Ann Inst Stat Math (2015) 67:335–357
DOI 10.1007/s10463-014-0454-0

Strong consistency of factorial *K*-means clustering

Yoshikazu Terada

Received: 26 June 2013 / Revised: 12 January 2014 / Published online: 27 March 2014

Abstract Factorial k-means (FKM) clustering is a method for clustering objects in a low-dimensional subspace. The advantage of this method is that the partition of objects and the low-dimensional subspace reflecting the cluster structure are obtained, simultaneously. In some cases that reduced k-means (RKM) clustering does not work well, FKM clustering can discover the cluster structure underlying a lower dimensional subspace. Conditions that ensure the almost sure convergence of the estimator of FKM clustering as the sample size increases unboundedly are derived. The result is proved for a more general model including FKM clustering. Moreover, it is also shown that there exist some cases in which RKM clustering becomes equivalent to FKM clustering as the sample size goes to infinity.

Keywords Subspace clustering · K-means

1 Introduction

If we apply a cluster analysis to data, it is highly unlikely that all variables relate to the same cluster structure. Hence, it is sometimes beneficial to regard the true cluster structure of interest as lying in a low-dimensional subspace of the data. In these cases, researchers often apply the following two-step procedure:

Step 1. Carry out principal component analysis (PCA) and obtain the first few components.
Step 2. Perform k-means clustering for the principal scores on the first few principal components, which are obtained in Step 1.

Y. Terada (✉)
Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka 560-0043, Japan
e-mail: terada@sigmath.es.osaka-u.ac.jp

This procedure is called "tandem clustering" by Arabie and Hubert (1994). Several authors warn against the use of tandem clustering (e.g., Arabie and Hubert (1994); Chang (1983); De Soete and Carroll (1994)). The first few principal components of PCA do not necessarily reflect the cluster structure in data. Thus, an appropriate clustering result might not be obtained using this procedure.

Instead of a two-step procedure, such as tandem clustering, some methods that perform cluster analysis and dimension reduction simultaneously have been proposed (e.g., De Soete and Carroll (1994); Vichi and Kiers (2001)). De Soete and Carroll (1994) proposed reduced k-means (RKM) clustering, which includes conventional k-means clustering as a special case. For given data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in $\mathbb{R}^p$, the fixed cluster number k and the dimension number of subspace q ($q < \min\{k-1, p\}$), the objective function of RKM clustering is defined by

$$RKM_n(F, A) := \frac{1}{n}\sum_{i=1}^{n} \min_{1\le j\le k} \|\boldsymbol{x}_i - A\boldsymbol{f}_j\|^2,$$

where $\boldsymbol{f}_j \in \mathbb{R}^q$, $F = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_k\} \subset \mathbb{R}^q$, A is a $p \times q$ column-wise orthonormal matrix, and $\|\cdot\|$ represents the Euclidean norm. Under certain regularity conditions, RKM clustering has strong consistency (Terada 2014). However, when the data have more variability in directions orthogonal to the subspace containing the cluster structure, RKM clustering may fail to find a subspace that reflects the cluster structure.

Example 1 Let $\boldsymbol{\mu}_1 = (4, 4, 0, \ldots, 0)$, $\boldsymbol{\mu}_2 = (4, -4, 0, \ldots, 0)$, $\boldsymbol{\mu}_3 = (-4, 4, 0, \ldots, 0)$ and $\boldsymbol{\mu}_4 = (-4, -4, 0, \ldots, 0)$ ($\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_4 \in \mathbb{R}^p$). Let Σ_p denote the $p \times p$ diagonal matrix with the elements $(1, 1, 20, \ldots, 20)$ on the diagonal. Observations $\boldsymbol{X}_i = [X_{i1}, \ldots, X_{ip}]^T$ ($i = 1, \ldots, n$) are generated as

$$\boldsymbol{X}_i := \sum_{k=1}^{4} u_{ik}\boldsymbol{\mu}_k + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{u}_i = (u_{i1}, \ldots, u_{i4})$ and $\boldsymbol{\epsilon}_i$ ($i = 1, \ldots, n$) are independently generated from the multinomial distribution for four trials with equal probabilities and the p-dimensional normal distribution $N_p(\boldsymbol{0}, \Sigma_p)$, respectively. In this setting, $\mathrm{Var}[X_{is}] = 17$ ($s = 1, 2$) and $\mathrm{Var}[X_{it}] = 20$ ($t = 3, \ldots, p$). Let $X = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]^T$. Then, the data matrix X has more variability in directions orthogonal to the subspace containing the cluster structure. Here, we set $n = 200$ and $p = 12$. RKM clustering has been applied to the data matrix X (Fig. 1). The result of RKM clustering for the data shown in Fig. 1 is given in Fig. 2. This result indicates that the low-dimensional subspace of RKM clustering does not reflect the actual cluster structure and that the clustering result is, in fact, incorrect.

Vichi and Kiers (2001) pointed out the possibility of such problems with RKM clustering and proposed a new clustering method, called factorial k-means (FKM) clustering. For the given data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in $\mathbb{R}^p$, the number of clusters k, and the number of dimensions of subspace q, FKM clustering is defined by the minimization of the following loss function:

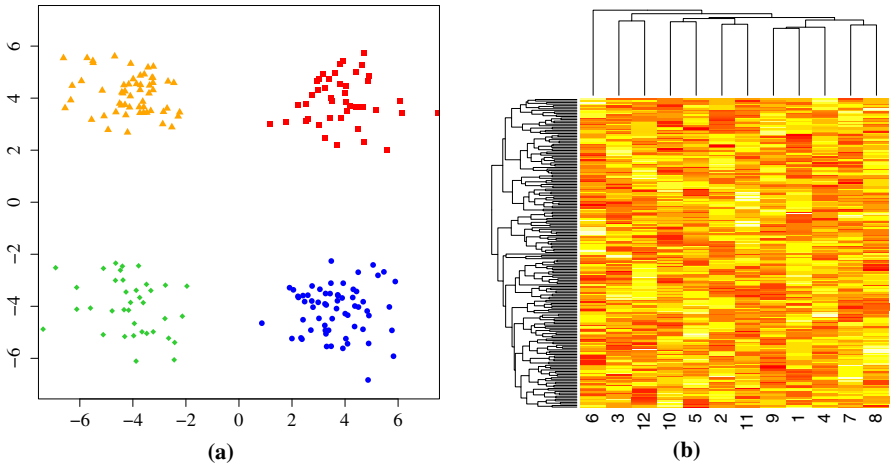


Fig. 1 Artificial data used to evaluate RKM clustering: **a** plot of the first two variables of the data matrix X and **b** heat map of X

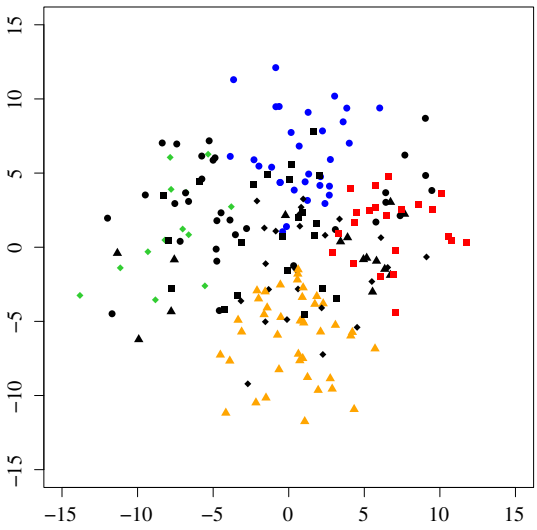


Fig. 2 Plot of the result of RKM clustering for the artificial data given in Fig. 1, where the *black points* represent misclassified objects

$$FKM_n(F, A \mid k, q) := \frac{1}{n} \sum_{i=1}^{n} \min_{1 \le j \le k} \left\| A^T \boldsymbol{x}_i - \boldsymbol{f}_j \right\|^2,$$

where $F := \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_k\}$, $\boldsymbol{f}_j \in \mathbb{R}^q$ and A is a $p \times q$ column-wise orthonormal matrix. When the given data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are independently drawn from a population distribution P, we can rewrite the FKM objective function as

$$FKM(F, A, P_n) := \int \min_{\boldsymbol{f} \in F} \left\| A^T \boldsymbol{x} - \boldsymbol{f} \right\|^2 P_n(d\boldsymbol{x}),$$

where P_n is the empirical measure of the data points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in $\mathbb{R}^p$. For each set of cluster centers F and each $p \times q$ orthonormal matrix A, we obtain

$$\lim_{n\to\infty} FKM(F, A, P_n) = FKM(F, A, P) := \int \min_{\boldsymbol{f}\in F} \left\| A^T\boldsymbol{x} - \boldsymbol{f} \right\|^2 P(d\boldsymbol{x}) \quad \text{a.s.}$$

by the strong law of large numbers (SLLN). Thus, besides k-means clustering and RKM clustering, the global minimizer of $FKM(\cdot, \cdot, P_n)$ is also expected to converge almost surely to the global ones of $FKM(\cdot, \cdot, P)$, say the population global minimizers.

In this paper, we derive sufficient conditions for the existence of population global minimizers and then prove the strong consistency of FKM clustering under some regularity conditions. The framework of the proof in this paper is based on ones of the proof of the strong consistency of k-means clustering (Pollard 1981; 1982) and RKM clustering (Terada 2014). In Pollard (1981), the proof of strong consistency of k-means clustering takes an inductive form. On the other hand, the proof of strong consistency of FKM clustering does not take such form as with Terada (2014). In the proof of main theorem, first we also show that the optimal sample centers eventually lie in some compact region on $\mathbb{R}^p$ as with Pollard (1981) and Terada (2014) and then prove the conclusion of the theorem in the same manner of the last part of the proof of the consistency theorem in Terada (2014). For an arbitrary $p \times q$ column-wise orthonormal matrix A ($A^TA = I_q,\ q < p$), an arbitrary p-dimensional point $\boldsymbol{x} \in \mathbb{R}^p$ and an arbitrary q-dimensional point $\boldsymbol{y} \in \mathbb{R}^q$, the key inequality in this paper is that $\|A^T\boldsymbol{x}\| \leq \|\boldsymbol{x}\|$ while the key equation in the strong consistency of RKM clustering (Terada 2014) is that $\|A\boldsymbol{y}\| = \|\boldsymbol{y}\|$.

The rest of the paper is organized as follows. In Sect. 2, we describe the algorithm of FKM clustering to get the local minimum and the relationship between RKM clustering and FKM clustering. We introduce prerequisites and notations in Sect. 3. In Sect. 4, we prove the uniform SLLN and the continuity of the objective function of FKM clustering. The sufficient condition for the existence of the population global minimizers and the strong consistency theorem for FKM clustering are stated, and we derive a rough large deviation inequality in Sect. 5. In Sect. 6, we provide the main proof of the consistency theorem.

