統計学第14回
「高度な解析法についての概説」

中澤　港
http://phi.ypu.jp/stat.html
<minato@ypu.jp>

高度な分析法

- 統計モデルが複雑
  - 多変量の整理
    - 次元の縮小
      - 主成分分析
    - 潜在変数の探索
      - 因子分析
      - 共分散構造分析
- データ構造が複雑
  - コンピュータ集約型統計学（シミュレーション）
  - オブザーベーション間の関係の整理
    - クラスター分析

主成分分析

- n 個体のサンプルがあって，それぞれについて，p 個の変数 x1, x2, ..., xp の観測値が得られているとする。一般に，p 個の変数の情報を全部一度に考えて n 個体の情報を把握することは難しい。そこで考えられるのが，p 個の変数を，もっと少ない数の，互いに独立な主成分 (principal component) で表せないかということ
- 主成分 ξ1, ξ2, ..., ξp を考え，これらを x の一次関数で表すことにする。つまり ξi=Σlijxj (j=1..p) とする。p×p 個の適当な係数 lij を求めることを考える。p 個の ξi は互いに無相関とし，この変換が直交変換であるとすると，行列の固有値と固有ベクトルを計算することで lij を求めることができる。詳しくは，M.G. ケンドール著（大橋靖雄・奥野忠一訳）「多変量解析」（培風館）などを参照。
- 通常，主成分を，固有値の大きい順に，第1主成分，第2主成分，... と呼ぶ。第1主成分は，あらゆる一次関数の中で可能な最大の分散をもつ。第2主成分は第1主成分と無相関な一次関数の中で可能な最大の分散をもつ。このようにして主成分を決めると，それぞれの固有値の，固有値の和に対する割合を使って，それぞれの主成分が全変動の何パーセントを説明するかを表すことができる。それを主成分の寄与率と呼ぶ。普通は，たくさんの変数から少数（例えば2つとか3つ）の主成分だけを使って全変動の 80% が説明できる，のように使う。R での関数は princomp() 。
- なお，それぞれの xj を変数ごとの平均からの偏差として lij を計算する方法を P モード，個体ごとの平均からの偏差とするのを Q モードと呼ぶ。Q モードは n 人の被験者がある特定の対象について p 個の測定をしたデータに使うのに適している。R では prcomp() を用いる。

