Strong Consistency of Factorial $K$-means Clustering

Yoshikazu Terada

Graduate School of Engineering Science, Osaka University, 1-3 Machikaneyama, Toyonaka, Osaka, Japan

e-mail: terada@sigmath.es.osaka-u.ac.jp

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 the reduced $k$-means clustering (RKM) 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.

Keywords and phrases: 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 the usual $k$-means clustering for the principal scores on the first few principal components, which are obtained in Step 1.

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 (1994); De Soete and Carroll (1994)). The first few principle 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 $x_1, \ldots, x_n$ 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
Fig 1. Artificial data used to evaluate RKM clustering: (a) plot of two variables related to a cluster structure and (b) heat map of 12 variables.

by

$$RKM_n(F, A) := \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq k} \| x_i - Af_j \|^2,$$

where $f_j \in \mathbb{R}$, $F = \{ f_1, \ldots, f_k \} \subset \mathbb{R}^q$, $A$ is a $p \times q$ column-wise orthonormal matrix, and $\| \cdot \|$ represents the usual norm. Under certain regularity conditions, RKM clustering has strong consistency (Terada (2012)). However, when the data matrix $X = (x_{ij})_{n \times p}$ has a full rank, i.e., $\text{rank}(X) = p$, RKM clustering may fail to find a subspace that reflects the cluster structure. Indeed, RKM clustering has been applied to data composed of a total of 12 independent variables (Figure 1), which consists of 2 variables actually related to the cluster structure and 10 noise variables. The result of RKM clustering for the data shown in Figure 1 is given in Figure 2. The results indicate that the low-dimensional subspace revealed 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 the RKM clustering method and proposed a new clustering method, called factorial $k$-means (FKM) clustering. For the given data points $x_1, \ldots, x_n$ 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:

$$FKM_n(F, A \mid k, q) := \frac{1}{n} \sum_{i=1}^{n} \min_{1 \leq j \leq k} \| A^T x_i - f_j \|^2,$$

where $F := \{ f_1, \ldots, f_k \}$, $f_j \in \mathbb{R}^q$ and $A$ is a $p \times q$ column-wise orthonormal matrix. When the given data points $x_1, \ldots, x_n$ are independently drawn from
a population distribution \( P \), we can rewrite the FKM objective function as
\[
FKM(F, A, P_n) := \int \min_{f \in F} \|A^T x - f\|^2 P_n(dx),
\]
where \( P_n \) is the empirical measure of the data points \( x_1, \ldots, x_n \) in \( \mathbb{R}^p \). For each set of cluster centers \( F \) and each \( p \times q \) orthonormal matrix \( A \), we obtain
\[
\lim_{n \to \infty} FKM(F, A, P_n) = FKM(F, A, P) := \int \min_{f \in F} \|A^T x - f\|^2 P(dx) \quad \text{a.s.}
\]
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 regular 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 (2012)). 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 and prove the consistency of FKM under the mild condition, as with Terada (2012). In the proof of main theorem, first we also show that the optimal sample centres eventually lie in some compact regions on \( \mathbb{R}^p \) as with Pollard (1981) and Terada (2012) 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 (2012). For an arbitrary \( p \times q \) column-wise orthonormal matrix \( A \) (\( A^T A = I_q, \ q < p \)), an arbitrary \( p \)-dimensional point \( x \in \mathbb{R}^p \) and an arbitrary \( q \)-dimensional point \( y \in \mathbb{R}^q \), the key inequality in this paper is that \( \|A^T x\| \leq \|x\| \) while the key
equation in the strong consistency of RKM clustering (Terada (2012)) is that \( \|A y\| = \|y\| \).

The rest of the paper is organized as follows. In Section 2, we describe the clustering algorithm of FKM to get the local minimum and the relationship between RKM clustering and FKM clustering. We introduce prerequisites and notation in Section 3. In Section 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 of FKM clustering are stated in Section 5. In Section 6, we provide the main proof of the theorem.

