0652	0.0028	0.1518	0.9475	0.0013	0.0513
4	あり	0.003	0.06	-0.0021	0.0388	-0.2364	0.0000	0.0058	0.2886	0.0001	0.0263	0.9736
5	ある	0.035	0.01	0.0973	-0.0123	0.0616	0.4137	0.0067	0.2253	0.7059	0.0113	0.2828
6	いい	0.004	0.31	-0.5515	-0.0858	-0.0433	1.6174	0.0395	0.0135	0.9705	0.0235	0.0060
7	いう	0.002	0.42	-0.2368	0.5798	0.1703	0.1704	1.0296	0.1198	0.1331	0.7980	0.0689
8	いかなければ	0.001	1.16	-1.0022	0.2741	-0.2796	1.5264	0.1151	0.1614	0.8676	0.0649	0.0675
9	いく	0.003	0.08	-0.1052	0.1284	0.2209	0.0421	0.0631	0.2520	0.1450	0.2159	0.6391
10	いけない	0.004	0.21	0.2953	0.3390	0.0940	0.4638	0.6158	0.0639	0.4135	0.5446	0.0419
11	いて	0.001	0.96	0.2761	0.4578	-0.8231	0.1158	0.3210	1.3992	0.0791	0.2176	0.7033
12	いない	0.002	0.12	-0.1876	-0.2405	0.1547	0.0802	0.1328	0.0741	0.3010	0.4944	0.2045
13	います	0.001	1.57	0.8845	0.3976	0.7959	1.1888	0.2421	1.3085	0.4970	0.1004	0.4025
14	いる	0.026	0.00	0.0511	-0.0343	0.0003	0.0844	0.0382	0.0000	0.6900	0.3100	0.0000
15	いろいろな	0.003	0.17	-0.1445	0.2340	-0.3067	0.0793	0.2095	0.4858	0.1230	0.3226	0.5544

構成要素変数の統計値(成分スコア、寄与度他)／多次元データ

図 18　構成要素（用いた単語群）の成分スコアほか

質的変数の統計値(成分スコア、寄与度他)／多次元データ解析の条件【[A0009] Trial-08_V639(Q19_4) × V631(k≧4)_自由回答全体】

	◆◆情報の「送り手と受け手の関係」について-情報の値打ちや価値は、情報の「受け手・受信者」が個々人の関心や好みによって自由に価値付けすればよい時代だ（選択肢）-質的変数	構成要素数構成比	距離	成分スコア1	成分スコア2	成分スコア3	絶対寄与度1	絶対寄与度2	絶対寄与度3	相対寄与度1	相対寄与度2	相対寄与度3
1	1. 非常にそう思う	0.243	0.23	0.2924	0.3054	0.2221	26.2483	28.8480	20.5883	0.3749	0.4088	0.2163
2	2. まあそう思う	0.463	0.09	-0.2821	0.0769	-0.0675	46.5516	3.4823	3.6224	0.8838	0.0656	0.0506
3	3. あまりそうは思わない	0.240	0.25	0.1185	-0.4702	0.1022	4.2503	67.4685	4.2968	0.0572	0.9003	0.0425
4	4. まったくそうは思わない	0.054	1.12	0.5826	0.0543	-0.8820	22.9498	0.2012	71.4924	0.3030	0.0026	0.6944

質的変数の統計値(成分スコア、寄与度他)／多次元データ解析の条件【[A0009] Trial-08_V639(Q19_4) × V631(k≧4)_自由回答全体】

図 19　質的変数（質問文の 4 つの選択肢）の成分スコアほか

ここで，はじめの 2 成分の同時布置図をみる（よって，寄与率を目安とすると全情報の約 73%をこの 2 次元空間内で観察している）. 図 20 がそれであるが，ここでは横軸を第 1 成分，縦軸が第 2 成分と指定した．またここでは，構成要素と質問文の 4 つの選択肢との成分スコアを**“同時布置図”**としてある．これで，質問文の 4 つの選択肢と，ここで用いた 213 語の単語とのおおまかな関連が見えてくる. 細かい単語・語句の拾い出しはここでは行わないが，たとえば，「そう思う」側に特徴的な語句と，「そうは思わない」側に分布する語句があることが見えてくる．ここらの対応をより詳しく調べるための情報を，WordMiner は用意してある．こうした視覚化は初動探査の基本ツールであって，これをガイドとして，さらに詳しい吟味を行う．しかし単語数が増えると，布置図での観察には限界がある．この図 20 でもすでにかなりの単語が重なり視認がむずかしい．そこで，WordMiner には別の視点からこれを観察するツールをいろいろ用意してある．

（注11）　構成要素（単語・語句）と質問文選択肢との同時布置図で，これら両者が同じ空間にあるものとして，互いの布置の点の関係を，距離的に近いあるいは遠いという見方は正しくはない．これの観察方法は実はやや面倒であり，注意を要する．これに

