A Unified View of Probabilistic PCA and Regularized Linear Fuzzy Clustering *

Yoshio Mori, Katsuhiro Honda, Akihiro Kanda and Hidetomo Ichihashi
Graduate School of Engineering, Osaka Prefecture University
1-1 Gakuen-cho Sakai Osaka JAPAN
honda@ie.osakafu-u.ac.jp

Abstract

FCM-type fuzzy clustering approaches are closely related to Gaussian Mixture Models (GMMs) and the objective function of Fuzzy c-Means with regularization by K-L information (KFCM) is optimized by an EM-like algorithm. In this paper, we propose to apply probabilistic PCA mixture models to linear clustering following the discussion on the relationship between Local PCA and linear fuzzy clustering. Although the proposed method is a kind of the constrained model of KFCM, the algorithm includes the Fuzzy c-Varieties (FCV) algorithm as a special case, and the algorithm can be regarded as a modified FCV algorithm with regularization by K-L information.

1 Introduction

Local Principal Component Analysis (Local PCA) is a useful tool for finding local features of large scale databases. The goal of Local PCA is to partition the data set into several small subregions and find linear expressions of the data subsets. For the task, several statistical approaches have been used. Fukunaga *et al.* [1] proposed local Karhunen-Loéve expansions that follows the clustering stage based on the similarities of data points. Kambhatla *et al.* [2] and Hinton *et al.* [3] used iterative algorithms that achieve the natural partitioning based on the reconstruction distances. And the "soft" version [3] is performed in an expectation-maximization (EM) framework [4] in which the partition assignments are considered as "missing data" and the responsibility of a principal component analyzer for each data point is estimated by using the corresponding reconstruction cost. Then the local models are determined by the maximization of a pseudo-likelihood function while no probability density is defined.

Roweis [5] and Tipping *et al.* [6] defined probabilistic models for PCA in which all of the model parameters are estimated through the maximization of a single likelihood function and the advantage of a probabilistic density function is available. The Probabilistic Principal Component Analysis (PPCA) mixture model [6] is regarded as a kind of constrained Gaussian Mixture Models (GMMs) and the flexibility for density estimation is better than that of GMMs with full covariance matrices when modeling high-dimensional data with a small number of examples [7].

Cluster analysis is also developed for capturing local substructures. Fuzzy c-Means (FCM) [8] and its derivatives [9, 10, 11] are closely related to GMMs and Ichihashi *et al.* [12] proposed a clustering algorithm, which is similar to the EM algorithm for GMMs, by using the fuzzification technique with K-L information. Fuzzy c-Varieties (FCV) [13] is a linear fuzzy clustering technique that captures the local linear structures of data sets and is regarded as a Local PCA technique because the prototypes of clusters are estimated by solving the eigenvalue problems of fuzzy scatter matrices.

In this paper, we discuss the relationship between Local PCA and linear fuzzy clustering and propose a new clustering technique that can capture the local linear structures flexibly. Although the objective functions of clustering techniques are based on the minimization of deviations between data points and prototypes, some linear fuzzy clustering algorithms are similar to that of Local PCA. The proposed algorithm is defined by expanding the PPCA mixture models and is also a constrained model of the FCM clustering with regularization by K-L information.

In the next section, we present a brief review of the relationship between mixture densities and fuzzy clustering followed by a unified view of Local PCA and linear fuzzy clustering.

*IJCNN 2003 Conference Proceedings, 541-546 (2003)

2 Mixture Densities and Fuzzy Clustering

2.1 Mixtures of Normal Densities

Let $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{im})^\top$, $i = 1, \cdots, n$ denote m dimensional observations of n samples. The mixture density for a sample $\boldsymbol{x}$ is given as the following probability density function:

$$p(\boldsymbol{x}) = \sum_{c=1}^{C} \pi_c p_c(\boldsymbol{x}), \tag{1}$$

where the conditional densities $p_c(\boldsymbol{x})$'s are the component densities and the mixing coefficients π_c's are the a priori probabilities. The most widely used model is the Gaussian Mixture Models (GMMs, e.g., [14]) in which the component densities are Gaussian distributions with covariance matrices Σ_c's that are chosen to be full, diagonal or spherical and means $\boldsymbol{b}_c$'s:

$$p_c(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{b}_c, \Sigma_c). \tag{2}$$

