高次元小標本における
非階層型クラスタリングの一致性について

東京理科大学・創域理工学部　　江頭　健斗（Kento Egashira）
Information Sciences
Tokyo University of Science
筑波大学・数理物質系　　矢田　和善（Kazuyoshi Yata）
Institute of Mathematics
University of Tsukuba
筑波大学・数理物質系　　青嶋　誠（Makoto Aoshima）
Institute of Mathematics
University of Tsukuba

1　はじめに

本論文では，非階層型クラスタリングとして，高次元小標本データに対する k-means 法を考える．高次元データに対するクラスタリングについて，Liu et al. (2008) は，“statistical significance of clustering(SigClust)” と呼ばれる 2 分割タイプのクラスタリングを提案した．Ahn et al. (2012) は，高次元における分割型の階層的クラスタリングを提案し，その漸近的性質を求めた．Huang et al. (2015) は，Liu et al. (2008) による SigClust を，ソフト閾値法によって発展させた．Yata and Aoshima (2010, 2020) は，高次元混合分布における幾何学的一致性を示し，それをクラスタリングに応用した．幾何学的一致性については，青嶋・矢田 (2019) も参照のこと．Nakayama et al. (2021) は，高次元データに対するカーネル主成分分析を用いたクラスタリングの漸近的性質を導出した．Borysov et al. (2014) は，2 クラスの高次元データに対する階層的クラスタリングの漸近的振舞いを定式化し，その性質を研究した．Egashira et al. (2023) は，3 クラス以上の高次元データに対する階層的クラスタリングの漸近的性質を導出した．一方で，高次元データに対する非階層型クラスタリングの理論的な研究が乏しいように思われる．本論文では，高次元小標本の枠組みにおける k-means 法の漸近的性質を導出する．2 節では，k-means 法を導入する．3 節では，

k-means 法の高次元漸近的性質を与える．4 節では，その漸近的性質を数値的に検証する．

2　k-means 法

本節では，k-means 法を導入する．与えられたデータセット $\boldsymbol{X}$ に対する k-means 法は，事前に設定したクラスター数 K を用いて，以下の最適化問題として定式化を与えることができる．

$$\{\hat{\boldsymbol{C}}_1, ..., \hat{\boldsymbol{C}}_K\} = \underset{\boldsymbol{C}_1,...,\boldsymbol{C}_K}{\operatorname{argmin}} \sum_{i=1}^{K} \sum_{\boldsymbol{x}\in\boldsymbol{C}_i} \|\boldsymbol{x} - \overline{\boldsymbol{C}}_i\|^2$$
$$\text{subject to } \cup_{i=1}^{K} \boldsymbol{C}_i = \boldsymbol{X},\ \boldsymbol{C}_i \cap \boldsymbol{C}_j = \emptyset \ \ (i \neq j).$$

ただし，$\overline{\boldsymbol{C}}_i$ は $\boldsymbol{C}_i$ に含まれるデータの算術平均とし，$\|\cdot\|$ はユークリッドノルムとする．k-means 法によるクラスタリング結果は，$\{\hat{\boldsymbol{C}}_1, ..., \hat{\boldsymbol{C}}_K\}$ として与えられる．一般的に，上記の k-means 法の最適化問題は，以下のような初期値に基づくアルゴリズムで局所解を求める．

k-means 法のアルゴリズム:

1. K 個の初期値 $\boldsymbol{c}_i$ $(\in \boldsymbol{X})$, $i = 1, ..., K$ を設定する．
2. 与えられた全てのデータ $\boldsymbol{x} \in \boldsymbol{X}$ に対して，$i = \underset{j=1}{\overset{K}{\operatorname{argmin}}} \|\boldsymbol{x} - \boldsymbol{c}_j\|$ ならば $\boldsymbol{x} \in \boldsymbol{C}_i$ とする．これをすべてのデータで繰り返し，集合 $\boldsymbol{C}_i, i = 1, ..., K$ を構成する．
3. 各 i で $\boldsymbol{C}_i$ に含まれるデータの算術平均を新たな初期値とみなし，ステップ 2 を行い集合 $\boldsymbol{C}_i, i = 1, ..., K$ を更新する．このとき，前のステップでの集合と更新した集合が一致するまで，このステップを繰り返す．
4. ステップ 3 で収束した集合 $\boldsymbol{C}_i, i = 1, ..., K$ を $\hat{\boldsymbol{C}}_i, i = 1, ..., K$ とおき，それらを k-means 法の結果とする．