ついては，文献を参照することを勧める．

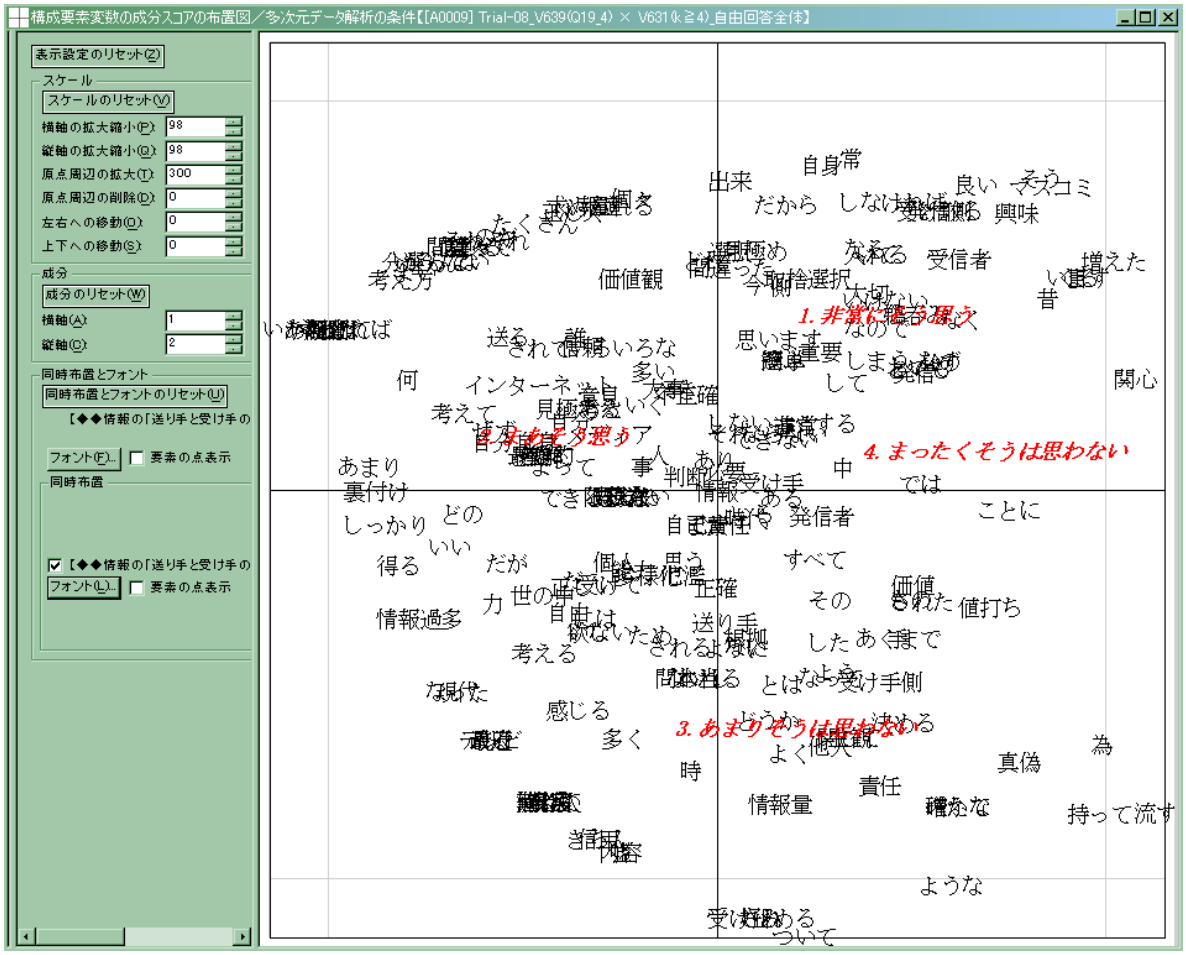


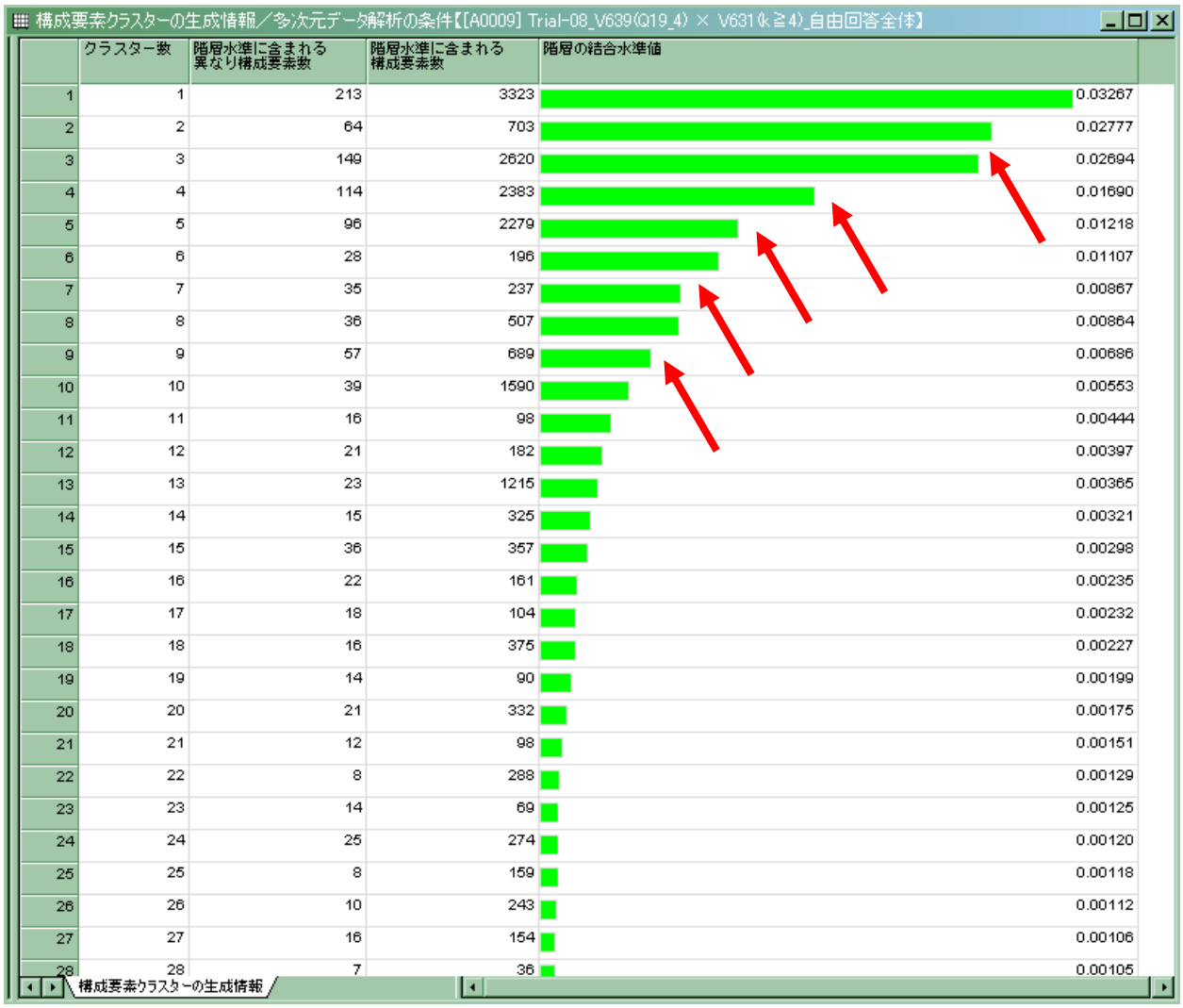
図 20　構成要素群と質問文の 4 つの選択肢の成分スコアの同時布置図

観察 4：構成要素のクラスター化

次に構成要素のクラスター化を調べる．**「構成要素のクラスター生成情報」**から，以下の出力情報が得られる（図 21）．ここで図の何カ所かに矢印を入れてみた．すでに述べたように，かりにクラスターらしい構造が予想されると，クラスター化過程でクラスター間変動の変化，つまり図内の棒グラフに大きな変化（ギャップ）があるだろうと考える．すでに述べたように，厳密なルールではないが，クラスター間変動や変動比なども参考にして，1 つの目安にする．

この例のようなやや大きな変化の部分があればよいが，かりにこの棒グラフの変化が滑らかで変化がない場合には，顕著なクラスター構造が存在しないかもしれないと考える．ここらはいずれも，発見的かつ経験則的であって，理論的に厳密な考えではない．

ここでは，クラスター化履歴の図の観察と，必要ならば前に述べたクラスター間変動，クラスター内変動の和，総変動との比（変動比）．などを参考にしてクラスター数を決める．この例では，たとえば「8 群」（クラスター数：$k=8$）としてみよう．もちろん探索的にクラスター数を替えて吟味することは望ましいことである．その意味で，計算を行う際に，この図を参考に複数のクラスター数を指定しておくのもよいだろう（この例では，このグラフの観察のあと．2 群から 18 群までを指定して再計算した）．



