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

広島大・理　　小田 凌也　　栁原 宏和

正規性を仮定した多変量線形回帰モデルにおいて, 有効な説明変数を選ぶ変数選択問題を扱う. そのような変数の選択法に用いられる手法として, 従来の手法である AIC や C_p 規準などの変数選択規準をすべての説明変数の組において計算し変数選択規準が最小となる変数の組み合わせを最適な変数とする変数選択規準総当たり法や, 最近発展してきたスパース推定による方法が挙げられる.

近年, 目的変数ベクトルの次元数 p や説明変数ベクトルの次元数 k が大きなデータ, いわゆる, 高次元データが解析対象となる場合が多くなっている. 特に, p と k は大きいと言えども, 標本数 n よりは小さいようなデータ, moderately high-dimensional data を考える. しかしながら, そのような高次元データにおいて従来の変数選択法では以下のような 2 つの問題点がある:

(1) k が大きなデータに対して変数選択規準総当たり法では物理的に計算が不可能である.

(2) k が大きな場合に適した漸近理論の下で, 真の変数の組み合わせが最適な変数として選ばれる確率 (選択確率) が漸近的に 1 となる性質, 即ち, 一致性の議論はされていない.

以上の問題点からこれまで, 高次元データに対して一致性を持つ高速な変数選択法は存在しなかった. そこで, 本発表では, 問題点 (1), (2) を同時に解決する変数選択法を提案する. まず, 問題点 (1) を解決するために, 重み付き残差平方和にモデルのパラメータ数の定数倍 α (>0) を加えた一般化 C_p (Generalized C_p; GC_p) 規準を用い, Zhao *et al.* (1986) で提案された以下のような変数選択アルゴリズムを適用することを考える.

Step 1. 各説明変数に対して, その説明変数のみを取り除いたモデルとすべての説明変数を用いたモデルであるフルモデルの GC_p 規準の差を計算する.

Step 2. Step 1 で求めた差が正になれば取り除いた説明変数を有効とし, 差が負となれば取り除いた説明変数を有効でないとする.

Step 3. Step 2 で有効と判断された説明変数の組を最適な変数の組み合わせとする.

次に, 問題点 (2) を解決するために, n のみを ∞ とする大標本漸近理論と n だけでなく p や k も ∞ とする高次元大標本漸近理論を統一的に扱う漸近理論: $n \to \infty,\ (p+k)/n \to c \in [0,1)$ を用いて, 上記の変数選択アルゴリズムの下で GC_p 規準が一致性を持つための α に関する条件を導出する. 実際の条件は以下のとおりである.

$$\alpha = \frac{n-k}{n-p-k+1} + \beta,\ \beta > 0\ s.t.\ \lim_{n\to\infty,\ (p+k)/n\to c} \frac{\sqrt{p}}{\sqrt[2r]{k}}\beta = \infty,\ \lim_{n\to\infty,\ (p+k)/n\to c} \frac{p\sqrt[2r]{k}}{n}\beta = 0.$$

ただし, r は任意の 2 以上の自然数である. 上記の条件式を満たす α を用いることで, p や k の大小に関わらず一致性を持つ高速な変数選択法を提案できる. 発表当日は, 本発表で提案される変数選択法とグループ Lasso による変数選択法を計算時間, 選択確率の観点から数値的に比較する.

参考文献

[1] Zhao, L. C., Krishnaiah, P. R. & Bai, Z. D. (1986). On detection of the number of signals in presence of white noise. *J. Multivariate Anal.*, **20**, 1–25.