ただし，初期値にクラスタリングの結果が依存することに注意する．

3　k-means 法の高次元漸近的性質

独立な d 次元の母集団が 2 個あると考える．$i = 1, 2$ に対して，母集団 π_i は平均に未知の d 次ベクトル $\boldsymbol{\mu}_i$，共分散行列に未知の d 次正定値対称行列 $\boldsymbol{\Sigma}_i$ をもつと仮定する．母集

団 π_i から n_i 個のデータ $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}$ を無作為に抽出する．

$$\boldsymbol{X}_i = \{\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{in_i}\}$$

とおく．$i = 1, 2$ に対して，以下を仮定として考える．

$$\limsup_{d\to\infty} \frac{\|\boldsymbol{\mu}_i\|^2}{d} < \infty, \quad \liminf_{d\to\infty} \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i)}{d} > 0, \quad \limsup_{d\to\infty} \frac{\mathrm{tr}(\boldsymbol{\Sigma}_i)}{d} < \infty.$$

また，$\Delta = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$，$L_i = \mathrm{Var}[\|\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i\|^2]$ とおき，$i = 1, 2$ に対して，以下を仮定する．

(A-i): $\quad \mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\Delta^2 \to 0,\ d \to \infty;$
(A-ii): $\quad L_i/\Delta^2 \to 0,\ d \to \infty.$

母集団に正規分布を仮定したとき，$L_i = 2\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ となるので (A-i) のもと (A-ii) が成り立つことに注意する．$K = 2$ とし，2 節で与えた k-means 法のアルゴリズムにおいて以下が成り立つ．

定理 1. *(A-i)* と *(A-ii)* を仮定する．さらに，$\boldsymbol{c}_i \in \boldsymbol{X}_i,\ i = 1, 2$ と仮定する．

$$\limsup_{d\to\infty} \frac{|\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)|}{\Delta} < 1 \tag{1}$$

のとき，$d \to \infty$ のもと $P(\hat{\boldsymbol{C}}_1 = \boldsymbol{X}_1,\ \hat{\boldsymbol{C}}_2 = \boldsymbol{X}_2) \to 1$ となる．

定理 1 より，初期値を適切に選べば，漸近的に確率 1 で $\hat{\boldsymbol{C}}_1 = \boldsymbol{X}_1,\ \hat{\boldsymbol{C}}_2 = \boldsymbol{X}_2$ と分類できる．

4 数値シミュレーション

本節では，高次元小標本のもとで，$K = 2$ と設定したときの k-means 法を数値的に検証する．母集団 π_1 と π_2 について，次の分布を考える．
(i) π_1 : $N_d(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$，$\pi_2$: $N_d(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$;
(ii) $\boldsymbol{x}_{1j} - \boldsymbol{\mu}_1$ は自由度 5 で共分散行列 $\boldsymbol{\Sigma}_1$ の d 次元 t 分布，$\boldsymbol{x}_{2j} - \boldsymbol{\mu}_2$ は自由度 5 で共分散行列 $\boldsymbol{\Sigma}_2$ の d 次元 t 分布にそれぞれ従う．
ここで，すべての成分が 1 である d 次元ベクトルを $\mathbf{1}_d$，最初の $\lceil d^{2/3} \rceil$ 個の成分が 1, それ以外が 0 である d 次元ベクトルを $\mathbf{1}_{2/3} = (1, ..., 1, 0, ..., 0)^T$ とする．ただし，$\lceil x \rceil$ は x 以上

表 1　k-means 法における誤分類確率

d	(I)	(II)	(III)	(IV)
16	0.956	0.930	0.949	0.940
32	0.925	0.908	0.909	0.927
64	0.807	0.892	0.827	0.928
128	0.646	0.925	0.687	0.947
256	0.383	0.967	0.529	0.971
512	0.235	0.991	0.425	0.989
1024	0.176	0.999	0.376	0.998
2048	0.157	1.000	0.358	1.000

