医学統計セミナー アドバンスコース

アンケート調査データの解析

下川 敏雄

和歌山県立医科大学 臨床研究センター

2016年度 医学統計セミナー

■ **ベーシック・コース**
- **基礎統計学** (6月15日・住金棟5F 大研修室)
- **量的データの解析** (7月27日・住金棟5F 大研修室)
- **質的データの解析** (8月24日・住金棟5F 大研修室)
- **共変量調整を伴う解析** (11月2日・病院棟4F 臨床講堂1)
- **生存時間・臨床検査データの解析**(11月16日・住金棟5F 大研修室)

■ **アドバンス・コース**
- **多群・経時データの解析と多重比較** (11月30日・病院棟4F 臨床講堂1)
- **臨床試験における症例数設定とガイドライン** (12月28日・住金棟5F 大研修室)
- **アンケート調査データの解析** (2月1日・病院棟4F 臨床講堂1)
- **統計的因果推論と傾向スコア** (2月22日・住金棟5F 大研修室)
- **メタアナリシス** (3月22日・病院棟4F 臨床講堂1)

アンケート調査の前に考えること：調査票に注意

アンケート調査用紙の作る前にしておくこと

1. アンケートを通して何をしたいかを予め考える.
   → 集計したいだけなのか(記述統計，次元縮約)？
   → 目的があって，その影響要因を探りたいのか(回帰分析)？
   → 潜在構造を知りたいのか(因子分析→潜在構造分析)？
   → グループ分けしたいのか(クラスター分析)？

2. 結果(応答)と原因(説明変数)がある場合には，原因を更に精査する.

   原因には2種類ある.
   - 制御因子(被験者が選択できる要因)
     → 喫煙・飲酒量，薬剤の投与量，睡眠の時間など
   - 非制御因子(被験者が選択できない要因)
     → 性別・年齢・疾患の進行度

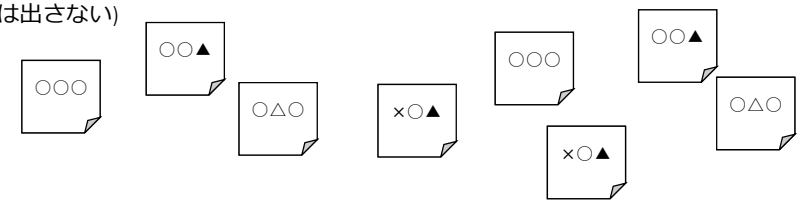
・本アンケートをもとに政策提言等をする場合には，制御因子がもとになる.
・非制御因子は制御因子の調整などに用いられる.

アンケート調査の前に考えること：調査票に注意

アンケート項目のブレインストーミング

Step.1：テーマに関連するキーワードを紙に記す

例えば，「緩和ケア」に影響する要因を探索．紙には，臨床実績，研修，医師とのコミュニケーション…といった用語を全員で書く(このとき，ネガティブな意見は出さない)

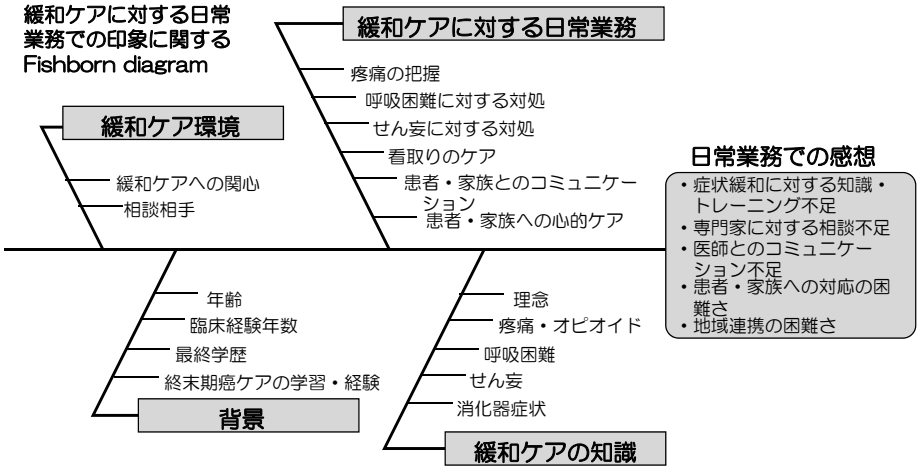


Step.2：キーワードを書いた紙のグループ分け及び重複した内容等を削除

キーワードをグループ分けすることで，「何のために聞くのか」を把握できる．また，問題意識を共有できる．さらに，重複した内容や不必要と考えられる内容を全員で削除することを検討する(場合によっては追加もある).

