多項式カーネルを用いたカーネル k-means の高次元漸近的性質について

東京理科大学・情報計算科学科　　江頭　健斗（Kento Egashira）
Department of 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，高次元小標本データに対するカーネル k-means を考える．従来の k-means を含むカーネル k-means は，その使用の簡易さから，遺伝子発現量データをはじめとする高次元小標本の解析に広く用いられている．高次元データに対するクラスタリングについて，これまでに多くの研究が行われてきた．Ahn et al. [1] は，高次元における階層的クラスタリング手法を提案した．Yata and Aoshima [11, 12] は，高次元における主成分スコアの一致性を示し，それをクラスタリングに応用した．Liu et al. [8] は，高次元小標本に対する特徴選択を伴うクラスタリング手法を検証した．Borysov et al. [3] は，さまざまな漸近的条件下における階層的クラスタリングの漸近的性質を示した．それに対し，Egashira et al. [5] は，緩い仮定のもとで階層的クラスタリングの理論的性質の証明を与え，多クラス設定における性質も導出した．高次元設定における従来の k-means アルゴリズムの性質については，Egashira et al. [6] により求められてる．さらに，カーネル k-means に関しても，ガウシアンカーネル関数を用いた場合には従来の k-means と同様の漸近的性質を有することが，Egashira et al. [6] によって示されている．一方で，例えば，Nakayama et al. [9] は，異なるカーネル関数を用いたカーネルサポートベクターマシンの高次元における性能を求め，多

項式カーネル関数はガウシアンカーネル関数とは異なる挙動を示すことを明らかにした．本論文では，多項式カーネル関数を用いたカーネル k-means の高次元における理論的性質を明らかにし，従来の k-means と比較を行う．

本論文の構成は以下の通りである．第 2 節では，カーネル k-means アルゴリズムを導入し，理論的性質を述べるための設定を導入する．第 3 節では，k-means アルゴリズムの理論的性質を導入する．第 4 節では，多項式カーネル関数を用いた場合のカーネル k-means の理論的性質を導出し，k-means アルゴリズムの漸近的性質と比較を行う．第 5 節では，数値実験により，従来の k-means およびカーネル k-means の挙動が，導出された理論結果と整合するかを検証する．数学的証明は付録に記載している．

2　カーネル k-means アルゴリズム

データセット $\boldsymbol{X}$ と事前に設定したクラスター数 K が与えられたとき，カーネル関数 $k(\boldsymbol{z}_1, \boldsymbol{z}_2)$ を用いたときのカーネル k-means アルゴリズムは，以下で与えられる．

カーネル k-means アルゴリズム

ステップ 1: K 個の初期値を $\boldsymbol{c}_{0i}$ $(\in \boldsymbol{X})$, $i \in \{1, \ldots, K\}$ として設定する．

ステップ 2: 各データ $\boldsymbol{x} \in \boldsymbol{X}$ を，$i = \underset{i' \in \{1,\ldots,K\}}{\operatorname{argmin}} \left(k(\boldsymbol{x}, \boldsymbol{x}) + k(\boldsymbol{c}_{0i'}, \boldsymbol{c}_{0i'}) - 2k(\boldsymbol{x}, \boldsymbol{c}_{0i'}) \right)$ を満たす $\boldsymbol{C}_i$ に割り当てる．この操作をすべてのデータに対して行い，集合 $\boldsymbol{C}_i$, $i \in \{1, \ldots, K\}$ を構成する．

ステップ 3: 各データ点 $\boldsymbol{x} \in \boldsymbol{X}$ を，

$$i = \underset{i' \in \{1,\ldots,K\}}{\operatorname{argmin}} \left(k(\boldsymbol{x}, \boldsymbol{x}) + \sum_{\boldsymbol{c}_{i'1}, \boldsymbol{c}_{i'2} \in \boldsymbol{C}_{i'}} \frac{k(\boldsymbol{c}_{i'1}, \boldsymbol{c}_{i'2})}{\#(\boldsymbol{C}_{i'})^2} - \sum_{\boldsymbol{c}_{i'1} \in \boldsymbol{C}_{i'}} \frac{2k(\boldsymbol{x}, \boldsymbol{c}_{i'1})}{\#(\boldsymbol{C}_{i'})} \right)$$