2. Factorial K-means clustering

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

\[
FKM_n(A, F, U | k, q) := \|XA - UF\|_F^2 = \sum_{i=1}^{n} \min_{1 \leq j \leq k} \|A^T x_i - f_j\|^2,
\]

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 \( f_j \) (\( j = 1, \ldots, k \)) are row vectors of \( F \) representing the \( j \)th 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

\[
u_{ij} = \begin{cases} 1 & \text{if } \|A^T x_i - f_j\|^2 < \|A^T x_i - f_{j'}\|^2 \text{ for each } j' \neq j, \\ 0 & \text{otherwise}. \end{cases}
\]

**Step 2.** \( A \) is updated by the first \( q \) eigenvectors of \( X^T [U(U^T U)^{-1} U^T - I_n] X \), where \( I_n \) is the \( n \)-dimensional identity matrix.

**Step 3.** \( F \) is updated using \( (U^T U)^{-1} U^T X A \).

**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 Figure 1.

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

\[
RKM_n(A, F, U) := \|X - UFA^T\|_F^2 = \sum_{i=1}^n \min_{1 \leq j \leq k} \|x_i - Af_j\|^2.
\]

This objective function can be decomposed into two terms:

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

The first term of equation (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 some cases. For more details about the relationship between RKM and FKM clustering, see Timmerman et al. (2010).

3. Preliminaries

In this paper, the similar notations as ones used in Pollard (1981) and Terada (2012) are used. Let \((\Omega, \mathcal{F}, P)\) be a probability space, and \(X_1, \ldots, X_n\) be i.i.d. \(p\)-dimensional random variables drawn from a distribution \(P\). Let \(P_n\) denote the empirical measure based on \(X_1, \ldots, X_n\). 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 \(\mathcal{R}_k := \{R \subset \mathbb{R}^q \mid \#(R) \leq k\}\), where \(\#(E)\) is the cardinality of \(E\). We will denote the parameter space by \(\Xi_k := \mathcal{R}_k \times \mathcal{O}(p \times q)\). For each \(M > 0\), \(\mathcal{R}_k^*(M) := \{E \subset \mathbb{R}^q \mid \#(E) \leq k\text{ and }E \subset B_q(M)\}\) 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 decreasing function. For each subset \(F \subset \mathbb{R}^q\)
and each $A \in \mathcal{O}(p \times q)$, the FKM clustering loss function with a probability measure $Q$ on $\mathbb{R}^p$ is defined by

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

Write

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

and

$$m^*_k(Q | M) := \inf_{(F, A) \in \Theta^*_k(M)} \Psi(F, A, Q).$$

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 $\Theta' := \{\theta \in \Xi_k | m_k(P) = \Psi(\theta, P)\}$ and $\Theta^*_n := \{\theta \in \Xi_k | m_k(P_n) = \Psi(\theta, P_n)\}$, respectively. For each $M > 0$, let $\Theta^*_n := \{\theta \in \Theta^*_n(M) | m_k(P_n | M) = \Psi(\theta, P_n)\}$ and $\Theta^*_n := \{\theta \in \Theta^*_n(M) | m_k(P_n | M) = \Psi(\theta, P_n)\}$. When we emphasize that $\Theta'$ and $\Theta^*_n$ are dependent on the index $k$, we write $\Theta'(k)$ and $\Theta^*_n(k)$ instead of $\Theta'$ and $\Theta^*_n$, respectively. One of the measurable estimators in $\Theta^*_n$ 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 $\Theta^*_n$. 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

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

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 (2012) the distance between $\hat{\theta}_n$ and $\Theta'$ is defined as

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

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

$$\int \psi(\|A^T x - f\|) P(dx) \leq \int \psi(\|A^T x\| + \|f\|) P(dx)$$

$$\leq \int \psi(\|x\| + \|f\|) P(dx)$$

$$\leq \int \psi(2\|f\|) P(dx) + \int_{\|f\| \leq \|x\|} \psi(2\|x\|) P(dx)$$

$$\leq \psi(2\|f\|) + \lambda \int \psi(\|x\|) P(dx)$$
for all \( f \in F \) 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 \( \int \psi(||x||)P(dx) < \infty \).