主成分分析の実際

- テキストに掲載している県別データの分析例
  - `library(mva)` によって多変量解析ライブラリを呼び出す
  - ```
    x <- read.delim("113-2003.dat")
    attach(x)
    mat <- cbind(CAR1990,TA1989,DIDP1985)
    res <- princomp(mat)
    summary(res)
    ```
  - 第1主成分と第2主成分のプロットは，biplot(res,xlabs=PREF) とすればよい。
- 組み込みデータの例
  - data(USArrests) によって 1973 年の米国の州別の人口 10 万当たり犯罪関連データを呼び出すことができる。変数は Murder が殺人逮捕者数，Assault が暴行逮捕者数，UrbanPop が都市人口割合，Rape が強姦逮捕者数
  - データフレーム内のデータ全部を使うなら，princomp(USArrests) のように，データフレーム名を与えればよい。
    ```

因子分析

- 因子分析では，主成分分析とは逆に，p 個の観測された変数 x があるときに，個々の x が m 個（m<p）の潜在因子の線型結合と誤差によって表されると考える。
- R では factanal(データ，factors= 因子数) という因子分析を行う関数がある。3変数や4変数では因子数は1しか指定できない。
- 少数の因子による累積寄与率を最大にするために varimax 回転や promax 回転を行うことがある。前者なら rotation=”varimax” という引数を与えればよい。生データからの計算の場合，因子得点の計算は scores=”Bartlett” などとすれば可能。

R

因子分析の実際

- 県別自動車保有データの例では，
  res <- factanal(mat,1) とする。
- 第1因子の因子負荷量は 1.764 であり，寄与率は 0.588 である。このことは，取り上げた3つの変数が，共通の潜在因子によって約 59% 説明されることを意味する。
- 組み込みデータでは USArrests を使う場合は同様。data(ability.cov) として6種類のテスト結果の分散共分散行列データを読めば，
  factanal(factors=3,covmat=ability.cov)
  のように covmat で指定できる。

共分散構造分析 (sem)

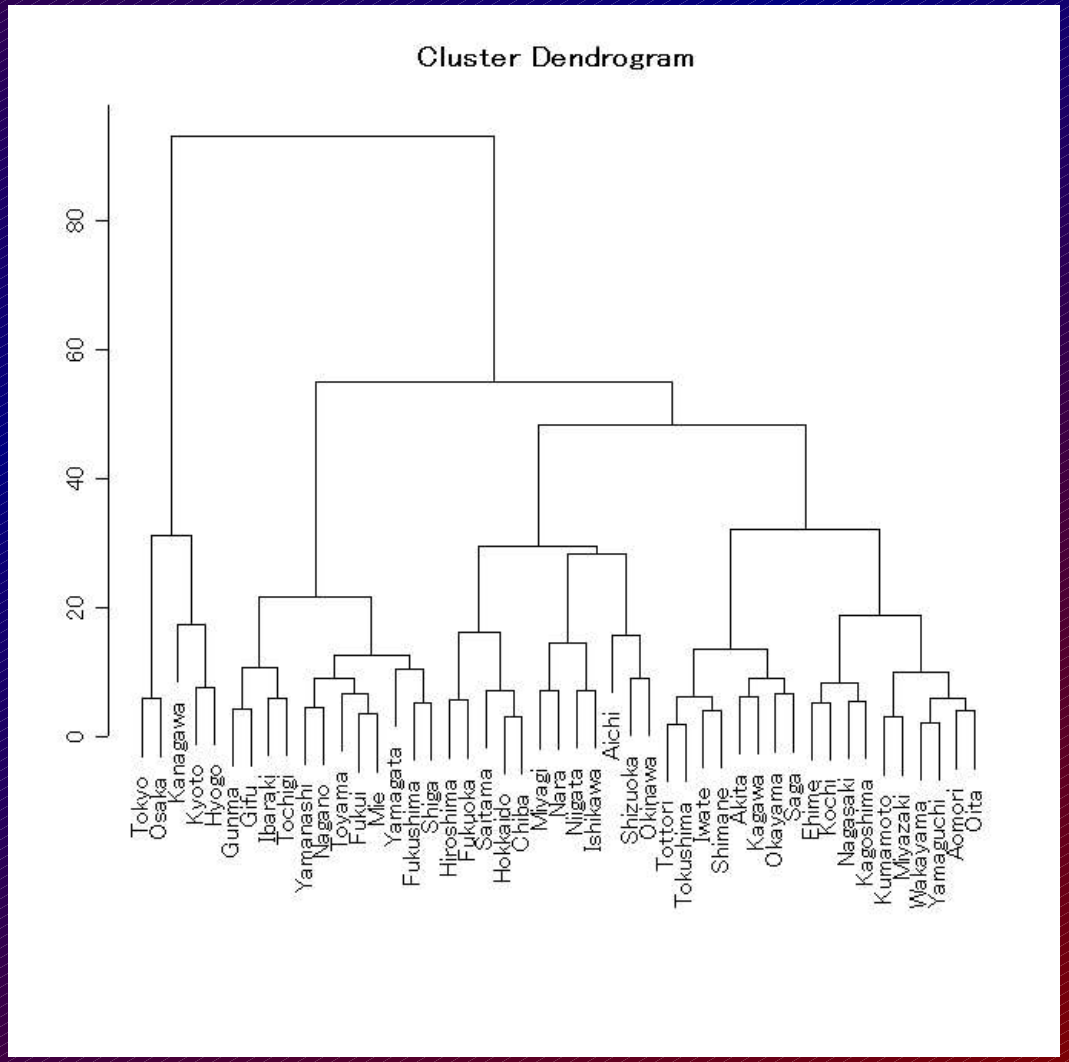
- 多数の実測された変数から因子としての構成概念を多数推測し(この部分は測定方程式と呼ばれる)，これら構成概念間でのパス解析を行い(この部分は構造方程式と呼ばれる)，モデル全体としての説明力が高ければ(適合度指標が高い値を示せば)，そのモデルが現実を反映していると考える
- 因子分析と回帰分析を同時に行うモデルであるという言い方もされる
- 最終的に得られたモデルはパス図として表示することが多い
