多変量解析
<初歩の初歩>に入門

多次元尺度法

なかもとけいこ
kenakamoto@nifty.com

多次元尺度法とは

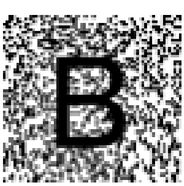
- 英語でいうと Multidimensional Scaling (MDS).
- 対象間の距離データが与えられたときに，多次元空間上に対象を上手く配置するための一手法.
  - データは距離として表せるものなら何でも良い
    - (非) 類似性 (similarity is proximity?)
    - 連想関係の強さ，人の仲良し度
- 計量的MDSと非計量的MDSがある.
  - 計量的MDSでは入力データがホントの意味で「距離」データ (つまり比率尺度)になってる必要がある.
  - 非計量MDSでは，入力データと距離との対応付けが単調増加(単調減少)になっていればよい.
  - よく使われるのは非計量の方.

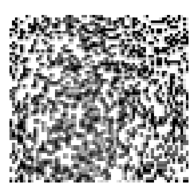
適用例

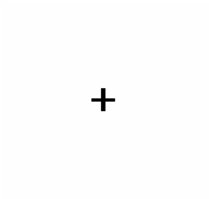
- 見間違いデータ (視覚呈示での混同率)

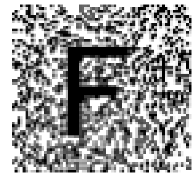
	B	C	D	E	F	G	H	L	O	P	Q	R
B												
C	.01											
D	.04	.01										
E	.04	.03	.03									
F	.03	.02	.02	.06								
G	.04	.11	.02	.02	.01							
H	.03	.01	.03	.04	.04	.06						
L	.02	.02	.04	.06	.05	.02	.07					
O	.01	.09	.02	.01	.01	.06	.01	.01				
P	.03	.02	.06	.02	.04	.01	.03	.03	.02			
Q	.01	.06	.01	.01	.01	.04	.01	.01	.10	.01		
R	.03	.01	.02	.01	.04	.02	.05	.02	.01	.06	.01	

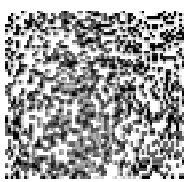
+





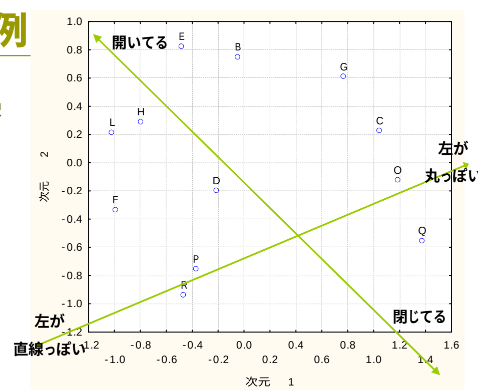






適用例

布置と解釈



MDSの利点

- そもそも，どういう変数が対象を特徴づけるのに有効か検討がつかないとき．
  - いろいろな変数は関与してそうなんだけど，具体的に何を測定すればいいか分からない．
  - 測定すべき変数が分かっているなら，それを測定して，主成分分析なり因子分析なりすればよい．
- 単に対象間の類似性データを得ることができれば，その類似性を決定している(心理的)次元を大づかみに把握することができる．
  - MDSの布置から得られた次元の解釈の妥当性をさらなる調査や実験で確かめればよい．

というわけで本日のお品書き

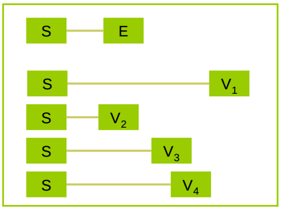
- (非)類似性データを得るための(心理学的)手法
- 非計量的MDSの概念的理解
  - 具体的な計算アルゴリズムには立ち入りません．

(非)類似性データの種類 (1)

- 直接判断
  - 評定法

    りんご
    バナナ　1 2 3 4 5 6 7 8 9 10 11

  - 多重比率判断法



(非)類似性データの種類 (2)

- 対比較法

  四つ組法: りんご バナナ ナシ 桃

  三つ組法: りんご バナナ 桃

- 順位法
- 分類法

- 間接的手法
  - 刺激の混同率 (刺激－刺激混同率, 刺激－反応混同率)
  - 刺激の汎化勾配

- もちろん, 変数(群)から非類似性を計算する方法もある.