の最小の整数を表す．次の 4 つの設定を考える．

(I)　(i)，$\boldsymbol{\mu}_1 = \mathbf{1}_d$，$\boldsymbol{\mu}_2 = \mathbf{0}_d$，$\boldsymbol{\Sigma}_1 = \boldsymbol{\Phi}$, $\boldsymbol{\Sigma}_2 = 1.5\boldsymbol{\Phi}$, $(n_1, n_2) = (8, 7)$;

(II)　(i)，$\boldsymbol{\mu}_1 = \mathbf{1}_{2/3}$，$\boldsymbol{\mu}_2 = \mathbf{0}_d$，$\boldsymbol{\Sigma}_1 = \boldsymbol{\Phi}$, $\boldsymbol{\Sigma}_2 = 1.5\boldsymbol{\Phi}$, $(n_1, n_2) = (8, 7)$;

(III)　(ii)，$\boldsymbol{\mu}_1 = \mathbf{1}_d$，$\boldsymbol{\mu}_2 = \mathbf{0}_d$，$\boldsymbol{\Sigma}_1 = \boldsymbol{\Phi}$, $\boldsymbol{\Sigma}_2 = 1.5\boldsymbol{\Phi}$, $(n_1, n_2) = (8, 7)$;

(IV)　(ii)，$\boldsymbol{\mu}_1 = \mathbf{1}_{2/3}$，$\boldsymbol{\mu}_2 = \mathbf{0}_d$，$\boldsymbol{\Sigma}_1 = \boldsymbol{\Phi}$, $\boldsymbol{\Sigma}_2 = 1.5\boldsymbol{\Phi}$, $(n_1, n_2) = (8, 7)$.

ただし，$\boldsymbol{\Phi} = \boldsymbol{B}(0.3^{|i-j|^{1/3}})\boldsymbol{B}$，$\boldsymbol{B} = \mathrm{diag}(\{0.5 + 1/(d+1)\}^{1/2}, ..., \{0.5 + d/(d+1)\}^{1/2})$ とする．ここで，$\mathrm{tr}(\boldsymbol{\Phi}) = d$ となる．(I) と (III) の場合は (1) を満たすが，(II) と (IV) の場合は (1) を満たさないことに注意する．次元を $d = 2^s$，$s = 4, ..., 11$ と設定する．各設定のもとでデータを発生させ，2 節で与えた k-means 法のアルゴリズムを実行のうえ，正しく分類されているかを確認した．実験を 2000 回繰り返し，誤分類確率を纏めたものが表 1 である．標準誤差は 0.011 以下であることに注意する．

定理 1 の (1) を満たす (I) と (III) の場合は誤分類確率が 0 に収束し，(1) を満たさない (II) と (IV) の場合は誤分類率が 0 に収束しないことが確認できた．特に，(I) と (III) の場合に，$\boldsymbol{c}_i \in \boldsymbol{X}_i$, $i = 1, 2$ として初期値を仮定せずとも誤分類確率が 0 に収束することを確認できた．

5　付録

一般性を失うことなく $\mathrm{tr}(\boldsymbol{\Sigma}_1) \leq \mathrm{tr}(\boldsymbol{\Sigma}_2)$ と仮定できる．$b_1 = 2\mathrm{tr}(\boldsymbol{\Sigma}_2) + \{\Delta - |\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)|\}/2$, $\Delta_* = \Delta + \mathrm{tr}(\boldsymbol{\Sigma}_1) + \mathrm{tr}(\boldsymbol{\Sigma}_2)$ とおく．$\Delta_* = \Delta - |\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)| + 2\mathrm{tr}(\boldsymbol{\Sigma}_2) =$

$\Delta + |\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)| + 2\mathrm{tr}(\boldsymbol{\Sigma}_1)$ となることに注意し，(A-i) と (1) のもとで，

$$\liminf_{d\to\infty}\{b_1 - 2\mathrm{tr}(\boldsymbol{\Sigma}_2)\}/d > 0,\ \liminf_{d\to\infty}\{\Delta_* - b_1\}/d > 0 \tag{2}$$