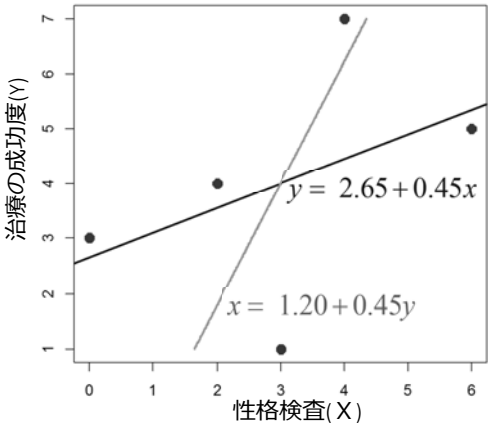
アンケート調査の前に考えること：調査票に注意

外的基準がある場合にはFishborn diagramを作る



相関関係と回帰関係は違う

相関分析：ＸとＹの関係性の強さを表わすもの (ＸとＹを入れ替えても同じ)
回帰分析：ＸからＹを予測するモデルを作るもの(ＸとＹを入れ替えると結果が違う)



回帰直線を引いて，寄与率r^2を計算して「高い相関関係がありました」と記載するのは間違い．

回帰直線は，残差(実測値-予測値)が最小になるように計算される．

相関係数(重相関係数)の2乗が寄与率になるが，目的が異なる点に注意が要る．

相関分析では，散布図だけで十分．回帰直線は誤りを招くもととなる．

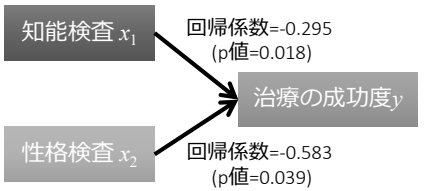
多変量解析概論

何故，多変量にするのか？：重回帰分析を例に

いま，ある精神疾患に対する治療法の成功度を治療前の治療検査と性格検査から予測したい．

単回帰のイメージ：片方の説明変数の応答への影響を考えずに計算

重回帰のイメージ：複数の説明変数の応答への影響を考慮して計算

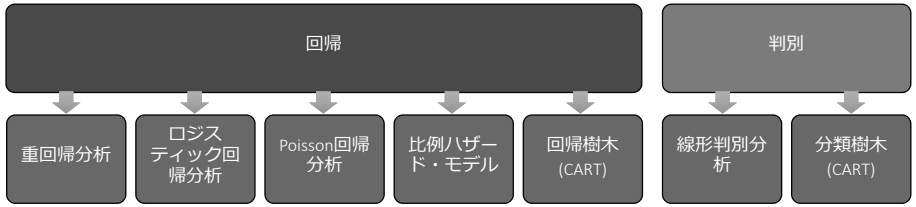


回帰分析とは「原因と結果」があり，説明変数が原因，結果が応答になる．重回帰分析とは，複数の原因から結果を見出そうということが目標になる．

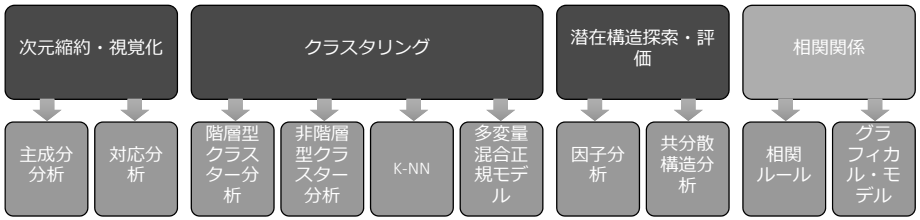
回帰分析で得られた回帰係数のもう一つの見方は，「性格検査の治療の成功度への影響を省いたときの知能検査の影響」を回帰係数で見ているともいえる．

多変量解析の概要

外的基準がある場合

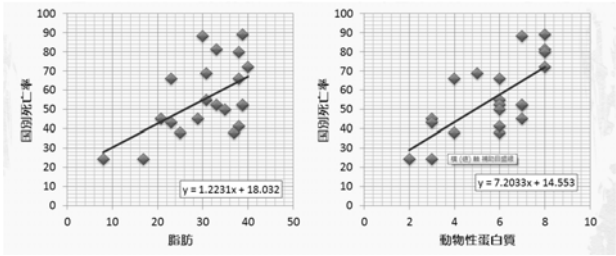


外的基準ない場合



変数の入れすぎに注意：多重共線性

各国の心臓病疾患と摂取エネルギーの関係(脂肪比率，動物性蛋白質摂取量)を調査した．脂肪および動物性蛋白質が心臓病疾患に及ぼす影響を調べなさい(Hilleboe, 1957)



重回帰分析の結果：多重共線性の発生

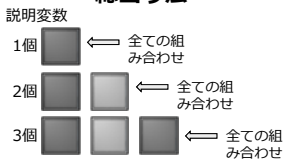
$$\hat{y} = 16.622 - 0.223x_1 + 8.044x_2$$

x_1：脂肪　x_2：動物性蛋白質

単回帰では脂肪が増加するにつれて死亡率が上昇していたが，重回帰分析では，脂肪が増加するにつれて死亡率が減少してしまう．すなわち，解釈が逆になってしまう．説明変数間の相関(関連性)が高い等の理由から悪影響を与え合うことを**多重共線性**という．