構成要素クラスターの生成情報／多次元データ解析の条件【[A0009] Trial-08_V639(Q19_4) × V631(k≧4)_自由回答全体】

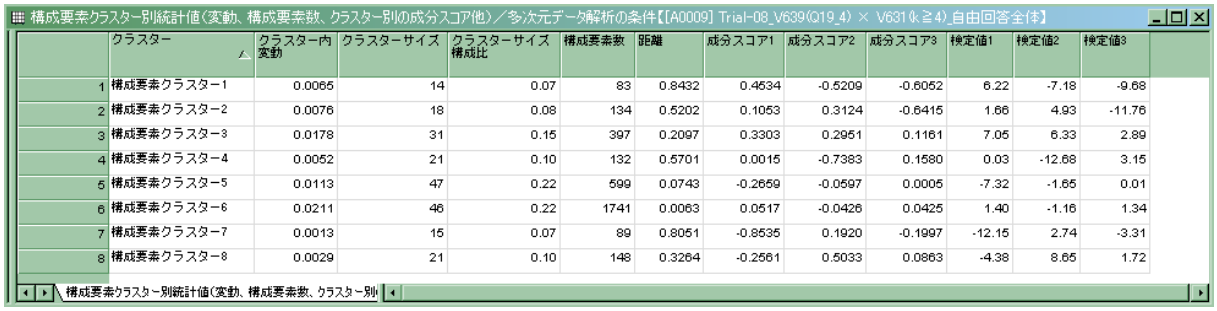
	クラスター数	階層水準に含まれる異なり構成要素数	階層水準に含まれる構成要素数	階層の結合水準値
1	1	213	3323	0.03267
2	2	64	703	0.02777
3	3	149	2620	0.02694
4	4	114	2383	0.01690
5	5	96	2279	0.01218
6	6	28	196	0.01107
7	7	35	237	0.00867
8	8	36	507	0.00864
9	9	57	689	0.00686
10	10	39	1590	0.00553
11	11	16	98	0.00444
12	12	21	182	0.00397
13	13	23	1215	0.00365
14	14	15	325	0.00321
15	15	36	357	0.00298
16	16	22	161	0.00235
17	17	18	104	0.00232
18	18	16	375	0.00227
19	19	14	90	0.00199
20	20	21	332	0.00175
21	21	12	98	0.00151
22	22	8	288	0.00129
23	23	14	69	0.00125
24	24	25	274	0.00120
25	25	8	159	0.00118
26	26	10	243	0.00112
27	27	16	154	0.00106
28	28	7	36	0.00105