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

\[
\int \psi(||A^T x - f||)P(dx) = \int \psi(||RA^T x - Rf||)P(dx).
\]

Hence, \( \Theta' \) 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 \( g(F, A)(x) := \min_{f \in F} \psi(||A^T x - f||) \), where \( (F, A) \) takes all values over \( \Theta^*_k(M) \). Suppose that \( \int \psi(||x||)P(dx) < \infty \). Then,

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

**Proof.** Dehuardt (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 \( \hat{g} \) and \( \bar{g} \) in \( \mathcal{G}_\epsilon \), with \( \hat{g} \leq g \leq \bar{g} \) and \( \int \hat{g}(x)P(dx) - \int \hat{g}(x)P(dx) < \epsilon \).

Choose an arbitrary \( \epsilon > 0 \). Let \( S_{p \times q}(\sqrt{\epsilon}) := \{ X \in \mathbb{R}^{p \times q} \mid \|X\|_F = \sqrt{\epsilon} \} \). We will denote by \( D_{\delta_1} \) the finite set on \( \mathbb{R}^q \) satisfying the condition that, for all \( f \in B_q(M) \), there exists \( g \in D_{\delta_1} \) such that \( \|f - g\| < \delta_1 \). Similarly, we will denote by \( A_{p \times q, \delta_2} \) the finite set on \( S_{p \times q}(\sqrt{\epsilon}) \) satisfying the condition that, for all \( A \in S_{p \times q}(\sqrt{\epsilon}) \), there exists \( B \in A_{p \times q, \delta_2} \) 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

\[
\min_{f \in F} \psi(||A^T x - f|| + \delta_1 + \delta_2\|x\|) \quad \text{or} \quad \min_{f \in F} \psi(||A^T x - f|| - \delta_1 - \delta_2\|x\|),
\]

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

For any \( F = \{ f_1, \ldots, f_k \} \in \mathcal{R}_k^*(M) \), there exists \( F_* = \{ f^*_1, \ldots, f^*_k \} \in \mathcal{R}_{k, \delta_1} \) with \( \|f_i - f^*_i\| < \delta_1 \) for each \( i \). In addition, since \( \mathcal{O}(p \times q) \subset \cup_{A_* \in A_{p \times q, \delta_2}} \{ A \mid \|A - A_*\|_F < \delta_2 \} \), for any \( A \in \mathcal{O}(p \times q) \), there exists \( A_* \in A_{p \times q, \delta_2} \) with \( \|A - A_*\|_F < \delta_2 \). Corresponding to each \( g(F, A) \in \mathcal{G} \), choose

\[
\hat{g}(F, A)(x) := \min_{f \in F_*} \psi(||A^T x - f|| + \delta_1 + \delta_2\|x\|)
\]

and

\[
\bar{g}(F, A)(x) := \min_{f \in F_*} \psi(||A^T x - f|| - \delta_1 - \delta_2\|x\|).
\]
Since $\psi$ is a monotone function and
\[ \| A_i^T x - f_i^* \| - \delta_1 - \delta_2 \| x \| \leq \| A_i^T x - f_i \| \leq \| A_i^T x - f_j^* \| + \delta_1 + \delta_2 \| x \| \]
for each $i$ and each $x \in \mathbb{R}^p$, we have $\hat{g}(F, A) \leq g(F, A) \leq \hat{g}(F, A)$.