2 Factorial k-means clustering

We will denote the number of objects and that of variables by n and p, respectively. Let $X = (x_{ij})_{n\times p}$ be a data matrix and $\boldsymbol{x}_i$ $(i = 1, \ldots, n)$ be row vectors of X. For given number of clusters k and given number of dimensions of a subspace q, the objective function of FKM clustering is defined by

$$FKM_n(A, F, U \mid k, q) := \|XA - UF\|_F^2 = \sum_{i=1}^{n} \min_{1\leq j\leq k} \left\| A^T\boldsymbol{x}_i - \boldsymbol{f}_j \right\|^2,$$

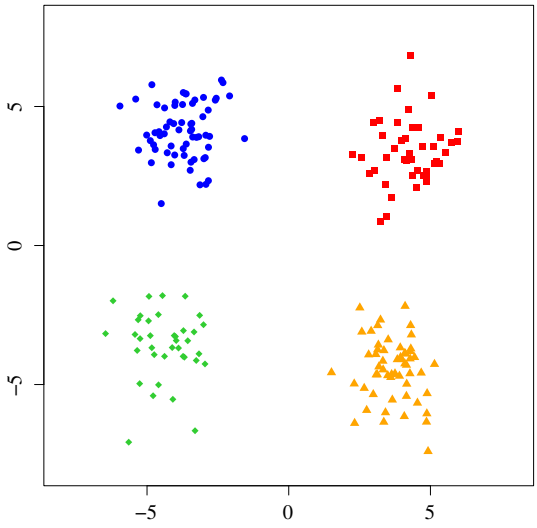


Fig. 3 Plot of the result of FKM clustering for the artificial data given in Fig. 1

where $\|\cdot\|_F$ denotes the Frobenius norm, $U = (u_{ij})_{n\times k}$ is a binary membership matrix, A is a $p \times q$ column-wise orthonormal loading matrix, $F = (f_{ij})_{k\times q}$ is a centroid matrix, and $\boldsymbol{f}_j$ $(j = 1, \ldots, k)$ are row vectors of F representing the jth cluster center. FKM$_n$ can be minimized by the following alternating least-squares algorithm:

Step 0. First, initial values are chosen for A, F, and U.
Step 1. For each $i = 1, \ldots, n$ and each $j = 1, \ldots, k$, we update u_{ij} by

$$u_{ij} = \begin{cases} 1 & \text{iff} \|A^T \boldsymbol{x}_i - \boldsymbol{f}_j\|^2 < \|A^T \boldsymbol{x}_i - \boldsymbol{f}_{j'}\|^2 \text{for each} j' \neq j, \\ 0 & \text{otherwise.} \end{cases}$$

Step 2. A is updated by the first q eigenvectors of $X^T \left[U(U^TU)^{-1}U^T - I_n\right] X$, where I_n is the n-dimensional identity matrix.
Step 3. F is updated using $(U^TU)^{-1}U^TXA$.
Step 4. Finally, the value of the function FKM_n for the present values of A, F, and U is computed. If the function value has decreased, the values of A, F, and U are updated in accordance with Steps 1–3. Otherwise, the algorithm has converged.

This algorithm monotonically decreases the FKM objective function and the solution of this algorithm will be at least a local minimum point. Thus, it is better to use many random starts to obtain the global minimum points.

Let $\hat{A}$, $\hat{F}$, and $\hat{U}$ denote the optimal parameters of FKM clustering. We can visualize the low-dimensional subspace that reflects the cluster structure by $X\hat{A}$. Figure 3 represents such a visualization of the optimal subspace that results from FKM clustering for the artificial data given in Fig. 1.

Next, we briefly discuss the relationship between FKM clustering and RKM clustering. The objective function of RKM clustering is defined by

$$RKM_n(A, F, U) := \|X - UFA^T\|_F^2 = \sum_{i=1}^{n} \min_{1 \le j \le k} \|\boldsymbol{x}_i - A\boldsymbol{f}_j\|^2.$$

This objective function can be decomposed into two terms:

$$RKM_n(A, F, U) = \|X - XAA^T\|_F^2 + \|XA - UF\|_F^2. \tag{1}$$

The first term of Eq. (1) is the objective function of the PCA procedure, and the second term is that of FKM clustering. Thus, FKM clustering reveals the low-dimensional subspace reflecting the cluster structure more clearly than the subspace of RKM clustering in the cases that the data have much variability in directions orthogonal to the subspace containing the cluster structure. For more details about the relationship between FKM and RKM clusterings, see Timmerman et al. (2010).

3 Preliminaries

In this paper, the similar notations as the ones used in Pollard (1981) and Terada (2014) are used. Let $(\Omega, \mathcal{F}, P)$ be the probability space, and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d. p-dimensional random variables drawn from the distribution P. Let P_n denote the empirical measure based on $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$. The set of all $p \times q$ column-wise orthonormal matrices will be denoted by $\mathcal{O}(p \times q)$. $B_q(r)$ denotes the q-dimensional closed ball of radius r centered at the origin. We will define $\mathcal{R}_k := \{R \subset \mathbb{R}^q \mid \#(R) \le k\}$, where $\#(R)$ is the cardinality of R. We will denote the parameter space by $\Xi_k := \mathcal{R}_k \times \mathcal{O}(p \times q)$. For each $M > 0$, $\mathcal{R}_k^*(M) := \{R \subset \mathbb{R}^q \mid \#(R) \le k \text{ and } R \subset B_q(M)\}$ and $\Theta_k^*(M) := \mathcal{R}_k^*(M) \times \mathcal{O}(p \times q)$. Let $\psi : \mathbb{R} \to \mathbb{R}$ denote a non-negative non-decreasing function. For each subset $F \subset \mathbb{R}^q$ and each $A \in \mathcal{O}(p \times q)$, the general loss function of FKM clustering with a probability measure Q on $\mathbb{R}^p$ is defined by

$$\Psi(F, A, Q) := \int \min_{\boldsymbol{f} \in F} \psi\left(\left\|A^T\boldsymbol{x} - \boldsymbol{f}\right\|\right) Q(d\boldsymbol{x}).$$

Write

$$m_k(Q) := \inf_{(F,A) \in \Xi_k} \Psi(F, A, Q)$$

and

$$m_k^*(Q \mid M) := \inf_{(F,A) \in \Theta_k^*(M)} \Psi(F, A, Q).$$

For $\theta = (F, A) \in \Xi_k$, we will use both descriptions $\Psi(\theta, Q)$ and $\Psi(F, A, Q)$. The set of population global optimizers and that of sample global optimizers will be denoted by $\Theta' := \{\theta \in \Xi_k \mid m_k(P) = \Psi(\theta, P)\}$ and $\Theta_n' := \{\theta \in \Xi_k \mid m_k(P_n) = \Psi(\theta, P_n)\}$, respectively. For each $M > 0$, let $\Theta^* := \{\theta \in \Theta_k^*(M) \mid m_k^*(P \mid M) = \Psi(\theta, P)\}$ and

$\Theta_n^* := \{\theta \in \Theta_k^*(M) \mid m_k^*(P_n \mid M) = \Psi(\theta, P_n)\}$. When we emphasize that Θ' and Θ'_n are dependent on the index k, we write $\Theta'(k)$ and $\Theta'_n(k)$ instead of Θ' and Θ'_n, respectively. One of the measurable estimators in Θ'_n will be denoted by $\hat{\theta}_n$ or $\hat{\theta}_n(k)$. Similarly, let $\hat{\theta}_n^*$ (or $\hat{\theta}_n^*(k)$) denote one of the measurable estimators in Θ_n^*. Existence of measurable estimators is guaranteed by the measurable selection theorem; see Section 6.7 of Pfanzagl (1994) for a detailed explanation.

Let $d_F(\cdot, \cdot)$ be the distance between two matrices based on the Frobenius norm and $d_H(\cdot, \cdot)$ be the Hausdorff distance, which is defined for finite subsets $A, B \subset \mathbb{R}^q$ as

$$d_H(A, B) := \max_{\boldsymbol{a} \in A} \left\{ \min_{\boldsymbol{b} \in B} \|\boldsymbol{a} - \boldsymbol{b}\| \right\}.$$

We will denote a product distance with d_F and d_H by d (e.g., $d := \sqrt{d_F^2 + d_H^2}$). As was done by Terada (2014), the distance between $\hat{\theta}_n$ and Θ' is defined as

$$d(\hat{\theta}_n, \Theta') := \inf\{d(\hat{\theta}_n, \theta) \mid \theta \in \Theta'\}.$$

Like in Pollard (1981), we assume that ψ is continuous and $\psi(0) = 0$. In addition, for controlling the growth of ψ, we assume that there exists $\lambda \geq 1$ such that $\psi(2r) \leq \lambda\psi(r)$ for all $r > 0$. Note that

$$\begin{aligned}
\int \psi\left(\left\|A^T\boldsymbol{x} - \boldsymbol{f}\right\|\right) P(\mathrm{d}\boldsymbol{x}) &\leq \int \psi\left(\left\|A^T\boldsymbol{x}\right\| + \left\|\boldsymbol{f}\right\|\right) P(\mathrm{d}\boldsymbol{x}) \\
&\leq \int \psi(\|\boldsymbol{x}\| + \|\boldsymbol{f}\|) P(\mathrm{d}\boldsymbol{x}) \\
&\leq \int_{\|\boldsymbol{f}\| > \|\boldsymbol{x}\|} \psi(2\|\boldsymbol{f}\|) P(\mathrm{d}\boldsymbol{x}) + \int_{\|\boldsymbol{f}\| \leq \|\boldsymbol{x}\|} \psi(2\|\boldsymbol{x}\|) P(\mathrm{d}\boldsymbol{x}) \\
&\leq \psi(2\|\boldsymbol{f}\|) + \lambda \int \psi(\|\boldsymbol{x}\|) P(\mathrm{d}\boldsymbol{x})
\end{aligned}$$

for all $\boldsymbol{f} \in F$ and all $A \in \mathcal{O}(p \times q)$. Thus, $\Psi(F, A, P)$ is finite for each $F \in \mathcal{R}_k$ and $A \in \mathcal{O}(p \times q)$ as long as $\int \psi(\|\boldsymbol{x}\|) P(d\boldsymbol{x}) < \infty$.

Let R be a $q \times q$ orthonormal matrix, i.e., $R^T R = RR^T = I_q$. For each $\boldsymbol{f} \in \mathbb{R}^q$ and each $A \in \mathcal{O}(p \times q)$, we have $AR^T \in \mathcal{O}(p \times q)$ and

$$\int \psi\left(\left\|A^T\boldsymbol{x} - \boldsymbol{f}\right\|\right) P(\mathrm{d}\boldsymbol{x}) = \int \psi\left(\left\|RA^T\boldsymbol{x} - R\boldsymbol{f}\right\|\right) P(\mathrm{d}\boldsymbol{x}).$$

Hence, Θ' is not a singleton when $\Theta' \neq \emptyset$; that is, FKM clustering has rotational indeterminacy, as well as RKM clustering.

