Sparse Group Lasso を用いたGMANOVAモデルの変数選択

中京大学 国際教養学部 永井 勇
広島大学 大学院理学研究科 小田 凌也
広島大学 大学院理学研究科 栁原 宏和

各個体に対して経時的に測定して得られるデータ (経時測定データ) は, 様々な分野で収集され分析されている. 本講演では, 全ての個体における測定時点が揃っている経時測定データに着目する. この経時測定データにおける測定時点を $t_1, \ldots, t_p$ とし, i 番目の個体の時点 t_j での測定値を (i,j) 成分に持つ $n \times p$ 行列を $\boldsymbol{Y}$ とする. また, 各行が各個体の k 個の説明変数からなる $n \times k$ 個体間説明変数行列を $\boldsymbol{A}$, 後述するように t_j に関する変動を表す $p \times q$ 個体内説明変数行列を $\boldsymbol{X}$ とする. これらを用いると, $\boldsymbol{Y}$ の分析でよく用いられるモデルである一般化多変量分散分析 (GMANOVA) モデル (Potthoff & Roy, 1964) は次の形で定義される;

$$\boldsymbol{Y} = \mathbf{1}_n \boldsymbol{\mu}' \boldsymbol{X}' + \boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}' + \boldsymbol{\mathcal{E}},$$

ここで $E[\boldsymbol{\mathcal{E}}] = \boldsymbol{O}_{n,p}$ ($n \times p$ ゼロ行列), $\mathrm{Cov}[\mathrm{vec}(\boldsymbol{\mathcal{E}})] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n$, $\boldsymbol{\Sigma}$ は $\mathrm{rank}(\boldsymbol{\Sigma}) = p$ の $p \times p$ 未知分散共分散行列, $\mathbf{1}_n$ は全ての成分が 1 の n 次元ベクトル, $\boldsymbol{\mu}$ は q 次元未知ベクトル, $\boldsymbol{\Xi}$ は $k \times q$ 未知回帰係数行列である. また, 本講演では $\boldsymbol{A}$ は中心化されていると仮定する (つまり, $\boldsymbol{A}'\mathbf{1}_n = \mathbf{0}_k$, $\mathbf{0}_k$ は k 次元ゼロベクトル). このモデルにおいて, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q)$ として $\boldsymbol{x}_i = (t_1^{i-1}, \ldots, t_p^{i-1})'$ とすることは, 測定時点 $t_1, \ldots, t_p$ の $(q-1)$ 次多項式を用いて $\boldsymbol{Y}$ の経時変動を推定することに対応している.

この GMANOVA モデルにおいても, 従来の多変量線型回帰モデルと同様に, ある推定関数に Lasso 型の罰則を加えたものを最小にする推定により, $\boldsymbol{\Xi}$ の一部の行や列の推定結果をゼロベクトルに縮小できると考えられる. ここで GMANOVA モデルにおいては, $\boldsymbol{A} = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k)$, $\boldsymbol{\Xi} = (\boldsymbol{\xi}^{(1)}, \ldots, \boldsymbol{\xi}^{(k)})' = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_q)$ を用いると, $\boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}'$ の項は次のように二通りの表現で表される;

$$\boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{X}' = \sum_{i=1}^{k} \boldsymbol{a}_i \boldsymbol{\xi}^{(i)'} \boldsymbol{X}' = \sum_{j=1}^{q} \boldsymbol{A}\boldsymbol{\xi}_j \boldsymbol{x}_j'.$$

したがって, $\boldsymbol{\xi}^{(i)}$ を $\mathbf{0}_q$ に縮小して推定することは $\boldsymbol{a}_i$ が不要であることに対応しているので, $\boldsymbol{A}$ に対する変数選択ができる. 一方で, $\boldsymbol{\xi}_i$ を $\mathbf{0}_p$ に縮小して推定することは $\boldsymbol{x}_i$ が不要であることに対応しているので, $\boldsymbol{X}$ に対する変数選択, すなわち推定に必要な次数の多項式のみを選択することができる. つまり, GMANOVA モデルにおいては, 多変量線型回帰モデルとは異なり, $\boldsymbol{\Xi}$ の行や列のどちらをゼロベクトルに縮小して推定するのかにより, $\boldsymbol{A}$ と $\boldsymbol{X}$ のどちらの変数を選択するのかが異なるという特徴がある.

そこで本講演では, 残差平方和に Sparse Group Lasso (Simon, Friedman, Hastie & Tibshirani, 2013) 型の罰則を加えた形で, $\boldsymbol{A}$ と $\boldsymbol{X}$ に対する変数選択を行う手法を提案する. また, この罰則付残差平方和を最小にする推定量を得るためのアルゴリズムとして Coordinate Descent Algorithm (座標降下法; 例えば Wu & Lange (2008) など参照) を用いた手法を提案する.

詳細や数値実験による比較については当日報告する.

引用文献:

[1] Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. (2013). A sparse group lasso. *J. Comput. Graph. Stat.*, **22**, 231–245.

[2] Potthoff, R. F. & Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, **51**, 2–326.

[3] Wu, T. T. & Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *Ann. Appl. Stat.*, **2**, 224-244.