# Extended GMANOVA モデルにおける Sparse Group Lasso を用いた変数などの選択

中京大学 国際教養学部 永井 勇

各個体に対して経時的に $p$ 回測定して得られるデータは経時測定データと呼ばれ, 様々な分野で収集・分析がされている. 本講演では, $n$ 個全ての個体に対する測定時点が全て揃っている経時測定データに着目する. ここで, 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, $i$ 番目の個体において時点 $t_j$ で測定した値を $(i, j)$ 成分に持つ $n \times p$ 行列を $\boldsymbol{Y}$ とする. このようなデータ $\boldsymbol{Y}$ に対しては, 各個体の性別といった説明変数も用いて, 一般化多変量分散分析 (GMANOVA) モデル (Potthoff & Roy, 1964) を用いることで, 経時的な変動 (経時変動) が推定できる. しかしながら, このモデルでは, 平均的な経時変動や全ての説明変数に対する経時変動に対し, 測定時点に基づく共通の次数の多項式などを用いて推定をしている (例えば, Nagai (2011) 参照).

そこで永井 (2017) では, 平均的な経時変動や各説明変数グループに対する経時変動に対して, それぞれ異なる次数の多項式などを用いて推定できるように, 次の形で定義される Kollo and von Rosen (2005) の Definition 4.1.3 に基づいた Extended GMANOVA モデルを提案した.

$$\boldsymbol{Y} = \boldsymbol{1}_n \boldsymbol{m}' \boldsymbol{X}_0' + \boldsymbol{A}_1 \boldsymbol{\Xi}_1 \boldsymbol{X}_1' + \boldsymbol{A}_2 \boldsymbol{\Xi}_2 \boldsymbol{X}_2' + \boldsymbol{\mathcal{E}},$$

ここで $\boldsymbol{1}_n$ は $n$ 次元 1 ベクトル, $\boldsymbol{m}$ は $q_0$ 次元未知ベクトル, $\boldsymbol{X}_i = (\boldsymbol{x}_{i\cdot 1}, \ldots, \boldsymbol{x}_{i\cdot q_i})$ は後述するように測定時点 $t_1, \ldots, t_p$ から構築される $p \times q_i$ 既知行列 (ただし $\mathrm{rank}(\boldsymbol{X}_i) = q_i$; $i = 0, 1, 2$), $\boldsymbol{A}_i$ は説明変数からなる $n \times k_i$ 既知行列 (ただし $\mathrm{rank}(\boldsymbol{A}_i) = k_i$, $\boldsymbol{A}_i' \boldsymbol{1}_n = \boldsymbol{0}_{k_i}$, $\boldsymbol{0}_r$ は $r$ 次元 0 ベクトル; $i = 1, 2$), $\boldsymbol{\Xi}_i$ は $k_i \times q_i$ 未知行列 $(i = 1, 2)$, $\boldsymbol{\mathcal{E}}$ は $E[\boldsymbol{\mathcal{E}}] = \boldsymbol{0}_n \boldsymbol{0}_p'$, $\mathrm{Cov}[\mathrm{vec}(\boldsymbol{\mathcal{E}})] = \boldsymbol{\Sigma} \otimes \boldsymbol{I}_n$ の $n \times p$ 誤差行列であり, $\boldsymbol{\Sigma}$ は分散共分散を表す $p \times p$ 未知行列 (ただし $\mathrm{rank}(\boldsymbol{\Sigma}) = p$) である. ここで, $\boldsymbol{X}_i$ の $j$ 行目を $(1, t_j, \ldots, t_j^{q_i - 1})$ とすると, 平均的な経時変動や各説明変数グループ $\boldsymbol{A}_i$ ごとの経時変動に対し, 測定時点の $(q_i - 1)$ 次多項式を用いた推定ができる.

このモデルにおいて, $\boldsymbol{A}_i = (\boldsymbol{a}_{i\cdot 1}, \ldots, \boldsymbol{a}_{i\cdot k_i})$, $\boldsymbol{\Xi}_i = (\boldsymbol{\xi}_i^{(1)}, \ldots, \boldsymbol{\xi}_i^{(k_i)})' = (\boldsymbol{\xi}_i^{[1]}, \ldots, \boldsymbol{\xi}_i^{[q_i]})$ とすると, GMANOVA モデルと同様に, $\boldsymbol{A}_1 \boldsymbol{\Xi}_1 \boldsymbol{X}_1'$ および $\boldsymbol{A}_2 \boldsymbol{\Xi}_2 \boldsymbol{X}_2'$ は次の二通りの表現ができる;

$$\boldsymbol{A}_i \boldsymbol{\Xi}_i \boldsymbol{X}_i' = \sum_{j=1}^{k_i} \boldsymbol{a}_{i\cdot j} \boldsymbol{\xi}_i^{(j)'} \boldsymbol{X}_i' = \sum_{\ell=1}^{q_i} \boldsymbol{A}_i \boldsymbol{\xi}_i^{[\ell]} \boldsymbol{x}_{i\cdot \ell}'.$$

この表現より, $\boldsymbol{\xi}_i^{(j)}$ を $\boldsymbol{0}_{q_i}$ と推定すること と $\boldsymbol{A}_i$ の中で $j$ 番目の説明変数が不要であること が対応し, $\boldsymbol{\xi}_i^{[\ell]}$ を $\boldsymbol{0}_{k_i}$ と推定すること と $\boldsymbol{X}_i$ の $\ell$ 列が不要であること が対応していることが分かる. つまり, 各 $\boldsymbol{\Xi}_i$ の列や行をゼロベクトルへ縮小推定することにより, 各 $\boldsymbol{A}_i$ に対しての変数の選択ができ, 各 $\boldsymbol{A}_i$ に対する経時変動の推定の際に用いる多項式の次数などが選択できる. 本講演では, これらの変数や次数などの選択のために, 永井ら (2018) と同様に, Sparse Group Lasso (Simon, Friedman, Hastie & Tibshirani, 2013) 型の罰則を用いた推定手法を提案する. また, この罰則を付加した残差平方和を最小にする $\boldsymbol{\Xi}_i$ の推定量を得るために, 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の講演で報告する予定である.

**引用文献**:


[1] Kollo, T. & von Rosen, R. (2005). *Advanced Multivariate Statistics and Matrices*, Springer.

[2] Nagai, I. (2011). Modified $C_p$ criterion for optimizing ridge and smooth parameters in the MGR estimator for the nonparametric GMANOVA model. *Open J. Stat.*, **1**, 1–14.

[3] 永井 (2017). Extended GMANOVA モデルにおける罰則付推定量の提案と罰則パラメータの最適化, 2017 年度統計関連学会連合大会 予稿集, 231 ページ.

[4] 永井・小田・柳原 (2018). Sparse Group Lasso を用いた GMANOVA モデルの変数選択. 2018 年度統計関連学会連合大会 予稿集, 294 ページ.

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

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

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