を満たす $\boldsymbol{C}_i$ に割り当てる．この操作を繰り返し，集合 $\boldsymbol{C}_i$, $i \in \{1, \ldots, K\}$ を更新する．ここで，$\#(A)$ は集合 A の基数とする．この操作を，更新前後の集合が一致するまで繰り返す．

ステップ 4: ステップ 3 において収束した集合を $\widehat{\boldsymbol{C}}_i$, $i \in \{1, \ldots, K\}$ と定義し，これをカーネル k-means アルゴリズムの結果とする．

線形カーネル関数 $k_L(\boldsymbol{z}_1, \boldsymbol{z}_2) = \boldsymbol{z}_1^\top \boldsymbol{z}_2$ を使用したカーネル k-means アルゴリズムは，従来の k-means アルゴリズムに一致する．カーネル法は広く応用されているが，Nakayama et al. [10] は，高次元小標本を対象の上，主成分分析を用いたクラスタリングに適用し，ガウ

シアンカーネル関数を用いたカーネル法が標本間の不均衡性を利用する上で有効であることを示している．しかし，Egashira et al. [6] は，ガウシアンカーネル関数を使用したカーネル k-means アルゴリズムにおいて，その漸近的性質が従来の k-means アルゴリズムと変わらないことを明らかにし，カーネル関数ごとにカーネル k-means が従来の k-means を上回るかどうかを確認する必要があることを示している．

本論文では，独立な d 次元の母集団が 2 個あると考え，各母集団 $\pi_i, i \in \{1,2\}$ 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．また，一般性を失うことなく，$\mathrm{tr}(\boldsymbol{\Sigma}_1) \leq \mathrm{tr}(\boldsymbol{\Sigma}_2)$ とする．$i \in \{1,2\}$ に対して，

$$\limsup_{d\to\infty} \|\boldsymbol{\mu}_i\|^2/d < \infty, \quad \liminf_{d\to\infty} \mathrm{tr}(\boldsymbol{\Sigma}_i)/d > 0, \quad \limsup_{d\to\infty} \mathrm{tr}(\boldsymbol{\Sigma}_i)/d < \infty \tag{1}$$

を仮定する．各 $i \in \{1,2\}$ について，母集団 π_i からの独立な n_i 個の観測値 $\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}$ が得られているとする．$i \in \{1,2\}$ に対して $\boldsymbol{X}_i = \{\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{in_i}\}$ とし，$\Delta = \|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|^2$ および $\Delta_\Sigma = |\mathrm{tr}(\boldsymbol{\Sigma}_1) - \mathrm{tr}(\boldsymbol{\Sigma}_2)|$ と定義する．さらに，$i \in \{1,2\}$ に対して，$L_i = \mathrm{Var}[\|\boldsymbol{x}_{ij} - \boldsymbol{\mu}_i\|^2]$ と定義する．本研究では，k-means クラスタリングの漸近的性質を検証するために，クラスタ数 K は 2 として事前に定められていると仮定する．次の節では，従来の k-means アルゴリズムの高次元における漸近的性質を導入する．

3　k-means アルゴリズムの漸近的性質

本節では，Egashira et al. [6] による k-means の高次元における漸近的性質を，2 母集団設定のもとで導入する．また，次元 d が無限大に発散し，標本数 N は固定されている漸近的枠組みを考える．以下の仮定を導入する．

(A-i): $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/\Delta^2 \to 0$ as $d \to \infty$, $i \in \{1,2\}$
(A-ii): $L_i/\Delta^2 \to 0$ as $d \to \infty$, $i \in \{1,2\}$.

母集団に正規分布を仮定したとき，$L_i = 2\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ が成立し，(A-i) と (A-ii) が同値であることに注意する．これらの仮定は，Aoshima and Yata [2] や Egashira et al. [4, 5, 6] においても用いられている．Egashira et al. [6] により，k-means アルゴリズムについて，以下の漸近的性質が示されている．

定理 1. $K = 2$ のもと，$i \in \{1,2\}$ について，初期値が $\boldsymbol{c}_{0i} \in \boldsymbol{X}_i$ と選ばれたとし，*(A-i)* と

