学位論文要旨

Consistent variable selection criteria in multivariate linear regression even when dimension exceeds sample size
(次元数が標本数を超えるときでも多変量線形回帰において一致性をもつ変数選択規準)

氏名　小田 凌也

本論文では, 目的変数が複数個ある場合の多変量線形回帰において, 有効な説明変数を選ぶための変数選択問題を扱う. 多変量線形回帰は複数の予測対象である変数（目的変数）と複数の目的変数に影響を与えると考えられる変数（説明変数）の関係を記述する多変量解析の手法の 1 つである. 実解析において, 目的変数に影響を与えている説明変数を特定することは重要であり, これは最適な説明変数の組み合わせを選択する変数選択問題として捉えられる. そのような変数選択問題において, 赤池情報量規準 (Akaike's Information Criterion; AIC) (Akaike, 1973, In *2nd International Symposium on Information Theory*; 1974, *IEEE Trans. on Automatic Control*) や C_p 規準 (Sparks *et al.*, 1983, *Comm. Statist. A-Theory Methods*) などの変数選択規準を用いる方法が挙げられる. さらに AIC, C_p 規準の拡張として, Nishii *et al.* (1988, *Hiroshima Math. J.*) により提案された 一般化情報量規準 (Generalized Information Criterion; GIC), Nagai *et al.* (2012, *Hiroshima Math. J.*) により提案された一般化 C_p (Generalized C_p; GC_p) 規準が挙げられる. GIC, GC_p 規準は, AIC, C_p 規準のそれぞれでモデルの複雑さを表す項 "2× (パラメータ数)" における 2 を任意の正数で置き換えることにより一般化された規準である.

変数選択では望ましい結果の 1 つとして, 目的変数に本当に影響を与える説明変数（真の説明変数）を特定することが挙げられる. これを漸近的に保証した性質が一致性であり, 一致性とは真の説明変数の組み合わせが最適な説明変数の組み合わせとして選ばれる確率 (選択確率) が漸近的に 1 となる性質のことをいう. そのため, 一致性は変数選択規準に望まれる性質の 1 つとして挙げられ, これまで多くの先行研究により標本数 n のみを無限大とする大標本 (LS) 漸近枠組みの下で様々な変数選択規準の一致性が調べられてきた.

一方, 近年では標本数 n のみでなく目的変数ベクトルの次元数 p も大きな高次元データを扱う需要が高まっている. しかし, そのような高次元データに対して LS 漸近枠組みを用いた一致性の評価は妥当でなく, 有限標本下における選択確率が低くなってしまう可能性がある. したがって, n のみでなく p も無限大とする高次元漸近理論を用いて一致性を評価する必要がある. Fujikoshi *et al.* (2014, *J. Multivariate Anal.*), Yanagihara *et al.* (2015, *Electron. J. Statist.*) では, 真の誤差ベクトルに正規性を仮定した下で, AIC や C_p 規準などの変数選択規準の一致性を主に以下の moderate-high-dimensional 漸近枠組みを用いて評価した:

$$(n, p) \to \infty,\ p/n \to c \in [0, 1),\ k\text{: fixed.} \tag{1}$$

ただし, k は全ての説明変数の個数を表す. 真の誤差ベクトルの正規性を緩めた非正規性の仮定の下では, Yanagihara (2015, *J. Japan Statist. Soc.*) により漸近枠組み (1) を用いて GIC が一致性をもつための条件が導出された. さらに, Yanagihara (2016, *Procedia Comput. Sci.*; 2019, *J. Jpn. Stat. Soc. Jpn. Issue*) は GIC もしくは GC_p 規準が一致性をもつための条件を導出し, その際, 漸近枠組み (1) の拡張である p は無限大でもそうでなくてもよい以下の hybrid-moderate-high-dimensional 漸近枠組みを用いた:

$$n \to \infty,\ p/n \to c \in [0, 1),\ k\text{: fixed.} \tag{2}$$

漸近枠組み (2) を用いて一致性を評価することで, p が n を超えない場合ではあるが, p が無限大の場合とそうでない場合の両方を統一的に扱うことを可能にしている. このような統一的な高次元漸近枠組みは他の多変量解析における統計量の高次元漸近性質の導出にも用いられており, 例えば, 正準相関分析では Yanagihara *et al.* (2017, *J. Multivariate Anal.*), Oda *et al.* (2019, *Random Matrices-Theo.*), Oda *et al.* (2019, *Sankhya A*), 正準判別分析では Oda *et al.* (2020, *J. Multivariate Anal.*), 多変量逆回帰では Oda *et al.* (2020, *Comm. Statist. Theory Methods*) が挙げられる.

