Robustness to Nonnormality on High-dimensionality-adjusted Consistent Generalized C_p Criterion for Multivariate Linear Regression Models

広島大学　大学院理学研究科　柳原 宏和

多変量線形回帰モデルにおいて, 推定された分散共分散行列により基準化された最小残差平方和に平均に関するパラメータ数の α 倍 ($\alpha > 0$) を加えることで定義される, 一般化 C_p (Generalized C_p; GC_p) 規準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. 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. 一致性は, 多くの場合, 標本数 n のみを無限大とする漸近理論である, 大標本漸近理論により評価されている. 一方で, 近年, ハードウェアの発展により, 蓄積・解析できるデータの数が爆発的に増大し, 目的変数の次元数 p が大きいデータである, 高次元データの解析の需要が高まっている. 本論文で取り扱う高次元データとは, 次元数 p は大きいが標本数 n よりも小さいとする適度な高次元 (moderately high-dimensional) データ (Yao *et al.* [3] 参照) である. このような高次元データでは, 大標本漸近理論ではなく, 次元数 p も p/n が 1 未満の定数に収束するという条件の下で n と共に無限大とする, 高次元大標本漸近理論により一致性を評価した方が妥当である.

近年, Yanagihara [2] で, 以下のような漸近理論に基に, 高次元性を調整した一致性を持つ一般化 C_p (High-dimensionality-adjusted Consistent Generalized C_p: $HCGC_p$) 規準が提案された.

$$n \to \infty, \quad p/n \to c_0 \in [0, 1). \tag{1}$$

上記の漸近理論は, 次元数 p を無限大にしてもしなくてもよいため, 大標本漸近理論と高次元大標本漸近理論の両方を特別な形として含むものになっている. $HCGC_p$ での実際の α は, 以下のような定数である.

$$\alpha = \frac{n}{n-p} + \beta, \quad \beta > 0 \quad s.t. \quad \lim_{n\to\infty, p/n\to c_0} \sqrt{p}\beta = \infty, \quad \lim_{n\to\infty, p/n\to c_0} \frac{p}{n}\beta = \infty. \tag{2}$$

ここでの lim は, (1) 式に基づく漸近理論の下での極限を表していることに注意する.

$HCGC_p$ は, (1) 式の漸近理論の下で一致性を持つが, 残念ながら, この性質は真のモデルの分布が多変量正規分布であるという仮定の下で保証されたものであるため, データが正規分布に従っていないときでも一致性を持つかどうかはわからない. 本発表では, データが正規分布に従っていなくても真のモデルの分布があるクラスに属しているのであれば, $HCGC_p$ は (1) 式の漸近理論の下で一致性を持つということを示す. また, Yanagihara [1] では, 候補のモデルに正規性を仮定した多変量線形回帰モデルにおける変数選択問題において, 最大対数尤度に基づく変数選択規準の一致性を, 候補のモデルに正規性を仮定したが真のモデルは正規分布ではないという状況下での一致性を高次元大標本漸近理論により評価している. しかしながら, この結果は, 非心パラメータ行列を n で割った行列の最大固有値のオーダーが $O(p)$ であるという強い仮定の下で導出されている. 本発表では, このような非心パラメータ行列の固有値の条件も緩めることも試みる.

引用文献：

[1] Yanagihara, H. (2015). Conditions for consistency of a log-likelihood-based information criterion in normal multivariate linear regression models under the violation of normality assumption. *J. Japan Statist. Soc.*, **45**, 21–56.

[2] Yanagihara, H. (2016). A high-dimensionality-adjusted consistent C_p-type statistic for selecting variables in a normality-assumed linear regression with multiple responses. *Procedia Computer Science*, **96**, 1096–1105.

[3] Yao, J., Zheng, S. & Bai, Z. (2015). *Large Sample Covariance Matrices and High-dimensional Data Analysis*. Cambridge University Press, New York.