が成立する．

定理 *1* の証明．(A-i)，(A-ii)，(1) を仮定する．以下の事象を定義する．

$$B = \{\max_{i,j}\|\boldsymbol{x}_{1i} - \boldsymbol{x}_{1j}\|^2 < \min_{i,j}\|\boldsymbol{x}_{1i} - \boldsymbol{x}_{2j}\|^2\},$$
$$C = \{\max_{i,j}\|\boldsymbol{x}_{2i} - \boldsymbol{x}_{2j}\|^2 < \min_{i,j}\|\boldsymbol{x}_{1i} - \boldsymbol{x}_{2j}\|^2\},$$

$$E_1 = \{\max_{i,j}\|\boldsymbol{x}_{1i} - \boldsymbol{x}_{1j}\|^2 < b_1\},$$
$$E_2 = \{\max_{i,j}\|\boldsymbol{x}_{2i} - \boldsymbol{x}_{2j}\|^2 < b_1\},$$
$$E_3 = \{\min_{i,j}\|\boldsymbol{x}_{1i} - \boldsymbol{x}_{2j}\|^2 > b_1\}.$$

このとき，包括関係 $E_1 \cap E_2 \cap E_3 \subset B \cap C$ が成立する．従って，以下を得ることができる．

$$P((B\cap C)^c) \le P(E_1^c) + P(E_2^c) + P(E_3^c).$$

次に，確率 $P(E_1^c)$，$P(E_2^c)$，$P(E_3^c)$ をそれぞれ評価する．チェビシェフの不等式を用いることで，(2) に基づき，

$$\begin{aligned}
P(E_1^c) &\le \sum_{i,j=1(i\neq j)}^{n_1} P(\|\boldsymbol{x}_{1i} - \boldsymbol{x}_{1j}\|^2 - 2\mathrm{tr}(\boldsymbol{\Sigma}_1) > b_1 - 2\mathrm{tr}(\boldsymbol{\Sigma}_1)) \\
&\le \sum_{i,j=1(i\neq j)}^{n_1} P(|\|\boldsymbol{x}_{1i} - \boldsymbol{x}_{1j}\|^2 - 2\mathrm{tr}(\boldsymbol{\Sigma}_1)| > b_1 - 2\mathrm{tr}(\boldsymbol{\Sigma}_1)) \\
&\le \sum_{i,j=1(i\neq j)}^{n_1} \frac{\mathrm{Var}[\|\boldsymbol{x}_{1i} - \boldsymbol{x}_{1j}\|^2]}{(b_1 - 2\mathrm{tr}(\boldsymbol{\Sigma}_1))^2} = O\{(L_1 + \mathrm{tr}(\boldsymbol{\Sigma}_1^2))/\Delta^2\} \to 0 \quad (d\to\infty)
\end{aligned}$$

を得る．同様に，$P(E_2^c) \to 0$ $(d\to\infty)$ を得る．ここで，

$$\begin{aligned}
\|\boldsymbol{x}_{1i} - \boldsymbol{x}_{2j}\|^2 =& \|\boldsymbol{x}_{1i} - \boldsymbol{\mu}_1\|^2 + \|\boldsymbol{x}_{2j} - \boldsymbol{\mu}_2\|^2 + \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2 - 2(\boldsymbol{x}_{1i} - \boldsymbol{\mu}_1)^T(\boldsymbol{x}_{2j} - \boldsymbol{\mu}_2) \\
&+ 2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\{\boldsymbol{x}_{1i} - \boldsymbol{\mu}_1 - (\boldsymbol{x}_{2j} - \boldsymbol{\mu}_2)\}
\end{aligned}$$

となり，$E(\|\boldsymbol{x}_{1i} - \boldsymbol{x}_{2j}\|^2) = \Delta_*$ である．さらに，$\mathrm{Var}\{(\boldsymbol{x}_{1i} - \boldsymbol{\mu}_1)^T(\boldsymbol{x}_{2j} - \boldsymbol{\mu}_2)\} = \mathrm{tr}(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2) \le \{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)\mathrm{tr}(\boldsymbol{\Sigma}_2^2)\}^{1/2} = o(\Delta^2)$，$\mathrm{Var}[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\{\boldsymbol{x}_{1i} - \boldsymbol{\mu}_1 - (\boldsymbol{x}_{2j} - \boldsymbol{\mu}_2)\}] = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{\Sigma}_1 +$