4 The uniform SLLN and the continuity of $\Psi(\cdot, \cdot, P)$

Lemma 1 *Let M be an arbitrary positive number. Let $\mathcal{G}$ be the class of all P-integrable functions on $\mathbb{R}^p$ of the form $g_{(F,A)}(\boldsymbol{x}) := \min_{\boldsymbol{f} \in F} \psi(\|A^T\boldsymbol{x} - \boldsymbol{f}\|)$, where (F, A) takes all values over $\Theta_k^*(M)$. Suppose that $\int \psi(\|\boldsymbol{x}\|) P(d\boldsymbol{x}) < \infty$. Then,*

$$\lim_{n\to\infty} \sup_{g\in\mathcal{G}} \left| \int g(\boldsymbol{x}) P_n(d\boldsymbol{x}) - \int g(\boldsymbol{x}) P(d\boldsymbol{x}) \right| = 0 \quad \text{a.s.}$$

Proof Dehardt (1971) provided a sufficient condition for the uniform SLLN. Thus, it is sufficient to prove that for all $\epsilon > 0$, there exists a finite class of functions $\mathcal{G}_\epsilon$ such that, for each $g \in \mathcal{G}$, there are $\dot{g}$ and $\bar{g}$ in $\mathcal{G}_\epsilon$ with $\dot{g} \le g \le \bar{g}$ and $\int \bar{g}(\boldsymbol{x}) P(d\boldsymbol{x}) - \int \dot{g}(\boldsymbol{x}) P(\boldsymbol{x}) < \epsilon$.

Choose an arbitrary $\epsilon > 0$. Let $S_{p\times q}(\sqrt{q}) := \{X \in \mathbb{R}^{p\times q} \mid \|X\|_F = \sqrt{q}\}$. We will denote by D_{δ_1} the finite set on $\mathbb{R}^q$ satisfying the condition that, for all $\boldsymbol{f} \in B_q(M)$, there exists $\boldsymbol{g} \in D_{\delta_1}$ such that $\|\boldsymbol{f} - \boldsymbol{g}\| < \delta_1$. Similarly, we will denote by $\mathcal{A}_{p\times q,\delta_2}$ the finite set on $S_{p\times q}(\sqrt{q})$ satisfying the condition that, for all $A \in S_{p\times q}(\sqrt{q})$, there exists $B \in \mathcal{A}_{p\times q,\delta_2}$ such that $\|A-B\|_F < \delta_2$. Let $\mathcal{R}_{k,\delta_1} := \{F \in \mathcal{R}_k^*(M) \mid F \subset D_{\delta_1}\}$. Take $\mathcal{G}_\epsilon$ as the finite class of functions of the form

$$\min_{\boldsymbol{f}\in F_*} \psi(\|A_*^T\boldsymbol{x} - \boldsymbol{f}\| + \delta_1 + \delta_2\|\boldsymbol{x}\|) \quad \text{or} \quad \min_{\boldsymbol{f}\in F_*} \psi(\|A_*^T\boldsymbol{x} - \boldsymbol{f}\| - \delta_1 - \delta_2\|\boldsymbol{x}\|),$$

where (F_*, A_*) takes all values over $\mathcal{R}_{k,\delta_1} \times \mathcal{A}_{p\times q,\delta_2}$ and $\psi(r)$ is defined as zero for all negative $r < 0$.

For any $F = \{\boldsymbol{f}_1, \ldots, \boldsymbol{f}_k\} \in \mathcal{R}_k^*(M)$, there exists $F_* = \{\boldsymbol{f}_1^*, \ldots, \boldsymbol{f}_k^*\} \in \mathcal{R}_{k,\delta_1}$ with $\|\boldsymbol{f}_i - \boldsymbol{f}_i^*\| < \delta_1$ for each i. In addition, since $\mathcal{O}(p\times q) \subset \bigcup_{A_*\in\mathcal{A}_{p\times q,\delta_2}} \{A \mid \|A - A_*\|_F < \delta_2\}$, for any $A \in \mathcal{O}(p\times q)$ there exists $A_* \in \mathcal{A}_{p\times q,\delta_2}$ with $\|A - A_*\|_F < \delta_2$. Corresponding to each $g_{(F,A)} \in \mathcal{G}$, choose

$$\bar{g}_{(F,A)}(\boldsymbol{x}) := \min_{\boldsymbol{f}\in F_*} \psi\big(\big\|A_*^T\boldsymbol{x} - \boldsymbol{f}\big\| + \delta_1 + \delta_2\|\boldsymbol{x}\|\big)$$

and

$$\dot{g}_{(F,A)}(\boldsymbol{x}) := \min_{\boldsymbol{f}\in F_*} \psi\big(\big\|A_*^T\boldsymbol{x} - \boldsymbol{f}\big\| - \delta_1 - \delta_2\|\boldsymbol{x}\|\big).$$

Since ψ is a monotone function and

$$\big\|A_*^T\boldsymbol{x} - \boldsymbol{f}_j^*\big\| - \delta_1 - \delta_2\|\boldsymbol{x}\| \le \big\|A^T\boldsymbol{x} - \boldsymbol{f}_j\big\| \le \big\|A_*^T\boldsymbol{x} - \boldsymbol{f}_j^*\big\| + \delta_1 + \delta_2\|\boldsymbol{x}\|$$

for each i and each $\boldsymbol{x} \in \mathbb{R}^p$, we have $\dot{g}_{(F,A)} \le g_{(F,A)} \le \bar{g}_{(F,A)}$.

Choosing $R > 0$ to be greater than $(M + \delta_1)/\sqrt{q}$ (or $(M + \delta_1)/(\sqrt{q} + \delta_2)$), we obtain

$$\begin{aligned}
&\int \big[\bar{g}_{(F,A)}(\boldsymbol{x}) - \dot{g}_{(F,A)}(\boldsymbol{x})\big] P(\mathrm{d}\boldsymbol{x}) \\
&\le \int \sum_{i=1}^{k} \big[\psi\big(\big\|A_*^T\boldsymbol{x} - \boldsymbol{f}_i^*\big\| + \delta_1 + \delta_2\|\boldsymbol{x}\|\big) - \psi\big(\big\|A_*^T\boldsymbol{x} - \boldsymbol{f}_i^*\big\| - \delta_1 - \delta_2\|\boldsymbol{x}\|\big)\big] P(\mathrm{d}\boldsymbol{x})
\end{aligned}$$

$$\leq k \sup_{\|\boldsymbol{x}\|\leq R} \sup_{\boldsymbol{f}\in B_q(M)} \sup_{A\in S_{p\times q}(\sqrt{q})} \left[\psi\left(\left\|A^T\boldsymbol{x}-\boldsymbol{f}\right\|+\delta_1+\delta_2\|\boldsymbol{x}\|\right)\right.$$
$$\left.-\psi\left(\left\|A^T\boldsymbol{x}-\boldsymbol{f}\right\|-\delta_1-\delta_2\|\boldsymbol{x}\|\right)\right]+2k\lambda^m\int_{\|\boldsymbol{x}\|\geq R}\psi(\|\boldsymbol{x}\|)P(\mathrm{d}\boldsymbol{x}),$$

where $m\in\mathbb{N}$ is chosen to satisfy the requirement that $\sqrt{q}+\delta_2\leq 2^{m-1}$. The second term in the last bound of the inequality directly above can be less than $\epsilon/2$ by choosing R to be sufficiently large. Note that ψ is uniformly continuous on a bounded set. The first term can be less than $\epsilon/2$ by choosing $\delta_1,\delta_2>0$ to be sufficiently small. Therefore, the sufficient condition of the uniform SLLN for $\mathcal{G}$ is satisfied, and the proof is complete. □

Lemma 2 *Let M be an arbitrary positive number. Suppose that $\int\psi(\|\boldsymbol{x}\|)P(d\boldsymbol{x})<\infty$. Then, $\Psi(\cdot,P)$ is continuous on $\Theta_k^*(M)$.*

Proof This lemma can be proven in a similar manner as the proof of Lemma 1. If $(F,A),(G,B)\in\Theta_k^*(M)$ is chosen to satisfy $d_H(F,G)<\delta_1$ and $\|A-B\|_F<\delta_2$, then for each $\boldsymbol{g}\in G$ there exists $\boldsymbol{f}(\boldsymbol{g})\in F$ such that $\|\boldsymbol{g}-\boldsymbol{f}(\boldsymbol{g})\|<\delta_1$. Choosing R to be larger than $M+\delta_1$, we obtain

$$\begin{aligned}
&\Psi(F,A,P)-\Psi(G,B,P)\\
&=\int\left[\min_{\boldsymbol{f}\in F}\psi\left(\left\|A^T\boldsymbol{x}-\boldsymbol{f}\right\|\right)-\min_{\boldsymbol{g}\in G}\psi\left(\left\|B^T\boldsymbol{x}-\boldsymbol{g}\right\|\right)\right]P(\mathrm{d}x)\\
&\leq\int\max_{\boldsymbol{g}\in G}\left[\psi\left(\left\|A^T\boldsymbol{x}-\boldsymbol{f}(\boldsymbol{g})\right\|\right)-\psi\left(\left\|B^T\boldsymbol{x}-\boldsymbol{g}\right\|\right)\right]P(\mathrm{d}\boldsymbol{x})\\
&\leq\int\sum_{\boldsymbol{g}\in G}\left[\psi\left(\left\|B^T\boldsymbol{x}-\boldsymbol{g}\right\|+\delta_1+\delta_2\|\boldsymbol{x}\|\right)-\psi\left(\left\|B^T\boldsymbol{x}-\boldsymbol{g}\right\|\right)\right]P(\mathrm{d}\boldsymbol{x})\\
&\leq k\sup_{\|\boldsymbol{x}\|\leq R}\max_{\boldsymbol{g}\in G}\left[\psi\left(\left\|B^T\boldsymbol{x}-\boldsymbol{g}\right\|+\delta_1+\delta_2\|\boldsymbol{x}\|\right)-\psi\left(\left\|B^T\boldsymbol{x}-\boldsymbol{g}\right\|\right)\right]\\
&\quad+2\sum_{\boldsymbol{g}\in G}\int_{\|\boldsymbol{x}\|\geq R}\psi\left(\left\|B^T\boldsymbol{x}-\boldsymbol{g}\right\|+\delta_1+\delta_2\|\boldsymbol{x}\|\right)P(\mathrm{d}\boldsymbol{x})\\
&\leq k\sup_{\|\boldsymbol{x}\|\leq R}\max_{\boldsymbol{g}\in G}\left[\psi\left(\left\|B^T\boldsymbol{x}-\boldsymbol{g}\right\|+\delta_1+\delta_2\|\boldsymbol{x}\|\right)-\psi\left(\left\|B^T\boldsymbol{x}-\boldsymbol{g}\right\|\right)\right]\\
&\quad+2k\lambda^m\int_{\|\boldsymbol{x}\|\geq R}\psi(\|\boldsymbol{x}\|)P(\mathrm{d}\boldsymbol{x}),
\end{aligned}\tag{2}$$

where $m\in\mathbb{N}$ is chosen to satisfy the condition that $2+\delta_2\leq 2^m$. By choosing R to be sufficiently large and $\delta_1,\delta_2>0$ to be sufficiently small, the last bound in the inequality (2) can be less than ϵ. Since for each $\boldsymbol{f}\in F$ there exists $\boldsymbol{g}(\boldsymbol{f})\in G$ such that $\|\boldsymbol{g}-\boldsymbol{g}(\boldsymbol{f})\|<\delta_1$, the other inequality needed for continuity is obtained by interchanging (F,A) and (G,B) in the inequality (2). □

5 Consistency theorem

5.1 Existence of population global optimizers

Our purpose is to prove that $\lim_{n\to\infty} d(\hat{\theta}_n, \Theta') = 0$ a.s. under some regularity conditions. However, there is a possibility that Θ' is empty. Therefore, first, we provide sufficient conditions for the existence of population global optimizers.