- R では sem というライブラリを追加インストールすれば実行可能だが難しい。example(sem) で様子はわかる。

クラスター分析

- オブザーベーション間の関係を表したいときに使う
- 距離行列に基づいて個体を結合しながらクラスターを積み上げていく（出力は樹状図またはネットワーク図になる）階層的手法と，予めいくつくらいの塊（クラスター）に分かれるかを決めて，データを適当に振り分ける非階層的手法がある
- 距離行列の計算法にも多々あり，結合法にも多々ある。いくつかの方法でやってみて，樹状図に差がなければ，そのクラスター分析の結果は安定していて，信頼できるといえる。樹状図が大きく変わるようなら信頼できない
-

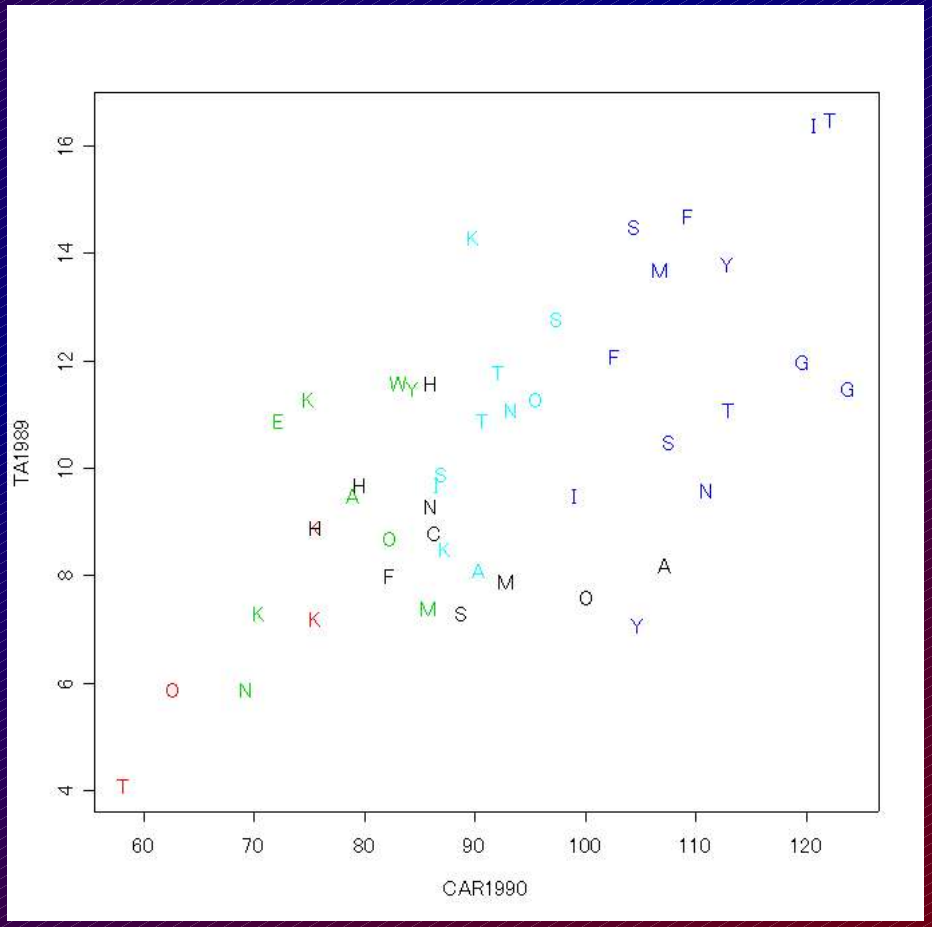
階層的クラスター分析の実際

- 県別自動車保有データの場合，データを読んでから，
  dis<-dist
  (mat,method="euclidean")
  clus<-hclust(dis)
  plot(clus,paste(PREF),
  xlab="",ylab="",sub="")
  とすれば樹状図がかける。
- 組み込みデータでも同様。USArrests で試してみるとよい。



非階層的クラスター分析の実際

- 県別自動車保有データでの k-means 法で5つのクラスタを指定すると,
clus5<-kmeans(mat,5)
plot
(mat,col=clus5$cluster,
pch=paste(PREF))
のようにすればよい。
- data(USArrests) でも同様にできる。



データマネジメント

- データをどうやって管理するか？
- 固定したデータ
  - CD-R や DVD-R など，書き換えできない形で保存すればよい
  - ソフトウェア独自形式でなく，データ構造そのものの説明を含む形か，あるいはテキスト形式で保存すべき。
- 継続蓄積されるデータ
  - どこに蓄積するか？　パソコン？　サーバ？
  - どうやって継続更新するか？
  - どうやってバックアップするか？

データマネジメントの実際

- 表計算ソフト（Microsoft EXCEL や OpenOffice.org calc など）に手入力→タブ区切りテキスト形式で保存
  - 小さければ見通しはいいが，sequential data になる
- パソコン内に DBMS（Microsoft Access や MySQL など）を起動して，それにデータ管理をさせる（通常，データは tree または list 構造をもつ）。この場合，パソコンのハードディスクの信頼性がかなり高くなくてはいけない。
  - DBMS を表計算ソフトやブラウザから呼び出す
  - DBMS 自身が内蔵しているユーザインターフェースを使う
- サーバに DBMS を起動し，データ管理はサーバ上で行う。DBMS の制御は web 経由で行うのが普通。無料のソフトならば，apache httpd ＋ php4 ＋ postgreSQL を使うのが流行。