Choosing $R > 0$ to be greater than $(M + \delta_1)/\sqrt{q}$ (or $(M + \delta_1)/(\sqrt{q} + \delta_2))$, we obtain
\[ \int \left[ \hat{g}(F, A)(x) - \hat{g}(F, A)(x) \right] P(dx) \]
\[ \leq \int \sum_{i=1}^k \left[ \psi(\| A_i^T x - f_i^* \| + \delta_1 + \delta_2 \| x \|) - \psi(\| A_i^T x - f_i^* \| - \delta_1 - \delta_2 \| x \|) \right] P(dx) \]
\[ \leq k \sup_{\| x \| \leq R} \sup_{f \in B_d(M)} \sup_{A \in S_p(x, \sqrt{q})} \left[ \psi(\| A_i^T x - f \| + \delta_1 + \delta_2 \| x \|) \right. \]
\[ \left. - \psi(\| A_i^T x - f \| - \delta_1 - \delta_2 \| x \|) \right] + 2k\lambda_m \int_{\| x \| \geq R} \psi(\| x \|) P(dx), \]
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 $\psi$ is uniform 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 $\int \psi(\| x \|) P(dx) < \infty$. 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 $(F, A) \in \Theta_k(M)$ is chosen to satisfy $d_H(F, G) < \delta_1$ and $\| A - B \|_F < \delta_2$, then for each $g \in G$ there exists $f(g) \in F$ such that $\| g - f(g) \| < \delta_1$. Choosing $R$ to be larger than $M + \delta_1$, we obtain
\[ \Psi(F, A, P) - \Psi(G, B, P) \]
\[ = \int \left[ \min_{f \in F} \psi(\| A_i^T x - f \|) - \min_{g \in G} \psi(\| B_i^T x - g \|) \right] P(dx) \]
\[ \leq \int \max_{g \in G} \left[ \psi(\| A_i^T x - f(g) \|) - \psi(\| B_i^T x - g \|) \right] P(dx) \]
\[ \leq \int \sum_{g \in G} \left[ \psi(\| B_i^T x - g \| + \delta_1 + \delta_2 \| x \|) - \psi(\| B_i^T x - g \|) \right] P(dx) \]
\[ \leq k \sup_{\| x \| \leq R} \max_{g \in G} \left[ \psi(\| B_i^T x - g \| + \delta_1 + \delta_2 \| x \|) - \psi(\| B_i^T x - g \|) \right] \]
\[ + 2 \sum_{g \in G} \int_{\| x \| \geq R} \psi(\| B_i^T x - g \| + \delta_1 + \delta_2 \| x \|) P(dx) \]
\[ \leq k \sup_{\| x \| \leq R} \max_{g \in G} \left[ \psi(\| B_i^T x - g \| + \delta_1 + \delta_2 \| x \|) - \psi(\| B_i^T x - g \|) \right] \]

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 $\epsilon$. Since for each $f \in F$ there exists $g(f) \in G$ such that $\|g - g(f)\| < \delta_1$, 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 $\Theta'$ is empty. Therefore, first, we provide sufficient conditions for the existence of population global optimizers.

**Proposition 1.** Suppose that $\int \psi(\|x\|)P(dx) < \infty$ 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 A.

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

**Corollary 1.** Suppose that $\int \psi(\|x\|)P(dx) < \infty$ 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 $M > M_0$

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

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

**Proof.** See Appendix A.

5.2. Strong consistency of FKM clustering

If the parameter space is restricted to $\Theta^*_\epsilon(M) \subset \Xi_k$, we easily obtain the strong consistency of FKM clustering. Since $\Theta^*_\epsilon(M)$ 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 $\Theta^*_\epsilon(M) := \{\theta \in \Theta^*_\epsilon(M) \mid d(\theta, \Theta') \geq \epsilon\}$.

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

$$\lim_{n \to \infty} d(\hat{\theta}^*_n, \Theta^*) = 0 \text{ a.s., and } \lim_{n \to \infty} m^*_k(P_n \mid M) = m^*_k(P \mid M) \text{ a.s.}$$
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 necessary take an inductive form with the number of cluster \( k \) and we can prove the consistency under the mild condition.