構成要素クラスターの生成情報

図 21　クラスター化の履歴の観察

(1) 成分スコアの観察，検定値の吟味など

すでに説明した知識の助けを借りて，どのクラスターがどの成分に関連があり，またクラスター相互の関係はどうなっているか，特徴的なクラスターはどれか，といった情報を観察する（たとえば，検定値の大きさを目安に観察，図内の楕円など）．ここでの注意点は，はじめのレストランの例と違って，データ表の次元数が増えていることである．よって，各成分軸を変えて図を観察することがコツである（図 22～24）．



構成要素クラスター別統計値(変動、構成要素数、クラスター別の成分スコア他)／多次元データ解析の条件【[A0009] Trial-08_V639(Q19_4) × V631(k≧4)_自由回答全体】

	クラスター	クラスター内変動	クラスターサイズ	クラスターサイズ構成比	構成要素数	距離	成分スコア1	成分スコア2	成分スコア3	検定値1	検定値2	検定値3
1	構成要素クラスター1	0.0065	14	0.07	83	0.8432	0.4534	-0.5209	-0.6052	6.22	-7.18	-9.68
2	構成要素クラスター2	0.0076	18	0.08	134	0.5202	0.1053	0.3124	-0.6415	1.66	4.93	-11.76
3	構成要素クラスター3	0.0178	31	0.15	397	0.2097	0.3303	0.2951	0.1161	7.05	6.33	2.89
4	構成要素クラスター4	0.0052	21	0.10	132	0.5701	0.0015	-0.7383	0.1580	0.03	-12.68	3.15
5	構成要素クラスター5	0.0113	47	0.22	599	0.0743	-0.2659	-0.0597	0.0005	-7.32	-1.65	0.01
6	構成要素クラスター6	0.0211	46	0.22	1741	0.0063	0.0517	-0.0426	0.0425	1.40	-1.16	1.34
7	構成要素クラスター7	0.0013	15	0.07	89	0.8051	-0.8535	0.1920	-0.1997	-12.15	2.74	-3.31
8	構成要素クラスター8	0.0029	21	0.10	148	0.3264	-0.2561	0.5033	0.0863	-4.38	8.65	1.72