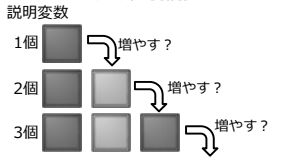
多重共線性への対処：変数選択

総当り法



説明変数が1個,2個,...のそれぞれのパターンを計算しその中から最適なものを選ぶ

変数増加法



説明変数が1個の場合からスタートして，変数を追加したほうが良ければ増やし，そうでなければ変数の追加をしない．

変数減少法

全ての説明変数からスタートして，変数を減らしても影響がなければ減らし，そうでなければ変数の削除をしない．

変数増減法

変数増加法では1度組み込まれた変数を削除できず，変数減少法では1度外された変数は追加されない．二つを組み合わせたのが変数増加法である．

縮小推定法

2000年代になって流行ってきた方法である．有名な方法としては，

・lasso法，　・ElasticNet法

がある．

古典的な4種類の変数選択法では，回帰係数の推定と変数の選択を別々にやっていたが，縮小推定法では，同時に行うことが出来る．

回帰分析の諸型

全ての回帰分析(基本的なものに限る)は一般化線形モデルという枠組みで計算される．

名前	応答の形式	例	係数の解釈
重回帰分析	計量	体脂肪率	回帰係数 (標準回帰係数)
ロジスティック回帰分析	2値	治療の成功/失敗	オッズ比
ー 名義ロジスティック	名義	疾患の種類	オッズ比
ー 比例オッズモデル	順序	疾患の進行程度	オッズ比
Poisson回帰分析	計数	ポリープの検出個数	率比
Cox比例ハザード・モデル	生存時間	がん患者の生存期間	ハザード比

- 回帰係数は説明変数の尺度に依存するため，すべての変数を標準化したもとで計算する回帰モデルの係数は標準回帰係数(標準化係数)と呼ばれる．標準化係数の絶対値の大きさを利用することで，応答に対する各説明変数の影響を評価できる．
- ロジスティック回帰，Poisson回帰，Cox比例ハザード・モデルでは，指数関数A=exp(β)を計算することで，それぞれ，オッズ比，率比，ハザード比を計算できる．それぞれの解釈は下記のとおり：
  - オッズ比：変数Xが1上がるとA倍y=1になる(例：A倍治療が成功する)．
  - 率比：変数Xが1上がるとA倍計数が上がる(例：A倍ポリープが検出される)．
  - ハザード比：変数Xが1上がるとA倍イベントリスクが高まる(例：A倍死亡リスクが挙がる)

外的基準がない方法：次元縮約・視覚化

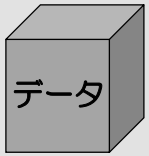
次元縮約・視覚化 | クラスタリング | 潜在構造探索・評価 | 相関関係

そのままの形でみることができる.

ただし

４次元以上のデータ

そのままの形でみることができない.



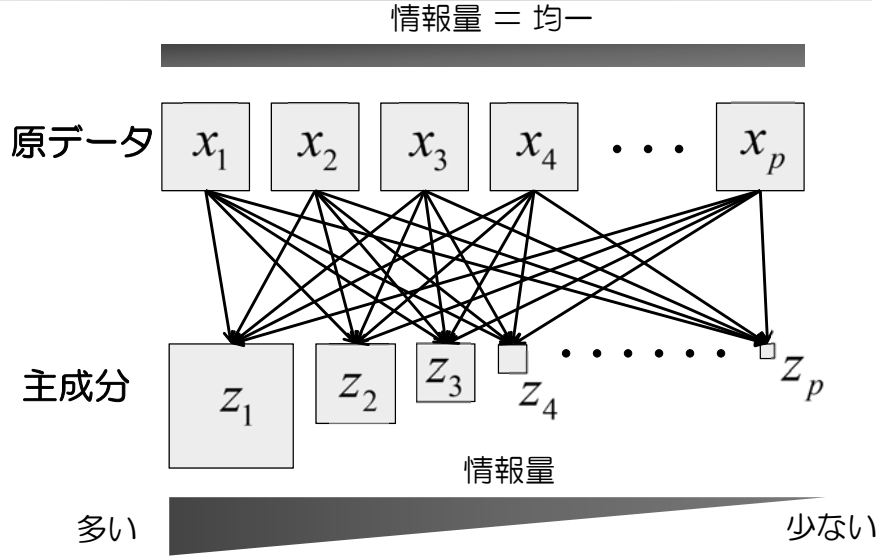
*P*次元空間

見える化

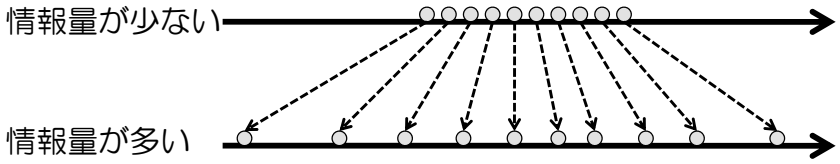
データの構造(情報量)をできる限り崩さずにに表現する

低次元
(できれば２次元)