類似性データが得られたら…

- 基本のMDSでは, 非類似性を距離で表せると考えるので, [aがbに似ている程度]と[bがaに似ている程度]は等しい (対称性)と仮定する.

- つまり(非)類似性データは対称行列で表される.
  - 対角要素の非類似性は0. (あるモノのそのもの自身への類似性は極めて高い (というか同じ) なので, 非類似性は0と考える)
  - データの取得法によっては, 対称にならないこともままあるが, そういうときは何らかの前処理で対称行列にするか, 非対称MDSを行う.
    - ただし非対称MDSは, 一般的な統計解析ソフトにパッケージとして含まれているわけではない (し, なかもと自身やったことはない)

非計量MDS (1)

- 非計量MDS nonmetric MDS で扱うデータは, (名義尺度や)順序尺度.
  - 類似性データを間隔尺度, 比率尺度でとるのは難しいので, 非計量で解析するのが無難.

- 非類似性の大小関係 (だけ) が距離の大小関係でも維持されるように, 非類似性(データ)→距離 (類似性空間上での布置) への変換を行う (単調変換)
  - つまり, 非類似度が $O_{ij} \geq O_{kl}$ ならば, 距離は $\hat{d}_{ij} \geq \hat{d}_{kl}$
  - $\hat{d}_{ij} = M(O_{ij})$ という変換で, Mが単調変換.

非計量MDS (2)

- $\hat{d}$ (d-hat)のうまい決め方として, 損失関数L(d, d-hat)を定義し, これを最小にするようなd_{ij}の最小2乗近似d_{ij}-hatを求める.

  - d_{ij}は空間上の座標値Xから求まる $d_{ij}=F(x_i,xj)$ Fは距離関数.
  - でも, 座標値自体分からないので, Lもまた分からない.
  - なので, 適当な初期値X(0)から初めて, 逐次的にXに到達するようにしなければならない (つまり, 反復推定).

非計量MDS (3)

- ここで, S回目の反復で, X(S)からD(S)を得たとき,

$$L = L(D^{(S)}, \hat{D}^{(s)})$$

- を最小にする変換 T が存在し,

$$\hat{D}^{(s)} = T(D^{(S)})$$

- によりD-hatの推定値が求められるとすると, Lは

$$L = L(D, T(D))$$

- であらわせる. Lを最小にするようなXを求めればいいのだから, (Lが連続で微分可能なら)LをXで微分して0とおき (つまり極値を求める), これに到達するようにすればよい.

非計量MDS (4)

- このL最小化のために絡んでくる式が，
  - 損失関数　$L = L(D^{(S)}, \hat{D}^{(s)})$
  - 距離関数　$D^{(s)} = F(X^{(s)})$
  - 推定関数　$\hat{D}^{(s)} = T(D^{(S)})$
  - 距離関数には，ユークリッド空間モデルが採用されることが多い．
  - 損失関数は，o_{ij}-> d_{ij}-hatにあうようにd_{ij}の変換を行ったときの誤差に相当するので，得られた解のデータへの当てはまりの良さに絡んでくる．
  - 推定関数は，d_{ij}-hatにd_{ij}を近づけていくときの具体的な方法に関連する．

非計量MDS (5)

- 損失関数とストレス

$$S_0^{\,2} = \sum\sum (d_{jk} - \hat{d}_{jk})^2$$

粗ストレス raw stress

$$S_1 = \sqrt{\frac{\sum\sum (d_{jk}^{\,2} - \hat{d}_{jk})^2}{\sum d_{jk}^{\,2}}}$$

粗ストレスを基準化したもの (stress 1)

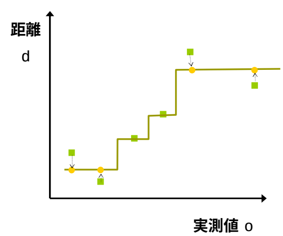
  - 0に近づくほど，当てはまりがよいことになる．
- Rule of thumb として，stressの値が .05以下だとgood，.10以下だとfair，.20以下だとpoor，それ以上はお話にならないっていう感じ，らしい．