(A-ii) を仮定する．

$$\limsup_{d\to\infty} \Delta_\Sigma/\Delta < 1 \tag{2}$$

であるとき，*k-means* アルゴリズムにおいて，$d \to \infty$ のもとで，

$$P\big(\widehat{\boldsymbol{C}}_1 = \boldsymbol{X}_1, \widehat{\boldsymbol{C}}_2 = \boldsymbol{X}_2\big) \to 1 \tag{3}$$

が成立する．

真のクラスタ数が 3 以上であるときの漸近的性質は，Egashira et al. [6] を参照のこと．次の節では，多項式カーネル関数を用いたカーネル k-means アルゴリズムの高次元における漸近的性質を示す．

4　多項式カーネルを用いたときの漸近的性質

多項式カーネル関数を，$k_P(\boldsymbol{z}_1, \boldsymbol{z}_2) = (\xi + \boldsymbol{z}_1^\top \boldsymbol{z}_2)^r$ として定義する．ここで，$\xi \geq 0$ 及び，$r \in \mathbb{N}$ はハイパーパラメータである．本節では，多項式カーネル関数を使用した場合の 2 母集団設定におけるカーネル k-means の理論的性質について検討する．特に，ハイパーパラメータ $\xi \geq 0$ が，

(C-i):　$\xi/d \to 0$ as $d \to \infty$.

を満たす場合を考え，以下の仮定

(C-ii):　$\mathrm{tr}(\boldsymbol{\Sigma}_i^2)/d^2 \to 0$, $L_i/d^2 \to 0$　　as $d \to \infty$, $i \in \{1, 2\}$
(C-iii):　$\liminf_{d\to\infty} \left(\|\boldsymbol{\mu}_1\|^{2r} + \|\boldsymbol{\mu}_2\|^{2r} - 2(\boldsymbol{\mu}_1^\top \boldsymbol{\mu}_2)^r\right)/d^r > 0$

を導入する．$d \to \infty$ において，$\xi/d \to (0, \infty)$ に対応する場合は，Egashira et al. [7] を参照のこと．

$$\liminf_{d\to\infty} \left| \|\boldsymbol{\mu}_1\|^2 - \|\boldsymbol{\mu}_2\|^2 \right| /d > 0 \tag{4}$$

であれば (C-iii) は成立する．(4) が成立しない場合でも，例えば，(C-i) のもとで $\|\boldsymbol{\mu}_1\|^2 = \|\boldsymbol{\mu}_2\|^2 = O(d)$ かつ $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2$ であり，r を奇数として選べば，(C-iii) は依然として成立する．

命題 1. N が固定のもとで，*(C-i)* から *(C-iii)* を仮定したとき，多項式カーネル関数について次の結果が得られる．

$$
\begin{aligned}
&k_P(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'}) = \kappa_i + o_P(\Delta_k), \quad \textit{for all } i, j, j' \ (j \neq j'), \\
&k_P(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij}) = \kappa_i^* + o_P(\Delta_k), \quad \textit{for all } i, j, \\
&k_P(\boldsymbol{x}_{ij}, \boldsymbol{x}_{i'j'}) = \kappa_3 + o_P(\Delta_k), \quad \textit{for all } i, i', j, j' \ (i \neq i').
\end{aligned}
$$

ここで，$\kappa_i = \|\boldsymbol{\mu}_i\|^{2r}$, $\kappa_i^* = (\|\boldsymbol{\mu}_i\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_i))^r$, $i \in \{1, 2\}$, $\kappa_3 = (\boldsymbol{\mu}_1^\top \boldsymbol{\mu}_2)^r$, $\Delta_k = \kappa_1 + \kappa_2 - 2\kappa_3$ とする．

命題 1 より，(C-i) から (C-iii) のもと，$\boldsymbol{c}_{0i} \in \boldsymbol{X}_i$, $i \in \{1, 2\}$ と $\boldsymbol{x}_{ij} \neq \boldsymbol{c}_{0i}$ について，

$$
\begin{aligned}
&\frac{\{k_P(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij}) + k_P(\boldsymbol{c}_{0i'}, \boldsymbol{c}_{0i'}) - 2k_P(\boldsymbol{x}_{ij}, \boldsymbol{c}_{0i'})\} - \{k_P(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij}) + k_P(\boldsymbol{c}_{0i}, \boldsymbol{c}_{0i}) - 2k_P(\boldsymbol{x}_{ij}, \boldsymbol{c}_{0i})\}}{\Delta_k} \\
&\approx \frac{(\kappa_{i'}^* - \kappa_{i'}) - (\kappa_i^* - \kappa_i)}{\Delta_k} + 1 \quad \text{for all } i, i', j \ (i \neq i')
\end{aligned}
$$