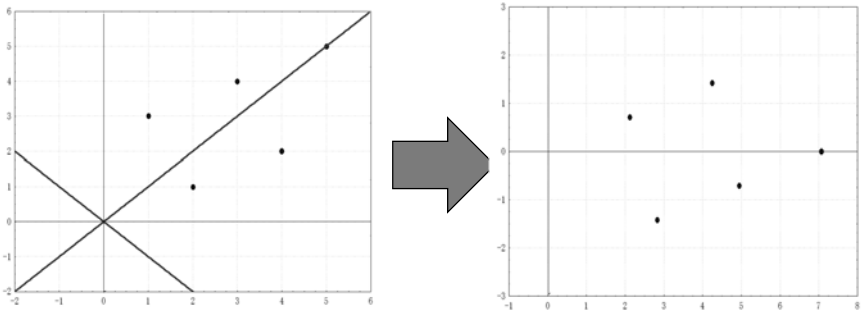
主成分分析：概念図



主成分分析における情報量が多いとは



主成分分析における情報量とは，バラツキ(すなわち分散)が大きいことを意味する.



主成分分析：例示

ある高校の学生５０人の６教科の試験成績がある.

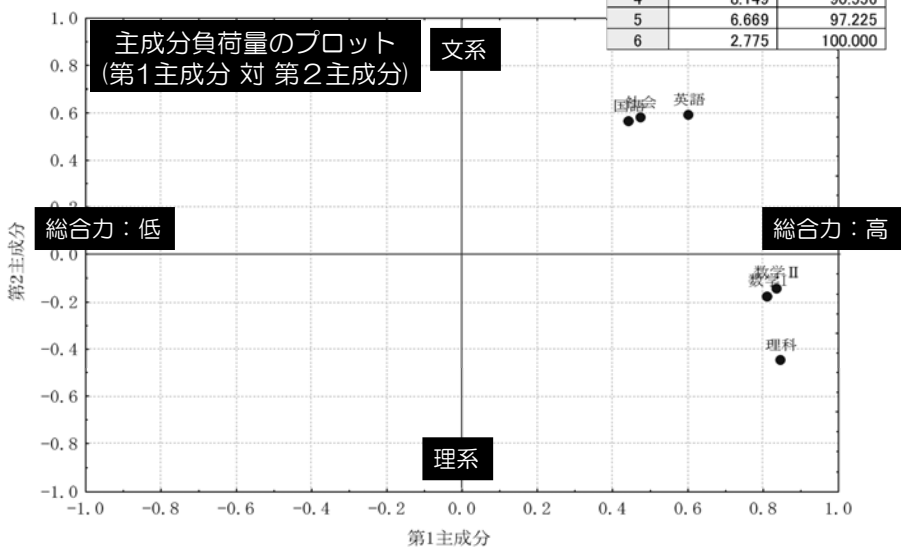
国語，英語，数学Ⅰ，数学Ⅱ，理科

主成分解析を用いて，成績の優劣を2次元平面上で省察したい.

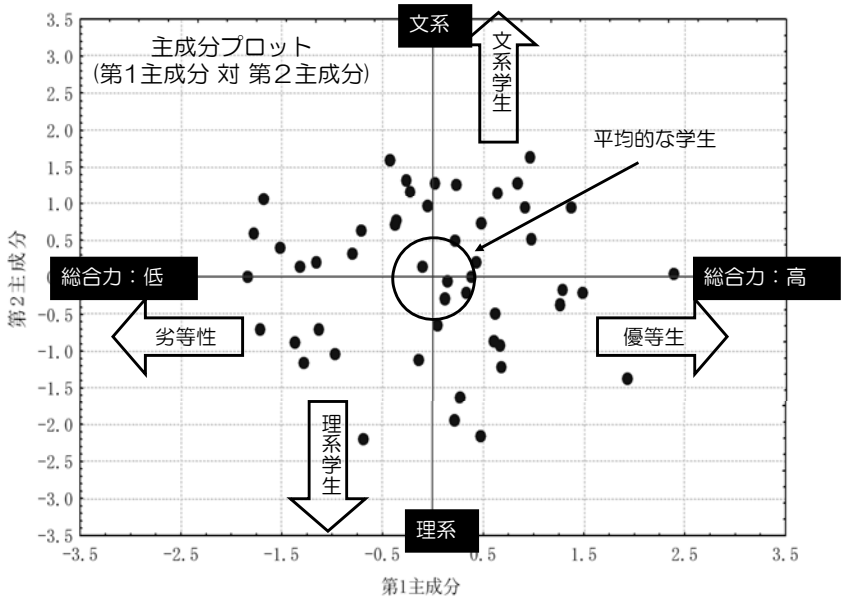
※ このとき，得られた新座標(主成分)は「どのように」解釈されるか.