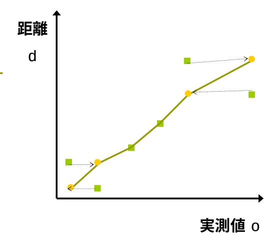
結果の読み方

- 当てはまりの検討
  - Stressの値
  - シェパード・ダイアグラム　(D-hatと距離のプロット)
- 布置の解釈
  - 解釈上，次元は回転させてもよい(し，歪んで解釈しても可)
  - MDSのx_1，x_2...の次元は計算上出てくるものなので，実際的な意味はない (入力データには非類似性　距離の情報しかないことに注意せよ)

非計量MDS (6)

- 推定関数
  - 単調回帰法 (Kruskal)
  - 順位イメージ法 (Guttman)





その他のMDS

- MDSには色々な種類がある．
  - INDSCAL
    - 個人差を扱うMDS
    - 類似性データを個人ごとにとり(たとえば，被験者n人にm個の単語の類似性を訪ねるなど)，m個の対象について標準的な類似性空間(布置)をつくり，その空間への重み付けの差として個人差を表現する．
  - ALSCAL
  - 非対称MDS
    - その名のとおり，類似性が非対称なことをそのまま表現できるように拡張，修正されたMDS

今回のまとめ

- 今回は次のことを学びました．
  - MDS は(非)類似性データをもとに，対象を(ユークリッド)空間上に配置する方法です．
  - 類似性データの取得方法には色々な方法があります．
    - 対比較系の手法はキレイなデータを取りやすいけど，組み合わせ爆発を起こす．
    - 他の方法も色々と考案されてるので，いい方法を選ぶとよいです．
  - 非計量MDSでは，非類似性の大小関係を維持するように，空間上の布置を決めます．
    - 実際の計算アルゴリズムはややこしいですが，計算はコンピュータがやってくれるので，まあそれはそれで．
    - 布置の解釈可能性を考えると，3次元解で何とかなる場合くらいまでが実際的な適用範囲でしょう．
  - 線形の空間に類似性が納まらないときには，自己組織化マップ(Self Organizing Map; SOM)という手もあります．
    - (最近，新版の訳もでたみたいですし)．

MDSを勉強するために…

- 多次元尺度法はちょっと特殊な多変量解析（なにせ，入力されるデータは(非)類似性の一変量だけだし）なので，一般の「多変量解析入門」みたいな本には載っていないことが多いです（普通に載っているのは，重回帰分析，主成分分析，因子分析，判別分析など）．
  - J･B･クラスカル，M･ウィッシュ著 ; 高根芳雄訳 (1980).多次元尺度法（人間科学の統計学1）朝倉書店
  - 高根芳雄 (1980). 多次元尺度法 東大出版
  - 齋藤尭幸 (1980). 多次元尺度構成法 朝倉書店
- なお，今回のスライド作成には下記を参考にしました．
  - 田中良明 (1977). 心理学的測定法 第2版　東京大学出版会 (第14章 多次元尺度法)
    - 中本が学部生のときに使っていた教科書ですが，未だにときどき参照します．やわらかい書き方ではないですが，とてもよい本だと思います．

マニアックなおまけ: 類似性と距離

- 通常のMDSでは，距離の公理が類似性にも成り立つと仮定している．
  - つまり，d(a,b)>=0，かつ d(a, b)=0ならばa=b　(最小性)
    d(a, b) = d(b, a)　(対称性)
    d(a, b)+d(b,c)>=d(a,c)　(三角不等式)
- しかし，これが必ずしも正しくないことは古くから知られている (少なくとも，Tversky, 1977から)
- また，類似性の知覚はたぶんに文脈依存的であり，課題要求，刺激セット，対象に対する知識などで変化する．
- 類似性を静的に捉えて，それをモトに心的表現のモデルを構築することには根本的に無理がある．
  - 自分のやってることを否定するようですが…
- 比較というプロセスの中で類似性を捉えた方が，おそらく望ましいでしょう．

最後のおまけ

- 工学的な観点から，パターンの分類や認識に「使える」方法を網羅的に紹介した本として，
- Duda, P.O., Hart, P.E., & Stork(著)尾上 守夫(監修) (2001) パターン識別　新技術コミュニケーションズ
- という本があります．分厚くて大きくて高い本ですが，(私もとても全部は読んでないけど)定評のある非常に良い本です．
- パターン認識に関わることは，MDS, SOM,判別分析，SVMなど手法の由来に関わらず一通り載っているバイブル的な本．
- 工学系の研究者，数理的な仕事をする心理学者と一緒に活動する予定があるなら，一度手にとって眺めるくらいはしてもいいでしょう，