主成分分析における次元数の推定

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，主成分分析における次元数の推定問題について関心がある．このとき, はじめのいくつかの異なる固有値は大きく, 残りの全ての固有値の小さく等しいモデルを考え, その大きな固有値の個数を次元数とする．このようなモデルは, Johnstone (2001) でも取り扱われ, そこではスパイクモデルと呼ばれている．スパイクモデルにおいて次元数の推定は重要な問題の 1 つである．

いま, 標本 $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N \overset{i.i.d.}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ が与えられているとする．このとき, 次元数を k とするスパイクモデル M_j は次のように表せる．

$$M_j : \lambda_j > \lambda_{j+1} = \cdots = \lambda_p = \lambda, \qquad \lambda_1, \ldots, \lambda_p \text{ は } \boldsymbol{\Sigma} \text{ の固有値}$$

このとき, 候補のモデル $\{M_0, M_1, \ldots M_{p-1}\}$ のから最適なモデルを選ぶ規準量として次の AIC 型・BIC 型の規準量を提案する．

$$\mathrm{AIC}_j = N \log \ell_1 \cdots \ell_j + N(p-j) \log \bar{\ell}_{(p-j)} + N \log \left(\frac{n}{N}\right)^p + Np \log 2\pi + 2\hat{b}_j$$

$$\mathrm{BIC}_j = N \log \ell_1 \cdots \ell_j + N(p-j) \log \bar{\ell}_{(p-j)} + N \log \left(\frac{n}{N}\right)^p + Np \log 2\pi + (\log n)\hat{b}_j$$

$$\bar{\ell}_{(p-j)} = \frac{1}{p-j}(\ell_{j+1} + \cdots + \ell_p), \quad \hat{b}_j = pj - \frac{1}{2}j(j+1) + j + 1 + p$$

ここで, $n = N - 1$, $\ell_1 > \cdots > \ell_p$ は $\mathbf{S}$ の固有値であり, $\mathbf{S}$ は次の標本共分散行列である．

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})', \quad \bar{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i$$

ここでは, 全ての固有値が異なるとするフルモデル M_{p-1} のときの値との差をとった次の規準量において考える．

$$\begin{aligned} A_j &= \mathrm{AIC}_j - \mathrm{AIC}_{p-1} \\ &= -N \sum_{i=j+1}^{p} \log \ell_j + N(p-j) \log \bar{\ell}_{(p-j)} - 2 \cdot \frac{1}{2}(p-j-1)(p-j+2) \\ B_j &= \mathrm{BIC}_j - \mathrm{BIC}_{p-1} \\ &= -N \sum_{i=j+1}^{p} \log \ell_j + N(p-j) \log \bar{\ell}_{(p-j)} - (\log n) \cdot \frac{1}{2}(p-j-1)(p-j+2) \end{aligned}$$

これを使い, 候補のモデルから次のモデルを最適なモデルとして選択する．

$$\hat{j}_A = \arg\min_j A_j, \qquad \hat{j}_B = \arg\min_j B_j$$

さらに, これらの規準量に対して大標本漸近枠組および高次元漸近枠組での真のモデルを選ぶ確率の漸近的な評価を与える．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と各規準量の比較を行い有効な規準量を明らかにする．

参考文献

1. Johnstone, I. M. (2001). On the distribution of the largest principal component. *Ann. Statist.*, **29**, 295–327.