Fuzzy PCA-Guided Robust k-Means Clustering

Katsuhiro Honda, *Member, IEEE*, Akira Notsu, and Hidetomo Ichihashi, *Member, IEEE*

***Abstract*—This paper proposes a new approach to robust clustering, in which a robust k-means partition is derived by using a noise-rejection mechanism based on the noise-clustering approach. The responsibility weight of each sample for the k-means process is estimated by considering the noise degree of the sample, and cluster indicators are calculated in a fuzzy principal-component-analysis (PCA) guided manner, where fuzzy PCA-guided robust k-means is performed by considering responsibility weights of samples. Then, the proposed method achieves cluster-core estimation in a deterministic way. The validity of the derived cluster cores is visually assessed through distance-sensitive ordering, which considers responsibility weights of samples. Numerical experiments demonstrate that the proposed method is useful for capturing cluster cores by rejecting noise samples, and we can easily assess cluster validity by using cluster-crossing curves.**

***Index Terms*—Clustering, data mining, kernel trick, principal-component analysis (PCA).**

I. Introduction

A DETERMINISTIC procedure for k-means clustering was proposed by Ding and He [1], based on the close relation between principal component analysis (PCA) and k-means clustering. k-means [2] is a popular clustering method that uses prototypes (centroids) to represent clusters by minimizing within-cluster errors. PCA has been often jointly used with clustering techniques, especially in pattern-recognition tasks. The combined applications can roughly be classified into three categories: 1) dimension reduction by PCA before clustering [3], [4]; 2) clusterwise local PCA after clustering [5]–[7]; and 3) initialization of model parameters in clustering by PCA [8], [9], i.e., PCA is performed to preprocess or postprocess clustering tasks. On the other hand, Ding and He introduced a new technique to directly apply PCA to cluster-indicator estimation in clustering tasks. In the PCA-guided k-means [1], the objective function of k-means clustering is redefined by a centroidless formulation, and the relaxed cluster-indicator vectors that represent cluster memberships are calculated by a PCA-like manner, in which the indicator vectors are identified with the eigenvectors of a within-cluster (inner product) similarity matrix, i.e., a continuous (relaxed) solution of the cluster membership indicators in k-means is identified with principal components in PCA.

This paper considers a new robust k-means algorithm that is based on a fuzzy PCA-guided clustering procedure. Fuzzy PCA [10] is a fuzzy version of the conventional PCA in which covariance structure of datasets are analyzed by considering the fuzzy-membership degree of data samples. In the proposed method, a responsibility weight of each sample for the k-means process is estimated based on the noise-fuzzy-clustering mechanism [11], [12] that is identified with the robust M-estimation technique [13] in the single-cluster case. Cluster membership indicators in the k-means process are derived as fuzzy principal components by considering the responsibility weights in fuzzy PCA. In this sense, the proposed method is a fuzzified PCA-guided robust k-means method.

The proposed method has some connections with the cluster-core concepts. Trimmed k-means [14], [15] extracts k distinct cluster cores by trimming noise samples previous to the conventional k-means process. Yang *et al.* [16] extended the cluster-core concept to fuzzy-clustering models with α-cut implementation. In the α-cut implementation, samples within cluster cores have full membership degree, while samples out of cluster cores have a relatively small (fuzzy) membership degree. Yang *et al.* demonstrated that the fuzzy approach outperforms the trimmed k-means in the sense of the sensitivity to initialization. Noise clustering [11] and possibilistic fuzzy c-means clustering [17] also extract cluster cores in a fuzzy manner by decreasing the responsibility of noise samples, while they do not extract crisp cores because even core samples have fuzzy-membership degrees. Possibilistic c-means [18] and its relatives [19] detect cluster cores independently in each cluster by giving up the probabilistic constraint in the k-means-type clustering techniques. These alternate optimization approaches, however, suffer from the initialization problem, and we often have several different results with the multistart strategy. Then, we need to evaluate the cluster validity to select the optimal one.

Several sequential approaches were also proposed to extract cluster cores one by one. Sequential fuzzy-cluster extraction (SFCE) [20] estimates memberships for each cluster core by using the eigenvector corresponding to the largest eigenvalue of a modified similarity matrix in each iteration step. A similar concept has been extended to spectral clustering with graph-based approaches [21], [22]. Although the sequential cluster extraction is performed in a deterministic manner, we need to iterate the cluster-core estimation k times to identify k clusters by using a different (modified) objective function in each step.

The proposed fuzzy PCA-guided robust k-means (FPR k-means) performs k cluster-core identification in a deterministic manner where cluster indicators for k clusters are calculated in a batch process considering responsibility weights of samples, while the responsibility weights are estimated in an iterative optimization frame.

The remainder of this paper is organized as follows: Section II briefly reviews the PCA-guided k-means and robust clustering algorithms. Section III proposes fuzzy PCA-guided robust

Manuscript received November 20, 2008; accepted August 26, 2009. First published November 13, 2009; current version published February 5, 2010. This work was supported in part by the Ministry of Internal Affairs and Communications, Japan, under the Strategic Information and Communications R&D Promotion Programme and in part by the Ministry of Education, Culture, Sports, Science and Technology, Japan, under Grant-in-Aid for Scientific Research (20700214).

The authors are with the Department of Computer Science and Intelligent Systems, Osaka Prefecture University, Osaka 599-8531 Japan (e-mail: honda@cs.osakafu-u.ac.jp; notsu@cs.osakafu-u.ac.jp; ichi@cs.osakafu-u.ac.jp).

Digital Object Identifier 10.1109/TFUZZ.2009.2036603

k-means (FPR k-means) that is a new approach to robust cluster-core identification in k-means clustering. Section IV presents several experimental results to reveal characteristic features of the proposed methods. Section V gives summary conclusions.

II. Principle-Component Analysis-Guided k-Means and Robust Clustering

A. k-Means by PCA-Guided Manner

Assume that we have n samples with s-dimensional observation $\boldsymbol{x}_i$, where $i = 1, \ldots, n$, and the goal is to partition the samples into several clusters, where samples belonging to same cluster are similar, while samples belonging to different clusters are dissimilar.

k-means [2] is a nonhierarchical prototype-based clustering method where prototypes (centroids) are used to represent clusters. The objective function is defined as the sum of within-cluster errors

$$L_{km} = \sum_{k=1}^{K} \sum_{i \in G_k} \|\boldsymbol{x}_i - \boldsymbol{b}_k\|^2 \tag{1}$$

where K is the predefined number of clusters, and $\boldsymbol{b}_k$ is the representative prototype (centroid) of cluster G_k. The k-means process is composed of two phases, i.e., prototype estimation and sample assignment, and the two phases are iterated until the solution is trapped in a local minimum. Although the process is very simple and useful in many cases, we suffer from the initialization problem, i.e., the greedy nature of the updating algorithm sometimes converges to different local optima with different initialization.

Recently, Ding and He [1] pointed out a close relation between PCA and k-means clustering and proposed an analytical (deterministic) means for k-means clustering in a PCA-guided manner. The k-means objective function of (1) can be redefined by a centroidless formulation as follows [23]:

$$L_{km} = \sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2 - \sum_{k=1}^{K} \frac{1}{n_k} \sum_{i,j \in G_k} \boldsymbol{x}_i^\top \boldsymbol{x}_j \tag{2}$$

where n_k is the number of samples belonging to cluster G_k, and $\top$ represents the transpose of a vector (or matrix). Here, the first term is a constant, while the second term is the sum of within-cluster (inner product) similarities. The solution of (hard) k-means clustering is represented by K nonnegative indicator vectors $H_K = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K)$ as

$$h_{ki} = \begin{cases} \dfrac{1}{n_k^{1/2}}, & \text{if sample } i \text{ belongs to cluster } G_k \\ 0, & \text{otherwise} \end{cases}$$

where $H_K^\top H_K = I_K$, and I_K is the $K \times K$ unit matrix. Because $\sum_{k=1}^{K} n_k^{1/2} h_{ki} = 1$, the indicator vectors have redundancies. In order to remove the redundancies and derive a unique solution, Ding and He introduced a $K \times K$ orthogonal transformation $T = \{t_{ij}\}$ as

$$Q_K = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K) = H_K T \tag{3}$$

and set the last column of T as

$$\boldsymbol{t}_K = \left(\sqrt{n_1/n}, \ldots, \sqrt{n_K/n}\right)^\top. \tag{4}$$

From the mutual orthogonality of $\boldsymbol{h}_k$, where $k = 1, \ldots, K$ and $\boldsymbol{q}_K = (\sqrt{1/n}, \ldots, \sqrt{1/n})^\top$, we have the following relations:

$$Q_{K-1}^\top Q_{K-1} = I_{K-1} \tag{5}$$

$$\sum_{i=1}^{n} q_{ki} = 0, \qquad k = 1, \ldots, K-1 \tag{6}$$

where $Q_{K-1} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{K-1})$, and $\boldsymbol{q}_k = (q_{k1}, \ldots, q_{kn})^\top$. Then, the k-means objective function can be written as

$$L_{km} = \sum_{i=1}^{n} \|\boldsymbol{x}_i\|^2 - \frac{1}{n} \boldsymbol{e}^\top X^\top X \boldsymbol{e} - \mathrm{Tr}(Q_{K-1}^\top X^\top X Q_{K-1}) \tag{7}$$

where $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, and $\boldsymbol{e}$ is the n-dimensional vector, whose all elements are 1.

Because the k-means problem does not distinguish the original data $\boldsymbol{x}_i$ and the centered data $\boldsymbol{y}_i$, the aforementioned objective function can be replaced with

$$L_{km} = \sum_{i=1}^{n} \|\boldsymbol{y}_i\|^2 - \mathrm{Tr}(Q_{K-1}^\top Y^\top Y Q_{K-1}) \tag{8}$$

where $Y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, and $Y\boldsymbol{e} = \boldsymbol{0}$. The optimal solutions for Q_{K-1} are derived by maximizing $\mathrm{Tr}(Q_{K-1}^\top Y^\top Y Q_{K-1})$, and continuous (relaxed) solutions are the eigenvectors corresponding to the $K-1$ largest eigenvalues of $Y^\top Y$, i.e., $K-1$-dimensional principal-component scores in PCA. This way, a continuous solution for k-means clustering is derived from a PCA-guided manner.

