Selective Inference を用いた不均一データ分析のための統計的推論

名古屋工業大学・工学研究科／理化学研究所・革新知能統合研究センター　竹内 一郎
名古屋工業大学・工学研究科　井上 茂乗
名古屋工業大学・工学研究科　梅津 佑太
名古屋大学・医学研究科　坪田 庄真

はじめに　生物医科学分野では背後にサブグループを持つ不均一なデータを分析することがある．例えば精密医療（precision medicine）では，これまで単一のものと考えられてきた病態から複数のサブタイプを同定する試みがなされている．また，細胞解析（single cell analysis）では細胞集団からいくつかのサブグループを同定したうえで各サブグループの遺伝的特徴が分析される．このような不均一データ分析では，まずクラスタリングなどによってサブグループを同定し，続いて各サブグループの特徴を分析するという二段階のアプローチが採用される．このような二段階アプローチでは，第一段階でデータに基づいてサブグループが同定されたことを適切に考慮し，第二段階の統計分析結果を補正する必要がある．本講演では，近年注目を集めている Selective Inference の考え方を用いることで，クラスタリングによって同定されたサブグループの統計的推論を適切に行えることを示す．

問題設定　本講演では不均一データ分析の具体例として，K 平均クラスタリングによってサブグループを同定した後，各サブグループに特異的な変数を選択する問題を考察する．観測データ $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n \times d}$ が与えられているとする．ただし，n は事例数，d は特徴数である．K 平均クラスタリングでは，K 個のクラスタ中心 $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_K$ を初期化し，K 個のクラスタの構成メンバ $\mathcal{C}_1, \ldots, \mathcal{C}_K$ を交互に更新することにより K 個のサブグループを同定する．続いて，K 平均クラスタリングによって同定されたサブグループにおいて，特徴的な変数の統計的推論を行う．同定された相異なる 2 つのサブグループ $\mathcal{C}_a, \mathcal{C}_b$ に対して，以下のような統計的仮説検定問題を考える

$$H_{0,j}^{(a,b)} : \mu_{a,j} = \mu_{b,j} \text{ vs. } H_{1,j}^{(a,b)} : \mu_{a,j} \neq \mu_{b,j}. \tag{1}$$

ただし，$\mu_{a,j}, \mu_{b,j}$ はクラスタ a, b それぞれの興味ある変数 j の母平均とする．この統計的推測問題に対する検定統計量として $\tau_{(a,b),j} = |m_{a,j}^{(T)} - m_{b,j}^{(T)}|$ を考える．これは，T 回のステップ後のクラスタリング結果における特徴 j のクラスタ中心の差を表している．

Selective Inference によるクラスタリングバイアス補正　クラスタリングによって同定された 2 群間の検定を行う場合，クラスタの選択がデータを用いて行われているため，選択バイアスの問題が生じてしまう．選択バイアスを考慮しない場合，誤検出率を正しく制御できず，統計的仮説検定のとしての妥当性を失ってしまう．この問題を回避するため，Selective Inference[1] を導入する．Selective Inference では，クラスタが K 平均クラスタリングによって選択された条件のもとでの検定統計量の条件付分布を考える．クラスタリングの各ステップで行った手順をイベント $\mathcal{E}$ と標記すると，Selective Inference では，以下のような信頼区間 $[\ell, u]$ を構成する：

$$\mathbb{P}_{H_{0,j}^{(a,b)}} \left(\tau_j^{(a,b)} \notin [\ell, u] \mid \hat{\mathcal{E}} = \mathcal{E} \right) < 0.05. \tag{2}$$

(2) 式の条件付信頼区間の計算は一般には難しいが，選択イベント $\mathcal{E}$ が特定の形式を持つ場合，正確に求めることができる．紙面の都合上詳細は省略するが，K 平均クラスタリングの選択イベント $\mathcal{E}$ は，データ行列 X の二次不等式の集合として表すことができ，これを利用すると (2) 式の条件付信頼区間の計算が可能となる．講演においては，本提案手法を一細胞解析データに適用した結果も示す．

参考文献　[1] Lee JD., Sun DL., Sun Y., Taylor J. Exact post-selection inference with the LASSO. Annals of Statistics 2016.