補足資料

# 対応分析法とクラスター化法

－ WordMiner™を理解するために －

大隅　昇
文部科学省統計数理研究所

# Ⅰ．対応分析法とは？

## １．対応分析法（CA）

Benzécri により提唱された方法で，質的データの主成分分析と考えることもできる．数理的には林の数量化法Ⅲ類と同等手法と考えてよい.

- Correspondence Analysis (CA)
- Analyse Factorielle des Correspondances (AFC)

J.-P. Benzécri の提案による（1962 年頃）
パリ第 6 大学，統計ラボラトリ主幹を務めた
仏国のデータ解析（analyse des données）の提唱者

## ２．同等手法，類似手法として

- 数量化法Ⅲ類（Quantification methods）
  （パターン分類；林知己夫，1952 年頃）
- 双対尺度法（dual scaling；西里静彦）
- 逆反復平均法，集群分析法
  （reciprocal averaging method；M. O. Hill 他）
- 等質性分析
  （homogenuity analysis；Gifi, J. Meulman 他）

## ３．関連手法

欧米，国内の研究ともに，多くの関連手法が登場した．とくに，フランスを中心とする欧州圏では，様々なデータ表形式に対応する対応分析が考案されてきた.

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

その他，無数にある.

## ４．対応分析法の要約　–仕組み–

（１）数量化法Ⅲ類との考え方の相違は何か

- 尺度化の発想から出発した（林知己夫）
- 質的データの主成分分析（Benzécri）
- ピアソンのカイ二乗統計量と密接に関連
  （クロス表の独立性の検定）

（２）対象とするデータ表

- 原則として二元のデータ表（行列型データ）
- 各要素が非負の数値
- 行または列のプロフィルが意味のあるデータ
- つまり，比率パターンが意味のあるデータ表ならよい

（３）これに含まれるデータ表として，例えば以下がある．

・通常の二元クロス表
・(0，1）型データ行列（二元クロス表の特別な場合）
・多重クロス表（バート表）(「多元」となってはいない）
・多くの統計表（数値が非負の集約データ）

［考え方］

通常の主成分分析（PCA）との併用に意味があることが多い

・ PCA との意味の違いを理解することが重要
・変量（特性）の測定単位に注意

## ５．対応分析法とは？

「二元のデータ表」，たとえば簡単に“ クロス表 ”を考える．

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

ここで，$I$ と $J$ は，それぞれ行と列のカテゴリーの集合

$$\boldsymbol{I} = \{1, 2, \cdots, i, \cdots, m\}$$

$$J = \{1, 2, \cdots, j, \cdots, n\}$$

（１）同時確率分布

$$\underset{m\times n}{\boldsymbol{P}_{IJ}} = \left(p_{ij}\right)\ \left(i \in I, j \in J\right)$$

（２）行の周辺確率分布

$$\underset{m\times m}{\boldsymbol{P}_{\boldsymbol{I}}} = diag\left(p_{i+}\right)\ \left(i \in I\right)$$

（３）列の周辺確率分布

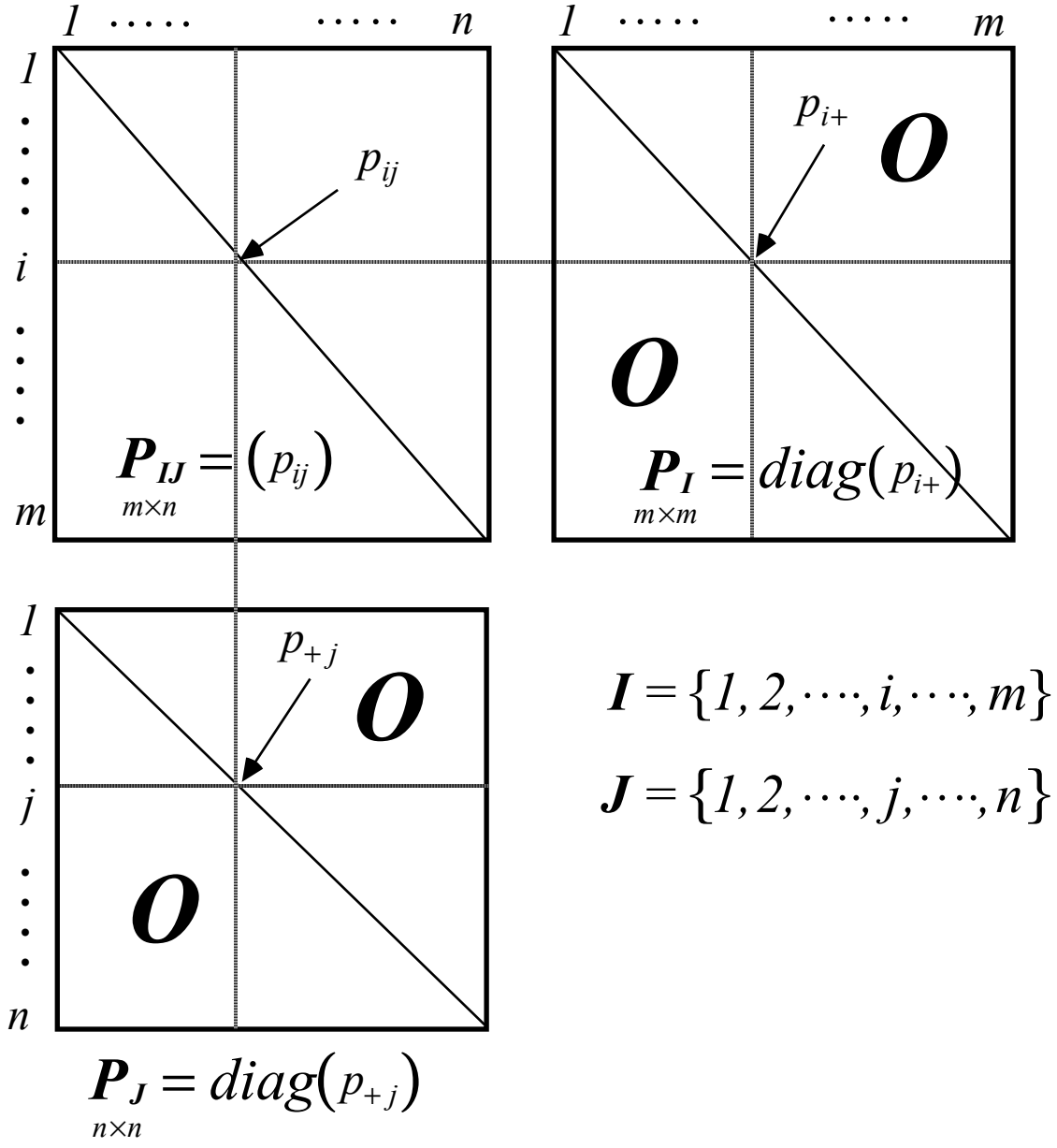
$$\underset{n\times n}{\boldsymbol{P}_{\boldsymbol{J}}} = diag\left(p_{+j}\right)\ \left(j \in J\right)$$

ここで，

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

$$N = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}$$

である．以上を模式的に表すと次の図のようになる．




（注）*diag(･)*は対角行列を意味する．

## ６．プロフィルについて

対応分析法あるいは数量化Ⅲ類では“プロフィル”の概念が重要である.

（１）行のプロフィル（行の比率パターン）

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

（２）列のプロフィル（列の比率パターン）

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

[考え方]

(1) プロフィルとは比率のパターンを考えることである.