Proposition 1 *Suppose that* $\int \psi(\|\boldsymbol{x}\|)P(d\boldsymbol{x}) < \infty$ *and that* $m_j(P) > m_k(P)$ *for* $j = 1, 2, \ldots, k-1$. *Then,* $\Theta' \neq \emptyset$. *Furthermore, there exists* $M > 0$ *such that* $F \subset B_q(5M)$ *for all* $(F, A) \in \Theta'$.

Proof See Appendix 8. □

Under the assumption of Proposition 1, we can prove that $\Psi(\cdot, P)$ ensures the identification condition, which is a requirement of the consistency theorem.

Corollary 1 *Suppose that* $\int \psi(\|\boldsymbol{x}\|)P(d\boldsymbol{x}) < \infty$ *and that* $m_j(P) > m_k(P)$ *for* $j = 1, 2, \ldots, k-1$. *Then, there exists* $M_0 > 0$ *such that for each* $M > M_0$

$$\inf_{\theta\in\Theta^*_\epsilon(M)} \Psi(\theta, P) > \inf_{\theta\in\Theta'} \Psi(\theta, P) \quad \textit{for all } \epsilon > 0,$$

where $\Theta^*_\epsilon(M) := \{\theta \in \Theta^*_k(M) \mid d(\theta, \Theta') \geq \epsilon\}$.

Proof See Appendix 8. □

5.2 Strong consistency of FKM clustering

If the parameter space is restricted to $\Theta^*_k(M) \subset \Xi_k$, we easily obtain the strong consistency of FKM clustering. Since $\Theta^*_k(M)$ is compact, we have $\Theta^* \neq \emptyset$ and the identification condition:

$$\inf_{\theta\in\Theta^*_\epsilon(M)} \Psi(\theta, P) > \inf_{\theta\in\Theta^*} \Psi(\theta, P) \quad \text{for all } \epsilon > 0$$

where $\Theta^*_\epsilon(M) := \{\theta \in \Theta^*_k(M) \mid d(\theta, \Theta^*) \geq \epsilon\}$.

Proposition 2 *Let* M *be an arbitrary positive number. Suppose that* $\int \psi(\|\boldsymbol{x}\|)P(d\boldsymbol{x}) < \infty$. *Then,*

$$\lim_{n\to\infty} d\big(\hat{\theta}^*_n, \Theta^*\big) = 0 \text{ a.s., } \textit{and } \lim_{n\to\infty} m^*_k(P_n \mid M) = m^*_k(P \mid M) \text{ a.s.}$$

Proof From Lemma 1 and Lemma 2, we already obtain the uniform SLLN and the continuity of $\Psi(\cdot, P)$ on $\Theta^*_k(M)$. Thus, the proof of this proposition is given by the similar argument of the last part of the proof of the consistency theorem. □

This fact is very important in the proof of Lemma 4. Using this fact, the proof of the main theorem does not necessarily take an inductive form with the number of cluster k.

We cannot assume the uniqueness condition since FKM clustering has rotational indeterminacy. In this study, as Terada (2014) did previously, we assume that $m_j(P) > m_k(P)$ for $j = 1, \ldots, k-1$. This condition implies that an optimal set $F(k)$ of cluster centers has k distinct elements. When we do not use the fact in Proposition 2, the proof of the main theorem takes an inductive form with the number of cluster k as with Pollard (1981) and becomes somewhat more complicated. The following theorem provides sufficient conditions for the strong consistency of FKM clustering.

Theorem 1 *Suppose that* $\int \psi(\|\boldsymbol{x}\|)P(d\boldsymbol{x}) < \infty$ *and that* $m_j(P) > m_k(P)$ *for* $j = 1, \ldots, k-1$. *Then,* $\Theta' \neq \emptyset$,

$$\lim_{n\to\infty} d(\hat{\theta}_n, \Theta') = 0 \text{ a.s., } \textit{and } \lim_{n\to\infty} m_k(P_n) = m_k(P) \text{ a.s.}$$

Proof See Sect. 6. □

Note that if there exists a specific A such that $\Psi(A, F, P) = 0$ for all F; that is, the population distribution, P, is degenerate and the number of dimensions with the support of P is given as $p - q$, $m_j(P) > m_k(P)$ for $j = 1, \ldots, k-1$ is not satisfied.

Based on the consistency of FKM clustering and RKM clustering, we can compare these methods more clearly. In the following example, we show that there exist some cases in which RKM clustering becomes equivalent to FKM clustering as n goes to infinity.

Example 2 (Asymptotic equivalence of FKM clustering and RKM clustering)

Let $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{0}$, $\mathrm{Var}(X_s) = \sigma^2$ and $\mathrm{Cov}(X_s, X_t) = 0$ $(s \neq t)$. For $A \in \mathcal{O}(p \times q)$, write $B = (b_{st})_{p\times p} := AA^T$. Then, we have

$$\begin{aligned}\int \left\|A^T\boldsymbol{x}\right\|^2 P(d\boldsymbol{x}) &= \int \boldsymbol{x}^T AA^T\boldsymbol{x}\, P(d\boldsymbol{x}) = \int \sum_{s=1}^{p}\sum_{t=1}^{p} b_{st}x_s x_t P(d\boldsymbol{x}) \\ &= \sum_{s=1}^{p} b_{ss}\int x_s^2 P(d\boldsymbol{x}) + 2\sum_{s=1}^{p-1}\sum_{t=s+1}^{p} b_{st}\int x_s x_t P(d\boldsymbol{x}) \\ &= \sum_{s=1}^{p}\left(\sum_{t=1}^{q} a_{st}^2\right)\sigma^2 = \sigma^2\mathrm{tr}(A^TA) = q\sigma^2.\end{aligned}$$

Thus, the objective function of RKM clustering can be decomposed into the following two terms:

$$\begin{aligned}&\int \min_{\boldsymbol{f}\in F} \|\boldsymbol{x} - A\boldsymbol{f}\|^2 P(d\boldsymbol{x}) \\ &= \int \left\|\boldsymbol{x} - AA^T\boldsymbol{x}\right\|^2 P(d\boldsymbol{x}) + \int \min_{\boldsymbol{f}\in F} \left\|A^T\boldsymbol{x} - \boldsymbol{f}\right\|^2 P(d\boldsymbol{x})\end{aligned}$$

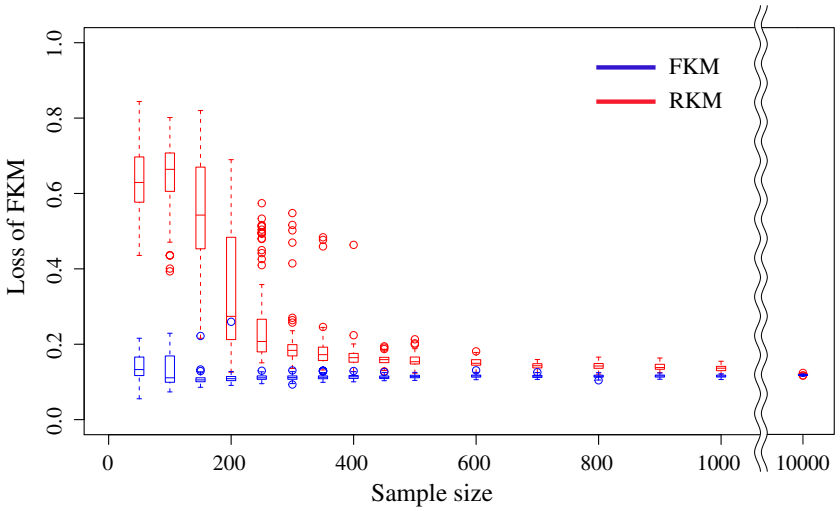


Fig. 4 Boxplots of $\Psi(\hat{\theta}_{\mathrm{FKM}}, P_n)$ and $\Psi(\hat{\theta}_{\mathrm{RKM}}, P_n)$ with the 100 data sets for each sample size

$$= \int \|\boldsymbol{x}\|^2 P(d\boldsymbol{x}) - \int \left\|A^T\boldsymbol{x}\right\|^2 P(d\boldsymbol{x}) + \int \min_{\boldsymbol{f}\in F} \left\|A^T\boldsymbol{x} - \boldsymbol{f}\right\|^2 P(d\boldsymbol{x})$$
$$= (p-q)\sigma^2 + \int \min_{\boldsymbol{f}\in F} \left\|A^T\boldsymbol{x} - \boldsymbol{f}\right\|^2 P(d\boldsymbol{x}).$$

Note that the first term is constant and the second term is the objective function of FKM clustering. In this setting, the set of population global optimizers of RKM clustering is same as that of FKM clustering. Let $\hat{\theta}_{\mathrm{FKM}} := (\hat{F}_{\mathrm{FKM}}, \hat{A}_{\mathrm{FKM}})$ and $\hat{\theta}_{\mathrm{RKM}} := (\hat{F}_{\mathrm{RKM}}, \hat{A}_{\mathrm{RKM}})$ denote the estimators of FKM and RKM clusterings, respectively. Here, we set $\psi(x) := x^2$. Both $\Psi(\hat{\theta}_{\mathrm{RKM}}, P_n)$ and $\Psi(\hat{\theta}_{\mathrm{FKM}}, P_n)$ converge to $m_k(P)$ almost surely as $n \to \infty$. Moreover, if the population global optimizers of FKM (or RKM) clustering are unique up to a rotation, then as $n \to \infty$,

$$\mathrm{Diff}(\hat{C}_{\mathrm{FKM}}, \hat{C}_{\mathrm{RKM}}) := \sum_{\boldsymbol{f}\in\hat{C}_{\mathrm{FKM}}} \min_{\boldsymbol{g}\in\hat{C}_{\mathrm{RKM}}} \|\boldsymbol{f}-\boldsymbol{g}\|^2 \to 0 \quad \text{a.s.},$$

where $\hat{C}_{\mathrm{FKM}} := \{\hat{A}_{\mathrm{FKM}}\boldsymbol{f} \mid \boldsymbol{f} \in \hat{F}_{\mathrm{FKM}}\}$ and $\hat{C}_{\mathrm{RKM}} := \{\hat{A}_{\mathrm{RKM}}\boldsymbol{f} \mid \boldsymbol{f} \in \hat{F}_{\mathrm{RKM}}\}$.

