Noise Fuzzy Clustering in Kernel Feature Spaces for Robust PCA

Hidetomo Ichihashi, Katsuhiro Honda
Graduate School of Engineering, Osaka Prefecture University
1-1 Gakuen-cho, Sakai, Osaka 599-8531 Japan
ichi@ie.osakafu-u.ac.jp

Abstract— **Inokuchi & Miyamoto studied the multidimensional data mapped by a Mercer kernel function using Kohonen's Self-Organizing Map. Along the line of Noise Clustering due to Davé, which is considered as a kind of robust M-estimation, we develop a robust kernel PCA method and clarify the data structure in a high dimensional feature space. Noise clustering approaches applied to a fuzzy counterpart of Gaussian mixture models plays a substantial role for finding that the data structure in the original input space is preserved to some extent in the feature space defined by dot products in terms of Gaussian kernel function.**

I. Introduction

Gaussian mixture models (GMM) [1] is well recognized as a statistical technique for density estimation, where a probability density function (PDF) is approximated by a mixture of Gaussian distribution functions rather than a single parametric function. The GMM uses the likelihood function as a measure of fit.

Hathaway [2] pointed out that there exists a close relationship between the Fuzzy c-Means (FCM) [3] clustering and the GMM. Entropy method that uses an additional term of entropy for fuzzification in the FCM was proposed by Miyamoto and Mukaidono [4]. A similar entropy term was considered by Davé and Krishnapuram [5] to prevent trivial solution within the scope of Possibilistic c-Means due to Krishnapuram and Keller [6]. A fuzzy counterpart of GMM, which has a modified FCM clustering objective function with an additional term known as Kullback-Leibler information is called FCM with K-L regularizer (KLFCM) [7]. We can also find a close relationship between the FCM and the deterministic annealing (DA) by Rose [8].

Like the fuzzy c-varieties and fuzzy c-elliptypes clustering [3], the GMM method is good only when a dataset contains clusters that are approximately the same shape, i.e., hyper elliptic. The fuzzy c-spherical shells and their modified methods [9], [10], [11] aim at detection of shell type clusters of the same class of models, for example, all are spherical shells.

In the passed several years, a number of powerful Mercer kernel-based learning machines [12], [13], e.g., support vector machines (SVMs), kernel Fisher discriminant (KFD), kernel principal component analysis (KPCA), kernel based clustering [14], kernel fuzzy c-means (KFCM) [15], and kernel-KLFCM [16] have been proposed. A common principle of these methods is to construct nonlinear variants of linear algorithms by substituting dot products by kernel functions. The resulting kernel algorithm can be interpreted as running the original algorithm on feature space mapped objects $\boldsymbol{\phi}(x_i)$.

In the kernel-KLFCM [16], when the number of feature vectors and clusters are n and c respectively, this kernel approach can find up to $c \times n$ nonzero eigenvalues. For reducing the number of parameters, a way to control the number in the mixture of probabilistic principal component analysis (PPCA) by Tipping and Bishop [17] was adopted. In [18], [19], a fuzzy counterpart of probabilistic PCA mixture models was proposed based on the relationship between Local PCA and linear fuzzy clustering. By further adopting this approach to kernel based clustering, the kernel-KLFCM [16] provides a partitioning with flexible shapes of clusters in the original input data space.

By using Kohonen Self-Organizing Map (SOM) [20], Inokuchi & Miyamoto [21] studied the data structure of multidimensional data mapped implicitly by $\boldsymbol{\phi}$ in terms of a kernel function. With this SOM approach, they discovered a typical data structure of the high dimensional feature space.

Any clustering algorithm needs to be robust against noise or outliers in order to be useful in practice. Davé's noise clustering (NC) [22] is one of the popular methods in fuzzy clustering. It was shown that the NC has a close relationship with a robust M-estimator [5]. Previously we applied the noise clustering to KLFCM in [23], and showed that the NC approach is quite robust for detecting linear clusters from heavily noisy data sets.

Along the line of NC approach due to Davé, which is considered as a kind of robust M-estimation, we develop a robust kernel based PCA method, which utilize Mahalanobis distances rather than Euclidian distances between feature vectors and linear prototypical subspaces. The proposed approach clarifies the data structure in a high dimensional feature space. The noise clustering approach applied to the fuzzy counterpart of GMM plays a substantial role for finding that the data structure in the original input space is preserved to some extent in the feature space.

II. Robust approach to kernel based KLFCM

In this section we will combine two of our previous approaches, which are Noise-KLFCM [23] and Kernel-KLFCM [16]. The former is related with robust M-estimation and the latter is related with kernel based PCA. We will first describe these approaches so that the

reader will be able to see that combining these two approaches is straight forward.

A. Noise clustering for KLFCM

Let s dimensional vector $\boldsymbol{x}_k$ represent the kth object or sample from a given set of n unlabelled objects. Each feature vector consists of s real-valued measurements describing the features of the object represented by $\boldsymbol{x}$.