が成立する．従って，$|(\kappa_1^* - \kappa_1) - (\kappa_2^* - \kappa_2)|/\Delta_k < 1$ であれば、カーネル k-means アルゴリズムのステップ 2 において，$K = 2$ の場合に $\boldsymbol{x}_{ij}$ を真の集合に分類することができる．多項式カーネル関数を用いたカーネル k-means アルゴリズムに関し，次の結果が得られる．

定理 2. $K = 2$ のもと，$i \in \{1, 2\}$ について，初期値が $\boldsymbol{c}_{0i} \in \boldsymbol{X}_i$ と選ばれたとし，*(C-i)* から *(C-iii)* を仮定する．

$$
\limsup_{d \to \infty} \frac{|(\kappa_1^* - \kappa_1) - (\kappa_2^* - \kappa_2)|}{\Delta_k} < 1 \tag{5}
$$

であるとき，多項式カーネルを用いたときのカーネル *k-means* アルゴリズムにおいて，$d \to \infty$ のもとで，*(3)* が成立する．

多項式カーネル関数を使用したカーネル k-means が，漸近的な観点から従来の k-means を上回るかを調査することが目的であった．その結果，2 母集団設定において，(5) が成立している場合でも (2) が成立しないことがあり，またその逆の場合も存在することが確認できた．数値実験を通じて，実際に多項式カーネル関数を使用したカーネル k-means が，従来の k-means よりも優れた性能を示す状況，および，その逆の状況を確認する．

5 数値実験

本節では，高次元小標本のもとで，多項式カーネル関数を用いたときのカーネル k-means アルゴリズムを数値的に検証する．母集団として正規分布を仮定する．ここで，全ての成分

が 0 である d 次元ベクトルを $\mathbf{0}_d$，全ての成分が 1 である d 次元ベクトルを $\mathbf{1}_d$ とする．次の 3 つの設定を考える．

(a) : $\boldsymbol{\mu}_1 = \mathbf{1}_d$，$\boldsymbol{\mu}_2 = 7/21_d$，$\boldsymbol{\Sigma}_1 = 11\boldsymbol{\Phi}$, $\boldsymbol{\Sigma}_2 = \boldsymbol{\Phi}$, $(n_1, n_2) = (10, 10)$;

(b) : $\boldsymbol{\mu}_1 = \mathbf{0}_d$，$\boldsymbol{\mu}_2 = \mathbf{1}_d$，$\boldsymbol{\Sigma}_1 = 5/3\boldsymbol{\Phi}$, $\boldsymbol{\Sigma}_2 = \boldsymbol{\Phi}$, $(n_1, n_2) = (10, 10)$;

(c) : $\boldsymbol{\mu}_1 = 11/101_d$，$\boldsymbol{\mu}_2 = -\mathbf{1}_d$，$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Phi}$, $(n_1, n_2) = (10, 10)$;

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．このとき，$\mathrm{tr}(\boldsymbol{\Phi}) = d$ が成立する．次元を $d = 2^s$，$s = 5, ..., 12$ と設定する．

各設定のもとでデータを発生させ，従来の k-means アルゴリズムと多項式カーネル関数を使用したカーネル k-means アルゴリズムを実行した．初期値は各クラスから選択をした．実験を 2000 回繰り返し，正しいクラスタリング結果を得られているかを確認し，その誤った割合を纏めたのが図 1 である．ここで，従来の k-means アルゴリズムを “k-means”，$r = 2$ とした多項式カーネル関数を用いた k-means アルゴリズムを “Polynomial 1”，同様に，$r = 3$ を “Polynomial 2”，$r = 4$ を “Polynomial 3”と表すことにする．ただし，多項式カーネル関数について，(C-i) を満たすために一貫して $\xi = 0$ とする．

