集約的シンボリックデータの多重対応分析

清水　信夫　　学際統計数理研究系　助教

• 研究の背景および動機

連続(実数)変数とカテゴリー変数が混在する大規模多変量データにおいて、自然に分けられた集団が存在し、それらに関する情報に興味がある場合を考えたい

- 各個体のデータ全てを網羅せずとも、各集団ごとの特徴的な値だけを使うことによりデータ全体の特徴を簡潔に捉えたい
- 各集団ごとに変数のいくつかの記述統計量(平均、分散、etc.)の集合を新たなデータと考える
- 各変数ごとの性質に加え、2つの変数間の関係を表す記述統計量についてもあわせて考える
  ⇒これらの記述統計量をまとめた新たなデータを**集約的シンボリックデータ(Aggregated Symbolic Data, ASD)**と呼ぶ
- 様々なデータ解析手法において、ASDの情報で結果を表すことを考えたい
  - 各集団ごとのASDのみを用いて、連続(実数)変数とカテゴリー変数が混在するデータの**多重対応分析(Multiple Correspondence Analysis, MCA)**を行いたい

• ASDの定義

p個の連続変数およびq個のカテゴリー変数(カテゴリー変数kにおけるカテゴリー値の数はm_k個)のデータ集合Xのうち、集団 $g(g = 1, \cdots, G)$ におけるデータ行列$X^{(g)}$ は

$$X^{(g)} = \begin{bmatrix} x_{11}^{(g)} & \cdots & x_{1p}^{(g)} & x_{11}^{(g,1)} & \cdots & x_{1m_1}^{(g,1)} & \cdots & x_{11}^{(g,q)} & \cdots & x_{1m_q}^{(g,q)} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ x_{n^{(g)}1}^{(g)} & \cdots & x_{n^{(g)}p}^{(g)} & x_{n^{(g)}1}^{(g,1)} & \cdots & x_{n^{(g)}m_1}^{(g,1)} & \cdots & x_{n^{(g)}1}^{(g,q)} & \cdots & x_{n^{(g)}m_q}^{(g,q)} \end{bmatrix}$$

$n^{(g)}$個のデータをもつ$X^{(g)}$において、左のp列がp個の連続変数値、それ以外がq個のカテゴリー変数ごとのダミー変数値である。ASDはこれらより以下の要素で表される。

- データの個数: $n^{(g)}$ (0次のモーメント)
- 列和 : $s^{(g)} = \mathbf{1}'_{n^{(g)}} X^{(g)}$ (1次のモーメント)
- 積和行列 : $N^{(g)} = X^{(g)\prime} X^{(g)}$ (2次のモーメント)

• 連続変数を含む場合の多重対応分析

各カテゴリー変数の各カテゴリー値にスコアを設定し、それらの各分散の総和がある種の条件下で最大となるようにスコアを求める方法が(一般的な)多重対応分析である

- 連続変数についてはn個の異なるカテゴリー値にそれぞれ1つずつデータが入るカテゴリー変数と置き換え、元からのカテゴリー変数と同様に考える
  ⇒連続変数をカテゴリー変数とみなした場合の各カテゴリー値に関するスコアは全て同じになる
- 各カテゴリー変数の各カテゴリー値、および各連続変数に対するスコアは、$m \times m$ 行列 $(m = p + m_1 + \cdots + m_q)$ の固有値分解の最大固有ベクトルに基づく値として求まる
  - この場合のスコアの第j主成分を $\boldsymbol{a}_j$ 、および各ダミー変数行列をスコアベクトルに掛けたものの総和を$\boldsymbol{h}_j$とする

• データ集合全体および各グループのASDに対する多重対応分析

ASDの多重対応分析においては各連続変数ごとのそれぞれの値に関するスコアは表せない(個々のデータはASDとして持っていないため)
ただし、各連続変数のスコアの平均や、そこを中心とする各グループの広がりを表す統計量はASDのみから表現することが可能である