主成分負荷量のプロット

主成分	寄与率	累積寄与率
1	50.157	50.157
2	21.633	71.790
3	10.617	82.407
4	8.149	90.556
5	6.669	97.225
6	2.775	100.000

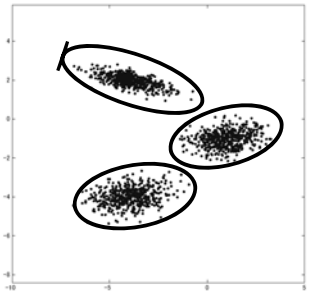


主成分分析：例示



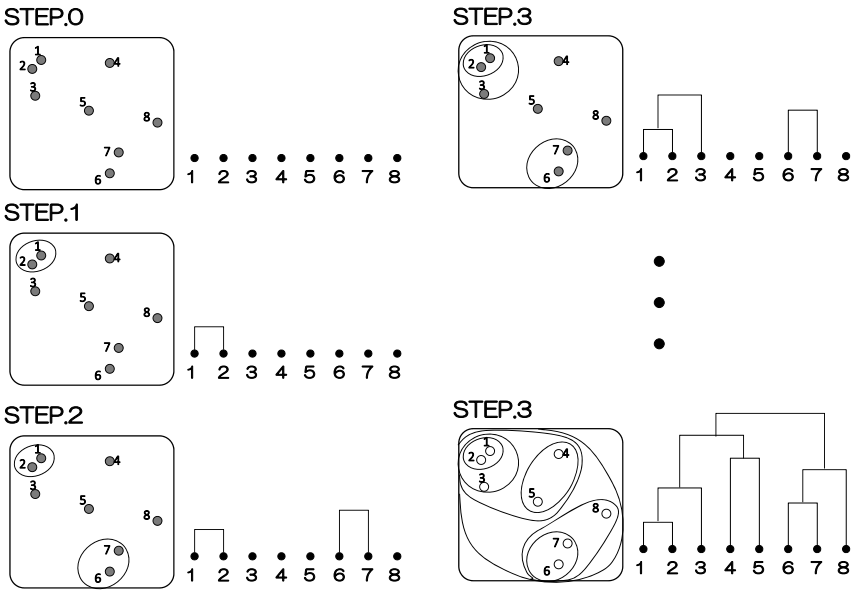
外的基準がない手法：クラスタリングとは？

次元縮約・視覚化 | クラスタリング | 潜在構造探索・評価 | 相関関係



- クラスター分析とは与えられた変数をいくつかのグループ(クラスター)に分類する方法である.
- 主成分分析が変数の圧縮ならば，クラスター分析は個体の圧縮と捉えることもできる.
- Classification(分類・判別分析)とClustering(クラスタリング)の違いは，前者はある基準に基づいて区別(分割)するのに対して，後者はひとまとまりにするという意味がある.

階層型クラスター分析の概念図



クラスター代表値について

個体間(あるいは**クラスター**)との非類似度が小さい

⇩

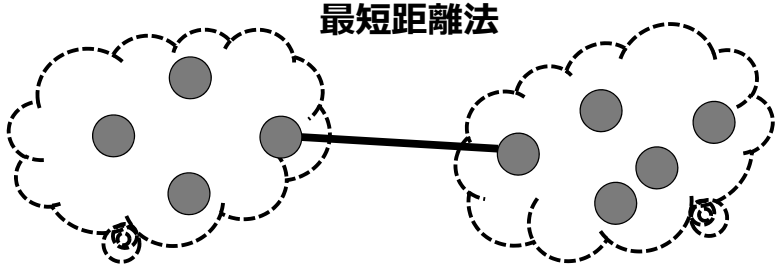
似ている個体(あるいは**クラスター**)
としてクラスタリングする

階層型クラスター分析では，逐次に一番非類似度が小さい個体あるいは**クラスター**を併合していく

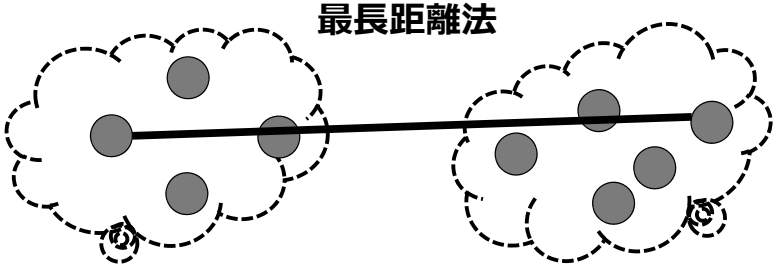
クラスターと個体，あるいはクラスターとクラスターを評価するには，クラスター代表値が必要