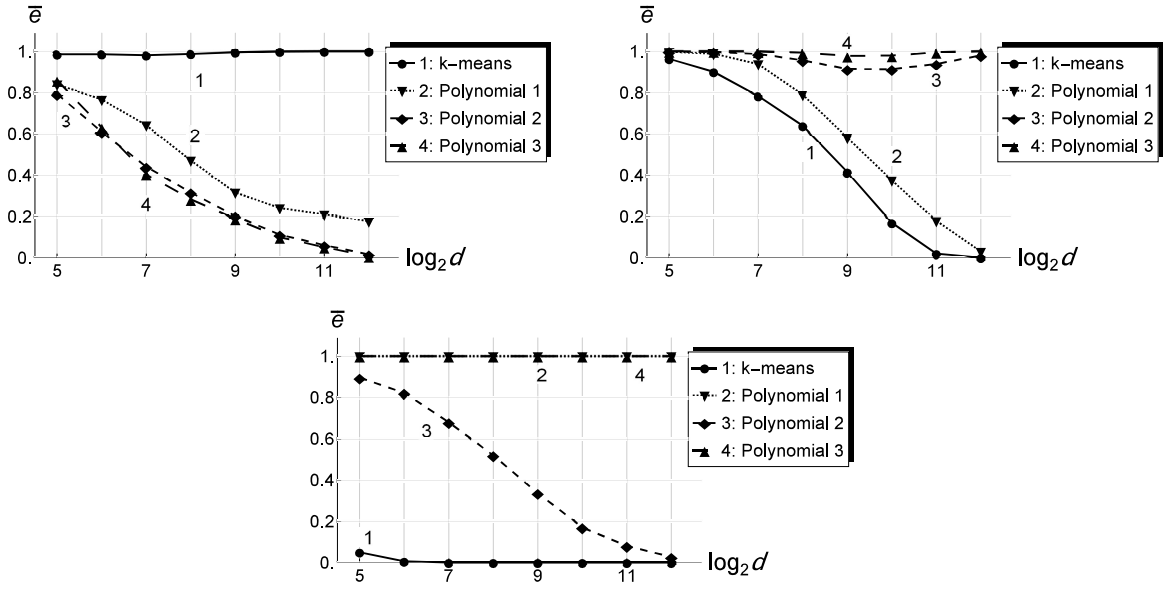

図 1　設定 (a), (b), (c) における “k-means”, “Polynomial 1”, “Polynomial 2”, “Polynomial 3”によるクラスタリングの誤り確率を表している．上段左は (a)，上段右は (b)，下段は (c) に対応している．

(a) において，(2) は成立しないが，$r = 2, 3, 4$ としたときの (5) は成立するように設定されている．実際に，“k-means”は誤り確率が 0 に収束せず，“Polynomial 1”，“Polynomial 2”，“Polynomial 3”の誤り確率は 0 に収束する様子を確認することができる．(2) は，(b) と (c) の両方の場合で成立する．(b) において，(5) は，$r = 2$ としたとき成立し，$r = 3, 4$ のときは成立しないように設定されている．定理 1 で示されている通り，“k-means”と“Polynomial 1”が良い性能を示している．(c) において，(5) は，$r = 3$ としたとき成立し，$r = 2, 4$ で偶数のときは成立しないように設定されている．“k-means”と比較すると収束速度に差はあるが，(5) を満たす “Polynomial 2”の誤り確率の 0 への収束を確認することができ，多項式カーネル関数の次数の選び方に依存して性能の振舞いが変わる例を表している．

6 付録

命題 *1* の証明. $d \to \infty$ としたとき，(1) と (C-ii) のもとで，$\boldsymbol{\mu}_i^\top \boldsymbol{\Sigma}_i \boldsymbol{\mu}_i \leq \|\boldsymbol{\mu}_i\|^2 \sqrt{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} = o(d^2)$ が成立することに注意する．(1) と (C-ii) を仮定すると，

$$
\begin{aligned}
\boldsymbol{x}_{ij}^\top \boldsymbol{x}_{ij'} =& \|\boldsymbol{\mu}_i\|^2 + o_P(d) \text{ for all } i, j, j' \ (j \neq j'), \\
\|\boldsymbol{x}_{ij}\|^2 =& \|\boldsymbol{\mu}_i\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_i) + o_P(d) \text{ for all } i, j, \\
\boldsymbol{x}_{ij}^\top \boldsymbol{x}_{i'j'} =& \boldsymbol{\mu}_i^\top \boldsymbol{\mu}_{i'} + o_P(d) \text{ for all } i, i', j, j' \ (i \neq i').
\end{aligned}
$$

