

Clustered Lasso と OSCAR に対する path algorithms

株式会社 ACCESS 高橋宏典, 統計数理研究所 野村俊一

本研究では, スパース正則化技法として知られる clustered Lasso と OSCAR に対する効率的な path algorithm を提案する. 高次元データの回帰分析において効率的に変数選択を行う方法として Lasso と呼ばれるスパース正則化技法が知られている. Lasso では, 回帰係数に対して L1 正則化を課すことにより, 効果の低い変数の回帰係数を厳密に零で推定して変数選択から外すことで, パラメータの自由度を削減することを狙いとしている. 変数選択以外でもパラメータの自由度を減らす方策として, fused Lasso では隣接関係にある変数について, その回帰係数の差に対する L1 正則化を導入することにより, 隣接した回帰係数の一部を等しい値へと統合することができる. 特に, She (2010) により提案された, 全ての回帰係数のペアを隣接関係に含めた fused Lasso は clustered Lasso と呼ばれ, 次式のように定式化される:

$$\min_{\beta} \frac{1}{2}(y - X\beta)^2 + \lambda_1 \sum_{i=1}^d |\beta_i| + \lambda_2 \sum_{1 \leq j < k \leq d} |\beta_j - \beta_k|.$$

ここで, $y \in \mathbb{R}^n$ は目的変数, $X \in \mathbb{R}^{n \times d}$ はデザイン行列, $\beta \in \mathbb{R}^d$ は回帰係数であり, λ_1 と λ_2 はそれぞれ回帰係数をスパースにする正則化項と回帰係数の値を揃える正則化項に対する重み付けを行う正則化パラメータである. さらに, Bondell and Reich (2008) は回帰係数の符号を区別せずにその絶対値を等しい値に揃えることを目的とした, 次の octagonal shrinkage and clustering algorithm for regression (OSCAR) を提案している:

$$\min_{\beta} \frac{1}{2}(y - X\beta)^2 + \lambda_1 \sum_{i=1}^d |\beta_i| + \lambda_2 \sum_{1 \leq j < k \leq d} \max\{|\beta_j|, |\beta_k|\}.$$

ここで, 上式の最後の項が回帰係数の絶対値を揃える役割を果たすことになる.

一般に, 正則化パラメータのチューニングにおいて, 正則化パラメータを変化させたときの解の軌跡を求める path algorithm と呼ばれる手法が用いられることがある. しかし, OSCAR に対する厳密な path algorithm は先行研究ではなく, また clustered Lasso でも正則化項の数が回帰係数の次元の二乗のオーダー $\mathcal{O}(d^2)$ となるために, 既存の path algorithm では非常に計算コストがかかってしまう. そこで, 本研究では, 正則化項の対称性を利用した clustered Lasso と OSCAR に対する効率的な path algorithm を提案する. 提案手法では, 正則化パラメータを $(\lambda_1, \lambda_2) = \eta(\bar{\lambda}_1, \bar{\lambda}_2)$ とおいて η を変化させたときの解 $\beta(\eta)$ の完全な軌跡を特定できる. 提案手法の計算量は, η を変化させたときに回帰係数の同じ値のグループが統合・分裂等に変化する回数を T とおいたとき $\mathcal{O}(nd^2 + ndT)$ となり, 既存手法に比べて計算時間が大幅に削減されることを数値実験により示す.

参考文献

- [1] She, Y. (2010). Sparse regression with exact clustering. *Electronic Journal of Statistics*, **4**, 1055–1096.
- [2] Bondell, H.D. and Reich, B.J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, **64**, 115–123.