対応分析手法を用いた 集約的シンボリックデータの表現

2020年10月27日　統計数理研究所　オープンハウス

清水　信夫　　データ科学研究系　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.)の集合をデータと考えて解析⇒**集約的シンボリックデータ（Aggregated Symbolic Data, ASD）**と呼ぶ

・変数の型や変数の組み合わせの方法によらず、ASDで使用される同種の記述統計量は同一の基準で求めたい

　・カテゴリー変数同士の関係性を分析する方法としてよく用いられる対応分析を連続変数が含まれる場合についても適用

　・対応分析における変数のスコアの概念を用いて同種の記述統計量を統一的に記述し、それらを用いて各集団を可視化

【変数型が混在する大規模データにおける集団の表現】

・p個の連続型変数および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

・連続変数およびカテゴリー変数に対しては、各々の変数内および異なる2変数間の関係の記述統計量を2次モーメントまでの範囲で定義

【カテゴリー変数を含む場合の変数間の相関】

2つの異なるカテゴリー変数間の相関が最大となるように各変数のカテゴリー値の最適な順番を導出し記述する方法として対応分析があり、その手法を用いてカテゴリー変数同士の相関およびそれを与える各変数のスコアを考える。

集団g における2つの名義変数 k_a, k_b のダミー変数行列をそれぞれ$X^{(g,k_a)}, X^{(g,k_b)}$ とすると$X^{(g,k_a)\prime}X^{(g,k_b)}=N^{(g,k_ak_b)}$ は2変数の分割表となる。ここで $a=\left[a_1\cdots a_{m_{k_a}}\right]', b=\left[b_1\cdots b_{m_{k_b}}\right]'$ としてスコア$X^{(g,k_a)}a$とスコア $X^{(g,k_b)}b$ の相関が最大となる場合を考える。名義変数同士の相関は $N^{(g,k_ak_b)}$ を標準化した行列を固有値分解した場合の最大固有値λ_1となり、それを与える(a,b) はλ_1に対応する最大固有ベクトルの組 (a_1,b_1) を用いて求まる。これは対応分析における各変数のスコアでもある。

分割表に順序変数が含まれる場合は、順序変数について順番を固定したまま、相関に対応する値および各変数のスコアを名義変数同士の相関と同様に考えることができる。2つのカテゴリー変数のうち1つの順序変数をk_a とするとき、対応するスコアを線形増加数列 $a=\left[1\cdots m_{k_a}\right]'$を標準化した形で表してもう1つの変数のスコアおよび2変数間の相関を計算できる。

連続変数とカテゴリー変数の相関については、連続変数のスコアを $n^{(g)}$ 個の連続変数の値を標準化したベクトルで表してもう1つの変数のスコアおよび2変数間の相関を計算できる。

【カテゴリー変数における平均および標準偏差に対応する値】

1つのカテゴリー変数についてはデータ集合Xおよび集団 $X^{(g)}$ いずれの場合も各カテゴリー値の個数の分布しか情報がない。そこで、カテゴリー変数k_aにおけるXのダミー変数行列を$X^{(k_a)}$とし、スコア$X^{(k_a)}a$の平均が0、標準偏差が1となるようなノルム1の係数ベクトルaを考え、これを集団 $g=1,\cdots,G$ におけるスコアに適用することにより各集団の平均および標準偏差を考える。名義変数の場合は条件をみたすa を、順序変数の場合は線形増加数列$a=\left[1\cdots m_{k_a}\right]'$を標準化した値をそれぞれ各集団に適用することにより、ASDが持っている情報から平均および標準偏差を計算できる。

【不動産情報データへの適用例】

表1はある不動産検索サイトにおける2013年時点の東京23区の賃貸住宅データ（有効総件数が約79万件）の一部である。このデータは2種類の連続変数、7種類の順序変数、55種類の名義変数を含む。

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）

No.	区	賃料	面積	物件種別	構造種別	・・・	管理形態
1	荒川区	8.25	26.83	マンション	鉄筋コン	・・・	記載なし
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
4588	港区	22.30	71.28	マンション	鉄筋コン	・・・	巡回管理
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
498088	足立区	6.40	33.34	アパート	軽量鉄骨	・・・	記載なし
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
714202	新宿区	16.40	55.64	マンション	鉄骨鉄筋	・・・	常駐管理
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

表2: 名義変数「物件種別」における各区の平均および標準偏差の値（一部）

物件種別	千代田	中央	・・・	目黒	・・・	足立	・・・	江戸川
平均	-0.478	-0.482	・・・	-0.004	・・・	0.339	・・・	0.529
標準偏差	0.131	0.085	・・・	0.996	・・・	1.198	・・・	1.244

