学位論文要旨

Consistency of log-likelihood-based information criteria for selecting variables in high-dimensional canonical correlation analysis under nonnormality

（非正規性の下での高次元正準相関分析における変数選択のための対数尤度関数に基づく情報量規準の一致性）

氏名　福井 敬祐

本論文では, 正準相関分析における冗長な変数の選択問題を考える. 正準相関分析とは変量ベクトル $\boldsymbol{x} = (x_1, \ldots, x_q)'$ と $\boldsymbol{y} = (y_1, \ldots, y_p)'$ の線形結合による合成変量の相関を分析することであり, 2 変量間の関係構造を把握する多変量解析の手法である. $\boldsymbol{z} = (\boldsymbol{x}', \boldsymbol{y}')'$ とすれば, $\boldsymbol{z}$ の平均および共分散行列は $\boldsymbol{x}$, $\boldsymbol{y}$ の平均 $\boldsymbol{\mu}_x$, $\boldsymbol{\mu}_y$ と共分散行列 $\boldsymbol{\Sigma}_{xx}$, $\boldsymbol{\Sigma}_{yy}$ を用いて,

$$E[\boldsymbol{z}] = \begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix} = \boldsymbol{\mu}, \quad Cov[\boldsymbol{z}] = \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}'_{xy} & \boldsymbol{\Sigma}_{yy} \end{pmatrix} = \boldsymbol{\Sigma},$$

と表される. ここで, $\boldsymbol{\Sigma}_{xy}$ は $\boldsymbol{x}$ と $\boldsymbol{y}$ の $q \times p$ 共分散行列である. このとき, 2 つの合成変量の相関係数の 2 乗は, $\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}'_{xy}$ の固有値として与えられ, s 番目に大きな固有値の平方根は第 s 正準相関係数と呼ばれる $(s = 1, \ldots, \min\{p, q\})$.

実解析において, 分析に対して有効でない変数 (冗長変数) の組を選別することは重要な問題であり, これは冗長性モデルの選択問題とみなすことができる. 本研究においては $\boldsymbol{x}$ の変数選択のみを考える. $j = \{j_1, \ldots, j_{q_j}\} \subseteq \omega = \{1, \ldots, q\}$ を $\boldsymbol{x}$ の変数の組み合わせを示す整数集合とし, q_j を j に含まれる要素の個数とする. 例えば $j = \{1, 2, 4\}$ のとき, $\boldsymbol{x}_j = (x_1, x_2, x_4)'$ であり, $q_j = 3$ となる. 今, q_j-次ベクトル $\boldsymbol{x}_j$ と $(q - q_j)$-次ベクトル $\boldsymbol{x}_{\bar{j}}$ により $\boldsymbol{x} = (\boldsymbol{x}'_j, \boldsymbol{x}'_{\bar{j}})'$ と分割されているとする. このとき, $\boldsymbol{\mu}_x$, $\boldsymbol{\Sigma}_{xy}$, $\boldsymbol{\Sigma}_{xx}$ も対応して

$$\boldsymbol{\mu}_x = \begin{pmatrix} \boldsymbol{\mu}_j \\ \boldsymbol{\mu}_{\bar{j}} \end{pmatrix}, \ \boldsymbol{\Sigma}_{xy} = \begin{pmatrix} \boldsymbol{\Sigma}_{jy} \\ \boldsymbol{\Sigma}_{\bar{j}y} \end{pmatrix}, \ \boldsymbol{\Sigma}_{xx} = \begin{pmatrix} \boldsymbol{\Sigma}_{jj} & \boldsymbol{\Sigma}_{j\bar{j}} \\ \boldsymbol{\Sigma}'_{j\bar{j}} & \boldsymbol{\Sigma}_{\bar{j}\bar{j}} \end{pmatrix},$$

と分割される. $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ を $\boldsymbol{z}$ と同一な分布からの独立標本であるとし, その不偏分散を $\boldsymbol{S} = (n-1)^{-1}\sum_{i=1}^n (\boldsymbol{z}_i - \bar{\boldsymbol{z}})(\boldsymbol{z}_i - \bar{\boldsymbol{z}})'$ とする. ここで, $\bar{\boldsymbol{z}} = n^{-1}\sum_{i=1}^n \boldsymbol{z}_i$ である. また, $\boldsymbol{S}$ は $\boldsymbol{\Sigma}$ の分割に対応して分割がされているとする. つまり, $\boldsymbol{S}$ が

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_{jj} & \boldsymbol{S}_{j\bar{j}} & \boldsymbol{S}_{jy} \\ \boldsymbol{S}_{\bar{j}j} & \boldsymbol{S}_{\bar{j}\bar{j}} & \boldsymbol{S}_{\bar{j}y} \\ \boldsymbol{S}_{yj} & \boldsymbol{S}_{y\bar{j}} & \boldsymbol{S}_{yy} \end{pmatrix},$$