However, if we want to know the cluster structure from the (continuous) solution of Q_{K-1}, we need to compute the optimal transformation T, although it is not easy in many case. In [1], cluster connectivity analysis is performed by calculating the following connectivity matrix $C = \{c_{ij}\}$:

$$C = H_K H_K^\top \cong Q_K Q_K^\top = \frac{1}{n} \boldsymbol{e}\boldsymbol{e}^\top + \sum_{k=1}^{K-1} \boldsymbol{q}_k \boldsymbol{q}_k^\top. \tag{9}$$

If $c_{ij} > 0$, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are in a same cluster. Then, a probability for the connectivity between samples i and j is given as

$$p_{ij} = \frac{c_{ij}}{\sqrt{c_{ii} c_{jj}}} \tag{10}$$

and c_{ij} may be set as 0 when $p_{ij} < 0.5$ in order to decrease noise influences.

B. Visual Assessment of Cluster Structure in Connectivity Matrix

A potential approach to visual assessment of cluster structure is distance-sensitive ordering of samples (objects) [24]. Assume that $P = \{p_{ij}\}$ is the similarity (connectivity) matrix among n objects and that the object arrangement is performed by the index permutation $\pi(1, 2, \ldots, n) = (\pi_1, \pi_2, \ldots, \pi_n)$. The goal of sample ordering is to estimate π_i so that adjacent objects are

similar, while the larger the distance between the objects, the less similar the two objects are.

The objective function for distance-sensitive ordering is defined as

$$J(\pi) = \frac{1}{2}\sum_{i,j}(\pi_i^{-1} - \pi_j^{-1})^2 p_{ij} \tag{11}$$

where π_i^{-1} is the inverse permutation. It was shown that a relaxed solution for the inverse permutation $\boldsymbol{r}$ is the solution of the following problem:

$$\begin{aligned} \min_{\boldsymbol{r}} \quad & \tilde{J} = \frac{\boldsymbol{r}^\top (G-P)\boldsymbol{r}}{\boldsymbol{r}^\top G \boldsymbol{r}} \\ \text{s.t.} \quad & \boldsymbol{r}^\top G\mathbf{1} = 0 \\ & \boldsymbol{r}^\top G \boldsymbol{r} = 1 \end{aligned} \tag{12}$$

where G is the diagonal matrix whose diagonal element is the sum of the corresponding row ($g_{ii} = \sum_j p_{ij}$). Then, the optimal $\boldsymbol{r}$ is the solution (the eigenvector with the smallest eigenvalue except for $\boldsymbol{r}_0 = \boldsymbol{e}$) of the eigenvalue problem

$$(G-P)\boldsymbol{r} = \zeta G \boldsymbol{r} \tag{13}$$

and, transforming as $\boldsymbol{r} = G^{-1/2}\boldsymbol{z}$, we have

$$G^{-1/2} P G^{-1/2} \boldsymbol{z} = \delta \boldsymbol{z}, \qquad \delta = 1 - \zeta. \tag{14}$$

The optimal $\boldsymbol{r}$ is derived by searching the largest eigenvalue of $G^{-1/2}PG^{-1/2}$, except for $\boldsymbol{z}_0 = \boldsymbol{e}$, and the inverse permutation is given as

$$r_i < r_j \rightarrow \pi_i^{-1} < \pi_j^{-1}. \tag{15}$$

After distance-sensitive ordering, cluster structure is shown in the diagonal block structure of the connectivity matrix. To find cluster boundaries, Ding and He [24] defined "cluster crossing" as

$$\tilde{\rho}(i) = \frac{1}{4}\rho\left(i + \frac{1}{2}\right) + \frac{1}{2}\rho(i) + \frac{1}{4}\rho\left(i - \frac{1}{2}\right) \tag{16}$$

where

$$\begin{aligned} \rho(i) &= \sum_{j=1}^{m} p_{\pi_{i-j}^{-1}, \pi_{i+j}^{-1}} \\ \rho\left(i \pm \frac{1}{2}\right) &= \sum_{j=1}^{m} p_{\pi_{i-j}^{-1}, \pi_{i+j\pm 1}^{-1}}. \end{aligned} \tag{17}$$

Cluster crossing is the sum along the antidiagonal direction in the connectivity matrix with a bandwidth m and takes a minimum at the cluster boundaries between clusters, i.e., each cluster forms a "mountain" in the cluster-crossing curve. By finding "mountains" and "valleys" in the curve, we can visually capture the cluster structures in the connectivity matrix. In the numerical experiments given in Section IV, m was set as 10 so that the minimum cluster volume is assumed to be 10.

C. Robust Clustering

Noise fuzzy clustering [11] is a robustified version of well-known fuzzy c-means (FCM) clustering [25] and is identified with robust possibilistic clustering [18] or robust M-estimation [13] in the case of a single cluster. In this paper, a single-cluster case is considered in order to remove noise samples from the modeling process, i.e., an alternative selection, whether a sample has some responsibility in the modeling process or not, is considered. In this paper, the degree of responsibility is called "responsibility weight" in order to distinguish it from the conventional cluster memberships (cluster indicators). Using a certain criterion d_i and a responsibility weight of sample i for prototype estimation $u_i \in [0,1]$, the objective function in noise fuzzy clustering is written as

$$L_{nfc} = \sum_{i=1}^{n}(1-u_i)^\theta \gamma + \sum_{i=1}^{n} u_i^\theta d_i \tag{18}$$

where θ is the weighting exponent used to estimate "fuzzy" memberships. The larger the θ, the fuzzier the memberships. A recommended value of θ is $\theta = 2$, and $\theta = 1.5$ can also be used for clear partitioning in real applications. γ is an additional penalty weight and tunes the noise sensitivity of solutions. The solution that satisfies the necessary condition for the optimality is given as follows:

$$u_i = \left[1 + \left(\frac{d_i}{\gamma}\right)^{1/\theta - 1}\right]^{-1}. \tag{19}$$

Since $u_i = 0.5$ for $d_i = \gamma$ and u_i becomes small as d_i increases, it is obvious that we have many noise samples with small γ. On the other hand, u_i becomes close to 1 with large γ. Therefore, γ is used to select the responsibility boundary. In possibilistic c-means [18], γ is often set as

$$\gamma = \beta \frac{\sum_{i=1}^{n} u_i^\theta d_i}{\sum_{i=1}^{n} u_i^\theta} \tag{20}$$

and we can tune the noise sensitivity by changing β, while a recommended value of β is $\beta = 1$.

III. Fuzzy Principle Component Analysis-Guided Robust k-Means Procedure

In this section, a new algorithm for robust k-means clustering is proposed by modifying the PCA-guided k-means algorithm. The k-means algorithm is sensitive to noise because of the probabilistic constraint for memberships that forces all samples (including even noise samples) to belong to a cluster. In this paper, a responsibility weight of each sample in k-means process is estimated based on the noise-fuzzy-clustering mechanism [11], [12], and cluster-membership indicators in k-means process are derived as fuzzy principal components by considering the responsibility weights in fuzzy PCA [10].

A. Robust k-Means by Fuzzy PCA-Guided Manner

Fuzzy PCA [10] is a fuzzified version of PCA, in which principal components are extracted by considering membership degree of samples. When we have fuzzy memberships of samples u_i, where $i = 1, \ldots, n$, and a (fuzzily) normalized data matrix Y, the principal-component vectors are the principal eigenvectors of fuzzy scatter matrix $YUY^\top$, where U is a diagonal matrix whose ith diagonal element is u_i.

Introducing the noise-clustering mechanism, the objective function for robust k-means clustering is defined as

$$L_{rkm} = \sum_{i=1}^{n}(1-u_i)^{\theta}\gamma + \sum_{k=1}^{K}\sum_{i\in G_k} u_i^{\theta}\|\boldsymbol{x}_i - \boldsymbol{b}_k\|^2 \tag{21}$$

where u_i is the responsibility degree of $\boldsymbol{x}_i$ for k-means clustering. If u_i is small, i.e., there is no cluster center in the neighborhood of $\boldsymbol{x}_i$, then $\boldsymbol{x}_i$ is classified as "noise" and is removed from the k-means process. The cluster centroid $\boldsymbol{b}_k$ satisfying the necessary condition for the optimality is calculated as

$$\boldsymbol{b}_k = \frac{\sum_{i\in G_k} u_i^{\theta}\boldsymbol{x}_i}{\sum_{i\in G_k} u_i^{\theta}}. \tag{22}$$

Considering the fuzzy-membership weights u_i, we can also obtain a centroidless formulation as follows:

$$\begin{aligned} L_{rkm} =& \sum_{i=1}^{n}(1-u_i)^{\theta}\gamma + \sum_{i=1}^{n} u_i^{\theta}\|\boldsymbol{x}_i\|^2 \\ &- \sum_{k=1}^{K}\frac{\sum_{i,j\in G_k} u_i^{\theta}\boldsymbol{x}_i^{\top}\boldsymbol{x}_j u_j^{\theta}}{\sum_{j\in G_k} u_j^{\theta}}. \end{aligned} \tag{23}$$

With fixed u_i, a robust k-means cluster assignment is derived in a similar manner with PCA-guided k-means. Assume that the solution is represented by K indicator vectors $H_K = (\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K)$

$$h_{ki} = \begin{cases} \dfrac{\left(u_j^{\theta}\right)^{1/2}}{\left(\sum_{j\in G_k} u_j^{\theta}\right)^{1/2}}, & \text{if } i \text{ belongs to } G_k \\ 0, & \text{otherwise} \end{cases}$$

and $H_K^{\top}H_K = I_K$. Here, the membership indicator h_{ki} represents the degree of responsibility for cluster G_k by considering noise degree. If $\boldsymbol{x}_i$ is a noise sample, then h_{ki} are small in all clusters. If $\boldsymbol{x}_i$ is not a noise sample, then h_{ki} have large value in the cluster to which $\boldsymbol{x}_i$ belongs. Using the transformed discrete-membership-indicator vectors $Q_K = H_K T$, (23) is written as

$$\begin{aligned} L_{rkm} =& \sum_{i=1}^{n}(1-u_i)^{\theta}\gamma + \sum_{i=1}^{n} u_i^{\theta}\|\boldsymbol{x}_i\|^2 \\ &- \mathrm{Tr}(H_K^{\top}U^{\theta/2}X^{\top}XU^{\theta/2}H_K) \\ =& \sum_{i=1}^{n}(1-u_i)^{\theta}\gamma + \sum_{i=1}^{n} u_i^{\theta}\|\boldsymbol{x}_i\|^2 \\ &- \mathrm{Tr}(Q_K^{\top}U^{\theta/2}X^{\top}XU^{\theta/2}Q_K) \end{aligned} \tag{24}$$

where U is a diagonal matrix whose ith diagonal element is u_i.

