

カーネル主成分分析に基づく高次元データのクラスタリングとチューニング

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 $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ とする。 $n_i = \#\{j | \mathbf{x}_j \in \Pi_i, j = 1, \dots, n\}$ とする。ここで、 $\#A$ は集合 A の要素の個数とし、 $n_i \geq 1$ とする。簡単のため、以下を仮定する。

$$\text{tr}(\boldsymbol{\Sigma}_1) \leq \text{tr}(\boldsymbol{\Sigma}_2), \quad \mathbf{x}_j \in \Pi_1 \text{ for } j = 1, \dots, n_1 \text{ and } \mathbf{x}_j \in \Pi_2 \text{ for } j = n_1 + 1, \dots, n.$$

本講演では、ガウシアンカーネルを用いたカーネル PCA の漸近的性質を考えることでクラスタリング手法を与える。 $n \times n$ のグラム行列 $\mathbf{K}$ の (j, j') 成分を

$$k(\mathbf{x}_j, \mathbf{x}_{j'}) = \exp(-\|\mathbf{x}_j - \mathbf{x}_{j'}\|^2 / \gamma) \quad (\gamma > 0)$$

とする。 $\mathbf{I}_n$ を n 次の単位行列、 $\mathbf{1}_n = (1, \dots, 1)^T$ として、 $\mathbf{P}_n = \mathbf{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n^T$ とする。中心化グラム行列を $\mathbf{K}_0 = \mathbf{P}_n \mathbf{K} \mathbf{P}_n$ とし、 $\mathbf{K}_0$ の固有値分解を $\mathbf{K}_0 = \sum_{i=1}^{n-1} \hat{\lambda}_i \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i^T$ ($\hat{\mathbf{u}}_i = (\hat{u}_{i1}, \dots, \hat{u}_{in})^T$, $\|\hat{\mathbf{u}}_i\|^2 = 1$) とする。 $\mathbf{x}_j$ の (基準化した) 第 i 主成分スコアを $s_{ij} = \sqrt{n} \hat{u}_{ij}$ とおく。

$\kappa_\mu = \exp(-\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 / \gamma)$, $\kappa_\Sigma = \exp(-|\text{tr}(\boldsymbol{\Sigma}_1) - \text{tr}(\boldsymbol{\Sigma}_2)| / \gamma)$. $\Delta_\kappa = 1 + \kappa_\Sigma^2 - 2\kappa_\mu \kappa_\Sigma$ とおき、以下を仮定する。

$$\limsup_{d \rightarrow \infty} \frac{(1 - \kappa_\Sigma^2)n}{\Delta_\kappa n_1 n_2} < 1 \quad (1)$$

第 1 主成分スコア s_{1j} , $j = 1, \dots, n$ について、以下の一致性が成り立つ。

定理 1. 適当な正則条件と (1) のもと、 $d \rightarrow \infty$ で以下が成り立つ。

$$s_{1j} = \begin{cases} \sqrt{n_2/n_1} + o_P(1) & \text{when } j = 1, \dots, n_1, \\ -\sqrt{n_1/n_2} + o_P(1) & \text{when } j = n_1 + 1, \dots, n. \end{cases} \quad (2)$$

定理 1 から第 1 主成分スコアの符号で高次元データを分類できることがわかる。Yata and Aoshima [1] は線形カーネルに対して、平均ベクトル間の距離 $\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$ が十分大きければ、(1) と同様な主成分スコアの一致性を得られることを示した。ガウシアンカーネルを用いたカーネル PCA は $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$ のときでさえも、共分散行列間の距離によって (2) を与えることができる。

当日は、従来の PCA とガウシアンカーネルを用いたカーネル PCA の理論的な比較を与える。さらに、(1) を満たすような γ の選択法を与え、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.