

マルチスケール・ブートストラップによる モデル選択後の selective inference

大阪大学 大学院基礎工学研究科, 理化学研究所 革新知能統合研究センター 寺田 吉孝
京都大学 大学院情報学研究科, 理化学研究所 革新知能統合研究センター 下平 英寿

従来の統計理論では, モデルの選択とデータ解析を分けて考える前提で統計的推測の妥当性を保証している. しかし, 実データ解析の場面においては, 予めモデルを固定するのではなく, データに基づきモデル選択を行い, 選ばれたモデルから統計的推測を導くことが多い. このような状況では, 従来の統計的推測の妥当性は保証されない. 近年, データに基づくモデル選択の影響を適切に扱った統計的推測は, selective inference や post-selection inference と呼ばれ, 注目を集めている (Taylor and Tibshirani, 2015).

本稿では, 回帰分析の枠組みにおいて forward stepwise selection や lasso によって変数選択を行った際の回帰係数に対する selective inference を考える. 目的変数 $y \in \mathbb{R}^n$ は, 以下の多変量正規モデルから得られていると仮定する.

$$y \sim N_n(\mu, \sigma^2 I_n)$$

$X = (x_{ij})_{n \times p} \in \mathbb{R}^{n \times p}$ をそれぞれの列が説明変数に対応する non-random な行列として, y の X による回帰分析を考える. Berk et al. (2013) では, 特定のモデル選択法を仮定せずに, どのモデル選択法を用いても妥当性が保証される保守的な selective inference の枠組みを提案している. また, Lee et al. (2016) や Tibshirani et al. (2016) では, forward stepwise selection や lasso によるモデル選択によって特定のモデルが選ばれる事象 (selection event) が, y の空間において polyhedral set で表現されることを明らかにした. そして, selection event が polyhedral set で表現されるモデル選択法を用いた際に不偏な selective inference を行う方法を提案している.

一方で, Terada and Shimodaira (2017) では, マルチスケール・ブートストラップを用いて, selection event や仮説が一般の領域で表現される一般的な状況で近似的に不偏な selective inference を行う方法を提案し, 階層的クラスタリングの信頼性評価に応用している. 本発表では, Terada and Shimodaira (2017) の枠組みを用いて, モデル選択後の回帰係数に対する selective inference を行う方法を提案する. 一般に, マルチスケール・ブートストラップを用いる場合, ブートストラップのサンプルサイズを変化させることで, 確率分布のスケールを変化させてブートストラップ確率の変化率から幾何的量を推定している. しかし, 上述の回帰分析の設定のもとでは, selection event をデータのサンプルサイズ n に依存しない空間で表現することは困難であり, サンプルサイズに応じて selection event の形状が変化してしまう. したがって, 単純に目的変数と説明変数のデータペア (y_i, x_i) をリサンプリングし, ブートストラップのサンプルサイズを変化させるアプローチは適切ではない. そこで, 回帰分析における従来のブートストラップ法と同様に残差をリサンプリングするブートストラップを考え, 残差を定数倍することでスケールを変化させ, ブートストラップ確率の変化率を得る方法を提案する. 詳細な方法や数値実験の結果については当日報告する.

参考文献

- [1] Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid Post-Selection Inference. *Annals of Statistics*, **41**, 802–837.
- [2] Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. (2016). Exact Post-Selection Inference, With Application to the Lasso. *Annals of Statistics*, **44**, 907–927.
- [3] Taylor, J. and Tibshirani, R. J. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 7629–7634.
- [4] Terada, Y. and Shimodaira, H. (2017). Selective inference for the problem of regions via multiscale bootstrap. *arXiv:1711.00949*.
- [5] Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2016). Exact Post-Selection Inference for Sequential Regression Procedures. *Journal of the American Statistical Association*, **111**, 600–620.