Here, assume that Y is a normalized data matrix so that $Y\boldsymbol{u} = \boldsymbol{0}$, where $\boldsymbol{u} = (u_1^{\theta}, \ldots, u_n^{\theta})^{\top}$. This condition is achieved by the centering process from the view point of the least-square method, where the mean vector is given as (22). Under the constraint of

$$\boldsymbol{t}_K = \left(\sqrt{\frac{\sum_{i\in G_1} u_i^{\theta}}{\sum_{i=1}^{n} u_i^{\theta}}}, \ldots, \sqrt{\frac{\sum_{i\in G_K} u_i^{\theta}}{\sum_{i=1}^{n} u_i^{\theta}}}\right)^{\top} \tag{25}$$

$\boldsymbol{q}_K$ is given as

$$\boldsymbol{q}_K = \left(\frac{u_1^{\theta/2}}{\left(\sum_{i=1}^{n} u_i^{\theta}\right)^{1/2}}, \ldots, \frac{u_n^{\theta/2}}{\left(\sum_{i=1}^{n} u_i^{\theta}\right)^{1/2}}\right)^{\top} \tag{26}$$

and we have $YU^{\theta/2}\boldsymbol{q}_K = \boldsymbol{0}$. Then, by using the normalized data matrix Y, (24) is written as

$$\begin{aligned} L_{rkm} =& \sum_{i=1}^{n}(1-u_i)^{\theta}\gamma + \sum_{i=1}^{n} u_i^{\theta}\|\boldsymbol{y}_i\|^2 \\ &- \mathrm{Tr}(Q_{K-1}^{\top}U^{\theta/2}Y^{\top}YU^{\theta/2}Q_{K-1}). \end{aligned} \tag{27}$$

Because the first and second terms of (27) are constant, the transformed discrete-membership-indicator vectors Q_{K-1} are derived by maximizing $\mathrm{Tr}(Q_{K-1}^{\top}U^{\theta/2}Y^{\top}YU^{\theta/2}Q_{K-1})$, and continuous (relaxed) solutions are the eigenvectors corresponding to the $K-1$ largest eigenvalues of $U^{\theta/2}Y^{\top}YU^{\theta/2}$. Here, $\boldsymbol{q}_k$ is identified with the fuzzy principal-component-score vector given in fuzzy PCA using a generalized membership weight u_i^{θ} instead of u_i.

B. Responsibility Weight for k-Means Process

Next, with fixed k-means cluster assignment, responsibility of each sample for the k-means process is estimated. In the noise-clustering formulation, the objective function is given as

$$L_{rkm} = \sum_{i=1}^{n}(1-u_i)^{\theta}\gamma + \sum_{i=1}^{n} u_i^{\theta} d_i \tag{28}$$

where d_i is the responsibility criterion for noise clustering. In order to calculate the criterion, the objective function of (23) is transformed as follows:

$$\begin{aligned} L_{rkm} =& \sum_{i=1}^{n}(1-u_i)^{\theta}\gamma + \sum_{i=1}^{n} u_i^{\theta}\|\boldsymbol{x}_i\|^2 \\ &- \sum_{k=1}^{K}\frac{\sum_{i,j\in G_k} u_i^{\theta}\boldsymbol{x}_i^{\top}\boldsymbol{x}_j u_j^{\theta}}{\sum_{j\in G_k} u_j^{\theta}} \\ =& \sum_{i=1}^{n}(1-u_i)^{\theta}\gamma + \sum_{i=1}^{n} u_i^{\theta}\|\boldsymbol{x}_i\|^2 \\ &- \sum_{k=1}^{K}\frac{\sum_{i,j\in G_k} (u_i^{\theta})^{1/2}(u_j^{\theta})^{1/2}(u_i^{\theta})^{1/2}(u_j^{\theta})^{1/2}\boldsymbol{x}_i^{\top}\boldsymbol{x}_j}{\left(\sum_{j\in G_k} u_j^{\theta}\right)^{1/2}\left(\sum_{j\in G_k} u_j^{\theta}\right)^{1/2}} \\ \cong& \sum_{i=1}^{n}(1-u_i)^{\theta}\gamma + \sum_{i=1}^{n} u_i^{\theta}\|\boldsymbol{x}_i\|^2 \\ &- \sum_{i=1}^{n}(u_i^{\theta})^{1/2}\sum_{j=1}^{n}\sum_{k=1}^{K} h_{ki}h_{kj}\boldsymbol{x}_i^{\top}\boldsymbol{x}_j(u_j^{\theta})^{1/2}. \end{aligned} \tag{29}$$

From $H_K H_K^\top = Q_K T^\top T Q_K^\top = Q_K Q_K^\top$, we have $\sum_{k=1}^K h_{ki}h_{kj} = \sum_{k=1}^K q_{ki}q_{kj}$. Then, L_{rkm} is reformulated as

$$\begin{aligned} L_{rkm} &= \sum_{i=1}^n (1-u_i)^\theta \gamma + \sum_{i=1}^n u_i^\theta \|\boldsymbol{x}_i\|^2 \\ &\quad - \sum_{i=1}^n (u_i^\theta)^{1/2} \sum_{j=1}^n \sum_{k=1}^K q_{ki}q_{kj} \boldsymbol{x}_i^\top \boldsymbol{x}_j (u_j^\theta)^{1/2} \\ &= \sum_{i=1}^n (1-u_i)^\theta \gamma + \sum_{i=1}^n u_i^\theta \|\boldsymbol{x}_i\|^2 \\ &\quad - \sum_{i=1}^n u_i^\theta \sum_{j=1}^n \sum_{k=1}^K q_{ki}q_{kj} \boldsymbol{x}_i^\top \boldsymbol{x}_j \left(\frac{u_j^\theta}{u_i^\theta}\right)^{1/2} \end{aligned} \tag{30}$$

and the criterion with fixed-weight ratio u_j/u_i is reduced to the following formula:

$$d_i = \|\boldsymbol{x}_i\|^2 - \sum_{k=1}^K \sum_{j=1}^n q_{ki}q_{kj} \boldsymbol{x}_i^\top \boldsymbol{x}_j \left(\frac{u_j}{u_i}\right)^{\theta/2}. \tag{31}$$

Then, u_i is estimated as

$$u_i = \left[1 + \left(\frac{d_i}{\gamma}\right)^{1/(\theta-1)}\right]^{-1} \tag{32}$$

so that the necessary condition for the optimality of (28) is satisfied.

C. Algorithm for Fuzzy PCA-Guided Robust k-Means

The aforementioned two phases of cluster-indicator estimation and responsibility estimation are repeated until convergence, and a connectivity matrix C is calculated as $C = Q_K Q_K^\top$. Then, a probability for the connectivity between samples i and j is given by considering the responsibilities of samples as

$$p_{ij} = u_i^\theta u_j^\theta \frac{c_{ij}}{\sqrt{c_{ii}c_{jj}}}. \tag{33}$$

Here, p_{ij} is large only when samples i and j are in a same cluster, and none of them is a noise sample, i.e., noise samples have small connectivity with all other samples.

Then, the proposed algorithm is written as follows:

Algorithm: Fuzzy PCA-guided Robust k-Means (FPR k-Means)

Step 1. Initialize responsibility weights u_i, $i = 1, \ldots, n$ as $u_i = 1$, and choose the noise sensitivity weight β and the termination condition ε.

Step 2. Calculate the normalized data matrix Y so that $Y\boldsymbol{u} = \boldsymbol{0}$ where $\boldsymbol{u} = (u_1^\theta, \ldots, u_n^\theta)^\top$.

Step 3. Calculate the transformed indicator vectors $Q_{K-1} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_{K-1})$ from the $K-1$ principal eigenvectors of $U^{\theta/2} Y^\top Y U^{\theta/2}$, and set $\boldsymbol{q}_K$ as (26).

Step 4. Calculate responsibility criteria d_i, $i = 1, \ldots, n$ and γ using (31) and (20), respectively. Update u_i using (32). (γ should be updated only in a first few iteration in the same manner with possibilistic clustering [18].)

Step 5. If $\max_i \mid u_i^{\mathrm{NEW}} - u_i^{\mathrm{OLD}} \mid < \varepsilon$, then output connectivity matrix C or $P = \{p_{ij}\}$. Otherwise, return to Step 2.

The proposed method is equivalent to the conventional PCA-guided k-means if all u_i are 1 (or γ is extremely large), i.e., the initial partition is given by a deterministic procedure based on the conventional PCA-guided k-means. Then, the following robustification process is also performed in a deterministic manner.