階層型クラスター分析の諸型

最短距離法

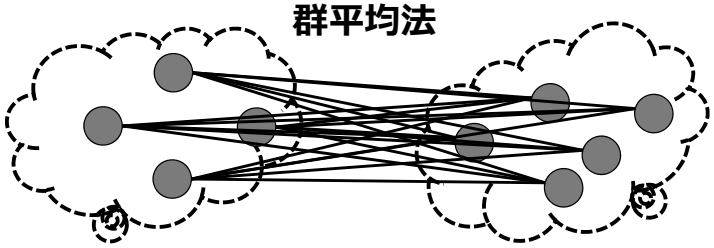


最長距離法



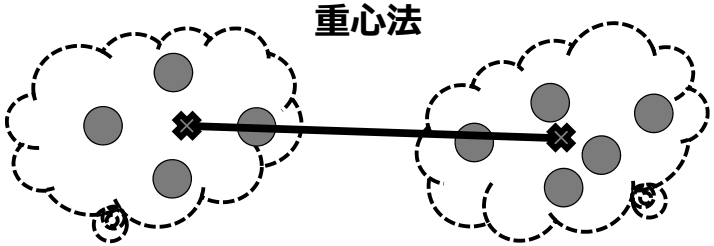
階層型クラスター分析の諸型

群平均法



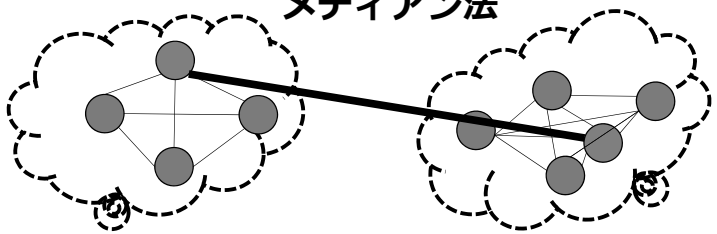
すべての個体間距離を平均したものを用いる

重心法

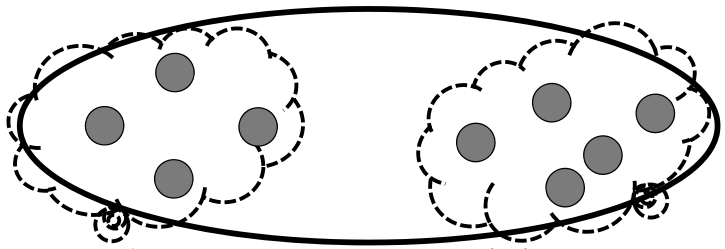


階層型クラスター分析の諸型

メディアン法

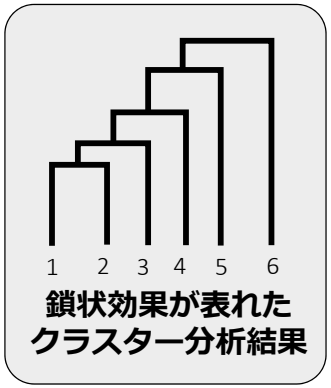


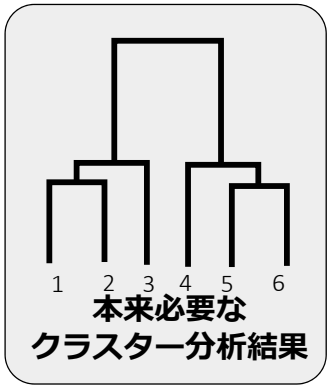
Ward法



１個のクラスターになったときの損失を用いる

最短距離法・メディアン法の問題点：鎖状効果





鎖状効果とは，クラスター内に1個づつ個体が連鎖的に追加される状態をいう．このような場合には，いくつのグループにクラスタリングすればよいかわからない．

非階層型クラスター分析

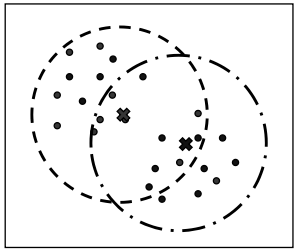
- 非階層型クラスター分析は，データを既知の個数の群にクラスタリングする方法である．
- 非階層型クラスター分析には，K平均法あるいは，Kメディアン法がある．

STEP1：個体を，目標とする群数に適当な方法で分割した結果を初期値とする．

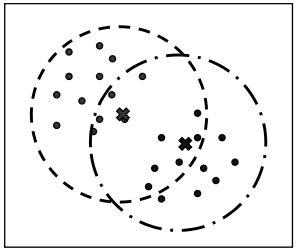
STEP2：データ点を順次に取り出して，k個のクラスターの代表値(K平均法ならば重心，Kメディアンほうならば中央値)との距離を計算する．もしも，最も近い代表値をもつクラスターが元の所属と異なる場合には，そのデータ点を再分類し，構成単位が変化したクラスターに関しては，その重心を再計算する．