We cannot assume the uniqueness condition since FKM clustering has rotational indeterminacy. In this study, as Terada (2012) 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 centres has \( k \) distinct elements. When we do not use the fact in Proposition 2, we may need more strict condition \( m_1(P) > m_2(P) > \cdots > m_k(P) \) and the proof of the main theorem takes an inductive form with the number of cluster \( k \) as with Pollard (1981). The following theorem provides sufficient conditions for the strong consistency of FKM clustering.

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

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

**Proof.** See Section 5.

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.

6. **Proof of the theorem**

Since the theorem deals with almost sure convergence, there might exist null subsets of \( \Omega \) on which the strong consistency does not hold. Therefore, throughout the proof, \( \Omega_1 \) denotes the set obtained by avoiding a possible null set from \( \Omega \).

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 \( \int \psi(\|x\|)P(dx) < \infty \). 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

\[
\psi(M - r)P(B_p(r)) > \int \psi(\|x\|)P(dx).
\]
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

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

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

$$
\int \psi(||A^T x||) P(dx) \leq \int \psi(||x||) P(dx)
$$

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

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

$$
\limsup_l \Psi(F_{n_l}, A_{n_l}, P_{n_l}) \geq \limsup_l \frac{1}{n_l} \sum_{i \in \{i \mid X_i \in K\}} \min_{f \in F_{n_l}} \psi(||A_{n_l}^T X_i - f||)
$$

$$
\geq \limsup_l \frac{1}{n_l} \sum_{i \in \{i \mid X_i \in K\}} \psi(M - r)
$$

$$
\geq \psi(M - r) P(B_p(r)).
$$

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

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

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

$$\lambda \int_{\|x\| \geq 2M} \psi(\|x\|) P_n(dx) < \epsilon. \quad (4)$$

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 $f_1 \in B_q(M)$. Since $\{x \mid \|A^T x\| \geq 2M\} \subset \{x \mid \|x\| \geq 2M\}$, we have

$$\int_{\|A^T x\| \geq 2M} \psi(\|A^T x - f_1\|) P_n(dx) \leq \int_{\|x\| \geq 2M} \psi(\|A^T x - f_1\|) P_n(dx)$$

$$\leq \int_{\|x\| \geq 2M} \psi(\|x\| + \|f_1\|) P_n(dx)$$

$$\leq \lambda \int_{\|x\| \geq 2M} \psi(\|x\|) P_n(dx)$$

for all $A \in O(p \times q)$. 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 $A \in O(p \times q)$, we have

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

for all $A \in O(p \times q)$. For each $x \in B_q(2M)$ and each $A \in O(p \times q)$, we have

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

and

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

Thus, we obtain

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

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

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 $F_{n_l}(\omega) \not\subset B_q(5M)$. 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

$$m_{k-1}(P) \leq m_{k-1}^*(P \mid 5M) \leq \lim_{l \to \infty} \Psi(F_{n_l}, A_{n_l}, P_{n_l}) \leq \lim\sup_n \Psi(F_{n_l}, A_{n_l}, P_{n_l})$$

$$\leq \lim\sup_n \left[ \int_{\|x\| < 2M} \min_{f \in F_n} \psi(\|A^T x - f\|) P_n(dx) \right]$$
\[ + \int_{\|x\| \geq 2M} \psi(\|A_n^T x - f_1\|) P_n(dx) \]
\[ \leq \limsup_n \left[ \Psi(F_n, A_n, P_n) + \lambda \int_{\|x\| \geq 2M} \psi(\|x\|) P_n(dx) \right] \]
\[ \leq \limsup_n \Psi(F, A, P_n) + \lambda \int_{\|x\| \geq 2M} \psi(\|x\|) P_n(dx). \]
\[(5)\]

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

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 (2012).

