SIRの改良について

秋田 智之（広島大・理・院）・若木 宏文（広島大・理）

0　概要

本発表では，以下のようなモデル

$$y = f(\boldsymbol{\beta}_1'\boldsymbol{x}, \boldsymbol{\beta}_2'\boldsymbol{x}, \ldots, \boldsymbol{\beta}_K'\boldsymbol{x}, \varepsilon), \quad \boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \Sigma), \quad \varepsilon \amalg \boldsymbol{x}$$

に従うデータ $\{(y_1,\boldsymbol{x}_1), \ldots, (y_n,\boldsymbol{x}_n)\}$ が得られたときの係数ベクトル$\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ の推定について考えた．1つの方法としてLi(1991) によって提案された層別化逆回帰法を用いればいくつかの場合によい推定が得られる．しかし，リンク関数が $y = (\boldsymbol{\beta}_1'\boldsymbol{x})^2 + \varepsilon$ のように対称な場合は推定に失敗することが知られている．今回はこの改良法として，スライスをさらに細分する方法を提案し，シミュレーションの結果を紹介し，ある条件下での推定量の収束の証明について言及した．

1　主成分分析を用いた層別化逆回帰法の改良 (PCA-SIR)

層別化逆回帰法 (SIR) のシミュレーションによるとリンク関数の形によっては y の値だけでグループ分けした方法はうまくいかないケースがあることが分かった．そこで$\boldsymbol{x}$の値によってもグループを分割する方法を考える．今回は次のような改良法を提案した．

(1) y の範囲を H 個のスライスに分割する．

(2) 各スライス内の$\boldsymbol{x}_i^{(h)}$ $(i = 1, \ldots, N; h = 1, \ldots, H)$ に対し，主成分分析を行い，第1主成分と第 p 主成分の値でデータをさらに $4H$ のグループに分割する．即ち各スライスで分散共分散行列

$$S_h = \frac{1}{N-1}\sum_{i=1}^{N}(\boldsymbol{x}_i^{(h)} - \bar{\boldsymbol{x}}^{(h)})(\boldsymbol{x}_i^{(h)} - \bar{\boldsymbol{x}}^{(h)})' \quad (\bar{\boldsymbol{x}}^{(h)} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i^{(h)})$$

の固有値を $\lambda_1^{(h)} > \cdots > \lambda_p^{(h)}$ とし，これに対応する固有ベクトルを$\boldsymbol{l}_1^{(h)}, \ldots, \boldsymbol{l}_p^{(h)}$ としたとき，第1主成分 $z_{1i}^{(h)}$ と第 p 主成分 $z_{pi}^{(h)}$ は

$$z_{1i}^{(h)} = \boldsymbol{l}_1^{(h)\prime}\boldsymbol{x}_i^{(h)}, \quad z_{pi}^{(h)} = \boldsymbol{l}_p^{(h)\prime}\boldsymbol{x}_i^{(h)}$$

で与えられ，それぞれの主成分の平均を $\bar{z}_1^{(h)}, \bar{z}_p^{(h)}$ とする．この2つの値によって h 番目のスライス内のデータを次のように分割する．

$$\{\boldsymbol{x}_i^{(h)} | z_{1i}^{(h)} \geq \bar{z}_1^{(h)}, z_{pi}^{(h)} \geq \bar{z}_p^{(h)}\} \quad \{\boldsymbol{x}_i^{(h)} | z_{1i}^{(h)} \geq \bar{z}_1^{(h)}, z_{pi}^{(h)} \leq \bar{z}_p^{(h)}\}$$
$$\{\boldsymbol{x}_i^{(h)} | z_{1i}^{(h)} \leq \bar{z}_1^{(h)}, z_{pi}^{(h)} \geq \bar{z}_p^{(h)}\} \quad \{\boldsymbol{x}_i^{(h)} | z_{1i}^{(h)} \leq \bar{z}_1^{(h)}, z_{pi}^{(h)} \leq \bar{z}_p^{(h)}\}$$

(3) 分割された $4H$ 個のグループ

$$\{\boldsymbol{x}_1^{(1)},\ldots,\boldsymbol{x}_{N_1}^{(1)}\},\{\boldsymbol{x}_1^{(2)},\ldots,\boldsymbol{x}_{N_2}^{(2)}\},\ldots,\{\boldsymbol{x}_1^{(4H)},\ldots,\boldsymbol{x}_{N_{4H}}^{(4H)}\}$$

に対し正準判別分析を行う．即ち

$$\hat{\Sigma}=\frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})'$$

$$\hat{\Gamma}=\frac{1}{n}\sum_{h=1}^{4H}N_h(\bar{\boldsymbol{x}}^{(h)}-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}^{(h)}-\bar{\boldsymbol{x}})'$$

$$(\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i,\quad \bar{\boldsymbol{x}}^{(h)}=\frac{1}{N_h}\sum_{i=1}^{N_h}\boldsymbol{x}_i^{(h)}\ (h=1,\ldots,4H))$$

として $\hat{\Sigma}^{-1/2}\hat{\Gamma}\hat{\Sigma}^{-1/2}$ の固有値を $\hat{\lambda}_1\geq\cdots\geq\hat{\lambda}_K\geq\cdots\geq\hat{\lambda}_p$ として，これに対応する固有ベクトルを $\hat{\boldsymbol{h}}_1,\ldots,\hat{\boldsymbol{h}}_K,\ldots,\hat{\boldsymbol{h}}_p$ として

$$\hat{\boldsymbol{\beta}}_j=\hat{\Sigma}^{-1/2}\hat{\boldsymbol{h}}_j$$

とする．$\hat{\boldsymbol{\beta}}_j$ で張られる空間を$\boldsymbol{\beta}_1$,$\boldsymbol{\beta}_2$,...,$\boldsymbol{\beta}_K$ が張る空間の推定量とする．

2　シミュレーション

従来の SIR と今回提案した PCA-SIR を比較する数値実験を行うと，PCA-SIR は従来の SIR が推定に失敗するケースにおいてよい推定を与え，更に従来の SIR がうまく働くケースにおいても比較的よい推定を与えることがわかる．実験の結果は公演で紹介した．

3　推定量の収束について

この節では PCA-SIR の理論的根拠について考える．現段階では $K=1$ で，リンク関数が対称である場合には推定量の一致性が示せていたのでこれを紹介した．

まずモデルとしては次のようなものを考える．

$$y=f(\boldsymbol{\beta}_1'\boldsymbol{x},\varepsilon)\quad \boldsymbol{x}\sim N_p(\boldsymbol{\mu},\Sigma)$$

このとき次が成り立つ．

定理 1. $f(z,\varepsilon)$ が対称であると仮定し，このモデルに従うデータ $\{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_n,y_n)\}$ が得られたとする．このとき *PCA-SIR* で作られた $\hat{\Sigma}^{-1/2}\hat{\Gamma}\hat{\Sigma}^{-1/2}$ の最大固有値に対応する固有ベクトルは $\Sigma^{1/2}\boldsymbol{\beta}_1$ に収束する．

この証明の概略についても当日紹介した．