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

(3) たとえば，試験成績データを考えたとき,

- 実得点の特徴，科目間の関連性や得点序列,（成績点の）高低を見るなら，主成分分析を使う.
- 科目のパターンや均衡,（相対的に）どの科目で浮き沈みがあるのか，成績得点の傾向（パターン）を見るなら，対応分析を使う.

(4) 従って，データ表のセル内の数値・頻度の大きさとプロフィルのバランスに敏感である（はずれ値などの影響）

## ７．データ行列

$$x_{ij} = \frac{p_{ij}}{p_{i+}\sqrt{p_{+j}}} - \sqrt{p_{+j}} = \frac{q_{ij}}{\sqrt{p_{+j}}} - \sqrt{p_{+j}}$$

または

$$x^{*}_{ij} = \frac{p_{ij}}{p_{j+}\sqrt{p_{i+}}} - \sqrt{p_{i+}} = \frac{q^{*}_{ij}}{\sqrt{p_{i+}}} - \sqrt{p_{i+}}$$

を要素とする行列

$$\boldsymbol{X} = (x_{ij})$$

を考える（なぜ，こうであるかが実は重要である）．あるいは，下記のように

$$y_{ij} = \frac{p_{ij}}{\sqrt{p_{i+}p_{+j}}} = \frac{f_{ij}}{\sqrt{f_{i+}f_{+j}}}$$

ここで

$$f_{i+} \neq 0,\ f_{+j} \neq 0$$

を要素とする次の行列を考えても良い（同じである）．
（注）行和，列和がゼロとなったときは該当列あるいは行のスクイズを行う．

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

前に用意した各行列を用いると，

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

$$\underset{n\times n}{\boldsymbol{V}} = \boldsymbol{Q}^t\,\boldsymbol{Q} = \boldsymbol{P}_J^{-1/2}\,\boldsymbol{P}_{JI}\,\boldsymbol{P}_I^{-1}\,\boldsymbol{P}_{IJ}\,\boldsymbol{P}_J^{-1/2}$$

ここで，$\boldsymbol{P}_{JI}$（$\boldsymbol{P}_{IJ}$の転置行列）である．この行列 $\boldsymbol{V}$（分散共分散行列）の固有値問題（あるいはスペクトル分解）となる．

$$\underset{n\times m}{\boldsymbol{P}_{JI}} = \underset{n\times m}{\boldsymbol{P}_{IJ}^{t}}$$

◆カイ二乗距離を用いること（加重付きの距離とすること）

⑴カテゴリー $i$ と $l$ との間の距離

$$d^2_{il} = \sum_{j=1}^{n} \frac{1}{p_{+j}} \left( \frac{p_{ij}}{p_{i+}} - \frac{p_{lj}}{p_{l+}} \right)^2$$

⑵カテゴリー $j$ と $t$ との間の距離

$$d^2_{jt} = \sum_{i=1}^{m} \frac{1}{p_{i+}} \left( \frac{p_{ij}}{p_{+j}} - \frac{p_{it}}{p_{+t}} \right)^2$$

ここで加重付きとなっていることで，ピアソンのカイ二乗統計量と関係することになる．

◆分布の同等性
⑴等値プロフィルの行（あるいは列）の併合は，列（あるいは行）の距離に影響を与えない．
⑵同じプロフィル（パターン）の併合は結果に影響を与えない．
⑶換言すると，対応分析はこうした考え方が当てはまる（そう考えたい）データに対して有効な方法である．

## ８．スコアの双対性（duality，transition equations）

（１）行のカテゴリー $i$ のスコア

$$z_{ik} = \frac{1}{\sqrt{\lambda_k}} \sum_{j=1}^{n} \left( \frac{p_{ij}}{p_{i+}} \right) z^*_{jk}$$

$$\left( \begin{array}{l} j \in J, k = 1, \cdots, K \\ \quad K = min\{m-1, n-1\} \end{array} \right)$$

（２）列のカテゴリー $j$ のスコア

$$z^*_{jk} = \frac{1}{\sqrt{\lambda_k}} \sum_{i=1}^{m} \left( \frac{p_{ij}}{p_{+j}} \right) z_{ik}$$

$$\left( \begin{array}{l} i \in I, k = 1, \cdots, K \\ \quad K = min\{m-1, n-1\} \end{array} \right)$$

[**考え方**]