For example, let $Z = [\boldsymbol{Z}_1, ..., \boldsymbol{Z}_n]^T$ be the normalized data matrix of the data matrix X in Example 1 with zero means and unit variances. Then we have $\mathbb{E}[\boldsymbol{Z}] = \boldsymbol{0}$, $\mathrm{Var}(Z_s) = 1$ and $\mathrm{Cov}(Z_s, Z_t) = 0$ $(s \neq t)$. Here, we set $p = 12$ and generated 100 data sets for each sample size. We applied FKM and RKM clusterings with 100 random starts for these data sets. Figure 4 shows the boxplots of $\Psi(\hat{\theta}_{\mathrm{FKM}}, P_n)$ and $\Psi(\hat{\theta}_{\mathrm{RKM}}, P_n)$ with the 100 data sets for each sample size. Figure 5 shows the boxplots of $\mathrm{Diff}(\hat{C}_{\mathrm{FKM}}, \hat{C}_{\mathrm{RKM}})$ with the 100 data sets for each sample size. These figures show that RKM clustering becomes equivalent to FKM clustering as n goes to infinity in this setting.

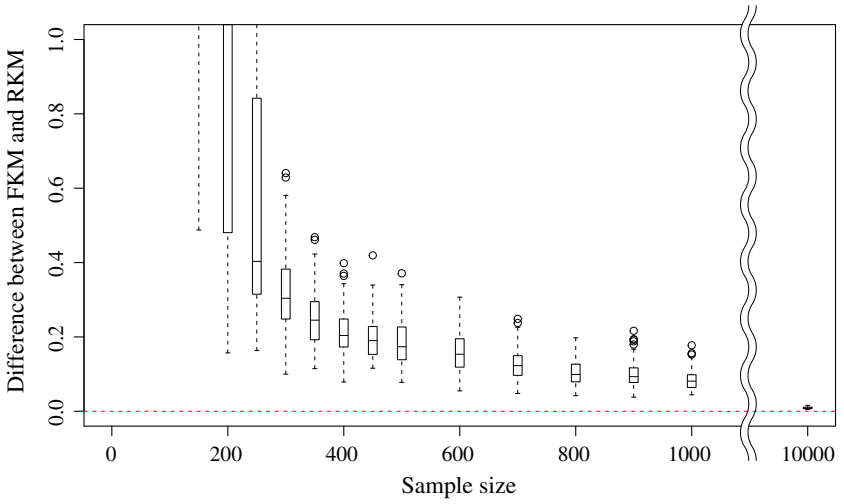


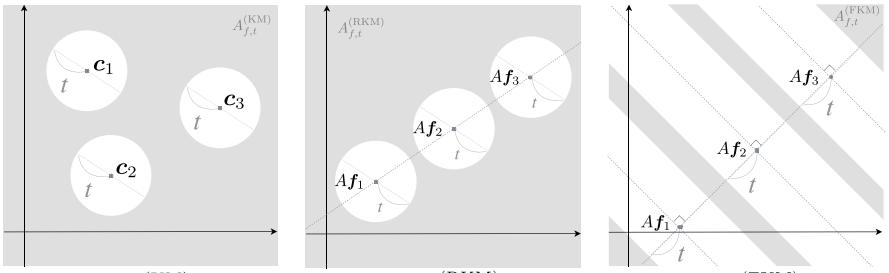
Fig. 5 Boxplots of $\mathrm{Diff}(\hat{C}_{\mathrm{FKM}}, \hat{C}_{\mathrm{RKM}})$ with the 100 data sets for each sample size

5.3 Large deviation inequality for FKM clustering

From Corollary 29.1 in Devroye et al. (1996), if the support of the population distribution is bounded, we have a non-asymptotic large deviation inequality. Before stating this theorem, we introduce some notations which are used in the proof of the theorem. In this subsection, we assume that the support of the population distribution is bounded; that is, $P(\|X_1\|^2 \leq B) = 1$ for some $B > 0$. Let $\mathcal{F}_{\mathrm{KM}}(k) := \{f(\boldsymbol{x}) = \min_{\boldsymbol{c}\in C} \|\boldsymbol{x} - \boldsymbol{c}\|^2 \mid C \subset B_p(\sqrt{B}), \#(C) \leq k\}$, $\mathcal{F}_{\mathrm{FKM}}(k, q) := \{f(\boldsymbol{x}) = \min_{\boldsymbol{f}\in F} \|A^T\boldsymbol{x} - \boldsymbol{f}\|^2 \mid (F, A) \in \Theta_k^*(\sqrt{B})\}$ and $\mathcal{F}_{\mathrm{RKM}}(k, q) := \{f(\boldsymbol{x}) = \min_{\boldsymbol{f}\in F} \|\boldsymbol{x} - A\boldsymbol{f}\|^2 \mid (F, A) \in \Theta_k^*(\sqrt{B})\}$. Moreover, for every $f \in \mathcal{F}_{\mathrm{KM}}(k)$ and $t \in [0, 4B]$, the set $A_{f,t}^{(\mathrm{KM})} \subset \mathbb{R}^p$ is defined as $A_{f,t}^{(\mathrm{KM})} := \{\boldsymbol{x} \mid f(\boldsymbol{x}) > t\}$. and define $\hat{\mathcal{F}}_{\mathrm{KM}} := \{A_{f,t}^{(\mathrm{KM})} \mid f \in \mathcal{F}_{\mathrm{KM}}(k, q), t \in [0, 4B]\}$. Similarly, we define $A_{f,t}^{(\mathrm{FKM})}$, $A_{f,t}^{(\mathrm{RKM})}$, $\hat{\mathcal{F}}_{\mathrm{FKM}}$ and $\hat{\mathcal{F}}_{\mathrm{RKM}}$. Figure 6 shows sets $A_{f,t}^{(\mathrm{KM})}$, $A_{f,t}^{(\mathrm{FKM})}$ and $A_{f,t}^{(\mathrm{RKM})}$ with $k = 3$, $p = 2$ and $q = 1$. From Fig. 6, we can clearly see the differences between k-means, FKM and RKM clusterings.

Theorem 2 *Let $\psi(x) := x^2$ and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n \in \mathbb{R}^p$ be i.i.d. random vectors such that $P(\|\boldsymbol{X}_1\|^2 \leq B) = 1$ for some $0 < B < \infty$. Then, for all $q < p$ and all $\epsilon > 0$,*

$$P\left(\Psi\left(\hat{\theta}_n, P\right) - m_k(P) > \epsilon\right) \leq P\left(2 \sup_{\theta\in\Theta_k^*(\sqrt{B})} |\Psi(\theta, P_n) - \Psi(\theta, P)| > \epsilon\right)$$
$$\leq 8n^{8.741k(p-q+1)(q+1)} \exp\left(-\frac{n\epsilon^2}{2048B^2}\right).$$



(a) $A^{(\mathrm{KM})}_{f,t}$ **(b)** $A^{(\mathrm{RKM})}_{f,t}$ **(c)** $A^{(\mathrm{FKM})}_{f,t}$

Fig. 6 Grey areas in three figures show the sets $A^{(\mathrm{KM})}_{f,t}$, $A^{(\mathrm{RKM})}_{f,t}$ and $A^{(\mathrm{FKM})}_{f,t}$ with $k = 3$, $p = 2$ and $q = 1$, respectively

Proof Since $P(\|\boldsymbol{X}_1\|^2 \leq B) = 1$, we clearly have $m_k(P_n) = m_k(P_n \mid \sqrt{B})$ a.s. for all $n \in \mathbb{N}$ and $m_k(P) = m_k(P \mid \sqrt{B})$. Let $\theta_* \in \Theta^*$, that is, $\Psi(\theta_*, P) = m_k\big(P \mid \sqrt{B}\big)$. Then, we have

$$\begin{aligned}
&\Psi\big(\hat{\theta}_n, P\big) - m_k(P \mid \sqrt{B}) \\
&= \Psi\big(\hat{\theta}_n, P\big) - \Psi\big(\hat{\theta}_n, P_n\big) + \Psi\big(\hat{\theta}_n, P_n\big) - \Psi(\theta_*, P_n) + \Psi(\theta_*, P_n) - \Psi(\theta_*, P) \\
&\leq \Psi\big(\hat{\theta}_n, P\big) - \Psi\big(\hat{\theta}_n, P_n\big) + \Psi(\theta_*, P_n) - \Psi(\theta_*, P) \\
&\leq 2 \sup_{\theta \in \Theta^*_k(\sqrt{B})} |\Psi(\theta, P_n) - \Psi(\theta, P)|.
\end{aligned}$$

Thus, we obtain

$$P\left(\Psi\big(\hat{\theta}_n, P\big) - m_k(P) > \epsilon\right) \leq P\left(2 \sup_{\theta \in \Theta^*_k(\sqrt{B})} |\Psi(\theta, P_n) - \Psi(\theta, P)| > \epsilon\right).$$

