集約的シンボリックデータの非類似度の利用

統計数理研究所　清水 信夫
統計数理研究所　中野 純司
徳島文理大学　　山本 由和

1　はじめに

大量の多変量データが自然なグループに分かれている場合、個体データそのものではなく、グループに関しての推論に興味がある場合が考えられる。このとき、そのようなグループを表すためのいくつかの記述統計量の集合をデータと考えたものを集約的シンボリックデータ (Aggregated Symbolic Data, ASD) と呼ぶ。その例として、連続変数値データの場合は、各グループをその平均および分散共分散行列を用いて表すことが考えられる。ただ実際のデータにおいては連続変数だけでなくカテゴリー変数も含まれている場合が多数ある。このような状況において、連続変数・カテゴリー変数いずれに対しても同じ規準で非類似度を考えるために、それぞれの 2 次のモーメントまでを考えた上で、連続変数をカテゴリー変数に変換してすべての変数をカテゴリー変数と考え、尤度比検定統計量の和として非類似度を構成することを提案する。そして、その尤度比検定統計量をさらに変数成分ごとに分解し、非類似度に大きな影響を与えている変数や、各変数における値に関する情報の関係性について考える。

2　集約的シンボリックデータ間の非類似度

データ集合全体を表す行列を X とし、その中におけるグループ g の個々のデータを表す行列 $X^{(g)}$ を

$$X^{(g)} = \begin{bmatrix} x_{11}^{(g)} & \cdots & x_{1p}^{(g)} & x_{11}^{(g,1)} & \cdots & x_{1m_1}^{(g,1)} & \cdots & x_{11}^{(g,q)} & \cdots & x_{1m_q}^{(g,q)} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ x_{n^{(g)}1}^{(g)} & \cdots & x_{n^{(g)}p}^{(g)} & x_{n^{(g)}1}^{(g,1)} & \cdots & x_{n^{(g)}m_1}^{(g,1)} & \cdots & x_{n^{(g)}1}^{(g,q)} & \cdots & x_{n^{(g)}m_q}^{(g,q)} \end{bmatrix} = [X_1^{(g)} X_2^{(g)}]$$

とする。ここで $X_1^{(g)}$ は p 個の連続変数に関する部分、$X_2^{(g)}$ は q 個のカテゴリー変数のダミー変数に関する部分であり、$n^{(g)}$ は g におけるデータの総数である。グループ g 内における変数同士の 2 次モーメントは

$$X^{(g)'}X^{(g)} = [X_1^{(g)} X_2^{(g)}]'[X_1^{(g)} X_2^{(g)}] = \begin{bmatrix} X_1^{(g)'}X_1^{(g)} & X_1^{(g)'}X_2^{(g)} \\ X_2^{(g)'}X_1^{(g)} & X_2^{(g)'}X_2^{(g)} \end{bmatrix} \equiv \begin{bmatrix} S_{11}^{(g)} & S_{12}^{(g)} \\ S_{21}^{(g)} & S_{22}^{(g)} \end{bmatrix}$$

と表せる。$S_{11}^{(g)}$,$S_{22}^{(g)}$,$S_{21}^{(g)}$ はそれぞれ連続変数間、カテゴリー変数間、連続変数とカテゴリー変数間の 2 次モーメントである。

異なるグループ g_1,g_2 が共通の連続変数およびカテゴリー変数をもつ場合の g_1 と g_2 の間の非類似度を考えるために、連続変数をカテゴリー変数に変換する。そのためには連続変数値が 1 個または 0 個含まれるような小さな等間隔の区間を考え、それをカテゴリー値とする。その場合の確率は連続変数の分布を正規分布と考えたときの確率密度関数で近似する。カテゴリー変数の確率はカテゴリー値ごとにパラメータとして設定する。このようにすると非類似度は、両グループが同じ多項分布に従うという帰無仮説をそれぞれ別の多項分布に従うという対立仮説に対して検定した尤度比検定統計量として考えられる。また尤度比検定統計量は変数の組み合わせごとの値にさらに分解できる。これにより、非類似度に大きな影響を与えている変数や、非類似度における各変数における値と別な変数における値との関係性についても考えることができる。詳細および適用例は当日に示す。