スパースな共分散構造を持つ多変量線形回帰モデルにおける高次元情報量規準

大阪大学大学院基礎工学研究科　片山 翔太
広島大学大学院理学研究科　伊森 晋平

次の多変量線形回帰モデルを考える．

$$\boldsymbol{y}_i = \boldsymbol{B}^{*T}\boldsymbol{x}_i + \boldsymbol{\varepsilon}_i = \sum_{j=1}^{K_n} x_{ij}\boldsymbol{b}_j^* + \boldsymbol{\varepsilon}_i, \quad i = 1, \ldots, n.$$

ここで，$\boldsymbol{y}_i$ は p_n 次元結果変数ベクトル，$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iK_n})^T$ は K_n 次元説明変数ベクトル，$\boldsymbol{B}^* = (\boldsymbol{b}_1^*, \ldots, \boldsymbol{b}_{K_n}^*)^T$ は未知の $K_n \times p_n$ 係数行列をそれぞれ表し，$\boldsymbol{\varepsilon}_i$ はそれぞれ独立に p_n 次元多変量正規分布 $N_{p_n}(\boldsymbol{0}, \boldsymbol{\Sigma}^*)$ に従う．いま，真のモデル $M_{k_n^*}^* = \{i | \boldsymbol{b}_i^* \neq \boldsymbol{0}\}$ をある候補モデルの集合 $\mathcal{M}$ から情報量規準によって選択することを考える．なお下添字 k_n^* はその集合の要素数を表す．この問題に対して，Nishii(1984) で提案された一般化情報量規準 (GIC) は，ある候補モデル $M_{k_n} \in \mathcal{M}$ に対して次で与えられる．

$$\mathrm{GIC}_{a_n}(M_{k_n}) = \log\det(\boldsymbol{S}_{k_n}) + a_n p_n k_n.$$

ここで，$\boldsymbol{S}_{k_n} = n^{-1}\boldsymbol{Y}^T(\boldsymbol{I}_n - \boldsymbol{P}_{k_n})\boldsymbol{Y}$, $\boldsymbol{P}_{k_n} = \boldsymbol{X}_{k_n}(\boldsymbol{X}_{k_n}^T\boldsymbol{X}_{k_n})^{-1}\boldsymbol{X}_{k_n}^T$, $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)^T$, $\boldsymbol{X}_{k_n} = (x_{ij} : 1 \leq i \leq n, j \in M_{k_n})$ であり，a_n はある正の調節パラメータである．

近年の情報技術の発展に伴い，p_n および K_n が標本サイズ n に比して大きなデータ (高次元データ) が，マーケティングリサーチ，ウェブ解析，生物情報など多種多様な分野で収集されている．しかしながら，このような高次元データに対して上記の GIC は，$\boldsymbol{X}_{k_n}^T\boldsymbol{X}_{k_n}$ の特異性および $\boldsymbol{S}_{k_n}$ の特異性といった 2 つの問題により定義出来ない．そこで本報告では，高次元データにおけるこれらの問題点を解決した新しい GIC を提案する．最初の問題点に対しては，Chen and Chen (2008) で導入された制限候補モデルを考える．すなわち候補モデルを，$k_n \leq \min(n, K_n)$ を満たすものに制限する．残りの問題点に対しては，Yuan and Lin (2007) で提案された精度行列のスパース推定量を用いる．具体的には，$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ として，係数行列 $\boldsymbol{B}^*$ と精度行列 $\boldsymbol{\Omega}^* = \boldsymbol{\Sigma}^{*-1}$ に対する ℓ_1 制約付き対数尤度関数

$$Q(\boldsymbol{B}, \boldsymbol{\Omega}) = \frac{1}{n}\mathrm{tr}\boldsymbol{\Omega}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}) - \log\det(\boldsymbol{\Omega}) + \tau_n\|\boldsymbol{\Omega}^-\|_1$$

を考え，次のように高次元 GIC (HGIC) を構成する．

$$\mathrm{HGIC}_{a_n}(M_{k_n}) = Q\big(\widehat{\boldsymbol{B}}_{k_n}, \widehat{\boldsymbol{\Omega}}_{k_n}\big) + a_n p_n k_n$$

なお，$\|\boldsymbol{\Omega}\|_1$ は $\boldsymbol{\Omega}$ の要素ごとの ℓ_1 ノルムを意味し，$\tau_n > 0$，$\boldsymbol{\Omega}^- = \boldsymbol{\Omega} - \mathrm{diag}(\boldsymbol{\Omega})$ であり，

$$\big(\widehat{\boldsymbol{B}}_{k_n}, \widehat{\boldsymbol{\Omega}}_{k_n}\big) = \underset{(\boldsymbol{B}, \boldsymbol{\Omega}) \in \Theta(M_{k_n}) \times \Xi}{\mathrm{argmin}} Q(\boldsymbol{B}, \boldsymbol{\Omega})$$

である．ここで，$\Theta(M_{k_n}) = \{\boldsymbol{B} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{K_n}) | \boldsymbol{b}_i = \boldsymbol{0}, i \in M_{k_n}^c\}$, $\Xi = \{\boldsymbol{\Omega} | \boldsymbol{\Omega} = \boldsymbol{\Omega}^T\}$ である．このようにして定義された HGIC は，調節パラメータ a_n を適切なレートで選べば $(n, p_n, K_n) \to \infty$ の下で一致性を持つことが示される．詳細は当日報告する．