が成立する．従って，(C-i) と (C-ii) のもとで，

$$
\begin{aligned}
&k_P(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij'}) = \kappa_i + o_P(d^r) \ \text{ for all } i, j, j' \ (j \neq j'), \\
&k_P(\boldsymbol{x}_{ij}, \boldsymbol{x}_{ij}) = \kappa_i^* + o_P(d^r) \ \text{ for all } i, j, \\
&k_P(\boldsymbol{x}_{ij}, \boldsymbol{x}_{i'j'}) = \kappa_3 + o_P(d^r) \ \text{ for all } i, i', j, j' \ (i \neq i').
\end{aligned}
$$

を得ることができ，(C-iii) のもとで，$\liminf_{d\to\infty} \Delta_k / d^r > 0$ であることから，命題 1 を示すことができる． □

定理 *2* の証明. 以下の事象を定義する．

$$
\begin{aligned}
A_K = \Big\{ &\max_{j,j'} \{k_P(\boldsymbol{x}_{1j}, \boldsymbol{x}_{1j}) - 2k_P(\boldsymbol{x}_{1j}, \boldsymbol{x}_{1j'}) + k_P(\boldsymbol{x}_{1j'}, \boldsymbol{x}_{1j'})\} \\
&< \min_{j,j'} \{k_P(\boldsymbol{x}_{1j}, \boldsymbol{x}_{1j}) - 2k_P(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j'}) + k_P(\boldsymbol{x}_{2j'}, \boldsymbol{x}_{2j'})\} \Big\}, \\
B_K = \Big\{ &\max_{j,j'} \{k_P(\boldsymbol{x}_{2j}, \boldsymbol{x}_{2j}) - 2k_P(\boldsymbol{x}_{2j}, \boldsymbol{x}_{2j'}) + k_P(\boldsymbol{x}_{2j'}, \boldsymbol{x}_{2j'})\} \\
&< \min_{j,j'} \{k_P(\boldsymbol{x}_{1j}, \boldsymbol{x}_{1j}) - 2k_P(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j'}) + k_P(\boldsymbol{x}_{2j'}, \boldsymbol{x}_{2j'})\} \Big\},
\end{aligned}
$$

$$E_{K1} = \Big\{ \max_{j,j'} \{k_P(\boldsymbol{x}_{1j}, \boldsymbol{x}_{1j}) - 2k_P(\boldsymbol{x}_{1j}, \boldsymbol{x}_{1j'}) + k_P(\boldsymbol{x}_{1j'}, \boldsymbol{x}_{1j'})\} < t_{K1} \Big\},$$
$$E_{K2} = \Big\{ \max_{j,j'} \{k_P(\boldsymbol{x}_{2j}, \boldsymbol{x}_{2j}) - 2k_P(\boldsymbol{x}_{2j}, \boldsymbol{x}_{2j'}) + k_P(\boldsymbol{x}_{2j'}, \boldsymbol{x}_{2j'})\} < t_{K1} \Big\},$$
$$E_{K3} = \Big\{ \min_{j,j'} \{k_P(\boldsymbol{x}_{1j}, \boldsymbol{x}_{1j}) - 2k_P(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2j'}) + k_P(\boldsymbol{x}_{2j'}, \boldsymbol{x}_{2j'})\} > t_{K1} \Big\},$$

ただし，$t_{K1} = \beta_1 + \beta_2 + (\Delta_k + \Delta_{k,\Sigma})/2$, $\Delta_{k,\Sigma} = |\beta_1 - \beta_2|$, $\beta_i = \kappa_i^* - \kappa_i$, $i \in \{1, 2\}$ とする．また，$\Delta_{k*} = \Delta_k + \beta_1 + \beta_2$ とおくと，(5) のもとで，$\limsup_{d\to\infty} \Delta_{k,\Sigma}/\Delta_k < 1$ が成立する．従って，(5) のもとで，

$$\liminf_{d\to\infty} \{t_{K1} - 2\beta_1\}/\Delta_k > 0, \quad \liminf_{d\to\infty} \{t_{K1} - 2\beta_2\}/\Delta_k > 0,$$
$$\liminf_{d\to\infty} \{\Delta_{k*} - t_{K1}\}/\Delta_k > 0$$