The negative log-likelihood to be minimized is defined as

$$L_{gmm} = -\sum_{i=1}^{n} \log\left\{\sum_{c=1}^{C} \pi_c p_c(\boldsymbol{x}_i)\right\}. \tag{3}$$

The optimal parameters are derived by EM algorithm that is an iterative algorithm composed of E-step (Expectation step) and M-step (Maximization step). In the case of full covariance matrices, the two steps are represented as follows:

- E-step Estimation of responsibility (posterior probability) of each data point for component densities:

$$\begin{aligned} u_{ci} &= \frac{\pi_c p_c(\boldsymbol{x}_i)}{\sum_{l=1}^{C} \pi_l p_l(\boldsymbol{x}_i)} \\ &= \frac{\pi_c \exp(-\frac{1}{2} d_{ci}) |\Sigma_c|^{-\frac{1}{2}}}{\sum_{l=1}^{C} \pi_l \exp(-\frac{1}{2} d_{li}) |\Sigma_l|^{-\frac{1}{2}}}, \end{aligned} \tag{4}$$

  where

$$d_{ci} = (\boldsymbol{x}_i - \boldsymbol{b}_c)^\top \Sigma_c^{-1} (\boldsymbol{x}_i - \boldsymbol{b}_c). \tag{5}$$

- M-step Estimation of the parameters of GMMs:

$$\pi_c = \frac{1}{n} \sum_{i=1}^{n} u_{ci}, \tag{6}$$

$$\boldsymbol{b}_c = \frac{\sum_{i=1}^{n} u_{ci} \boldsymbol{x}_i}{\sum_{i=1}^{n} u_{ci}}, \tag{7}$$

$$\Sigma_c = \frac{\sum_{i=1}^{n} u_{ci} (\boldsymbol{x}_i - \boldsymbol{b}_c)(\boldsymbol{x}_i - \boldsymbol{b}_c)^\top}{\sum_{i=1}^{n} u_{ci}}. \tag{8}$$

2.2 FCM-type Fuzzy Clustering Approaches

FCM-type fuzzy clustering is closely related to mixtured density models in its algorithmic framework. In the standard FCM algorithm [8], the clustering criterion is the distance between the data point and the prototype of cluster and the objective function to be minimized is defined as

$$L_{fcm1} = \sum_{c=1}^{C} \sum_{i=1}^{n} (u_{ci})^\theta ||\boldsymbol{x}_i - \boldsymbol{b}_c||^2, \tag{9}$$

where u_{ci} represents the membership value of i-th data sample for c-th cluster and the weighting exponent ($\theta > 1$) is added for the fuzzification of memberships. The larger θ is, the fuzzier the memberships are. For deriving a clustering partition, an iterative algorithm is used. From the necessary conditions for the optimality, the new prototypes are derived as the weighted centers of clusters and the memberships are calculated by using the functions of the clustering criteria. Usually, the sum of u_{ci} with respect to c is constrained to be 1. The constraint is called "the probabilistic constraint" [15] and the memberships are obtained by a similar formula for updating posterior probabilities in the EM algorithm with GMMs.

