高次元カーネル主成分分析の漸近的性質とその応用

筑波大学・数理物質科学　　中山　優吾
筑波大学・数理物質系　　矢田　和善
筑波大学・数理物質系　　青嶋　誠

本講演では，高次元データのクラスタリングを考える．高次元 PCA によって潜在空間を推定し，そこへデータを射影させることで高精度なクラスタリング手法を提案する．

2 つの d 次元分布を Π_1, Π_2 と名付け，それぞれ平均 $\boldsymbol{\mu}_1$，$\boldsymbol{\mu}_2$ と，共分散行列 $\boldsymbol{\Sigma}_1$，$\boldsymbol{\Sigma}_2$ をもつと仮定する．いま，データは，p.d.f.

$$f(\boldsymbol{x}) = \varepsilon_1 f_1(\boldsymbol{x}; \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \varepsilon_2 f_2(\boldsymbol{x}; \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2),\ \varepsilon_1 + \varepsilon_2 = 1\ (\varepsilon_i > 0)$$

をもつ混合分布からの標本とみなす．ここで，$f_i(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ は Π_i の p.d.f. である．この母集団から n (≥ 2) 個のデータを無作為に抽出し，データ行列を $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$ とする．そのとき，$\mathrm{Var}(\boldsymbol{x}_i) = \varepsilon_1 \boldsymbol{\Sigma}_1 + \varepsilon_2 \boldsymbol{\Sigma}_2 + \varepsilon_1 \varepsilon_2 (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T$ $(= \boldsymbol{\Sigma})$ である．$\Delta = ||\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2||^2$ とおく．$\boldsymbol{\Sigma}$ の固有値を $\lambda_1 \geq \cdots \geq \lambda_d (\geq 0)$ とし，適当な直交行列 $\boldsymbol{H} = [\boldsymbol{h}_1, ..., \boldsymbol{h}_d]$ で $\boldsymbol{\Sigma} = \boldsymbol{H} \boldsymbol{\Lambda} \boldsymbol{H}^T$，$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, ..., \lambda_d)$ と分解する．さらに，$\boldsymbol{X} - [\boldsymbol{\mu}, ..., \boldsymbol{\mu}] = \boldsymbol{H} \boldsymbol{\Lambda}^{1/2} \boldsymbol{Z}$ として，$\boldsymbol{Z} = (z_{ij})$ と表記する．

Yata and Aoshima [1] は，第 1 主成分スコア s_{1j} $(= \sqrt{\lambda_1} z_{1j})$, $j = 1, ..., n$ について，平均ベクトル間の距離 Δ に関する条件

$$\frac{\lambda_{\max}(\boldsymbol{\Sigma}_i)}{\Delta} \to 0 \ \text{ as } d \to \infty \text{ for } i = 1, 2 \tag{1}$$

のもとで

$$\operatorname*{plim}_{d \to \infty} \frac{s_{1j}}{\sqrt{\lambda_1}} = \begin{cases} \sqrt{\varepsilon_2 / \varepsilon_1}, & \boldsymbol{x}_j \in \Pi_1, \\ -\sqrt{\varepsilon_1 / \varepsilon_2}, & \boldsymbol{x}_j \in \Pi_2 \end{cases} \tag{2}$$

なる一致性を示した．ただし，$\lambda_{\max}(\boldsymbol{\Sigma}_i)$ は $\boldsymbol{\Sigma}_i$ の最大固有値を表す．つまり，第 1 主成分スコアを精度よく推定できれば，その符号から高次元データを分類することができる．しかしながら，条件 (1) を満たすほど Δ が十分に大きくなければ，従来の PCA を用いて高次元データを分類することは困難である．

一方で，非線形な構造をもつ場合はカーネル PCA が有効であることが知られている．そこで，高次元非線形構造まで加味したクラスタリング手法を与えるために，ガウシアンカーネル

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp(-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 / \gamma) \ \ (\gamma > 0)$$

を用いた高次元カーネル PCA を考える．高次元におけるカーネル主成分スコアの漸近的性質を導出し，平均ベクトル間の距離 Δ だけでなく，共分散行列間の距離にもよって (2) のような主成分スコアの高次元一致性をもつことを示す．当日は，カーネル PCA と従来の PCA との理論的な比較を行い，その性能を数値実験と実データ解析を用いて検証する．

[1] Yata, K., Aoshima, M. (2015). Principal component analysis based clustering for high-dimension, low-sample-size data. arXiv:1503.04525.