構成要素クラスター別統計値(変動、構成要素数、クラスター別

図 22　クラスター別の情報（成分スコア，検定値ほか）

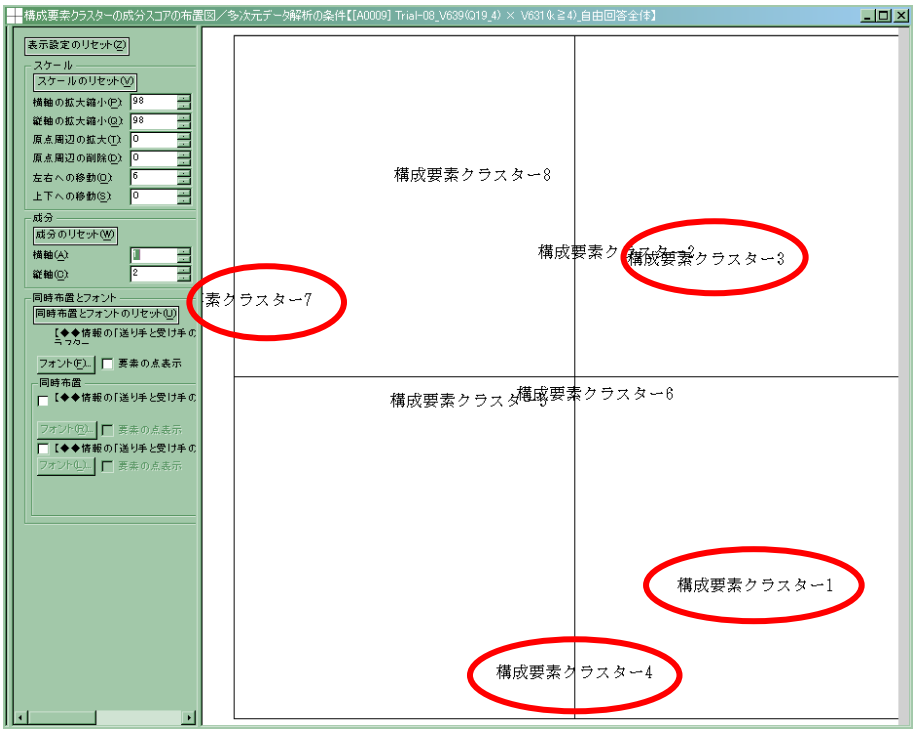


図 23　第 1 成分と第 2 成分の観察

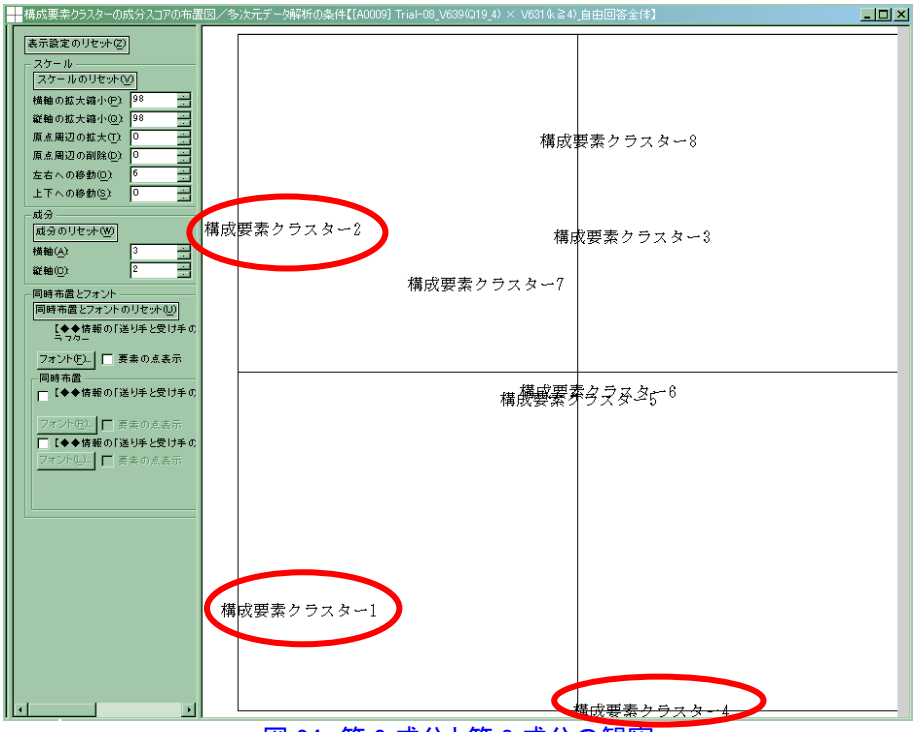
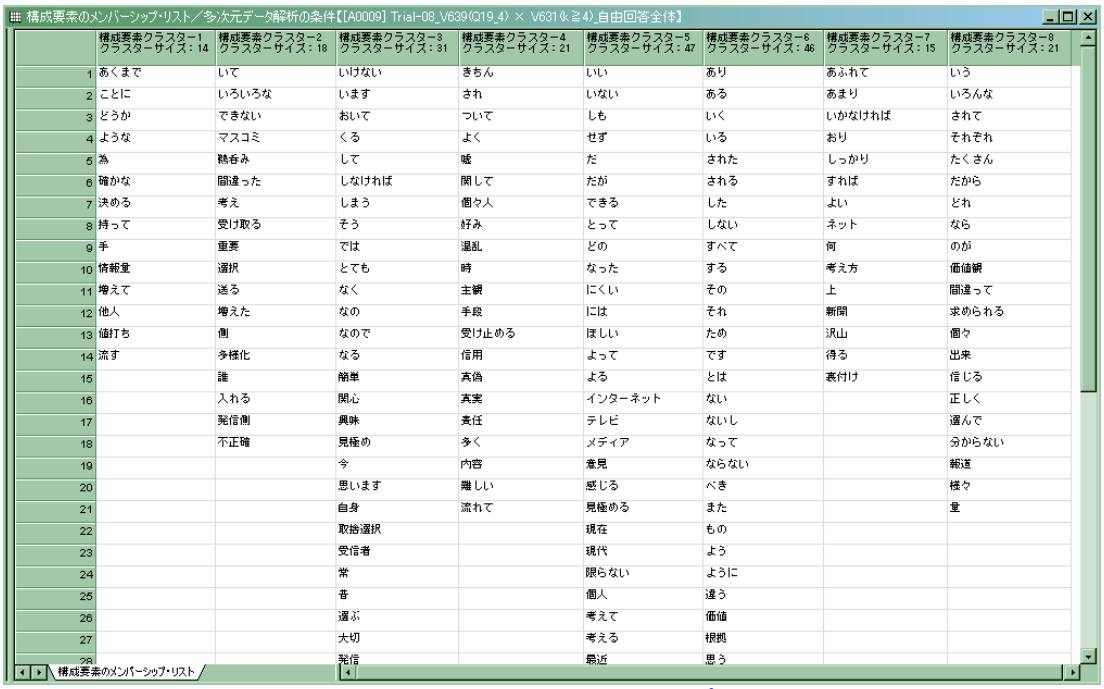


図 24　第 2 成分と第 3 成分の観察

(2)各クラスターのメンバーシップの確認

さらに，求めた 8 群の各クラスターにはどのような構成要素（単語，語句）が含まれるのか，これを**「構成要素のメンバーシップリスト」**で観察する（図 25）．各クラスターのクラスターサイズをみて，クラスター間の分布の様子と図 22 の表と併せて観察する．たとえば，クラスター1 には14 個の構成要素があってその内容は図 25 の左のはじめの欄のようになる．