Choose \(\theta_0 \in \Theta_k^*(5M)\) such that \(d(\theta_0, \Theta') > 0\). Write
\[ \hat{\theta}_n = \begin{cases} \hat{\theta}_n & \text{if } \hat{\theta}_n \in \Theta_k^*(5M) \\ \theta_0 & \text{if } \hat{\theta}_n \notin \Theta_k^*(5M) \end{cases}. \]

By Lemma 4, we have \(\hat{\theta}_n = \hat{\theta}_n\) for a sufficiently large \(n\). Since \(\Psi(\hat{\theta}_n, P_n) = \inf_{\hat{\theta} \in \Xi_k} \Psi(\hat{\theta}, P_n)\), we have
\[ \limsup_n \left[ \Psi(\hat{\theta}_n, P_n) - \inf_{\hat{\theta} \in \Theta'} \Psi(\hat{\theta}, P_n) \right] \leq 0 \quad \text{a.s.} \]

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
\[ 0 \geq \limsup_n \Psi(\hat{\theta}_n, P_n) - \limsup_n \inf_{\theta \in \Theta'} \Psi(\theta, P_n) \]
\[ \geq \limsup_n \Psi(\hat{\theta}_n, P_n) - m_k(P) \quad \text{a.s.} \]
Let $\Theta^*_M := \{ \theta \in \Theta^*_k(5M) \mid d(\theta, \Theta') \geq \epsilon \}$. By the uniform SLLN applied to $\Theta^*_M$, we obtain

$$\liminf_{n} \inf_{\theta \in \Theta^*_M} \Psi(\theta, P_n) \geq \inf_{\theta \in \Theta^*_M} \Psi(\theta, P) \quad \text{a.s.}$$

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

$$\liminf_{n} \inf_{\theta \in \Theta^*_M} \Psi(\theta, P_n) > \limsup_{n} \Psi(\tilde{\theta}_n, P_n) \quad \text{a.s.}$$

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

$$\inf_{\theta \in \Theta^*_M} \Psi(\theta, P_n) > \Psi(\tilde{\theta}_n, P_n)$$

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

$$\inf_{\theta \in \Theta^*_M} \Psi(\theta, P_n) = \Psi(\tilde{\theta}_n, P_n),$$

which is a contradiction. Thus, we obtain

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

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

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

Moreover, by the continuity of $\Psi(\cdot, P)$ on $\Theta^*_M(5M)$, 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 by using the frameworks of the proof for the consistency of $k$-means clustering (Pollard (1981)) and the consistency of RKM clustering (Terada (2012)). The compactness of 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 the $k$-means and RKM clustering, the proof is based on Blum-DeHardt uniform SLLN (Peskir (2000)). Thus, for the consistency of FKM clustering, stationarity and ergodicity is 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 proved the uniform SLLN and continuity of the FKM objective function in the proof of the consistency theorem.

In the future, we will derive the rate of convergence of FKM clustering estimators.
References

[1] 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.

[2] Chang, W. (1994). On using principal components before separating a mixture of two multivariate normal distributions. *Applied Statistics*. 32, 267–275.

[3] 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 and B. Burtschy (Eds.), *New Approaches in Classification and Data Analysis* (pp. 212–219). Berlin, Springer-Verlag.

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

[5] Peskir, G. (2000). From Uniform Laws of Large Numbers to Uniform Ergodic Theorems. *Lecture notes series*. 66, University of Aarhus, Department of Mathematics.

[6] Pfanzagl, J. (1994). *Parametric Statistical Theory*. de Gruyter, Berlin.

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

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

[9] Terada, Y. (2012). Strong consistency of reduced k-means clustering. *arXiv*.

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

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

Appendix A: Existence of $\Theta'$

Here we prove the existence of population global optimizers.

