

高次元多変量回帰モデルにおける変数と次元の同時選択法

広島大・理・名誉教授

藤越 康祝

公立諏訪東京理科大・共通・マネジメント教育センター

櫻井 哲朗

p 個の目的変数 $\mathbf{y} = (y_1, \dots, y_p)'$ と k 個の説明変数 $\mathbf{x} = (x_1, \dots, x_k)'$ に関する大きさ n の観測行列を $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_n)' : n \times p$, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)' : n \times k$ とし, 多変量回帰モデル

$$\mathbf{Y} = \mathbf{X}\Theta + \mathbf{E}$$

を想定する. ここに, Θ は $k \times p$ の未知パラメータ行列で, 誤差行列 $\mathbf{E}$ の各行は互いに独立で, それぞれ平均 $\mathbf{0}$, 共分散行列 Σ の p 次元正規分布に従うとする. より簡単な線形構造を探索するため, 説明変数の選択, および, Θ のランクの選択の同時選択問題を考える. 第 i 説明変数 x_i をモデルに取り込むかどうかは Θ の第 i 行 $\theta_{(i)}$ がゼロベクトルかどうかで定式化される. Θ のランクが r であることは, k 個の説明変数が r 個の潜在説明変数で表されることを意味している. このような同時選択問題は, (1) $\Sigma = \sigma^2 \mathbf{I}_p$, および, (2) $\Sigma > \mathbf{O}$ の場合に研究されている. (1) の場合には, Bunea et al. (2012), Chen et al. (2012), 藤越 (2016) による方法がある. 藤越 (2016) の方法は, (2) の場合でも考察されている. 同時選択法として, 最初に変数選択をし, 次に, 選ばれた変数に関する次元を選択する 2 段階法を考えるが, 本報告では, 各段階で藤越 (2016) を改良した選択法を提案する. 具体的には, 変数選択に関して, 計算量が格段に少なくてよい KOO(Kick-One-Out) 法を用い, ランク選択法に関しては一般化情報量を考える.

本報告で提案する同時選択法の各段階での変数選択法および次元選択法は次のように与えられる. 以下では, (1) の場合を考えるが, (2) の場合も同様である. 変数選択基準は, 変数 x_i の有意性検定統計量に基づく統計量

$$T_{d,i} = n \log(1 + s_{h,i}^2/s_e^2) - dp, \quad s_e^2 = \text{tr} \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_\mathbf{X})\mathbf{Y}, \quad s_{h,i}^2 = \text{tr} \mathbf{Y}'(\mathbf{P}_\mathbf{X} - \mathbf{P}_{\mathbf{X}_{(-i)}})\mathbf{Y}$$

に基づいて. ここに, $\mathbf{P}_\mathbf{X}$ は $\mathbf{X}$ の列ベクトルによって張られる空間への射影行列, $\mathbf{X}_{(-i)}$ は $\mathbf{X}$ から i 列を除いた行列である. $T_{d,i} > 0$ あるいは $T_{d,i} \leq 0$ によって x_i を選択するかどうかを決める選択方式を $\hat{T}_d$ とする. d は調整項である. 次元が $r \leq m = \min(k, p)$ であるモデルの一般化情報量規準量は

$$\text{IC}_{b,r} = np \log \left(1 + \frac{\ell_{r+1} + \dots + \ell_m}{\text{tr} \mathbf{Y}'(\mathbf{I}_n - \mathbf{P}_\mathbf{X})\mathbf{Y}} \right) - b(p-r)(k-r)$$

で与えられる. ここに, $\text{IC}_{b,m} = 0$. b は調整項で, $b = 2$ の場合の IC は AIC である. この規準量を用いて, $\text{IC}_{b,r}$ が最小になる次元を選ぶ. このような選択法が高次元漸近的枠組み $p/n \rightarrow c$ のもとで, (各段階で) 一致性を持つための十分条件について考察し, その結果に対し数値実験によって妥当性を検証する. なお, $\Sigma > \mathbf{O}$ の場合の $\text{IC}_{b,r}$ に相当する規準量の高次元一致性は Fujikoshi and Sakurai (2016) により調べられている.

参考文献

1. BUNEA, F., SHE, Y. and WEGKAMP, M. H. (2012). Joint variable and renk selection for parsimonious estimation of high-dimensional matrices. *Ann. Statist.*, **40**, 2359–2388.
2. CHEN, L. and HUANG, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journ. Amer. Statist. Assoc.*, **107**, 1533–1545.
3. 藤越 康祝 (2016). 多変量線形モデルにおける次元および変数選択のための情報量規準の高次元性質. 2016 年度日本統計関連学会連合大会報告集.
4. FUJIKOSHI, Y. and SAKURAI, T. (2016). High-dimensional consistency of rank estimation criteria in multivariate linear Model. *Journal of Multivariate Analysis*, **149**, 199–212.