Best subset selection in multivariate linear regressions via discreate first-order algorithms

広島大・理　　鈴木 裕也　大石 峰暉　小田 凌也　柳原 宏和

回帰モデルを扱う際に, 最適な説明変数の組み合わせを決定することは重要である. その際, 全ての組み合わせを考える変数総当たり法を実行できるのであればそれに越したことはないが, 説明変数の数が多くなると計算量的に実行できなくなってしまう. その問題を解決できる方法として, Bertsimas *et al*. (2016) より discrete first-order algorithms を用いた重回帰モデルの変数選択法が提案されている. この方法は, 以下の最小化問題を解くことで変数選択を行うもので, 条件付最小化問題を効率的に解くためのアルゴリズムを提案している.

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 \quad s.t. \quad \|\boldsymbol{\beta}\|_0 \le q.$$

ここで, $\boldsymbol{y}$ は n 次元目的変数ベクトル, $\boldsymbol{X}$ は $n \times k$ 説明変数行列, $\boldsymbol{\beta}$ は k 次元回帰係数ベクトル, q は $q \in \{1,\dots,k\}$ である. この最小化問題は q を決めると選択される説明変数が制限されるので, 全ての q で条件付最小解 $\hat{\boldsymbol{\beta}}_q$ を求め, それぞれの $\hat{\boldsymbol{\beta}}_q$ に基づく情報量規準の最小化により, 最適な変数の組み合わせを求めることができる. この方法は説明変数の数の繰り返し計算を行うだけなので, 総当たり法では計算できない数の説明変数でも変数選択が可能となり計算速度も速い. そこで, 本研究ではこの手法を多変量線形回帰にも適応できるように, 以下のような最小化問題に拡張した.

$$\min_{\boldsymbol{\Theta}} g(\boldsymbol{\Theta}) \quad s.t. \quad \|\boldsymbol{\Theta}\|_0 \le q,$$
$$g(\boldsymbol{\Theta}) = \frac{1}{2n}\mathrm{tr}\{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\Theta})\boldsymbol{S}^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\Theta})'\}.$$

ここで, $\boldsymbol{Y}$ は $n \times p$ 目的変数行列, $\boldsymbol{\Theta}$ は $k \times p$ 回帰係数行列, $\boldsymbol{S}$ は標本共分散行列とする. また, $\boldsymbol{\theta}_j'$ を $\boldsymbol{\Theta}$ の第 j 行ベクトルとすると $\|\boldsymbol{\Theta}\|_0 = \sum_{j=1}^k I(\|\boldsymbol{\theta}_j\|_2 \ne 0)$ と定義する. ただし $I(A)$ は定義関数である. このとき, q を固定した下での最適な $\boldsymbol{\Theta}$ の推定アルゴリズムを以下に示す. ただし, $L = \lambda_{\max}(\boldsymbol{X}'\boldsymbol{X}/n)/\lambda_{\min}(\boldsymbol{S})$, $\boldsymbol{D}(\boldsymbol{\Theta}) = \partial g(\boldsymbol{\Theta})/\partial\boldsymbol{\Theta}$, $\boldsymbol{d}(\boldsymbol{\Theta})_j$ は $\boldsymbol{D}(\boldsymbol{\Theta})$ の第 j 列ベクトルとする.

1. $\boldsymbol{\Theta}$ の初期値 $\boldsymbol{\Theta}_0 = (\boldsymbol{\theta}_{0,1},\dots,\boldsymbol{\theta}_{0,k})'$ を決める.

2. m 回目の繰り返しにおいて, $\boldsymbol{\eta}_{m,j} = \boldsymbol{\theta}_{m-1,j} - \boldsymbol{d}(\boldsymbol{\Theta}_{m-1})_j/L$ $(j = 1,\dots,k)$ を計算する.

3. 各 $j = 1,\dots,k$ に対して $\|\boldsymbol{\eta}_{m,j}\|_2$ を大きい順に並び替え, q 番目に大きい値を c_m とすると $\boldsymbol{\Theta}_m$ を以下の式で更新する.

$$\boldsymbol{\theta}_{m,j} = I(\|\boldsymbol{\eta}_{m,j}\|_2 \ge c_m)\boldsymbol{\eta}_{m,j}$$

4. $g(\boldsymbol{\Theta}_m)$ が収束するまで 2, 3 を繰り返し, 収束した時の $\boldsymbol{\Theta}_m$ を 最適な $\boldsymbol{\Theta}$ とし, $\hat{\boldsymbol{\Theta}}$ とする.

上記のアルゴリズムを全ての q で行い, 各 q における $\hat{\boldsymbol{\Theta}}$ を $\hat{\boldsymbol{\Theta}}_q$ と書く. このとき, 情報量規準最小化により最適な q を選択する. その際, 上記のアルゴリズムの初期値は $\hat{\boldsymbol{\Theta}}_{q-1}$(ただし $\hat{\boldsymbol{\Theta}}_0 = \boldsymbol{0}_{k\times p}$) とする. 発表当日は, 提案手法と他の手法の数値比較の結果も紹介する.

引用文献

[1] Bertsimas, D., King, A. and Mazumder, R. (2016), Best subset selection via a modern optimization lens, *Ann. stat.*, **35**(6), 2313-2351.