構成要素のメンバーシップ・リスト／多次元データ解析の条件【[A0009] Trial-08_V639(Q19_4) × V631(k≧4)_自由回答全体】

	構成要素クラスター1 クラスターサイズ：14	構成要素クラスター2 クラスターサイズ：18	構成要素クラスター3 クラスターサイズ：31	構成要素クラスター4 クラスターサイズ：21	構成要素クラスター5 クラスターサイズ：47	構成要素クラスター6 クラスターサイズ：46	構成要素クラスター7 クラスターサイズ：15	構成要素クラスター8 クラスターサイズ：21
1	あくまで	いて	いけない	きちん	いい	あり	あふれて	いう
2	ことに	いろいろな	います	され	いない	ある	あまり	いろんな
3	どうか	できない	おいて	ついて	しも	いく	いかなければ	されて
4	ような	マスコミ	くる	よく	せず	いる	おり	それぞれ
5	為	鵜呑み	して	嘘	だ	された	しっかり	たくさん
6	確かな	間違った	しなければ	関して	だが	される	すれば	だから
7	決める	考え	しまう	個々人	できる	した	よい	どれ
8	持って	受け取る	そう	好み	とって	しない	ネット	なら
9	手	重要	では	混乱	どの	すべて	何	のが
10	情報量	選択	とても	時	なった	する	考え方	価値観
11	増えて	送る	なく	主観	にくい	その	上	間違って
12	他人	増えた	なの	手段	には	それ	新聞	求められる
13	値打ち	側	なので	受け止める	ほしい	ため	沢山	個々
14	流す	多様化	なる	信用	よって	です	得る	出来
15		誰	簡単	真偽	よる	とは	裏付け	信じる
16		入れる	関心	真実	インターネット	ない		正しく
17		発信側	興味	責任	テレビ	ないし		選んで
18		不正確	見極め	多く	メディア	なって		分からない
19			今	内容	意見	ならない		報道
20			思います	難しい	感じる	べき		様々
21			自身	流れて	見極める	また		皇
22			取捨選択		現在	もの		
23			受信者		現代	よう		
24			常		限らない	ように		
25			昔		個人	違う		
26			選ぶ		考えて	価値		
27			大切		考える	根拠		
28			発信		最近	思う		

構成要素のメンバーシップ・リスト

図 25　各クラスターのメンバーシップの観察

観察 5：構成要素と質問文の関係ほか

以上で，WordMiner が提供するクラスタリングに関連する主な情報の観察の手順を示した．いきなり大きな寸法のデータ表の分析に取り組むのではなく，データ表の構造や分析結果がある程度みえるようなミニチュア・データを用意して，クラスター化が何を行っているのかを体験することがよいだろう．

また，WordMiner はクラスター化だけでなく，対応分析を巡るさまざまな応用機能があって，それらを含めて総合的に分析を進めるとよい．たとえば，対応分析とクラスター化のあと，処理結果がフローティング・フレーム内にすべて表示されている．いまの例であると，図 26 のようなフレームが得られる．とくにここで，**「質的変数の構成要素の有意性テスト」**の項の情報は，WordMiner の備える有用なツール群で，ここの利用方法を理解することが，WordMiner を使いこなす 1 つの鍵である．

① カテゴリー別の情報要約：ここで用いた質的変数つまり質問文の内容確認
② 頻度による有意性テスト要約：有意なサンプルの要約
③ 頻度による有意性テスト要約：有意な構成要素の要約
④ 距離による有意性テスト要約：有意なサンプルの要約
⑤ 頻度による有意性テスト要約：サンプル別一覧
⑥ 頻度による有意性テスト要約：構成要素別一覧
⑦ 距離による有意性テスト要約：サンプル別一覧

これらの情報の見方，解釈については，別の資料を用意したので，それを参照していただきたい（[7]）．ここでは①と③の簡単な例を示そう．

(1)「カテゴリー別の情報要約」の観察

ここでは，①の**「カテゴリー別の情報要約」**の出力を調べる（図 27）．ここにみるように，分析に用いたサンプル数は，はじめに集計でみた（表 21 の）結果とは異なる．これは，構成要素の編集や，出現構成要素数の選別，条件を満たさないサンプルの除外などの理由から，サンプル数が目減りしているからである．また，質問文の選択肢別に，編集前後の構成要素数や異なり構成要素数，構成比率などが集計されている．全体の構成要素のうちのどの程度

の構成要素を分析に用いたのか，それらが質問文にどう反映されたかなどが分かる．

たとえばここで，対象となった「サンプル数」（点線枠の丸数字①）にある構成要素つまり回答者の発語（点線枠の丸数字②）は，それぞれの選択肢に対してどのような内容だったのか，これを次に調べよう．