To assess the cluster validity, we should take into account the responsibilities of samples in the visual-assessment approach [24]. Because the derived connectivity matrix P reveals the probability of the mutual connectivity among samples by considering the responsibility weights for k-means process, noise samples that have small responsibility weights may be inserted in irrelevant positions without significant loss of cluster-crossing values. In this paper, samples having smaller responsibility weights than a predefined threshold are removed before distance-sensitive ordering. Then, the remaining samples are arranged by considering the following objective function:

$$J(\pi) = \frac{1}{2} \sum_{i,j} u_i^\theta u_j^\theta \left(\pi_i^{-1} - \pi_j^{-1}\right)^2 p_{ij} \tag{34}$$

and a relaxed solution for the inverse permutation $\boldsymbol{r}$ is the solution of the following problem:

$$\begin{aligned} \min_{\boldsymbol{r}} \quad & \tilde{J} = \frac{\boldsymbol{r}^\top U^{\theta/2}(G-P)U^{\theta/2}\boldsymbol{r}}{\boldsymbol{r}^\top U^{\theta/2} G U^{\theta/2} \boldsymbol{r}} \\ \text{s.t.} \quad & \boldsymbol{r}^\top G U^{\theta/2} \boldsymbol{1} = 0 \\ & \boldsymbol{r}^\top U^{\theta/2} G U^{\theta/2} \boldsymbol{r} = 1 \end{aligned} \tag{35}$$

where the normalization constraints are also modified by considering responsibility weights. Then, the optimal $\boldsymbol{r}$ is the solution (the eigenvector with the smallest eigenvalue except for $\boldsymbol{r}_0 = \boldsymbol{e}$) of the eigenvalue problem

$$U^{\theta/2}(G-P)U^{\theta/2}\boldsymbol{r} = \zeta U^{\theta/2} G U^{\theta/2} \boldsymbol{r}. \tag{36}$$

Transforming as $\boldsymbol{r} = G^{-1/2} U^{-\theta/2} \boldsymbol{z}$, we have

$$G^{-1/2} P G^{-1/2} \boldsymbol{z} = \delta \boldsymbol{z}, \qquad \delta = 1 - \zeta. \tag{37}$$

The optimal $\boldsymbol{r}$ is derived by searching the largest eigenvalue of $G^{-1/2} P G^{-1/2}$, except for $\boldsymbol{z}_0 = \boldsymbol{e}$, and the inverse permutation is given as

$$r_i < r_j \rightarrow \pi_i^{-1} < \pi_j^{-1}. \tag{38}$$

Here, it should be noted that the main difference of the conventional distance-sensitive ordering method is to consider the responsibility in recovering the inverse permutation as $\boldsymbol{r} = G^{-1/2} U^{-\theta/2} \boldsymbol{z}$, i.e., samples having small responsibility weights are pulled away from center position. This is because they have small connectivity values with all samples so that they might be gathered into center position.

D. Application of Kernel Trick

The conventional PCA technique derives at most the same number of significant principal components with the data

dimension and is not applicable to capture nonlinear cluster boundaries. In [1], the solution for kernel k-means was also given by kernel PCA, in which data points are mapped into a higher dimensional space via kernels. In this section, the kernel method is applied to FPR k-means in order to extract a larger number of clusters than the dimensionality of a dataset having nonlinear boundaries.

The following nonlinear transformation (mapping) to the higher dimensional space is considered:

$$\boldsymbol{x}_i \rightarrow \phi(\boldsymbol{x}_i). \tag{39}$$

After mapping, the objective function for FPR k-means clustering is given as

$$\begin{aligned} L_{krkm} =& \sum_{i=1}^{n}(1-u_i)^{\theta}\gamma + \sum_{i=1}^{n} u_i^{\theta}\|\phi(\boldsymbol{x}_i)\|^2 \\ & - \sum_{k=1}^{K} \frac{\sum_{i,j\in G_k} u_i^{\theta}\phi(\boldsymbol{x}_i)^{\top}\phi(\boldsymbol{x}_j)u_j^{\theta}}{\sum_{j\in G_k} u_j^{\theta}} \\ =& \sum_{i=1}^{n}(1-u_i)^{\theta}\gamma + \sum_{i=1}^{n} u_i^{\theta} w_{ii} \\ & - \mathrm{Tr}(Q_K^{\top} U^{\theta/2} W U^{\theta/2} Q_K) \end{aligned} \tag{40}$$

where $W = \{w_{ij}\}$ is the kernel matrix whose element is $w_{ij} = \phi(\boldsymbol{x}_i)^{\top}\phi(\boldsymbol{x}_j)$. With fixed u_i, the first and second term of (40) are constant, and the clustering problem is reduced to the maximization of the (negative) third term.

In the kernel method (or also called kernel trick) [26], we do not have an exact form of function $\phi(\boldsymbol{x}_i)$ but assume that the scalar product of kernel function $K(\boldsymbol{x}, \boldsymbol{y})$ is given as

$$K(\boldsymbol{x}, \boldsymbol{y}) = < \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) > \tag{41}$$

where $< \cdot, \cdot >$ is the inner product. Without having the exact form of function $\phi(\boldsymbol{x}_i)$ (or constructing the exact high-dimensional feature space), we can apply several types of analysis such as k-means (or FCM) [27] and PCA [28]. The polynomial kernel

$$K(\boldsymbol{x}, \boldsymbol{y}) = (< \boldsymbol{x}, \boldsymbol{y} > + c)^d \tag{42}$$

and the Gaussian kernel

$$K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\lambda\|\boldsymbol{x} - \boldsymbol{y}\|^2) \tag{43}$$

are widely used. Here, a kernel matrix may not be centered, while the PCA-guided k-means is formulated for centered data. Therefore, in [1], centering of the kernel was performed as

$$W \rightarrow SWS \tag{44}$$

$$S = I_n - \boldsymbol{e}e^{\top}. \tag{45}$$

After centering of the kernel, all indicator vectors $\boldsymbol{q}_k$ satisfy $\boldsymbol{q}_k^{\top}\boldsymbol{e} = 0$. Then, the solution to kernel k-means is given by kernel PCA components.

In the same way, with kernel PCA-guided k-means, we can derive the optimal solution to minimize (40) by maximizing $\mathrm{Tr}(Q_K^{\top} U^{\theta/2} W U^{\theta/2} Q_K)$. In order to normalize the feature vectors in the high-dimensional feature space by considering responsibility weights for k-means process, the kernel matrix is centered as follows:

$$W \rightarrow SWS \tag{46}$$

$$S = I_n - \frac{1}{\boldsymbol{u}^{\top}\boldsymbol{e}}\boldsymbol{u}e^{\top} \tag{47}$$

where $\boldsymbol{u} = (u_1^{\theta}, \ldots, u_n^{\theta})^{\top}$. After the normalization, we can derive the optimal Q_{K-1} by calculating the eigenvectors corresponding to the $K-1$ largest eigenvalues of $U^{\theta/2} W U^{\theta/2}$ because feature vectors $\phi(\boldsymbol{x}_i)$ in the high-dimensional feature space are centered as $\Phi S U^{\top}\boldsymbol{e} = \tilde{\Phi} U^{\top}\boldsymbol{e} = \boldsymbol{0}$, where $\Phi = (\phi(\boldsymbol{x}_1), \ldots, \phi(\boldsymbol{x}_n))$. Here, $\boldsymbol{q}_K$ is also given by (26).

Next, the responsibility weight of each sample is estimated using the kernel matrix. The responsibility criterion d_i is calculated as follows:

$$d_i = w_{ii} - \sum_{k=1}^{K}\sum_{j=1}^{n} q_{ki} q_{kj} w_{ij} \left(\frac{u_j}{u_i}\right)^{\theta/2}. \tag{48}$$

Then, u_i is updated using (32) with fixed weight u_j/u_i.

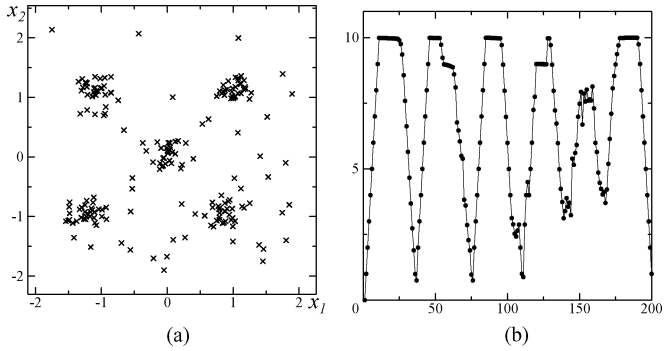

Fig. 1. Artificial dataset. (a) 2-D plots of dataset. (b) Cluster-crossing curve derived by PCA-guided k-means with $K = 5$.

IV. Numerical Experiments

This section shows several experimental results to demonstrate the characteristic features of the proposed methods.

A. Artificial Dataset

A numerical experiment was performed with an artificially generated dataset shown in Fig. 1(a), in which five cluster cores with 30 samples, each drawn from spherical normal distributions having equal variances, are buried in 50 noise samples from uniform distribution.

1) Cluster Validation: First, the conventional PCA-guided k-means was applied with $K = 5$, and the cluster-crossing curve shown in Fig. 1(b) was derived. The kernel trick was applied with a Gaussian kernel ($\lambda = 5.0$) in order to capture the nonlinear cluster boundary. Because of many noise samples, five cluster cores were concealed, and the figure indicates that there are six clusters. Then, the clustering algorithm was reapplied with $K = 6$, and the cluster-crossing curve shown in Fig. 2(a) was derived. The figure implies that there are six clusters, although we have only five cluster cores. Fig. 2(b) shows the derived six clusters,

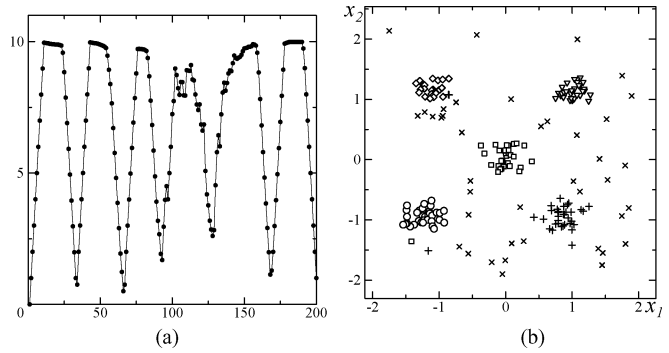

Fig. 2. Cluster-crossing curve and cluster partition derived by PCA-guided k-means with $K = 6$. (a) Cluster-crossing curve. (b) Cluster partition.