The FCM clustering partition a data set by introducing memberships to fuzzy clusters. p dimensional vector $\boldsymbol{v}_i$ denotes prototype parameter (i.e., cluster centroid). u_{ik} denotes the membership of the kth data to the ith cluster. The clustering criterion used to define good clusters for fuzzy c-means partitions is the FCM objective function as:

$$J_m = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m d_{ik}, \tag{1}$$

where m is the weighting exponent on each fuzzy membership. The larger m is, the fuzzier the partition becomes. The nonnegative membership u_{ik} sum to one with respect to c clusters for each object.

$$d_{ik} = (\boldsymbol{x}_k - \boldsymbol{v}_i)^\top A_i^{-1}(\boldsymbol{x}_k - \boldsymbol{v}_i) \tag{2}$$

is a measure of the distance from $\boldsymbol{x}_k$ to the ith cluster prototype. The Euclidean distance metric is often used where A_i is a unit matrix. In the modified FCM by Gustafson and Kessel [24], the matrices A_i are also decision variables and each size of $|A_i|$ is constrained to a certain value. The optimal u_{ik} and $\boldsymbol{v}_i$ for all i and k are sought using a fixed-point iteration scheme, which is similar to the GMM algorithm.

Noise fuzzy clustering (NC) was proposed by Davé [22] so that noise data or outliers will be included in the noise cluster. Noise is considered to be a separate class. Noise prototype is an entity such that it is always at the same distance from every point in the noise cluster. Let $c+1$ th cluster be a noise cluster, then the objective function is defined as:

$$J_{m\delta} = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m d_{ik} + \delta\sum_{k=1}^{n}(u_{c+1\ k})^m \tag{3}$$

with a constraint $\sum_{i=1}^{c+1} u_{ik} = 1$. By minimizing this objective function, if $d_{ik} > \delta$, $\boldsymbol{x}_k$ tends to belong in the $c+1$ th cluster. Hence the δ should be positive small number. If δ is large, the number of data points in the noise cluster becomes small. If it is set to a negative number then all data will be included in the noise cluster.

Hathaway [2] provided an interpretation of the optimization problem of negative log-likelihood in the GMM and regarded the EM algorithm as a penalized version of the hard means clustering algorithm. The negative log-likelihood to be minimized is written as:

$$\begin{aligned} J_H &= \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}\log u_{ik} \\ &+ \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}\log(1/(\pi_i p_i(\boldsymbol{x}_k))), \end{aligned} \tag{4}$$

where p_i denotes a Gaussian density function. In [4], an entropy term K and a positive parameter λ are introduced and $J_\lambda = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}d_{ik} + \lambda K$ is minimized instead of J_m. This approach is referred to as entropy regularization. By further extension, replacing the entropy term with K-L information, we considered the minimization of the following objective function under the constraints that both the sum of u_{ik} and the sum of π_i with respect to i equal one [23].

$$\begin{aligned} J_{\lambda\gamma} &= \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}d_{ik} + \delta\sum_{k=1}^{n} u_{c+1\ k} \\ &+ \lambda\sum_{i=1}^{c+1}\sum_{k=1}^{n} u_{ik}\log\frac{u_{ik}}{\pi_i} \\ &+ \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}\log|A_i| \\ &+ \sum_{k=1}^{k} u_{c+1\ k}\ \log\alpha \\ &= \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}d_{ik} + \gamma\sum_{k=1}^{n} u_{c+1\ k} \\ &+ \lambda\sum_{i=1}^{c+1}\sum_{k=1}^{n} u_{ik}\log\frac{u_{ik}}{\pi_i} \\ &+ \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}\log|A_i|, \end{aligned} \tag{5}$$

where $\gamma = \delta + \log\alpha$. δ and α are predetermined values and correspond to $d_{c+1\ k}$ and $\log|A_{c+1}|$ respectively.

When the number c of ordinary clusters is given, the $c+1$th cluster plays the role of the noise cluster.

From the necessary condition of optimality, for $i \le c$, we have

$$\boldsymbol{v}_i = \frac{\sum_{k=1}^{n} u_{ik}\boldsymbol{x}_k}{\sum_{k=1}^{n} u_{ik}}. \tag{6}$$