Another alternative for the fuzzification of the memberships is the regularization with entropy [11]. Using the regularization technique, the objective function is defined as

$$\begin{aligned} L_{fcm2} &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}||\boldsymbol{x}_i - \boldsymbol{b}_c||^2 \\ &\quad +\lambda\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci}, \end{aligned} \tag{10}$$

where the entropy term works like the weighting exponent in the standard FCM algorithm. The larger λ is, the fuzzier the memberships are. The updating rules for memberships and cluster centers are derived as follows:

$$u_{ci} = \frac{\exp(-\frac{1}{\lambda}||\boldsymbol{x}_i - \boldsymbol{b}_c||^2)}{\sum_{l=1}^{c}\exp(-\frac{1}{\lambda}||\boldsymbol{x}_i - \boldsymbol{b}_l||^2)}, \tag{11}$$

$$\boldsymbol{b}_c = \frac{\sum_{i=1}^{n} u_{ci}\boldsymbol{x}_i}{\sum_{i=1}^{n} u_{ci}}. \tag{12}$$

These rules are equivalent to the two steps of the EM algorithm with GMMs under the condition that the unknown parameters of Gaussian components are only the mean vectors and $\pi_c = 1/C$ (const.). In the model, the fuzzifier λ corresponds to the variance of the Gaussian density function and the other elements of the covariance matrix are 0.

The FCM clustering with K-L information term (KFCM) [12] is the fuzzy counterpart of the GMMs with full unknown parameters. By replacing the entropy term in Eq.(10) with K-L information, the objective function of the KFCM clustering is defined as follows:

$$\begin{aligned} L_{kfcm} &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}d_{ci} + \lambda\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log\frac{u_{ci}}{\pi_c} \\ &\quad +\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log|\Sigma_c| \end{aligned} \tag{13}$$

where d_{ci} is the mahalanobis distance

$$d_{ci} = (\boldsymbol{x}_i - \boldsymbol{b}_c)^\top\Sigma_c^{-1}(\boldsymbol{x}_i - \boldsymbol{b}_c), \tag{14}$$

and the matrices Σ_c's are also decision variables. Eq.(13) is minimized under the condition that both the sum of u_{ci} and the sum of π_c with respect to c equal 1 respectively. As the entropy term in Eq.(10) forces memberships u_{ci} to take similar values, i.e., to obtain fuzzy clusters, the K-L information term of Eq.(13) becomes 0 if u_{ci}, $i = 1, \cdots, n$ take the same value π_c within the c-th cluster for all c. If $u_{ci} \simeq \pi_c$ for all i and c, partition becomes very fuzzy, but when λ is 0 the optimization problem reduces to a linear one, and the solution u_{ci} are obtained at an extremal point (i.e., 0 or 1). Fuzziness of the clusters can be controlled by λ. From the necessary conditions, the updating rules in the fix-point iteration algorithm are given as follows.

$$u_{ci} = \frac{\pi_c\exp\big(-\frac{1}{\lambda}d_{ci}\big)|\Sigma_c|^{-\frac{1}{\lambda}}}{\sum_{l=1}^{C}\pi_l\exp\big(-\frac{1}{\lambda}d_{li}\big)|\Sigma_l|^{-\frac{1}{\lambda}}}, \tag{15}$$

$$\boldsymbol{b}_c = \frac{\sum_{i=1}^{n} u_{ci}\boldsymbol{x}_i}{\sum_{i=1}^{n} u_{ci}}, \tag{16}$$

$$\pi_c = \frac{1}{n}\sum_{i=1}^{n} u_{ci}, \tag{17}$$

$$\Sigma_c = \frac{\sum_{i=1}^{n} u_{ci}(\boldsymbol{x}_i - \boldsymbol{b}_c)(\boldsymbol{x}_i - \boldsymbol{b}_c)^\top}{\sum_{i=1}^{n} u_{ci}}. \tag{18}$$

The algorithm is equivalent to the EM algorithm with GMMs only if the fuzzification coefficient $\lambda = 2$. When $\lambda \neq 2$, there is no corresponding mixture density.