表3: データ全体における各変数間の相関係数値の2乗値（一部）

2 乗相関	賃料（連続）	面積（連続）	部屋数（順序）	地上階（順序）	物件種別（名義）	構造種別（名義）
部屋数（順序）	0.192	0.483	1.000	0.000	0.041	0.005
地上階（順序）	0.189	0.042	0.000	1.000	0.904	0.795
物件種別（名義）	0.139	0.088	0.054	0.904	1.000	0.932
構造種別（名義）	0.166	0.054	0.008	0.795	0.932	1.000

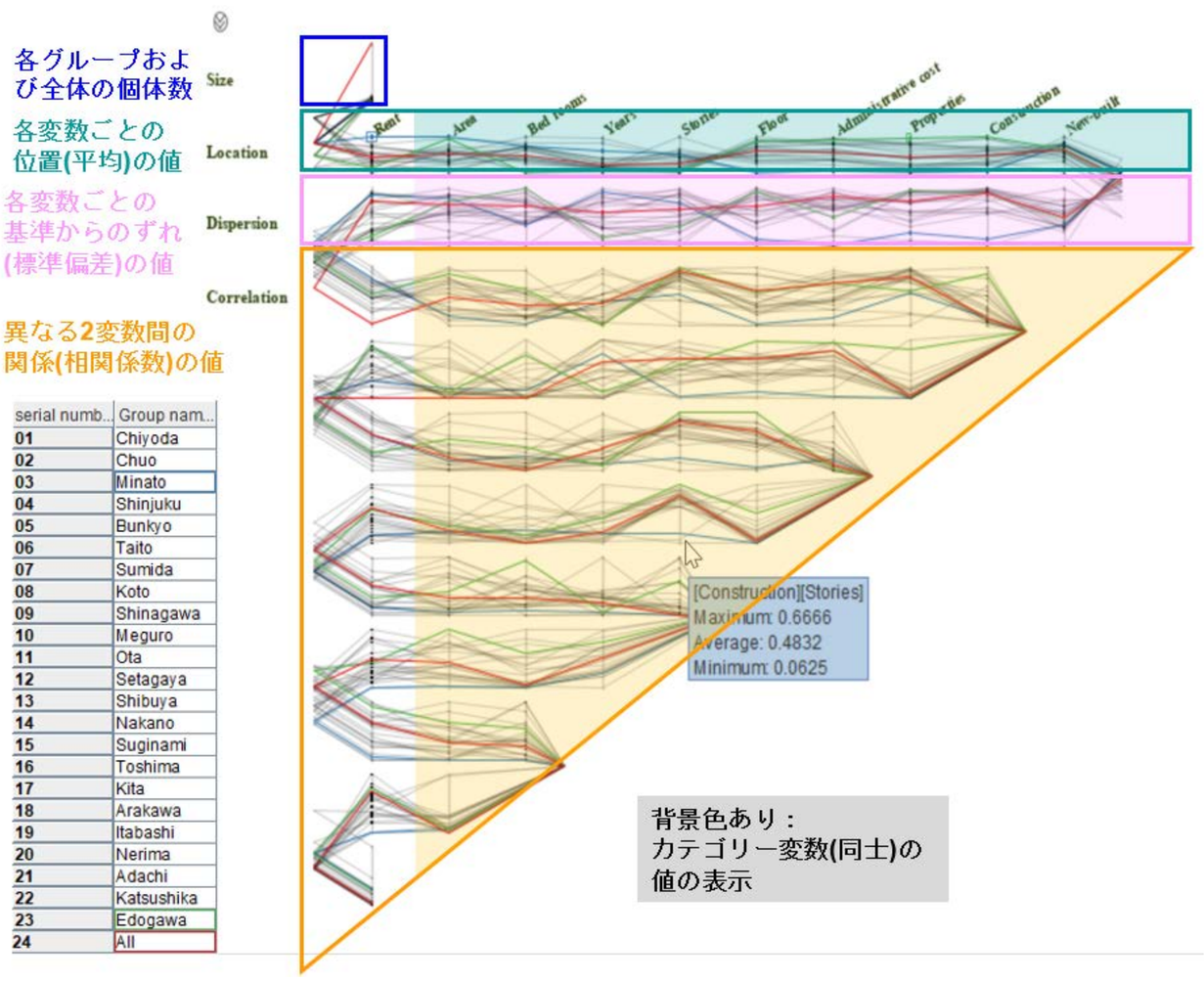

図1: 元データから10変数(連続変数2,順序変数3,名義変数5)のみ抜粋した場合のデータ全体および各区のASDを平行座標プロットにより表示

表2は名義変数「物件種別」におけるデータ全体の平均を0、標準偏差を1と考えた時の各区の平均および標準偏差を示したものである。また表3はデータ全体における各変数間の相関係数の2乗値を示したものである。これらはいずれも各変数に関するASDの値から対応分析による各変数のスコアを用いて求められる。このような各変数の平均および標準偏差、および2変数間の相関係数の2乗値をデータ全体および各区において求め、それらを平行座標プロットにより表示したものを図1に示す。最上段から最下段まで繋がっている1本の線が各々の集団を表し、赤い線がデータ全体を示す。