Lasso の誤差分散の推定について

中央大学大学院理工学研究科　保科架風
中央大学理工学部　酒折文武

目的変数 y と p 個の説明変数 $x_1, \ldots, x_p$ について観測されたデータ $\{(y_i, \boldsymbol{x}_i\ ;\ i = 1, \ldots, n)\}$ に対する線形回帰モデル $\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ ($\boldsymbol{y} = (y_1, \ldots, y_n)^T$, $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^T$ $(i = 1, \ldots, n)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$, $\boldsymbol{\varepsilon} \sim N_p(\boldsymbol{0}, \sigma^2 I_p)$) において, lasso (Least Absolute Shrinkage and Selection Operator; Tibshirani, 1996)

$$\hat{\boldsymbol{\beta}} := \arg\min_{\boldsymbol{\beta}} \left[\frac{1}{n} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right]. \tag{1}$$

に代表されるスパース回帰モデリングの諸手法は, 近年のコンピューター技術や ICT 技術の発展に伴って大規模・複雑化されたデータから有益な情報を効率的に抽出する手法として注目を集めている. Lasso では, モデルの推定とモデルに含まれる変数の選択を同時に行うことが可能である. これにより, 一般的にデータの次元が高まるほど困難になる変数選択が, モデリングプロセス全体をコントロールする調整パラメータ λ (> 0) の値の決定問題となる.

この問題に対し, モデル選択基準を適用することが一般的であり, モデルの有効自由度を考慮したモデル選択基準の有効性が知られている (e.g., Zou *et al.*, 2007; Hirose *et al.*, 2013). ただし, lasso では推定に $\beta_j = 0$ $(j = 1, \ldots, p)$ で微分不可能な正則化項を含む目的関数の最小化が必要となり, $\hat{\boldsymbol{\beta}}$ が解析的に陽な形で与えらない. これにより, モデル選択基準においてモデルの複雑さを評価する “モデルの有効自由度” (DF) を求めることが困難となる. これに対し, Zou *et al.* (2007) は $\hat{\boldsymbol{\beta}}$ の非ゼロ成分の数がモデルの有効自由度の不偏推定量であることを示し, また, Hirose *et al.* (2013) では lasso を含むいくつかの Sparse 推定のモデルの有効自由度を数値的に求める方法を提案している.

しかし, 多くのモデル選択基準では値を求めるのに誤差分散の真の値, あるいはその推定値が必要となる. 一般には誤差分散の値は未知であり, また, lasso を含むスパース回帰モデリングの諸手法では局外母数である誤差分散の推定量については言及していないものが多い. このため, スパース回帰モデリングに関する多くの研究では誤差分散の最尤推定値 $\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2/(n - p - 1)$ やモデルの有効自由度を考慮した $\|\boldsymbol{y} - X\hat{\boldsymbol{\beta}}\|^2/(n - \mathrm{DF} - 1)$ などを用いるなど, 各研究によって採用する誤差分散の推定値が異なる.

これに対し, Reid *et al.* (2014) はいくつかの誤差分散の推定量の性質を数値的に比較した. しかし, 誤差分散の推定量として何を選択すべきなのかという問題の完全な解決には至っていない. そこで本研究では, モデル選択の点において適切な誤差分散の推定量が何か検証を行い, 推定量の選択について考察する.

参考文献

[1] Hirose, K., Tateishi, S. and Konishi, S. (2013). Tuning parameter selection in sparse regression modeling, *Computational Statistics and Data Analysis*, 59, 28–40.

[2] Reid, S., Tibshirani, R., and Friedman, J. (2014). A study of error variance estimation in lasso regression. *arXiv*, 1311.5274v2.

[3] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society*, Ser. B, 58, 267–288.

[4] Zou, H., Hastie, T. and Tibshirani, R. (2007). On the degrees of freedom of the lasso. *The Annals of Statistics*, 35, 2173–2192.