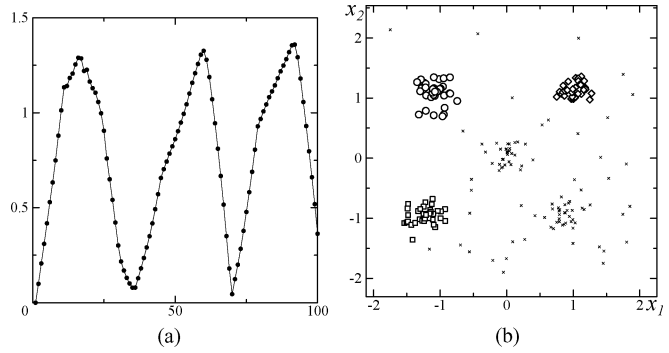

Fig. 3. Cluster-crossing curve and cluster partition derived by proposed method with $K = 3$. (a) Cluster-crossing curve. (b) Cluster partition.

in which the noise cluster ($\times$) was also extracted as a proper (nonnoise) cluster. This way, the conventional PCA-guided k-means is easily influenced by noise samples and is not useful to capture cluster cores.

Next, the proposed FPR k-means was applied with various cluster numbers. In the robust k-means, the weight for noise penalty β was set to 1.0, and the same kernel function was used. Before cluster arrangement, noise samples whose responsibility weights are lower than 0.4 were rejected, and the cluster-crossing curves were constructed by using the remaining samples. Figs. 3–6 show the derived cluster-crossing curves and cluster partitions where small $\times$ are noise (rejected) samples. When K is 3–5, we can find the corresponding clusters in the cluster-crossing curves. Here, it should be noted that several cluster cores were rejected, as shown in Figs. 3(b) and 4(b), when K was smaller than 5. This is because the proposed method rejects the samples out of K cores. On the other hand, when K is 6, we cannot find the six cluster structures in the cluster-crossing curve shown in Fig. 2, i.e., the proper cluster number is $K = 5$.

However, when K is 2, only one cluster core is extracted, as is shown in Fig. 7, while all other samples were assigned relatively smaller responsibilities ($0.4 < u_i < 0.5$), although no sample was rejected. Using the constraint of (26), the case of $K = 2$ uses only the most principal eigenvector (a single vector) that is responsible for alternative selection, and the process worked for noise selection. It may be because the initial partition given by

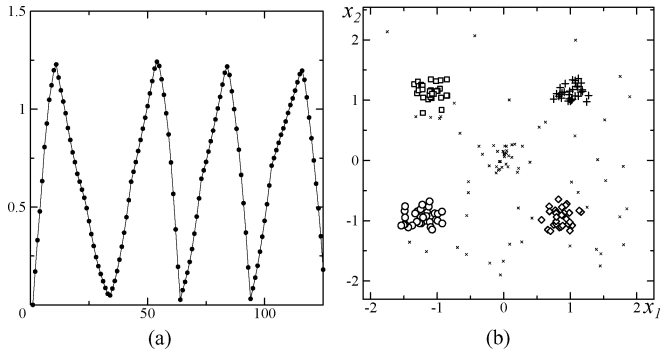

Fig. 4. Cluster-crossing curve and cluster partition derived by proposed method with $K = 4$. (a) Cluster-crossing curve. (b) Cluster partition.

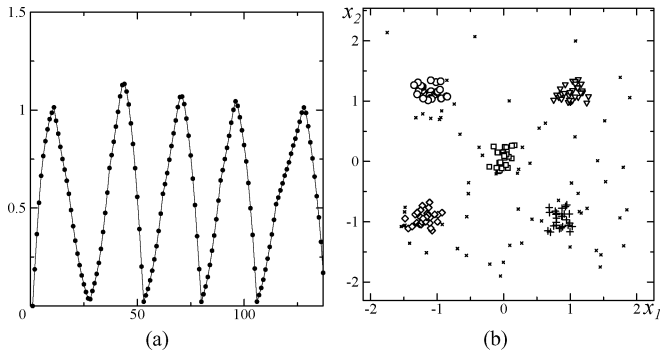

Fig. 5. Cluster-crossing curve and cluster partition derived by proposed method with $K = 5$. (a) Cluster-crossing curve. (b) Cluster partition.

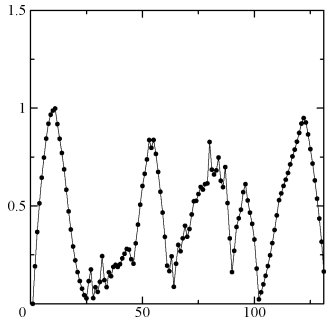

Fig. 6. Cluster-crossing curve derived by proposed method with $K = 6$.

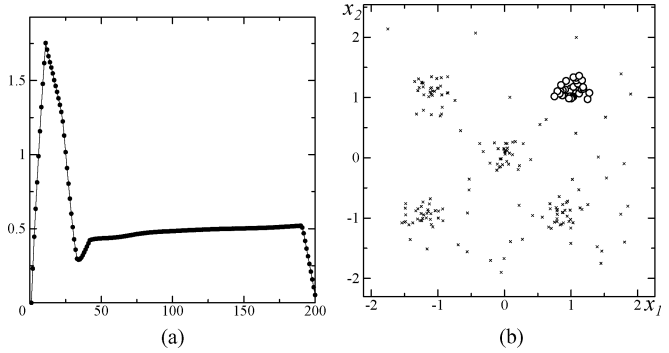

Fig. 7. Cluster-crossing curve and cluster partition derived by proposed method with $K = 2$. (a) Cluster-crossing curve. (b) Cluster partition.

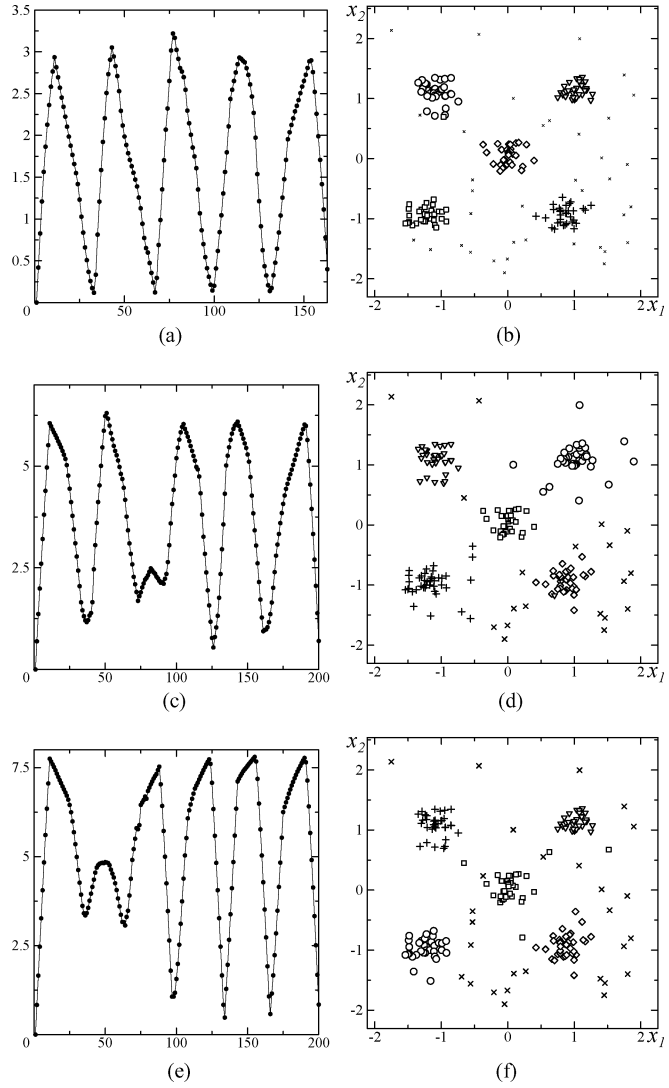

Fig. 8. Cluster-crossing curve and cluster partition derived by the proposed method with several noise weights. (a) Cluster crossing ($\beta = 2$). (b) Cluster partition ($\beta = 2$). (c) Cluster crossing ($\beta = 5$). (d) Cluster partition ($\beta = 5$). (e) Cluster crossing ($\beta = 10$). (f) Cluster partition ($\beta = 10$).

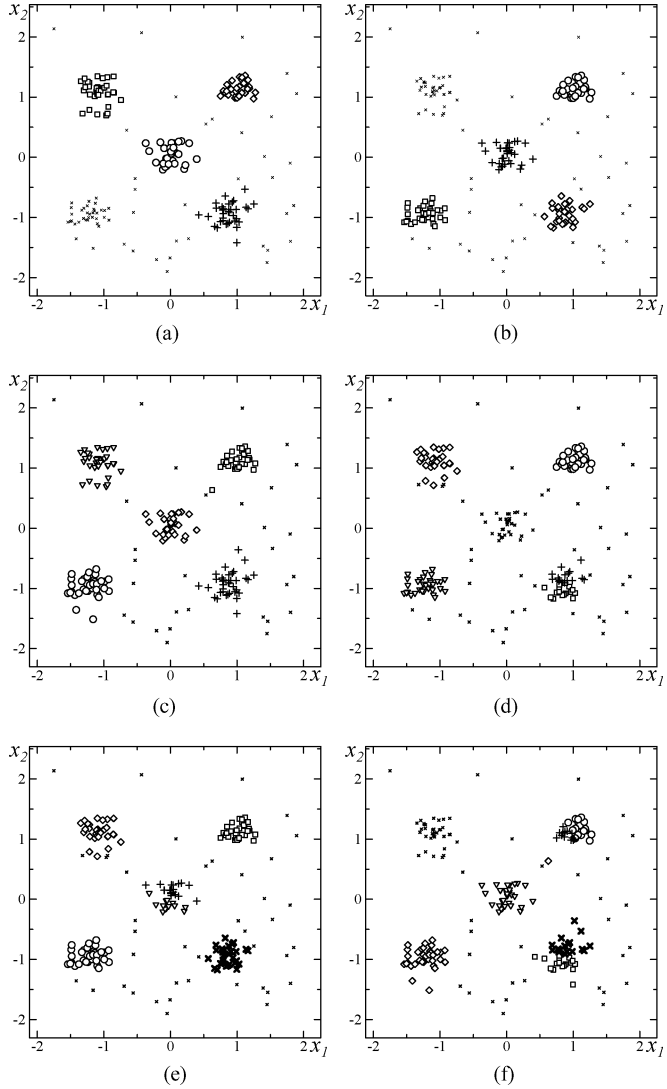

Fig. 9. Comparison of cluster partitions derived by noise fuzzy clustering with different initializations. (a) $K = 4$ (type I). (b) $K = 4$ (type II). (c) $K = 5$ (type I). (d) $K = 5$ (type II). (e) $K = 6$ (type I). (f) $K = 6$ (type II).

the conventional PCA-guided k-means regarded the all samples out of the cluster core as one cluster. This process implies that the proposed method can be applied to sequential cluster extraction [20], while the application is remained in future work.