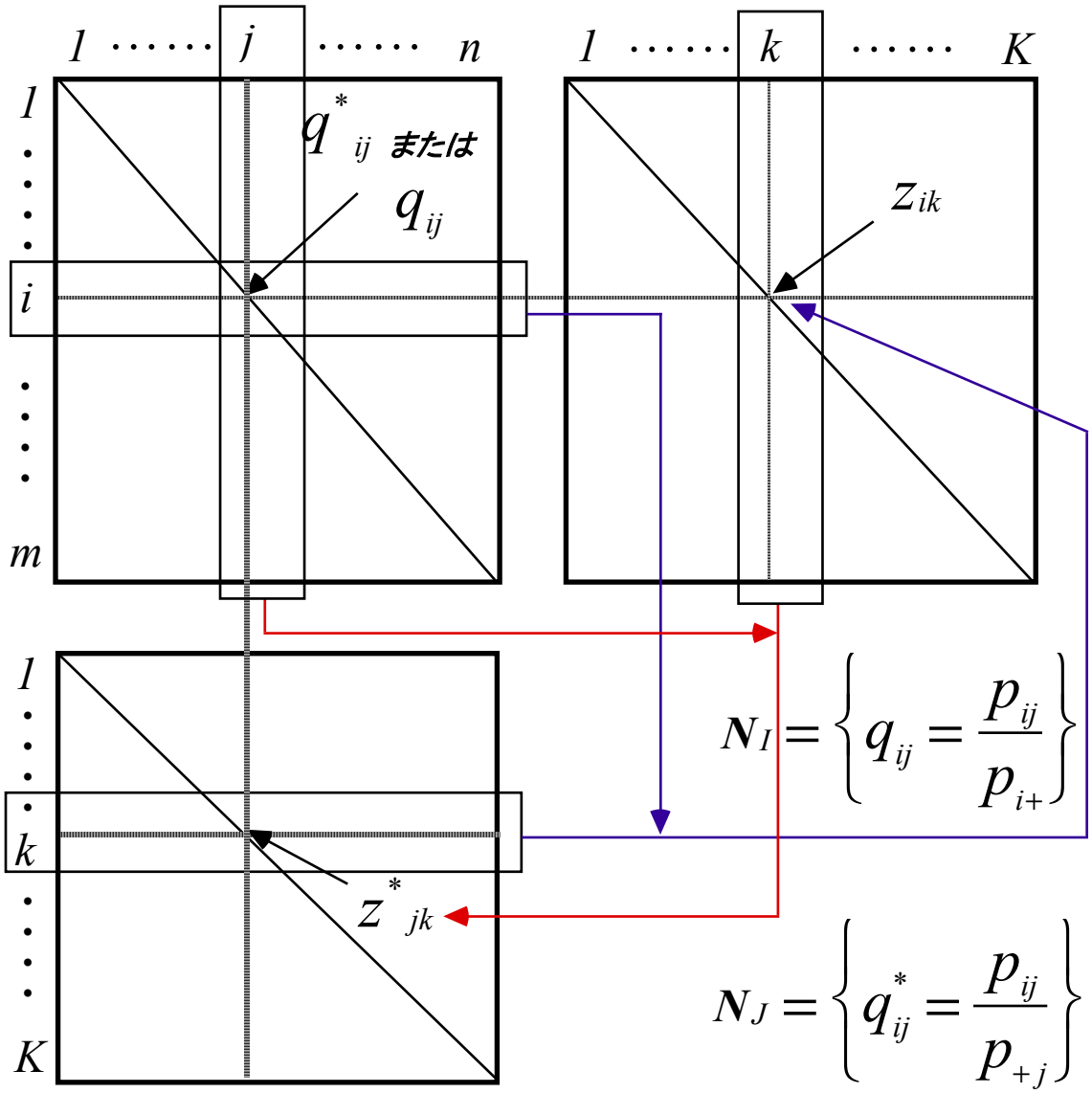
(1)　ここで，$\lambda_k$ は第 $k$ 成分の固有値

(2)　スコアは相互のプロフィルの加重平均となっていることに注意する（次の図を参照）．これが，数量化Ⅲ類・対応分析の本質的な意味である．

(3)　固有の数はデータ表の次元数の小さい方から１を引いた数となる（換言すると，その個数の固有値しかでない）．

(4)　この双対性があることが対応分析の重要な性質となっている．

★ 双対性の関係を図で表すと次のようになる．




（注）行（列）のあるカテゴリーのスコアは，列（行）のスコアの加重平均となっていることが重要である．このことを，同時布置図を考えるときに思い出す必要がある．

## ９．スコアの解釈

得られたスコア（数量化得点）について「布置図」や「同時布置図」を描いて観察する．

◆スコアの散布図（布置図）

行あるいは列のカテゴリーに対するスコアを散布図として観察する．

行のスコア：を図式化

$$(z_{ik},\ z_{ik'})\ (k,k' = 1,2,\cdots,K)$$

列のスコア：を図式化

$$(z^{*}_{jk},\ z^{*}_{jk'})\ (k,k' = 1,2,\cdots,K)$$

**［スコアを観察する際の注意事項］**

（1） スコアの布置の相対的な位置関係に注目する．
（2） 軸の解釈は，場合に応じて考慮する（通常はあまり重要でない）．
（3） 「多重クロス表」から求めたサンプル・スコアの解釈は「元の変量・項目のカテゴリーのスコア」であるから意味理解に注意する．
（4） 固有値，寄与率の解釈は，多重クロス表から出発の場合は注意する．
（5） カテゴリーが順序尺度の場合には図中の序列・並びに注意する．
（6） この意味でスコアを用いたクラスター化操作には十分な注意が必要である（単純な k-means 法や階層的分類ではうまく対応できないことが多い）．
（7） また「はずれ値」の存在に注意する．はずれ値は元のデータ表の中の頻度分布の不均衡から生じる．対応分析の特徴でもある．
（8） 行あるいは列の各カテゴリーに付与されたスコアについては，同時布置を考えたとき，それらの標準化に際しては，それぞれを標準化する場合（平均をゼロ，分散を１とする），しない場合があるので，４通りの可能性がある．なお，WordMiner では，いずれも標準化しない（分散は固有値のまま）を用いている．その理由については大隅他（1994）を参照のこと．

◆ スコアの同時 布置図

行，列それぞれのカテゴリーのスコア

$$(z_{ik},\ z_{ik'}),(z^{*}_{jk},z^{*}_{jk'})\ (k,k'=1,2,\cdots,K)$$

の散布図の同時布置を作ることをいう.

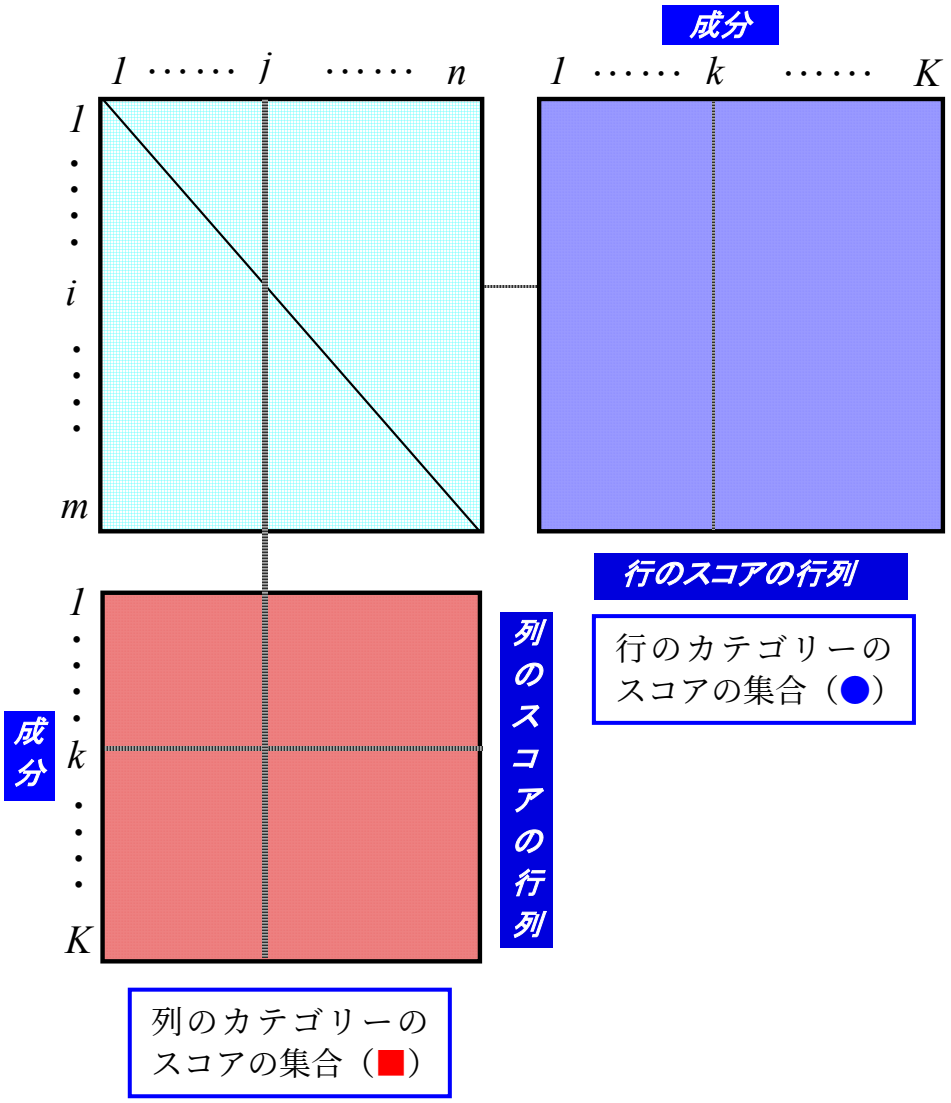
ここで，成分 ***k***，***k'*** について，

$$(z_{ik},\ z_{ik'}),(z^{*}_{jk},z^{*}_{jk'})\ (k,k'=1,2,\cdots,K)$$

を同じ散布図内の打点として図式化する.

元のデータ表と行のカテゴリーのスコア，列のカテゴリーのスコアの関係は次の図のようになる.