By denoting

$$\begin{aligned} W_k &= \sum_{j=1}^{c}\pi_j\exp\Big(-\frac{1}{\lambda}d_{jk}\Big)|A_j|^{-1/\lambda} \\ &+ \pi_{c+1}\exp\Big(-\frac{\gamma}{\lambda}\Big), \end{aligned} \tag{7}$$

we have for $i \le c$

$$u_{ik} = \pi_i\exp\Big(-\frac{1}{\lambda}d_{ik}\Big)|A_i|^{-1/\lambda}/W_k \tag{8}$$

and for $i = c+1$

$$u_{ik} = \pi_{c+1}\exp\Big(-\frac{\gamma}{\lambda}\Big)/W_k. \tag{9}$$

Matrix A_i is for $i \leq c$

$$A_i = \frac{\sum_{k=1}^{n} u_{ik}(\boldsymbol{x}_k - \boldsymbol{v}_i)(\boldsymbol{x}_k - \boldsymbol{v}_i)^\top}{\sum_{k=1}^{n} u_{ik}}, \quad (10)$$

and for $i \leq c+1$

$$\pi_i = \frac{\sum_{k=1}^{n} u_{ik}}{\sum_{j=1}^{c+1}\sum_{k=1}^{n} u_{jk}} = \frac{1}{n}\sum_{k=1}^{n} u_{ik}. \quad (11)$$

The clustering algorithm in this case is also the fixed point iteration as in the conventional FCM.

B. KLFCM with kernel trick

The following theory used as a basis for applying the kernel trick is based upon Reproducing Kernel Hilbert Spaces (RKHS). A dot product in feature space has an equivalent kernel in input space,

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^\top \boldsymbol{\phi}(\boldsymbol{x}') \quad (12)$$

provided certain conditions (Mercer's Conditions). These kernel functions can be interpreted as representing the dot product of data objects implicitly mapped into a nonlinear related feature space. A typical example of Mercer kernel is the Gaussian kernel as:

$$k(\boldsymbol{x}, \boldsymbol{x}_i) = \exp(-(\boldsymbol{x} - \boldsymbol{x}_i)^\top (\boldsymbol{x} - \boldsymbol{x}_i)/\beta). \quad (13)$$

Let us consider that the original input data space will be mapped into a high dimensional feature space through some nonlinear mapping $\boldsymbol{\phi}$. Then the fuzzy covariance matrix for the ith cluster is written in the matrix form as:

$$S_i = ((n\pi_i)^{-\frac{1}{2}} M_i^{\frac{1}{2}} \Phi_i)^\top ((n\pi_i)^{-\frac{1}{2}} M_i^{\frac{1}{2}} \Phi_i), \quad (14)$$

where $\Phi_i = (\boldsymbol{\phi}(\boldsymbol{x}_1) - \boldsymbol{v}_i^\phi, ..., \boldsymbol{\phi}(\boldsymbol{x}_n) - \boldsymbol{v}_i^\phi)^\top$ and $M_i = \mathrm{diag}(u_1, ..., u_n)$. $\boldsymbol{v}_i^\phi$ is a cluster centroid in the extended feature space. The dimension of $\boldsymbol{\phi}$ and $\boldsymbol{v}$ is assumed to be r in the extended feature space. Eigenvalue decomposition of S_i may be written as:

$$S_i = W_i \Delta_i^2 W_i^\top, \quad (15)$$

where $W_i = (\boldsymbol{w}_{i1}, ..., \boldsymbol{w}_{ir})$ is an $r \times r$ matrix and $\boldsymbol{w}_{i1}, ..., \boldsymbol{w}_{ir}$ are eigenvectors associated with positive eigenvalues $(\delta_{i1}^2, ..., \delta_{ir}^2)$ of S_i. The vectors are normalized as $\boldsymbol{w}_{il}^\top \boldsymbol{w}_{il} = 1$. $\Delta^2 = \mathrm{diag}(\delta_{i1}^2, ..., \delta_{ir}^2)$ is a diagonal matrix of the eigenvalues. $\boldsymbol{v}_i^\phi$ is a centroid for the ith cluster.

$$\boldsymbol{v}_i^\phi = \frac{\sum_{k=1}^{n} u_{ik}\boldsymbol{\phi}(\boldsymbol{x}_k)}{\sum_{k=1}^{n} u_{ik}}. \quad (16)$$

By the singular value decomposition, we have

$$(n\pi_i)^{-\frac{1}{2}} M_i^{\frac{1}{2}} \Phi_i = F_i \Delta_i W_i^\top, \quad (17)$$

where F_i is an $n \times r$ matrix. If the left-hand side of Eq.(17) is non-singular, the distance between $\boldsymbol{\phi}(\boldsymbol{x}_k)$ and cluster centroid $\boldsymbol{v}_i^\phi$ can be written as:

$$\begin{aligned} d_{ik} &= (\boldsymbol{\phi}(\boldsymbol{x}_k) - \boldsymbol{v}_i^\phi)^\top W_i \Delta_i^{-2} W_i^\top (\boldsymbol{\phi}(\boldsymbol{x}_k) - \boldsymbol{v}_i^\phi) \\ &= n\pi_i u_{ik}^{-1} \boldsymbol{f}_{ik}^\top \boldsymbol{f}_{ik}, \end{aligned} \quad (18)$$

where $F_i = (\boldsymbol{f}_{i1}, ..., \boldsymbol{f}_{in})^\top$. $\boldsymbol{f}_{ik}$'s are r dimensional vectors.

