多変量回帰分析や判別分析などにおける新たな変数選択法の提案

公立諏訪東京理科大・共通・マネジメント教育センター　櫻井　哲朗
広島大・理・名誉教授　藤越　康祝

本報告では，多変量回帰分析や判別分析などの多変量分析での変数選択問題について取り扱う．このような問題に対して, 仮説検定やモデル選択規準量などから最適なモデルを与えることができる．仮説検定では尤度比検定統計量が, モデル選択規準量としては AIC や C_p などがよく用いられている．モデル選択規準の最近の研究として, 標本数と次元数がともに大きくなる高次元漸近枠組のもとで真のモデルを選択する確率が 1 となる性質, すなわち, 規準量の一致性, が知られている．しかし, AIC や C_p から最適なモデルを求めるためには全ての変数の組み合わせを計算する必要があり変数が多くなりすぎると計算が困難となる．小田・柳原 (2017), 櫻井・藤越 (2017) では, この困難を克服する方法として, AIC や C_p の差に基づく変数選択法について考察している．本報告では，このような変数選択法の考えを拡張したり, 判別法などに適用することを考える．

ここでは, 次のような方法により最適なモデルを決める．いま k 個の変数があり, この変数の選択により最適なモデルを構築することを考える．このとき, 各変数に対して次の規則を用いて最適なモデルを決める．i 番目の変数について,

$$\begin{cases} \mathrm{MV}_{(-i)} - \mathrm{MV}_{\mathrm{Full}} > d \quad \Rightarrow \quad i \text{ 番目の変数を取り入れる} \\ \mathrm{MV}_{(-i)} - \mathrm{MV}_{\mathrm{Full}} \leq d \quad \Rightarrow \quad i \text{ 番目の変数を取り除く} \end{cases} \qquad i = 1, \ldots, k$$

ここで, $\mathrm{MV}_{\mathrm{Full}}$ は全ての変数を用いたモデルから計算される値であり, $\mathrm{MV}_{(-i)}$ は全ての変数から i 番目の変数を除いたモデルから計算される値である．このモデル間の差 $\mathrm{MV}_{(-i)} - \mathrm{MV}_{\mathrm{Full}}$ は, 一つの見方として i 番目の変数の影響度として捉えることができる．また, d は定数で MV の形などによって決まる値である．例えば, 多変量回帰分析における AIC や C_p の差に基づく基準では, $d = 0$ として最適なモデルを構築する．

本報告では，まず, 二群の判別分析において, 冗長性モデルに基づく変数選択問題を考える．このとき, AIC を用いたモデル間の差によるモデル選択は

$$\mathrm{AIC}_{(-i)} - \mathrm{AIC}_{\mathrm{Full}} = N \log\left\{1 + \frac{\frac{n_1 n_2}{n}(D^2 - D^2_{(-i)})}{n - 2 + \frac{n_1 n_2}{n} D^2_{(-i)}}\right\} - 2, \qquad d = 0$$

として与えられる．ここに, $D^2 = (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)' \mathbf{S}^{-1} (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2)$ であり, $D^2_{(-i)}$ は i 番目の変数を除いた場合での標本マハラノビス距離である．また, AIC を一般化した規準量でのモデル間の差によるモデル選択や, 通常はモデル選択に使用されない標本マハラノビス距離に基づくモデル間の差によるモデル選択も考え, これらの一致性について考察する．さらに, 多変量回帰分析において共分散行列に仮定をおいたもとでのモデル間の差によるモデル選択についても議論する．これらの結果に対して, 数値シミュレーションによって妥当性を検証する．

参考文献

1. 小田凌也, 柳原宏和. (2017). 多変量線形回帰モデルにおいて目的変数と説明変数が高次元の場合でも一致性を持つ高速な変数選択法．日本統計学会関連学会報告集.
2. 櫻井哲朗, 藤越康祝. (2017). 共分散構造をもつ多変量線形回帰モデルにおける C_p 型の変数選択規準の高次元一致性．日本統計学会関連学会報告集.