$\boldsymbol{\Sigma}_2)(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \leq \Delta\{\mathrm{tr}(\boldsymbol{\Sigma}_1^2)^{1/2}+\mathrm{tr}(\boldsymbol{\Sigma}_2^2)^{1/2}\}$ となることに注意すれば，$\mathrm{Var}(\|\boldsymbol{x}_{1i}-\boldsymbol{x}_{2j}\|^2) = o(\Delta^2)$ を得る．よって，

$$
\begin{aligned}
P(E_3^c) &\leq \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} P(|\|\boldsymbol{x}_{1i}-\boldsymbol{x}_{2j}\|^2 - \Delta_*| > \Delta_* - b_1) \\
&= O\left(\frac{\mathrm{Var}(\|\boldsymbol{x}_{1i}-\boldsymbol{x}_{2j}\|^2)}{(\Delta_*-b_1)^2}\right) \to 0 \ (d\to\infty)
\end{aligned}
$$

となる．以上より，$P((B\cap C)) \to 1 \ \ (d\to\infty)$ を得る．これは，$\boldsymbol{c}_i \in \boldsymbol{X}_i$, $i=1,2$ であれば，ステップ 2 で漸近的に確率 1 で $\boldsymbol{C}_1 = \boldsymbol{X}_1$, $\boldsymbol{C}_2 = \boldsymbol{X}_2$ となることを意味する．

各 i で $\overline{\boldsymbol{x}}_i = \sum_{j=1}^{n_i}\boldsymbol{x}_{ij}/n_i$ とおく．さらに，$\Delta_\star = \Delta + \mathrm{tr}(\boldsymbol{\Sigma}_1)/n_1 + \mathrm{tr}(\boldsymbol{\Sigma}_2)/n_2$ とおく．次に，任意の $\boldsymbol{x}_1 \in \boldsymbol{X}_1$ と $\boldsymbol{x}_2 \in \boldsymbol{X}_2$ に対して，以下の事象を定義する．

$$
\begin{aligned}
\hat{B} &= \{\|\boldsymbol{x}_1-\overline{\boldsymbol{x}}_1\|^2 < \|\boldsymbol{x}_1-\overline{\boldsymbol{x}}_2\|^2\}, \\
\hat{C} &= \{\|\boldsymbol{x}_2-\overline{\boldsymbol{x}}_2\|^2 < \|\boldsymbol{x}_2-\overline{\boldsymbol{x}}_1\|^2\},
\end{aligned}
$$

$$
\begin{aligned}
\hat{E}_1 &= \{\|\boldsymbol{x}_1-\overline{\boldsymbol{x}}_1\|^2 < b_2\}, \\
\hat{E}_2 &= \{\|\boldsymbol{x}_2-\overline{\boldsymbol{x}}_2\|^2 < b_3\}, \\
\hat{E}_3 &= \{\|\boldsymbol{x}_1-\overline{\boldsymbol{x}}_2\|^2 > b_2\}, \\
\hat{E}_4 &= \{\|\boldsymbol{x}_2-\overline{\boldsymbol{x}}_1\|^2 > b_3\}.
\end{aligned}
$$

ただし，$b_2 = (n_1-1)\mathrm{tr}(\boldsymbol{\Sigma}_1)/n_1 + \Delta_\star/2$, $b_3 = (n_2-1)\mathrm{tr}(\boldsymbol{\Sigma}_2)/n_2 + \Delta_\star/2$ とする．このとき，包括関係 $\hat{E}_1\cap\hat{E}_2\cap\hat{E}_3\cap\hat{E}_4 \subset \hat{B}\cap\hat{C}$ が成立する．従って，以下を得ることができる．

$$
P((\hat{B}\cap\hat{C})^c) \leq P(\hat{E}_1^c) + P(\hat{E}_2^c) + P(\hat{E}_3^c) + P(\hat{E}_4^c).
$$

次に，確率 $P(\hat{E}_1^c)$, $P(\hat{E}_2^c)$, $P(\hat{E}_3^c)$, $P(\hat{E}_4^c)$ をそれぞれ評価する．チェビシェフの不等式を用いることで，

$$
\begin{aligned}
P(\hat{E}_1^c) &= P(\|\boldsymbol{x}_1-\overline{\boldsymbol{x}}_1\|^2 - (n_1-1)\mathrm{tr}(\boldsymbol{\Sigma}_1)/n_1 > b_2 - (n_1-1)\mathrm{tr}(\boldsymbol{\Sigma}_1)/n_1) \\
&\leq P(|\|\boldsymbol{x}_1-\overline{\boldsymbol{x}}_1\|^2 - (n_1-1)\mathrm{tr}(\boldsymbol{\Sigma}_1)/n_1| > \Delta_\star/2) \\
&= O\{(L_1 + \mathrm{tr}(\boldsymbol{\Sigma}_1^2))/\Delta^2\} \to 0 \ \ (d\to\infty)
\end{aligned}
$$