## 対応分析における同時布置の考え方

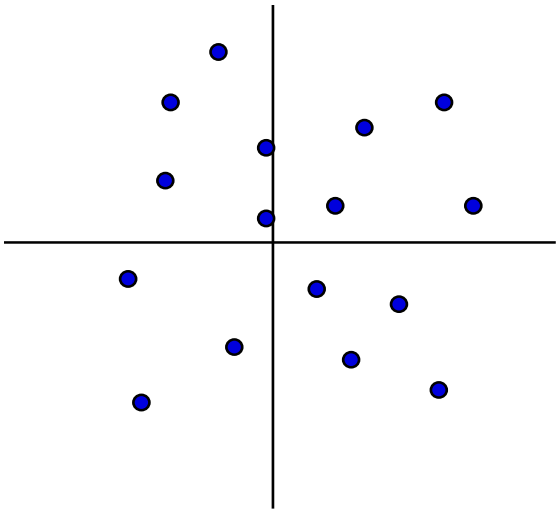


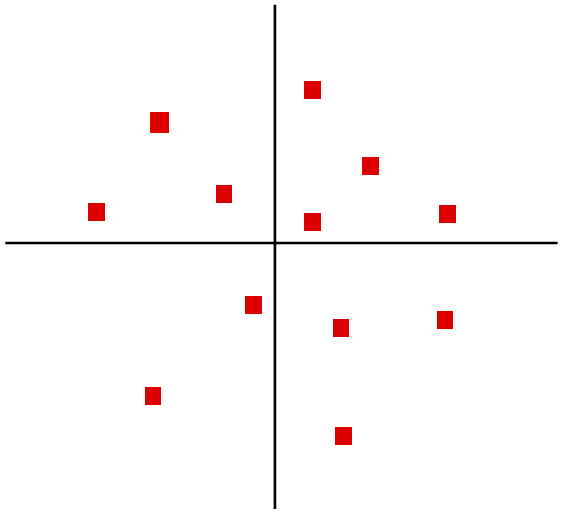

（注）行と列との，同じ成分に対応するスコアを取り出して図式化する．

[考え方]

(1) ここで，行のスコアと列のスコアについて，布置の位置が近いからといって，そのまま「類似している，近い」と判断してはいけない．これは双対性の原理から明らかである．

(2) このことから，両者のスコアを同時的には括れない．つまり，クラスター化を両者のスコアについて（同時布置内で）同時的には行えない．



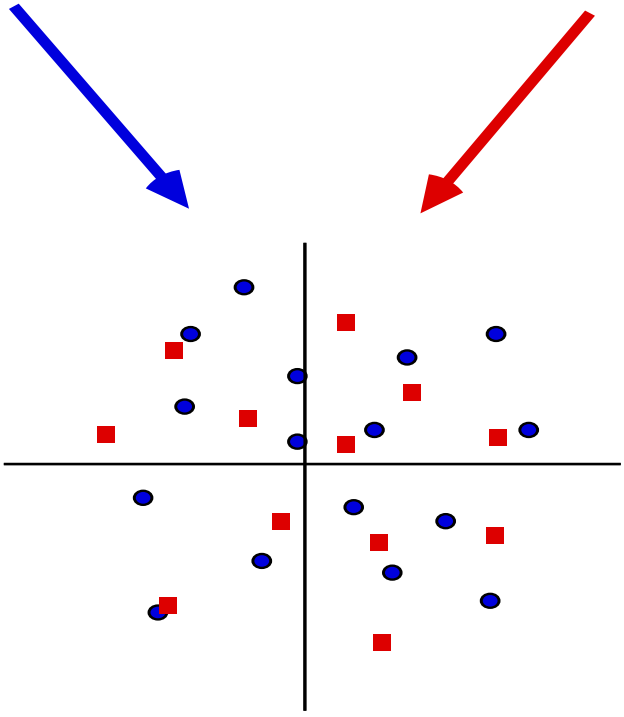


$(z_{ik}, z_{ik'})$

<行のスコアの関係>

$(z^*_{jk}, z^*_{jk'})$

<列のスコアの関係>



$(z_{ik},\ z_{ik'}), (z^*_{jk}, z^*_{jk'})\ (k, k' = 1, 2, \cdots, K)$

<スコアの同時布置図>

**◆データ表の基本的な組み合わせ**

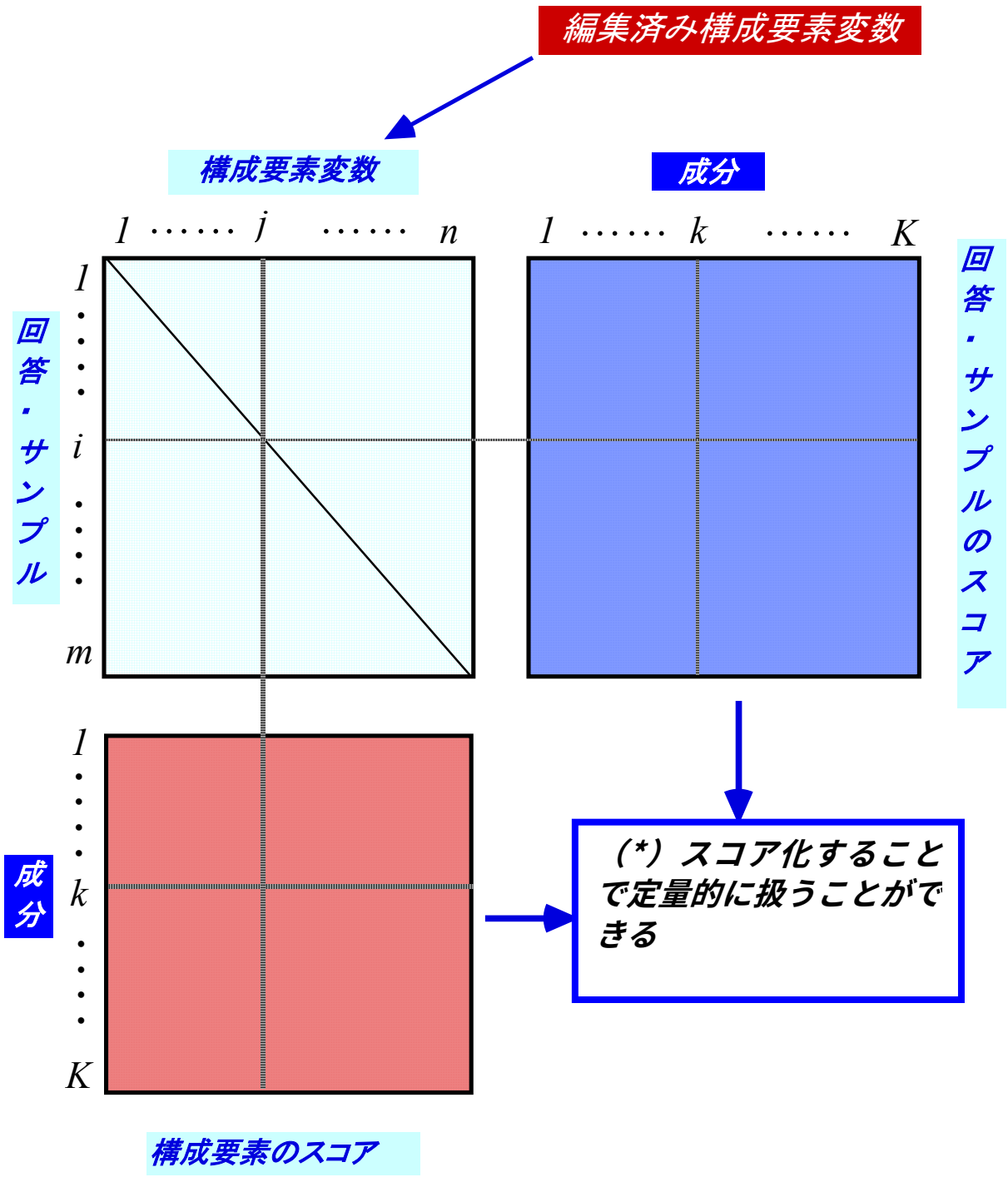
データ表の，表頭（列）と表側（行）に以下のように対応させることで，各種のデータ表の解析を行う．また，双対性から行と列を入れ替えても結果は変わらない．このように組み合わせながら，どれが有意，意味あるかを“探査的”に調べる．