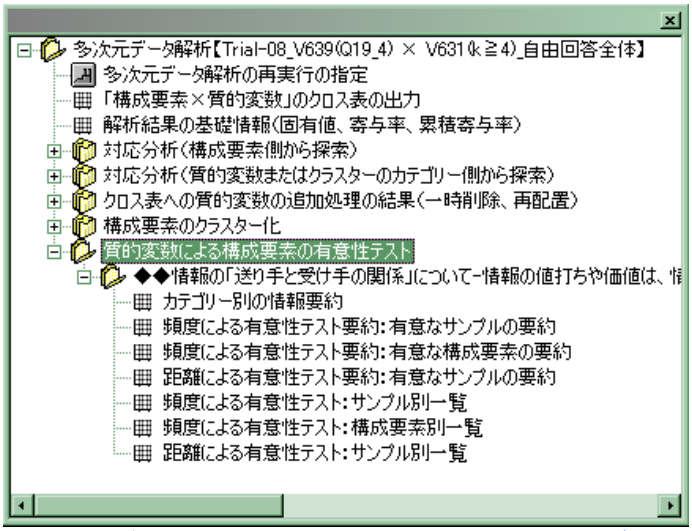


図 26　多次元データ解析の処理結果を占めるフローティング・フレーム

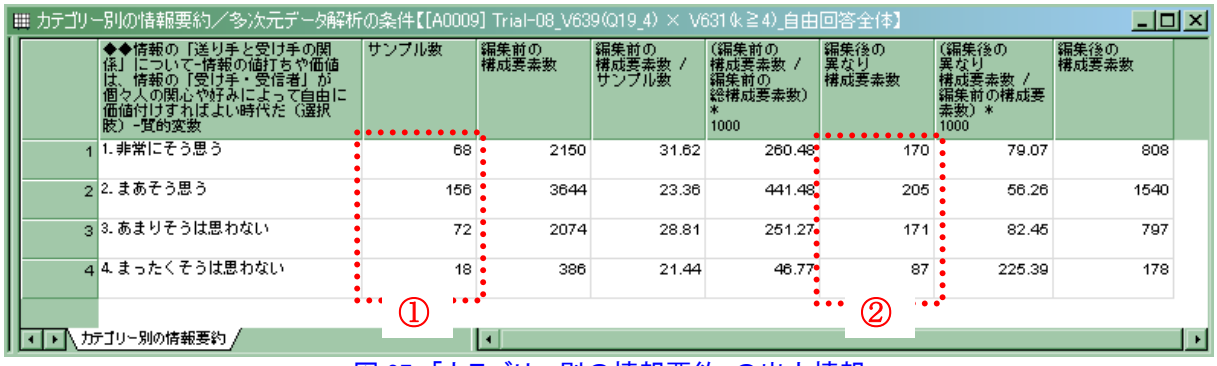


図 27「カテゴリー別の情報要約」の出力情報

(2)質問文の観察(質的変数の観察)

すでに述べたように，所与のデータ表の“行側のクラスター化”（ここでは構成要素）と“列側のクラスター化”のいずれも WordMiner では対応できる（対応分析の仕組みから当たり前のこと）．この例では，質的変数の質問文の選択肢は少ないので分類する意味はあまりない．一方，構成要素の回答分布（発言の内容）の構成を観測することは有効である．

たとえば，図 27 でみた集計結果の具体的な内容を**「質的変数による構成要素の有意性テスト」**の中で得られる**「頻度による有意性テストの要約」**情報から拾ってみる．質問文（Q19_4）の **4 つの選択肢**のそれぞれで意味のある（有意となりそうな）**“上位の語句群”**と反対にあまり寄与しないと思われる**“下位の語句群”**とを観察する．

WordMiner では，たとえばこれを図 28 のように要約情報として提供する（他の探査ツールもいろいろあるが，まずはこの要約から観察する）．

これに関連の情報の解釈についてはここでは述べない（別に資料が用意されているのでそれをみていただきたい）．ここでは，クラスター化と関連させてこれら情報の観察も必要であることだけを指摘しておこう．

この図 29 の情報をテキスト・ファイルとしてエクスポートして表 22 として再編集してみた．ここから，4 つの選択肢（「非常にそう思う」「まあそう思う」「あまりそうは思わない」「まったくそうは思わない」）のそれぞれに特徴的な語句の傾向がみえてくる．細かい説明は

省くが，この各選択肢の中の語句の関係を詳細に示す情報も得られるのである．語句をみると，それだけでは若干意味不明のものや，類似の語句の言い替えがあったりする．しかしはじめに述べたように，ここでは“ほとんど単語・語句の編集を行っていないこと”を思い出そう．しかし一方，“出現頻度が 3 語以下の単語・語句は分析では用いていないこと”にも注意しよう．つまりはそのようなある種の網をかけた（篩（ふるい）にかけた）情報からデータの特徴探査・抽出を行ったことになる．こうした探査的なアプローチを行えること（マイニングを行うこと）が WordMiner の特徴である．

図 28　質問文の 4 つの選択肢と関連のある単語群（上位と下位）

表 22　質問文の 4 つの選択肢について意味ある構成要素（単語群）の内容［図 28 から］

有意の順位	1.非常にそう思う サンプル数：68 異なり構成要素数：170	2.まあそう思う サンプル数：156 異なり構成要素数：205	3.あまりそうは思わない サンプル数：72 異なり構成要素数：171	4.まったくそうは思わない サンプル数：18 異なり構成要素数：87
上位 1	そう	すれば	責任	増えた
上位 2	興味	沢山	ついて	流す
上位 3	受信者	自分	ような	持って
上位 4	なる	何	され	値打ち
上位 5	する	ネット	好み	誰
上位 6	発信	あまり	受け止める	側
上位 7	います	考え方	ない	なく
上位 8	くる	裏付け	なって	鵜呑み
上位 9	目	あふれて	きちん	重要
上位 10	取捨選択	いかなければ	持って	不正確
上位 11	では	おり	信用	発信
上位 12	して	上	もの	価値
上位 13	思います	新聞	その	どうか
上位 14	自身	いい	嘘	選択
上位 15	なく	見極める	真偽	
上位 16	必要	それぞれ	内容	
上位 17	常	しっかり	よく	
上位 18	いけない	よい	時	
上位 19	見極め	できる	情報量	
上位 20	だから	情報過多	難しい	
上位 21	中	誰	ように	
上位 22	昔	どの	多く	
上位 23	大切	だ		
上位 24	良い	いろんな		
上位 25		選んで		
上位 26		送る		
上位 27		得る		
上位 28		分からない		