- 第(j_1, j_2) 主成分におけるグループgの平均$[h_{j_1}^{(g)}, h_{j_2}^{(g)}]'$および分散共分散行列 $\begin{bmatrix} \sigma_{j_1 j_1}^{(g)} & \sigma_{j_1 j_2}^{(g)} \\ \sigma_{j_2 j_1}^{(g)} & \sigma_{j_2 j_2}^{(g)} \end{bmatrix}$ をもつ2変量正規分布の50%楕円として表す
- 第(j_1, j_2) 主成分における各カテゴリー変数の各カテゴリー値、および各連続変数に対するスコア $[\boldsymbol{a}_{j_1}, \boldsymbol{a}_{j_2}]$の各行を原点からの矢印で示す

• 実データへの適用例

表1はある不動産検索サイトにおける2017年時点の東京23区の賃貸住宅データ(有効総件数が約419万件)の一部である。このデータは2種類の連続変数、173種類のカテゴリー変数を含む。
ここではカテゴリー変数について14種類を選び、23の区ごとにASDを使用して多重対応分析を行った結果を示す。

表1: 不動産検索サイトにおける東京23区の賃貸住宅データ（一部）

Ward	Log.Rent	Log.Area	Dep.Mos	...	Rms.	Struct.	...	Loc.Fl.	ELV	...	CATV
Chuo	5.18	1.61	2	...	1	RC	...	7	1	...	1
⋮	⋮	⋮	⋮	...	⋮	⋮	...	⋮	⋮	...	⋮
Minato	5.72	2.09	3	...	3	SRC	...	3	1	...	1
⋮	⋮	⋮	⋮	...	⋮	⋮	...	⋮	⋮	...	⋮
Ota	5.00	1.65	1	...	2	Steel	...	3	0	...	1
⋮	⋮	⋮	⋮	...	⋮	⋮	...	⋮	⋮	...	⋮
Nakano	4.89	1.31	1	...	1	RC	...	2	1	...	0
⋮	⋮	⋮	⋮	...	⋮	⋮	...	⋮	⋮	...	⋮
Adachi	4.83	1.35	0	...	0	Wood	...	1	0	...	0
⋮	⋮	⋮	⋮	...	⋮	⋮	...	⋮	⋮	...	⋮

図1は連続変数とカテゴリー変数を合わせたデータ集合について、第1主成分および第2主成分における50%の確率楕円およびその中心を23区ごとにプロットしたものである。この図においては千代田・中央・港の都心3区の楕円の傾きがその他の区と大きく異なっており、3区の物件の状況が他の区のそれとは大きく異なっていることを示す。また図2はそれに**各連続変数(黒)**および**各カテゴリー変数の各カテゴリー値(赤)**の各主成分のスコアを矢印で示し中心部を拡大表示したものである。ここでは矢印の長さおよび向きではなく、それぞれの矢印の間の角度に注目する。その時に異なる変数およびカテゴリー値の間の角度が小さい場合は両者の相関が高めであることを示す。

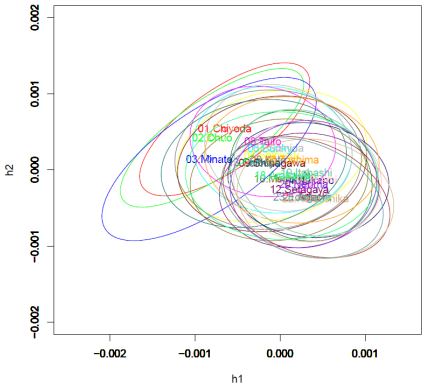

図1. MCAにおける第1&第2主成分の表示（軸なし）

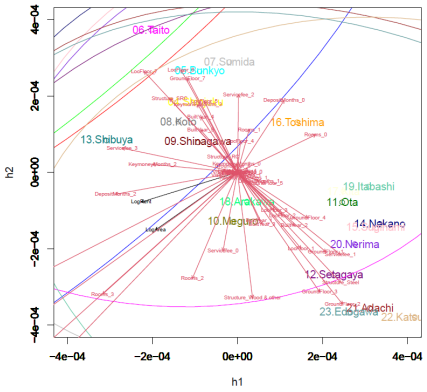

図2. MCAにおける第1&第2主成分の表示（軸あり&中心部拡大）