周辺分布と関連性が与えられた場合の複数のカテゴリー変数の同時確率の生成

2021年6月18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

・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求められる

・カテゴリー変数が3つ以上の場合の相互間の関係については2変数ごとの分割表の集合(Burt表)を用いて表される

　・全体的な関係が解りにくく、より明快な関係を導き出すための理論や手法が必要

　・2つの名義変数間の相関に相当する統計量が与えられた場合に、その情報が保持された3つ以上のカテゴリー変数に関する確率の組み合わせをテンソルとして表現したい

【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

【 2つの名義変数間の関係性を表す値】

・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

・$a = \left[a_1 \cdots a_{m_{k_a}}\right]', b = \left[b_1 \cdots b_{m_{k_b}}\right]'$ 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、この値は名義変数間の相関に対応する値となる

・この値は $N^{(g,k_ak_b)}$ を標準化した行列を特異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

【 3つ以上の名義変数による確率テンソルのASDによる表現】

・q個の名義変数（変数kにおけるカテゴリー値の数は m_k個）の確率テンソルPを考える

・名義変数 $k = 1,\cdots,q$ の周辺分布ベクトルを $p_1,\cdots,p_q$、ベクトルの各要素を対角成分とする行列を$D_1,\cdots,D_q$ とする

・名義変数 k_1, k_2 に関する確率行列を$P_{k_1k_2}$とすると $S_{k_1k_2} = D_{k_1}^{-1/2}(P_{k_1k_2} - p_{k_1}p_{k_2}')\,D_{k_2}^{-1/2}$ は

$$S_{k_1k_2} = U_{(k_1k_2)}^{(k_1)}\,\Lambda^{(k_1k_2)}U_{(k_1k_2)}^{(k_2)\prime}$$

と特異値分解が可能である（ $\Lambda^{(k_1k_2)}$:固有値を対角成分とする行列、$U_{(k_1k_2)}^{(k_1)}, U_{(k_1k_2)}^{(k_2)}$:各固有値に対応する k_1 および k_2の固有ベクトルを含む行列）

・ $\Lambda^{(k_1k_2)}$ の最大固有値 $\lambda_1^{(k_1k_2)}$ はk_1とk_2の相関係数に相当する値であり、$U_{(k_1k_2)}^{(k_1)}, U_{(k_1k_2)}^{(k_2)}$ はそれぞれに直交するベクトルとして定められる

・Pが名義変数k_1, k_2, k_3 による確率テンソルである場合 $P = \Lambda^{(k_1k_2k_3)} + p_{k_1}\otimes p_{k_2}\otimes p_{k_3}$

ただし$\Lambda^{(k_1k_2k_3)}$ および Δ は

$$\Lambda^{(k_1k_2k_3)} = (\Lambda^{(k_1k_2)} \times_{k_1} U_{(k_1k_2)}^{(k_1)} \times_{k_2} U_{(k_1k_2)}^{(k_2)} + \Lambda^{(k_1k_3)} \times_{k_1} U_{(k_1k_3)}^{(k_1)} \times_{k_3} U_{(k_1k_3)}^{(k_3)} + \Lambda^{(k_2k_3)} \times_{k_2} U_{(k_2k_3)}^{(k_2)} \times_{k_3} U_{(k_2k_3)}^{(k_3)} + \Delta)/3$$

かつ

$$\Delta \times_{k_3} 1'_{m_{k_3}} = 2\Lambda^{(k_1k_2)} \times_{k_1} D_{k_1}^{1/2}U_{(k_1k_2)}^{(k_1)} \times_{k_2} D_{k_2}^{1/2}U_{(k_1k_2)}^{(k_2)},\ \Delta \times_{k_2} 1'_{m_{k_2}} = 2\Lambda^{(k_1k_3)} \times_{k_1} D_{k_1}^{1/2}U_{(k_1k_3)}^{(k_1)} \times_{k_3} D_{k_3}^{1/2}U_{(k_1k_3)}^{(k_3)},\ \Delta \times_{k_1} 1'_{m_{k_1}} = 2\Lambda^{(k_2k_3)} \times_{k_2} D_{k_2}^{1/2}U_{(k_2k_3)}^{(k_2)} \times_{k_3} D_{k_3}^{1/2}U_{(k_2k_3)}^{(k_3)}$$

をみたす3次元テンソルである

【周辺分布と2変数ごとの相関係数に相当する値が与えられた場合の数値例】

3個の名義変数の周辺分布ベクトル: $p_1 = (0.45,0.35,0.15,0.05)', p_2 = (0.3,0.2,0.1,0.25,0.15)', p_3 = (0.45,0.3,0.25)'$

2変数間の相関係数に相当する値: $r_{12} = \lambda_1^{(12)} = 0.6, r_{13} = \lambda_1^{(13)} = 0.5, r_{23} = \lambda_1^{(23)} = 0.32$

が与えられているとき、確率テンソル

$$P_{[,,1]} = \begin{bmatrix} 8.7e-02 & 5.8e-02 & 2.9e-02 & 3.6e-02 & 2.2e-02 \\ 6.8e-02 & 4.5e-02 & 2.3e-02 & 2.8e-02 & 1.7e-02 \\ 5.4e-04 & 3.6e-04 & 1.8e-04 & 1.7e-02 & 1.0e-02 \\ 1.8e-04 & 1.2e-04 & 5.9e-05 & 5.7e-03 & 3.4e-03 \end{bmatrix}$$

$$P_{[,,2]} = \begin{bmatrix} 5.8e-02 & 3.9e-02 & 1.9e-02 & 2.4e-02 & 1.4e-02 \\ 4.5e-02 & 3.0e-02 & 1.5e-02 & 1.9e-02 & 1.1e-02 \\ 3.6e-04 & 2.4e-04 & 1.2e-04 & 1.1e-02 & 6.9e-02 \\ 1.2e-04 & 7.9e-05 & 4.0e-05 & 3.8e-03 & 2.3e-03 \end{bmatrix}$$

$$P_{[,,3]} = \begin{bmatrix} 2.3e-02 & 1.5e-02 & 7.7e-03 & 1.1e-02 & 6.6e-03 \\ 1.8e-02 & 1.2e-02 & 6.0e-03 & 8.6e-03 & 5.1e-03 \\ 1.7e-05 & 1.1e-05 & 5.7e-06 & 6.4e-02 & 3.8e-02 \\ 5.7e-06 & 3.8e-06 & 1.9e-06 & 2.1e-02 & 1.3e-02 \end{bmatrix}$$

は上記の条件をみたすものの1つである。$P_{[,,i]}$ は変数3のi番目のカテゴリー値における変数1×変数2の行列である。