When the explicit form of $\boldsymbol{\phi}$ is not known, for obtaining the values of F_i and Δ_i, let us define an $n \times n$ matrix K_i and rewrite it by using Eq.(17) as:

$$\begin{aligned} K_i &= (n\pi_i)^{-1} M_i^{\frac{1}{2}} \Phi_i \Phi_i^\top M_i^{\frac{1}{2}} \\ &= F_i \Delta_i^2 F_i^\top. \end{aligned} \quad (19)$$

Let $\Phi^* \Phi^{*\top}$ be a matrix in terms of dot product

$$\Phi^* \Phi^{*\top} = (\boldsymbol{\phi}(\boldsymbol{x}_k)^\top \boldsymbol{\phi}(\boldsymbol{x}_l))_{n \times n}, \quad (20)$$

where $(\ )_{n\times n}$ denotes a matrix of $n \times n$ dimension. The elements of $\Phi^* \Phi^{*\top}$ are dot products of non-centered vectors $\phi(\boldsymbol{x})$. Then centered $\Phi_i \Phi_i^\top$ can be written as:

$$\begin{aligned} \Phi_i \Phi_i^\top &= \Phi^* \Phi^{*\top} - \Phi^* \Phi^{*\top} \overline{\boldsymbol{u}_i} \mathbf{1}_{n\times 1}^\top \\ &\quad - \mathbf{1}_{n\times 1} \overline{\boldsymbol{u}_i}^\top \Phi^* \Phi^{*\top} \\ &\quad + (\mathbf{1}_{n\times 1} \overline{\boldsymbol{u}_i}^\top) \Phi^* \Phi^{*\top} (\overline{\boldsymbol{u}_i} \mathbf{1}_{n\times 1}^\top). \end{aligned} \quad (21)$$

$\mathbf{1}_{n\times 1}$ denotes the vector of dimension $n \times 1$ with all entries equal to 1.

$$\overline{\boldsymbol{u}_i} = (u_{i1} / \sum_{k=1}^{n} u_{ik}, ..., u_{in} / \sum_{k=1}^{n} u_{ik})^\top. \quad (22)$$

Thus $\Phi_i \Phi_i^\top$ can be calculated by replacing the dot product in $\Phi^* \Phi^{*\top}$ with a kernel function $k(\boldsymbol{x}_k, \boldsymbol{x}_l)$, such that

$$\Phi^* \Phi^{*\top} = (\boldsymbol{\phi}(\boldsymbol{x}_k)^\top \boldsymbol{\phi}(\boldsymbol{x}_l))_{n\times n} = (k(\boldsymbol{x}_k, \boldsymbol{x}_l))_{n\times n}. \quad (23)$$

F_i and Δ_i are obtained from the eigenvalue decomposition of K_i as in Eq.(19). The remaining value that we need for updating u_{ik} is $|S_i|$. If $\mathrm{rank}(S_i) = r$,

$$|S_i| = |W_i||\Delta_i^2||W_i^\top| = |\Delta_i^2| = \prod_{l=1}^{r} \delta_{il}^2. \quad (24)$$

C. Robust local kernel PCA

In the feature space extended by some kernel function, the dimensionality r or the number of positive eigenvalues of S_i is unknown but usually large, and the number of observation n may exceed r. Because this kernel approach can find up to $c \times n$ nonzero eigenvalues, reduction of the number of decision variables (e.g., F_i) is significant.

Unlike the global nonlinear approaches, GMM or KLFCM is to model nonlinear structure with a collection, or mixture, of local linear sub-models of PCA. When estimating covariance structures in high dimensions, while not over-constraining the model flexibility, Tipping and Bishop proposed a way to control the number of parameters in the mixture of probabilistic principal component analysis (PPCA) [17]. In [18], [19], a fuzzy counterpart of probabilistic PCA mixture models was proposed based on the relationship between Local PCA and linear fuzzy clustering. The algorithm is regarded as a modified FCV

algorithm with regularization by K-L information. Although in both of the approaches the number r should be known, by making this parameter an adjustable one, we apply the kernel trick to the clustering method in a high dimensional feature space.