3 Local Principal Component Analysis and Linear Fuzzy Clustering

3.1 Local PCA and Probabilistic Models

Despite the flexibility of the GMMs, they are not always useful for local subspace learning. For the non-linear dimension reductions, Local PCA techniques are used. Kambhatla *et al.* [2] proposed an iterative algorithm composed of the (hard) clustering of data sets and the estimation of local principal components in each cluster. Hinton *et al.* [3] extended the idea to "soft version". In the "soft version", the responsibility of each data point for its generation is shared amongst all of the principal component analyzers instead of being assigned to only one analyzer. The objective function to be minimized is a negative pseudo-likelihood function

$$L_{lpca} = \frac{1}{\lambda}\sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}E_{ci} + \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci}\log u_{ci}, \tag{19}$$

where E_{ci} is the squared reconstruction error, i.e., the distance between i-th data point and c-th principal subspace. The responsibility of c-th analyzer for reconstructing data point $\boldsymbol{x}_i$ is given by

$$u_{ci} = \frac{\exp(-\frac{1}{\lambda}E_{ci})}{\sum_{l=1}^{c}\exp(-\frac{1}{\lambda}Eli)}. \tag{20}$$

Although the optimal local models are derived by minimizing a single negative likelihood function, no probability density is defined.

Recently, probabilistic models for PCA have been proposed [5, 6]. In the latent variable models, the prior distributions of the latent variables are given as Gaussian distributions and the single model is easily extended to a mixture of Local PCA models in which all of the model parameters are estimated through the maximization of a single likelihood function. Mixture of probabilistic PCA (MPCA) [6] defines the linear latent models where an m dimensional observation vector $\boldsymbol{x}$ is related to a p dimensional latent vector $\boldsymbol{f}_c$ in each probabilistic model,

$$\boldsymbol{x} = A_c\boldsymbol{f}_c + \boldsymbol{b}_c + \boldsymbol{\epsilon}_c, \quad ; c = 1, \cdots, C. \tag{21}$$

The $(m \times p)$ matrix A_c is the principal component matrix composed of p local principal component vectors and the vector $\boldsymbol{b}_c$ is the mean vector of c-th probabilistic model. The density distribution of the latent variables are assumed to be a simple Gaussian, $\boldsymbol{f}_c \sim \mathcal{N}(\boldsymbol{0}, I)$. When the error model $\boldsymbol{\epsilon}_c \sim \mathcal{N}(\boldsymbol{0}, R_c)$ is restricted to $R = \sigma_c^2 I$, the probabilistic model is associated to PCA and the conventional PCA is recovered with $\sigma_c \to 0$.

Using the isotropic Gaussian noise model $\boldsymbol{\epsilon}_c \sim \mathcal{N}(\boldsymbol{0}, \sigma_c^2 I)$, the $\boldsymbol{f}_c$ conditional probability distribution over $\boldsymbol{x}$ space is given by

$$p_c(\boldsymbol{x}|\boldsymbol{f}_c) \sim \mathcal{N}(A_c\boldsymbol{f}_c + \boldsymbol{b}_c, \sigma_c^2 I). \tag{22}$$

The marginal distribution for the observation $\boldsymbol{x}$ is also Gaussian:

$$p_c(\boldsymbol{x}) \sim \mathcal{N}(\boldsymbol{b}_c, W_c), \tag{23}$$

where $W_c = A_c A_c^\top + \sigma_c^2 I$. The negative log-likelihood function to be minimized is defined as

$$L_{mpca} = -\sum_{i=1}^{n} \log\left\{\sum_{c=1}^{C} \pi_c p_c(\boldsymbol{x}_i)\right\}. \tag{24}$$

The parameters of these linear models are estimated by EM algorithm.

- E-step Estimation of responsibility of each data point for component densities:

$$u_{ci} = \frac{\pi_c \exp(-\frac{1}{2}E_{ci})|W_c|^{-\frac{1}{2}}}{\sum_{l=1}^{C} \pi_l \exp(-\frac{1}{2}E_{li})|W_l|^{-\frac{1}{2}}}, \tag{25}$$

where

$$E_{ci} = (\boldsymbol{x}_i - \boldsymbol{b}_c)^\top W_c^{-1}(\boldsymbol{x}_i - \boldsymbol{b}_c). \tag{26}$$

