共分散構造をもつ多変量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

諏訪東京理科大・共通教育センター　櫻井　哲朗
広島大・理・名誉教授　藤越　康祝

本報告では，多変量回帰分析の変数選択問題について取り扱う．一般の多変量回帰分析では, 標本数 n として $n \times p$ の目的変数 $\mathbf{Y}$, $n \times k$ の説明変数 $\mathbf{X}$ が与えられ, $\mathbf{X}$ から j 個の列ベクトルを取り出した $\mathbf{X}_j$ を説明変数とした多変量線形回帰モデルを

$$M_j : \mathbf{Y} \sim \mathrm{N}_{n,p}(\mathbf{X}_j\boldsymbol{\mathcal{B}}_j, \mathbf{I}_n \otimes \boldsymbol{\Sigma}), \quad \boldsymbol{\mathcal{B}}_j,\ \boldsymbol{\Sigma}:\ \text{未知パラメータ}$$

を考える．このとき, 数ある候補のモデルの中から最適なモデルを選ぶ問題が変数選択問題である．このような変数選択問題に対して, マローの C_p などが広く使われている．

$$\mathrm{C}_p = (n-k)\mathrm{tr}\hat{\boldsymbol{\Sigma}}_\omega^{-1}\hat{\boldsymbol{\Sigma}}_j + 2pj, \quad \hat{\boldsymbol{\Sigma}}_j = \frac{1}{n}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_j)\mathbf{Y}, \quad \hat{\boldsymbol{\Sigma}}_\omega = \frac{1}{n}\mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_\omega)\mathbf{Y},$$
$$\mathbf{P}_j = \mathbf{X}_j(\mathbf{X}_j'\mathbf{X}_j)^{-1}\mathbf{X}_j', \quad \mathbf{P}_\omega = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

これは, 次のリスクの漸近的不偏推定量として提案されている．

$$R_{\mathrm{C}} = \mathrm{E}_{\mathbf{Y}}\mathrm{E}_{\mathbf{Y}_F}[\mathrm{tr}\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_F - \hat{\mathbf{Y}}_j)'(\mathbf{Y}_F - \hat{\mathbf{Y}}_j)],$$
$$\mathbf{Y}_F \perp \mathbf{Y}, \quad \mathbf{Y}_F \sim \mathrm{N}_{n,p}(\mathbf{X}_j\boldsymbol{\mathcal{B}}_j, \mathbf{I}_n \otimes \boldsymbol{\Sigma})$$

このように定義された規準量において, Fujikoshi, Sakurai and Yanagihara (2014) や Yanagihara (2016) などの研究結果から, C_p は標本数・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, 規準量の一致性, を持つことが知られている．

以上のような一般の共分散行列 $\boldsymbol{\Sigma}$ のもとでは, 豊富なモデルを表現する一方で標本数と次元数の問題がある．標本数と次元数の問題とは, 標本数は次元数を超えなければないという問題である．具体的な問題として, 規準量を求めるにあたり $\boldsymbol{\Sigma}$ の最尤推定量の逆行列や行列式の対数などを求める必要がある．標本数が次元数を超えていないとき, これらの値は求めることができない．この標本数と次元数の問題を解決する方法として, Ridge などの様々な手法が提案されている．ここでは, 共分散行列 $\boldsymbol{\Sigma}$ に構造を仮定することで対処する．ここで取り扱う構造は次のような構造である．

独立構造:$\boldsymbol{\Sigma} = \sigma^2\mathbf{I}_p$,　一様構造:$\boldsymbol{\Sigma} = \sigma^2\{(1-\rho)\mathbf{I}_p + \rho\mathbf{1}\mathbf{1}'\}$,　自己回帰構造:$\boldsymbol{\Sigma} = \sigma^2(\rho^{|i-j|})$

このような構造のもとで, C_p 規準を導出し, 規準量の一致性について考察する．また, 数値シミュレーションによって今回求めた結果の妥当性を検証する．

参考文献

1. Fujikoshi, Y. , Enomoto, R. and Sakurai, T. (2014). Consistency of high-dimensional AIC-type and C_p-type criteria in multivariate linear regression. *Journal of Multivariate Analysis*, **123**, 184-200.

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.