2) Noise Sensitivity: Noise sensitivity of the proposed method is compared with several noise weights β. Fig. 8 shows the clustering results derived with $\beta = 2, 5,$ and 10 and indicates that we have fewer noise samples (larger cluster cores) with larger β, i.e., noise sensitivity of the proposed algorithm is tuned by the noise weight. However, the result of $\beta = 5.0$ implies that a larger β generates a noise cluster [$\times$ in Fig. 8(d)] with small cluster-crossing mountain [samples 73–91 in Fig. 8(c)]. Furthermore, when β is too large (e.g., $\beta = 10.0$), we have a similar result with that of the conventional PCA-guided k-means.

3) Comparison With Noise Fuzzy Clustering: The result given by the proposed method is compared with that of the conventional noise fuzzy clustering [11]. When we consider the multicluster (K clusters) case, the objective function of noise fuzzy clustering is defined as follows:

$$L_{nfc} = \sum_{i=1}^{n}(1 - u_{*i})^{\theta}\gamma + \sum_{k=1}^{K}\sum_{i=1}^{n} u_{ki}^{\theta}\|\boldsymbol{x}_i - \boldsymbol{b}_k\|^2 \qquad (49)$$

where $u_{*i} = \sum_{k=1}^{K} u_{ki}$, and "$1 - u_{*i}$" means the membership degree to the noise cluster. Fig. 9 compares the clustering results (maximum membership classification) with different initializations (types I and II). In the figure, small $\times$ means the noise samples. Fig. 9(a) and (b) indicates that the noise fuzzy clustering also has the ability of capturing fewer numbers of cluster cores in the same manner with the proposed method when the cluster number is smaller than the appropriate one. However, the missing cores are dependent on initialization, i.e., different cores are missing with different initializations. Fig. 9(c) and (d) compares the two partitions for the $K = 5$ case with different

TABLE I
COMPARISON OF CLUSTER-CORE PURENESS (PURE.) AND REJECTION RATIO (REJECT.) WITH WINE DATA

		pure.	reject.	freq.
FPR	$\beta = 1.0$	1.000	0.545	
	$\beta = 1.1$	0.992	0.343	
	$\beta = 1.2$	0.979	0.152	
	$\beta = 1.3$	0.965	0.000	
PCA		0.964	0.000	
FCM$_\alpha$ (ave.)	CL=0.7	1.000	0.507	
	CL=0.6	0.992	0.193	
	CL=0.5	0.993	0.010	
FCM$_\alpha$ (best)	CL=0.7	1.000	0.507	1.00
	CL=0.6	0.992	0.193	1.00
	CL=0.5	0.993	0.007	0.57

initializations and show that the optimal cluster cores can be extracted with the appropriate initialization, while inappropriate partition may be given with bad initialization in which one core was missing and another core was divided into two subcores. Fig. 9(e) and (f) shows the result with the larger cluster-number case. If the cluster number is larger than the optimal one, cluster cores would be divided. Additionally, it is also the case that some cluster cores are illegally divided into subcores with missing several cores, as is shown in Fig. 9(f).

This way, the proposed method can derive the best solution of the conventional noise fuzzy clustering in a deterministic procedure, although noise fuzzy clustering may generate illegal solutions with bad initializations. Additionally, the noise fuzzy clustering also suffers from the partition-validation problems that are used to search the optimal cluster number and initialization.

B. Real-World Datasets

In this section, comparative experiments on cluster-core estimation are performed with three real-world datasets. The pureness of the cluster cores estimated by the proposed method is compared with Yang's α-cut FCM [16]. α-cut FCM is a k-means-type cluster-core-estimation (robust clustering) technique based on the alternate optimization scheme. In this experiment, the performance is compared with α-cut FCM because Yang *et al.* reported that α-cut FCM outperformed the other methods in robust cluster-core estimation. In α-cut FCM, cluster-core samples are estimated by α-cut implementation where samples having larger memberships than a cutting level (CL) (α) are assigned a full membership (membership degree is 1). The CL is usually set as larger than or equal to 0.5 in order to avoid overlapping of cluster cores.

1) Wine Dataset: In the first experiment, wine dataset, which is available from the University of California (UC) Irvine Machine Learning Repository [29], was used. The dataset includes 178 samples with 13 attributes drawn from three classes (class 1: 59 samples, class 2: 71 samples, and class 3: 48 samples), and it is known that the three classes are almost linearly separable. In this experiment, the cluster number was fixed as $K = 3$, and the Gaussian kernel with $\lambda = 0.1$ was used in PCA-guided methods.

Table I shows the pureness ratio of cluster cores and noise-rejection ratio given by the proposed method with various noise weights (β), PCA-guided k-means, and α-cut FCM with various CLs. [FPR: FPR k-means, PCA: PCA-guided k-means, FCM$_\alpha$ (ave.): average performance in α-cut FCM, and FCM$_\alpha$ (best): best performance in α-cut FCM with its frequency (freq.)] The degree of fuzziness in noise-rejection mechanism and the noise threshold were set as 1.5 and 0.5, respectively, for clearly distinguishing noise samples. The pureness ratio is the average pureness of clusters after cluster labeling based on maximum numbers of sample classes in each cluster. Here, each class label was assigned only to one cluster. The noise-rejection ratio was the proportion of samples having responsibility memberships smaller than noise threshold, i.e., samples out of cores. In α-cut FCM, 100 trials were performed based on the multistart strategy, where initial cluster centers are randomly generated.

The proposed method gave 100 % pure cluster cores with noise weight $\beta = 1$, while 54.5 % samples were rejected as noise. The larger the noise weight, the smaller the pureness and rejection ratio. The cluster-core pureness without rejection (the case of $\beta = 1.3$) was 96.5% and is very similar to the conventional PCA-guided k-means.

Here, the performance of α-cut FCM seems to be almost equal to or slightly better than that of the proposed method. For example, the proposed method achieved 99.2% pureness with 34.3% rejection ($\beta = 1.1$), while α-cut FCM achieved the pureness with only 1% rejection (CL = 0.5). When all clusters are clearly isolated, like Wine dataset, alternate optimization approaches derive a single optimal solution in almost every trial and work well. Therefore, the performance of the proposed method is comparative with that of alternate optimization approaches when all clusters are clearly isolated.

2) Iris Dataset: In the second experiment, Iris dataset composed of 150 samples with four attributes drawn from three classes (Setosa: 50 samples, Versicolour: 50 samples, and Virginica: 50 samples) are used. Setosa is well isolated from other two classes, while the boundary between Versicolour and Virginica is very ambiguous. Therefore, the dataset is sometimes partitioned into two clusters by several cluster-validity measures, although the three classes form three masses. This experiment is performed with the goal being to capture the three cluster cores corresponding to the three classes, and the cluster number was fixed as $K = 3$. The Gaussian kernel with $\lambda = 1.0$ was used in PCA-guided methods.

Table II compares the results and implies that the proposed method estimated pure cluster cores with small noise weights ($\beta = 1.0$ or 1.1). Furthermore, the cluster cores still have 79.6% pureness without noise rejection ($\beta = 1.5$) and is larger than that of the conventional PCA-guided k-means because the proposed responsibility weights emphasize the purely core samples in cluster-center estimation and then derive robust cluster centers. The robustification mechanism is useful to improve cluster pureness in PCA-guided k-means, i.e., cluster pureness is improved even when no sample is rejected.

Next, the performance is compared with α-cut FCM. The table implies that the proposed method achieved 100% pureness with lower rejection ratio, while the intermediate model of α-cut FCM with around 20% rejection ratio slightly outperformed the proposed method with $\beta = 1.3$ from the pureness ratio viewpoint. Here, it should be noted that α-cut FCM with

TABLE II
COMPARISON OF CLUSTER-CORE PURENESS (PURE.) AND REJECTION RATIO (REJECT.) WITH IRIS DATA

		pure.	reject.	freq.
FPR	$\beta = 1.0$	1.000	0.633	
	$\beta = 1.1$	1.000	0.547	
	$\beta = 1.2$	0.933	0.387	
	$\beta = 1.3$	0.841	0.220	
	$\beta = 1.4$	0.794	0.053	
	$\beta = 1.5$	0.796	0.000	
PCA		0.671	0.000	
FCM_α (ave.)	CL=0.9	1.000	0.673	
	CL=0.8	0.968	0.453	
	CL=0.7	0.899	0.202	
	CL=0.6	0.840	0.092	
	CL=0.5	0.814	0.031	
FCM_α (best)	CL=0.9	1.000	0.673	1.00
	CL=0.8	0.971	0.440	0.02
	CL=0.7	0.900	0.200	0.97
	CL=0.6	0.872	0.093	0.21
	CL=0.5	0.847	0.033	0.50

0.5 CL failed to capture the three cluster cores in about 10% trials, i.e., bad initialization led to inappropriate solutions, in which Setosa was shared by two clusters, and the remaining cluster included both of Versicolour and Virginica, although the best performance is slightly better than the proposed method. This way, when cluster boundaries are ambiguous, alternate optimization approaches cannot derive appropriate solutions without good initialization, while the proposed method always derives a proper result in a deterministic manner.