- M-step Estimation of the parameters of the latent models:

$$\pi_c = \frac{1}{n}\sum_{i=1}^{n} u_{ci}, \tag{27}$$

$$\boldsymbol{b}_c = \frac{\sum_{i=1}^{n} u_{ci}\boldsymbol{x}_i}{\sum_{i=1}^{n} u_{ci}}, \tag{28}$$

$$A_c = U_{pc}(\Delta_{pc} - \sigma_c^2 I)^{1/2} V, \tag{29}$$

$$\sigma_c^2 = \frac{1}{m-p}\sum_{j=p+1}^{m} \delta_{cj} \tag{30}$$

where U_{pc} is $(m \times p)$ matrix composed of eigenvectors corresponding to the largest eigenvalues of S_c, Δ_{pc} is the $(p \times p)$ diagonal matrix of the largest eigenvalues, V is an arbitrary $(p \times p)$ orthogonal matrix. $\delta_{c,p+1}, \cdots, \delta_{cm}$ are the smallest $m - p$ eigenvalues of S_c, where S_c is a local responsibility-weighted covariance matrix

$$S_c = \frac{1}{\pi_c n}\sum_{i=1}^{n} u_{ci}(\boldsymbol{x}_i - \boldsymbol{b}_c)(\boldsymbol{x}_i - \boldsymbol{b}_c)^\top. \tag{31}$$

Here, the model can be seen as a method for capturing the covariance structure of the m dimensional observation using $A_c A_c^\top + \sigma^2 I$ that has only $(m \times p + 1)$ free parameters while the full covariance matrix used in GMMs has m^2 parameters and the complexity can be tuned by the dimension of latent space. Moerland [7] performed a comparison of mixture models for density estimation and reported that the mixtures of latent variable models outperformed GMMs in terms of generalization though we have to choose the extra parameter (the dimensions of latent space).

3.2 Linear Fuzzy Clustering with Entropy Regularization

FCM-type fuzzy clustering algorithm can be also extended to local subspace learning. Fuzzy c-Varieties (FCV) [13] is a linear fuzzy clustering technique that captures the local linear structures by using linear varieties as the prototypes of clusters. In the FCV algorithm, the clustering criterion d_{ci} is replaced with the distance between i-th data point and c-th linear variety as follows:

$$E_{ci} = ||\boldsymbol{x}_i - \boldsymbol{b}_c||^2 - \sum_{k=1}^{p} |\boldsymbol{a}_{ck}^\top(\boldsymbol{x}_i - \boldsymbol{b}_c)|^2, \tag{32}$$

where p dimensional linear variety spanned by normal vectors $\boldsymbol{a}_{ck}$'s is the prototype of c-th cluster. Because the optimal $\boldsymbol{a}_{ck}$'s are derived by solving the eigenvalue problems of fuzzy scatter matrices, they are regarded as the local principal component vectors. Note that the objective function with entropy regularization is equivalent to the negative log-likelihood of Hinton's "soft version" Local PCA.

In this way, linear fuzzy clustering algorithms are closely related to Local PCA techniques.

3.3 Fuzzy c-Varieties with Regularization by K-L Information

In this subsection, we propose a new linear fuzzy clustering method that corresponds to a fuzzy version of MPCA. In the following, we capture the local linear structures with a constrained model of the KFCM clustering. Replacing the full rank matrix A_c with the constrained matrix $W_c = A_c A_c^\top + \sigma_c^2 I$, the objective function of the KFCM algorithm is given as

$$\begin{aligned} L_{kfcv} &= \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci} E_{ci} + \lambda \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci} \log \frac{u_{ci}}{\pi_c} \\ &\quad + \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci} \log |A_c A_c^\top + \sigma_c^2 I| \end{aligned} \tag{33}$$

where E_{ci} is the generalized mahalanobis distance

$$E_{ci} = (\boldsymbol{x}_i - \boldsymbol{b}_c)^\top (A_c A_c^\top + \sigma_c^2 I)^{-1} (\boldsymbol{x}_i - \boldsymbol{b}_c). \tag{34}$$