From Corollary 29.1 in Devroye et al. (1996), if we find an upper bound of the shutter coefficient $s(\hat{\mathcal{F}}_{\mathrm{FKM}}, n)$, then we obtain an uniform deviation inequality for FKM clustering. Thus, we derive a upper bound of $s(\hat{\mathcal{F}}_{\mathrm{FKM}}, n)$. Let $\hat{\mathcal{F}}^c_{\mathrm{FKM}} := \{A^{(\mathrm{FKM})^c}_{f,t} \mid A_{f,t} \in \hat{\mathcal{F}}_{\mathrm{FKM}}\}$. From Theorem 13.5 (ii) in Devroye et al. (1996), we have that $s(\hat{\mathcal{F}}_{\mathrm{FKM}}, n) = s(\hat{\mathcal{F}}^c_{\mathrm{FKM}}, n)$. For every $f(\boldsymbol{x}) := \min_{\boldsymbol{f} \in F} \|A^T \boldsymbol{x} - \boldsymbol{f}\|^2 \in \hat{\mathcal{F}}_{\mathrm{FKM}}$, $\mathcal{H}_l$ denotes the $(p-q)$-dimensional affine subspace which contains $A\boldsymbol{f}_l$ and is orthogonal to the q-dimensional subspace spanned by the k cluster centers $A\boldsymbol{f}_1, \ldots, A\boldsymbol{f}_k$. Then, a set $A \in \hat{\mathcal{F}}^c_{\mathrm{FKM}}$ is the union of the following k hyperbands

$$\mathcal{B}(\mathcal{H}_l, t) := \{\boldsymbol{x} \in \mathbb{R}^p \mid d_O(\boldsymbol{x}, \mathcal{H}_l) \leq t\} \quad (l = 1, \ldots, k),$$

where $d_O(\boldsymbol{x}, \mathcal{H}_l)$ is the orthogonal distance between $\boldsymbol{x}$ and $\mathcal{H}_l$. We have

$$\hat{\mathcal{F}}^c_{\mathrm{FKM}} \subset \left\{\bigcup_{l=1}^{k} \mathcal{B}_l \mid \mathcal{B}_1, \ldots, \mathcal{B}_k \in \mathcal{C}^p_{(p-q)}\right\},$$

where $\mathcal{C}^p_{(p-q)} := \{\mathcal{B}(\mathcal{H}, t) \mid \mathcal{H} \subset \mathbb{R}^p$ is a $(p-q)$-dimensional affine subspace and $t \geq 0\}$. By Theorem 13.5 (iv) in Devroye et al. (1996),

$$s\big(\hat{\mathcal{F}}_{\mathrm{FKM}}, n\big) = s\big(\hat{\mathcal{F}}^c_{\mathrm{FKM}}, n\big) \leq s\big(\mathcal{C}^p_{(p-q)}, n\big)^k.$$

Akama et al. (2010) provides the lower and upper bounds of the VC dimension of $\mathcal{C}^p_{(p-q)}$:

$$(p-q+1)(q+1) \leq \mathrm{VCdim}\big(\mathcal{C}^p_{(p-q)}\big) \leq 8.741(p-q+1)(q+1).$$

Since $\mathrm{VCdim}(\mathcal{C}^p_{(p-q)}) > 2$ for all $q < p$, by Theorem 13.3 in Devroye et al. (1996), we have

$$s\big(\mathcal{C}^p_{(p-q)}, n\big) \leq n^{\mathrm{VCdim}\left(\mathcal{C}^p_{(p-q)}\right)} \leq n^{8.741(p-q+1)(q+1)}$$

and

$$s\big(\hat{\mathcal{F}}_{\mathrm{FKM}}, n\big) \leq s\big(\mathcal{C}^p_{(p-q)}, n\big)^k \leq n^{8.741k(p-q+1)(q+1)}.$$

Since $0 \leq f(\boldsymbol{X}_1) \leq 4B$ a.s. for all $f \in \mathcal{F}_{\mathrm{FKM}}(k, q)$, by Corollary 29.1 in Devroye et al. (1996) we obtain

$$\begin{aligned} P\left(\Psi\big(\hat{\theta}_n, P\big) - m_k(P) > \epsilon\right) &\leq P\left(2 \sup_{\theta \in \Theta^*_k(\sqrt{B})} |\Psi(\theta, P_n) - \Psi(\theta, P)| > \epsilon\right) \\ &\leq 8s\big(\hat{\mathcal{F}}_{\mathrm{FKM}}, n\big) \exp\left(-\frac{n\epsilon^2}{2048B^2}\right) \\ &\leq 8n^{8.741k(p-q+1)(q+1)} \exp\left(-\frac{n\epsilon^2}{2048B^2}\right). \end{aligned}$$

□

6 Proof of the consistency theorem

Since the theorem deals with almost sure convergence, there might exist null subsets of Ω on which the strong consistency does not hold. Therefore, throughout the proof, Ω_1 denotes the set obtained by avoiding a proper null set from Ω.

First, we prove that there exists $M > 0$ such that, for sufficiently large n, at least one center of the estimator $F_n \in \mathcal{R}_k$ is contained in $B_q(M)$.

Lemma 3 *Suppose that* $\int \psi(\|\boldsymbol{x}\|) P(d\boldsymbol{x}) < \infty$. *Then, there exists* $M > 0$ *such that*

$$P\left(\bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} \{\omega \mid \forall (F_m, A_m) \in \Theta'_m;\ F_m(\omega) \cap B_q(M) \neq \emptyset\}\right) = 1.$$

Proof Choose an $r > 0$ to satisfy the condition that $P(B_p(r)) > 0$. Let us take M to be sufficiently large to ensure that $M > r$ and

$$\psi(M-r)P(B_p(r)) > \int \psi(\|\boldsymbol{x}\|)P(d\boldsymbol{x}). \tag{3}$$

Note that $m_k(P_n) \leq \Psi(F, A, P_n)$ for all $F \in \mathcal{R}_k$ and all $A \in \mathcal{O}(p \times q)$. Let F_0 be the singleton that consists of only the origin. By the SLLN, we obtain

$$\Psi(F_0, A, P_n) = \int \psi\left(\left\|A^T\boldsymbol{x}\right\|\right)P_n(\mathrm{d}\boldsymbol{x}) \to \int \psi\left(\left\|A^T\boldsymbol{x}\right\|\right)P(\mathrm{d}\boldsymbol{x}) \quad \text{a.s.}$$

for all $A \in \mathcal{O}(p \times q)$. Since $\|A^T\boldsymbol{x}\| \leq \|\boldsymbol{x}\|$, we have

$$\int \psi\left(\left\|A^T\boldsymbol{x}\right\|\right)P(d\boldsymbol{x}) \leq \int \psi(\|\boldsymbol{x}\|)P(d\boldsymbol{x})$$

for all $A \in \mathcal{O}(p \times q)$.

Let $\Omega' := \{\omega \in \Omega_1 \mid \forall n \in \mathbb{N};\ \exists m \geq n;\ F_m(\omega) \cap B_q(M) = \emptyset\}$. For all $\omega \in \Omega'$, there exists a subsequence $\{n_l\}_{l\in\mathbb{N}}$ such that $F_{n_l}(\omega) \cap B_q(M) = \emptyset$. Since $\|A^T\boldsymbol{x} - \boldsymbol{f}\| \geq \|\boldsymbol{f}\| - \|\boldsymbol{x}\| > M - r$ for all $\boldsymbol{x} \in B_p(r)$, all $\boldsymbol{f} \notin B_q(M)$, and all $A \in \mathcal{O}(p \times q)$, we have

$$\begin{aligned}
\limsup_l \Psi(F_{n_l}, A_{n_l}, P_{n_l}) &\geq \limsup_l \frac{1}{n_l} \sum_{i\in\{i\mid \boldsymbol{X}_i\in B_p(r)\}} \min_{\boldsymbol{f}\in F_{n_l}} \psi\left(\left\|A_{n_l}^T\boldsymbol{X}_i - \boldsymbol{f}\right\|\right) \\
&\geq \limsup_l \frac{1}{n_l} \sum_{i\in\{i\mid \boldsymbol{X}_i\in B_p(r)\}} \psi(M-r) \\
&\geq \psi(M-r)P(B_p(r)).
\end{aligned}$$

From the assumptions made on the values of M, we have

$$\limsup_l \Psi(F_{n_l}, A_{n_l}, P_{n_l}) > \int \psi(\|\boldsymbol{x}\|)P(d\boldsymbol{x}),$$

which contradicts $m_k(P_n) \leq \Psi(F, A, P_n)$ for all $F \in \mathcal{R}_k$ and all $A \in \mathcal{O}(p \times q)$. Therefore, we obtain $P(\Omega') = 0$; that is,

$$P\left(\bigcup_{n=1}^{\infty}\bigcap_{m=n}^{\infty}\{\omega \mid \forall(F_m, A_m) \in \Theta'_m;\ F_m(\omega) \cap B_q(M) \neq \emptyset\}\right) = 1.$$

□

By Lemma 3, without loss of generality, we can assume that each F_n contains at least one element of $B_q(M)$ when n is sufficiently large. The next lemma indicates that there exists $M > 0$ such that $B_q(5M)$ contains all the estimators of centers when n is sufficiently large.

Lemma 4 *Under the assumption of the theorem, there exists $M > 0$ such that*

$$P\left(\bigcup_{n=1}^{\infty}\bigcap_{m=n}^{\infty}\{\omega \mid \forall (F_m, A_m) \in \Theta'_m;\ F_m(\omega) \subset B_q(5M)\}\right) = 1.$$

Proof Choose $\epsilon > 0$ sufficiently small such that $\epsilon + m_k(P) < m_{k-1}(P)$. Let us take $M > 0$ to satisfy the inequality (3) and

$$\lambda \int_{\|\boldsymbol{x}\| \geq 2M} \psi(\|\boldsymbol{x}\|) P(d\boldsymbol{x}) < \epsilon. \tag{4}$$

Suppose that F_n contains at least one center outside $B_q(5M)$. By Lemma 3, when n is sufficiently large, F_n must contain at least one center in $B_q(M)$, say $\boldsymbol{f}_1 \in B_q(M)$. Since $\{\boldsymbol{x} \mid \|A^T\boldsymbol{x}\| \geq 2M\} \subset \{\boldsymbol{x} \mid \|\boldsymbol{x}\| \geq 2M\}$, we have

$$\begin{aligned}\int_{\|A^T\boldsymbol{x}\| \geq 2M} \psi\left(\left\|A^T\boldsymbol{x} - \boldsymbol{f}_1\right\|\right) P_n(\mathrm{d}\boldsymbol{x}) &\leq \int_{\|\boldsymbol{x}\| \geq 2M} \psi\left(\left\|A^T\boldsymbol{x} - \boldsymbol{f}_1\right\|\right) P_n(\mathrm{d}\boldsymbol{x}) \\ &\leq \int_{\|\boldsymbol{x}\| \geq 2M} \psi(\|\boldsymbol{x}\| + \|\boldsymbol{f}_1\|) P_n(\mathrm{d}\boldsymbol{x}) \\ &\leq \lambda \int_{\|\boldsymbol{x}\| \geq 2M} \psi(\|\boldsymbol{x}\|) P_n(\mathrm{d}\boldsymbol{x})\end{aligned}$$

for all $A \in \mathcal{O}(p \times q)$. Let F_n^* denote the set obtained by deleting all centers lying outside $B_q(5M)$ from F_n. Since $(F_n^*, A) \in \Theta_{k-1}^*(5M)$ for all $A \in \mathcal{O}(p \times q)$, we have

$$\Psi\left(F_n^*, A, P_n\right) \geq m_{k-1}^*(P_n \mid 5M) \geq m_{k-1}(P_n)$$

for all $A \in \mathcal{O}(p \times q)$. For each $\boldsymbol{x} \in B_p(2M)$ and each $A \in \mathcal{O}(p \times q)$, we have

$$\|A^T\boldsymbol{x} - \boldsymbol{f}\| \geq \|\boldsymbol{f}\| - \|\boldsymbol{x}\| > 3M \quad \text{for all } \boldsymbol{f} \notin B_q(5M)$$

and

$$\left\|A^T\boldsymbol{x} - \boldsymbol{g}\right\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{g}\| < 3M \quad \text{for all } \boldsymbol{g} \in B_q(5M).$$

Thus, we obtain

$$\int_{\|x\| < 2M} \min_{\boldsymbol{f} \in F_n} \psi\left(\left\|A^T\boldsymbol{x} - \boldsymbol{f}\right\|\right) P_n(\mathrm{d}\boldsymbol{x}) = \int_{\|x\| < 2M} \min_{\boldsymbol{f} \in F_n^*} \psi(\|A^T\boldsymbol{x} - \boldsymbol{f}\|) P_n(\mathrm{d}\boldsymbol{x})$$

for all $A \in \mathcal{O}(p \times q)$.