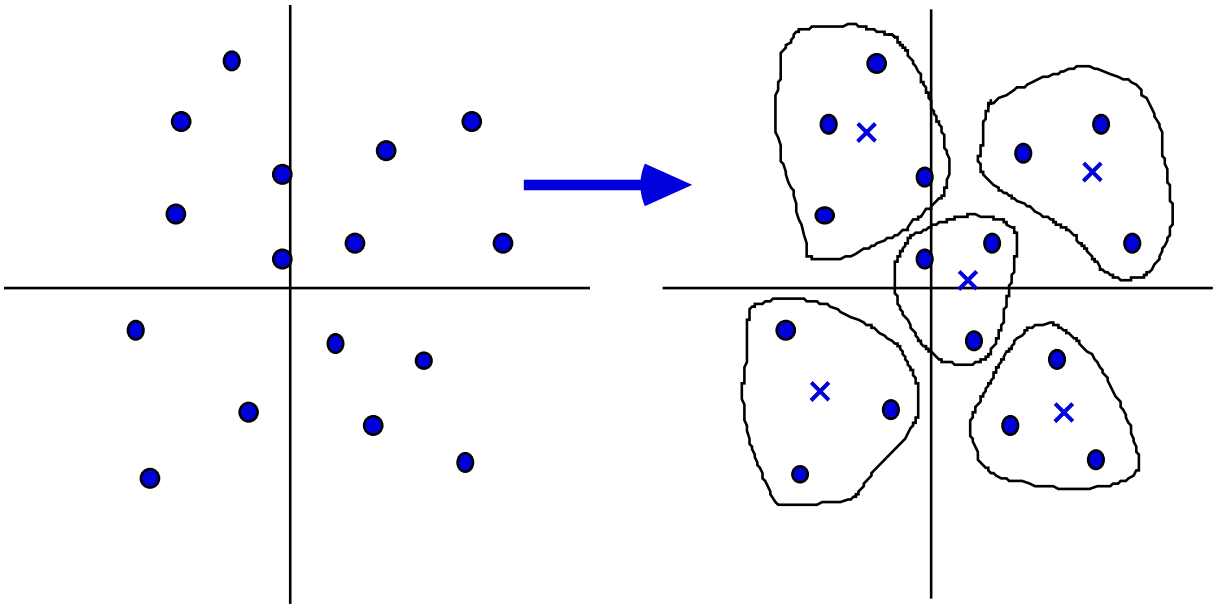
| 表側項目：$I$ | 表頭項目：$J$ |
|---|---|
| ● 構成要素，キーワード | ● 回答，個体 |
| ● 構成要素，キーワード | ● 質的変数<br>● （選択肢型設問・属性等） |
| ● 構成要素，キーワード | ● ｸﾗｽﾀｰ･ﾒﾝﾊﾞｰｼｯﾌﾟ情報<br>（クラスター変数） |