From the necessary condition for the optimality, new memberships are derived as follows:

$$u_{ci} = \frac{\pi_c \exp\left(-\frac{1}{\lambda} E_{ci}\right) |A_c A_c^\top + \sigma_c^2 I|^{-\frac{1}{\lambda}}}{\sum_{l=1}^{C} \pi_l \exp\left(-\frac{1}{\lambda} E_{li}\right) |A_l A_l^\top + \sigma_l^2 I|^{-\frac{1}{\lambda}}}. \tag{35}$$

In the same way, $\boldsymbol{b}_c$'s and π_c's are updated by using Eq.(16) and (17). To calculate new A_c's and σ_c's, the objective function is rewritten as

$$\begin{aligned} L_{kfcv'} &= \sum_{c=1}^{C} \left\{ \sum_{i=1}^{n} u_{ci} \right\} \mathrm{tr}(W_c^{-1} S_c) \\ &\quad + \lambda \sum_{c=1}^{C}\sum_{i=1}^{n} u_{ci} \log \frac{u_{ci}}{\pi_c} \\ &\quad + \sum_{c=1}^{C} \left\{ \sum_{i=1}^{n} u_{ci} \right\} \log |W_c|, \end{aligned} \tag{36}$$

where S_c is the fuzzy covariance matrix in c-th cluster that is calculated by the same equation as Eq.(31). From the necessary condition $\partial L_{kfcv'} / \partial A_c = O$,

$$-W_c^{-1} S_c W_c^{-1} A_c + W_c^{-1} A_c = O. \tag{37}$$

Then, the local principal component matrix A_c is derived as

$$A_c = U_{pc} (\Delta_{pc} - \sigma_c^2 I)^{1/2} V. \tag{38}$$

This is the same equation as Eq.(29) and the optimal A_c's are given by eigenvectors corresponding to largest eigenvalues.

While this constrained KFCM algorithm is equivalent to the MPCA algorithm in the case of $\lambda = 2$, there is no corresponding probabilistic model when $\lambda \neq 2$. Then the proposed method isn't a probabilistic approach but a sort of fuzzy modeling techniques where the parameter λ determines the degree of fuzziness. If the proportions and the error model parameters are restricted to $\pi_c = 1/C$ and $\sigma_c \to 0$, the model derives the FCV clustering with entropy regularization. In this sense, this constrained model is regarded as the modified FCV algorithm with regularization by K-L information.

The proposed algorithm can be written as follows.

Fuzzy c-Varieties with Regularization by K-L Information (KFCV) Algorithm

Step 1 Initialize u_{ci}'s randomly and normalize them so that they satisfy "the probabilistic constraint".

Step 2 Calculate π_c's using Eq.(17).

Step 3 Calculate $\boldsymbol{b}_c$'s using Eq.(16).

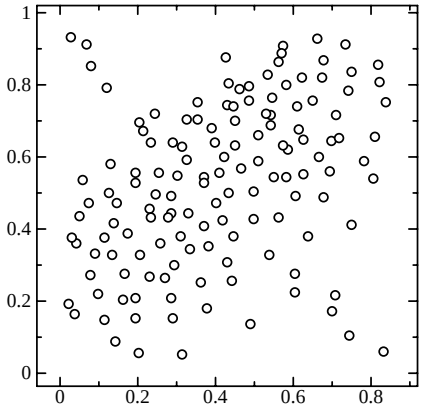


Figure 1: Artificial Data Set

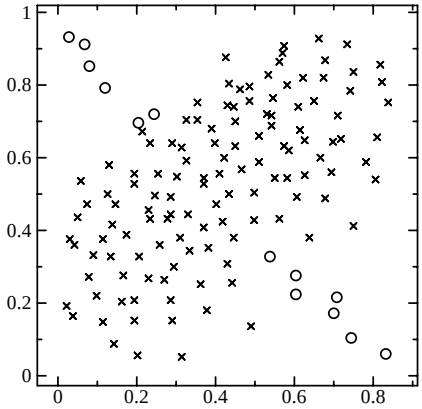


Figure 2: Result of MPCA

Step 4 Calculate A_c's and σ_c's by solving the eigenvalue problems of fuzzy covariance matrices.