Let S_i' denotes an approximation of S_i in Eq.(14) and (15) for $p < r$ as:

$$S_i' = W_i^p((\Delta_i^p)^2 - \sigma_i^2 I_r)W_i^{p\top} + W_i(\sigma_i^2 I_r)W_i^\top, \qquad (25)$$

where W_i^p is an $r \times p$ matrix and Δ_i^p is a $p \times p$ diagonal matrix.

$$\sigma_i^2 = \frac{1}{r-p}(\mathrm{trace}(K_i) - \sum_{l=1}^{p}\delta_{il}^2), \qquad (26)$$

Unfortunately r is an unknown dimensionality of the feature space mapped by an unknown function $\boldsymbol{\phi}$, so we make r an adjustable parameter. As we will see in the numerical example, this parameter does not significantly affect the clustering results. The squared fuzzy Mahalanobis distance between a point $\boldsymbol{\phi}(\boldsymbol{x}_k)$ and a cluster centroid $\boldsymbol{v}_i^\phi$ can be approximated as:

$$\begin{aligned} d_{ik} &= n\pi_i u_{ik}^{-1}\Big(\boldsymbol{f}_{ik}^{p\top}\boldsymbol{f}_{ik}^p \\ &+ \frac{1}{\sigma_i^2}(\frac{u_{ik}}{n\pi_i}(\boldsymbol{\phi}(\boldsymbol{x}_k) - \boldsymbol{v}_i^\phi)^\top(\boldsymbol{\phi}(\boldsymbol{x}_k) - \boldsymbol{v}_i^\phi) \\ &- \boldsymbol{f}_{ik}^{p\top}(\Delta_i^p)^2\boldsymbol{f}_{ik}^p)\Big), \end{aligned} \qquad (27)$$

where $\boldsymbol{f}_{ik}^p = (f_{ik1}, ..., f_{ikp})^\top$ is p dimensional, and the k th diagonal element of Eq.(21) is

$$\begin{aligned} &(\boldsymbol{\phi}(\boldsymbol{x}_k) - \boldsymbol{v}_i^\phi)^\top(\boldsymbol{\phi}(\boldsymbol{x}_k) - \boldsymbol{v}_i^\phi) = \boldsymbol{\phi}(\boldsymbol{x}_k)^\top\boldsymbol{\phi}(\boldsymbol{x}_k) \\ &-2\overline{\boldsymbol{u}_i}^\top\Phi^{*\top}\boldsymbol{\phi}(\boldsymbol{x}_k) + \overline{\boldsymbol{u}_i}^\top\Phi^*\Phi^{*\top}\overline{\boldsymbol{u}_i}. \end{aligned} \qquad (28)$$

$\overline{\boldsymbol{u}_i}$ is defined in Eq.(22).

It should be noted that u_{ik} must be positive in Eq.(18) and (27). Thus we add a very small positive number as 1×10^{-7} to u_{ik} in each repetition of update.

We approximate $|S_i|$ by the product of largest p eigenvalues of K_i and $\sigma_i^{2(r-p)}$ where r is some estimated positive integer.

$$|S_i'| \simeq (\prod_{l=1}^{p}\delta_{il}^2)\sigma_i^{2(r-p)}. \qquad (29)$$

When σ_i is set to a small positive number, this method reduces to a FCV type clustering by the property shown in [18], [19]. By denoting

$$\begin{aligned} W_k &= \sum_{j=1}^{c}\pi_j \exp\big(-\frac{1}{\lambda}d_{jk}\big)|S_i'|^{-1/\lambda} \\ &+ \pi_{c+1}\exp\big(-\frac{\gamma}{\lambda}\big), \end{aligned} \qquad (30)$$

the membership to cluster u_{ik} is written for $i \le c$ as:

$$u_{ik} = \pi_i \exp\big(-\frac{1}{\lambda}d_{ik}\big)|S_i'|^{-1/\lambda}/W_k, \qquad (31)$$

and for $i = c+1$

$$u_{ik} = \pi_{c+1}\exp\big(-\frac{\gamma}{\lambda}\big)/W_k. \qquad (32)$$

and the ratio π_i may be written as:

$$\pi_i = \frac{\sum_{k=1}^{n}u_{ik}}{\sum_{j=1}^{c}\sum_{k=1}^{n}u_{jk}} = \frac{1}{n}\sum_{k=1}^{n}u_{ik}. \qquad (33)$$

The algorithm is the repetition of these update for all clusters, i.e., $i = 1, ..., c$ and may be described as:

Noise Kernel-KLFCM algorithm

Step 1: Initialize u_{ik} for all i and k with random numbers.

Step 2: Calculate π_i for all i by Eq.(33).

Step 3: Calculate K_i and its eigenvalue decomposition for all i using Eqs.(19)-(23).

Step 4: Calculate u_{ik} for all i and k by Eqs.(26)-(32).

Step 5: If

$$\max_{i,\ k}\|u_{ik}^{NEW} - u_{ik}^{OLD}\| < \epsilon$$

is satisfied, then terminate, else go to *Step 2*.

III. Numerical Example

To provide some intuition on how proposed robust kernel PCA approach clarifies the data structure in feature space, we show a set of experiments with an artificial 2-D data, using a Gaussian kernel and a polynomial kernel. PCA is an orthogonal transformation of the coordinate system in which we describe our data. It is often the case that a small number of principal components is sufficient to account for most of the structure in the data. Fig.1 shows the projection of data obtained by PCA on the full dataset in feature space defined by dot products in terms of the Gaussian kernel. From left to right and from top to bottom, the first 4 principal components are plotted, in order of decreasing eigenvalue size. The 2-D data in input space is shown on the left bottom of the figure. Fig.2 shows the 8 largest eigenvalues.