一方, 高次元データの中でも p が n を超えるようなデータを扱う需要も増えている. しかし, $p > n$ のとき標本共分散行列が特異となるため, GIC の値が常に $-\infty$ となってしまい, また GC_p 規準は定義できない. そこで, $p > n$ のときでも計算可能な変数選択規準として, Prediction Error 規準 (Fujikoshi *et al.*, 2011, *Ann. I. Stat. Math.*), 標本共分散行列の逆行列を用いる代わりにリッジ型標本共分散行列の逆行列を用いた規準 (Yamamura *et al.*, 2010, *Japan. J. Appl. Stat.*; Kubokawa and Srivastava, 2012, *Comm. Statist. A-Theory Methods*) が提案された. さらに, Katayama and Imori (2014, *J. Multivariate Anal.*) は, 共分散行列の逆行列に対する lasso 型の推定量を用いた規準を提案し, 真の誤差ベクトルに正規性を仮定した下で提案規準が一致性をもつための条件を導出した. このとき, 全ての説明変数の組み合わせを候補の組み合わせとすると, p が n を超えて無限大としてよい高次元漸近枠組み: $(n, p) \to \infty,\ \log p/n \to 0,\ k/n \to 0$ を用いた. しかし, 真の誤差ベクトルの非正規性の下で p が n を超えて無限大となってもよい高次元漸近枠組みを用いたときに一致性をもつ変数選択規準はこれまで存在しなかった.

そこで本論文では, $p > n$ でも計算可能とするため, 特定の重み行列を伴う重み付き残差平方和に任意の正数 α で調整したモデルの複雑さを表す項 "$\alpha\times$ (パラメータ数)" を足し合わせた 2 つの変数選択規準を考える. 具体的には, 重み行列としてスカラー行列を用いて定義されるスカラー型一般化 C_p (Scalar-type GC_p; SGC_p) 規準, リッジパラメータ λ によるリッジ型標本共分散行列を用いて定義されるリッジ型一般化 C_p (Ridge-type GC_p; RGC_p) 規準を使用する. このとき, SGC_p 規準, RGC_p 規準が一致性をもつための条件を真の誤差ベクトルの非正規性の下で以下の hybrid-ultra-high-dimensional 漸近枠組みを用いて導出する:

$$n \to \infty,\ p/n \to c \in [0, \infty],\ k\text{: fixed.} \tag{3}$$

ただし, 漸近枠組み (3) 内の $c = \infty$ は $p/n \to \infty$ を表す. 漸近枠組み (3) は漸近枠組み (1), (2) を含んでおり, もし k が固定であれば Katayama and Imori (2014, *J. Multivariate Anal.*) で用いられた漸近枠組みも含んでいる. 実際に, 漸近枠組み (3) を用いたときに SGC_p 規準, RGC_p 規準が一致性をもつための条件は以下の通りである:

- SGC_p 規準が一致性をもつための α に関する条件:

$$\lim_{n\to\infty, p/n\to c} \alpha = \infty, \quad \lim_{n\to\infty, p/n\to c} \frac{\alpha}{n} = 0.$$

- RGC_p 規準が一致性をもつための λ, α に関する条件:

$$\lim_{n\to\infty, p/n\to c} \frac{p\alpha}{\lambda} = \infty, \quad \lim_{n\to\infty, p/n\to c} \frac{(1+\lambda^{-1})p\alpha}{n} = 0.$$

さらに本論文では, 真の共分散行列と誤差ベクトルに追加の仮定をした下で, p は必ず無限大とするよう漸近枠組み (3) を制限した高次元漸近枠組みを用いたときに SGC_p 規準, RGC_p 規準が一致性をもつための条件も導出する. 数値実験により一致性をもつための条件を満たす λ または α を用いた SGC_p 規準, RGC_p 規準は n のみが大きい場合だけでなく n と p が大きい場合でも選択確率が高いことを確認している.