STEP3：STEP2が収束するまで続ける．

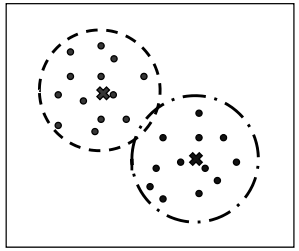
非階層型クラスター分析の例示：K平均法の場合



ステップ1：適当に初期クラスターを作り，重心(平均値)を計算



ステップ2(a)：個体を重心に近いほうに割り当てる



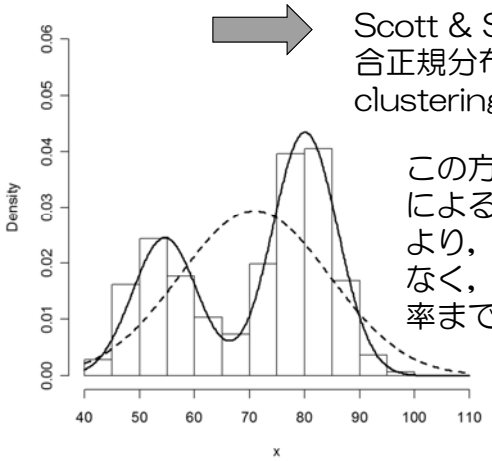
ステップ2(b)：重心を再計算

この作業をすべての個体にクラスター変更がなくなるまで続ける．

最近のクラスター分析手法：Model based clustering

階層型クラスター解析やK平均法では，定式化されたモデルに基づいていないため，最適なクラスター数の選択問題が多分に困難

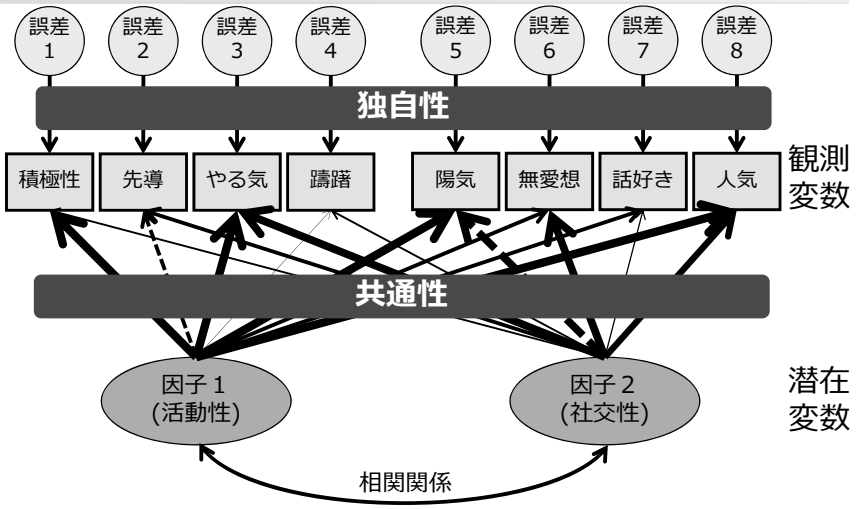
Scott & Symons(1971)は，多変量混合正規分布をあてはめるModel-based clusteringを提案している．



この方法では，多変量混合正規分布によるパラトリック・アプローチにより，クラスター平均，分散だけでなく，各クラスターに対する帰属確率までも計算できる．

外的基準がない手法：因子分析

次元縮約・視覚化　クラスタリング　潜在構造探索・評価　相関関係



個々の因子に対する，すべての観測変数の因子負荷量が計算できる(上図のパスの太さが因子負荷量を表す)．太いパスの先にある変数に共通する因子として影響を与える．

(探索的)因子分析の流れ

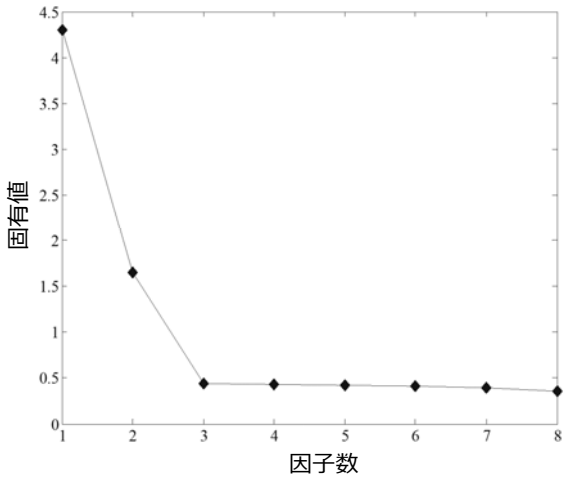
因子数の選定
- 分析前に決定する
- 初期解を先に求めて，因子負荷量を選定

初期解を求める
- 探索的因子分析では，諸種の因子解を求める方法がある．
- 初期解とは解ではないが解に近い結果を表す

軸を回転させる．
- 因子の因子負荷量(共通性)が高くなるように，決定した因子数のもとで軸を回転させる

初期解(いいかえれば因子負荷量)の計算には，主因子法，最尤法，最小2乗法がある．
- 主因子法(古典的．簡便法)
- 最尤法(比較的よく用いられている)
- 最小2乗法(あたりさわりがない)

