Selective Inference under the Local Alternative

梅津 佑太 [1]　　竹内 一郎 [1,2,3]

[1] 名古屋工業大学 大学院工学研究科
[2] 理化学研究所 革新知能統合研究センター
[3] 物質・材料研究機構 情報統合型物質・材料研究拠点

1　はじめに

探索的なデータ解析では, あらかじめ検証すべき仮説が定まっていることは少ない. そのため, 通常は, データを独立な二つのデータセットに分割して, 探索と仮説をそれぞれのデータセットで段階的に行われる. あるいは, 遺伝子データ解析など, データの分割が困難である場合には, false discovery rate (FDR) のように, 選択された仮説を考慮した過誤を利用することで, データを分割することなく統計的な信頼性を評価する.

選択された仮説に基づく, 統計推論は post-selection inference として知られている. 近年, この問題に対するの新たなフレームワークとして selective inference と呼ばれる手法が提案された [1]. selective inference では, FDR のように探索的に仮説が選択されたことを考慮するものである. 具体的には, ある仮説が選択されたということを確率的な事象ととらえ, その事象を条件付けたもとで, type I error に類似する過誤をコントロールする. このような過誤を利用して, 選択的な意味での type I error をコントロールすることそのものは探索的データ解析では有用ではあるものの, 一方で, 検出力についての議論はそれほど多くない. 本研究では, selective inference において, local alternative のもとでの検出力について議論する.

2　Selective Inference

選択写像とは, データ $\boldsymbol{y} = (y_1, \ldots, y_n)$ から仮説 $\mathcal{H} = \{h_1, h_2, \ldots, \}$ を対応付ける写像 $h = h(\boldsymbol{y}) \in \mathcal{H}$ であるとする. このとき, 選択写像の逆像 $\mathcal{H}^{-1}(h) = \{\boldsymbol{y} \mid h(\boldsymbol{y}) \in \mathcal{H}\}$ は selection event とよばれる. 仮説 h が選ばれたとき, 興味のあるパラメータ θ_h に関する検定問題

$$\mathrm{H}_0 : \theta_h = 0 \quad \text{vs.} \quad \mathrm{H}_1 : \theta_h = \frac{\delta}{\sqrt{n}} \quad (\delta > 0)$$

を考える. θ_h の推定量を $\hat{\theta}_h = \hat{\theta}_h(\boldsymbol{y})$ とし, その実現値を $\theta_{h,0}$ とすれば, 適当な条件のもとで

$$p_h = \mathrm{P}_{\mathrm{H}_0}(\hat{\theta}_h \geq \theta_{h,0} \mid \boldsymbol{y} \in \mathcal{H}^{-1}(h)) \sim \mathrm{Unif}(0, 1)$$

となる. p_h は selective p-value とよばれ, 適当な有意水準 α に対して, $p_h \leq \alpha$ ならば帰無仮説を棄却する. 検定統計量 $\hat{\theta}_h$ に対する棄却点を z_α とする. このとき, 適当な条件のもと, 検定統計量 $\hat{\theta}_h$ と独立な L, U が存在して, 対立仮説のもとで以下が成り立つ:

$$\begin{aligned}\mathrm{P}_{\mathrm{H}_1}(\hat{\theta}_h \geq z_\alpha \mid \boldsymbol{y} \in \mathcal{H}^{-1}(h)) &= \alpha + \frac{\kappa}{v} n^{-1/2} + \frac{\kappa}{v^2} \frac{\phi(U/v) - \phi(L/v)}{\Phi(U/v) - \Phi(L/v)} n^{-1} + O(n^{-3/2}) \\ &\geq \frac{3}{4}\alpha + \frac{1}{4} \frac{\phi(z_\alpha/v) - \phi(U/v)}{\phi(L/v) - \phi(U/v)} + O(n^{-3/2}).\end{aligned}$$

ただし, Φ および ϕ は標準正規分布の累積分布関数および確率密度関数であり

$$\kappa = \frac{\phi(z_\alpha/v) - \{\phi(U/v) - \alpha(\phi(U/v) - \phi(L/v))\}}{\Phi(U/v) - \Phi(L/v)} \leq 0$$

である. したがって, 対立仮説のもとでは, selective inference は近似的に 1 次の不偏検定になっていることがわかる. また, 3 次のオーダーまで評価すると, 不偏ではないものの, 検出力の下界を評価できる.

References

[1] Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso, *The Annals of Statistics*, **44**, 907–927.