Let $\Omega^* := \{\omega \in \Omega_1 \mid \forall n \in \mathbb{N};\ \exists m \geq n;\ \exists (F_m, A_m) \in \Theta'_m;\ F_m(\omega) \not\subset B_q(5M)\}$. By the axiom of choice, for an arbitrary $\omega \in \Omega^*$, there exists a subsequence $\{n_l\}_{l \in \mathbb{N}}$

such that $F_m(\omega) \not\subset B_q(5M)$. By Proposition 2, we have

$$\lim_{n\to\infty} m^*_{k-1}(P_n \mid 5M) = m^*_{k-1}(P \mid 5M) \quad \text{a.s.}$$

For any $(F, A) \in \Xi_k$, we have

$$\begin{aligned} m_{k-1}(P) \le m^*_{k-1}(P \mid 5M) &\le \liminf_l \Psi\left(F^*_{n_l}, A_n, P_n\right) \le \limsup_l \Psi(F^*_{n_l}, A_{n_l}, P_{n_l}) \\ &\le \limsup_n \left[\int_{\|\boldsymbol{x}\|<2M} \min_{\boldsymbol{f}\in F_n} \psi\left(\left\| A_n^T \boldsymbol{x} - \boldsymbol{f} \right\|\right) P_n(d\boldsymbol{x}) \right. \\ &\qquad \left. + \int_{\|\boldsymbol{x}\|\ge 2M} \psi\left(\left\| A_n^T \boldsymbol{x} - \boldsymbol{f}_1 \right\|\right) P_n(d\boldsymbol{x}) \right] \\ &\le \limsup_n \left[\Psi(F_n, A_n, P_n) + \lambda \int_{\|\boldsymbol{x}\|\ge 2M} \psi(\|\boldsymbol{x}\|) P_n(d\boldsymbol{x}) \right] \\ &\le \limsup_n \Psi(F, A, P_n) + \lambda \int_{\|\boldsymbol{x}\|\ge 2M} \psi(\|\boldsymbol{x}\|) P_n(d\boldsymbol{x}). \end{aligned} \tag{5}$$

Choose $(\bar{F}, \bar{A}) \in \Theta'$ as $(F, A) \in \Xi_k$ in the last bound of the above inequality. By the assumption of $M > 0$ and the SLLN, for a sufficiently large n, the last bound of the inequality (5) can be less than $m_k(P) + \epsilon$, which is a contradiction. Therefore, we obtain

$$P\left(\bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} \{\omega \mid \forall (F_m, A_m) \in \Theta'_m;\ F_m(\omega) \subset B_q(5M)\} \right) = 1.$$

□

Hereafter, M denotes a positive value satisfying inequalities (3) and (4). According to Lemma 4, for all $(F_n, A_n) \in \Theta'_n$, $F_n \in \mathcal{R}^*_k(5M)$ when n is sufficiently large. Since $\mathcal{R}^*_k(5M)$ is compact, $\Theta^*_k(5M)$ is also compact.

By the uniform SLLN, the continuity of $\Psi(\cdot, \cdot, P)$ on $\Theta^*_k(5M)$ and Lemma 4, the conclusion of the theorem for the cluster number k can be proved in the same manner as was done for the last part of the proof of the consistency theorem in Terada (2014).

Choose $\theta_* \in \Theta^*_k(5M)$ such that $d(\theta_*, \Theta') > 0$. Write

$$\tilde{\theta}_n = \begin{cases} \hat{\theta}_n & \text{if } \hat{\theta}_n \in \Theta^*_k(5M) \\ \theta_* & \text{if } \hat{\theta}_n \notin \Theta^*_k(5M) \end{cases}.$$

By Lemma 4, we have $\tilde{\theta}_n = \hat{\theta}_n$ for a sufficiently large n. Since $\Psi(\hat{\theta}_n, P_n) = \inf_{\theta\in\Xi_k} \Psi(\theta, P_n)$, we have

$$\limsup_n \left[\Psi\left(\tilde{\theta}_n, P_n\right) - \inf_{\theta\in\Theta'} \Psi(\theta, P_n) \right] \le 0 \quad \text{a.s.}$$

Since $\limsup_n \psi(\theta_0, P_n) = m_k(P)$ for any $\theta_0 \in \Theta'$,

$$\limsup_n \inf_{\theta \in \Theta'} \Psi(\theta, P_n) \leq \limsup_n \Psi(\theta_0, P_n) = m_k(P) \quad \text{a.s.}$$

Hence, we have

$$\begin{aligned} 0 &\geq \limsup_n \Psi\left(\tilde{\theta}_n, P_n\right) - \limsup_n \inf_{\theta \in \Theta'} \Psi(\theta, P_n) \\ &\geq \limsup_n \Psi\left(\tilde{\theta}_n, P_n\right) - m_k(P) \quad \text{a.s.} \end{aligned}$$

Let $\Theta_\epsilon^*(5M) := \{\theta \in \Theta_k^*(5M) \mid d(\theta, \Theta') \geq \epsilon\}$. By the uniform SLLN applied to $\Theta_k^*(5M)$, we obtain

$$\liminf_n \inf_{\theta \in \Theta_\epsilon^*(5M)} \Psi(\theta, P_n) \geq \inf_{\theta \in \Theta_\epsilon^*(5M)} \Psi(\theta, P) \quad \text{a.s.}$$

for all $\epsilon > 0$. Fix an arbitrary $\epsilon > 0$. By Corollary 1,

$$\liminf_n \inf_{\theta \in \Theta_\epsilon^*(5M)} \Psi(\theta, P_n) > \limsup_n \Psi\left(\tilde{\theta}_n, P_n\right) \quad \text{a.s.}$$

Thus, for any $\omega \in \Omega_1$ there exists $n_0 \in \mathbb{N}$ such that

$$\inf_{\theta \in \Theta_\epsilon^*(5M)} \Psi(\theta, P_n) > \Psi\left(\tilde{\theta}_n, P_n\right)$$

for all $n \geq n_0$. Conversely, suppose that $\mathrm{d}\left(\tilde{\theta}_n, \Theta'\right) \geq \epsilon$ for some $n \geq n_0$. Then, we have

$$\inf_{\theta \in \Theta_\epsilon^*(5M)} \Psi(\theta, P_n) = \Psi\left(\tilde{\theta}_n, P_n\right),$$

which is a contradiction. Thus, we obtain

$$\lim_{n \to \infty} \mathrm{d}\left(\tilde{\theta}_n, \Theta'\right) = 0 \quad \text{a.s.}$$

By $\tilde{\theta}_n = \hat{\theta}_n$ for a sufficiently large n, it follows that

$$\lim_{n \to \infty} \mathrm{d}\left(\hat{\theta}_n, \Theta'\right) = 0 \quad \text{a.s.}$$

Moreover, by the continuity of $\Psi(\cdot, P)$ on $\Theta_k^*(5M)$, we obtain

$$\lim_{n \to \infty} m_k(P_n) = m_k(P) \quad \text{a.s.}$$

7 Conclusion

In this study, we proved the strong consistency of FKM clustering under i.i.d. sampling using the frameworks of the proof for the consistency of k-means clustering (Pollard 1981) and the consistency of RKM clustering (Terada 2014). Based on these facts, we showed that there exists some cases in which FKM and RKM clusterings become equivalent as n goes to infinity. The compactness of the parameter space is not a requirement for the sufficient condition of the strong consistency for FKM clustering, as well as k-means clustering and RKM clustering. As with k-means clustering and RKM clustering, the proof of the consistency theorem is based on Blum–DeHardt uniform SLLN (Peskir 2000). Thus, for the consistency of FKM clustering, stationarity and ergodicity are only required and the i.i.d. condition is also not necessary. We also derived the sufficient condition for ensuring the existence of population global optimizers of FKM clustering. Moreover, we provided a rough large deviation inequality for FKM clustering.

Finally, as with Timmerman et al. (2010), we note that RKM clustering works well and FKM clustering does not work when the subspace containing the cluster structure has more variability than the orthogonal subspace. On the other hand, FKM clustering works well and RKM clustering does not work when the data have much variability in directions orthogonal to the subspace containing the cluster structure. Moreover, we mention that, since for all $k \in \mathbb{N}$ and all $q < p$

$$\inf_{(F,A)\in\Xi_k} \int \min_{\boldsymbol{f}\in F} \left\| A^T\boldsymbol{x} - \boldsymbol{f} \right\|^2 P_n(d\boldsymbol{x}) \leq \inf_{A\in\mathcal{O}(p\times q)} \int \left\| A^T\boldsymbol{x} \right\|^2 P_n(d\boldsymbol{x}),$$

FKM clustering does not work for a data matrix which is rank deficient or nearly rank deficient.

8 Appendix: Existence of Θ'

Here, we prove the existence of population global optimizers.

Lemma 5 *Suppose that* $\int \psi(\|\boldsymbol{x}\|)P(d\boldsymbol{x}) < \infty$. *There exists* $M > 0$ *such that*

$$\inf_{A\in\mathcal{O}(p\times q)} \Psi(F', A, P) > \inf_{\theta\in\Theta_k^*(M)} \Psi(\theta, P)$$

for all $F' \in \mathcal{R}_k$ *satisfying* $F' \cap B_q(M) = \emptyset$.