3) Document Classification: In the third experiment, a Japanese novel "Kokoro" written by S. Natsume that can be downloaded from Aozora Bunko (http://www.aozora.gr.jp), which is a Web library containing copyright-expired Japanese books, is used. The English version of "Kokoro" is also available from Eldritch Press (http://www.ibiblio.org/eldritch/). The novel is composed of three chapters (Sensei and I, My Parents and I, and Sensei and His Testament), and the chapters include 36, 18, and 56 sections, respectively. In this experiment, the sections were considered as individual text documents (number of samples is $n = 110$), and the cluster number was set as $K = 3$. The text documents were preprocessed using a morphological analysis system software "Chasen" (http://chasen.naist.jp/hiki/ChaSen/), which segments Japanese text string into morphemes, and tags those morphemes with their parts of speech and pronunciations. Then, the 83 most frequently used substantives and verbs (they were used more than 50 times in the novel) were given as attributes to be analyzed with their tf-idf weights.

Fig. 10 shows the 2-D document map constructed by PCA, where 2-D coordinates indicate the 2-D principal-component scores of each document and implies that the three chapters do not have clear boundaries. The Gaussian kernel with $\lambda = 0.01$ was used in PCA-guided methods. As shown in Fig. 11, the PCA-guided k-means failed to construct a cluster-crossing curve that has three clear mountains. Then, the proposed method was applied with several noise-weights values. The performance is shown in Table III. By emphasizing only the core documents, the proposed method could capture the three cluster cores, while

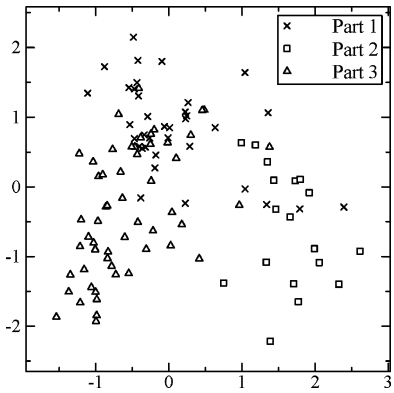

Fig. 10. Two-dimensional document map with "Kokoro."

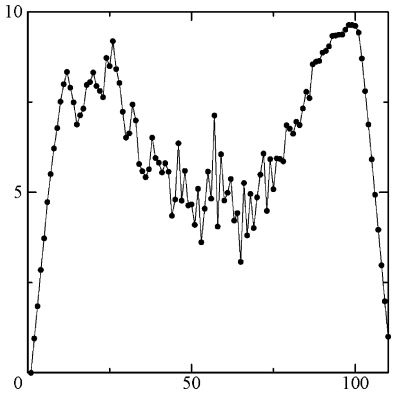

Fig. 11. Cluster crossing derived by PCA-guided k-means with "Kokoro."

TABLE III
COMPARISON OF CLUSTER-CORE PURENESS (PURE.) AND REJECTION RATIO (REJECT.) WITH DOCUMENTS OF "KOKORO"

		pure.	reject.
FPR	$\beta = 1.0$	0.949	0.691
	$\beta = 1.01$	0.941	0.564
	$\beta = 1.02$	0.924	0.418
	$\beta = 1.03$	0.841	0.345

the clustering model with β larger than 1.03 failed to reveal the three chapter structures.

Although α-cut FCM with 0.5 CL was also applied, no core samples were extracted, i.e., there were no samples that have memberships larger than 0.5, when fuzzifier $\theta = 1.5$. Then, the algorithm was reapplied with $\theta = 1.2$, which derives almost crisp (nonfuzzy) partition-rejecting noise samples. In 100 trials with various initializations, three cluster cores were captured in only 35 times, and the average pureness and rejection ratio in 35 trials were 85.6% and 41.6%, respectively. Although the best performance achieved 91.3% pureness and 38.0% rejection, the result was derived only in one trial during 100 trials. This way, when cluster boundaries are very unclear, alternate optimization approaches are very sensitive to initial cluster assignment and fail to capture cluster cores in almost every trial. On the other hand, the proposed method captures cluster cores in a deterministic manner by emphasizing only the core samples.

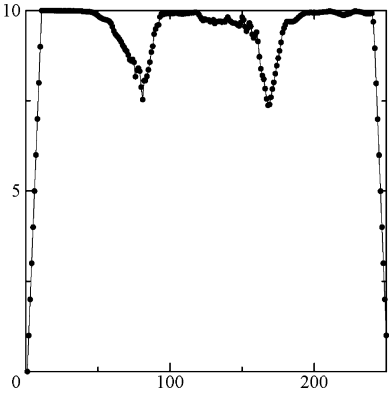

Fig. 12. Cluster crossing derived by PCA-guided k-means with COIL-20.

TABLE IV
COMPARISON OF CLUSTER-CORE PURENESS (PURE.) AND REJECTION RATIO (REJECT.) WITH COIL-20

		pure.	reject.
FPR	$\beta = 1.0$	0.921	0.604
	$\beta = 1.1$	0.956	0.476
	$\beta = 1.2$	0.906	0.268
	$\beta = 1.3$	0.812	0.000

C. High-Dimensional Dataset

In this section, the proposed method is applied to a high-dimensional dataset. The Columbia University Image Library (COIL-20) [30] is a database of 128×128 grayscale images of 20 objects, in which each object has 72 images taken by rotating through 360° with 5° steps. In this experiment, three objects (obj1, obj3, and obj4) were used with their 72 samples, each composed of 128×128 grayscale elements, i.e., the dataset includes three clusters with 72 samples each in the 128×128 dimensional data space. Additionally, the remaining 17 objects with their two samples each (the front and back images only) were added as noise, i.e., 17×2 noise images were added. Then, the dataset to be analyzed is composed of $3 \times 72 + 17 \times 2 = 250$ samples with $128 \times 128 = 16\,384$ attribute values. In the proposed method with Gaussian kernel, the parameters were set as $(\theta, K, \lambda) = (1.5, 3, 1.0 \times 10^{-8})$. Here, we must note that the computational cost is not influenced much by the dimensionality of data, i.e., the computational cost is mainly dependent on the sample size because the time-consuming part of the proposed method is the eigendecomposition of the inner product dissimilarity matrix whose size is (number of samples)×(number of samples).

Fig. 12 shows the cluster-crossing curve given by the conventional PCA-guided k-means and indicates that all noise samples were buried in the three clusters, i.e., we cannot recognize the noise, while the three data masses were captured. Then, the proposed method was applied in conjunction with noise-rejection mechanism. Table IV compares the cluster-core pureness ratio. The table indicates that the proposed method still work well, even with high-dimensional datasets in the same way as with the previous benchmark datasets. Therefore, in this experiment, the pureness ratio of $\beta = 1.0$ was inferior to $\beta = 1.1$. It may be because the sparseness of core might bring an ambiguous boundary, as shown in Fig. 13, where the boundary between the left and center clusters has a very small mountain (samples 26–28) when $\alpha = 1.0$, while the small mountain disappeared when $\alpha = 1.1$. This way, a clear mountain structure without small mountains tends to bring pure cluster cores.

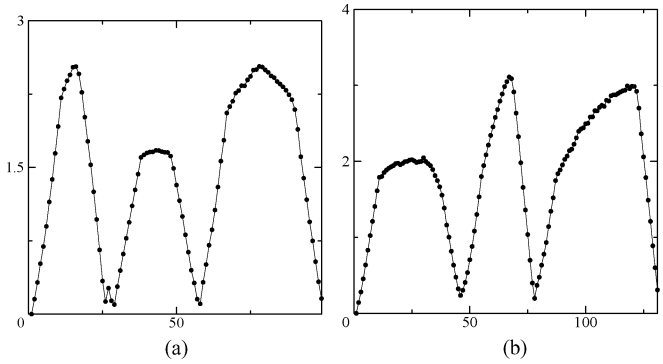

Fig. 13. Comparison of cluster-crossing curve derived by proposed method with COIL-20. (a) $\alpha = 1.0$. (b) $\alpha = 1.1$.

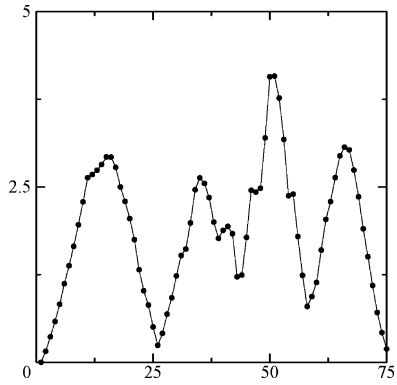

Fig. 14. Cluster crossing derived by proposed method without kernel trick with COIL-20.

Finally, the effectiveness of kernel trick is investigated. Fig. 14 shows the cluster-crossing curve given by the proposed method without the kernel trick, where $(\theta, K, \beta) = (1.5, 3, 1.0)$. Although the cluster number was set as $K = 3$, the cluster-crossing curve indicates as if there are four (or more) clusters. It may be because a cluster core having nonlinear boundaries was illegally divided into subcores. We can see that the unkernelized model may fail to find proper cluster boundaries, even when we have enough large-dimensional feature space, i.e., cluster boundaries are often nonlinear, even when they are characterized by high-dimensional observations, and the kernelized model seems to be plausible for most real applications.

V. CONCLUSION

This paper proposed a new approach to robust clustering in which robust k-means partition is derived using noise-rejection mechanism based on noise-clustering approach. In the proposed iterative algorithm, the responsibility weight of each sample for the k-means process is estimated by considering the noise degree of the sample. Then, cluster indicators are calculated in a fuzzy PCA-guided manner where fuzzy PCA-guided k-means is

performed by considering responsibility weights of samples. Therefore, the proposed method achieves cluster-core estimation in a deterministic way. The validity of the derived cluster cores is visually assessed through the distance-sensitive ordering that considers responsibility weights of samples.