と分割されているとする. このとき, $\boldsymbol{x}_{\bar{j}}$ に含まれる変数に冗長性を仮定したモデルは冗長性モデルと呼ばれる. 特に, $\boldsymbol{z}$ に正規性を仮定すれば,

$$M_j : (n-1)\boldsymbol{S} \sim W_{p+q}(n-1, \boldsymbol{\Sigma}) \ s.t. \ \mathrm{tr}(\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}'_{xy}) = \mathrm{tr}(\boldsymbol{\Sigma}_{jj}^{-1}\boldsymbol{\Sigma}_{jy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}'_{jy}), \quad (1)$$

と定義される. Fujikoshi (1985) により冗長性モデルの選択問題は共分散構造の選択問題として定式化され, Akaike (1973, 1974) により提案された AIC などの情報量規準により

共分散構造を選択するという方法が提案された. 本研究においては特に, 正準相関分析における冗長性モデル選択に関して次の対数尤度関数に基づく情報量規準 IC_m を考える.

$$\mathrm{IC}_m(j) = (n-1)\log\frac{|\boldsymbol{S}_{yy\cdot j}|}{|\boldsymbol{S}_{yy\cdot x}|} + m(j). \tag{2}$$

ここで, $\boldsymbol{S}_{yy\cdot j} = \boldsymbol{S}_{yy} - \boldsymbol{S}'_{jy}\boldsymbol{S}_{jj}^{-1}\boldsymbol{S}_{jy}$ であり, 特に, $j = \omega$ のとき $\boldsymbol{S}_{yy\cdot x}$ とかく. $m(j)$ はモデルの複雑さに対する罰則を表現する項であり, (2) は AIC や AIC_c (Fujikoshi, 1985), CAIC (Bozdogan, 1987), BIC (Schwarz, 1978), HQC (Hannan & Quinn, 1979) などの対数尤度関数に基づく情報量規準を含んだ情報量規準であり, $m(j)$ によって特徴付けられる. 最適な変数の組み合わせ $\hat{j}_m$ は IC_m の最小化により決定される, つまり,

$$\hat{j}_m = \arg\min_{j\subseteq\omega} \mathrm{IC}_m(j).$$

情報量規準を用いてモデル選択を行う場合, 使用する情報量規準が一致性を持つことが重要な特性とされる. 一致性とは, 真のモデルにおける変数の組み合わせを j_* としたとき, $\hat{j}_m = j_*$ となる確率が漸近的に 1 に収束することである.

Yanagihara *et al.* (2014) は $\boldsymbol{y}$ の次元数 p が大きいような高次元データに対する正準相関分析, すなわち, 高次元正準相関分析における冗長性モデル選択のための情報量規準について, 標本数 n と p が $c_{n,p} = p/n \to c_0 \in (0, 1]$ の仮定を満たしながら同時に ∞ に近づく高次元大標本漸近理論の枠組みで, IC_m が一致性を持つための $m(j)$ の条件を導出している. 情報量規準の特性の評価は次元数 p を固定し標本数 n のみを ∞ に近づける大標本漸近理論の枠組みで行われることが一般的であったが, 次元数 p が大きい高次元データに関しては, p を固定した大標本漸近理論による議論よりも次元数 p も同時に ∞ に近づく高次元大標本漸近理論の枠組みでの議論が適当であると考えられる. また, Yanagihara *et al.* (2014) において高次元大標本漸近理論の枠組みで導出された条件は, 大標本漸近理論の下で導出される条件とは異なった結果を与えており, 非常に興味深いものである.

高次元漸近理論を用いる場合には, 共分散行列の推定量の行列の大きさが p とともに増加してしまう, ことにより 通常の行列の大数の法則が使用できないという問題を回避する必要がある. 正準相関分析における冗長性モデル選択のための情報量規準の評価においても例外でなく, $\boldsymbol{S}_{yy\cdot j}$ の大きさが ∞ に発散してしまう問題点がある. Yanagihara *et al.* (2014) においては, $\boldsymbol{z}$ を発生させる真のモデルの分布に正規性を仮定することで, Wishart 分布の特性を $\boldsymbol{S}_{yy\cdot j}$ に適用し, その問題点を回避している. しかし, 実解析の場面においては正規性が仮定できるとは限らない. そこで本研究においては, 候補のモデルには正規性を仮定するが, $\boldsymbol{z}$ を発生させる真のモデルの分布に正規性を仮定しない下で高次元漸近漸近理論の枠組みによる一致性成立の条件の導出を行う.

正規性の仮定がない場合には, Wishart 分布の特性を利用することはできない. そこで, Yanagihara (2013) で用いられた, 基本的な線形代数の行列変換を用いる方法を使用する. Yanagihara (2013) では多変量線形回帰モデルの変数選択について, 情報量規準の一致性成立の条件が導出されているが, 正準相関分析においては多変量線形回帰モデルと異なり, $\boldsymbol{y}$ と $\boldsymbol{x}$ の両方が確率変数となるため, Yanagihara (2013) とは異なる仮定の下での導出を行う必要がある.