（注）

- 構成要素＝語 句，単語群あるはそれに相当の情報
- クラスター・メンバーシップ＝クラスター生成で得られる質的変数

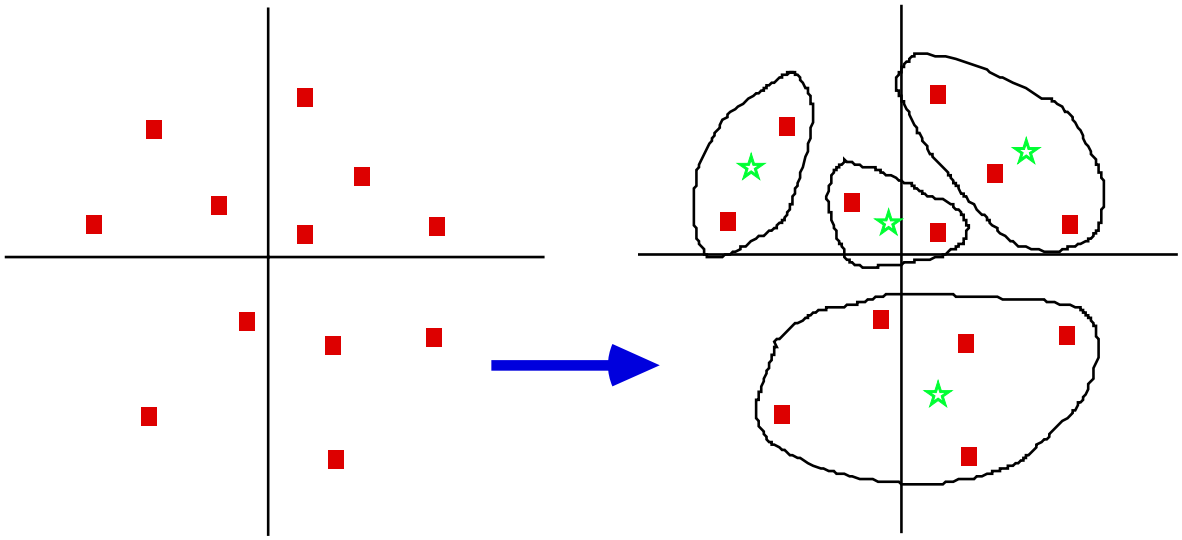
例えば，「（回答・サンプル）×（構成要素）」の場合を考えると，…





回答・サンプルのスコア

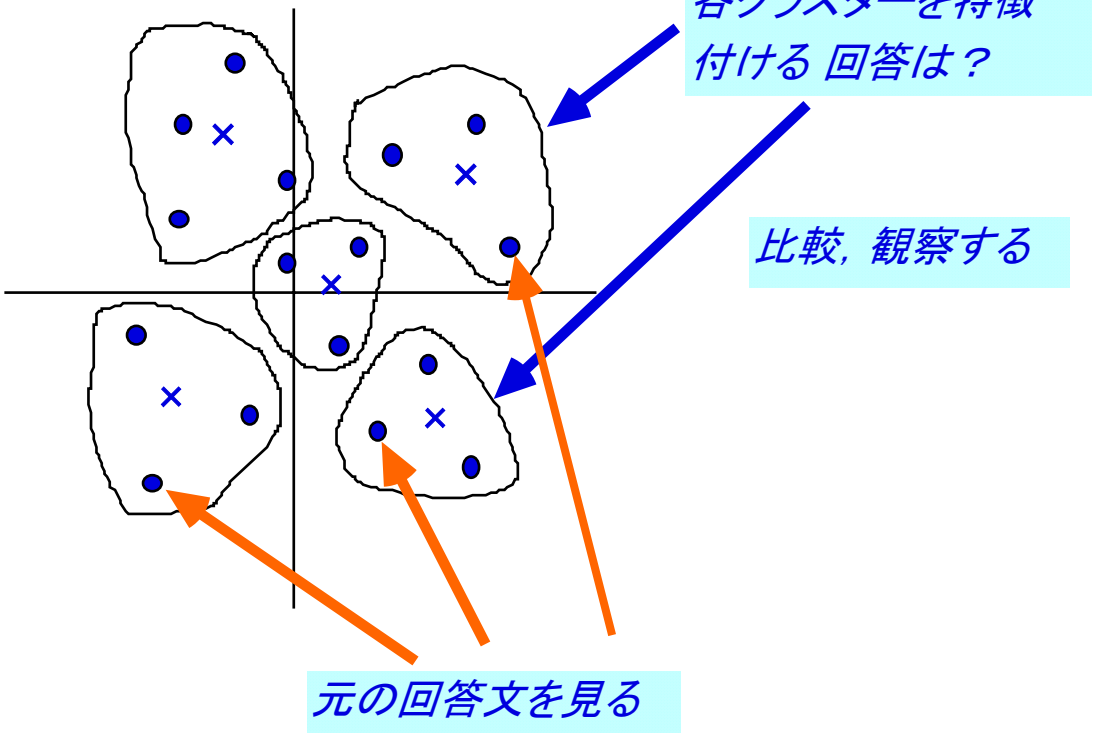
回答・サンプルのスコア
のクラスタリング



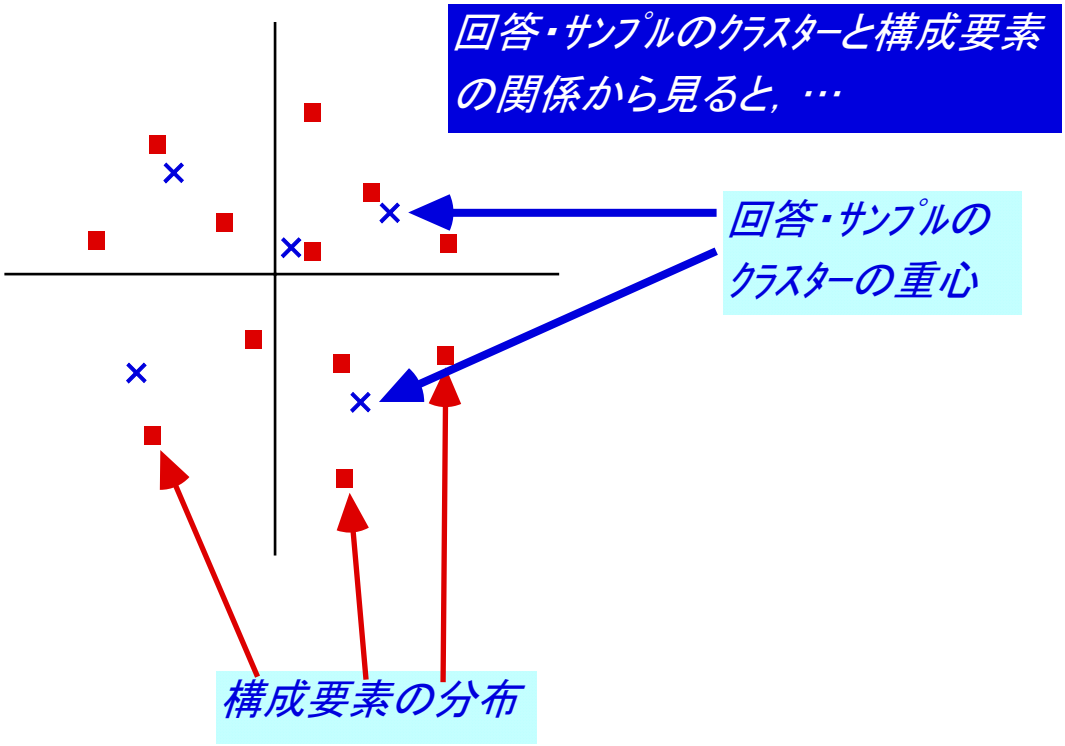
構成要素のスコア

構成要素のスコア
のクラスタリング

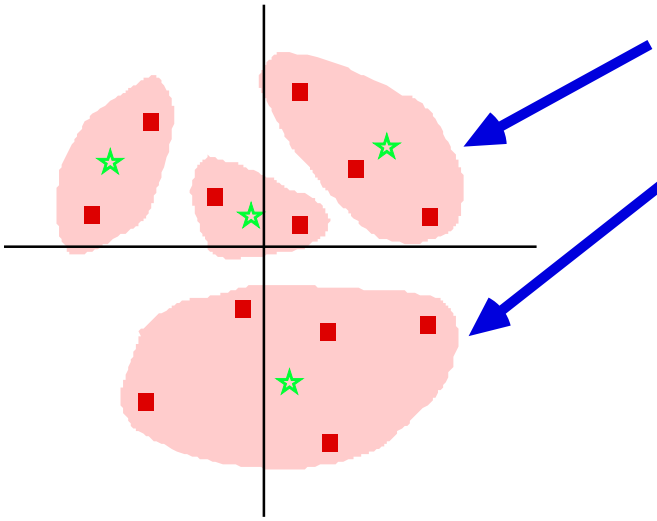
回答･ｻﾝﾌﾟﾙの側から探査すると，…




回答･ｻﾝﾌﾟﾙのｸﾗｽﾀｰと構成要素の関係から見ると，…



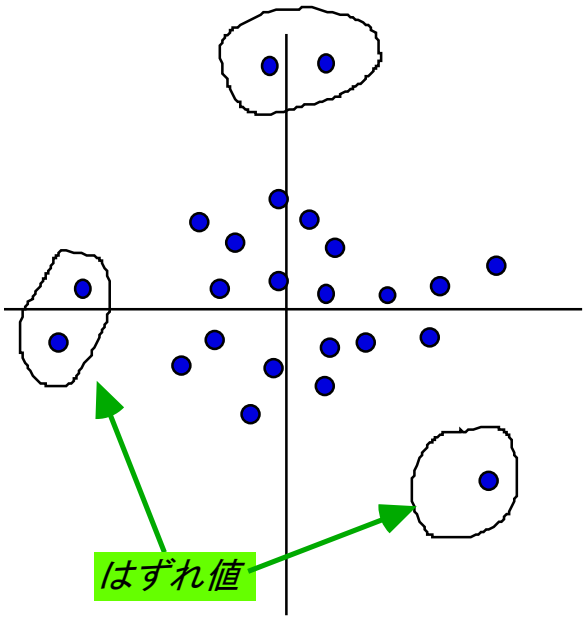
構成要素の側から探査すると，…



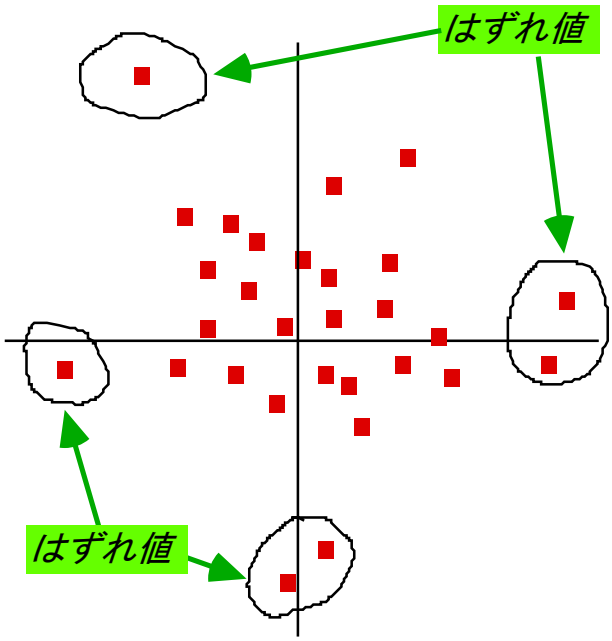
構成要素の各クラスターを特徴は？