因子数の決定



探索的因子分析では，観測変数の数だけ因子を構成できる．

このとき，最適な因子数を決めるためにスクリープロットを構成することがある．

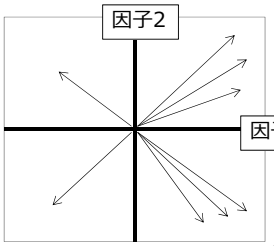
スクリープロットの中で，

- 固有値が1になる．
- 固有値の減少が飽和する直前

などの方法により選択数する．

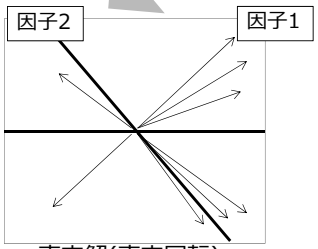
因子軸の回転

観測値の因子負荷量のプロット(緑の矢印は因子負荷量を表している)

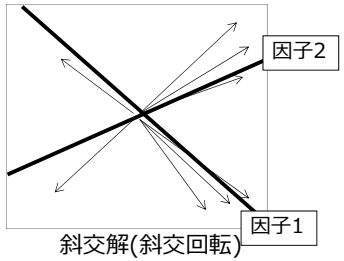


初期解

どの回転方法を用いるかは，得られた因子負荷量の解釈のし易さで選択される傾向にある．



直交解(直交回転)



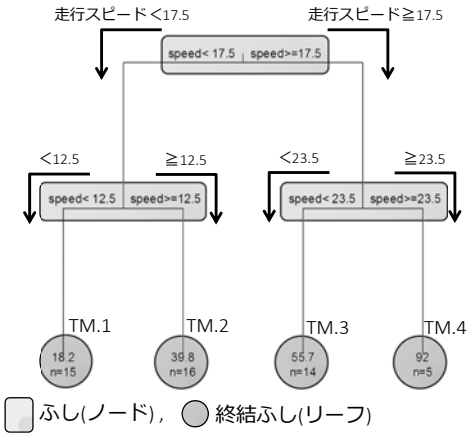
斜交解(斜交回転)

回帰・分類分析における新たな潮流：Regression trees / Classification trees

■自動車の走行スピードと走行距離の関係を調査したデータ(McNeil, 1977)

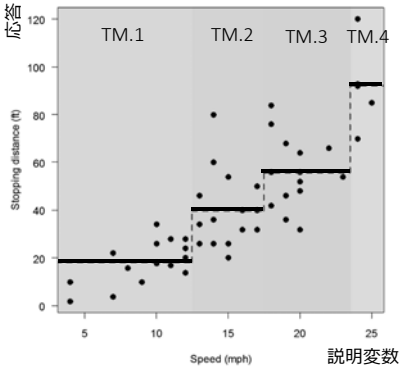
- 回帰を意図した場合には回帰樹木(回帰木)と呼ばれる

樹木図



ふし(ノード)，終結ふし(リーフ)

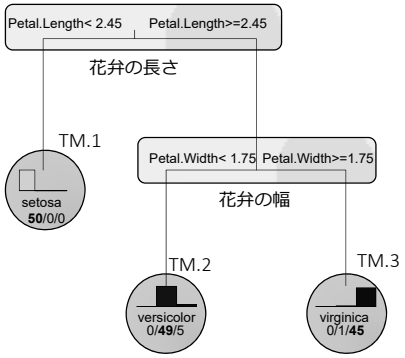
散布図



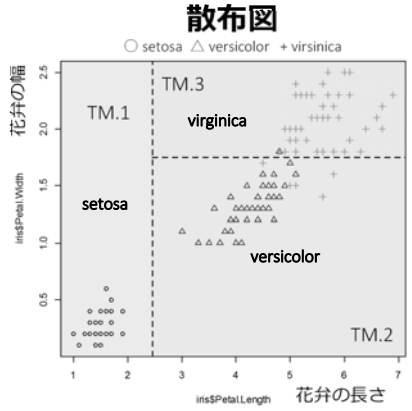
回帰・分類分析における新たな潮流：Regression trees / Classification trees

■あやめのデータ(Fisher, 1936) - 判別・分類を意図した場合には分類樹木(分類木)と呼ばれる

樹木図



散布図



終結ふし内で最も多いカテゴリが予測値になる．下記に**プロダクション・ルール**で表す．

TM.1 $(\texttt{Petal.Length} < 2.45)$, $\hat{y} = \texttt{setosa}$

TM.2 $(\texttt{Petal.Length} \geq 2.45) \cap (\texttt{Petal.Width} < 1.75)$, $\hat{y} = \texttt{versicolor}$

TM.3 $(\texttt{Petal.Length} \geq 2.45) \cap (\texttt{Petal.Width} \geq 1.75)$, $\hat{y} = \texttt{virsinica}$

ご清聴ありがとうございました