2つのカテゴリー変数間の相関係数

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
- 各変数ごとの性質だけでなく、2つの変数間の関係を表す記述統計量についても考えたい
  - 2つの連続変数同士であれば**Pearson相関係数**が定義されている
  - 2つのカテゴリー変数同士の各カテゴリー値ごとの組み合わせは分割表で表され、これより2つのカテゴリー変数間の相関に相当する統計量が考えられる
  - 順序変数同士の場合における分割表から求められる関係として**ポリコリック相関**、順序変数と連続変数の組み合わせにおける関係として**ポリシリアル相関**が存在
    ⇒Pearson相関係数などとの整合性がない
    ⇒名義変数の場合が考慮されておらず、カテゴリー値の全ての並べ替えの中での最大相関を考えると計算時間が長くなる
  - 名義変数が含まれる場合における関係の値について、既存の相関係数と対応が取れる形の指標を定義したい

• 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

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

- $n^{(g)}$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

• カテゴリー変数同士の相関の定義

2つの異なるカテゴリー変数についてそれぞれのスコアを考え、それらの間の相関が最大となる場合をカテゴリー変数間の相関とする
⇒連続変数における、標準化された数値のPearson相関に対応

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。
スコアに条件をつけない場合、名義変数同士の相関は確率行列$P^{(g,k_ak_b)} = N^{(g,k_ak_b)}/n^{(g)}$を標準化した行列を特異値分解した場合の最大特異値$\lambda_1$となる。
それを与える(a, b) はλ_1に対応する最大特異ベクトルの組 (a_1, b_1)を用いて$(a, b) = (D_{k_a}^{-1/2}a_1, D_{k_b}^{-1/2}b_1)$と求まる。
（D_{k_a}, D_{k_b}は、それぞれ$P^{(g,k_ak_b)}$ のk_a, k_b 方向の周辺分布ベクトル $\boldsymbol{v}_{k_a}, \boldsymbol{v}_{k_b}$ の各成分を対角成分とする正方行列とする）

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。
2つのカテゴリー変数のうち順序変数が1つだけでそれをk_a とするとき、対応するスコアを線形増加数列を成分とするベクトル $a_{ns} = \left[1, 2, \cdots, m_{k_a}\right]'$に関して

$$\tilde{a} = \frac{(I_{m_{k_a}} - \mathbf{1}_{m_{k_a}}\boldsymbol{v}_{k_a}')a_{ns}}{\sqrt{a_{ns}'(D_{k_a} - \boldsymbol{v}_{k_a}\boldsymbol{v}_{k_a}')a_{ns}}}$$

と変換した形で表し、2変数間の相関が最大となるようにもう1つの変数のスコアbを一意に計算できる。
2つとも順序変数のときは、それぞれのスコアを線形増加数列の変換の形で表し、相関を求めることができる。

• 名義変数が含まれる場合の相関についての検証と修正

名義変数においてはカテゴリー値の並び順は任意に決定可能だが、相関を考えるには全ての並びの中での特定の並び順の場合における順序変数同士の場合と一致させないとカテゴリー変数を含む全体の様々な場合を考える上で整合性が取れない
⇒スコアに何らかの条件をつける必要がある

- どんな条件をつけるか？
  ⇒条件をつけない場合において求められた最適なスコアにおける成分が昇順となるようにカテゴリー値の順番を並べ替える
  ⇒条件をつけない場合の並べ替え済みの最適なスコアを、単調増加数列を成分とするベクトルの変換により求められたスコアで置き換える
- これにより名義変数のカテゴリー値の最適な並べ替えをした場合の相関を、順序変数同士の相関と同様に考えられる

• 名義変数同士の相関係数の導出例

元々の分割表は適切に並べ替えるとブロック対角行列に近い形となり、スコアに条件をつけない場合だと1にかなり近い第1特異値が相関として導出される。
これに対し、スコアに特定の並べ替えに基づいた場合という条件を付けると、ブロックの分布を反映して1よりもある程度小さな値を相関係数として返しており、
より妥当な値に近いと考えられる。

$$N = \begin{bmatrix} 35 & 20 & 1 & 30 & 0 & 25 \\ 20 & 30 & 1 & 20 & 0 & 20 \\ 1 & 1 & 40 & 1 & 50 & 1 \\ 20 & 20 & 1 & 30 & 0 & 40 \end{bmatrix} \quad a = \begin{bmatrix} -0.549 \\ -0.544 \\ 1.825 \\ -0.550 \end{bmatrix} \quad b = \begin{bmatrix} -0.543 \\ -0.540 \\ 1.742 \\ -0.545 \\ 1.916 \\ -0.547 \end{bmatrix}$$

$$\lambda_1 = 0.952$$

図1．元々の分割表と各変数ごとのスコアと第1特異値

$$\widetilde{N} = \begin{bmatrix} 40 & 30 & 20 & 20 & 1 & 0 \\ 25 & 30 & 35 & 20 & 1 & 0 \\ 20 & 20 & 20 & 30 & 1 & 0 \\ 1 & 1 & 1 & 1 & 40 & 50 \end{bmatrix} \quad \tilde{a} = \begin{bmatrix} -1.264 \\ -0.369 \\ 0.525 \\ 1.419 \end{bmatrix} \quad \tilde{b} = \begin{bmatrix} -1.291 \\ -0.685 \\ -0.080 \\ 0.525 \\ 1.130 \\ 1.736 \end{bmatrix}$$

$$\rho = 0.657$$

図2．並べ替えた分割表と置き換えたスコアと相関係数ρ