が成立し，$E_{K1} \cap E_{K2} \cap E_{K3} \subset A_K \cap B_K$ が示される．命題 1 より，(5) が成立するとき，N が固定で $d \to \infty$ のもと，(C-i) から (C-iii) を仮定すると，$\Pr(E_{Ki}) \to 1$, $i \in \{1, 2, 3\}$, as $d \to \infty$ が成立する．これは，$i \in \{1, 2\}$ について $\boldsymbol{c}_{0i} \in \boldsymbol{X}_i$ であるとき，ステップ 2 において，漸近的に確率 1 で $\boldsymbol{C}_i = \boldsymbol{X}_i$, $i \in \{1, 2\}$ が成立することを意味している．次に，以下の事象を定義する．

$$\widehat{A}_K = \Bigg\{ \max_j \Big\{ k_P(\boldsymbol{x}_{1j}, \boldsymbol{x}_{1j}) - \frac{2}{n_1} \sum_{l=1}^{n_1} k_P(\boldsymbol{x}_{1j}, \boldsymbol{x}_{1l}) + \frac{1}{n_1^2} \sum_{l,l'=1}^{n_1} k_P(\boldsymbol{x}_{1l}, \boldsymbol{x}_{1l'}) \Big\}$$
$$< \min_j \Big\{ k_P(\boldsymbol{x}_{1j}, \boldsymbol{x}_{1j}) - \frac{2}{n_2} \sum_{l=1}^{n_2} k_P(\boldsymbol{x}_{1j}, \boldsymbol{x}_{2l}) + \frac{1}{n_2^2} \sum_{l,l'=1}^{n_2} k_P(\boldsymbol{x}_{2l}, \boldsymbol{x}_{2l'}) \Big\} \Bigg\},$$
$$\widehat{B}_K = \Bigg\{ \max_j \Big\{ k_P(\boldsymbol{x}_{2j}, \boldsymbol{x}_{2j}) - \frac{2}{n_2} \sum_{l=1}^{n_2} k_P(\boldsymbol{x}_{2j}, \boldsymbol{x}_{2l}) + \frac{1}{n_2^2} \sum_{l,l'=1}^{n_2} k_P(\boldsymbol{x}_{2l}, \boldsymbol{x}_{2l'}) \Big\}$$
$$< \min_j \Big\{ k_P(\boldsymbol{x}_{2j}, \boldsymbol{x}_{2j}) - \frac{2}{n_1} \sum_{l=1}^{n_1} k_P(\boldsymbol{x}_{2j}, \boldsymbol{x}_{1l}) + \frac{1}{n_1^2} \sum_{l,l'=1}^{n_1} k_P(\boldsymbol{x}_{1l}, \boldsymbol{x}_{1l'}) \Big\} \Bigg\},$$

$$\widehat{E}_{K1}=\left\{ \max_j \left\{ k_P(\boldsymbol{x}_{1j},\boldsymbol{x}_{1j}) - \frac{2}{n_1}\sum_{l=1}^{n_1} k_P(\boldsymbol{x}_{1j},\boldsymbol{x}_{1l}) + \frac{1}{n_1^2}\sum_{l,l'=1}^{n_1} k_P(\boldsymbol{x}_{1l},\boldsymbol{x}_{1l'}) \right\} < t_{K2} \right\},$$

$$\widehat{E}_{K2}=\left\{ \max_j \left\{ k_P(\boldsymbol{x}_{2j},\boldsymbol{x}_{2j}) - \frac{2}{n_2}\sum_{l=1}^{n_2} k_P(\boldsymbol{x}_{2j},\boldsymbol{x}_{2l}) + \frac{1}{n_2^2}\sum_{l,l'=1}^{n_2} k_P(\boldsymbol{x}_{2l},\boldsymbol{x}_{2l'}) \right\} < t_{K3} \right\},$$

$$\widehat{E}_{K3}=\left\{ \min_j \left\{ k_P(\boldsymbol{x}_{1j},\boldsymbol{x}_{1j}) - \frac{2}{n_2}\sum_{l=1}^{n_2} k_P(\boldsymbol{x}_{1j},\boldsymbol{x}_{2l}) + \frac{1}{n_2^2}\sum_{l,l'=1}^{n_2} k_P(\boldsymbol{x}_{2l},\boldsymbol{x}_{2l'}) \right\} > t_{K2} \right\},$$

$$\widehat{E}_{K4}=\left\{ \min_j \left\{ k_P(\boldsymbol{x}_{2j},\boldsymbol{x}_{2j}) - \frac{2}{n_1}\sum_{l=1}^{n_1} k_P(\boldsymbol{x}_{2j},\boldsymbol{x}_{1l}) + \frac{1}{n_1^2}\sum_{l,l'=1}^{n_1} k_P(\boldsymbol{x}_{1l},\boldsymbol{x}_{1l'}) \right\} > t_{K3} \right\},$$