比較，観察する

はずれ値を探査すると，…

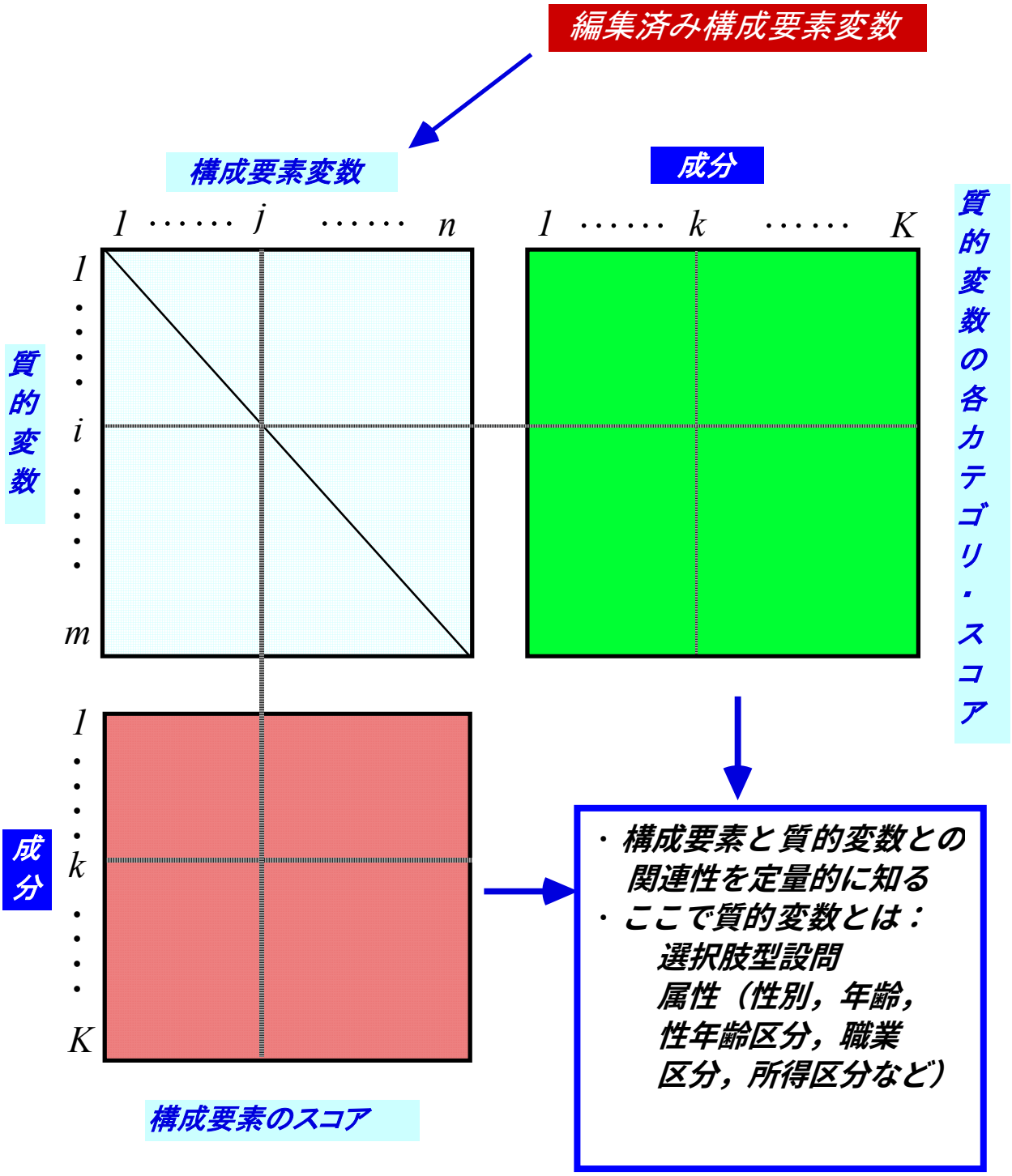


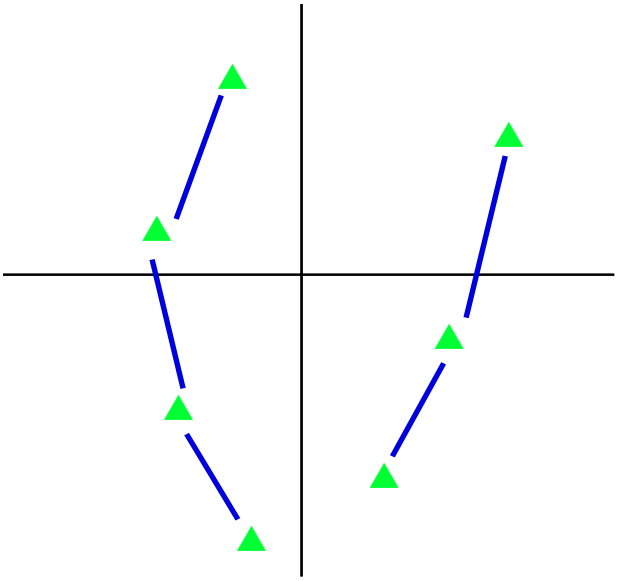

どの回答･ｻﾝﾌﾟﾙかを探査する

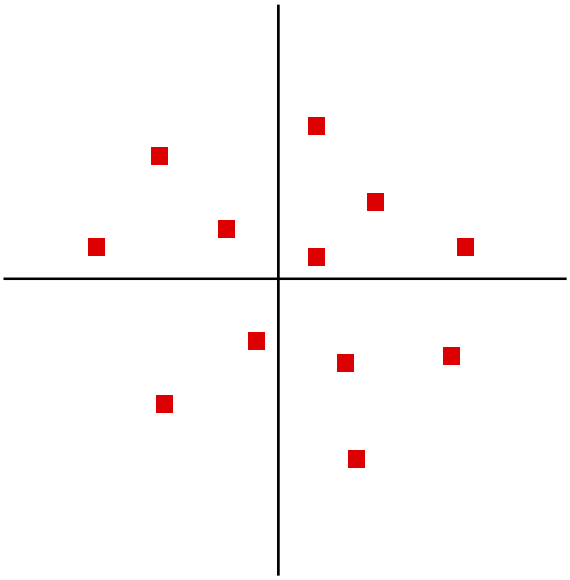



どの構成要素かを探査する

「(構成要素) × (質的変数)」の場合には，…







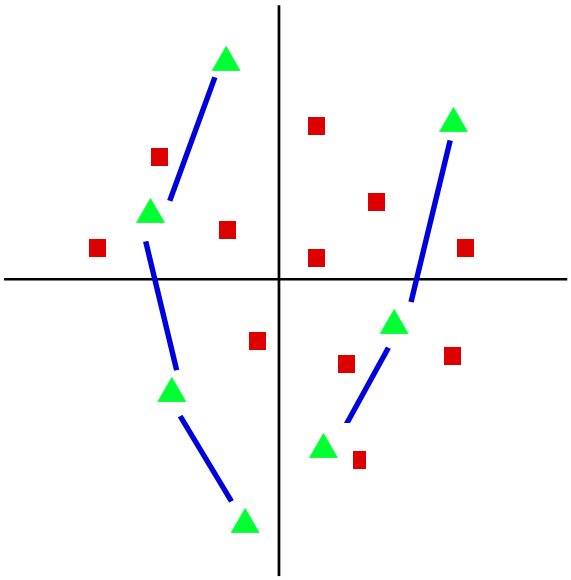
**質的変数のカテゴリ・スコア**

●例えば，性年齢区分であれば，「20代女性，30代女性，…，60代男性」など
●段階尺度の例であると，「満足」「まあ満足」「あまり満足でない」…など
●ｸﾗｽﾀｰ化で得たｸﾗｽﾀｰ変数であれば，そのｸﾞﾙｰﾌﾟ変数情報（名義尺度）

**構成要素のスコア**

**同時布置すると**

●構成要素と質的変数の関係が分かる
●有意性ﾃｽﾄで，客観的に各ｶﾃｺﾞﾘに有意な構成要素の関係が分かる
●構成要素の各ｶﾃｺﾞﾘ内での寄与の程度が分かる
●その原文（回答）を観察する