を得る．同様に，$P(\hat{E}_2^c) \to 0 \ \ (d\to\infty)$ を得る．ここで，$\mathrm{Var}(\|\boldsymbol{x}_{1i}-\boldsymbol{x}_{2j}\|^2) = o(\Delta^2)$ と同様に $\mathrm{Var}(\|\boldsymbol{x}_1-\overline{\boldsymbol{x}}_2\|^2) = o(\Delta^2)$ となることを注意すれば，

$$
\begin{aligned}
P(\hat{E}_3^c) &\leq P(|\|\boldsymbol{x}_1-\overline{\boldsymbol{x}}_2\|^2 - (\Delta + \mathrm{tr}(\boldsymbol{\Sigma}_1) + \mathrm{tr}(\boldsymbol{\Sigma}_2)/n_2)| > \Delta_\star/2) \\
&= O\left(\frac{\mathrm{Var}(\|\boldsymbol{x}_1-\overline{\boldsymbol{x}}_2\|^2)}{\Delta_\star^2}\right) \to 0 \ \ (d\to\infty)
\end{aligned}
$$

を得る．また，$P(\hat{E}_4^c) \to 0 \ \ (d \to \infty)$ を得る．以上より，$P((\hat{B} \cap \hat{C})) \to 1 \ \ (d \to \infty)$ を得る．これは，$\boldsymbol{c}_i = \overline{\boldsymbol{x}}_i$, $i = 1, 2$ としてもステップ 2 の結果が漸近的に確率 1 で $\boldsymbol{C}_1 = \boldsymbol{X}_1$, $\boldsymbol{C}_2 = \boldsymbol{X}_2$ となることを意味する．よって，定理 1 を示すことができる． □

謝辞

科学研究費補助金 基盤研究 (A) 20H00576 研究代表者: 青嶋 誠「大規模複雑データの理論と方法論の革新的展開」，学術研究助成基金助成金 挑戦的研究 (萌芽) 22K19769 研究代表者: 青嶋 誠「テンソル構造をもつ巨大データの統計的圧縮技術の開発」，および，科学研究費補助金 基盤研究 (C) 22K03412 研究代表者: 矢田 和善「非線形特徴量に基づく新たな高次元統計理論の開発とその応用」から研究助成を受けています．また，京都大学数理解析研究所の国際共同利用・共同研究拠点事業により研究助成を受けています．

参考文献

[1] 青嶋 誠，矢田和善 (2019). 高次元の統計学. 共立出版.

[2] Ahn, J., Lee, M.H., Yoon, Y.J. (2012). Clustering high dimension, low sample size data using the maximal data piling distance. *Statistica Sinica* **22**. 443–464.

[3] Borysov, P., Hannig, J., Marron, J.S. (2014). Asymptotics of hierarchical clustering for growing dimension. *Journal of Multivariate Analysis* **124**. 465–479.

[4] Egashira, K., Yata, K., Aoshima, M. (2023). Asymptotic properties of hierarchical clustering under high dimensional settings, submitted.

[5] Huang, H., Liu, Y., Yuan, M., Marron, J.S. (2015). Statistical Significance of Clustering using Soft Thresholding. *Journal of Computational and Graphical Statistics* **24**. 975–993.

[6] Liu, Y., Hayes, D.N., Nobel, A., Marron, J.S. (2008). Statistical significance of clustering for high-dimension, low-sample size data. *Journal of the American Statistical Association* **103**. 1281–1293.

[7] Nakayama, Y., Yata, K., Aoshima, M. (2021). Clustering by principal component analysis with Gaussian kernel in high-dimension, low-sample-size settings. *Journal of Multivariate Analysis* **185**. 104779.

[8] Yata, K., Aoshima, M. (2010). Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix. *Journal of Multivariate*

Analysis, **101**, 2060–2077.

[9] Yata, K., Aoshima, M. (2020). Geometric consistency of principal component scores for high-dimensional mixture models and its application. *Scandinavian Journal of Statistics* **47**. 899–921.