Proof Conversely, suppose that, for all $M > 0$, there exists $F' \in \mathcal{R}_k$ such that $F' \cap B_q(M) = \emptyset$ and

$$\inf_{A\in\mathcal{O}(p\times q)} \Psi(F', A, P) \leq \inf_{\theta\in\Theta_k^*(M)} \Psi(\theta, P).$$

Choose $r > 0$ to satisfy that the ball $B_p(r)$ has a positive P measure; that is $P(B_p(r)) > 0$. Let M be sufficiently large such that $M > r$ and that it satisfies inequality (3). Since $\|A^T\boldsymbol{x} - \boldsymbol{f}\| \geq \|\boldsymbol{f}\| - \|A^T\boldsymbol{x}\| > M - r$ for all $\boldsymbol{f} \notin B_q(M)$ and all $\boldsymbol{x} \in B_p(r)$, we have

$$\begin{aligned}\int \psi(\|\boldsymbol{x}\|)P(\boldsymbol{x}) &\geq \inf_{\theta\in\Theta_k^*(M)} \Psi(\theta, P) \geq \inf_{A\in\mathcal{O}(p\times q)} \Psi(F', A, P)\\ &\geq \inf_{A\in\mathcal{O}(p\times q)} \int_{\boldsymbol{x}\in B_p(r)} \min_{\boldsymbol{f}\in F'} \psi\left(\left\|A^T\boldsymbol{x} - \boldsymbol{f}\right\|\right) P(\mathrm{d}\boldsymbol{x})\\ &\geq \phi(M - r)P(B_p(r)).\end{aligned}$$

This is a contradiction. □

Lemma 6 *Suppose that* $\int \psi(\|\boldsymbol{x}\|)P(d\boldsymbol{x}) < \infty$, *and for* $j = 2, 3, \ldots, k-1$, $m_j(P) > m_k(P)$. *There exists* $M > 0$ *such that, for all* $F' \in \mathcal{R}_k$ *satisfying* $F' \not\subset B_q(5M)$,

$$\inf_{A\in\mathcal{O}(p\times q)} \Psi(F', A, P) > \inf_{\theta\in\Theta_k^*(5M)} \Psi(\theta, P).$$

Proof Choose $M > 0$ to be sufficiently large to satisfy inequalities (3) and (4). Suppose that, for all $M > 0$, there exists $F' \in \mathcal{R}_k$ satisfying $F' \not\subset B_q(5M)$ and

$$\inf_{A\in\mathcal{O}(p\times q)} \Psi(F', A, P) \leq \inf_{\theta\in\Theta_k^*(5M)} \Psi(\theta, P).$$

Let $\mathcal{R}_k'$ be the set of such F' and then

$$m_k(P) = \inf_{\theta\in\mathcal{R}_k'\times\mathcal{O}(p\times q)} \Psi(\theta, P).$$

According to Lemma 5, each $F' \in \mathcal{R}_k'$ includes at least one point on $B_q(M)$, say $\boldsymbol{f}_1$. For all $\boldsymbol{x}$ satisfying $\|\boldsymbol{x}\| < 2M$ and all $A \in \mathcal{O}(p \times q)$, we obtain

$$\left\|A^T\boldsymbol{x} - \boldsymbol{f}\right\| > 3M \quad \text{for all } \boldsymbol{f} \notin B_q(5M)$$

and

$$\left\|A^T\boldsymbol{x} - \boldsymbol{g}\right\| < 3M \quad \text{for all } \boldsymbol{g} \in B_q(M).$$

Thus,

$$\int_{\|\boldsymbol{x}\|<2M} \min_{\boldsymbol{f}\in F'} \psi\left(\left\|A^T\boldsymbol{x} - \boldsymbol{f}\right\|\right) P(d\boldsymbol{x}) = \int_{\|\boldsymbol{x}\|<2M} \min_{\boldsymbol{f}\in F^*} \psi\left(\left\|A^T\boldsymbol{x} - \boldsymbol{f}\right\|\right) P(d\boldsymbol{x}),$$

where the set F^* is obtained by deleting all points outside $B_q(5M)$ from F'. Since $\int_{\|\boldsymbol{x}\|\geq 2M} \psi(\|A^T\boldsymbol{x} - \boldsymbol{f}_1\|)P(d\boldsymbol{x}) \leq \lambda \int_{\|\boldsymbol{x}\|\geq 2M} \psi(\|\boldsymbol{x}\|)P(d\boldsymbol{x})$, we obtain that

$$\begin{aligned}
&\Psi(F_k', A, P) + \lambda \int_{\|\boldsymbol{x}\| \geq 2M} \psi(\|\boldsymbol{x}\|) P(\mathrm{d}\boldsymbol{x}) \\
&\geq \int_{\|\boldsymbol{x}\| < 2M} \min_{\boldsymbol{f} \in F^*} \psi\left(\left\|A^T \boldsymbol{x} - \boldsymbol{f}\right\|\right) P(\mathrm{d}\boldsymbol{x}) + \int_{\|\boldsymbol{x}\| \geq 2M} \psi\left(\left\|A^T \boldsymbol{x} - \boldsymbol{f}_1\right\|\right) P(\mathrm{d}\boldsymbol{x}) \\
&\geq \Psi(F^*, A, P) \geq m_{k-1}(P)
\end{aligned}$$

for all $A \in \mathcal{O}(p \times q)$. It follows that $m_k(P) + \epsilon \leq m_{k-1}(P)$, which is a contradiction. □

Let us consider $M > 0$ to be sufficiently large to satisfy inequalities (3) and (4). Write $\Theta_k := \mathcal{R}_k^*(5M) \times \mathcal{O}(p \times q)$. Proposition 1 and Corollary 1 can be proved in the same way as Proposition 1 and Corollary 1 in Terada (2014).

Proof of Proposition 1 According to Lemma 6,

$$\inf_{\theta \in \Xi_k} \Psi(\theta, P) = \inf_{\theta \in \Theta_k} \Psi(\theta, P).$$

Moreover, for any $\theta \in (\mathcal{R}_k \setminus \mathcal{R}_k^*(5M)) \times \mathcal{O}(p \times q)$, $m_k(P) < \Psi(\theta, P)$. Thus, we only have to prove $\Theta' \neq \emptyset$.

Let $C := \{\Psi(\theta, P) \mid \theta \in \Theta_k\}$ and then $m_k(P) = \inf C$. By the definition of the infimum, for all $x > m_k(P)$, there exists $c \in C$ such that $c < x$. By the axiom of choice, we can obtain a sequence $\{c_n\}_{n \in \mathbb{N}}$ such that $c_n \to m_k(P)$ as $n \to \infty$. Using the axiom of choice again, we can obtain a sequence $\{\theta_n\}_{n \in \mathbb{N}}$ such that $\Psi(\theta_n, P) \to m_k(P)$ as $n \to \infty$.

By the compactness of Θ_k, there exists a convergent subsequence of $\{\theta_n\}_{n \in \mathbb{N}}$, say $\{\theta_{n_i}\}_{i \in \mathbb{N}}$. Let $\theta_* \in \Theta_k$ denote the limit of subsequence $\{\theta_{n_i}\}_{i \in \mathbb{N}}$, i.e., $\theta_{m_i} \to \theta_*$ as $i \to \infty$. Since $\Psi(\cdot, P)$ is continuous on Θ_k, $\Psi(\theta_*, P) = m_k(P)$. Hence, we obtain $\Theta' \neq \emptyset$. □

Proof of Corollary 1 Let $\Theta_\epsilon := \{\theta_k \in \Theta_k \mid \Psi(\theta_k, P) = m_k(P)\}$. Conversely, suppose that there exists $\epsilon > 0$ such that $\inf_{\theta \in \Theta_\epsilon} \Psi(\theta, P) = \inf_{\theta \in \Theta'} \Psi(\theta, P)$. By the definition of the infimum, there exists a sequence $\{\theta_n\}_{n \in \mathbb{N}}$ on Θ_ϵ such that $\Psi(\theta_n, P) \to m_k(P)$ as $n \to \infty$. By compactness of Θ_k, there exists a convergent subsequence of $\{\theta_n\}_{n \in \mathbb{N}}$, say $\{\theta_{m_i}\}_{i \in \mathbb{N}}$. Let $\theta_* \in \Theta_k$ denote the limit of subsequence $\{\theta_{m_i}\}_{i \in \mathbb{N}}$. Since $\theta_{m_i} \to \theta_*$ as $i \to \infty$, we have $d(\theta_{m_i}, \theta_*) < \epsilon$ for a sufficiently large i, which is a contradiction. □

Acknowledgments The author wishes to express his thanks to Dr. Michio Yamamoto for his helpful comments and discussions related to Example 2. The author gratefully acknowledges the helpful comments and suggestions of the two anonymous reviewers. Moreover, the author also thank Professor Yutaka Kano for his supervision of this research. This work was supported by Grant-in-Aid for JSPS Fellows Number 24 · 2466.

References

Akama, Y., Irie, K., Kawamura, A., Uwano, Y. (2010). VC dimensions of principal component analysis. *Discrete and Computational Geometry*, *44*(3), 589–598.

Arabie, P., Hubert, L. (1994). Cluster Analysis in Marketing Research. In R. P. Bagozzi (Ed.), *Advanced methods of marketing research* (pp. 160–189). Oxford: Blackwell.

Chang, W. (1983). On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*, *32*(3), 267–275.

De Soete, G., Carroll, J. D. (1994). K-means clustering in a low-dimensional Euclidean space. In E. Diday, Y. Lechevallier, M. Schader, P. Bertrand, & B. Burtschy (Eds.), *New approaches in classification and data analysis* (pp. 212–219). Berlin: Springer.

Dehardt, J. (1971). Generalizations of the Glivenko-Cantelli theorem. *Annals of Mathematical Statistics*, *42*(6), 2050–2055.

Devroye, L., Györfi, L., Lugosi, G. (1996). *A probabilistic theory of pattern recognition*. New York: Springer.

Peskir, G. (2000). *From uniform laws of large numbers to uniform ergodic theorems*. Lecture notes series. 66, University of Aarhus, Department of Mathematics.

Pfanzagl, J. (1994). *Parametric statistical theory*. Berlin: de Gruyter.

Pollard, D. (1981). Strong consistency of k-means clustering. *Annals of Statistics*, *9*(1), 135–140.

Pollard, D. (1982). Quantization and the method of k-means. *IEEE Transactions of Information Theory*, *28*(2), 199–205.

Terada, Y. (2014). Strong consistency of reduced k-means clustering. To appear in *Scandinavian Journal of Statistics*.

Timmerman, M. E., Ceulemans, E., Kiers, H. A. L., Vichi, M. (2010). Factorial and reduced K-means reconsidered. *Computational Statistics and Data Analysis*, *54*(7), 1858–1871.

Vichi, M., Kiers, H. A. L. (2001). Factorial k-means analysis for two-way data. *Computational Statistics and Data Analysis*, *37*(1), 49–64.