［考え方］
　こうした情報を，インタラクティブな操作で探索的に分析を進めることが肝要である．

⑴布置図で視覚的に観察する

⑵高次元空間内のおよその情報を知る
（*）非常に疎なデータ表を扱うので，少数次元内に布置が難しいのであくまでも目安とする
（*）同時に無数の構成要素（単語，語句など）を布置した図の視認には限界がある（煩雑になる）ので別の方法と併用する（有意性テストの結果を検討）
（*）はずれ値の検出に有効である

⑶回答・サンプルと構成要素の関係を知る
（*）どの回答にはどんな構成要素が使われているかなど
（*）クラスター化で得た類型を特徴付ける構成要素を有意性テストで客観的に調べ，要約する

⑷構成要素と質的変数や属性との関係を知る
（*）このときは，比較的少数次元の空間内に布置できるので，布置図をしっかりと見る
（*）どの変数が，構成要素に強く関係するかを有意性テストで客観的に調べ，要約する

⑸はずれ値やまれな回答例，構成要素を探査する
（*）これを知って除去したり，その除去効果を知る

⑹構成要素の再編集を繰り返し，その効果を知る
（*）編集によって，その影響が，どこにどう現れるかを知る

⑺基本的には「視認できる情報の範囲，限界」をよく知ったうえで用いることが必要である

**10. 追加処理と追加要素**

次のような場面で，いわゆる「追加処理」を行う．
（supplementary elements, supplementary treatment）

◆外れ値の一時除去と再配置

（1） 一時的に除去した「外れ値」の再布置

（2） 外れ値の影響の程度を見るとき

◆判別分析的，グループ間類似・差異を見る

（1） 層別変数や属性などで，複数のグループに分けられるデータセットを，層やグループ単位で「追加処理」する．

◆データ表の，行の追加と列の追加，あるいは，一次除去と再配置を行うこと

（1） （回答）×（構成要素）に，構成要素を追加

（2） （構成要素）×（質的変数）に，構成要素群を追加，別の質的変数を追加

## Ⅱ．クラスター化法について

### １．用語・呼称

・クラスター分析（cluster analysis）
・クラスタリング（clustering）
・自動分類法（automatic classification）
・数量分類法，数値分類法（numerical taxonomy）
・教師なし分類（unsupervised classification）

いろいろな呼称があることが特徴の一つである．よって研究分野によって無数の jargon が多いのが特徴でもある（とくに，階層的分類法にはこれが多い）．また，一つの手法を示す用語でもなく，分類手法の総称である．

「クラスター化」という呼称が適切な表現だと思う（クラスターは生成するものと考えること），またはコンピュータ処理の観点からは自動分類法が適当である．クラスターは存在するものとの仮説を立ててみても，その検証はきわめて難しい．

パターン認識の分野では，クラスター化法を教師なし分類，判別分析型の手法を教師あり分類と呼ぶことがある．

### ２．手法・技法の分類

「階層的分類法」と「非階層的分類法」に分けることが多い．

・階層的分類法
–凝集型階層的分類法
（AHC：agglomerative hierarchical clustering）
（組み合わせ的手法，ウォード法など）
–分枝型・分岐型階層的分類法（divisive type）
（AID ・ CAID ・ CHAID ・ THAID, CART など）

・非階層的分類法
–分割最適化型分類法（partitioning type）
（SAS/FASTCLUS，k-means 型手法，
それらのファジイ・ヴァージョンなど）
–分布混合型（mixture distributions など）

## ３．ハイブリッド型分類法の要点

### ３．１階層的分類法の主な特徴

（1） 同じ手法の別名が多いこと
（2） 用いる類似度・非類似度と算法（クラスター化アルゴリズム）との組み合わせが無数となること
（3） よって同一データを用いても結果（解）が異なること
（4） 多様な類似度・非類似度に対応できる
（5） 質的データ等にも対応できる
（6） 大量データの処理に不向き（せいぜい 3,000ー5,000 サンプル程度まで）
（7） 手法と生成クラスターの関連の解釈に面倒がある
（8） 多くは，排反的分割となる（重なりがない分類）
（9） 手法によってははずれ値検出に有効な方法もある（single link，MST）

などの性質がある．要するに，同一データを用いても「適用する類似度・非類似度と適用アルゴリズムの組み合わせ」を考えただけでも無数の“異なる解”があるから，クラスター化評価を慎重に行うことが求められる．つまり，どのような方法，手順で行ったかを明示的に示すべきである．

### ３．２非階層的分類法の主な特徴

（1） 大量データの分類に適する（と言われている）
（2） 殆どが「量的データ」向きの手法
（3） 等質性規準，クラスター化規準に依存する（平方和規準，行列式規準など）
（4） つまり，生成クラスターに「くせ」がある
　　–クラスター・サイズがそろう傾向
　　–はずれ値の検出に弱い，これの影響を受けやすい

非階層的手法についても，生成クラスターの評価規準を設けて客観的に検討する必要がある．

クラスター化の共通する問題に「クラスター数はいくつと考えるか」がある．このことは，クラスター化生成規準やクラスター評価規準と連動させて考えるべき問題である．

### ３．３ハイブリッド法

階層的分類法，非階層的分類法の両者の利点をつなぎ合わせることが考えられる．これはきわめて自然かつ容易な発想である．また，

（1） はずれ値への対応
（2） クラスター間のデータの移動（メンバーシップ調整，更新）による安定化
（3） 計算処理時間の短縮化
（4） クラスター化過程の追跡と評価方法の容易性

などの利点がある．

## ４．WordMiner で用いるクラスター化法

・独自のハイブリッド法を用いている．
・原則として「量的データ」に適用する手法である．
・したがって，対応分析で求めた数量化スコアに対して用いる．
・クラスター数はユーザ指定されるが，判定の目安としての規準を表示する．

なお，「クラスター数」を決める適切かつ決定的な方法や指標・規準は，いまのところないといってよい．しかし，WordMiner ではとりあえずある規準を用意し，およそのクラスター数の目安を与えるようにした．

◆初期分類
（1） 対象データの一部を用いて，階層的分類法を行う．
（2） 階層的分類法は，ウォード法に近い考え方（平方和和規準）を使う．

◆二次分類
（1） すべてのデータを用いて，階層的分類法で作った各クラスターに配置する（クラスター化規準の最適化を行う）．
（2） 同時に，必要に応じて，再配置や刈り込み処理を行う．

◆再配置とクラスター規準の改善（細分類）
（1） 各プロセスで，クラスター規準の改善チェックを行う．
（2） 最終的に収束するまで，改善を図る．

◆細分類，再分類
クラスター化の結果，等質性をさらに確保するための二次的処理を必要に応じて行う．

【参考文献】


[1] .B.D. Ripley (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press.

[2] J. C. Gower and D. J. Hand（1996）：*Biplot*, Monographs on Statistics and Applied Probability 54, Chapman & Hall.

[3] M. J. Greenacre (1993): *Correspondence Analysis in Practice*, Academic Press.

[4] R.O. Duda, P.E. Hart, and D.G. Stork (2001), *Pattern Classification*, second edition, John Wiley

[5] 岩坪秀一（1987）：数量化法の基礎，朝倉書店.

[6] 水野欽司（1996）：多変量データ解析講義，朝倉書店.

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

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

[9] 大隅昇（2000），多次元データ解析における分類手法の役割−分けて知ることの難しさ−，エストレーラ（ESTRELA），2000 年 10 月号〔通巻 79 号〕，10-20.

[10] 林知己夫（2000），多変量解析と多次元データ解析−データの科学の中で見る−，エストレーラ（ESTRELA），2000 年 10 月号〔通巻 79 号〕，2-9.