Figs.3 and 4 show the projection of data obtained by Noise Kernel-KLFCM on the data respectively in the 1st (c=1) and the noise (c=2) clusters in feature space. The obtained 4 principal components are plotted. The dimensionality of the feature space was assumed to be 9 (r=9), and the number of eigenvectors p=4. Other parameters were chosen as λ=2.0, γ=10, β=70. The 2-D data in input space was clustered as shown on the left bottom of the figures in which the data in the 1st and noise clusters are depicted crisply by the largest membership values and represented by circles and triangles respectively. Fig.5 shows the 8 eigenvalues for the 1st cluster (left) and the noise cluster (right) obtained by the same algorithm as above,

in which r=13 and p=8. Though the number of used eigenvectors (p=8) was greater than the result in Figs.3 and 4 (p=4), we had similar results. Thus the number p is not deemed to significantly affect the results. The largest 2 eigenvalues contribute greatly for both the clusters.

From Eq.(13), we see that these results are invariant with respect to origins of coordinates $\boldsymbol{x}$, but are not scale invariant. Although the circular cluster is clearly detected as in Fig.4, slightly vague clusters are obtained in Figs.6 and 7 for the data $\boldsymbol{x}$ multiplied by 0.01. Fig.8 shows the result with polynomial kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i^\top \boldsymbol{x}_j)^4$, $\lambda = 2$, $\gamma = 0$, $p = 2$ and $r = 3$. The eigenvalues explain that the data distribution in feature space is almost within two dimensions. The learning curve is shown on the right bottom of the figure.

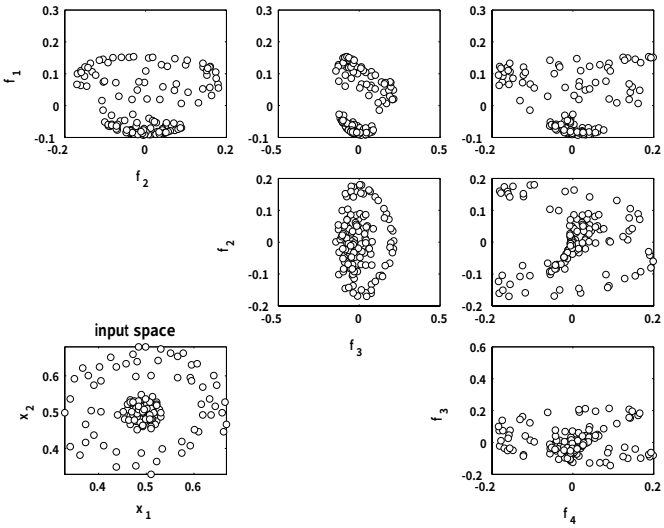

Fig. 1. Projections of data in feature space by PCA on full dataset.

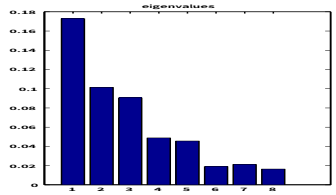

Fig. 2. Eigenvalues by PCA on full dataset.

IV. Conclusion

We have proposed a kernel based noise fuzzy clustering for local robust PCA on the feature space mapped objects. The data structure is clarified by the robust PCA. The common case such as in computer vision applications involves intra-sample outliers which affect some, but not all, of the variables in a data sample. Taking this point into consideration, comparisons with the de-noising by kernel based PCA [12], [13] is left for future study.

References

[1] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

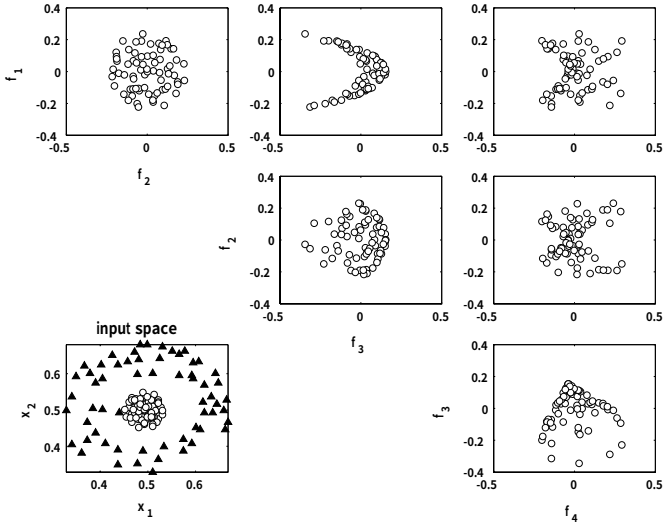

Fig. 3. Projections of data by Noise Kernel-KLFCM on 1st cluster.