**Lemma 5.** Suppose that $\int \psi(\|x\|)P(dx) < \infty$. 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 \( ||A^T x - f|| \geq ||f|| - ||A^T x|| > M - r \) for all \( f \not\in B_q(M) \) and all \( x \in B_p(r) \), we have

\[
\int \psi(||x||)P(x) \geq \inf_{\theta \in \Theta^*_1(M)} \Psi(\theta, P) \geq \inf_{A \in \mathcal{O}(p \times q)} \Psi(F', A, P) \geq \inf_{A \in \mathcal{O}(p \times q)} \int_{x \in B_p(r)} \min_{f \in F'} \psi(||A^T x - f||)P(dx) \geq \phi(M - r)P(B_p(r)).
\]

This is a contradiction. \( \square \)

**Lemma 6.** Suppose that \( \int \psi(||x||)P(dx) < \infty \), 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 \( F' \not\subset B_q(5M) \),

\[
\inf_{A \in \mathcal{O}(p \times q)} \Psi(F', A, P) \geq \inf_{\theta \in \Theta^*_1(5M)} \Psi(\theta, P).
\]

**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

\[
\inf_{A \in \mathcal{O}(p \times q)} \Psi(F', A, P) \leq \inf_{\theta \in \Theta^*_1(5M)} \Psi(\theta, P).
\]

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 \( f_1 \). For all \( x \) satisfying \( ||x|| < 2M \) and all \( A \in \mathcal{O}(p \times q) \), we obtain

\[
||A^T x - f|| > 3M \quad \text{for all} \quad f \not\in B_q(5M)
\]

and

\[
||A^T x - g|| < 3M \quad \text{for all} \quad g \in B_q(M).
\]

Thus,

\[
\int_{||x|| < 2M} \min_{f \in F'} \psi(||A^T x - f||)P(dx) = \int_{||x|| < 2M} \min_{f \in F^*} \psi(||A^T x - f||)P(dx),
\]

where the set \( F^* \) is obtained by deleting all points outside \( B_q(5M) \) from \( F' \). Since \( \int_{||x|| \geq 2M} \psi(||A^T x - f_1||)P(dx) \leq \lambda \int_{||x|| \geq 2M} \psi(||x||)P(dx) \), we obtain that

\[
\Psi(F', A, P) + \lambda \int_{||x|| \geq 2M} \psi(||x||)P(dx) \geq \int_{||x|| < 2M} \min_{f \in F^*} \psi(||A^T x - f||)P(dx) + \int_{||x|| \geq 2M} \psi(||A^T x - f_1||)P(dx) \geq \Psi(F^*, A, P) \geq m_k(P)
\]

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. \( \square \)
Let us consider $M > 0$ to be sufficiently large to satisfy inequalities (3) and (4). Write $\Theta_k := R_k^* (5M) \times O(p \times q)$. Proposition 1 and Corollary 1 can be proved in the same way as Proposition 1 and Corollary 1 in Terada (2012).

**Proof of Proposition 1.** According to Lemma 6,

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

Moreover, for any $\theta \in (R_k \setminus R_k^* (5M)) \times O(p \times q)$, $m_k(P) < \Psi(\theta, P)$. 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 $\Theta_k$, 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 $\{\theta_{m_i}\}_{i \in \mathbb{N}}$, i.e., $\theta_{m_i} \to \theta_*$ as $i \to \infty$. Since $\Psi(\cdot, P)$ is continuous on $\Theta_k$, $\Psi(\theta_*, P) = m_k(P)$. Hence, we obtain $\Theta' \neq \emptyset$. 

**Proof of Corollary 1.** Let $\Theta_\varepsilon := \{\theta_k \in \Theta_k \mid \Psi(\theta_k, P) = m_k(P)\}$. Conversely, suppose that there exists $\epsilon > 0$ such that $\inf_{\theta \in \Theta_k} \Psi(\theta, P) = \inf_{\theta \in \Theta_\epsilon} \Psi(\theta, P)$. By the definition of the infimum, there exists a sequence $\{\theta_n\}_{n \in \mathbb{N}}$ on $\Theta_k$ such that $\Psi(\theta_n, P) \to m_k(P)$ as $n \to \infty$. By compactness of $\Theta_k$, 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.