Numerical experiments demonstrated that the proposed method is useful to capture cluster cores by rejecting noise samples, and we can easily assess cluster validity using cluster-crossing curves. Comparative experiments using several real-world datasets revealed that the proposed method can derive proper results without initialization problems, although alternate optimization approaches with bad initialization often fail to capture cluster structures. However, the best results in multiple trials of α-cut FCM sometimes outperformed that of the proposed method. It may be because PCA-guided k-means derives just a relaxed clustering solution that is not necessarily the optimal solution for k-means objective function when all clusters are not clearly isolated and do not form very small separate masses. On the other hand, when cluster boundaries are too ambiguous to capture the cluster cores by α-cut FCM, the relaxed cluster solution contributes to roughly capture the cluster cores, as shown in Section IV-B3. In several experiments, the proposed method without noise rejection outperformed the conventional PCA-guided k-means from the view point of pureness in cluster cores. Therefore, the responsibility weights can contribute to improve the relaxed solution by weakening the influences of noise samples that are distant from cluster centers in PCA-guided k-means.

A potential future work is an extension to sequential cluster extraction [20]–[22]. In numerical experiments, it was shown that the proposed method extract only a few cluster cores when the cluster number was set as smaller numbers. Especially when cluster number was two, only one proper cluster core was extracted. This may imply that we can extract proper cluster cores one by one by iteratively applying the proposed method with a small cluster number. In another direction, it is also possible to apply other visual-assessing methods, such as the visual assessment of cluster tendency (VAT) algorithm [31], to connectivity analysis. The VAT algorithm is a practical technique for visual assessment of cluster tendencies in relational data with lower computational cost than the distance-sensitivity ordering [24]. However, if we want to apply the VAT algorithm to the robust k-means problems, some modifications must be done in order to consider responsibility weights for the k-means process. Another potential future work is comparative study with other fuzzy PCA-based clustering methods [32]–[34].

ACKNOWLEDGMENT

The authors would like to thank the anonymous referees for their valuable comments.

REFERENCES

[1] C. Ding and X. He, "K-means clustering via principal component analysis," in *Proc. Int. Conf. Mach. Learning*, 2004, pp. 225–232.

[2] J. B. MacQueen, "Some methods of classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Stat. Prob.*, 1967, pp. 281–297.

[3] A. Nasser and D. Hamad, "K-means clustering algorithm in projected spaces," in *Proc. 9th Int. Conf. Inf. Fusion*, 2006, pp. 1–6.

[4] K. Weike, P. Azad, and R. Dillmann, "Fast and robust feature-based recognition of multiple objects," in *Proc. 6th IEEE-RAS Int. Conf. Humanoid Robots*, 2006, pp. 264–269.

[5] K. Fukunaga and D. R. Olsen, "An algorithm for finding intrinsic dimensionality of data," *IEEE Trans. Comput.*, vol. C-20, no. 2, pp. 176–183, Feb. 1971.

[6] G. E. Hinton, P. Dayan, and M. Revow, "Modeling the manifolds of images of handwritten digits," *IEEE Trans. Neural Netw.*, vol. 8, no. 1, pp. 65–74, Jan. 1997.

[7] J.-H. Na, M.-S. Park, and J.-Y. Choi, "Pre-clustered principal component analysis for fast training of new face databases," in *Proc. Int. Conf. Control, Autom. Syst.*, 2007, pp. 1144–1149.

[8] M. Xu and P. Fränti, "A heuristic K-means clustering algorithm by kernel PCA," in *Proc. Int. Conf. Image Process.*, 2004, vol. 5, pp. 3503–3506.

[9] T. Su and J. Dy, "A deterministic method for initializing K-means clustering," in *Proc. 16th IEEE Int. Conf. Tools Artif. Intell.*, 2004, pp. 784–786.

[10] Y. Yabuuchi and J. Watada, "Fuzzy principal component analysis and its application," *Biomed. Fuzzy Human Sci.*, vol. 3, pp. 83–92, 1997.

[11] R. N. Davé, "Characterization and detection of noise in clustering," *Pattern Recognit. Lett.*, vol. 12, no. 11, pp. 657–664, 1991.

[12] R. N. Davé and R. Krishnapuram, "Robust clustering methods: A unified view," *IEEE Trans. Fuzzy Syst.*, vol. 5, no. 2, pp. 270–293, May 1997.

[13] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.

[14] J. A. Cuesta-Albertos, A. Gordaliza, and D. C. Matrán, "Trimmed k-means: An attempt to robustify quantizers," *Ann. Stat.*, vol. 25, no. 2, pp. 553–576, 1997.

[15] L. A. Garcia-Escudero and A. Gordaliza, "Robustness properties of k-means and trimmed k-means," *J. Amer. Stat. Assoc.*, vol. 94, no. 447, pp. 956–969, Sep. 1999.

[16] M.-S. Yang, K.-L. Wu, J.-N. Hsieh, and J. Yu, "Alpha-cut implemented fuzzy clustering algorithms and switching regressions," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 3, pp. 588–603, Jun. 2008.

[17] N. R. Pal, K. Pal, J. M. Keller, and J. C. Bezdek, "A possibilistic fuzzy c-means clustering algorithm," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 4, pp. 508–516, Aug. 2005.

[18] R. Krishnapuram and J. M. Keller, "A possibilistic approach to clustering," *IEEE Trans. Fuzzy Syst.*, vol. 1, no. 2, pp. 98–110, May 1993.

[19] F. Masulli and S. Rovetta, "Soft transition from probabilistic to possibilistic fuzzy clustering," *IEEE Trans. Fuzzy Syst.*, vol. 14, no. 4, pp. 516–527, Aug. 2006.

[20] K. Tsuda, M. Minoh, and K. Ikeda, "Extracting straight lines by sequential fuzzy clustering," *Pattern Recognit. Lett.*, vol. 17, pp. 643–649, 1996.

[21] K. Inoue and K. Urahama, "Sequential fuzzy cluster extraction by a graph spectral method," *Pattern Recognit. Lett.*, vol. 20, pp. 699–705, 1999.

[22] U. Luxburg, "A tutorial on spectral clustering," *Stat. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[23] H. Zha, C. Ding, M. Gu, X. He, and H. Simon, "Spectral relaxation for K-means clustering," in *Proc. Adv. Neural Inf. Process. Syst. 14*, 2002, pp. 1057–1064.

[24] C. Ding and X. He, "Linearized cluster assignment via spectral ordering," in *Proc. Int. Conf. Mach. Learning*, 2004, pp. 233–240.

[25] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*. New York: Plenum, 1981.

[26] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer-Verlag, 2006.

[27] S. Miyamoto, H. Ichihashi, and K. Honda, *Algorithms for Fuzzy Clustering*. Berlin Heidelberg: Springer-Verlag, 2008.

[28] B. Schölkopf, A. Smola, and K. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.

[29] A. Asuncion and D. J. Newman. (2007). UCI machine learning repository. School Inf. Comput. Sci., Univ. Calif., Irvine [Online]. Available: http://www.ics.uci.edu/mlearn/MLRepository.html

[30] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (COIL-20)," Dept. Comput. Sci., Columbia Univ., New York, Tech. Rep. CUCS-005-96, 1996.

[31] J. C. Bezdek, R. J. Hathaway, and J. M. Huband, "Visual assessment of clustering tendency for rectangular dissimilarity matrices," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 5, pp. 890–903, Oct. 2007.

[32] K. Honda and H. Ichihashi, "Linear fuzzy clustering techniques with missing values and their application to local principal component analysis," *IEEE Trans. Fuzzy Systems*, vol. 12, no. 2, pp. 183–193, Apr. 2004.

[33] K. Honda and H. Ichihashi, "Regularized linear fuzzy clustering and probabilistic PCA mixture models," *IEEE Trans. Fuzzy Syst.*, vol. 13, no. 4, pp. 508–516, Aug. 2005.
[34] K. Honda, H. Ichihashi, F. Masulli, and S. Rovetta, "Linear fuzzy clustering with selection of variables using graded possibilistic approach," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 5, pp. 878–889, Oct. 2007.

Katsuhiro Honda (M'01) received the B.E., M.E., and D.Eng. degrees in industrial engineering from Osaka Prefecture University, Osaka, Japan, in 1997, 1999 and 2004, respectively.

From 1999 to 2009, he was a Research Associate and an Assistant Professor with Osaka Prefecture University, where he is currently an Associate Professor with the Department of Computer Sciences and Intelligent Systems. His research interests include hybrid techniques of fuzzy clustering and multivariate analysis, data mining with fuzzy data analysis, and neural networks.

Dr. Honda received the Paper Award and the Young Investigator Award from the Japan Society for Fuzzy Theory and Intelligent Informatics in 2002 and 2005, respectively, and delivered a lecture on "Introduction to Clustering Techniques" at the 2004 IEEE International Conference on Fuzzy Systems.

Akira Notsu received the B.E., M.I., and D.Inf. degrees from Kyoto University, Kyoto, Japan, in 2000, 2002, and 2005, respectively.

He is currently an Assistant Professor with the Department of Computer Sciences and Intelligent Systems, Osaka Prefecture University, Osaka, Japan. His research interests include agent-based social simulation, communication networks, game theory, human–machine interface, and cognitive engineering.

Hidetomo Ichihashi (M'94) received the B.E. and D.Eng. degrees in industrial engineering from Osaka Prefecture University, Osaka, Japan, in 1971 and 1986, respectively.

From 1971 to 1981, he was with the Information System Center, Matsushita Electric Industrial Corporation, Ltd., Tokyo, Japan. Since 1981, he has been with Osaka Prefecture University, where he was a Research Associate, an Assistant Professor, and an Associate Professor and is currently a Professor with the Department of Computer Sciences and Intelligent Systems. His research interest include adaptive modeling of group-method-of-data-handling-type neural networks, fuzzy c-means clustering and classifiers, data mining with fuzzy data analysis, human–machine interfaces, and cognitive engineering.