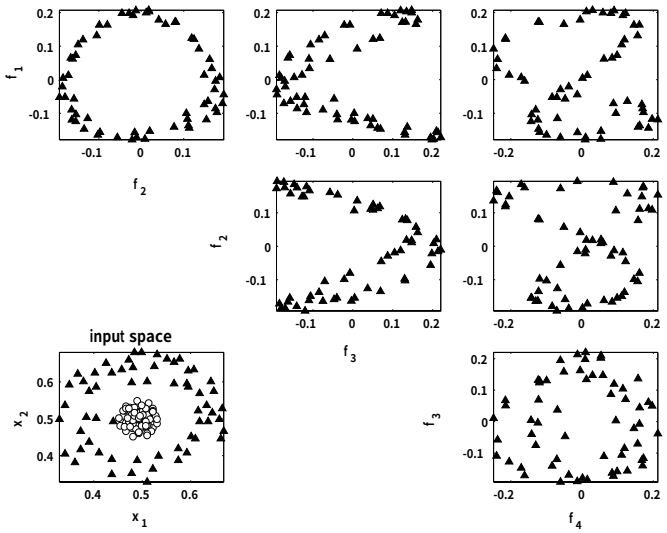

Fig. 4. Projections of data by Noise Kernel-KLFCM on noise cluter (c=2).

[2] R. J. Hathaway, Another interpretation of the EM algorithm for mixture distributions, *Statistics and Probability Letters*, Vol.4, pp.53-56, 1986.

[3] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, 1981.

[4] S. Miyamoto and M. Mukaidono, Fuzzy c-means as a regularization and maximum entropy approach, *Proc. of the 7th International Fuzzy Systems Association World Congress(IFSA '97)*, Vol.II, pp.86-92, 1997.

[5] R. N. Davé and R. Krishnapuram, Robust clustering methods-A unified approach, *IEEE Transactions on Fuzzy Systems*, Vol.5, No.2, pp.270-293, 1997.

[6] R. Krishnapuram and J. Keller, A possibilistic approach to clustering, *IEEE Transactions on Fuzzy Systems*, Vol.1, pp.98-110, 1993.

[7] H. Ichihashi, K. Miyagishi and K. Honda, Fuzzy c-means clustering with regularization by K-L information, *Proc. of 10th IEEE International Conference on Fuzzy Systems*, Melbourne, November, 2001.

[8] K. Rose, Deterministic annealing for clustering, compression, classification, regression, and related optimization problems, *Proc. of the IEEE*, vol.86, no.11, pp.2210-2239, 1998.

[9] R. N. Davé, Fuzzy-shell clustering and applications to circle de-

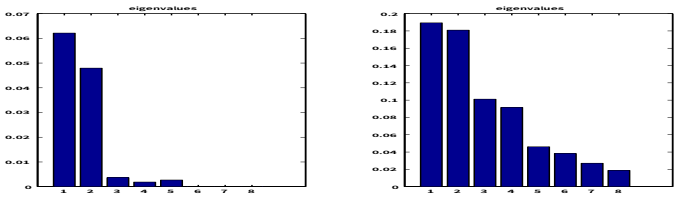

Fig. 5. Eigenvalues by Noise Kernel-KLFCM.

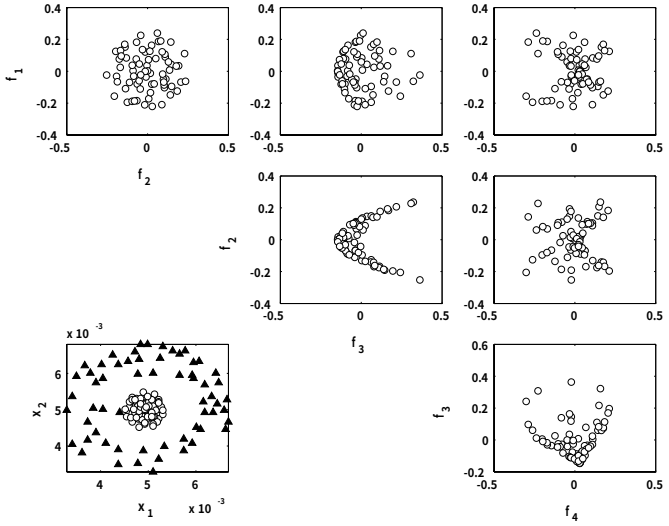

Fig. 6. Projections of data by Noise Kernel-KLFCM on 1st cluster multiplied by 0.01.

