Consistent Information Criterion for Normal Multivariate Linear Regression Model under High-dimensionality

広島大学　大学院理学研究科　　柳原 宏和
大阪大学　大学院基礎工学研究科　下平 英寿

正規性を仮定した多変量線形回帰モデルにおいて, Nishii [1] により提案された, $-2\times$ 最大対数尤度にモデルのパラメータ数の定数倍 m (>0) を加えることで定義される一般化情報量規準 (Generalized Information Criterion; GIC) の最小化により最適な変数を選ぶ変数選択法を取り扱う. GIC は, 既に提案されている様々な情報量規準を含むものであり, n を標本数とすると, 具体的な情報量規準と m の関係は, 以下の通りである.

$$\mathrm{AIC}: m=2, \quad \mathrm{BIC}: m=\log n, \quad \mathrm{CAIC}: m=1+\log n, \quad \mathrm{HQC}: m=2\log\log n. \tag{1}$$

変数選択法の望ましい性質の一つとして, 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が漸近的に 1 となる性質, 即ち, 一致性がある. n のみを ∞ とする大標本漸近理論で評価すれば, GIC が一致性を持つための m の条件は以下のようになる.

$$\lim_{n\to\infty} m=\infty, \quad \lim_{n\to\infty}\frac{m}{n}=0. \tag{2}$$

この結果から, (1) 式での AIC は一致性を持たず, BIC, CAIC, HQC は一致性を持つことがわかる.

近年, 目的変数ベクトルの次元数 p 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 は大きいと言えども, n よりも小さいようなデータ, いわゆる, moderately high-dimensional data (例えば, Zheng *et al.* [3] 参照) を考える. そのような高次元データおいて, 大標本漸近理論による漸近近似では精度が悪くなるが, n だけではなく p も ∞ とする高次元大標本漸近理論による漸近近似では, 近似の精度も悪くならないことが知られている. Yanagihara *et al.* [2] では, $p/n\to c_0\in[0,1)$ の条件の下で, n と p を同時に ∞ とする漸近枠組みにおいて情報量規準が一致性持つための条件を導出しており, その結果から, AIC が一致性を持ち, BIC, CAIC, HQC が一致性を持たない場合があることが報告されている. Yanagihara *et al.* [2] の条件では, 一致性を持つか持たないかは非心パラメータ行列の最大固有値の発散速度に依存しており, 残念ながら, 現状提案されている規準量の中で, どのような非心パラメータ行列でも一致性を持つような規準量を見つけられない. そこで, 本発表では, どのような非心パラメータ行列でも, さらに使用する漸近理論が大標本でも高次元大標本のどちらでも, 一致性を持つような規準量を, 一致性を持つための条件を再評価することにより提案する. ここでは, 大標本漸近理論と高次元大標本漸近理論を統一的に取り扱うため, $p/n\to c_0\in[0,1)$ の条件の下で, n を ∞ とする漸近理論により一致性を評価する. この場合, p は ∞ となってもならなくてもどちらでも良いことに注意する. 再評価した条件により, 一致性を持つ規準量として, 以下の条件を満たす規準量が提案できる.

$$m=-\frac{n}{p}\log\left(1-\frac{p}{n}\right)+\alpha, \quad \lim_{n\to\infty,p/n\to c_0}\sqrt{p}\alpha=\infty, \quad \lim_{n\to\infty,p/n\to c_0}\frac{p\alpha}{n}=0. \tag{3}$$

p を固定して, n のみを ∞ とすれば, $m=1+\alpha+O(n^{-1})$ となり, (3) 式の条件式は (2) の条件式と同値になることがわかる.

引用文献：

[1] Nishii, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.*, **12**, 758–765.

[2] Yanagihara, H., Wakaki, H. & Fujikoshi, Y. (2015). A consistency property of the AIC for multivariate linear models when the dimension and the sample size are large. *Electron. J. Statist.*, **9**, 869-897.

[3] Zheng, S., Jiang, D., Bai, Z. & He, X. (2014). Inference on multiple correlation coefficients with moderately high dimensional data. *Biometrika*, **101**, 748–754.