下位 25		中		
下位 24		ような		
下位 23		昔		
下位 22		して		
下位 21		ある		
下位 20		なって		
下位 19		なく		
下位 18		受信者		
下位 17		興味	しなければ	
下位 16	送り手	います	のが	
下位 15	きちん	くる	個々	
下位 14	信用	為	考え方	
下位 13	裏付け	関心	自身	
下位 12	だ	目	いう	
下位 11	できる	流す	すれば	
下位 10	いい	その	正しく	
下位 9	あまり	する	誰	
下位 8	すれば	真偽	なる	
下位 7	情報過多	発信	自分	
下位 6	情報量	価値	たくさん	
下位 5	誰	そう	どれ	
下位 4	難しい	責任	それぞれ	
下位 3	どうか	では	側	
下位 2	正しい	持って	だから	だ
下位 1	には	もの	選択	必要

（*）表頭のセル内に，各選択肢に含まれる回答者数と異なり構成要素数の情報がある．図 27 の要約表と比較しよう．

【キーワード】
対応分析法（Correspondence Analysis，Analyse des Correspondances），慣性（inertia），カイ二乗距離（Chi-square distance），ピアソンのカイ二乗統計量（Chi-square statistic），自動分類（automatic classification），クラスター分析（cluster analysis），凝集型階層的分類法（AHC : agglomerative hierarchical classification），分割型分類法（partitioning-type classification），*k*-平均法（*k*-means method），混合方式（mixed clustering approaches），成分スコア”（principal coordinates, coordinates），相互最近隣の規則（RNN : reciprocal nearest neighbours rule），二元のデータ表（two-way data table），クロス表，構成要素，構成要素変数，質的変数，プロフィル（profile），行のプロフィル（row profile），列のプロフィル（column profile），固有値，寄与率，累積寄与率，総変動・全分散（total inertia，total variance），デンドログラム・樹形図（dendrogram），階層の結合水準（hierarchical indices），クラスター内変動・クラスター内分散（within-cluster variances），クラスター間変動・クラスター間分散（between-cluster variances），クラスター間変動比，検定値と検定統計量（test statistic），単純無作為抽出（SRS : simple random sampling），復元抽出（SWR : sampling with replacement）と非復元抽出（SWOR : sampling without replacement），標本平均の分布，母集団と標本，標準化，正規分布，正規近似，クラスター数を決める目安，同時布置図，布置図

【参考文献】

[1] Brigitte Le Roux and Henry Rouanet (2004): *Geometrical Data Analysis – From Correspondence Analysis to Structural Data*, Dordrecht Kluwer.
[2] Brigitte Le Roux and Henry Rouanet (2010): *Multiple Correspondence Analysis*, Series: Quantitative Applications in the Social Sciences No.163, Sage Publications, Inc.
[3] Ludovic Lebart（1998）: *Exploring Textual Data*, .Kluwer Academic Publishers.
[4] Michael J. Greenacre (1984): *Theory and Applications of Correspondence Analysis*, Academic Press.
[5] Michael J. Greenacre (2007): *Correspondence Analysis in Practice* (second edition), Academic Press.
[6] Michael J. Greenacre (ed.) (2006): *Multiple Correspondence Analysis and Related Methods*, Chapman & Hall/CRC.
[7] 大隅昇，Ludovic Lebart 他（1994）: 記述的多変量解析法，日科技連出版社.
[8] 岩坪秀一（1987）: 数量化法の基礎，朝倉書店.

この他，テキスト・マイニング研究会ホームページから提供される各種の情報がある．とくに，対応分析法については，ホームページの「技術解説」の項から，解説文の pdf 形式のファイルがダウンロードできる．

◆テキスト・マイニング研究会ホームページ：
http://wordminer.comquest.co.jp/
◆技術解説： http://wordminer.comquest.co.jp/wmtips/analysis.html
1)「対応分析法･数量化法 III 類の考え方」http://wordminer.comquest.co.jp/wmtips/pdf/20060910_3.pdf
2)「よくある質問へのヒント」 http://wordminer.comquest.co.jp/wmtips/pdf/20060910_a_kaitei.pdf
・構成要素，異なり構成要素の分布の特性
・有意性テスト（とくに頻度による有意性テスト）
◆レシピ： http://wordminer.comquest.co.jp/wmtips/index.html

2012 年 11 月 23 日作成
2012 年 11 月 27 日更新
2013 年 1 月 19 日更新
2013 年 3 月 20 日更新

資料作成：大隅　昇　（ohsumi@ss.iij4u.or.jp）