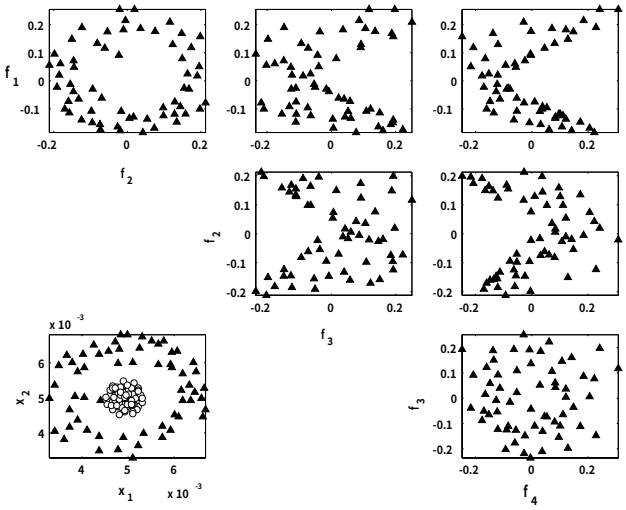

Fig. 7. Projections of data by Noise Kernel-KLFCM on noise cluster (c=2) multiplied by 0.01.

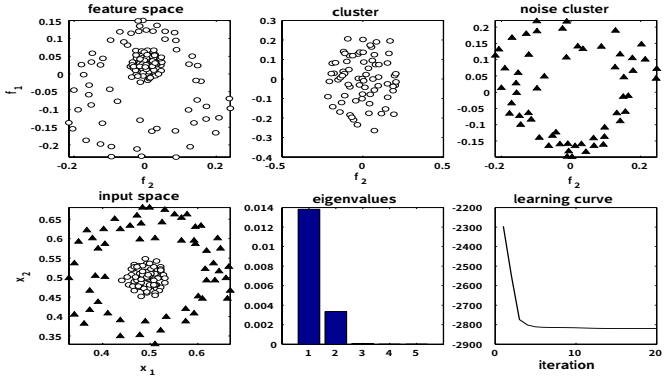

Fig. 8. Result by Noise Kernel-KLFCM with polynomial kernel.

tection in digital images, *Internat. J. General Systems*, Vol.16, pp.343-345, 1990.

[10] R. N. Davé, Generalized fuzzy c-shells clustering and detection of circular and elliptical boundaries, *Pattern Recognition*, Vol.25(7), pp.713-722, 1992.

[11] R. Krishinapuram, H. Frigui and O. Nasraoui, Quadric shell clustering algorithms and the detection of second degree curves, *Pattern Recognition Letter*, Vol.14, No.7, pp.545-552, 1993.

[12] B. Schölkopf, S. Mika, C.J.C. Burges, P. Knirsch, K.-R. Müller, Rätsch, G. and A.J. Smola, Input space versus feature space in kernel-based methods. *IEEE Transactions on Neural Networks* Vol.10, No.5, pp.1000-1017, 1999.

[13] K-R. Müller, S.Mika, G. Ratsch, K. Tsuda, and B. Schölkopf, An Introduction to Kernel-Based Learning Algorithms, *IEEE Transactions on Neural Networks*, Vol.12, No.2, pp.181-201, 2001.

[14] M.Girolami, Mercer kernel based clustering in feature space,*I.E.E.E. Transaction on Neural Networks*, Vol.13, No.4, pp.780-784,2002.

[15] S. Miyamoto and D.Suizu, Fuzzy c-means clustering using kernel functions in support vector machines, *J. of Advanced Computational Intelligence*, Vol.7, No. 1, pp.25-30, 2003.

[16] H. Ichihashi and K. Honda, Application of kernel trick to fuzzy c-means with regularization by K-L information *Proc. of the Fourth International Conference on Intelligent Technologies*, pp.626-635, 2003.

[17] M.E. Tipping, C.M. Bishop, Mixtures of probabilistic principal component analysers, *Neural Computation*, Vol.11, pp.443-482, 1999.

[18] Y. Mori, K. Honda, A. Kanda, H. Ichihashi, A unified view of probabilistic PCA and regularized linear fuzzy clustering, *Proc. the 2003 International Joint Conference on Neural Networks*, pp.541-546, 2003.

[19] K. Honda, H. Ichihashi, Regularized linear fuzzy clustering and probabilistic PCA mixture models, *IEEE Transactions on Fuzzy Systems* (to appear).

[20] T. Kohonen, *Self-Organizing Maps*, Springer-Verlag, Heidelberg, 1995.

[21] R. Inokuchi and S. Miyamoto, LVQ Clustering and SOM using a kernel function, *Proc. of 12th IEEE International Conference on Fuzzy Systems*, 2004.

[22] R. N. Davé, Characterization and detection of noise in clustering, *Pattern Recognition Letters*, Vol.12, pp.657-664, 1991.

[23] H. Ichihashi and K. Honda, On parameter setting in applying Davé's noise fuzzy clustering to Gaussian mixture models, *Proc. of IEEE International Conference on Fuzzy Systems*, Budapest,July, 2004.

[24] D. E. Gustafson and W. C. Kessel, Fuzzy clustering with a fuzzy covariance matrix, *Proc. IEEE CDC*, Vol.2, pp.761-766, 1979.