Step 5 Update u_{ci}'s using Eq.(35).

Step 6 If

$$\max_{i,c} \mid u_{ci}^{NEW} - u_{ci}^{OLD} \mid < \epsilon,$$

then stop. Otherwise, return to Step 2.

4 Experimental Results

4.1 Comparison between Probabilistic Mixture Models and Fuzzy Clustering

First, we discuss the difference between probabilistic mixture models and fuzzy clustering. The artificial data set shown in Fig. 1 is composed of the samples from two different generative models. One is the linear model whose error variance is small and the data points are distributed forming a thin line. The other set is generated with larger error model and the data points form a rectangle. The goal of the analysis is to reveal these two linear latent models.

Fig. 2 shows the result of MPCA that is equivalent to the KFCV model with $\lambda = 2$. The data set was classified into two clusters represented by $\circ$ and $\times$ in the sense of maximum membership (posterior probability) and the a priori probabilities were $\pi_1 = 0.1$ and $\pi_2 = 0.9$ respectively. Because $\pi_1 << \pi_2$, the probabilistic model regarded 2nd cluster ($\times$) as the meaningful cluster and classified the overlapped region into 2nd cluster.

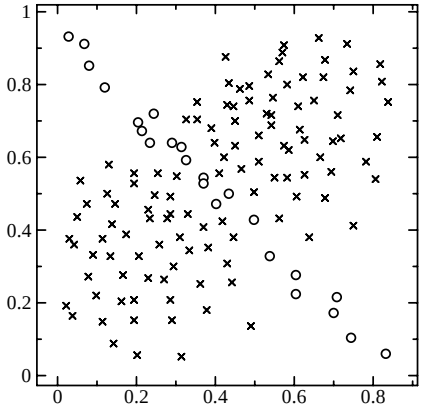
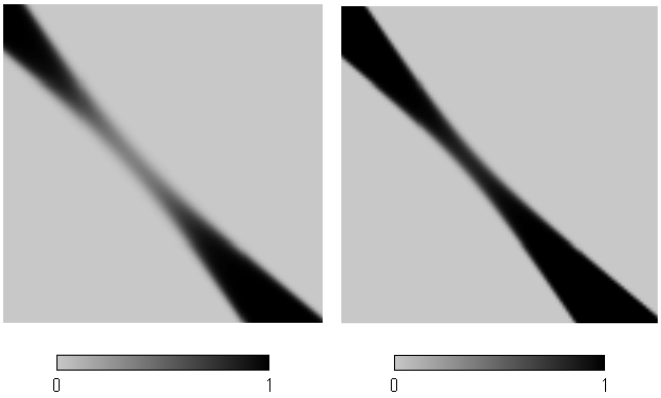


Figure 3: Result of KFCV with $\lambda = 1$



(a) MPCA (b) KFCV with $\lambda = 1$

Figure 4: Comparison of Classification Functions

Next, we performed the KFCV algorithm with $\lambda = 1$. Fig. 3 shows the clustering result where $\pi_1 = 0.15$ and $\pi_2 = 0.85$ respectively. Fig. 4 shows the comparison of the classification function (Eq.(25) and Eq.(35)) in the data space. Because $\lambda < 2$, the derived partitioning was not so fuzzy as that of MPCA and tended toward the crisp one emphasizing the smaller cluster. Although the derived partitioning did not correspond to the maximum likelihood model, the classification function suited for the intuitive partitioning. In this way, fuzzy clustering can derive more flexible partitioning by tuning parameter λ.

4.2 Comparison between KFCM and KFCV

MPCA is a constrained model of GMMs, in which the covariance matrices are approximated by using fewer parameters, and outperforms GMMs in terms of generalization [7]. In this subsection, we compare the proposed KFCV algorithm with the KFCM algorithm using Ionosphere database [16] composed of 351 instances with 34 numeric attributes. The data set can be classified with over 90% accuracy by a linear perceptron [17], i.e., the data set forms two principal masses in the multivariate data space. The goal of this experiment is to capture the characteristics of the two masses without a priori class information. The 5-fold cross-validation was used for testing the validity of the derived local models. Table 1 shows the averages of the objective function derived by the KFCV algorithm varying the number of the latent variables. The KFCM algorithm was performed using full covariance matrices and the result corresponded to that of the KFCV algorithm with $p = m - 1$.