ただし，$t_{K2}=(n_1-1)\beta_1/n_1+\Delta_{k\star}/2$，$t_{K3}=(n_2-1)\beta_2/n_2+\Delta_{k\star}/2$，$\Delta_{k\star}=\Delta_k+\beta_1/n_1+\beta_2/n_2$ とする．(5) のもとで，

$$\liminf_{d\to\infty}\left\{t_{K2}-(n_1-1)\beta_1/n_1\right\}/\Delta_k>0, \liminf_{d\to\infty}\left\{(\Delta_k+\beta_1+\beta_2/n_2)-t_{K2}\right\}/\Delta_k>0,$$
$$\liminf_{d\to\infty}\left\{t_{K3}-(n_2-1)\beta_2/n_2\right\}/\Delta_k>0, \liminf_{d\to\infty}\left\{(\Delta_k+\beta_1/n_1+\beta_2)-t_{K3}\right\}/\Delta_k>0$$

が成立することに注意すると，$\widehat{E}_{K1}\cap\widehat{E}_{K2}\cap\widehat{E}_{K3}\cap\widehat{E}_{K4}\subset\widehat{A}_K\cap\widehat{B}_K$ が示される．命題 1 より，(5) が成立するとき，N が固定で $d\to\infty$ のもと，(C-i) から (C-iii) を仮定すると，$\Pr(\widehat{E}_{Ki})\to 1$, $i\in\{1,\ldots,4\}$ が成立する．これは，$i\in\{1,2\}$ について $\boldsymbol{c}_{0i}\in\boldsymbol{X}_i$ であるとき，ステップ 3 において，漸近的に確率 1 で $\boldsymbol{C}_i=\boldsymbol{X}_i$, $i\in\{1,2\}$ が成立することを意味している．従って，定理 2 を示すことができた． □

謝辞

本研究は，科学研究費 若手研究 24K20748 研究代表者: 江頭　健斗「高次元小標本におけるクラスタリング手法とカーネル法の有効性に関する理論と応用」，科学研究費 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」科学研究費 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」，科学研究費 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，京都大学数理解析研究所の国際共同利用・共同研究拠点事業より研究助成を受けています．

参考文献

[1] 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.

[2] Aoshima, M., Yata, K. (2014) "A distance-based, misclassification rate adjusted classifier for multiclass, high-dimensional data." *Annals of the Institute of Statistical Mathematics*, **66**, 983–1010.

[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. (2021) "Asymptotic properties of distance weighted discrimination and its bias correction for high-dimension, low-sample-size data." *Japanese Journal of Statistics and Data Science*, **4**, 821–840.

[5] Egashira, K., Yata, K., Aoshima, M. (2023) "Asymptotic properties of hierarchical clustering in high-dimensional settings." *Journal of Multivariate Analysis*, **199**, 105251.

[6] Egashira, K., Yata, K., Aoshima, M. "Asymptotic properties of k-means and its bias correction for high-dimension, low-sample-size data," submitted.

[7] Egashira, K., Yata, K., Aoshima, M. "Asymptotic properties of kernel k-means and its comparison with conventional k-means in high-dimensional settings," submitted.

[8] Liu, T., Lu, Y., Zhu, B., Zhao, H. (2023) "Clustering high-dimensional data via feature selection." *Biometrics*, **79**, 940–950.

[9] Nakayama, Y., Yata, K., Aoshima, M. (2020) "Bias-corrected support vector machine with Gaussian kernel in high-dimension, low-sample-size settings." *Annals of the Institute of Statistical Mathematics*, **72**, 1257–1286.

[10] 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.

[11] 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.

[12] Yata, K., Aoshima, M. (2025) "Automatic sparse PCA for high-dimensional data." *Statistica Sinica*, **35**, 1069-1090.