In this experiment, the KFCM algorithm was so complicated that the local models overfitted to the training set because the models could not represent the test set. On the other hand, the KFCV algorithm with a restricted number of parameters had better generalization ability. In this way, the proposed method is an attractive alternative of the KFCM algorithm when we deal with a high-dimensional data set with a small number of samples.

Table 1: Averages of Objective Function

algorithm	train	test	difference
KFCV($p = 2$)	-40.4	-39.0	1.4
KFCV($p = 5$)	-69.3	-67.3	2.0
KFCV($p = 9$)	-73.2	-70.6	2.6
KFCM	-61.9	-45.3	16.6

5 Conclusion

This paper discussed the relationship between Local PCA techniques and linear fuzzy clustering algorithms and proposed a modified linear clustering algorithm that can capture the local substructures flexibly. Considering the membership value of each data point as the responsibility (posterior probability) for component densities, some clustering algorithms are associated with mixture density models. However, the objective function methods are generally more flexible than the maximum likelihood approaches and it is easy to introduce additional objectives or constraints. The introduction of the mechanism such as the annealing of fuzzifier [12] or the noise clustering approach [18] is remained in future works.

References

[1] K. Fukunaga, and D. R. Olsen, "An algorithm for finding intrinsic dimensionality of data," *IEEE Trans. on Computers*, vol. C-20, pp. 176-183, 1971.

[2] N. Kambhatla, and T. K. Leen, "Dimension reduction by local principal component analysis," *Neural Computation*, vol. 9, no. 7, pp. 1493-1516, 1997.

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

[4] A. P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. of the Royal Statistical Society*, Series B, vol. 39, pp. 1–38, 1977.

[5] S. Roweis, "EM algorithms for PCA and SPCA," *Advances in Neural Information Processing Systems 10*, Eds. M. I. Jordan, M. J. Kearns and S. A. Solla, MIT Press, pp. 626–632, 1998.

[6] M. E. Tipping, and C. M. Bishop, "Mixtures of probabilistic principal component analysers," *Neural Computation*, vol. 11, no. 2, pp. 443-482, 1999.

[7] P. Moerland, "A comparison of mixture models for density estimation," *Proc. of 9th Int. Conf. Artificial Neural Networks (ICANN'99)*, vol. 1, pp 25-30, 1999.

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

[9] D. E. Gustafson, and W. C. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," *Proc. of the IEEE Conf. Decision and Control*, vol. 2, pp. 761-766, 1979.

[10] I. Gath, and A. B. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp 773-781, 1989.

[11] 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*, vol. 2, pp 86-92, 1997.

[12] 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*, 2001.

[13] J. C. Bezdek, C. Coray, R. Gunderson, and J. Watson, "Detection and characterization of cluster substructure 2. fuzzy c-varieties and convex combinations thereof," *SIAM J. Appl. Math.*, vol. 40, no. 2, pp 358-372, 1981.

[14] C. M. Bishop, *Neural Networks for Pattern Recognition*, Clarendon Press, 1995.

[15] F. Höppner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis*, Jhon Wiley & Sons, 1999.

[16] P. M. Murphy, and D. W. Aha, *UCI Repository of machine learning databases* [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science, 1994.

[17] V. G. Sigillito, S. P. Wing, L. V. Hutton, and K. B. Baker, "Classification of radar returns from the ionosphere using neural networks," Johns Hopkins APL Technical Digest, vol. 10, pp. 262-266, 1989.

[18] H. Ichihashi, and K. Honda, "Robust clustering in fuzzy c-means with regularization by cross entropy," *Proc. of 1st International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 1, pp 471-475, 2002.